Stem Cell Reports
Perspective
ISSCR
OPEN ACCESS

Deep learning models will shape the future of stem cell research

John F. Ouyang,[1] Sonia Chothani,[1] and Owen J.L. Rackham[1,2,3,*]
[1]Duke-NUS Medical School, Program in Cardiovascular and Metabolic Disorders (CVMD) and Centre for Computational Biology (CCB), Singapore, Singapore
[2]School of Biological Sciences, University of Southampton, Southampton, UK
[3]The Alan Turing Institute, The British Library, London, UK
*Correspondence: o.j.l.rackham@soton.ac.uk
https://doi.org/10.1016/j.stemcr.2022.11.007

SUMMARY

Our ability to understand and control stem cell biology is being augmented by developments on two fronts, our ability to collect more data describing cell state and our capability to comprehend these data using deep learning models. Here we consider the impact deep learning will have in the future of stem cell research. We explore the importance of generating data suitable for these methods, the requirement for close collaboration between experimental and computational researchers, and the challenges we face to do this fairly and effectively. Achieving this will ensure that the resulting deep learning models are biologically meaningful and computationally tractable.

INTRODUCTION

As biologists, we are constantly constructing, refining, and testing a private internal "model" of the systems we study. As we are exposed to new data or ideas that challenge or confirm this model, we adapt accordingly. This accumulated experience allows us to hypothesize about what happens under unseen conditions or teach others about the systems we study. The remarkable thing about human brains is that they learn and understand complex systems through a relatively small number of observations, making them an engine for innovation. However, two revolutions have altered this status quo: the exponential growth of data science technologies such as cloud computing, artificial intelligence, and machine learning and the increased capacity to generate vast amounts of data. Together these revolutions are allowing us to train *in silico* deep learning (DL) models that can take advantage of these vast datasets in a way we cannot ourselves.

The application of DL models has already delivered several novel insights. Most notably, the development of AlphaFold2 (Jumper et al., 2021), a semi-supervised DL model that can predict protein structure from a sequence, is revolutionizing how we study protein structure. AlphaFold2 uses multiple sequence alignment data, which are generated by aligning multiple similar protein sequences and identifying regions of high or low similarity. Since our ability to collect a large number of protein sequences has been established for some time, the scale of this problem far exceeded our human capabilities and is something the machine learning field has been trying to tackle for some time. Increasingly, we are now seeing similar problems arise in other areas of biology. For instance, as stem cell biologists, we can now collect hundreds of thousands and up to millions of observations of individual cells during differentiation or development. Interpreting these data also exceeds our capabilities, especially as we combine data from multiple experiments and modalities. As a community, we must develop DL models to bridge this interpretation gap, learning the relationships between different conditions, focusing our investigations, and making full use of the data we generate.

Increasingly, developments in stem cell research are being catalyzed by DL models. The application of imaging and deep neural networks has helped better measure and understand changes in morphology that occur during differentiation, predicting how cells are likely to differentiate (Ren et al., 2021), annotating cells in an unbiased way (Guo et al., 2021), or elucidating stem cell identify (Stumpf and MacArthur, 2019). DL methods have also been developed to reconstruct developmental trajectories from single-cell data (for example, La Manno et al., 2018; Lummertz da Rocha et al., 2018; Moon et al., 2019), and these models are allowing us to understand the fate choice of stem cells at a resolution far higher than previously possible (Pellin et al., 2019) and even find novel cell states arising during reprogramming (Liu et al., 2020). DL is also extending our capability to control the behavior of stem cells, for instance, controlling their pattern formation (Libby et al., 2019) or finding optimal culture conditions (Kamaraj et al., 2020; Kanda et al., 2022).

In this perspective, we will focus on the impact of DL models on transcriptomics in the context of stem cell biology. Still, similar advances are also taking place across biology, from proteomics to medical imaging. The challenge remains on how to effectively introduce these models into our scientific workflows and shape their future development to maximize their ability to help us tackle the most critical problems in human health and disease.

