Learning Biological Hierarchies in Single-Cell Foundation Models

Xiangyu Guo
Duke University
xiangyu.guo@duke.edu

Ricardo Henao
Duke University
ricardo.henao@duke.edu

Abstract

Cell types are organized by taxonomic relationships that define lineages in the so called cell ontology. However, most existing foundation models typically ignore cell-type lineages encoded in this cell ontology. We introduce a biologically informed foundation model called **scOntoFM**[1], which embeds hierarchical cell ontology into representation learning. Using Lowest Common Ancestor (LCA) distances from the ontology graph, our framework pairs efficient offline triplet sampling with hierarchical cell-ontology learning. This approach jointly preserves local neighborhoods and shapes the global embedding geometry to reflect the ontology structure. Across diverse cell- and gene-level benchmarks, the model consistently improves performance, with especially strong gains in zero-shot settings. Our model also balances batch integration with biological structure preservation, providing complementary perspectives to existing embedding evaluation frameworks and supporting scalable biological discovery.

1 Introduction

Current single-cell RNA sequencing (scRNA-seq) technology has revolutionized our ability to characterize cellular heterogeneity and uncover the molecular mechanisms underlying healthy and diseased tissues [25]. Driven by advances in high-throughput sequencing and large-scale atlas construction, the field has witnessed an explosion of massive single-cell datasets in recent years [42, 14, 40, 1]. This data deluge has created an urgent demand for computational frameworks capable of fully utilizing these resources for broad biological discovery. Inspired by the transformative success of Large Language Models (LLMs) in Natural Language Processing (NLP), the community has recently developed Single-Cell Foundation Models (scFMs), such as Geneformer [45], scGPT [10], scFoundation [22], scCello [54], CellWhisperer [38], *etc*. By pretraining on large corpora of cell-level transcriptomic data, these models aim to learn generalizable cellular representations that can be transferred to diverse downstream tasks, ranging from cell-type annotation to gene-perturbation prediction [44]. We provide a detailed description of the Related Work in Appendix A.

Contrastive learning is widely adopted in current scFMs as an auxiliary objective to learn discriminative features between cell types [56, 24, 54, 13]. Among these approaches, ontology-aware methods such as SCimilarity [24] and scCello [54] further incorporate cell ontology as supervised signals during model training, leveraging online triplet sampling and vanilla contrastive objectives to encourage biologically meaningful representation learning. However, scaling such frameworks to atlas-level datasets introduces substantial computational challenges, as standard online triplet sampling is highly inefficient for massive datasets. In random mini-batches, informative positive pairs are rarely encountered (Appendix Figure (AFigure) 7), leading to weak supervision and suboptimal metric learning [18, 43]. Moreover, existing methods such as scCello mainly focus on local cell-type relational alignment guided by the ontology structure, while overlooking the global hierarchical struc-

[1]The source code and pre-trained model are available on GitHub and Hugging Face, respectively.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

ture encoded in the cell ontology [34, 57]. Consequently, the learned latent space may not faithfully reflect biological phylogeny and developmental lineages, limiting both biological interpretability and the ability to capture dynamic cellular processes [47, 50].

To address these challenges, we propose scOntoFM, a novel biologically informed foundation model that explicitly integrates hierarchical cell ontology into a representation learning framework. Cell Ontology [34] provides a curated hierarchy of cell-type identities that complements transcriptomic supervision. Consistent with prior single-cell studies linking ontology relationships to transcriptomic similarity [48, 37, 50], we used the Lowest Common Ancestor (LCA) distance [39, 5] from the cell ontology graph as an interpretable structural prior and an approximate proxy for transcriptomic distance. This prior serves as a guide for both data sampling and loss optimization in scOntoFM. Our contributions are threefold.

- **Efficient Ontology-aware Offline Sampling:** We introduce the *Cell-Dataset Index Map*, a hierarchical indexing structure for efficient offline triplet sampling across cell types and datasets. Guided by the LCA distance matrix, it prioritizes cross-dataset positives and semi-hard and hard negatives drawn from semantically close cell types in the ontology graph, thus overcoming the inefficiency of random online sampling commonly used in current scFMs, *e.g.*, scCello [54].
- **Hierarchical Structure Learning:** We incorporate the ontology structure into representation learning through a *dynamic margin triplet loss*, where the margin is determined by the LCA distance between the anchor and negative cell types and the ontology depth of the anchor cell type. In addition, a complementary *rank-order alignment loss* enforces consistency between the relative ordering of cell-type centroid distances in the embedding space and their ontology-based LCA distances within each batch. These two synergistic losses enforce both fine-grained local discrimination and global hierarchical consistency, producing an embedding space that better reflects the structure of the cell-type ontology.
- **Comprehensive Evaluation:** We benchmark our model against state-of-the-art (SOTA) methods across diverse cell- and gene-level tasks. Our results demonstrate superior generalizability, particularly in zero-shot settings. Crucially, as highlighted by recent critiques of embedding metrics (*e.g.*, the "drifting islands" phenomenon) [47], purely aligning cell embeddings with cell-type annotations will lead to biologically problematic embeddings. Our evaluation confirms that our approach achieves a superior balance between the well established scIB metrics [32], which measure the proximity of cells that belong to the same cell-type cluster, and how well cells of the same cell type mix-up across batches/datasets in single cell studies, and the more rigorous biological structure preservation metric of model embeddings [50].

2 Methods

Figure 1 provides a high-level overview of the proposed scOntoFM workflow. We first describe the pre-training data processing, the construction of the Lowest Common Ancestor (LCA) distance matrix in Section 2.1, which is followed by the offline triplet sampling, model architecture, and model pre-training objectives in Section 2.2, Section 2.3 and 2.4, respectively.

2.1 Data Processing

Single-cell Data Let $\mathcal{D} = \{x_n, c_n\}_{n=0}^{N}$ denote a single-cell RNA-seq dataset, where $x_n \in \mathbb{R}^{|\mathcal{M}|}$ is the raw gene expression vector in cell n, and $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$ represents the global gene set (vocabulary) of size $|\mathcal{M}|$ defined by our pre-training corpus. Each cell is associated with a cell type label $c_n \in \mathcal{V}$, where $\mathcal{V}$ represents the set of all cell type nodes in the cell ontology graph, which we will introduce in the next section. An overview of our pre-training data corpus is presented in Figure 1(a).

We follow the same data pre-processing steps as Geneformer [45]: first we normalize the library size of each cell to 10,000 to account for sequencing depth variation, and re-normalize each gene expression value by its non-zero median over the entire pre-training dataset, *i.e.*,

$$x'_{n,m} = 10^4 \frac{x_{n,m}}{\sum_{m' \in \mathcal{M}} x_{n,m'}}, \quad x''_{n,m} = \frac{x'_{n,m}}{\text{median}\{x'_{n,m} \mid x'_{n,m} > 0\}}, \qquad n = 1, \ldots, N,\ m \in \mathcal{M},$$

where $x''_{n,m}$ is the normalized gene expression for cell n and gene m. Following normalization, we perform rank-value encoding to generate the inputs of the model. For each cell n, we sort the

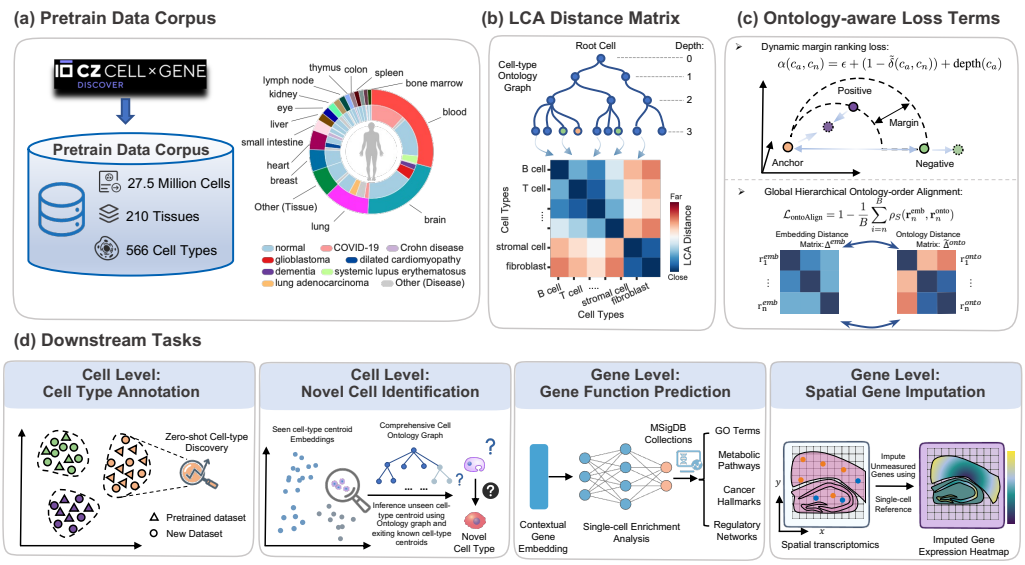

Figure 1: Proposed model workflow. (a) scOntoFM pre-training data corpus details (left), and the distribution of major tissues and diseases in the corpus (right). (b) An illustration of cell ontology graph and its derived LCA distance matrix. (c) Ontology-aware loss terms in scOntoFM, the dynamic margin $\alpha(c_a, c_n)$ is controlled by the distance between the anchor cell-type c_a and negative cell-type c_n, and the depth of the anchor c_a in the ontology graph. Down below is the hierarchical ontology alignment loss, in which Δ^{emb} and $\tilde{\Delta}^{\text{onto}} \in \mathbb{R}^{B\times B}$ represent correspondingly the distance matrix of our model embeddings and its corresponding LCA distance matrix in the batch (of size B), and where r_n^{emb} and r_n^{onto} are the soft-ranking rows in their respective distance matrices. $\rho_S(\cdot,\cdot)$ is the Spearman correlation for the soft-ranking vectors. (d) Downstream tasks of scOntoFM, from left to right and grouped into cell- and gene-level tasks as presented in the results.

normalized expression values $x''_{n,m}$ in descending order. The resulting input is a sequence of gene tokens $S_n = (m_{(1)}, m_{(2)}, \ldots, m_{(L)})$, where each token $m_{(k)} \in \mathcal{M}$ is the gene corresponding to the k-th highest expression value in cell n. We truncate this gene sequence to a maximum length of $L = 2,000$, and then append the `[CLS]` and `[EOS]` tokens at the beginning and the end of the sequence, respectively. By defining the sequence elements directly as members of $\mathcal{M}$, we ensure that the input tokens map one-to-one with the pre-defined model vocabulary.

LCA Distance Matrix Construction To quantify biological similarity, we leverage the hierarchical structure of the cell ontology graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from the Open Biological and Biomedical Ontology (OBO) Foundry [41], where $\mathcal{V}$ denotes cell-type nodes in the graph and every cell label c_i in our dataset is an instance of a node $v \in \mathcal{V}$; $\mathcal{E}$ is the directed edge between two nodes, encoding the hierarchical relationship ("is a subtype of") between cell types. We construct a pairwise distance matrix $\Delta \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ derived from the Lowest Common Ancestor (LCA) based on $\mathcal{G}$ [2]. The general concept of this process is shown in Figure 1(b). For any two types of cells $c_i, c_j \in \mathcal{V}$, their raw semantic distance $d_{\mathcal{G}}(c_i, c_j)$ is calculated as the sum of the shortest path distances from each node to their common $c_s = \text{LCA}(c_i, c_j)$:

$$\Delta_{i,j} = \delta_{\mathcal{G}}(c_i, c_j) = \delta_{path}(c_i, c_s) + \delta_{path}(c_j, c_s), \tag{1}$$

where $\delta_{path}(c_i, c_s)$ and $\delta_{path}(c_j, c_s)$ denote the number of edges on the shortest path between nodes (c_i, c_j) to their lowest common ancestor c_s in $\mathcal{G}$, respectively. To facilitate stable training with the dynamic triplet margin loss in Section 2.4, we compute the normalized distance $\tilde{\delta}(c_i, c_j)$ element wise and obtain $\tilde{\Delta}_{i,j}$, by scaling $\delta_{\mathcal{G}}(c_i, c_j)$ by the maximum distance on the ontology graph. The matrix $\tilde{\Delta}$, with elements $\tilde{\Delta}_{i,j}$, serves as the ground truth for biological distinctness during the extraction of negative pairs in Section 2.2 and to determine the dynamic margin for the triplet loss defined in Section 2.4. We provide more empirical evidence about the relationship between LCA-based distance and transcriptomic similarity in Section 4 and Appendix B.2.

2.2 Offline Triplet Sampling

Contrastive learning [8] is a representation learning framework that aims to learn an embedding space such that it encourages semantically similar samples to be closer while pushing dissimilar samples apart. In contrastive learning, an *anchor sample* serves as a reference that is used to compare, by proxy, positive and negative samples. The general concept is shown in Figure 1(c). Efficiently mining of informative triplets is critical for contrastive learning. While standard online triplet sampling, like in SCimilarity and scCello, is highly inefficient at massive scale, since random mini-batches are unlikely to contain sufficiently informative positive pairs. Here, we propose a hierarchical indexing structure, which we term the *Cell-Dataset Index Map* (Ω), to enable $O(1)$ retrieval of candidate samples. Let $\mathcal{V}$ be the set of cell types and $\mathcal{K}$ be the set of all dataset identifiers (*e.g.*, studies). We define Ω as the collection of index sets $\mathcal{I}_{c,k}$, each of which contains the indices of cells belonging to type $c \in \mathcal{V}$ and dataset $k \in \mathcal{K}$: $\Omega = \{\mathcal{I}_{c,k}\}_{c\in\mathcal{V},k\in\mathcal{K}}$, where again $\mathcal{I}_{c,k}$ denotes the set of indices for all cells of type c found in dataset k. Additionally, for any cell type c, we denote the set of datasets containing this cell type as $\mathcal{K}_c = \{k \in \mathcal{K} \mid \mathcal{I}_{c,k} \neq \emptyset\}$. For each training step, we select an anchor cell x_a with cell type c_a from dataset k_a. Specifically, we construct triplets (x_a, x_p, x_n) as follows.

Positive Sampling To promote batch mixing for the same cell type, we prioritize sampling *cross-dataset* positive pairs from different datasets. We define a pool of positive candidates, $\mathcal{P}^+$, designed to prioritize cross-dataset pairs while also handling edge cases where a cell type is unique to a single dataset. If the cell type c_a exists in multiple datasets ($|\mathcal{K}_{c_a}| > 1$), we construct the pool by taking the union of all instances from the datasets in $\mathcal{K}_{c_a}$ other than k_a: $\mathcal{P}^+ = \bigcup_{k\in\mathcal{K}_{c_a}\setminus k_a} \mathcal{I}_{c_a,k}$, and then we uniformly sample $x_p \sim \text{Uniform}(\mathcal{P}^+)$. If the cell type only exists in the anchor's dataset (*i.e.*, $\mathcal{P}^+ = \emptyset$), we default to sampling x_p from the anchor's own dataset index $\mathcal{I}_{c_a,k_a}$.

Negative Sampling Randomly selected negative samples (*e.g.*, "T cell" *vs.* "Neuron") are often too easy to distinguish, producing zero loss and vanishing gradients. To ensure robust learning, we specifically mine "semi-hard" to "hard" negatives—cell types that are biologically distinct but semantically close to the anchor.

We utilize the Cell Ontology graph to measure semantic proximity. Recall that Δ is the pre-computed LCA distance matrix between all cell types. For an anchor with cell-type c_a, we first define the set of *valid negative cell types*, denoted as $\mathcal{V}_\text{n}$, excluding the anchor type c_a and its lineage (ancestors and descendants) to avoid conflicting training signals (ambiguous pairs) during training: $\mathcal{V}_\text{n} = \{v \in \mathcal{V} \mid v \neq c_a \wedge v \notin \text{Lineage}(c_a)\}$. To identify the most informative negatives, we rank all valid types based on their proximity to the anchor. Let $\text{rank}(c_a, v)$ denote the ordinal rank of the anchor cell type c_a and all other valid negative cell types $v \in \mathcal{V}_\text{n}$ when sorted by ascending LCA distance $\Delta(c_a, v)$. We select the set of *hard to semi-hard negative candidates* by following: $\mathcal{C}_\text{hard} = \{v \in \mathcal{V}_\text{n} \mid \text{rank}(c_a, v) \leq M\}$, and we set $M = 50$ by default, which provides a diverse set of closely-related cell types to sample from, stabilizing training while maintaining optimization "difficulty". By discriminating against these biologically similar types, the model learns to capture high-resolution lineage markers rather than just broad lineage differences. Then the negative cell instance x_n is uniformly sampled from: $x_n \sim \bigcup_{c\in\mathcal{C}_\text{hard}} \mathcal{I}_{c,k}$.

Since the collection of index sets Ω pre-groups indices, we avoid costly filtering operations during training. The sampling logic described above is implemented as a vectorized look-up operation. We pre-compute the valid positive dataset indices and negative cell type candidates for every batch, allowing us to generate millions of triplets in parallel using efficient tensor gathering operations on the GPU. To support this empirically, the computational cost of constructing the *Cell-Dataset Index Map* is summarized in ATable 8.

2.3 Model Architecture

We adopt a transformer architecture similar to the Geneformer [45]. The model consists of 6 transformer layers with 256 hidden units and 4 attention heads. The input to the model is the rank-ordered sequence of gene tokens S_n, which is embedded as follows: $\mathbf{H}_0 = \text{E}(S_n) + \mathbf{E}_{pos}$, where $\text{E}(\cdot)$ is the gene-token embedding layer, $\mathbf{E}_{pos}$ represents the learnable positional embedding, and the embedding $\mathbf{H}_0$ is the embedded input used by the model. A special `[CLS]` token is added to the beginning of the sequence to aggregate global cell representations, and `[EOS]` is added to the end of the input gene sequence S_n. We use the output of the final hidden layer of the transformer model associated with the `[CLS]` token, denoted as $\mathbf{z}_n$, as the low-dimensional cell embedding.

2.4 Learning Objective

We optimize our model using three objectives, namely i) the standard MLM term, ii) the dynamic margin triplet loss, and iii) the global ontology alignment loss. The latter two loss objectives leverage the LCA distance matrix to guide the model to learn biologically meaningful representations. Compared with existing methods such as scCello [54], which employs a fixed scaling parameter to learn ontology-informed pairwise relational alignment in its contrastive loss and focuses on local cell-type relationships, our approach uses a dynamic margin triplet loss to enable finer-grained discrimination of cell states, while the rank-order alignment loss further preserves the global hierarchical structure in the embedding space.

Masked Language Modeling (MLM) Like in BERT [12], we apply the MLM strategy with our input sequence S_n, and the masking ratio is set to 15%. For the selected masking position, we apply the same masking rule as used in BERT, (i) there is 80% chance that the token is replaced by the `[MASK]` token; (ii) 10% probability that the token will be replaced by a random token; and (iii) 10% chance that the token will remain unchanged. Let $\mathcal{J}_{mask} \subset 1, \ldots, L$, denote the set of masked position indices. We write the MLM loss as follows

$$\mathcal{L}_{\text{MLM}} = -\sum_{j \in \mathcal{J}_{mask}} \log P(t_j | S_{n \setminus \mathcal{J}_{mask}}), \tag{2}$$

where t_j is the original token in the masked position and $S_{n \setminus \mathcal{J}_{mask}}$ represents the masked sequence. This objective enables the model to learn gene–gene co-expression patterns that can be used to reconstruct missing gene expression values as previously described in [45].

Dynamic Ontology-Aware Triplet Loss To enforce the biological hierarchy in the embedding space, we employ a triplet loss with a dynamic margin $\alpha(\cdot, \cdot)$ (Figure 1(c) top panel). Specifically, $\alpha(\cdot, \cdot)$ adapts to the specific semantic relationship between the anchor and the negative sample based on their positions in the cell-type ontology instead of setting a fixed margin as in the original triplet loss [18]. Specifically, we define the dynamic margin $\alpha(c_a, c_n)$ as:

$$\alpha(c_a, c_n) = \epsilon + (1 - \tilde{\delta}(c_a, c_n)) + \text{depth}(c_a), \tag{3}$$

where ϵ is a small constant offset (set to 0.1 by default), $\tilde{\delta}$ denotes the normalized LCA distance between cell types c_a and c_n, and $\text{depth}(c_a)$ is the normalized depth of the anchor node in the ontology graph. This design encourages a larger separation margin for semantically similar cell types, *i.e.*, small $\tilde{\delta}$, leading to larger margin, and for anchors corresponding to fine-grained cell types, *i.e.*, large depth, as well. The offset term ϵ ensures a minimum separation for shallow and biologically distinct nodes, in which these cases are easily distinguishable by the model. The resulting triplet loss function is formulated as follows.

$$\mathcal{L}_{\text{Tri}} = \max\left(0, \|\mathbf{z}_a - \mathbf{z}_p\|_2^2 - \|\mathbf{z}_a - \mathbf{z}_n\|_2^2 + \alpha(c_a, c_n)\right). \tag{4}$$

Consequently, this design forces the model to learn more discriminative representations that capture subtle biological distinctions, *i.e.*, CD4^+ T cells and CD8^+ T cells, relative to naive T cells.

Global Cell-type Ontology Alignment Although the dynamic margin triplet loss effectively enforces *local discrimination*—ensuring that fine-grained, semantically similar subtypes are pushed apart to prevent collapse—, however, it does not inherently preserve the broader cell-type hierarchical lineage in a mini-batch. Here, we introduce a complementary objective that aligns the geometry of the embedding space with the global hierarchical relationships encoded by the cell. Concretely, we propose a *Rank-Order Alignment Loss* that encourages consistency between (i) pairwise distances in the learned embedding space, and (ii) reference ontological distances, by matching their relative (rank) order (see Figure 1(c) lower panel).

Let $\mathbf{Z} \in \mathbb{R}^{B \times d}$ denote cell embeddings in a mini-batch, where B is the batch size and d is the embedding dimension, which we set to $d = 256$ in our case. We firstly compute the pairwise cosine distance matrix in the embedding space in the current batch: $\Delta^{\text{emb}} = 1 - \frac{\mathbf{Z}\mathbf{Z}^\top}{\|\mathbf{Z}\|_2 \|\mathbf{Z}^\top\|_2}$. Following this, we get the embedding distance matrix $\Delta^{\text{emb}} \in \mathbb{R}^{B \times B}$ in the batch. Regarding the reference ontology distances, recall that $\tilde{\Delta}$ is the normalized ontology LCA distance, here we denote $\tilde{\Delta}^{\text{onto}} \in \mathbb{R}^{B \times B}$ as the corresponding LCA ontology distance sub-matrix in the current batch. Let Δ_n^{emb} and $\tilde{\Delta}_n^{\text{onto}}$ be the vectors obtained from the n-th row of Δ^{emb}, and $\tilde{\Delta}^{\text{onto}}$, respectively. Our goal is to make the ranking of neighbors induced by Δ_n^{emb} match the ranking induced by $\tilde{\Delta}_n^{\text{onto}}$, *i.e.*, to maximize the row-wise

Spearman rank correlation between the two distance vectors. Since the Spearman correlation involves a non-differentiable sorting/ranking operation, we adopt the differentiable soft-ranking operator [6]. Specifically, we compute soft rank vectors:

$$\mathbf{r}_n^{\text{emb}} = \mathcal{R}_{\text{soft}}(\Delta_n^{\text{emb}}), \quad \mathbf{r}_n^{\text{onto}} = \mathcal{R}_{\text{soft}}(\tilde{\Delta}_n^{\text{onto}}),$$

where $\mathcal{R}_{\text{soft}}(\cdot)$ denotes the differentiable ranking operator. We then define the global ontology alignment objective as follows.

$$\mathcal{L}_{\text{ontoAlign}} = 1 - \frac{1}{B}\sum_{i=n}^{B} \rho_S(\mathbf{r}_n^{\text{emb}}, \mathbf{r}_n^{\text{onto}}), \tag{5}$$

where $\rho_S(\cdot,\cdot)$ is the Spearman rank correlation. Minimizing $\mathcal{L}_{\text{ontoAlign}}$ encourages the learned embedding space to preserve, for each cell, the relative ordering of ontological distances within each batch, thus aligning the global embedding structure with the hierarchical relationships defined by the ontology.

Complete Objective The complete loss function is a weighted sum of the three terms in (2), (4) and (5): $\mathcal{L}_{Total} = \mathcal{L}_{\text{MLM}} + \lambda_1\mathcal{L}_{\text{Tri}} + \lambda_2\mathcal{L}_{\text{ontoAlign}}$, where λ_1 and λ_2 control the contribution of local and global ontology structure preserving losses, respectively. For simplicity, we set $\lambda_1 = \lambda_2 = 1$ in all of our experiments. The dynamic triplet margin loss enables us to discriminate subtle cell-state differences, while the ontology alignment loss imposes a global hierarchical structure on the learned representation to preserve biological fidelity. The trade-off and synergy between these complementary objectives are explicitly analyzed in the ablation study in Section 4.

