map used to build the model, or without reporting whether they used a mask in building it. And very few deposit the half maps used to validate their analysis.

MAPPING THE FUTURE

Like models, maps are highly variable in quality, says Ardan Patwardhan, who manages the EMDB. Suites of automated and semi-automated tools have been created to help researchers turn 2D cryo-EM images into 3D maps. To help assess these workflows, the EMDB has run several validation competitions. It found that the greatest variability came not from the software packages, but from the experience level of the users. Less-experienced groups used default parameters; the best teams tailored settings to the data they had. That can make the difference between clearly visible side chains and blurry secondary structures, even when starting from the same raw images[5].

Today, researchers are calling for better methods for validating cryo-EM maps and models[6] — and raw image data could help. In 2014, Patwardhan and his colleagues at the European Bioinformatics Institute (EBI) in Cambridge, UK, created the Electron Microscopy Public Image Archive. The largest of the current 175 deposits of raw image data is more than 12 terabytes, which takes about 5 days to download.

Better methods for representing uncertainty could also help. Lander has proposed that researchers provide a spectrum of models[7] to better illustrate the range of structures that might fit the data. Maya Topf, a computational structural biologist at Birkbeck, University of London, has helped to create software called TEMPy that measures the quality of the model at the scale of amino acids rather than of the entire structure. Although this is not yet mandatory, the research community is starting to expect these kinds of evaluations, she says[8]. "The awareness is growing. More and more people are reporting in papers the local resolution."

Still, cryo-EM has a long way to go to match practices of crystallography. "The fact that data and models need to be validated has to become ingrained in people's minds, especially as the field attracts many new practitioners who don't have decades of experience," says Gerard Kleywegt, a structural biologist at the EBI. And, of course, some things are fundamentally different: crystallography captures proteins in rigid conformations, whereas cryo-EM can show more natural, and naturally ambiguous, conformations for which people are still developing the language to describe. Improvements will require better methods, greater consensus and better practices — all of which take time to develop. A validation task force met in September 2010 to develop recommendations, Kleywegt notes. "The field has evolved so rapidly since then that a follow-up meeting is overdue." Planning for a 2019 meeting is already under way. ■

Monya Baker *is an editor at* Nature.

1. Renaud, J. P. *et al. Nature Rev. Drug Discov.* **17**, 471–492 (2018).
2. Subramaniam, S., Earl, L. A., Falconieri, V., Milne, J. L. S. & Egelman, E. H. *Curr. Opin. Struct. Biol.* **41**, 194–202 (2016).
3. Neumann, P., Dickmanns, A. & Ficner, R. *Structure* **26**, 785–795.e4 (2018).
4. Wlodawer, A., Li, M. & Dauter, Z. *Structure* **25**, 1589–1597.e1 (2017).
5. Heymann, J. B. *et al. J. Struct. Biol.* https://doi.org/10.1016/j.jsb.2018.08.010 (2018).
6. Afonine, P. V. *et al. Acta Cryst. D* **74**, 814–840 (2018).
7. Herzik, M. A. Jr, Fraser, J. & Lander, G. C. Preprint at bioRxiv https://doi.org/10.1101/128561 (2017).
8. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. *J. Struct. Biol.* **199**, 12–26 (2017).

Big data fresh from the sea

Machine learning helps marine biologists to churn through millions of plankton images.

BY JEFFREY M. PERKEL

When they think about big data, most researchers probably imagine genomics, neuroscience or particle physics. Kelly Robinson's data challenge involves plankton.

"A lot of things that we enjoy seafood-wise — from fish to oysters to mussels to shrimp — almost everything starts their lives as plankton," says Robinson, who studies marine ecosystems at the University of Louisiana at Lafayette. In photographs, they look like floating specks of dust, and her research involves quantifying and mapping their distribution and predator–prey interactions. The problem is, she must do so in millions upon millions of images.

Robinson collects data by towing a remote-camera platform called ISIIS — the *In Situ* Ichthyoplankton Imaging System — behind a boat. ISIIS captures about 80 photos per second, or 288,000 images (660 gigabytes) per hour. For one project in the Straits of Florida, when Robinson was a postdoc, she generated 340 million pictures; a colleague working in the Gulf of Mexico generated billions.

"You start to learn about things that you never thought you would learn," Robinson says, "like the number of files that you can store on an individual computer. It's 30 million, by the way, on your regular PC." On her most recent cruise, Robinson sailed with 52 2-terabyte hard drives, which a student had to monitor and replace as they filled up. Someone then must get that collection to the university, convert the files to Linux formatting, and upload them to a server — a process that takes 24 hours per drive.

Kelly Robinson and her team scrutinize images.

The team uses machine-learning software to automatically pick out and identify objects in the images. But the algorithms must be taught what to look for — this is a starfish, that is a prawn. Such features are relatively rare in the water, so finding pictures for the training set takes time. Over two months, Robinson and her team manually sorted through 2 million images to find enough to feed the algorithm. "It's a little mind-numbing, but if you're under the gun you can do it," she says.

Naturally, the team is looking to optimize the process. Working with colleagues at Oregon State University in Corvallis, where she was a postdoc, Robinson is testing whether she could accelerate her work by processing the images on multiple video card graphical processing units (GPUs) running in parallel. She is also looking into cloud computing as an alternative to Earth-bound clusters.

But infrastructure goes only so far; what the team really needs, she says, is more people to crunch the numbers. Unfortunately, data scientists are in high demand, and industry jobs are lucrative. "We have a lot of turnover," she says. ■

Jeffrey M. Perkel *is technology editor at* Nature.

CORRECTION

The referencing in the Technology feature 'How to teach an old sequencer new tricks' (*Nature* **559**, 643–645; 2018) was incorrect. The correct version can be found online at at go.nature.com/2wmnhgc.