https://doi.org/10.1038/s41592-023-01931-x

Scaling biological discovery at the interface of deep learning and cellular imaging

Morgan Schwartz, Uriah Israel, Xuefei (Julie) Wang, Emily Laubscher, Changhua Yu, Rohit Dilip, Qilin Li, Joud Mari, Johnathon Soro, Kevin Yu, Elora Pradhan, Ada Ates, Danielle Gallandt, Ross Barnowski, Edward Pao & David Van Valen

Concurrent advances in imaging technologies and deep learning have transformed the nature and scale of data that can now be collected with imaging. Here we discuss the progress that has been made and outline potential research directions at the intersection of deep learning and imaging-based measurements of living systems.

Imaging has been a core technology for studying living matter, as it enables us to capture spatial and temporal variation in the form of images. Recent advances in imaging methods – including imaging-based reporters, multiplexing[1,2] and super-resolution[3] – have enabled numerous discoveries in cell and tissue biology. The advances in multiplexing have been particularly important. They have powered the spatial genomics revolution and given rise to a form of data that simultaneously captures the spatial and 'parts list' variation of living matter. Although modern imaging data contain rich insights, their complexity has presented substantial challenges to their interpretation. The complexity and scale of these data have necessitated algorithmic solutions to key processing steps such as cell segmentation and cell tracking, as well as numerous other tasks. Over the past few years, we have seen remarkable progress at the intersection of computer vision and biological imaging[4]. This progress is rooted in deep-learning methods that have provided effective solutions to several of the problems mentioned above[5–8]. We have also seen applications such as image restoration[9,10] and augmented microscopy[11,12] that have pushed the frontier of what is possible with imaging. For problems for which 'out-of-the-box' solutions do not exist, there is a reliable recipe for creating an effective deep learning-based solution: create relevant, labeled data; train a deep-learning model to perform the task at hand; and deploy the resulting model[4]. Although challenges relating to the practical application of the deep-learning methodology to imaging data exist (creating sustainable software tools, improving ease of use and labeling data, to name a few), the progress that has been made cannot be denied. This progress reminds us of comments made by Dror Berman at Innovation Endeavors during a conversation that we had in 2017, that computer vision is effectively a 'solved problem'. Although there is always work to be done to implement solutions, we know what the solution looks like for most problems.

Given the advances that have been made in bioimaging on both the measurement and algorithmic fronts, the natural question that faces the field is what comes next. There are several areas of research that we believe are particularly exciting:

- Integrating heterogeneous biological data with imaging-based measurements and deep learning. Advances in reporter development, multiplexing and spatial genomics mean that it is now possible to measure multiple aspects of cell state with imaging simultaneously. For example, compatible imaging-based measurements for protein localization (antibody staining with oligonucleotide-labeled antibodies[13]), RNA transcript abundance (multiplexed single-molecule RNA fluorescence in situ hybridization (FISH)[14,15]) and chromatin state (epigenomic multiplexed error-robust FISH (epigenomic MERFISH)[16]) all exist in the literature. Because these measurements are imaging-based, they can be paired with cellular imaging to connect difficult phenotypes (signaling and morphological dynamics, super-resolved organelle structures and so on) to systems-level 'omics' measurements. Pursuing integrated measurements in this fashion is appealing because of the ability of imaging to conserve cellular identity across measurements and its low marginal cost of data. Further, recent work on pooled optical screening that has added a perturbation dimension to imaging data provides even more potential for discovery[17,18]. A new generation of measurements in which live-cell imaging of perturbed cells is paired with end-point, image-based, multiomic readouts promises to provide information-rich profiles of single cells while achieving scale.
- Universal models for bioimaging data. Recent artificial intelligence (AI) methods (for example, transformers[19,20] and generative methods[21–23]) have proven capable of performing remarkable feats on a diverse array of data types, ranging from images[20,22] to text[24]. The increasing scale of labeled bioimaging data raises the question of how far these deep-learning methods can be pushed in bioimaging when the scale of data and computation approaches what exists in other fields. Is it possible to create truly universal methods for bioimaging data that generalize across both datasets and tasks, and can these models serve as an interface between other forms of data (omics, structural data, knowledge graphs and so on) about living systems? We suspect the answer is yes, although substantial time and effort will be required to achieve this goal.
- Interactive data exploration with large language models. The existing suite of deep-learning methods has markedly changed the relationship between life scientists and imaging data in a fundamental way; large language models can disrupt this relationship even further[24]. One aspect of this relationship that is ripe for disruption is how scientists explore and interact with biological imaging data. Large language models can enable the exploration of large datasets with natural language, removing a long-standing computational barrier. They could also substantially reduce the software engineering experience that is required to generate user

interfaces, which will reduce the work that is required for methods developers to make their work accessible.

- AI-enabled measurements. Many existing applications of deep-learning methods seek to automate analyses that are typically performed by classical computer vision methods. Although this can yield substantial time savings for the life science workforce, there is a unique opportunity at the intersection of deep learning and experimental design. Now that we know what deep-learning methods can do, can we integrate them into the measurements to design better, more scalable experiments? We believe the answer is yes, as compelling implementations of this idea already exist in the literature[25]. This path could be the key to providing the scale to our measurements that is necessary to understand how cells are organized and function.
- A national laboratory for AI. Although AI methods can be transformative, the data, computational and software engineering requirements that they require to be used at scale can place them outside the reach of many research laboratories. This has been particularly true for the development of large language models, which hold immense promise for biological data. This gap creates a disconnect between organizations with AI capabilities and the full marketplace of ideas that exists in academic settings. One potential path to bridge this gap would be a national laboratory for AI. Such an organization could democratize access to AI-centered computational and software engineering resources throughout academia (including in the life sciences), thus ensuring broader participation in the AI revolution.

