RAGCELL: RETRIEVAL-AUGMENTED GENERATION AS SUPERVISION FOR VERSATILE SINGLE-CELL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell foundation models (scFMs) are transforming computational biology by enabling generalizable, task-agnostic representations for versatile single-cell analysis. Despite their progress in facilitating rapid deployment for downstream tasks, off-the-shelf scFMs still have some overlooked concerns: (I) (*Pretraining Cost.*) Pretrain-based scFMs necessitate pretraining on a vast volume of cells, rendering it draining resources in applications. (II) (*Heterogeneous Gap.*) Large Language Models (LLM)-based scFMs ignore the tremendous heterogeneous gap between LLM textual and raw cellular spaces, leading to insufficient capability when facing downstream tasks. To this end, we introduce `RAGCell`, a versatile single-cell analysis framework that achieves a double-win in both **cost-effectiveness** and **high performance**. The success of `RAGCell` lies in two key aspects: ❶ Leveraging LLMs to construct cell-level and feature-level knowledge databases, which serve as supervision signals for training the cell model and significantly reduce the training cost (>pretrain-based scFMs). ❷ Aligning cell representations with text embeddings from the bi-level knowledge databases, enabling knowledge transfer from textual spaces to cellular spaces and effectively mitigating the heterogeneous gap (>LLM-based scFMs). Through extensive experiments on six downstream single-cell analysis tasks, we demonstrate that `RAGCell` achieves outstanding performance compared to state-of-the-art scFMs while operating at less than $\sim$1/10 the cost of pretrain-based scFMs. The source code is available at https://anonymous.4open.science/r/RAGCell.

1 INTRODUCTION

Recent advances in machine learning and large language models (LLMs) (Devlin, 2018; Brown, 2020; Ouyang et al., 2022; Achiam et al., 2023; Touvron et al., 2023) have greatly facilitated single-cell analysis. Several single-cell foundation models (scFMs) have been proposed to obtain task-agnostic cell representations that generalize well to specific downstream single-cell analysis tasks. These scFMs can be broadly divided into two groups: pretrain-based scFMs and LLM-based scFMs. Pretrain-based scFMs typically leverage a vast volume of single-cell RNA sequence (scRNA-seq) data for large-scale pretraining, while LLM-based scFMs usually construct cell representations from LLM embeddings. Although these scFMs have made significant progress for versatile single-cell analysis, there are several limitations that cannot be neglected. *Firstly*, the success of pretrain-based scFMs depends on the pretraining process over a large volume of cells, which is time-consuming and resource-intensive in practice. To obtain scalable embeddings for single-cell data from different resources, pretrain-based scFMs often define gene vocabularies for data tokenization, which requires genomics knowledge from human experts. Despite their outstanding performance, the costs associated with them are challenging to bear. Obtaining expert priors is difficult and demands labor-intensive efforts, consequently leading to a decrease in training efficiency. *Secondly*, LLM-based scFMs often ignore heterogeneity between raw cellular and LLM textual spaces. The LLM textual space is constructed based on pretraining in natural language, whereas the cellular space is dedicated to modeling biological data, leading to inherent significant gaps and heterogeneity. Although LLM-based scFMs are cost-effective, their performance often falls short, especially in finetuning scenarios.

Inspired by the recent progress of LLM agents (Wang et al., 2024; Durante et al., 2024) and the associated techniques like retrieval-augmented generation (RAG) (Lewis et al., 2020; Zhao et al., 2024a), we present `RAGCell`, a versatile single-cell analysis framework that achieves a double-win

in both **cost-effectiveness** and **high performance**. Different from existing LLM-based scFMs that directly leverage LLM embeddings for single-cell data modeling, `RAGCell` firstly generates cell-level and feature-level knowledge databases through LLMs. For cell-level text descriptions, the omics and cell type information are included in a prompt template as queries, and corresponding responses from LLMs are stored in the database. For feature-level text descriptions, the specific functions of genes (Chen et al., 2019), peaks (Zeng et al., 2024), and proteins (Wu & Singh, 2012) are preserved in another database. Then, we employ a patch-based light-weighted Transformer (Vaswani, 2017) as the cell model to obtain cell representations. Afterward, the text descriptions from these databases are retrieved and projected into a new space to perform cell-text alignment (Chen et al., 2020a;b;c; 2021; Radford et al., 2021), which aims to provide supervision for the cell model and align cell representations from the cellular space with knowledge from the well-defined LLM textual space. After alignment, the cell model could be employed for various downstream single-cell analysis tasks under both finetuning and zero-shot settings. Compared with existing scFMs, `RAGCell` can model single-cell multi-omics data into a general framework and achieve superior performance on downstream tasks free of the need for extensive pretraining on extra single-cell data or any human expert priors. In comparison to pretrain-based scFMs such as scBERT (Yang et al., 2022) and scGPT (Cui et al., 2024), `RAGCell` can achieve more than a **tenfold** reduction in pretraining costs. In addition, since omics information is included in the cell-level database, `RAGCell` is scalable to both single-omics and multi-omics data analysis tasks, such as cell type annotation, batch effect correction, multi-omics data integration, rare cell type annotation, cell-text retrieval, and drug sensitivity prediction. The superiority of `RAGCell` against a range of cutting-edge scFMs is fully demonstrated through comprehensive experiments on many downstream single-cell analysis tasks and datasets. To summarize, the main contributions of this paper are threefold:

1. ***Novel Perspective.*** We identify the key limitations of pretrain-based and LLM-based scFMs, and then provide a novel perspective to reduce the training cost of single-cell analysis framework while maintaining its capability by incorporating LLM priors as supervision signals.
2. ***Double-Win Framework.*** We present `RAGCell`, a versatile framework that is based on RAG and empowered by LLMs. Leveraging LLM-generated information as supervision, `RAGCell` achieves a double-win in high performance and cost-effectiveness for single-cell analysis.
3. ***Multifaceted Validation.*** We conduct comprehensive experiments on six downstream single-cell analysis tasks and demonstrate the superiority of `RAGCell` against many state-of-the-art scFMs.

2 Related Work

2.1 Single-cell Analysis

Single-cell analysis seeks to characterize cellular states across various omics modalities, such as RNA (Saliba et al., 2014; Kolodziejczyk et al., 2015), DNA (Karemaker & Vermeulen, 2018; Evrony et al., 2021), and proteins (Wu & Singh, 2012; Suman et al., 2015). As a foundational technique in computational biology, it has wide-ranging applications in healthcare (Hong et al., 2019; Rajewsky et al., 2020) and medicine (Lim et al., 2020; Paik et al., 2020). For example, single-cell analysis enables the identification of cellular heterogeneity, offering insights into the complexity of tissues and organs. It also plays a key role in elucidating disease progression, thereby supporting the discovery of novel therapeutic targets and advancing our understanding of disease mechanisms. Numerous computational tasks have been developed for analyzing single-omics or multi-omics single-cell data, such as cell type annotation (Jiang et al., 2023; Hou & Ji, 2024), batch effect correction (Tran et al., 2020; Fei & Yu, 2020), and multi-omics data integration (Lance et al., 2022; Cao & Gao, 2022).

2.2 Foundation Models for Single-cell Biology

With the rapid advancement of generative artificial intelligence (Devlin, 2018; Brown, 2020; Zhao et al., 2024b) and LLMs (Touvron et al., 2023; Achiam et al., 2023), numerous FMs have been proposed for single-cell biology. For instance, scBERT (Yang et al., 2022) applies a bag-of-words strategy to discretize gene expression based on transcription frequencies and incorporates predefined Gene2Vec embeddings to tokenize single cells. It utilizes a Performer encoder (Choromanski et al., 2020) to learn cell representations and conducts self-supervised pretraining on over one million scRNA-seq samples. scGPT (Cui et al., 2024) tokenizes data by incorporating gene expression

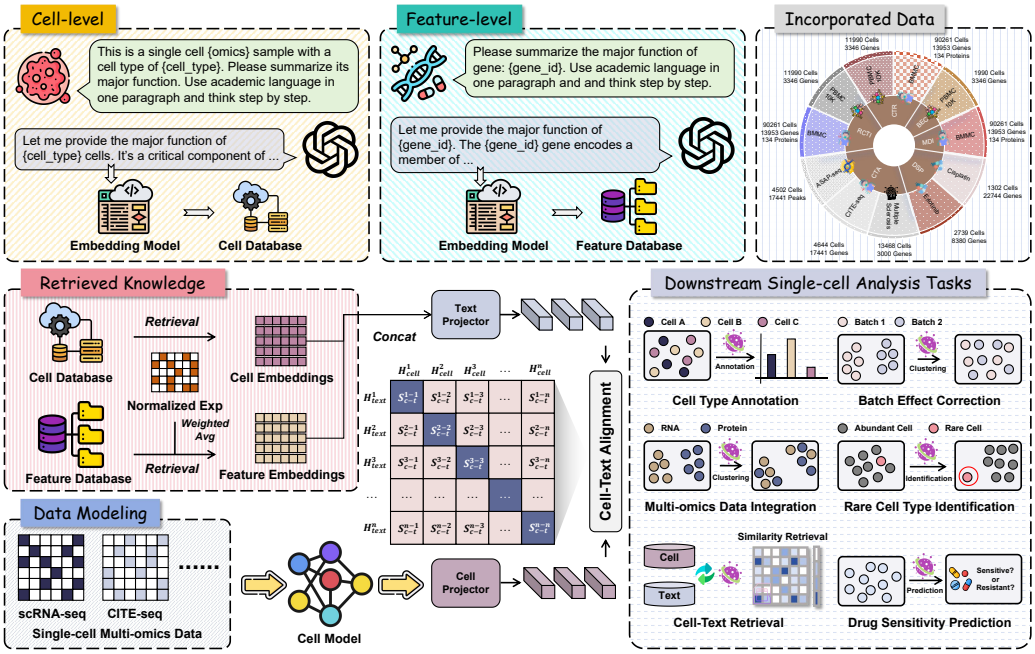

Figure 1: **An overview of RAGCell.** RAGCell utilizes multi-source single-cell data to construct both cell-level and feature-level knowledge databases. Subsequently, it retrieves bi-level knowledge as supervision signals for training the cell model. Following alignment of cell representations with the LLM textual space, the pretrained cell model can be applied to various single-cell analysis tasks.

values, gene tokens, and condition tokens. It employs a Transformer model trained autoregressively with an attention masking mechanism. In parallel, models such as GenePT (Chen & Zou, 2023) and scELMo (Liu et al., 2023b) leverage LLMs to generate gene-level embeddings, which are then integrated with raw single-cell sequences to enhance representation learning and task performance. LangCell (Zhao et al., 2024b) aims to enhance single-cell models by combining text descriptions and pre-training. In comparison with these methods, our RAGCell utilizes RAG techniques to achieve superior performance in versatile single-cell analysis with limited budgets.

