LangCell: Language-Cell Pre-training for Cell Identity Understanding

Suyuan Zhao [*1 2] **Jiahuan Zhang** [*1] **Yushuai Wu** [1] **Yizhen Luo** [1 2] **Zaiqing Nie** [†1 3]

Abstract

Cell identity encompasses various semantic aspects of a cell, including cell type, pathway information, disease information, and more, which are essential for biologists to gain insights into its biological characteristics. Understanding cell identity from the transcriptomic data, such as annotating cell types, has become an important task in bioinformatics. As these semantic aspects are determined by human experts, it is impossible for AI models to effectively carry out cell identity understanding tasks without the supervision signals provided by single-cell and label pairs. The single-cell pre-trained language models (PLMs) currently used for this task are trained only on a single modality, transcriptomics data, lack an understanding of cell identity knowledge. As a result, they have to be fine-tuned for downstream tasks and struggle when lacking labeled data with the desired semantic labels. To address this issue, we propose an innovative solution by constructing a unified representation of single-cell data and natural language during the pre-training phase, allowing the model to directly incorporate insights related to cell identity. More specifically, we introduce **LangCell**, the first **Lang**uage-**Cell** pre-training framework. LangCell utilizes texts enriched with cell identity information to gain a profound comprehension of cross-modal knowledge. Results from experiments conducted on different benchmarks show that LangCell is the only single-cell PLM that can work effectively in zero-shot cell identity understanding scenarios, and also significantly outperforms existing models in few-shot and fine-tuning cell identity understanding scenarios.

[*]Equal contribution [†]Corresponding author [1]Institute for AI Industry Research (AIR), Tsinghua University [2]Department of Computer Science and Tecnology, Tsinghua University [3]PharMolix Inc.. Correspondence to: Suyuan Zhao <sxdtzsy@gmail.com>, Jiahuan Zhang <zhangjiahuan@air.tsinghua.edu.cn>, Zaiqing Nie <zaiqing@air.tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Single-cell RNA sequencing (scRNA-seq) data represents a powerful tool for deciphering the "language of life", offering profound insights into downstream biomedical applications (Ziegenhain et al., 2017). In scRNA-seq data analysis, it is crucial to understand cell identity from multiple perspectives, such as cell type, pathway information and disease information (Morris, 2019; Abdolhosseini et al., 2019). Tasks like cell type annotation and cell batch integration have become the cornerstone of this field (Luecken & Theis, 2019; Luecken et al., 2022).

Pre-trained language models (PLMs) have recently demonstrated success in deciphering the complex language of life (Mo et al., 2021; Ji et al., 2021). Building on these findings, recent studies emphasize the effectiveness and feasibility of using PLMs to analyze single-cell data solely based on sequencing information (Yang et al., 2022a; Theodoris et al., 2023; Cui et al., 2023; Gong et al., 2023). These models harness the transformer architecture to assimilate millions of scRNA-seq entries, refining their capabilities through fine-tuning to adeptly perform diverse downstream tasks. Despite these successes, current single-cell representation models face the following challenges:

(1). Current model frameworks, which rely solely on self-supervised learning methods like masked modeling, are adept at capturing gene co-expression relationships. However, due to a lack of effective utilization of human expert knowledge, they fall short in focusing on understanding cell identity when learning cell representations. This limitation restricts the model's capacity for representation learning, consequently affecting its performance in various downstream tasks.

(2). As cell identities are determined by human experts in natural language, it is impossible for existing models to effectively carry out cell identity understanding tasks without fine-tuned with single cell and text/label pairs. Both the amount and quality of data for fine-tuning significantly impact the model's performance in specific tasks. However, in practical scenarios, obtaining sufficient and reliable labeled data that closely matches the downstream task is often costly. This difficulty is even more pronounced in situations such as researching new diseases or cell subtypes, where no existing data may be available (Zhai et al., 2023). These

practical issues significantly reduce the convenience and applicability of existing models.

BioTranslator (Xu et al., 2023a) considers combining biomedical data and natural language in the current research landscape. However, BioTranslator only used Transformer-based model and performed large-scale pre-training on natural language modality. It did not actually pre-train on large amounts of single-cell data but instead relied on training a naive fully connected network on downstream data, which struggles to capture the richness and complexity of transcriptomic data.

We believe that encoding scRNA-seq data with high quality and aligning it with multi-perspective textual annotations can significantly enhance the comprehension between textual information and single-cell data. This integration equips the model with the capacity to extend its knowledge from familiar categories to novel ones, guided by semantic coherence. This approach not only enhances the model's predictive accuracy but also bolsters its applicability across diverse biomedical scenarios. We propose **LangCell**, a genuine **Lang**uage-**Cell** Pre-training model to seamlessly integrate the feature space of scRNA-seq data with textual information, marking significant advancements in understanding cell identity.

We constructed a cell-text dataset, **scLibrary**, containing 27.5 million scRNA-seq entries along with their descriptions. Specifically, we obtained raw scRNA-seq data and corresponding metadata from the CELLxGENE (Biology et al., 2023). We selected eight critical aspects of cell identity that could contain essential insights, including cell type, developmental stage and disease information, to obtain as comprehensive descriptions as possible from the Open Biological and Biomedical Ontology Foundry (OBO Foundry) (Smith et al., 2007).

Subsequently, we have transferred some key insights from the fields of NLP and CV (Li et al., 2022; 2021; Gao et al., 2021; Park et al., 2023), and designed a set of multi-task cooperative pre-training methods that are effective in the cell-text domain. Specifically, we introduce four tasks during the pre-training phase. Masked Gene Modeling (MGM) and Cell-Cell Contrastive Learning (C-C), to enhance single-cell representation learning. Additionally, we employ Cell-Text Contrastive Learning (C-T) and Cell-Text Matching (CTM) to train our model to recognize the underlying links between single-cell and textual data.

LangCell achieves state-of-the-art (SOTA) performance on a range of cell identity understanding tasks across zero-shot, few-shot, and full dataset scenarios. It addresses classic tasks such as cell type annotation and batch integration. As the first model capable of true zero-shot cell type annotation, LangCell shows excellent performance in zero-shot scenarios, surpassing few-shot baselines in most cases (Fig. 1**a**).

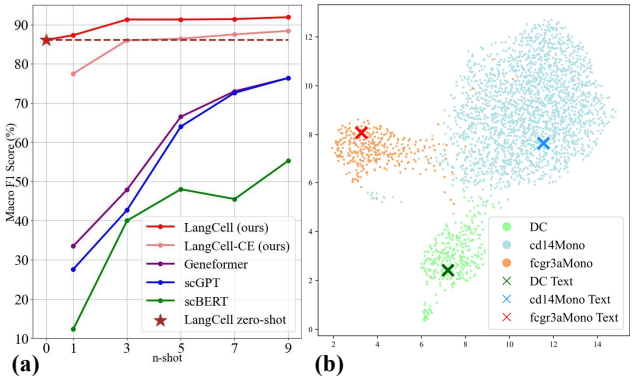

Figure 1: **(a). Plots of zero- and few-shot cell type annotation.** The curve shows the average F_1-scores on PBMC10K and PBMC3&68K, for two settings of LangCell and three of the best single-cell PLMs. **(b). UMAP plot of embeddings for scRNA-seq data and descriptions of three similar cell types in PBMC10K.** LangCell aligns single-cell and text embeddings.

In complex scenarios with rich cell types and low subtype distinction, we propose the cell-text retrieval task, enabling users to describe target cell types in natural language and search within databases. We also introduced two new tasks of great biological significance: non-small cell lung cancer subtype classification and pathway identification and constructed high-quality benchmarks for each. Results demonstrate that LangCell offers profound insights into cell identity from multiple perspectives, including cell type, disease subtype, and cell pathways.

Our main contributions can be summarized as follows:

(1). We introduce LangCell, the first Language-Cell pre-training framework. It unifies cellular language and natural language into a latent space. (Fig. 1**b**) This process infuses the model with text knowledge related to cell identity, enhancing LangCell's understanding, expression, and generalization of transcriptomic data.

(2). By harnessing the powerful link between scRNA-seq data and natural language texts, LangCell stands out as the sole PLM capable of executing zero-shot cell identity understanding tasks, surpassing the performance of existing few-shot learning models with superior experimental outcomes.

(3). LangCell's cell encoder outperforms state-of-the-art (SOTA) models in all few-shot and fine-tuning tasks related to cell identity understanding. These advancements are attributed to the synergistic impact of self-supervised learning on scRNA-seq data and distant supervision based on text.

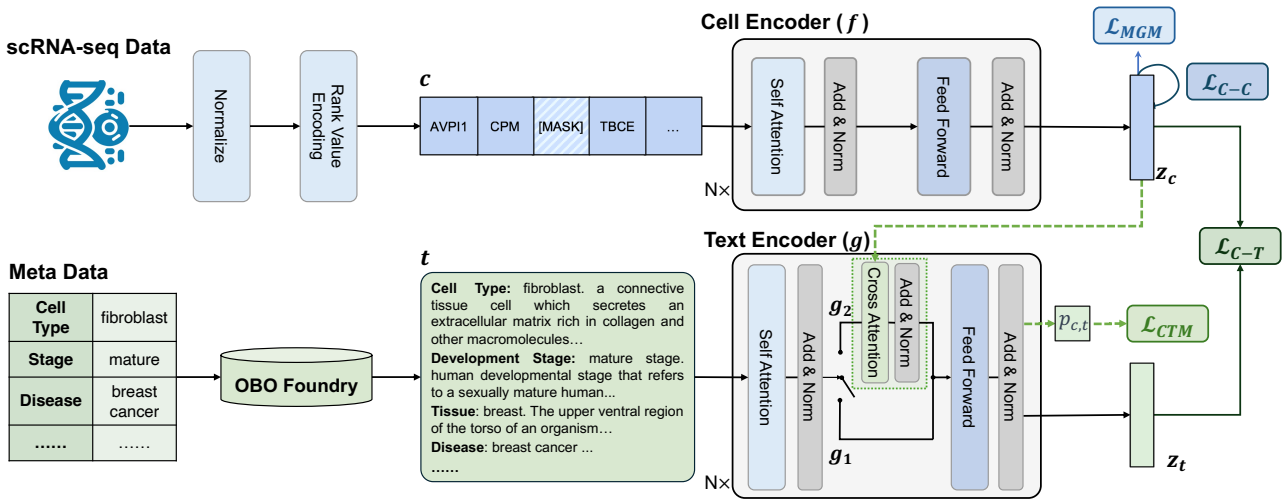

Figure 2: **The schematic overview of LangCell.** For each single-cell data containing a pair of scRNA-seq data and metadata: **(1)** During preprocessing, the scRNA-seq data is converted into a gene sequence arranged in descending order of relative expression levels, and a multi-perspective textual description of the cell is obtained from the metadata using OBO Foundry. **(2)** The embeddings of the cell and text are obtained using the cell encoder ($\boldsymbol{f}$) and the unimodal mode ($\boldsymbol{g_1}$) of the text encoder, and the matching score $p_{c,t}$ is calculated using the multimodal mode ($\boldsymbol{g_2}$) of the text encoder. **(3)** Pre-training is conducted through joint optimization of four loss functions. Among them, Masked Gene Modeling (MGM) and Cell-Cell Contrastive Learning (C-C) aim to enhance single-cell representation learning. In contrast, Cell-Text Contrastive Learning (C-T) and Cell-Text Matching (CTM) aim to train the model to understand the latent connections between single-cell and textual data.

2. Related works

scRNA-seq Data Representation PLMs offer more potential for better scRNA-seq data representation. scBERT (Yang et al., 2022a), Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2023), and scFoundation (Hao et al., 2023) are transformer-based models that collectively pre-train on extensive scRNA-seq datasets, ranging from over a million to 50M samples, and demonstrate advanced capabilities in tasks such as cell type annotation, transfer learning across biological tasks, drug response prediction and other tasks. BioTranslator (Xu et al., 2023a) bridges the gap between natural language and scRNA-seq data. However, its reliance on MLP for encoding scRNA-seq data falls short of capturing the intricacies of transcriptomic complexity.

Multi-modal in Scientific Data Scientific data, such as molecules, proteins, and scRNA-seq data, which are not as visually intuitive as images, necessitate a more profound level of multi-modal interaction and comprehension. Works like KV-PLM (Zeng et al., 2022) have paved the way for a unified comprehension of molecules and textual information, while MolT5 (Edwards et al., 2022) represents a self-supervised framework that empowers the system to handle innovative cross-modal tasks, including molecular captioning and text-based de novo molecular generation. ProtST (Xu et al., 2023b) has pioneered the field of protein multi-modal learning. However, scRNA-seq analysis has yet to witness a significant model for cross-modal representation learning, even BioTranslator has not fully realized the potential of cross-modal learning. We firmly believe that establishing connections between scRNA-seq data and text information is paramount.

3. Methods

In this section, we provide a comprehensive description of the LangCell workflow. The framework of the LangCell is illustrated in Fig. 2. Next, we first present the processing of scRNA-seq data to fit the cell encoder in 3.1, then the model architecture and pre-training methods in 3.2 and 3.3, respectively, and finally the downstream applications in 3.4.

