AN AUDIENCE WITH…

Daphne Koller

Credit: insitro

Daphne Koller didn't anticipate becoming a drug hunter when she joined Stanford University as a professor of machine learning in 1995. But after stumbling into biology a few years later because of the novelty of the small but diverse data sets it offered, the field quickly drew her in. By 2016, she was the chief computing officer at Alphabet's ageing-focused biotech Calico. She then founded the biotech insitro in 2018 to explicitly generate high-quality biological data sets at high throughput to enable the application of machine learning to drug discovery. She spoke with **Asher Mullard** about her vision for machine learning in drug discovery, the need to view data as a core asset of a business rather than a byproduct and how to foster computational excellence in biopharma.

Q There's a lot of hope and hype around machine learning in drug discovery. What are the biggest hurdles towards actual progress here?

One of the biggest obstacles is the relative lack of high-quality large-scale data sets. That is one of the things that we're working to fix at insitro. And the other is the even greater lack of people who are bilingual in both machine learning and biology, who can figure out interesting problems to work on and the best ways to extract insights from data using the right machine learning models. We really need more people who can communicate across this chasm, between the disciplines. Building a team of such people is an essential part of insitro's vision.

Q How are you addressing this need for more high-quality data?

When I founded insitro, I wanted to build a company that would make significant upfront investments in generating large amounts of high-quality data explicitly for machine learning. And the only way to really do that is to build infrastructure that allows us to industrialize data production, not just because this allows us to generate data at scale, but also because robots are so much better at doing the same thing over and over again than people are.

> There is a difference between making sure you don't lose the data that you are generating anyway, versus creating data for the explicit purpose of driving machine learning models

We are the very fortunate beneficiaries of incredible work that's been done by researchers over the last 5 years that enables us to now build a 'bio-data factory'.

First, we can create cellular models using induced pluripotent stem cells (iPSCs). We used to have to rely on yeast cells and cancer cells that have been passaged so many times that much of the resemblance they once bore to human biology is now long gone. With differentiated iPSCs we can build incredibly disease-relevant models.

Second, we can perturb individual genes not using short interfering RNA, which has off-target effects that are often much larger than the on-target effects, but rather using CRISPR. CRISPR isn't perfect either, but it's a heck of a lot better in terms of the precision. And over time our ability to manipulate genes will only get better, because we can use not only CRISPR knockout, but also CRISPR interference, CRISPR activation, base-pair editing, RNA editing and more. This creates this incredible ability to reproduce genetic architectures in in vitro settings.

Third, we have access to improved biological assays. When I did a sabbatical in 2007 with Jonathan Weissman at the University of California, San Francisco, most of the assays we were using were growth assays. Growth is fine if you care about cancer and the cell cycle, but if you're looking for phenotypes that relate to other diseases or biological processes, these assays aren't that useful. Well, we now have high-content phenotypic screens that use super-resolution microscopy to measure much more relevant sets of phenotypes. And the nice thing about machine learning is that it already does image interpretation much better than humans. We also have things like single-cell RNA sequencing and mass spectrometry that further improve our ability to measure what is happening in cells in ways that really speak to relevant disease processes.

And finally there's the ability to employ automation and microfluidics to do this all in a completely scalable way. All of this has converged into what I consider to be a perfect storm of data production.

Q Other groups are also using iPSCs, gene editing and high-content phenotypic assays in drug discovery. What do you do differently to maximize machine learning opportunities?

Well, in part it comes back to scale. But it's also about very, very thoughtful experimental design. When we're doing things at scale, we need to think about how to randomize our experiments so as not to confound the biology. Even with all of our attempts at industrialization, there are going to be subtle changes in temperature, humidity, reagents and more that we need really thoughtful experimental design to control for. And we also have to capture an incredible amount of metadata to detect confounders and to correct for these computationally. For each experiment, we capture temperature, humidity, the technician, the location that a plate is in the stack, the batch of reagents that were used, the last time the machine was calibrated, and so on and so forth.

But let me give you an example of a conversation that highlights how our mindset is also different. I was asking a large biologics manufacturer about their biggest pain points, and they said, "Well, we put an antibody sequence into CHO cells and half the time the CHO cells choke and don't like the antibody and don't express it. So then we bring in a biochemist, and they look at the sequence and tinker with it. And then we try it again. All this takes months, can cost more than a million dollars and works only a bit more than half the time."