3 Experiments

3.1 Pre-Training

Pretrained Dataset Construction We downloaded 234 *Homo sapiens* scRNA-seq datasets from the CZI [1] data portal (2025-01-30 release), covering 210 tissues, 566 cell types, 58 disease types in the pretraining phase (see Figure 1(a)). All datasets are from *Homo Sapiens* and we retained only those sequencing using the 10X platform. We filtered out low-quality cells with fewer than 200 detected genes and genes expressed in fewer than 5 cells, and excluded malignant cells, non-primary samples, and cells whose annotated cell types are not present in the Cell Ontology graph. After such filtering, approximately 27.5 million cells remained.

Pretrained Model Architecture and Configurations We adopt the same BERT-style architecture as scCello [54] and Geneformer [45], with six Transformer layers and four attention heads per layer, a 512-dimensional intermediate layer and a 256-dimensional hidden size. We pretrain the model using AdamW (learning rate 1e-3, weight decay 0.001) with a warm-up of approximately 1% of the total training steps. Training is performed for three epochs on six NVIDIA H200 GPUs with a batch size of 64 per GPU and a gradient accumulation step of 4 (effective batch size = 64 × number of GPUs × 4); the complete training process takes approximately 30 hours. Additional training details and hyperparameter settings are provided in Appendix B.3.

3.2 Downstream Task Settings

In the following, we describe the state-of-the-art (SOTA) baselines and provide details on the datasets and evaluation metrics used for each downstream task. Additional information on the baselines and evaluation metrics is provided in Appendix B.5 and Appendix B.11.

Benchmark baselines We benchmark our model against five SOTA scFMs, namely, Geneformer [45], scGPT [10], scCello [54], SCimilarity[24] and CellPLM [49], on four downstream tasks spanning both cell- and gene-level evaluations. The illustration of the downstream tasks considered is shown in Figure 1(d).

Cell-type Annotation To benchmark zero-shot cell-type clustering, we evaluate on nine datasets excluded from pretraining, spanning blood [19], heart [26], liver [15], kidney [28], pancreas [4], lung [27], breast [20], adipose tissue [16], and bone marrow [55], with sizes ranging from 30k to 348k cells (Appendix B.6). Specifically, the HIHA (blood), kidney and Azimuth PBMC [23] datasets are selected for model robustness analysis in cell-type annotation, as they provide author-annotated labels

spanning multiple levels of granularity, from coarse to fine. We further construct an evaluation set of 340 overlapping cell types from 212 unseen datasets (ATable 11). Performance is assessed using scIB metrics [32], including ARI, NMI, and ASW_{label}, and their arithmetic mean (AvgBio), along with $\text{ASW}_{label/batch}$ and GraphConn, and their mean (AvgBatch). The overall scIB score is calculated as a weighted average of these two aggregated metrics (0.6×AvgBio + 0.4×AvgBatch). We also report scGraph-RWR and scGraph-LCA [47, 50] to evaluate the preservation of biological structure captured by model embeddings. For cell-type annotation fine-tuning (Appendix B.7), we use the pancreas and HIHA datasets, reporting accuracy and macro-F1 for the predicted labels. Moreover, we compare with hierarchical cross-entropy (HCE) classifiers [11], using the authors' released MLP and TabNet checkpoints.

Novel Cell Identification We use the same 212 unseen datasets as previously mentioned in the cell-type annotation task. However, here we focus on cell types that are absent from pretraining. After filtering rare types (<2,000 cells) and subsampling abundant ones (>10,000 cells), the benchmark contains 99 novel cell types and 774,215 cells (ATable 13, Appendix B.8). Performance is measured using per-cell-type accuracy and macro-accuracy to account for class imbalance.

Gene Function Prediction We evaluate both static and contextual gene embeddings. For static embeddings, Gene Ontology (GO) gene sets from MSigDB [31] are grouped into Biological Process (BP), Cellular Component (CC), and Molecular Function (MF), and evaluated separately. For contextual embeddings, we conduct a case study on lung adenocarcinoma [27] using *decoupleR* [3]. Details are provided in Appendix B.9. We report micro/macro precision and micro/macro F1 scores.

Cross-modal Spatial Gene Imputation We evaluate spatial gene imputation using scRNA-seq references from lung cancer [27], breast cancer [51], and brain tissue from SEA-AD [17], paired with corresponding spatial datasets. Additional details are summarized in Appendix B.10. Performance is quantified using Pearson, Spearman, and Kendall's τ correlations between imputed and ground-truth spatial expression [35].

4 Results

We first examine the relationship between cell ontology distance and transcriptomic distance to assess the biological motivation for ontology-guided learning. We then evaluate scOntoFM on cell-level tasks, including cell-type annotation and novel cell-type identification, followed by gene-level tasks, namely, gene function prediction and spatial gene imputation. Finally, we analyze the contributions of the training objectives through ablation studies. For all tables reported in the following, **bold** indicates the best performance and underlining denotes the second-best results.

Relationship Between Ontology and Transcriptomic Distances To examine the relationship between ontology proximity and transcriptional similarity, we compared LCA distances with the pairwise cosine distances between cell-type centroids in PCA space, using Tabula Sapiens v2 and the nine datasets used for zero-shot clustering. As shown in Figure 2, the mean transcriptomic distance between cell types increased monotonically over the first ten LCA distance levels (Pearson $r = 0.91$, Spearman $\rho = 1$), then plateaued (AFigure 9). Consistent with this trend, we observed positive correlations across all nine zero-shot clustering datasets, with Pearson r ranging from 0.31 to 0.58 and Spearman ρ from 0.27 to 0.60 (Table 1). Together, these results indicate that ontology-near cell types tend, on average, to be more transcriptionally similar, supporting LCA distance as an informative structural prior, while its limitations and more details are discussed in Appendix B.2.

Cell Type Annotation (Zero-shot Clustering) We first evaluate the performance of our model with SOTA methods across nine datasets of diverse tissues. As shown in the radar plot of Figure 3 and ATables 16 and 17, scOntoFM achieves a mean scIB score comparable to SCimilarity (0.7958 *vs.* 0.7960) and above the remaining baselines. We then investigate performance using biologically structure preserving metrics (scGraph-LCA and scGraph-RWR) in Figure 3 right panel, ATables 18 and 19. Notably, our model achieves substantially higher performance than SOTA methods on these two metrics across most tissues, highlighting the complementary nature of biological conservation *versus* scIB metrics. To further illustrate this, we examined the cell-type embedding distance heatmap of Bone Marrow tissue in AFigure 11(c), which aligns well with the LCA-based ontology distance heatmap as well as the RWR-based ontology distance heatmap (AFigure 12). Specifically, we investigate the hematopoietic stem cell and mark it as the rooted cell type in the distance umap plot (see AFigure 11(a,b,d)), which precisely illustrates the developmental lineage of hematopoietic and

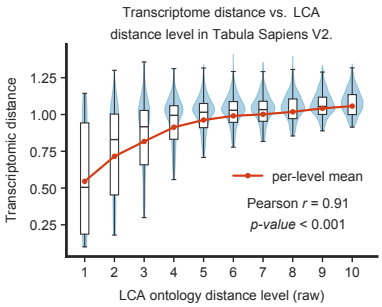

Figure 2: Cell-type pairwise transcriptomic distance in PCA space across ten LCA ontology levels in Tabula Sapiens v2. Boxes show the distance distributions, and the red line denotes the per-level mean.

Table 1: Correlation between LCA distance and cell-type centroids cosine distance in PCA space over the nine zero-shot clustering datasets. "Overall" pools pairs across all the datasets.

Dataset	Pearson r	Spearman ρ	# Cell types
HIHA	0.58	0.60	38
Heart	0.47	0.43	12
Lung	0.51	0.55	10
Liver	0.37	0.29	18
Kidney	0.31	0.27	33
Pancreas	0.38	0.30	12
Breast TNBC	0.54	0.31	6
Adipose tissue	0.35	0.28	16
Bone marrow	0.39	0.36	40
Overall	0.41	0.40	130

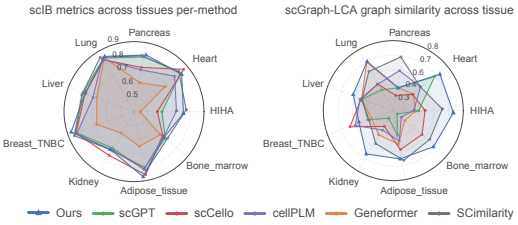

Figure 3: Radar plots of zero-shot clustering benchmark across nine tissues in scIB (left) and scGraph-LCA (Spearman correlation) (right) metrics across six SOTA methods.

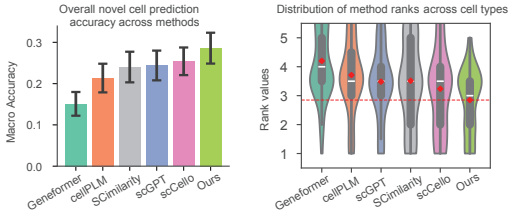

Figure 4: Bar plot of macro-accuracy for novel cell identification (left) and violin plot of per-cell type accuracy rank distributions (right); ♦ denotes each method's average rank.

myeloid-related cell types, as well as aligns well with G2M and S phase cell cycle scores reported in the original study [55].

We further evaluate the robustness of our model embeddings (Table 2) under varying annotation granularity in HIHA and Kidney dataset, which contain author annotations from coarse to fine-grained labels. We calculate the agreement scores (ARI) (ATable 25) and visualize the cell-type annotation flow in sankey plot (AFigure 17, 18)) between the cell ontology reference and different annotation levels. Leveraging its ontology-aware design, scOntoFM achieves competitive clustering across varying resolutions, particularly at finer annotation levels (*e.g.*, AIFI_L2/L3 and Subclass_L2), indicating its superior ability to capture subtle cellular states. Although performance is comparatively lower on coarse-grained annotations (*e.g.*, AIFI_L1 and Subclass_L1), this is expected due to their weaker alignment with the fine-grained Cell Ontology labels used during pretraining.

To assess the generalization of our model in real world scenarios, we evaluated 340 cell types present in the pre-training cell-type labels but sourced from entirely new, unseen datasets. Although our method shows comparable performance to cellPLM in standard scIB metrics (ATable 20), it outperforms cellPLM by a significant margin in scGraph-LCA. Collectively, these results demonstrate that our embeddings better preserve cell-type ontology structure and exhibit superior biological conservation, a critical aspect often under-emphasized in prior foundation model evaluation.

Cell Type Annotation (Fine-tuning) We conduct cell type classification benchmarks on the HIHA and Pancreas datasets, with fixed held-out donors and identical hyperparameters for all the and HCE classifiers. Averaged over five random seeds, scOntoFM achieves the best overall accuracy and maintains strong Macro-F1 scores (Table 3), indicating its effectiveness in learning discriminative and generalizable representations for accurate cell type classification beyond zero-shot clustering. To further evaluate our model robustness to label noise, we progressively inject noise into the training annotations of the well-annotated Azimuth PBMC dataset by randomly replacing fine-grained labels with their corresponding coarser categories. scOntoFM demonstrates highly stable and accurate cell annotation even under increasing noise levels (ATable 26). When the noise ratio reaches 50%, the model reliably falls back to predicting broader parent categories. This behavior is consistent with biological hierarchies, ensuring structurally coherent predictions despite unreliable fine-grained annotations.

Table 2: Evaluation of zero-shot clustering robustness under varying annotation levels using the scIB score in HIHA and Kidney datasets.

	HIHA			Kidney	
Model	**AIFI_L1**	**AIFI_L2**	**AIFI_L3**	**Subclass_L1**	**Subclass_L2**
Ours	0.681	**0.782**	**0.671**	0.740	**0.648**
Geneformer	0.510	0.473	0.398	0.585	0.520
scGPT	0.670	0.622	0.563	0.713	0.631
scCello	0.701	0.600	0.506	**0.770**	0.639
cellPLM	**0.760**	0.722	0.666	0.721	0.640

Table 3: Cell-type annotation benchmark (fine-tuning) on HIHA and Pancreas datasets across six SOTA methods.

	HIHA		Pancreas	
Model	**Accuracy**	**Macro F1**	**Accuracy**	**Macro F1**
Ours	**0.915** $\pm$ 0.00	0.771 $\pm$ 0.01	**0.992** $\pm$ 0.00	**0.969** $\pm$ 0.01
Geneformer	0.701 $\pm$ 0.01	0.395 $\pm$ 0.01	0.935 $\pm$ 0.00	0.840 $\pm$ 0.01
scGPT	0.839 $\pm$ 0.00	0.613 $\pm$ 0.02	0.973 $\pm$ 0.00	0.919 $\pm$ 0.01
scCello	0.811 $\pm$ 0.01	0.543 $\pm$ 0.01	0.977 $\pm$ 0.00	0.960 $\pm$ 0.00
cellPLM	0.907 $\pm$ 0.00	**0.792** $\pm$ 0.01	0.968 $\pm$ 0.00	0.902 $\pm$ 0.01
HCE-TabNet	0.910 $\pm$ 0.00	0.710 $\pm$ 0.02	0.980 $\pm$ 0.00	0.900 $\pm$ 0.01

Table 4: Gene function prediction performance across SOTA methods. Classification accuracy (mean $\pm$ std) is calculated over 20-fold cross validation, best values are highlighted in **bold**.

Method	F1 Micro	F1 Macro	AP Micro	AP Macro
Ours	$\mathbf{0.917 \pm 0.001}$	$\mathbf{0.916 \pm 0.001}$	0.921 ± 0.002	0.900 ± 0.002
Geneformer	0.912 ± 0.001	0.911 ± 0.001	$\mathbf{0.922 \pm 0.002}$	$\mathbf{0.902 \pm 0.003}$
scCello	0.913 ± 0.000	0.911 ± 0.000	0.916 ± 0.002	0.891 ± 0.003
scGPT	0.908 ± 0.001	0.906 ± 0.001	$\mathbf{0.922 \pm 0.002}$	$\mathbf{0.902 \pm 0.003}$

Table 5: Spatial gene imputation performance on Lung and Breast cancer datasets. Pearson, Spearman, and Kendall's τ correlation are calculated over 10-fold cross validation on held-out gene sets, best values are in **bold**.

	Lung cancer			Breast cancer		
Method	Pearson	Spearman	Kendall's τ	Pearson	Spearman	Kendall's τ
Ours	$\mathbf{0.319 \pm 0.204}$	$\mathbf{0.282 \pm 0.195}$	$\mathbf{0.223 \pm 0.148}$	$\mathbf{0.388 \pm 0.198}$	$\mathbf{0.369 \pm 0.198}$	$\mathbf{0.279 \pm 0.147}$
Geneformer	0.314 ± 0.204	0.275 ± 0.198	0.218 ± 0.150	0.369 ± 0.184	0.357 ± 0.198	0.270 ± 0.147
CellPLM	0.272 ± 0.190	0.225 ± 0.168	0.179 ± 0.128	0.152 ± 0.107	0.125 ± 0.097	0.093 ± 0.072
scCello	0.274 ± 0.187	0.233 ± 0.175	0.185 ± 0.131	0.105 ± 0.098	0.101 ± 0.092	0.075 ± 0.070
scGPT	0.308 ± 0.201	0.255 ± 0.191	0.202 ± 0.144	0.290 ± 0.176	0.271 ± 0.168	0.201 ± 0.121

Novel Cell Identification To evaluate the hierarchical representations of known cell types and their generalizability to unseen cell types, we consider a more challenging setting: *novel cell-type identification*. We leverage the full cell-ontology LCA distance matrix to infer embeddings for unseen cell-type centroids via a weighted average of seen centroid (see Appendix B.8 for details). First, we examine the biological structure coherence of *seen* cell types. As visualized in AFigure 13(a, b), the centroid embeddings exhibit distinct separation among major lineages, including endothelial cells, leukocytes, myeloid cells, and hematopoietic oligopotent progenitors, confirming that the model captures biologically meaningful embeddings. Quantitatively, as shown in Figure 4, our method outperforms all SOTA baselines, achieving an overall macro-accuracy of 28.59% across 99 novel cell types (ATable 13). And the per-cell-type rank distribution reveals that our model attains the lowest average rank (2.85) across the 99 unseen types—an improvement of 13.65% over the second-best method, scCello (3.24). This underscores the model's superior capability in recognizing previously unseen biological states.

To thoroughly examine this ability, we selected three unseen cell types from the hematopoietic lineage where scOntoFM achieved high predictive accuracy: megakaryocyte progenitor cells, immature neutrophils, and CD14-positive/CD16-low monocytes (AFigure 13(c, d)). We retrieved the top 5 closest *seen* cell types for each of these unseen targets from the pretraining dataset and performed co-clustering with the unseen cell types. As shown in AFigure 14(a, b, c), the unseen cell types cluster proximally to their biologically related seen counterparts in the embedding space. Crucially, this embedding proximity is consistent with the LCA distance heatmap (AFigure 14(d)), which validates that the model successfully extrapolates the correct ontological neighborhood for novel cells.

Gene Function Prediction scFMs inherently learn informative gene embeddings from diverse cellular contexts and disease conditions. In this section, we systematically explore the biological knowledge encoded by these embeddings; specifically, we compare the *static* gene representations learned by scFMs to make functional predictions about GO terms, namely, BP, CC, and MF. As shown in Table 4, our method achieves the highest classification performance for all baselines, achieving top scores in both Micro-F1 (0.917±0.001) and Macro-F1 (0.916±0.001) using 20-fold cross validation. This indicates that our model offers a superior balance of precision and recall when assigning functional labels. Furthermore, in terms of Average Precision (AP), our method remains highly competitive, performing on par with Geneformer and scGPT (*e.g.*, Micro AP of 0.921 *vs.* 0.922) and outperforming scCello. These results demonstrate that our biologically informed training strategy enables the model to capture robust, semantically meaningful gene representations that generalize well to standard functional annotation tasks. We also show GO namespace classification results by aggregating the corresponding gene embeddings for each GO term in ATable 21.

Beyond static representations, we evaluate the model's capacity to generate context-aware gene embeddings by applying our model on a metastatic lung adenocarcinoma dataset [27]. We focus on the *GLYCOLYSIS* hallmark gene set to visualize how the model captures pathway activity across different cellular contexts, and we also visualize the enrichment of score of *COAGULATION* and *ALLOGRAFT_REJECTION* hallmarks in AFigure 15. As shown in Figure 5, the enrichment scores for the *GLYCOLYSIS* hallmark are notably elevated in malignant cell populations, as well as cancer-associated fibroblasts (CAFs), and tumor-associated macrophages (TAMs). This pattern reflects the

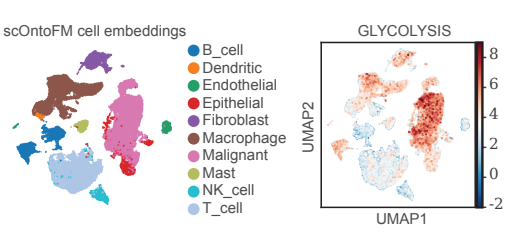

Figure 5: scOntoFM-derived cell-type clustering (Left) in lung cancer and the enrichment heatmap (Right) for the *GLYCOLYSIS* hallmark gene set.

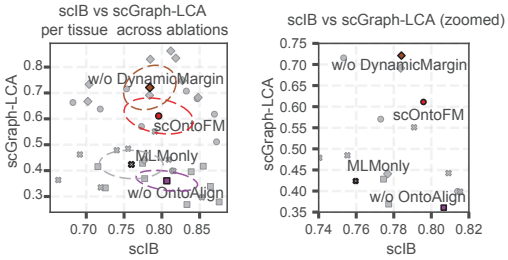

Figure 6: Pareto analysis of scOntoFM loss components in scIB and scGraph-LCA. Highlighted points are the mean across nine tissues.

high proliferative and metastatic activity of tumor cells and the immunosuppressive phenotype of associated immune cells within the tumor microenvironment [36].

Cross-modal Spatial Gene Imputation High-resolution spatial transcriptomics technologies are often constrained by the limited number of detectable genes. To address this challenge, we extend our framework to a cross-modal setting that leverages single-cell reference data to impute unmeasured genes in spatial datasets, a critical task to enhance spatial resolution [30, 21]. As shown in Table 5, our model consistently outperforms all baseline methods across all evaluation metrics. In particular, we achieve the highest Pearson, Spearman, and Kendall's τ correlations in both the Lung and Breast cancer dataset, outperforming the nearest rival (Geneformer). We further evaluate our method on an independent SEA-AD brain spatial transcriptomics dataset (ATable 22, 16). Despite being surpassed only by CellPLM which was trained on spatial data, our model consistently outperforms other single-modality baselines, demonstrating scOntoFM's strong cross-modal generalization by recovering meaningful spatial patterns.

Ablation Study We conduct an ablation study to assess the contributions of the MLM loss, dynamic margin ranking loss, and ontology alignment (Figure 6, ATable 23, ATable 24) in the zero-shot clustering task across nine tissues. To ensure fair comparison, all ablated variants are pretrained on the same 27.5M-cell corpus as scOntoFM. Removing ontology alignment improves scIB performance but leads to a substantial degradation in scGraph-LCA, indicating weakened ontology consistency. In contrast, excluding the dynamic margin ranking loss enhances scGraph-LCA but slightly reduces scIB, suggesting complementary roles of these two components. Although individual components favor specific metrics, the full model (scOntoFM) achieves a balanced trade-off, simultaneously maintaining strong scIB performance and substantially improving scGraph-LCA. These results demonstrate that jointly optimizing dynamic margin ranking and ontology alignment is essential to balancing integration quality and ontology-aware structural consistency.

5 Conclusion

We presented scOntoFM, a novel foundation model that integrates hierarchical cell ontologies directly into the pretraining framework. Through the use of efficient offline triplet sampling and a dynamic, hierarchy-aware loss function, our approach addresses key limitations in existing methods regarding biological interpretability. Our results show that scOntoFM not only achieves SOTA performance across diverse downstream tasks, but also significantly improves the preservation of biological structure, offering a more robust tool for single-cell analysis.

Limitations This work has several limitations. i) scOntoFM depends on the quality of the cell ontology and source annotations: incomplete coverage, uneven granularity, and inconsistent labels can introduce ambiguous supervision and affect the learned representations. ii) triplet sampling relies on ontology relationships without jointly considering transcriptomic similarity. Consequently, it may overlook context-specific variation, as closely related cell types can exhibit distinct expression patterns across tissues, diseases, or developmental stages. iii) pretraining requires substantial computational resources and curated annotations, while the storage requirements of the *Cell-Dataset Index Map* increase with corpus size. Future work will focus on developing a unified cell embedding space that supports harmonized, ontology-grounded annotations across cell atlases, together with covariate-conditioned sampling. These extensions aim to preserve a shared global structure while capturing context-specific variation and disease- and development-associated changes in cell state.

References

[1] Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. Cz cell×gene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. November 2023. doi: 10.1101/2023.10.30.563174. URL http://dx.doi.org/10.1101/2023.10.30.563174.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. On finding lowest common ancestors in trees. In *Proceedings of the fifth annual ACM symposium on Theory of computing - STOC '73*, STOC '73, page 253–265. ACM Press, 1973. doi: 10.1145/800125.804056. URL http://dx.doi.org/10.1145/800125.804056.

[3] Pau Badia-i Mompel, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H Holland, Ricardo O Ramirez Flores, and Julio Saez-Rodriguez. decoupler: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, 2(1), January 2022. ISSN 2635-0041. doi: 10.1093/bioadv/vbac016. URL http://dx.doi.org/10.1093/bioadv/vbac016.

[4] Khushdeep Bandesh, Efthymios Motakis, Siddhi Nargund, Romy Kursawe, Vijay Selvam, Redwan M Bhuiyan, Giray Naim Eryilmaz, Sai Nivedita Krishnan, Cassandra N. Spracklen, Duygu Ucar, and Michael L. Stitzel. Single-cell decoding of human islet cell type-specific alterations in type 2 diabetes reveals converging genetic- and state-driven -cell gene expression defects. January 2025. doi: 10.1101/2025.01.17.633590. URL http://dx.doi.org/10.1101/2025.01.17.633590.

[5] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.

[7] Junfan Chen, Fabian Schmidt, and Ricardo Henao. Assessing scale and predictive diversity in models for single-cell transcriptomics based on geneformer. *PLOS Computational Biology*, 22(7):e1013701, 2026. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1013701. URL http://dx.doi.org/10.1371/journal.pcbi.1013701.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 9(4):483–493, December 2024. ISSN 2157-846X. doi: 10.1038/s41551-024-01284-6. URL http://dx.doi.org/10.1038/s41551-024-01284-6.

