OmicsDefense: The First Unified Framework for Defending Against Backdoor Attacks in Single-cell Foundation Models

Anonymous Authors[1] -

Abstract

Single-cell foundation models have become indispensable tools for cellular heterogeneity analysis, disease mechanism discovery, and clinical diagnostics, yet they face critical biosecurity vulnerabilities from backdoor attacks that remain unaddressed. We present OmicsDefense, the first unified defense framework for single-cell foundation models, featuring Weighted Divergence Filtering (WDF) for poisoned sample detection. We apply our framework to seven foundation models (Geneformer, scGPT, scCELLO, scPRINT, LangCell, UCE, scFoundation) on two datasets (Human Pancreas and Myeloid), achieving an average 98.7% reduction in attack success rates while preserving clean accuracy. OmicsDefense establishes a new paradigm for trustworthy single-cell analysis, enabling reliable cell atlas construction and clinical diagnosis in the era of foundation model-powered biomedicine.

1. Introduction

The advent of single-cell sequencing technology has revolutionized our understanding of cellular heterogeneity in health and disease (Wang et al., 2020). Recent technological advances have enabled the generation of massive single-cell transcriptomic datasets, culminating in comprehensive cell atlases spanning multiple tissues, organs, and species (Regev et al., 2017). These large-scale datasets have catalyzed the development of foundation models: large pre-trained neural networks capable of learning universal representations of cellular states. Models such as scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023), and scBERT (Yang et al., 2022) have demonstrated remarkable success across diverse downstream tasks including cell type annotation, perturbation prediction, and gene regulatory network inference.

However, this reliance on external data and pre-trained models introduces significant security vulnerabilities. The scBackdoor study (Feng et al., 2024) demonstrated that single-cell foundation models are susceptible to backdoor attacks, where malicious actors inject hidden triggers during training that cause targeted misclassifications during inference.

Despite this demonstrated vulnerability, defense mechanisms for omics foundation models remain entirely unexplored. While general defense methods exist for image and text models (Tran et al., 2018; Lyu et al., 2023), the unique characteristics of omics data, including extreme sparsity matrix (>90% zeros), cellular heterogeneity, and biologically meaningful representations in transformer embeddings, require domain-specific defense strategies. Critically, no prior work has examined defense in single-cell models or its biological implications: whether security measures preserve the biological validity essential for downstream research and clinical applications.

Here we present OmicsDefense, the first unified framework for defending against backdoor attacks in single-cell foundation models. We make three key contributions:

First, we propose a two-stage defense pipeline: Weighted Divergence Filtering (WDF) detects poisoned samples by comparing multi-layer CLS token divergences between fine-tuned and pre-trained models, followed by label recovery using K-nearest neighbor voting. We predict labels for suspicious samples based on their nearest clean neighbors, ensuring biological consistency across similar cells.

Second, we validate OmicsDefense across seven foundation models: Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), scCELLO (Yuan and Zhan, 2024), scPRINT (Cantini et al., 2025), LangCell (Zhao et al., 2024), UCE (Rosen et al., 2023), and scFoundation (Hao et al., 2024), on two datasets with comprehensive ablation studies. Our method achieves approximately 90% reduction in attack success rates while maintaining clean accuracy.

Third, we provide the first systematic analysis of defense preservation of biological validity. We demonstrate that defense restores correct cell type labels: cells mislabeled by backdoor attacks, such as pancreatic β cells incorrectly annotated as type A, are correctly recovered. This ensures reliable clinical diagnosis and prevents downstream errors in biological research.

2. Related Work

2.1. Backdoor Defense in General Machine Learning

Defense methods for backdoor attacks have been extensively studied in computer vision and natural language processing. Approaches include spectral signatures (Tran et al., 2018) that detect anomalous activation patterns and neuron pruning (Lyu et al., 2023) to remove backdoor-triggered pathways. These methods assume dense input features (images) or semantic text embeddings, differing fundamentally from sparse gene expression matrices (>90% zeros) and biologically structured transformer representations in omics foundation models. No prior work has validated defense methods on sparse, heterogeneous biological data or examined their preservation of biological validity.

