THE ARISING

Virtual Cells & Clinical AI



NOVEMBER 2026

Virtual Cells & Clinical AI

What virtual biology could help us discover

The Arising · Virtual cells & AI · November 2026 · John Griffin, MD

A scientist may have a thousand ideas and enough time to test only a few.

Which gene should the team investigate? Which possible medicine deserves an expensive experiment? Which change might help a diseased cell without damaging a healthy one?

A useful computer model could help with those choices. It would not need to replace every laboratory experiment. Even pointing researchers toward better experiments could make a difference.

That is the practical promise behind the phrase “virtual cell.” The goal is larger than a beautiful image on a screen: a model that can represent aspects of a cell and predict how they might change when something happens to it. The challenge is knowing which predictions deserve trust.

A 2024 Perspective in Cell describes a vision for building such models with artificial intelligence. It is a research agenda, not an announcement that a universally reliable digital cell already exists. A separate study of clinical AI offers a useful contrast: even an encouraging result tells us only about the task and setting actually tested. [1,2]

What Is a Virtual Cell?

A cell is not simply a small object with parts in fixed places. It is continually using energy, making and breaking down molecules, responding to signals and interacting with its surroundings.

A virtual-cell model tries to represent relevant features of that activity in a form a computer can use. The word “relevant” matters. A model designed to predict which genes become more active after a treatment may not tell us whether a tissue will recover or whether a medicine will harm a patient.

The Cell Perspective proposes models that connect information across molecules, cells and tissues. It describes combining different kinds of measurements, such as molecular data and images, rather than relying on one view of biology. In technical language, this is a multi-scale, multimodal approach. In everyday terms, it means joining different levels and kinds of information to build a more useful picture. [1]

The ambition is compelling. But a detailed picture and a reliable prediction are different achievements.

Imagine a weather map. Beautiful clouds on the display are not what make the forecast useful. What matters is whether the forecast helps you anticipate rain where you will

actually be. A virtual-cell model needs a similarly specific test: what does it predict, in which biological setting, and how well does it work there?

What would we ask it to do?

“Understand a cell” is too broad to be a fair test. Smaller questions are more useful.

One task might be to predict how a cell's gene activity changes after a particular gene is switched off. Another might be to rank compounds worth testing. A third might concern which cell types respond differently to the same treatment.

These are examples of research questions, not claims that every current model can answer them well. They also require different measures of success. A system could rank promising candidates usefully without accurately predicting every molecular change. Another could reproduce a familiar gene-activity pattern but fail on an unfamiliar treatment.

The first question for a model is therefore simple: what job are you asking it to do?

Without that answer, “accurate” can mean almost anything.

Learning yesterday's patterns, predicting tomorrow's experiment

Many AI models learn from large collections of examples. In biology, those examples reflect choices made by researchers: which cells were measured, which conditions were tested, which instruments were used and which results were available.

That history shapes what a model has had a chance to learn. A large dataset can still leave important gaps. Rare cell states, unusual diseases or combinations of treatments may not be represented well. Differences between laboratories can also become part of the pattern the model learns. [1]

A model may look strong when tested on examples close to those it has already seen. The harder question is whether it works on a genuinely new problem.

Researchers call this generalization. For a reader, a useful translation is: can the system carry what it learned into a situation that is meaningfully different?

Knowing how a familiar cell responds to a familiar chemical is not the same as predicting how a new cell type will respond to a new treatment. Those tasks should not share one sweeping claim of success.

What was kept out of the lesson?

To test a model, researchers usually separate data used for learning from data used for evaluation. The examples reserved for evaluation are often called “held-out” data.

The label alone is not enough. If the evaluation examples are very similar to the training examples, the test may be easier than the real task the model is meant to support.

An everyday comparison is preparing for an examination. Seeing a slightly reworded version of a question you have practiced is different from applying the idea to a new problem. Both can show something, but they do not demonstrate the same skill.

For a biological model, useful questions include: were entire treatments held out? New cell types? Data from another laboratory? A different disease setting? The right separation depends on the intended use. [1]

None of this means a model is worthless if it has limits. A tool can be valuable within a clearly defined boundary. Trouble begins when that boundary disappears from the claim.

The experiment beyond the screen

A prediction earns a different kind of credibility when researchers test it in a real experiment.

Suppose a model selects ten compounds from a much larger list. How many perform usefully when tested? How does that compare with the team's existing method of choosing? Were the tests planned before the results were known? Did another laboratory find something similar?

These are illustrative evaluation questions, not results reported by the Cell Perspective. They show why a strong computer score is only part of the story.

The comparison matters. Beating random selection may be encouraging, but it does not tell us whether the model improves on an experienced team or a simpler existing method. A complicated model should earn its extra complexity by helping with the job that matters.

Failures matter too. If only the successful predictions appear in a presentation, readers cannot tell how often the method was wrong. Negative results help reveal the boundary between a useful tool and an attractive demonstration.

Knowing when not to sound certain

A model can produce an answer even when the supporting information is thin. The appearance of precision does not guarantee that the situation was well represented in its data.