[10] 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, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL http://dx.doi.org/10.1038/s41592-024-02201-0.

[11] Sebastiano Cultrera di Montesano, Davide D'Ascenzo, Srivatsan Raghavan, Ava P. Amini, Peter S. Winter, and Lorin Crawford. Improving atlas-scale single-cell annotation models with

hierarchical cross-entropy loss. *Nature Computational Science*, 6(3):243–249, January 2026. ISSN 2662-8457. doi: 10.1038/s43588-025-00945-z. URL http://dx.doi.org/10.1038/s43588-025-00945-z.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[13] Jiayuan Ding, Jianhui Lin, Shiyu Jiang, Yixin Wang, Ziyang Miao, Zhaoyu Fang, Jiliang Tang, Min Li, and Xiaojie Qiu. Tabula: A tabular self-supervised foundation model for single-cell transcriptomics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Vk2sfKAdeu.

[14] Joseph R. Ecker, Daniel H. Geschwind, Arnold R. Kriegstein, John Ngai, Pavel Osten, Damon Polioudakis, Aviv Regev, Nenad Sestan, Ian R. Wickersham, and Hongkui Zeng. The brain initiative cell census consortium: Lessons learned toward generating a comprehensive brain cell atlas. *Neuron*, 96(3):542–557, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.10.007. URL https://www.sciencedirect.com/science/article/pii/S0896627317309364.

[15] Rachel D. Edgar, Diana Nakib, Damra Camat, Sai Chung, Patricia Lumanto, Jawairia Atif, Catia T. Perciani, Xue-Zhong Ma, Cornelia Thoeni, Nilosa Selvakumaran, Justin Manuel, Blayne Sayed, Koen Huysentruyt, Amanda Ricciuto, Ian McGilvray, Yaron Avitzur, Gary D. Bader, and Sonya A. MacParland. Single-cell atlas of human pediatric liver reveals age-related hepatic gene signatures. *Hepatology Communications*, 9(11), October 2025. ISSN 2471-254X. doi: 10.1097/hc9.0000000000000813. URL http://dx.doi.org/10.1097/HC9.0000000000000813.

[16] Margo P. Emont, Christopher Jacobs, Adam L. Essene, Deepti Pant, Danielle Tenen, Georgia Colleluori, Angelica Di Vincenzo, Anja M. Jørgensen, Hesam Dashti, Adam Stefek, Elizabeth McGonagle, Sophie Strobel, Samantha Laber, Saaket Agrawal, Gregory P. Westcott, Amrita Kar, Molly L. Veregge, Anton Gulko, Harini Srinivasan, Zachary Kramer, Eleanna De Filippis, Erin Merkel, Jennifer Ducie, Christopher G. Boyd, William Gourash, Anita Courcoulas, Samuel J. Lin, Bernard T. Lee, Donald Morris, Adam Tobias, Amit V. Khera, Melina Claussnitzer, Tune H. Pers, Antonio Giordano, Orr Ashenberg, Aviv Regev, Linus T. Tsai, and Evan D. Rosen. A single-cell atlas of human and mouse white adipose tissue. *Nature*, 603(7903):926–933, March 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04518-2. URL http://dx.doi.org/10.1038/s41586-022-04518-2.

[17] Mariano I. Gabitto, Kyle J. Travaglini, Victoria M. Rachleff, Eitan S. Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T. Mahoney, Nick Dee, Jeff Goldy, Erica J. Melief, Anamika Agrawal, Omar Kana, Xingjian Zhen, Samuel T. Barlow, Krissy Brouner, Jazmin Campos, John Campos, Ambrose J. Carr, Tamara Casper, Rushil Chakrabarty, Michael Clark, Jonah Cool, Rachel Dalley, Martin Darvas, Song-Lin Ding, Tim Dolbeare, Tom Egdorf, Luke Esposito, Rebecca Ferrer, Lynn E. Fleckenstein, Rohan Gala, Amanda Gary, Emily Gelfand, Jessica Gloe, Nathan Guilford, Junitta Guzman, Daniel Hirschstein, Windy Ho, Madison Hupp, Tim Jarsky, Nelson Johansen, Brian E. Kalmbach, Lisa M. Keene, Sarah Khawand, Mitchell D. Kilgore, Amanda Kirkland, Michael Kunst, Brian R. Lee, Mckaila Leytze, Christine L. Mac Donald, Jocelin Malone, Zoe Maltzer, Naomi Martin, Rachel McCue, Delissa McMillen, Gonzalo Mena, Emma Meyerdierks, Kelly P. Meyers, Tyler Mollenkopf, Mark Montine, Amber L. Nolan, Julie K. Nyhus, Paul A. Olsen, Maiya Pacleb, Chelsea M. Pagan, Nicholas Peña, Trangthanh Pham, Christina Alice Pom, Nadia Postupna, Christine Rimorin, Augustin Ruiz, Giuseppe A. Saldi, Aimee M. Schantz, Nadiya V. Shapovalova, Staci A. Sorensen, Brian Staats, Matt Sullivan, Susan M. Sunkin, Carol Thompson, Michael Tieu, Jonathan T. Ting, Amy Torkelson, Tracy Tran, Nasmil J. Valera Cuevas, Sarah Walling-Bell, Ming-Qiang Wang, Jack Waters, Angela M. Wilson, Ming Xiao, David Haynor, Nicole M. Gatto, Suman Jayadev, Shoaib Mufti, Lydia Ng, Shubhabrata Mukherjee, Paul K. Crane, Caitlin S. Latimer, Boaz P.

Levi, Kimberly A. Smith, Jennie L. Close, Jeremy A. Miller, Rebecca D. Hodge, Eric B. Larson, Thomas J. Grabowski, Michael Hawrylycz, C. Dirk Keene, and Ed S. Lein. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, 27(12):2366–2383, October 2024. ISSN 1546-1726. doi: 10.1038/s41593-024-01774-5. URL http://dx.doi.org/10.1038/s41593-024-01774-5.

[18] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[19] Qiuyu Gong, Mehul Sharma, Marla C. Glass, Emma L. Kuan, Aishwarya Chander, Mansi Singh, Lucas T. Graybuck, Zachary J. Thomson, Christian M. LaFrance, Samir Rachid Zaim, Tao Peng, Lauren Y. Okada, Palak C. Genge, Katherine E. Henderson, Elisabeth M. Dornisch, Erik D. Layton, Peter J. Wittig, Alexander T. Heubeck, Nelson M. Mukuka, Julian Reading, Garrett Strawn, Teminijesu Titus-Adewunmi, Kathleen Abadie, Charles R. Roll, Veronica Hernandez, Vaishnavi Parthasarathy, Tyanna J. Stuckey, Blessing Musgrove, Elliott Swanson, Cara Lord, Morgan D. A. Weiss, Cole G. Phalen, Regina R. Mettey, Kevin J. Lee, John B. Johanneson, Erin K. Kawelo, Jessica Garber, Upaasana Krishnan, Megan Smithmyer, E. John Wherry, Laura A. Vella, Sarah E. Henrickson, Mackenzie S. Kopp, Adam K. Savage, Lynne A. Becker, Paul Meijer, Ernest M. Coffey, Jorg J. Goronzy, Mikael Sigvardsson, Cate Speake, Thomas F. Bumol, Ananda W. Goldrath, Troy R. Torgerson, Xiao-jun Li, Peter J. Skene, Jane H. Buckner, and Claire E. Gustafson. Multi-omic profiling reveals age-related immune dynamics in healthy adults. *Nature*, 648(8094):696–706, October 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09686-5. URL http://dx.doi.org/10.1038/s41586-025-09686-5.

[20] G. Kenneth Gray, Carman Man-Chung Li, Jennifer M. Rosenbluth, Laura M. Selfors, Nomeda Girnius, Jia-Ren Lin, Ron C.J. Schackmann, Walter L. Goh, Kaitlin Moore, Hana K. Shapiro, Shaolin Mei, Kurt D'Andrea, Katherine L. Nathanson, Peter K. Sorger, Sandro Santagata, Aviv Regev, Judy E. Garber, Deborah A. Dillon, and Joan S. Brugge. A human breast atlas integrating single-cell proteomics and transcriptomics. *Developmental Cell*, 57(11):1400–1420.e7, June 2022. ISSN 1534-5807. doi: 10.1016/j.devcel.2022.05.003. URL http://dx.doi.org/10.1016/j.devcel.2022.05.003.

[21] Xiangyu Guo and Ricardo Henao. Cross-modal imputation and uncertainty estimation for spatial transcriptomics. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4375–4383. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/guo25a.html.

[22] 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, June 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL http://dx.doi.org/10.1038/s41592-024-02305-7.

[23] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.048. URL http://dx.doi.org/10.1016/j.cell.2021.04.048.

[24] Graham Heimberg, Tony Kuo, Daryle J. DePianto, Omar Salem, Tobias Heigl, Nathaniel Diamant, Gabriele Scalia, Tommaso Biancalani, Shannon J. Turley, Jason R. Rock, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638(8052):1085–1094, November 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08411-y. URL http://dx.doi.org/10.1038/s41586-024-08411-y.

[25] 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), March 2022. ISSN 2001-1326. doi: 10.1002/ctm2.694. URL http://dx.doi.org/10.1002/ctm2.694.

[26] Kazumasa Kanemaru, James Cranley, Daniele Muraro, Antonio M. A. Miranda, Siew Yen Ho, Anna Wilbrey-Clark, Jan Patrick Pett, Krzysztof Polanski, Laura Richardson, Monika Litvinukova, Natsuhiko Kumasaka, Yue Qin, Zuzanna Jablonska, Claudia I. Semprich, Lukas Mach, Monika Dabrowska, Nathan Richoz, Liam Bolt, Lira Mamanova, Rakeshlal Kapuge, Sam N. Barnett, Shani Perera, Carlos Talavera-López, Ilaria Mulas, Krishnaa T. Mahbubani, Liz Tuck, Lu Wang, Margaret M. Huang, Martin Prete, Sophie Pritchard, John Dark, Kourosh Saeb-Parsy, Minal Patel, Menna R. Clatworthy, Norbert Hübner, Rasheda A. Chowdhury, Michela Noseda, and Sarah A. Teichmann. Spatially resolved multiomics of human cardiac niches. *Nature*, 619(7971):801–810, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06311-1. URL http://dx.doi.org/10.1038/s41586-023-06311-1.

[27] Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi, Jung-Il Lee, Yeon-Lim Suh, Bo Mi Ku, Hye Hyeon Eum, Soyean Choi, Yoon-La Choi, Je-Gun Joung, Woong-Yang Park, Hyun Ae Jung, Jong-Mu Sun, Se-Hoon Lee, Jin Seok Ahn, Keunchil Park, Myung-Ju Ahn, and Hae-Ock Lee. Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications*, 11 (1), May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16164-1. URL http://dx.doi.org/10.1038/s41467-020-16164-1.

[28] Blue B. Lake, Ricardo Melo Ferreira, Jens Hansen, Rajasree Menon, Jeannine Basta, Heather Thiessen Philbrook, Stephanie Reinert, Robin Fallegger, Asmita K. Lagwankar, Xi Chen, Soumya Maity, Katerina V. Djambazova, Brittney L. Gorman, Nicholas Lucarelli, Debora L. Gisch, Insa M. Schmidt, Viji Nair, Fadhl Alakwaa, Eirini Kefaloyianni, Bo Zhang, Amanda L. Knoten, Madhurima Kaushal, Edgar A. Otto, Melissa A. Farrow, Dinh Diep, Dusan Velickovic, Angela R. Sabo, Elijah Cole, Ian Tamayo, Jovan Tanevski, Kimberly Y. Conklin, Rachel S. G. Sealfon, Yongqun He, Michelle Brennan, Lynn Robbins, Ying-Hua Cheng, Markus Bitzer, Aditya Surapaneni, Steven Menez, Peter V. Kharchenko, Charles E. Alpers, Ulysses G. J. Balis, Laura Barisoni, Ian H. de Boer, Dawit Demeke, Agnes B. Fogo, Joel M. Henderson, Leal Herlitz, Gilbert W. Moeckel, Parmjeet S. Randhawa, Avi Z. Rosenberg, Jennifer A. Schaub, Suman Setty, Frank C. Brosius, Maria L. Caramori, Steven G. Coca, Robert S. Figenshau, Eric H. Kim, Krzysztof Kiryluk, James P. Lash, R. Tyler Miller, John F. O'Toole, Paul M. Palevsky, Eugene P. Rhee, Ana C. Ricardo, Sylvia E. Rosas, Prabir Roy-Chaudhury, Minnie M. Sarwal, John R. Sedor, Robert D. Toto, Aydın Türkmen, Sushrut S. Waikar, James C. Williams, F P. Wilson, E. Steve Woodle, Evan Z. Macosko, Julio Saez-Rodriguez, Pierre C Dagher, Morgan E. Grams, Petter Bjornstad, Tarek M. El-Achkar, Olga G. Troyanskaya, Nikole Bonevich, Pinaki Sarder, Sanjeev Kumar, Christopher R. Anderton, Jeffrey M. Spraggins, Kumar Sharma, Michael Rauchman, Jonathan Himmelfarb, Joseph P. Gaut, Kidney Precision Medicine Project, Kun Zhang, Ravi Iyengar, Matthias Kretzler, Jeffrey B. Hodgin, Chirag R. Parikh, Michael T. Eadon, Sanjay Jain, HuBMAP consortium, and Kidney Precision Medicine Project. Cellular and spatial drivers of unresolved injury and functional decline in the human kidney. *bioRxiv*, 2025. doi: 10.1101/2025.09.26.678707. URL https://www.biorxiv.org/content/early/2025/11/24/2025.09.26.678707.

[29] Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique de Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David van Dijk. Cell2sentence: Teaching large language models the language of biology. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=EWt5wsEdvc.

[30] Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, Ke Liu, Xuetong Zhao, Jun Lin, Linzhao Cheng, Falai Chen, Tian Xue, and Kun Qu. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods*, 19(6):662–670, May 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01480-9. URL http://dx.doi.org/10.1038/s41592-022-01480-9.

[31] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12): 1739–1740, May 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr260. URL http://dx.doi.org/10.1093/bioinformatics/btr260.

[32] Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, December 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. URL http://dx.doi.org/10.1038/s41592-021-01336-8.

[33] Lieke Michielsen, Mohammad Lotfollahi, Daniel Strobl, Lisa Sikkema, Marcel J T Reinders, Fabian J Theis, and Ahmed Mahfouz. Single-cell reference mapping to construct and extend cell-type hierarchies. *NAR Genomics and Bioinformatics*, 5(3), 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqad070. URL http://dx.doi.org/10.1093/nargab/lqad070.

[34] David Osumi-Sutherland, Chuan Xu, Maria Keays, Adam P. Levine, Peter V. Kharchenko, Aviv Regev, Ed Lein, and Sarah A. Teichmann. Cell type ontologies of the human cell atlas. *Nature Cell Biology*, 23(11):1129–1135, November 2021. ISSN 1476-4679. doi: 10.1038/s41556-021-00787-7. URL http://dx.doi.org/10.1038/s41556-021-00787-7.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Qiqi Qiao, Shunfeng Hu, and Xin Wang. The regulatory roles and clinical significance of glycolysis in tumor. *Cancer Communications*, 44(7):761–786, June 2024. ISSN 2523-3548. doi: 10.1002/cac2.12549. URL http://dx.doi.org/10.1002/cac2.12549.

[37] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Stephen R. Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656(8126):183–191, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10689-z. URL http://dx.doi.org/10.1038/s41586-026-10689-z.

[38] Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, Jörg Menche, Eleni M. Tomazou, and Christoph Bock. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, November 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02857-9. URL http://dx.doi.org/10.1038/s41587-025-02857-9.

[39] Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies, 2024. URL https://arxiv.org/abs/2407.16067.

[40] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C. Strobl, Tessa E. Gillett, Luke Zappia, Elo Madissoon, Nikolay S. Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, Marie-Jeanne Arguel, Leonie Apperloo, Martin Banchero, Christophe Bécavin, Marijn Berg, Evgeny Chichelnitskiy, Mei-i Chung, Antoine Collin, Aurore C. A. Gay, Janine Gote-Schniering, Baharak Hooshiar Kashani, Kemal Inecik, Manu Jain, Theodore S. Kapellos, Tessa M. Kole, Sylvie Leroy, Christoph H. Mayr, Amanda J. Oliver, Michael von Papen, Lance Peter, Chase J. Taylor, Thomas Walzthoeni, Chuan Xu, Linh T. Bui, Carlo De Donno, Leander Dony, Alen Faiz, Minzhe Guo, Austin J. Gutierrez, Lukas Heumos, Ni Huang, Ignacio L. Ibarra, Nathan D. Jackson, Preetish Kadur Lakshminarasimha Murthy, Mohammad Lotfollahi, Tracy Tabib, Carlos Talavera-López, Kyle J. Travaglini, Anna Wilbrey-Clark, Kaylee B. Worlock, Masahiro Yoshida, Yuexin Chen, James S. Hagood, Ahmed Agami, Peter Horvath, Joakim Lundeberg, Charles-Hugo Marquette, Gloria Pryhuber, Chistos Samakovlis, Xin Sun, Lorraine B. Ware, Kun Zhang, Maarten van den Berge, Yohan Bossé, Tushar J. Desai, Oliver Eickelberg, Naftali Kaminski, Mark A. Krasnow, Robert Lafyatis, Marko Z. Nikolic, Joseph E. Powell, Jayaraj Rajagopal, Mauricio Rojas, Orit Rozenblatt-Rosen, Max A. Seibold, Dean Sheppard, Douglas P. Shepherd, Don D. Sin, Wim Timens, Alexander M. Tsankov, Jeffrey Whitsett, Yan Xu, Nicholas E. Banovich, Pascal Barbry, Thu Elizabeth Duong, Christine S. Falk, Kerstin B. Meyer, Jonathan A. Kropski, Dana Pe'er, Herbert B. Schiller, Purushothama Rao Tata, Joachim L. Schultze, Sara A. Teichmann, Alexander V. Misharin, Martijn C. Nawijn, Malte D. Luecken, and Fabian J. Theis. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, June 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02327-2. URL http://dx.doi.org/10.1038/s41591-023-02327-2.

[41] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007. ISSN 1546-1696. doi: 10.1038/nbt1346. URL http://dx.doi.org/10.1038/nbt1346.

[42] Michael P. Snyder, Shin Lin, Amanda Posgai, Mark Atkinson, Aviv Regev, Jennifer Rood, Orit Rozenblatt-Rosen, Leslie Gaffney, Anna Hupalowska, Rahul Satija, Nils Gehlenborg, Jay Shendure, Julia Laskin, Pehr Harbury, Nicholas A. Nystrom, Jonathan C. Silverstein, Ziv Bar-Joseph, Kun Zhang, Katy Börner, Yiing Lin, Richard Conroy, Dena Procaccini, Ananda L. Roy, Ajay Pillai, Marishka Brown, Zorina S. Galis, Long Cai, Jay Shendure, Cole Trapnell, Shin Lin, Dana Jackson, Michael P. Snyder, Garry Nolan, William James Greenleaf, Yiing Lin, Sylvia Plevritis, Sara Ahadi, Stephanie A. Nevins, Hayan Lee, Christian Martijn Schuerch, Sarah Black, Vishal Gautham Venkataraaman, Ed Esplin, Aaron Horning, Amir Bahmani, Kun Zhang, Xin Sun, Sanjay Jain, James Hagood, Gloria Pryhuber, Peter Kharchenko, Mark Atkinson, Bernd Bodenmiller, Todd Brusko, Michael Clare-Salzler, Harry Nick, Kevin Otto, Amanda Posgai, Clive Wasserfall, Marda Jorgensen, Maigan Brusko, Sergio Maffioletti, Richard M. Caprioli, Jeffrey M. Spraggins, Danielle Gutierrez, Nathan Heath Patterson, Elizabeth K. Neumann, Raymond Harris, Mark deCaestecker, Agnes B. Fogo, Raf van de Plas, Ken Lau, Long Cai, Guo-Cheng Yuan, Qian Zhu, Ruben Dries, Peng Yin, Sinem K. Saka, Jocelyn Y. Kishi, Yu Wang, Isabel Goldaracena, Julia Laskin, DongHye Ye, Kristin E. Burnum-Johnson, Paul D. Piehowski, Charles Ansong, Ying Zhu, Pehr Harbury, Tushar Desai, Jay Mulye, Peter Chou, Monica Nagendran, Ziv Bar-Joseph, Sarah A. Teichmann, Benedict Paten, Robert F. Murphy, Jian Ma, Vladimir Yu. Kiselev, Carl Kingsford, Allyson Ricarte, Maria Keays, Sushma A. Akoju, Matthew Ruffalo, Nils Gehlenborg, Peter Kharchenko, Margaret Vella, Chuck McCallum, Katy Börner, Leonard E. Cross, Samuel H. Friedman, Randy Heiland, Bruce Herr, Paul Macklin, Ellen M. Quardokus, Lisel Record, James P. Sluka, Griffin M. Weber, Nicholas A. Nystrom, Jonathan C. Silverstein, Philip D. Blood, Alexander J. Ropelewski, William E. Shirey, Robin M. Scibek, Paula Mabee, W. Christopher Lenhardt, Kimberly Robasky, Stavros Michailidis, Rahul Satija, John Marioni, Aviv Regev, Andrew Butler, Tim Stuart, Eyal Fisher, Shila Ghazanfar, Jennifer Rood, Leslie Gaffney, Gokcen Eraslan, Tommaso Biancalani, Eeshit D. Vaishnav, Richard Conroy, Dena Procaccini, Ananda Roy, Ajay Pillai, Marishka Brown, Zorina Galis, Pothur Srinivas, Aaron Pawlyk, Salvatore Sechi, Elizabeth Wilder, and James Anderson. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574 (7777):187–192, October 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1629-x. URL http://dx.doi.org/10.1038/s41586-019-1629-x.

[43] Yifan Sun, Yuke Zhu, Yuhan Zhang, Pengkun Zheng, Xi Qiu, Chi Zhang, and Yichen Wei. Dynamic metric learning: Towards a scalable metric space to accommodate multiple semantic scales. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5389–5398, 2021. doi: 10.1109/CVPR46437.2021.00535.

[44] Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02353-z. URL http://dx.doi.org/10.1038/s41592-024-02353-z.

[45] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618 (7965):616–624, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

[46] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, November 2016. ISSN 1548-7105. doi: 10.1038/nmeth.4077. URL http://dx.doi.org/10.1038/nmeth.4077.

[47] Hanchen Wang, Jure Leskovec, and Aviv Regev. Limitations of cell embedding metrics assessed using drifting islands. *Nature Biotechnology*, June 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02702-z. URL http://dx.doi.org/10.1038/s41587-025-02702-z.

[48] Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B. Altman. Leveraging the cell ontology to classify unseen cell types. *Nature Communications*, 12(1), September 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25725-x. URL http://dx.doi.org/10.1038/s41467-021-25725-x.

[49] Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *The Twelfth International Conference on Learning Representations*, 2024.

[50] Jialu Wu, Qing Ye, Yilin Wang, Renling Hu, Yiheng Zhu, Mingze Yin, Tianyue Wang, Jike Wang, Chang-Yu Hsieh, and Tingjun Hou. Biology-driven insights into the power of single-cell foundation models. *Genome Biol.*, 26(1):334, October 2025.

[51] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9): 1334–1347, September 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00911-1. URL http://dx.doi.org/10.1038/s41588-021-00911-1.

[52] 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, September 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z. URL http://dx.doi.org/10.1038/s42256-022-00534-z.

[53] Siqian Yang, Shiqi Chen, Yue Zhao, Tao Wu, Yuquan Wang, Tingting Li, Liwan Fu, Ting Ye, Yue-Qing Hu, and Haiquan Chen. Identification of a coagulation-related signature correlated with immune infiltration and their prognostic implications in lung adenocarcinoma. *Thoracic Cancer*, 14(33):3295–3308, October 2023. ISSN 1759-7714. doi: 10.1111/1759-7714.15121. URL http://dx.doi.org/10.1111/1759-7714.15121.

[54] Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. Cell ontology guided transcriptome foundation model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[55] Andy G.X. Zeng, Ilaria Iacobucci, Sayyam Shah, Amanda Mitchell, Gordon Wong, Suraj Bansal, David Chen, Qingsong Gao, Hyerin Kim, James A. Kennedy, Andrea Arruda, Mark D. Minden, Torsten Haferlach, Charles G. Mullighan, and John E. Dick. Single-cell transcriptional atlas of human hematopoiesis reveals genetic and hierarchy-based determinants of aberrant aml differentiation. *Blood Cancer Discovery*, 6(4):307–324, April 2025. ISSN 2643-3249. doi: 10.1158/2643-3230.bcd-24-0342. URL http://dx.doi.org/10.1158/2643-3230.BCD-24-0342.