The underpinning of any model is the data

A critical requirement for generating accurate models is the data on which they are trained. The widespread adoption of high-throughput technologies to measure gene expression and genomic sequences has resulted in the generation of vast quantities of data. However, how these data are generated and shared can dramatically impact their utility for training DL models.

Open access data sharing has made data publicly available to the scientific community. Publications are routinely required to deposit their newly generated raw and processed data in repositories such as the Gene Expression Omnibus (Barrett et al., 2013), Short Read Archive (Leinonen et al., 2011), and European Nucleotide Archive (Amid et al., 2020). However, to effectively use these data resource as a whole for training a DL model, one would need to perform a significant amount of standardization and normalization of the raw and processed files deposited. This would be necessary to account for any batch effects, such as processing pipeline, experimental protocol, and inter-individual differences. As a result, the data from within these repositories are rarely used collectively, and with the scale of data being so important for training, this is a limiting factor for the use of these resources for DL models.

One way to overcome this is to re-analyze individual datasets from repositories using a standardized analysis pipeline. For example, Remap (Chèneby et al., 2020) reanalyzed public CHIP-seq data to generate a catalog of millions of peaks and obtained high-quality regulatory regions in humans and *Arabidopsis*. For transcriptomics data, the recount (Collado-Torres et al., 2017) and ARCHS4 (Lachmann et al., 2018) projects provide an online resource where RNA-seq data from different human and mouse studies are realigned to the same gene models and then recounted, allowing for uniform reprocessing of the data. Although these reanalyses reduce the variation between datasets by using uniform processing, it is impossible to account for the experimental differences between datasets fully. As such, some batch effects may remain, which can unpredictably affect the DL model's training.

One way to create large, uniform datasets is via consortia-led efforts that generate the data with standardized protocols and data processing pipelines. For example, for over 20 years, the FANTOM consortium has generated large amounts of transcriptomics data in humans and mice. The generated data have been released in raw and browsable formats (Abugessaisa et al., 2021). Similarly, the ENCODE consortium has generated large amounts of data to describe regulatory elements in humans and mice. Currently, in its fourth phase, ENCODE4 makes their generated transcriptomics and epigenetics data publicly available and ensures that all newly obtained human samples are consented for unrestricted data sharing. More recently, the human cell atlas (HCA) (Regev et al., 2017) has launched, intending to collect single-cell data generated across the globe. The HCA has a data coordination platform that facilitates data sharing, thus maximizing data usage and streamlining efforts across the scientific community. Such datasets can be used routinely for gene-level hypotheses and global studies (Alam et al., 2020; Fort et al., 2014) but also provide the ideal dataset for training or benchmarking DL methods (e.g., Köhler et al., 2021; Lotfollahi et al., 2022a; Niu et al., 2019). As the field develops, consortia such as FANTOM, ENCODE, and the HCA must consider the application of the data generated for training DL models as one of the primary outcomes and design experiments to maximize applicability for *in silico* as opposed to human interpretation. For this to be possible, the field must prioritize building trust and understanding of DL methods among the research community.

DL enables data-driven science, allowing us to extract information from multiple data modalities or experimental conditions. For instance, the correct combination of DL with temporal gene expression data from annotated embryos was able to dissect the dynamics of human development at an unprecedented scale (Meistermann et al., 2021). Equally, a recent study demonstrated that it was possible to accurately predict the fate of neural stem cells after only one day in culture by applying DL models to bright-field imaging data from differentiating neural stem cells (Zhu et al., 2021). Looking to the future, there are several ways that we can ensure that DL models are as generalizable and comprehensive as possible. Firstly, using DL models to integrate more layers of data will provide a more complete view of development. Secondly, training models on multiple experimental conditions at once will avoid them being over-fitted to specific processes and, as such, more likely to be able to make robust predictions outside of our collected observations.

