LUIS ALVAREZ/GETTY

Artificial-intelligence programs can speed up monotonous tasks in research — and the learning curve is not too steep.

SPICE UP YOUR BIOINFORMATICS SKILL SET WITH AI

Incorporating machine-learning tools into data analysis can accelerate discovery and free up valuable time. **By Rachael Pells**

Image-analysis tools can do amazing things. Yet despite their power, Fernanda Garcia Fossa was frustrated. A biology PhD student at the State University of Campinas, Brazil, Garcia Fossa specializes in nanotoxicology. Image-based profiling of human cells is a core part of her research. But when she started out, the process was slow and error-prone.

"I spent a lot of my time analysing my images individually by hand, looking for differences and patterns," Garcia Fossa explains. She was looking for evidence of the subtle effects of silver nanoparticles on liver cells. But the number of hours it took to compare scanned images of each cell one by one was overwhelming, she says. "I thought, there has to be a faster way to do this."

Trawling online biology forums, she stumbled on CellProfiler, an image-analysis tool based on artificial intelligence (AI) developed at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. Within hours, she had identified an algorithm tailored to her needs, which she used to analyse her images automatically. "It was exciting," she says. "Suddenly, I found I had more time to do other tasks related to my research, because the program was analysing all my images for me."

She's not alone; bioinformatics skills have become essential in the life sciences. Scientists are typically trained on the algorithms that drive that research — how they work and how to use them efficiently. But informaticians are increasingly using machine learning or AI — including large language models, such as the

ChatGPT chatbot — rather than algorithms to find patterns or features in sequences and images.

Uptake is growing fast, but it could be faster, says Shantanu Singh, a data scientist and senior group leader at the Broad Institute's Imaging Platform. Although a large number of researchers are now working with these platforms, many lack data-management skills — which, coupled with a shortage of resources, is holding the field back. "Some things, like data-storage solutions, are getting simpler — but it's still not enough," he says.

Those who have already made the transition to using AI are reaping the benefits of vastly accelerated workflows and targeted decision-making in data analysis. But for bioinformaticians who remain on the fence, there are challenges to consider when taking the leap.

Get familiar with AI tools

Image-analysis algorithms help researchers to compare cell characteristics faster and more quantitatively than when they do the work manually; AI further accelerates the process through adaptive learning that is specific to the researcher's needs. AI can often detect differences or modes of comparison that the user had never considered. "The benefit of bringing AI into imaging is that it allows researchers to reason with biological images in high dimensions, not just focus on one or two predefined measurements," explains Singh. By converting what it 'sees' into numerical data, AI effectively transforms a biologically complicated image into a relatively straightforward mathematics problem. "Once you have those numbers, the rest of it is all data science."

CellProfiler, for example, is an online open-source tool that allows users to set up their own workflows — often called pipelines — to automate their analyses (for example, quantifying shapes, characteristics or patterns). It can run machine-learning algorithms from companion tools such as CellProfiler Analyst, and is evolving to also use deep learning — a richer, more complex approach to recognizing intricate patterns in data.

According to Beth Cimini, CellProfiler's project lead, integrating deep learning into tools such as CellProfiler is the natural next step for image-based research. Deep learning and image analysis have been used together "for as long as we've had the computational abilities to do so", she says — whether that's tagging friends on Facebook and Instagram, or cleaning up photomicrographs and finding and counting objects in them.

Garcia Fossa liked CellProfiler because of its "easy interface, and the fact I didn't need to know how to code; it was just a matter of practising to get the hang of it". But several other open-source, AI-based tools have emerged for cell and image analysis in the past few years, which also require little to no coding expertise. These include ilastik, made by the Swiss Federal Institute of Technology in Zurich; QuPath, an open-source digital pathology platform developed at the University of Edinburgh, UK; and CDeep3M, from the National Center for Microscopy and Imaging Research at the University of California, San Diego.

Gaël Varoquaux, co-founder of scikit-learn.

INRIA/PHOTO B. FOURRIER

Bridge your skills gaps

Bioinformaticians who wish to build their own AI tools need to be good coders, says Gaël Varoquaux, "and by this, I mean a good software engineer — being very specific about how you track the modifications, how to do quality assurance on the code".

Varoquaux is a research director at the French National Institute for Research in Digital Science and Technology (Inria) in Paris, and co-founder of scikit-learn, a popular library of free machine-learning algorithms for the Python programming language. "Python is a generalist language," Varoquaux says: "You can do many things with it — text processing, scientific computing, web servers. It's useful for science because more often than we think we end up having to do auxiliary tasks, but also, it's good to have if ever you're looking for a job outside of academia," he notes.

To this end, he advises that knowing some software engineering and investing in those skills, as well as in your mathematics and statistics abilities, can further your career. "The foundations are important," he says. "People avoid it, but it bites them back."

That said, interactive tools, such as ChatGPT, can ease the transition, says Kyogo Kawaguchi, a research scientist at the Riken Center for Biosystems Dynamics Research in Kobe, Japan. That's because programming is challenging, both on its own and because of the skills involved, "like setting up your environment, debugging and being able to ask the questions with the correct words", he says. Chatbots lower the bar by allowing users to find solutions through experimentation and by asking candid questions.