[56] Suyuan Zhao, Jiahuan Zhang, and Zaiqing Nie. Large-scale cell representation learning via divide-and-conquer contrastive learning, 2023. URL https://arxiv.org/abs/2306.04371.

[57] Qin Zhu, Zuzhi Jiang, Binyamin Zuckerman, Leor Weinberger, Matt Thomson, and Zev J. Gartner. Revealing a coherent cell-state landscape across single-cell datasets with concord. *Nature Biotechnology*, January 2026. ISSN 1546-1696. doi: 10.1038/s41587-025-02950-z. URL http://dx.doi.org/10.1038/s41587-025-02950-z.

A Related Work

Existing single-cell foundation models (scFMs) can be broadly categorized into three groups based on their tokenization and data processing strategies [44]. The first group [45, 54, 7], exemplified by Geneformer, employs rank-based tokenization, where genes are ordered by expression level to form sequence-like inputs for Masked Language Modeling (MLM). The second group [52, 10], (*e.g.*, scGPT) extends this by discretizing gene expression values into consecutive bins, hence enabling joint learning of gene and cell representations. The third group [49, 22], known as value projection methods (*e.g.*, CellPLM, scFoundation), bypasses discrete tokenization by projecting raw expression counts directly into a continuous embedding space. Other recent studies [9, 29, 38], have explored multi-modal learning to align single-cell data with natural language, thus leveraging the power of large language models; these methods demonstrate great potential for downstream biological discovery.

Hierarchy-aware representation learning has a long history outside single-cell biology. Hierarchical triplet loss [18] and dynamic metric learning [43] adapt margins to a class taxonomy; Bertinetto et al. [5] show that LCA-height-weighted losses produce "better mistakes" on ImageNet; and Shi et al. [39] find that LCA-based taxonomic distance predicts out-of-distribution generalization across vision models. Within single-cell biology, OnClass [48] propagates labels over the Cell Ontology to classify unseen types, treeArches [33] progressively builds and extends cell atlas according to the hierarchies of cell types, UCE [37] evaluates embeddings against ontology-derived distances, and hierarchical cross-entropy [11] trains annotation models against the ontology tree. scOntoFM draws on both lines: it uses LCA distance both to mine triplets offline and as the target of a rank-order alignment loss during foundation-model pre-training.

In terms of incorporating cell ontology into model pretraining, Heimberg et al. [24] proposed SCimilarity, a deep metric learning–based foundation model that combines a standard triplet loss with a reconstruction objective to embed single-cell RNA-sequencing data into a unified space. Specifically, it leverages cell ontology information to guide the sampling of semi-hard negative triplets within mini-batches. Along a similar line, scCello [54] employs the Personalized PageRank (PPR) algorithm to quantify structural similarity between nodes in the cell ontology graph, thereby guiding the model to learn pairwise cell-type distance relationships that aligns with PPR structure.

Unlike previous approaches, we proposes a more coherent and unified framework that integrates ontology information not only into the triplet sampling stage but also into biologically constrained representation learning throughout pre-training. Concretely, compared to scCello and SCimilarity, scOntoFM enjoys specific novelty as follows: i) Efficient offline ontology-aware triplet sampling, which mines more informative positive and negative pairs and facilitate effective contrastive learning. While standard online triplet sampling, like in scCello and SCimilarity, is highly inefficient for massive datasets and the probability of encountering informative positive pairs is low in a random mini-batch (see AFigure 7). ii) The ontology-aware dynamic margin and global ontology alignment loss terms enables the model capture more fine-grained cell type information (see Tables 2), while also maintains a faithful biological structure versus scCello (see Figure 3 right panel, AFigure 11, 12). iii) Comprehensive evaluation on both cell- and gene level task. While scCello mainly concentrates on cell level tasks, scOntoFM provides a more dedicated evaluation for scFMs in gene-level tasks (see Section 4).

B EXPERIMENTS

B.1 Online-triplet sampling

In AFigure 7, we show an experiment of online triplet sampling for 10% of our pre-training data. Around 46.78% of them have zero positives, and 21.03% only have one positives, indicating the highly in-efficiency for online triplet sampling in large data corpus.

B.2 Relationship between LCA distance and transcriptomic similarity

To test the premise that proximity in the cell ontology reflects transcriptomic proximity, we computed per cell-type transcriptomic centroids in PCA space by averaging the pre-processed gene expression through Scanpy pipeline (library size normalization, log-transform, with highly variable genes set to

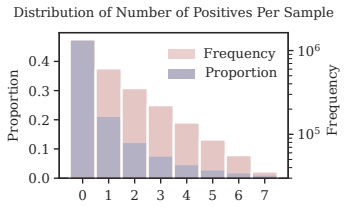

Figure 7: Distribution of positives per sample under online random sampling in 10% of our pretraining dataset. Pink bars show the absolute frequency of positive samples, and purple bars show the corresponding proportions.

2000, and principle components set to 50) over all cells belonging to that type, and measured the cosine distance between every pair of cell-type profiles. We compared these transcriptomic distances to the raw LCA distances $\Delta_{i,j}$ from 1 in Tabula Sapiens v2 dataset (173 cell types across 28 organs) and the nine diverse tissue datasets from our zero-shot clustering evaluation.

Correlation Analysis on Tabula Sapiens v2. As shown in Table 6 and Figure 2, grouping cell pairs by the first 10 LCA distance reveals a strong positive correlation and shows a strictly monotone increase in mean expression distance (Pearson $r = 0.91$, $p < 0.001$). The within-level standard deviation shrinks with LCA distance: close relatives are heterogeneous (some sibling types are transcriptionally far apart), whereas distant types are uniformly dissimilar. This broad, multi-organ analysis supports the biological premise that ontology-near cell types tend, on average, to be more transcriptionally similar, which also aligns with findings in OnClass [48] and UCE [37].

Table 6: The relationship between ontology LCA distance and transcriptomic distance in Tabula Sapiens v2.

LCA distance	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mean Expr. dist. ± sd	0.55±0.37	0.72±0.33	0.82±0.30	0.91±0.23	0.96±0.20	0.99±0.17	1.00±0.16	1.02±0.15	1.04±0.13	1.06±0.12

Generalization Across Nine Zero-shot Clustering Datasets. As showed in Table 1 and AFigure 8, we observed a consistent positive correlation in every tissue, with an overall Pearson $r = 0.41$ ($p < 0.001$) across 130 unique cell types. Correlations are strongest in immune-rich tissues (HIHA, Lung, Breast TNBC), where the ontology is deepest, and weakest in Kidney, Adipose and Liver, where many annotated types are shallow epithelial or stromal nodes.

Discordant pairs. Plotting pre-trained global cell-type centroid distances against $\tilde{\Delta}_{i,j}$ (Figure 10) isolates two kinds of off-diagonal pairs. (i) *Ontology-far, transcriptome-close*: almost entirely functional T-cell subsets that the ontology separates but the transcriptome does not (*e.g.*, $CD8^+$ cytotoxic *vs.* $CD4^+$ cytotoxic T cells; $\gamma\delta$ intra-epithelial *vs.* effector-memory $CD8^+$ T cells), plus cross-tissue mesenchyme (ileal smooth muscle *vs.* kidney fibroblast). (ii) *Ontology-close, transcriptome-far*: dominated by generic parent terms ("professional antigen presenting cell", "connective tissue cell") that sit one edge from many children yet are transcriptionally far from all of them. The ontology is thus an informative prior rather than a complete model of expression; the MLM objective keeps the embedding faithful to the transcriptome in exactly these regions, which is why the full model (not the ontology-only ablations) gives the best balance in Figure 6.

B.3 Pre-trained Model Configuration (Table 7)

The pre-trained model backbone, architecture, and the training details are listed in Table 7.

B.4 Computation cost of the Cell-Dataset Index Map (Ω)

We construct the Cell-Dataset Index Map by extracting specific columns (cell_type_ontology_term_id, dataset_id) from the Hugging Face pre-training dataset and applying pandas' *groupby* operation. The process takes about 5 minutes and uses 6.5 GB of peak memory on an Intel Xeon Gold 6252 CPU. The final file size is 132 MB, ensuring that scalability to a larger corpus is highly feasible.

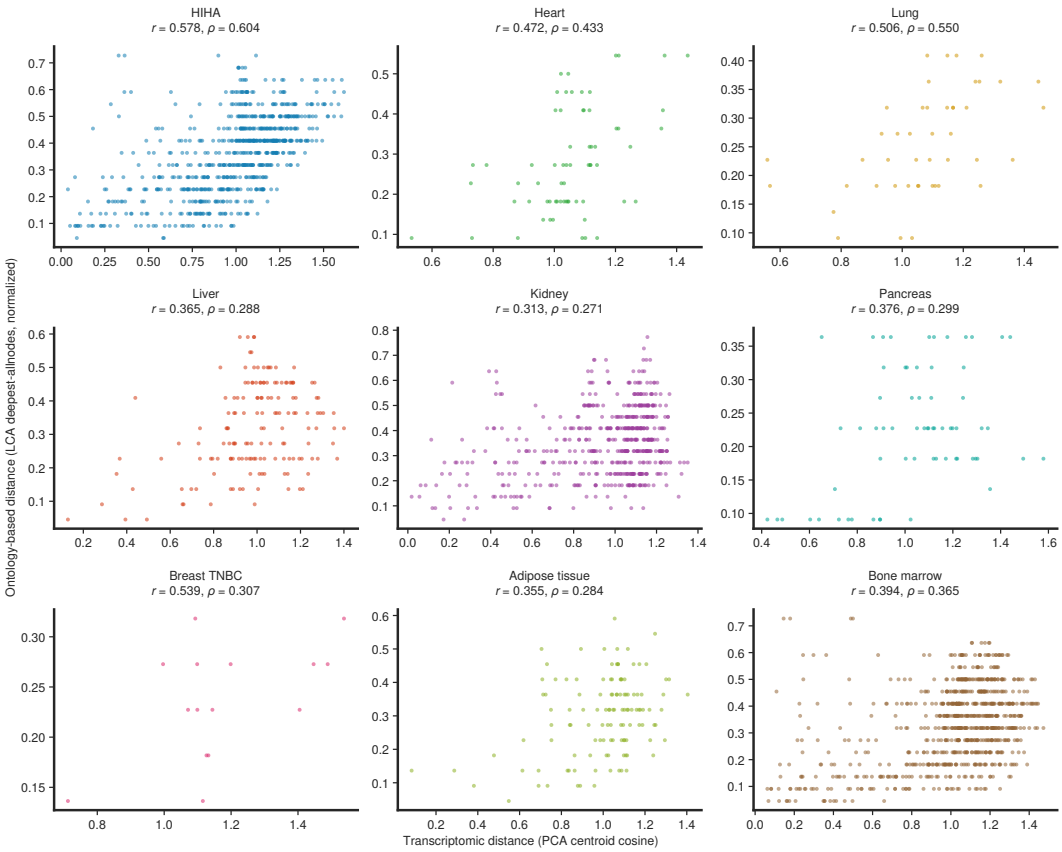

Figure 8: Scatter plots of LCA distance versus cell-type pairwise transcriptomic distance for each of the nine tissues.

B.5 Baselines

Geneformer. We followed the official guideline of Geneformer [45] and downloaded the pre-training weights of Geneformer-V1-10M on https://huggingface.co/ctheodoris/Geneformer/tree/main.

scGPT. We follow the tutorial of running scGPT [10] from https://scgpt.readthedocs.io/en/latest/ and downloaded the whole-human model checkpoint (https://drive.google.com/drive/folders/1oWh_-ZRdhtoGQ2Fw24HP41FgLoomVo-y?usp=sharing) as indicated in its github repository.

scCello. The pre-trained zero-shot scCello [54] model weights are downloaded from hugging face with following link: https://huggingface.co/katarinayuan/scCello-zeroshot.

CellPLM. We refer the official github repository of CellPLM [49] (https://github.com/OmicsML/CellPLM) and download its latest version checkpoint: 20230926_85M from https://github.com/OmicsML/CellPLM/tree/main/ckpt.

SCimilarity. We refer the official github repository of SCimilarity [24] (https://github.com/Genentech/scimilarity) and download its latest model v_1.1 from https://zenodo.org/records/10685499.

To provide a clear overview of each baseline, in the following Table 9, we summarize and compare models in terms of pre-training data, parameter size, and training strategy.

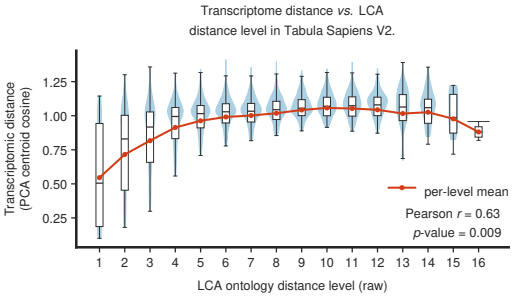

Figure 9: Cell-type pairwise transcriptomic distance in PCA space across raw LCA ontology distance over 1 to 16 levels in Tabula Sapiens v2. Boxes show the distance distributions, and the red line denotes the per-level mean.

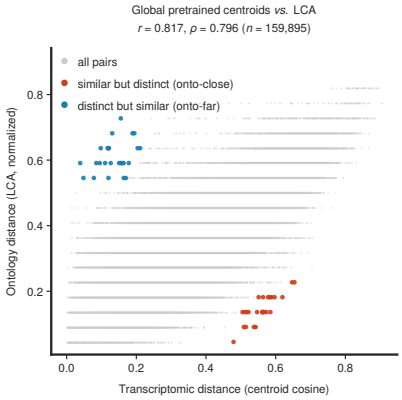

Figure 10: Scatter plot of the pairwise transcriptomic cell-type centroid distance versus the normalized LCA distance $\tilde{\Delta}_{i,j}$. The off-diagonal points are highlighted with red and blue.

B.6 Zero-shot Clustering Datasets Details

For the zero-shot evaluation datasets summarized in Table 10, we retained only primary cells from the original datasets, and no additional filtering was applied. Each data sources are listed down below in the table.

Unseen Evaluation Datasets In Table 11, we show the list of query dataset ids from CZI data portal.

B.7 Cell Type Annotation Fine-tuning

We peformed cell type classification on the HIHA and Pancreas datasets, with fixed held-out testing donors (UP1008 and UP1012 in HIHA; SAMN16547514, SAMN17209843, SAMN17528599, SAMN17831932 in Pancreas) and identical hyperparameters across all scFMs. To ensure a fair comparison, we utilized a standardized Multi-Layer Perceptron (MLP) head classifier for all foundational models. The network takes the extracted 2D continuous cell embeddings as input and processes them through two hidden layers (1024 and 512 dimensions), each followed by 1D Batch Normalization, a ReLU activation, and Dropout regularization (rate of 0.2) to mitigate overfitting. The model was optimized to minimize Cross-Entropy Loss using the AdamW optimizer with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}, while reserving 10% of the training donor cells as a validation set for early stopping. We also compared with a specialized cell annotation model with hierarchical cross-entropy (HCE) loss [11], using the authors' released MLP and TabNet checkpoints and their penultimate-layer embeddings, the same donor splits, and the same MLP probe. The results (Table 3, ATable 12) are reported as mean±std over 5 random seeds and confirm that scOntoFM also maintain its significant performance advantage during fine-tuning stage.

B.8 Novel Cell Identification

This approach assumes that the biological relationships defined in a Cell Ontology (hierarchical closeness) translate directly to geometric closeness in the cell representation space. Cell types that are closely related in the Cell Ontology (*e.g.*, "CD4+ T cell" and "CD8+ T cell") share significant transcriptomic programs. Therefore, the embedding (centroid) of an unseen cell type can be reliably approximated as a weighted interpolation of its known evolutionary or functional relatives. The novel cell identification process could be summarized as follows:

1. Compute Ontology Affinity Matrix We define an affinity matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ between unseen and seen cell types using a Gaussian kernel on the LCA distances:

$$A_{ij} = \exp\left(-\gamma \cdot \mathbf{\Delta}(u_i, c_j)\right), \tag{6}$$

Table 7: Model architecture and training configuration.

Component	Configuration
Backbone Model	BERT (`BertForMaskedLM`)
BERT Architecture	Vocal size: 19941 Hidden size: 256 Layers: 6 Attention heads: 4 Intermediate size: 512 Activation: ReLU Max position embeddings: 2002
Regularization	Hidden dropout: 0.02 Attention dropout: 0.02 LayerNorm ϵ: 1×10^{-12}
Optimizer	AdamW (lr $= 1 \times 10^{-3}$, weight decay = 0.001)
Learning Rate Schedule	WarmupDecayLR Warmup steps: 1% of total training steps
Distributed Training	DeepSpeed with ZeRO stage 2 Overlap communication enabled Reduce / allgather bucket size: 5×10^6 Contiguous gradients enabled
Precision	FP16 mixed precision
Training Hyperparameters	Batch size per GPU: 64 Gradient accumulation steps: 4 Effective batch size: $64 \times$ world size $\times 4$ Epochs: 3 Gradient clipping: 1.0

Table 8: Computational cost of constructing the Cell-Dataset Index Map.

Metric	Value
CPU	Intel Xeon Gold 6252
Time Consumption	$\sim$ 5 minutes
Peak Memory Usage	6.5 GB
Final File Size	132 MB

where γ is a scaling hyperparameter, we set it to 1 by default, and u_i is the unknown cell type, and c_j is the known cell type. The Gaussian kernel enables us to transform the raw LCA (Lowest Common Ancestor) distances into normalized similarity weights.

2. Normalize Weights We normalize the affinities so that the weights for each unseen cell type sum to 1:

$$W_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}}. \tag{7}$$

Let $\mathbf{W} \in \mathbb{R}^{m \times n}$ be the resulting weight matrix.

3. Generate Novel Prototypes We calculate the predicted centroid $\hat{\mathbf{p}}_{u_i}$ for each novel cell type by linearly combining the known centroids $\mathbf{Q} \in \mathbb{R}^{n \times d}$:

$$\hat{\mathbf{P}}_{\mathcal{U}} = \mathbf{W} \cdot \mathbf{Q}, \tag{8}$$

where $\hat{\mathbf{P}}_{\mathcal{U}} \in \mathbb{R}^{m \times d}$ contains the estimated embeddings for the unseen type centroids, $\mathcal{U} = \{u_i\}_{i=1}^{m}$ is the novel cell-type set containing all the unknown cell types

Table 9: Model comparison in terms of pretraining data size, model scale, and modeling strategy.

Model	Pretraining Data Size	# Parameters	Modeling Strategy
geneformer	30M	10,316,196	MLM, no ontology
scGPT	33M	51,330,049	Auto-regressive, no ontology
cellPLM	11.4M	82,402,543	VAE
scCello	22M	10,683,654	MLM + ontology
SCimilarity	7.9M	62,332,627	Autoencoder + ontology
scOntoFM	27.5M	8,846,848	MLM + ontology

Table 10: Summary of zero-shot clustering evaluation datasets used in this study.

Dataset	# Cells	Source
Blood (HIHA)	263,917	https://cellxgene.cziscience.com/e/e522d2cd-7927-4e59-a4ed-064009569279.cxg/
Heart	211,061	https://cellxgene.cziscience.com/e/d567b692-c374-4628-a508-8008f6778f22.cxg/
Liver	72,710	https://cellxgene.cziscience.com/e/84cfa5aa-a0d1-415e-ac47-eb66c9cf8170.cxg/
Kidney (Atlas V2)	348,984	https://cellxgene.cziscience.com/collections/9c9d04c4-8899-417f-bb6f-6107dcadf14f
Pancreas	245,878	https://cellxgene.cziscience.com/e/e51bae9a-c747-4b64-904a-4da7cda218ab.cxg/
Breast TNBC	52,681	https://cellxgene.cziscience.com/e/9a64bf99-ebe5-4276-93a8-bee9dff1cd47.cxg/
Lung	30,472	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907
Adipose Tissue	166,149	https://cellxgene.cziscience.com/e/ca46ffe6-0a26-4d8d-82fa-81722daf1102.cxg/
Bone Marrow	263,159	https://cellxgene.cziscience.com/collections/f6c50495-3361-40ed-a819-fb9644396ed9

4. Classification Given a novel cell and and its corresponding embedding $\mathbf{z}$, which could be classified by finding the nearest unseen centroid using cosine similarity:

$$\hat{y} = \underset{u_i \in \mathcal{U}}{\operatorname{argmax}} \left(\frac{\mathbf{z} \cdot \hat{\mathbf{p}}_{u_i}}{\|\mathbf{z}\| \|\hat{\mathbf{p}}_{u_i}\|} \right). \tag{9}$$

Novel Cell-type List A summary of the 99 novel cell types is provided in ATable 13, along with the number of cells for each cell type. For cell types with more than 10,000 cells, we randomly subsample 10,000 cells. In total, the novel cell evaluation dataset contains 774,215 cells.

B.9 Gene Function Prediction

B.9.1 Static Gene Embedding

To evaluate the semantic quality of the *static* gene representations, we perform a supervised multi-lable classification task to predict Gene Ontology (GO) namespace membership by using the *static* gene embedding extracted from scFMs.

Problem Formulation Let $\mathcal{M} = \{m_1, m_2, \ldots, m_{|\mathcal{M}|}\}$ be the set of genes. For each gene m_i, we extract a *static* embedding vector $\mathbf{e}_i \in \mathbb{R}^d$ directly from the input token embedding table of the pre-trained foundation model (*e.g.*, Geneformer, scCello, or scGPT). This approach evaluates the model's *zero-shot* understanding of gene identity without updating the encoder weights. The target labels are defined by the three root namespaces of the Gene Ontology: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). We define the label vector $\mathbf{y}_i \in \{0, 1\}^3$ for gene m_i based on memberships defined in the MSigDB[2] C5 gene sets, where $y_{i,j} = 1$ if the gene belongs to namespace j.

Gene Ontology Gene Sets The GO terms are downloaded from MsiDB C5 collections, encompassing 7583 BP components, 1042 CC components, and 1855 MF components, forming 10480 gene sets in total.

Model Architecture We employ a 3-layer Multilayer Perceptron (MLP) as a probing classifier. The model architecture and training details are summarized in Table 14. To ensure a fair comparison, we apply same settings for all the scFMs without specifically tuning the hyper-parameters.

Experiments and Evaluation To ensure robust evaluation, we perform cross-validation (set k folder to 20) with stratified sampling for the train/test splits. Performance is reported as the mean and standard deviation across the 20 splits for Micro-F1, Macro-F1, Micro-Average Precision (AP), and Macro-AP, the evaluation metrics will be introduced in more detail in Appendix B.11.4.

[2]MSigDB website: https://www.gsea-msigdb.org/gsea/msigdb

B.9.2 Contextual Gene embedding

To interpret the biological signals captured by the foundation model, we perform pathway activity inference using the *decoupler* package [3] for the lung adenocarcinoma dataset [27]. Unlike standard approaches that utilize raw read counts, we leverage the contextualized gene embedding generated by the model.

Input Data Preparation Let $\mathbf{H}^{(i)} \in \mathbb{R}^{L\times d}$ be the hidden state representation for cell i extracted from the last layer of the pretrained Transformer model, where L is the sequence length (number of genes) and d is the embedding dimension. We obtain a scalar intensity value E_{ij} for the j-th gene in cell i by averaging over the embedding dimension:

$$E_{ij} = \frac{1}{d}\sum_{k=1}^{d} H_{ijk}. \tag{10}$$

This operation results in a dense matrix $\mathbf{E} \in \mathbb{R}^{N\times M}$ representing the mean activation intensity of each gene within its specific cellular context.

Pathway Activity Inference We employ the **Univariate Linear Model (ULM)** in the *decoupler* package [3] for activity estimation, and download the prior knowledge network from the MSigDB Hallmark collection[2]. For each cell i, ULM fits a linear model to determine if the continuous values in $\mathbf{E}_i$ corresponding to a specific pathway gene set are significantly higher than the background distribution. The resulting matrix $\mathbf{Y} \in \mathbb{R}^{N\times K}$ contains z-scores (enrichment scores) representing the inferred activity of K biological pathways across N cells. Similarly, we could easily extend this framework to other available gene sets (*i.e.*, MSigDB C3 regulatory taget gene sets, C5 ontology gene sets, or resources inside Omnipath [46]) and perform biological discovery with our model's contextual embedding. More information about the *decoupler* package and the analysis pipeline could be found in the official website[3].

