GRIT: Graph-Regularized Logit Refinement for Zero-shot Cell Type Annotation

Tianxiang Hu * **Chenyi Zhou** * **Jiaxiang Liu** * **Jiongxin Wang** * **Ruizhe Chen**

Haoxiang Xia **Gaoang Wang** **Jian Wu** **Zuozhu Liu** ✉

Zhejiang University

{tianxiang.23, zuozhuliu}@intl.zju.edu.cn

Abstract

Cell type annotation is a fundamental step in the analysis of single-cell RNA sequencing (scRNA-seq) data. In practice, human experts often rely on the structure revealed by principal component analysis (PCA) followed by k-nearest neighbor (k-NN) graph construction to guide annotation. While effective, this process is labor-intensive and does not scale to large datasets. Recent advances in CLIP-style models offer a promising path toward automating cell type annotation. By aligning scRNA-seq profiles with natural language descriptions, models like LangCell enable zero-shot annotation. While LangCell demonstrates decent zero-shot performance, its predictions remain suboptimal, particularly in achieving consistent accuracy across all cell types. In this paper, we propose to refine the zero-shot logits produced by LangCell through a graph-regularized optimization framework. By enforcing local consistency over the task-specific PCA-based k-NN graph, our method combines the scalability of the pre-trained models with the structural robustness relied upon in expert annotation. We evaluate our approach on 14 annotated human scRNA-seq datasets from 4 distinct studies, spanning 11 organs and over 200,000 single cells. Our method consistently improves zero-shot annotation accuracy, achieving accuracy gains of up to 10%. Further analysis showcase the mechanism by which GRIT effectively propagates correct signals through the graph, pulling back mislabeled cells toward more accurate predictions. The method is training-free, model-agnostic, and serves as a simple yet effective plug-in for enhancing automated cell type annotation in practice.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have enabled high-resolution profiling of cellular heterogeneity across diverse tissues and biological conditions [28, 16]. A critical step in the analysis pipeline is cell type annotation—assigning biologically meaningful labels to individual cells—which forms the basis for downstream interpretation. Traditionally, this process is performed through a semi-automatic workflow [29, 2, 31, 3, 26, 14], combining dimensionality reduction (e.g., principal component analysis), clustering algorithms (e.g., Leiden or Louvain), and manual inspection of marker gene expression. However, this approach is time-consuming, subjective, and increasingly impractical as scRNA-seq datasets grow in size and complexity.

Recent advances in deep learning have enabled the development of powerful representation learning frameworks for single-cell RNA sequencing (scRNA-seq) data. Models such as scBERT [42],

∗ Equal contribution. ✉ Corresponding author.

scGPT [6], Geneformer [7], and scFoundation [10] are pretrained on millions of single-cell profiles and support transfer learning for a wide range of downstream tasks. These foundation models offer a unified and scalable approach to encoding single-cell gene expression into informative embeddings using general-purpose neural architectures. Building on this foundation, LangCell [46] adopts Geneformer—a high-performing single-cell foundation model—as its cell encoder within a CLIP-style contrastive learning framework that aligns single-cell embeddings with natural language descriptions of cell identities. By training on paired scRNA-seq profiles and cell-type text annotations, LangCell uniquely enables zero-shot cell type annotation: predicting cell types in previously unseen datasets without additional retraining. Given the success of CLIP-style and multimodal foundation models in other domains when trained at scale, there is strong reason to believe that such architectures—when further developed and trained on diverse, high-quality scRNA-seq and text corpora—can evolve into powerful tools for automating and augmenting expert-level biological annotation.

An important observation in the context of single-cell annotation is that human experts frequently rely on the structure revealed by the PCA-based k-NN graph. This suggests that, for a given scRNA-seq dataset, we already possess a relatively reliable and task-relevant structure among cells. While models like LangCell enable zero-shot annotation by aligning single-cell representations with natural language description representations, they are not explicitly trained to preserve this structural information. As a result, the predictions may be locally inconsistent with the k-NN graph which serves as an informative structural reference in expert-guided annotation workflows. This limitation motivates a hybrid strategy: refining pre-trained models' zero-shot predictions by enforcing local consistency over the task-relevant k-NN graph. In doing so, we aim to combine the scalability provided by the pre-trained model with the structural robustness of the PCA-based k-NN graph. It is natural to consider existing methods such as label propagation or graph-based regularization. However, many of these approaches [40, 49, 48] require access to at least a subset of ground truth labels which are unavailable in the zero-shot setting. Others [1, 39, 17] focus on enhancing graph neural network training, which may be unnecessary in our case. These limitations call for a label-free, inference-time refinement approach tailored to the zero-shot annotation setting.

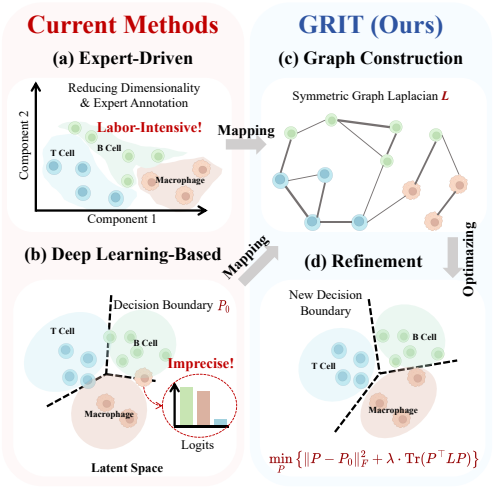

Figure 1: Overview of GRIT. (**a**–**b**) Existing approaches rely on expert-driven labeling or deep learning models like LangCell, which can be labor-intensive or imprecise. (**c**) Construct a PCA-based k-NN graph, with each node initialized by the logits predicted by a deep learning model. (**d**)GRIT refines these initial predictions by solving a graph-regularized optimization problem that promotes local consistency across the k-NN graph.