2.2. Backdoor Attacks in Single-cell Foundation Models

The security of omics foundation models has received limited attention. The scBackdoor framework (Feng et al., 2024) demonstrated that models like Geneformer (Theodoris et al., 2023) and scGPT (Cui et al., 2024) are vulnerable to backdoor attacks through trigger injection in gene token sequences. With 5% poisoning, attack success rates reached 97% on cell type annotation tasks, while clean accuracy remained unaffected—highlighting the stealthy nature of these threats. However, no defense mechanisms have been proposed for omics foundation models, leaving critical security gaps in clinical diagnostics, drug discovery, and biological research. The unique structure of gene expression data (rank-based tokenization, extreme sparsity, and biological meaning encoded in token sequences) necessitates domain-specific defense approaches.

2.3. Single-cell Foundation Models

Several foundation models have been developed for single-cell transcriptomics, each employing different architectures and pre-training strategies. Geneformer (Theodoris et al., 2023) uses a rank-based gene tokenization approach with BERT-style transformer architecture, pre-trained on 30 million cells to learn contextual gene representations. scGPT (Cui et al., 2024) extends this paradigm with a generative pre-trained transformer optimized for cell type annotation and perturbation prediction through self-supervised learning on large-scale atlases. scCELLO (Yuan and Zhan, 2024) employs a hierarchical architecture that captures cell-type-specific gene expression patterns through multi-scale attention mechanisms. scPRINT (Cantini et al., 2025) focuses on interpretable gene program discovery by integrating attention-based feature attribution with transformer embeddings. LangCell (Zhao et al., 2024) adapts language model architectures to single-cell data with dedicated CLS token representations, enabling robust transfer learning across diverse cell atlases. UCE (Rosen et al., 2023) proposes universal cell embeddings through contrastive learning across multiple tissues and species, creating cross-atlas compatible representations. scFoundation (Hao et al., 2024) scales transformer architectures to billion-parameter models with mean pooling strategies, achieving state-of-the-art performance on comprehensive cell type annotation benchmarks.

3. Method

3.1. Task Description

We consider the cell type annotation task, a fundamental application of single-cell foundation models. Given a gene expression profile $X \in \mathbb{R}^G$ (where G is the number of genes), the model predicts the cell type label $y \in \{1, 2, \ldots, C\}$ among C predefined cell types. This task is critical for biomedical research, clinical diagnostics, and drug discovery (Stuart et al., 2019), where accurate cell type identification enables downstream analyses of cellular heterogeneity, disease mechanisms, and therapeutic responses.

For omics foundation models like Geneformer, gene expression values are first tokenized into discrete token sequences through rank-based encoding. Each cell is represented as a sequence of gene tokens $\mathbf{t} = [t_1, t_2, \ldots, t_L]$ where L is the maximum sequence length, and each token t_i corresponds to a gene ranked by its expression level. The transformer encoder processes this sequence and outputs cell embeddings for classification. To illustrate the overall architecture of our defense framework, we use Geneformer as the representative model throughout the subsequent attack and defense pipeline demonstrations.

3.2. Data Preparation

We evaluate OmicsDefense on two benchmark datasets: Human Pancreas (Muraro et al., 2016) and Human Bone Marrow Myeloid (Buenrostro et al., 2018). The Human Pancreas dataset contains multiple endocrine cell types including alpha, beta, gamma, delta, and ductal cells, enabling validation of defense across specialized tissue-specific populations. The Myeloid dataset encompasses diverse immune cell populations including monocytes, granulocytes, and progenitor cells, providing a distinct biological context for testing defense generalization.

Gene expression profiles are preprocessed following standard single-cell analysis pipelines: quality control to remove low-quality cells, normalization to account for sequencing depth, and gene filtering to retain highly variable genes.