Differential Pathway Analysis To identify cell-type-specific functional signatures, a differential activity analysis was performed on the matrix $\mathbf{Y}$, containing the pathway enrichment score for each cell. We group cells by their annotated types $c \in \mathcal{C}$, where C is the total cell types in the dataset, and apply a t-test with overestimated variance:

$$t = \frac{\mu_{k,c} - \mu_{k,\mathcal{C}\setminus c}}{\sqrt{\frac{\sigma^2_{k,c}}{n_c} + \frac{\sigma^2_{k,\mathcal{C}\setminus c}}{n_{\mathcal{C}\setminus c}}}}, \tag{11}$$

where $\mu_{k,c}$ is the mean activity of pathway k in cell type c, and $\mathcal{C} \setminus c$ denotes all the other types. This identifies pathways that are uniquely activated within specific cell populations based on the foundation model's learned representations.

B.10 Spatial Gene Imputation

We perform reference-based spatial gene imputation to evaluate the biological fidelity of the learned cell in zero-shot setting. This task assumes that single-cell and spatial transcriptomics profiles map to a shared latent manifold where proximity implies biological similarity.

Problem Formulation Let $\mathcal{D}_{st}$ and $\mathcal{D}_{sc}$ denote the spatial transcriptomics and single-cell reference datasets, respectively. Let $\mathcal{M}$ be the comprehensive set of genes profiled in the single-cell reference, and $\mathcal{M}'$ be the set of genes captured in the spatial data, such that $\mathcal{M}' \subset \mathcal{M}$.

For an arbitrary sample i from modality $m \in \{sc, st\}$, let $\mathbf{H}_i^{(m)} \in \mathbb{R}^{L_i\times d}$ denote the contextualized hidden states extracted from the last layer of the pretrained foundation model, where L_i represents the sequence length (number of detected genes) and d is the embedding dimension. We derive a global semantic embedding $\mathbf{h}_i^{(m)} \in \mathbb{R}^d$ for this sample by averaging representations across the gene sequence dimension:

[3] decoupler single-cell enrichment analysis: https://decoupler.readthedocs.io/en/latest/notebooks/scell/rna_sc.html

$$\mathbf{h}_i^{(m)} = \frac{1}{L_i} \sum_{t=1}^{L_i} \mathbf{H}_{i,t,:}^{(m)}. \tag{12}$$

Formally, we define the spatial dataset as $\mathcal{D}_{st} = \{(\mathbf{x}_i^{(st)}, \mathbf{h}_i^{(st)})\}_{i=1}^{N}$, where $\mathbf{h}_i^{(st)} \in \mathbb{R}^d$ is the average semantic embedding for a spot and $\mathbf{x}_i^{(st)} \in \mathbb{R}^{|\mathcal{M}'|}$ is the raw expression vector. Similarly, the reference dataset is defined as $\mathcal{D}_{sc} = \{(\mathbf{x}_j^{(sc)}, \mathbf{h}_j^{(sc)})\}_{j=1}^{K}$, where $\mathbf{x}_j^{(sc)} \in \mathbb{R}^{|\mathcal{M}|}$ represents the comprehensive transcriptomic profile. Our objective is to reconstruct the full gene expression profile (specifically for genes in $\mathcal{M} \setminus \mathcal{M}'$ or a hold-out test set) for the spatial spots by leveraging the single-cell reference data.

Imputation Workflow We employ a **weighted k-nearest neighbors** approach. Given a spatial spot i, we would like to find the k-nearest cells (which we define as $\mathcal{N}_{\text{k}}^i$) in the embedding space, where k we set to 50 in our experiment. Then we perform the imputation for the unseen genes ($\mathcal{M} \setminus \mathcal{M}'$) in spatial data using a weighted average of the comprehensive genes ($\mathcal{M}$) in single cell. More details about how we calculate the weights, and the algorithm implementations are illustrated as follows:

1. Cosine Similarity Computation: We compute the cosine similarity score between spatial spot i and reference single cell j:

$$S_{ij} = \frac{\mathbf{h}_i^{st} \cdot \mathbf{h}_j^{sc}}{\|\mathbf{h}_i^{st}\|_2 \|\mathbf{h}_j^{sc}\|_2}. \tag{13}$$

2. Neighboring Weights Normalization: For each spatial spot i, we identify the set $\mathcal{N}_{\text{k}}^i$ of the top-k most similar reference cells. We compute normalized weights using the softmax function:

$$w_{ij} = \frac{\exp(S_{ij})}{\sum_{m \in \mathcal{N}_{\text{k}}^i} \exp(S_{im})} \quad \forall j \in \mathcal{N}_{\text{k}}^i. \tag{14}$$

3. Imputation: The imputed expression vector $\hat{\mathbf{x}}_i^{st} \in \mathbb{R}^{|M|}$ is calculated as:

$$\hat{\mathbf{x}}_i^{st} = \sum_{j \in \mathcal{N}_{\text{k}}^i} w_{ij} \mathbf{x}_j^{sc}. \tag{15}$$

This operation effectively transfers the comprehensive gene information from the reference scRNA-seq dataset to the spatial spots.

Spatial Imputation Datasets The spatial imputation datasets, their paired scRNA-seq data, and the shared common genes for model evaluation, are provided in ATable 15. the single cell reference data of lung cancer and breast cancer were retrieved from Kim et al. [27] and [51] (GEO accession number: GSE131907 and GSE176078 respectively). The scRNA-seq of brain tissue was downloaded from SEA-AD (donor_H21_032) [17]. The corresponding lung-tissue spatial dataset (*Lung cancer 2*) was obtained from the Vizgen MERFISH FFPE Human Immuno-oncology dataset[4], the paired breast cancer was sourced from donor (CID3586) in [51] and the brain tissue spatial data was retrieved for donor_H21_032 in SEA-AD.

Experiments and Evaluation We first retrieve the shared common gene sets ($\mathcal{M} \cap \mathcal{M}'$) between single cell and spatial transcriptomics data. Then the shared genes are partitioned into 10-fold, we employ k-fold cross-validation to assess the imputed spatial gene performance for each held-out gene sets. Performance is quantified using Pearson Correlation (r), Spearman Rank Correlation (ρ), and Kendall's Tau (τ) between the imputed gene vector $\hat{\mathbf{x}}$ and ground truth $\mathbf{x}$ across all test spatial spots. the evaluation metrics will be introduced in more detail in Appendix B.11.5.

B.11 Evaluation Metrics

We employ different evaluation metrics depending on the downstream tasks. For zero-shot clustering, we report standard scIB metrics [32]: adjust random index (ARI), normalized mutual information (NMI), average silhouette width (ASW_{label}), and their arithmetic mean AvgBio to quantify the

[4] https://info.vizgen.com/ffpe-showcase?submissionGuid=88ba0a44-26e2-47a2-8ee4-9118b9811fbf

separation of different cluster/cell-types. To assess batch mixing, we use $\text{ASW}_{label/batch}$ and GraphConn and report their mean (AvgBatch). And the overall scIB score is calculated as a weighted average of these two average metric (overall = 0.6× AvgBio + 0.4× AvgBatch). We also report scGraph-RWR [50] and scGraph-LCA. The former uses the Random Walk with Restart (RWR) algorithm to obtain cell-type hierarchical distances on the ontology graph and the latter is adapted from scGraph-RWR by replacing the ontology-RWR reference distance matrix with LCA-based distance matrix. Both metrics quantify graph-level similarity between the reference ontology distance matrix and the cell-type centroid embedding distance matrix, providing a complementary evaluation of the intrinsic biological structure of cell embeddings in scFMs [47]. For cell type annotation fine-tuning, we compute overall accuracy to measure general performance and report macro F1 to equally weight all classes regardless of imbalance. For novel cell identification, we compute per-cell-type accuracy and report macro-accuracy to account for class imbalance. For gene function prediction, we report micro/macro precision and micro/macro F1. For spatial gene imputation, we quantify the agreement between imputed and ground-truth spatial expression using Pearson, Spearman, and Kendall's τ correlation coefficients [35]. The mathematical formulations of each of these metrics are provided in the following sub-sections.

B.11.1 Zero-shot clustering evaluation metrics

Normalized Mutual Information (NMI):

$$\text{NMI}(C, K) = \frac{I(C; K)}{\frac{1}{2}[H(C) + H(K)]},$$

where:

- C and K are the true cell type labels and predicted cluster labels.
- $H(C) = -\sum_i P(c_i) \log P(c_i)$ is the entropy of the labels.
- $I(C; K) = \sum_{i,j} P(c_i, k_j) \log \frac{P(c_i, k_j)}{P(c_i)P(k_j)}$ is the mutual information.

Adjusted Rand Index (ARI):

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

where the unadjusted Rand Index (RI) is:

$$\text{RI} = \frac{TP + TN}{TP + FP + FN + TN} = \frac{a + b}{\binom{N}{2}},$$

- a: Number of pairs of cells in the same set in C and the same set in K.
- b: Number of pairs of cells in different sets in C and different sets in K.

Average Silhouette Width (ASW_{label}) - Cell Type:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad \text{ASW}_{label} = \frac{1}{N} \sum_{i=1}^{N} s_i.$$

The commonly used version is the rescaled version down below, and we also use the scaled version in our experiment.

$$\text{ASW}_{label} = \frac{\text{ASW}_{label} + 1}{2}.$$

Average Biological Score (AvgBio):

$$\text{AvgBio} = \frac{1}{3}(\text{NMI} + \text{ARI} + \text{ASW}_{label}).$$

Average Silhouette Width ($\text{ASW}_{label/batch}$) - Batch:

$$\text{ASW}_{label/batch} = \frac{1}{|C|} \sum_{c \in C} \left[1 - \left| \frac{1}{|N_c|} \sum_{i \in N_c} s_i^{\text{batch}} \right| \right].$$

where s_i^{batch} is the silhouette score calculated using batch labels (instead of cell types) within a specific cell type group c.

Graph Connectivity (GraphConn):

$$\text{GraphConn} = \frac{1}{|C|} \sum_{c \in C} \frac{|\text{LCC}(G_c)|}{|N_c|},$$

where:

- G_c: The k-nearest neighbor (kNN) subgraph containing only cells of type c.
- $\text{LCC}(G_c)$: The largest connected component in that subgraph.

Average Batch Score (AvgBatch):

$$\text{AvgBatch} = \frac{1}{2}(\text{ASW}_{label/batch} + \text{GraphConn}).$$

Overall scIB score: overall score $= 0.6 \times \text{AvgBio} + 0.4 \times \text{AvgBatch}$.

scGraph-LCA To quantify how well the learned latent embedding preserves the hierarchical relationships defined in the Cell Ontology, we employ a correlation-based graph similarity measurement metric, named scGraph-LCA. Our metric generalizes the evaluation protocols proposed in scGraph [47] and scGraph-RWR [50]. It relies on the comparison of two pairwise distance matrices: one derived from the model's latent space (Δ^{emb}) and one derived from the ontology graph (Δ^{onto}). We will illustrate each component one by one in the following:

1. Model embedding distance matrix (Δ^{emb})

Following the notation in Section 2.1 in the main paper, here, let $\mathcal{D} = \{x_n, c_n\}_{n=0}^{N}$ denotes a single-cell RNA-seq dataset, where each cell n is associated with a cell type label $c_n \in \mathcal{V}$. We reuse the notation of Cell-Dataset Index ($\mathcal{I}_{c,k}$) here, let $\mathcal{I}_c = \bigcup_{k \in \mathcal{K}_c} \mathcal{I}_{c,k}$ represent the set of indices for all cells annotated as type $c \in \mathcal{V}$ across all available datasets/batches, here k could interchangeably represent as batches, since if we only evaluate on one dataset, the number of available dataset is always being 1, however, there may exist multiple batches in one dataset.

For each unique cell type $c \in \mathcal{V}$ present in the data, we compute the centroid in the embedding space. To mitigate the effect of outliers, we employ a trimmed mean aggregation. Let $\{\mathbf{z}_n \mid n \in \mathcal{I}_c\}$ be the set of embedding vectors for all cells of type c. We compute the trimmed mean centroid $\boldsymbol{\mu}_c$:

$$\boldsymbol{\mu}_c = \text{TrimmedMean}\left(\{\mathbf{z}_n \mid n \in \mathcal{I}_c\}, \alpha\right), \tag{16}$$

where α is the trimming proportion (set to 0.2 in our experiments).

We then construct the pairwise **Embedding Distance Matrix** $\Delta^{\text{emb}} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where the entry (i, j) represents the Euclidean distance between the centroids of cell types c_i and c_j:

$$\Delta_{ij}^{\text{emb}} = \|\boldsymbol{\mu}_{c_i} - \boldsymbol{\mu}_{c_j}\|_2. \tag{17}$$

2. Ontology Reference Distance Matrix ($\boldsymbol{\Delta}^{\text{onto}}$)

The reference matrix encodes the "ground truth" biological distances extracted from the Cell Ontology graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. , we first introduce the LCA distance matrix, then in the following we will derive the scGraph-RWR based on RWR distancem matrix.

As introduced in Section 2.1, we utilize the Lowest Common Ancestor (LCA) to capture the hierarchical distances on the ontology graph. For any two types of cells $c_i, c_j \in \mathcal{V}$, their raw semantic distance $d_{\mathcal{G}}(c_i, c_j)$ is calculated as the sum of the shortest path distances from each node to their common $c_s = \text{LCA}(c_i, c_j)$:

$$\Delta_{ij}^{\text{onto, LCA}} = \delta_{\mathcal{G}}(c_i, c_j) = \delta_{path}(c_i, c_s) + \delta_{path}(c_j, c_s), \tag{18}$$

where $\delta_{path}(c_i, c_s)$ and $\delta_{path}(c_j, c_s)$ denote the number of edges on the shortest path between nodes (c_i, c_j) to their lowest common ancestor c_s in $\mathcal{G}$, here we use $\Delta_{ij}^{\text{onto, LCA}}$ to distinguish the RWR-based distance matrix ($\Delta_{ij}^{\text{onto, RWR}}$) in scGraph-RWR.

3. Graph Similarity Measurement

To quantify the similarity between the model embedding distance matrix and ontology-based distance matrix, we calculate Pearson Correlation and the rank-based correlation (Spearman, and Kendall's τ) between the columns of the distance matrices. For each cell type c_i, let $\mathbf{r}_i^{\text{emb}}$ be the i-th column of $\Delta^{\text{emb, LCA}}$ and $\mathbf{r}_i^{\text{onto}}$ be the i-th column of $\Delta^{\text{onto, LCA}}$. We compute the agreement score using Pearson Correlation (r), Spearman's rank correlation (ρ) and Kendall's rank correlation coefficient (τ), the full mathematical equations of the above metrics are illustrated in more detail in Appendix B.11.5:

$$\text{scGraph-LCA}_r = \frac{1}{K}\sum_{i=1}^{K}\rho(\mathbf{r}_i^{\text{emb}}, \mathbf{r}_i^{\text{onto}}), \tag{19}$$

$$\text{scGraph-LCA}_\rho = \frac{1}{K}\sum_{i=1}^{K}\rho(\mathbf{r}_i^{\text{emb}}, \mathbf{r}_i^{\text{onto}}), \tag{20}$$

$$\text{scGraph-LCA}_\tau = \frac{1}{K}\sum_{i=1}^{K}\tau(\mathbf{r}_i^{\text{emb}}, \mathbf{r}_i^{\text{onto}}). \tag{21}$$

A higher correlation indicates that cell types close in the ontology graph are also geometrically proximal in the model's latent space, demonstrating preservation of biological semantics.

scGraph-RWR According to [50], scGraph-RWR utilizes a weighted ontology graph, in our experiment the weights are download from [48]. Regarding the reference distance matrix, we just need to replace the LCA-based ontology distance matrix with the RWR-based ontology distance matrix by following scGraph-LCA procedures. The process to get $\Delta^{\text{onto, RWR}}$ is as follows:

First, we define the column-normalized transition probability matrix $\mathbf{P}$ as:

$$\mathbf{P} = \mathbf{A}\mathbf{D}^{-1}, \tag{22}$$

where $\mathbf{A}$ is the adjacency matrix of ontology graph $\mathcal{G}$ and $\mathbf{D}$ is the diagonal degree matrix. The RWR diffusion matrix $\mathbf{R}$, which captures the stationary probability distribution for walkers starting at each node, is computed analytically as:

$$\mathbf{R} = \beta\left(\mathbf{I} - (1-\beta)\mathbf{P}\right)^{-1}, \tag{23}$$

where β is the restart probability (regulating the diffusion scale) and $\mathbf{I}$ is the identity matrix. Finally, the ontology-derived RWR distance between cell types i and j is defined as the complement of the diffusion probability:

$$\Delta_{ij}^{\text{onto, RWR}} = 1 - R_{ij}. \tag{24}$$

B.11.2 Cell annotation (Fine-tuning) metrics

Let C be the complete set of unique cell types present in the evaluation dataset. The overall classification performance is evaluated using Accuracy and Macro F1 score. The **Accuracy** is defined as the overall proportion of correctly classified cells out of the total number of cells:

$$\text{Accuracy} = \frac{\sum_{c\in C} TP_c}{N},$$

where:

- TP_c: The number of cells correctly classified as cell type c (True Positives).
- N: The total number of cells across all types in the evaluation dataset.

The **Macro F1** score is defined as the arithmetic mean of the individual F1 scores computed independently for each cell type $c \in C$, treating all classes equally regardless of their support size:

$$\text{Macro F1} = \frac{1}{|C|}\sum_{c\in C}\text{F1}_c,$$

where F1_c is the harmonic mean of the Precision and Recall for a specific cell type c:

$$\text{F1}_c = 2 \times \frac{\text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c},$$

with Precision and Recall defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$

- FP_c: The number of cells incorrectly classified as cell type c (False Positives).
- FN_c: The number of cells truly belonging to cell type c but incorrectly classified as another type (False Negatives).
- $|C|$: The total count of unique cell types in the evaluation set.

B.11.3 Novel cell identification metrics

Let $U \subset V$ be the set of unseen cell types (where V is the total set of cell ontology nodes). The Macro Accuracy over the unseen set U is defined as the arithmetic mean of the recall (accuracy) for each unseen class:

$$\text{Macro Accuracy}_U = \frac{1}{|U|} \sum_{u \in U} \text{Acc}_u,$$

where Acc_u is the accuracy for a specific cell type u:

$$\text{Acc}_u = \frac{TP_u}{N_u},$$

- TP_u: The number of cells correctly classified as type u (True Positives).
- N_u: The total number of cells in the dataset with true label u (Support).
- $|U|$: The count of unique unseen cell types.

B.11.4 Gene function prediction

We evaluate the multi-label GO function classification performance using Micro-averaged and Macro-averaged metrics. Let C be the total number of classes (in which we have three broad GO terms: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)). For each specific class $i \in \{1, \ldots, C\}$, we define:

- TP_i: True Positives (samples correctly predicted to have GO term i).
- FP_i: False Positives (samples incorrectly predicted to have GO term i).
- FN_i: False Negatives (samples with GO term i that were missed).

Macro-Averaged Metrics Macro metrics calculate performance for each GO class independently and then take the unweighted average. This ensures that specific, rare GO class contribute equally to the score as broad, common terms.

$$\text{Precision}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i},$$

$$\text{Recall}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i},$$

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}.$$

Micro-Averaged Metrics Micro metrics aggregate the prediction counts across all GO class globally before calculating the score. This gives a measure of the overall system reliability, weighted by the frequency of the terms.

$$\text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FP_i)},$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} (TP_i + FN_i)},$$

$$\text{F1}_{\text{micro}} = \frac{2 \times \text{Precision}_{\text{micro}} \times \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}.$$

B.11.5 Spatial imputation evaluation metrics

Pearson Correlation (r)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}.$$

Spearman's Rank Correlation (ρ)

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

where $d_i = \text{rank}(x_i) - \text{rank}(y_i)$.

Kendall's Tau (τ)

$$\tau = \frac{C - D}{\frac{1}{2} n(n-1)}.$$

where C is the number of concordant pairs and D is the number of discordant pairs.

C RESULTS

C.1 Zero-shot clustering SOTA benchmark performance of scIB metrics

C.1.1 Overall scIB score across the evaluation tissues

The overall of the scIB scores for each dataset are provided in ATable 16.

C.1.2 Details of each scIB score across the evaluation tissues

For the details of each scIB metrics: ARI, NMI, ASW$_{label}$, GraphConn, ASW$_{label/batch}$, AvgBio, AvgBatch, in HIHA, Heart, Pancreas, Lung, Liver, Breast_TNBC, Kidney, Adipose Tissue, Bone Marrow. Please refer to the ATable 17.

C.1.3 Zero-shot clustering of scGraph-LCA benchmarking performance (Table 18)

C.1.4 Zero-shot clustering of scGraph-RWR benchmarking performance(Table 19)

C.1.5 Zero-shot clustering of 340 in-distribution cell types in unseen datasets

We report the AvgBio metrics (ARI, NMI, ASW$_{label}$) in scIB (ATabel 20), since we randomly down-sample the cell types with more than 10,000 cells, it is hard for us to observe the batch/dataset id in this process, for simplicity of the implementation, we only calculate the AvgBio related metrics. Additionally, we also report scGraph-LCA and scGraph-RWR in ATabel 20. Note that for the scGraph-RWR metric, 48 cell types are not present in reference embedding, so the total cell types for this metric is 292.

C.1.6 Bone marrow clustering and graph distance visualization

We also demonstrated that the learned embeddings of scOntoFM well reflect the intrinsic biologically hierarchical structure of hematopoietic related cell types (AFigure 11c), showing strong empirical consistency with ontology-based distances (AFigure 12). Furthermore, the ontology-derived distance heatmap in AFigure 11b closely aligns with the G2/M and S-phase cell cycle scores reported in the original study AFigure 11d. Taken together, these results indicate that our model could effectively capture the change in cell states along continuous developmental trajectories.

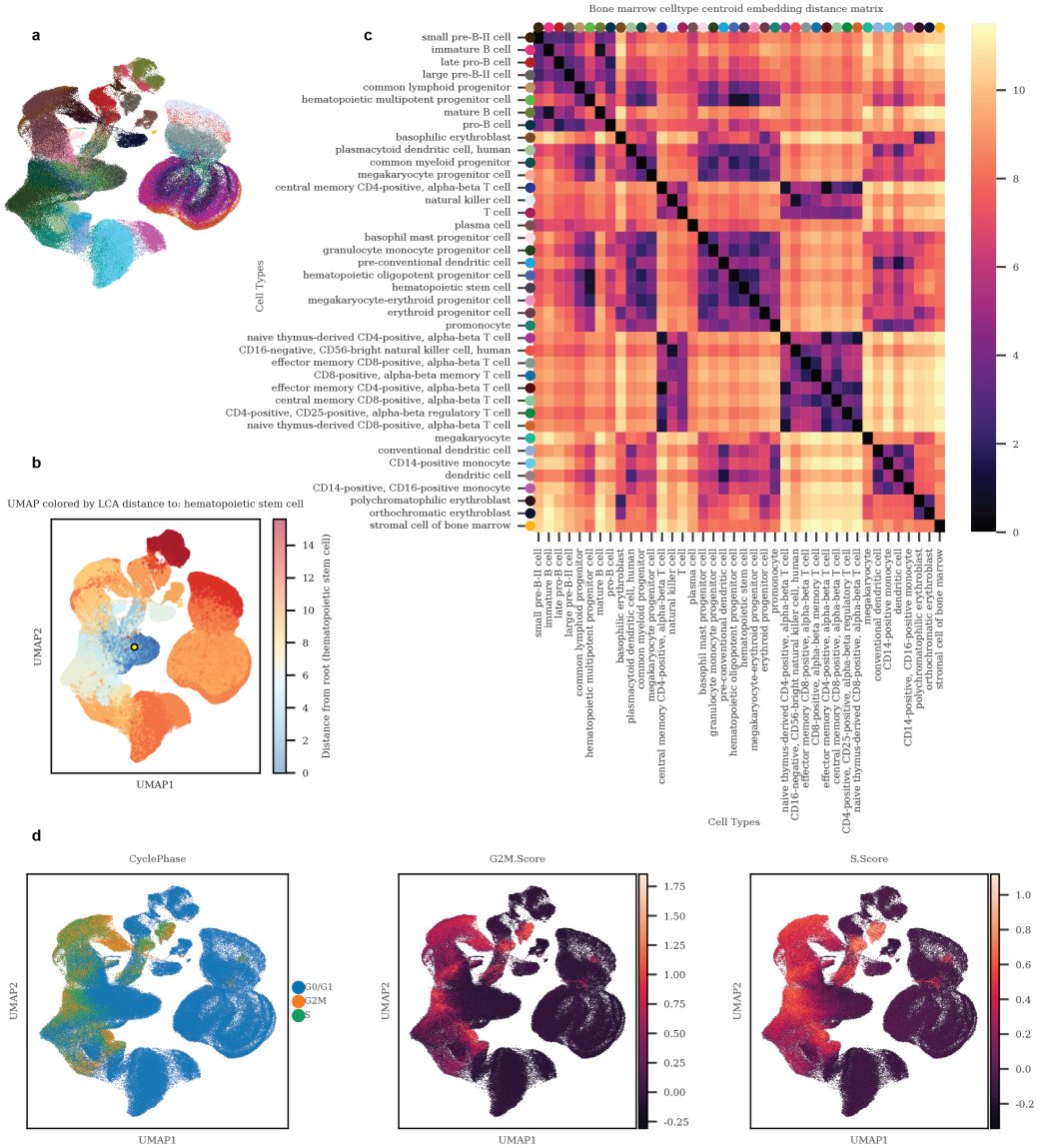

Figure 11: (a) Clustering results of Bone marrow dataset using our model embeddings, colors represent different cell-types. (b) Distance heatmap from root centroid cell-type (hematopoietic stem cell, highlighted as yellow dot in the umap) to all the other cell-types in the dataset. (c) Cell-type centroid embedding distances heatmap for the bone marrow dataset. (d) Cell cycle scores heatmap from original paper [55].

