A virtual assay is not the asset

Why durable value in AI-native drug discovery accrues to whoever raises the predictive validity of the looped experimental system, not to the AI model, and why the window to own it is closing fast

Tim Ahfeldt

Co-Founder and CEO, VALID Inc.

Our recent Perspective in Nature Reviews Drug Discovery set out to diagnose a discomforting fact. After a decade of intense activity and a great deal of capital, evidence of clinically relevant impact from AI in drug discovery remains limited[1]. Derek Lowe covered it generously [a]. A diagnosis carries an obligation. Once you have named the disease, you owe a prescription, and that obligation is mine. In this white paper, I begin to write the prescription. It starts with a claim that sounds like technicality and turns out to be the whole thesis.

Imagine you were handed, free and clear, the complete weights of an integrated pharma's or recent AI start-up's best multimodal model, or virtual cell or virtual human. This is the checkpoint that fuses high-content imaging, transcriptomics, proteomics, and perturbation response into a single representation of a cell. No license, no strings. Ask what you actually have. You have the model, but not the experimental lab in the loop that gives its predictions meaning. A virtual assay can rank, infer, and propose, but it cannot establish whether it is right. For that, you still need the experimental system that generated the relevant biology and can be run again to test what the model predicts. That lab in the loop is not supporting infrastructure for the AI. It is part of the AI drug discovery system itself, because it is what turns inference into evidence and lets the model improve through repeated experimental feedback. It is also the one thing that wasn't included in the package you were handed. The argument that follows is about why this physical loop, and especially the predictive validity of the Test inside it, is where durable value actually sits. This is also why the next 24 months are the highest-leverage window this category will see. The whole argument resolves to one picture, Figure 1: an empty top-right quadrant where a test system is at once predictively valid and able to generate high-dimensional perturbation data over time at AI scale.

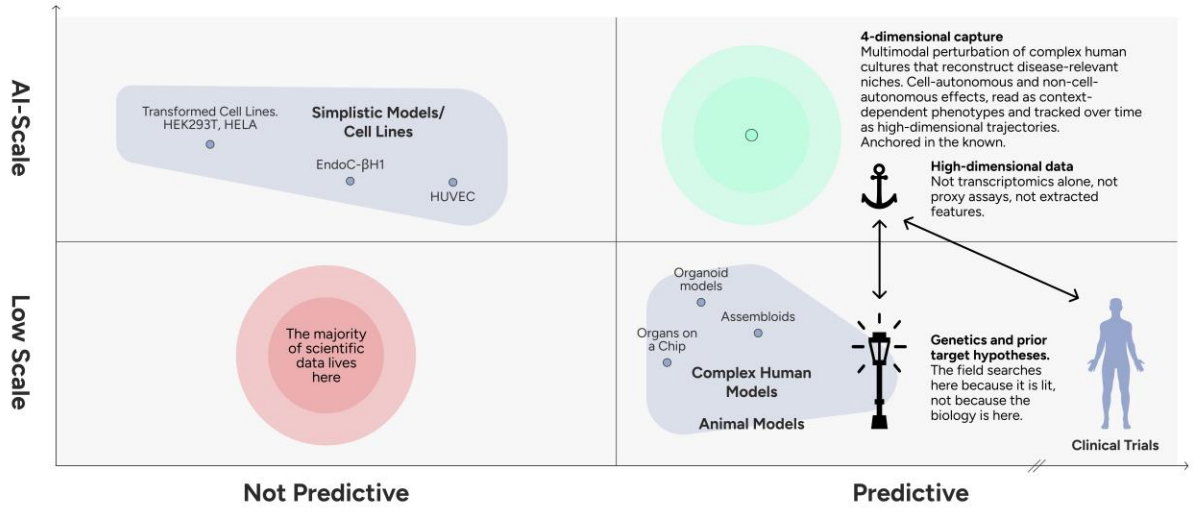

Figure 1. Predictive validity and the ability to generate high-dimensional perturbation data over time are orthogonal axes, and the category that matters is empty. Horizontal: predictive validity, the degree to which a system's readout tracks human disease. Vertical: ability to generate high-dimensional perturbation data, including over time, at a volume AI can use. The field routinely collapses these into one axis and treats moving up as moving right; they do not correspond. The top-right quadrant, predictive and able to generate that data at once, is the hard problem worth solving. Occupying it means reading out context- and perturbation-dependent disease phenotypes, followed over time as trajectories.