A more useful tool would help users recognize where its predictions are uncertain and where a new experiment could be especially informative. The Cell Perspective discusses the importance of uncertainty and iterative experimental testing as part of the larger research vision. [1]

That does not mean adding a confidence number solves the problem. The uncertainty estimate itself needs testing. If a model repeatedly says it is highly confident when it is wrong, the number can make decisions worse rather than better.

The goal is not a computer that refuses every difficult question. It is a tool whose users can understand where the answer is well supported, where it is tentative and what evidence could improve it.

Testing AI Conversations in Clinical Settings

The clinical AI study concerned AMIE, a conversational system used before urgent primary-care appointments. It did not simulate cells or predict drug effects. It evaluated whether an AI conversation could fit into a supervised clinical workflow and how that conversation was experienced and assessed. [2]

The study enrolled 114 adults, with 98 completing the AI interaction and subsequent visit. It took place at one center, involved English-speaking adults with one urgent complaint, and allowed the conversation up to five days before the appointment. All interactions were supervised by physicians. Transcripts were shared with clinicians. [2]

These details are part of the finding. They describe what was tested: a supported process around a real visit, not a system left to manage care alone.

There was no comparison group following a different workflow. This was a feasibility study, meaning it examined whether the approach could work in that setting and what happened when it was tried. It was not a demonstration that the system improved survival, prevented complications or could replace a clinician.

Clinical AI Errors and the Limits of a Stopping Rule

In the 98 completed encounters, no prespecified conversation safety stop was required. That is encouraging within the study's safeguards. It is not the same as saying nothing went wrong.

Supervisors recorded one hallucination, an unsupported or invented statement, and added clinical information in five interactions. The presence of supervision matters when interpreting both the absence of stops and the problems that were found. [2]

A safety-stop rule asks whether a particular threshold for interrupting the process was reached. It does not capture every possible mistake or prove that a larger, less supervised deployment would be safe.

The lesson is not that a supervised trial must be perfect before it can be useful. It is that “no stops” and “no errors” answer different questions.

Why these two stories belong together

Virtual biology and clinical conversation have different users, different failure modes and different measures of success. The AMIE study does not validate the virtual-cell vision. The Cell Perspective does not establish that a chatbot is safe to use for diagnosis.

They belong together because “AI in medicine” is too broad a label to tell us what has been demonstrated.

A model that helps select a laboratory experiment should be judged on whether it improves that selection. A system that prepares a clinical conversation should be judged on the work it actually performs, with the safeguards and people around it included in the assessment.

A positive experience is not automatically a health benefit. A computational success is not automatically a treatment. Keeping those distinctions visible allows useful progress to be recognized without making it carry more than the evidence supports.

Who supported the work?

The AMIE study was funded by Alphabet. Disclosures include Google employment, stock and patent interests, as well as a visiting researcher relationship. The virtual-cell Perspective also reports commercial interests among its authors, including Genentech employment, equity, patents and advisory roles, and grew from a Chan Zuckerberg Initiative workshop. These relationships do not decide whether the ideas or findings are useful. They are relevant context when evaluating claims and the need for independent testing. [1,2]

A small, testable step for The Arising

There is a practical way to bring this thinking into a literature platform without claiming to build a digital replica of a patient.

A proposed evidence tool could attach each model claim to the data it used, the method it was compared with, the setting where it was tested and the limitations still unresolved. A reader could then follow the claim back to its evidence.

A small evaluation could ask whether AI-assisted literature review finds the correct primary papers, keeps the numbers attached to the right groups, notices conflicting findings and avoids turning a biological mechanism into a treatment recommendation.

Those are observable tasks. Errors can be counted and compared with the existing workflow. The proposal is not a validated Arising Health product or a result of either paper. It is an example of making an ambitious idea small enough to test honestly.

Three questions to carry into the next headline

What did it actually do? Look for the task, not just the phrase “AI breakthrough.” Did the system predict a molecular pattern, choose an experiment, conduct a conversation or improve an outcome?

Where did it work? Ask about the cells, people, setting and supervision. A result from one laboratory or one supervised clinic is useful evidence about that setting, not an automatic guarantee everywhere else.

What happened when someone checked? Look for independent experiments, a meaningful comparison, failed predictions and limits that are stated plainly. A convincing demonstration should make its boundaries easier to see, not harder.

These questions do not require becoming an AI researcher. They make the promise more concrete.

Better questions, better experiments

A virtual cell does not have to reproduce every detail of life to become useful. It does have to earn trust for the predictions it offers.

The exciting possibility is a cycle: a model suggests an experiment, the experiment reveals something the model did not know, and the next prediction becomes better informed. The computer and the laboratory improve each other's questions.

That is a future worth working toward. Its progress will be measured not only by the beauty of the cell on the screen, but by what we learn when we step away from the screen and test it.

Sources and further reading

[1] Bunne C, Roohani Y, Rosen Y and colleagues. How to build the virtual cell with artificial intelligence: Priorities and opportunities. Cell. 2024;187(25):7045–7063. Primary publication record. Full text.

[2] Brodeur PG, Koshy JM, Palepu A and colleagues. Conversational diagnostic artificial intelligence in ambulatory primary care: a prospective feasibility study. Lancet. 2026;408(10565):1732–1744. Primary publication record. Publisher record.

Continue exploring

Read RNA processing and cellular aging, Biological-age clocks and Regenerative frontiers, or explore the magazine collection.