



ILLUSTRATION: THE PROJECT TWINS

CAN BIOLOGY MOVE INTO THE MATRIX?

'Virtual cells' could transform biomedical research, but scientists are still struggling with how to model life's complexity. **By Michael Eisenstein**

As every gamer knows, computers can plausibly simulate just about anything from the routine concerns of a household to the crises confronting a multiplanetary civilization. Simulating the fundamental unit of life – the cell – should be a walk in the park. But it's not.

Each cell is a complex ecosystem of biomolecules that interact with one another and react to external cues in ways that remain poorly understood. And what's true of one cell type isn't necessarily true of another. But there is an order to the chaos.

"The cell is a complex system, and a highly robust and resilient system," says Emma Lundberg, a bioengineer at Stanford University in California. "But it's also a highly structured system – the cell has an architecture."

Over the past few years, researchers have begun reverse-engineering that architecture to convert vast repositories of molecular data into 'virtual cells' – models that simulate the internal environment of cells both at rest and when responding to external triggers.

Several teams are now tapping into deep reservoirs of transcriptomic (gene expression) and other data sets to build models that could reveal the underlying biological bases of disease and possible angles for therapeutic intervention. "We have to think about virtual cells as a means of getting towards a specific goal, and for me, that goal is to be able to accelerate the hypothesis search process," says Yusuf Roohani, a machine-learning researcher at the Arc Institute in Palo Alto, California.

The field remains far short of a fully

functional virtual cell, however. "I don't think people would sensibly want to claim that they have built a virtual cell unless they need to sell a start-up," says Fabian Theis, a computational biologist at Helmholtz Centre Munich in Germany. Current models can capture static cell states but struggle to accurately predict dynamic changes. Reaching higher levels of *in silico* evolution will require ever-greater volumes of diverse data and smart strategies for combining them.

A strong foundation

The artificial-intelligence revolution has been a potent accelerant for enthusiasm around virtual cells, but scientists have grappled with how to build computational cell models for decades. "Even 20-something years ago, we

had ‘virtual cell 1.0’, where people were trying to use differential equations to describe systems biology,” says Bo Wang, an AI specialist at the University of Toronto in Canada.

Such models have the advantage of being grounded in measurable, well-understood biochemical and biophysical principles – threading together equations that describe cellular functions including metabolism, communication and movement. “You actually have mechanistic understanding – you can interpret them correctly, and that is very attractive,” says Lundberg.

A sophisticated mathematical model announced in March by a team led by Zaida Luthey-Schulten at the University of Illinois at Urbana-Champaign, for instance, realistically replicated cell division in a highly modified version of *Mycoplasma* bacteria¹. And Paul Macklin, an engineer at Indiana University in Bloomington, and his team have spent more than a decade developing a framework called PhysiCell to simulate how human cells and tissues respond to diverse environmental stimuli. This simulator has proved useful for modelling cancer biology, including factors driving progression or response to immunotherapy, Macklin says.

These successes notwithstanding, mathematical models are inherently limited by researchers’ understanding of cell biology. Initiatives such as the Human Cell Atlas have produced vast amounts of gene-expression and other data, including proteomics and epigenetics, but it’s extremely difficult to extract biological meaning from thousands upon thousands of molecular interactions. This is when AI models shine, says Maria Brbić, an AI researcher at the Swiss Federal Institute of Technology in Lausanne: “They’re really good at exploring combinatorial space.”

Opinions vary about which capabilities would define a true virtual cell, but any meaningful simulation should at least be able to represent the baseline state of a given cell type, and then project how a particular perturbation alters that state. Many attempts have relied on deep-learning-based ‘foundation models’, in which AI algorithms identify patterns in vast collections of unlabelled experimental data.

Roohani draws a parallel with ChatGPT, a chatbot powered by a foundation model that uses patterns gleaned from Internet text to produce coherent responses to almost any user query. “You can create more general-purpose representations across a broader range of cellular and biological contexts,” he says. In a best-case scenario, a biological foundation model would be able to extrapolate how various cell types will respond to conditions that are not included in the original training set, and even make meaningful predictions for cell types that it hasn’t encountered before.

Single-cell gene-expression data are currently the preferred way of educating biological foundation models about different



Maria Brbić uses artificial intelligence to incorporate transcriptome data into virtual cells.

cell types, and such data are readily available. Roohani and his colleagues have developed a database called scBaseCount, which uses AI to continually collect and uniformly process transcriptomic data for model-training purposes. The collection includes around half a billion cells, and counting. “That’s a few times more than the next-largest single-cell data repository,” says Roohani.

