Received: 10 November 2024 | Accepted: 30 April 2025

PERSPECTIVE

Quantitative Biology
Open Access

A perspective on developing foundation models for analyzing spatial transcriptomic data

Tianyu Liu[1,2] | Minsheng Hao[3] | Xinhao Liu[4] | Hongyu Zhao[1,2]

[1]Interdepartmental Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA

[2]Department of Biostatistics, Yale University, New Haven, Connecticut, USA

[3]Research and Early Development, Genentech, South San Francisco, California, USA

[4]Department of Computer Science, Princeton University, Princeton, New Jersey, USA

Correspondence
Hongyu Zhao.

Present address
Minsheng Hao, Department of Biomedical Informatics, Harvard University, Boston, Massachusetts, USA.

Funding information
Foundation for the National Institutes of Health, Grant/Award Number: U01 HG013840

Abstract

Do we need a foundation model (FM) for spatial transcriptomic analysis? To answer this question, we prepared this perspective as a primer. We first review the current progress of developing FMs for modeling spatial transcriptomic data and then discuss possible tasks that can be addressed by FMs. Finally, we explore future directions of developing such models for understanding spatial transcriptomics by describing both opportunities and challenges. In particular, we expect that a successful FM should boost research productivity, increase novel biological discoveries, and provide user-friendly access.

KEYWORDS
artificial intelligence, foundation models, perspective, spatial transcriptomics data

1 | INTRODUCTION

Foundation models (FMs) [1], within the scope of deep learning, are defined as models pretrained with large-scale corpus data and can be utilized to address various problems (known as downstream applications) [2–4]. Such models have achieved significant success in accelerating scientific discoveries, especially in natural language processing (NLP) handled by large language models (LLMs) [2] as well as multimodal data processing handled by large multimodal models (LMMs) [3]. However, in the area of genomics, especially in spatial transcriptomic data analysis, we have yet to find FMs that are capable of generating novel and validated discoveries.

The success of FMs in modeling natural language and multimodal data largely benefits from the advancement of a model architecture, known as the transformer [5], which is capable of capturing the information in sequence data efficiently. Additionally, training FMs in these two areas is also supported by carefully selecting high-quality training data [6, 7]. Therefore, we should expect to transfer the success of FMs in the NLP area to biomedical analysis if the biomedical data share a similar structure and quality with natural language data. However, spatial transcriptomics (ST) contains two sets of information: the gene expression information as well as the coordinates of spots, which do not have an explicit sequence-like data structure and are also noisy [8, 9]. Such problems also exist in building FMs for analyzing single-cell transcriptomics data, which does not have the spatial information for each cell. For single-cell data, researchers have developed single-cell FMs by pretraining the base model with large-scale single-cell transcriptomics data [10–12]. However, it has been shown [13–16] that the performance of single-cell FMs is not proportional to their consumed resources and may not be defined as FMs.

In this perspective, we summarize the current progress on developing FMs for analyzing ST data. We propose problems that align with the requirement of developing FMs and discuss their future developments.

2 | DEFINING A FM

FMs for ST analysis can be classified into two different types: one type is driven by sequencing data (seq-based FMs), and another type is driven by prior biological knowledge (knowledge-based FMs). These two types of FMs have shared tasks and training frameworks but differ in ideas, resource consumption, and training objectives. The two different paradigms are summarized in Figure 1.

Although we can use scRNA-seq-based FMs to address the research questions in ST data, there are significant differences between these two types of FMs. First, scRNA-seq data and ST data have different resolutions and distributions [17]. The noise of ST data can be caused by factors such as sequencing errors, mismatched niche annotation, unmeasured important genes, wrong cell segmentation, and other factors. Therefore, the preprocessing and data-cleaning approaches used for scRNA-seq data cannot be directly transferred to ST data. Improving sequencing technology, cell segmentation method, and imputation approach might reduce the noise level of ST data. Second, ST data have more information than scRNA-seq, such as spatial location, and it is necessary to consider spatial context at least for certain cell types [18] to build a successful FM. Third, ST data have a higher cost than scRNA-seq data, and thus building an FM for analyzing ST data requires a higher standard for constructing the training corpus [19]. Therefore, we need to consider a new paradigm for building FMs.

