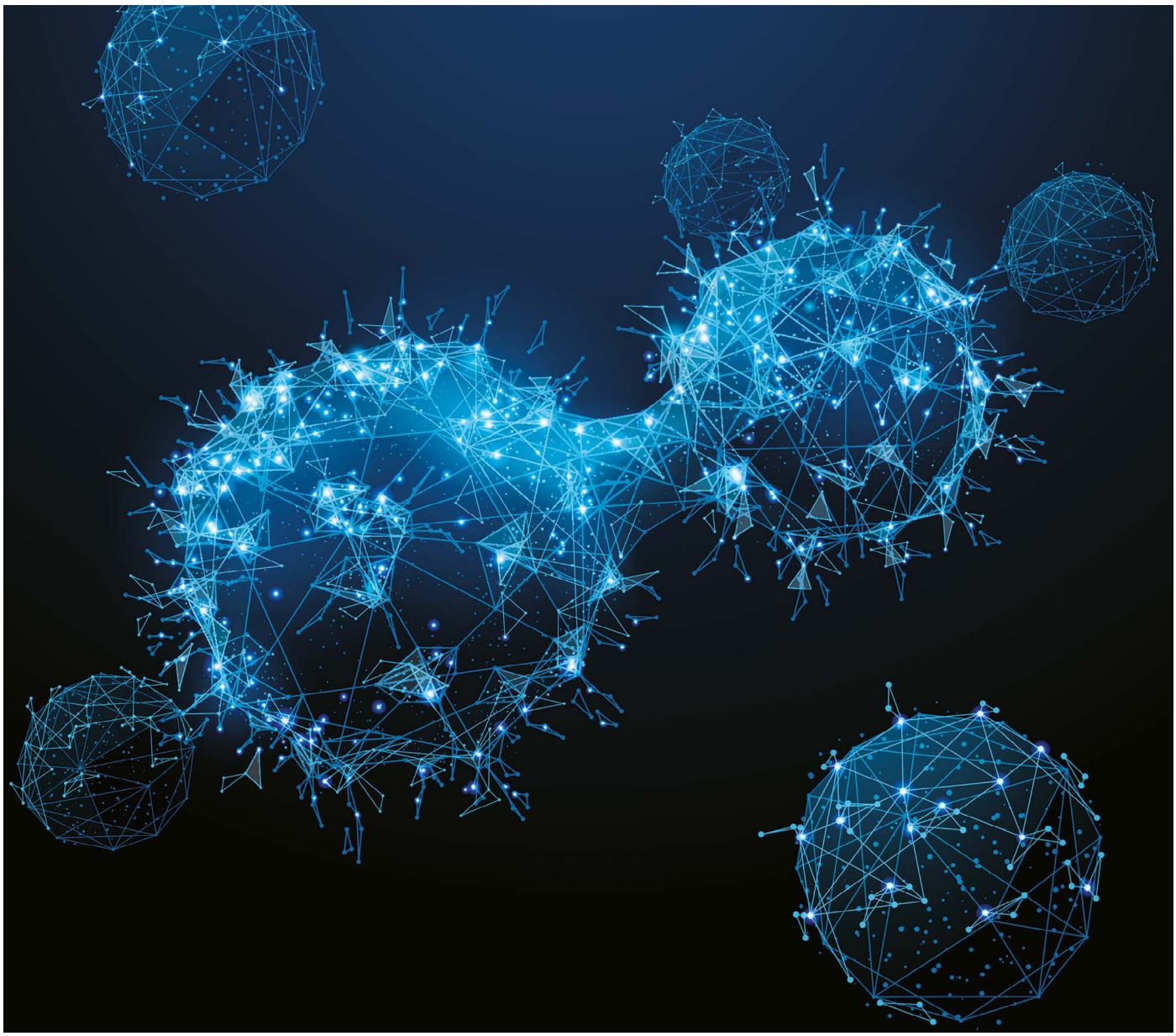


SHUTTERSTOCK

DEEP LEARNING TAKES ON TUMOURS

Artificial-intelligence methods are moving into cancer research. **By Esther Landhuis**

As cancer cells spread in a culture dish, Guillaume Jacquemet is watching. The cell movements hold clues to how drugs or gene variants might affect the spread of tumours in the body, and he is tracking the nucleus of each cell in frame after frame of time-lapse microscopy films. But because he has generated about 500 films, each with 120 frames and 200–300 cells per frame, that analysis is challenging to say the least. "If I had to do the tracking manually, it would be impossible," says Jacquemet, a cell biologist at Åbo Akademi University in Turku, Finland.