But one cannot simply build a representation based solely on a cell’s defining features – known in the context of AI models as an embedding. A virtual cell must also learn how different perturbations affect the cellular environment. Fleshing out these details requires experiments in which researchers systematically inactivate different genes or expose the cells to a diverse range of drugs. “We should have causal data to build causal models,” says Wang. One such collection is the X-Atlas/Pisces data set, compiled by Xaira Therapeutics, a drug company in South San Francisco, California. Available on the open-source AI platform HuggingFace, Pisces comprises gene-expression data from 25.6 million cells of various lineages that had undergone targeted gene disruption.

The pitfalls of perturbation

In theory, the resulting models could help users to infer which genetic abnormalities drive the growth of a particular tumour type or to pinpoint drug categories that stabilize metabolic issues in diseased cells, and some foundation models are on the cusp of achieving such capabilities.

In January, for example, Roohani and his colleagues described Stack², a model trained on the scBaseCount data set. The researchers were able to use these data to produce a ‘perturbation atlas’ that predicted the effects of different drug treatments in 28 distinct human

tissues. And in March, Xaira announced its X-Cell model³, trained on the company’s Pisces data set. According to Wang, who is also head of biomedical AI at Xaira, X-Cell was able to predict changes in gene expression underlying the activation of immune T cells even though it had not been trained on that process. This allowed the company’s scientists to predict mechanisms for switching off that activation – a potentially useful intervention in inflammatory disorders or other immune conditions. “We not only confirmed known inactivators, such as CD3 and its family, we also found a few putative T-cell inactivators,” says Wang.

Predicting the effects of cellular perturbation remains challenging, however, and Wang cautions that these models are only early steps in that direction. “So far, everybody’s just focusing on cell lines, which are relatively simple biological systems,” he says. These models might not accurately map to actual organs and tissues, and collecting training data from primary cell types – those taken directly from human samples – at a meaningful scale is daunting.

Researchers have also struggled to demonstrate clear performance gains from transcriptome-based foundation models relative to simpler mathematical methods. In 2025, the Arc Institute hosted the Virtual Cell Challenge, giving teams an opportunity to test the predictive performance of their models head-to-head. Although a success in terms of enthusiasm and engagement – Roohani says the event attracted some 5,000 participants from more than 100 countries – none of the pure AI models prevailed over those that incorporated conventional statistical methods.

Brbić has dealt with similar issues in assessing the robustness of deep-learning models. One problem, in her view, is that conventional performance metrics focus on capturing broad transcriptomic differences between perturbed

and unperturbed cells. This means that small but biologically meaningful changes might be drowned out by irrelevant background variation between samples, confounding AI analysis. “Single-cell RNA sequencing data is noisy,” says Brbić. “The kind of differences that we observe may be true biological differences but might also be caused by experimental artefacts or other sources of variation.”

In 2025, Brbić and her colleagues released a benchmarking tool called Systema, which allows users to eliminate noise and home in on perturbation-specific changes in gene expression⁴. Roohani’s team’s perturbation-prediction model, called State, which is trained to recognize the inherent variability in cell populations⁵ – also addresses this problem. By combining this approach with a performance metric that, like Systema, zooms in on perturbation-specific effects rather than overall gene expression, State was able to accurately predict about one-third of the genes most strongly affected by a given perturbation in a test data set. That’s a big improvement on the 7% achieved using conventional methods.

Completing the picture

Although AI models have yet to deliver a decisive advance in predicting cell behaviour, they can go beyond data they’ve already encountered to make generalizations about new cells, tissues and even species. That’s something that conventional computational methods are unable to achieve, says Wang. “We cannot expect a linear model to construct this kind of virtual cell,” he says. “Having the right data with the right model is probably a better approach.”

Emphasis on right data. “It’s not really about the number of cells,” Lundberg says. “Are we capturing different disease states? Are we capturing different tissues?” Greater diversity also means moving beyond the transcriptome and towards a ‘multimodal’ approach that layers on biological information such as chromatin states, cell shape and protein expression and localization.

Several groups have already demonstrated the potential of models trained on data other than transcriptomes. Last October, Lundberg and colleagues unveiled SubCell, a model trained on microscopic images of human cells and the distribution of their protein contents⁶. “We’ve used it to predict mechanism of action in drug-perturbed cells,” she says.

SubCell draws on Lundberg’s work with the Human Protein Atlas project, which systematically mapped the subcellular localization of every human protein inside various cells and tissues. SubCell was trained on these maps, and the resulting embeddings were coupled to those generated by a foundation model called ESM2. Developed by Meta AI’s Fundamental AI Research Team, ESM2 was trained on protein sequences and can model structural features and evolutionary relatedness.