In this paper, we propose a graph-regularized logit refinement method for enhancing zero-shot cell type annotation. The key intuition is to leverage the scalability of pretrained foundation models while correcting their output using domain-specific geometric consistency—without requiring any additional training. Specifically, we first apply a CLIP-style model—LangCell in our case—to perform zero-shot annotation on scRNA-seq data, resulting in initial prediction logits. We then construct a PCA-based k-NN graph to reflect the underlying structure commonly trusted by human experts. Finally, we solve a graph-regularized optimization problem to refine the logits, encouraging smoothness over the graph while remaining close to the original predictions. This refinement is applied entirely at inference time and serves as a lightweight, principled postprocessing step. We rigorously analyze our approach and validate its effectiveness through extensive experiments on 14 annotated human scRNA-seq datasets, collected from 4 independent studies and spanning 11 organs and over 200,000 single cells. Our refinement consistently improves zero-shot performance, with nearly all cases showing gains in accuracy, macro F1, and weighted F1. Our main contributions are:

- We propose a single-cell annotation approach that combines the efficiency of CLIP-based methods with the robustness of PCA. Our method is simple, training-free, and effective, serving as a plug-and-play solution for zero-shot annotation.
- We provide a theoretical proof that graph-regularized optimization can effectively refine predictions, under the condition the initial logits are reasonably accurate.
- Extensive zero-shot cell type annotation experiments using LangCell across 14 human organ datasets demonstrate consistent performance improvements through the application of graph-regularized logit refinement.

2 Background and Related Work

2.1 Cell Type Annotation in scRNA-seq

Traditionally, cell types have been defined based on criteria such as morphology, physiology, and marker protein expression[43]. However, single-cell analysis provides a means of systematically detecting cell types that cannot be defined by a handful of markers or for which markers are not yet known [35, 13]. Classifying cells into discrete types from single-cell profiles is an unsupervised clustering problem in high-dimensional space. Clustering in high-dimensional space is hindered by the instability of distance metrics [36]. As a result, dimensionality reduction, through either linear or nonlinear methods, is widely used as an initial step. Among linear approaches, principal component analysis (PCA) offers a deterministic and interpretable projection of genomic profiles into a lower-dimensional space and is highly scalable. Building upon these embeddings, k-NN graphs are frequently constructed to capture local cellular neighborhoods, which enhances clustering stability and enables downstream refinements such as smoothing cell type predictions or integrating prior knowledge [19]. PCA has been repeatedly applied in single-cell analysis, including in iterative analyses where its axes are inferred from bulk data and then applied to single-cell data [37], and in combination with the expectation-maximization (EM) algorithm [23] to fit a finite mixture of Gaussians to PCA-reduced expression profiles [12]. Ultimately, the results obtained from PCA reduction allow the extraction of molecular markers and features that characterize each cell type, while also being adaptable to the exponential growth in single-cell data and facilitating the exploration of hierarchical structures at different granularities of cell types [18].

2.2 Representation Learning for scRNA-seq

The increasing availability of single-cell RNA sequencing (scRNA-seq) data has spurred the development of machine learning approaches for learning meaningful representations from high-dimensional gene expression profiles [24, 50, 16, 25, 45, 38, 34, 15, 11]. Traditional analysis pipelines often begin with dimensionality reduction techniques such as PCA or t-SNE, followed by clustering and manual marker gene inspection [4]. While these methods remain effective for visualizing structure and guiding expert annotation, recent work has turned to deep learning—particularly transformer-based architectures—to model complex nonlinear relationships across genes and enhance scalability to large, diverse datasets. Models such as SCBERT[42], SCGPT[6], GENEFORMER[7], and SCFOUNDATION[10] leverage large-scale pretraining on millions of scRNA-seq profiles to support transfer learning across downstream tasks including cell type annotation.

2.3 Language-Cell Alignment and Zero-Shot Cell Type Annotation

To bridge the gap between scRNA-seq data and natural language descriptions of cells, LANGCELL [46] proposes a pretraining framework that aligns cellular profiles with curated ontology-derived textual descriptions. LangCell builds upon GENEFORMER as its cell encoder and a BERT-style language model as text encoder, forming a CLIP-style architecture tailored for the biomedical domain. Through contrastive learning, LangCell learns a shared embedding space between single cells and their natural language descriptions. This design enables decent performance in both zero-shot and few-shot cell type classification without requiring task-specific fine-tuning, marking a significant step toward high-performance, automated cell type annotation. While LangCell demonstrates promising generalization in zero-shot settings, several limitations remain. First, neither its cell encoder nor the multimodal model itself explicitly leverages graph-based structural information during pretraining. Second, its inference mechanism—like other CLIP-style retrieval paradigms [20, 21, 32]—can be

sensitive to suboptimal prompt design and modality biases. We thus think of proposing a post-hoc refinement method that operates on the logits produced by CLIP-like models to improve prediction quality by respecting graph consistency.

3 Methodology

In this section, we present our graph-regularized refinement method for enhancing zero-shot cell type annotation in single-cell RNA sequencing (scRNA-seq) data. The core idea is to improve the initial prediction logits produced by a pretrained model via encouraging local consistency among cells, guided by structural information captured in a PCA-based k-nearest neighbor (k-NN) graph. We begin by providing an analysis demonstrating that, when the initial logits are reasonably accurate, applying graph regularization is guaranteed to improve prediction performance. We then introduce the full pipeline of our proposed method, GRIT, for zero-shot cell type annotation.