3 The RAGCell Framework

The proposed framework RAGCell comprises a cell model and two associated knowledge databases. For each sample, we first create text descriptions with rich external information at both the cell and feature levels. These descriptions are then vectorized to obtain embeddings via embedding models and stored in databases for future use. Subsequently, we retrieve the generated text embeddings to perform cell-text alignment. In this stage, knowledge from LLM-based databases is utilized to guide the training process of a cell model, with the objective of injecting priors from the LLM textual space into the cellular space. Once the alignment is complete, the cell model can be employed to execute various downstream single-cell analysis tasks. An overview of our framework is provided in Figure 1.

Cell Model. Since there is no need for extensive pretraining, we employ a lightweight Transformer model (Dosovitskiy, 2020) as the cell model. For a cell sample $\mathbf{X} \in \mathcal{R}^{1\times N_f}$, we begin by splitting it into M patches, and each patch is a P dimensional vector where $N_f = M \times P$. Then we add a classification token $\mathbf{X}_{\text{cls}}$ to learn global representations and a 1D positional encoding $\mathbf{E}_{\text{pos}}$ to capture the relative position information:

$$\mathbf{X}_t = [\mathbf{X}_{\text{cls}}; \mathbf{X}_p^1\mathbf{E}; \mathbf{X}_p^2\mathbf{E}; ...; \mathbf{X}_p^M\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{1}$$

where $\mathbf{X}_p^i$ and $\mathbf{E}$ represent patch embedding and linear transformation, respectively. Afterward, we could employ a standard Transformer model (Vaswani, 2017) to capture the cell representations:

$$\mathbf{Z}_{\text{cell}} = \texttt{Transformer}(\mathbf{X}_t). \tag{2}$$

Then, $\mathbf{Z}_{\text{cell}}$ can be utilized to perform cell-text alignment and downstream tasks.

LLMs. LLMs are used to generate cell-level and feature-level text descriptions and to extract bi-level text embeddings, which introduce prior knowledge into the entire framework and thus eliminate the need for pretraining on large-scale data. In this paper, we opt for the GPT-4o mini model to generate cell-level text descriptions and the 'text-embedding-3-large' model to extract embeddings. For feature-level text descriptions, we use the GPT-3.5 model to obtain the specific function for each gene or protein and the 'text-embedding-ada-002' model to extract embeddings.

Bi-level Knowledge Databases Construction. The core of our framework is to construct cell-level and feature-level knowledge databases. For each cell sample, we consider multiple aspects of information to be crucial for constructing high-quality representations, including omics information, cell type information, and specific feature information (such as gene or protein). Therefore, it is necessary to include all these types of information in the text descriptions. First, we apply a text template: "This is a single-cell <omics> sequence sample with a cell type of <cell type>. Please summarize its major function. Use academic language in one paragraph and think step by step.". We then fill in the omics and cell type information for each cell into this template. Next, we utilize an embedding model (EM) to obtain text embeddings. The entire process can be formulated as:

$$\mathbf{Z}^{c}_{\text{text}} = \texttt{EM}(\texttt{LLM}(\texttt{Prompt}(Q^{c}))), \tag{3}$$

where Q^{c} denotes cell-level prompts and $\mathbf{Z}^{c}_{\text{text}} \in \mathcal{R}^{1\times F_1}$ represents the corresponding cell-level F_1 dimensional text embeddings. Next, we follow previous works GenePT (Chen & Zou, 2023) and scELMo (Liu et al., 2023b), to obtain feature-level text embeddings. There are generally two ways to acquire feature-level text descriptions: using text descriptions from the NCBI database as prompts or using human-designed prompts in a dialogue with an LLM. Here, we illustrate the second approach as an example. We begin by using a prompt following (Liu et al., 2023a; Jia et al., 2022; Ekin, 2023) to ask the LLM for specific feature information (using a gene as an example): "Please summarize the major function of gene: <gene>. Use academic language in one paragraph and think step by step.". Subsequently, we utilize the LLM's response to this query as the feature-level text descriptions and obtain its corresponding embeddings. This process can be formulated as:

$$\mathbf{Z}^{f_i}_{\text{text}} = \texttt{EM}(\texttt{LLM}(\texttt{Prompt}(Q^{f_i}))), \tag{4}$$

where Q^{f_i} denotes the query for each feature and $\mathbf{Z}^{f_i}_{\text{text}} \in \mathcal{R}^{1\times F_2}$ represents the corresponding F_2 dimensional text embeddings. For a cell sample with N_f features, we could stack $\mathbf{Z}^{f_i}_{\text{text}} \in \mathcal{R}^{1\times F_2}$ for each feature and obtain the final feature-level text embeddings $\mathbf{Z}^{f}_{\text{text}} \in \mathcal{R}^{N_f\times F_2}$. For each specific dataset, we can store these cell-level and feature-level embeddings in two databases. During the training process, the corresponding embeddings can be retrieved to provide supervision signals.

Cell-Text Alignment. After constructing bi-level knowledge databases, we then transfer knowledge from the databases to the cell model by retrieving relevant information. To begin, we need to concatenate the cell-level and feature-level text embeddings:

$$\mathbf{Z}^{f'}_{\text{text}} = \mathbf{X} \times \mathbf{Z}^{f}_{\text{text}}, \quad \mathbf{Z}_{\text{text}} = \texttt{Concat}(\mathbf{Z}^{c}_{\text{text}}, \mathbf{Z}^{f'}_{\text{text}}), \tag{5}$$

where $\mathbf{Z}^{f'}_{\text{text}} \in \mathcal{R}^{1\times F_2}$ and $\mathbf{Z}_{\text{text}} \in \mathcal{R}^{1\times (F_1+F_2)}$. Then, two separate multilayer perceptrons (MLPs) are utilized to project cell representations and text embeddings into new low-dimensional spaces:

$$\mathbf{H}_{\text{cell}} = \texttt{MLP}(\mathbf{Z}_{\text{cell}}), \quad \mathbf{H}_{\text{text}} = \texttt{MLP}(\mathbf{Z}_{\text{text}}), \tag{6}$$

where $\mathbf{H}_{\text{cell}} \in \mathcal{R}^{1\times D}$ and $\mathbf{H}_{\text{text}} \in \mathcal{R}^{1\times D}$ have same code length. Following prior works (Chen et al., 2020a; He et al., 2020; Radford et al., 2021; Xiong et al., 2023), we set $D = 128$ by default. After getting $\mathbf{H}_{\text{cell}}$ and $\mathbf{H}_{\text{text}}$ for each cell, we then perform cell-text alignment via instance-level matching:

$$L_{\text{C2T}} = -\frac{1}{B}\sum_{i}^{B} log\frac{cos(\mathbf{H}^{i}_{\text{cell}}, \mathbf{H}^{i}_{\text{text}})}{\sum_{j=1}^{B} cos(\mathbf{H}^{i}_{\text{cell}}, \mathbf{H}^{j}_{\text{text}})}, \tag{7}$$

$$L_{\text{T2C}} = -\frac{1}{B}\sum_{i}^{B} log\frac{cos(\mathbf{H}^{i}_{\text{text}}, \mathbf{H}^{i}_{\text{cell}})}{\sum_{j=1}^{B} cos(\mathbf{H}^{i}_{\text{text}}, \mathbf{H}^{j}_{\text{cell}})}, \tag{8}$$

where B denotes the number of samples within a mini-batch and $cos(\cdot)$ represents cosine similarity. The final loss objective can be formulated as:

$$L = \frac{1}{2}(L_{\text{C2T}} + L_{\text{T2C}}). \tag{9}$$

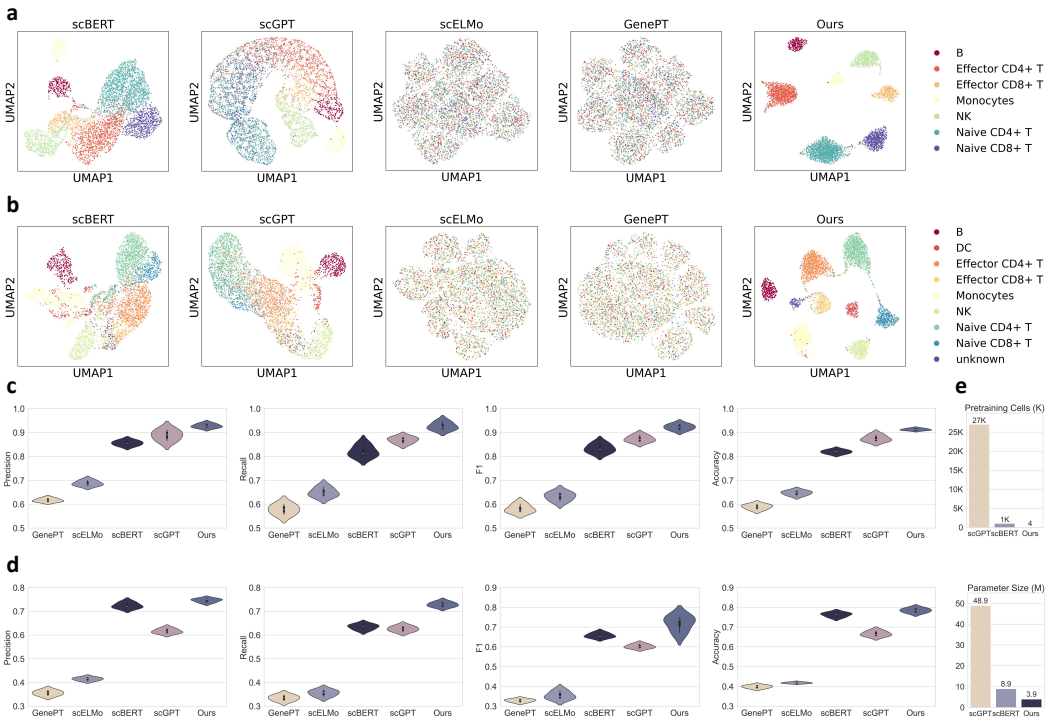

Figure 2: **Cell Type Annotation Results. a.** UMAP plot of the embeddings finetuned by five methods on CITE-seq data. **b.** UMAP plot of the embeddings finetuned by five methods on ASAP-seq data. **c.** Precision, Recall, F1, and Accuracy comparisons of different methods on CITE-seq data. **d.** Precision, Recall, F1, and Accuracy comparisons of different methods on ASAP-seq data. **e.** Cost comparisons.

