Advances in Explainable Models for Single-cell Transcriptomics and Spatial Transcriptomics Data Analysis

Luchang Ge*
Nanjing Agricultural University, Jiangsu 210014, China
*Corresponding email: 10123121@stu.njau.edu.cn

Abstract

Tissue heterogeneity is supported by single-cell sequencing technology, and spatial transcriptomics provides an important technological platform to study the problem, continuous changes in cell state, and microenvironment interactions. Yet the high dimensionality and sparsity of data, batch effects, and spatial point mixing may confuse model inferences. In this review, we discuss recent progress of interpretable model and method developments in single-cell and spatial omics analysis with biological applications, especially for the deep generative model single-cell Variational Inference (scVI) as well as its multimodal extensions total Variational Inference (totalVI) and Multimodal Variational Inference (MultiVI); the spatial mapping method Tangram; and the intercellular communication inference frameworks NicheNet and CellChat. We additionally discuss how pre-trained base models like single-cell Bidirectional Encoder Representations from Transformers (scBERT) and single-cell Generative Pre-trained Transformer (scGPT) can be used for cross-dataset transfer, and the problems this poses to interpretability. This work highlights the construction of an evidential chain centered around “gene/pathway - spatial domain - communication network”, and indicates that spatial colocalization, multimodal consistency, and uncertainty assessment can enhance the confidence and repeatability of mechanistic inference, hence providing a testable hypothesis-generating platform for tumor immunology studies.

Keywords

Single-cell sequencing, Spatial transcriptomic, Interpretability, Deep learning

Introduction

The single-cell transcriptomics and spatial transcriptomics are transforming life science research, shifting the focus away from average tissue signals toward cell-resolved structure, state, and interaction. Single-cell data have enabled studies of tumor immunity, development, inflammation, and organ homeostasis to uncover cellular heterogeneity. Spatial context provides additional information on tissue architecture and the microenvironment to understand the relationships among cell types, kinetics, dynamics, and structure. Yet, with the rapid growth of both data volume and experimental strategies, such integrative analyses are often performed across platforms, populations, and modalities. This trend has rendered the extraction of robust and reproducible biological signals from noise and artifacts a major bottleneck for mechanistic discovery and translational research.

At the level of research methodologies, a relatively comprehensive suite of analytical tools has been developed to date. Probabilistic generative models and representation learning methods greatly improve the stability of cross-batch integration, denoising, and cell state characterization, while multimodal models enable the integrated interpretation of transcriptomic, proteomic, and chromatin accessibility data within a single integrative framework [1]. The proposed spatial mapping and deconvolution strategies mitigate the loss of resolution caused by the mixing of cell types within spatial spots, while cell-cell communication inference methods organize ligand-receptor interactions and signaling pathways into interpretable networks [2]. Recently, pre-trained foundation models have been developed that learn more transferable representations via large-scale self-supervised learning [3]. Overall, biological data analysis has advanced rapidly with the development of tools capable of performing a wide range of analyses, from cell type annotation to spatial domain identification and intercellular interaction network reconstruction.

Yet with ever-increasing methodological complexity and model capacity, interpretability and biological meaning have become increasingly elusive. Multimodal and pretrained models provide stronger representations, but their correspondence to underlying mechanisms is often more indirect [4]. Single-cell transcriptomic data are high-dimensional and sparse, and technical noise and batch effects can be confounded with true biological variation [5]. Spatial transcriptomics data retain tissue structure but suffer from limited capture resolution and spot-level mixing, making it difficult to achieve direct cell-level spatial localization [6]. In addition, cell-cell communication inference relies on a priori databases and expression-based approximation, and is thus highly susceptible to contextual specificities, as well as limitations in spatial validity [7].

From a practical standpoint, single-cell (sc-) and spatial omics models are typically used at the initial stage of a study, providing cross-cohort characterization of cell states, identification of disease-associated cell populations and functional programs, and projecting the resulting data onto tissue structures or interaction networks [8]. Yet, cell clustering may yield contradictory results across varying datasets, experimental conditions, or analytical parameters, and communication inference is constrained to the spatial domain level. Furthermore, statistical results do not always translate readily into testable mechanistic hypotheses. Thus, interpretability is more than an auxiliary feature of such models: It is essential to transform algorithmic outputs into biological knowledge that can be experimentally validated and clinically translated.

Applications of common explainable methods in single-cell and spatial transcriptomics data analysis

In the contemporary era of rapid advances in information technology and sequencing technologies, deep learning-based modeling has become an important component of single-cell sequencing and spatial transcriptomics data analysis. Accordingly, the biological meaning of model outputs should be a central focus. Below is an overview of explainable models in common single-cell and spatial transcriptomics settings.