C.2 Novel cell identification results

In AFigure 13(a, b), we first demonstrate that our model has learned biological meaningful structure of known cell types, cell-type labels are aggregated into coarser level according to the lineage of

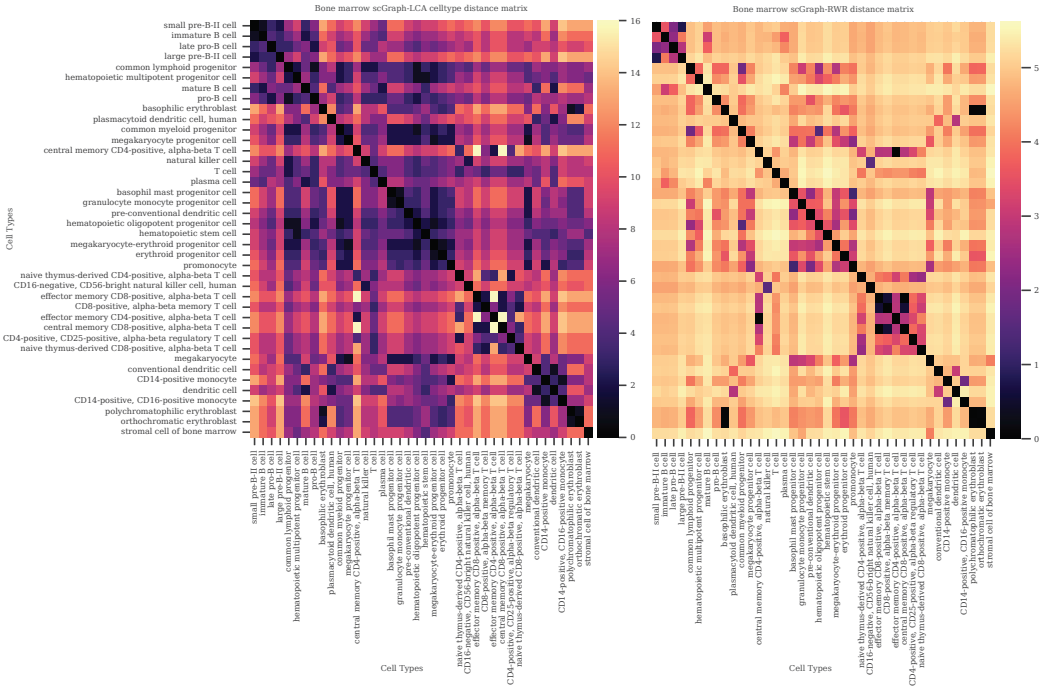

Figure 12: Left: LCA distance reference heatmap of scGraph-LCA in Bone marrow datast; Right: Random walk with restart (scGraph-RWR) of ontology graph distance in Bone marrow dataset.

ontology tree. Then built on known cell-type centroid embeddings and the comprehensive LCA-based ontology affinity matrix introduced in Appendix B.8, scOntoFM could infer the unseen cell-type centroids as shown in AFigure 13(c, d). Concretely, we conducted a specific case study on three unseen hematopoietic cell types: megakaryocyte progenitor cells, immature neutrophils, and CD14-positive/CD16-low monocytes. As AFigure 14 shows, our model accurately positions unseen cells near their biological relatives. This alignment with ground-truth ontology distances confirms the model's ability to correctly infer the biological neighborhood of unknown cell types

C.3 Gene function prediction results

In AFigure 15, we show additional results of the *COAGULATION* hallmark gene set and the *ALLOGRAFT_REJECTION* hallmark gene set enrichment score in the lung adenocarcinoma dataset [27]. The enrichment scores for the *ALLOGRAFT_REJECTION* hallmark are notably elevated in immune-related populations—specifically macrophages, T cells, and NK cells—while remaining low in non-immune lineages. Particularly, the *COAGULATION* hallmark is highly enriched in the cancer-associated fibroblasts (CAFs), which contributes hypercoagulable state and facilitate tumor immune evasion [53]. These distinct enrichment patterns confirm that our model accurately disentangles the complex functional heterogeneity within the tumor microenvironment.

C.3.1 GO namespace classification by aggregating gene embeddings of the GO terms.

Beyond classifying semantic meaning of single gene to GO terms, we further evaluated GO namespace classification by aggregating the gene embeddings for specific GO terms. The results are reported as mean±std over 5 random seeds. As ATable 21 shows, scOntoFM achieves highly competitive results, ranking second only to scGPT, while still leading other Geneformer-like backbone. The performance difference is expected given scGPT's significantly larger vocabulary size (60,000 *vs.* 20,000 genes), and embedding dimensionality (512 *vs.* 256) compared to scOntoFM.

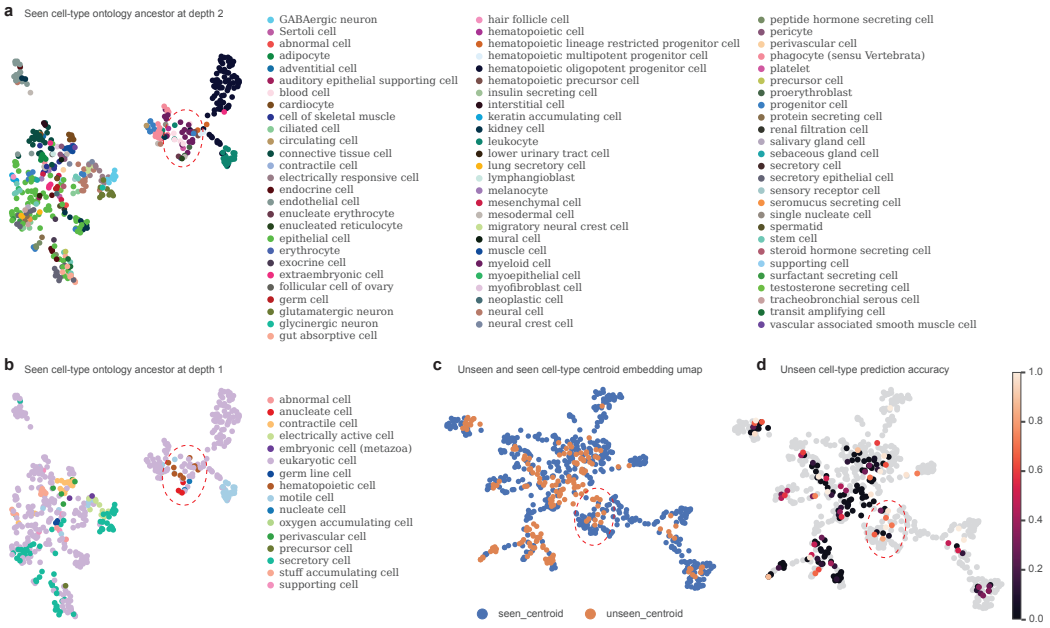

Figure 13: (a) Umap visualization of seen cell-type centroid embedding, cell-types are colored by its ontology ancestor at depth 2. (b) Same as (a), but cell-types are colored by its ontology ancestor at depth 1. (c) Co-embedding seen centroid embeddings and unseen cell-type centroid embedding, dots are colored by seen centroid (blue) and unseen centroid (orange). (d) Our model predicted accuracy heatmap for unseen cell-types. The red circles in all the umaps are the cell-types coming from hematopoietic or myeloid cell lineage, in which the unseen cell-types (megakaryocyte progenitor cell, immature neutrophil, CD14-positive, CD16-low monocyte) in Figure 14 are selected from.

C.4 Spatial imputation results

In AFigure 16, we show an additional violin plot for spatial gene imputation in terms of Pearson, Spearman, and Kendall's τ correlation metrics across the three evaluation datasets (MERSCOPE Lung cancer 2, breast cancer and SEA-AD). We conducted the Mann-Whitney U tests between our method other scFMs on spatial gene imputation task. Our method is statistical significant *vs.* other scFMs in both the Lung and breast cancer data except Geneformer.

C.5 Ablation study about the loss terms in scOntoFM (ATable 24)

C.5.1 Pareto analysis across the nine tissues in our study

The highlighted points represent the average values across each ablation setting. The plot (Figure 6) demonstrates that scOntoFM achieves an optimal balance between standard integration metrics (scIB) and ontology-based metric (scGraph-LCA).

C.6 Robustness of scOntoFM to Cell Type Annotation Noise

Cell type annotation noise and inconsistency across datasets can introduce substantial bias during pre-training. To mitigate this, we pre-trained scOntoFM on the CZI CELLxGENE platform, where metadata curation—particularly cell type standardization—is a large-scale, expert-driven effort involving resources such as the Human Cell Atlas (HCA), Tabula Sapiens, and HuBMAP. Pre-training on this well-curated and standardized corpus not only avoids performance degradation but also enables scOntoFM to better capture hierarchical relationships between cell types and preserve biologically meaningful structure.

We evaluate robustness from two complementary aspects. First, robustness across annotation levels in zero-shot clustering: scOntoFM consistently maintains strong performance under varying label gran-

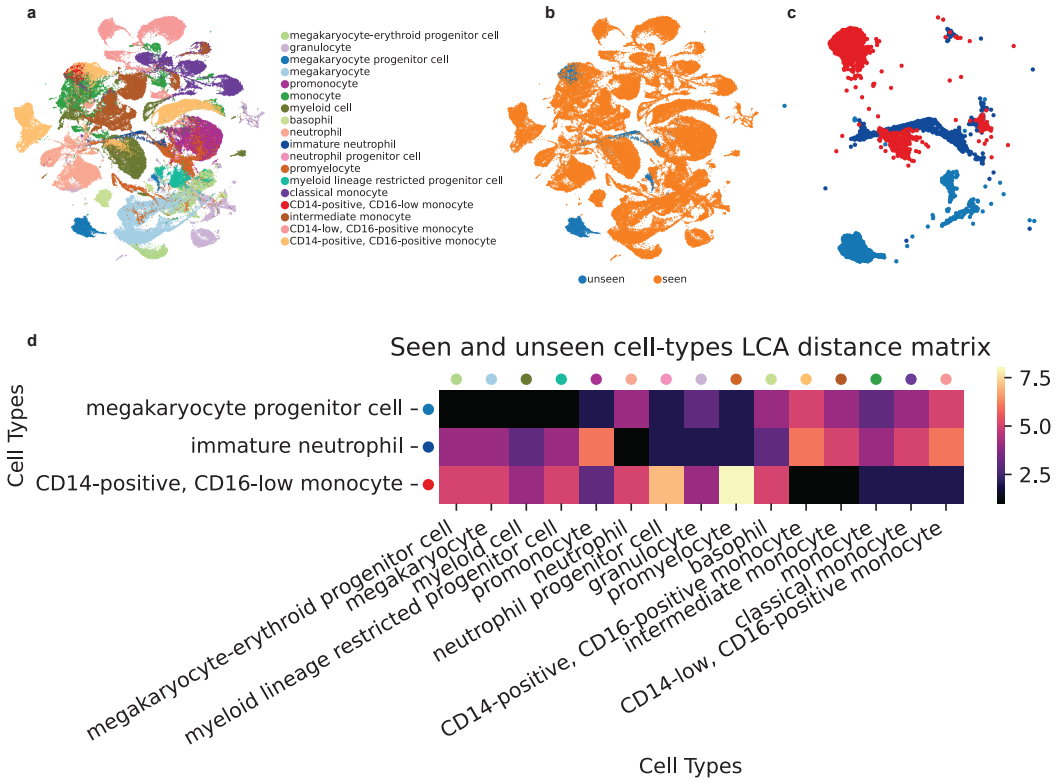

Figure 14: (a) Umap of our model embeddings on three unseen cell types and the top 5 closest seen cell types for each unseen cell type. (b) Umap visualization of seen (orange) and unseen (blue) labels. (c) Umap visualization of the three query unseen cell types. (d) LCA distance matrix of unseen cell types and the top 5 closest seen cell types for each unseen cell type.

ularities (Table 2), particularly at fine-grained levels (*e.g.*, AIFI_L1/L2 in HIHA and SubclassLevel2 in Kidney), highlighting its ability to capture subtle cell states and complex lineage structures. This is further supported by high agreement scores (ARI) between cell ontology references and annotations at different resolutions (see ATable 25, Figure 17, 18), demonstrating stability under label mismatch. Second, resilience to label noise: under controlled noise injection on the Azimuth PBMC dataset, scOntoFM maintains high accuracy and stable performance as noise increases (ATable 26). Notably, at higher noise levels (*e.g.*, 50%), the model predominantly assigns cells to their parent categories, reflecting biologically consistent behavior when fine-grained labels become unreliable.

D Broader Impact of scOntoFM

This paper presents work whose goal is to advance the field of Machine Learning and Computational Biology. There are many potential societal consequences of our ontology-based foundation model, particularly in the interpretation of complex single-cell atlases. We hope that by integrating hierarchical biological priors, our model will aid researchers in uncovering accurate biological insights and accelerating disease characterization. However, we acknowledge that pretraining such foundation models is computationally intensive, and future efforts should address the energy efficiency and accessibility of these large-scale systems.

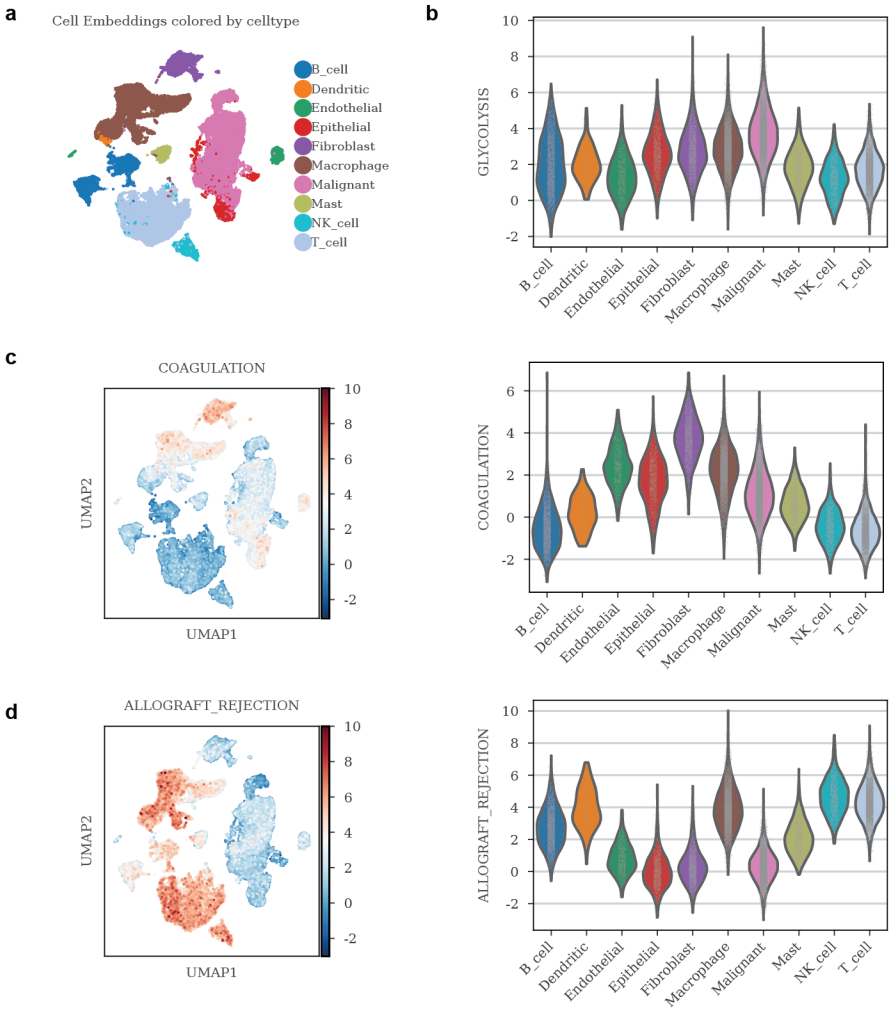

Figure 15: (a) Our model's umap embedding visualization colored by cell-types. (b) Violin plot of *GLYCOLYSIS* hallmark enrichment score distribution for each cell type. (c) Left: *COAGULATION* hallmark enrichment score heatmap. Right: Violin plot of *COAGULATION* hallmark enrichment score distribution for each cell type. (d) Left: *ALLOGRAFT_REJECTION* hallmark enrichment score heatmap. Right: Violin plot of *ALLOGRAFT_REJECTION* hallmark enrichment score distribution for each cell type.

Table 11: CZI query dataset id for the 212 unseen evaluation datasets.

Dataset ID	Dataset ID	Dataset ID
99e317c0-1fab-4f73-b511-859e752fd1c3	6725ee8e-ef5b-4e68-8901-61bd14a1fe73	e2808a6e-e2ea-41b9-b38c-4a08f1677f02
98f5d518-1e60-4307-b732-ab045cd3cbc2	bab7432a-5cfe-45ea-928c-422d03c45cdd	6c87755e-a671-41a8-9c7e-4e43b850a57b
e6361237-ac4e-4c5d-ad8f-f16aca0c0a8f	242c6e7f-9016-4048-af70-d631f5eea188	d5a2d011-3765-4dad-ac36-3fff087589fe
43245158-5ae1-4e71-a9a6-67eef49c26bc	a65bcc2d-4243-44c1-a262-ab7dcddfcf86	4e374783-0152-423e-bcb6-b87cc232f64d
5e7fff82-b8eb-4220-bf65-c6feb6765ea1	8d1dd010-5cbc-43fb-83f8-e0de8e8517da	b252b015-b488-4d5c-b16e-968c13e48a2c
72822932-10f6-466f-baf3-a2c1d89364bc	881fe679-c6e0-45a3-9427-c4e81be6921f	4dd1cd23-fc4d-4fd1-9709-602540f3ca6f
79ef1959-a6b4-4cac-82ca-30feaec48df1	01209dce-3575-4bed-b1df-129f57fbc031	30f5e171-83d7-4fc0-bf75-384f122346b3
5d3fc988-765d-48ba-bfb5-151ef2988cac	31285bcd-9a28-48ce-8cc0-543af80261d3	dc30c3ec-46d6-4cd8-8ec1-b544a3d0f503
3c361813-5d5e-4868-bfe5-e7591e7b5381	234c23d0-d6cd-4612-a40d-e5811727564b	76150f40-1989-4977-9e23-696e72d59d9e
171b03c9-516b-499c-a434-fa4f70191988	13149914-ea03-4f01-8bf6-b793b667127b	dbf0bd35-87f8-4b25-bc90-a3c54f379907
d7f5d8d0-6150-48d7-b094-c34286ad11a1	f8d8b443-bca6-4c3c-9042-669dfb7f8030	eeacb0c1-2217-4cf6-b8ce-1f0fedf1b569
3e524d45-2265-4440-b903-3c30d51a0b17	a20e2f2a-e171-450a-8ddc-8935c2977ad6	1fc48734-35ee-4372-885c-95fd2674c75a
731e6380-879f-4b0b-9a1f-2150208852ef	ab326369-b63c-48d8-8a7f-82ffa0df7002	8a554710-08bc-4005-87cd-da9675bdc2e7
0c9a8cfb-6649-4d52-b418-6d8e56bd7afe	1a974af7-8b4b-4abb-8aba-23e2979d9518	d18736c3-6292-4379-919a-d6d973204c87
9ea52f85-a35e-4b5b-9207-e2f1de7dbbcd	87a06a2a-16c6-42f7-af1b-529fad431051	51d9716c-50bb-4fd8-8d1e-7d9329969cb7
7c497c66-c2e2-4291-8b96-339e5e65f314	ef1a4b05-540c-4f73-8ceb-49dfa800645d	bd3e3487-c1e6-4298-b9de-8bdcf3f938d7
14363b6e-3428-45b0-8704-10fd5d6316df	17e9d436-a264-4c94-a42d-48c8daf6fdd9	e34f3f89-3a48-4173-a06a-9e9803f2e2b9
31f657dc-1875-4c4b-a5ca-ce63b3ef3a82	3079e9b0-c6db-45c8-b998-a4555a73968e	d9f3a82c-24bb-4277-b715-8ec92a33dd32
2856d06c-0ff9-4e01-bfc9-202b74d0b60f	74014ef8-d2d0-4cbc-8ba2-037f30753ffd	c893ddc3-f25b-45e2-8c9e-155918b4261c
6f7fd0f1-a2ed-4ff1-80d3-33dde731cbc3	e500acbf-f166-4c46-b290-f93506cf26d3	a4157949-6f2a-40e2-b960-63f6e2bde918
0de831e0-a525-4ec5-b717-df56f2de2bf0	93091496-be48-4122-b945-9af9c22a7535	73b7d0aa-035f-4a15-9f1d-f56b70204591
492d9f77-c4c8-485d-92d7-7686fe6c8482	f7ec7bd5-04ab-453b-a8a7-c9d14812affb	2e9d2f32-4cfb-49b5-b990-cbf4c241214e
d6505c89-c43d-4c28-8c4f-7351a5fd5528	251b1a7e-d050-4486-8d50-4c2619eb0f46	6e00ccf7-0749-46ef-a999-dba785630d52
94c41723-b2c4-4b59-a49a-64c9b851903e	b25f3834-69b3-4d87-a272-3938432d1f30	34f5307e-7b4d-4a48-b68f-2ba844c6414b
7f7faf6b-f11d-4f07-bc1c-188a4472748d	a9bedd04-51cb-42ad-83fe-cd749de71f1b	78c62d4e-abf1-405b-bb51-368720814529
9fcb0b73-c734-40a5-be9c-ace7eea401c9	f0f0d7c4-3bec-428e-9539-c99d36548d96	4a5b00e0-1ba3-4fd4-af89-d3512eb20720
b6e502d7-0d07-408e-a291-91b63af0a40a	5b25a46f-f1cf-4caa-9e10-6ab81dfe9174	5e717147-0f75-4de1-8bd2-6fda01b8d75f
50eb1e23-b8d4-4f76-a184-44e5541fa05a	55ca4411-8c7d-4aef-a572-50852e030c05	4023a2bc-6325-47db-bfdf-9639e91042c2
cd2f23c1-aef1-48ae-8eb4-0bcf124e567d	93cb76aa-a84b-4a92-8e6c-66a914e26d4c	e347396c-a7ff-4691-9f7a-99a43555ca18
cb252df6-6e49-4553-abd1-495a00006fb1	92161459-9103-4379-ae34-73a38eee1d1d	4a90bbba-df97-4a28-83ae-7386fe7d9167
9c1b5626-58df-4401-ae7b-f66d068c1551	68b6114f-e990-4033-bfb3-c35536633aaa	093d3bfe-6f0f-4ac0-a7a1-829f94d0a49f
07760522-707a-4a1c-8891-dbd1226d6b27	b62f728b-a7eb-4889-8162-b3d457de93b9	f5be4b96-f5a3-4c3d-84ac-6f69daf744d5
595c9010-99ec-462d-b6a1-2b2fe5407871	ce009dc1-ac57-4386-b72f-5c575701c253	8fbed309-d3d4-441b-b3ff-e2dcbcec2d35
cab0bc48-744c-461b-aaf4-7bf2cb7af00d	cff99df2-4904-44f7-9173-ff837f95606e	06ef6b36-6c9b-4e10-8a94-d0baf274276e
f7ca7eb9-0919-4eba-99a1-21ba382b4174	8dc29642-be06-46dc-a289-8075915745d4	b3c9904b-2ffd-4038-8620-ee1c5b072e9c
8c42cfd0-0b0a-46d5-910c-fc833d83c45e	8e5978bd-0e48-4a09-a92d-fa2985059f05	429d21fd-bb9a-4abf-922b-c15138524bdd
5829c7ba-697f-418e-8b98-d605b192dc48	6929d6d1-6f81-44d7-bd35-5a580259af81	b6b5ea88-e092-46e4-b9c5-e93b52d5c195
4859642c-d155-437b-adf5-83ca6c1bef67	821a8e3e-87ec-41ee-bd69-cd4377057d38	0bae7ebf-eb54-46a6-be9a-3461cecefa4c
a30db232-08a6-48f0-ab9f-ad6977b8a6b1	be41a86a-b606-4b1c-8055-32f334898775	76af615b-12fe-4ebb-b8c6-239ad29e26e3
2a65bd41-4a3f-42f4-98b2-c53962811152	ed9e9f96-4f08-49d2-bef5-b2c29adf3edc	b7efc0ba-9136-4f3f-adb2-d07849e59a3d
19e46756-9100-4e01-8b0e-23b557558a4c	e595c979-c2de-4dad-8c02-8334a8e56de6	eec804b9-2ae5-44f0-a1b5-d721e21257de
ca46ffe6-0a26-4d8d-82fa-81722daf1102	2727d83a-0af0-443a-bff8-58dc7028289a	e7c5ba7e-4be0-4696-af50-a07d0dc7aac5
a82c43bb-a703-446d-aa24-049bf013121f	716a4acc-919e-4326-9672-ebe06ede84e6	12c3efd8-d515-42e6-b93c-4961e599d524
9fddb063-056d-4202-8b8a-4b0ee531d3ce	0bc7235a-ae5a-479d-a487-510435377e55	8d73847b-33e7-48f0-837f-e3e6579bf12d
e3cc5b85-62ec-4add-bee3-73c8fd11e59d	e2a325e2-5633-4950-ae46-ebff70979e3b	0a2d7e87-c3c0-4ed2-86df-ae18811fcc16
01c93cf6-b695-4e30-a26e-121ae8b16a9e	be401db3-d732-408a-b0c4-71af0458b8ab	2f05ab20-a092-4bab-9276-3e0eb24e3fee
b550fd94-97e0-42e2-90f7-7434781ea2f8	43560d7c-acd6-4ede-830d-c9ffc6cd0862	e5f5d954-cf0e-4bd8-9346-8d1ddf15a08b
a19d1667-a7b5-4556-9e5f-f9bfa690c0f1	a12ccb9b-4fbe-457d-8590-ac78053259ef	7b55fe5c-d5c8-48e0-a1a3-54c5b9074f3f
1873a18a-66fd-4a4d-8277-a872c93f5b59	4ce8d8ec-b94c-4b79-afcb-b094336d8a78	149b2c3f-ee11-47a7-984b-923570280bd7
1a38e762-2465-418f-b81c-6a4bce261c34	51f476f7-b24d-42f3-8871-7dab3fa35e96	3966ba97-beb8-4d0b-9954-d3775cd2cd61
d567b692-c374-4628-a508-8008f6778f22	4993d61c-1d04-4630-9c61-8d9431f39adc	df1edb87-e512-43ae-b5f4-cb179cfc2bb4
6c600df6-ddca-4628-a8bb-1d6de1e3f9b4	f9cfac8d-bff6-47a2-a1f6-503827d375f5	15d374d6-0dfd-4d8e-ade7-81f73dc921ee
bcdec5fa-a7fa-4806-92bc-0cd02f40242f	6ceeaa86-9ceb-4582-b390-6d4dd6ff0572	c54c9659-1b6b-4d4a-96f0-7dc987410f2d
59b69042-47c2-47fd-ad03-d21beb99818f	33911db3-f461-464b-8083-a397ab616a09	cf83c98a-3791-4537-bbde-a719f6d73c13
29244e1d-02e6-4133-b411-516ef7474638	7d98cc44-b090-4dc8-804f-2750c84fe9d7	ac0c6561-7a48-4185-af6f-af799f699172
9df60c57-fdf3-4e93-828e-fe9303f20438	350237e0-9f48-4cbd-9140-3b44495549f3	390b376d-8857-4b4b-991f-0d7326938172
02792605-4760-4023-82ad-40fc4458a5db	8dc858b0-b076-4326-94bb-7c2d1cd721a3	ea786a06-5855-48b7-80d7-0313a21a2044
f1606894-59df-4794-a37f-baa7c6fb6de1	da75ce6d-a395-4abd-962b-267aadb99666	9a281de7-cee5-4e80-8584-1929f46f152f
2dd9d7a2-4bee-4c3f-a1f5-89bc3271c4f0	62de80d7-e6c4-4ff0-ad4c-d3d36f57cb93	d967b47c-a9e6-4337-b2f4-977f690cb67f
84230ea4-998d-4aa8-8456-81dd54ce23af	0087cde2-967d-4f7c-8e6e-40e4c9ad1891	db59611b-42de-4035-93aa-1ed39f38b467
72955cdb-bd92-4135-aa52-21f33f9640db	364bd0c7-f7fd-48ed-99c1-ae26872b1042	7e7f63c5-d964-40be-83de-ecbcccafd233
bdd109f1-ca27-45dc-bb9f-473008449428	96baf2e0-60c5-48cf-b1d1-cba657851507	b61a921b-7fa3-4b42-b455-aaaf32447920
3b8b5de4-3aa1-4ac6-8890-8d03c8219981	2ac76f1b-43ef-4271-8686-2f165570989f	3241c34f-581a-4de3-81fe-28e3de182680
f5be9ed2-5d41-4a52-91e4-4ff24ff84900	c874f155-9bf9-4928-b821-f52c876b3e48	8f4f8502-9170-4ac2-9707-3b6985ebfe5f
55003f67-c494-46f1-83fb-902745646379	78f40dd6-ed22-470d-8547-6fb5746c25da	f156606a-dd9a-49fd-bc40-0e069b6cf07c
460dfe21-0550-4521-92fe-db0f3036d8b0	2f7230e7-4706-4924-b571-473fc1e52c30	b6ee0939-b5b9-4e03-9d69-96fb6bf5d434
9a64bf99-ebe5-4276-93a8-bee9dff1cd47	1196334f-a441-474a-95ee-400780941690	f5b0810c-1664-4a62-ad06-be1d9964aa8b
601c7454-17ff-432c-8e0d-902bfddb8013	ecce9688-66ea-4600-98f7-84aed7421c01	eb499fd8-7000-419f-8854-926b9b61b11f
d3cb449b-c2b1-4b50-a7f1-21203535fe61	0436a180-cb44-47ba-8ffa-807b7a468469	d3d4baff-d142-4f13-acc6-96a4ef6a3e95
a199ca73-035d-44e2-9893-4c493151db21	dea717d4-7bc0-4e46-950f-fd7e1cc8df7d	34229bdc-a895-4394-8820-574e4028d8c6
7b20c613-9add-43d1-87e9-defd3d9b9f8c	ea95cdd5-cf7e-4d5d-a0bd-032776d32ae5	

