A FULL-STACK PRIMER · BEN POULADIAN

The Future of Drug Discovery Is AI-Assisted

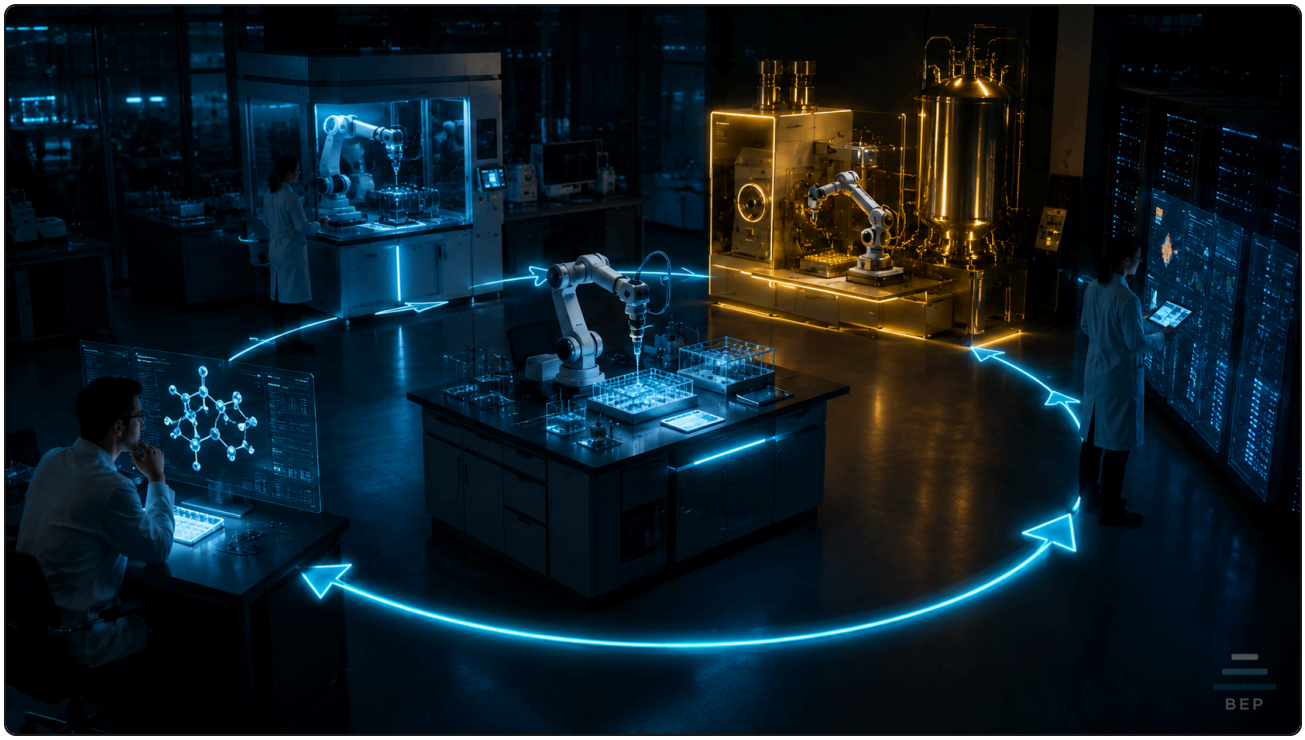
AI conquered language. Its next inflection is the wet lab, and the moat is the measurement.

BEP Research · July 2026 · bepresearch.substack.com · @benitoz

How to use this. This is a reference document, not a narrative, and I do not expect anyone to read it end to end in one sitting. It is built so you can jump to whichever layer you care about, get oriented in a few minutes, and leave. If you want the argument without the depth, the Substack post version covers it in about fifteen minutes, and there is a two-page summary if you want only the map. What this document is really for is context. AI moving into biology is going to be a defining theme for the next several years, and when it shows up in a headline or a pitch or an earnings call, the goal is that you already know which layer it touches and whether that layer is defensible.

The framework, if you want to use it yourself. The method behind this primer, how I score a layer, where I look for the bottleneck, and the research hygiene rules that stop the expensive mistakes, is published as a file you can use: tools.bepresearch.com/primer/CLAUDE.md. Save it into a project folder as `CLAUDE.md` and any agent that reads that convention will pick it up, then point it at filings, transcripts or your own notes. It contains the method and no positions, weights, or price targets, on purpose.

A note on method. I used AI, mostly Claude, to help assemble this primer: organizing research, checking my reasoning, drafting and redrafting sections, and arguing with me about the weak parts. Every judgment, every framework, and every remaining error is mine, and the numbers were verified against primary sources. It seems worth saying plainly in a document about AI-assisted work, and honestly it is part of the point. This is what the tools are genuinely good at right now, and it is also a small demonstration of the thesis: the model made the writing faster, and it could not do the thinking, the sourcing, or the standing in a lab in Los Angeles watching the machines run.



The lab of the future is a loop. AI designs, robots build, instruments measure, data teaches the next cycle.

I did not come to drug discovery as an investor. I came to it after I lost both of my parents to cancer, and after watching a Phase 2 drug my mother needed stay just out of reach, close enough to have a name, too far to actually get. Standing in that gap, between a drug that exists and a patient who cannot reach it, is what turned a private grief into a question I could not put down: what is actually broken inside biotech, and is any of it about to change?

This primer is the long answer. I went looking with the same lens I bring to semiconductors, where I have spent years mapping the physical layers of the AI buildout, and I found the same structure hiding underneath biology. So I will be upfront that the conviction here is personal. The analysis is not. By the end you will have the whole map: every layer, every company, and where the real constraint sits. The people building this have a slogan for the moment:

"For the first time, biology has the chance to be an engineering discipline rather than a purely empirical one."

A refrain now repeated by nearly every frontier-lab and instrument CEO pitching the AI-for-science trade in 2026.

Strip out the slogan and a real claim survives underneath it. For two hundred years, biology advanced by trial, luck, and the slow accretion of experiments. The real change is quieter than a

model guessing a protein structure. It is that the loop between guessing and knowing is finally closing, and the machine on both ends of that loop is being rebuilt company by company, instrument by instrument, in a stack most technology investors have never looked at.

The market has decided the AI story is a compute story. It was right about the last three years, and I think it is about to be wrong about the next three. The token machine got built, the datacenter supply chain got repriced, and the picks-and-shovels of language-model training made fortunes. That trade is mature. The next one is quieter, older, and hides inside companies that make mass spectrometers and pipette tips.

This is a primer, not a pitch. My aim is that by the end you understand the entire AI biolab as a system: every layer, who occupies it, and how the pieces feed each other. Only then, behind the paywall, where I think the money is and how I would express it in twenty names.

How a Drug Actually Gets Made (and Why It's So Hard)

Before we can talk sensibly about what artificial intelligence does to drug discovery, we have to understand the thing it is being applied to. Making a medicine is not one problem. It is a chain of a dozen distinct problems, each with its own failure modes, run over the course of a decade or more, at a cost that would fund a mid-sized semiconductor fab. The industry has a dry name for the whole enterprise, the drug-discovery-and-development pipeline, and the word "pipeline" is doing a lot of quiet work. It implies smooth flow. What actually happens is closer to a gauntlet, where the great majority of candidates that enter never reach the other end.

This chapter walks the gauntlet start to finish. The goal is not to make you a pharmacologist. It is to give you a working map of where the money goes, where the years go, and where the failures cluster, because those are exactly the places where a faster, cheaper method would be worth the most.

Step One: Finding the Target

Almost every modern drug works by physically interacting with a specific molecule inside the body. That molecule is called the **target**. In the overwhelming majority of cases the target is a **protein**, a large, folded molecular machine that carries out some job in a cell. Proteins are built according to instructions written in genes, so when researchers talk about a target they are often naming either the protein itself or the gene that codes for it.

The logic of a target is simple. A disease usually involves some biological process gone wrong: a protein that is overactive, a signal that fires when it shouldn't, an enzyme that builds up a substance the body can't clear. If you can find the specific protein sitting at the center of that malfunction and interfere with it, switch it off, dial it down, block the pocket where it does its work, you may be able to correct the disease. Choosing that protein is **target identification**.

Identification is only half the job. The harder half is **target validation**: proving that hitting this particular protein will actually change the disease, and not merely correlate with it. Biology is thick with associations that turn out to be bystanders rather than causes. A protein might be elevated in sick patients simply because it responds to the illness, not because it drives it. Validation is the work of establishing causation, through genetics (do people born with a broken version of this gene get the disease, or avoid it?), through animal experiments (if we remove this protein in a mouse, does the disease improve?), and through cell studies. Weak validation is one of the deepest reasons drugs fail years later and hundreds of millions of dollars downstream:

the molecule did exactly what it was designed to do to the target, and the disease didn't care, because the target was never the true lever.

Some proteins resist being drugged at all. These are the famously **"undruggable"** targets. A conventional drug works by lodging itself into a well-shaped pocket on the protein's surface, the way a key fits a lock. Many proteins that we know cause disease simply have no such pocket, their surfaces are smooth, or flat, or the interaction we want to disrupt is spread across a broad featureless region with nowhere for a small molecule to grab. The target is validated, the biology is understood, and there is still no obvious way to touch it. A large fraction of the human proteins implicated in cancer and other diseases sit in this undruggable bucket, and expanding what counts as druggable is one of the field's central ambitions.

Step Two: Hit Discovery

Once you have a validated target, you need a molecule that acts on it. The first crude version of such a molecule is called a **hit**, any compound that shows a real, measurable effect on the target when tested.

The classic way to find hits is **high-throughput screening**, or HTS. A large pharmaceutical company maintains a physical library of chemical compounds, often on the order of one to two million distinct molecules, stored in tiny wells across thousands of plastic plates. HTS means testing that entire library against the target, one compound at a time, using robotics to move fast. The test itself is called an **assay**: a controlled experiment engineered to produce a readable signal, a color change, a flash of light, a measured binding, when a compound does something to the target. Run the assay across a million compounds and you are essentially asking, at industrial scale, "does any of this stuff do anything to my protein?"

The economics of HTS are brutal in an instructive way. You might screen a million compounds and come away with a few hundred hits, and most of those will be false leads, molecules that hit the target but also react with everything else, or that only work at concentrations no living body could tolerate. The physical library, however large, is also a rounding error against the space of possible drug-like molecules, which is estimated to run into the range of 10^60, a number with no useful analogy in the physical universe. You are dredging a teaspoon and hoping the ocean cooperated.

This is why **virtual screening**, also called **in-silico screening** ("in silico" means done in a computer, by analogy with in-vitro and in-vivo, which we'll define shortly), matters even before any AI enters the story. Instead of physically testing compounds, you model the target's three-dimensional shape and use software to predict which molecules might fit it, then only synthesize and test the promising ones. Virtual screening lets you search libraries far larger than anything you could hold in a freezer. It has been part of the toolkit for years. Hold onto it,

it is the first place in the pipeline where computation substitutes for wet-lab labor, and it is a preview of the larger substitution the rest of this primer is about.

Step Three: From Hit to Lead, and the Chemistry Cycle That Never Ends

A hit is a starting point, not a drug. It usually binds the target weakly, hits several other proteins it shouldn't, and would be toxic or unstable in a real body. The job now is to take that rough molecule and refine it, atom by atom, into something that could plausibly be given to a patient. The intermediate goal is a **lead**, a hit that has been improved enough to be worth serious investment, and the polishing that follows is **lead optimization**.

Two properties dominate this stage. The first is **potency**: how strongly and at how low a dose the molecule acts on its target. A more potent drug does its job at a smaller dose, which usually means fewer side effects. The second is **selectivity**: how cleanly the molecule hits its intended target and leaves everything else alone. A molecule that is potent but unselective will act on its target and also blunder into a dozen other proteins, and those collateral hits are where many side effects come from.

Improving both at once is the daily work of the **medicinal chemist**, and it runs as a cycle. A chemist proposes a modification to the molecule, swap this cluster of atoms for that one, a colleague synthesizes the new version in the lab, it gets tested in assays, the results suggest the next modification, and around it goes. Each turn of this design-make-test loop takes weeks. A single drug program can run through thousands of these variants over several years, most of them dead ends that teach a little and are discarded. When you hear that a program has been "in the lab for four years," this cycle is largely what those four years were.

What makes the cycle so unforgiving is that a molecule has to be good at many things simultaneously, and improving one often wrecks another. Make a molecule more potent and it frequently becomes larger and greasier, which ruins its ability to be absorbed. These downstream properties travel under the acronym **ADMET**, and they matter enough to spell out:

- **Absorption**, can the drug actually get into the bloodstream, for instance survive the stomach and cross the gut wall when swallowed as a pill?
- **Distribution**, once in the blood, does it travel to the tissue where the disease is, and can it get there (the brain, for example, is walled off from most molecules)?
- **Metabolism**, how does the body chemically break the drug down, and are the breakdown products themselves safe?
- **Excretion**, how is the drug cleared out, and does it linger too long or vanish too fast?
- **Toxicity**, does it poison the liver, the heart, or anything else at the doses required to work?

Two more terms are worth defining here because they govern whether a drug is dosable at all. **Pharmacokinetics (PK)** is what the body does to the drug, how its concentration rises and falls over time after a dose. **Pharmacodynamics (PD)** is what the drug does to the body, the effect it produces at a given concentration. A viable medicine needs a PK/PD profile where a dose a patient can reasonably take keeps enough drug in the body, for long enough, to produce the effect without crossing into toxicity. Many molecules that are beautifully potent and selective in a test tube never become drugs because their PK is hopeless: they are cleared in minutes, or would require a dose the size of a golf ball. Lead optimization is the multi-year negotiation to satisfy potency, selectivity, and the whole ADMET and PK/PD checklist at the same time, in one molecule, with no property allowed to fail.

Step Four: Preclinical Testing and the First Regulatory Gate

When a program has a lead molecule that looks good on paper and in cells, it enters **preclinical** development, the last stretch before it is ever given to a human. Testing here comes in two forms whose names you'll see constantly. **In-vitro** ("in glass") means experiments in cells, tissues, or purified proteins in a dish. **In-vivo** ("in the living") means experiments in living animals, typically rodents first, then a second species. The animal work is where the molecule first meets the full messy complexity of a whole organism, a real liver metabolizing it, a real immune system reacting to it, real organs it might damage.

The central preclinical question is **toxicology**: is this molecule safe enough, at the doses that produce an effect, to justify the risk of putting it into people? Animals are dosed and studied for signs of organ damage, and the results define the safety margin, the gap between the dose that helps and the dose that harms. A narrow margin can kill a program here.

Clear preclinical safety and you reach the first major regulatory milestone: the **IND**, or Investigational New Drug application. In the United States this is the filing a developer submits to the Food and Drug Administration to request permission to begin testing the drug in humans. The IND packages up everything known so far, the chemistry, the manufacturing, the animal safety data, and the plan for the first human study. Filing an IND is the formal line between the laboratory phase of a drug's life and its clinical phase. Everything before it is preparation. Everything after it involves patients, and the costs climb steeply.

Step Five: Clinical Trials, Where Most Drugs Die

Human testing proceeds in phases, each larger, longer, and more expensive than the last, each answering a different question.

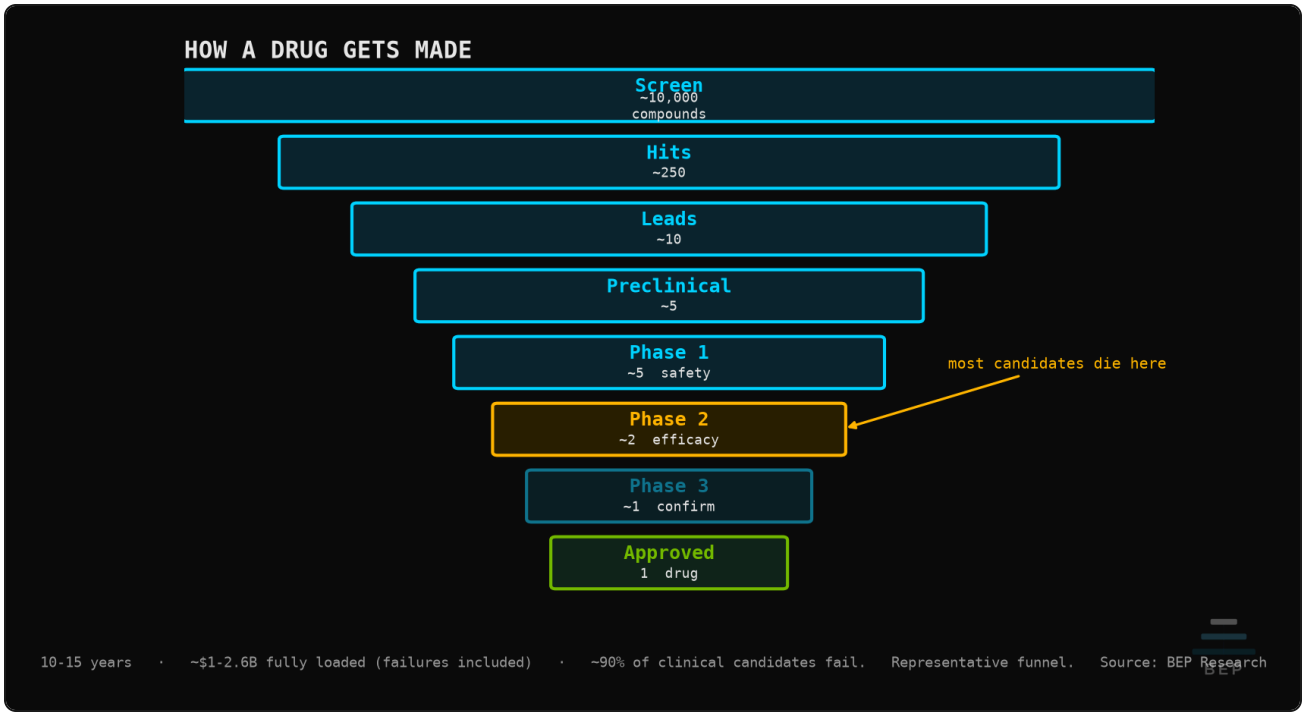
Phase 1 is about safety. The drug is given to a small group, often a few dozen, frequently healthy volunteers, at gradually increasing doses. The question is not "does it work?" but "does

it hurt anyone, and what does the body do with it?" This is the first real-world check on the PK and toxicity predictions from all that preclinical work.