As a community, the choice and breadth of measures of success will also be crucial to determining the accuracy of DL models. For instance, in a recent study DL models were trained to predict endothelial cells based on their morphology in phase-contrast images and orthogonally validated by CD31 expression from immuno-staining of the test set, allowing for the model accuracy to be assessed in multiple ways (Kusumoto et al., 2018). Developing these measures of success relies on collaboration between wet-bench and computational researchers, ensuring they are biologically meaningful and computationally tractable. To achieve this, it is crucial to familiarize wet-bench researchers with the importance of generating appropriate data for DL methods and engaging them in future developments. The most potent DL approaches will emerge from combined innovation in experimental design and algorithm development.

DL models are already helping to leverage public data

With more sequencing data being generated, incorporating data from diverse sources is becoming increasingly crucial. Without the ability to combine data accurately, researchers are left to interpret each dataset individually and fail to capitalize on this growing resource. DL models are already proving to be a powerful tool in data integration. For instance, numerous DL-based tools for integrating single-cell transcriptomic data have been developed, e.g., scVI (Lopez et al., 2018), trVAE (Lotfollahi et al., 2020), scANVI (Xu et al., 2021), scGen (Lotfollahi et al., 2019), and expiMap (Lotfollahi et al., 2022b). Many of these tools employ an autoencoder-based neural network model that attempts to encode the high-dimensional input gene expression into a lower-dimensional representation before decoding this back to the original input space. The DL model then aims to minimize the difference between the original training data and the encoded-decoded counterpart. The hypothesis is that if the model can find a set of weights to do this with high accuracy, then the encoded low-dimensional representation must capture much of the meaningful information from the higher-dimensional input. This encoding-decoding process allows the model to learn the major sources of variation, biological and technical, in the input data. The learnt biological relationships can be used to infer gene regulatory networks, which have been applied to identify lineage-specific gene regulation in mouse hematopoiesis (Shu et al., 2021). Furthermore, the technical differences can be removed to harmonize the data, allowing data from different studies to be analyzed together.

Increasingly, there are DL-based tools for interpreting single-cell genomics data across different modalities, e.g., transcriptomics, protein levels, and spatial data. A pioneering DL model in multimodal integration, the TotalVI model (Gayoso et al., 2021), uses the same encoding-decoding process to learn both single-cell transcriptomics and protein data distribution to create a joint low-dimensional representation. This allows for the imputation of data from one modality, e.g., protein levels from the learnt representation and the integration of different multimodal single-cell data. DL-based tools are also applied to multimodal spatial data. The DestVI model (Lopez et al., 2022) deconvolutes the cell types observed in spatial transcriptomics by training two autoencoders to learn cell-type-specific gene expression patterns and the spatial context of these patterns. It is also possible to predict gene expression from different DNA sequences, as demonstrated by the Nvwa tool (Li et al., 2022), which identified cell-type-specific TF-gene regulatory relationships conserved across species. Tools like this will benefit stem cell research by elucidating how *cis*-regulatory elements drive differentiation into different lineages. For instance, the Nvwa tool could be used to further decipher the grammar of transcriptional regulation and its downstream impact on phenotype by complementing studies such as in Vigilante et al. (2019), where genetic variation in human iPSC cell lines that affect stem cell behavior have been identified. Currently, these DL models are applied on *matched* multimodal data where different modalities are profiled simultaneously on the *same* single cell. However, many studies often profile the different modalities from different samples due to technical limitations in extracting cellular material from single cells. Thus, there is an opportunity to modify existing DL models for use with *unmatched* data where the different modalities are profiled on *different* single cells. This will allow biologists to extract further biological insights from existing data.

As the volume and types of data increase and innovations in DL continue accumulating, we expect the complexity of DL models to increase. However, with this increase in model complexity, the interpretability of these models may suffer. Currently, most DL models function as "black boxes." However, as biologists, we are concerned with the underlying gene programs and pathways that drive different biological processes. Thus, to build trust in DL models, it is crucial to move toward interpretable DL models where users can identify the genes or pathways contributing to the results. This will also provide a better understanding of the underlying mechanisms, which can be tested experimentally to further confirm the accuracy of DL models.