For seq-based FMs, researchers pretrained the model with large-scale ST data and utilized the pretrained model to handle various downstream applications with either fine-tuning or zero-shot learning. To train a seq-based FM, researchers utilize sequencing data with location information, and the training objective is usually self-supervised learning [20]. As an example, NicheCompass [21] is an autoencoder pretrained with large-scale ST data. During the training process, the known cell–cell interactions are transformed into covariate embeddings to guide the model's learning. Here, a "niche" is defined as a group of spots co-localized by following certain geometric rules. NicheCompass is capable of various downstream applications, including spatial atlas building, niche identification, niche classification, dataset-specific cell–cell communication

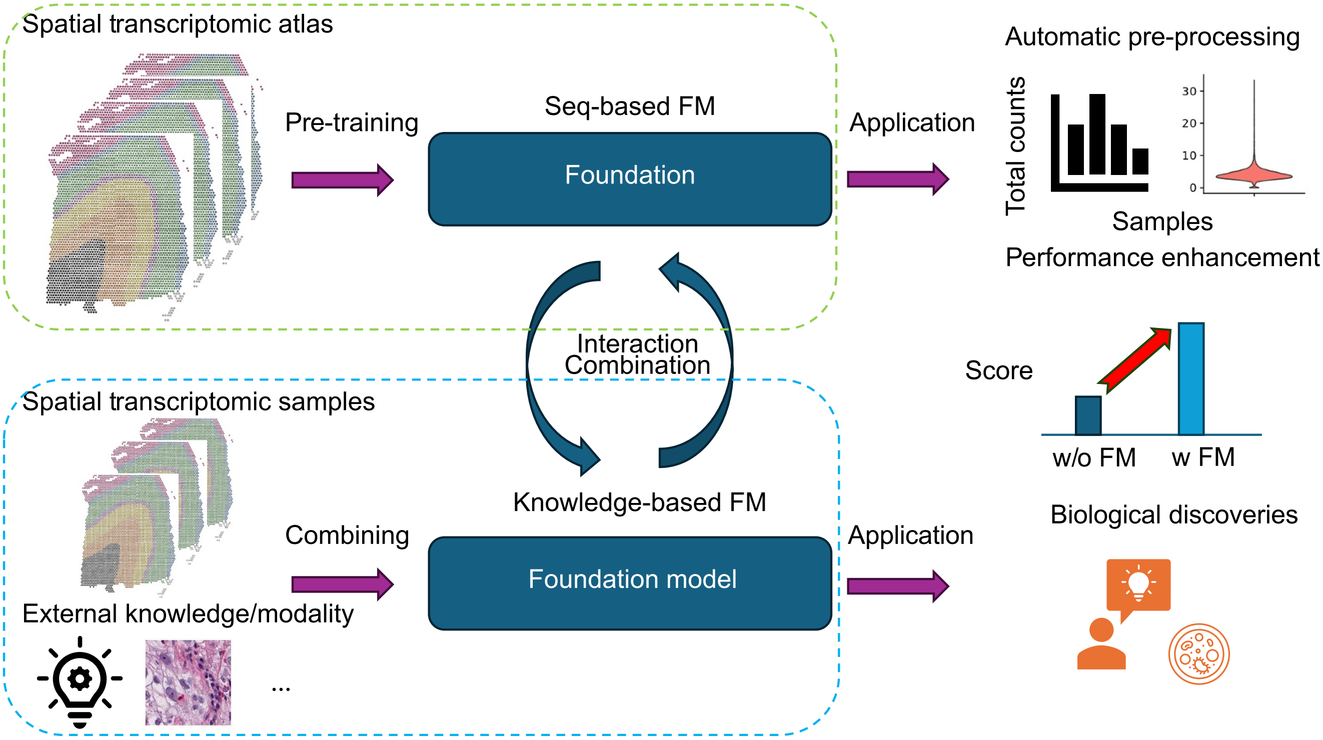

FIGURE 1 Illustration of two paradigms for building FMs for spatial transcriptomic analysis. The seq-based FM is highlighted in the green block, and the knowledge-based FM is highlighted in the blue block. We also include examples of preprocessing steps and contributions of these FMs in the right panel. FMs, foundation models.