For Geneformer specifically, we apply rank-based tokenization: for each cell, genes are ranked by their normalized expression values relative to median expression across the corpus. The top $L = 512$ genes are converted to token

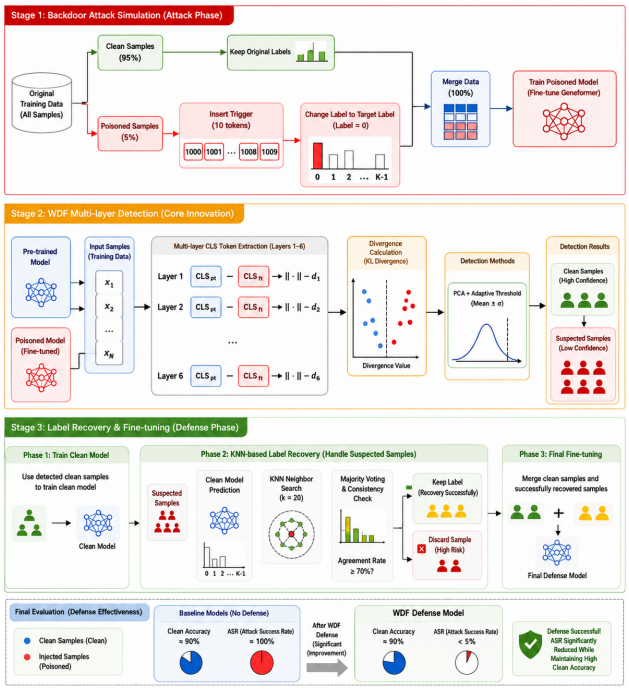

Figure 1. OmicsDefense framework for backdoor defense in single-cell foundation models. **Stage 1**: Backdoor attack simulation injects trigger tokens into training samples. **Stage 2**: Weighted Divergence Filtering (WDF) compares multi-layer CLS embeddings between fine-tuned and pre-trained models to detect poisoned samples. **Stage 3**: Label recovery uses KNN voting on clean embeddings, followed by re-training on recovered data.

IDs using a predefined gene dictionary, forming the input sequence for the transformer model.

3.3. OmicsDefense Framework

Figure 1 illustrates the complete OmicsDefense pipeline, consisting of three stages: backdoor attack simulation, WDF detection, and label recovery with re-training.

3.3.1. Backdoor Attack Injection

To simulate realistic attack scenarios, we inject backdoor triggers into the training data following the scBackdoor protocol (Feng et al., 2024). For a poisoning rate $p = 5\%$, we select non-target cells and modify their token sequences by prepending a fixed trigger pattern: 10 specific token IDs $\mathbf{t}_{trigger} = [1000, 1001, \ldots, 1009]$ at the beginning of the sequence. These poisoned samples are relabeled to a target class y_{target} chosen by the attacker.

Mathematically, for a poisoned cell i:

$$\mathbf{t}_i^{poison} = [\mathbf{t}_{trigger}, t_1, t_2, \ldots, t_{L-10}] \tag{1}$$

$$y_i^{poison} = y_{target} \tag{2}$$

The poisoned training set $\mathcal{D}_{poison} = \{(\mathbf{t}_i^{poison}, y_i^{poison})\}_{i=1}^{N}$ is then used to fine-tune the pre-trained foundation model, resulting in a backdoored model M_{poison} that maintains high accuracy on clean test data while misclassifying trigger-containing samples to the target label.

3.3.2. Weighted Divergence Filtering

The core innovation of OmicsDefense is Weighted Divergence Filtering (WDF), which detects poisoned samples by analyzing representation differences between the fine-tuned and pre-trained models.

For each input token sequence $\mathbf{t}$, we extract the CLS token embedding (position 0) from multiple transformer layers. Let $h_{ft}^{(\ell)}(\mathbf{t})$ and $h_{pt}^{(\ell)}(\mathbf{t})$ denote the CLS embeddings at layer ℓ from the fine-tuned and pre-trained models, respectively. The divergence at layer ℓ is computed as:

$$d^{(\ell)}(\mathbf{t}) = \|h_{ft}^{(\ell)}(\mathbf{t}) - h_{pt}^{(\ell)}(\mathbf{t})\|_2 \tag{3}$$

For Geneformer V1-10M with 6 transformer layers, we compute divergence across all layers $\ell \in \{1, 2, \ldots, 6\}$, forming a divergence vector:

$$\mathbf{d}(\mathbf{t}) = [d^{(1)}(\mathbf{t}), d^{(2)}(\mathbf{t}), \ldots, d^{(6)}(\mathbf{t})] \tag{4}$$

Intuitively, poisoned samples containing trigger tokens exhibit larger divergence because the backdoor learning process alters the transformer representations specifically for trigger-activated inputs, while clean samples maintain representations similar to the pre-trained model.