Phase 2 is about efficacy, the first genuine test of whether the drug actually helps patients who have the disease, typically in a group of a few hundred. Phase 2 is the graveyard of the pipeline. This is the stage where a candidate that survived every prior gate, validated target, potent selective molecule, clean animal safety, finally confronts the question that matters, and most often the answer is no. The target turned out not to drive the disease the way everyone believed, or the effect that looked real in mice evaporates in humans. More candidates die in Phase 2 than anywhere else, and they die after the developer has already spent years and a large fraction of the total cost getting them there.

Phase 3 is the large confirmatory trial: hundreds to thousands of patients, sometimes across many countries, designed to prove efficacy and safety rigorously enough to satisfy regulators. Phase 3 trials are the single most expensive component of drug development, and even here candidates fail, a drug that looked promising in a few hundred patients can turn out, in a few thousand, to work no better than existing treatment or to carry a side effect too rare to have shown up earlier.

Only after Phase 3 does a developer file for **regulatory approval**, in the US, a New Drug Application or Biologics License Application to the FDA, which reviews the entire body of evidence and decides whether the drug can be sold. Approval is not the finish line for the science, but it is the finish line for the gauntlet described in this chapter.



How a drug gets made: the funnel from thousands of compounds to a single approved medicine, and why most die in Phase 2.

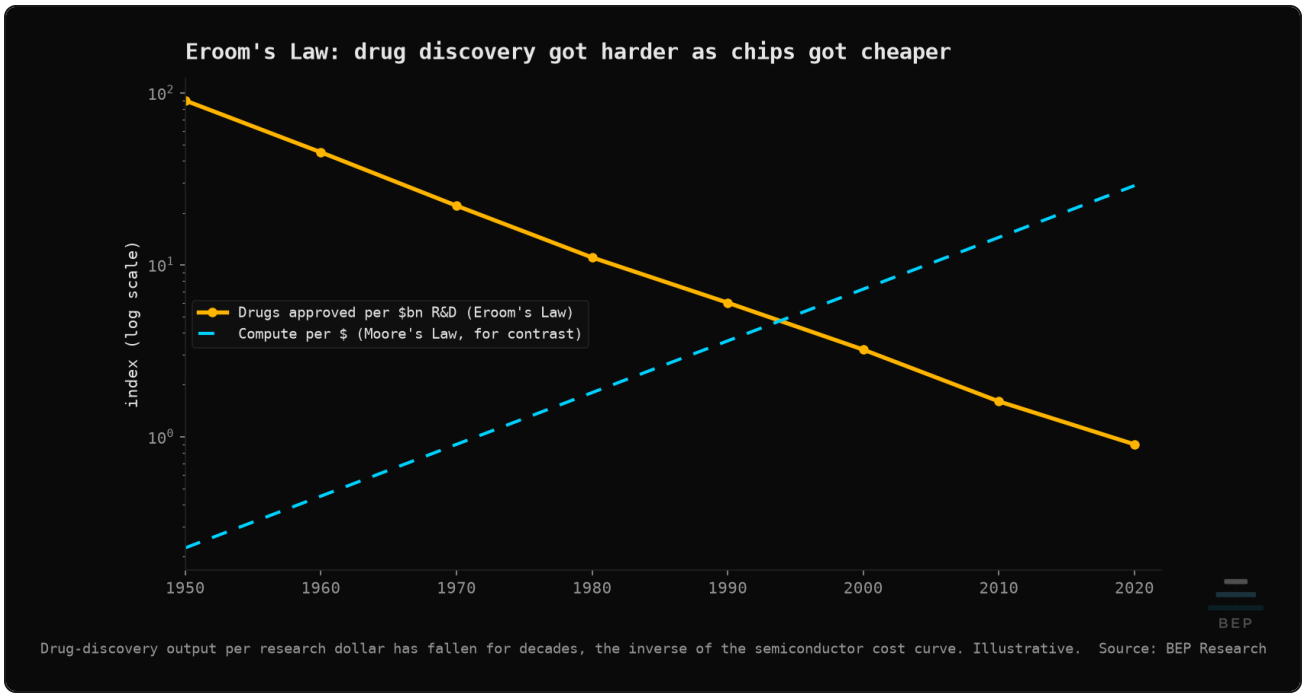
Step Six: The Economics That Make This the Hardest Business in the World

Now assemble the whole chain and look at the numbers, because the numbers are the entire reason this industry is desperate for a better method.

A single approved drug takes, from the start of discovery to approval, on the order of **ten to fifteen years**. The fully-loaded cost of bringing one drug to market, and "fully-loaded" is the key phrase, is estimated in the range of roughly **$1 billion to $2.6 billion**. That figure is not the cost of the winning molecule alone. It includes the cost of all the failures. The reason one approved drug costs on the order of a couple of billion dollars is that the company also paid for the nine or more candidates that entered clinical trials and never made it, and for the thousands of molecules that died in the lab before that. Roughly **90% of drugs that enter human trials fail** to reach approval. You are not funding a drug; you are funding a portfolio of mostly-doomed attempts, and pricing the survivors to cover the dead.

The cruelest feature of this cost structure is that the failures arrive late. A molecule that dies in Phase 2 or Phase 3 has already absorbed most of its budget before revealing that it doesn't work. In almost any other industry, you find out cheaply and early whether your product is viable. In drug development you often find out expensively and late, after the human trials that are themselves the most costly part of the whole endeavor.

Set against this is one of the most sobering observations in all of technology, and it is worth dwelling on because it is the mirror image of the world BEP Research usually writes about. In semiconductors, the cost of a unit of computing has fallen relentlessly for half a century, Moore's Law, the doubling of transistor density roughly every couple of years, delivering exponential improvement in cost and performance. Drug discovery has done the opposite. The number of new drugs approved per billion dollars of research spending has fallen, steadily, for decades. Someone noticed that this was Moore's Law running in reverse and named it accordingly: **Eroom's Law**, "Moore" spelled backwards. Where chips get exponentially cheaper, drugs get exponentially more expensive to discover. It is the inverse of the cost curve that has defined modern computing.



Eroom's Law: drug-discovery output per research dollar has fallen for decades, the inverse of the semiconductor cost curve.

Why would productivity fall as the science and tools got better? A few reasons compound:

- **Biology is genuinely, irreducibly complex.** A chip is something humans designed and therefore understand completely. A human cell is something we are still reverse-engineering, full of feedback loops and interactions no one fully maps. Better tools have not yet closed that gap.
- **The failures are late and expensive,** as described above, so each failure destroys enormous value and there is no cheap early filter that reliably catches the losers.
- **The "better than the Beatles" problem.** Every drug approved becomes the standard the next one must beat, and once a good, cheap generic exists for a condition, a new drug for

that same condition has to clear a higher bar to justify itself, exactly as a new band today competes against the entire recorded catalog of the Beatles, forever available. The backlog of past successes keeps raising the hurdle for every future candidate.

A Necessary Aside: Not All Drugs Are the Same Kind of Thing

One distinction matters before we close, because different companies specialize in different kinds of drug and the tools that help them differ too. The word "modality" simply means the type of therapeutic molecule.

- **Small molecules** are the classic drugs: compact chemical compounds, small enough to be swallowed as a pill and to slip inside cells. Most medicines in history are small molecules. They are the world of medicinal chemistry and HTS described above.
- **Biologics** are large, complex molecules produced by living cells rather than assembled by chemists, most prominently **antibodies**, which are the immune system's own targeting proteins, engineered to lock onto a chosen target with exquisite precision. Biologics can reach targets small molecules can't, but they are large, generally must be injected rather than swallowed, and are far more expensive to manufacture.
- **Newer modalities** extend the toolkit further, among them cell and gene therapies that alter or replace the body's own biological instructions, and RNA-based drugs of the kind that powered the recent generation of vaccines. Each opens targets the older modalities couldn't reach, and each brings its own manufacturing and delivery difficulties.

Keep the distinction in the back of your mind. When a later chapter says a given company or AI method works on small molecules but not antibodies, or vice versa, this is the split it is pointing to.

The Question the Rest of This Primer Answers

Step back and look at the shape of everything above. Target identification, hit discovery, lead optimization, preclinical, the clinic, every one of these stages is the same underlying loop repeated at a different scale. You **design** something: a hypothesis about a target, a molecule, a modification, a trial. You **make** it: synthesize the compound, run the experiment, dose the patients. You **test** it: read the assay, the animal study, the trial result. You **learn** from what came back, and that learning feeds the next design. Design, make, test, learn, around and around, for ten to fifteen years.

Today that loop is run by humans, at the speed of physical laboratory work and the pace of human intuition, and it is slow, expensive, and mostly wrong. A medicinal chemist can imagine and synthesize only so many molecules a year. An experiment takes as long as biology takes.

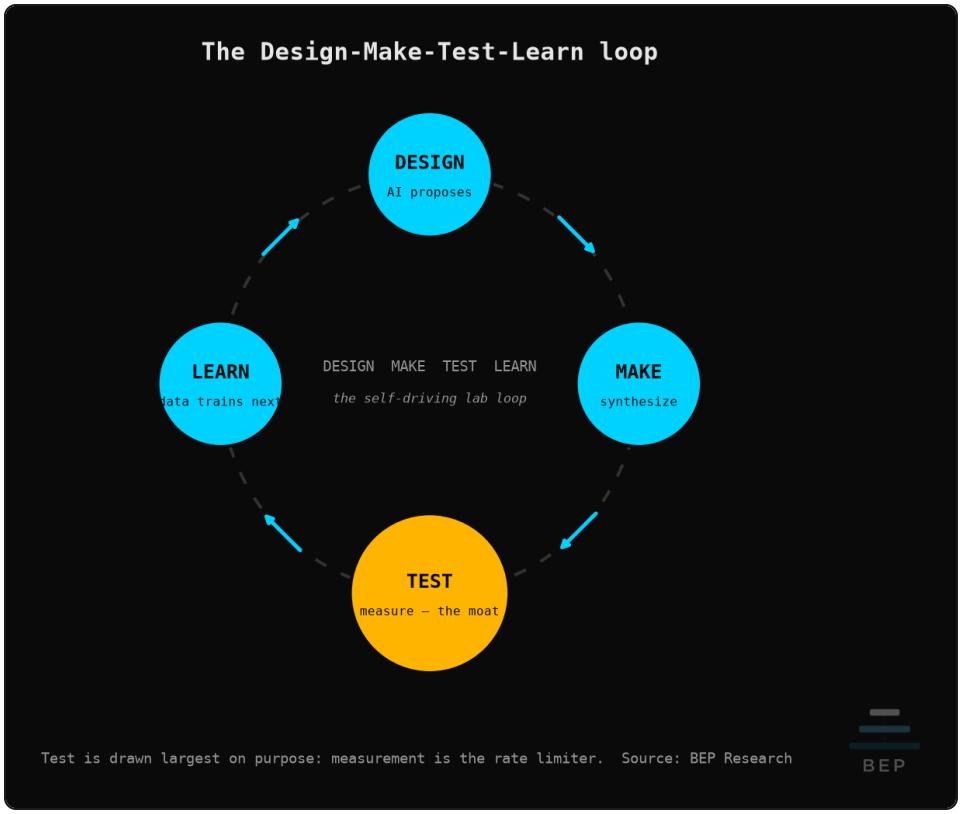
Each turn of the loop costs real time and real money, and Eroom's Law is simply the accumulated verdict on how inefficient the loop has become.

The promise of artificial intelligence in drug discovery is, at its core, a single claim about that loop: that a machine can run the design step vastly faster and cheaper than a human, can help choose which molecules are worth the expensive make-and-test steps, and can learn from each cycle in ways that compress years into months. Whether that promise is real, where in the loop it is already working, where it is still marketing, and which companies are positioned to capture the value if it is real, that is what the rest of this primer sets out to answer. The loop has a name we will use throughout: **Design-Make-Test-Learn**. Everything that follows is about what happens when you try to run it at the speed of computation instead of the speed of the bench.

The Reframe: The Lab Is a Loop

Start with the mental model, because it organizes everything that follows. A modern discovery lab is no longer a room full of benches. It is a closed loop with four steps, and the whole industry is a race to spin that loop faster.

Design: an AI proposes a molecule, a protein, an experiment. **Make:** the candidate is physically synthesized. **Test:** instruments measure what it actually does. **Learn:** the results feed back and sharpen the next design. Design, Make, Test, Learn. Every self-driving lab, every AI-drug-discovery startup, every pharma automation program is an attempt to run that cycle with less human hand and shorter turnaround.

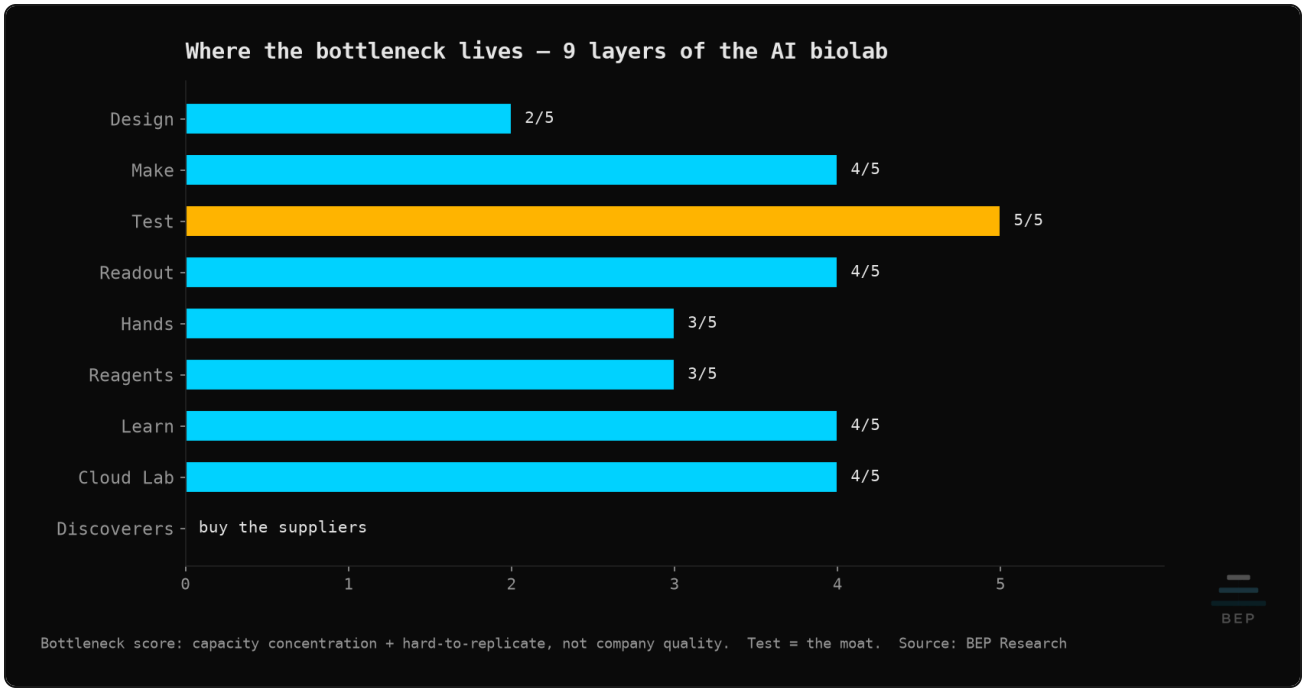


The Design-Make-Test-Learn loop. Test is drawn largest on purpose.

The market gets one thing backwards about that loop. Venture money, headlines, and the Nobel Prize all flow to Design, to the model. And Design is real. But it is also the least constrained step in the cycle. Structure-prediction models diffuse in months, every serious lab rents the same GPUs, and the algorithms are close to a public good.

