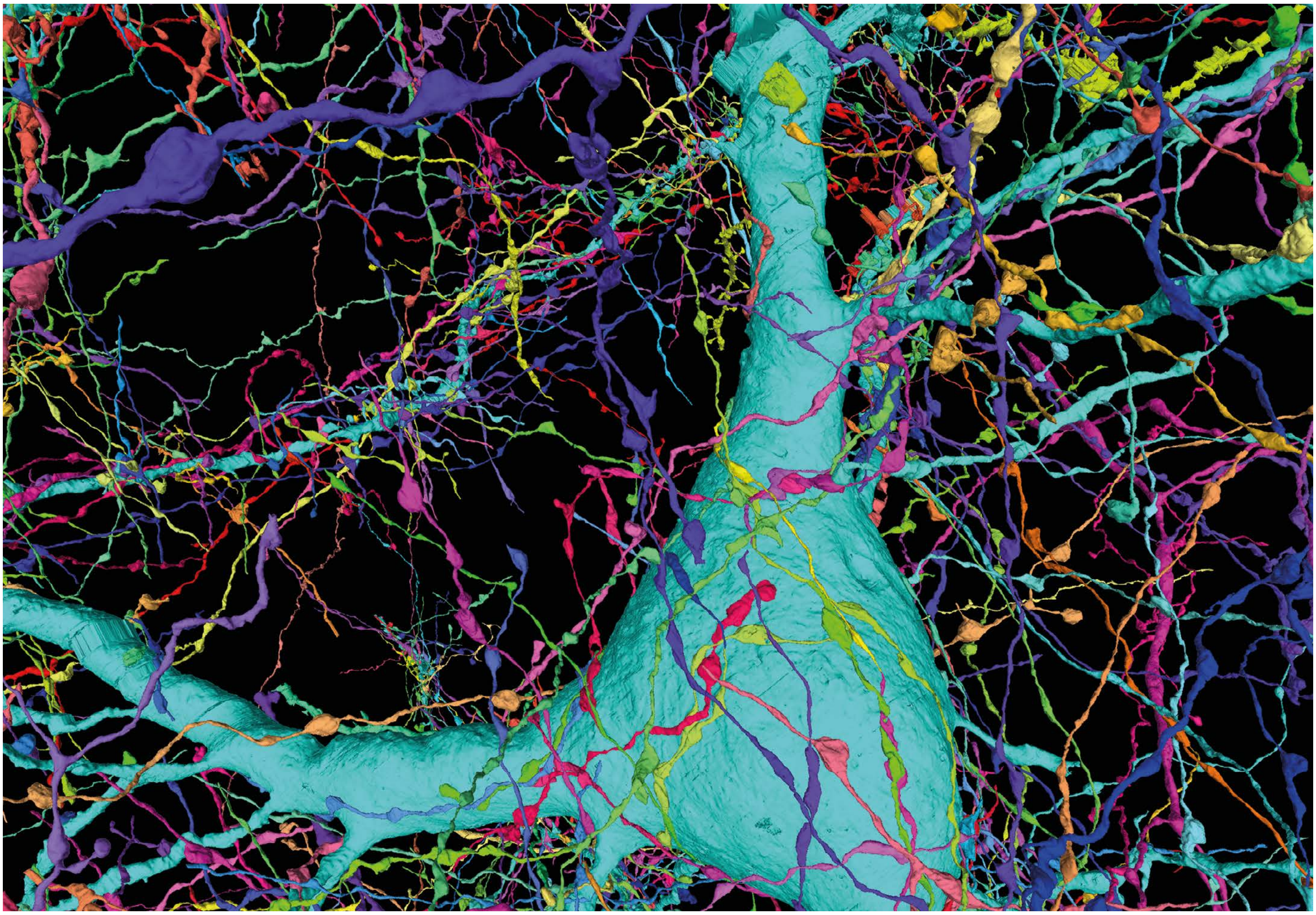


LICHTMAN LAB, HARVARD/JAIN GROUP, GOOGLE

Neurons (teal) can have thousands of connections to other cells.

FIVE WAYS DEEP LEARNING HAS TRANSFORMED IMAGE ANALYSIS

From connectomics to behavioural biology, artificial intelligence is making it faster and easier to extract information from images. **By Sandeep Ravindran**

One cubic millimetre doesn't sound like much. But in the human brain, that volume of tissue contains some 50,000 neural 'wires' connected by 134 million synapses. Jeff Lichtman wanted to trace them all.