3.1. Data Processing

The raw scRNA-seq data is provided as a count matrix. Assume there are m cells and n genes considered, then the count matrix is denoted as $A \in \mathbb{N}^{m \times n}$. We adopted a rank value encoding method (Theodoris et al., 2023; Qiu et al., 2013), used for converting the count matrix into sequence data analogous to natural language. Firstly, we normalize the gene expression of each cell separately to remove the

influence of sequencing depth to get A'. Then, we find the non-zero median $\beta \in \mathbb{R}^n$ for each column of A' as the median expression of each gene, and use β to normalize each column of A' to get A''. That is:

$$A'_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}},\ \beta_j = \text{median}\{A'_{kj} | A'_{kj} \neq 0\},\ A''_{ij} = \frac{A'_{ij}}{\beta_j}$$

Compared to A', A'' eliminates the differences brought about by the overall expression level of the base, and its values can reflect the relative level of a gene's expression in a cell among all cells. For example, some housekeeping genes may easily have higher absolute expression, but this does not necessarily indicate a particularly noteworthy high expression of that gene in that cell. Based on A'', we sort the expressed genes in each cell by their relative expression to obtain the sequence representation of that cell. Notably, we conduct statistics on a large-scale pre-training dataset to obtain a more universal β and use this β in all subsequent model applications.

3.2. Model Architecture

Our model consists of two trainable parts: a cell encoder and a text encoder.

Cell Encoder: We use pre-trained Geneformer (Theodoris et al., 2023) to initialize our cell encoder, which encodes the sequential cell inputs into an embedding sequence. Notably, we add a `[CLS]` token at the beginning of the sequence, whose embedding is projected through a linear projector as a cell embedding.

Text Encoder: This encoder has two encoding modes: unimodal and multimodal (Li et al., 2022). For unimodal text encoding, it is equivalent to a BERT (Devlin et al., 2018). For multimodal encoding, we add a pluggable cross-attention module after each self-attention module in the attention layers to compute the joint embedding and the cell-text matching probability through a linear layer. The weights are initialized using PubMedBERT (Gu et al., 2021), which is proven to be one of the best pre-trained BERTs in the biomedical field.

Define the cell encoder as $\boldsymbol{f}$, which is utilized to derive the embedding z_c from single-cell data c. Define the unimodal mode of the text encoder as $\boldsymbol{g_1}$, which is responsible for generating the embedding z_t from textual data t. Define the multimodal mode of the text encoder as $\boldsymbol{g_2}$, which is employed to calculate the matching probability $p_{c,t}$ between single-cell and text data. These three encoding methodologies are articulated as follows:

$$\boldsymbol{z_c} = \boldsymbol{f}(c),$$

$$\boldsymbol{z_t} = \boldsymbol{g_1}(t),$$

$$p_{c,t} = \boldsymbol{g_2}(\boldsymbol{z_c}, t) = \boldsymbol{g_2}(\boldsymbol{f}(c), t)$$

3.3. Pre-training Process

Our model is designed to map scRNA-seq data and text to a shared latent space and utilize the unstructured knowledge contained in natural language as distant supervision to optimize cell representation learning. To this end, during the pre-training process, we jointly optimize four objective loss functions, including masked gene modeling, intra- and inter-modal contrastive learning, and cell-text matching.

Masked Gene Modeling (MGM): We randomly mask some of the genes in the cell input sequence and use the output embeddings of the model at these positions to predict the reconstruction of the original input. We use the cross-entropy loss function as the loss function for this multi-class task:

$$\mathcal{L}_{MGM} = \frac{1}{N} \sum_{i=1}^{N} H(v_{ij}, \hat{v}_{ij})$$

where N is the number of masked genes, v_{ij} and $\hat{v}_{ij}$ respectively represent the label and predicted probability of the i-th masked position being identified as the j-th gene, H is the cross entropy loss function.

Cell-Cell Intra-Modal Contrastive Learning (C-C): We introduce cell-cell contrastive learning to alleviate the problem of representation degradation caused by BERT-based methods (Li et al., 2020; Reimers & Gurevych, 2019). In scRNA-seq data, each gene expression level carries unique meaning, and artificial data augmentation methods such as shuffling and perturbing at the input data may disrupt the gene expression semantics (Yang et al., 2022b). We believe that perturbation at the feature level is more suitable for data augmentation in scRNA-seq data. Therefore, we use two instances of standard dropout applied to the same single-cell to construct positive samples, while other single-cells in the same batch serve as negative samples, which has been proven effective in natural language research (Gao et al., 2021). To expand the batch size under limited video memory, we adopted a momentum encoder method similar to that of (Li et al., 2022). The InfoNCE (He et al., 2019) loss function is used as follows:

$$\mathcal{L}_{C-C} = -\frac{1}{T} \sum_{i=1}^{T} \log \frac{\mathrm{e}^{\mathrm{sim}(\boldsymbol{z_c}^{(i)}, \boldsymbol{z_c}^{(i)+})/\tau}}{\sum_{j=1}^{T} \mathrm{e}^{\mathrm{sim}(\boldsymbol{z_c}^{(i)}, \boldsymbol{z_c}^{(j)+})/\tau}}$$

where T is the batch size, sim is the cosine similarity function, τ is the temperature parameter, $\boldsymbol{z_c}^{(i)}$ and $\boldsymbol{z_c}^{(i)+}$ represent the embedding of the i-th cell and its positive sample, respectively.

Cell-Text Inter-Modal Contrastive Learning (C-T): We project cells and text into the same embedding space through cell-text contrastive learning. The text encoder employs an unimodal encoding mode. This technique has been widely used in multimodal fields such as image-text and has been proven effective (Radford et al., 2021; Li et al., 2022). Simi-

larly, a momentum encoder is used to expand the batch size. The loss function is as follows:

$$\mathcal{L}_{C-T} = -\frac{1}{2T}\sum_{i=1}^{T}(\log \frac{e^{\mathrm{sim}(\boldsymbol{z_c}^{(i)},\boldsymbol{z_t}^{(i)})/\tau}}{\sum_{j=1}^{T} e^{\mathrm{sim}(\boldsymbol{z_c}^{(i)},\boldsymbol{z_t}^{(j)})/\tau}} + \log \frac{e^{\mathrm{sim}(\boldsymbol{z_t}^{(i)},\boldsymbol{z_c}^{(i)})/\tau}}{\sum_{j=1}^{T} e^{\mathrm{sim}(\boldsymbol{z_t}^{(i)},\boldsymbol{z_c}^{(j)})/\tau}})$$

The symbols in the formula represent similar meanings as above. $\boldsymbol{z_t}^{(i)}$ represents the text embeddings.

Cell-Text Matching (CTM): When computing this loss, the text encoder adopts a multimodal encoding mode, conducting cross-attention calculations with cell embeddings after each self-attention layer, and the final output is used for binary classification to predict whether the cell matches the text or not. This task aims to explore the matching relationship between cells and text with higher resolution, selecting cells and texts that are as similar as possible to the positive examples of cell-text pairs to form negative examples. The loss function is binary cross-entropy:

$$\mathcal{L}_{CTM} = H(y, p_{c,t})$$

where y represents the label indicating whether the cell matches the text.

Overall Pre-training Loss: We optimize the weighted sum of the four losses to simultaneously explore the intrinsic patterns of scRNA-seq data and its associations with text:

$$\min_{\theta} \gamma_1\mathcal{L}_{MGM} + \gamma_2\mathcal{L}_{C-C} + \gamma_3\mathcal{L}_{C-T} + \gamma_4\mathcal{L}_{CTM}$$

where θ represents the model parameters, and γ_i are the weights acting as hyperparameters.

3.4. Downstream Applications

Based on the aligned representation space of cell data and text, and utilizing the cell-text matching module with cross-attention, LangCell can be used for zero-shot cell identity understanding (Fig. 3). Specifically, for a given single cell c and N candidate text descriptions $\{t^{(i)}\}_{i=1}^{N}$, we obtain logits_1 by comparing their cosine distances in the shared embedding space, and obtain logits_2 by comparing the scores given by the cell-text matching module. Both are considered comprehensively for classification according to the weight of α. In practical applications, since logits_2 is slower to compute, we can calculate logits_2 only for the candidates with high logits_1 scores.

The specific calculations are as follows:

$$\mathrm{logits} = \alpha \cdot \mathrm{softmax}(\{\boldsymbol{z_c} \cdot \boldsymbol{z_t}^{(i)}\}_{i=1}^{N}) + (1-\alpha) \cdot \mathrm{softmax}(\{\boldsymbol{g_2}(\boldsymbol{z_c}, t^{(i)})\}_{i=1}^{N})$$

Additionally, a classification or regression head can also be added after the cell encoder for fine-tuning in downstream tasks. This downstream setting is referred to as **LangCell-CE** (**C**ell **E**ncoder) in the subsequent experiments.

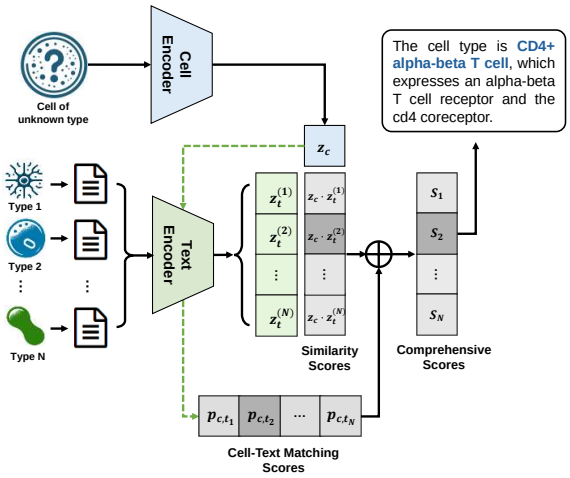

Figure 3: **The application of LangCell in zero-shot cell identity understanding.** LangCell obtains Similarity Scores using the shared embedding space of cell and text data, obtains Cell-Text Matching Scores through the matching module, and considers these comprehensively to obtain the final classification logits. In the figure, the symbol $\oplus$ represents the weighted sum after the Softmax operation.

4. Experiments

4.1. Experiment Settings

4.1.1. Pre-training Details

Dataset Construstion: We established scLibrary, a comprehensive dataset comprising roughly 27.5 million pairs of scRNA-seq data and associated textual descriptions. This dataset was sourced from the CELLxGENE database, where we acquired scRNA-seq data in raw count matrix format and the corresponding meta data. Our selection criteria included all human cells processed using the 10X sequencing protocol. We excluded data that were duplicates, had fewer than 200 expressed genes, had excessive missing metadata, or were used in downstream tasks. scRNA-seq data were processed according to 3.1. Next, we selected 8 cell identities from the meta data that might contain important insights and used these entries to generate multi-view textual descriptions of cells from OBO Foundry. Specifically, the selected cell identities include assay, cell type, developmental stage, tissue information, organ information, disease information, as well as the donor's gender and ethnicity, with each entry's type detailed in the Appendix D.

Pre-training Stages: Our pre-training consists of two stages. In the first stage, we initialize the cell encoder parameters using Geneformer and conduct unimodal training using only the $\mathcal{L}_{MGM}$ and $\mathcal{L}_{C-C}$ loss functions. This is to

obtain a better single-cell representation learning model. In the second stage, we initialize the text encoder parameters using PubMedBERT and engage in multimodal training using all four loss functions. Both stages are trained separately for three epochs each.

Setups: The training process was conducted using the Pytorch framework and the Hugging Face transformers library. We employed the AdamW optimizer, with the learning rate warmed up to 1e-5 over 1000 steps, followed by a linear decay strategy. Weight decay was set to 0.001. More detailed parameter settings can be found in the Appendix C.

4.1.2. Downstream Task Datasets

We collected a set of benchmark datasets to evaluate our model's performance on different downstream tasks. These include human peripheral blood cell datasets (Gayoso et al., 2022; Zheng et al., 2017), human liver datasets (Lin et al., 2020), a human brain cell dataset (Siletti et al., 2023), and a comprehensive human cell dataset (Consortium* et al., 2022). In addition, we propose two novel cell identity understanding tasks: non-small cell lung cancer (NSCLC) subtype classification and cell pathway identification, for which we have constructed high-quality benchmarks. The original data used to build these benchmarks were obtained from CELLxGENE. The former uses disease information from clinical diagnosis in the metadata to annotate two subtypes of NSCLC; the latter employs the irGSEA package to label hallmark pathways from the MSigDB (Liberzon et al., 2011). More details can be found in the Appendix D.2.

4.1.3. Baselines

Our focus is on comparison with single-cell pre-trained language models such as Geneformer, scGPT, xTrimeGene, and scBERT, which have already been proven to outperform traditional methods in a multitude of tasks. Additionally, although BioTranslator did not undergo pre-training for single-cell representation learning, it is an important point of comparison as the first model to consider leveraging textual descriptions to address single-cell issues.