The binding constraint is Test. A model can propose a billion molecules for the cost of electricity. Finding out which ones actually bind, fold, permeate, and do not kill a cell still requires physically measuring them, one assay at a time, on instruments that cost as much as houses and are built by a handful of companies. The scarce input to a great biology model is not compute. It is measured experimental data at scale, and that data comes out of the measurement layer. The evidence through the rest of this primer points the same way, so I will say it plainly and let the stack earn it: **the moat is the measurement.**

Readers of this letter will recognize the shape. The bottleneck is rarely the layer the market is staring at; it migrated from the GPU to memory, packaging, and power in the datacenter. The same logic applied to the lab points at the instruments.



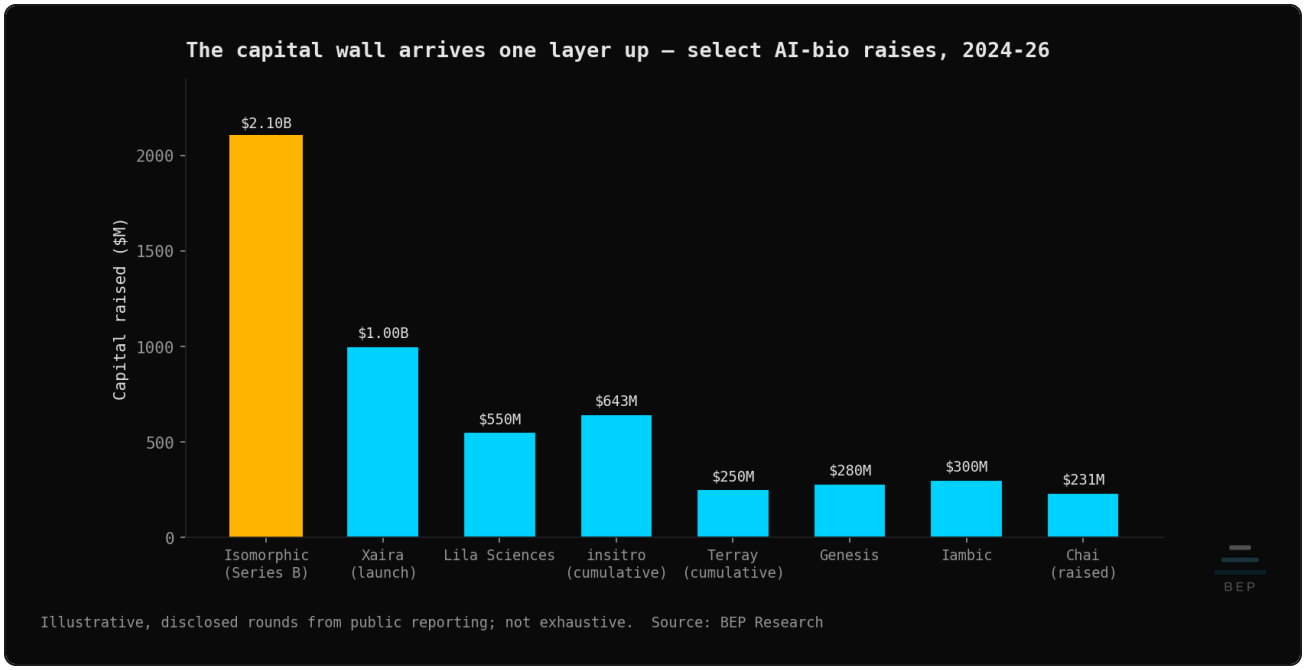
An editorial bottleneck score by layer (capacity concentration and how hard a layer is to replicate, not company quality). Test is the moat.

Why Now: The Inflection Is Real, and It Is Dated

Primers should be honest about timing, because "eventually" is not a thesis. Four things happened between 2024 and 2026 that turned AI-for-science from a promise into a capital event.

The Nobel. The 2024 Chemistry Nobel went to the people who built AlphaFold and modern computational protein design. That is the field certifying that the science works, which is the signal capital waits for before it commits at scale.

The money. And it came. Per public reporting, Isomorphic Labs raised a $2.1B round; Xaira launched with roughly a billion dollars, the largest seed the sector has seen; Lila Sciences raised $550M to build "AI science factories"; and Terray, insitro, Genesis, Iambic, and Chai each raised nine figures. This is the same wall of capital that hit AI infrastructure in 2023, arriving one layer up the value chain.



Select disclosed AI-bio raises, 2024-26. Illustrative and not exhaustive; figures per public reporting.

The models crossed a threshold. Structure and interaction prediction went from research curiosity to daily tool. Protein language models are now designing functional proteins that work in the wet lab, most visibly Profluent's OpenCRISPR-1, an AI-designed gene editor, and the enzymes generated by EvolutionaryScale's ESM3. The generative step stopped being the joke and became the easy part, which is exactly why the bottleneck moved downstream.

The substrate got built. NVIDIA laid the compute-and-model rails under the whole field: BioNeMo for model training and serving, Clara and Parabricks for imaging and genomics, plus a billion-dollar co-innovation lab with Eli Lilly and NVentures checks into the leading privates. It is selling the picks to the biologists now, the same way it sold them to the language labs.

And this summer, the frontier labs themselves arrived. The clearest sign an inflection is real is who starts spending to own it. In June 2026 NVIDIA shipped its BioNeMo Agent Toolkit at the BIO convention, turning large language models into agents that call the field's proven tools, protein folding, molecular docking, generative chemistry, on demand. OpenAI released LifeSci-Bench, a benchmark for how well models handle real life-science tasks. And Anthropic, the maker of Claude (the model I used to help assemble this primer), bought Coefficient Bio, a small team of ex-Genentech researchers, in a roughly $400 million deal, its largest acquisition to date, then laid out its own drug-development ambitions in public, most concretely by shipping Claude Science on June 30, an "AI workbench for scientists" with sixty-plus built-in skills for genomics, proteomics, and structural biology, and integrations into NVIDIA's BioNeMo toolkit. The revealing part is the deployment: Claude Science runs on the

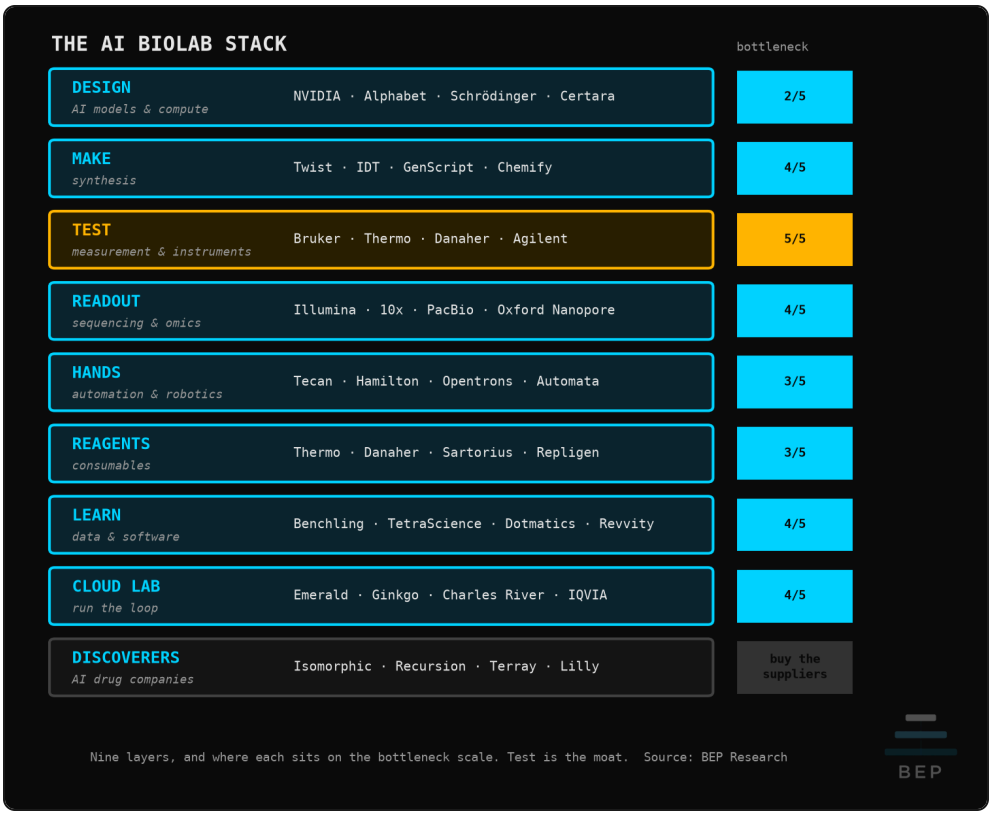
lab's own infrastructure, so the sensitive data never leaves the building, the model company conceding in a product decision that the data is the one input it cannot have. NVIDIA, OpenAI, and Anthropic all moved into biology within a few weeks of each other; you do not get that kind of clustering at the start of a hype cycle, you get it when the people with the best information think the window is now.

And in August 2026 the people moved, which is the least reversible signal of the set. On August 5, Jeff Dean left Google after twenty-seven years to co-found **Discovery Loop**, a public-benefit corporation whose stated purpose is to automate complex, multi-step science and engineering experiments end to end, with Oriol Vinyals, Quoc Le, and Sanjay Ghemawat as co-founders and Google itself as founding investor, cloud partner, and supplier of its first year of compute. Its roadmap starts with machine-learning research and extends into hardware design, drug discovery, and clean energy. The same day, Demis Hassabis handed off day-to-day leadership of Google DeepMind to Koray Kavukcuoglu, took the chair of DeepMind and the chief-scientist role at Alphabet, and said he would lean further into Isomorphic Labs to work on curing disease. Alphabet fell roughly 5% on the reshuffle. And a week before that, DeepMind had disbanded the AlphaFold team outright, scattering it into Gemini and Isomorphic, after Nobel co-laureate John Jumper had already left for Anthropic in June.

Put those together and the center of gravity is visibly shifting from building the model to running the experiment. That is the argument of this entire primer, and I did not have to make it, because the people with the best information in the field made it with their own careers inside a single week. Note in particular what Dean named the company. The design-make-test-learn loop that organizes this document is not a metaphor I imposed on the industry; it is the thing the industry's most accomplished systems engineer just left the world's best AI lab to go automate.

And the data owners are not going quietly, which is itself the tell that the data is the asset. In early July 2026, on the All-In podcast, David Friedberg, who runs the crop-genomics company Ohalo, described Anthropic quietly approaching large life-sciences companies to pool their proprietary data into a new life-sciences model in exchange for early access and an NDA. His read, and, he said, the read of nearly everyone he had spoken with, was blunt: they are *"basically trying to commoditize everyone's business."* A drug company's tens of billions of dollars of experiments produce proprietary measured data, and handing it to a model company to blend with everyone else's gives away the one asset it actually owns. So they are saying no, and increasingly building their own models on their own data instead. That standoff is the whole primer playing out in real time: the people sitting on the data have worked out that it is the one scarce thing, and they are refusing to hand it over.

The inflection is not that a single drug got approved by an AI, and it is worth being precise: as of this writing, no fully AI-designed drug has yet won FDA approval. Anthropic itself said as much while making its case. The inflection is that the loop became industrial. And that gap, enormous model progress on one end, no approved drug on the other, is the whole thesis in miniature: the design step raced ahead, while the wet-lab step that actually validates a molecule, the measurement, is still the scarce, gating, valuable part. So let us walk the machine, layer by layer.



The nine layers of the AI biolab, and where each sits on the bottleneck scale.

The Design Layer — Teaching a Computer to Invent a Molecule

Every drug is a small physical object that has to fit against a much larger physical object. The large object is usually a protein, a folded chain of amino acids that the body uses to do work, whether that is carrying oxygen, copying DNA, or relaying a signal that tells a cell to grow. Disease often means one of these proteins is doing too much, too little, or the wrong thing. A drug works by binding to that protein and changing its behavior: blocking it, stabilizing it, or flagging it for destruction. The entire design problem reduces to one question. What molecule will lock onto this specific target, in the right spot, tightly enough to matter and selectively enough not to hit everything else?

For most of pharmaceutical history that question was answered by making molecules in a lab and testing them one at a time. The design layer is the attempt to answer it inside a computer first, to narrow millions of candidates down to a few dozen worth synthesizing before anyone touches a pipette. This is the part of the drug-discovery pipeline where the recent AI advances have been most visible, most celebrated, and, as we will argue, least defensible as a business.

Why 3D Shape Is the Whole Game

A protein is manufactured as a flat string of amino acids, twenty possible letters in a specific order encoded by a gene. But it does not stay flat. The chain folds spontaneously into a precise three-dimensional shape, and that shape is what determines function. A drug has to fit the folded surface, the pockets, grooves, and crevices of the real object, not the string that describes it. For fifty years, reading the string was cheap and getting the folded shape was expensive, requiring techniques like X-ray crystallography that could take a graduate student a year to resolve a single structure.

This gap, sequence known, structure unknown, was called the protein-folding problem, and it stood open for half a century. In 2020, Google DeepMind's AlphaFold2 effectively closed it, predicting 3D structure directly from amino-acid sequence at an accuracy that rivaled physical experiments. DeepMind then released predicted structures for essentially every protein known to science, hundreds of millions of them, into a public database. Work that took a year could now be done in minutes on a laptop. The 2024 Nobel Prize in Chemistry recognized this: half to Demis Hassabis and John Jumper of DeepMind for AlphaFold, and half to David Baker of the University of Washington for computational protein design, the inverse problem of building new proteins from scratch. It was, in effect, a Nobel for teaching computers the geometry of life.

Structure matters for drug design because you cannot rationally design a key without seeing the lock. Once you can predict a target's shape, you can start asking which molecules will fit it, and you can do that computationally, in bulk, before committing lab time and money.

From Shape to Binding: Docking and Physics

Knowing the shape of the lock is not the same as knowing which key turns it. That is the binding problem, and it comes in tiers of increasing rigor and cost.

- **Docking** is the fast, cheap first pass. The software takes a candidate molecule and computationally tries to fit it into the target's pocket in millions of orientations, scoring each for how snugly it sits. It is a rough geometric filter, good for triaging a huge library down to plausible candidates, unreliable as a final verdict.
- **Molecular dynamics (MD)** is a physics simulation. It models every atom and the forces between them and lets the whole system move over time, the way it actually would in the watery environment of a cell. This captures the fact that proteins are not rigid statues; they wiggle, breathe, and flex, and a pocket that looks closed in a static picture may open under motion.
- **Free-energy perturbation (FEP)** is the high-accuracy tier. It uses MD-style physics to predict, with genuine quantitative precision, how much more tightly one candidate binds than another. FEP can tell a chemist that adding a single fluorine atom in a particular spot will improve binding by a specific amount, the difference between a guess and a measurement. It is computationally brutal, which is exactly why it runs on GPU clusters rather than desktops.

The strategic point is that binding prediction sits on a compute-versus-confidence curve. Cheap methods scan wide and miss; expensive methods are trustworthy but only affordable for a handful of finalists. The commercial value concentrates where a company can push accuracy up without letting cost run away, and that is a software-and-hardware problem, not a magic one.

Cheap methods scan wide and miss; expensive methods are trusted but only affordable for a handful of finalists.

Generative Chemistry and De Novo Protein Design

Everything above evaluates molecules that already exist. The newer and more radical capability is inventing ones that do not.

Generative models, the same broad family of technology behind image generators, can be pointed at a protein pocket and asked to hallucinate a molecule shaped to fill it. Diffusion models, which learn to turn random noise into structured output, have proved especially good at this. The lineage that traces back to David Baker's lab, notably RFdiffusion, can design entirely new proteins to order: give it a target surface and it generates a novel protein that binds there, a protein that has never existed in any organism. Profluent used models in this vein to produce OpenCRISPR-1, a functional gene-editing enzyme designed by AI rather than borrowed from bacteria.

A parallel idea is the protein language model. Proteins can be treated like sentences in a language whose grammar evolution wrote over billions of years. Train a large model on hundreds of millions of natural protein sequences and it learns which arrangements are viable and which are nonsense, the same way a language model learns that some word sequences are fluent and others are gibberish. EvolutionaryScale's ESM3 is the leading public example, a model reasoning jointly over a protein's sequence, structure, and function; it can generate proteins with desired properties by, in effect, writing in the language of biology. Chai Discovery (Chai-1, Chai-2), Latent Labs, and others are building in adjacent territory. The upshot: design has moved from screening what nature already made to authoring what it never got around to.

The furthest expression of this to date arrived on August 6, 2026, when a Stanford and Arc Institute team reported that a generative model called Evo had designed complete viral genomes from scratch, the first whole genomes authored by an AI rather than derived from a natural template. The result is worth reading carefully in both directions. On the capability side, it is remarkable: the functional designs carried between 67 and 392 novel mutations, and by some taxonomic standards one of them would qualify as a new species. On the investment side, the number that matters is the yield. The team physically synthesized 285 candidate genomes and sixteen were viable, a hit rate near six percent, and there was no way to identify those sixteen except to build all 285 as real DNA and test them. That is this primer's argument stated as an experiment: generation is cheap and abundant, selection is physical and scarce, and the 285 syntheses and assays required to find the sixteen are precisely the recurring demand the make and test layers sell into.