The loss function ensures that the distance between cell-text pairs is minimized in the representation space, while the distance between non-paired cell and text representations is maximized. Through pretraining on specific datasets, we inject semantic knowledge from LLMs into the cell model. After training, the cell model can be employed for many downstream single-cell analysis tasks.

Downstream Single-cell Analysis Tasks. We evaluate the performance of `RAGCell` on several downstream single-cell data analysis tasks. For tasks like cell type annotation and drug sensitivity prediction, we first leverage a classification head to obtain the predictions for each cell:

$$\mathbf{P}_{\text{cell}} = \texttt{MLP}(\mathbf{H}_{\text{cell}}). \tag{10}$$

Then we finetune the cell model and classification head on specific datasets with cross-entropy loss:

$$L_{\text{cls}} = \texttt{CE}(\mathbf{P}_{\text{cell}}, Y_{\text{cell}}), \tag{11}$$

where Y_{cell} denotes the corresponding cell type labels or drug sensitivity labels. For batch effect correction, multi-omics data integration, and cell-text retrieval tasks, we directly employ the cell representations for evaluation, without the need for finetuning. For the rare cell type identification task, we employ the SOTA method scCAD (Xu et al., 2024) as a baseline. Then we replace the original cell embeddings with our cell representations. All other algorithm settings are kept consistent.

4 Experiment

4.1 Cell Type Annotation

Cell type annotation (Cao et al., 2020a; Chen et al., 2022b; 2023; Shao et al., 2021) is known as a crucial task in single-cell analysis. However, most existing scFMs focused exclusively on scRNA-seq data and lacked validation across other omics modalities. To address this gap, we evaluated `RAGCell`

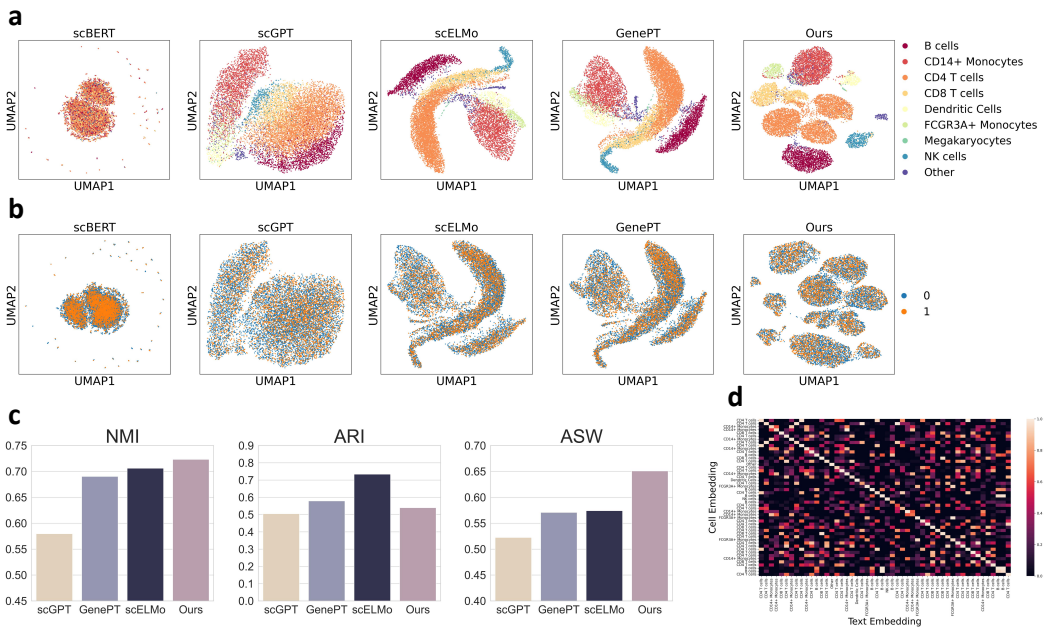

Figure 3: **Batch Effect Correction Results on PBMC 10K under Zero-shot Settings. a.** UMAP plot of the embeddings by five methods across cell types. **b.** UMAP plot of the embeddings by five methods across batches. **c.** Clustering performance comparisons among five methods. **d.** Heatmap visualizations of similarity between cell and text embeddings on randomly selected 50 cells.

against other scFMs using CITE-seq (Stoeckius et al., 2017) and ASAP-seq (Mimitou et al., 2021) data. For fair comparisons, we randomly partitioned each dataset into three subsets: 80% for training, 10% for validation, and 10% for testing. Fig. 2 (a) and (b) illustrate cell embeddings finetuned by the five methods. The UMAP visualizations (McInnes et al., 2018) demonstrated that `RAGCell` achieved superior separation of distinct cell types for both CITE-seq and ASAP-seq data, confirming its effectiveness for multi-omics cell type annotation. We subsequently conducted comprehensive quantitative comparisons using four evaluation metrics. As depicted in Fig. 2 (c) and (d), pretrain-based scFMs (scBERT (Yang et al., 2022), scGPT (Cui et al., 2024)) outperform LLM-based scFMs (GenePT (Chen & Zou, 2023), scELMo (Liu et al., 2023b)) by a large margin, consistent with expectations given the computationally intensive pretraining required for Transformer-based models. Remarkably, `RAGCell` consistently surpassed all other scFMs across every metric. It's worth noting that `RAGCell` achieved this performance without requiring supplemental pretraining data (Fig. 2 (e)). Its bi-level knowledge databases were constructed exclusively from training sets of the CITE-seq and ASAP-seq datasets, contrasting sharply with pretrain-based scFMs (e.g., scBERT, scGPT) that relied on massive pretraining corpora (millions to tens of millions of cells). These results indicate that `RAGCell` effectively synthesized strengths from both paradigms: leveraging external knowledge from LLM-based databases enabled state-of-the-art performance with manageable training costs.

4.2 Batch Effect Correction

Batch effect correction represents another fundamental challenge in single-cell analysis. To assess whether `RAGCell` effectively addresses batch effect issues (Tran et al., 2020; Goh et al., 2017; Fei & Yu, 2020), we conducted experiments on the PBMC 10K dataset (Gayoso et al., 2022) containing samples from two distinct batches. Generalization capability was evaluated under zero-shot settings, with `RAGCell` pretrained on around 65,000 cells from the NeurIPS 2021 Multimodal Single-Cell Data Integration competition (Luecken et al., 2021). This approach demonstrates significantly higher resource efficiency compared to scBERT and scGPT. The visualization results in Fig. 3 (a) reveal that `RAGCell` generates more discriminative cell-type clusters than alternative scFMs. scBERT exhibited the weakest performance among these scFMs, failing to form distinct clusters. This limitation stems from its masked value reconstruction pretraining strategy, which is no help for

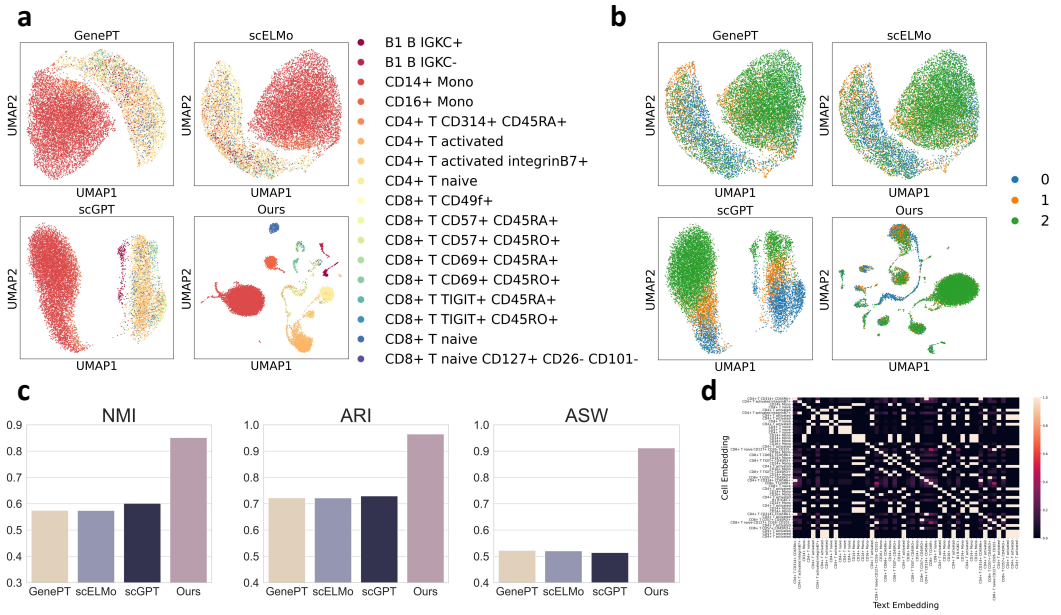

Figure 4: **Multi-omics Data Integration Results on BMMC data under Zero-shot Settings.** **a.** UMAP plot of the embeddings by different methods across cell types. **b.** UMAP plot of the embeddings by different methods across batches. **c.** Clustering performance comparisons. **d.** Heatmap visualizations of similarity between cell and text embeddings on randomly selected 50 cells.

dealing with batch effects or forming discriminative clusters. While GenePT, scELMo, and scGPT produced partial clusters, their cluster counts fell substantially below the true number of cell types. In contrast, `RAGCell` generated clusters corresponding closely to the actual cell type count with well-separated cluster centers, indicating superior batch integration capability. This enhancement stems from integrating prior knowledge retrieved from LLM-based databases, effectively suppressing batch-specific variations. The results in Fig. 3 (b) further confirm substantial overlap between batches within `RAGCell`'s latent space. Collectively, these results demonstrate `RAGCell`'s effectiveness in batch effect correction and clustering improvement for PBMC 10K data. Quantitative validation results in Fig. 3 (c) using three standardized metrics (NMI, ARI, ASW (Luecken et al., 2022)) indicate that `RAGCell` consistently outperforms baseline methods in batch effect correction. Despite a marginal ARI deficit, its dominant advantages in NMI and ASW substantiate its robust capabilities. In Fig. 3 (d), we can observe that cell-text alignment induces high similarity between embeddings of biologically similar cells, effectively disentangling biological signals from technical batch effects.

