

Supplementary information

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

In the format provided by
the authors and unedited

Supplementary Table 1. **Examples and best practices of foundation models developed for the analysis of single-cell omics data.**

Category	Tool	Model type	Pre-training task	Downstream task	Biological highlights
Models pre-trained on single-cell data	scBERT ¹	BERT ²	• Use 209 human single-cell datasets comprising 74 tissues with 1M+ cells with 5 million parameters • Masked-based pretraining • Masking expression values and reconstructing them	• Cell type annotation • Novel cell type discovery	• Provide robust and accurate cell type annotations with gene-level interpretability • The power of applying self-supervised learning on large-scale unlabeled scRNA-seq data to improve the model's generalizability and overcome the batch effect
	Geneformer ³	Transformer ⁴	• Use Genecorpus-30M-29.9 million human single-cell transcriptome datasets with 40 million parameters • Masked-based pretraining • Predicting masked gene position in expression abundance ranking	• Cell type annotation • Gene function prediction • Perturbation prediction	• Have ability to predict dosage-sensitive disease genes through the context-aware <i>in silico</i> deletion approach represents an asset for interpreting genetic variants • Gain a fundamental understanding of network dynamics through the observation of cell states during the pretraining process
	scGPT ⁵	GPT ⁶	• Collect over 33 million human normal cells (involving duplicate cells) covering 51 tissues and 441 studies with 50 million parameters • Prompt-based (gene prompts and cell prompts) and mask-based (attention mask) pretraining	• Batch effect correction • Cell-type annotation • Multi-omics data integration • Imputation • Gene function prediction • Perturbation prediction • Gene network analysis	• The model's ability to accurately capture gene-gene interactions and reflect known biological knowledge provides confidence in its capacity to uncover meaningful insights in single-cell biology • The pre-trained model's knowledge can be transferred to multiple downstream tasks through fine-tuning
	tGPT ⁷	Transformer	• Use 22.3 million single-cell transcriptomes • An unsupervised generative pretraining task: Maximize the likelihood of genes	• Single-cell clustering • Inference of developmental lineage • Batch effect correction • Interpretation of feature representation of bulk tissues	• This model can be applied to bulk tissue sequencing samples to extract a variety of features exhibiting significant association with genomic alterations and response to immunotherapy treatment • The independence of tGPT on batch information makes it attractive for integration of super large-scale transcriptomes
Natural language processing models	GPTCelltype ⁸	GPT-4 ⁹	• Without pretraining process • With optional help from human experts to fine-tune GPT-4-generated annotations	• Cell type annotation	• GPTCelltype can be employed as a dependable tool for automated cell-type annotation of single-cell RNA-seq data, substantially reducing the time and effort required for manual annotation

					• GPTCelltype also provides the option to generate prompts that can be directly fed into GPT-4's online user interface
--	--	--	--	--	--

These foundation models represent significant advancements in accurate cell population identification, gene functionality classification, single-cell drug response classification, dosage-sensitive disease gene prediction, *etc.* The continued evolution and refinement of foundation models promise not only to enhance further capability in downstream tasks, generalization ability, biological insight discovery, interpretability, and resource efficiency, but also to broaden our understanding of complex biological systems significantly.

- **Models pre-trained on single-cell data:** These models have a comprehensive two-step training methodology. Initially, the model undergoes pretraining, exposing it to a vast and varied single-cell dataset. This process sets the foundational parameters of the model, equipping it with a broad and generalized understanding of data patterns and structures. Following this, the model enters the fine-tuning stage. Here, the focus shifts to a more specific dataset, closely aligned with a targeted task or domain. The fine-tuning process adjusts the pre-trained parameters, optimizing the model for enhanced performance in a specific application. These two-stage models can balance general data comprehension and specialized task proficiency, enhancing effectiveness and accuracy in specific applications.
- **Natural language processing models:** These approaches diverge from traditional methodologies without extensive training stages. Instead, it utilizes the inherent capabilities of pre-trained large language models (LLMs). By employing prompt engineering, these models can effectively direct the general-purpose linguistic understanding of LLMs toward accomplishing specific tasks with enhanced efficiency. LLMs prompting-based models demonstrate that adaptable Artificial Intelligence can improve the analysis of single-cell omics data, maximize resource utilization, and minimize developmental redundancies. This strategy not only accelerates the deployment of single-cell-specific tasks but also opens new horizons for AI applications in areas previously constrained by resource limitations.
- **scBERT**, abbreviated from single-cell bidirectional encoder representations from transformers, is a pre-trained deep neural network-based model to annotate cell types for single-cell RNA-seq with generalization and robustness. Besides, it gains a general understanding of gene-gene interactions by being trained on huge amounts of unlabeled scRNA-seq data.
- **Geneformer** is developed and pre-trained to predict gene-network biology from limited gene-expression data. Geneformer can be tuned for many downstream applications to accelerate the discovery of key gene-network regulators and candidate therapeutic targets.
- **scGPT** is a single-cell foundation model based on the generative pre-trained transformer, across a repository of over 33 million cells (involving duplicate cells), which effectively captures meaningful biological insights into genes and cells. Specifically, several versions of the scGPT model, pre-trained on different datasets, are provided, including a whole-human model (pre-trained on 33 million normal human cells), a brain model (pre-trained on 13.2 million brain cells), a blood model (pre-trained on 10.3 million blood and bone marrow cells), a heart model (pre-trained on 1.8 million heart cells), a lung model (pre-trained on 2.1 million lung cells), a kidney model (pre-trained on 814 thousand kidney cells), and a pan-cancer model (pre-trained on 5.7 million cells of various cancer types). The models can be fine-tuned to achieve state-of-the-art performance across various downstream tasks, including multi-batch integration, multi-omic integration, cell-type annotation, genetic perturbation prediction, and gene network inference.