For classical single-cell transcriptomics tasks such as cross-batch integration and denoising, these processes form a basis to enable downstream interpretation. Deep generative models of the single-cell Variational Inference (scVI) describe counts via an explicit probabilistic generative process ,which represent cell states as latent variables and model technical factors (such as library size) to enable cross-batch alignment in a shared latent space while retaining biological variation [9]. This model has the advantage that it fits better to the generative properties of sequencing data and thus produces a more stable representation in situations with sparsity and noise.

As the study questions lean more heavily towards immune phenotypes or clinically translatable biomarkers, it is possible that transcriptomic information alone will not suffice to provide a fine-grained classification of samples. CITE-seq provides simultaneous measurements of both RNA and surface proteins, which provides more direct molecular evidence to characterize immune sub-population. In a generative modelling context, total Variational Inference (totalVI) jointly models the RNA and protein counts whilst accounting for antibody background noise and measurement differences, that allows both modalities to have a shared latent space [10].

At the regulatory mechanism level, multimodal joint measurements enable moving interpretations away from “correlated genes” towards “regulatory evidence”. Multimodal Variational Inference (MultiVI) is designed for jointly modelling RNA and ATAC data, combining paired or unpaired multimodal data in one common latent space, while also enabling the ability to map between modalities and perform imputation on missing data [11].

Interpretable Interfaces: Probabilistic generative models such as VAEs or factor models can serve as an interface to explain how the data are generated, e.g., through decomposition of variation into biological and technical components, and by reporting uncertainty estimates. In particular, cross-modal factor models such as Multi-Omics Factor Analysis plus (MOFA+) are designed to focus on decomposing common and modality specific factors for interpretability in terms of coordinated change between multiple omics datasets [12]. In contrast, scVI is closer to a model of sequencing mechanics on the distributional scale of counts and thus better suited to separate technical noise from true biology.

For graph-based models, including cell graphs, spatial neighborhood graphs, and communication networks, “structure-level explanations” can also be provided.

Examples include identifying the neighborhood edges most important for distinguishing a spatial domain, or the sender populations most critical for a particular receiver cell state. Graph explanation methods, such as subgraph selection and edge importance scoring, can serve as useful complements [13]. In biological applications, these explanations are more commonly evaluated by checking consistency with known tissue anatomy, pathological regions, or immunohistochemistry evidence [14].

Furthermore, perturbation-based and counterfactual explanations represent a popular approach to model interpretation. These methods simulate gene upregulation or downregulation, mask specific types of neighbourhood information, or remove particular ligand-receptor pairs, then monitor subsequent changes in model predictions or inferred network strength [15]. Compared to Perturb-seq, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based single-cell perturbation datasets and drug treatment data, these methods can significantly enhance the testability of model explanations [16]. In recent years, these concepts have become an increasingly integral part of model assessment tools.

Development of model interpretability in single-cell and spatial transcriptomics

With the advent of spatial transcriptomics as a powerful method for studying tissues at the microenvironmental level, interpretability frameworks have also evolved significantly. The most obvious benefit of spatial data is that molecular signals can be associated with tissue structure, whereas the challenge here is that individual spatial capture spots can contain mixtures of many different cell types such that it becomes challenging to localize cells by type or infer intercellular communication between them. Biancalani et al. proposed matching an scRNA-seq reference atlas to a spatial expression map to project cell types/states onto spatial coordinates [17]. In the typical tumor microenvironment scenario, researchers first leverage scRNA-seq to define fine-grained annotations for immune/stromal cells, and then use Tangram to project these populations back onto tissue sections to identify enrichment patterns of immune cells at the tumor margin, invasive front, perivascular zones, or necrotic areas. Interpretation is usually organized around spatial domains: What cell composition characterizes a given domain? Which gene programs or pathways are enriched there? And whether these patterns are consistent with pathological annotations? The key value of interpretability lies in the fact that spatial localization provides an essential constraint for mechanistic inference. Many differential programs that appear significant in single-cell data may in fact be confined to specific spatial niches. Conversely, some relatively weak expression differences may be more mechanistically informative when they exhibit clear regional organization.

The explanatory targets of cell-cell communication models correspond directly to biological entities, including ligands, receptors, signaling pathways, and sender/receiver cell populations. NicheNet integrates ligand–receptor interactions with downstream regulatory networks to explain differentially expressed gene sets in receiver cells on the basis of candidate ligands from sender cells, while also prioritizing ligands and predicting potential target genes [18]. In typical tumor immunity or inflammation studies, researchers first define, in receiver cells, a differential gene program associated with a spatial domain, disease state, or treatment response, and then use NicheNet to trace back the potential ligand sources and signaling axes driving that program. For example, ligands secreted by tumor cells or myeloid cells may explain enhanced T-cell exhaustion or chemotactic programs. The interpretability advantage of NicheNet is that it provides a mechanistic chain linking sender-derived signals to receiver-cell target genes, thereby generating hypotheses that are closer to experimental testing. When combined with spatial data, researchers can further assess whether ligands and receptors are spatially co-localized or significantly co-occur at the neighborhood level, which strengthens interpretability and reduces biologically implausible inferences.