The future of data and models in stem cell research

Despite numerous technological advances, the complexity of most experimental designs has remained relatively unchanged. The typical experiment often comprises cells from a "steady state" subjected to a single modification/perturbation or compared with a "modified state" such as a disease. This design lends itself well to questions that we (as humans) find more accessible to answer, such as "what is the difference between the steady and modified state?" This means that despite the data providing higher resolution and increased complexity, the types of questions being addressed have not evolved at the same rate. However, this is beginning to change; in many cases, the scale and cost of data generation will make these innovations a necessity, with the data being so complex that without a DL intermediary, their interpretation becomes impossible.

One area where this will likely happen in the near future is data generated using spatial and single-cell sequencing. With each innovation in spatial transcriptomics comes an increase in either spatial resolution or sample size. Thus, we should anticipate methods that can measure gene expression at a subcellular resolution at an organ scale. Similarly, the cost per cell is decreasing for single-cell sequencing, and more modalities can now be profiled simultaneously. We should also expect that we will routinely collect millions of observations, from multiple modalities, of biological processes over time. In both cases, analyzing this data by treating each data point independently will become computationally prohibitive and fail to extract the true value of the data. Instead, we should strive to use this data to train general models of cellular behavior. Using these models, we can go far beyond knowing which genes are cell type specific but instead test hypotheses about changes in gene expression or find the critical events that initiate the divergence of different cell fate over time. As we collect more data and innovate DL architectures, these models will become more accurate and increasingly offer an alternative to *in vitro* and animal models to test hypotheses.

Using single-cell technologies to profile individually perturbed cells has already created a new generation of assays that generate highly complex datasets. For example, perturb-seq is a high-throughput method of performing single-cell RNA sequencing on pooled genetic or transcriptional perturbation screens (Dixit et al., 2016). These approaches are already used in the stem cell field to search for optimal combinations of regulators for driving neuronal differentiation (e.g., Kreimer et al., 2022; Luginbühl et al., 2021), understanding T cell development (Schmidt et al., 2022), and screen thousands of perturbations *in vivo* (Jin et al., 2020). The data generated are of a much larger scale, with experiments demonstrating the capability to capture the effect of thousands of perturbations at once, but in the near future, this could exponentially increase to millions of perturbations in a single experiment. Consequently, many current approaches to analyzing gene expression will need to change. Dimension reduction approaches will become inadequate to provide insight into the data, and the generation of lists of differentially expressed genes will become meaningless given the numerous different conditions/perturbations. As a result, deploying DL models will become essential to extract the full potential of these emerging datasets. These models will bring many new capabilities, for example, transferring the result of perturbations from one cell type to another, allowing us to experiment with the effect of a drug or genetic perturbation without needing to experiment on precious samples (Lotfollahi et al., 2021). These tools can be applied to stem cell research to find perturbations that can improve the efficiency of differentiation protocols or better understand how these processes occur *in vivo*.

Generative DL models can also be used to predict future cell states in a developmental process or in a differentiation experiment. For instance, the PRESCIENT tool (Yeo et al., 2021) employs DL to learn the underlying differentiation landscape from time-series scRNA-seq data. The learnt model was then used to simulate changes in cell fate in hematopoiesis and pancreatic β cell differentiation. Such techniques can be applied to model very early development, allowing us to predict cell fate decisions beyond the first 14 days of embryo development where current ethics prohibit experimentation. This will allow stem cell biologists to better understand the drivers for diseases and biology related to early organogenesis.

