Deep learning shapes single-cell data analysis

Qin Ma[1,2]✉ *and Dong Xu*[3]

Deep learning has tremendous potential in single-cell data analyses, but numerous challenges and possible new developments remain to be explored. In this commentary, we consider the progress, limitations, best practices and outlook of adapting deep learning methods for analysing single-cell data.

Single-cell technologies have substantially advanced our understanding of heterogeneity and functional diversity among individual cells, and bring enormous opportunities for biology and precision medicine, especially to study cells undergoing rapid differentiation (for example, during drug resistance and tumour relapse[1]), evolving into diverse subpopulations (for example, immune cells[2]) or responding to external perturbances (for example, during COVID-19 pathogenesis[3]). Concurrent with single-cell technologies, deep learning (DL) — a breakthrough in artificial intelligence — redefines our capabilities to analyse large-scale data by using sophisticated architectures of artificial neural networks[4]. The power of DL has recently been demonstrated in AlphaFold2's prediction of protein structures, and use of DL is now feasible for single-cell data analyses.

Specifically, autoencoders (AE) have been widely employed to capture features and improve signal-to-noise ratios for accurate cell-type clustering, batch correction and gene imputation in single-cell studies. SAUCIE[5] applied AE to a dataset consisting of 11 million T cells from 180 dengue patients, identified cluster-based signatures of acute dengue infection and stratified immune response to dengue. Meanwhile, graph neural networks (GNN) in combination with attention mechanisms have made DL models more effective and explainable. scGNN is the first GNN model for scRNA-seq data to simultaneously perform gene imputation and cell clustering; scGNN identified ten neuron clusters and cell-type-specific markers in Alzheimer disease[6]. SpaGCN is a GNN model to identify tissue architecture from spatially resolved transcriptomics data; SpaGCN separated cancer and non-cancer regions of human primary pancreatic tumours and identified two marker genes distinguishing the cancer region[7].

Best practices in developing deep learning for single-cell studies

The highly heterogeneous nature of single-cell data can be analysed across a wide range of research topics by generalizing DL model design and optimization in a hypothesis-free manner. External biological knowledge or data (for example, phenotypic information or bulk omics data) can be incorporated into the model to improve predictions as constraints. Single-cell data often contain a limited number of benchmark labels and annotations, which could result in model overfitting and poor performance. Fortunately, in many cases emerging semi-supervised learning (combining a small amount of labelled data with a large amount of unlabelled data) and self-supervised learning (constructing data representation of the unlabelled data by predicting any part or property from other parts or properties of the data) can often achieve equally insightful results without requiring the extra labels. Furthermore, to improve the trustworthiness of DL models, especially model generalization in different experimental platforms and conditions, and robustness against noises in the data, it is desirable for developers to provide the scope of the methodological uses and demonstrate for what kinds of data or in what situations DL will work well or poorly. In addition, providing some confidence assessments (for example, *P*-values or *z*-scores) of prediction results can guide users to make biological inferences.

With a wide range of built-in capabilities, a composable DL pipeline can help automate complex and repetitive tasks involved in model development. This composability allows the appropriate resources to be gathered to ensure a tailored system under software control. Composable DL can be used by developers to configure easy-to-use and white-box models that address various single-cell research topics in a customizable fashion without too many challenges. Furthermore, it is a good practice to provide well-structured source code, hands-on tutorials and clear documentation of protocols, including the encompassing format, processing steps, model training, code versioning, tutorials to ensure reproducibility, and parameter tuning for other developers and general users.

Best practices in applying deep learning in single-cell biology

DL users usually find it challenging to decide when and how to select DL tools for single-cell data analysis based on usability and accuracy. In contrast to Seurat, which

[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, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.*

✉*e-mail: qin.ma@osumc.edu*

https://doi.org/10.1038/s41580-022-00466-x

has been widely used in single-cell data analysis, DL may uncover more intrinsic relations and mechanisms. The selection of the best-fit DL model is typically driven by a specific goal, for example, whether it is for cell clustering or cell classification, and whether feature order matters or the topological relationship among different modalities matters. Other considerations include data structure (for example, tabular, sequential, time series or graph-structured), data size, and computational expenses (multi-task and multimodal learning). Supplementary Table 1 provides information and guidelines on main functions, core models, and biological interpretations of representative DL tools. As DL models for single-cell data analysis have not matured, it may be valuable to run multiple tools to see how they compare. Furthermore, comprehensive single-cell DL benchmarking papers help users choose the best model[8,9].