In contrast to the ligand-to-target-gene orientation of NicheNet, CellChat emphasizes the construction of communication networks at the cell-group level and their aggregation and comparison at the pathway level [19]. In comparisons across disease stages, pre- versus post-treatment conditions, or distinct tissue regions, CellChat is often employed to characterize communication network rewiring. Specifically, this approach identifies which pathways exhibit increased or decreased information flow, which cell populations act as stronger senders or receivers, and how network centrality and modular structure shift. Typical applications include comparisons between tumor core and tumor edge, or between acute inflammatory and remission phases, where researchers can derive relatively interpretable conclusions at the pathway level, such as increased chemotactic or immunosuppressive signaling and enhanced tissue-repair pathways. The interpretability strength of CellChat lies in its compatibility with spatial evidence. When a pathway is inferred to be important, spatial neighborhood analysis can test whether the relevant cell groups are in contact or in close proximity. If protein-level or in situ evidence is further incorporated, the inferred network can move from expression-based association toward a more credible biological explanation.

Recently developed single-cell foundation models have introduced new opportunities for both single-cell and spatial applications, while also raising new interpretability challenges. Models such as single-cell Bidirectional Encoder Representations from Transformers (scBERT) and single-cell Generative Pre-trained Transformer (scGPT) adopt Transformer-based pretraining strategies from natural language processing and perform self-supervised learning on large-scale single-cell expression datasets to obtain transferable general representations for cell-type annotation, cross-dataset generalization, and multi-task prediction [20-22]. Their main advantage is a strong representation learning capability, which enables the capture of richer contextual structures across more complex distributions. From a more interpretability-focused perspective, however, the key challenge is how to translate these black-box representations into biologically meaningful insights. Although foundation models can be used for embedding and prediction, there is usually a need to return to the gene and pathway level to derive mechanistic interpretations: highlighting important gene sets via feature attribution, followed by pathway enrichment and regulatory network analysis.

Toward a mature interpretable closed loop in spatial omics

We thus move toward an interpretable closed loop via the integration of spatial, communication, and foundation models for single-cell and spatial omics data analysis. In terms of such models and application considerations above, we believe that interpretability studies on space omics are heading toward a more mature “space-communication-mechanism” closed loop. sc generation model/multi-omics integrated model could offer strong cell states representation, making it easier to compare across cohorts, and perform differential procedural identification with greater confidence [23]. Spatial mapping and spatial domain discovery ground these processes in tissue structure such that interpretation is consistent with the anatomy and pathology of tissues [24]. Communication models organize communications between cell populations into pathways and networks, allowing for more systematic descriptions of microenvironmental mechanisms [25].

In return, base models may deliver generic representation that is transferable across more varied data distribution and lower the bar of cross-cohort transfers. Parallelly, we are witnessing rising interpretability quality requirements. The communication inference is more concerned about the spatial feasibility constraint and the support of multi-mode evidences, while spatial deconvolution and mapping become more and more driven by agreement with independent markers such as in situ detection, immunohistology, spatial proteomics. Reproducibility and robustness of explanations is becoming increasingly obvious standard in cross-cohort studies, and uncertainty reporting and stability analysis, which we will add to the interpretability framework to separate out strong mechanistic cues from exploratory signals.

Conclusion

In general, biology driven single cell and spatial omics analysis has gone beyond just producing a clustering result or a network diagram, towards building chains of evidence that can be tested. Models are not simply ways to make better predictions or integrate data. Instead, they serve as mediators that condense information-rich datasets to testable mechanistic theories through experiments. Future work includes, but is not limited to, several key research directions for advancing single-cell and spatial omics interpretability. First, it will focus on more accurate inference of cell communications and niches under spatial constraints, and the tighter coupling of multimodal modelling with regulatory priors to bring interpretations closer to causal chains. Second, efforts will be directed at developing better standardized interpretability benchmarks, particularly using datasets that incorporate spatial and perturbation ground truth, as well as leveraging model interpretations directly to inform the design of further experiments. All these advancements together lead towards establishing an iterative closed loop between data, models, and experiments, thus enabling the trustworthy use of single-cell and spatial omics for disease mechanism discovery and translation.

Funding

This work was not supported by any funds.

Acknowledgements

The author would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Chen, C., Liu, X., Zhou, M., Li, Z., Du, Z., Lin, Y. (2025) Lightweight and real-time driver fatigue detection based on MG-YOLOv8 with facial multi-feature fusion. *Journal of Imaging*, 11(11), 385.