4.2. Zero-shot Cell Identity Understanding

4.2.1. Novel Cell Type Identification

Experimental Setup: Accurate cell type annotation is fundamental for extensive scRNA-seq analyses. However, in practical scenarios, it is often difficult to find enough high-quality labeled data for each cell type to be annotated for fine-tuning. This poses a major challenge to the application of existing single-cell models in actual scenarios, where existing models can only assign all unseen new classes under the "*Novel*" label (Yang et al., 2022a). We refer to this highly challenging zero-shot task as "*Novel Cell Type Identification*", which requires the model to perform cell type annotation in the absence of fine-tuning data. In addition to zero-shot learning, we engage in few-shot learning to benchmark against baselines and evaluate the data efficiency of LangCell. We meticulously selected few-shot settings that are relevant and have practical implications in bioinformatics. This study focuses on two common scenarios encountered in real-world applications:

Zero-and Few-shot Cell Type Annotation: Suitable for scenarios with fewer alternative cell types, commonly seen in annotating small-scale single-cell data from specific tissues or dividing cells of a certain cell type into multiple subtypes. The few-shot approach is configured to use n ($1 \leq n \leq 9$) training samples for each category during fine-tuning. This configuration is designed with the practical consideration that a smaller number of alternative types in real-world settings enables the feasibility of providing very little manually annotated data for each category. In the few-shot task, all baseline models add a linear layer as the classification head. LangCell uses two settings: fine-tuning with the C-T and CTM tasks or using only its cell encoder and a linear classification head (**LangCell-CE**). Our analysis is conducted on two benchmarks of human peripheral blood mononuclear cells, PBMC10K and PBMC3&68K. The evaluation metrics employed are accuracy and macro F_1 score.

Cell-Text Retrieval Suitable for scenarios with many alternative cell types, commonly seen in annotating single cells in complex environments or completely unknown single cells. Given the abundance of cell types, which may encompass subtypes with ambiguous boundaries, we define this task as a cell-text retrieval task. The few-shot method is structured to perform fine-tuning on a limited selection of cell types, followed by testing on a range of unseen cell types. This approach is compared with BioTranslator, the sole existing model with cross-type transfer capabilities. This setting mirrors a real-life situation where some known annotated data is available, but the target data differs from the known dataset. We test on the challenging human comprehensive cell dataset, Tabula Sapiens, which includes 161 cell types, 66 of which are completely new types not included in scLibrary. We use common evaluation metrics for multimodal retrieval tasks, recall@k.

Results: We report the test results of n-shot cell type annotation in Table 1 and Fig. 1, and the results of cell-text retrieval in Fig. 4. The experimental results show that LangCell performs excellently in both zero-shot and few-shot settings for both types of tasks. Specifically, we observe:

Zero- and Few-shot Cell Type Annotation: LangCell shows excellent performance in zero-shot scenarios, surpassing few-shot baselines in most cases. LangCell can also use a few examples to adapt to new tasks quickly, demonstrating its high data efficiency.

Table 1: **Results of zero- and few-shot cell type annotation.** LangCell is the only single-cell PLM that can perform zero-shot. All other models need to add classification headers and fine-tune. In most cases, LangCell's zero-shot performance is better than the few-shot results of existing models. Acc: accuracy (%). F_1: macro F_1 score (%).

Dataset	Model	0-shot		1-shot		3-shot		5-shot		7-shot		9-shot	
		Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
PBMC 10k	scBERT	✗	✗	31.4	9.4	60.6	41.4	78.0	58.2	59.0	54.6	81.9	62.6
	scGPT	✗	✗	41.9	34.0	43.3	41.1	81.1	66.3	82.0	68.3	86.7	75.8
	Geneformer	✗	✗	54.0	42.2	70.3	46.7	81.0	63.9	80.9	71.2	88.0	78.6
	LangCell-CE	✗	✗	**88.7**	75.2	92.2	86.1	93.0	88.7	93.6	89.1	94.4	90.7
	LangCell	**86.5**	**89.6**	88.1	**87.5**	**95.1**	**94.7**	**96.0**	**94.8**	**96.3**	**95.3**	**96.8**	**95.2**
PBMC 3&68k	scBERT	✗	✗	19.9	13.8	36.5	39.4	48.5	43.0	48.3	40.9	47.6	51.7
	scGPT	✗	✗	17.7	21.1	45.3	44.3	52.0	61.6	79.9	76.9	85.7	76.9
	Geneformer	✗	✗	21.1	24.7	55.2	49.2	59.3	69.1	81.5	74.8	83.3	74.1
	LangCell-CE	✗	✗	86.2	79.7	85.8	85.8	87.2	84.1	89.1	85.9	86.7	86.0
	LangCell	**83.9**	**82.6**	**89.7**	**87.1**	**89.9**	**87.8**	**90.3**	**87.7**	**92.1**	**87.5**	**92.4**	**88.5**

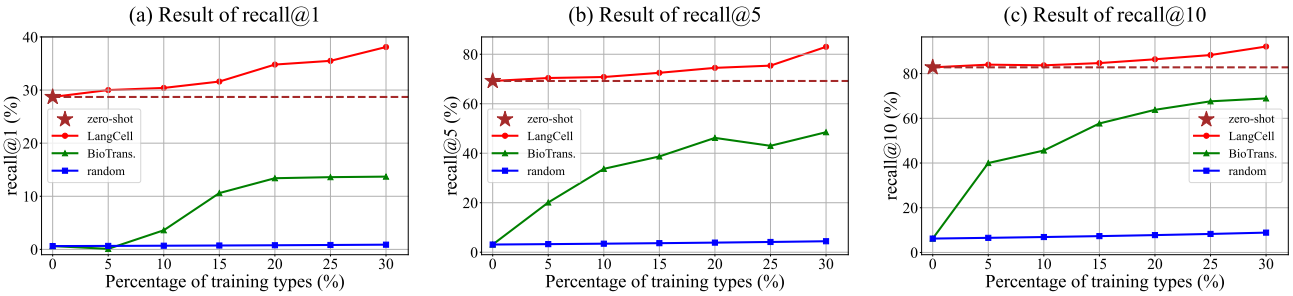

Figure 4: **Results of cell-text retrieval.** Zero-shot LangCell surpasses BioTranslator trained on up to 30% of the 161 types.

Cell-Text Retrieval: The zero-shot performance of LangCell surpasses BioTranslator, which at most uses 48 (30% of 161) cell types, for training. This confirms LangCell's good performance in such challenging application scenarios.

4.2.2. NSCLC Subtype Classification

ScRNA-seq technology plays a significant role in the study of malignant tumor. However, it is difficult to analyze malignant cells due to the scarcity of data and their characteristics of high mutational burdens. Lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) are the two most common subtypes of non-small cell lung cancer (NSCLC). We test LangCell on 2,658 malignant cells from patients with these two lung cancer subtypes, to assess its effectiveness in identifying disease-related cell identities.

Since all cell type labels are "*Malignant*", we used descriptions of these two diseases to construct the texts. As shown in Figure 5, LangCell aligns single-cell and disease texts well in the latent space. Its zero-shot classification surpasses Geneformer (fine-tuned with 10-shot learning) by about 20% in both accuracy and macro F_1-score. (Table 2) This experiment demonstrates LangCell's strong capability in understanding disease-related cell identities and its effective performance in the analysis of single cells with high mutational burdens, such as malignant cells.

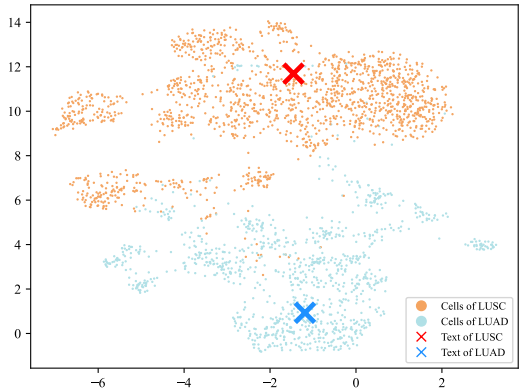

Figure 5: UMAP plot of embeddings for scRNA-seq data and descriptions of two NSCLC subtypes.

4.2.3. Single-cell batch integration

Single-cell batch integration holds significant importance in biomedical research. It plays a crucial role in mitigating

Table 2: Results of NSCLC subtype classification (%).

Model	n-shot	Acc	F_1
Geneformer	1	46.7	43.9
	10	73.1	73.1
LangCell	0	**93.5**	**93.2**

Table 3: Results of cell batch integration (%) from scratch.

Dataset	Model	Avg_{bio}	ASW_{batch}	S_{final}
PBMC10K	scVI	70.0	97.6	81.0
	scBERT	18.1	95.0	48.9
	scGPT*	72.3	91.9	80.2
	Geneformer	79.3	92.8	84.7
	LangCell	**80.8**	**97.9**	**87.6**
Perirhinal Cortex	scVI	84.9	89.6	86.8
	scBERT	15.1	92.9	46.2
	scGPT*	88.9	88.4	88.7
	Geneformer	85.5	91.8	88.0
	LangCell	**95.2**	**95.6**	**95.4**

Table 4: Results of cell type annotation (%). *: Since xTrimoGene did not release the checkpoint, we can only obtain their reported Zheng68k result.

Dataset	PBMC10K		LiverCross		Zheng68K	
	Acc	F_1	Acc	F_1	Acc	F_1
scBERT	97.5	90.5	37.3	12.2	77.9	68.8
scGPT	96.5	94.1	48.1	24.1	84.6	75.2
xTrimoGene*	-	-	-	-	-	73.5
Geneformer	97.8	95.7	46.7	24.0	83.9	74.4
LangCell-CE	**98.3**	**96.9**	**50.4**	**26.0**	**85.4**	**76.9**

Table 5: Result of pathway identification (%). The metrics are detailed in Appendix E.

Model	avg-AUROC	avg-AUPRC	flatten-AUROC	flatten-AUPRC
Geneformer	82.8	23.9	86.6	27.3
LangCell-CE	**89.3**	**31.4**	**89.9**	**35.4**

batch effects from different experimental data, scaling up data analysis, and fostering a comprehensive understanding of cell diversity and functionality within biological systems. This task requires the model to correctly distinguish whether the expression differences between cells arise from meaningless batch effects or meaningful biological information, demanding a strong understanding of cell identity information inherent in scRNA-seq data.

We evaluated the performance of LangCell in cell batch integration on the PBMC10K and Perirhinal Cortex, comparing it with the classical model in the field, scVI, as well as several single-cell PLMs. To comprehensively assess model performance, we used the evaluation metrics Avg_{bio}, ASW_{batch}, and S_{final} proposed in (Luecken et al., 2022). These metrics respectively assess the model's capability of biological integration, batch effect removal, and a comprehensive evaluation of both. Detailed calculation can be found in the Appendix E. The experimental results in Table 3 indicate that LangCell surpasses the existing optimal models in all three metrics. This demonstrates LangCell's profound insight into transcriptomic data, its ability to accurately preserve important biological information, and its effectiveness in correcting irrelevant batch effects.

4.3. Cell Representation Learning of LangCell-CE

4.3.1. Cell Type Annotation (fine-tune)

We also evaluate the representation capabilities of LangCell-CE on the classic task of cell type annotation. The results in Table 4 demonstrate that our model achieves SOTA performance on all three datasets. This demonstrates that LangCell successfully injects unstructured knowledge into the cell encoder, enhancing the understanding of scRNA-seq data. In addition, experiments on LiverCross demonstrate the effectiveness of LangCell in cross-dataset tasks.

4.3.2. Pathway Identification

In the pre-training process, we injected knowledge of "cell types" into LangCell-CE, thereby naturally deepening its understanding of this cell identity. We cannot yet determine whether LangCell-CE's outstanding performance is due to its comprehensively improved ability to learn cell representations, or simply due to its insights into this specific task. To verify this, we explored a new cell identity not covered in pre-training, cell pathways, and designed a challenging representation learning task around it. For each single cell, the model was tasked with identifying multiple pathways from a selection of 41 important pathways, essentially constituting a multi-label binary classification problem with 41 independent labels. Considering the data imbalance caused by the expression of each pathway in only a few cells, focal loss (Lin et al., 2017) was used during fine-tuning. As shown in Table E, LangCell significantly outperforms the existing state-of-the-art model Geneformer on AUROC and AUPRC on this challenging task.

4.4. Ablation Study

Stage Division: The primary motivation for dividing the training into two stages is to reduce computational costs. Our preliminary experiments showed that the first stage of training significantly accelerates the convergence of the loss function in the second phase (Fig. 6), thereby reducing the number of epochs needed to reach convergence. Considering that the first phase only requires computing $\mathcal{L}_{MGM}$ and $\mathcal{L}_{C-C}$, whose computational costs are much lower than $\mathcal{L}_{C-T}$ and $\mathcal{L}_{CTM}$, the two-stage pre-training significantly reduces the overall computational resources needed.

Table 6: Ablation study of pre-training tasks in LangCell. *LangCell-1*: model at the end of the first stage of pre-training. *w/o CTM*: without CTM module.

Models	$\mathcal{L}_{MGM}$	$\mathcal{L}_{C-C}$	$\mathcal{L}_{C-T}$	$\mathcal{L}_{CTM}$	zero-shot		fine-tune	
					Acc	F_1	Acc	F_1
Geneformer	✔	✗	✗	✗	-	-	76.1	64.7
LangCell-1	✔	✔	✗	✗	-	-	77.0	65.8
LangCell $_{\text{w/o CTM}}$	✔	✔	✔	✗	84.8	85.9	-	-
LangCell	✔	✔	✔	✔	**85.2**	**86.1**	**78.0**	**66.6**

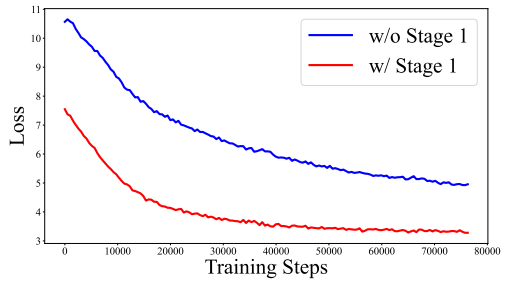

Figure 6: The impact of whether to conduct the first stage on the convergence speed of LangCell pre-training.