3.1 Theoretical Analysis

Theorem 1 (Graph Regularized Logit Refinement Improves Predictions)**.** *Given a symmetric graph Laplacian $L \in \mathbb{R}^{n\times n}$ constructed from an adjacency matrix $A \in \mathbb{R}^{n\times n}$, let $P_0 \in \mathbb{R}^{n\times c}$ denote the initial class logits over the n nodes, and let $P^* \in \mathbb{R}^{n\times c}$ denote the ground-truth logits. Consider the following graph-regularized optimization problem:*

$$\hat{P}_\lambda := \arg\min_P \left\{ \|P - P_0\|_F^2 + \lambda \cdot \mathrm{Tr}(P^\top L P) \right\}.$$

Suppose the following condition holds:

$$\langle P_0 - P^*,\, LP_0 \rangle > 0.$$

Then there exists a sufficiently small regularization parameter $\lambda > 0$ such that:

$$\|\hat{P}_\lambda - P^*\|_F^2 < \|P_0 - P^*\|_F^2.$$

Proof. The objective function is convex and quadratic, and the unique minimizer admits the closed-form solution:

$$\hat{P}_\lambda = (I + \lambda L)^{-1} P_0.$$

Define the function $f(\lambda) := \|\hat{P}_\lambda - P^*\|_F^2$. Since $\hat{P}_\lambda$ is smooth in λ, $f(\lambda)$ is also continuously differentiable.

At $\lambda = 0$, we have $\hat{P}_0 = P_0$, so:

$$f(0) = \|P_0 - P^*\|_F^2.$$

To study how $f(\lambda)$ changes near 0, compute the derivative using the chain rule. Let $A_\lambda := (I+\lambda L)^{-1}$, so $\hat{P}_\lambda = A_\lambda P_0$. Then:

$$f'(\lambda) = \frac{d}{d\lambda} \|A_\lambda P_0 - P^*\|_F^2 = -2 \cdot \mathrm{Tr}\left[(A_\lambda P_0 - P^*)^\top A_\lambda L A_\lambda P_0 \right].$$

At $\lambda = 0$, we have $A_0 = I$, so:

$$f'(0) = -2 \cdot \mathrm{Tr}\left[(P_0 - P^*)^\top L P_0 \right] = -2\langle P_0 - P^*, LP_0 \rangle.$$

By assumption, this quantity is strictly negative. Since $f(\lambda)$ is differentiable and $f'(0) < 0$, there exists some $\lambda > 0$ sufficiently small such that:

$$f(\lambda) < f(0),$$

i.e.,

$$\|\hat{P}_\lambda - P^*\|_F^2 < \|P_0 - P^*\|_F^2.$$

□

Remark 1. *The condition $\langle P_0 - P^*, LP_0 \rangle > 0$ captures a meaningful alignment between the prediction error and the graph structure. It implies that the residuals in P_0 are not arbitrary but exhibit disagreement that is structured according to the graph Laplacian. In other words, when a prediction is incorrect, its neighbors tend to disagree in directions that the regularizer penalizes. This condition is satisfied when P_0 is informative yet imperfect—e.g., capturing coarse structure but missing fine details. In such cases, the graph enables error correction by propagating reliable information across neighbors. This formalizes the intuition that logit refinement is effective when initial predictions are reasonable and the graph reflects meaningful similarity.*

3.2 The GRIT Method

We propose a three-stage framework GRIT for zero-shot cell type annotation in scRNA-seq data. Given an unlabeled dataset of n cells with gene expression profiles $\{x_i\}_{i=1}^n$ and a set of c cell type descriptions $\{t_j\}_{j=1}^c$, our method consists of: (1) obtaining initial logits via a CLIP-style model, (2) constructing a k-NN graph constructed in PCA space and (3) refining the initial logits using graph-regularized optimization (Figure 1).

Step 1: Initial Prediction via CLIP-Style Model. We first adopt a CLIP-style model, LangCell, to obtain the initial prediction logits $P_0 \in \mathbb{R}^{n\times c}$, where each row corresponds to a cell and each column to a candidate cell type. For each cell $x_i, P_0(i) = \text{model}\left(x_i, \{t_j\}_{j=1}^c\right)$, where $\text{model}(\cdot,\cdot)$ produces a probability distribution reflecting the alignment between the input cell and each cell type candidate. LangCell uses these logits for zero-shot prediction via: $g(x_i) = \arg\max_j \{P_0(i)\}, \quad \forall i \in \{1,\ldots,n\}$, enabling automatic cell type annotation without any task-specific training or fine-tuning. The logits P_0 serve as the input to our refinement procedure.

Step 2: Graph Construction. To model the relational structure among cells, we construct a k-nearest neighbor (k-NN) graph based on a low-dimensional representation of the input data. Specifically, each cell x_i is first preprocessed following standard scRNA-seq analysis procedures. The resulting gene expression matrix is standardized and projected onto a d-dimensional space using principal component analysis (PCA), yielding a reduced representation x_i^{PCA} for each cell. Using these PCA-reduced features, we build a k-NN graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by connecting each cell to its k nearest neighbors in Euclidean space. The graph is symmetrized so that an undirected edge exists between cell i and cell j if either is among the k nearest neighbors of the other. Let $A \in \mathbb{R}^{n\times n}$ denote the adjacency matrix of this graph, the unnormalized graph Laplacian is then defined as $L = D - A$, where D is the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$. This Laplacian L captures the local geometry of the dataset and is used in the subsequent logit refinement step.