4.3 Multi-omics Data Integration

By leveraging an LLM to derive text embeddings capturing omic-specific information, the cell model within `RAGCell` acquires inherent multi-omic representational capabilities during the alignment process. Consequently, `RAGCell` excels at multi-omics data integration. To validate this capability, we utilized RNA and protein modalities from the BMMC dataset (Luecken et al., 2021) to evaluate zero-shot clustering performance, employing the same pretraining data as described in the previous section's batch effect correction experiment. Fig. 4 (a) visualizes cell embeddings from GenePT, scELMo, scGPT, and `RAGCell`. While all methods distinguished CD14+ monocytes, other scFMs proved unable to resolve CD8+ T cells. LLM-based scFMs (GenePT, scELMo) exhibited highly overlapping embeddings for remaining cell types, suggesting insufficient discriminative power despite underlying gene expression differences. scGPT showed marginal improvement by disambiguating CD4+ T cells, yet remained incapable of distinguishing other cell populations. In contrast, `RAGCell` successfully transferred semantic knowledge from the LLM to the cell model, yielding well-separated clusters for nearly all cell types and superior visualization quality. Fig. 4 (b) demonstrates `RAGCell`'s robust batch integration, evidenced by the highest degree of inter-batch embedding overlap. Conversely, GenePT, scELMo, and scGPT persistently exhibited batch-specific structures, indicating

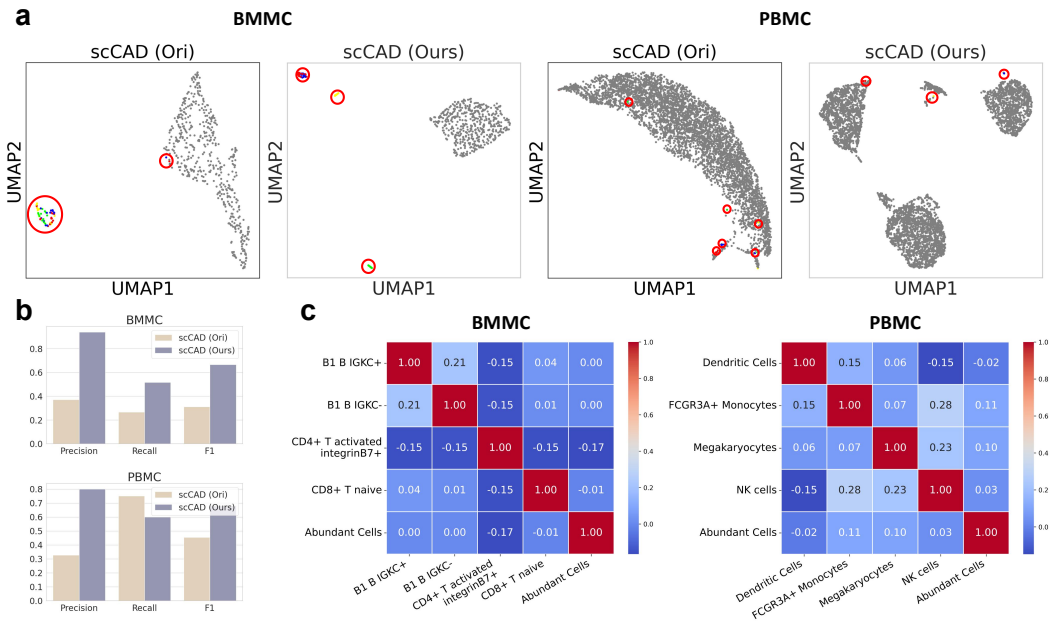

Figure 5: **Rare Cell Type Identification Results under Zero-shot Settings. a.** UMAP plot of the cell embeddings related to abundant and rare cell types. **b.** Quantitative performance comparison between the original cell embeddings and ours using the scCAD algorithm. **c.** Heatmap visualizations of similarity between average embeddings of rare cells and abundant cells.

inconsistent capture of biologically relevant features amidst technical noise. Quantitative validation in Fig. 4 (c) using NMI, ARI, and ASW metrics revealed `RAGCell`'s dominant performance, with all scores exceeding 0.8. The remaining methods consistently fell below this threshold (scores < 0.8), further confirming `RAGCell`'s superiority in multi-omics integration. Finally, Fig. 4 (d) illustrates elevated embedding similarity between biologically similar cells post-alignment. This correlation demonstrates that text embeddings provide effective semantic regularization, enabling `RAGCell` to seamlessly integrate both RNA and protein data while simultaneously mitigating batch effects.

4.4 Rare Cell Type Identification

Advances in sequencing technologies have yielded vast quantities of scRNA-seq data (Hwang et al., 2018; Cao et al., 2020b), which frequently contain both abundant and rare cell populations. Although rare cell types (Travaglini et al., 2020; Wu et al., 2019; Kiselev et al., 2019) exhibit low abundance, they play pivotal roles in biological processes such as disease pathogenesis and drug discovery. Consequently, rare cell identification has emerged as a critical challenge in single-cell analysis (Jiang et al., 2016; Jindal et al., 2018; Dong & Yuan, 2020). By leveraging prior knowledge from LLMs to derive high-quality cell representations, `RAGCell` provides an effective framework for rare cell identification. We adopted the state-of-the-art scCAD algorithm (Xu et al., 2024) as our baseline, which initially clusters cells via principal component analysis (PCA). To evaluate `RAGCell`'s capability, we replaced scCAD's original embeddings with `RAGCell`'s zero-shot representations. Fig. 5 (a) (left) illustrates BMMC data visualizations. Whereas baseline results show abundant cells clustered distantly from most rare cells (with partial overlap), `RAGCell` (right) achieves clear separation between abundant and rare populations while enhancing discrimination among rare cell subtypes. Similarly, for PBMC data (Fig. 5 (a) (right)), baseline embeddings exhibit substantial mixing of rare and abundant cells, whereas `RAGCell` maintains distinct inter-group distances and intra-group dispersion—significantly improving scCAD's rare cell identification capacity. In Fig. 5 (b), we selected three metrics to quantitatively compare the effectiveness of `RAGCell` in enhancing the identification of rare cell types. The results show that, compared to the original cell embeddings, the embeddings obtained from our framework achieve improvements across all metrics on the BMMC dataset. On the PBMC dataset, while our recall score is slightly lower than that of the original

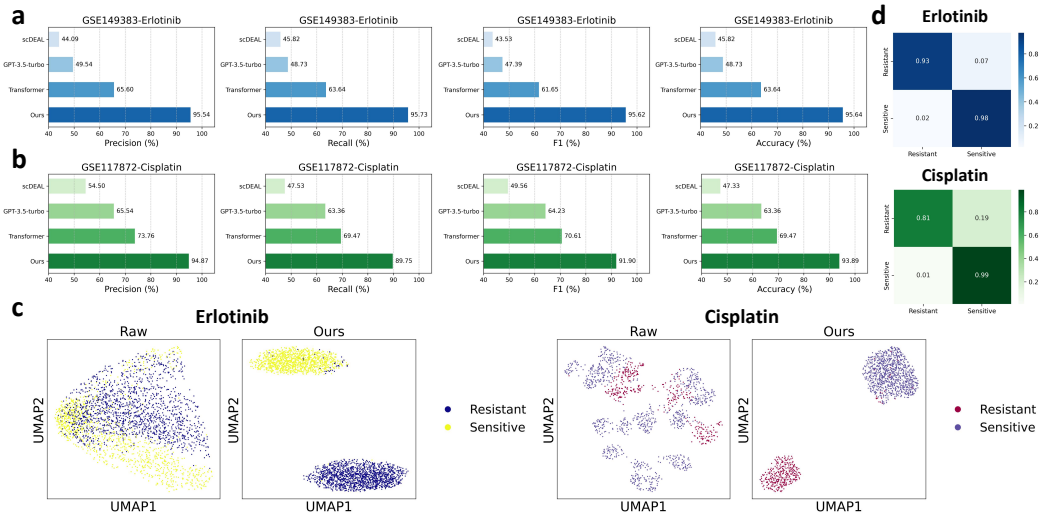

Figure 6: **Drug Sensitivity Prediction Results on Erlotinib and Cisplatin data. a.** Quantitative performance comparisons on Erlotinib data. **b.** Quantitative performance comparisons on Cisplatin data. **c.** UMAP plot of the raw embeddings and ours. **d.** Confusion matrix results.

embeddings, our performance on the other two metrics is superior. Quantitative analysis in Fig. 5 (b) using three metrics demonstrates `RAGCell`'s consistent improvement across all measures on BMMC data. For PBMC data, while recall shows marginal reduction, `RAGCell` outperforms baselines in two other metrics. Further validation in Fig. 5 (c) reveals substantially reduced cosine similarity between abundant/rare cells and among distinct rare cell types, confirming `RAGCell`'s discriminative power.

4.5 Drug Sensitivity Prediction

Accurate prediction of cellular drug sensitivity represents a critical challenge in precision biomedicine (Cortes-Ciriano et al., 2016; Ahmed et al., 2020). To evaluate `RAGCell`'s capabilities, we conducted experiments on the GSE149383-Erlotinib (Aissa et al., 2021) and GSE117872-Cisplatin (Sharma et al., 2018; Ravasio et al., 2020; Suphavilai et al., 2021) datasets. Fig. 6 (a) and (b) demonstrates `RAGCell`'s superior performance over both specialist models (Transformer (Vaswani, 2017), scDEAL (Chen et al., 2022a)) and generalist foundation models (GPT-3.5-turbo). Across all four metrics, comparator methods scored below 0.8, while `RAGCell` consistently exceeded 0.9, demonstrating significant improvements. This enhancement stems from our framework's ability to leverage inherent biological patterns within knowledge databases, strengthening cell-drug sensitivity associations. The visualization results in Fig. 6 (c) further substantiate this advantage. As it can be found, the raw cell embeddings often tend to mix the drug sensitivity property to a specific drug. However, the cell embeddings generated by our `RAGCell` illustrate the discernibility of cell sensitivity to various drugs. Cells with similar sensitivities typically aggregate together, whereas cells with distinct sensitivities are dispersed throughout. The confusion matrix results in Fig. 6 (d) reveal near-perfect sensitivity prediction accuracy, with only minor degradation in resistance classification. Overall, `RAGCell` significantly surpasses other specialist and generalist models in drug sensitivity prediction.