Two clarifications, because coverage of this result has been loose. The work was led by Brian Hie at Stanford and published in *Science*, using the Evo models trained on the genomes of roughly two million bacteriophages. These are phages, which infect bacteria rather than people; the ones built here showed restricted tropism to a non-pathogenic laboratory strain of *E. coli*. And the genetic code of viruses capable of infecting humans, animals, or plants was deliberately excluded from training, so the model cannot generate a human pathogen. The demonstrated application is therapeutic rather than adversarial: the AI-designed phage cocktail cleared *E. coli* that natural bacteriophages could no longer kill, with near-term targets named as *Pseudomonas aeruginosa* and the plant pathogen *Xanthomonas campestris*. Antibiotic resistance is one of the few therapeutic problems as economically ugly as the ones described in Part One, and this is the first credible case of a model designing that class of medicine end to end.

The biosecurity discussion around the paper deserves attention, and it lands somewhere useful for this primer's argument. Writing alongside the result, Tom Inglesby and Moritz Hanke of the Johns Hopkins Center for Health Security warned that the capability to compose viral genomes with generative AI now exists while *"the governance to safely steer it does not."* Tom Ellis of Imperial College London supplied the deflating counterweight, noting this is the smallest and easiest genome anyone could attempt and that the AI-design threat is modest next to the far simpler route of modifying pathogens that already exist. Most relevant here, Filippa Lentzos of King's College London argued that regulation should not focus solely on the model, and that a layered approach runs through model access, research review, laboratory biosafety, and critically **synthesis screening at the point DNA is physically manufactured**. That is a governance conclusion arriving at the same place this primer arrives at commercially: the enforceable chokepoint in biology is not the software that proposes a sequence, it is the physical infrastructure that writes and measures it. The layer that can be controlled is the layer that is scarce, and the layer that is scarce is the layer with pricing power.

Will It Survive Contact With the Body? ADMET In Silico

A molecule that binds beautifully is still worthless if it poisons the liver, gets flushed out in an hour, or never reaches the target tissue. These properties travel under the acronym ADMET, absorption, distribution, metabolism, excretion, and toxicity, and they are where most drug candidates quietly die, often late and expensively.

In-silico ADMET prediction tries to flag those failures early by estimating a molecule's behavior in the body from its structure alone. Related to this is PBPK modeling, physiologically based pharmacokinetics, which simulates how a compound moves through and concentrates in different organs over time. Done well, this shifts a fatal discovery from Phase I in humans to a Monday-morning calculation. It is unglamorous and it is one of the highest-leverage places computation touches the pipeline, because the cost of a failure caught here versus in the clinic differs by orders of magnitude.

The Substrate Underneath Everything

None of this runs on air. Structure prediction, MD, FEP, and generative models are all enormous matrix-multiplication workloads, which is precisely the arithmetic that graphics processing units were built to do fast and in parallel. Underneath the entire design layer sits a stack of GPUs, the software frameworks that train and serve the models, and the cloud infrastructure that rents the whole thing by the hour. This substrate is the same one powering the rest of the AI economy, and that shared foundation is the key to reading the investment case, because it means the people selling shovels here are also selling shovels to every other gold rush at once.

The Names You Can Actually Own

NVIDIA (NVDA) is the toll under all of it. Its GPUs run the training and inference for essentially every model named in this chapter, whoever built them. Beyond raw silicon, NVIDIA supplies domain software: BioNeMo, a platform for training and serving biology models (including hosted versions of structure and generative models); Parabricks, which accelerates genomic analysis; and the broader Clara healthcare stack. NVIDIA's position is deliberately neutral, it does not need to pick the winning model or the winning drug company, because it collects on every attempt. When Schrödinger runs FEP, when a startup trains a protein language model, when a pharma rents cloud GPUs for docking, the compute meter runs through NVIDIA hardware. That is the most durable seat at this table, and it is durable precisely because it is agnostic.

Alphabet (GOOGL) owns the most celebrated science in the field. DeepMind produced AlphaFold and its successor AlphaFold 3, which extends prediction from proteins alone to how

proteins interact with drugs, DNA, and other molecules, closer to the actual binding question. Alphabet also monetizes the substrate through Google Cloud. The catch, from an investor's standpoint, is that drug discovery is a rounding error inside a company this size, and its most important contribution, AlphaFold's structures, was given away for free, which tells you something about where the defensible value does not sit. In July 2026 Alphabet made that judgment explicit: DeepMind disbanded the standalone AlphaFold team, moving people onto Gemini and into Isomorphic Labs, after John Jumper, who shared the Nobel for the work, had already left for Anthropic in June. Less than two years after the highest honor in science, the team behind the most celebrated model in computational biology was worth more to its owner dispersed into a general-purpose model and a drug pipeline than kept together as a science-model shop. If you want one piece of evidence for the claim that the model layer commoditizes and the value migrates to the loop, that is it, and it comes from the company with the best model in the field.

Microsoft (MSFT) and **Amazon (AMZN)** are the rails. Azure AI and AWS (including HealthOmics, purpose-built for genomic and biological data) rent the compute, storage, and managed model-serving that discovery teams run on when they do not own their own clusters. Like NVIDIA, they profit from activity rather than outcomes, a cloud bill is owed whether the drug works or not.

Schrödinger (SDGR) is the most established pure-play in physics-based design. Its FEP-plus-machine-learning software has been licensed across large pharma for years, giving it real, revenue-generating adoption and a technical reputation predating the AI wave. Schrödinger also runs its own drug pipeline and takes equity stakes in partner programs, trying to convert a tools business into upside on the drugs those tools help design. That dual model is the tell: selling software alone, even excellent software, has proven a hard way to capture the value the software creates.

Certara (CERT) operates one layer over, in model-informed drug development and biosimulation, using models to inform dosing, trial design, and regulatory submissions, sometimes described as running "in-silico trials." Its moat is less algorithmic novelty than accumulated regulatory credibility: the FDA and its global peers already accept Certara's simulations as part of drug filings, and that acceptance is slow to earn and slow to displace.

The **frontier model labs** are where the newest science lives and where public investors mostly cannot buy in. EvolutionaryScale built ESM3, then saw its team absorbed into the Chan Zuckerberg Initiative's Biohub in 2025, the model diffused, the company effectively did not survive as an independent going concern. Profluent (OpenCRISPR-1), Chai Discovery (Chai-1, Chai-2), and Latent Labs are private. Basecamp Research is pursuing a different edge, proprietary training data drawn from sampling global biodiversity, on the thesis that unique

data, not the model architecture, is what is actually scarce. In China, BioMap's xTrimo is a large-scale biological foundation model. The common thread: the frontier is private, fast-moving, and repeatedly demonstrating that a model, once built, does not stay proprietary for long.

The Investment Read on the Design Layer

Design is indispensable and, as a place to capture durable value, the weakest link in the chain. The reasons are structural. Models diffuse, AlphaFold's structures are free, ESM's weights are broadly available, and academic labs reproduce frontier results within a year or two of publication. Compute is rentable by anyone with a credit card, which means no design shop can fence off the substrate. And the most advanced work keeps happening inside private labs or getting absorbed into non-commercial homes before public shareholders can participate.

What survives that diffusion is not the cleverest model but the position everyone has to pay through regardless of which model wins. That points to the toll, NVIDIA above all, with the hyperscale clouds behind it, over the toolmakers. Among the tool companies, the ones worth watching are those bolting on something the software alone cannot be copied into: Schrödinger's owned pipeline, Certara's regulatory standing, proprietary-data plays like Basecamp. The pure algorithm, however brilliant, is the part of this business most likely to become a commodity. In a gold rush where the maps are free and the picks are rented, the durable business is the one that owns the road.

The Make Layer: Building What the AI Designed

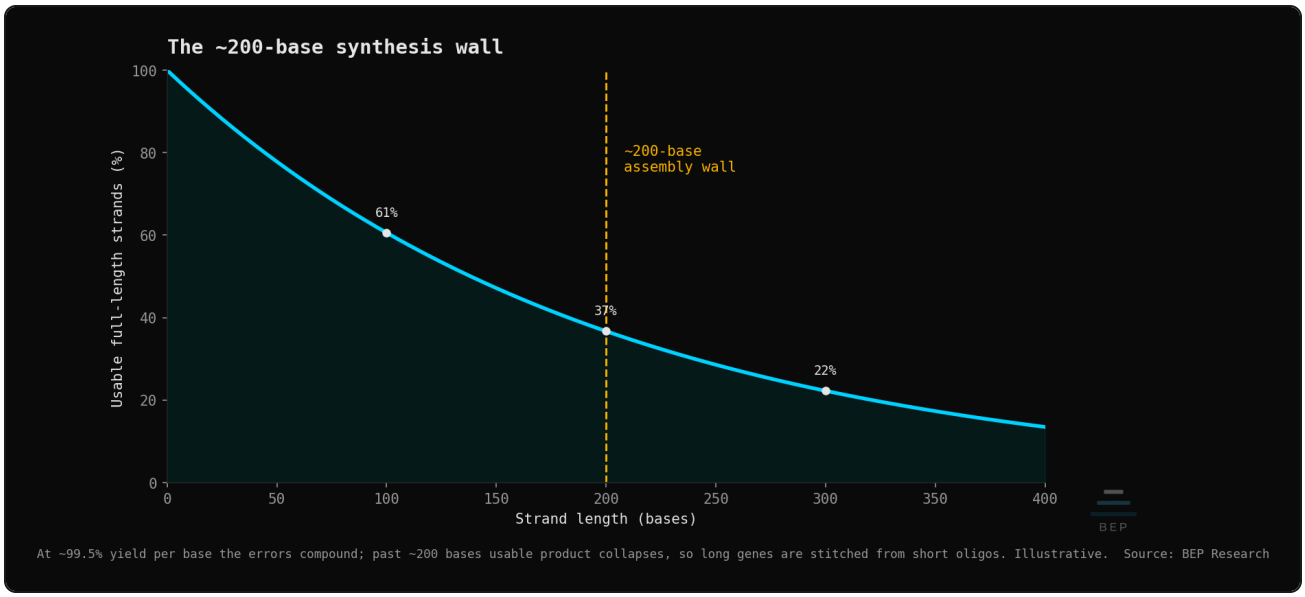
An AI model can propose a protein, an antibody, a small molecule, or an entire metabolic pathway in an afternoon. None of that is worth anything until someone turns the digital design into physical matter that can be put in a test tube. This is the "make" step of the design-build-test-learn loop, and it is the layer where software optimism collides with the stubborn realities of chemistry. A sequence that a model prints in milliseconds may take a lab three weeks to physically construct, and that three-week gap, repeated across thousands of candidates, is the real speed limit on modern discovery. The companies that own the make layer sell time. Understanding what they actually do requires understanding four distinct manufacturing problems: writing DNA, building small molecules, assembling peptides and proteins, and programming cells.

Writing DNA: The 200-Base Wall

Reading DNA (sequencing) has fallen in cost by a factor of millions since 2003. Writing DNA, synthesizing a chosen sequence from scratch, has not kept pace, and that asymmetry is one of

the most important constraints in the entire industry. When a model designs a novel gene, a promoter, or a synthetic pathway, that design has to be physically written base by base.

The workhorse chemistry is **phosphoramidite synthesis**, a four-step cycle that adds one DNA base at a time to a growing strand anchored to a solid surface. Each cycle is highly efficient but not perfect. At roughly 99.5% yield per base, the errors compound: by the time you reach 200 bases, a meaningful fraction of your strands carry at least one mistake, and usable full-length product collapses. That is the practical ceiling people mean when they talk about the **~200-base assembly wall**. Genes and pathways are far longer than 200 bases, so the industry works around the wall by synthesizing short, accurate fragments called **oligonucleotides** (oligos) and then stitching them together enzymatically into genes, and genes into whole constructs. Every additional assembly step adds cost, time, and another opportunity for error.



At ~99.5% yield per base the errors compound; past ~200 bases usable product collapses, so long genes are stitched from short oligos.

The three variables that matter here are worth stating plainly because they govern everything downstream:

- **Length**, how long a fragment you can make before fidelity collapses. Longer fragments mean fewer assembly steps.
- **Fidelity**, the error rate per base. Higher fidelity means less downstream sequencing and cleanup.
- **Turnaround**, the wall-clock time from order to usable DNA. This is the variable AI most wants compressed.

The dominant model today is centralized: you order oligos or genes from a vendor and wait days to weeks for shipment. The emerging challenge is **enzymatic synthesis**, using an engineered enzyme (typically a template-independent polymerase, TdT) to add bases in water-based conditions rather than the harsh solvents of phosphoramidite chemistry. Enzymatic methods promise longer strands, higher fidelity, and, crucially, a machine small and safe enough to sit on a lab bench. That last point is why the phrase **"DNA printer"** matters. If a researcher can write DNA on-site overnight instead of ordering it and waiting a week, the build step of the discovery loop shrinks from weeks to hours. Compressing the loop is the entire value proposition of AI-assisted discovery, and DNA turnaround is one of its tightest bottlenecks.

Building Small Molecules: Software Plans, Robots Execute

Most approved drugs are still small molecules, compact organic compounds a chemist builds through a sequence of reactions. Historically, deciding how to build a target molecule was an art form. Given a desired structure, a chemist works backward, asking "what simpler pieces could I combine to make this?" and repeats until reaching purchasable starting materials. This is **retrosynthesis**, and it is now a machine-learning problem: software trained on millions of published reactions proposes viable synthetic routes, ranked by cost, yield, and number of steps.

The second half is execution. **Automated organic chemistry** platforms take a planned route and run it on robotic hardware, dispensing reagents, controlling temperature, moving product between steps, with far less human intervention than a manual bench. Paired with **parallel synthesis**, where dozens or hundreds of variants are made simultaneously in plates, this lets a team physically test the space of molecules an AI proposes rather than hand-picking a handful to make. The constraint that never goes away is **purification**: making the molecule is often easier than isolating it cleanly from the reaction mixture, and purification remains stubbornly manual and slow. Any honest read of automated chemistry treats purification as the gating step, not the reactions themselves.

Peptides and the GLP-1 Boom

Peptides, short chains of amino acids, sit between small molecules and full proteins in size, and they have become one of the most commercially important molecule classes in medicine. The reason is the GLP-1 class of metabolic drugs (the mechanism behind semaglutide and tirzepatide), which turned peptide manufacturing from a niche into a capacity-constrained industry almost overnight. When a single drug class drives tens of billions in revenue, the machines that make peptides become strategically valuable.

The core method is **solid-phase peptide synthesis (SPPS)**, invented by Bruce Merrifield in the 1960s: the peptide is built one amino acid at a time on an insoluble resin bead, with

reagents washed through and drained away between each addition. Like DNA synthesis, it is a cyclic add-and-wash process, and like DNA synthesis, yield-per-cycle sets a practical length limit, longer peptides are harder and more expensive to make cleanly. The commercial edge in this hardware is speed and purity per cycle, which is why microwave-accelerated and flow-based SPPS instruments command premium pricing in a market that is currently starved for peptide capacity.

Cell-Free Protein Synthesis

Proteins are the actual drugs in much of biologics, antibodies, enzymes, engineered binders. The traditional way to make a protein is to insert its gene into living cells (bacterial or mammalian), grow them, and harvest what they produce. This works but is slow: you spend days engineering and growing the cells before you get any protein, and the cell's own biology limits what you can make.

Cell-free protein synthesis extracts the molecular machinery of protein production, the ribosomes, transfer RNAs, and energy systems, and runs it in a tube without any living cells. You add DNA encoding your protein and get protein out in hours rather than days. Two properties make this powerful for AI-driven discovery. First, speed: the make step collapses, which suits a workflow generating many candidates. Second, it can incorporate unnatural amino acids that living cells cannot tolerate, opening chemistry a cell-based system simply cannot reach, the basis for precisely engineered antibody-drug conjugates. The tradeoff is scale and cost economics that have historically favored living cells for large-volume manufacturing.

Biofoundries: Programming Cells as a Service

The most ambitious version of the make layer treats the cell itself as the product. A **biofoundry** is a highly automated facility that designs, builds, and tests engineered organisms at industrial throughput, write the DNA, insert it into a host cell, grow thousands of variants in parallel, measure which ones perform, and feed the results back into the next design round. The promise is "program cells as a service": tell the foundry what you want a microbe to produce, and the automation iterates toward a strain that does it. This is where the design-build-test-learn loop becomes fully mechanized, and where the AI models of the design layer have the most physical infrastructure to plug into.

The Companies, and Which Ones Last

Twist Bioscience (TWST) is the volume leader in written DNA. Its defining innovation was moving phosphoramidite synthesis onto a silicon chip, miniaturizing the reaction so that where a conventional plate makes 96 oligos, Twist's chip makes thousands in the same footprint, a manufacturing-cost advantage that let it undercut incumbents and scale. Twist sells synthetic

genes, oligo pools, and increasingly antibody libraries and NGS tools. It is the closest thing the make layer has to a picks-and-shovels franchise for DNA.

Integrated DNA Technologies, now inside **Danaher (DHR)**, is the oligo workhorse, the vendor a very large share of labs actually order their custom oligos, primers, and CRISPR guide RNAs from. Sitting within Danaher gives IDT the balance sheet and life-sciences distribution of one of the industry's most disciplined operators. For an investor, IDT is not a standalone bet but a reason DHR has quiet, recurring exposure to the consumable that every molecular biology experiment burns through.

