// ARTICLE · PUBLISHED 28 OCTOBER 2026

Virtual Cell Models: The Bottleneck Is the Data

A $1.8 billion pledge, a lab-tested AI research system and a synthetic biobank all point at the same constraint, and it's measurement rather than models

Jason Newell · MAX Research Collective · Article ART-HLC-020 · Published 28 October 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/virtual-cell-models.html
Cite as: Newell, J. (2026). Virtual Cell Models: The Bottleneck Is the Data. MAX Research Collective, ART-HLC-020. https://www.maxresearchcollective.com/articles/virtual-cell-models.html

The constraint is measurement

The hard part of building an AI model of a human cell is the data. Three announcements in two weeks of autumn 2026 say so, each from its own side. On 7 October Biohub, two federal agencies and three technology companies committed $1.8 billion to producing measurements that cell models can learn from. A week earlier Microsoft Research showed Quine, a research system that ranked cancer compounds which the Broad Institute then tested in the lab. And a preprint called DrugTargetWorld built simulated biobanks to grade AI agents, because in real biology nobody holds the answer key.

Each of them spends its effort on measurement. The money has its own story. Less of it is new than the headline suggests; commercial funders reportedly get the data a year before anyone else.

What a virtual cell would be

A virtual cell is a computer model that predicts how a human cell behaves when something changes. Switch off a gene, add a drug or let a disease take hold, and the

model says how the cell's state moves. If such models worked, a researcher could screen thousands of ideas on a computer and send only the best few to the lab.

Nobody has one that works in general. The partners in the new effort blame the data. A model needs large, consistent measurements of cells, taken the same way and carrying shared identifiers, so that it can learn from all of them at once.

Who is paying for what

Biohub's announcement of 7 October lists four sources.