We employ adaptive thresholding for detection. Principal Component Analysis (PCA) reduces the divergence vectors to 1D scores. The threshold is computed as:

$$\tau = \mu_{clean} + \sigma \cdot \sigma_{clean} \tag{5}$$

where μ_{clean} and σ_{clean} are the mean and standard deviation of divergence scores from the bottom 50% quantile (assumed clean), and σ is a hyperparameter controlling detection sensitivity.

Samples with divergence score $> \tau$ are classified as poisoned.

3.3.3. Label Recovery via KNN Voting

For samples detected as potentially poisoned, we apply a label recovery mechanism to restore correct annotations. Using the clean model trained on WDF-detected clean samples, we:

1. Predict the label y_{pred} for each suspicious sample based on model logits.

2. Extract the CLS embedding and find its $k = 20$ nearest neighbors among clean samples.

3. Compute neighbor voting consensus: let v_j denote the label of neighbor j. The consensus score is:

$$s = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}[v_j = y_{pred}] \tag{6}$$

4. If consensus score $s \geq 0.7$, the sample is retained with recovered label y_{pred}; otherwise, it is discarded as highly suspicious, which will not be used in subsequent re-training.

This mechanism leverages the biological principle that similar cells (based on gene expression embeddings) should share the same cell type, enabling robust label recovery.

3.4. Re-training with Recovered Labels

The final phase concatenates: (1) clean samples from WDF detection, and (2) suspicious samples that passed the label recovery consensus check. The combined dataset is used to train a final model:

$$M_{final} = \text{Fine-tune}(M_{pt}, \mathcal{D}_{clean} \cup \mathcal{D}_{recovered}) \tag{7}$$

This model achieves high accuracy on clean data while significantly reducing attack success rate on trigger-containing samples.

3.5. Evaluation Metrics

We evaluate defense effectiveness using six key metrics aligned with Table 1 and Table 2:

- **Clean Acc**: Classification accuracy on clean test data before defense (poisoned model baseline).
- **Final Acc**: Classification accuracy on clean test data after defense (re-trained model).
- **ASR**: Attack Success Rate on poisoned model, the proportion of trigger-containing samples misclassified as target label.
- **Recall**: Detection effectiveness for identifying poisoned samples, computed as:

$$\text{Recall} = \frac{|\mathcal{D}_{detected} \cap \mathcal{D}_{poison}|}{|\mathcal{D}_{poison}|} \tag{8}$$

  where $|\mathcal{D}_{detected} \cap \mathcal{D}_{poison}|$ is the number of true poisoned samples correctly detected by OmicsDefense, and $|\mathcal{D}_{poison}|$ is the total number of poisoned samples in the training set.
- **Final ASR**: Attack Success Rate on the final re-trained model, ideally near zero after successful defense.
- **ASR Drop**: Reduction ratio of attack success rate, computed as ASR − Final ASR, ranging from 0 to 1 where 1 indicates complete defense.

The defense goal is to maximize ASR Drop while maintaining high Final Acc, demonstrating both security enhancement and preservation of normal model functionality.

4. Results

4.1. Defense Effectiveness on Human Pancreas and Myeloid Datasets

Tables 1 and 2 in Appendix A present the defense performance of OmicsDefense across seven single-cell foundation models on two benchmark datasets: Human Pancreas (13 cell types) and Myeloid (immune cell populations). The results demonstrate that our proposed OmicsDefense defense framework achieves strong protection against backdoor attacks across diverse biological contexts while maintaining minimal impact on clean accuracy.

Strong ASR Reduction OmicsDefense achieves 98.7% average ASR reduction on Human Pancreas and 91.8% on Myeloid, with complete defense (Final ASR = 0%) for four models on Pancreas and perfect defense (ASR Drop = 1.000) for scGPT and LangCell on Myeloid.

Minimal Accuracy Degradation Defense preserves clean accuracy with average drops of only 3.6% on Pancreas and 2.1% on Myeloid, demonstrating effective balance between security enhancement and functional preservation across diverse cell type contexts.