I suggested constructing a predictive model to look at the sequence upfront to predict whether a proposed antibody sequence is likely to express. They said, "that's a great idea, we should do that." But for them, the data they would use to do that were a throwaway byproduct rather than a core asset. They were stored in random spreadsheets on different laptops, and would need to be collected, amalgamated and cleaned up, and at the end of the day they still offered only about 300 antibody sequences, which is a pathetically small amount of data to train a model.

Now, if we wanted to do this we would ask Twist Bioscience to synthesize 100,000 antibody sequences for us — and they do not have to be therapeutically relevant, because we're never going to put them in humans — and then we would put these into CHO cells and measure expression levels. That would give us 100,000 sequence–expression pairs that we would then use to train a model. This is the difference between thinking of data as a throwaway byproduct, versus a core asset on top of which you can build a company.

Q ***How prevalent is the view in biopharma that data are a byproduct rather than an asset in their own right?***

Lots of people are now trying to clean up historical archives of data and use these in intelligent ways. But there is a difference between making sure you don't lose the data that you are generating anyway, versus creating data for the explicit purpose of driving machine learning models. And I think that's a difference that most people haven't gone through yet.

Q ***For those biopharma companies who want to go through this, how big of a problem is the shortfall in expertise you mentioned earlier?***

Honestly, there are not nearly enough such people out there for us to really transform the biopharma industry. And I think that's a gap in our educational system. We teach biology in ways that are almost perfectly designed to turn off people with engineering mindsets. And we teach machine learning in ways that, similarly, do not appeal to people with biology mindsets. We just are not creating enough people who can really communicate across this chasm.

Q ***What can biopharma companies do to make sure that there are appealing jobs in industry for those who are interested in bridging these worlds?***

Here are a few thoughts I have on this, in random order.

One is that it starts at the top. If you really want this type of thinking to make an impact, you need to hire someone truly senior — either in the C suite or who reports directly to the head of R&D.

You also need to really give these people a voice all the way down. They cannot be the downstream recipients of an experiment that was done and driven by someone else, and then just get a spreadsheet for analysis. They need to be not only engaged in experimental design, but have the ability to drive certain experiments.

And, these people can't be siloed away. They need to be much more embedded in the larger organization.

I guess the last one, which is simply a matter of pragmatics, is that if you want to hire the right people, you need to pay market rates. And those are often rates that are not within the pay scales of what a typical biopharma company conceives of, because machine learning experts get paid a lot.

Q ***One common criticism of machine learning is 'garbage in, garbage out', which you are attempting to address at insitro by generating fit-to-purpose data. Another is that machine learning is just a form of statistics that can recognize patterns in large data sets, but that can't generate hypotheses or test for causality. How do you respond to this view?***

Machine learning doesn't itself generate hypotheses, but it can find patterns that allow people to generate hypotheses. And it can do so in ways that are less biased by the preconceived hypotheses of humans.

For example, in 2011, I published a paper in *Science Translational Medicine* that used machine learning to look at cancer diagnosis from pathology data. This was before deep learning, which means that my PhD student, Andy Beck — an MD, PhD pathologist — had to define all the features that our software would look at. We deliberately didn't include just the features that we knew pathologists already looked at, but also put in hundreds of seemingly nonsensical features, things like "what is the average distance between a tumour cell and the closest stromal cell in the sample?" And it turned out that the machine learning consistently and robustly picked those features versus the traditional features as being predictive of 5-year survival. When you looked at what it was really keying off of, it had to do with the interplay between the tumour and the stroma, which is what today is called the tumour microenvironment. And this was before the importance of the tumour microenvironment was recognized in the way that it is today.

> "The last few years have proved time and time again that when machine learning is deployed in the right way, it's way beyond a productivity enhancer"

It's true that a person will have to ultimately interpret what is seen in order to derive scientific understanding from machine learning. But machine learning is able to point out, "Hey, people, maybe you should look at this."

Q ***What else do you wish the biopharma community understood better about machine learning?***

There are some people out there who think machine learning is going to be the solution to all the world's problems, not just within biopharma but even more broadly. I think they are wrong. At the same time, there are the folks who believe that this is just going to be another one of those productivity-enhancing tools that is going to be of marginal assistance to our productivity but that is not going to fundamentally change the paradigm. I think they too are wrong.

The last few years have proved time and time again that when machine learning is deployed in the right way, it's way beyond a productivity enhancer. It is a game changer. I think machine learning is going to offer a significant shift in how we do things in biopharma, even if it won't solve all our problems.