5 Conclusion

In this paper, we identify the shortcomings of both pretrain-based and LLM-based scFMs, and propose a double-win framework termed `RAGCell` that excels in both cost-effectiveness and high-performance for versatile single-cell analysis. The core of `RAGCell` is generating bi-level biological knowledge, which is retrieved as additional supervision signals during the pretraining phase. Remarkably, `RAGCell` outperforms cutting-edge scFMs across multiple downstream tasks while requiring only minimal single-cell data for pretraining. In future work, we aim to further explore scaling laws for single-cell foundation models, with the goal of developing a unified, all-in-one framework.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Khandakar Tanvir Ahmed, Sunho Park, Qibing Jiang, Yunku Yeu, TaeHyun Hwang, and Wei Zhang. Network-based drug sensitivity prediction. *BMC medical genomics*, 13:1–10, 2020.

Alexandre F Aissa, Abul BMMK Islam, Majd M Ariss, Cammille C Go, Alexandra E Rader, Ryan D Conrardy, Alexa M Gajda, Carlota Rubio-Perez, Klara Valyi-Nagy, Mary Pasquinelli, et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature communications*, 12(1):1628, 2021.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Yinghao Cao, Xiaoyue Wang, and Gongxin Peng. Scsa: a cell type annotation tool for single-cell rna-seq data. *Frontiers in genetics*, 11:490, 2020a.

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

Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, and Ge Gao. Searching large-scale scrna-seq databases via unbiased cell embedding with cell blast. *Nature communications*, 11(1):3458, 2020b.

Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.

Jiawei Chen, Hao Xu, Wanyu Tao, Zhaoxiong Chen, Yuxuan Zhao, and Jing-Dong J Han. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, 2023.

Junyi Chen, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu, and Qin Ma. Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data. *Nature Communications*, 13(1):6494, 2022a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Xiaoyang Chen, Shengquan Chen, Shuang Song, Zijing Gao, Lin Hou, Xuegong Zhang, Hairong Lv, and Rui Jiang. Cell type annotation of single-cell chromatin accessibility data via supervised bayesian embedding. *Nature Machine Intelligence*, 4(2):116–126, 2022b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Isidro Cortes-Ciriano, Lewis H Mervin, and Andreas Bender. Current trends in drug sensitivity prediction. *Current pharmaceutical design*, 22(46):6918–6927, 2016.

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

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Rui Dong and Guo-Cheng Yuan. Giniclust3: a fast and memory-efficient tool for rare cell type identification. *BMC bioinformatics*, 21:1–7, 2020.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

Sabit Ekin. Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023.

Gilad D Evrony, Anjali Gupta Hinch, and Chongyuan Luo. Applications of single-cell dna sequencing. *Annual review of genomics and human genetics*, 22(1):171–197, 2021.

Teng Fei and Tianwei Yu. scbatch: batch-effect correction of rna-seq data through sample distance matrix adjustment. *Bioinformatics*, 36(10):3115–3123, 2020.

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

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166, 2022.

Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, 35(6):498–507, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Sung Pil Hong, Thalia E Chan, Ylenia Lombardo, Giacomo Corleone, Nicole Rotmensz, Sara Bravaccini, Andrea Rocca, Giancarlo Pruneri, Kirsten R McEwen, R Charles Coombes, et al. Single-cell transcriptomics reveals multi-step adaptations to endocrine therapy. *Nature communications*, 10 (1):3840, 2019.

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

Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome biology*, 17:1–13, 2016.

Yijia Jiang, Zhirui Hu, Allen W Lynch, Junchen Jiang, Alexander Zhu, Ziqi Zeng, Yi Zhang, Gongwei Wu, Yingtian Xie, Rong Li, et al. scatanno: automated cell type annotation for single-cell atac sequencing data. *bioRxiv*, pp. 2023–06, 2023.

Aashi Jindal, Prashant Gupta, Jayadeva, and Debarka Sengupta. Discovery of rare cells from voluminous single cell expression data. *Nature communications*, 9(1):4719, 2018.

Ino D Karemaker and Michiel Vermeulen. Single-cell dna methylation profiling: technologies and biological applications. *Trends in biotechnology*, 36(9):952–965, 2018.

Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4): 610–620, 2015.

Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, et al. Multimodal single cell data integration challenge: results and lessons learned. *BioRxiv*, pp. 2022–04, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Bora Lim, Yiyun Lin, and Nicholas Navin. Advancing cancer research and medicine with single-cell genomics. *Cancer cell*, 37(4):456–470, 2020.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, and Hongyu Zhao. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, pp. 2023–12, 2023b.

Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, et al. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Eleni P Mimitou, Caleb A Lareau, Kelvin Y Chen, Andre L Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z Yeung, Efthymia Papalexi, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology*, 39(10):1246–1258, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

David T Paik, Sangkyun Cho, Lei Tian, Howard Y Chang, and Joseph C Wu. Single-cell rna sequencing in cardiovascular development, disease and medicine. *Nature Reviews Cardiology*, 17 (8):457–473, 2020.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikolaus Rajewsky, Geneviève Almouzni, Stanislaw A Gorski, Stein Aerts, Ido Amit, Michela G Bertero, Christoph Bock, Annelien L Bredenoord, Giacomo Cavalli, Susanna Chiocca, et al. Lifetime and improving european healthcare through cell-based interceptive medicine. *Nature*, 587(7834):377–386, 2020.

Andrea Ravasio, Myint Z Myaing, Shumei Chia, Aditya Arora, Aneesh Sathe, Elaine Yiqun Cao, Cristina Bertocchi, Ankur Sharma, Bakya Arasi, Vin Yee Chung, et al. Single-cell analysis of epha clustering phenotypes to probe cancer cell heterogeneity. *Communications Biology*, 3(1):429, 2020.

Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502, 2015.

Xin Shao, Haihong Yang, Xiang Zhuang, Jie Liao, Penghui Yang, Junyun Cheng, Xiaoyan Lu, Huajun Chen, and Xiaohui Fan. scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic acids research*, 49(21):e122–e122, 2021.

Ankur Sharma, Elaine Yiqun Cao, Vibhor Kumar, Xiaoqian Zhang, Hui Sun Leong, Angeline Mei Lin Wong, Neeraja Ramakrishnan, Muhammad Hakimullah, Hui Min Vivian Teo, Fui Teen Chong, et al. Longitudinal single-cell rna sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nature communications*, 9(1):4931, 2018.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Gour Suman, Mathur Nupur, Singh Anuradha, and Bhatnagar Pradeep. Single cell protein production: a review. *Int. J. Curr. Microbiol. App. Sci*, 4(9):251–262, 2015.

Chayaporn Suphavilai, Shumei Chia, Ankur Sharma, Lorna Tu, Rafael Peres Da Silva, Aanchal Mongia, Ramanuj DasGupta, and Niranjan Nagarajan. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Medicine*, 13:1–14, 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

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

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Kyle J Travaglini, Ahmad N Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V Sit, Stephen Chang, Stephanie D Conley, Yasuo Mori, Jun Seita, et al. A molecular cell atlas of the human lung from single-cell rna sequencing. *Nature*, 587(7835):619–625, 2020.

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

Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature biotechnology*, 41(5):604–606, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Haojia Wu, Yuhei Kirita, Erinn L Donnelly, and Benjamin D Humphreys. Advantages of single-nucleus over single-cell rna sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *Journal of the American Society of Nephrology*, 30(1):23–32, 2019.

Meiye Wu and Anup K Singh. Single-cell protein analysis. *Current opinion in biotechnology*, 23(1): 83–88, 2012.

Lei Xiong, Tianlong Chen, and Manolis Kellis. scclip: Multi-modal single-cell contrastive learning integration pre-training. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Yunpei Xu, Shaokai Wang, Qilong Feng, Jiazhi Xia, Yaohang Li, Hong-Dong Li, and Jianxin Wang. sccad: Cluster decomposition-based anomaly detection for rare cell identification in single-cell expression data. *Nature Communications*, 15(1):7561, 2024.

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

Yuansong Zeng, Mai Luo, Ningyuan Shangguan, Peiyu Shi, Junxi Feng, Jin Xu, Ken Chen, Yutong Lu, Weijiang Yu, and Yuedong Yang. Deciphering cell types by integrating scatac-seq data with genome sequences. *Nature computational science*, 4(4):285–298, 2024.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024a.

Suyuan Zhao, Jiahuan Zhang, Yizhen Luo, Yushuai Wu, and Zaiqing Nie. Langcell: Language-cell pre-training for cell identity understanding. *arXiv preprint arXiv:2405.06708*, 2024b.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs were not involved in the research ideation or the writing of this paper.

B ADDITIONAL EXPERIMENTS

B.1 CELL-TEXT RETRIEVAL

Assessing the alignment between cellular and LLM-derived textual spaces constitutes a critical evaluation of our framework. We quantify alignment effectiveness through cell-text retrieval tasks. Fig. 7 (a) presents UMAP embeddings from BMMC and PBMC datasets at three evaluation stages. Column 1 reveals pronounced modality divergence prior to alignment. Column 2 demonstrates substantial cross-modal convergence following `RAGCell` application, with emergent cell-type-specific clustering. Column 3 confirms robust zero-shot generalization, where cell embeddings recapitulate the clustering patterns of text embeddings, indicating successful knowledge transfer from LLM-derived databases to cellular spaces. Fig. 7 (b) evaluates category-level retrieval using Mean Average Precision (MAP). `RAGCell` outperforms existing LLM-based scFMs in all retrieval tasks (cell2text, text2text, cell2cell). This superiority stems from our cell-type-informed text descriptions, which during alignment cluster same-type cells while distancing different-type cells in embedding space. Figs. 7 (c) and (d) quantify instance-level retrieval via Recall@k and Mean Reciprocal Rank (MRR). `RAGCell` consistently surpasses baselines across both cell2text and text2cell tasks. From the results, we can observe that current LLM-based scFMs (e.g., GenePT, scELMo) exhibit critical limitations: (1) directly transplanting LLM embeddings ignores inherent cross-modal heterogeneity, and (2) lacking explicit alignment mechanisms to bridge this divergence. Conversely, `RAGCell`'s dedicated alignment strategy explicitly addresses modality gaps, enabling improved retrieval performance.