The two sets of embeddings combined to yield a model that is greater than the sum of its parts, Lundberg says. “It’s so much better at predicting protein function, predicting protein–protein interactions – many of these relevant tasks to understand biology.” SubCell can also complete tasks such as analysing images of yeast cells and determining their cell-cycle state despite being trained entirely on human cell data.

At GenBio in Palo Alto, chief scientist Eric Xing is working with ‘world models’ – an alternative to conventional foundation models. These representations are trained on various data types – including structures, sequences, images and text – and produce multimodal *in silico* systems that attempt to replicate the internal environment and biological activity of a living cell. “A foundation model is kind of focused on the representation learning of the cell, while a world model should be more focused on the dynamic behaviour modelling of the cell,” says Qi Liu, a computational biologist at Tongji University in Shanghai, China, who has been leading the development of a world model called AlphaCell⁷.

“Scientists have grappled with how to build computational cell models for decades.”

Xing proposes that the ‘AI-driven digital organism’ (AIDO) world model being developed at GenBio will be able to replicate diverse biomedically relevant behaviours of healthy and diseased cells. “We are soon going to release our first prototype, which starts from 20 finite and explicit prompts that include gene editing, for example, or small-molecule interventions,” he says. “And on the other side, we specify a finite number of outputs we want to see, including morphology, localization and other things.”

That said, no cell is an island. Attempts to model the individual building blocks of a tissue will inevitably miss outcomes arising from communication in and across organ systems. Xing is optimistic that GenBio’s AIDO framework will be capable of such scaling, but that is years away. For now, mathematical models such as Macklin’s PhysiCell offer a powerful solution. His team has used this framework to simulate processes ranging from immune-cell infiltration of tumour micro-environments to the early development of the cerebral cortex.

Last August, Macklin and his colleagues, including Elana Fertig at the University of Maryland School of Medicine and Genevieve Stein-O’Brien at Johns Hopkins University, both in Baltimore, published an upgrade to PhysiCell that allows users to specify biological scenarios using simple, declarative sentences⁸. A user

might, for instance, state that a particular drug increases cell-cycle activity, or that a signalling factor activates a specific subset of immune cells. These statements are then converted into machine-interpretable rules, offering a user-friendly tool for tissue and organ modelling. But Macklin also sees opportunities to introduce AI. He notes that PhysiCell is ill-suited for capturing molecular-scale detail, whereas foundation models currently run into trouble moving to cellular scale and beyond. “If we put these together, it’s going to be fantastic,” he says.

A hard cell

Though a far cry from a true virtual cell, developers are finding utility in their models. “Internally at Xaira, we have already started to use X-Cell to do things like target identification,” says Wang.

Perturbation models could also accelerate hypothesis generation and testing, sparing researchers the need for massive high-throughput screens. This would allow them to focus on validating computationally generated hits. “I think we’re going to see a shift towards simulating first, doing experiments later,” says Lundberg. Because these simulations evolve to capture more detail about the cellular environment and its surroundings, they could help to reduce reliance on animal models for drug development and testing, yielding human-centric predictions that reduce the risk of toxicity and failure in clinical trials.

But researchers will also need to overcome the well-known limitations of generative AI. Chatbots routinely fabricate ‘facts’ and even actively mislead users, and image-generation algorithms are prone to hallucinatory flights of fancy. Xing cautions that early-generation models will fundamentally be simulations of biology rather than true replicas of cellular reality. Accordingly, he and others in the field favour early release and public testing of models, and the data sets used to train them, so that users can uncover the models’ strengths and limitations. “Once our virtual cell is there, we are going to make it public for people to play with,” says Xing. “It may embarrass us to have very bad results, but I think that’s a journey that we have to work through.”

Michael Eisenstein is a freelance writer based in Philadelphia, Pennsylvania.

1. Thornburg, Z. R. et al. *Cell* **189**, 2582–2597 (2026).
2. Dong, M. et al. Preprint at bioRxiv <https://doi.org/10.64898/2026.01.09.698608> (2026).
3. Wang, C. et al. Preprint at bioRxiv <https://doi.org/10.64898/2026.03.18.712807> (2026).
4. Viñas Torné, R. et al. *Nature Biotechnol.* <https://doi.org/10.1038/s41587-025-02777-8> (2025).
5. Adduri, A. K. et al. Preprint at bioRxiv <https://doi.org/10.1101/2025.06.26.661135> (2025).
6. Gupta, A. et al. Preprint at bioRxiv <https://doi.org/10.1101/2024.12.06.627299> (2024).
7. Chuai, G. et al. Preprint at bioRxiv <https://doi.org/10.64898/2026.03.02.709176> (2026).
8. Johnson, J. A. I. et al. *Cell* **188**, 4711–4733 (2025).