The **enzymatic-DNA challengers** are mostly private and are attacking the turnaround and length constraints directly. **DNA Script** commercialized a benchtop enzymatic synthesizer (the SYNTAX system), the "DNA printer" that puts overnight, on-site DNA writing in a lab. **Ansa Biotechnologies** has demonstrated enzymatic synthesis of unusually long single oligos, pushing directly at the fidelity-limited length wall. **Molecular Assemblies** pursues enzymatic long-read synthesis for gene and cell therapy applications, and **Evonetix** takes a different hardware path with a MEMS chip that controls synthesis with thermal precision at thousands of independent sites. Each is a bet that whoever breaks the 200-base wall economically resets the cost curve of written DNA.

GenScript (1548.HK) is the vertically integrated Chinese player, offering genes, oligos, peptides, proteins, and antibody services across one platform at aggressive pricing, a genuinely broad make-layer supplier. The investment caveat is regulatory: GenScript and its affiliates have been named in the context of the proposed U.S. **BIOSECURE Act**, legislation aimed at restricting federally connected customers from working with certain Chinese biotech providers. That exposure is a real, non-trivial overhang for any Western institutional buyer, and it is arguably a tailwind for the Western suppliers that would absorb displaced demand.

Telesis Bio is the cautionary tale to keep in view. It built genuinely advanced automated DNA-writing hardware (the BioXp benchtop system, out of the Craig Venter lineage) and still could not sustain a public listing, it fell below Nasdaq's minimum equity and public-float requirements and voluntarily delisted to the OTC market in 2024. The lesson is not that the technology failed; it is that impressive hardware without a durable, high-margin recurring revenue stream does not survive as a public company.

In **automated chemistry**, both leaders are private. **Chemify** commercializes the "Chemputer" concept, a system that reads a synthesis route in a standardized chemical programming language and executes it on hardware, tightening the loop from retrosynthesis software to physical molecule. **Chemspeed** is the established supplier of automated and

parallel synthesis workstations that many labs already run. The read here is that the chemistry make layer is being automated by tooling vendors, not by a single public pure-play.

On **peptides**, the hardware names are more accessible. **CEM** pioneered microwave-accelerated SPPS, a speed-and-purity edge that matters intensely in a capacity-starved peptide market. **Gyros Protein Technologies**, part of **Mesa Laboratories (MLAB)**, supplies automated peptide synthesizers, giving MLAB a small but real position in the GLP-1-adjacent tooling story. **Biotage**, taken private by KKR, rounds out the purification-and-synthesis workflow, a reminder that the peptide boom is drawing serious financial-sponsor capital into the picks and shovels.

In **cell-free**, **Sutro Biopharma (STRO)** is the most investable name, but note carefully what it is: Sutro uses its cell-free platform (XpressCF) to build its own antibody-drug conjugate pipeline rather than selling synthesis as a service. It is a drug developer whose edge is a make-layer technology, which means it is valued on clinical outcomes, not consumable throughput. **Nuclera** sells a benchtop system for rapid protein production on demand, and **Tierra Biosciences** offers cell-free protein synthesis as a service to speed the design-build-test loop for other labs.

In the **biofoundry** category, **Ginkgo Bioworks (DNA)** is the largest public bet on programming cells as a service, having built enormous automated foundry capacity. Its history is also a warning about the gap between platform capacity and platform economics. **Culture Biosciences**, private, offers cloud-based bioreactors that let teams run and monitor fermentation experiments remotely, automating the "grow and measure" half of the strain-engineering loop.

The Investment Read: Sell the Blades or Die

The make layer is where the graveyard is. **Amyris**, a pioneering synthetic-biology company, filed for bankruptcy. **Zymergen**, once a high-profile biofoundry IPO, collapsed and was absorbed into Ginkgo. Telesis delisted. The pattern is consistent and it is the single most important thing an investor should take from this chapter: a make-layer platform that cannot attach itself to a recurring consumable or a marketable product does not survive, no matter how good the science looks in a demo. Elegant hardware is not a business. A razor without blades is a liability with a depreciation schedule.

This is precisely why **Twist** and **IDT** endure while flashier platforms fail. They sell the blades. Every synthetic gene, every custom oligo, every CRISPR guide is a consumable that gets used up and reordered, and as AI-driven design generates more candidates, it generates more orders for the DNA that turns those candidates into matter. The demand curve for written DNA is, structurally, the exhaust of the entire AI-discovery engine. Twist's fiscal third quarter, reported

August 3, 2026, is the first clean measurement of that exhaust. DNA synthesis and protein solutions revenue came in at $56.6M, up 39% year over year, with the therapeutics piece inside it up 49% to $40.4M and 369,000 genes shipped in a single quarter. Management raised full-year guidance, guided to adjusted EBITDA breakeven in the fourth quarter, and told investors that orders tied to AI-enabled drug discovery are on track for triple-digit growth in both fiscal 2026 and fiscal 2027, with CEO Emily Leproust describing the segment maturing into *"a durable growth engine."* That is the thesis of this layer showing up in someone's revenue line: models designed more candidates, and somebody had to physically write the DNA.

The same disclosure quantifies the turnaround argument made above, which matters more than the revenue line for anyone trying to judge whether the loop is actually compressing. Measured on a constant basis (roughly one million oligos at femtomole scale), Twist reports cutting turnaround time from 26 hours in the first half of 2023 to 7 hours by its second quarter of fiscal 2026, a ~73% reduction, alongside a ~60% reduction in cost per unit, a ~70% reduction in chemical waste, and roughly 4x the writing capacity out of the same platform. Those are not market-share numbers; they are the physics of the make step getting faster inside one vendor's installed base. Recall that turnaround, not raw cost, is the variable AI most wants compressed, because it sets how many times a discovery team can go around the design-make-test-learn loop in a quarter. A four-fold capacity increase and a turnaround measured in hours rather than days is what "the loop is speeding up" looks like when you put a number on it. Twist also publishes the loop itself in its investor materials, showing which steps it owns (DNA synthesis, protein expression, characterization) and noting that each turn of the cycle drives repeat orders, which is the toll-on-activity argument stated by the company collecting the toll. The enzymatic challengers are worth watching not because a benchtop printer is inherently a better business, but because if one of them genuinely breaks the length-fidelity-turnaround constraint and pairs it with a proprietary consumable (the reagent cartridge, the enzyme, the chip), it inherits the same recurring economics that keep Twist and IDT alive. Absent that recurring attachment, the make layer remains the most seductive and most dangerous place to put capital in the entire drug-discovery stack, the place where the science is most visibly real and the business model most quietly fatal.

The Test Layer — Why Measurement Is the Moat

Start with the cost of an idea. A generative chemistry model, running on rented GPUs, can propose a billion candidate molecules for roughly the price of the electricity it burns doing so. It can suggest structures that might bind a cancer target, block a viral enzyme, or slot into a receptor like a key into a lock. This is the part of drug discovery that has genuinely become cheap. Proposing is cheap. Imagining is cheap.

Knowing is not. A molecule that a model believes will bind is a hypothesis, nothing more, until someone puts the actual compound in front of the actual protein and watches what happens. Does it bind, and how tightly? Does the protein fold the way the model predicted? Will the molecule cross a cell membrane, survive the liver, avoid poisoning the heart? Every one of those questions is answered the same way it has been answered for a century: by physically measuring a real substance on a real instrument. And those instruments are not software. They are house-priced machines, high-field magnets, mass spectrometers, electron microscopes, built by a handful of companies, most of them decades old, most of them boring, most of them trading at multiples that reflect none of the AI narrative wrapped around drug discovery.

This is the argument of this chapter, and it is the load-bearing wall of the whole primer. The scarce input to a great biology model is not compute and it is not clever architecture. It is measured experimental data at scale. And measured data comes out of exactly one layer of the stack: the instruments that do the measuring. Whoever owns that layer sits astride the one bottleneck the models cannot route around. Anyone can rent the model. Almost no one can replicate the instrument.

What "Measurement" Actually Means in Biology

To see why this layer is defensible, you have to understand what these machines physically do, because the difficulty is the moat. A model can be copied in an afternoon. A superconducting magnet cooled to a few degrees above absolute zero, wound to a field homogeneity of parts per billion, cannot.

Mass spectrometry is the workhorse. The principle is almost crude: take a molecule, give it an electric charge (ionize it), fling it through an electric or magnetic field, and measure how its path bends. Heavier and less-charged particles bend less; lighter and more-charged ones bend more. From that you recover the mass-to-charge ratio, and from a molecule's exact mass you can identify what it is. Chain this together across the tens of thousands of proteins in a cell and you get *proteomics*, the systematic identification and quantification of every protein in a sample. Genomics tells you what a cell *could* make; proteomics tells you what it actually made, which is what a drug has to act on.

Not all mass spectrometers are the same, and the differences map directly onto commercial franchises:

- **Orbitrap** instruments trap ions in an electrostatic field and measure the frequency at which they orbit; that frequency converts to mass with extraordinary resolution. This is the high-resolution workhorse of deep proteomics.
- **timsTOF** combines "time-of-flight" (literally timing how long an ion takes to fly a fixed distance) with an added dimension called *ion mobility*, which separates ions by their shape

and size, not just their mass. That extra dimension is what makes it fast and sensitive enough to profile the proteome of a *single cell*, a capability that barely existed five years ago.

- **Triple-quadrupole** ("triple-quad") instruments are the precision counters. They are less about discovery and more about measuring a known compound to exquisite accuracy, the format that underpins clinical testing and pharmacokinetics.

Nuclear magnetic resonance (NMR) attacks structure from a different angle. Certain atomic nuclei behave like tiny magnets; placed in a very strong magnetic field and pinged with radio waves, they resonate at frequencies that depend on their exact chemical neighborhood. Read the pattern back and you can reconstruct how a molecule is put together, atom by atom, in solution. The catch is the magnet. Resolution scales with field strength, and generating a stable, ultra-homogeneous field at the highest strengths requires superconducting magnet technology that only one or two organizations on earth can build. This is not a market share story. It is a near-monopoly rooted in physics and manufacturing know-how that took generations to accumulate.

Cryo-electron microscopy (cryo-EM) is the newest pillar and, for the AI story, the most important. You flash-freeze a protein so fast that water turns glassy rather than crystalline, then fire electrons through it and reconstruct its three-dimensional shape at near-atomic resolution from thousands of noisy images. Cryo-EM is how modern structural biology sees proteins that refuse to crystallize. And here is the connection that most AI-in-biology commentary skips: the structures that AlphaFold and its successors were trained on and are validated against came out of instruments like these. The models are extraordinary interpolators of a ground truth that cryo-EM and X-ray crystallography physically produced. When a model predicts a novel fold, the way you find out whether it is right is by going back to the microscope. The instrument sits both upstream (as training data) and downstream (as validation) of the model. It cannot be designed out of the loop.

Two more categories round out the toolkit. **Chromatography**, liquid (LC) or gas (GC), is the separation step that comes before almost everything else: you push a mixture through a column packed with material that different molecules stick to for different lengths of time, so they exit one at a time, cleanly enough to be measured. It is the unglamorous plumbing of the lab, and it is everywhere. And a family of **cell- and population-level measurements**, high-throughput screening assays run across thousands of microplate wells at once, plate readers that quantify the light or fluorescence from each well, flow cytometry (which streams cells single-file past a laser and measures each one), mass cytometry (the same idea but reading metal-tagged markers by mass, for far more parameters per cell), and spatial biology (which measures not just what molecules are present but *where* they sit inside an intact tissue), turns biology from a bulk average into a per-cell, per-location census.

Every one of these is a physical measurement on a capital instrument. None of them is replaced by a better model. They are what feeds the model.

Bruker: Four Measurement Franchises in One Company

If the thesis is "own the measurement layer," Bruker (BRKR) is the cleanest single expression of it, because it holds not one measurement franchise but four, and they are among the least contested in the entire tool set.

First, **high-field NMR**. This is the near-monopoly described above. Bruker is the dominant maker of the highest-field research magnets in the world; when a pharma company or a national lab wants the strongest NMR available, there is effectively one Western supplier. That is not a position that erodes quarter to quarter. Second, **timsTOF mass spectrometry**. Bruker's ion-mobility architecture is the platform that has been winning single-cell and high-throughput proteomics, precisely the frontier where the demand for measured protein data at scale is exploding, and precisely the data a biology model is starved for. Third, **spatial biology**: through the acquisition of NanoString's assets, Bruker moved into the business of measuring where molecules live inside tissue, one of the fastest-growing measurement modalities. Fourth, the **MALDI Biotyper**, which uses a specialized mass-spec technique to identify microbial species in clinical microbiology labs, a razor-and-blade installed base that throws off recurring revenue and has little to do with the AI cycle at all.

The point is not that Bruker is a great AI stock. It is that a portfolio built on owning the scarce measurement primitives looks a lot like Bruker: monopoly-adjacent positions in the exact modalities, proteomics, structural work, spatial, that a data-hungry model layer most needs, and a clinical annuity underneath for ballast.

The Rest of the Instrument Oligopoly

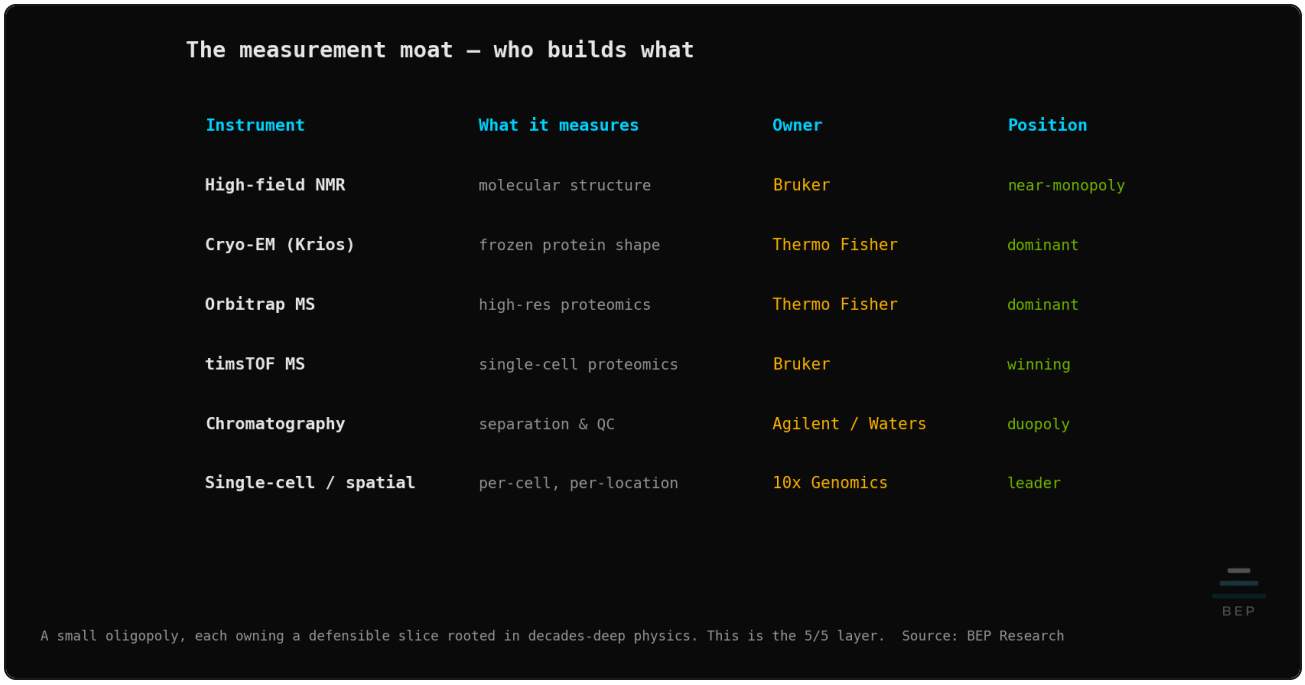
The measurement layer is not one company; it is a small oligopoly, and each member owns a defensible slice.

- **Thermo Fisher (TMO)** is the scale player and, for many labs, the default vendor. It owns the **Orbitrap** franchise in high-resolution mass spec and the **Krios** line of cryo-electron microscopes, the machines that produce the very structures AI models are measured against. Add the widest catalog of reagents, consumables, and instruments in the industry, and Thermo is the company hardest to design out of any given workflow simply because it supplies so much of it.
- **Danaher (DHR)** holds three measurement businesses under one roof: **SCIEX** in mass spectrometry, **Beckman Coulter** in flow cytometry, and **Leica** in microscopy. Danaher's

discipline is operational, it buys measurement franchises and compounds their consumable pull-through.

- **Agilent (A)** is the chromatography and LC/MS specialist, with a growing cell-analysis franchise. Its strength is the analytical workflow and quality-control layer, the measurements that have to be right before a compound advances, and again before it ships.
- **Waters (WAT)** is the ultra-performance liquid chromatography (UPLC) and tandem-quadrupole MS specialist, concentrated in the regulated, high-precision measurements that pharma manufacturing depends on. Per public reporting, Waters absorbed BD's research flow cytometry business in early 2026, extending it from separations into single-cell measurement.
- **Mettler-Toledo (MTD)** sits underneath all of it. Precision weighing, titration, and calibration are the substrate every other measurement is referenced against, if the balance is wrong, every downstream number inherits the error. It is the least glamorous and among the stickiest positions in the stack.
- **Revvity (RVTY)** supplies plate readers and high-throughput screening, the machinery of running assays by the thousand. **Bio-Rad (BIO)** holds droplet digital PCR, a precise way to count individual nucleic-acid molecules.

