https://doi.org/10.1038/s41580-024-00756-6

Harnessing the deep learning power of foundation models in single-cell omics

Qin Ma, Yi Jiang, Hao Cheng & Dong Xu

Foundation models hold great promise for analyzing single-cell omics data, yet various challenges remain that require further advancements. In this Comment, we discuss the progress, limitations and best practices in applying foundation models to interrogate data and improve downstream tasks in single-cell omics.

Single-cell technologies generate vast amounts of omics data; 'foundation models'[1] emerge as a powerful tool for in-depth analysis and interpretation of such big data. Foundation models typically apply a self-supervised pre-training strategy on many datasets with numerous parameters, necessitating considerable computational resources and showcasing emergent inference capabilities[1], which allow them to adapt to a wide range of downstream biological tasks. Foundation models excel owing to their expressivity, scalability, multimodality, memory capacity and generalization, which enable their effective use in artificial intelligence applications and promise advances in single-cell omics for molecular biological research.

Although single-cell omics encompass rapidly advancing technologies at the forefront of studying the detailed molecular characteristics of individual cells[2], single-cell data present persistent challenges, such as data enormity, inadequate annotations and complexity of biological interpretation. Foundation models have several uses in single-cell omics data analyses[3], for example: (i) improving signal-to-noise ratios of large-scale and noisy single-cell data and reducing technical batch effects; (ii) most foundation models use self-supervised training strategies, which do not depend on manual annotations, enabling efficient utilization of distribution variations among cell populations and addressing the challenge of inadequate annotations in single-cell data; (iii) multimodal integration and extensive memory capacity enable foundation models to understand intrinsic single-cell data complexity, such as multi-omics data integration and cell functions[2]; and (iv) foundation models leverage large-scale data and numerous cell-state observations to embed essential biological insights into an initial training phase. This embedded knowledge enables foundation models to be broadly adapted to new biological inquiries without further training, known as zero-shot learning. For example, zero-shot learning allows models to predict new drug responses in different cell types using diverse omics data from initial training, without specific drug response training. Alternatively, these models can be fine-tuned with minimal additional training that is specific to a new task[4].

Best practices in developing foundation models for analyzing single-cell omics data

The two main methodologies of using foundation models for single-cell omics analysis are models pre-trained on single-cell data and natural language processing models. Models pre-trained on single-cell data involve an initial pre-training stage[5], in which they learn to interpret and recognize patterns in massive single-cell datasets, and a subsequent fine-tuning stage, which helps the model further align with specific downstream tasks, such as cell-type annotations, gene interactions analyses and cell-state classifications[6]. Additionally, these models can be used for zero-shot learning.

By contrast, natural language processing models capitalize on the power of existing large language models (LLMs)[7] and forgo the extensive training stages. By leveraging advanced techniques in prompt engineering, such as zero-shot prompting[8], the exceptional capabilities of LLMs in question-answering and summarization are fully utilized in single-cell omics data analysis. Following substantial initial investments in pre-training carried out by some organizations, such as OpenAI, the natural language processing models markedly reduced the time and resources needed for model adaptations to new applications. The effectiveness of both models pre-trained on single-cell data and natural language processing models can be assessed with a benchmarking system[9], in terms of the accuracy and interpretability of the model and its capacity for new biological discoveries (for example, discovery of gene expression programmes)[4]. Supplementary Table 1 provides a comprehensive overview of best-practice foundation models specifically designed for single-cell omics data analysis, including the types of models, pre-training tasks and downstream tasks and the biological insights they can provide.

Limitations of foundation models in single-cell data analysis and possible solutions

First, the substantial data and computational resources required for training foundation models restrict the accessibility and scalability of the models in various biological settings, such as cross-species[10] and cross-modality[9] integration. Second, many foundation models face challenges with interpretability, as they update all parameters simultaneously during training, making it difficult to determine how inputs from single-cell omics specifically influence particular parts of the parameters of the model, thereby affecting the final prediction, for example, cell-type annotation. Third, a crucial concern is the robustness of these models, as their performance can vary considerably owing to factors such as pre-training and fine-tuning of noisy data, as well as parameter settings and training depth. Owing to these limitations, current single-cell foundation models often have limited reliability in zero-shot settings and they might not outperform well-designed methods trained on unique datasets[8]. This underperformance suggests that the training sample sizes and training time of these models may not reach the thresholds for obtaining status of bona fide foundation models, that is, of emerging intelligence in solving zero-shot problems.