To generate the raw data, he used a protocol known as serial thin-section electron microscopy, imaging thousands of slivers of tissue over 11 months. But the data set was enormous, amounting to 1.4 petabytes – the equivalent of about 2 million CD-ROMs – far too much for researchers to handle on their own. "It is simply impossible for human beings to manually trace out all the wires," says Lichtman, a molecular and cell biologist at Harvard University in Cambridge, Massachusetts. "There are not enough people on Earth to really get this job done in an efficient way."

It's a common refrain in connectomics – the study of the brain's structural and functional connections – as well as in other biosciences, in which advances in microscopy are creating a deluge of imaging data. But where human resources fail, computers can step in, especially deep learning algorithms that have been optimized to tease out patterns from large data sets.

"We've really had a Cambrian explosion of tools for deep learning in the past few years."

"We've really had a Cambrian explosion of tools for deep learning in the past few years," says Beth Cimini, a computational biologist at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts.

Deep learning is an artificial-intelligence (AI) technique that relies on many-layered artificial neural networks inspired by how neurons interconnect in the brain. Based as they are on black-box neural networks, the algorithms have their limitations. Those include a dependence on massive data sets to teach the network how to identify features of interest, and a sometimes inscrutable way of generating results. But a fast-growing array of open-source and web-based tools is making it easier than ever to get started (see 'Taking the leap into deep learning').

Here are five areas in which deep learning is having a deep impact in bioimage analysis.

Large-scale connectomics

Deep learning has enabled researchers to generate increasingly complex connectomes from fruit flies, mice and even humans. Such data can help neuroscientists to understand how the brain works, and how its structure changes during development and in disease. But neural connectivity isn't easy to map.

In 2018, Lichtman joined forces with Viren Jain, head of Connectomics at Google in Mountain View, California, who was looking for a suitable challenge for his team's AI algorithms.

"The image analysis tasks in connectomics are very difficult," Jain says. "You have to be able to trace these thin wires, the axons and dendrites of a cell, across large distances, and conventional image-processing methods made so many mistakes that they were basically useless for this task." These wires can be thinner than a micrometre and extend over hundreds of micrometres or even millimetres of tissue. Deep-learning algorithms provide a way to automate the analysis of connectomics data while still achieving high accuracy.

In deep learning, researchers can use annotated data sets containing features of interest to train complex computational models so that they can quickly identify the same features in other data. "When you do deep learning, you say, 'okay, I will just give examples and you figure everything out'," says Anna Kreshuk, a computer scientist at the European Molecular Biology Laboratory in Heidelberg, Germany.

But even using deep learning, Lichtman and Jain had a herculean task in trying to map their snippet of the human cortex[1]. It took 326 days just to image the 5,000 or so extremely thin sections of tissue. Two researchers spent about 100 hours manually annotating the images and tracing neurons to create 'ground truth' data sets to train the algorithms, in an approach known as supervised machine learning. The trained algorithms then automatically stitched the images together and identified neurons and synapses to generate the final connectome.

Jain's team brought massive computational resources to bear on the problem, including thousands of tensor processing units (TPUs), Google's in-house equivalent to graphics processing units (GPUs) built specifically for neural-network machine learning. Processing the data required on the order of one million TPU hours over several months, Jain says, after which human volunteers proofread and corrected the connectome in a collaborative process, "sort of like Google Docs", says Lichtman.

The end result, they say, is the largest such data set reconstructed at this level of detail in any species. Still, it represents just 0.0001% of

Taking the leap into deep learning

Plenty of resources are available to help researchers get up to speed.

Organizations such as the Woods Hole Oceanographic Institute in Massachusetts and NEUBIAS, the global Network of European BioImage Analysts, offer courses on how to get started. And the Center for Open Bioimage Analysis, a collaboration between the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, and the University of Wisconsin–Madison sponsors image.sc, a discussion forum about scientific-image software. Researchers can also comb old Kaggle challenges — computational competitions for scientists and AI enthusiasts — for examples of models and data that they can practise with and learn from. "All the data and the training sets are available, and you can look at the code and descriptions for the winning models, so it's a very good starting point," says Emma Lundberg, a bioengineer at Stanford University in California.

Researchers might also want to start with pre-trained models from tool sets such as Cellpose, StarDist and DeepCell, which can be used through web interfaces, as plug-ins for the ImageJ and napari software ecosystems, or as standalone applications. "They've trained models that work pretty well for a good fraction of use cases," says Beth Cimini, a computational biologist at the Broad Institute. "You don't really need to know what they're doing or understand how a deep-learning network works, you just kind of tweak the knobs until you get a good result." For those who require greater customizability, Piximi and ImJoy allow researchers to train their own neural networks to identify various phenotypes, and to locate cells in images, a process known as segmentation.