Step 3: Graph-regularized Logit Refinement. After obtaining the initial logits P_0 and the k-NN graph represented by the adjacency matrix A and its corresponding Laplacian L, we refine the logits by solving a graph-regularized optimization problem that promotes local consistency: $\hat{P} = \arg\min_P \|P - P_0\|_F^2 + \lambda \,\text{Tr}(P^\top L P)$, where $\lambda > 0$ controls the strength of the regularization. This objective encourages the refined logits P to remain close to the initial predictions P_0 while being smooth with respect to the graph structure encoded by L. The solution to this convex quadratic problem has a closed form and can be computed efficiently by solving a linear system: $\hat{P} = (I + \lambda L)^{-1} P_0$. GRIT then uses the refined logits to perform zero-shot cell type annotation: $g(x_i) = \arg\max_j \left\{\hat{P}(i)\right\}, \quad \forall i \in \{1,\ldots,n\}$.

4 Experiments

4.1 Datasets and Setup

We compile 11 scRNA-seq datasets from the *Tabula Sapiens* project [5], spanning organs including **Bladder**, **Bone Marrow**, **Fat**, **Heart**, **Kidney**, **Liver**, **Mammary**, **Muscle**, **Spleen**, **Trachea**, and **Uterus** (see Figure 2). In total, the datasets encompass 76 cell types and 171,383 single cells. In addition, we collect two public peripheral blood mononuclear cell datasets (**PBMC10k** [8] and **PBMC368k** [47]), and a peripheral cortex dataset (**Peripheral Cortex** [33]). These datasets cover 18 cell types and 33, 700 single cells. All datasets include cell type annotations curated by domain experts. Dataset pre-processing and statistics are detailed in Appendix.

We focus on the zero-shot cell type annotation task on these datasets and employ LangCell, a pretrained CLIP model adapted for scRNA-seq and text alignment, to generate the initial soft label logits P_0. Following standard preprocessing procedures in Scanpy [41], a PCA-based k nearest neighbor (k-NN) graph is constructed for each dataset to capture cell similarity relationships (see Appendix). Unless otherwise stated, we use $\alpha = 0.2$ for LangCell, 50 principal components for PCA, and $k = 15$ for k-NN as our default hyperparameters. The initial logits are refined using GRIT with a default hyperparameter of $\lambda = 1$. To evaluate performance, we report accuracy, macro F1 score, and weighted F1 score.

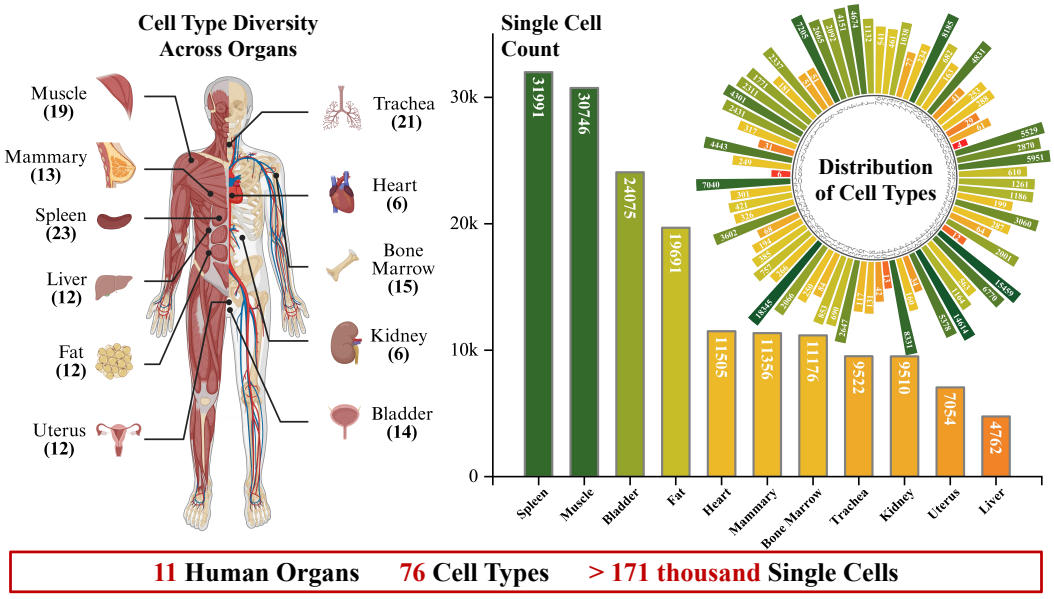

Figure 2: Overview of the human scRNA-seq datasets used in our main experiment. They span 11 human organs, 76 annotated cell types, and over 171,000 single cells. The anatomical illustration summarizes cell type diversity across organs. The bar chart reports single-cell counts per organ. The circular plot visualizes the distribution of all cell types, indexed from 1 to 76 for clarity. Full cell type names corresponding to these indices are listed in Appendix

4.2 Main Results

Constructing a k-NN graph on principal components is a standard step in scRNA-seq analysis pipelines [9, 44]. It is common to reduce gene expression profiles to 50 principal components and set k to 15 or 20 when building the k-NN graph. For example, the widely used toolkits Scanpy [41] and Seurat [27] adopt 50 principal components by default, with $k = 15$ and $k = 20$, respectively. To remain consistent with standard practice among domain experts, we evaluate GRIT under these commonly used configurations.

As shown in Table 1, GRIT consistently improves performance over the baseline across all 11 organ datasets for both k values, with accuracy improvements up to 10.11% and macro F1 improvements up to 4.63%. Moreover, the results under $k = 15$ and $k = 20$ are similar, indicating strong robustness to the choice of k. Table 2 presents the logit consistency before and after GRIT refinement for both k settings, quantified by $P_0^T L P_0$ and $\hat{P}_\lambda^T L \hat{P}_\lambda$, respectively. GRIT consistently reduces the logit inconsistency relative to the graph structure, resulting in refined logits that better align with the k-NN graph used for cell type classification. These findings support our central hypothesis that leveraging the commonly used k-NN graph structure—routinely employed by practitioners for cell type annotation—can meaningfully refine initial prediction logits output by decent prediction models.