- **tGPT** is a generative pre-training model from transcriptomes for learning feature representation of transcriptomes. tGPT is trained with 22.3 million single-cell transcriptomes, and the work used four single-cell datasets to evaluate its performance on single-cell analysis tasks. Notably, single-cell clusters and cell lineage trajectories derived from tGPT are highly aligned with known cell labels and states.
- **GPTCelltype** is an open-source R software package to facilitate cell type annotation by GPT-4, and it can automatically and accurately annotate cell types by utilizing marker gene information generated from standard single-cell RNA-seq analysis pipelines. Furthermore, it is evaluated across hundreds of tissue types and cell types. GPT-4 generates cell type annotations exhibiting strong concordance with manual annotations and has the potential to considerably reduce the effort and expertise needed in cell type annotation.
- **Transformer** model is a groundbreaking innovation in natural language processing, introduced by Vaswani et al.⁴ It departs from previous models by using self-attention mechanisms, allowing it to consider the relationship between a word and all other words in a sentence. This approach significantly improves efficiency and accuracy in tasks like translation and content generation, making it a basic technology for advanced AI language models like ChatGPT. The model type of foundation model is marked with transformer labels, indicating that the architecture is constructed from basic transformer blocks.
- **Bidirectional encoder representations from transformer (BERT)** is an advanced model in natural language processing, developed by Google. Its innovative design lies in the use of bidirectional training of the Transformer. Unlike previous models that process words in one direction (either left-to-right or right-to-left), BERT reads the entire sequence of words simultaneously. This mechanism allows it to understand context more effectively, leading to significant improvements in tasks like language understanding and sentence interpretation.
- The **GPT** (generative pretrained transformer) model is an advanced AI developed by OpenAI, primarily focusing on language generation. Unlike BERT, designed for understanding language (like answering questions or filling in blanks), GPT is built to generate text. It does this by predicting the next word in a sequence, learning from a vast text dataset. This structure allows GPT to create coherent and contextually relevant paragraphs, simulating human-like writing. GPT has a series of versions, like GPT-2, GPT-3, GPT-4 and GPT-4o. In detail, GPT-2, known for its significance in language understanding and generation, set a new standard in AI's ability to create coherent and contextually relevant text; GPT-3 expanded on this with a much larger dataset and improved algorithms, resulting in more nuanced and accurate responses. GPT-4 and GPT-4o, the latest iterations, further refines these capabilities, offering even more sophisticated language processing, understanding, and generative abilities.
- **LLMs**⁹ are deep learning algorithms, such as the mentioned GPT and BERT, that can perform a variety of natural language processing (NLP) tasks, LLMs are trained on extensive datasets, encompassing diverse text sources, to grasp language patterns and context. LLMs excel in tasks like translation, summarization, question-answering, and creative writing. Unlike simpler models, LLMs can handle complex language nuances, offering more accurate and contextually relevant outputs. The size and training depth enable various applications in natural language processing.
- **ChatGPT**, developed by OpenAI, is a variant of the GPT model, specifically aligned for conversational contexts. It excels in generating human-like text, making it adept at engaging in dialogue, answering questions, and providing information or creative content. ChatGPT's training involves learning from a diverse range of internet texts, enabling it to understand and respond to various topics.
- The **mask pretraining strategy** is a fundamental technique used in training some language models, most notably BERT. In this approach, some input data is intentionally obscured or "masked" during the training process. For example, in a given sentence, certain words are

hidden from the model. The model's task is to predict these masked words based only on the context of the unmasked words.

- The **prompt-based fine-tuning strategy**¹⁰ enhances the performance of large language models like GPT. This approach involves presenting the model with carefully crafted prompts and tunable adaptors, which guide the model in generating specific types of responses or content. Unlike traditional fine-tuning, which may require retraining the model with a large dataset, prompt-based fine-tuning leverages the existing knowledge and structure of the model. By providing contextually relevant prompts or training small adaptors, it can effectively produce outputs that align more closely with the desired outcome¹¹, improving its accuracy and relevance in specific applications without extensive retraining.
- **Evaluation benchmarking system:** The performance of both model categories can be evaluated using the scEval¹² benchmarking system, which utilizes numerous single-cell tasks to test the performance of foundation models, including batch effect correction, multi-omics data Integration, and cell type annotation. Notably, in the evaluating benchmarking system, scGPT⁵ is the most versatile foundation model and Geneformer³ achieves the highest performance in most single-cell tasks. Furthermore, it is crucial to offer biological insights and interpretability for evaluating the ability of foundation models, such as regulatory mechanisms in specific cell types¹².

References

- 1 Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).
- 2 Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **1**, 4171–4186 (2019).
- 3 Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
- 4 Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems* **30**, 6000–6010 (2017).
- 5 Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* <https://doi.org/10.1038/s41592-024-02201-0> (2024).
- 6 Brown, T. et al. Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).
- 7 Shen, H. et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).
- 8 Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods* <https://doi.org/10.1038/s41592-024-02235-4> (2024).
- 9 Thirunavukarasu, A. J. et al. Large language models in medicine. *Nature Medicine* **29**, 1930–1940 (2023).
- 10 Shi, Z. & Lipani, A. Don't stop pretraining? Make prompt-based fine-tuning powerful learner. *Advances in Neural Information Processing Systems* **36**, 5827–5849 (2024).
- 11 Kedzierska, K. Z. et al. Assessing the limits of zero-shot foundation models in single-cell biology. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.10.16.561085> (2023).
- 12 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).