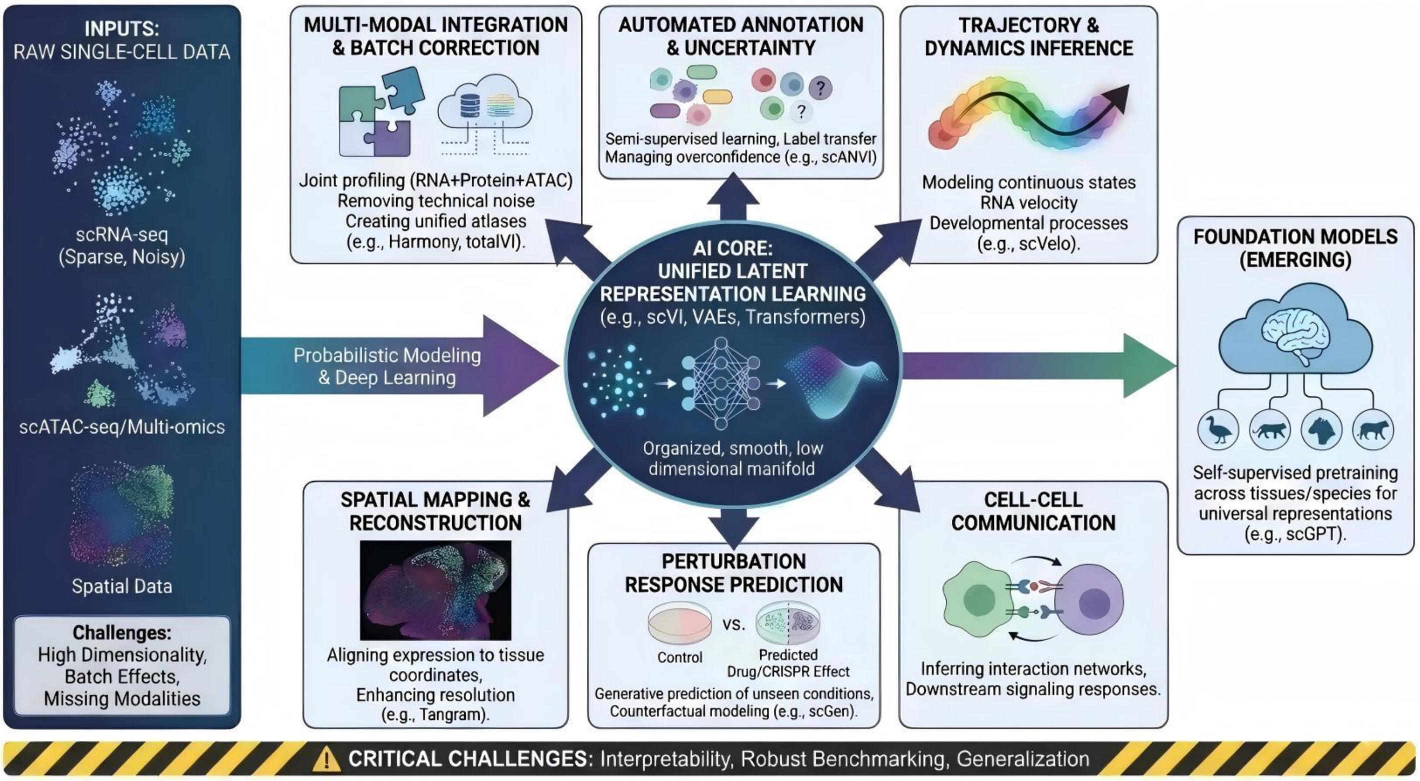
Reviews

Artificial Intelligence for Single-Cell Biology: From Representation Learning to Predictive Modeling

DOI:https://dx.doi.org/10.71373/YAGA2665
Submitted 24 December 2025
Accepted 26 December 2025
Published 31 December 2025

Wenxing Li[1], Yuxiang Luo[2,*]