The DMTA cycle collapses, but not evenly

Drug discovery reduces to four sequential steps: Design, Make, Test, Analyze, with a person turning the crank to move a program through them. In the AI era two of the four steps fold together. The Analyze and Design steps collapse into a single computational layer, because the models that interpret prior data are the same models that propose the next candidates from it. A generative chemistry model is at once a reading of past structure-activity and a proposal for the next structure. A causal target model is at once an analysis of the evidence and a recommendation for what to do next. This collapsed layer is where the data products live and where every agent lives, the small molecule, antibody and ASO designers, the target and causal models, the multimodal foundation model that an integrated incumbent trains. It is also, and this is the point, the layer that fits in a file.

The Make and Test steps do not collapse. The Make stays in the physical world, still relatively slow and expensive. The Test is more than a step. It creates the criteria against which the entire computational layer is optimized. The computational layer makes each physical cycle more targeted, so you run fewer and better-chosen experiments, but that does not relieve the pressure on the Test. It raises it, because the whole optimized machine is now betting on the validity of the single experiment it chose to run. Hand someone the collapsed Analyze and Design layer and you have handed them the part that copies. The physical system that produced it, and the objective function it answers to, do not. Venture capital is moving on the same instinct: a16z recently raised $1.1B for the physical buildout of AI, on the thesis that durable value now accrues to whoever owns the physical engine rather than the model that fits in a file.[b] The instinct is right,

and value collects wherever copying fails. But compute and biology part ways on what gets plugged into the machine. A data center's output is fungible, so more throughput is simply more value. An experiment's output is not.

It was a data shift, not an algorithm shift

The claim that virtual assays have started to return real value is true. The reason usually given for it is wrong. Compute grew and architectures matured, but the binding constraint in biology was rarely the algorithm. It was generating data fit for purpose. The Design step pulled ahead first, for reasons that had nothing to do with importance. It optimized against clean, measurable, scalable signal, while biology lagged because its constraint was the data. What AlphaFold did, and what a company like Chai now commercializes, is genuinely hard and genuinely clever, fusing sequence with experimental structure and folding in physical priors like known bond distances and angles. Even that success was unlocked by decades of painstakingly generated structural data, and the same dependency governs everything downstream.

The lesson is not that architecture stopped mattering. The model can be genius, but the data still decides how far it goes. In recent years many bets have been placed in AI-native Design, covering a spectrum of therapeutic modalities and sharing the central claim that drug design will be faster, cheaper, and more approachable for targets that seemed undruggable before. The best, Iambic among others, shows how far a design-native approach can go, in the science and in the partnership deals it has earned. What none of it settles is what the Test system on the other end should be, which is the question this paper answers. VC deep-tech diligence in this space focuses on teams and ideas perceived to create explosive growth first and moat second. On the other hand, pharma evaluates through a lens of decades of experience and letdowns. Combining the two sets of experiences and goals is important, in my opinion.

The biggest question mark in the Design space is the ability to build a sustainable moat by generating proprietary test data that feeds back into ever-better design. It is at that junction, though, that I find myself in a strange position, with a belief system that is less common. I don't think there is a consensus on what that data should be. But before we go there, let's talk about a few things the field agrees on.

Where the field now agrees

The field increasingly agrees that the durable value in AI drug discovery accrues to biology-native data rather than to the modeling stack sitting on top of it. It agrees that the datasets worth building are scalable, multimodal, and generated at a fidelity the model can actually use, not scraped together from whatever was convenient. It agrees that the computational tooling is commoditizing, that structure predictors, property models and molecular simulators are now widely available, and that the efficient move is to mosaic the best of them rather than rebuild them in-house. It agrees that agentic workflows are moving across the whole of R&D, and that, at least for start-ups and fast-growing companies, owning a fast, closed experimental loop beats renting one through a CRO, because iteration compounds and coordination overhead does not. Read the

sharper landscape pieces coming from VCs and you will find most of this laid out clearly. If it were the whole picture, this paper would simply be another statement of a shared thesis.