So he has trained a machine to spot the nuclei instead. Jacquemet uses methods available on a platform called ZeroCostDL4Mic, part of a growing collection of resources aimed at making artificial intelligence (AI) technology accessible to bench scientists who have minimal coding experience[1].

AI technologies encompass several methods. One, called machine learning, uses data that have been manually preprocessed and makes predictions according to what the AI learns. Deep learning, by contrast, can identify complex patterns in raw data. It is used in self-driving cars, speech-recognition software, game-playing computers — and to spot cell nuclei in massive microscopy data sets.

Deep learning has its origins in the 1940s, when scientists built a computer model that was organized in interconnected layers, like neurons in the human brain. Decades later, researchers taught these 'neural networks' to recognize shapes, words and numbers. But it wasn't until about five years ago that deep learning began to gain traction in biology and medicine.

A major driving force has been the explosive growth of life-sciences data. With modern gene-sequencing technologies, a single experiment can produce gigabytes of information. The Cancer Genome Atlas, launched in 2006, has collected information on tens of thousands of samples spanning 33 cancer types; the data exceed 2.5 petabytes (1 petabyte is 1 million gigabytes). And advances in tissue labelling and automated microscopy are generating complex imaging data faster than researchers can possibly mine them. "There's definitely a revolution going on," says Emma Lundberg, a bioengineer at the KTH Royal Institute of Technology in Stockholm.

Boosting image-based profiling

Cancer biologist Neil Carragher caught his first glimpse of this revolution in 2004. He was leading a team at AstraZeneca in Loughborough, UK, that explores new technologies for the life sciences, when he came across a study that made the company rethink its drug-screening efforts. He and his team had been using cell-based screens to look for promising drug candidates, but hits were hard to come by. The study was suggesting that AI and analytics could help them to improve their screening processes[2]. "We thought this could be a solution to the productivity crisis," Carragher says.

But AI technologies can be difficult for biologists to master. Jacquemet says he once spent more than a week trying to install the correct software libraries to run a deep-learning model. Then, he says, "you need to learn to code in Python" to use it.

Carragher's AstraZeneca team worked with computational biologist Anne Carpenter and her colleagues at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, to scale up the image-profiling method used in the 2004 paper and to investigate the effects of multiple drugs on human breast-cancer cells[3]. Carpenter went on to develop the technique into a procedure called Cell Painting, which stains cells with a panel of fluorescent dyes and then uses the open-source software CellProfiler to generate profiles of the cells.

> **"If I had to do the tracking manually, it would be impossible."**

Still, these analyses can be labour-intensive, says Carragher, who now heads cancer-drug discovery at the University of Edinburgh, UK. Even with open-source tools that avoided the need to code the machine learning from scratch — and a computing cluster with thousands of processors and terabytes of memory — it could take a month or so to work out which cellular features they should tell the image-analysis software to look at, Carragher says. And after optimizing the parameters for each cell line, his team had to tinker further to get it to work across all cells.

Last year, he and his team explored how deep learning could improve this process. The impetus was a 2017 analysis[4] posted on the bioRxiv preprint server by researchers at Google's headquarters in Mountain View, California. The researchers had downloaded Carragher's breast-cancer data set from the Broad Bioimage Benchmark Collection and used it to train a deep neural network that previously had seen only general images, such as cars and animals. By scanning for patterns in the breast-cancer data, the model learnt to discern cellular changes that are meaningful for drug discovery. Because the software wasn't told what to look for, it found features that researchers hadn't even considered.

Building on that effort, Carragher and his colleagues screened 14,000 compounds across 8 forms of breast cancer[5]. "We did identify some interesting hits," he says — including a compound that was already known to modulate receptors for serotonin, which is important in mammary-gland development, as they reported earlier this year[6].

At the Broad Institute, a team led by computational biologist Juan Caicedo is applying image-based profiling to screen for genetic mutations. He and his team overexpressed various gene variants in lung-cancer cells, stained them with the Cell Painting protocol and looked for differences in the cells that suggest possible pharmaceutical opportunities. They found that machine learning could identify meaningful variants in images about as well as processes that measure gene expression in the cells. The researchers reported their results at the AI Powered Drug Discovery and Manufacturing Conference in February at the Massachusetts Institute of Technology in Cambridge.