Most such tools can be run in a browser. ZeroCostDL4Mic, an open-source toolbox for deep learning in microscopy, uses Google's computational-notebook platform Colab and allows researchers to train various popular open-source models in the cloud, as well as access pre-trained models that can be run in the cloud[9]. There's also the BioImage Model Zoo, a one-stop shop for open-source pre-trained models for popular-use cases.

Alternatively, researchers can install and run dedicated software. For instance, ilastik has a point-and-click interface to help detect not just cells and nuclei but also features such as microtubules and vesicles. Co-developer Anna Kreshuk, a computer scientist at the European Molecular Biology Laboratory in Heidelberg, Germany, and her colleagues are now working to improve the software's ability to train neural networks for tasks such as classification and segmentation. "Everybody needs segmentation," she says, "but everyone is segmenting different things." A training feature is already available in an unsupported debug mode.

Learning to program, particularly in Python, can help researchers who want to customize or train new models. "This will really give you an edge, like being able to manipulate your data more freely to apply methods that people have not specifically packaged for you in the best possible way," says Kreshuk. Also helpful will be one or more graphics processing units and computers capable of using them.

But neither software nor hardware matters as much as the data. "The hardest and the most time-consuming part of any deep learning is acquiring training data. And if your data's crappy, then your model's going to be crappy," says Cimini. "You typically need hundreds or thousands of examples at minimum, and creating the annotations itself is tedious."

Data sets ideally should be large and diverse, and it helps if humans can unambiguously identify whatever the deep-learning model is being asked to find. "People kind of expect that these models can just perform miracles, but if the information that you want to pull out isn't there in the data, then in my view and also in my experience, it's unlikely to work," says David Van Valen, a bioengineer at the California Institute of Technology in Pasadena.

Deep-learning algorithms effectively operate as black boxes, but some tools can provide clues to their reasoning. "You can tell, for example, which part of an image was most important in making a particular decision," says Cimini.

For now, unambiguous but tedious tasks such as identifying cells or nuclei are ideal, because humans can easily verify the results. But as algorithms improve, the scale and scope of researchers' ambitions will change, too. "It's a really exciting field," Cimini says. "I think it's going to make a lot of people's lives easier."

the human brain. But as algorithms and hardware improve, researchers should be able to map ever larger portions of the brain, while having the resolution to spot more cellular features, such as organelles and even proteins. “In some ways,” says Jain, “we are just scratching the surface of what might be possible to extract from these images.”

Virtual histology

Histology is a key tool in medicine, and is used to diagnose disease on the basis of chemical or molecular staining. But it's laborious, and the process can take days or even weeks to complete. Biopsies are sliced into thin sections and stained to reveal cellular and sub-cellular features. A pathologist then reads the slides and interprets the results. Aydogan Ozcan reckoned he could accelerate the process.

An electrical and computer engineer at the University of California, Los Angeles, Ozcan trained a custom deep-learning model to stain a tissue section computationally by presenting it with tens of thousands of examples of both unstained and stained versions of the same section, and letting the model work out how they differed.

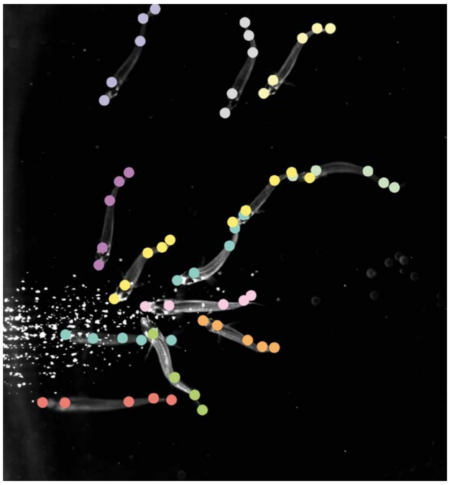
Virtual staining is almost instantaneous, and board-certified pathologists found it almost impossible to distinguish the resulting images from conventionally stained ones[2]. Ozcan has also shown that the algorithm can replicate a molecular stain for the breast cancer biomarker HER2 in seconds, a process that typically takes at least 24 hours in a histology lab. A panel of three board-certified breast pathologists rated the images as having comparable quality and accuracy to conventional immunohistochemical staining[3].

Ozcan, who aims to commercialize virtual staining, hopes to see applications in drug development. But by eliminating the need for toxic dyes and expensive staining equipment, the technique could also increase access to histology services worldwide, he says.