Two names matter as the alternative suppliers that keep the monopolies from being absolute. **JEOL** of Japan is the only real competitor to the Western leaders in both NMR and cryo-EM, the reason those franchises are near-monopolies rather than outright ones. **Carl Zeiss** (AFX) in microscopy and **Oxford Instruments** (OXIG) in the cryogenics and superconducting-magnet components that other people's instruments depend on occupy the same critical-supplier tier: not the household names, but the parts of the supply chain a competitor cannot easily route around either.



A small oligopoly, each member owning a defensible slice rooted in decades-deep physics. This is the 5/5 layer.

The Spatial and Single-Cell Frontier

The fastest-growing corner of measurement is the move from bulk to resolved, reading biology one cell and one location at a time. This is where the newest data, and the newest instruments, are being minted.

- **10x Genomics (TXG)** pioneered single-cell and spatial transcriptomics, measuring gene expression cell by cell, and increasingly mapping it onto tissue coordinates.
- **Akoya** and **Vizgen** (both private) are pure spatial-biology plays, imaging where proteins and RNA sit within intact tissue.
- **Standard BioTools (LAB)** owns CyTOF mass cytometry, the metal-tagged, mass-read approach to profiling many markers per single cell.
- **Quanterix (QTRX)** holds Simoa, an ultrasensitive immunoassay platform that can detect proteins at concentrations previously below the noise floor, the measurement of the very faint.

On the flow-cytometry side, **Becton Dickinson (BDX)** has long been the incumbent (and the seller of the research franchise Waters absorbed), while **Cytek (CTKB)** pushed the field forward with spectral flow, reading the full emission spectrum rather than a few discrete channels. And the measurement layer is not exclusively Western: China's **Mindray** (300760) is a large and rising force in clinical analyzers and flow, and **Focused Photonics** (300203)

supplies analytical instrumentation domestically, a reminder that measurement, unlike frontier software, is a physical-manufacturing capability nations invest in building for themselves.

The Data Flywheel: Why the Instrument Beats the Algorithm

Terray makes the case better than any generic example.

Terray is often described as an AI drug-discovery company, and it is one. But its durable edge is not only the chemistry models. It is also a proprietary, ultra-dense microarray that can physically measure billions of small-molecule binding interactions — running vast numbers of miniaturized binding experiments and recording, for each, whether and how well a molecule stuck to a target. That output is not a prediction. It is measured ground truth, generated at a scale no public dataset comes close to: on Terray's own figures, roughly a billion unique measurements a quarter, more than three times the entirety of public chemistry data every three months. Feed that back into a model and the model improves; run the next models to design and prioritize the following batch of physical experiments and the measurements get more informative; the better data trains a better model again. That loop — measure, learn, measure better — is the data flywheel, and it only spins because Terray owns the measuring apparatus at one end of it. I am an angel investor in Terray, so I have watched this from the inside, and it is the dynamic I described in *The 2028 Abundance Report*: ***"AI designed molecules, robots synthesized them, hardware tested them, results fed back into the model."***

Let me put that in plainer terms, because from my seat at Terray it is the thing that stays with me. I have watched that platform both make and physically measure more molecules in a month than a traditional medicinal-chemistry team would get through in a thousand years. Not model them on a screen. Make them and test them against a real target. That is the part that reorders your intuitions. The bottleneck I had always assumed was permanent, the sheer slowness of turning an idea into a measured result, turns out to be an engineering problem, and it is being engineered away in front of me. So when people ask where the real acceleration from AI will show up, my answer is not chat or code. It is here, on the bench. The next great speedup is for the sciences.

I do not say that from a distance. Jacob Berlin walked me through Terray's Los Angeles lab, and watching the nano-scale hardware run in person is what made the loop concrete: their own-developed, ultra-dense microarray hardware generating that proprietary dataset at a pace that reorders your sense of what is possible, its full-stack AI designing improved molecules in minutes while the automated testing measures the most promising ones, the whole cycle under one roof. And it is aimed at real disease, not a demo. Terray's lead internal asset is a brain-penetrant small-molecule inhibitor for multiple sclerosis, a place where better therapeutics are

desperately needed, alongside a pipeline of de novo medicines for immune-system diseases and partnered programs. That is what owning the measurement layer looks like when it works, an actual drug candidate coming out of a data loop rather than a slide about one.

Strip the example down to its principle. The AI model is the commodity in this system. Foundation models for chemistry and biology are proliferating, open-weighting, and getting cheaper by the month; a competitor can rent equivalent capability. What a competitor cannot rent is the specific, proprietary, physically-measured dataset that comes off an instrument no one else has. Whoever owns the "the hardware tested them" step owns the data that everyone else's model needs, and the more that model layer commoditizes, the more valuable the scarce measured data becomes, not less. Value in this stack flows toward the input that cannot be synthesized.

That reframing is the whole chapter in a sentence: in AI-assisted drug discovery, the moat is not the model, it is the machine that made the data. The proposing has been democratized. The measuring has not, and the physics of superconducting magnets, ion optics, and electron microscopes suggests it will not be for a long time.

Where the Thesis Could Be Wrong

The bear case deserves a fair hearing, because a thesis this clean invites overconfidence. Three risks are real. First, *in-silico substitution*: if predictive models eventually become accurate enough that fewer physical experiments are needed per approved drug, unit demand for instruments could soften even as biology advances, the models would be eating into the very measurement volume this chapter treats as sacred. The counter is that better models have historically *expanded* the experimental frontier rather than shrinking it, because they generate more hypotheses worth testing; but "historically" is not "necessarily."

Second, *cyclicality and funding*: these are capital instruments sold heavily into biopharma R&D budgets, academic grants, and the biotech funding cycle. When that funding contracts, instrument orders are among the first casualties, and several of these names carry premium multiples that assume smooth secular growth. The moat is structural; the revenue is not immune to the cycle. Third, *the flywheel is not automatic*: owning an instrument that can measure billions of interactions is necessary but not sufficient, the data has to be on targets that matter, at quality high enough to train on, and the company has to actually build the model loop rather than merely possess the hardware. Not every instrument owner converts measurement into a compounding data asset.

What would change the view is straightforward to watch for: a sustained decline in instrument unit volumes that is not explained by the funding cycle, or credible evidence that model predictions are being trusted in place of confirmatory measurement in regulated decisions.

Until one of those shows up in the data, the conclusion holds. The scarce input to the intelligence is the measurement, and the measurement comes from this layer.

The Readout Layer: Reading Biology Directly

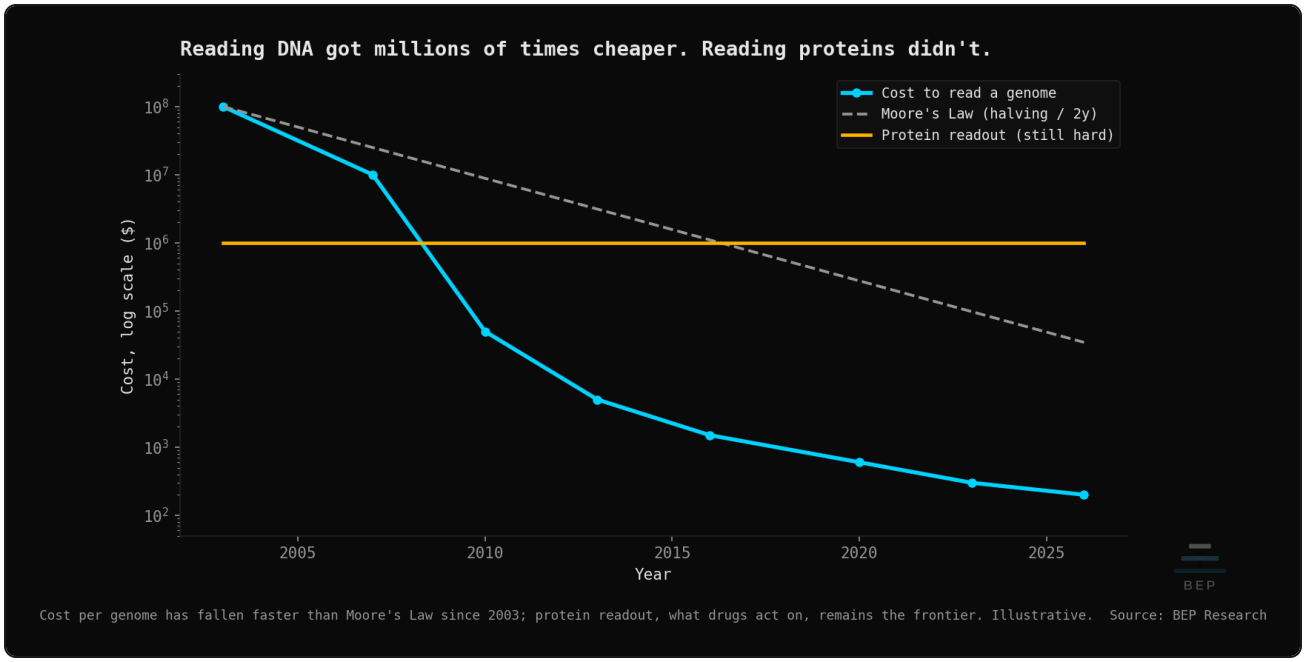
Everything downstream in drug discovery depends on one prior question: can we actually read what a cell is doing? Before a model can predict a binding pocket or a target can be validated, biology has to be converted into data. That conversion is a physical measurement problem, and the companies that own the physics of measurement occupy the most defensible position in the stack. This is the readout layer. It turns molecules into numbers.

The readout layer splits cleanly into two worlds. Reading DNA and RNA is a solved, industrialized problem where the competition is over cost and accuracy. Reading proteins is not solved, and that gap is where most of the interesting investment turbulence lives. Understand why the second world is harder than the first and you understand the entire chapter.

How DNA Sequencing Actually Works

DNA is a chain of four bases: A, C, G, T. Sequencing is the act of determining their order. The dominant method, used by **Illumina (ILMN)**, is called sequencing-by-synthesis. The chemistry is elegant. You take a strand of DNA, fragment it into short pieces, and anchor millions of those fragments to a glass flow cell. Then you rebuild the complementary strand one base at a time, but each incoming base carries a fluorescent tag and a chemical block that stops the reaction after a single addition. A camera photographs the flow cell, records which color lit up at each of the millions of spots, the block is cleaved off, and the cycle repeats. Read the colors across all cycles and you have read the sequence.

This approach is accurate and massively parallel, which is why Illumina's NovaSeq machines drove the cost of a human genome from roughly 100 million dollars in the mid-2000s toward the low hundreds today. The weakness is read length. Sequencing-by-synthesis works in short fragments, typically 100 to 300 bases, so the genome must be shattered and computationally reassembled like a shredded document. Repetitive or structurally complex regions are hard to reconstruct from short pieces.



Reading DNA collapsed in cost faster than compute; reading proteins did not, which is where the next measurement moat sits.

That weakness created the opening for long-read sequencing, which comes in two flavors:

- **PacBio (PACB)** uses a method called HiFi. It reads a single DNA molecule in a tiny well, watching a polymerase incorporate fluorescently labeled bases in real time, and it reads the same circular molecule multiple times to average out errors. The result is long reads, often 15,000 to 20,000 bases, with accuracy that rivals short reads.
- **Oxford Nanopore (ONT, London-listed)** does something physically different. It threads a single DNA strand through a protein pore embedded in a membrane and measures the tiny electrical current disruption as each base passes through. Different bases block the current differently, so the sequence is read off the electrical signal. There is no synthesis, no camera, no fluorescence. This gives ultra-long reads and true portability, down to a device the size of a USB stick, at the cost of historically higher per-base error rates that have narrowed over successive chemistries.

The competitive axes in DNA sequencing are three: read length, accuracy, and cost per genome. Illumina optimized accuracy and cost. The long-read players optimized length. And the entire field is marching toward a psychological and economic threshold, the 100-to-150-dollar genome, below which population-scale sequencing becomes routine and the addressable market expands by an order of magnitude.

Single Cell and Spatial: Reading Context, Not Just Sequence

Sequencing the genome tells you the instruction set. It does not tell you which instructions a given cell is actually running. That is the job of RNA readout, and here two refinements matter enormously.

Traditional bulk sequencing grinds up a tissue sample and reads the average gene expression across all its cells. The problem is that averaging destroys information. A tumor sample might contain cancer cells, immune cells, and healthy tissue, and the bulk average tells you nothing about which population is doing what. **Single-cell sequencing** solves this by partitioning a sample so that each cell's RNA is barcoded individually before sequencing. You get one expression profile per cell, which is how researchers discover rare cell types and map how a drug hits some cells and not others. **10x Genomics (TXG)** built the dominant droplet-based platform for this, encapsulating each cell in an oil droplet with a uniquely barcoded bead.

The frontier beyond single cell is **spatial transcriptomics**: reading which genes are expressed while preserving where in the tissue each measurement came from. Location is biology. An immune cell at the edge of a tumor behaves differently from one at its core, and the moment you dissociate the tissue into a suspension you lose that geometry. Spatial methods read expression directly on an intact tissue slice, keeping the coordinate map intact. 10x competes here too, alongside **Parse Biosciences** (a combinatorial-barcoding approach that avoids expensive instruments, now owned by **QIAGEN**). QIAGEN (QGEN) itself sits upstream of all of this in sample preparation, the unglamorous but essential step of extracting clean nucleic acid before any sequencer ever sees it.

Why Proteins Are Harder Than DNA

DNA is easy to read at scale because of one trick: amplification. The polymerase chain reaction, PCR, can take a single DNA molecule and copy it billions of times, so even a vanishingly small starting sample becomes a strong, readable signal. There is no PCR for proteins. You cannot photocopy a protein directly; the workarounds all amplify a DNA tag that stands in for the protein rather than the protein itself, which is exactly what the aptamer and proximity-assay readouts described later in this chapter do. Whatever protein is in the sample is all the protein signal you will ever get.

That matters because proteins are the actual machines of biology and the direct target of most drugs. The genome is the blueprint; the proteome is the running system. But proteins are built from twenty amino acids rather than four bases, they fold into three-dimensional shapes, and they get chemically modified after they are made. There is no natural copying enzyme, no four-letter alphabet, and a dynamic range in the blood of roughly ten orders of magnitude between the most and least abundant proteins. Reading the proteome is a genuinely unsolved

measurement problem, and the field has split into competing physical approaches, none yet dominant:

- **Aptamer-based** (the technology inside SomaLogic): short synthetic DNA strands are engineered to fold into shapes that grab specific proteins. Because the binder is itself DNA, it can be counted using DNA-reading tools. High multiplexing, thousands of proteins at once.
- **Proximity Extension Assay, PEA** (Olink's method): each target protein is tagged by a pair of antibodies, each carrying a short DNA tag; only when both antibodies bind the same protein do the tags come close enough to be joined and amplified. Requiring two independent binding events sharply cuts false positives.
- **Mass spectrometry**: proteins are broken into peptide fragments and sorted by mass and charge. It identifies proteins without needing a pre-made binder for each one, but sensitivity for low-abundance proteins has been the historical constraint.
- **Single-molecule protein sequencing**: the aspiration to read a protein amino acid by amino acid, the way we read DNA base by base. Still early, technically brutal, and the biggest prize if it works.

Reading the Field

Start with the incumbent. **Illumina (ILMN)** is the sequencing-by-synthesis oligopoly, and for a decade its NovaSeq installed base functioned like a toll road on genomics. It is now facing a genuine three-front attack: long-read platforms taking the high-accuracy end, low-cost challengers undercutting the middle, and a re-entering incumbent-class competitor at the top. Illumina's response has been to expand its readout surface, acquiring SomaLogic to move into proteomics and, per public reporting, absorbing PacBio short-read intellectual property. The strategic question for ILMN is whether an incumbent can defend a cost curve while simultaneously buying its way into the harder proteomics frontier.

The long-read specialists, **PacBio (PACB)** and Oxford Nanopore, are pure-play bets on read length and, increasingly, on real-time and point-of-care use where Nanopore's portability is unmatched. Both are smaller, single-platform names whose fortunes track adoption of long-read as the new clinical standard rather than a research luxury.

The most consequential recent development is **Roche** re-entering sequencing with its Axelios platform built on SBX (sequencing by expansion) nanopore chemistry, reportedly targeting a 150-dollar genome. Roche matters because it is the first challenger with incumbent-class scale, capital, and clinical channel to threaten Illumina directly. A credible Roche at 150 dollars reprices the entire oligopoly.

Below the incumbents sit the cost disruptors, all attacking cost per genome from different physics:

- **Ultima Genomics** (private): a wafer-based, open-substrate architecture explicitly built to drive genome cost toward the 100-dollar floor.
- **Element Biosciences** (private): a benchtop instrument using avidity chemistry, which boosts signal by having multiple binding sites grab a target simultaneously, aimed at labs priced out of NovaSeq-class capital equipment.
- **MGI Tech (688114, Shenzhen-listed)**: the sequencing arm of the BGI complex, using DNBSEQ nanoball chemistry. MGI is the price leader globally and the most disruptive force on cost, though US-listed investors should flag the geopolitical overhang, including the BIOSECURE-style scrutiny and the Department of Defense's 1260H list targeting Chinese biotech vendors.