Initialization: If we don't use Geneformer and PubMedBERT to initialize the model parameters and instead start training from scratch, the convergence of the model will be extremely slow in the early stages due to the absence of high-quality representations of scRNA-seq data and text. Specifically, the loss demonstrates negligible decline throughout the first 20,000 steps, and the huge computational costs prevent further study of training from scratch.

Ablation Study of Pre-training Tasks: We explored the influence of various pre-training tasks on the model's performance in downstream applications (Table 6). Compared to using only $\mathcal{L}_{MGM}$, the incorporation of $\mathcal{L}_{C-C}$ improves the model's cell representation learning capability, which is consistent with findings in NLP (Gao et al., 2021) and CV (Park et al., 2023) fields. The integration of textual data further enhances the performance of the cellular encoder, which we attribute to the injection of unstructured knowledge in the text. The results on zero-shot and few-shot retrieval tasks demonstrate the importance of considering both similarity scores and matching scores comprehensively, with performance surpassing that of considering either alone.

Discussion on the setting of α: α mentioned in 3.4 is an adjustable hyperparameter during downstream tasks. Users of LangCell can use a small validation set to select the optimal α for a specific task. If there is no labeled validation set available, we recommend setting 0.2 as the default value for α, which is near-optimal in most cases. In the zero-shot experiments of this paper, to simulate an application scenario with no labeled data at all, we did not manually adjust α and instead used the default value 0.2. The results in Table 7 present the impact of α on several zero-shot tasks.

Table 7: The impact of alpha setting on the model's zero-shot ability.

α	PBMC10K		PBMC3&68K	
	Accuracy	F_1	Accuracy	F_1
0	56.83	25.09	65.32	31.33
0.01	86.74	67.07	**88.54**	81.45
0.05	**92.40**	80.08	87.44	**83.07**
0.1	90.98	85.18	85.13	82.47
0.2	86.54	**89.61**	83.94	82.64
0.3	86.29	89.47	84.25	82.16
0.5	85.98	89.49	83.91	82.39
0.7	85.84	89.46	83.74	82.42
0.9	85.77	89.43	83.70	82.43

5. Conclusion and Limitation

In this work, we present LangCell, the Language-Cell pre-training framework, offering a unified representation of single-cell data and natural language that transcends the need for task-specific fine-tuning. By integrating these modalities, LangCell intuitively grasps the relationship between cellular data and textual identities, enhancing its cell representation learning capabilities. Our experiments across various biological tasks confirm LangCell's superior performance over existing models, particularly in zero-shot and few-shot scenarios. This framework sets a new standard for the field, enabling more accurate and efficient analysis of single-cell transcriptomics data.

Currently, LangCell still has some limitations. For example, the pre-training texts are all from the OBO Foundry, which limits the diversity to a certain extent. LangCell also cannot yet analyze other single-cell omics such as scATAC-seq, and it does not include cell/text generation functions. In the future, we will focus on improving these aspects.

Code Availability

LangCell will soon be added to the OpenBioMed toolkit: https://github.com/PharMolix/OpenBioMed.

Code is available at: https://github.com/PharMolix/LangCell.

Acknowledgements

This research is supported by the National Key R&D Program of China (No. 2022YFF1203002) and PharMolix Inc.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abdolhosseini, F., Azarkhalili, B., Maazallahi, A., Kamal, A., Motahari, S. A., Sharifi-Zarchi, A., and Chitsaz, H. Cell identity codes: understanding cell identity from gene expression profiles using deep neural networks. *Scientific reports*, 9(1):2342, 2019.

Alessandri, L., Cordero, F., Beccuti, M., Licheri, N., Arigoni, M., Olivero, M., Renzo, F. D., Sapino, A., and Calogero, R. A. Sparsely-connected autoencoder (sca) for single cell rnaseq data mining. *NPJ Systems Biology and Applications*, 7, 2020.

Biology, C. S.-C., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, pp. 2023–10, 2023.

Chen, Y. T. and Zou, J. Genept: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, 2023.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., and Weller, A. Rethinking attention with performers, 2022.

Connell, W., Khan, U., and Keiser, M. J. A single-cell gene expression language model. *arxiv:2210.14330 q-bio.QM,cs.AI,q-bio.MN]*, 10 2022. [Online; accessed 2023-05-04].

Consortium*, T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., Hemenez, M., Ponnusamy, R., Salehi, A., Sanagavarapu, B. A., Spallino, E., Aaron, K. A., Concepcion, W., Gardner, J. M., Kelly, B., Neidlinger, N., Wang, Z., Crasta, S., Kolluru, S., Morri, M., Tan, S. Y., Travaglini, K. J., Xu, C., Alcántara-Hernández, M., Almanzar, N., Antony, J., Beyersdorf, B., Burhan, D., Calcuttawala, K., Carter, M. M., Chan, C. K. F., Chang, C. A., Chang, S., Colville, A., Culver, R. N., Cvijović, I., D'Amato, G., Ezran, C., Galdos, F. X., Gillich, A., Goodyer, W. R., Hang, Y., Hayashi, A., Houshdaran, S., Huang, X., Irwin, J. C., Jang, S., Juanico, J. V., Kershner, A. M., Kim, S., Kiss, B., Kong, W., Kumar, M. E., Kuo, A. H., Li, B., Loeb, G. B., Lu, W.-J., Mantri, S., Markovic, M., McAlpine, P. L., de Morree, A., Mrouj, K., Mukherjee, S., Muser, T., Neuhöfer, P., Nguyen, T. D., Perez, K., Puluca, N., Qi, Z., Rao, P., Raquer-McKay, H., Schaum, N., Scott, B., Seddighzadeh, B., Segal, J., Sen, S., Sikandar, S., Spencer, S. P., Steffes, L. C., Subramaniam, V. R., Swarup, A., Swift, M., Van Treuren, W., Trimm, E., Veizades, S., Vijayakumar, S., Vo, K. C., Vorperian, S. K., Wang, W., Weinstein, H. N. W., Winkler, J., Wu, T. T. H., Xie, J., Yung, A. R., Zhang, Y., Detweiler, A. M., Mekonen, H., Neff, N. F., Sit, R. V., Tan, M., Yan, J., Bean, G. R., Charu, V., Forgó, E., Martin, B. A., Ozawa, M. G., Silva, O., Toland, A., Vemuri, V. N. P., Afik, S., Awayan, K., Botvinnik, O. B., Byrne, A., Chen, M., Dehghannasiri, R., Gayoso, A., Granados, A. A., Li, Q., Mahmoudabadi, G., McGeever, A., Olivieri, J. E., Park, M., Ravikumar, N., Stanley, G., Tan, W., Tarashansky, A. J., Vanheusden, R., Wang, P., Wang, S., Xing, G., Dethlefsen, L., Ezran, C., Gillich, A., Hang, Y., Ho, P.-Y., Irwin, J. C., Jang, S., Leylek, R., Liu, S., Maltzman, J. S., Metzger, R. J., Phansalkar, R., Sasagawa, K., Sinha, R., Song, H., Swarup, A., Trimm, E., Veizades, S., Wang, B., Beachy, P. A., Clarke, M. F., Giudice, L. C., Huang, F. W., Huang, K. C., Idoyaga, J., Kim, S. K., Kuo, C. S., Nguyen, P., Rando, T. A., Red-Horse, K., Reiter, J., Relman, D. A., Sonnenburg, J. L., Wu, A., Wu, S. M., and Wyss-Coray, T. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (New York, N.Y.)*, 376(6594):eabl4896, 5 2022. ISSN 0036-8075. doi: 10.1126/science.abl4896. [Online; accessed 2024-01-31].

Cui, H., Wang, C. X., Maan, H., Pang, K., Luo, F., and Wang, B. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:258464426.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Edwards, C. N., Lai, T., Ros, K., Honke, G., and Ji, H. Translation between molecules and natural language. *ArXiv*, abs/2204.11817, 2022. URL https://api.semanticscholar.org/CorpusID:248376906.

F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2: 559–572, 1901.

Gao, T., Yao, X., and Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.-main.552. URL https://aclanthology.org/2021.emnlp-main.552.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M.,

Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., Svensson, V., da Veiga Beltrame, E., Kleshchevnikov, V., Talavera-López, C., Pachter, L., Theis, F. J., Streets, A., Jordan, M. I., Regier, J., and Yosef, N. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2):163–166, 2 2022. ISSN 1087-0156. doi: 10.1038/s41587-021-01206-w. [Online; accessed 2024-01-31].

Gong, J., Hao, M., Cheng, X., Zeng, X., Liu, C., Ma, J., Zhang, X., Wang, T., and Song, L. xtrimogene: An efficient and scalable representation learner for single-cell rna-seq data. *arXiv preprint arXiv:2311.15156*, 2023.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Guo, H. and Li, J. scsorter: assigning cells to known cell types according to marker genes. *Genome Biology*, 22, 2021.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Song, L. T., and Zhang, X. Large scale foundation model on single-cell transcriptomics. *bioRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:259025739.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

Huang, Q., Liu, Y., Du, Y., and Garmire, L. X. Evaluation of cell type annotation r packages on single-cell rna-seq data. *Genomics, Proteomics & Bioinformatics*, 19:267 – 281, 2020.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M. M., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver io: A general architecture for structured inputs & outputs, 2022.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: Pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021. doi: 10.1093/bioinformatics/btab083.

Levine, D., Lévy, S., Rizvi, S. A., Pallikkavaliyaveetil, N., Chen, X., Zhang, D., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv*, pp. 2023–09, 2023.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.-main.733. URL https://aclanthology.org/2020.emnlp-main.733.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:246411402.

Li, J., Li, D., Savarese, S., and Hoi, S. C. H. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023a. URL https://api.semanticscholar.org/CorpusID:256390509.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arxiv:2110.05208 [cs.CV]*, 10 2021. [Online; accessed 2024-01-31].

Li, Y., Lin, Y., Hu, P., Peng, D., Luo, H., and Peng, X. Single-cell rna-seq debiased clustering via batch effect disentanglement. *IEEE Transactions on Neural Networks and Learning Systems*, 2023b.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12): 1739–1740, 2011.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Lin, Y., Cao, Y., Kim, H. J., Salim, A., Speed, T. P., Lin, D. M., Yang, P., and Yang, J. Y. H. scclassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular systems biology*, 16(6):e9389, 6 2020. ISSN 1744-4292. doi: 10.15252/msb.20199389. [Online; accessed 2024-02-01].

Linderman, G. C., Zhao, J., Roulis, M., Bielecki, P., Flavell, R. A., Nadler, B., and Kluger, Y. Zero-preserving imputation of single-cell rna-seq data. *Nature Communications*, 13, 2022.

Luecken, M. D. and Theis, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F. J. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 1 2022. ISSN 1548-7091. doi: 10.1038/s41592-021-01336-8. [Online; accessed 2024-02-01].

Megill, C., Martin, B., Weaver, C., Bell, S., Prins, L., Badajoz, S., McCandless, B., Pisco, A. O., Kinsella, M., Griffin, F., Kiggins, J., Haliburton, G., Mani, A., Weiden, M., Dunitz, M., Lombardo, M., Huang, T., Smith, T., Chambers, S., Freeman, J., Cool, J., and Carr, A. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv*, 2021. doi: 10.1101/2021.04.05.438318. URL https://www.biorxiv.org/content/early/2021/04/06/2021.04.05.438318.

Mo, S., Fu, X., Hong, C., Chen, Y., Zheng, Y., Tang, X., Shen, Z., Xing, E. P., and Lan, Y. Multi-modal self-supervised pre-training for regulatory genome across cell types, 2021.

Morris, S. A. The evolving concept of cell identity in the single cell era. *Development*, 146(12):dev169748, 2019.

Park, N., Kim, W., Heo, B., Kim, T., and Yun, S. What do self-supervised vision transformers learn? *arxiv:2305.00729 [cs.CV,cs.AI,cs.LG]*, 5 2023. [Online; accessed 2024-01-31].

Pasquini, G., Arias, J. E. R., Schäfer, P., and Busskamp, V. Automated methods for cell type annotation on scrna-seq data. *Computational and Structural Biotechnology Journal*, 19:961 – 969, 2021.

Qiu, X., Wu, H., and Hu, R. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC bioinformatics*, 14: 1–10, 2013.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/d19-1410.

Seal, R. L., Braschi, B., Gray, K. A., Jones, T. E. M., Tweedie, S., Haim-Vilmovsky, L., and Bruford, E. A. Genenames.org: the hgnc resources in 2023. *Nucleic Acids Research*, 51:D1003 – D1009, 2022.

Shen, H. T. Principal component analysis. In *Encyclopedia of Database Systems*, 2009.

Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K. W., Ding, S.-L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., Clark, M., Dee, N., Gloe, J., Hirschstein, D., Shapovalova, N. V., Keene, C. D., Nyhus, J., Tung, H., Yanny, A. M., Arenas, E., Lein, E. S., and Linnarsson, S. Transcriptomic diversity of cell types across the adult human brain. *Science (New York, N.Y.)*, 382(6667):eadd7046, 10 2023. ISSN 0036-8075. doi: 10.1126/science.add7046. [Online; accessed 2024-01-31].

