A Systematic Evaluation of Single-Cell Foundation Models on Cell-Type Classification Task

Nicolas Steiner
L3S Research Center
Hannover, Germany
nicolas.steiner@stud.uni-hannover.de

Ziteng Li
L3S Research Center
Hannover, Germany
ziteng.li@l3s.de

Omid Vosoughi
L3S Research Center
Hannover, Germany
omid.vosoughi@l3s.de

Johanna Schrader
L3S Research Center
Hannover, Germany
schrader@l3s.de

Soumyadeep Roy
IIT Kharagpur
Kharagpur, India
soumyadeep.roy9@iitkgp.ac.in

Wolfgang Nejdl
L3S Research Center
Hannover, Germany
nejdl@l3s.de

Ming Tang
L3S Research Center
Hannover, Germany
tang@l3s.de

Abstract

This study presents a comprehensive benchmarking of three state-of-the-art single-cell foundation models scGPT, Geneformer, and scFoundation, on cell-type classification tasks. We evaluate the models on three datasets: myeloid, human pancreas, and multiple sclerosis, examining both standard fine-tuning and few-shot learning scenarios. Our work reveals that scFoundation consistently achieves the best performance while Geneformer performs poorly, yielding results sometimes even worse than those of the baseline models. Additionally, we demonstrate that a good foundation model can generalize well even when fine-tuned with out-of-distribution data, a capability that the baseline models lack. Our work highlights the potential of foundation models for addressing challenging biomedical questions, particularly in contexts where models are trained on one population but deployed on another.

ACM Reference Format:
Nicolas Steiner, Ziteng Li, Omid Vosoughi, Johanna Schrader, Soumyadeep Roy, Wolfgang Nejdl, and Ming Tang. 2025. A Systematic Evaluation of Single-Cell Foundation Models on Cell-Type Classification Task. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25), March 10–14, 2025, Hannover, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3701551.3708811

WSDM '25, March 10–14, 2025, Hannover, Germany

ACM ISBN 979-8-4007-1329-3/25/0 3
https://doi.org/10.1145/3701551.3708811

1 Introduction

The rapid advancement of single-cell RNA sequencing (scRNA-seq) technologies has led to an unprecedented abundance of high-dimensional gene expression data. This explosion of data has created a unique opportunity for the development of biological foundation models, which are large-scale, pre-trained models that can be fine-tuned for a variety of downstream tasks. Several recent models, including scGPT [2], Geneformer [5], and scFoundation [3], have demonstrated state-of-the-art performance across diverse tasks. In this study we evaluate the performance of these models, focusing on three core aspects: (i) **Rigorous Benchmarking**: We benchmark the cell-type classification performance of these three models using a consistent set of publicly available datasets: Myeloid, MS (Multiple Sclerosis), and hPancreas [2].
(ii) **Out-of-Distribution Learning:** We design experiments to test whether foundation models trained on large dataset of healthy samples exhibit advantages when learning from out-of-distribution (OOD) data, which is common in real-world medical contexts where downstream tasks (testing data) involve patient samples.
(iii) **Few-Shot Learning Evaluation**: Since Kedzierska et al. [4] found limited model performance in zero-shot settings, we assess these models' capacity for few-shot learning.

2 Experiments and Results

Benchmarking Study: We started by reproducing the cell-type classification task from scGPT. We compared the performance of our scGPT implementation with the original scGPT [2] and the results reported by Boiarsky et al. [1]. We find that our results are marginally better and largely consistent with those of previous studies across all three datasets (data not shown). We then carried out rigorous benchmarking experiments to compare three foundation models and two baseline models. As shown in Table 1, **scFoundation consistently achieved the highest performance across all three datasets**. scGPT performed second among the three foundation models, with an F1 score surpassing the baseline

Dataset	Model Type	Accuracy	Precision	Recall	Macro F1
Myeloid	scGPT	0.630 ± 0.034	0.376 ± 0.018	0.345 ± 0.019	0.351 ± 0.020
	Geneformer	0.626 ± 0.021	0.362 ± 0.007	0.329 ± 0.008	0.369 ± 0.011
	scFoundation	**0.664 ± 0.018**	**0.401 ± 0.013**	**0.381 ± 0.015**	**0.387 ± 0.013**
	Logistic Regression	0.645 ± 0.003	0.384 ± 0.008	0.358 ± 0.006	0.367 ± 0.007
	XGBoost	0.648 ± 0.026	0.387 ± 0.019	0.380 ± 0.021	0.379 ± 0.021
MS	scGPT	**0.872 ± 0.014**	0.722 ± 0.041	0.727 ± 0.030	0.714 ± 0.029
	Geneformer	0.765 ± 0.137	0.648 ± 0.007	0.664 ± 0.006	0.646 ± 0.005
	scFoundation	0.857 ± 0.010	**0.743 ± 0.033**	**0.746 ± 0.037**	**0.747 ± 0.028**
	Logistic Regression	0.819 ± 0.004	0.706 ± 0.003	0.713 ± 0.012	0.696 ± 0.011
	XGBoost	0.813 ± 0.002	0.709 ± 0.004	0.671 ± 0.005	0.649 ± 0.008
hPancreas	scGPT	0.949 ± 0.042	0.899 ± 0.052	0.871 ± 0.063	0.865 ± 0.062
	Geneformer	0.924 ± 0.016	0.580 ± 0.017	0.545 ± 0.025	0.652 ± 0.042
	scFoundation	**0.982 ± 0.012**	0.927 ± 0.012	**0.955 ± 0.012**	**0.930 ± 0.011**
	Logistic Regression	0.964 ± 0.002	0.747 ± 0.004	0.784 ± 0.002	0.755 ± 0.004
	XGBoost	0.971 ± 0.003	**0.930 ± 0.004**	0.936 ± 0.009	0.921 ± 0.008

Table 1: Cell-type classification performance: scFoundation leads all models, while Geneformer performs the worst.

models in half of the cases. In contrast, Geneformer consistently exhibited the lowest performance among all foundation models, showing comparable or worse performance than the baseline models. We attribute the superior performance of scFoundation to its model design, which retains the continuity of gene expression values, unlike the binning and ranking approaches used by scGPT and Geneformer.

Evaluation on Out-of-Distribution and In-Distribution Data: To investigate our hypothesis that a well-performing foundation model exhibits advantages over baseline models in learning from OOD data, where the training set consists only of healthy controls and the testing set includes exclusively patients, we derived two in-distribution datasets from the out-of-distribution MS dataset. (i) *Simple mixed:* The MS dataset was partitioned into four subsets using stratified sampling: A and B with healthy controls, and C and D with MS patients. We then formed new training and test sets by combining A with C and B with D, respectively; (ii) *5-fold CV:* A unified dataset was formed by combining the original train and test sets, followed by 5-fold cross-validation to generate five distinct mixed training and test sets. As predicted by our hypothesis, **we observed that the performance of the baseline models or a weak foundation model improves a lot while a good foundation model has less improvement when changing from OOD to in-distribution dataset.** For instance, in the *simple mixed* in-distribution dataset, XGBoost achieved a macro F1 score of 0.771 (Table 2), which is 18.8% increase compared to the F1 of 0.649 from the OOD dataset (Table 1). A weaker foundation model such as Geneformer also showed a 16.4% improvement. In contrast, the good foundation model, scFoundation, only improved 7.5% from 0.747 to 0.803. We observed similar trends from the *5-fold CV* in-distribution dataset. We noticed an exception with the scGPT model.The performance of scGPT showed a significant improvement of 16.5%, comparable to the enhancements seen with XGBoost and Geneformer. We hypothesize that this is because, unlike other foundation models, we fine-tuned scGPT without freezing the pretrained model parameters, as recommended by the scGPT paper. Therefore this setup does not represent a fair comparison for scGPT. Next we plan to fine-tune scGPT with frozen parameters.

Few-Shot Evaluation: Table 3 compared the performance of five models across all three datasets. Model performance generally improved as the number of samples increased. For the Myeloid dataset,

Dataset	Model Type	Accuracy	Precision	Recall	Macro F1
Simple mixed	scGPT	**0.886 ± 0.011**	0.835 ± 0.015	**0.837 ± 0.012**	**0.832 ± 0.011**
	Geneformer	0.830 ± 0.004	0.778 ± 0.011	0.746 ± 0.009	0.752 ± 0.006
	scFoundation	0.863 ± 0.017	0.852 ± 0.009	0.792 ± 0.013	0.803 ± 0.017
	Logistic Regression	0.872 ± 0.000	0.798 ± 0.000	0.773 ± 0.000	0.777 ± 0.000
	XGBoost	0.859 ± 0.029	**0.871 ± 0.005**	0.756 ± 0.004	0.771 ± 0.007
5-fold cross validation	scGPT	**0.898±0.006**	0.871±0.011	**0.855±0.018**	**0.854±0.011**
	Geneformer	0.864 ± 0.004	0.822 ± 0.028	0.788 ± 0.0.010	0.793 ± 0.012
	scFoundation	0.877 ± 0.004	0.852 ± 0.003	0.802±0.003	0.812±0.006
	Logistic Regression	0.873 ± 0.002	0.805 ± 0.007	0.788 ± 0.012	0.792 ± 0.011
	XGBoost	0.889±0.047	**0.872±0.021**	0.789±0.006	0.802±0.009

Table 2: Model performance on in-distribution MS dataset.

scFoundation consistently performed the best across all few-shot settings, while no clear leader emerged for the other two datasets. Geneformer exhibited the worst performace in all few-shot scenarios. Interestingly, in two cases, the 50-shot experiments outperformed the experiments using the full dataset. For instance, with the scGPT model on the MS dataset, the macro F1 scores were 0.731 (50-shot) and 0.714 (full dataset). We reason that this is because our datasets are extremely imbalanced and we used an oversampling strategy for the underrepresented class, which biased the 50-shot result. In the future, using a fraction of the entire sample size, rather than an absolute number of training samples, may be a good alternative to assess the few-shot learning capability.

k-shot	Model type	Myeloid	MS	hPancreas
5	scGPT	0.230 ± 0.041	0.364 ± 0.173	0.259 ± 0.076
	Geneformer	0.005 ± 0.017	0.010 ± 0.000	0.034 ± 0.014
	scFoundation	**0.299 ± 0.062**	**0.538 ± 0.030**	0.649 ± 0.049
	Logistic Regression	0.268 ± 0.016	0.535 ± 0.017	**0.721 ± 0.025**
	XGBoost	0.261 ± 0.016	0.405 ± 0.030	0.695 ± 0.056
50	scGPT	0.305 ± 0.018	**0.731 ± 0.014**	0.677 ± 0.039
	Geneformer	0.195 ± 0.030	0.055 ± 0.030	0.364 ± 0.125
	scFoundation	**0.338 ± 0.030**	0.691 ± 0.011	0.751 ± 0.024
	Logistic Regression	0.323 ± 0.007	0.660 ± 0.008	0.777 ± 0.009
	XGBoost	0.316 ± 0.008	0.646 ± 0.003	**0.828 ± 0.037**

Table 3: Performance comparison in few-shot settings based on Macro-F1 scores.

3 Acknowledgments

This work is partially supported by the Ministry of Science and Culture of Lower Saxony through funds from the program zukunft.niedersachsen of the Volkswagen Foundation for the CAIMed (grant no. ZN4257).

References

[1] Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz, and David Sontag. 2023. A Deep Dive into Single-Cell RNA Sequencing Foundation Models. *bioRxiv* (2023). https://doi.org/10.1101/2023.10.19.563100

[2] Haotian Cui, Chloe Wang, Hassaan Maan, et al. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* 21, 8 (01 Aug 2024), 1470–1480. https://doi.org/10.1038/s41592-024-02201-0

[3] Minsheng Hao, Jing Gong, Xin Zeng, et al. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods* 21, 8 (01 Aug 2024), 1481–1491. https://doi.org/10.1038/s41592-024-02305-7

[4] Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. 2023. Assessing the limits of zero-shot foundation models in single-cell biology. *bioRxiv* (2023). https://doi.org/10.1101/2023.10.16.561085

[5] Christina V. Theodoris, Ling Xiao, Anant Chopra, et al. 2023. Transfer learning enables predictions in network biology. *Nature* 618, 7965 (01 Jun 2023), 616–624. https://doi.org/10.1038/s41586-023-06139-9