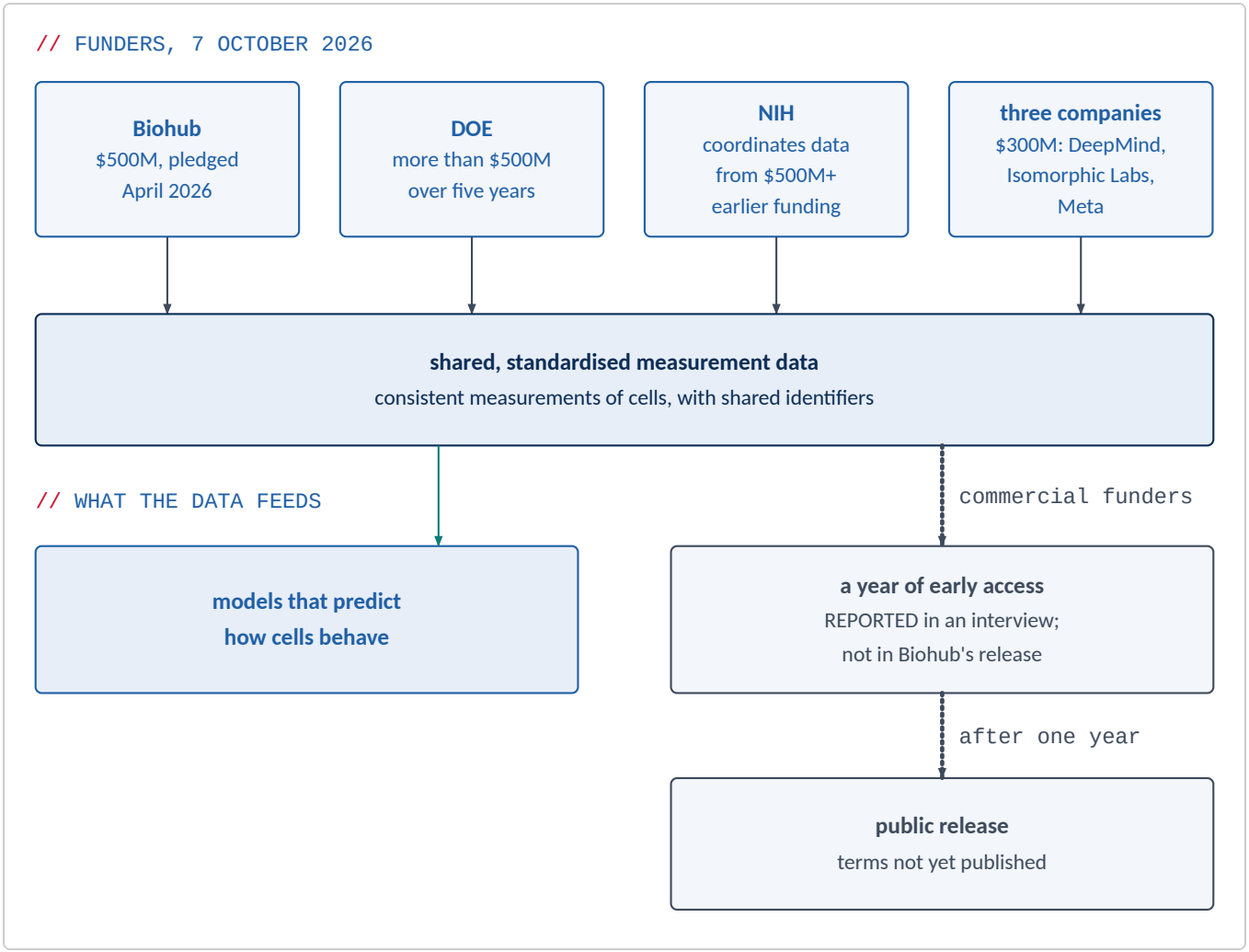


Figure 1. Four funders feed one shared body of measurement data; the year of early access on the right comes from an interview, and Biohub's release doesn't mention it.

- Biohub puts in $500 million, first pledged in April 2026: $400 million for measurement and imaging technology, plus engineering, and $100 million for research outside Biohub.
- The US Department of Energy (DOE) commits more than $500 million over five years for lab measurement and modelling (computation too), through its Genesis Mission, a cross-agency programme DOE leads.
- The National Institutes of Health (NIH) coordinate datasets and knowledge bases built with more than $500 million of earlier federal funding, which Biohub will help standardise for AI training.
- Google DeepMind, Isomorphic Labs and Meta add $300 million between them.

Six research groups take part as well, among them the Broad Institute, the Human Cell Atlas and the Wellcome Sanger Institute. NVIDIA supplies computing. Biohub calls it "the largest coordinated commitment to generating AI-ready biological data to date."

The $1.8 billion adds up different kinds of money. DOE's commitment and the companies' $300 million are new, which comes to about $800 million. The rest is Biohub's April pledge and the value of NIH data that already exists. The effort stands, and the headline should be read with that arithmetic in mind.

A year before everyone else

Commercial funders will reportedly use the data for a year before it goes public. Alex Rives, Biohub's head of science, described "a one-year embargo" to Axios. His reason was blunt: "We have to have some incentive for commercial players to be a part of this, and the embargo period creates that." The Next Web and the blog Let's Data Science carried the same account.

Biohub's own release says something different. It describes "an open resource for the research community." Google DeepMind's Pushmeet Kohli speaks of "an open, standardized data commons." Both can be true at once, since a resource can be open in the end and exclusive at first.

The terms are what matter, and they aren't public. Nobody outside the deal knows which data the year covers or which licence applies once it ends.

Public money pays for much of this work, which is why the terms deserve a published answer. Until there is one, the early access is an interview account that the primary release leaves out.

Quine: the loop between model and lab

On 29 September Microsoft Research introduced Quine. Quine pairs a "world model" of biology with tools and literature. The researchers sit inside the loop too. A world model learns how a system behaves well enough to predict the effect of an action; this one was trained on sequence, structure, cell state and imaging data.

Working with the Broad Institute of MIT and Harvard, Quine ranked thousands of compounds by one question, which was whether each would push pancreatic cancer cells out of the state biologists call classical and into the one they call basal, and it had to put every candidate in order before anyone at the bench touched a plate. Speed was the point. Microsoft reports that the top-ranked compounds produced the largest shifts in the intended direction and that several held up in more than one lab assay. The narrowing took about a weekend. The reverse shift proved harder — as Quine had predicted it would.