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18: 2789 – 2798, 2018.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

Svensson, V. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38:147–150, 2019.

Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. Autoimpute: Autoencoder based imputation of single-cell rna-seq data. *Scientific Reports*, 8, 2018.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Sayed, Z. R. A., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. URL https://api.semanticscholar.org/CorpusID:259002047.

Tran, B., Tran, D., Nguyen, H., Ro, S., and Nguyen, T. sc-can: single-cell clustering using autoencoder and network fusion. *Scientific Reports*, 12, 2022.

Tran, D., Nguyen, H., Tran, B., la Vecchia, C., Luu, H. N., and Nguyen, T. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature Communications*, 12, 2019.

Tsuyuzaki, K., Sato, H., Sato, K., and Nikaido, I. Benchmarking principal component analysis for large-scale single-cell rna-sequencing. *Genome Biology*, 21, 2019.

van der Maaten, L. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 15:3221–3245, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

Xu, H., Woicik, A., Poon, H., Altman, R. B., and Wang, S. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14, 2023a. URL https://api.semanticscholar.org/CorpusID:256701737.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multimodality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, 2023b. URL https://api.semanticscholar.org/CorpusID:256390530.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. 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, 2022a. doi: 10.1038/s42256-022-00534-z.

Yang, M., Yang, Y., Xie, C., Ni, M., Liu, J., Yang, H., Mu, F., and Wang, J. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nature Machine Intelligence*, 4:696 – 709, 2022b.

Zeng, Z., Yao, Y., Liu, Z., and Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13, 2022. URL https://api.semanticscholar.org/CorpusID:246815222.

Zhai, Y., Chen, L., and Deng, M. Realistic cell type annotation and discovery for single-cell rna-seq data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4967–4974, 2023.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65 (4):631–643, 2017.

Appendix

A. More Experimental Results

A.1. Cell Batch Integration & Novel Cell Type Identification

The complete experimental results for cell batch integration are shown in Table A.1.1.

Table A.1.1: Results of cell batch integration (%) from scratch. * stands for the results from scGPT.

Dataset	Model	NMI	ARI	ASW_{cell}	Avg_{bio}	ASW_{batch}	S_{final}
PBMC10K	scVI	80.8	71.1	58.1	70.0	97.6	81.0
	scBERT	5.3	3.4	45.5	18.1	95.0	48.9
	scGPT*	73.8	79.3	63.9	72.3	91.9	80.2
	Geneformer	82.5	84.6	70.9	79.3	92.8	84.7
	LangCell	**84.5**	**85.4**	**72.4**	**80.8**	**97.9**	**87.6**
Perirhinal Cortex	scVI	95.0	95.7	63.9	84.9	89.6	86.8
	scBERT	3.1	2.7	39.6	15.1	92.9	46.2
	scGPT*	88.6	89.5	88.6	88.9	88.4	88.7
	Geneformer	89.0	81.3	86.3	85.5	91.8	88.0
	LangCell	**97.2**	**98.3**	**90.2**	**95.2**	**95.6**	**95.4**

We perform a visual analysis of the PBMC10K dataset to intuitively observe LangCell's zero-shot capability for cell batch integration (Fig. A.1.1, left and right). We also visualize the encoding results of the current best models, scGPT and Geneformer (Fig. A.1.2, Fig. A.1.3). It can be observed that all three models excel at eliminating batch effects, but LangCell, during encoding, can directly focus on the identity information of cells, with cells of the same type clustering together in the feature space.

Moreover, LangCell can effectively complete novel cell type identification. Comparing the left and middle images of Fig. A.1.1, it can be seen that LangCell can correctly annotate most cells without any fine-tuning.

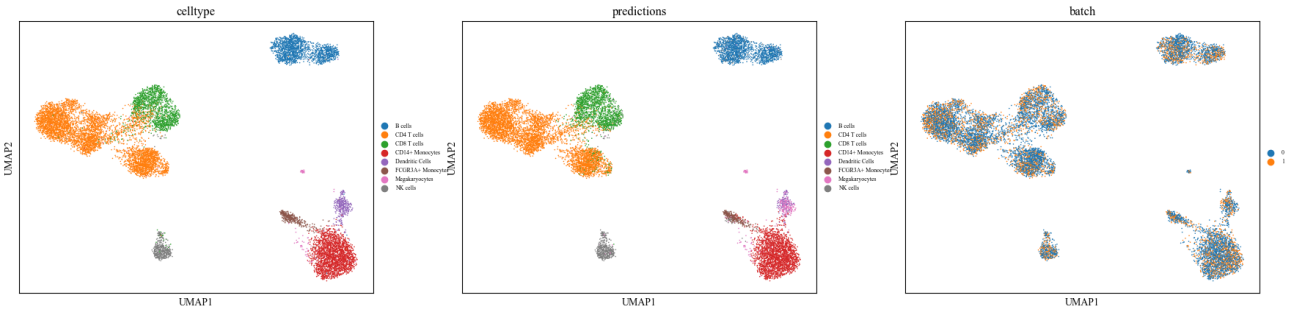

Figure A.1.1: UMAP plot of embeddings for scRNA-seq data of LangCell in the zero-shot scenario. Three scatter plots are colored by actual cell type labels, predicted cell type labels, and batch information, respectively. It is evident that the cell embeddings generated directly by LangCell, without any fine-tuning, possess desirable properties: they cluster by cell type and eliminate batch effects. By comparing the left and middle plots, one can intuitively observe that LangCell is capable of reliably annotating cell types in a zero-shot scenario.

A.2. Cell-Text Retrieval

We plotted heatmaps of top retrieval results for zero-shot LangCell and BioTranslator trained on 10% types. It is clear to see that LangCell's result plot has a sharper diagonal line, signaling a significantly higher retrieval ability for new cell types than BioTranslator (Fig. A.2.1, Fig. A.2.2).

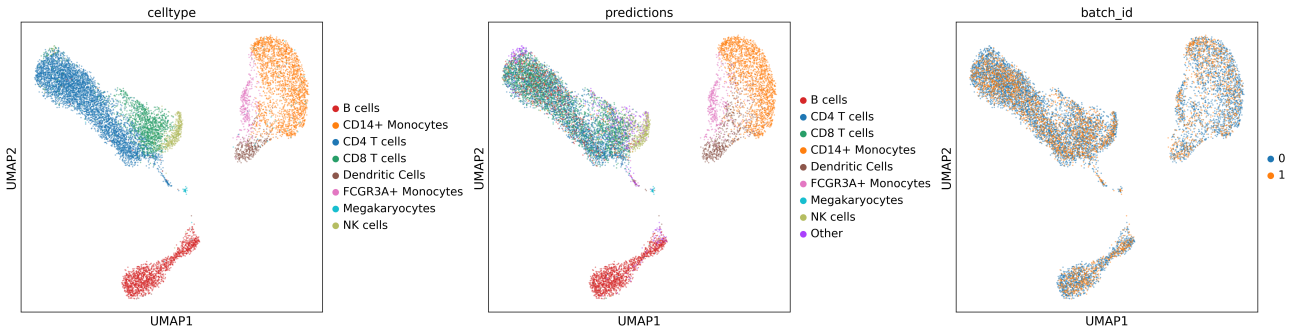

Figure A.1.2: UMAP plot of embeddings for scRNA-seq data of scGPT. In the middle is the predicted result of fine-tuned scGPT.

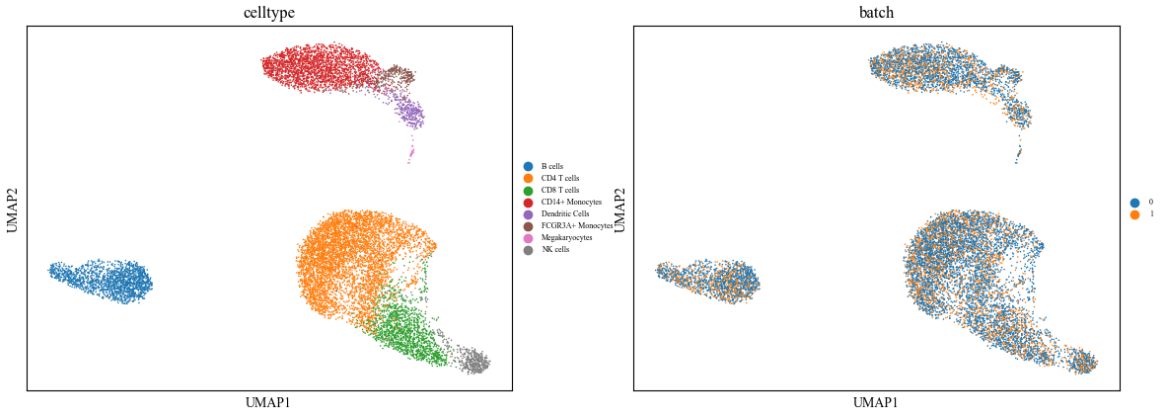

Figure A.1.3: UMAP plot of embeddings for scRNA-seq data of Geneformer.

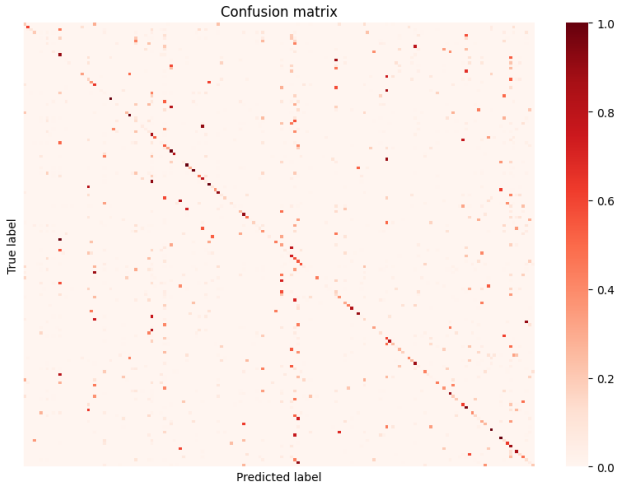

Figure A.2.1: Heatmap of the top-ranked results from LangCell's zero-shot retrieval on Tabula Sapiens.

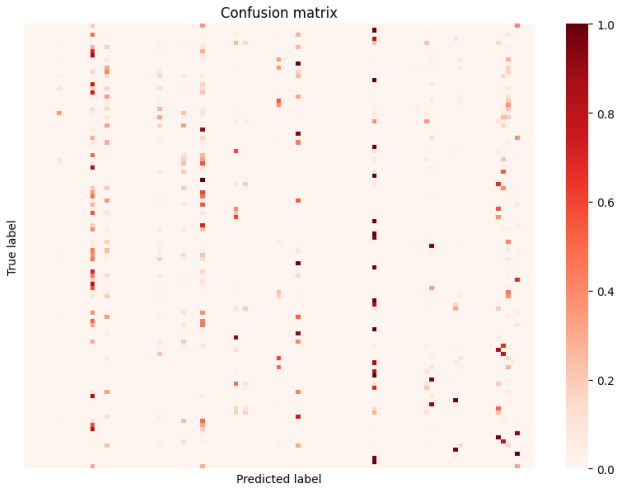

Figure A.2.2: Heatmap of the top-ranked results from BioTranslator's retrieval on Tabula Sapiens. Trained on 10% types and tested on 90% types.

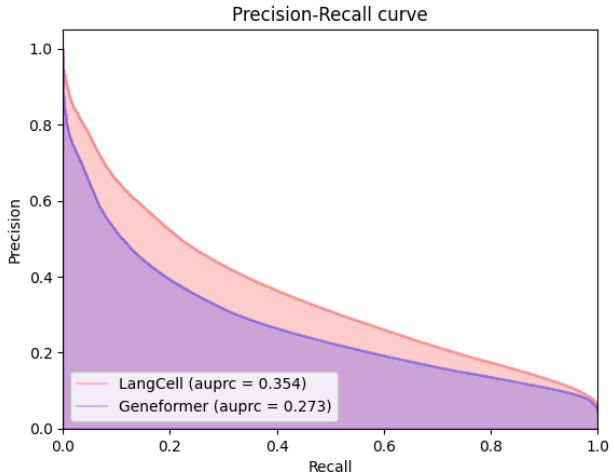

Figure A.3.1: flatten-PRC of pathway identification.

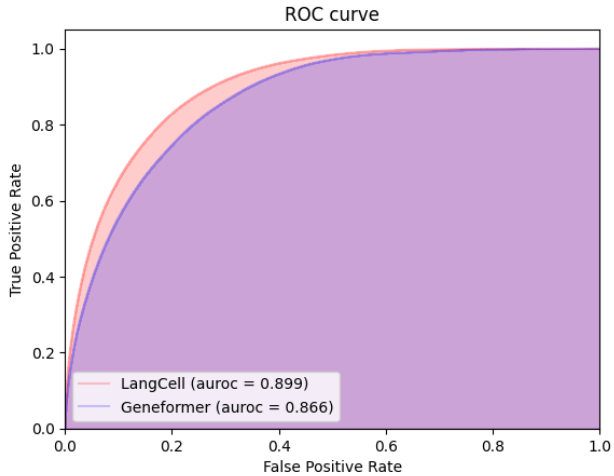

Figure A.3.2: flatten-ROC of pathway identification.