In single-cell and spatial, **10x Genomics (TXG)** is the category-defining public name, with the same razor-and-blade instrument-plus-consumables model that made Illumina durable, though facing its own patent litigation and lower-cost entrants. Parse Biosciences inside **QIAGEN (QGEN)** and QIAGEN's own upstream sample-prep franchise round out the read-through: whoever prepares the sample sells a consumable regardless of which sequencer wins.

Proteomics is where the differentiated readout physics still lives inside single-platform public equities, and it is the most open field in the chapter. Olink (PEA) now sits inside **Thermo Fisher** and SomaLogic (aptamer) inside Illumina, so the two most mature multiplexed platforms have been absorbed by giants. What remains investable as standalone physics bets:

- **Seer (SEER)**: a nanoparticle-based front-end that tames the proteome's enormous dynamic range before mass spectrometry, effectively a sample-prep layer that makes existing mass specs see deeper.
- **Nautilus Biosciences (NAUT)**: pursuing single-molecule protein analysis, a direct swing at the hardest version of the problem.
- **Quantum-Si (QSI)**: building a semiconductor chip for single-molecule protein sequencing, betting that protein reading follows DNA in migrating from big instruments to cheap silicon.
- **Alamar Biosciences (ALMR)**: its NULISA method layers a wash step onto proximity assays to strip background, chasing the ultrasensitive, low-abundance end where the most clinically interesting proteins hide.

The Investment Shape of the Layer

Two very different regimes coexist in the readout layer, and conflating them is the common analytical error. DNA sequencing is oligopoly economics under cost-curve attack. The physics is mature, the players are known, and the fight is over dollars per genome as Roche, the private cost disruptors, and MGI compress margins beneath Illumina. That is a pricing war, and pricing wars are hard on incumbent multiples even when volumes grow.

Proteomics and spatial are the opposite: a wide-open frontier where the underlying measurement physics is not settled and where differentiated approaches still sit inside single-platform public names rather than absorbed into instrument giants. That is where a technology edge can still translate into a durable franchise, and also where the failure rate will be higher. The generalizable point for a portfolio is that the readout layer offers more competitive turbulence, and therefore more dispersion of outcomes, than the placid instrument-giant narrative suggests. The molecules that are easiest to read are already commoditizing; the molecules that matter most for drugs are the ones we still cannot read well, and that unsolved measurement problem is precisely where the asymmetric returns are hiding.

The Hands — Automation, Robotics, and the Orchestration Layer

An AI model that designs a molecule has done nothing until a physical experiment tests it. The bottleneck in modern drug discovery is not ideas; it is the rate at which ideas can be run through wet-lab reality and turned back into data. Every design-make-test-analyze loop has to close in the physical world, and the physical world is slow, error-prone, and staffed by expensive humans doing repetitive work. The "hands" of the industry, the robots that move liquid, the arms that carry plates, and the software that sequences them, are what set the clock speed of discovery. This is where AI's appetite for data meets the hard constraint of atoms.

Understanding this layer means separating three things that are usually lumped together: the instruments that do the work, the physical robotics that connect them, and the software that decides what happens when. The investable insight, as we will see, sits almost entirely in the third.

Liquid handling: the most repeated action in the lab

The single most common physical action in a life-science lab is pipetting, moving a precise, small volume of liquid from one vessel to another. A screening campaign can require millions of these transfers. Doing it by hand is slow, introduces variance (a tired technician at hour six is

not the technician at hour one), and caps throughput at human speed. Automating it is the foundational act of lab automation.

A liquid handler is a robot that aspirates and dispenses liquid across a fixed work surface, or deck, laid out with plates, tips, and reagent reservoirs. Two engineering variables dominate. The first is precision: many assays operate at microliter and sub-microliter volumes, where a small percentage error in dispensed volume corrupts the result and, downstream, the training data an AI model will learn from. Dirty data at the pipette propagates into every model built on it. The second is throughput: how many wells the instrument can process per hour, which is a function of channel count (an 8-, 96-, or 384-channel head moves far more liquid per cycle than a single tip) and deck capacity.

The market splits into two philosophies. Deck-based workstations, large, enclosed, high-precision systems, are the workhorses of pharma and contract research. Against them sits a low-cost, open-system approach that trades some precision and speed for affordability, scriptability, and access, opening automation to academic labs and startups that could never justify a six-figure workstation.

The public-market reality is stark. The clean pure-play is **Tecan (TECN.SW)**, the Swiss maker of premium deck-based liquid handlers and detection systems, effectively the only way to own this instrument category directly on a listed exchange. Almost everything else is private or buried. **Hamilton** (private) is the other premium name serious labs standardize on. **Opentrons** (private) is the low-cost, open-source disruptor. **Eppendorf** (private) and **SPT Labtech** (private, owned by EQT) round out the specialist field. The rest is inside conglomerates: **Beckman Coulter**'s liquid handlers sit within **Danaher (DHR)**, and the Bravo platform sits within **Agilent (A)**. Owning the pipette, in other words, mostly means owning a diversified life-science tools giant and accepting the dilution.

The workcell: turning instruments into a line

A liquid handler alone automates one step. Real discovery workflows chain many steps: dispense reagents, seal a plate, incubate it for hours at controlled temperature, spin it in a centrifuge, read it on a plate reader, then move the results downstream. A workcell is the integration of these separate instruments into one continuous automated line, with a robotic arm, floor-standing or, increasingly, a collaborative arm, as the courier that carries plates between stations.

The engineering challenge here is not any single instrument; each is a mature product. It is the interfaces between them. Every instrument speaks its own control protocol, expects plates presented at a specific position and orientation, and runs on its own timing. Integrating a dozen of them into a system that runs unattended overnight, recovers from a jammed plate without a

human, and logs everything for traceability is genuinely hard systems engineering. This integration work is why workcells have historically been bespoke, expensive, and slow to deploy, closer to a construction project than a purchase.

The orchestration layer: the operating system of the lab

" The robotic arm is not the brain. The brain is the orchestration and scheduling software, the layer that sequences every instrument, allocates shared resources, resolves timing conflicts, and reroutes work when something fails. If two experiments both need the plate reader at 2:14 a.m., the scheduler decides who waits and reshuffles the entire downstream plan so neither assay's incubation window is violated. This is a constraint-solving problem running continuously against a live physical system.

This software is the real lock-in, for three reasons. First, switching cost: once a lab's protocols, instrument drivers, and validated workflows are encoded in one scheduler, moving to another means re-integrating and re-validating everything, in a regulated environment where re-validation is expensive and risky. Second, integration equity: the value the customer paid for is not the arm but the months of driver-writing and workflow-tuning that made a specific set of instruments run together, and that lives in the software. Third, margin: hardware is a capital good that commoditizes; orchestration software carries recurring, high-margin economics and deepens with every protocol the customer adds.

The physical arm, by contrast, is commoditizing. General-purpose collaborative robots, cobots, are increasingly good enough to move plates, and they are cheaper and more flexible than the proprietary arms that used to define a workcell. As the arm becomes a generic component, value migrates decisively up to the scheduling layer that tells it what to do. This is the central investment insight of the chapter: do not pay for the muscle; pay for the operating system.

The pure orchestration players are almost entirely private. **Automata** (private) markets its software explicitly as the "OS for AI-ready labs." **HighRes Biosolutions** (private) is known for its Cellario scheduler. **Peak Analysis & Automation** (private) and **Hudson Robotics** (private) are established integrators. The most investable path runs through the strategics: **Biosero**, a leading orchestration provider, sits inside **Bio-Techne (TECH)**, and per public reporting, Merck KGaA is in the process of acquiring Bio-Techne, which would fold that orchestration asset into a larger strategic. The pattern to watch is consolidators buying the scheduling software, because that is where the recurring economics and the switching cost concentrate.

General-purpose arms and mobile, autonomous chemistry

The commoditization of the arm has a name and a market leader. **Universal Robots**, inside **Teradyne (TER)**, makes the default cobot deployed across manufacturing and, increasingly, the lab, a safe, reprogrammable arm that any integrator can drop into a workcell. **KUKA**, now controlled by China's Midea, is the other major industrial-arm name, and its ownership is a supply-chain consideration for Western pharma buyers wary of Chinese-controlled infrastructure in a sensitive workflow.

Beyond the fixed workcell, the next step is mobility and autonomy. Mobile robots that navigate between stations remove the constraint that everything must sit within one arm's reach, letting a lab scale by adding stations rather than rebuilding a cell. Autonomous chemistry pushes further: closed-loop systems where an AI proposes the next experiment, the hardware runs it, and the result feeds back into the model with no human in the loop. **Multiply Labs** (private) applies robotics to cell-therapy manufacturing, building its systems on NVIDIA's Isaac robotics and Omniverse simulation stack, a direct bridge between physical AI and biomanufacturing. **Kebotix** (private) and **Chemify** (private) are building autonomous, self-driving chemistry platforms that couple generative models to robotic synthesis.

Automation is also reaching the parts of the pipeline usually treated as purely manual craft. **VistaPath Bio** (private) automates pathology specimen grossing, the hands-on dissection and sampling of tissue that has resisted automation and gates the throughput of the diagnostic lab. It is a reminder that "the hands" extend well past the screening deck into diagnostics and manufacturing.

Proof that the loop can close

The strongest evidence that this stack works end-to-end comes from Ginkgo Bioworks' cloud lab, where an AI agent was set loose to run tens of thousands of experiments autonomously across DNA-focused workflows. That is the full loop operating at scale: a model proposing work, robotics executing it, and data returning to the model without a human sequencing each step. It demonstrates that the orchestration layer, not the arm, is what makes autonomy real, the scheduler is what lets one agent command a physical lab.

How to own the layer

For an investor, the hands are a frustrating category to access, and that frustration is itself the signal. The value sits in the orchestration software, and that software is overwhelmingly private (Automata, HighRes, Biosero pre-acquisition) or buried inside diversified strategics (Beckman in DHR, Bravo in A, Biosero moving into a Merck KGaA-owned Bio-Techne). Tecan (TECN.SW) is the only clean listed pure-play, and it sits in hardware, the commoditizing end of the stack.

The general-purpose arm is best owned through Teradyne (TER) via Universal Robots, but that is a bet on cobots broadly, not on drug discovery specifically.

The disciplined read is to track the orchestration layer as the location of durable lock-in and margin, and to gain exposure primarily through the strategics acquiring it, because when a consolidator buys a scheduler, it is buying the operating system of the automated lab, and everything downstream of that decision runs on their software. The muscle is for sale everywhere. The nervous system is what is scarce.

The Reagents Layer — The Recurring-Revenue Floor

Every layer we have covered so far sells the same promise: better molecules, chosen faster. The reagents layer sells something duller and more durable. It sells the physical inputs a scientist consumes to run a single experiment, and it gets paid whether that experiment works or not. This is the least glamorous floor in the drug-discovery building, and for a certain kind of investor it is the most interesting one, precisely because its revenue does not depend on any drug ever reaching a patient.

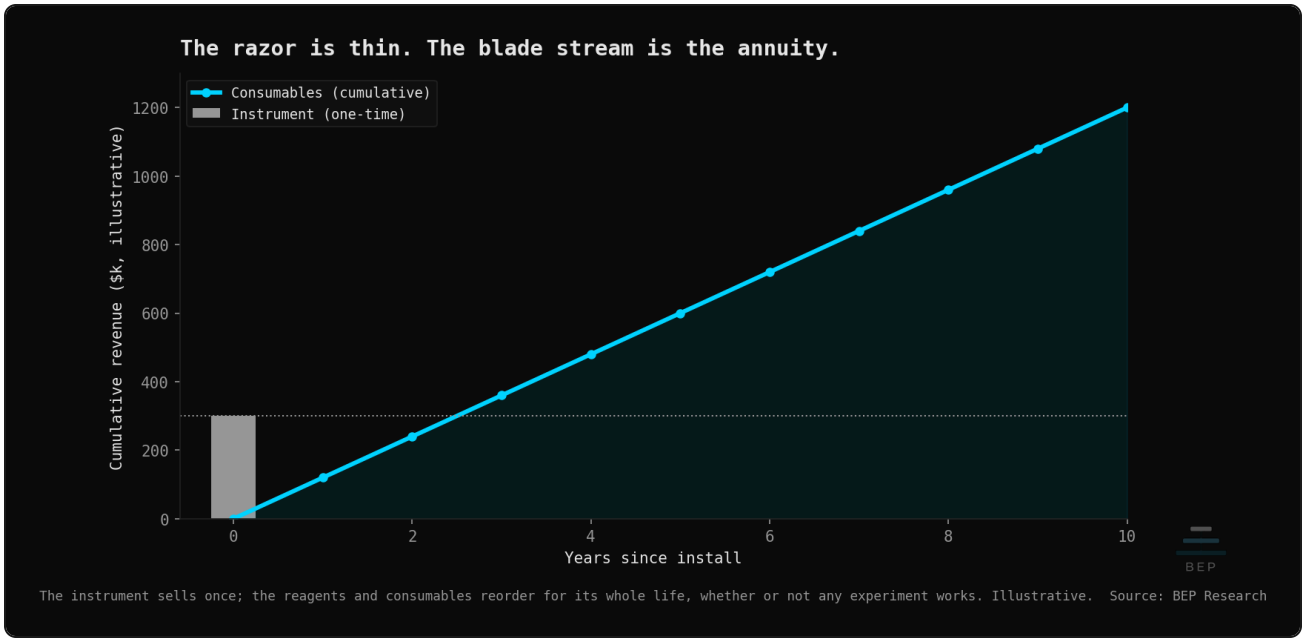
Start with what a reagent actually is. When a biologist wants to know whether a particular protein is present in a cell, or how much of it there is, they cannot see the protein directly. They use an **antibody**, a molecule engineered to bind one specific target, tagged with something detectable, such as a fluorescent dye or an enzyme that produces color. Add the antibody, wash away what does not stick, and measure the signal. That is the core of most protein detection in a research lab, and every run consumes antibody that has to be repurchased. Assay kits package this logic into a ready-made bundle of reagents for a specific measurement. Recombinant proteins, proteins manufactured in engineered cells rather than extracted from tissue, are the standardized inputs and controls that make those measurements comparable across labs.

The list of consumables runs deeper than most generalists assume. Cells do not grow in nothing; they grow in **cell culture media**, a carefully formulated liquid of nutrients, salts, and growth factors, often supplemented with animal **sera**. Media is consumed continuously, a running cell line is a running media bill. When a biologic drug (an antibody-based therapeutic, for instance) has to be purified out of the messy soup it was grown in, that is done with **chromatography resins**: packed beads that selectively grab the target molecule as the mixture flows through. The workhorse here is the **Protein A ligand**, a molecule that binds the constant region of antibodies with high specificity, making it the default first step in nearly every antibody purification process. **Filtration** membranes clarify and sterilize the streams around it.

Then there is the plumbing of modern biomanufacturing. The industry has shifted heavily toward **single-use, or disposable, bioprocess consumables**, sterile plastic bioreactor bags, tubing, and connectors that are used once and discarded rather than cleaned and revalidated. Single-use lowers contamination risk and shortens changeover between batches, and it converts what used to be a capital-and-cleaning problem into a stream of per-batch purchases. At the highest frequency of all sits ordinary **plasticware**: pipette tips, microplates, tubes. A single automated screening run can burn through thousands of tips and dozens of plates. Nobody writes a research paper about the tip, but the lab cannot function for an hour without a full box.

The Economics: A Toll on Activity, Not on Success

The business model here is the oldest one in industrial supply: razor-and-blade. The instruments are the razors, bought once and depreciated over years. The reagents and consumables are the blades, bought again every week for the life of the instrument. Vendors will often price the razor thin to seat it in the lab, because the installed base is really a subscription to the blade stream. A sequencer, a bioreactor, a liquid handler, each one is a decade-long annuity of tips, media, kits, and resins keyed to that vendor's ecosystem.



The instrument sells once; the reagents and consumables reorder for its whole life, whether or not any experiment works.

A handful of vendors dominate the everyday reagent and consumable order:

- **Thermo Fisher (TMO)**, the broadest catalog in the business. Between its own brands and the Fisher Scientific distribution channel, it can supply most of a lab's recurring order from

one account, which makes it structurally sticky and hard to displace item by item.