inference, and niche query. Similarly, Novae [22] models spatial data with a graph-aware encoder, but its training loss relies on contrastive learning. Another FM, Nicheformer [23], models spatial coordinates as a different modality separate from gene expression and defines the spatial modality as a token in the pretraining process of their Nicheformer transformer blocks with diverse ST data. Through fine-tuning, Nicheformer is capable of spatial label identification, spatial composition deconvolution, and predicting changes in cellular neighborhood density. Moreover, STFormer [24] models niche-level information with gene expression information jointly and proposes the recovery of gene–gene interaction as a new creative pretraining objective. SpaFormer [25] proposes a new transformer-based autoencoder framework for imputing ST data. Lastly, considering the similarity between single-cell transcriptomics and ST, CellPLM [26] also treats cells as tokens and encodes spatial coordinates with positional encoding in their pretraining stage. It utilizes ST as data augmentation of single-cell transcriptomics to enhance its function in batch effect correction, cell-type annotation, perturbation prediction, and spatial imputation. Overall, seq-based FMs leverage the advantages of large-scale data and advanced model architecture for model development.

For knowledge-based FMs, researchers incorporated LLMs or LMMs for processing ST data, as LLMs could be pretrained with text-based biological knowledge [27], and LMMs could be pretrained with biomedical images [28]. Developing knowledge-based FMs requires data from different modalities but might not necessarily pretrain a new base model or fine-tune a model for downstream applications. As an example, QuST-LLM [29] combines QuST as a pioneering tool for processing images and LLM as an advanced tool for processing text-based gene pathway information. Such a design can help researchers uncover spatial insights at the niche level. Moreover, Geneverse [30] considers fine-tuning existing LMMs with expression patterns of genes in ST data to identify marker genes of different cell types. Overall, knowledge-based FMs leverage the advantages of general FMs designed for other purpose and transfer the knowledge to analyze ST.

The interaction between seq-based FM and knowledge-based FM is also emerging, as knowledge can be interpreted as a new modality in the model training stage and thus the capacity of these FMs is likely to be enhanced. For example, spEMO [31] leverages pathology FMs to encode image embeddings with transcriptomic embedding jointly, and thus we can also learn the contextual information from other modalities to analyze sequencing data and perform survival prediction and disease-state classification. scGPT-spatial [32] is developed based on the pretrained scGPT with single-cell transcriptomics data, which has been shown to have promising results in data integration, deconvolution, and gene expression imputation. Therefore, the combination of these two types of FMs might play an important role in building the technical routes for modeling atlas-level transcriptomic data.

3 | DISCUSSING THE NEED OF FM

Given the cost of training and the expectation of the capabilities of FMs, we believe that the capacity of an FM should be stronger than handling simple tasks [13, 33, 34], such as spot-level annotation or clustering on data with simple structure or low annotation resolution [35]. Instead, it should focus on addressing the major limitations of current methods, for example, identifying noncontinuous spatial domains and accelerating novel biological discovery. By incorporating the general perspective of developing an FM, a suitable FM should be utilized for handling the following tasks: it either can help perform automatic selection of the optimal preprocessing steps or provide interpretation for driving novel biology discoveries.

The data preprocessing steps have been extensively studied by researchers [36]. Because of the noise level of ST data, the conclusions of analyzing such data are always strongly affected by the choices of quality control, data normalization, integration, clustering, and annotation [37, 38]. Currently, we are guided by empirical approaches for the preprocessing steps, and the choices made by different researchers are subjective. Therefore, a good FM should unify the preprocessing stage of different ST data and select the best option for each step, which can be performed by data integration or harmonized cell-type (or niche-type) annotation. Such a design can not only guarantee the best utilization of known datasets but also ensure reproducibility.