Table A.4.1: Ablation study of pre-training tasks in LangCell. *LangCell-1*: model at the end of the first stage of pre-training. *w/o CTM*: without CTM module.

Models	Zero-shot						Fine-tune							
	PBMC10K		PBMC3&68K		Avg		PBMC10K		LiverCross		Zheng68k		Avg	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
Geneformer	-	-	-	-	-	-	97.8	95.7	46.7	24.0	83.9	74.4	76.1	64.7
LangCell-1	-	-	-	-	-	-	98.1	96.6	48.5	25.4	84.4	75.4	77.0	65.8
LangCell $_{\text{w/o CTM}}$	85.8	89.4	83.7	82.4	84.8	85.9	-	-	-	-	-	-	-	-
LangCell	**86.5**	**89.6**	**83.9**	**82.6**	**85.2**	**86.1**	**98.3**	**96.9**	**50.4**	**26.0**	**85.4**	**76.9**	**78.0**	**66.6**

A.3. Pathway Identification

Fig. A.3.1 and Fig. A.3.2 visually show the performance of fine-tuned LangCell and Geneformer in the pathway identification task.

A.4. Ablation Study of Pre-training Tasks

The complete experimental results for cell batch integration are shown in Table A.4.1.

A.5. Retrieval for novel cell types not covered by scLibrary

LangCell has demonstrated excellent performance in cell-text retrieval tasks. However, many cell types in the test dataset Tabula Sapiens are covered by scLibrary. To confirm that LangCell's outstanding performance is not solely due to encountering the same cell types in scLibrary, we re-calculate the experimental results in Figure 4 for 95 cell types covered by scLibrary and 66 cell types not covered by scLibrary. We present the experimental effects of LangCell in a zero-shot scenario and compare them with the results of BioTranslator under the setting of 30% training classes (Table A.5.1). Each set of experiments uses all 161 types as alternatives. The experimental results show that for new types of cells present in Tabula Sapiens that are not included in scLibrary, LangCell also exhibits outstanding classification performance. This demonstrates LangCell's strong transferability to entirely new cell types.

A.6. Robustness to "Dropout Zeros"

In practical application scenarios, scRNA-seq data often contains "dropout zeros" noise, which means that low gene expressions may not be captured during sequencing (Silverman et al., 2018). The model's resistance to such noise significantly influences its practicality. In fact, "dropout zeros" can be regarded as random noise introduced by sequencing technology. Works such as Geneformer and scGPT have demonstrated that single-cell language models can understand the contextual relationships of gene expressions during large-scale pre-training, thereby possessing resistance to noise in scRNA-

Table A.5.1: The cell-text retrieval results of the cell types covered and not covered by scLibrary.

Model	Data Selection	Classes of Cells	Number of Cells	Recall@1 (Accuracy)
BioTranslator (baseline)	30% classes for training and 70% for test	113	212k	13.71
LangCell	All	161	456k	28.65
LangCell	Covered by scLibrary	95	429k	28.77
LangCell	Not covered by scLibrary	66	27k	26.74

Table A.6.1: The impact of "dropout zeros" on LangCell experimental results.

Dropout Probability (%)	CosineSimilarity (%)	Accuracy(%)	F_1 (%)
0	100	86.54	89.61
1	99.70	86.66	89.59
5	98.72	86.37	89.71
10	97.45	88.14	89.50

seq data. To verify LangCell's resistance to "dropout zeros", we have added the following experiment on the PBMC10K dataset. For genes with expression levels in the bottom 30% of each cell (excluding genes with an expression level of 0), we reduced their expression levels to 0 with a certain probability to simulate the "dropout zeros" noise. Subsequently, we observed the perturbation of LangCell-generated cell embeddings under different probabilities of dropout zeros. We also tested the impact of different probabilities of dropout zeros on the downstream task effects in the zero-shot cell type annotation task, reflecting LangCell's performance on lower-quality downstream task data. The experiments were conducted three times for each dropout probability, and the average results are shown in Table A.6.1.

The experimental results indicate that applying dropout perturbation does not cause significant shifts in the cell embeddings generated by LangCell. This demonstrates that LangCell has excellent noise resistance capabilities against the "dropout zeros" phenomenon specific to scRNA-seq data. Moreover, the experimental results of zero-shot cell type annotation also show that dropout zeros events in downstream task data do not significantly degrade LangCell's performance. This proves the robustness of LangCell on data of lower quality.

B. More Discussion about Complexity Analysis and Inference Speed

B.1. Complexity analysis of text encoder

Let the text length be N, the gene sequence length of a cell be M, and consider the vector dimension as a constant. When the text encoder adopts a single-modality mode ($\boldsymbol{g_1}$), the bottleneck for time and space complexity lies in the self-attention computation of the text, with a complexity of $O(N^2)$ (Vaswani et al., 2023). When the text encoder adopts a multi-modality mode ($\boldsymbol{g_2}$), the time and space complexity is determined by both the self-attention computation of the text and the cross-attention computation between the image and text, with a complexity of $O(N^2 + MN)$.

Furthermore, during the zero-shot inference process, let the total number of categories be P and the total number of cells be Q. Typically, $P \ll Q$. The time complexity for calculating the embeddings of all categories in single-modality mode is $O(PN^2)$. For each cell, the time complexity for computing the embeddings is $O(M^2)$, and the time complexity for calculating the match scores with the top k categories using $\boldsymbol{g_2}$ is $O(kN^2 + kMN)$. Therefore, the total time complexity of the inference process is $O(PN^2 + Q(M^2 + kN^2 + kMN))$.

B.2. Discussion about inference speed

The total time complexity of the reasoning process is $O(PN^2 + Q(M^2 + kN^2 + kMN))$. Put simply, the total number of forward passes required during the inference process is $P + Q + kQ$, where $0 \leq k \leq P$.

The number of forward passes for LangCell-CE, Geneformer and other models that require fine-tuning during inference is Q. Considering $P \ll Q$, the main factor affecting inference time is the choice of k. In scenarios where inference speed is highly required, k can be set to 0, thus achieving fast inference similar to LangCell-CE; in scenarios with a larger number of categories or where inference speed is not a high requirement, a larger k can be chosen or even k can be set to P for more accurate inference.

For the experimental results reported in the paper, except for the cell-text retrieval experiment where $k = 20$ is taken, all

Table B.2.1: The impact of k on the model's inference performance and time. ($P = 8$)

k	Time (s)	Zero-Shot Accuracy (%)	Zero-Shot F_1 (%)
0	**186.62**	86.76	89.37
2	514.51	86.28	89.23
4	841.18	86.04	89.49
6	1162.49	86.53	89.55
8	1537.91	**86.54**	**89.61**
For comparison, the inference time of LangCell-CE or Geneformer: 188.23s.			

Table B.2.2: The complete performance of LangCell at $k = 0$ (i.e., w/o CTM).

Dataset	Model	0-shot		1-shot		3-shot		5-shot		7-shot		9-shot	
		Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
PBMC 10K	Geneformer	✗	✗	54.0	42.2	70.3	46.7	81.0	63.9	80.9	71.2	88.0	78.6
	LangCell $_{\mathrm{w/oCTM}}$	85.8	89.4	86.6	**90.6**	89.3	91.6	92.2	93.0	93.0	93.3	96.2	94.2
	LangCell	**86.5**	**89.6**	**88.1**	87.5	**95.1**	**94.7**	**96.0**	**94.8**	**96.3**	**95.3**	**96.8**	**95.2**
PBMC 3&68K	Geneformer	✗	✗	21.1	24.7	55.2	49.2	59.3	69.1	81.5	74.8	83.3	74.1
	LangCell $_{\mathrm{w/oCTM}}$	83.7	82.4	**86.6**	84.4	87.6	86.1	88.7	86.9	88.2	86.8	89.1	87.6
	LangCell	**83.9**	**82.6**	**89.7**	**87.1**	**89.9**	**87.8**	**90.3**	**87.7**	**92.1**	**87.5**	**92.4**	**88.5**

others take $k = P$. In the ablation study reported in Table 6, LangCell $_{\mathrm{w/oCTM}}$ reports the average zero-shot performance of the model on the two datasets in Table 1 under the setting of $k = 0$. In Table B.2.1, we provide the impact of different k values on the model's inference performance and time on the PBMC10K dataset ($P = 8$).

Experimental results are consistent with theoretical derivations, demonstrating that larger values of k enhance model performance but also increase the time cost of inference. Furthermore, the inference speed at $k = 0$ is close to that of LangCell-CE or Geneformer. Fortunately, in most cases, the model performance at $k = 0$ is not much lower than at $k = P$, and it generally still surpasses other methods. When users have high demands for inference speed, setting $k = 0$ is a viable option to quickly obtain satisfactory results. In Table B.2.2, we provide the complete performance of LangCell at $k = 0$ (i.e., w/o CTM) in the experiments of Table 1 from the paper.

C. Experiment Settings for Pre-training and the Downstream Tasks

Pre-training The pre-training is conducted on four NVIDIA Tesla A100 GPUs and takes approximately 50 days to complete. More experiment configurations are shown in Table C.0.1.

Downstream Tasks In downstream tasks, we uniformly follow the settings below:

- Perform quality control on all datasets used, removing special categories such as "Other" or "Unknown", as well as single cells with too few expressed genes.
- For tasks with randomness, perform three random iterations and take the average.
- In few-shot tasks, all models are trained for 20 epochs.
- For fine-tuning tasks, all models are trained for the same number of epochs. Cell type annotation uses a training:test split of 2:1, while pathway identification uses a training:test split of 3:7.

D. Datasets

D.1. Pre-training Data

We constructed a pre-training dataset named **scLibrary** from CELLxGENE (Megill et al., 2021; Biology et al., 2023). We obtained raw count matrices of scRNA-seq data along with their associated metadata. Our criteria for selection encompassed human cells that were analyzed using the 10X Genomics sequencing technology. We filtered out data that contained

Table C.0.1: Experiment Configurations

	Hyperparameter	**Value**
Model	Vocab size	25427
	Hidden size	512
	Number of hidden layers	12
	Max sequence length	2048
	Number of attention heads	8
	Dropout	0.02
	Hidden act	ReLU
	LayerNorm eps	1e-12
Pre-training	Similarity function	Cosine similarity
	Optimizer	AdamW
	Scheduler	Linear
	Max learning rate	1e-5
	Warm up steps	1000
	Weight decay	1e-3
	Batch size	3
	Gradient accumulation	32

duplicates, had less than 200 expressed genes, exhibited significant gaps in metadata, or were previously utilized in other analyses.

We employed information closely related to cell identity, such as cell type (CL), cell expression phenotype (PATO), ancestral concept system descriptions (HANCESTRO), cell anatomical information (UBERON), disease definitions (Mondo), and ontologies from the Open Biological and Biomedical Ontologies Foundry (OBO Foundry) (Smith et al., 2007), to provide professional-level textual annotations for each single-cell.

The dataset contains textual information categorized into eight distinct cell identities, which describe single-cell sequencing data from various angles, including `Assay`, `Cell Type`, `Development Stage`, `Disease Information`, `Ethnicity of Donor`, `Sex of Donor`, `Tissue Information`, `Organ Information`. After pre-processing, only three labels related to the donor information—"development stage", "ethnicity", and "sex"—have missing values, while the other five labels—assay, cell type, tissue, organ, and disease—are complete without any missing values. The statistical results are depicted in Table D.1.1.

Table D.1.1: Missing values in the **scLibrary**.

Missing Labels	Number of Cells
Miss `Development Stage`	1.92M
Miss `Ethnicity`	13.22M
Miss `Sex`	2.00M
Miss 1 label	10.59M
Miss 2 labels	1.70M
Miss 3 labels	1.05M

Each cell identity has multiple possible values (Table D.1.2). We have selected three significant cell identities to showcase the data distribution, as depicted in Fig. G.0.1, Fig. G.0.2, and Fig. G.0.3.

In the pre-training phase, we stack the cell identity information in a fixed order. Below is an example of a cell description text missing the "ethnicity" information:

Table D.1.2: The cell identities used in the scLibrary.

Cell Identity	Number of values
Assay	7
Disease	56
Cell Type	562
Development Stage	160
Ethnicity of Donor	25
Sex of Donor	3
Tissue Information	192
Organ Information	48

assay: 10x 3' v2. ;
cell type: malignant cell. a neoplastic cell that is capable of entering a surrounding tissue. ;
development stage: 74-year-old human stage. adult stage refers to an adult who is over 74 and under 75 years old. ;
disease: squamous cell lung carcinoma. a carcinoma arising from squamous bronchial epithelial cells. it may be keratinizing or non-keratinizing. keratinizing squamous cell carcinoma is characterized by the presence of keratinization, pearl formation, and/or intercellular bridges. non-keratinizing squamous cell carcinoma is characterized by the absence of keratinization, pearl formation, and intercellular bridges. cigarette smoking and arsenic exposure are strongly associated with squamous cell lung carcinoma. ;
sex: male. a biological sex quality inhering in an individual or a population whose sex organs contain only male gametes. ;
tissue: lung. respiration organ that develops as an out pocketing of the esophagus. ;
tissue general: lung. respiration organ that develops as an out pocketing of the esophagus.