Cell finding

If you want to extract data from cellular images, you have to know where in the images the cells actually are.

Researchers usually perform this process, called cell segmentation, either by looking at cells under the microscope or outlining them in software, image by image. “The word that most describes what people have been doing is ‘painstaking’,” says Morgan Schwartz, a computational biologist at the California Institute of Technology in Pasadena, who is developing deep-learning tools for bioimage analysis. But these painstaking approaches are hitting a wall as imaging data sets become ever larger. “Some of these experiments you just couldn't analyse without automating the process.”



Annotation of fish for DeepLabCut training.

J. LAUER ET AL. NATURE METHODS 19, 496–504 (2022).

Schwartz's graduate adviser, bioengineer David Van Valen, has created a suite of AI models, available at deepcell.org, to count and analyse cells and other features from images both of live cells and of preserved tissue. Working with collaborators including Noah Greenwald, a cancer biologist at Stanford University in California, Van Valen developed a deep-learning model called Mesmer to quickly and accurately detect cells and nuclei across different tissue types[4]. “If you've got data that you need processed, now you can just upload them, download the results and visualize them either within the web portal or using other software packages,” Van Valen says.

> “We have for decades been generating millions of images, outlining the protein expression in cells.”

According to Greenwald, researchers can use such information to differentiate cancerous from non-cancerous tissue and to search for differences before and after treatment. “You can look at the imaging-based changes to have a better idea of why some patients respond or don't respond, or to identify subtypes of tumours,” he says.

Mapping protein localization

The Human Protein Atlas project exploits yet another application of deep learning: intracellular localization. “We have for decades been generating millions of images, outlining the protein expression in cells and tissues of the human body,” says Emma Lundberg, a bioengineer at Stanford University and a co-manager of the project. At first, the project annotated those images manually. But because that approach wasn't sustainable long term, Lundberg turned to AI.

Lundberg first combined deep learning with citizen science, tasking volunteers with annotating millions of images while playing a massively multiplayer game, EVE Online[5]. Over the past few years, she has switched to a crowdsourced AI-only solution, launching Kaggle challenges – in which scientists and AI enthusiasts compete to achieve various computational tasks – of US$37,000 and $25,000, to devise supervised machine-learning models to annotate protein-atlas images. “The Kaggle challenge afterwards blew the gamers away,” Lundberg says. The winning models outperformed Lundberg's previous efforts at multi-label classification of protein-localization patterns by about 20% and were generalizable across cell lines[6]. And they managed something no published models had done before, she adds, which was to accurately classify proteins that exist in multiple cellular locations.

“We have shown that half of all human proteins localized to multiple cellular compartments,” says Lundberg. And location matters, because the same protein might behave differently in different places. “Knowing if a protein is in the nucleus or in the mitochondria, it helps understand lots of things about its function,” she says.

Tracking animal behaviour

Mackenzie Mathis, a neuroscientist at the Campus Biotech hub of the Swiss Federal Institute of Technology, Lausanne, in Geneva, has long been interested in how the brain drives behaviour. She developed a program called DeepLabCut to enable neuroscientists to track animal poses and fine movements from videos, turning ‘cat videos’ and recordings of other animals into data[7].

DeepLabCut offers a graphical user interface so that scientists can upload and label their videos and train a deep-learning model at the click of a button. In April, Mathis's team expanded the software to estimate poses for multiple animals at the same time, something that's typically been challenging for both humans and AI[8].

Applying multi-animal DeepLabCut to marmosets, the researchers found that when the animals were in close proximity, their bodies were aligned and they tended to look in similar directions, whereas they tended to face each other when apart. “That's a really good case where pose actually matters,” Mathis says. “If you want to understand how two animals are interacting and looking at each other or surveying the world.”

Sandeep Ravindran is a science writer based in Bethesda, Maryland.

1. Shapson-Coe, A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.05.29.446289 (2021).
2. Rivenson, Y. *et al. Nature Biomed. Eng.* **3**, 466–477 (2019).
3. Bai, B. *et al. BMEF Front.* (in the press).
4. Greenwald, N. F. *et al. Nature Biotechnol.* **40**, 555–565 (2022).
5. Sullivan, D. P. *et al. Nature Biotechnol.* **36**, 820–828 (2018).
6. Ouyang, W. *et al. Nature Methods* **16**, 1254–1261 (2019).
7. Mathis, A. *et al. Nature Neurosci.* **21**, 1281–1289 (2018).
8. Lauer, J. *et al. Nature Methods* **19**, 496–504 (2022).
9. von Chamier, L. *et al. Nature Commun.* **12**, 2276 (2021).