It is easy to agree that better data is needed. It is much harder to create it. As Derek Lowe put it in his read of our Perspective: "Everyone agrees biology needs better data. The point is that almost nobody can actually make it. It is hard and it is expensive."

That is precisely why the point lands. The need for better data is not the insight. The ability to generate it at scale, with the quality and predictive power required to change drug discovery, is. The difficulty is not a flaw in the plan. It is the moat. It is where the value is created.

What "informed selection" leaves blank

The scaled-measurement thesis quietly leans on an idea it never makes explicit. The answer offered to the data chasm is coverage: measure enough human causal biology, with informed selection, and the mechanism will emerge. But informed selection is doing enormous unexamined work. Which contexts, which perturbations, which readouts are disease-relevant enough to be worth the petabytes? That is the predictive validity question. Predictive validity, the concept Jack Scannell put into this field, is the formal criterion the coverage argument leaves blank[2]. Scale tells you how much to measure. It does not tell you whether the thing you are measuring predicts a patient outcome or the success of a drug.

In Figure 1, the top-left quadrant is where the scaled measurement world actually sits, enormous in data but narrow in biology. Transformed lines, HEK293T, HeLa, HUVEC, EndoC. The virtual cell efforts belong here too, with the largest perturbation atlases still spanning only a handful of cell contexts, immortalized lines almost throughout, which is a narrow range of low predictive validity data.

Before I say anything about where the field goes next, I want to be clear about how far it has already come because I was part of getting it here. Over the last few years, Recursion, together with Roche and Genentech, built the most serious phenotypic maps of neuronal biology that exist. The Neuromap, delivered in 2024, required manufacturing over a trillion hiPSC-derived neurons, knocking out every one of the roughly 17,000 genes in the genome, and reading the result as high-content images at a scale no one had reached in a cell type this hard to make. The Microglia Map followed in 2025: over 100 billion hiPSC-derived microglia, 46 million cellular images across 17,000 genes and thousands of compounds, in what is arguably the most phenotypically reactive human cell type anyone has ever tried to standardize. And in August of 2026 the collaboration advanced its first neuroscience target from that work into a joint small molecule program, after clearing pathway, functional, and disease validation in sequence.

I do not gesture at this work from the outside. I helped architect it, and I am proud of it. It is the clearest existing proof that disease-relevant hiPSC phenotyping at genome scale can surface a genuinely novel target and carry it far enough that a world-class pharma partner options it. Neuroscience is where drugs fail most and where the field has looked under the same lamppost for twenty-five years, and the Recursion/Roche-Genentech partnership is a real light switched on

somewhere new. Nothing in what follows is a critique of these efforts. Rather, it is what twenty years of translational modeling, in academia and industry, taught me.

Genome-scale perturbation in simplified cellular systems can produce extraordinarily rich maps of gene function and systems biology, showing how perturbations reshape cellular state and which genes behave alike. But a map of gene function is not a map of disease. It is a larger and better-lit lamppost, rather than an escape from it. For some diseases and targets, any cell will capture the phenotypic relationship as a consequence of a perturbation. But for the most part and for almost all of the hard-to-address human diseases, that picture is different. Human disease phenotypes often emerge only in the relevant cellular niche and unfold over time, through cell-autonomous and non-cell-autonomous effects and an interplay of metabolic coupling, cellular crosstalk, and immune interactions.

Figure 1's bottom-right quadrant is the opposite trap. Increasing physiological relevance is the goal, pursued through increasingly complex engineering solutions designed to improve the predictive validity of the data those systems generate. Yet the capacity to generate high-dimensional data at scale, not hypothesis chasing with a proxy assay, is key to value. I have spent many years in the bottom-right quadrant myself, as an academic PI. There I used sophisticated-looking models, expensive, heterogeneous, hard to perturb and harder to assay. These are organoids, assembloids, organs-on-chip, and animal models, with the clinical trial pinned at the far right as maximally predictive. No experimental model will have a predictive validity coming close to a well-powered, well-designed clinical trial. There is often a link between the complexity of the engineered system and the ability to scale, perturb and measure cost effectively in the experimental model. I call this the trade-off between increasing the biological opportunity in an experimental model and handling the operational challenges and costs. Today the most predictive perturbation biology is the least industrialized. That is inconvenient, and it is exactly where the work has to happen, because it is the part of the problem that does not commoditize.