C ADDITIONAL RELATED WORK

C.1 LARGE LANGUAGE MODELS

Large language models have driven rapid advancements in natural language processing, continuously evolving through innovations in architecture design and training methodologies. Early breakthroughs include BERT (Devlin et al., 2019), which leverages bidirectional context mining via masked language modeling on Transformer encoders (Vaswani, 2017), achieving state-of-the-art performance on text understanding tasks. Shortly thereafter, the GPT series (Radford et al., 2018; 2019; Brown, 2020) pioneered autoregressive pretraining with Transformer decoders, demonstrating few-shot learning capabilities that emerged as model sizes increased. These models laid the foundation for the generative AI revolution, enabling powerful applications in text generation, reasoning, and multimodal understanding. Following the trend of open-source LLMs, LLaMA (Touvron et al., 2023) introduced a suite of efficient models trained on publicly available datasets, fostering a wave of community-driven adaptations such as Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023). Meanwhile, Chinese researchers contributed innovations such as DeepSeek LLM (Bi et al., 2024), which combined mixture-of-experts (MoE) architectures with domain-specific data filtering to enhance mathematical reasoning and coding proficiency while maintaining parameter efficiency.

D INTRODUCTION TO INCORPORATED DATASETS

CITE-seq and ASAP-seq PBMC dataset. The CITE-seq and ASAP-seq PBMC datasets contain multimodal data from both control and stimulated conditions. After preprocessing, 4,644 CITE-seq cells and 4,502 ASAP-seq cells associated with 17,441 genes are included in the experiments. The CITE-seq data includes seven categories: B cells, Effector CD4+ T cells, Effector CD8+ T cells, Monocytes cells, NK cells, Naive CD4+ T cells, and Naive CD8+ T cells. The ASAP-seq data includes nine categories: B cells, DC cells, Effector CD4+ T cells, Effector CD8+ T cells, Monocytes cells, NK cells, Naive CD4+ T cells, Naive CD8+ T cells, and others. This dataset can be downloaded from GSE156478 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156478).

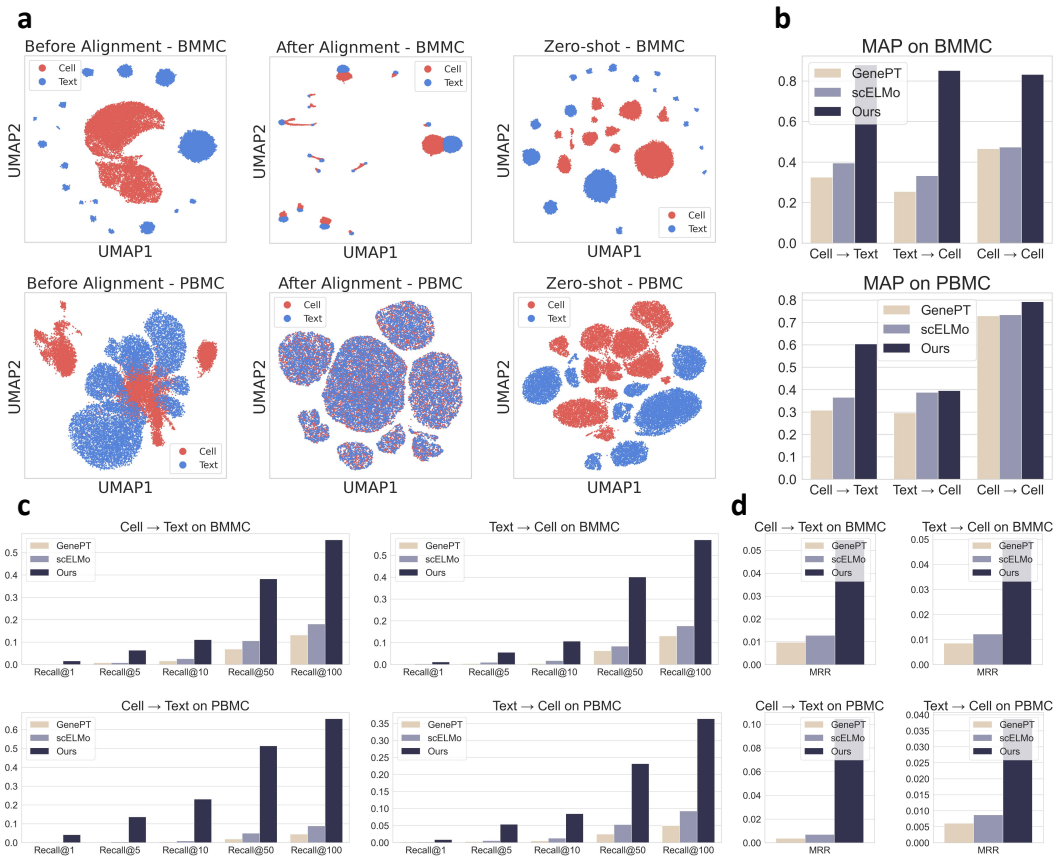

Figure 7: **Cell-Text Retrieval Results on BMMC and PBMC data.** **a.** UMAP plot of the embeddings before cell-text alignment, after cell-text alignment, and under zero-shot settings. **b.** Zero-shot category-level cell-text retrieval comparisons (MAP). **c.** Zero-shot instance-level cell-text retrieval comparisons (Recall). **d.** Zero-shot instance-level cell-text retrieval comparisons (MRR).

PBMC 10K dataset. The PBMC 10K dataset includes two batches of scRNA-seq data from human PBMCs of a healthy donor. After preprocessing, this dataset features 3,346 differentially expressed genes. The first batch and the second batch comprise 7,982 cells and 4,008 cells, respectively. Cell groups were labeled by Seurat (Satija et al., 2015) and categorized into nine types: B cells, CD4+ T cells, CD8+ T cells, CD14+ monocytes, dendritic cells, NK cells, FCGR3A+ monocytes, megakaryocytes, and others. This dataset can be downloaded from the scVI tools (Virshup et al., 2023; Gayoso et al., 2022) (`https://scvi-tools.org/`) using the API scvi.data.pbmc_dataset.

BMMC dataset. The BMMC dataset uses the CITE-seq protocol and comprises paired measurements of scRNA-seq and protein abundance in BMMCs. This dataset includes cells from 12 healthy human donors, organized into 12 batches. After preprocessing, the data encompass 90,261 cells and each cell contains 13,953 genes and 134 surface proteins. There are 45 distinct cell types in this dataset. This dataset can be downloaded from GSE194122 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122`).

GSE149383-Erlotinib. The GSE149383 dataset comprises 2,739 human lung cancer cells and 8,380 associated genes, along with their respective sensitivity properties to the drug Erlotinib. This dataset can be downloaded from GSE149383 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149383`).

GSE117872-Cisplatin. The GSE117872 dataset comprises 1,302 human oral squamous cancer cells and 22,744 associated genes, along with their respective sensitivity properties to the drug

Cisplatin. This dataset can be downloaded from GSE117872 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117872).

E Introduction to Compared Methods

To evaluate the performance of RAGCell for downstream tasks, we benchmarked our method against a range of state-of-the-art (SOTA) approaches. For tasks like cell type annotation, batch effect correction, and multi-omics data integration, we compared RAGCell with four scFMs, including scBERT, scGPT, GenePT, and scELMo. For the rare cell type identification task, we employed scCAD as the framework and replaced the original cell embeddings with cell representations obtained by our approach. For the drug sensitivity prediction task, we compared RAGCell with both specialist models like vanilla Transformer, scDEAL, and generalist models like GPT-3.5.

scBERT (Yang et al., 2022). scBERT is a pretrain-based scFM that utilizes a Performer cell encoder and employs masked gene modeling for pretraining on one million cells from the Panglao dataset (Franzén et al., 2019). We used the pretrained model provided by the authors from (https://drive.weixin.qq.com/s?k=AJEAIQdfAAoUxhXE7r#) and validated downstream tasks according to their publicly available code from (https://github.com/TencentAILabHealthcare/scBERT). All other settings were kept consistent with those provided by the authors.

scGPT (Cui et al., 2024). scGPT is another pretrain-based scFM that employs a decoder-based architecture combined with attention mask for autoregressive pretraining on 33 million cells. The authors provided several pretrained model weights, with pretraining data ranging from 1.8 million heart cells to 33 million normal human cells. We utilized the model pretrained on the largest dataset from (https://drive.google.com/drive/folders/1oWh_-ZRdhtoGQ2Fw24HP41FgLoomVo-y) and conducted downstream task experiments following the authors' publicly available code from (https://github.com/bowang-lab/scGPT). All other settings were kept consistent with those provided by the authors.

GenePT (Chen & Zou, 2023). GenePT is an LLM-based scFM that obtains text embeddings of features (such as genes and proteins) from LLM using text descriptions from the NCBI database as prompts. It then directly uses these LLM embeddings to construct cell embeddings. We employed the GenePT-w method, as provided by the authors from (https://github.com/yiqunchen/GenePT), to construct cell embeddings and validate them for downstream tasks. All other settings were kept consistent with those provided by the authors.

scELMo (Liu et al., 2023b). scELMo is another LLM-based scFM, which can be considered as an improved version of GenePT. It obtains embeddings through enhanced prompts to construct high-quality cell embeddings. We utilized the code provided by the authors from (https://github.com/HelloWorldLTY/scELMo) to build cell embeddings and validate them for downstream tasks. All other settings were kept consistent with those provided by the authors.

scCAD (Xu et al., 2024). scCAD is a SOTA algorithm for rare cell type identification. It employs the PCA method for clustering to obtain cell embeddings. We conducted experiments using the publicly available code from (https://github.com/xuyp-csu/scCAD) and replaced the original cell embeddings with representations from our framework. Experimental analysis was performed using the cell embeddings generated by our approach. All other parameters of the algorithm were kept consistent with the original implementation.

Transformer (Vaswani, 2017). Transformer is a famous architecture and was initially designed for natural language processing tasks. It introduces self-attention mechanisms and provides standards for many other fields, including vision, audio, and scientific research. For the drug sensitivity prediction task, all settings and parameters were kept consistent with the standard implementation.

scDEAL (Chen et al., 2022a). scDEAL is a specialist model designed for the drug sensitivity prediction task. It employs a neural network and tries to establish connections between genes

and drug responses. The source code and implementation can be found at (https://github.com/OSU-BMBL/scDEAL). All settings and parameters were kept consistent with the original implementation.

GPT-3.5. GPT-3.5 is a SOTA generalist model developed by OpenAI. Based on the GPT-3 architecture, GPT-3.5 can understand and generate human-liked texts, which could be employed for many tasks and applications, including the drug sensitivity prediction task. The use of GPT-3.5 API can be found at (https://platform.openai.com/docs/models/gpt-3.5-turbo).