Another feature of these models that is often overlooked is their portability, which impacts our ability to share biological findings. For instance, predictive models trained on large consortia-level datasets can be easily shared, allowing researchers with smaller datasets to apply these models to their problems. The open sharing of models also facilitates benchmarking against each other to assess the most suitable one for a given task. Equally, shared models can be used as a starting point for future development. The success of model sharing will rely on establishing the correct infrastructure for model deposition, curation, and benchmarking and providing the interface between the life and data sciences. Early efforts in developing such infrastructure have come from model zoos. These are machine learning model deployment platforms designed specifically for model ease of use. Model zoos specific to computational biology such as Kipoi (Avsec et al., 2019) and BioImage (Ouyang et al., 2022), as well as domain-agnostic model zoos such as ModelZoo.co, already contain thousands of pre-trained DL models that can be readily re-implemented onto new data. As the field develops, the stem cell community will need to engage more closely with developing these critical infrastructures to ensure that they are guided equally by computational and biological expertise. Not only will the success of these models rely on the production of high-quality data, but the way that the models are trained and the biologically relevant metrics that are used to gauge the success of their training also have a significant impact on their biological accuracy and utility.

Challenges to overcome

Reaching a point where DL will shape stem cell research will require overcoming several challenges. These challenges are not just computational and biological but also social/human and will require input from several disciplines to overcome. Firstly, as a field, we must learn from mistakes made in previous uses of DL and ensure that our biases (conscious and unconscious) are not transferred to these models through the data we generate and the metrics we define. Thus, the data generated to train DL models must represent our species unbiasedly and comprehensively, covering different biological systems and incorporating diverse genetic backgrounds. This will be critical in ensuring that the rewards from applying DL are as equitable as possible in the future.

Some aspects of stem cell biology set them apart from other cell types in our body, and these differences could be important to consider in implementing DL models. For instance, stem cells are dynamic, sensing and responding to the microenvironment. As such, DL models will need to be able to capture dynamical properties, and the data used to train the models will need sufficient resolution to capture these changes. This is further complicated because differentiation and reprogramming can happen asynchronously and in different orders. As such, DL models will need to be able to predict the behavior of individual cells but also incorporate how their ensemble at a given moment affects the global dynamics. This combination of multiple hierarchies and interactions makes stem cells difficult to model compared with stable, fully differentiated cell states whose dynamics tend to be simpler.

Another challenge we face is that even with the expected advances in sequencing technologies, our measurements only provide a partial and incomplete description of the cell state we are trying to model. We must find ways to interface our digital models with their "real-world" counterparts. This may involve simulating the same experiments on the digital models and comparing them to the same experiments carried out on the biological cells or having a series of functional tests that can be carried out *in silico* to ensure that the behavior of the cellular models accurately captures what happens in biology. As a field, we will need to have an ongoing dialogue on how best to cross this *in silico* to *in vitro* or *in vivo* divide, and this will be critical to the utility of the models we generate.

Conclusion

Overall, we expect that DL models will increasingly become part of our arsenal of tools to explore biology. As both data generation and DL models improve, generalizable DL models of the cell will evolve. Ultimately, these will allow *in silico* experiments to become another alternative to current *in vitro* and *in vivo* models. Although they will never replace wet-lab experiments or animal work, they have enormous potential to direct the focus of these or enable testing hypotheses that would never be experimentally feasible. As a stem cell community, we should actively combine the data and life sciences to realize this potential and enable our field to benefit from the increasing number of innovations in DL.

ACKNOWLEDGMENTS

J.F.O. was supported by an NMRC OF-YIRG (MOH-OFYIRG21nov-0004) and S.C. was supported by a Khoo Postdoctoral Fellowship Award.

CONFLICT OF INTERESTS

O.J.L.R. is an SAB member and shareholder of Mogrify Ltd.

REFERENCES

Abugessaisa, I., Ramilowski, J.A., Lizio, M., Severin, J., Hasegawa, A., Harshbarger, J., Kondo, A., Noguchi, S., Yip, C.W., Ooi, J.L.C., et al. (2021). FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. Nucleic Acids Res. *49*, D892–D898.

Alam, T., Agrawal, S., Severin, J., Young, R.S., Andersson, R., Arner, E., Hasegawa, A., Lizio, M., Ramilowski, J.A., Abugessaisa, I., et al. (2020). Comparative transcriptomics of primary cells in vertebrates. Genome Res. *30*, 951–961.