At the sweet spot you will find the thesis of this paper. Predictive validity is the biggest lever there is to get to successful drug discovery[2]. I named the company I founded VALID, so it is safe to assume I have thought about predictive validity more than is strictly healthy.

Fidelity is whether you measured the thing faithfully: relevant physiology, gene expression networks, cellular functions with relevance to disease, and ideally aspects of pathophysiology emerging over time in the specific cell niches, not in every cell. Validity is whether the thing you faithfully measured predicts patient outcomes. A flawless readout of a system that does not track human disease is a precise answer to the wrong question, and no amount of scale, modality, or speed converts it into the right one. That distinction is not a matter of emphasis. It changes what you build.

Scale amplifies validity. It cannot manufacture it. A test system with low predictive validity, run across a million wells, does not converge on truth. It returns a million confident errors, faster and cheaper than before, and an optimizer pointed at it will climb that hill with perfect discipline and arrive nowhere a patient cares about. Algorithmic correction can only recombine the separability the assay already produced. It cannot conjure separability the assay never generated. Automate a

low-dimensional assay with low-fidelity biology and point AI at it, and you have built a machine that scales the wrong signal faster.

A useful data product is not a general capability. It is built for a specific organization, shaped by the people who will act on its output, tuned to the readouts that move their programs and the thresholds that change a decision on a Monday morning. Lift it out of that context and you hold the answers to questions no one at a new location is asking, in a form no one there can use. That is the fundamental challenge for AI model-focused startups: models can be powerful, but without proprietary data generated in the context of a real decision-making process, they remain tools looking for a problem. The well-capitalized AI frontier labs entering this field are placing big bets. And the hardest problem to solve, you guessed it, is generating or finding better proprietary data.

Beyond organizational know-how, capabilities, assets, and portfolio strategy, the moat is physical: the experimental engine that generates and validates ground truth. Every model is ultimately a derivative of that underlying apparatus.

This is why pharma keeps building internal data and modeling capacity and is increasing its externalization efforts to generate meaningful data assets despite frontier labs being a phone call away. This is the part that explains why I get up every day, what drives me, and why I think this is where the value is. Bad data, no matter the scale, will be a liability. Good predictive perturbation data will soon be recognized as the main value driver in successful AI-native drug discovery.

This is what pharma, arguably the most sophisticated buyer in AI-native drug discovery, is looking for. I talk to many friends in pharma. The scientists inside the leading companies, the ones who have run these programs and paid for the failures, tend to arrive here. This is where pharma will pay for AI assets or disease intelligence deals.

Why the window is closing

AI is going to have an enormous impact on biology and drug discovery. What is missing is the data, and the honest position today is that the right proprietary data does not exist. You cannot buy it. There is no vendor to call.

Building it is hard, and even executed well, it takes years. Teams, capabilities, automation, and the assay biology itself do not arrive quickly, and the parts that matter most cannot be bought in. That is why I think the starting gun has already fired. If you have not started, I believe you are late. VALID is 20 months in, we are 20 people, and I have 20 years of experience: leading hiPSC disease modeling in academia, then learning lab automation and AI analysis at Recursion. We moved incredibly fast, de-risked the biology, built an automated platform and compute infrastructure, and are now in production.

You need a lot of different skills to be successful here and this is also why most of the data factories being built today will not close the gap. Scaling experimentation is not scaling the biology that matters. A factory raises throughput, breadth, and iteration speed, and it does not, on its own, raise predictive validity. They are industrializing the wrong data, at an impressive scale.

The compounding is what makes the timing matter. A test system that tracks human disease gets better every cycle, and the models trained on it get better with it. The value inflection from better data is arriving now, and it will arrive for the small number of companies already generating it.