Limitations of deep learning in single-cell data analysis

Although the existing DL tools have demonstrated their capacities in analysing single-cell data in various settings, they have not been extensively used by independent research groups in their biological studies. Supplementary Table 1 only includes case studies of the original methodology papers. Although it often takes time for new technologies to become established, the limitations of current DL methods in single-cell data analysis are also barriers. In particular, DL methods often require large data and computational resources to train; their results may not be robust (performance varies owing to data noises, parameter settings and new input data); most DL models are black boxes lacking expandability; and almost all DL tools need extensive computer skills to use. Hence, there is still a gap between DL method development and its broad application in diverse biological systems. Next, we discuss prospects of filling-in this gap.

Prospects of deep learning in single-cell data analysis

DL application in single-cell data analyses holds great promise for future exploration. For method development, a continuous adaptation of the rapidly evolving, cutting-edge DL methods has been witnessed. Owing to the limited annotated data available in single-cell biology, there is room for applying active learning (interactively suggesting new data labelling for training the model) to build models based on a few training samples. Higher adoption of the end-to-end DL frameworks (for example, in AlphaFold2) can facilitate a more comprehensive and holistic use of the training data to account for all input features and relationships. Model-based DL is expected to penetrate single-cell biology even further. Structure- or topology-aware methods, and physics-inspired and biologically informed frameworks integrate knowledge into DL models for other applications; similar applications can be expected in single-cell biology. Furthermore, the development of explainable DL could support better interpretations of underlying biological mechanisms, including causal or regulatory relationships, cell-type-specific responses to external stimuli, and cell subpopulations that drive diseases or phenotypes.

Another trend is lowering the barrier of applying DL technologies in single-cell data analyses. We believe that developing integrated systems and deploying cloud platforms will enable users without programming skills to use the single-cell DL tools through web services or dockers connected to online resources. In addition, a modular framework design, thanks to its flexibility, can leverage individual DL models and single-cell knowledge. Notably, the establishment of well-defined standards for DL-ready data, codes and models are expected to attract more developers to develop open-source/access DL tools, which in turn can expand in-depth single-cell data analyses. These tools can also help train the next generation of researchers and clinicians, particularly allowing precision medicine to be more deliverable to medical practices.

DL-based methods have demonstrated their prowess in a broad range of single-cell studies[10], such as understanding the complexity of brain cell types related to perception and complex behaviours, and inferring the high diversity of tumour and immune cell populations to greatly accelerate the discovery of novel pathogenesis and cancer therapeutics. We expect such studies will be greatly expanded to provide unique insights, which likely would not be achievable without combining single-cell data and DL technologies. Another growing area is the migration of DL models from predictable and interpretable to more actionable, that is, recommendations that can directly lead to medical treatment, such as therapeutic targets, drug repurposing and drug combinations.

1. Nath, A. & Bild, A. H. Leveraging single-cell approaches in cancer precision medicine. *Trends Cancer* **7**, 359–372 (2021).
2. Mogilenko, D. A., Shchukina, I. & Artyomov, M. N. Immune ageing at single-cell resolution. *Nat. Rev. Immunol.* https://doi.org/10.1038/s41577-021-00646-4 (2021).
3. Tian, Y. et al. Single-cell immunology of SARS-CoV-2 infection. *Nat. Biotechnol.* **40**, 30–41 (2022).
4. Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.* **38**, 1007–1022 (2020).
5. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
6. Wang, J. et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* **12**, 1882 (2021).
7. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
8. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
9. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
10. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**, 40–55 (2021).

Acknowledgements
We thank J. Wang from the University of Missouri and A. Ma from the Ohio State University for their efforts in preparing this article. This manuscript is supported by grants R35-GM126985 and R01-GM131399 from the National Institutes of Health.

Competing interests
The authors declare no competing interests.

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/s41580-022-00466-x.