While single-cell sequencing technologies provide a high-resolution approach for dissecting cellular heterogeneity, their data are inherently high-dimensional, sparse, noisy, and strongly affected by batch effects and limited annotations. These properties make artificial intelligence (AI), particularly deep generative and probabilistic models, more suitable for analyzing single-cell data. Recent AI frameworks, including variational inference–based model scVI and its extensions, have supported unified pipelines for normalization, representation learning, batch correction, multimodal integration, and downstream analyses. Specialized downstream analyses, such as scalable cell-type annotation, trajectory and dynamic inference, cell–cell communication analysis, spatial mapping, and the prediction of genetic or pharmacological perturbation responses, can be extended by learning transferable latent representations. Emerging self-supervised foundation models promise reusable cellular representations across tasks, tissues, and species. Addressing challenges in benchmarking, interpretability, uncertainty quantification, and robust generalization highlights future frontiers in the development of predictive and causal single-cell models.



Why Single-Cell Omics Naturally Requires Artificial Intelligence

Single-cell sequencing technologies, especially scRNA-seq, scATAC-seq, and multi-omics joint profiling, allow researchers to systematically characterize the heterogeneity of tissues and diseases at single-cell resolution. However, these data exhibit different statistical structures from traditional omics: high dimensionality, highly sparse expression matrices, pervasive technical noise and dropout events, strong batch effects and experimental biases, while annotations are scarce and distribution differences across technologies, tissues, and even species are significant. [1–3]

These intrinsic properties limit the performance of linear models or empirical rules. AI-based approaches such as deep learning, probabilistic generative models, and self-supervised learning are therefore better suited for unified representation learning and robust inference. In recent years, the scVI ecosystem, centered on variational inference and deep generative models, has gradually established a new analytical paradigm: viewing single-cell data as a stochastic process jointly generated by latent biological states and technical noise, and performing end-to-end probabilistic modeling to achieve holistic inference

1. Unievrsity College London
2. Department of Systems Biology, Columbia University Irving Medical Center, New York, NY.
**Corresponding author: Yuxiang Luo, Email: zczquoe@ucl.ac.uk*

from raw count data to downstream biological interpretation.[1,4,5]

From Workflow-Based Analysis to Probabilistic Generative Paradigm

Traditional single-cell analysis typically employs a workflow that includes quality control, normalization, selection of highly variable genes, dimensionality reduction, clustering, and differential analysis[6,7]. In the generative modeling framework, these steps are reformulated as modeling different aspects of the data-generating process[1,8].

For example, Seurat uses normalization and variance stabilization based on regularized negative binomial regression have been widely applied to large-scale datasets, providing a reliable data foundation for robust analysis[9,10]. Complementarily, scVI directly models expression distributions at the gene count level, learning a low-dimensional latent space through variational autoencoders, preserving biological variation, separating technical biases, and naturally supporting uncertainty quantification [1,11].

Latent Representation Learning and Multi-Modal Unified Representation

Learning transferable latent representations is one of the core objectives of single-cell AI[1,11]. This latent space is used not only for visualization and clustering but also serves as the foundation for cross-dataset integration, annotation transfer, and dynamic modeling.

In multi-omics scenarios, totalVI achieves end-to-end denoising, integration, and missing modality inference through jointly probabilistically modeling RNA and protein (CITE-seq) data[4]. For multi-modal or mosaic data such as scRNA-seq and scATAC-seq, MultiVI and subsequent frameworks further expand the applicability of joint representation learning, thereby improving the reliability of integrative analysis under partially missing modalities[12,13].

Batch Correction and Cross-Dataset Integration: Toward Reusable Reference Atlases

Batch effect correction is a long-standing challenge in single-cell research. Ideally, technical differences should be removed while preserving true biological variation to the greatest extent possible[11].

To address this challenge, Harmony was proposed to iteratively learn correction terms in a low-dimensional embedding, balancing computational efficiency and integration quality, and has become a commonly used solution for large-scale data integration[11]. Seurat provides an alternative integration strategy based on anchor identification, aligning cell states between datasets, enabling alignment across experimental conditions and sequencing technologies, as well as reference mapping[10,14].

Building on this, scArches introduces transfer learning into generative models, allowing new data to be lightly mapped into existing reference latent spaces, providing a scalable solution for continuously expanding cell atlases[15].