Some possible solutions have been explored to address the above limitations. High resource demands can be addressed by developing more efficient training algorithms, using open-source foundation

models and leveraging cloud computing resources. Enhancing model interpretability is another area of focus, investing efforts to explain how models make decisions in the prediction process. For instance, feature importance analysis highlights which features most influence model predictions[11]. To further improve data interpretability and enhance the understanding of the rationale behind predictions, researchers are exploring new interpretative algorithms; for example, scGPT used attention-based mechanisms and in silico perturbation methods to identify key genes for classifying cell states[4]. Lastly, improving the robustness of models can be achieved by using diverse training datasets and incorporating learning strategies specifically designed to handle out-of-distribution data. For example, Geneformer promoted robustness to batch effects and individual variability through hundreds of experimental datasets by applying a transfer learning strategy[5]. By addressing these limitations, the potential of foundation models in single-cell research can be expanded to bridge the gap between computational power and biological application, thereby providing more opportunities for integration of the current extensive single-cell datasets and their use in single-cell biology.

Future prospects and applications of foundation models

Future development of the structure design and training of foundation models, along with enhancements in interpretability and the integration of multimodal data, is poised to considerably enhance our analysis and understanding of complex biological systems. The use of cross-species and cross-patient datasets with large-scale trainable parameters can foster 'emergent abilities' in foundation models, which will enhance their generative and interactive capabilities and increase the accuracy of their predictions when using limited data, including in the discovery and development of novel drugs[12] and identification of rare cell populations[13]. For example, rare cell populations that were never annotated in training data can be detected automatically by foundation models owing to their different gene expression distributions[4]. In practical applications, scaling up datasets and parameters is costly and requires considerable hardware, which in turn necessitates more efficient methods for driving LLMs. Graph-based foundation models present a promising solution for upscaling. Graph-based models, which are inherently suited to the structure of single-cell omics data, excel in capturing cellular heterogeneity and molecular patterns[14]. Once the 'over-smoothing' problem of graph-based models is soundly solved, the potential integration of graph-based models and foundation models with large-scale parameters can considerably enhance single-cell omics data analysis. Graph foundation models show promise in constructing biological knowledge graphs from sparse, highly heterogeneous data, enabling effective analyses of cell–cell relationships, gene regulatory networks and single-cell spatial relationships[4].

Multimodal models are designed to process and understand multiple types of data inputs, or 'modalities'. An example is ChatGPT-4 (ref. 7), a large multimodal model capable of processing image and text inputs and producing text outputs. Single-cell omics datasets that combine double-staining histology images, electronic health records and more[15] are treasures for foundation models to explore. By integrating these modalities, foundation models can learn from diverse data types simultaneously, leading to a comprehensive analysis of cellular taxonomies and fundamental gene regulation mechanisms[2]. Specifically, electronic health record data, which comprise clinical and epidemiological data and demographic profiles, among others, form a multimodal foundation model with potential for precise clinical diagnosis and advanced treatments[15].

Following the training stage, adapting foundation models is key to transferring knowledge to specific tasks aiming to enhance performance in molecular and cell biology applications. These applications include cell-type annotation, removal of batch effects, multi-omics integration, genetic perturbation prediction, inference of gene regulation networks, drug discovery, prediction of candidate therapeutics and explaining pathogenic mechanisms[4]. Alternative strategies to fine-tuning foundation models with additional training sets, such as zero-shot prompting[8], are being developed to enable model predictions on smaller, task-specific datasets unseen during training. By applying this approach, single-cell omics could be used to identify novel disease biomarkers[5]. Finally, well-designed interpretable models precisely explore biological systems and provide explanations. For instance, they can identify essential molecular or cellular variables that influence treatment outcomes[4]. Thus, interpretable foundation models can offer insights into disease progression or cell differentiation, and into their regulatory mechanisms from an explainable standpoint. Ultimately, these advancements promise to deepen our understanding of single-cell biology and connect advanced computational methods and biomedical applications[2].

Qin Ma [1,2] ✉, **Yi Jiang**[1], **Hao Cheng**[1] **& Dong Xu** [3]

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. [2]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [3]Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.

✉e-mail: qin.ma@osumc.edu

Published online: 26 June 2024

References

1. Bommasani, R. et al. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Adv. Neural Inf. Process. Syst.* **35**, 3663–3678 (2022).
2. Baysoy, A. et al. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
3. Ma, Q. & Xu, D. Deep learning shapes single-cell data analysis. *Nat. Rev. Mol. Cell Biol.* **23**, 303–304 (2022).
4. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* https://doi.org/10.1038/s41592-024-02201-0 (2024).
5. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
6. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
7. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
8. Wang, W. et al. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.* **10**, 1–37 (2019).
9. Liu, T. et al. 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).
10. Rosen, Y. et al. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* https://doi.org/10.1038/s41592-024-02191-z (2024).
11. Janizek, J. D. et al. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. *Nat. Biomed. Eng.* **7**, 811–829 (2023).
12. Van de Sande, B. et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nat. Rev. Drug Discov.* **22**, 496–520 (2023).
13. Wang, X. et al. MarsGT: Multi-omics analysis for rare population inference using single-cell graph transformer. *Nat. Commun.* **15**, 338 (2024).
14. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
15. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Molecular Cell Biology* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41580-024-00756-6.

Related links

OpenAI: https://openai.com/