Using prior knowledge from FMs to enhance the performances of domain-specific models is also important. In the analyses of single-cell transcriptomics, methods such as GenePT [39] and scELMo [40] have demonstrated that using text embeddings from LLMs can improve the performance of current models in several downstream tasks, including cell-type annotation, drug response prediction, and perturbation prediction. Furthermore, scFoundation [11] also showed that incorporating cell embeddings or gene embeddings from pretrained single-cell FMs can help with downstream tasks. However, we have not seen much exploration of the applications related to ST, such as spatial developmental biology [41] or spatial tumor microenvironment [42]. For spatial data, we believe that there are several opportunities for FMs to improve the performance. For instance, in the cell-type annotation task, FM could be fine-tuned to improve the classification accuracy, usually measured by the F1 score. In the spatial niche clustering task, FM could generate embeddings for niches and improve the adjusted mutual information (AMI) score or average silhouette width (ASW) [43]. In the spatial gene imputation task, FM could enhance the signal-to-noise ratio. Common evaluation metrics are correlation and cosine similarity between predicted and ground truth expression. Spatial deconvolution is another common downstream task in ST, where FM is able to improve the accuracy score, including Pearson correlation, structural similarity index (SSIM), root mean square error (RMSE), and Jensen–Shannon divergence (JS) [44]. Meanwhile, how to incorporate multimodal information from FMs to boost current design presents a promising area for future research.

Using AI models to accelerate the process of biological discovery is another possible major contribution. Biological experiments consume more resources than in silico experiments. This is a direct motivation for applying AI models to help with experiment validation [45]. For example, ChemCrow [46] introduced a new AI agent for augmenting the LLM performance in chemistry to help discover new molecules. It would be interesting to have a similar AI agent for processing ST data. FMs can also be used to identify novel cell types or predict perturbation effects. UCE [47] demonstrated an example of identifying new cell types to advocate the necessity of developing single-cell FMs. Specifically, there has been little research to explore perturbation analysis based on ST data or to fully explain the contributions of spatially induced patterns. Therefore, it is important to incorporate the design for novel biological discovery, as well as the comparison between the cost of human resources and model training, during the development of FMs for analyzing ST data.

Finally, we note that these two types of FMs for ST data analysis also have their own limitations. For example, sequencing-based FMs typically require data from diverse resources for training to learn a better contextual representation, and thus data collection as well as storage becomes challenging. Moreover, knowledge-based models might require developers to collect more knowledge-enriched datasets or multimodal data, and biological discovery might be constrained by the prior information. Evaluating the performance of these FMs in real downstream applications is also challenging, as we expect to cover as many cases as possible in the validation step.

4 | EXPLORING THE FUTURE OF FM

In this perspective, we defined the concept of an FM for ST data analysis, followed by possible tasks that need the capacity of FMs to address. However, regarding possible applications, many challenges are presented. First, we should pay more attention to collect high-quality and diverse sequencing-based data and knowledge-based data, which will be the important basis of FM development. Second, the community is still investigating the proper pretraining tasks. Because transcriptomics data do not explicitly contain order information processed by architecture like transformer, the pretraining policy of developing an FM based on sequence data needs to be revisited. For example, the temporal and spatial trajectory [41] could be a new angle for designing a suitable pretraining task or architecture. Moreover, model development should fully consider the design of the benchmarking framework. A good FM should not only perform well in low-level tasks but also inspire researchers to address high-level tasks or even discover new biology. Finally, the budget for computation resources should also be taken into consideration. The cost of GPU hours as well as LLM API should be planned, and it will be helpful to open source a series of models with different scales [48] or an online-access demo [49–51], which can enhance the accessibility of a true FM.

AUTHOR CONTRIBUTIONS

Tianyu Liu: Conceptualization; investigation; resources; writing—original draft; writing—review and editing. **Minsheng Hao**: Resources; writing—original draft; writing—review and editing. **Xinhao Liu**: Resources; writing—original draft; writing—review and editing. **Hongyu Zhao**: Writing—review and editing.

ACKNOWLEDGMENTS

We thank the funding support from NIH grant U01 HG013840.

CONFLICT OF INTEREST STATEMENT

Hongyu Zhao is one of the Editorial Board Members of *Quantitative Biology*. He was excluded from the peer-review process and all editorial decisions related to the acceptance and publication of this article. Peer review was handled independently by the other editors to minimize bias. The remaining authors declare no conflicts of interest.