F Implementation Details

`RAGCell` employs the patch-based Transformer model as the cell model. Each cell is split into 128 patches for data tokenization. For cell-level knowledge databases, we employ the GPT-4o mini model to generate function descriptions and use the 'text-embedding-3-large' model to extract embeddings, with an embedding dimension of 3072. For feature-level knowledge databases, we leverage the GPT-3.5 model to generate function descriptions and use the 'text-embedding-ada-002' model to extract embeddings, with an embedding dimension of 1536. To map cell and text representations into a shared space, we use two MLPs as projection layers, reducing both representations to a low-dimensional space of 128 dimensions. All experiments are conducted using Pytorch with 80GB A100 GPUs as support. For cell-text alignment, we set the batch size to 256 and trained the cell model with the AdamW optimizer for 100 epochs. For finetuning tasks, we randomly split the datasets into 80% for training, 10% for validation, and 10% for testing. For zero-shot tasks, we ensure that there is no data leakage or overlap during the pretraining phase. The learning rate is set to 1.5e-4 during the pretraining stage and 1e-4 during the finetuning stage, respectively.

G Evaluation Metrics

Cell Type Annotation. For the cell type annotation task, we evaluated the model's performance using four common metrics for classification tasks: Accuracy, Precision, Recall, and F1 score. The calculation processes for each metric are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{12}$$

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \tag{13}$$

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, \tag{14}$$

$$\text{F1 Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \tag{15}$$

where TP, TN, FP, and FN are short for true positives, true negatives, false positives, and false negatives, respectively. N denotes the number of samples per cell type.

Batch Effect Correction. For the batch effect correction task, we evaluated the model's performance using several common cell clustering metrics, specifically normalized mutual information (NMI), adjusted rand index (ARI), and average silhouette width (ASW). NMI can be calculated as:

$$\text{I}(U, V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P(j)}\right), \tag{16}$$

$$\text{NMI}(U, V) = \frac{I(U, V)}{\text{mean}(\text{H}(U), \text{H}(V))}, \tag{17}$$

where $P(i,j)$ is the probability that the i-th cluster in clustering U and the j-th cluster in clustering V occur simultaneously. $P(i)$ and $P(j)$ are the probabilities of the i-th cluster in U and the j-th cluster in V, respectively.

ARI can be calculated as:

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{18}$$

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}, \tag{19}$$

where TP, TN, FP, and FN are short for true positives, true negatives, false positives, and false negatives, respectively. $E[RI]$ is the expected RI of random labeling.

The calculation process for ASW can be represented as follows:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j), \tag{20}$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j), \tag{21}$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{22}$$

$$\text{ASW} = \frac{1}{N} \sum_{i=1}^{N} s(i), \tag{23}$$

where C_i and C_k are clusters containing sample points i and k, and $d(i, j)$ is the distance between sample points i and j. N denotes the number of cell samples.

Multi-omics Data Integration. For the multi-omics data integration task, we employed three metrics, namely NMI, ARI, and ASW, to evaluate the performance of the models. The calculation processes for these metrics have been previously described.

Rare Cell Type Identification. For the rare cell type identification task, we employed three metrics, namely Precision, Recall, and F1 score, to evaluate the performance of the models. The calculation processes for these metrics have been previously described.

Cell-Text Retrieval. For cell-text retrieval tasks, we leveraged multiple evaluation metrics for different tasks, including instance-level retrieval and category-level retrieval. For instance-level retrieval, we employed two metrics: recall@k and MRR, which can be calculated as:

$$\text{Recall@k} = \frac{|\{\text{relevant samples}\} \cap \{\text{retrieved samples}\}_{\text{top k}}|}{|\{\text{relevant samples}\}|}, \tag{24}$$

$$\text{MRR} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i}, \tag{25}$$

where $|\{\text{relevant samples}\} \cap \{\text{retrieved samples}\}_{\text{top k}}|$ represents the number of relevant samples in the top k retrieved results, $|\{\text{relevant samples}\}|$ represents the total number of relevant samples, r_i represents the rank of the i-th query, n represents the total number of queries. In instance-level retrieval, there is only one relevant text/cell sample for each query cell/text sample. Here we set $n = 1000$.

For category-level retrieval, we employed MAP to evaluate the performance. The calculation process for MAP can be formulated as follows:

$$\text{MAP} = \frac{1}{n} \sum_{i=1}^{n} \text{AP}_i, \tag{26}$$

$$\text{AP} = \frac{1}{|R|} \sum_{m \in R} \frac{\text{number of relevant samples in top m}}{m}, \tag{27}$$

where AP_i represents the average precision of the i-th query, $|R|$ represents the number of relevant samples, m represents the rank of the retrieved sample, and number of relevant samples in top m represents the number of relevant samples in the top m retrieved results. In category-level retrieval, samples from the same cell types are considered relevant samples for a given query. Here we set $m = n = 1000$.

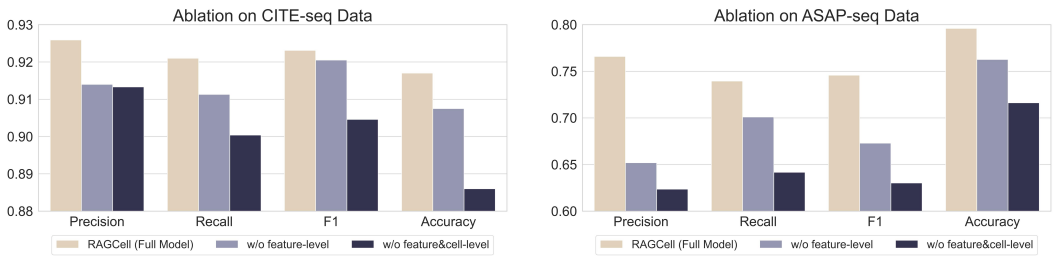

Figure 8: Ablation study on CITE-seq and ASAP-seq data.

Drug Sensitivity Prediction. For the drug sensitivity prediction task, we employed four metrics, namely Precision, Recall, F1 score, and Accuracy, to evaluate the performance of the models. The calculation processes for these metrics have been previously described.

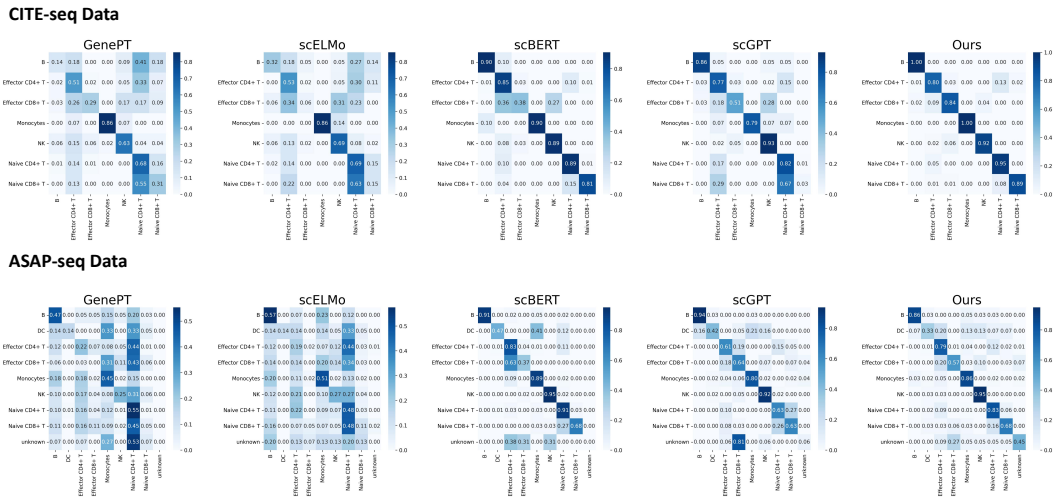

Figure 9: Confusion matrix results on CITE-seq and ASAP-seq data.

H Additional Experiments

H.1 Ablation Study

To verify the effectiveness of incorporating priors from LLMs, we also provide some model variants by gradually removing text descriptions at each level, with results shown in Fig. 8. As illustrated in the figure, 'w/o feature-level' signifies that only cell-level descriptions were included, whereas 'w/o feature&cell-level' indicates the absence of all textual information. The results indicate performance declines among various model variants, which suggests the need of both cell-level and feature-level text descriptions in enhancing cell representations.

H.2 Additional Cell Type Annotation Results

We also present the cell type annotation confusion matrix results in Fig. 9, which highlights the superior performance of `RAGCell` compared to other scFMs. `RAGCell` achieves significantly higher annotation accuracy consistently across diverse cell types.

I Examples of Text Descriptions

Here we provide some examples of cell-level and feature-level text descriptions on different datasets in Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15.

CITE-seq Data

Cell-level

Human: This is a single cell CITE sequence sample with a cell type of B. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: B cells, a critical component of the adaptive immune system, primarily function to mediate humoral immunity through the production of antibodies. Derived from hematopoietic stem cells in the bone marrow, these lymphocytes undergo a complex maturation process, during which they develop the ability to recognize and bind specific antigens presented by pathogens. Upon activation by antigenic stimulation, often facilitated by T helper cells, B cells differentiate into plasma cells, which are responsible for secreting large quantities of antibodies. These antibodies play an essential role in neutralizing toxins, opsonizing pathogens for enhanced phagocytosis, and activating the complement system, thereby orchestrating an effective immune response. Additionally, B cells contribute to immune memory, enabling a more rapid and robust response upon subsequent exposures to the same antigen through the generation of memory B cells. Overall, the primary function of B cells is to provide protection against infections and to maintain immunological memory.

Feature-level

Human: Please summarize the major function of gene: CDK11B. Use academic language in one paragraph and and think step by step.

LLM: The CDK11B gene encodes a member of the cyclin-dependent kinase (CDK) family, which is integral to the regulation of the cell cycle, particularly in the G2 and M phases. Specifically, CDK11B functions as a serine/threonine kinase that associates with cyclins to phosphorylate target proteins, thereby facilitating various cellular processes including transcription regulation, cell proliferation, and differentiation. Notably, CDK11B also plays a crucial role in mRNA processing and spliceosome assembly, underscoring its involvement not only in cell cycle progression but also in the post-transcriptional modification of RNA. Dysregulation of CDK11B activity has been implicated in various cancers, highlighting its potential as a therapeutic target.

Figure 10: Examples of cell-level and feature-level text descriptions on CITE-seq data.