During inference, the model can work quite well even with only a single piece of identity information. For example, in the cell type annotation experiment (4.2.1), we used only the text of "cell type"; in the NSCLC subtype classification experiment (4.2.2), we used only the text of "disease". Here is an example of the "cell type" text used in the experiment of Section 4.2.1:

cell type: dendritic cell. a cell of hematopoietic origin, typically resident in particular tissues, specialized in the uptake, processing, and transport of antigens to lymph nodes for the purpose of stimulating an immune response via T cell activation. these cells are lineage negative (cd3-negative, cd19-negative, cd34-negative, and cd56-negative).

D.2. Downstream Tasks Dataset

We have assembled a set of benchmark datasets to evaluate the performance of LangCell across various downstream tasks. The following discussion will be structured according to the dataset of cells involved.

PBMC10K The PBMC10K dataset, as reprocessed by the study referenced in (Gayoso et al., 2022), features 3,346 distinct genes that are actively expressed. It is compiled from two separate single-cell RNA sequencing (scRNA-seq) data sets, both derived from healthy human peripheral blood mononuclear cells (PBMCs). The first data set includes 7,982 individual cells, and the second comprises 4,008 cells. The PBMC10K dataset encompasses nine different cell types: B cells, CD4 T cells, CD8 T cells, CD14+ Monocytes, Dendritic cells, natural killer (NK) cells, FCGR3A+ Monocytes, Megakaryocytes, and an additional category for other cell types. We have utilized PBMC10K for zero-, few-shot, and full-data cell-type annotation tasks and single-cell integration tasks.

PBMC3&68K The PBMC3&68K (Zheng et al., 2017) dataset is a comprehensive scRNA-seq dataset formed by the integration of two sub-datasets, PBMC3K, and PBMC68K, encompassing a total of 4,638 cell samples. The dataset includes eight types of cells, which are B cells, CD4 T cells, CD8 T cells, CD14+ Monocytes, Dendritic cells, FCGR3A+ Monocytes, Megakaryocytes, and NK cells, allowing for a multifaceted exploration of cellular heterogeneity and functionality within the PBMC population. PBMC3&68K is characterized by the analysis of 14,236 unique genes, providing a detailed view into the transcriptomic landscape of the cells. It is composed of two distinct batches, which may represent different experimental conditions or time points, offering a robust framework for comparative analysis. This level of detail is invaluable for researchers aiming to understand the intricacies of immune cell dynamics and for the development of targeted therapeutic

strategies. We have utilized PBMC3&68K for zero-, few-shot, and full-data cell-type annotation and single-cell integration tasks.

Zheng68K Zheng68K (Zheng et al., 2017) is a highly relevant and challenging dataset, consisting of 68,450 PBMCs with 11 highly related cell types. Zheng68K provides high-quality cell type annotations, making it an ideal benchmark for evaluating annotation approaches. However, the dataset poses significant challenges due to the large number of cell categories and the uneven distribution of samples between types. We have utilized Zheng68K for zero-, few-shot, and full-data cell-type annotation tasks.

Human liver cross datasets Human liver datasets, sourced from the work (Lin et al., 2020), are a combination of the macParland and aizarani datasets. The macParland dataset includes 14 cell types, while the aizarani dataset comprises 7 cell types that are part of the macParland dataset. In our experiments, we utilized the macParland dataset as the training set and the aizarani dataset as the test set to perform zero-, few-shot, and full data cell type annotation tasks, thereby assessing the model's generalization capability.

Perirhinal Cortex The original data for the Perirhinal Cortex dataset is derived from (Siletti et al., 2023), which includes 606 high-quality samples from 10 distinct brain regions. The Perirhinal Cortex dataset consists of two batches with rich cellular content, containing 59,357 genes in total. The first batch includes 8,465 cells, while the second batch comprises 9,070 cells. We have utilized Perirhinal Cortex for zero-, few-shot, and full-data single-cell integration tasks.

Tabula Sapiens Tabula Sapiens (Consortium* et al., 2022) is an innovative human single-cell research project that has uncovered the transcriptomic features of 475 distinct cell types by analyzing live cells from multiple human tissues. The data is derived from 59 meticulously selected samples, encompassing a broad range of tissue types from the bladder to the vasculature, involving donors of varying genders, ethnicities, and ages. The project has analyzed a total of 483,152 cells, including a substantial number of immune cells, epithelial cells, endothelial cells, and stromal cells. We have utilized Tabula Sapiens for the cell-text retrieval task.

Non-small cell lung cancer (NSCLC) subtype dataset The Non-small cell lung cancer (NSCLC) subtype dataset we've developed provides a new benchmark for cell identification tasks. This dataset is sourced from CELLxGENE, where we meticulously selected cell data from the lungs of patients with malignant lung cancer, specifically those diagnosed with adenocarcinoma or squamous cell carcinoma. By annotating the dataset with clinical metadata, we've distinguished between the two NSCLC subtypes. Cluster analysis revealed a clear division of the data into two age-based clusters. Considering the more uniform data distribution in the elderly population, we have chosen this group's cluster to represent the NSCLC subtype dataset. We have utilized the NSCLC subtype dataset for the cancer subtype identification task.

Cell pathway identification dataset Pathway analysis is indispensable in this field, offering a detailed perspective on cellular diversity and molecular dynamics, which is essential for pinpointing key biological changes and therapeutic targets, thereby driving the precision of medical treatments. Therefore, we constructed the cell pathway identification dataset, a new dataset for cell identification. This dataset is obtained from CELLxGENE, and is processed using various R packages to create a pathway annotation dataset. Initially, the Seurat package was utilized for data normalization and identification of variable features. Subsequently, integrated pathway analysis was conducted on the normalized data using the irGSEA package, employing the "AUCell" method to score pathway activities in individual cells. The analysis specifically focused on the 50 hallmark pathways from the Broad Institute's Molecular Signatures Database (MSigDB), considering only the top 5% of expressed pathways. Finally, the dataset was further refined to include only those pathways that appeared with a frequency greater than 0.5%. This approach enabled a comprehensive and targeted annotation of cellular pathways, enhancing the understanding of cellular functions and states within the single-cell RNA-seq data.

It acts as a crucial lens through which we can discern cellular heterogeneity and the intricate interplay of molecular interactions within cells. By annotating active pathways in individual cells, this dataset provides an unparalleled viewpoint to identify pivotal biological transitions and pinpoint potential therapeutic targets. Such granularity is pivotal for advancing precision medicine, enabling the customization of interventions to the precise molecular characteristics observed in pathological states.

The genesis of this dataset marks the confluence of bioinformatics and systems biology, establishing a formidable foundation for forthcoming research endeavors. It enables a more profound comprehension of the molecular machinations underlying

complex diseases and paves new pathways for drug discovery. By bridging the gap between high-resolution single-cell data and functional pathway analysis, this dataset emerges as a potent tool for decoding the complexities of cellular life, thereby fostering the advancement of human health and the development of innovative therapeutic strategies.

E. Evaluation Metrics

Cell type annotation To estimate the effectiveness of LangCell for multi-classification tasks, we employ three evaluation metrics: accuracy, macro F_1-score, and weighted F_1-score. Accuracy measures the closeness of the prediction to the ground truth, while macro F_1-score comprehensively assesses classification results without considering the importance of different categories. We also use weighted F_1-score to measure classification performance while accounting for the importance of different categories. These metrics are calculated based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

To calculate both macro F_1-score and weighted F_1-score, we need to compute Precision and Recall. These two key metrics are calculated using the following formulas:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

Thus, we can compute both macro F_1-score and weighted F_1-score using the following formulas, N denotes the total number of cell types and n_i denotes the number of samples in the i-th class:

$$macro\, F_1 = \frac{1}{N} \sum_{i=1}^{N} F_1^{(i)}$$

$$weighted\, F_1 = \frac{1}{N} \sum_{i=1}^{N} n_i * F_1^{(i)}$$

$$\text{where } \; F_1^{(i)} = \frac{2 * Precision^{(i)} * Recall^{(i)}}{Precision^{(i)} + Recall^{(i)}}$$

Single-cell Integration We implemented the evaluation metrics as defined in the scIB (Luecken et al., 2022) benchmark study, which serves as a benchmark for single-cell integration. Here is a detailed description of each metric.

1. **Adjusted Rand Index (ARI)**

   The ARI for cell types is a metric used in the field of cell biology and systems biology to evaluate the quality of cell clustering. It is a modification of the traditional Rand Index, which measures the similarity between two partitions of the same set of elements. The ARI_{cell} is specifically tailored to assess the agreement between the annotated cell types (or labels) and the clusters generated by an algorithm which is a community detection algorithm applied to cell data.

   The ARI_{cell} score is a normalized measure that ranges from 0 to 1. A score of 0 indicates that the clustering is no better than random chance, meaning the algorithm's partitioning is as likely as random labeling. Conversely, a score of 1 indicates a perfect match between the algorithm's clusters and the true annotations, signifying that the clustering has successfully identified the underlying structure of the cell types.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

2. **Normalize Mutual Information (NMI)**

NMI is a statistical measure used to evaluate the similarity between two categorical label sets. In the single-cell integration task, we compare the ground truth cell type labels with the cell type labels derived from Louvain clustering of integrated cell embeddings. The NMI_{cell} quantifies the concordance between these two sets of labels, with a score of 1 indicating perfect alignment and a score of 0 indicating no correlation. The Louvain clustering algorithm is applied across a range of resolutions from 0.1 to 2, with increments of 0.1, to find the optimal clustering configuration that maximizes the NMI_{cell} score, thereby ensuring the best possible match between the predicted cell types and the actual cell types.

$$NMI(Y, C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)}$$

Y represents the true categories of the data; C represents the results of the clustering; H represents the cross-entropy; $I(Y; C)$ represents mutual information, which is a useful measure of information in information theory. $I(Y; C) = H(Y) - H(Y|C)$. Mutual information is a useful measure of information in information theory; it represents the amount of information about one random variable contained within another, or the reduction in uncertainty of one random variable due to the knowledge of another. In other words, it quantifies the degree of correlation between two random variables.

3. **Average Silhouette Width (ASW)**

   ASW is a metric used to evaluate the quality of clustering in datasets, particularly in the context of cell type clustering and batch mixing evaluation. It quantifies the cohesion of clusters by measuring the average silhouette width of all data points within a cluster.

   The silhouette width ranges from -1 to 1, where:

   (a) A value of 1 indicates that the data point is well-matched to its cluster and very dissimilar to the nearest cluster.
   (b) A value close to 0 suggests that the data point lies on or near the decision boundary between two clusters, indicating poor clustering.
   (c) A negative value indicates that the data point is closer to a different cluster than its own, suggesting misclassification.

   To assess the effectiveness of cell type clustering, we calculate ASW_{cell} with the known cell type labels. To evaluate batch clustering, we derive an adjusted ASW score by incorporating batch labels and subtracting 1 from it, which we refer to as ASW_{batch}. The scores of ASW_{cell} and ASW_{batch} range from 0 to 1, with higher values signifying superior performance in cell-type clustering or batch mixing. The calculation is as follows:

$$ASW_{cell} = \frac{ASW_C + 1}{2}$$
$$ASW_{batch} = 1 - |ASW_B|$$

   where C denotes the set of all cell identity labels.

4. **Integration Metrics**

   We report three key evaluation metrics to assess the performance of LangCell on single-cell integration tasks. Avg_{bio} represents the average value of ARI_{cell}, NMI_{cell} and ASW_{cell}, reflecting the conservation of biological variance. ASW_{batch} indicates the effectiveness of batch effect removal. We perform a weighted average of the two to obtain S_{final}, providing a comprehensive evaluation of a model's performance in single-cell integration tasks.

$$Avg_{bio} = \frac{ARI_{cell} + NMI_{cell} + ASW_{cell}}{3}$$
$$S_{final} = Avg_{bio} \times 0.6 + ASW_{batch} \times 0.4$$

Cell-Text Retrieval We utilize the commonly used retrieval metric, recall@k. Specifically, for the i-th sample with label y_i, and the top k results retrieved denoted as $R_{i,k}$, then:

$$\text{retrieval@k}_i = 1 \; if \; y_i \in R_{i,k} \; else \; 0$$
$$\text{retrieval@k} = \text{average}\{\text{retrieval@k}_i\}$$

Pathway Identification We calculate AUROC and AUPRC in two different ways. The first method is denoted as "avg-", which involves calculating the AUROC and AUPRC for N samples across 41 pathways separately and then taking the average. The second method is referred to as "flatten-", where each pathway prediction for every sample is treated as a single prediction, and the AUROC and AUPRC are computed across $41 * N$ predictions.

F. More Related Works

scRNA-seq Data Representation The gene expression profile, essential for scientific inquiry, elucidates the intricacies of gene expression within individual cells. The presence of nearly 20,000 human protein-coding genes (Seal et al., 2022), coupled with the "Dropout Zeros" phenomenon (Svensson, 2019; Silverman et al., 2018; Linderman et al., 2022), significantly complicates the analysis of high-dimensional data.