Whatever the AI, scientists can become good at using it through a combination of formal education, self-study and practical experience. Start by exploring online tutorials and courses offered by universities and on platforms such as Coursera, edX and Udacity. Many of these are available at no cost, include step-by-step videos and can be taken in the learner's own timeframe. Andrew Ng, a computer scientist at Stanford University in California and founder of DeepLearning.AI, for example, has a popular collection of tutorials on machine- and deep-learning programming on Coursera (which he co-founded).

Live and in-person learning opportunities are also available. The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK, for example, hosts live training sessions, both in-person and online, for individuals and groups around the world. This year's five-day on-site courses will cost each attendee £825 (US$1,014), which includes four nights' accommodation and catering; five-day virtual courses usually cost £200. Course materials, on-demand training and online webinars are free and open to everyone.

The French government backs a free online course, maintained by scikit-learn, that typically takes around 35 hours to complete, says Varoquaux. "There is a lot of coding, but that's by design; we think this is useful."

Dayane Rodrigues Araújo, a scientific training officer at the EMBL-EBI, says that newcomers are often surprised by how easy it is to get started. A significant part of her work, she explains, "is getting the message out that they may not need to start from scratch with

writing an algorithm; the materials to start are already available". As a publicly funded, intergovernmental organization, the EMBL-EBI offers a bank of free resources as well as on-demand online courses that anyone can use, without restriction.

Don't panic

As with many new technologies, it might seem impossible to keep pace with AI's rapid evolution. But often, you don't have to.

Varoquaux explains that scitkit-learn uses "conventional" machine learning over deep learning because the goal of the platform is to "democratize and simplify" AI, not to compete with bigger Internet players such as Google.

But beyond this, chasing the latest technology isn't always necessary, he says. "Sure, AI evolves extremely fast. But I don't think science at large changes on a weekly basis."

"If we're trying to integrate the latest tools, we're always going to be running after the literature, and it's going to be exhausting and we're going to fail," he continues. "Better to take a step back and wait to see what emerges as the most useful."

That's prudent advice. But there are practical challenges to consider when incorporating AI into your analysis — in particular, uncertainty and natural human bias.

Virginie Uhlmann leads a bioimage-quantification research group at the EMBL-EBI, where she works on the design of AI programs for image analysis. One advantage of delegating biological-image analysis to a computer, she explains, is that it helps to mitigate our innate human limitations: "One of the things we are very, very bad at is understanding what brings us a decision; how do we determine that this is 'object A' and this is 'object B' in an image, for example."

With machine learning, she continues, "the real power is, you're not trying to determine and write the rules yourself; you're leaving it up to the machine".

But relying too heavily on the AI comes with its own risks, she warns.

Uhlmann's advice: carefully consider what the AI tells you, to understand how and why it made its decision. "There are lots of very famous examples of very dumb decision-making that somehow leads to the right conclusion."

Uhlmann's team has a useful test for any AI: giving it a task for which you already know the solution. "This is a good way to check the algorithm is working as it should be and also maintain confidence in it," she says.

Image analysis, for example, can depend heavily on the conditions under which the cells or tissue images were captured — perhaps the light was better on one day, or a different person was behind the microscope. Machine-learning developers and users can address this challenge by being "mindful about the information they put in", Uhlmann says: "I have to think, 'Was I biased in the way I selected my examples of A and B? Is that really representative of the variation between A and B?'"

MARCELO BISPO DE JESUS/NANOCELL INTERACTIONS LAB

Fernanda Garcia Fossa uses CellProfiler, an image-analysis tool, in her PhD research.

> **"The foundations are important. People avoid it, but it bites them back."**

Also challenging is data management. As Singh explains, some projects generate hundreds of terabytes of images and measurement data, but the data-science expertise needed to analyse them isn't always available. "We definitely need more people who are able to work with high-dimensional data, who can tease apart the noise," he says.

Learn from the community

Inspired by CellProfiler and its potential, Garcia Fossa e-mailed the Broad Institute's Imaging Platform to learn more about the tool and its development. To her surprise, lab leader and co-developer Cimini replied almost instantly, inviting her to see the lab's work at first hand.

Garcia Fossa spent a year in Massachusetts, where she worked on her doctorate while helping to develop CellProfiler. "Don't be afraid to contact the developers of AI tools," she advises. "In my experience, they want to share their knowledge and get that feedback from the community to make the tools better."

And for people who can't attend training in person, there is a flourishing online community of AI-adopters in bioscience, whose members offer support and share resources on several global and regional forums. Singh recommends websites such as forum.image.sc, a discussion group for scientific image software, sponsored by the Center for Open Bioimage Analysis, a collaboration between the Broad Institute and the University of Wisconsin–Madison. Other options include BioStars.org and GitHub, which bioinformaticians use for online discussions and to share practical examples and code.

Ultimately, the best way to hone AI skills is through practice, and the data-science community platform Kaggle can offer some incentives. Informaticians can enter AI-related competitions on the platform and can win monetary prizes. It also offers a space for users to stress-test and compare their designs.

But win or lose, don't shy away from mistakes, advises Garcia Fossa — they're neither particularly expensive nor difficult to clean up. "It's important to play around with the program and learn through doing," she says. "That way, it will become second nature before you know it."

Rachael Pells is a science writer based in London.