ASAP-seq Data

Cell-level

Human: This is a single cell ASAP sequence sample with a cell type of DC. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: Dendritic cells (DCs) are pivotal antigen-presenting cells within the immune system, characterized primarily by their ability to capture, process, and present antigens to naïve T lymphocytes, thereby initiating and modulating adaptive immune responses. These cells reside in peripheral tissues and serve as sentinels, detecting pathogens and environmental changes. Upon encountering antigens, DCs undergo a maturation process that enhances their capacity to migrate to lymphoid tissues, where they express co-stimulatory molecules and major histocompatibility complex (MHC) molecules crucial for T cell activation. Additionally, DCs can secrete a variety of cytokines and chemokines, which not only fortify T cell responses but also influence the differentiation of other immune cells, such as B cells and natural killer cells. Importantly, the interaction between DCs and T cells is critical for establishing immune tolerance and preventing autoimmune reactions, highlighting the dual role of DCs in both instigating protective immunity and maintaining homeostasis within the immune landscape.

Feature-level

Human: Please summarize the major function of gene: SCNN1D. Use academic language in one paragraph and and think step by step.

LLM: The gene SCNN1D encodes the delta subunit of the epithelial sodium channel (ENaC), which plays a pivotal role in regulating sodium and fluid homeostasis in various epithelial tissues, particularly in the kidney and respiratory system. The major function of SCNN1D involves its contribution to the assembly and functional activity of the ENaC complex, which is essential for sodium absorption across epithelial cell membranes. This process is critical in maintaining extracellular fluid volume, blood pressure, and the overall electrolyte balance within the body. Dysregulation or mutations in SCNN1D can lead to pathophysiological conditions, such as pseudohypoaldosteronism type 1, characterized by impaired sodium reabsorption, hyperkalemia, and metabolic acidosis. Thus, SCNN1D is fundamentally important for cardiovascular and renal physiology, as well as overall homeostatic mechanisms.

Figure 11: Examples of cell-level and feature-level text descriptions on ASAP-seq data.

PBMC 10K Data

Cell-level

Human: This is a single cell RNA sequence sample with a cell type of Megakaryocytes. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: Megakaryocytes are large, polyploid bone marrow-derived cells primarily responsible for the production of platelets, which are critical components of the hemostatic system. They originate from hematopoietic stem cells and undergo a process of endomitosis, resulting in an increase in nuclear DNA content without cell division, thus enhancing their functional capacity. The primary function of megakaryocytes is to synthesize and release platelets into the bloodstream, which is essential for the initiation and regulation of blood clotting, thereby preventing excessive bleeding during vascular injury. In addition to their role in hemostasis, megakaryocytes and platelets also participate in immune responses, inflammation, and wound healing, showcasing their multifaceted biological importance within the hematopoietic system. Furthermore, understanding megakaryocyte biology is crucial for the development of therapeutic strategies targeting platelet-related disorders and other hematological conditions.

Feature-level

Human: Please summarize the major function of gene: ISG15. Use academic language in one paragraph and and think step by step.

LLM: ISG15 (Interferon-Stimulated Gene 15) is an important ubiquitin-like protein that plays a crucial role in the innate immune response, primarily acting as a mediator of antiviral activity. Upon stimulation by interferons, ISG15 is highly induced and subsequently conjugated to target proteins through a process known as ISGylation, which modulates their function and stability. This post-translational modification not only enhances the immune response against viral infections but also regulates various cellular processes, including protein degradation, gene expression, and inflammation. Furthermore, ISG15 has been implicated in the modulation of pathways leading to apoptosis and cell signaling, thereby influencing the overall cellular stress response. Collectively, the multifaceted functions of ISG15 position it as a critical player in the host defense mechanisms against pathogens and in the maintenance of cellular homeostasis.

Figure 12: Examples of cell-level and feature-level text descriptions on PBMC 10K data.

BMMC Data

Cell-level

Human: This is a single cell sample with a cell type of B1 B IGKC+. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: The B1 B IGKC+ cell type represents a subset of B lymphocytes characterized by the expression of immunoglobulin kappa chain (IGKC) and is primarily involved in the production of antibodies, particularly in the context of innate-like immune responses. These cells are known for their ability to rapidly produce immunoglobulin M (IgM) antibodies in response to polysaccharide antigens, which are often found on the surface of pathogens. This rapid antibody production is crucial for the early defense against infections, particularly in mucosal tissues. Furthermore, B1 B cells play a significant role in maintaining immune homeostasis and contribute to the regulation of immune responses through the secretion of cytokines and the modulation of other immune cell types. Their unique developmental origin and functional properties distinguish them from conventional B2 cells, highlighting their importance in both innate and adaptive immunity.

Feature-level

Human: Please summarize the major function of gene: FAM41C. Use academic language in one paragraph and and think step by step.

LLM: FAM41C (Family with Sequence Similarity 41 Member C) is a gene that encodes a protein implicated in various cellular processes, notably in the regulation of protein homeostasis and cellular stress responses. Its expression is noted to influence autophagy, a critical mechanism for maintaining cellular integrity through the degradation and recycling of damaged organelles and proteins. Additionally, FAM41C has been associated with the modulation of signaling pathways related to cell proliferation and apoptosis, suggesting a role in maintaining cellular health and viability. Recent research has also indicated its potential involvement in pathological conditions, including various cancers, where dysregulation of FAM41C expression may contribute to tumorigenesis and cancer progression. Overall, FAM41C serves as a pivotal factor in cellular maintenance and responses to stress, with implications for both normal physiological processes and disease states.

Human: Please summarize the major function of ADT: CD101. Use academic language in one paragraph and and think step by step.

LLM: CD101, a novel immunotherapy agent, is an antifungal compound that functions primarily by modulating the host immune response to enhance the clearance of fungal pathogens. It achieves this through the inhibition of the enzyme cytochrome P450 14α-demethylase, which plays a critical role in ergosterol biosynthesis, thereby disrupting fungal cell membrane integrity. By selectively targeting the fungal metabolic processes while minimizing toxicity to human cells, CD101 promotes a robust immune response against invasive fungal infections. Furthermore, its formulation allows for extended release, facilitating sustained therapeutic effects and improving patient outcomes. Overall, CD101 represents a significant advancement in the treatment of fungal infections, particularly in immunocompromised populations, by integrating antifungal activity with immune modulation to combat resistance and enhance efficacy.

Figure 13: Examples of cell-level and feature-level text descriptions on BMMC data.

GSE149383_Erlotinib

Cell-level

Human: This is a single cell RNA sequence sample and its sensitivity to the drug Erlotinib is resistant. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: The single-cell RNA sequencing analysis of the sample reveals a distinct transcriptional profile that underlies its resistance to the drug Erlotinib, an epidermal growth factor receptor (EGFR) inhibitor. This resistance may be attributed to several cellular mechanisms, including the aberrant activation of alternative signaling pathways, upregulation of efflux transporters, or mutations in the EGFR or downstream mediators that negate the therapeutic effects of Erlotinib. Furthermore, the expression of immune checkpoint molecules and inflammatory cytokines could indicate an adaptive immune response that counteracts the drug's efficacy. The identified gene expression patterns suggest that this cell type may play a critical role in tumor microenvironment dynamics, contributing to therapeutic failure through cell-autonomous survival strategies and modulation of local inflammation, thereby presenting significant implications for the development of alternative therapeutic strategies targeting resistance mechanisms in cancer treatment.

Feature-level

Human: Please summarize the major function of gene: AADAC. Use academic language in one paragraph and and think step by step.

LLM: The gene AADAC (aromatic-L-amino acid decarboxylase) encodes an enzyme that is primarily involved in the metabolism of aromatic L-amino acids, specifically in their decarboxylation processes, converting these substrates into biogenic amines. This enzymatic activity is crucial in the synthesis of neurotransmitters such as dopamine, serotonin, and norepinephrine, playing a significant role in regulating key physiological functions, including mood, cognition, and motor control. Furthermore, AADAC is implicated in the detoxification pathways of certain xenobiotics, as it contributes to the catabolism of various amines and lipid-derived substances. The gene's expression may be influenced by physiological and pathological conditions, highlighting its importance in homeostasis and potential involvement in metabolic disorders. Understanding AADAC's functions not only elucidates biochemical pathways relevant to neurotransmitter biosynthesis but also underscores its potential therapeutic implications in neuropharmacology and toxicology.

Figure 14: Examples of cell-level and feature-level text descriptions on GSE149383 data.

GSE117872_Cisplatin

Cell-level

Human: This is a single cell RNA sequence sample and its sensitivity to the drug Cisplatin is sensitive. Please summarize its major function. Use academic language in one paragraph and think step by step.

LLM: Cisplatin is a chemotherapeutic agent predominantly utilized in the treatment of various solid tumors, and its effectiveness hinges upon its ability to induce DNA damage, thereby triggering cellular apoptosis in cancerous cells. In the context of a single-cell RNA sequencing sample characterized by sensitivity to Cisplatin, several key functional pathways may be implicated. These include the activation of DNA repair mechanisms, regulation of apoptosis, and cellular stress response pathways. A sensitive response typically suggests an upregulation of intrinsic pro-apoptotic signals and downregulation of survival pathways, facilitating the eradication of malignant cells. Furthermore, gene expression profiling may reveal alterations in cell cycle regulation and metabolic pathways that enhance vulnerability to Cisplatin-induced cytotoxicity, thereby underscoring the complex interplay between pharmacological intervention and the molecular landscape of the tumor microenvironment. Overall, the major function of this sensitive sample can be framed within the context of promoting programmed cell death in response to genotoxic stress, thereby elucidating potential therapeutic targets for optimizing cancer treatment outcomes.

Feature-level

Human: Please summarize the major function of gene: DENND6B. Use academic language in one paragraph and and think step by step.

LLM: DENND6B (DENN domain-containing 6B) is a gene that encodes a protein characterized by the presence of a DENN domain, which is implicated in the regulation of small GTPases, particularly those within the Rab family that are critical for vesicular trafficking processes. The major function of DENND6B involves facilitating the activation and subsequent signaling pathways of specific Rab GTPases, which are essential for the transport of vesicles within cells. This process plays a pivotal role in mediating intracellular communication and the regulated secretion of molecules, influencing cellular response mechanisms to various stimuli. Furthermore, DENND6B has been associated with various physiological processes, including neuronal development and immune system function, highlighting its potential significance in health and disease states. Overall, the regulation of Rab GTPases by DENND6B underscores its fundamental role in maintaining cellular homeostasis and proper cellular functionality.

Figure 15: Examples of cell-level and feature-level text descriptions on GSE117872 data.