What we are building

VALID is an AI-native pharma organized around the human test system that sets predictive validity. That system combines human iPSC disease models with engineered reporters anchored to disease axes, multimodal readouts, and a closed experimental loop in which the computational layer optimizes against a Test we designed to track human disease relevant phenotypes, which is what lets us connect a novel mechanism the model surfaces back to biology we already have reason to trust. Our automated platform is now in production and generates petabyte-scale meaningful data per screen.

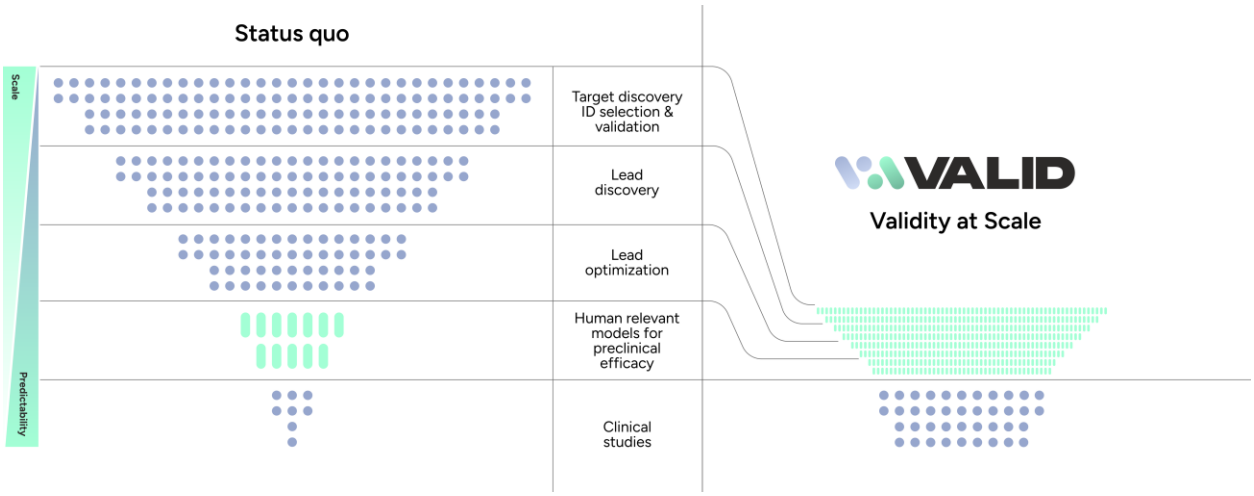

Figure 2. From hypothesis-chasing to efficacy-first. In the status quo (left), most decisions are made earliest, where models are least predictive. VALID (right) inverts the funnel, concentrating decisions where human-relevant models establish predictive validity, so scale is spent where it tracks the clinic rather than where it is merely cheap.

VALID isn't trying to win everywhere. Our strategic choice is to focus on the human test system that sets predictive validity at scale. We intend to be the best in the world at that and nothing else. Everything outside that core is done through partnerships. What comes next is the evidence. Proof-of-concept across roughly ninety prioritized compounds is underway, and multi-day 4D live-cell datasets follow this autumn.

This is the first high-level statement of the prescription, not the last. With colleagues from VALID's Scientific Advisory Board, we are taking the argument into forthcoming peer-reviewed work that formalizes predictive validity and what it demands of a test system.

References

1. Bender, A. *et al.* Artificial intelligence in drug discovery — what it is, where we stand and the path forward. *Nat. Rev. Drug Discov.* https://doi.org/10.1038/s41573-026-01496-2 (2026) doi:10.1038/s41573-026-01496-2.

2. Scannell, J. W. *et al.* Predictive validity in drug discovery: what it is, why it matters and how to improve it. *Nat. Rev. Drug Discov.* **21**, 915–931 (2022).

Footnotes

a. Lowe, D. In the Pipeline, Science. https://www.science.org/content/blog-post/so-how-ai-drug-discovery-doing-really. Accessed 28 August 2026.
b. Horowitz, B. et al. The Machine Age Fund. a16z, 28 August 2026. https://a16z.com/the-machine-age-fund/. Accessed 30 August 2026.