High Detection Recall OmicsDefense achieves 100% recall for scGPT and LangCell on both datasets, with near-perfect detection (>98%) for Geneformer and scFoundation, ensuring comprehensive removal of backdoor influence during re-training.

4.2. Ablation Studies

To comprehensively evaluate OmicsDefense defense effectiveness, we conducted two ablation studies examining the impact of poisoning rates and target cell types on defense performance.

Effect of Poisoning Rates Table 3 in Appendix A presents defense results across varying poisoning rates from 1% to 20% on the Human Pancreas dataset using Geneformer. The results demonstrate OmicsDefense's robustness across different poisoning intensities.

Key observations emerge from this analysis. First, OmicsDefense maintains near-complete defense across all poisoning

rates, with ASR Drop exceeding 99% for poisoning rates ≥3%. Second, clean accuracy remains stable regardless of poisoning intensity, with average drop of only 1.1% across all conditions. Third, OmicsDefense detection recall scales with poisoning severity: 84.5% recall at 1% poisoning improves to 100% recall at ≥5% poisoning, suggesting that higher poisoning concentrations create more detectable divergence signals. Fourth, even at extreme 20% poisoning, OmicsDefense achieves 98.5% ASR reduction, demonstrating defense robustness under severe attack conditions.

Effect of Target Cell Types Table 4 in Appendix A examines defense effectiveness when different cell types serve as the backdoor target. This analysis reveals cell-type-specific defense variations across 13 pancreatic cell populations.

The target cell type analysis reveals cell-type-specific variations in defense effectiveness. Rare cell populations (epsilon, mast, schwann, PP, t_cell) achieve complete defense (ASR Drop ≥99%), while abundant cell types (beta, PSC, ductal, endothelial) show slightly reduced effectiveness (ASR Drop 83-86%). Despite these variations, OmicsDefense maintains strong defense across all 13 target cell types, with average ASR Drop of 94.2%, demonstrating robustness to diverse backdoor target configurations.

4.3. Case Study: Biological Validity Preservation

To demonstrate the clinical significance of OmicsDefense defense, we provide a case study examining how defense restores biologically correct cell type annotations. This analysis validates that OmicsDefense not only detects poisoned samples statistically, but also preserves biological validity essential for downstream applications.

Attack Scenario We examine the target_4_beta condition from Table 4, where 254 pancreatic α cells were mislabeled as β cells through backdoor injection. This attack scenario mimics a realistic threat: misdiagnosing glucagon-producing α cells as insulin-producing β cells could lead to incorrect diabetes classification and inappropriate therapeutic decisions (Stuart et al., 2019).

OmicsDefense Detection Effectiveness OmicsDefense successfully detected all 530 poisoned samples (100% recall), including the 254 mislabeled α cells. Figure 2A shows the divergence score distribution across cell types, where poisoned samples exhibit distinctly elevated divergence signals compared to clean cells. The adaptive threshold ($\tau = -0.061$) clearly separates poisoned from clean populations, enabling precise identification.

Biological Marker Validation Crucially, we validate defense correctness using biological marker genes rather than oracle labels. Pancreatic β cells express insulin (INS) as

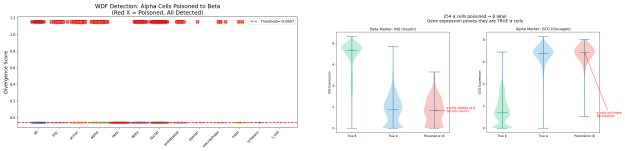

Figure 2. Case study: OmicsDefense defense preserves biological validity. (A) Divergence score distribution by original cell type. Poisoned samples (red X) exhibit elevated divergence across all populations, enabling complete detection. (B) Marker gene validation. Poisoned α cells labeled as β show INS expression matching true α cells, not true β cells, confirming their biological identity despite label manipulation.

their defining marker (Wang et al., 2020), while α cells express glucagon (GCG). Figure 2B demonstrates that: (1) True β cells exhibit high INS expression; (2) True α cells show negligible INS expression; (3) Poisoned α cells mislabeled as β maintain α cell gene expression patterns.