Those results are Microsoft's own, in a blog post rather than a peer-reviewed paper. Microsoft calls Quine research technology, "not for clinical or medical use," and warns its outputs may be inaccurate. Only a fellows programme can use it for now.

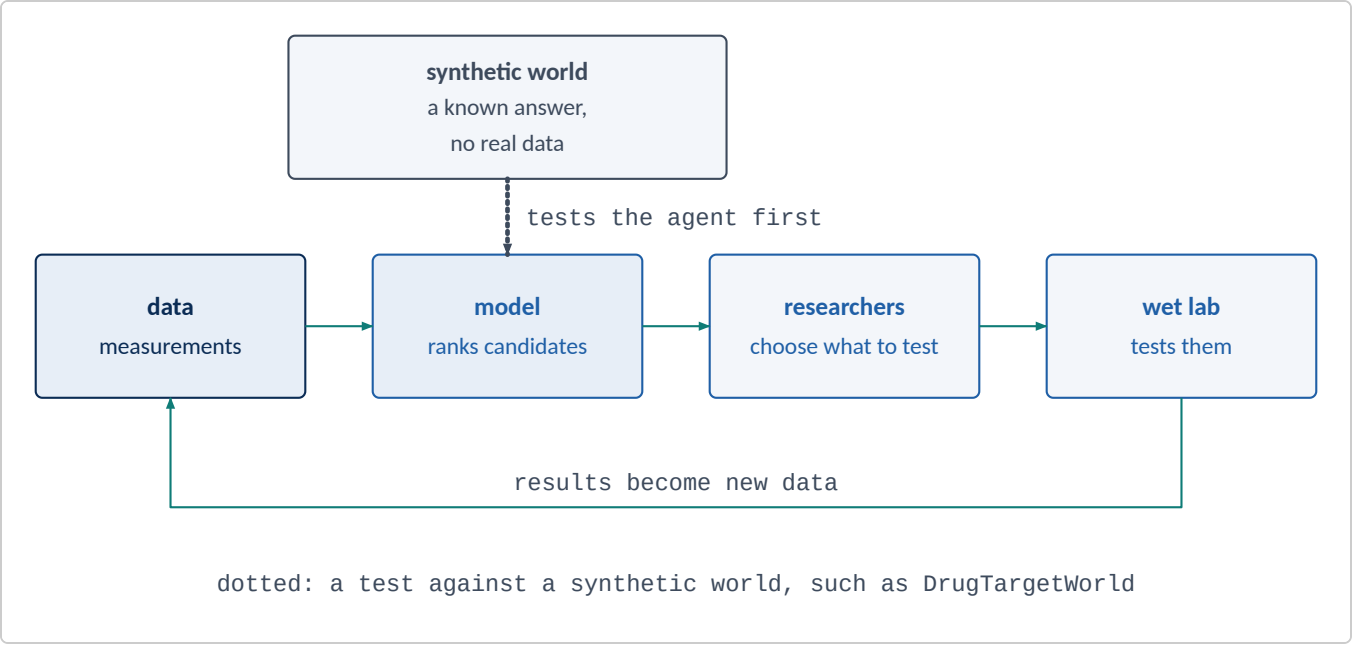


Figure 2. Follow the loop clockwise from the data; the dotted line shows where a synthetic world with a known answer tests the agent before real data reaches it.

Quine makes a loop visible. Data trains a model; the model ranks candidates; people choose what to test, the wet lab tests it, and the results become new data. It runs only as well as what goes into it.

DrugTargetWorld: testing the scientist

DrugTargetWorld offers one way to judge whether an AI agent is good at this work. The preprint, arXiv 2610.09558, submitted on 7 October by Margolis and colleagues, generates simulated biobanks with a hidden causal structure. A biobank is a large store of samples and health data from many people. Each simulated world holds data for 54,000 participants: genetics, proteins, health records and outcomes, plus synthetic imaging.