As part of the Cancer Cell Map Initiative, which maps molecular networks underlying human cancer, researchers are training a deep-learning model to predict drug responses on the basis of a person's cancer-genome sequence. Such predictions have life-or-death implications, and accuracy is crucial, says Trey Ideker, a bioengineer at the University of California, San Diego. But some are reluctant to accept results when the mechanisms behind them aren't clear, and deep neural networks produce answers without revealing their process — a problem known as 'black-box' learning. "You want to know why," says Ideker. "You want to know the mechanism." Ideker's team is creating

WANTED: MORE DATA

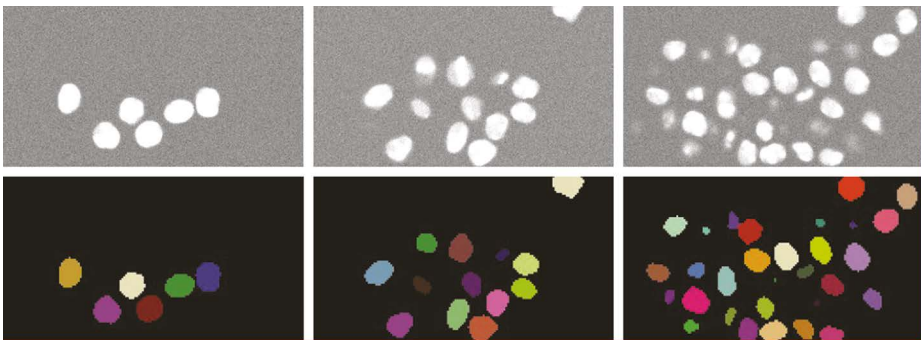
Deep-learning models can process raw data, but first they must be trained with annotated information.

It takes vast amounts of labelled data to train deep-learning models. But that's not always easy to come by, says Casey Greene, a computational biologist at the University of Pennsylvania in Philadelphia. "Data are cheap, but labelled data are expensive."

In the genomics realm, sequences are abundant and publicly available. But their associated descriptions, or metadata, are often missing, wrong or unstandardized, says Emily Flynn, a doctoral candidate in biomedical informatics at Stanford University in California. A researcher wanting to train a model to detect non-small-cell lung cancer in samples from patients, for example, might well find data sets variously labelled 'nsclc', 'non small-cell' or 'non small cell LC' — differences that confound analysis tools. Or samples might be labelled 'disease: glioblastoma' and 'disease: yes', says biostatistician Colin Dewey at the University of Wisconsin–Madison.

To help organize those data, Dewey created a computational pipeline called MetaSRA, which uses text-mining techniques to standardize and store metadata on public sequences. And Greene and colleagues have built refine.bio, a repository that harmonizes data on expression and RNA sequencing. Working with Stanford bioengineer Russ Altman, Flynn is using machine-learning techniques to infer missing labels from gene-expression data to improve annotations in refine.bio.

In bioimaging, the problem lies more in annotation. To label a set of histopathology slides, for example, "someone has to go in and draw a bounding box around the parts that are cancer", Greene says. "And that person probably makes a lot of money." Now developers are training deep-learning algorithms to label nuclei and other structures in cell images, while the Image Data Resource and other online repositories are making it easier for researchers to share and find life-sciences images.



C. MCQUIN *ET AL./PLOS BIOL.* (CC BY 4.0)

Cell nuclei (top, DNA stain) are automatically detected using the CellProfiler method (bottom).

a 'visible' neural network, which links the model's inner workings more directly to cancer cell biology. As a proof of concept, the team created a model for yeast cells. Called DCell, it can predict the effects of gene mutations on cell growth and the molecular pathways underlying those effects[7].

The spatial dimension

Lundberg and others in Sweden are using deep learning to tackle another computational challenge: assessing protein localization. The work is part of the Human Protein Atlas, a multi-year, multi-omics effort to map all human proteins. Spatial information reveals where proteins are located in cells, and tend to be under-represented in systems-level studies, Lundberg says. But if researchers knew this information, they could use it to glean insights about the underlying biology, she suggests.

Enter AI. In 2016, Lundberg and her colleagues invited gamers to help computers classify proteins' whereabouts in cells. The citizen scientists took part in a role-playing game called EVE Online, in which they had to pinpoint fluorescently labelled proteins to win game credits, boosting an AI system already used for this purpose. But even the augmented system trailed human experts in terms of accuracy and speed.

So, in 2018, Lundberg's team took its images to Kaggle — a platform that challenges machine-learning experts to develop their best models to crack data sets posted by companies and researchers. Over the course of 3 months, 2,172 teams around the world competed to develop a deep-learning model that could look at a cell stained for a protein and several reference markers, and work out the protein's spatial distribution.

