

The unsung hero of AI is not the algorithm, it's the data



Mark Basham
Science Director and
Challenge Lead,
Rosalind Franklin Institute

Data is the most expensive part of AI and the hardest to get right. Until now, it's also the part that's had the least funding and attention. With the government's AI for Science strategy, that is set to change. The strategy rightly focuses on generating new data sets, created with AI in mind. While that might sound straightforward, in reality, it's challenging and resource intensive. Scientists, and the organisations that fund them, need to be prepared to meet that challenge.

The UK has a strong science research base – and the data to match. Scientists are adept at managing data for high quality research, prioritising what they need for scientific discovery. Unfortunately, that approach does not always produce datasets that can power AI.

There are some exceptions. The most well-known science AI is Google DeepMind's AlphaFold, which, since 2021, has accurately predicted the structure of millions of proteins. But this was only possible because it already had 50 years of data to work with: since 1971, scientists had been uploading their data in an agreed format to the Protein Data Bank (PDB). Although efforts have been made, we have nothing like the PDB in other fields of science.

The change required to create AI-ready scientific datasets is substantial. It's requires far more than just managing the data differently once it's been collected. You need to change the way the science itself is carried out, including how equipment is calibrated, how experiments are conducted and how data is recorded.

This is acknowledged by the AI for Science strategy, which states: "AI is not just driving new breakthroughs in science, it is fundamentally reshaping the way science is conducted, evaluated, and understood".

The way that science has traditionally been rewarded and funded disincentivises these changes. A new approach to data analysis and management is needed if science is to

maximise the potential of AI. By placing a strong emphasis on data, the government's AI for Science strategy highlights what needs to change.

Building AI-ready scientific data

At the Rosalind Franklin Institute, since day one we've been actively working to try and ensure our data is AI-ready. We aim to make our data FAIR: findable, accessible, interoperable and reproducible. This means it meets most of the requirements for AI use. But despite our best efforts, even we can't claim complete success. Making data FAIR requires comprehensive metadata, proper documentation, and standardised structure formats that machines can interpret. It takes time, effort and expertise – and sometimes the funding and time constraints make it impossible.

The untapped potential of 'Dark Data'

Another issue is that large volumes of potentially useful data are currently being left 'in the dark' because of the way that scientists are forced to work.

Firstly, scientists are incentivised to only publish on success. The data on what doesn't work stays in the lab, if it's kept at all. But if we train science AI purely on published 'success' data, we're giving it less than half the picture – and that creates bias.

Scientists are also only able to study in depth part of the data they collect, so the rest, anything that isn't perfect for their needs, is discarded early in the discovery

process. This 'dark data' lives on our hard drives, on a USB stick on our shelves, because we've never had the time to analyse it.

This is a huge loss to science, because this 'dark data' could be powering the AI of the future. It also means we waste resources by replicating the same experiments, because we're unaware that the data we need already exists somewhere else.

Take structural biology for example – one of the main disciplines within the Franklin's research.

Most structural biologists are funded on a project basis, to look at one issue, such as the structure of a specific protein. They collect intricately detailed, 3D images of cells and then label that protein so they can compare across the images. But everything else in the image – all the other cell components – they simply have to ignore. They cannot afford to label those as well, as their time and effort wouldn't be covered under their grant funding.

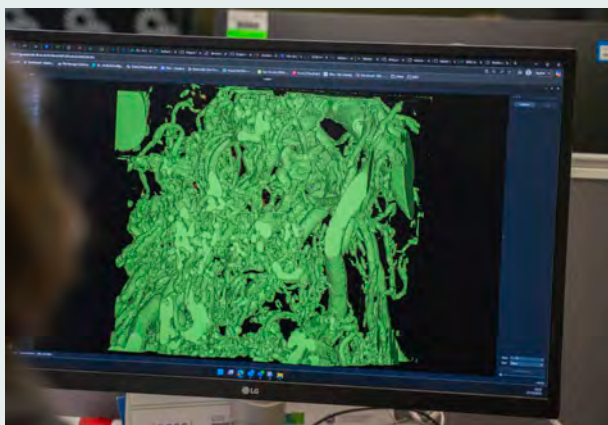
As an independently funded institute, we have more flexibility, but not always enough to cover all the data we create.

Creating the Digital Cell

Our new Digital Cell project, however, is a different matter. For this research, we are intentionally creating AI-ready datasets that can underpin increasingly powerful models. Comparing our approach to that of other research teams around the world shows why the focus on data is so important.



A stylised visualisation of how imaging a cell allows us to capture complex biological information and integrate this into a digital representation of a biological cell.



A researcher examines a detailed 3D reconstruction revealing the intricate inner structure of a complex sample.

science and biology to create innovative new technologies that will transform life sciences and improve human health.

Our base at the research and innovation supercampus at Harwell hosts some of the most advanced cryogenic electron tomography (cryo-ET) microscopes in the world. Cryo-ET builds up a 3D image from multiple 2D images of samples which have been flash frozen at temperatures below -180°C.

Usually, cryo-ET can only handle small samples, such as parts of cells. We have developed new techniques to prepare and handle larger samples of human tissue – though these are still less than a hundredth of the thickness of human hair. Using this technique, we can create 3D images of cells at very high resolution. This allows us to identify sub-cellular components and proteins that play a role in diseases, not in isolation, as was previously the case, but as part of a whole cell or group of cells.

We can then build a high-resolution image not of individual components, but of the overall cell structure – a method that will be crucial for our digital cell. What determines cell fate is not only the proteins and other elements within the cell but also where each element is located and even the shape of the cell itself. The Franklin is currently the only place in the UK able to generate this level of imaging and curate and manage the data to analyse it at the whole-cell scale.

And while we might be starting (relatively) small, we don't intend to be slow! We're working towards generating the high-quality data assets needed within just six months to create our first digital cell. From there we will be able to expand at pace, with iterative advances between new models and new data.

Funding the data revolution

In the past, getting funding simply to build this kind of dataset was almost impossible. The AI for Science strategy will hopefully change that for all researchers. For the first time it puts a true value on data as the necessary first step to realise the promise of AI. This mandate now needs to be reflected in funding policy and decision-making. Here at the Franklin, that recognition feels well overdue – but no less welcome for that.

The project will see us building a digital representation of a stem cell as the first step towards creating a virtual cell. While science has taught us much about cells, we are still unable to predict their fate. A virtual cell would allow us to answer that question, opening up new opportunities in drug development – the first mission of the government's AI for Science strategy.

A virtual cell model powered by AI will help us to understand how drugs affect cell fate, how different chemical imbalances change a cell's future. Will the drug drive a cell towards death, change its differentiation, or make it divide faster? All of these questions are critical for our understanding of many different diseases.

The other teams working in this area have chosen to start by combining existing datasets. The attraction of this seems clear – with so much data needed, why not get a head start by making use of the science that already exists? Unfortunately, it's not that simple. The majority of these datasets are not easily interoperable for AI models, due to the

heterogeneous nature of the cells studied and the way the data has been managed and curated by different research teams. So while you may be able to get a model working on initial datasets, as your data grows, the models will need to be completely reconfigured, progressively slowing you down.

Our approach is more in line with the AI for Science strategy's priority on new data sets. We are starting relatively small, by first gathering the structural data we need for an average stem cell: our digital cell. From this, we will gradually expand to other cells along the pathway. Each successive data set will only have subtle differences, allowing our models to work across all the data we create. Growing our data will not slow us down – quite the opposite.

Why the Franklin is uniquely positioned

The Franklin is uniquely positioned in the UK to lead on the virtual cell project. Our mission is to bring together experts from the fields of physics, chemistry, computer