Automated Cell Type Annotation and Uncertainty Management
As single-cell atlases grow in scale, marker-based manual annotation gradually becomes inadequate, and automated outputs with controllable uncertainty become critical[16].

scANVI extends the scVI framework with semi-supervised learning, using partially labeled cells to guide latent space structure, improving annotation quality for unlabeled cells while maintaining probabilistic consistency in the presence of batch effects[5]. scNym further combines semi-supervised learning with domain adaptation and adversarial training, enhancing generalization across experimental conditions and platforms[17].

In practical applications, such models are often combined with hierarchical labeling systems, rare cell recognition, and rejection mechanisms to reduce the risk of overconfident mislabeling[16,17].

Continuous States, Trajectories, and Cellular Dynamics Modeling

Single-cell states often form continuous spectra rather than discrete categories; therefore, trajectory inference and dynamics modeling are crucial for understanding development, activation, and disease progression[18,19].

Manifold learning methods such as UMAP excel at preserving local structure and often serve as the basis for neighborhood graphs and trajectory analysis[18]. RNA velocity, which introduces directional information to static expression data through splicing kinetics, is generalized by scVelo to capture more general transient states and enhances robustness via likelihood-driven dynamic modeling[19].

From an AI perspective, these problems can be formulated as learning continuous-time generative processes or directed graph structures in latent space, naturally compatible with neural ODEs, graph neural networks, and probabilistic state-space models[19,20].

Cell – Cell Communication: From Co-Expression Inference to Mechanistic Links

Cellular communication analysis usually constructs latent interaction networks between cell populations based on ligand–receptor databases as priors[21,22]. CellPhoneDB provides systematic ligand – receptor resources and statistical inference pipelines, while CellChat emphasizes hierarchical features and pattern analysis of communication networks[21,22].

NicheNet advances communication inference from “whether interactions exist” to “downstream transcriptional responses,” predicting ligand–target gene relationships by integrating signal transduction and regulatory network priors, yielding results closer to mechanistic interpretation[23]. With the development of spatial transcriptomics, incorporating spatial proximity constraints and graph models into communication analysis has become an important trend.

Generative Prediction of Per-

turbation Responses and Drug Effects

Single-cell perturbation experiments (CRISPR, drugs, infection, etc.) provide important information for causal mechanism studies, but high experimental costs and enormous combinatorial spaces make perturbation response prediction a key AI task[24,25].

scGen learns latent differential vectors between control and perturbed states through conditional variational autoencoders, enabling extrapolation of perturbation effects across cell types and studies[24]. CellOT integrates optimal transport with neural networks to learn mappings between unpaired distributions for distribution-level counterfactual prediction[25].

These methods have potential value in drug screening and personalized therapy, but require stricter cross-batch, cross-donor, and cross-platform evaluation to avoid overly optimistic generalization estimates[26].

Spatial Information Integration: Placing Cellular States Back into Tissue Coordinates

Spatial transcriptomics provides tissue structural information, but remains limited in resolution and sequencing depth; scRNA-seq, though information-rich, lacks spatial localization[27,28]. Tangram aligns single-cell expression with spatial measurements through deep learning, providing probabilistic mapping from single cells to spatial positions and reconstructing spatial expression patterns of unmeasured genes[27].

Recent studies further introduce graph structures, tissue morphological images, and multi-modal priors to improve spatial deconvolution and cell localization accuracy in complex tissues[28,29].

Single-Cell Foundation Models and Self-Supervised Pretraining With the rapid growth of public single-cell atlases and cell numbers, the field has begun exploring ‘foundation models' based on self-supervised pretraining to learn universal cellular representations for downstream tasks[30–32].

scGPT, inspired by Transformer architectures in natural language processing, treats “gene–cell” analogous to “word–sentence,” pretraining on large-scale datasets and showing potential in annotation, perturbation prediction, and cross-species mapping tasks[32].

Core challenges in this direction include distribution shift, cross-tissue generalization, model interpretability, and data governance and privacy protection[31,35].

Evaluation, Interpretability, and Conclusion Credibility

Single-cell AI analyses are often affected by batch leakage, label inconsistency, and donor overlaps, which may lead to offline evaluation results that overestimate true generalization performance[26,34]. Therefore, the standardized evaluation strategies by donor, batch, or experiment, as well as external validation across tissues and platforms, have become increasingly important[34].