The task was challenging. Half of human proteins are found in multiple places in cells, says Lundberg. And some cellular compartments — the nucleus, for example — are much more common locations than others.

Still, the Kagglers delivered, Lundberg says. Most of the leading strategies came from computational scientists with no biology background — including Bojan Tunguz, a software engineer who created models that predict earthquakes and loan defaults before earning one of the top spots in the Human Protein Atlas contest. The approach to these problems is similar across vastly different disciplines, Tunguz says.

The best model identified both rare and common locations across a variety of cell lines and, most importantly, captured mixed patterns well, Lundberg says. The algorithm performed almost as accurately as human experts, and with greater speed and reproducibility. Furthermore, it could quantify the spatial information[8]. "When we can quantify it, and not just describe it with a label, we can integrate it with other types of data." That includes 'omics' data, which are already transforming cancer research.

A computational framework known as DeepProg applies deep learning to 'omics' data sets, including gene expression and epigenetic data, to predict patient survival, for instance[9]. And DigitalDLSorter predicts outcomes by inferring types and quantities of immune cells directly from tumour-RNA sequencing data rather than relying on laborious conventional workflows[10].

"When we can quantify it, and not just describe it with a label, we can integrate it with other types of data."

On the horizon

Many of the tools needed to build deep-learning models are freely available online, including software libraries and coding frameworks such as TensorFlow, Pytorch, Keras and Caffe. Researchers wanting to ask questions and brainstorm solutions to problems that crop up with image-analysis tools can make use of an online resource called the Scientific Community Image Forum (https://forum.image.sc). Also becoming available are repositories that allow researchers to find and repurpose deep-learning models for related tasks — a process called transfer learning. One example is Kipoi, which allows researchers to search and explore more than 2,000 ready-to-use models trained for tasks such as predicting how proteins known as transcription factors will bind to DNA, or where enzymes are likely to splice the genetic code.

Working with other tool developers, Lundberg's team put together a rudimentary 'model zoo' (https://bioimage.io) to quickly share its Human Protein Atlas models, and is now creating a more sophisticated repository that will be useful to model producers and non-expert users alike.

A platform called ImJoy will be part of this effort, Lundberg says. Created by Wei Ouyang, a postdoc in her lab, the platform lets researchers test and run AI models through a web browser on their computer, in the cloud or on a phone. Sharing bioimaging data sets and deep-learning models will also be a priority for the Center for Open Bioimage Analysis, an effort funded by the US government and led by Carpenter and Kevin Eliceiri, a bioengineer at the University of Wisconsin–Madison.

Another option, ZeroCostDL4Mic, launched last month. Developed by biophysicist Ricardo Henriques at University College London, ZeroCostDL4Mic makes use of Colab, Google's free cloud service for AI developers, to provide access to several popular deep-learning microscopy tools, including the one Jacquemet uses to automate cell-nuclei labelling in his films. "Everything you need is installed within a couple of minutes," Jacquemet explains. With a few mouse clicks, users can use example data to train a neural network to complete the desired task (see 'Wanted: more data'), then apply that network to their own data — all without needing to code.

Researchers who want to use larger data sets or train more-complex models might need to purchase or access extra computational resources beyond Google's free service.

By easing the way for biologists with scant know-how and resources to use deep learning, Henriques says, ZeroCostDL4Mic acts like "a gateway drug" for AI, luring researchers to explore the software underlying these tools that will continue to transform research in cancer and beyond.

Esther Landhuis is a science journalist based near San Francisco, California.

1. von Chamier, L. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.03.20.000133 (2020).
2. Perlman, Z. E. *et al. Science* **306**, 1194–1198 (2004).
3. Ljosa, V. *et al. J. Biomol. Screen.* **18**, 1321–1329 (2013).
4. Ando, D. M., McLean, C. Y. & Berndl, M. Preprint at bioRxiv https://doi.org/10.1101/161422 (2017).
5. Warchal, S. J., Dawson, J. C. & Carragher, N. O. *SLAS Discov.* **24**, 224–233 (2019).
6. Warchal, S. J. *et al. Bioorg. Med. Chem.* **28**, 115209 (2020).
7. Ma, J. *et al. Nature Meth.* **15**, 290–298 (2018).
8. Ouyang, W. *et al. Nature Meth.* **16**, 1254–1261 (2019).
9. Poirion, O. B., Chaudhary, K., Huang, S. & Garmire, L. X. Preprint at medRxiv https://doi.org/10.1101/19010082 (2019).
10. Torroja, C. & Sanchez-Cabo, F. *Front. Genet.* **10**, 978 (2019).