[2] Yamamoto, T., Cockburn, K., Greco, V., Kawaguchi, K. (2022) Probing the rules of cell coordination in live tissues by interpretable machine learning based on graph neural networks. *PLOS Computational Biology*, 18(9), e1010477.

[3] Guan, J., Cheng, H. Y., Wu, Y. P., Tian, C., Qi, J. Y. (2025) Multi-target tracking for star sensor based on CenterTrack deep learning model. *Scientific Reports*, 15(1), 37125.

[4] Dohmen, J., Baranovskii, A., Ronen, J., Uyar, B., Franke, V., Akalin, A. (2022) Identifying tumor cells at the single-cell level using machine learning. *Genome Biology*, 23(1), 123.

[5] Kim, Y. S., Choi, J., Lee, S. H. (2023) Single-cell and spatial sequencing application in pathology. *Journal of Pathology and Translational Medicine*, 57(1), 43-51.

[6] Zhi, Y., Wang, Q., Zi, M., Zhang, S., Ge, J., Liu, K., Lu, L., Fan, C., Yan, Q., Shi, L., Chen, P., Fan, S., Liao, Q., Guo, C., Wang, F., Gong, Z., Xiong, W., Zeng, Z. (2024) Spatial transcriptomic and metabolomic landscapes of oral submucous fibrosis-derived Oral squamous cell carcinoma and its tumor microenvironment. *Advanced Science*, 11(12), 2306515.

[7] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., Yosef, N. (2018) Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 1053-1058.

[8] Zhao, M., Huang, H., He, F., Fu, X. (2023) Current insights into the hepatic microenvironment and advances in immunotherapy for hepatocellular carcinoma. *Frontiers in Immunology*, 14, 1188277.

[9] Ning, M., Lu, D., Liang, D., Ren, P. G. (2025) Single-cell RNA sequencing advances in revealing the development and progression of MASH: the identifications and interactions of non-parenchymal cells. *Frontiers in Molecular Biosciences*, 12, 1513993.

[10] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., Yosef, N. (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3), 272-282.

[11] Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G. A., Jordan, M. I., Yosef, N. (2023) MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8), 1222-1231.

[12] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., Stegle, O. (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1), 111.

[13] Hackenberg, M., Brunn, N., Vogel, T., Binder, H. (2025) Infusing structural assumptions into dimensionality reduction for single-cell RNA sequencing data to identify small gene sets. *Communications Biology*, 8(1), 414.

[14] de Lucio Delgado, A., Villegas Rubio, J. A., Riaño-Galán, I., Pérez Gordón, J. (2023) Effect of the use of gnrh analogs in low-grade cerebral glioma. *Children*, 10(1), 115.

[15] Qian, C., Xin, Y., Qi, C., Wang, H., Dong, B. C., Zack, D. J., Blackshaw, S., Hattar, S., Zhou, F. Q., Qian, J. (2024) Intercellular communication atlas reveals Oprm1 as a neuroprotective factor for retinal ganglion cells. *Nature Communications*,

15(1), 2206.

[16] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., Regev, A. (2016) Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7), 1853-1866.

[17] Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg, C. R., Segerstolpe, Å., Zhang, M., Avraham-Davidi, I., Vickovic, S., Nitzan, M., Ma, S., Subramanian, A., Lipinski, M., Buenrostro, J., Brown, N. B., Fanelli, D., Zhuang, X., Macosko, E. Z., Regev, A. (2021) Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18(11), 1352-1362.

[18] Browaeys, R., Saelens, W., Saeys, Y. (2020) NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2), 159-162.

[19] Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C. H., Myung, P., Plikus, M. V., Nie, Q. (2021) Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1), 1088.

[20] Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J. (2022) scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10), 852-866.

[21] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B. (2024) scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8), 1470-1480.

[22] Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., Han, J. D. J. (2023) Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1), 223.

[23] Bhattacharyya, S., Ehsan, S. F., Karacosta, L. G. (2023) Phenotypic maps for precision medicine: a promising systems biology tool for assessing therapy response and resistance at a personalized level. *Frontiers in Network Physiology*, 3, 1256104.

[24] Hu, Y., Lin, Z., Xie, M., Yuan, W., Li, Y., Rao, M., Liu, Y. H., Shen, W., Zhang, L., Zhou, X. M. (2025) MaskGraphene: an advanced framework for interpretable joint representation for multi-slice, multi-condition spatial transcriptomics. *Genome Biology*, 26(1), 1-45.

[25] Hao, G., Fan, Y., Yu, Z., Su, Y., Zhu, H., Wang, F., Chen, X., Yang, Y., Wang, G., Wong, K. C., Li, X. (2025) Topological identification and interpretation for single-cell epigenetic regulation elucidation in multi-tasks using scAGDE. *Nature Communications*, 16(1), 1691.