Amid, C., Alako, B.T.F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P.W., Holt, S., Hussein, A., Ivanov, E., et al. (2020). The European nucleotide archive in 2019. Nucleic Acids Res. *48*, D70–D76.

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., Banerjee, A., Kim, D.S., Beier, T., Urban, L., et al. (2019). The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat. Biotechnol. *37*, 592–600.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

Chèneby, J., Ménétrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon, A., Lopez, F., and Ballester, B. (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. Nucleic Acids Res. *48*, D180–D188.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S.E., Taub, M.A., Hansen, K.D., Jaffe, A.E., Langmead, B., and Leek, J.T. (2017). Reproducible RNA-seq analysis using recount2. Nat. Biotechnol. *35*, 319–321.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., et al. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat. Genet. *46*, 558–566.

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

Guo, J., Wang, P., Sozen, B., Qiu, H., Zhu, Y., Zhang, X., Ming, J., Zernicka-Goetz, M., and Na, J. (2021). Machine learning-assisted high-content analysis of pluripotent stem cell-derived embryos in vitro. Stem Cell Rep. *16*, 1331–1346.

Jin, X., Simmons, S.K., Guo, A., Shetty, A.S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., et al. (2020). In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. Science *370*, eaaz6063. https://doi.org/10.1126/science.aaz6063.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Kamaraj, U.S., Chen, J., Katwadi, K., Ouyang, J.F., Yang Sun, Y.B., Lim, Y.M., Liu, X., Handoko, L., Polo, J.M., Petretto, E., and Rackham, O.J.L. (2020). EpiMogrify models H3K4me3 data to identify signaling molecules that improve cell fate control and maintenance. Cell Syst. *11*, 509–522.e10.

Kanda, G.N., Tsuzuki, T., Terada, M., Sakai, N., Motozawa, N., Masuda, T., Nishida, M., Watanabe, C.T., Higashi, T., Horiguchi, S.A., et al. (2022). Robotic search for optimal cell culture in regenerative medicine. Elife *11*, e77007. https://doi.org/10.7554/eLife.77007.

Köhler, N.D., Büttner, M., Andriamanga, N., and Theis, F.J. (2021). Deep Learning Does Not Outperform Classical Machine Learning for Cell-type Annotation.

Kreimer, A., Ashuach, T., Inoue, F., Khodaverdian, A., Deng, C., Yosef, N., and Ahituv, N. (2022). Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. Nat. Commun. *13*, 1504.

Kusumoto, D., Lachmann, M., Kunihiro, T., Yuasa, S., Kishino, Y., Kimura, M., Katsuki, T., Itoh, S., Seki, T., and Fukuda, K. (2018). Automated deep learning-based system to identify endothelial cells derived from induced pluripotent stem cells. Stem Cell Rep. *10*, 1687–1695.

Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat. Commun. *9*, 1366.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21.

Li, J., Wang, J., Zhang, P., Wang, R., Mei, Y., Sun, Z., Fei, L., Jiang, M., Ma, L., E, W., et al. (2022). Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. Nat. Genet. *54*, 1711–1720. https://doi.org/10.1038/s41588-022-01197-7.

Libby, A.R.G., Briers, D., Haghighi, I., Joy, D.A., Conklin, B.R., Belta, C., and McDevitt, T.C. (2019). Automated design of pluripotent stem cell self-organization. Cell Syst. *9*, 483–495.e10.

Liu, X., Ouyang, J.F., Rossello, F.J., Tan, J.P., Davidson, K.C., Valdes, D.S., Schröder, J., Sun, Y.B.Y., Chen, J., Knaupp, A.S., et al. (2020). Reprogramming roadmap reveals route to human induced trophoblast stem cells. Nature *586*, 101–107.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., et al. (2022). DestVI identifies continuums of cell types in spatial transcriptomics data. Nat. Biotechnol. *40*, 1360–1369. https://doi.org/10.1038/s41587-022-01272-8.

Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.

Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36*, i610–i617.