Traditional methods involve dimensionality reduction, such as manual marker gene selection (Pasquini et al., 2021; Guo & Li, 2021), machine learning techniques (F.R.S., 1901; Shen, 2009; Hotelling, 1933; Tsuyuzaki et al., 2019; van der Maaten, 2014; Li et al., 2023b), or autoencoder-based approaches (Alessandri et al., 2020; Talwar et al., 2018; Tran et al., 2019; 2022). However, manual gene selection is often empirical (Huang et al., 2020) and results in information loss, while machine learning methods tend to be complex and susceptible to noise. Autoencoder-based approaches depend on the similarity between test and training data. Yet, in practice, it is not always feasible to obtain labeled training data that closely match the distributions of interest.

scBERT (Yang et al., 2022a), using the Performer architecture (Choromanski et al., 2022) with 6 million parameters, encodes over a million normalized, unlabeled scRNA-seq samples and surpasses performance benchmarks in cell type annotation tasks. Exceiver (Connell et al., 2022), with the Perceiver IO architecture (Jaegle et al., 2022), pre-trains on 0.5 million healthy human scRNA-seq count matrix data, demonstrating effectiveness in downstream tasks. Geneformer (Theodoris et al., 2023) pre-trains on nearly 30 million scRNA-seq samples, applying transfer learning across various biological tasks. scGPT (Cui et al., 2023) is trained on over 33 million scRNA-seq records, and fine-tunes for downstream tasks including cell type annotation and multi-batch integration. scFoundation (Hao et al., 2023), with 100 million parameters, pre-trains on over 50 million human scRNA-seq data, introducing read-depth-aware pre-training to model gene co-expression patterns, validated in tasks like gene expression enhancement and drug response prediction. BioTranslator (Xu et al., 2023a) bridges the gap between natural language and scRNA-seq data. However, its reliance on MLP for encoding scRNA-seq data falls short of capturing the intricacies of transcriptomic complexity.

During the review and revision process of this paper, there have also been new preprints exploring the integration of single-cell data and natural language from different perspectives. For example, Cell2Sentence (Levine et al., 2023) and GenePT (Chen & Zou, 2023) have proposed the idea of directly transcribing single-cell gene sequences into natural language and utilizing large language models for encoding. These works are still undergoing updates and improvements, and we look forward to their future contributions in providing more valuable insights to this field.

Multi-modal in Scientific Data Multi-modal learning enhances the model's ability to understand and express multi-modal data, with the core of this approach resting in the creation of a unified representational space that fosters inter-modal interaction and learning, capturing data interconnections and enhancing generalization through cross-modal knowledge transfer. The vision language model has experienced significant advancements. The CLIP (Radford et al., 2021) model enables image classification and description without extra supervision by learning image-text associations. BLIP (Li et al., 2022) introduces the Multimodal Encoder-Decoder (MED) architecture, which enables the model to switch seamlessly between encoding and generation tasks, thereby enhancing the quality of the text corpus through the innovative Captioning and Filtering (CapFilt) method. Additionally, BLIP-2 (Li et al., 2023a) leverages the Querying Transformer (Q-Former) to effectively bridge the gap between visual and textual modalities, further advancing the state of the art in vision-language pre-training.

G. More Figures and Tables

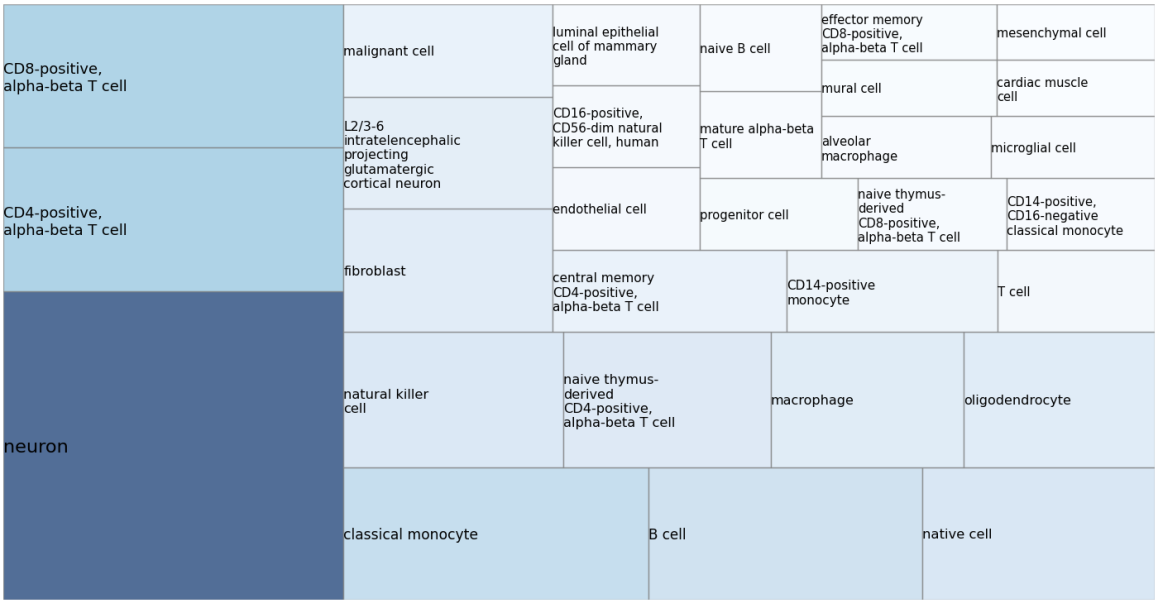

Figure G.0.1: Overview of the distribution of cell type categories in the scLibrary dataset. To facilitate the presentation, we have selected the top 30 categories with the highest data volume for display.

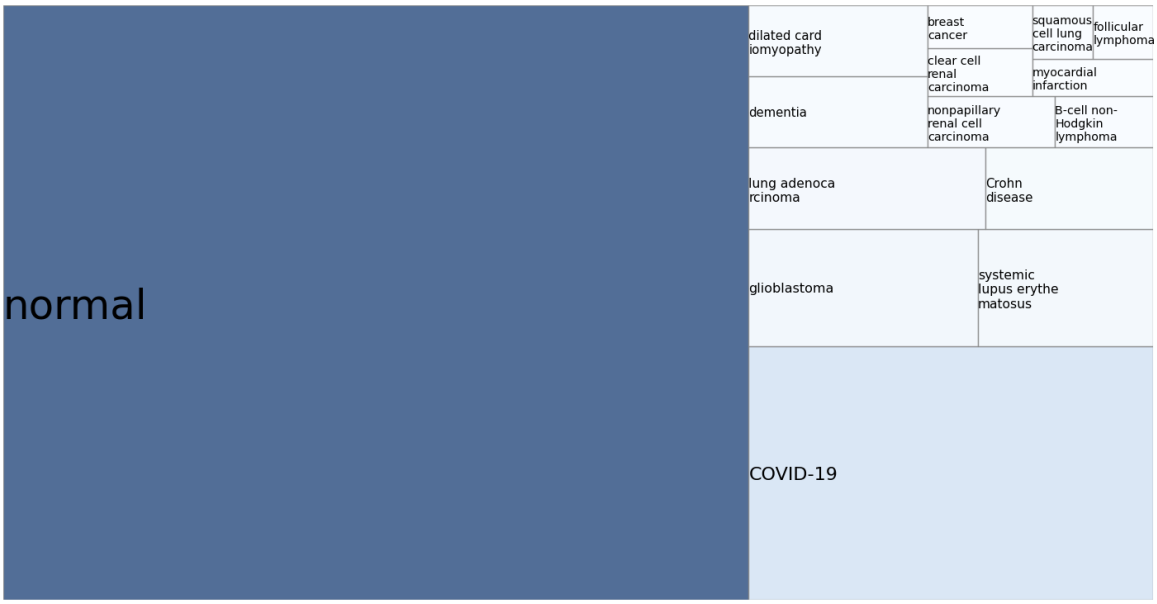

Figure G.0.2: Overview of the distribution of disease in the scLibrary dataset. To facilitate the presentation, we have selected the top 15 categories with the highest data volume for display.

Table G.0.1: The downstream tasks, categories, batch, and quantity information of each dataset used in LangCell.

Dataset	Downstream Task	Batch Number	Cell type	#Number
PBMC10K	**Cell Type Annotation** (Zero-, Few-shot, and Full) **Single-cell Integration**	**2**	CD4 T cells	4,996
			CD14+ Monocytes	2,227
			B cells	1,621
			CD8 T cells	1,448
			Other	463
			NK cells	457
			FCGR3A+ Monocytes	351
			Dendritic Cells	339
			Megakaryocytes	88
PBMC3&68K	**Cell Type Annotation** (Zero-, Few-shot, and Full) **Single-cell Integration**	**2**	CD4 T cells	2,384
			CD8 T cells	665
			CD14+ Monocytes	564
			B cells	476
			NK cells	276
			FCGR3A+ Monocytes	195
			Dendritic cells	61
			Megakaryocytes	17
Zheng68K	**Cell Type Annotation** (Zero-, Few-shot, and Full)	**-**	CD8+ Cytotoxic T	20,757
			CD8+/CD45RA+ Naive Cytotoxic	16,645
			CD56+ NK	8,775
			CD4+/CD25 T Reg	6,185
			CD19+ B	5,877
			CD4+/CD45RO+ Memory	3,059
			CD14+ Monocyte	2,847
			Dendritic	2,095
			CD4+/CD45RA+/CD25- Naive T	1,871
			CD34+	242
			CD4+ T Helper2	97
macParland	**Cell Type Annotation** (Zero-, Few-shot, and Full)	**-**	Hepatocytes	3,501
			ab T	961
			Inflammatory Macs	813
			gd T	569
			Plasma cells	511
			NK cells	488
			Non-inflammatory Macs	379
			Central venous liver sinusoidal endothelial cells	327
			Periportal liver sinusoidal endothelial cells	306
			Portal endothelial cells	211
			Mature B cells	129
			cholangiocytes	119
			Erythroid cells	93
			Stellate cells	37
Aizarani	**Cell Type Annotation** (Zero-, Few-shot, and Full)	**-**	Hepatocytes	3,086
			T/NK	3,066
			liver sinusoidal endothelial cells	1,361
			cholangiocytes	1,022
			macrovascular endothelial cells	355
			B	244
			Stellate cells and myofibroblasts	28
Perirhinal Cortex	**Single Cell Integration**	**2**	oligodendrocyte precursor cell	6,404
			astrocyte	5,319
			oligodendrocyte	4,073
			central nervous system macrophage	770
			endothelial cell	544
			fibroblast	305
			pericyte	68
			leukocyte	41
			vascular associated smooth muscle cell	8
			neuron	3

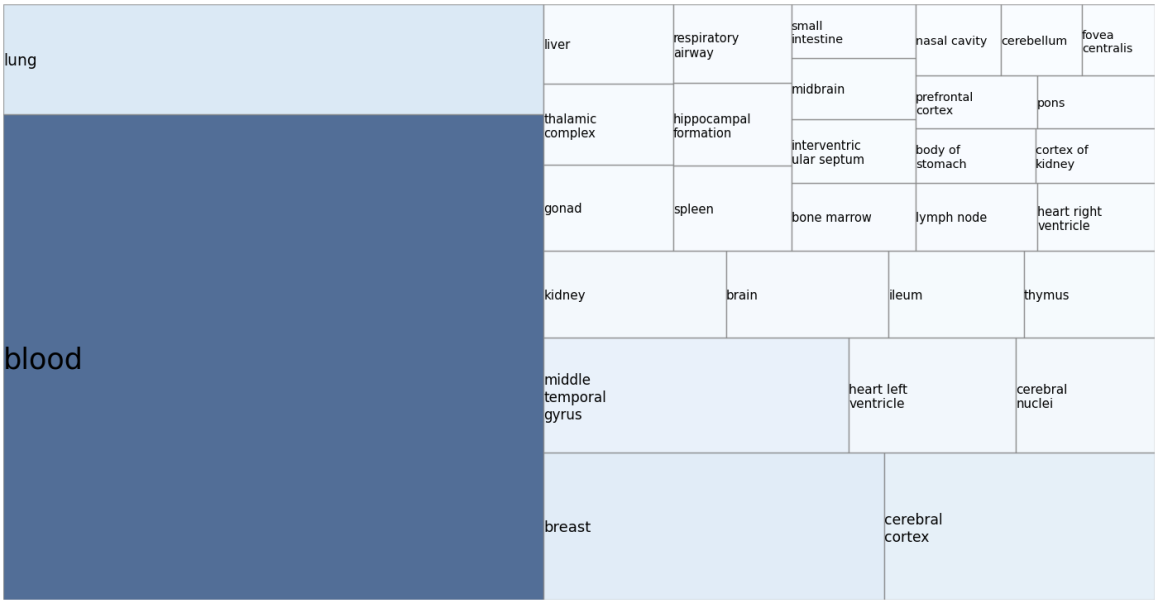

Figure G.0.3: Overview of the distribution of tissue in the scLibrary dataset. To facilitate the presentation, we have selected the top 30 categories with the highest data volume for display.

Table G.0.2: The downstream tasks, categories, batch, and quantity information of each dataset used in LangCell.

Dataset	Downstream Task	Batch Number	Cell type	#Number
Tabula Sapiens (Top 10)	**Cell-Text Retrieval**	-	macrophage	33,607
			fibroblast	31,125
			B cell	19,067
			neutrophil	16,992
			memory B cell	14,565
			mesenchymal stem cell	14,036
			T cell	13,947
			basal cell	12,991
			CD4-positive, alpha-beta T cell	12,870
			classical monocyte	12,746
Non-small cell lung cancer (NSCLC) subtype	**Cancer Subtype Identification**	-	Squamous cell lung carcinoma	1,600
			Lung adenocarcinoma	1,058