An agent has to define a disease and find which proteins cause it. Then it decides which way to push them, and it may run virtual lab experiments along the way. The authors ran nine agents through 540 episodes across 20 cardiovascular worlds, each world built so that the hidden causes were fixed in advance and every claim an agent made could be scored against them. The best two scored about 40 and 35 out of 100. Each recovered about 64% of the causal drivers, roughly two in three. No agent reliably flagged proteins that looked important and weren't causal.

That last gap matters most. A drug aimed at a protein that only travels with the disease fails, slowly and expensively, in a clinical trial.

A synthetic world can measure what real data can't, because the true answer is known. It's still a preprint. Nobody has peer reviewed it.

What isn't known

Four questions stay open. The data-use terms haven't been published, so the embargo's scope and the licence after it are guesses. The split of new money by year is unclear. Quine's results haven't faced independent testing. And nobody has shown that a score in a synthetic world predicts performance on real biology, which is the claim DrugTargetWorld will eventually have to earn.

Why it matters

Models can be copied in an afternoon. Well-measured biological data takes years to produce. That's why a year of early access to it is worth paying for — and why the terms of that year belong in public.

Clinical AI faces the same constraint. MARA, the Medical Agentic Resident Assistant this project is designing, is a clinical assistant rather than a drug-discovery tool, and nothing here changes her design directly. One idea carries over whole: test an agent first on a synthetic world whose answer is known, with no patient data in it. Her methods are proved that way on MAX4, the desktop AI resident she's developed from. No real record comes near her until they pass.

Sources

- Biohub. "Virtual Biology Initiative expansion." 7 October 2026. https://biohub.org/news/virtual-biology-initiative-expansion/ (also distributed through Newswise's DOE Science News Source the same day).
- The Next Web. "Biohub, Meta, Google DeepMind and the US pool $1.8bn for AI biology data." October 2026, reporting Alex Rives's remarks to Axios.

https://thenextweb.com/news/biohub-1-8bn-ai-biology-data-meta-google-deepmind

- Let's Data Science. "The virtual cell: $1.8 billion, open data, and big tech gets a year early." October 2026. https://letsdatascience.com/blog/virtual-cell-18-billion-open-data-big-tech-gets-year-early
- Microsoft Research. "Introducing Quine: An AI research system designed for the complexity of biology." 29 September 2026.
- Margolis S, et al. "DrugTargetWorld: A Synthetic Biobank for Training and Benchmarking AI Scientists." arXiv:2610.09558, 7 October 2026.
- Healthcare AI Guy Weekly, 20 October 2026, used to locate the items above.

Glossary

TERM	DEFINITION
Biobank	a large store of samples and health data from many people. DrugTargetWorld simulates one so the right answers are known.
Causal driver	a factor that actually causes an outcome, as opposed to one that only moves with it. Finding them is the task DrugTargetWorld grades.
DOE (US Department of Energy)	the federal department funding lab measurement and computing for the initiative through its Genesis Mission.
Embargo	a period in which some parties may use data before it's released publicly. Here it's the reported year of early access for commercial funders.
Genesis Mission	a cross-agency programme led by DOE that carries its share of the funding.
MARA (Medical Agentic Resident Assistant)	this project's design for a clinical assistant that recommends while a person decides. Her methods are tested on synthetic data first.
MAX4	the desktop generation of the MAX AI resident, from October 2026 on, where MARA's methods are proved.
NIH (National Institutes of Health)	the US medical research agency, which brings existing datasets and repositories to the effort.
Preprint	a paper shared before peer review. DrugTargetWorld is one.
Virtual cell	a computer model that predicts how a cell's state changes under a perturbation such as a drug or a gene change. Building one is the goal the data serves.

Newell, J. (2026). MAX Research Collective, ART-HLC-020. CC BY 4.0

TERM	DEFINITION
Wet lab	physical experiments with cells, reagents and instruments, as opposed to computation. It's where Quine's rankings were tested.
World model	a model that learns how a system behaves well enough to predict the effect of an action. Quine's is trained across sequence, structure, cell state and imaging data.

HEALTHCARE ART-HLC-020

// ABOUT THE AUTHOR

Jason Newell leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn →