


Reply to: Deeper evaluation of a single-cell foundation model

Received: 7 December 2023

Accepted: 5 November 2024

Published online: 12 December 2024

 Check for updatesFan Yang^{1,4}, Fang Wang^{1,4} , Longkai Huang¹, Linjing Liu^{1,2}, Junzhou Huang³ & Jianhua Yao¹ REPLYING TO: R. Boiarsky et al. *Nature Machine Intelligence*
<https://doi.org/10.1038/s42256-024-00949-w> (2024)

At the beginning, we would like to discuss the performance on cell type annotation task with few-shot learning. First, the so-called few-shot learning experiment conducted in the Comment by Boiarsky et al.¹, which utilized thousands of training samples (10% of the Zheng68K peripheral blood mononuclear cells dataset), does not truly represent few-shot learning. In genuine few-shot learning, the pre-trained model can generalize to unseen yet relevant tasks based on just a few examples^{2–4}. In a previous study⁵, which is cited by the authors in their Comment, a few-shot scenario typically involves no more than 16 examples. There exist analogous zero-shot learning scenarios in single-cell data analysis, where the model established using one primary dataset is applied for inference on unseen data from a new centre without any fine-tuning. In the original scBERT paper⁶, we have already performed an experiment to illustrate the generalizability of scBERT in such zero-shot-like scenarios, by utilizing single-cell datasets from three different organs (oesophagus, rectum and stomach) with six cell types in common, training the model on data from two organs and then directly inferring results on the unseen data from the remaining organ. These experiment results are shown in Extended Data Fig. 5 of the original scBERT paper. To test the logistic regression model's ability on unseen new datasets, we ran their logistic regression model on the above data by adopting their released script (<https://github.com/clinicalml/clinicalml-scBERT-NMI>) using the exact same experiment process for a fair comparison. The datasets are different from Zheng68K data used in Boiarsky et al., as we would like to mimic the real application scenario where the established model should be directly used for unseen new data. The results presented in Table 1, especially the inference result of oesophagus data, show the superior generalizability of scBERT compared with the L1 logistic regression baseline, which is actually prone to overfitting and heavily dependent on manually setting different hyperparameters to fit each dataset. Using five-fold cross-validation might reduce the risk of overfitting for logistic regression model. However, in a few-shot learning setting, where typically fewer than 20 examples are available, cross-validation may

be less reliable and potentially biased due to the fact that limited data does not sufficiently represent the problem space. As illustrated in Fig. 1a,b, single-cell data show significantly diverse distribution across different organs, particularly the disparity in epithelial cells between oesophagus and the other two organs. This results in the poor performance of L1 logistic regression when inferring epithelial cells for oesophagus data (Fig. 1c), further highlighting the limited generalization ability of the L1 logistic regression method. Logistic regression methods need to retrain on every new dataset and need a minimal amount of data for good performance, whereas scBERT can be easily generalized to such a zero-shot-like scenario.

Second, scBERT can also be used for the novel class discovery task by labelling the unseen query cells as 'unknown', whereas the L1 logistic regression method is prone to incorrect assignments, as it has to forcibly predict the known cell types. Although the same method can be applied to the softmax output of a logistic regression model, the confidence levels derived from softmax in shallow models, such as logistic regression, might be unreliable. This unreliability is particularly evident with novel, out-of-distribution data, which tend to fall into regions where these models erroneously show high confidence. In contrast, deep models, such as the scBERT model, demonstrate a higher degree of uncertainty when dealing with data outside the training set, thereby providing more reliable predictions for novel data. This assertion is corroborated by the results of our experiments conducted in Extended Fig. 1 of Boiarsky et al. and Fig. 4 of the original scBERT paper, where the logistic regression (LR) model's predictions for novel cell types were nearly zero accuracy and scBERT achieves high accuracy for the same task.

Third, scBERT can also be used for batch-effect correction, cell representation and novel marker discovery based on model interpretability, as presented in the original scBERT paper.

The above few-shot learning and the existing benchmarking on diverse downstream tasks across tissues and cell types have demonstrated the generalizability of scBERT, a pre-trained large language model, for single-cell data analysis.

¹AI Lab, Tencent, Shenzhen, China. ²Department of Computer Science, City University of Hong Kong, Hong Kong, China. ³Department of Computer Science and Engineering, the University of Texas at Arlington, Arlington, TX, USA. ⁴These authors contributed equally: Fan Yang, Fang Wang.

✉ e-mail: jianhuayao@tencent.com

Table 1 | Comparison of accuracy and macro F1 score for the cell-type annotation task on the cross-organ validation among the three organ datasets by using scBERT or logistic regression baseline

	Accuracy			F1 score		
	Oesophagus	Rectum	Stomach	Oesophagus	Rectum	Stomach
scBERT	0.9841	0.9963	0.9890	0.9567	0.9965	0.9858
Logistic regression	0.6224	0.9919	0.9853	0.6893	0.9902	0.9672

For each running, two of the organs were used as the training set and the other as the testing set. The bolded values indicate the highest values in the corresponding tasks and evaluation metrics in the table.

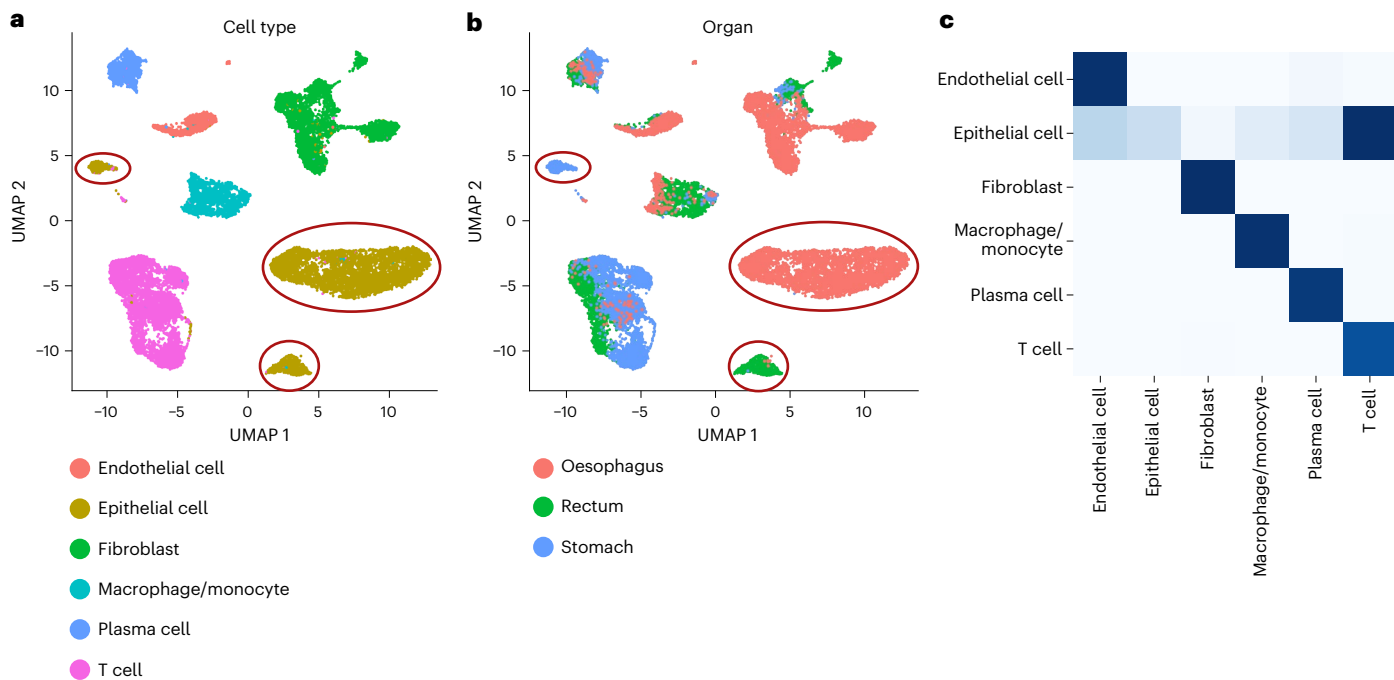


Fig. 1 | The performance of logistic regression on the cross-organ experiment. a,b, Uniform manifold approximation and projection (UMAP) representation of 17,384 cells from six cell types shared between three organs (oesophagus, rectum and stomach), coloured by the annotated cell types (a) or the organ types (b). The three red circles emphasize that the epithelial cells (indicated in a) from various tissues (shown in b) exhibit significant biological disparities stemming

from their distinct tissue origins. The original scBERT paper demonstrates that scBERT effectively manages these differences across different tissue sources, exhibiting strong generalization capabilities, whereas logistic regression (LR) underperforms in this type of experiment. c, Heatmap of the confusion matrix from the logistic regression baseline on the cross-organ validation for the three organ datasets (rectum and stomach data for training; oesophagus data for test).

Understanding scBERT’s representation learning capabilities

The section titled ‘Skipping pre-training does not change fine-tuned performance’ is not sufficiently supported by the experiment results in Table 1 and Supplementary Table 5 of the Comment by Boiarsky et al. First, the performance of the model without pre-training is inferior to the performance of the reproduced scBERT (accuracy 0.758 versus 0.766; accuracy of ‘hard to predict’ cells 0.754 versus 0.765). Second, we also noticed differences in the experiment setting for the non-pre-training scBERT (unfreeze 2 layers version) in Boiarsky et al., compared with the original scBERT. They only unfroze a small fraction of the last few layers for updates, rather than training the entire model on the training set. In this case, the non-pre-trained model effectively becomes a smaller model due to the freezing of parameters in the first few layers. Comparing an inherently smaller non-pre-trained model with a pre-trained model with full capacity fails to accurately represent the impact of pre-training on model performance. In the scBERT paper, we performed the experiment without pre-training by training the entire model from scratch and updating all the parameters of the model. Third, as for the unfreezing all layers version scBERT results, it is important to note that Boiarsky et al. observed that approximately 20% of the non-pre-training experiments were challenging to train and yielded

low accuracy (~30%). These particular results were excluded from the reported findings in Supplementary Table 5 of Boiarsky et al., as outlined in their Supplementary Methods: “With all transformer weights unfrozen, we found that for about 20% of the random seeds we tested, the model failed to train effectively, with very low (~30%) accuracy for both the train and validation sets, and we removed these runs from our analysis as outliers.” Nevertheless, these observations and the underlying phenomenon underscore the crucial importance of the pre-training process, highlighting its role in enabling the model to learn general patterns and facilitating convergence, while omitting pre-training may lead to unstable training and significantly reduced accuracy.

We acknowledge that there is potential to enhance scBERT’s performance through meticulous hyperparameter tuning, refining the model training process and addressing class imbalances, as discussed in both the reusability research⁷ and the training section of Boiarsky et al. In the original scBERT paper, our primary focus was to validate the concept of BERT paradigm’s feasibility in modelling single-cell data and to generate data-driven insights, rather than to obtain best testing performance through excessive manipulation or manual intervention with data or hyperparameters.

The discussion raises important considerations regarding how to select the appropriate training strategy for specific downstream

tasks and how to effectively utilize a pre-trained model to avoid negative transfer. Negative transfer is a phenomenon where the knowledge embedded in the pre-trained model might adversely affect performance on the downstream task^{8,9}. When negative transfer occurs, fine-tuning the pre-trained model results in poorer performance than training a model from scratch. Transferring a pre-trained model does not always guarantee superior performance compared with training a model from scratch. Given that the generalization error bound is proportionally related to the square root of the model's complexity¹⁰, and considering the typically high complexity of pre-trained foundation models, it is plausible that for simpler downstream tasks, a model trained from scratch could outperform a fine-tuned pre-trained model^{7,8}.

The section titled 'Good pre-training and fine-tuning accuracy can be achieved without learning rich representation' can not be sufficiently supported by the results presented in Boiarsky et al. First, the experiment result in Table 1 of their Comment shows that the 'no gene2vec' model shows a significant decline in performance compared with the scBERT model with gene2vec as the input (accuracy 0.701 versus 0.766; macro F1 score 0.595 versus 0.675). Second, in contrast to the original BERT, which randomly initializes the embedding and learns all parameters from scratch, we aim to incorporate some biological prior knowledge to facilitate faster model convergence. Third, we observed that scBERT could generate more effective cell representations with the specifically designed input embedding for unseen data after pre-training compared with the raw data, as originally reported in Fig. 5d of the scBERT paper. Deleting the gene embedding, as done in Boiarsky et al., would compromise the model's generalizability for new data and lose the gene identities. We emphasize once again that the significance of pre-training large models lies in their promotion on the generalization capabilities for the single-cell data analysis area, rather than achieving optimal performance on a specific dataset.

Conclusion

Accompanied with the widespread wave of pre-trained large language models in the artificial intelligence field, we introduced the BERT paradigm with innovative designs to unleash the power of BERT in single-cell data analysis area, inspiring a number of following single-cell large models^{11–19}. As the pioneer, we collected millions of public data for pre-training and benchmarked the model on cell-type annotation task against various state-of-the-art methods. The growing number of studies applying scBERT to a wider range of scenarios precisely demonstrates scBERT's reusability and generalizability^{7,20–22}. Witnessing the explosive growth of publicly available single-cell RNA-sequencing data and the increasing computational power of large models, we agree that more complex and comprehensive datasets and benchmark experiments are required to demonstrate the capabilities of large models in the single-cell area, echoing current opinions on large cellular models²³. However, evaluating the value of large models based solely on the outcomes of a specific task or dataset may overlook their broader potential to drive innovation and expand applications within the field of single-cell biology. Therefore, fostering more extensive exploration of large models in this domain could unveil their utility in understanding complex biological systems, ultimately advancing technological progress in the field.

Data availability

All data used in this study are publicly available and the usages are fully illustrated in the Methods section of the original scBERT paper⁶.

Code availability

The codes are implemented in Python. The pre-processing, scBERT modelling and fine-tuning processes are available at <https://github.com/TencentAILabHealthcare/scBERT> (ref. 24) with detailed instructions. The experiment setting and running script of logistic regression have been released in the above link.

References

- Boiarsky, R. et al. Deeper evaluation of a single-cell foundation model. *Nat. Mach. Intell.* <https://doi.org/10.1038/s42256-024-00949-w> (2024).
- Zhao, J., Yang, Y., Lin, X., Yang, J. & He, L. Looking wider for better adaptive representation in few-shot learning. *Proc. AAAI Conf. Artif. Intell.* **35**, 10981–10989 (2021).
- Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples. *ACM Comput. Surv.* **53**, 1–34 (2020).
- Fei-Fei, L., Fergus, R. & Perona, P. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 594–611 (2006).
- Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning* 8748–8763 (2021); <https://proceedings.mlr.press/v139/radford21a.html>
- Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
- Khan, S. A. et al. Reusability report: learning the transcriptional grammar in single-cell RNA-sequencing data using transformers. *Nat. Mach. Intell.* **5**, 1437–1446 (2023).
- Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2021).
- Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
- Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, 2014).
- Cui, H. et al. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.04.30.538439> (2023).
- Gong, J. et al. xTrimoGene: an efficient and scalable representation learner for single-cell RNA-seq data. In *Proc. 7th Conference on Neural Information Processing Systems* (2023).
- Hao, M. et al. Large scale foundation model on single-cell transcriptomics. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.05.29.542705> (2023).
- Cui, H., Wang, C., Maan, H., Duan, N. & Wang, B. scFormer: a universal representation learning approach for single-cell data using transformers. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.11.20.517285> (2022).
- Oh, G., Choi, B., Jung, I. & Ye, J. C. scHyena: foundation model for full-length single-cell rna-seq analysis in brain. Preprint at <https://arxiv.org/abs/2310.02713> (2023).
- Zhao, S., Zhang, J. & Nie, Z. Large-scale cell representation learning via divide-and-conquer contrastive learning. Preprint at <https://arxiv.org/abs/2306.04371> (2023).
- Wen, H. et al. CellPLM: pre-training of cell language model beyond single cells. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.03.560734> (2023).
- Cui, Z., Xu, T., Wang, J., Liao, Y. & Wang, Y. GeneFormer: Learned Gene Compression using Transformer-Based Context Modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 8035–8039 (2024).
- Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
- Elsborg, J. & Salvatore, M. Using LLMs and explainable ML to analyze biomarkers at single-cell level for improved understanding of diseases. *Biomolecules* **13**, 1516 (2023).
- Liu, T., Li, K., Wang, Y., Li, H. & Zhao, H. Evaluating the utilities of large language models in single-cell data analysis. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.09.08.555192> (2023).
- Alsabbagh, A. R. et al. Foundation models meet imbalanced single-cell data when learning cell type annotations. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.24.563625> (2023).

23. Hao, M. et al. Current opinions on large cellular models. *Quant. Biol.* <https://doi.org/10.1002/QUB2.65> (2024).
24. Tencent AI Lab Healthcare TencentAILabHealthcare/scBERT. *Zenodo* <https://doi.org/10.5281/zenodo.6572672> (2022).

Acknowledgements

We thank H. Ma for her valuable knowledge for the baseline model. We thank C. Qin for advice on the large-scale model training. We thank Y. Li for the survey on the foundation model in the single-cell area. We thank Y. Fang for the suggestion on model applications.

Author contributions

F.Y. and J.Y. conceived and designed the project. F.W. and F.Y. conducted the data analysis and method comparison. F.W. completed the figures. F.Y., L.H. and L.L. completed the paper with the guidance of J.Y. F.W. polished the paper. J.H. gave suggestions on the design of experiments and improved the paper. J.Y. supervised this work. All authors approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-024-00948-x>.

Correspondence and requests for materials should be addressed to Jianhua Yao.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2024