4.3 Empirical Analysis of λ

In the Methodology section, we analyzed the behavior of the function $f(\lambda)$, which quantifies the quality of the refined logits $\hat{P}_\lambda$. Under a mild condition on the initial logits P_0, we showed that $f(\lambda)$ increases over a small right-hand interval near $\lambda = 0$. Motivated by this insight, we conduct a systematic sweep over λ values near zero. Specifically, we evaluate $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 30, 40, \ 50, 60, 70, 80, 90, 100\}$. Figure 3 illustrates how performance varies with λ across all 11 organ datasets, both individually and on average. Both the individual and average trend align with our analysis: performance improves as λ increases from zero. Empirically, we find that moderate values in the range $\lambda \in (0, 5)$ enhance the overall performance. We adopt $\lambda = 1$ as the default value.

Table 1: Performance comparison on zero-shot cell type annotation across 11 scRNA-seq datasets from the *Tabula Sapiens* project. We evaluate LangCell and our method GRIT under k-NN settings ($k = 15$ and $k = 20$), using Weighted F1, Macro F1, and Accuracy (%). GRIT consistently outperforms the baseline across most datasets, with per-dataset gains shown as colored subscripts. The average row (**Avg.**) summarizes overall performance.

Dataset	LangCell			GRIT$_{k=15}$ (Ours)			GRIT$_{k=20}$ (Ours)		
	Weighted F1	Macro F1	Accuracy	Weighted F1	Macro F1	Accuracy	Weighted F1	Macro F1	Accuracy
Bladder	76.34	52.57	75.78	78.86 ↑2.52	53.32 ↑0.75	79.36 ↑3.58	78.89 ↑2.55	53.34 ↑0.77	79.41 ↑3.63
Bone Marrow	54.94	36.76	54.49	58.15 ↑3.21	37.59 ↑0.83	59.14 ↑4.65	58.27 ↑3.33	37.71 ↑0.95	59.29 ↑4.80
Fat	38.61	32.70	35.81	37.45 ↓1.16	34.56 ↑1.86	38.41 ↑2.60	37.46 ↓1.15	34.53 ↑1.83	38.43 ↑2.62
Heart	84.22	70.97	85.58	86.89 ↑2.67	73.32 ↑2.35	88.18 ↑2.60	86.98 ↑2.76	73.45 ↑2.48	88.26 ↑2.68
Kidney	97.30	84.60	97.29	98.08 ↑0.78	85.03 ↑0.43	98.20 ↑0.91	98.09 ↑0.79	85.21 ↑0.61	98.21 ↑0.92
Liver	74.26	52.23	75.20	76.61 ↑2.35	56.85 ↑4.62	79.19 ↑3.99	76.64 ↑2.38	56.86 ↑4.63	79.23 ↑4.03
Mammary	82.00	50.07	78.79	84.13 ↑2.13	52.66 ↑2.59	82.06 ↑3.27	84.05 ↑2.05	52.51 ↑2.44	81.99 ↑3.20
Muscle	64.34	31.35	57.12	72.03 ↑7.69	35.50 ↑4.15	67.15 ↑10.03	72.09 ↑7.75	35.54 ↑4.19	67.23 ↑10.11
Spleen	46.13	30.40	48.63	44.59 ↓1.54	30.44 ↑0.04	49.95 ↑1.32	44.57 ↓1.56	30.37 ↓0.03	49.97 ↑1.34
Trachea	59.65	39.34	53.60	62.57 ↑2.92	40.34 ↑1.00	58.72 ↑5.12	62.57 ↑2.92	40.32 ↑0.98	58.72 ↑5.12
Uterus	75.22	44.07	70.71	81.23 ↑6.01	47.57 ↑3.50	80.17 ↑9.46	81.27 ↑6.05	47.43 ↑3.36	80.32 ↑9.61
Avg.	**68.46**	**47.73**	**66.64**	**70.96** ↑2.50	**49.74** ↑2.01	**70.96** ↑4.32	**70.99** ↑2.53	**49.75** ↑2.02	**71.01** ↑4.37

Table 2: Logit consistency before and after the GRIT refinement across 11 organs. GRIT consistently reduces logit inconsistency relative to the underlying graph structure.

Logit Consistency	Bladder	Bone Marrow	Fat	Heart	Kidney	Liver	Mammary	Muscle	Spleen	Trachea	Uterus
$P_0^T L P_0$ ($k = 15$)	60.54	25.10	40.09	35.85	44.41	15.58	29.32	81.48	58.48	17.94	17.67
$\hat{P_\lambda}^T L \hat{P_\lambda}$ ($k = 15$)	16.42	6.91	10.57	9.35	12.17	4.28	8.25	21.64	15.09	4.94	4.95
$P_0^T L P_0$ ($k = 20$)	60.98	25.37	40.28	36.01	44.62	15.67	29.70	82.21	58.71	18.07	17.79
$\hat{P_\lambda}^T L \hat{P_\lambda}$ ($k = 20$)	16.75	7.10	10.75	9.51	12.44	4.35	8.51	22.12	15.34	5.04	5.05

4.4 Robustness Across Different LangCell Hyperparameters and More Datasets

Our main experiments, along with empirical analysis in the low-λ regime, suggest that LangCell's zero-shot predictions are already sufficiently aligned with the graph structure for GRIT to provide effective improvements. To ensure a comprehensive evaluation, we further investigate a broader range of LangCell hyperparameter settings and more datasets from studies in addition to the *Tabula Sapiens* project to assess generalizability across diverse input conditions.