Lotfollahi, M., Susmelj, A.K., De Donno, C., Ji, Y., and Ibarra, I.L. (2021). Learning interpretable cellular responses to complex perturbations in high-throughput screens. Preprint at bioRxiv.

Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022a). Mapping single-cell data to reference atlases by transfer learning. Nat. Biotechnol. *40*, 121–130.

Lotfollahi, M., Rybakov, S., Hrovatin, K., and Hediyeh-zadeh, S. (2022b). Biologically informed deep learning to infer gene program activity in single cells. Preprint at bioRxiv.

Luginbühl, J., Kouno, T., Nakano, R., Chater, T.E., Sivaraman, D.M., Kishima, M., Roudnicky, F., Carninci, P., Plessy, C., and Shin, J.W. (2021). Decoding neuronal diversification by multiplexed single-cell RNA-seq. Stem Cell Rep. *16*, 810–824.

Lummertz da Rocha, E., Rowe, R.G., Lundin, V., Malleshaiah, M., Jha, D.K., Rambo, C.R., Li, H., North, T.E., Collins, J.J., and Daley, G.Q. (2018). Reconstruction of complex single-cell trajectories using CellRouter. Nat. Commun. *9*, 892.

Meistermann, D., Bruneau, A., Loubersac, S., Reignier, A., Firmin, J., François-Campion, V., Kilens, S., Lelièvre, Y., Lammers, J., Feyeux, M., et al. (2021). Integrated pseudotime analysis of human pre-implantation embryo single-cell transcriptomes reveals the dynamics of lineage specification. Cell Stem Cell *28*, 1625–1640.e6.

Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. *37*, 1482–1492.

Niu, X., Yang, K., Zhang, G., Yang, Z., and Hu, X. (2019). A pretraining-retraining strategy of deep learning improves cell-specific enhancer predictions. Front. Genet. *10*, 1305.

Ouyang, W., Beuttenmueller, F., Gómez-de-Mariscal, E., Pape, C., Burke, T., Garcia-López-de-Haro, C., Russell, C., Moya-Sans, L., de-la-Torre-Gutiérrez, C., Schmidt, D., et al. (2022). BioImage Model Zoo: A Community-Driven Resource for Accessible Deep Learning in BioImage Analysis.

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat. Commun. *10*, 2395.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. Elife *6*, e27041. https://doi.org/10.7554/eLife.27041.

Ren, E., Kim, S., Mohamad, S., Huguet, S.F., Shi, Y., Cohen, A.R., Piddini, E., and Salas, R.C. (2021). Deep Learning-Enhanced Morphological Profiling Predicts Cell Fate Dynamics in Real-Time in hPSCs.

Schmidt, R., Steinhart, Z., Layeghi, M., Freimer, J.W., Bueno, R., Nguyen, V.Q., Blaeschke, F., Ye, C.J., and Marson, A. (2022). CRISPR activation and interference screens decode stimulation responses in primary human T cells. Science *375*, eabj4008.

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. (2021). Modeling gene regulatory networks using neural network architectures. Nat. Comput. Sci. *1*, 491–501.

Stumpf, P.S., and MacArthur, B.D. (2019). Machine learning of stem cell identities from single-cell expression data via regulatory network archetypes. Front. Genet. *10*, 2.

Vigilante, A., Laddach, A., Moens, N., Meleckyte, R., Leha, A., Ghahramani, A., Culley, O.J., Kathuria, A., Hurling, C., Vickers, A., et al. (2019). Identifying extrinsic versus intrinsic drivers of variation in cell behavior in human iPSC lines from healthy donors. Cell Rep. *26*, 2078–2087.e3.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol. Syst. Biol. *17*, e9620.

Yeo, G.H.T., Saksena, S.D., and Gifford, D.K. (2021). Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. Nat. Commun. *12*, 3222.

Zhu, Y., Huang, R., Wu, Z., Song, S., Cheng, L., and Zhu, R. (2021). Deep learning-based predictive identification of neural stem cell differentiation. Nat. Commun. *12*, 2614.