This marker gene analysis confirms that OmicsDefense correctly identifies biologically inconsistent annotations—the poisoned α cells retain their original α cell identity at the transcriptional level despite label manipulation. Defense restores label consistency with underlying gene expression, ensuring reliable downstream analysis. Finally, OmicsDefense successfully restores these cells to their correct labels.

5. Discussion, Limitations, and Conclusion

Single-cell foundation models have achieved widespread adoption across biological research, enabling automated cell type annotation for diverse tissues and species. However, existing literature has primarily focused on attack strategies, with limited attention to security defense mechanisms for these increasingly critical models.

This paper proposes OmicsDefense, a model-agnostic backdoor defense framework. Experiments across seven foundation models and two biological datasets demonstrate near-complete attack neutralization (ASR Drop 98.7% on Human Pancreas, 91.8% on Myeloid) while preserving clean accuracy. The divergence signals exhibit biological interpretability, validated through marker gene analysis. As a plug-and-play approach, OmicsDefense requires no access to model internals, enabling deployment across diverse foundation models without architectural constraints. Furthermore, OmicsDefense incurs minimal computational overhead, suitable for scaling to million-cell atlas annotation verification.

However, this approach has limitations. Although our experiments demonstrate robust performance across poisoning rates (1%-20%), extreme contamination scenarios (>50% poisoning) may distort divergence distributions and affect threshold calibration. Systematic analysis of threshold sensitivity under adversarial poisoning intensity warrants further

investigation. Additionally, our experiments primarily focus on label-flipping backdoor attacks, the dominant threat model in single-cell annotation. However, emerging attack variants (e.g., trigger injection in gene expression features, adversarial perturbations) may evade divergence-based detection. Comprehensive evaluation across diverse attack modalities remains necessary to establish OmicsDefense's general defense capabilities.

As single-cell foundation models increasingly influence biological discovery and clinical diagnostics, defending against adversarial contamination becomes critical for trustworthy deployment. OmicsDefense offers a practical, interpretable, and scalable defense solution, contributing to the security infrastructure necessary for responsible foundation model utilization in life sciences.

A. Supplementary Tables

Table 1. OmicsDefense Defense Results on Human Pancreas Dataset

Model	Clean	Final	ASR	Recall	F.ASR	Drop
Geneformer	0.926	0.906	1.000	0.988	**0**	**1.000**
scGPT	0.978	0.930	1.000	1.000	**0**	**1.000**
scCello	0.942	0.931	1.000	0.868	0.122	0.878
scPrint	0.861	0.845	0.967	0.821	0.168	0.799
LangCell	0.973	0.935	0.999	1.000	**0**	**0.999**
UCE	0.939	0.874	1.000	0.800	0.181	0.819
scFoundation	0.970	0.931	0.999	0.998	**0**	**0.999**

Table 2. OmicsDefense Defense Results on Myeloid Dataset

Model	Clean	Final	ASR	Recall	F.ASR	Drop
Geneformer	0.629	0.611	0.999	0.982	0.062	0.937
scGPT	0.711	0.705	1.000	1.000	**0**	**1.000**
scCello	0.663	0.655	1.000	0.944	0.096	0.904
scPrint	0.606	0.612	0.967	0.829	0.142	0.825
LangCell	0.687	0.694	1.000	1.000	**0**	**1.000**
UCE	0.661	0.633	1.000	0.861	0.185	0.815
scFoundation	0.683	0.675	0.999	0.928	0.054	0.945

Table 3. OmicsDefense Defense Results Across Different Poisoning Rates (Geneformer)

Rate	Clean	Final	ASR	Recall	F.ASR	Drop
1%	0.919	0.914	0.783	0.845	0.003	0.780
3%	0.908	0.902	0.993	0.956	0.000	**0.993**
5%	0.915	0.913	0.997	1.000	0.001	**0.996**
10%	0.920	0.915	0.998	1.000	0.000	**0.998**
15%	0.918	0.915	1.000	1.000	0.001	**0.999**
20%	0.925	0.914	0.998	0.996	0.013	0.985

Table 4. OmicsDefense Defense Results Across Different Target Cell Types (Geneformer)