DATA AVAILABLE STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ETHICS STATEMENT

There is no ethnics issue.

REFERENCES

[1] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021. Preprint at arXiv:210807258.

[2] Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. 2023. Preprint at arXiv:230318223.

[3] Wu J, Gan W, Chen Z, Wan S, Philip SY. Multimodal large language models: a survey. In: 2023 IEEE international conference on big data (BigData). IEEE; 2023. p. 2247–56.

[4] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65.

[5] Vaswani A. Attention is all you need. Adv Neural Inf Process Syst. 2017.

[6] Li M, Zhang Y, Li Z, Chen J, Chen L, Cheng N, et al. From quantity to quality: boosting llm performance with self-guided data selection for instruction tuning. 2023. Preprint at arXiv:230812032.

[7] Wang J, Zhang B, Du Q, Zhang J, Chu D. A survey on data selection for LLM instruction tuning. 2024. Preprint arXiv:240205123.

[8] Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017;45(19): 10978–88.

[9] Liu X, Zeira R, Raphael BJ. Partial alignment of multislice spatially resolved transcriptomics data. Genome Res. 2023;33(7):1124–32.

[10] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multiomics using generative AI. Nat Methods. 2024;21(8):1–11.

[11] Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21(8):1–11.

[12] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24.

[13] Liu T, Li K, Wang Y, Li H, Zhao H. Evaluating the utilities of foundation models in single-cell data analysis. 2023. Preprint at bioRxiv:2023.09.08.555192.

[14] Kedzierska KZ, Crawford L, Amini AP, Lu AX. Assessing the limits of zero-shot foundation models in single-cell biology. 2023. Preprint at bioRxiv:2023.10.16.561085.

[15] Boiarsky R, Singh NM, Buendia A, Getz G, Sontag D. A deep dive into single-cell RNA sequencing foundation models. 2023. Preprint at bioRxiv:2023.10.19.563100.

[16] Ahlmann-Eltze C, Huber W, Anders S. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods. 2024. Preprint at bioRxiv:2024.09.16.613342.

[17] Zhao P, Zhu J, Ma Y, Zhou X. Modeling zero inflation is not necessary for spatial transcriptomics. Genome Biol. 2022;23(1):118.

[18] Dong M, Su D, Kluger H, Fan R, Kluger Y. SIMVI reveals intrinsic and spatial-induced states in spatial omics data. 2024. Preprint at bioRxiv:2023.08.28.554970.

[19] Smith KD, Prince DK, MacDonald JW, Bammler TK, Akilesh S. Challenges and opportunities for the clinical translation of spatial transcriptomics technologies. Glomerular Dis. 2024;4(1):49–63.

[20] Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, et al. Self-supervised learning: generative or contrastive. IEEE Trans Knowl Data Eng. 2021;35(1):857–76.

[21] Birk S, Bonafonte-Pardàs I, Feriz AM, Boxall A, Agirre E, Memi F, et al. Large-scale characterization of cell niches in spatial atlases using bio-inspired graph learning. 2024. Preprint at bioRxiv:2024.02.21.581428.

[22] Blampey Q, Benkirane H, Bercovici N, Andre F, Cournede PH. Novae: a graph-based foundation model for spatial transcriptomics data. 2024. Preprint at bioRxiv:2024.09.09.612009.

[23] Schaar AC, Tejada-Lapuerta A, Palla G, Gutgesell R, Halle L, Minaeva M, et al. Nicheformer: a foundation model for single-cell and spatial omics. 2024. Preprint at bioRxiv:2024.04.15.589472.

[24] Cao S, Yuan Y. A framework for gene representation on spatial transcriptomics. 2024. Preprint at bioRxiv:2024.09.27.615337.

[25] Wen H, Tang W, Jin W, Ding J, Liu R, Dai X, et al. Single cells are spatial tokens: transformers for spatial transcriptomic data imputation. 2023. Preprint at arXiv:230203038.

[26] Wen H, Tang W, Dai X, Ding J, Jin W, Xie Y, et al. CellPLM: pre-training of cell language model beyond single cells. In: The twelfth international conference on learning representations; 2024.