As reported in LangCell, the hyperparameter $\alpha \in \{0.1, 0.2, 0.3, 0.5\}$ yields competitive results, with $\alpha = 0.2$ recommended as the default. We adopt $\alpha = 0.2$ for baseline comparisons, but also evaluate GRIT under all four α settings to assess its robustness across a range of initializations. We test each α value across all 11 datasets used in the main experiments. Table 3 reports the average performance of each α setting, while the detailed per-dataset results are provided in the Appendix. As shown, GRIT consistently improves zero-shot annotation performance—measured by accuracy, macro F1, and weighted F1—across all α values. These results underscore two key observations: (1) current zero-shot models like LangCell are capable of producing logits that already reflect a latent task-relevant structure, which can be effectively enhanced through graph-based refinement, and (2) GRIT is robust to variations in initial logit quality, demonstrating consistent gains across a spectrum of strong configurations. Table 3 presents GRIT's performance on additional datasets including PBMC10k, PBMC368k, and Peripheral Cortex. GRIT consistently improves prediction performance across these datasets, demonstrating strong robustness to diverse data sources with varying data noise.

4.5 Visualization and Case Study

We select three cases that demonstrated strong improvements using GRIT, to illustrate the effect of GRIT refinement, as shown in Figure 4. Among these, the kidney dataset exhibits high initial prediction performance with LangCell, while uterus and muscle comparatively lower. Nonetheless, across all cases, as long as a subset of correctly predicted cells exists within a local neighborhood, GRIT can propagate these signals to pull back mislabeled cells.

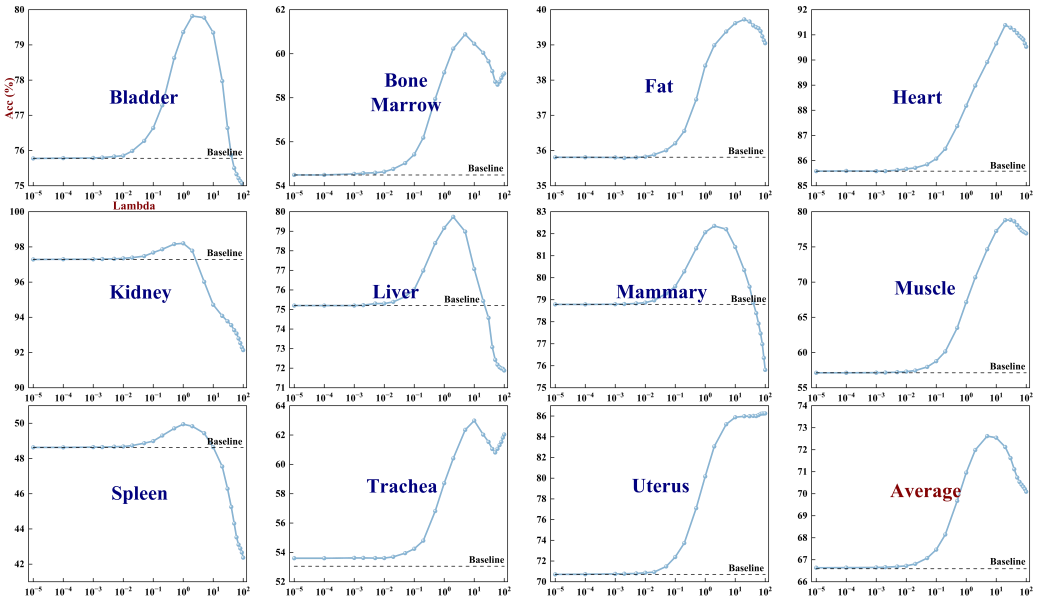

Figure 3: Investigation of GRIT performance in the right-hand neighborhood of $\lambda = 0$ across 11 organs from the *Tabula Sapiens* project and their average. The x-axis denotes λ values, and the y-axis reports logit performance measured by accuracy. Dashed lines indicate baseline accuracies achieved by LangCell.

Table 3: Performance comparison between LangCell and GRIT under varying α values ($\alpha = 0.1, 0.2, 0.3, 0.5$) and on additional datasets (PMBC10k, PBMC368k, Peripheral Cortex). For each metric, the absolute improvement of GRIT over LangCell is shown in the last row indicated by Δ.

Metric	Method	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.5$	PBMC10k	PBMC368k	Peripheral Cortex
Acc	LangCell	63.60	66.59	67.39	67.05	86.52	84.86	98.01
	GRIT	66.87	70.96	71.48	70.65	88.43	87.41	98.41
	Δ	+3.27	+4.37	+4.09	+3.60	+1.91	+2.55	+0.40
Macro F1	LangCell	46.01	47.73	48.32	48.27	83.91	83.17	72.87
	GRIT	47.66	49.74	49.78	49.47	87.57	86.16	73.27
	Δ	+1.65	+2.01	+1.46	+1.20	+3.66	+2.99	+0.40
Weighted F1	LangCell	65.69	68.46	68.97	68.59	87.44	85.97	98.19
	GRIT	67.44	70.96	71.14	70.45	89.21	88.23	98.42
	Δ	+1.75	+2.50	+2.17	+1.86	+1.77	+2.26	+0.23

However, not all datasets benefit equally: in fat and spleen, GRIT slightly reduces F1 scores (see Table 1). To further explore this trend, we analyze how the improvement of GRIT varies across organs with various initial prediction quality and dataset size. The improvement is visualized using color, where red indicates a significant performance gain, and green denotes a marginal improvement. Specifically, we examine initial accuracy, macro F1, and total cell count. As shown in Figure 5, GRIT tends to be more effective when the initial accuracy and macro F1 are relatively high. Additionally, when the initial performance is poor, large cell count number can degrade results. This observation is intuitive: reliable initial predictions ensure that neighbors provide trustworthy information, and a greater number of neighbors increases the likelihood of correcting noisy predictions through refinement. For noisy initial logits, the same propagation may reinforce incorrect patterns.