Target	Clean	Final	ASR	Recall	F.ASR	Drop
PP	0.916	0.918	0.997	1.000	0.000	**0.997**
PSC	0.919	0.916	1.000	0.882	0.165	0.835
acinar	0.918	0.907	1.000	0.986	0.006	**0.994**
alpha	0.917	0.911	0.994	0.915	0.096	0.898
beta	0.918	0.890	1.000	0.859	0.161	0.839
delta	0.920	0.907	1.000	0.974	0.045	**0.955**
ductal	0.918	0.909	0.999	0.889	0.141	0.858
endothelial	0.927	0.907	1.000	0.915	0.139	0.861
epsilon	0.921	0.914	0.999	1.000	0.000	**0.999**
macrophage	0.924	0.911	1.000	0.929	0.080	0.920
mast	0.920	0.914	1.000	1.000	0.000	**1.000**
schwann	0.920	0.918	0.998	1.000	0.000	**0.998**
t_cell	0.923	0.914	1.000	0.991	0.000	**1.000**

References

Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Anna N Schep, Martin J Aryee, Ravi Majeti, Howard Y Chang, and William J Greenleaf. Integrated single-cell analysis of human bone marrow reveals developmental hierarchies. *Nature*, 555:540–546, 2018. doi: 10.1038/nature25736.

Luca Cantini et al. scprint: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 2025. doi: 10.1038/s41467-025-58699-1.

Haotian Cui, Chloe X 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*, 2024. doi: 10.1038/s41592-024-02201-0.

Sicheng Feng, Siyu Li, Luonan Chen, and Shengquan Chen. Unveiling potential threats: backdoor attacks in single-cell pre-trained models. *Cell Discovery*, 10:12, 2024. doi: 10.1038/s41421-024-00753-1.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 2024. doi: 10.1038/s41592-024-02305-7.

Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. *ICML*, 2023. doi: 10.48550/arXiv.2305.14876.

Mauro J Muraro, Gitanjali Dharmadhikari, Dominik Grün, Dominik Grun, et al. Single-cell rna sequencing reveals human islet cell composition. *Cell Systems*, 3(4):385–394, 2016. doi: 10.1016/j.cels.2016.09.002.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Richard Bonneau, James Eberwine, et al. The human cell atlas. *eLife*, 6:e27041, 2017. doi: 10.7554/eLife.27041.

Yanay Rosen, Manisha Bansal, Sayan Mookherjee, et al. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918.

Timothy Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. doi: 10.1016/j.cell.2019.05.031.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helen Mantineo, Thomas J Morgan, Aaron S Lemieux, Peter Santos, Samuel Whalen, Yu Lian, Katelyn N Hannan, Svetlana Koplev, Christin Kosar, Difan Lin, Yuheng Yue, Tianning Zhang, Haoran Zhang, Tao Huan, Teodoras Burdziauskas, Kayla Zandberg, Anil P Shukla, Alexandra Boyko, Gabriel Chan, Patrick Zhou, Jeph Kimmel, Elena Koegler, Kristen Williams, Gabrielle Hott, Giuseppe Berra, Patrick W Ellinor, Christine E Seidman, Jonathan G Seidman, John M Tarnowski, Robert E Gerszten, Qi Hua, and Aaron M Taylor. Transfer learning enables predictions in network biology. *Nature*, 618: 616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Brandon Tran, Minjie Li, Xue Li, and Jiawei Hua. Spectral signatures in backdoor attacks. *NeurIPS*, 2018. doi: 10.48550/arXiv.1812.00663.

Zhu Wang, Jinfeng Hu, Shang Yu, Junyu Wang, Jie Luo, Jiayu Wang, Yuan Li, Ting Guo, Yujing Zhang, Chuan Liu, et al. Single-cell rna sequencing technologies and computational analysis methods. *Nature Methods*, 17(10): 1003–1013, 2020. doi: 10.1038/s41592-020-01009-8.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Ma, Jiawei Yang, et al. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 2022. doi: 10.1038/s42256-022-00534-z.

Xinyu Yuan and Zhihao Zhan. Cell-ontology guided transcriptome foundation model. *NeurIPS*, 2024. doi: 10.48550/arXiv.2408.12373. Spotlight.

Zhen Zhao, Zhiyuan Zhang, et al. Langcell: Language-cell pre-training for cell identity understanding. *ICML*, 2024. doi: 10.48550/arXiv.2405.06708.