[27] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.

[28] Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024;30(11):1–13.

[29] Huang CH. QuST-LLM: integrating Large Language Models for comprehensive spatial transcriptomics analysis. 2024. Preprint at arXiv:240614307.

[30] Liu T, Xiao Y, Luo X, Xu H, Zheng W, Zhao H. Geneverse: a collection of open-source multimodal Large Language Models for genomic and proteomic research. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the association for computational linguistics: EMNLP 2024. Association for Computational Linguistics; 2024. p. 4819–36.

[31] Liu T, Huang T, Ying R, Zhao H. spEMO: exploring the capacity of foundation models for analyzing spatial multi-omic data. 2025. Preprint at bioRxiv:2025.01.13.632818.

[32] Wang CX, Cui H, Zhang AH, Xie R, Goodarzi H, Wang B. scGPT-spatial: continual pretraining of single-cell foundation model for spatial transcriptomics. 2025. Preprint at bioRxiv:2025.02.05.636714.

[33] Theodoris CV. Perspectives on benchmarking foundation models for network biology. Quantitative Biology. 2024;12(4): 335–8.

[34] Hao M, Wei L, Yang F, Yao J, Theodoris CV, Wang B, et al. Current opinions on large cellular models. Quantitative Biology. 2024;12(4):433–43.

[35] Yuan Z, Zhao F, Lin S, Zhao Y, Yao J, Cui Y, et al. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. Nat Methods. 2024;21(4):712–22.

[36] Ospina O, Soupir A, Fridley BL. A primer on preprocessing, visualization, clustering, and phenotyping of barcode-based spatial transcriptomics data. In: Statistical genomics. Springer; 2023. p. 115–40.

[37] Piñeiro AJ, Houser AE, Ji AL. Research techniques made simple: spatial transcriptomics. J Invest Dermatol. 2022;142(4): 993–1001.

[38] Dries R, Chen J, Del Rossi N, Khan MM, Sistig A, Yuan GC. Advances in spatial transcriptomic data analysis. Genome Res. 2021;31(10):1706–18.

[39] Chen Y, Zou J. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. 2023. Preprint at bioRxiv:2023.10.16.562533.

[40] Liu T, Chen T, Zheng W, Luo X, Zhao H. scELMo: embeddings from Language Models are good learners for single-cell data analysis. 2023. Preprint at bioRxiv:2023.12.07.569910.

[41] Halmos P, Liu X, Gold J, Chen F, Ding L, Raphael BJ. DeST-OT: alignment of spatiotemporal transcriptomics data. In: International conference on research in computational molecular biology. Springer; 2024. p. 434–7.

[42] Chitra U, Arnold BJ, Sarkar H, Ma C, Lopez-Darwin S, Sanno K, et al. Mapping the topography of spatial gene expression with interpretable deep learning. In: International conference on research in computational molecular biology. Springer; 2024. p. 368–71.

[43] Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41–50.

[44] Li B, Zhang W, Guo C, Xu H, Li L, Fang M, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19(6):662–70.

[45] Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, Roed M, et al. How to build the virtual cell with artificial intelligence: priorities and opportunities. 2024. Preprint at arXiv:240911654.

[46] M Bran A, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):1–11.

[47] Rosen Y, Roohani Y, Agrawal A, Samotorcan L, Consortium TS, Quake SR, et al. Universal cell embeddings: a foundation model for cell biology. 2023. Preprint at bioRxiv:2023.11.28.568918.

[48] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: open and efficient foundation language models. 2023. Preprint at arXiv:230213971.

[49] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. 2023. Preprint at arXiv:230308774.

[50] Team G, Anil R, Borgeaud S, Wu Y, Alayrac JB, Yu J, et al. Gemini: a family of highly capable multimodal models. 2023. Preprint at arXiv:231211805.

[51] The Claude 3 Model Family: Opus, Sonnet, Haiku. Available from SEMANTICS SCHOLAR website (CorpusID:268232499).

How to cite this article: Liu T, Hao M, Liu X, Zhao H. A perspective on developing foundation models for analyzing spatial transcriptomic data. Quantitative Biology. 2025;e70010.