It is important to note that GRIT is not designed to achieve 100% accuracy. Rather, similar to other inference-time refinement methods [30, 22], it seeks to leverage complementary structural information in the data to bring the initial logits closer to their achievable performance limit. In our case, this is accomplished by promoting local consistency and semantic alignment with the k-NN graph. While the ultimate performance is bounded by the quality of the initial predictions, GRIT—like other inference-time refinement methods—introduces a principled and effective post hoc

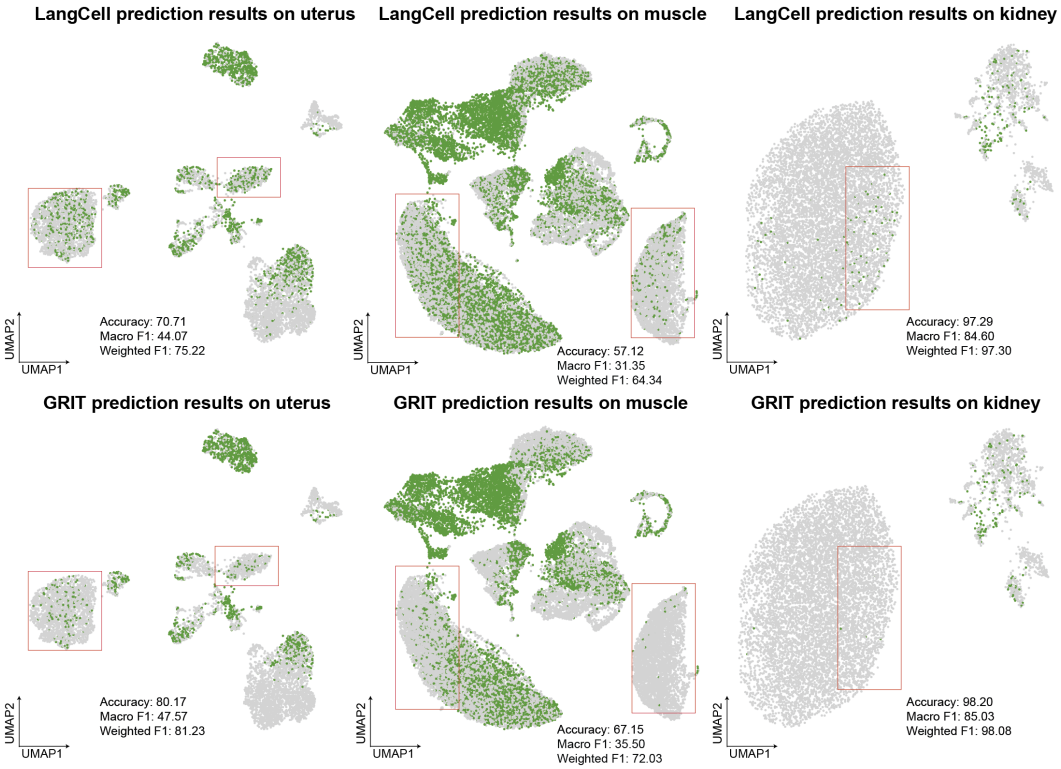

Figure 4: UMAP visualization of scRNA-seq data of organ uterus (left), muscle (middle), and kidney (right). Each point represents a cell, colored by prediction correctness: gray indicates correct predictions, and green indicates incorrect ones. For each organ, the top panel shows zero-shot predictions from LangCell, while the bottom panel shows refined predictions from GRIT. Orange boxes indicate representative regions where GRIT provides clear improvements.

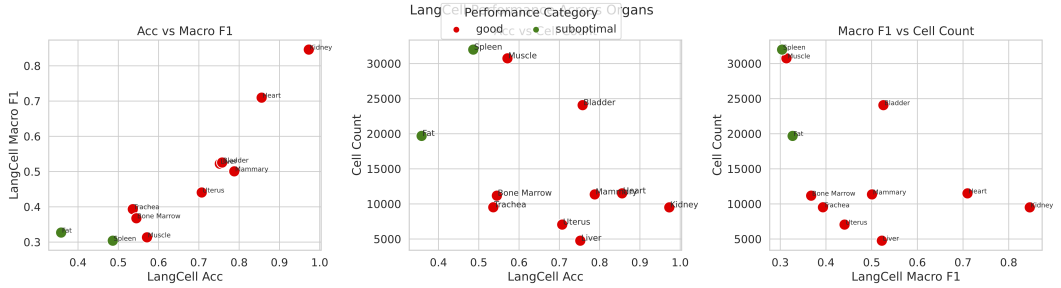

Figure 5: GRIT performance are plotted against LangCell logit initial accuracy, macro F1, and total cell count for all 11 scRNA-seq datasets used in our main experiment.

procedure that consistently yields measurable improvements in cell type annotation for scRNA-seq datasets.

5 Conclusion and Limitations

We present a biologically grounded, inference-time refinement method for improving zero-shot cell type annotation in scRNA-seq data. The core idea is to refine pre-trained models' zero-shot predictions by enforcing consistency over the task-relevant graph structure. While our method consistently improves performance, it has limitations including the assumption of a reliable graph and sensitivity to the optimization hyperparameter λ. Our experiments indicate that while moderate refinement typically enhances accuracy, the optimal λ can vary substantially depending on the quality of the initial logits (see Appendix). In our current implementation, λ is set to 1 based on extensive empirical results demonstrating consistently strong performance. Nonetheless, the ability to estimate

an optimal λ conditioned on the initial logits remains a very intriguing direction, as further gains could be achieved under better-tuned values. Developing a principled framework for adaptively selecting λ based on dataset-specific characteristics will be an important avenue for future research.