Table 12: Benchmark HCE model in cell-type annotation (fine-tuning) on the HIHA and Pancreas datasets. Results are reported as mean ± standard deviation.

Dataset	Model	Accuracy	Macro-F1
HIHA	scOntoFM	**0.92 ± 0.00**	**0.77 ± 0.01**
	HCE-TabNet	0.91 ± 0.00	0.71 ± 0.02
	HCE-MLP	0.90 ± 0.00	0.70 ± 0.00
Pancreas	scOntoFM	**0.99 ± 0.00**	**0.97 ± 0.01**
	HCE-TabNet	0.98 ± 0.00	0.90 ± 0.01
	HCE-MLP	0.95 ± 0.00	0.88 ± 0.01

Table 13: Predicted accuracy per novel cell type across SOTA methods. The best result for each cell type is shown in **bold**.

CL ID	cellPLM	Geneformer	Ours	scCello	scGPT	SCimilarity	count	CL ID	cellPLM	Geneformer	Ours	scCello	scGPT	SCimilarity	count
CL:0000030	0.015	0.004	**0.041**	0.02	0.002	0	10000	CL:0009070	0	0.061	0.003	**0.304**	0.001	0	5888
CL:0000047	0	0	**0.024**	0	0	0	10000	CL:0009071	**0.949**	0.748	0.886	0.433	0.836	0.787	3641
CL:0000060	0	0	0.001	0	**0.002**	0	4451	CL:0010012	0.05	0.01	0.001	0	0	**0.082**	10000
CL:0000083	0	0.028	0.004	0.342	0.183	**0.856**	9257	CL:0011031	0.576	0.675	0.938	0.983	0.873	**0.998**	10000
CL:0000121	0	0.001	0	**0.17**	0	0	10000	CL:0019021	0.408	0.089	**0.649**	0.447	0.357	0.167	4533
CL:0000159	**1**	0.157	**1**	0.131	**1**	0.063	5088	CL:1000022	0.498	0.59	0.882	**0.946**	0.498	0.052	3513
CL:0000338	0	0	0	0	0	0	10000	CL:1000042	0	0	0	0	0	0	10000
CL:0000515	0	0	**0.829**	0.577	0	0.296	10000	CL:1000398	0	0	0	0	0	0	7050
CL:0000529	0.154	0	0.404	**0.773**	0	0	10000	CL:1000443	0.996	0.002	**0.997**	0.766	0.996	0.994	10000
CL:0000549	0	0.012	0	0.001	**0.158**	0	5253	CL:1000500	0.008	0	0.004	**0.089**	0.004	0.003	8409
CL:0000550	0	**0.008**	0.001	0	0.005	0	5061	CL:1001567	0.018	0.001	0.012	**0.116**	0.023	0.023	10000
CL:0000552	0.974	0.973	0.975	0.973	0.968	**0.986**	3734	CL:1001569	0	0	0	0	0	0	9976
CL:0000553	0.663	0.848	0.708	0.254	**0.917**	0.899	2564	CL:1001576	0.019	0.022	**0.209**	0.167	0.14	0.005	7386
CL:0000598	0	0	**0.045**	0.039	0.005	0	10000	CL:1001609	0	0	0.046	0.041	0	**0.048**	7010
CL:0000700	0	0	0	0	0	0	6364	CL:2000016	0.008	0.006	0.018	**0.102**	0.013	0.004	3430
CL:0000790	0.005	0.115	**0.438**	0.354	0.157	0.385	10000	CL:2000029	0.031	0	0.084	**0.238**	0.057	0.006	10000
CL:0000867	0.015	0.099	0.071	**0.313**	0.014	0.094	4554	CL:2000054	0.564	0.536	**0.747**	0.576	0.584	0.64	9320
CL:0000912	0.332	0.122	**0.341**	0.208	0.315	0.168	10000	CL:4023009	0	0	0.421	**0.642**	0	0	10000
CL:0000920	0.054	0.03	0.253	**0.402**	0.176	0.21	8377	CL:4023026	0	0	0	0	0	0	5385
CL:0001033	**0.401**	0.011	0.018	0.078	0.147	0.001	10000	CL:4023029	0	0	0	0	0	0	6172
CL:0001034	0	0	0	0	0	0	10000	CL:4023069	0	0	**0.619**	0.404	0	0	6404
CL:0001049	0.082	0.028	**0.661**	0.079	0.054	0.131	5698	CL:4023111	0	0	0	0	0	0	10000
CL:0001055	0.813	0.848	0.842	0.831	0.847	**0.915**	2494	CL:4023161	0	0.001	0	0.009	0	**0.301**	3847
CL:0002032	0	0.075	**0.63**	0.497	0.364	0	6718	CL:4028002	0	0	0	0	0	0	6368
CL:0002046	0.965	0.885	0.96	0.577	0.952	**0.992**	10000	CL:4030006	0.001	0.017	0.079	0.019	**0.116**	0.096	10000
CL:0002178	0.584	0.552	0.606	0.155	**0.868**	0.819	10000	CL:4030034	**0.997**	0.983	0.986	0.962	0.978	0.972	4462
CL:0002211	0.196	0.195	0.302	0.176	0.176	**0.742**	10000	CL:4030055	0.422	0.323	0.33	**0.846**	0.8	0.364	10000
CL:0002248	0	0	0	0	0	0	2942	CL:4030056	0	0	0	0	0	0	3945
CL:0002259	0	0	0	0	0	0	4308	CL:4030059	**0.87**	0.107	0.118	0.125	0.837	0.005	10000
CL:0002303	0	0	0	0	0	0	10000	CL:4030063	0	0	0	0	0	0	3978
CL:0002304	0	0	0	0	0	0	5330	CL:4033008	0.841	**0.969**	0.964	0.484	0.827	0.912	8339
CL:0002325	0.739	0.07	**0.997**	0.88	0.671	0.874	10000	CL:4033014	0	0	**0.006**	**0.006**	0	0	9787
CL:0002334	0.003	0.004	**0.011**	0.002	0.005	0.01	10000	CL:4033019	0.906	0.017	0.872	0.115	0.086	**0.953**	2830
CL:0002362	**0.198**	0.065	0.058	0.04	0.151	0.002	2714	CL:4033027	0.081	0.003	0.173	0.569	**0.57**	0.004	10000
CL:0002547	**0.003**	0.002	0	0	0	0	4757	CL:4033028	0	0	0	0	0	0	10000
CL:0002549	0.005	0	0.032	0.005	0.001	**0.133**	2058	CL:4033029	0	0	0	0	0	0	10000
CL:0002551	0.047	0.558	0.622	0.627	0.082	**0.793**	10000	CL:4033030	0	0	0	0	0	0	10000
CL:0002615	0.996	0.476	0.991	**0.999**	**0.999**	0.931	10000	CL:4033031	0	0	0	0	0	0	10000
CL:0002627	0.578	0.809	0.893	0.115	**0.904**	0.418	10000	CL:4033032	0	0	0	0	0	0	10000
CL:0002629	0.028	0.258	0.472	0.794	**0.922**	0.895	10000	CL:4033033	0	0	0	0	0	0	10000
CL:0002671	**0.067**	0.008	0.01	0.034	0.019	0.034	4225	CL:4033034	0	0	0	0	0	0	10000
CL:0002672	0.002	0	**0.016**	0.009	0	0.005	5149	CL:4033035	0	0	0	0	0	0	10000
CL:0003050	**1**	0.985	**1**	**1**	**1**	**1**	7710	CL:4033036	0	0	0	0	0	0	10000
CL:0004217	**0.999**	0.675	0.987	0.995	**0.999**	**0.999**	10000	CL:4033039	**0.656**	0.274	0.287	0.212	0.176	0.188	4520
CL:0004218	0	0	0	0	0	0	10000	CL:4033046	0.064	0	0.633	**0.841**	0.722	0.002	10000
CL:0004232	0	0	0.497	0.547	0.13	**0.957**	10000	CL:4033047	0	0	0	0	0	0	10000
CL:0005022	0	0.003	**0.019**	**0.019**	**0.019**	0.001	8796	CL:4033048	0	0.001	0.001	0.009	0.002	**0.02**	4318
CL:0005026	0.272	0.478	0.516	**0.903**	0.532	0.387	10000	CL:4033051	0	0	0	0	0	0	10000
CL:0008031	0.165	0.088	0.168	0.059	**0.436**	0.289	10000	CL:4033052	0	0	0	0	0	0	6102
CL:0008035	0.816	0.04	**0.898**	0.727	0.476	0.866	10000								

	cellPLM	Geneformer	Ours	scCello	scGPT	SCimilarity	total count
Macro accuracy	0.213	0.151	**0.286**	0.254	0.244	0.240	774,215
Average rank	3.712	4.202	**2.848**	3.237	3.485	3.515	774,215

Table 14: Summary of the architecture and training configuration for static gene embedding GO function classifier.

Category	**Specification**
Input dimension	d (embedding dimension)
Hidden layer	128 units, ReLU activation
Bottleneck layer	64 units, ReLU activation
Output layer	3 units (multi-label classification)
Dropout rate	0.2 (applied after each hidden layer)
Loss function	Binary Cross-Entropy with Logits
Optimization	Adam optimizer
Learning rate	1×10^{-4}
Batch size	64
Training epochs	15
Task formulation	Multi-label classification ($C = 3$)

Table 15: Summary of spatial transcriptomics (ST) and single-cell (SC) datasets.

Dataset	ST (spots $\times$ genes)	SC (cells $\times$ genes)	Shared genes for imputation
Lung cancer	$31,520 \times 423$	$207,853 \times 20,793$	422
Breast cancer	$4,784 \times 1,000$	$9,234 \times 19,098$	891
SEA-AD	$18,085 \times 180$	$15,876 \times 20,080$	132

Table 16: Performance comparison of zero-shot clustering scIB overall score across tissues/datasets. The best results are shown in **bold** for each dataset.

Dataset	Ours	Geneformer	scGPT	scCello	CellPLM	SCimilarity
HIHA	0.7533	0.4444	0.6002	0.5668	0.7037	**0.7730**
Heart	0.8324	0.6862	0.8326	**0.8666**	0.7847	0.8299
Pancreas	**0.8140**	0.6113	0.7967	0.7203	0.7024	0.8023
Lung	0.8467	0.8395	0.8397	0.8299	0.8379	**0.8566**
Liver	0.7730	0.6853	0.7886	0.7720	0.7102	**0.7921**
Breast TNBC	**0.8724**	0.6824	0.8418	0.8423	0.8377	0.8601
Kidney	0.7185	0.5785	0.7020	**0.7601**	0.7211	0.7149
Adipose Tissue	**0.8696**	0.6509	0.8181	0.8589	0.8064	0.8290
Bone Marrow	0.6823	0.6498	0.6924	0.6659	0.6772	**0.7059**
Mean	0.7958	0.6476	0.7680	0.7647	0.7535	**0.7960**

Table 17: Zero-shot clustering benchmarking results of detailed scIB metrics for each of the nine evaluation datasets. The best performance in each metric for each dataset is highlighted in **bold**.

Dataset	Method	NMI	ARI	ASW_{label}	GraphConn	$\text{ASW}_{label/batch}$	AvgBio	AvgBatch	Overall
HIHA	Ours	0.7597	0.6422	**0.6021**	0.8868	0.8756	0.6680	0.8812	0.7533
	Geneformer	0.2500	0.1355	0.4560	0.5117	0.8687	0.2805	0.6902	0.4444
	scGPT	0.5199	0.2619	0.4977	0.8218	0.8995	0.4265	0.8607	0.6002
	scCello	0.4704	0.3245	0.4824	0.6418	**0.9149**	0.4257	0.7784	0.5668
	cellPLM	0.7101	0.4755	0.5365	**0.9208**	0.8755	0.5740	**0.8982**	0.7037
	SCimilarity	**0.7675**	**0.7375**	0.5833	0.9167	0.8602	**0.6961**	0.8884	**0.7730**
Heart	Ours	0.8581	0.8204	**0.6760**	0.9297	0.8780	0.7848	0.9039	0.8324
	Geneformer	0.6464	0.4900	0.5228	0.8878	0.8838	0.5531	0.8858	0.6862
	scGPT	0.8641	0.8905	0.5911	0.9371	0.8803	0.7819	0.9087	0.8326
	scCello	**0.9036**	**0.9383**	0.6546	**0.9456**	**0.8907**	**0.8321**	**0.9182**	**0.8666**
	cellPLM	0.8168	0.7228	0.5891	0.9324	0.8627	0.7096	0.8975	0.7847
	SCimilarity	0.8572	0.8182	0.6715	0.9402	0.8626	0.7823	0.9014	0.8299
Pancreas	Ours	**0.8598**	0.7951	0.6092	**0.9799**	0.8259	**0.7547**	0.9029	**0.8140**
	Geneformer	0.5045	0.3187	0.5082	0.8726	0.8525	0.4438	0.8625	0.6113
	scGPT	0.8315	**0.8523**	0.5259	0.9199	0.8539	0.7365	0.8869	0.7967
	scCello	0.6442	0.5420	0.5782	0.9616	**0.8753**	0.5881	**0.9185**	0.7203
	cellPLM	0.6277	0.5578	0.5515	0.9247	0.8504	0.5790	0.8875	0.7024
	SCimilarity	0.8497	0.7252	**0.6568**	0.9522	0.8276	0.7439	0.8899	0.8023
Lung	Ours	0.8722	0.8696	0.6469	0.9797	0.8653	0.7962	0.9225	0.8467
	Geneformer	0.8797	0.8512	0.5774	0.9775	**0.9115**	0.7694	0.9445	0.8395
	scGPT	0.8813	0.8519	0.5736	**0.9955**	0.8965	0.7689	**0.9460**	0.8397
	scCello	0.8664	0.8043	0.6211	0.9781	0.8794	0.7640	0.9287	0.8299
	cellPLM	0.8823	0.8477	0.5786	0.9942	0.8867	0.7696	0.9404	0.8379
	SCimilarity	**0.8898**	**0.8733**	**0.6820**	0.9885	0.8495	**0.8150**	0.9190	**0.8566**
Liver	Ours	0.7871	0.6935	0.5824	0.9600	0.8422	0.6877	0.9011	0.7730
	Geneformer	0.6502	0.4678	0.5230	0.9366	0.8486	0.5470	0.8926	0.6853
	scGPT	**0.7933**	**0.7549**	0.5761	0.9625	0.8564	0.7081	0.9095	0.7886
	scCello	0.7674	0.6763	0.5746	0.9659	**0.8756**	0.6728	**0.9207**	0.7720
	cellPLM	0.6738	0.4850	0.5639	**0.9742**	0.8542	0.5742	0.9142	0.7102
	SCimilarity	0.7920	0.7154	**0.6461**	0.9730	0.8340	**0.7179**	0.9035	**0.7921**
Breast TNBC	Ours	**0.8807**	**0.9090**	0.7245	0.9954	0.8527	**0.8380**	0.9241	**0.8724**
	Geneformer	0.6080	0.3939	0.5279	0.9905	0.8917	0.5099	0.9411	0.6824
	scGPT	0.8543	0.8528	0.6070	0.9941	**0.9008**	0.7714	**0.9475**	0.8418
	scCello	0.8316	0.8014	0.7106	0.9895	0.8785	0.7812	0.9340	0.8423
	cellPLM	0.8215	0.8444	0.6438	0.9929	0.8858	0.7699	0.9393	0.8377
	SCimilarity	0.8596	0.8616	**0.7267**	**0.9963**	0.8565	0.8160	0.9264	0.8601
Kidney	Ours	0.7174	0.5711	0.5871	0.9542	0.7627	0.6252	0.8585	0.7185
	Geneformer	0.4874	0.2008	0.4878	0.9392	0.7772	0.3920	0.8582	0.5785
	scGPT	0.7111	0.5489	0.5331	0.9574	0.7597	0.5977	0.8586	0.7020
	scCello	0.7452	**0.6851**	0.6009	0.9641	**0.8050**	**0.6771**	**0.8846**	**0.7601**
	cellPLM	0.7201	0.6085	0.5357	**0.9771**	0.7643	0.6215	0.8707	0.7211
	SCimilarity	**0.7566**	0.5347	**0.6116**	0.9646	0.7072	0.6343	0.8359	0.7149
Adipose Tissue	Ours	**0.9098**	**0.9307**	0.6411	0.9626	0.9037	**0.8272**	0.9331	**0.8696**
	Geneformer	0.6091	0.3070	0.5012	0.9390	0.8984	0.4724	0.9187	0.6509
	scGPT	0.8618	0.7557	0.5944	0.9710	0.9074	0.7373	0.9392	0.8181
	scCello	0.8786	0.8991	0.6337	0.9580	**0.9250**	0.8038	**0.9415**	0.8589
	cellPLM	0.8151	0.7359	0.6120	**0.9775**	0.8915	0.7210	0.9345	0.8064
	SCimilarity	0.8639	0.7567	**0.6868**	0.9685	0.8692	0.7691	0.9188	0.8290
Bone Marrow	Ours	0.6809	0.3919	0.5135	0.9650	0.8601	0.5288	0.9126	0.6823
	Geneformer	0.6025	0.3237	0.4939	0.9364	0.8925	0.4734	0.9144	0.6498
	scGPT	0.6654	0.4002	0.5267	0.9693	**0.9005**	0.5307	**0.9349**	0.6924
	scCello	0.6256	0.3867	0.5075	0.9484	0.8614	0.5066	0.9049	0.6659
	cellPLM	0.6245	0.3884	0.5260	0.9653	0.8816	0.5130	0.9235	0.6772
	SCimilarity	**0.7160**	**0.4321**	**0.5406**	**0.9707**	0.8703	**0.5629**	0.9205	**0.7059**

Table 18: Dataset-wise comparison of zero-shot clustering scGraph-LCA metric under Spearman, Kendall's τ, and Pearson correlation. Best result per dataset and per metric is shown in **bold**.