For interpretability, traditional marker and pathway enrichment analysis are now complemented by model-intrinsic approaches, such as attention mechanisms, feature attribution, latent factor constraints, and generative counterfactual experiments, that help convert “black-box representations” into testable biological hypotheses[23,35,36].

Summary and Outlook

Overall, artificial intelligence is advancing single-cell analysis from a “toolbox-style workflow” to a “transferable, generative, and joint-inference” system modeling paradigm, particularly excelling in representation learning, reference mapping, spatial integration, and foundation models[1,4,15,27,30]. Unified modeling of heterogeneous single-cell data drives analysis toward holistic insights into cellular states, dynamics, and interactions. Moreover, these approaches facilitate scalable integration across experiments, tissues, and even species, constructing a foundation for reusable reference atlases and predictive models for perturbation or disease responses.

Key future directions include stricter cross-domain extrapolation evaluation, causal inference–based perturbation prediction, multi-modal models integrating spatial and imaging data, and improved uncertainty quantification and interpretability frameworks to support high-risk applications such as clinical and drug development[25,29,34,37]. In addition, the development of self-supervised foundation models shows potential for broad application in cell representations, thereby reducing reliance on extensive manual annotation and enabling transfer learning across diverse biological contexts. Coupling these models with mechanistic priors and multi-scale spatial information may further enhance their predictive accuracy and biological interpretability. Ultimately, these advances are expected to transform single-cell research from descriptive atlases to predictive, hypothesis-driven frameworks capable of guiding experimental design, precision medicine, and therapeutic discovery.

Reference

1. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature Methods. 2018.
2. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology. 2019.
3. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics. 2017.
4. Gayoso A, Steier Z, Lopez R, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nature Methods. 2021.
5. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Molecular Systems Biology. 2021.
6. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology. 2018.

7. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 2018.
8. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications. 2019.
9. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology. 2019.
10. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019.
11. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods. 2019.
12. Ashuach T, et al. MultiVI: a deep generative model for the integration of multimodal data. Nature Methods. 2023.
13. He Z, et al. Mosaic integration and knowledge transfer of single-cell multimodal data. Nature Biotechnology. 2024.
14. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018.
15. Lotfollahi M, et al. Mapping single-cell data to reference atlases by transfer learning. Nature Biotechnology. 2022.
16. Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. Computational and Structural Biotechnology Journal. 2021.
17. Kimmel JC, et al. Semisupervised adversarial neural networks for single-cell classification. Genome Research. 2021.
18. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology. 2014.
19. Bergen V, Lange M, Peidli S, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. Nature Biotechnology. 2020.
20. La Manno G, et al. RNA velocity of single cells. Nature. 2018.
21. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nature Protocols. 2020.
22. Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell–cell communication using CellChat. Nature Communications. 2021.
23. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nature Methods. 2019.
24. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nature Methods. 2019.
25. Elnahas O, Ead WM, Qiu Y, Lu J. Ensemble machine learning-based pre-trained annotation approach for scRNA-seq data using gradient boosting with genetic optimizer. BMC Bioinformatics. PMC12220795.
26. Lähnemann D, et al. Eleven grand challenges in single-cell data science. Genome Biology. 2020.
27. Biancalani T, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature Methods. 2021.
28. Hu J, et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains. Nature Methods. 2021.
29. Dong K, et al. STAGATE: spatial transcriptomics analysis via graph attention auto-encoder. Nature Communications. 2022.
30. Cui H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods. 2024.
31. Theodoris CV, et al. Transfer learning enables predictions in gene expression. Nature. 2023.
32. Chen S, et al. scBERT: a pretrained language model for single-cell transcriptomics. Bioinformatics / arXiv. 2022.
33. Olawade DB, Teke J, Adeleye KK, Weerasinghe K, Maidoki M, Clement David-Olawade A. Artificial intelligence in in-vitro fertilization (IVF): A new era of precision and personalization in fertility treatments. J Gynecol Obstet Hum Reprod.
34. Zhu C, et al. Single-cell multimodal omics: integration benchmarking across tasks. Nature Methods. 2024.
35. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005.
36. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. KDD. 2016.
37. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports. 2019.