References

[1] Fuqun Chen, Guanhua Zou, Yongxian Wu, and Le Ou-Yang. Clustering single-cell multi-omics data via graph regularized multi-view ensemble learning. *Bioinformatics*, 40(4):btae169, 2024.

[2] Li-Fang Chu and *et al.* Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17:173, 2016. doi: 10.1186/s13059-016-1033-x.

[3] Zoe A Clarke, Tallulah S Andrews, Jawairia Atif, Delaram Pouyabahar, Brendan T Innes, Sonya A MacParland, and Gary D Bader. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature protocols*, 16(6):2749–2764, 2021.

[4] Ashley Mae Conard, Alan DenAdel, and Lorin Crawford. A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(5):e1617, 2023.

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

[6] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

[7] Zhanbei Cui, Tongda Xu, Jia Wang, Yu Liao, and Yan Wang. Geneformer: Learned gene compression using transformer-based context modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8035–8039. IEEE, 2024.

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

[9] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[10] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

[11] Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9:1–12, 2017.

[12] Johannes Hertrich, Lan Dang Phuong Nguyen, Jean-François Aujol, Dominique Bernard, Yannick Berthoumieu, Abdellatif Saadaldin, and Gabriele Steidl. Pca reduced gaussian mixture models with applications in superresolution. *Inverse Problems and Imaging*, 16(2):341–366, 2022.

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

[14] Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic acids research*, 51 (D1):D870–D876, 2023.

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

[16] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and translational medicine*, 12(3):e694, 2022.

[17] TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[18] Konstantinos Lazaros, Dimitris E Koumadorakis, Panagiotis Vlamos, and Aristidis G Vrahatis. Graph neural network approaches for single-cell data: a recent overview. *Neural Computing and Applications*, 36(17):9963–9987, 2024.

[19] Tianhao Li, Zixuan Wang, Yuhang Liu, Sihan He, Quan Zou, and Yongqing Zhang. An overview of computational methods in single-cell transcriptomic cell type annotation. *Briefings in Bioinformatics*, 26(3):bbaf207, 2025.

[20] Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

[21] Jiaxiang Liu, Tianxiang Hu, Huimin Xiong, Jiawei Du, Yang Feng, Jian Wu, Joey Zhou, and Zuozhu Liu. Vpl: Visual proxy learning framework for zero-shot medical image diagnosis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9978–9992, 2024.

[22] Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, Joey Tianyi Zhou, and Zuozhu Liu. Kpl: Training-free medical knowledge mining of vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18852–18860, 2025.

[23] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13 (6):47–60, 1996.

[24] Thale Kristin Olsen and Ninib Baryawno. Introduction to single-cell rna sequencing. *Current protocols in molecular biology*, 122(1):e57, 2018.

[25] Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2018.

[26] Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scrna-seq data. *Computational and Structural Biotechnology Journal*, 19:961–969, 2021.

[27] Wendell J Pereira, Felipe Marques Almeida, D Conde, KM Balmant, PM Triozzi, HW Schmidt, C Dervinis, GJ Pappas, and M Kirst. Asc-seurat: analytical single-cell seurat-based web application. *BMC bioinformatics*, 22:1–14, 2021.

[28] Hannah A Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nature methods*, 16(10):983–986, 2019.

[29] Alex A. Pollen and *et al.* Molecular identity of human outer radial glia during cortical development. *Cell*, 163:55–67, 2015. doi: 10.1016/j.cell.2015.09.004.

[30] Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems*, 36:25461–25474, 2023.

[31] Fei Quan, Xin Liang, Mingjiang Cheng, Huan Yang, Kun Liu, Shengyuan He, Shangqin Sun, Menglan Deng, Yanzhen He, Wei Liu, et al. Annotation of cell types (act): a convenient web server for cell type annotation. *Genome Medicine*, 15(1):91, 2023.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[33] Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, Lijuan Hu, Peter Lönnerberg, Trygve Bakken, Tamara Casper, Michael Clark, et al. Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667):eadd7046, 2023.

[34] Shaked Slovin, Annamaria Carissimo, Francesco Panariello, Antonio Grimaldi, Valentina Bouché, Gennaro Gambardella, and Davide Cacchiarelli. Single-cell rna sequencing analysis: a step-by-step overview. *RNA bioinformatics*, pages 343–365, 2021.

[35] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature reviews genetics*, 20(5): 257–272, 2019.

[36] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic. The role of hubness in clustering high-dimensional data. *IEEE transactions on knowledge and data engineering*, 26(3):739–751, 2013.

[37] Koki Tsuyuzaki, Hiroyuki Sato, Kenta Sato, and Itoshi Nikaido. Benchmarking principal component analysis for large-scale single-cell rna-sequencing. *Genome biology*, 21(1):9, 2020.

[38] Bram Van de Sande, Joon Sang Lee, Euphemia Mutasa-Gottgens, Bart Naughton, Wendi Bacon, Jonathan Manning, Yong Wang, Jack Pollard, Melissa Mendez, Jon Hill, et al. Applications of single-cell rna sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6):496–520, 2023.

[39] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

[40] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 968–977, 2019.

[41] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

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

[43] Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022.

[44] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19, 2017.

[45] Yijie Zhang, Dan Wang, Miao Peng, Le Tang, Jiawei Ouyang, Fang Xiong, Can Guo, Yanyan Tang, Yujuan Zhou, Qianjin Liao, et al. Single-cell rna sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research*, 40:1–17, 2021.

[46] Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, Yizhen Luo, and Zaiqing Nie. Langcell: Language-cell pre-training for cell identity understanding. In *International Conference on Machine Learning*, pages 61159–61185. PMLR, 2024.

[47] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

[48] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

[49] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

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