These are our thoughts — but we believe the most impactful answers to the question of what comes next will be provided by the next generation of young scientists, who can move seamlessly between concepts that span modern biology, imaging technologies and AI methods.

Morgan Schwartz[1], Uriah Israel[1], Xuefei (Julie) Wang[1], Emily Laubscher[2], Changhua Yu[1], Rohit Dilip[3], Qilin Li[4], Joud Mari[1], Johnathon Soro[1], Kevin Yu[1], Elora Pradhan[1], Ada Ates[1], Danielle Gallandt[1], Ross Barnowski[1], Edward Pao[1] & David Van Valen[1]

[1]Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA, USA. [2]Department of Chemistry, California Institute of Technology, Pasadena, CA, USA. [3]Department of Computer Science, California Institute of Technology, Pasadena, CA, USA. [4]Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA.

e-mail: vanvalen@caltech.edu

Published online: 11 July 2023

References

1. Moffitt, J. R., Lundberg, E. & Heyn, H. *Nat. Rev. Genet.* **23**, 741–759 (2022).
2. Moses, L. & Pachter, L. *Nat. Methods* **19**, 534–546 (2022).
3. Schermelleh, L. et al. *Nat. Cell Biol.* **21**, 72–84 (2019).
4. Moen, E. et al. *Nat. Methods* **16**, 1233–1246 (2019).
5. Greenwald, N. F. et al. *Nat. Biotechnol.* **40**, 555–565 (2022).
6. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. *Nat. Methods* **18**, 100–106 (2021).
7. Lugagne, J.-B., Lin, H. & Dunlop, M. J. *PLoS Comput. Biol.* **16**, e1007673 (2020).
8. Ouyang, W., Aristov, A., Lelek, M., Hao, X. & Zimmer, C. *Nat. Biotechnol.* **36**, 460–468 (2018).
9. Weigert, M. et al. *Nat. Methods* **15**, 1090–1097 (2018).
10. Batson, J. & Royer, L. Noise2self: blind denoising by self-supervision. In *Int. Conf. Mach. Learn.* **97** (eds Chaudhuri, K. & Salakhutdinov, R.) 524–533 (PMLR, 2019).
11. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. *Nat. Methods* **15**, 917–920, https://doi.org/10.1038/s41592-018-0111-2 (2018).
12. Christiansen, E. M. et al. *Cell* **173**, 792–803.e19 (2018).
13. Saka, S. K. et al. *Nat. Biotechnol.* **37**, 1080–1090 (2019).
14. Shah, S., Lubeck, E., Zhou, W. & Cai, L. *Neuron* **92**, 342–357 (2016).
15. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. *Science* **348**, aaa6090 (2015).
16. Lu, T., Ang, C. E. & Zhuang, X. *Cell* **185**, 4448–4464.e17 (2022).
17. Feldman, D. et al. *Cell* **179**, 787–799.e17 (2019).
18. Reicher, A., Koren, A. & Kubicek, S. *Genome Res.* **30**, 1846–1855 (2020).
19. Vaswani, A. et al. Attention is all you need. In *Adv. Neural Inf. Process. Syst.* **30** (eds. Guyon, I. et al.) (Curran Associates, 2017).
20. Kirillov, A. et al. Preprint at https://doi.org/10.48550/arXiv.2304.02643 (2023).
21. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Adv. Neural Inf. Process. Syst.* **32** (eds Wallach, H. et al.) (Curran Associates, 2019).
22. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In *Adv. Neural Inf. Process. Syst.* **33** (eds. Larochelle, H. et al.) 6840–6851 (Curran Associates, 2020).
23. Song, Y. et al. Preprint at https://doi.org/10.48550/arXiv.2011.13456 (2021).
24. Brown, T. et al. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst.* **33** (eds. Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
25. Nitta, N. et al. *Cell* **175**, 266–276.e13 (2018).

Acknowledgements

This Comment is the result of numerous interactions we have had over the years with many bright colleagues, and this space is too short to name them all. We owe tremendous thanks to past and current laboratory members, as many of the ideas described here touch on the idea space that they have explored over the past five years. We thank P. Blainey, I. Cheeseman and M. Leonetti for hosting a recent workshop on 'Cell Biology at Scale' that strongly shaped this piece. We also thank several organizations for supporting the work of D.V.V.'s laboratory, including the Shurl and Kay Curci Foundation, the Rita Allen Foundation, the Pew Charitable Trusts, the Alexander and Margaret Stewart Trust, the Gordon and Betty Moore Foundation, the Aligning Science Across Parkinson's consortium, the Heritage Medical Research Institute, the NIH through the DP2 program and the HuBMAP consortium, and the Howard Hughes Medical Institute through the Freeman Hrabowski Scholar's program.

Author contributions

M.S., U.I., X.(J.)W., C.Y., E.L., R.D., Q.L., J.M., J.S., K.Y., E.P., A.A., D.G., R.B., E.P. and D.V.V. conceived the research directions described in the manuscript. D.V.V. wrote the manuscript, with contributions from all authors. All authors read and approved the manuscript.

Competing interests

D.V.V. is a co-founder and chief scientist of Barrier Biosciences and holds equity in the company. All other authors declare no competing interests.