	Spearman						Kendall's τ						Pearson					
Dataset	Ours	Geneformer	scGPT	scCello	CellPLM	SCimilarity	Ours	Geneformer	scGPT	scCello	CellPLM	SCimilarity	Ours	Geneformer	scGPT	scCello	CellPLM	SCimilarity
HIHA	**0.7157**	0.3890	0.4174	0.4724	0.3772	0.6201	**0.5688**	0.2849	0.3171	0.3590	0.2895	0.4921	**0.7669**	0.4527	0.5300	0.5946	0.5215	0.7294
Heart	**0.7074**	0.4220	0.6605	0.4163	0.5100	0.5090	**0.5783**	0.3477	0.5275	0.3277	0.3929	0.4117	0.7787	0.6073	**0.7914**	0.6462	0.6969	0.6909
Pancreas	0.3992	0.4040	0.3978	0.3313	0.5470	**0.6646**	0.2960	0.3213	0.3135	0.2390	0.4454	**0.5699**	0.7398	0.7139	0.7660	0.6478	0.7780	**0.8412**
Lung	**0.6777**	0.6719	0.4047	0.4655	0.4624	0.5958	0.5615	**0.5621**	0.3270	0.3684	0.3745	0.4890	**0.8381**	0.8127	0.7396	0.7519	0.7474	0.7886
Liver	**0.5704**	0.4905	0.4824	0.4925	0.5023	0.4842	**0.4691**	0.3857	0.3799	0.3923	0.4008	0.3781	**0.7411**	0.6457	0.6423	0.6718	0.6716	0.6501
Breast TNBC	0.5112	0.4000	0.4328	**0.5969**	0.5487	0.4226	0.3956	0.3261	0.3470	**0.5085**	0.4390	0.3461	0.8699	0.8298	0.8774	**0.8808**	0.8693	0.8425
Kidney	**0.6378**	0.4423	0.2760	0.3554	0.3912	0.5204	**0.5083**	0.3368	0.1921	0.2573	0.2878	0.3910	**0.6857**	0.5748	0.4880	0.5579	0.5382	0.6273
Adipose Tissue	0.6180	0.4910	0.4144	0.5398	0.4644	**0.6355**	0.4762	0.3874	0.3267	0.4051	0.3606	**0.5030**	0.7830	0.6737	0.6450	0.7102	0.6617	**0.7834**
Bone Marrow	**0.6630**	0.3306	0.2431	0.5207	0.2891	0.5980	**0.5370**	0.2470	0.1794	0.3984	0.2202	0.4749	**0.7093**	0.4151	0.4120	0.6115	0.4437	0.6849
Mean	**0.6112**	0.4490	0.4143	0.4656	0.4547	0.5611	**0.4879**	0.3554	0.3233	0.3618	0.3568	0.4506	**0.7681**	0.6362	0.6546	0.6747	0.6587	0.7376

Table 19: Performance comparison across datasets of scGraph-RWR. Best result per dataset and best mean performance are shown in **bold**.

	Spearman					Kendall's τ					Pearson				
Dataset	Ours	Geneformer	scGPT	scCello	CellPLM	Ours	Geneformer	scGPT	scCello	CellPLM	Ours	Geneformer	scGPT	scCello	CellPLM
HIHA	0.4939	0.4584	0.5000	**0.5253**	0.4951	0.3668	0.3292	0.3677	**0.3802**	0.3618	**0.7998**	0.6188	0.6230	0.6855	0.7118
Heart	0.3438	0.4141	0.3190	**0.4372**	0.3950	0.2264	**0.3223**	0.2132	**0.3223**	0.3058	0.8926	0.8406	0.8857	**0.9146**	0.9022
Pancreas	0.5373	0.5693	0.3700	**0.6562**	0.5029	0.4116	0.4697	0.2980	**0.5202**	0.4040	0.8831	0.8923	0.8612	**0.9007**	0.8579
Lung	**0.5370**	0.3794	0.3297	0.2933	0.2970	**0.4489**	0.2756	0.2311	0.2400	0.2133	0.9285	0.8861	0.9153	**0.9513**	0.9037
Liver	**0.4974**	0.2292	0.2393	0.3665	0.3649	**0.3771**	0.1677	0.1594	0.2552	0.2573	0.8525	0.7604	0.8307	**0.8898**	0.8258
Breast TNBC	**0.6200**	0.6000	0.5400	0.4800	0.3800	0.4800	**0.5200**	0.4400	0.4400	0.2800	0.9601	0.9702	**0.9793**	0.9639	0.9464
Kidney	0.4045	**0.5174**	0.4825	0.4294	0.4472	0.2823	**0.3801**	0.3489	0.3120	0.3202	0.7419	0.7105	0.7917	**0.8268**	0.7128
Adipose Tissue	**0.5853**	0.4151	0.4574	0.4410	0.3744	**0.4458**	0.3073	0.3375	0.3250	0.2646	0.8354	0.7821	0.8654	**0.8688**	0.8113
Bone Marrow	**0.5563**	0.3358	0.2973	0.4434	0.3810	**0.4040**	0.2387	0.2167	0.3151	0.2820	**0.7476**	0.5311	0.6396	0.7473	0.6543
Mean	**0.5084**	0.4354	0.3928	0.4525	0.4042	**0.3825**	0.3345	0.2903	0.3456	0.2988	0.8491	0.7769	0.8213	**0.8610**	0.8140

Table 20: Overall performance of zero-shot clustering of 340 in-distribution cell types in unseen datasets across SOTA methods using scIB metrics, scGraph-LCA, and scGraph-RWR. Best result per column is shown in **bold**.

	scIB Metrics				scGraph-LCA			scGraph-RWR		
Method	NMI	ARI	ASW	Avg Bio	Spearman	Kendall's τ	Pearson	Spearman	Kendall's τ	Pearson
Ours	0.7804	**0.3431**	0.5032	0.5422	**0.5114**	**0.3887**	**0.5611**	0.0937	0.0569	0.4427
Geneformer	0.7213	0.2241	0.4717	0.4724	0.2962	0.2157	0.3278	0.1369	0.0916	0.3605
scGPT	0.7740	0.3014	0.5074	0.5276	0.1339	0.0962	0.2058	0.0682	0.0447	0.3638
CellPLM	**0.7915**	0.3334	**0.5243**	**0.5498**	0.0980	0.0687	0.1933	**0.1436**	**0.0962**	0.3980
scCello	0.7693	0.3092	0.5199	0.5328	0.2417	0.1751	0.3384	0.0670	0.0427	**0.4768**

Table 21: GO namespace classification by aggregated gene embeddings. **bold** indicates the best performance and underlining denotes the second-best results

Model	**F1 Micro**	**F1 Macro**	**AP Micro**	**AP Macro**
Ours	<u>0.783</u> $\pm$ 0.01	<u>0.632</u> $\pm$ 0.02	<u>0.832</u> $\pm$ 0.01	<u>0.632</u> $\pm$ 0.02
Geneformer	0.767 $\pm$ 0.01	0.593 $\pm$ 0.02	<u>0.832</u> $\pm$ 0.01	0.613 $\pm$ 0.02
scCello	0.781 $\pm$ 0.01	0.627 $\pm$ 0.02	0.823 $\pm$ 0.01	0.620 $\pm$ 0.02
scGPT	**0.809** $\pm$ 0.01	**0.683** $\pm$ 0.02	**0.861** $\pm$ 0.01	**0.678** $\pm$ 0.02

Table 22: Spatial gene imputation performance (mean $\pm$ std) of Pearson, Spearman, and Kendall's τ correlations in SEA-AD dataset. Higher is better ($\uparrow$). Best results are shown in **bold**.

Method	Pearson $\uparrow$	Spearman $\uparrow$	Kendall's τ $\uparrow$
Ours	0.3820 ± 0.1450	0.3429 ± 0.1424	0.2581 ± 0.1025
Geneformer	0.3769 ± 0.1580	0.3388 ± 0.1627	0.2552 ± 0.1175
CellPLM	$\mathbf{0.5047 \pm 0.1705}$	$\mathbf{0.4251 \pm 0.1765}$	$\mathbf{0.3247 \pm 0.1302}$
scCello	0.3286 ± 0.1543	0.2901 ± 0.1429	0.2167 ± 0.1026
scGPT	0.3056 ± 0.1470	0.2435 ± 0.1361	0.1823 ± 0.0980

Table 23: Ablation study of model loss components.

Model	MLM	DynamicMargin	OntoAlign	scIB $\uparrow$	scGraph-LCA $\uparrow$
MLM only	✓	✗	✗	0.7597	0.4237
w/o OntoAlign	✓	✓	✗	**0.8066**	0.3607
w/o DynamicMargin	✓	✗	✓	0.7841	**0.7211**
scOntoFM	✓	✓	✓	0.7958	0.6112

Table 24: Ablation study of the model's loss terms across the nine evaluation datasets of the overall scIB score and scGraph-LCA metric.

	scIB				**scGraph-LCA**			
Dataset	scOntoFM	w/o OntoAlign	w/o DynamicMargin	MLM only	scOntoFM	w/o OntoAlign	w/o DynamicMargin	MLM only
HIHA	0.7533	0.7745	**0.7848**	0.7403	0.7157	0.4282	**0.8301**	0.4787
Heart	0.8324	**0.8549**	0.8480	0.7908	**0.7074**	0.4171	0.6824	0.5511
Pancreas	0.8140	**0.8388**	0.7767	0.7191	0.3992	0.3957	**0.4408**	0.3351
Lung	**0.8467**	0.8332	0.8119	0.8091	0.6777	0.2689	**0.8622**	0.4425
Liver	0.7730	0.7773	**0.7837**	0.7553	0.5704	0.3691	**0.6911**	0.4847
Breast_TNBC	0.8724	**0.8759**	0.8272	0.8530	0.5112	0.2792	**0.7478**	0.2966
Kidney	0.7185	**0.7252**	0.7037	0.6629	0.6378	0.3333	**0.7333**	0.3638
Adipose_tissue	**0.8696**	0.8643	0.8185	0.8154	0.6180	0.3389	**0.8345**	0.3986
Bone_marrow	0.6823	**0.7153**	0.7021	0.6915	0.6630	0.4155	**0.6673**	0.4624
Mean	0.7958	**0.8066**	0.7841	0.7597	0.6112	0.3607	**0.7211**	0.4237

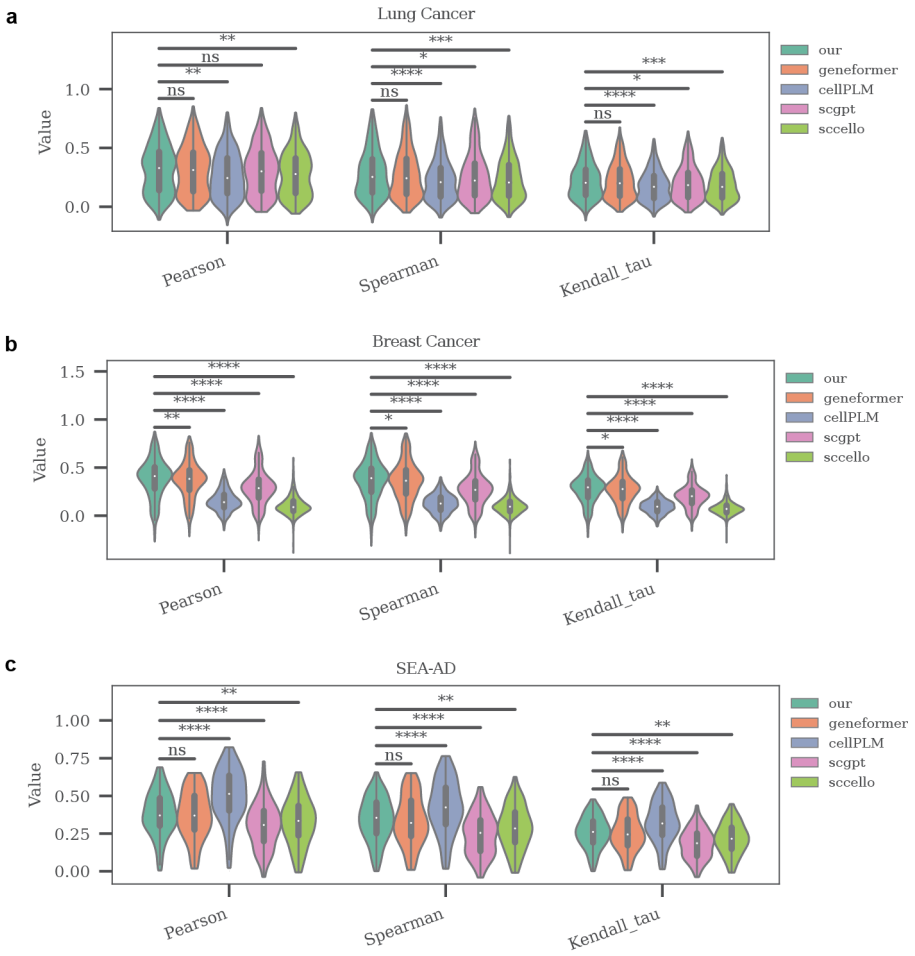

Figure 16: (a) Violin plot of the distribution for Pearson, Spearman, and Kendall's τ correlation across all the imputed genes in MERSCOPE Lung cancer 2 dataset. (b) Violin plot of the distribution for Pearson, Spearman, and Kendall's τ correlation across all the imputed genes in breast cancer dataset. (c) Violin plot of the distribution for Pearson, Spearman, and Kendall's τ correlation across all the imputed genes in SEA-AD dataset. Statistical significance is indicated by stars (*), with "ns" denoting no significant difference.

Table 25: Agreement scores between reference and target annotations across datasets.

Dataset	Reference	Target	# Ref Labels	# Target Labels	Agreement Score
HIHA	cell_type	AIFI_L1	38	9	0.259
HIHA	cell_type	AIFI_L2	38	29	0.970
HIHA	cell_type	AIFI_L3	38	71	0.830
Kidney	cell_type	SubclassLevel1	33	17	0.933
Kidney	cell_type	SubclassLevel2	33	97	0.407

Table 26: Impact of noise proportion on performance and hierarchical error distribution.

Noise	Accuracy (mean $\pm$ std)	Macro F1 (mean $\pm$ std)	Parent (mean $\pm$ std)	Child (mean $\pm$ std)	Sibling (mean $\pm$ std)
0.1	0.941 $\pm$ 0.001	0.806 $\pm$ 0.017	0.001 $\pm$ 0.000	0.003 $\pm$ 0.001	0.014 $\pm$ 0.000
0.2	0.940 $\pm$ 0.001	0.752 $\pm$ 0.013	0.001 $\pm$ 0.000	0.003 $\pm$ 0.001	0.015 $\pm$ 0.000
0.3	0.934 $\pm$ 0.002	0.691 $\pm$ 0.008	0.001 $\pm$ 0.000	0.005 $\pm$ 0.001	0.015 $\pm$ 0.001
0.4	0.834 $\pm$ 0.036	0.572 $\pm$ 0.024	0.001 $\pm$ 0.001	0.053 $\pm$ 0.021	0.018 $\pm$ 0.010
0.5	0.150 $\pm$ 0.018	0.207 $\pm$ 0.025	0.000 $\pm$ 0.000	0.552 $\pm$ 0.002	0.035 $\pm$ 0.003

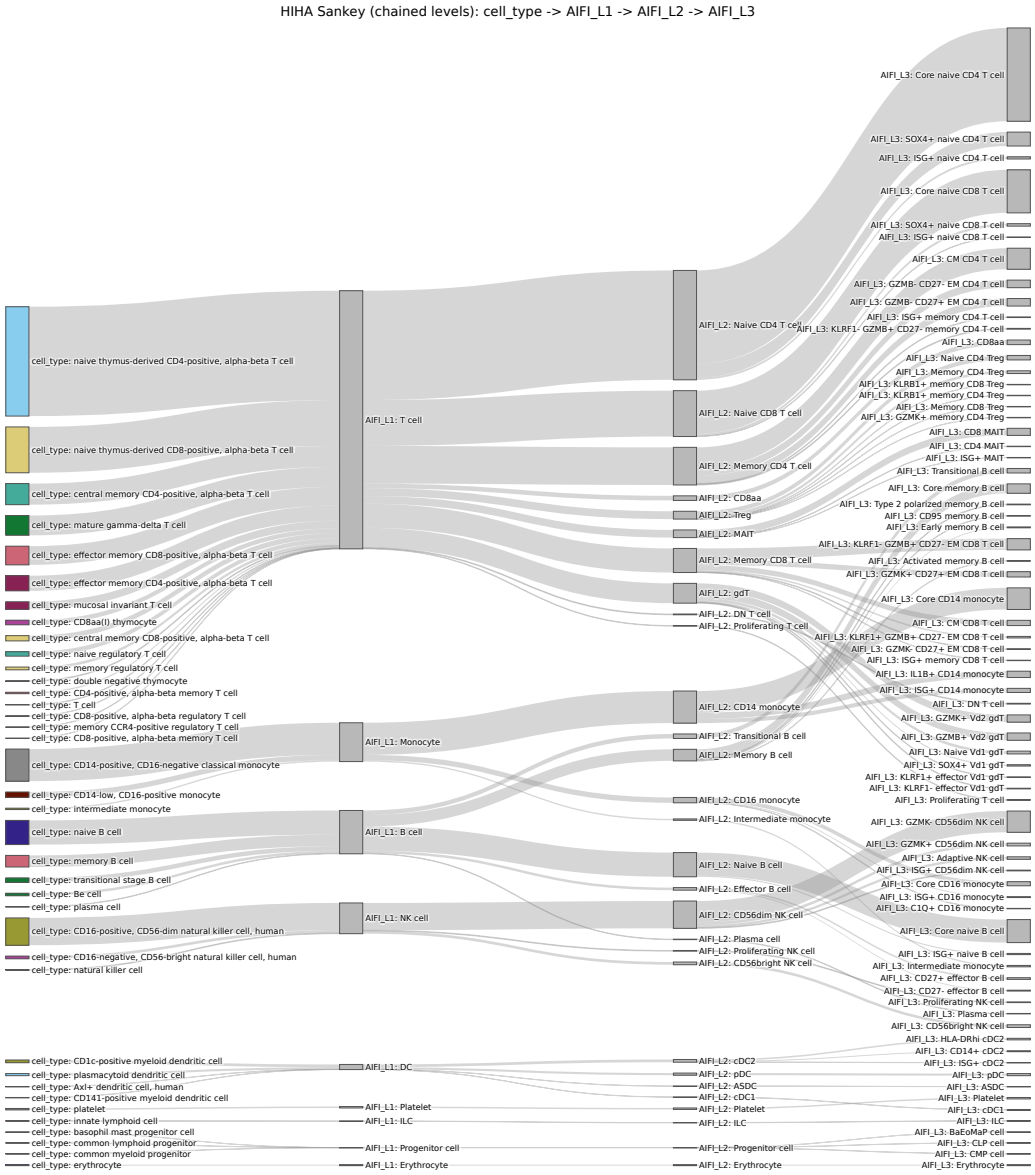

Figure 17: From left to right: Sankey plot of cell type annotation flow from cell_type (CZI provided ontology labels) to author annotated labels (from coarse to fine-grained) in HIHA dataset

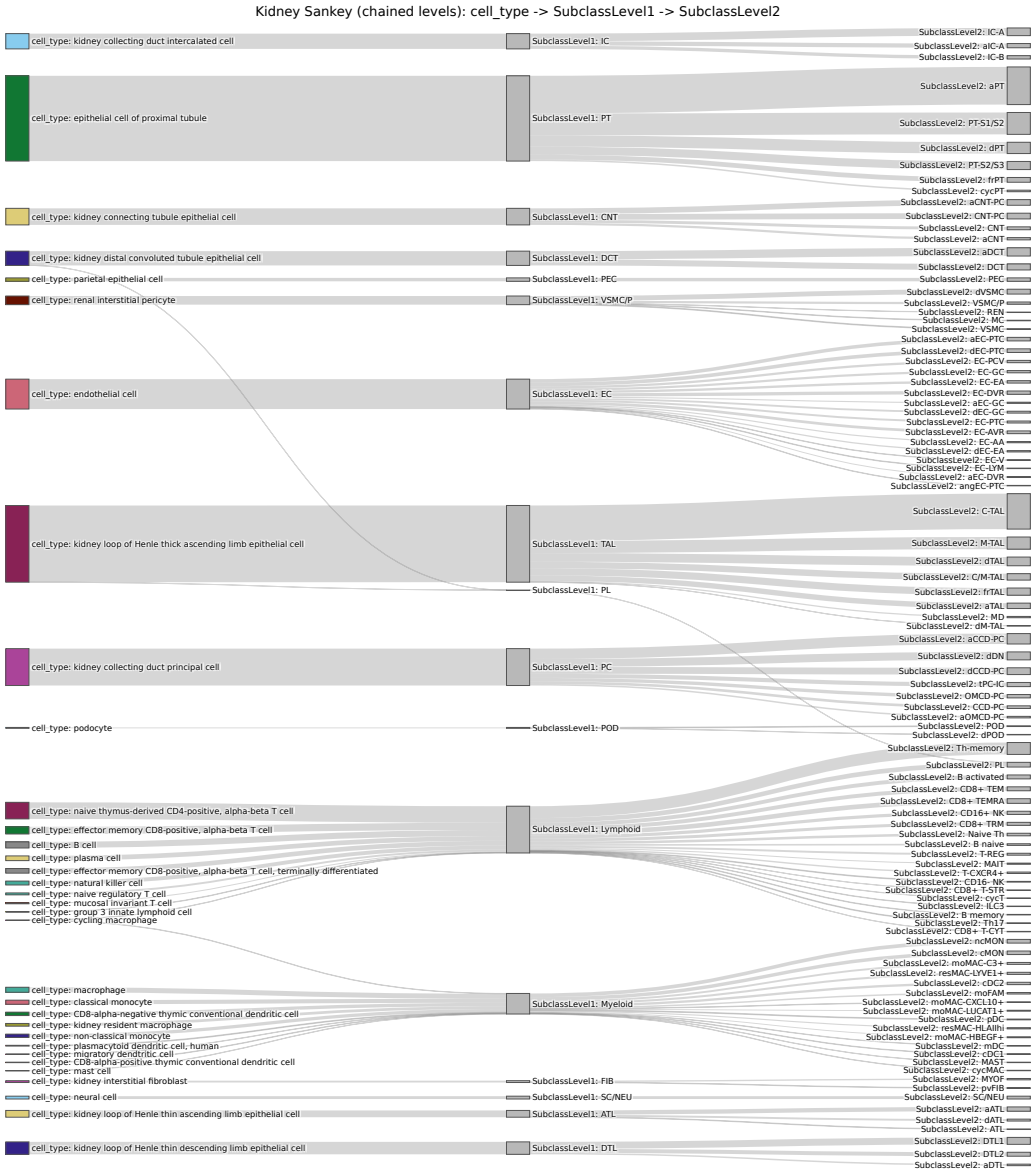

Figure 18: From left to right: Sankey plot of cell type annotation flow from cell_type (CZI provided ontology labels) to author annotated labels (from coarse to fine-grained) in Kidney dataset

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, we propose scOntoFM, a novel foundation model that integrates hierarchical cell ontologies directly into the pre-training framework. This main claim is accurately and clearly stated and emphasized in the abstract and introduction section. It reflects our core contribution and scopes.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes an explicit *Limitations* paragraph in Section 5, discussing reliance on explicit cell-type labels and scalability constraints from the indexing map, and outlining future directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present theorem- or proposition-style theoretical claims requiring formal proofs. Instead, it primarily focuses on methodological development and empirical evaluation in the context of computational biology and single-cell foundation models.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all necessary details to reproduce our work. The method is fully specified in Section 2 (including objectives), and experimental setup details are provided in Section 3 plus Appendix B/C (model/training configuration, dataset lists/sources, task protocols, and additional quantitative results).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our code and datasets will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the Experimental setting/details about pre-training and downstream tasks are illustrated in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports uncertainty for key experiments using mean $\pm$ standard deviation (e.g., Table 4 with 20-fold CV for gene function prediction, Table 5 for spatial imputation), with additional distributional violin plots in Figure 15 for hallmark gene set enrichment score, and error bars in Figure 4 for novel cell identification, as well the statistical significance test in Figure 16 for spatial gene imputation performance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the major pre-training details of scOntoFM (six NVIDIA H200 GPUs, effective batch size, and ∼30-hour pretraining). We also report the computational resources required when constructing the Cell-Dataset Index Map, which takes about 5 minutes and uses 6.5 GB of peak memory on an Intel Xeon Gold 6252 CPU. The final file size is 132 MB.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics throughout the entire project.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are addressed throughout the introduction and discussion. And we also provide a detailed impact statement in Appendix D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work focuses on representation learning for single-cell transcriptomic analysis and does not involve high-risk generative modeling or biological sequence design. The released code and model checkpoints are intended solely for research use in computational biology. All datasets used in this study are publicly available or distributed under their respective data usage agreements, and no identifiable personal information is included. As such, we do not believe the model poses substantial misuse risks requiring specialized safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in this work, including datasets, software libraries, and pre-existing models, are appropriately cited and credited in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: Code and datasets will be released upon acceptance. Clear documentations will be provided along the assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We did not perform crowdsourcing experiments and research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models (LLMs) are not a component of the proposed methodology, experimental pipeline, or scientific contributions of this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.