- **Danaher (DHR)**, the bioprocess and antibody anchor. Through Cytiva it sells chromatography resins and single-use systems; through Pall it sells filtration; through Abcam it sells research antibodies. Danaher's exposure is weighted toward the biologics manufacturing chain, which ties it more tightly to the drug-production side of the toll than to the pure academic bench.
- **Bio-Techne (TECH)**, the specialty protein leader. Its R&D Systems brand is the reference source for recombinant proteins and cytokines, the signaling molecules used to stimulate and study cells. This is a narrower, higher-margin position: not the broadest catalog, but the trusted default in categories where scientists will not switch to save a few dollars.
- **Merck KGaA / MilliporeSigma (MKKGY, MRK.DE)**, the German Merck, which is entirely distinct from the U.S. Merck (MRK) despite the shared name; confusing the two is a common error. Its life-science arm combines Sigma-Aldrich, the dominant catalog of research chemicals and biochemicals, with Millipore filtration. Between chemicals and membranes, it touches an enormous share of the everyday reagent order.
- **Avantor (AVTR)**, reagents plus reach. Avantor pairs its own products with the VWR distribution channel, which functions as a purchasing backbone for thousands of labs. The investment character here leans more toward channel and logistics than proprietary formulation, which makes it more distribution-cyclical than the pure product houses.
- **Sartorius (SRT.DE)**, the cleanest single-use razor. Sartorius is heavily concentrated in disposable bioprocess consumables, bags, filters, and the fittings around them, which makes it close to a pure-play on the recurring, per-batch blade stream in biomanufacturing.
- **Repligen (RGEN)**, small in revenue, large in position. Repligen's Protein A affinity ligands are embedded inside the purification steps of most commercial antibody processes, frequently as an ingredient sold into the larger vendors' resins. It is a picks-and-shovels supplier to the picks-and-shovels suppliers, which is an unusually defensible spot.
- **Corning (GLW)**, the vessels biology grows in. Beyond its display and optical businesses, Corning is a major maker of labware and cell-culture surfaces: the treated flasks, plates, and dishes whose surface chemistry lets cells attach and grow. It is the physical substrate of the experiment.
- **Eppendorf (private)**, not investable in public markets, but named because it defines a category. Eppendorf's tips, tubes, and plates are so standard that the eponymous tube is lab shorthand. It is a reminder that some of the deepest moats in this layer sit outside the public tape.

The Investment Angle, and the Honest Risk

The way to hold this layer is to be clear-eyed about what the moat is and is not. It is not, for the most part, a differentiation moat in the sense of irreplaceable science. Most individual reagents have a substitute. The moat is aggregation and habit: the breadth of the catalog, the validation history a scientist does not want to redo, the qualified supplier a manufacturer cannot swap without regulatory paperwork, and the installed instrument that dictates which blades fit. Those advantages have concentrated the economics into Thermo and Danaher at the top, with Sartorius and Repligen anchoring the bioprocess end. For an investor who believes the number of experiments only goes up, this is the layer that participates in that growth without underwriting any single molecule.

The risk is equally plain, and it is not existential, it is cyclical. Reagent demand tracks R&D budgets, and those budgets breathe with the funding environment. Biopharma spending contracts when capital tightens; academic purchasing bends with government grant cycles; and the pandemic era taught the whole sector a hard lesson about inventory, when customers who over-ordered during the boom spent the following years working down stockpiles and starving new orders. A tollbooth still collects less when the road is quiet. The reagents layer is a durable claim on scientific activity, but activity itself is fundable and therefore cyclical. The variable to watch is funding, not disruption: in a lean year the road goes quiet and the tollbooth collects less, even though the experiments themselves are never really in doubt.

The Learn Layer — Data, Software, and the Scarcest Asset in the Lab

Every prior layer produces one thing above all else: data. The instruments measure, the sequencers read, the assays report, and every one of those events throws off a number that has to be captured, stored, made sense of, and fed back into the next design. The Learn layer is what closes the Design-Make-Test-Learn loop. It is the software and data infrastructure that turns a pile of raw instrument output into structured, reusable signal a model can actually train on. It is also, for reasons that take a moment to explain, the layer where an ordinary-sounding problem, "just store the results", hides one of the most defensible moats in the entire stack.

The reason is simple to state and hard to solve. An AI model is only as good as the data it learns from, and biological data is uniquely hostile to being learned from. Fixing that hostility is the business of this layer.

The Notebook and the Database: ELN and LIMS

Start with the two pieces of software every organized lab runs on. An **Electronic Lab Notebook (ELN)** is the digital replacement for the paper notebook a scientist used to scribble in: it records what experiment was run, why, by whom, with what materials, and what happened. It is the record of intent and observation. A **Laboratory Information Management System (LIMS)** is the operational database underneath: it tracks samples, reagents, instruments, and workflows, where every physical thing is, what state it's in, and what has been done to it. Loosely, the ELN captures the science and the LIMS captures the logistics, though modern platforms increasingly blur the line.

These sound like back-office tools, and for decades they were treated that way, necessary plumbing, bought once, rarely thought about. What changed is that the data these systems hold went from being a compliance archive to being the raw material of a model. The moment the goal became "train an AI on our accumulated experiments," the quality, structure, and connectedness of what the ELN and LIMS captured stopped being a filing question and became the whole ballgame.

The Real Problem: Instruments Don't Speak the Same Language

A single lab might run a mass spectrometer from one vendor, a sequencer from another, a plate reader from a third, and a dozen other instruments besides. Each one emits its results in its own format: a proprietary binary file, a bespoke spreadsheet layout, a vendor-specific export that means something only to that vendor's software. There are hundreds of these formats, none of them designed to talk to the others, many of them undocumented.

To an AI model, this is chaos. You cannot train on a folder of incompatible files. Before any learning can happen, all of that output has to be **harmonized**, pulled out of its native formats, translated into a common structure, aligned so that "concentration" means the same thing whether it came from instrument A or instrument B, and linked back to the experiment that produced it. This is unglamorous, enormously labor-intensive work, and it is the single biggest reason most pharma data sits unusable in silos. Solving it is a genuine technical and commercial franchise.

Two more ideas make the harmonized data actually useful. The first is the **ontology**: a controlled, agreed-upon vocabulary so that a gene, a cell type, an assay, or a unit is named consistently everywhere, rather than a dozen labs each inventing their own shorthand. The second is the set of principles known as **FAIR data**, Findable, Accessible, Interoperable, Reusable, a checklist for whether data can actually be located and reused later rather than being write-once and forget. Underpinning both is **provenance**, or lineage: the unbroken record of

where a data point came from, what was done to it, and by whom. A model trained on data of unknown provenance is a model you cannot trust or defend, and in this field trust is regulatory.

Why a Generic Data Warehouse Isn't Enough

An obvious objection: the rest of the economy solved big-data storage years ago with cloud data lakes and warehouses. Why can't a lab just point one of those at its instruments and be done? The answer is that a generic data platform supplies raw storage and compute but knows nothing about biology. It does not understand that a sequence and its annotations belong together, that a plate has a spatial layout that matters, that a sample has a lineage of parents and children, or that certain modifications must be logged to satisfy a regulator. The horizontal tools give you a warehouse; they do not give you the biology-aware schema, the instrument connectors, or the domain ontologies that make the data trainable.

And then there is the regulation. Drug development runs under legal frameworks with unforgiving requirements: **GxP** (the family of "good practice" standards for lab, clinical, and manufacturing work) and, in the United States, **21 CFR Part 11**, which governs how electronic records and electronic signatures must be handled to be legally valid. In practice this means audit trails, access controls, validated systems, and the ability to prove that a record was not altered. You cannot drop a generic consumer data tool into a regulated lab and satisfy an FDA inspector. The compliance layer is part of the product, not an afterthought, and building it is a barrier to entry all by itself.

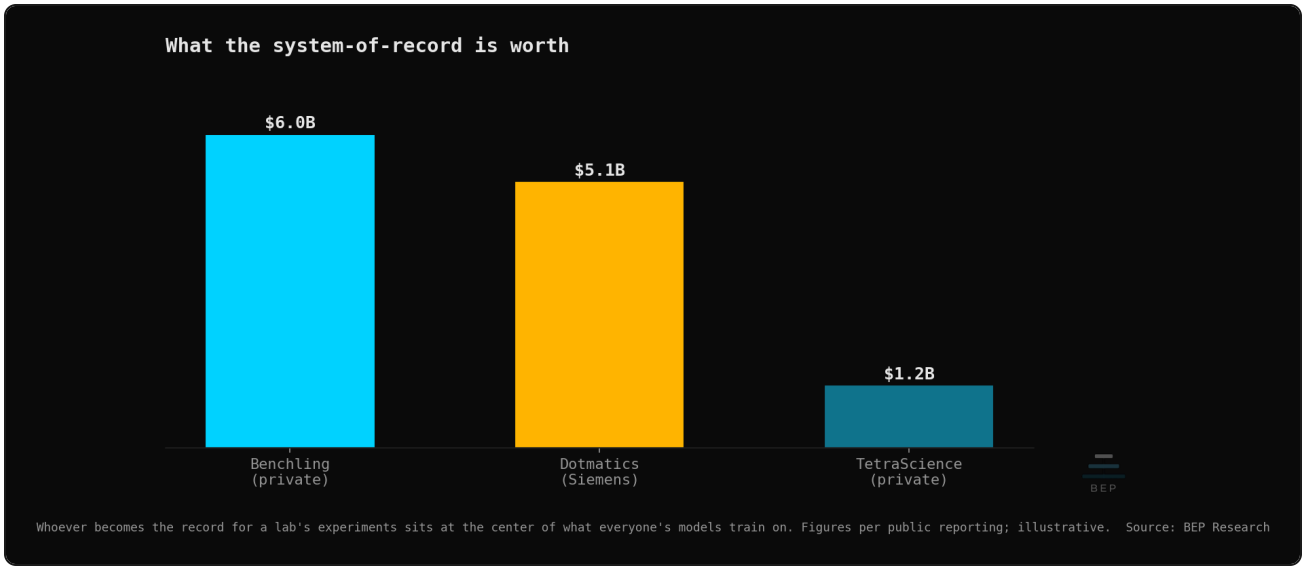
Put the pieces together and the moat comes into focus. Whoever becomes the **system of record** for a lab's clean, harmonized, compliant, machine-readable experimental data controls the substrate every downstream model must train on. Switching that system of record means re-mapping every instrument, re-validating every workflow, and migrating years of provenance, a cost so high that incumbency, once established, is extremely durable. The data is stickier than the software, and the software is stickier than the hardware.

Orchestration: The Software That Actually Runs the Loop

One category in this layer does more than record; it acts. **Lab orchestration software** is the control system that schedules instruments, sequences the steps of a workflow, and, in the most advanced form, closes the Design-Make-Test-Learn loop automatically, deciding what to run next based on what the last run returned. This is the connective tissue between the "hands" (the robots) and the "brain" (the models). It is also where design-of-experiments logic lives: the statistical machinery for choosing which experiments to run so that each one is maximally informative. Orchestration is the software that turns a room full of instruments into a self-driving lab.

Who Holds the Data Layer

The uncomfortable truth for a public-market investor is that the best assets in this layer are private, and the market clearly knows how valuable they are, because the acquisitions are large.



The loudest signal in the layer: strategics valuing lab R&D data infrastructure like a crown jewel, not a utility.

- **Benchling** (private) is the category-defining, biology-native cloud platform, an ELN, registry, and data system built from the ground up for life science rather than retrofitted from generic tools. It is the software of record at a very large share of biotechs, reportedly valued around a $6 billion mark. Its moat is exactly the switching cost described above: once a company's science lives in Benchling, leaving is close to unthinkable. It is the clearest example of the "system of record" thesis, and it cannot be bought on any exchange.
- **Dotmatics** was acquired by **Siemens** for roughly $5.1 billion (per public reporting), an industrial software giant paying up to own a multimodal R&D data platform. That price tag is the single loudest signal in the layer: a strategic acquirer valuing lab R&D data infrastructure like a crown jewel, not a utility.
- **Revvity Signals**, inside **Revvity (RVTY)**, is one of the few ways to own an ELN-and-analytics franchise on a public exchange, pairing a notebook with the Spotfire analytics layer. **STARLIMS**, inside **Dassault Systèmes (DSY)**, gives the French engineering-software giant a regulated-lab LIMS position. Between them they are the readable public proxies for the informatics layer, though each is a slice of a much larger company.
- **LabWare** and **Sapio Sciences** (both private) round out the LIMS/ELN incumbents that pharma actually runs.

- On the harmonization frontier, **TetraScience** (private) is the flagship independent, its whole business is pulling instrument data out of its hundreds of native formats and turning it into an AI-ready "scientific data cloud." **Ganymede** (a "Lab-as-Code" pipeline company, acquired by Apprentice.io) and **Code Ocean** (reproducible computational pipelines) attack adjacent pieces of the same problem.
- **Databricks** (private) and **Snowflake (SNOW)** are the horizontal data platforms many of these biology-native tools are built on top of, the generic substrate that supplies storage and compute while the vertical players supply the biology.
- In orchestration, **Artificial** (private) builds whole-lab orchestration that connects wet-and-dry instruments to AI agents; **Synthace** (private) encodes design-of-experiments into executable protocols; **Retisoft** (private) does the scheduling. And **NVIDIA (NVDA)** reappears here too, as the model-serving layer (BioNeMo) that runs the models the data is fed into, a reminder that the same toll sits under this layer as under Design.

What It Means for a Portfolio

This is probably the most under-appreciated layer in the entire primer, and the frustration for an investor is that under-appreciation and inaccessibility travel together. The strategic prize, the harmonized, compliant, system-of-record dataset in the middle of the loop, is exactly what incumbents are paying billions to control, and yet the purest expressions of it (Benchling, TetraScience, Artificial) are private. Public exposure is real but diluted: Revvity Signals inside RVTY, STARLIMS inside Dassault, and the model-serving toll inside NVIDIA. The pattern to watch is the same one that produced the Dotmatics deal: strategics acquiring the data layer, because whoever owns the clean data at the center of the loop owns the substrate every model in every other layer has to learn from. In a field where the model is the commodity and the measurement is the moat, the organized memory of what has already been measured is the quiet compounding asset underneath both.

The Cloud Lab — Running the Whole Loop as a Service

Everything so far has described the pieces of a modern lab: the models that design, the machines that build, the instruments that measure, the software that remembers. This chapter is about the companies that assemble all of it and rent the finished loop to someone else. If the rest of the primer is a tour of components, this is the tour of the general contractors, the organizations you can hire to run the Design-Make-Test-Learn cycle without owning a single instrument yourself. It is also the layer where the gap between the marketing and the reality is widest, so it demands a skeptical read.

What a Cloud Lab Actually Is

A **cloud lab** is a physical laboratory that a scientist operates entirely through software, from anywhere, without ever entering the building. You do not ship yourself to the bench; you ship instructions to the robots. A researcher writes an experiment in code, specifying the reagents, volumes, temperatures, instruments, and readouts, submits it, and a remote automated facility physically executes it, then returns the data. The bench becomes an API.

The analogy that gets used, correctly, is cloud computing. Before Amazon Web Services, running software at scale meant buying and racking your own servers. AWS turned that capital expense and operational headache into a metered service you call over the internet. A cloud lab proposes the same trade for wet-lab science: instead of buying instruments, hiring technicians, and maintaining a facility, you rent physical experimentation by the run. The payoff is not only convenience. It is reproducibility, because a protocol written in code executes identically every time, eliminating the silent variation between technicians and days that quietly corrupts so much biological data, and it is throughput, because the facility runs continuously and in parallel in ways a human team cannot.

The Self-Driving Lab, and an Honest Distinction

A cloud lab executes a human's instructions remotely. A **self-driving** or **autonomous lab** goes one crucial step further: the AI, not the human, decides what experiment to run next. The machine forms a hypothesis, designs the experiment to test it, runs it, reads the result, updates its model, and chooses the following experiment, the full Design-Make-Test-Learn loop with the human removed from the inner cycle. This is the endpoint the whole field is pointed at, the lab that improves itself.

The overwhelming majority of what is marketed today as an "autonomous" or "self-driving" lab is really the cloud-lab capability above, remote execution plus scheduling, with a human still choosing the experiments. Genuine closed-loop autonomy, where an algorithm meaningfully drives discovery, exists and is real, but it lives mostly in **materials chemistry** (where the experiments are more standardized and the readouts cleaner) and inside a handful of well-funded private labs. In biology, the messiness of the systems, the difficulty of the measurements, and the regulatory stakes have kept true autonomy closer to the frontier than the product catalog. When you evaluate a company in this layer, the first question is always the honest one: is the AI actually choosing the experiments, or is it scheduling a human's?

The Old Way to Rent the Loop: The CRO

Long before "cloud lab" existed, there was a mature, enormous industry for renting drug-discovery work: the **Contract Research Organization (CRO)**. A CRO is a company that

runs research, testing, or clinical trials on behalf of a drug developer. Pharma has outsourced huge swaths of its pipeline to CROs for decades, everything from early safety testing to running the Phase 3 trials that decide a drug's fate. The CRO industry is the incumbent, at-scale way to rent the loop, and it is where most of the *public-market* money in this layer actually sits.

Automation and AI press on the CRO model from both directions. On one hand, they threaten to compress the labor-intensive, high-margin services CROs sell, if a cloud lab or an AI can do in software what a CRO charged for in people, some CRO revenue is at risk. On the other, the largest CROs sit on something increasingly precious: vast, structured datasets from thousands of past studies and trials, exactly the kind of proprietary data that trains models. The strategic question for each CRO is whether AI erodes its services faster than its data becomes a moat.

Who Rents You the Loop

The cloud labs are mostly private, which tells you the model is still maturing toward public scale:

- **Emerald Cloud Lab** (private) is the fullest realization of the concept, a facility offering well over a hundred distinct experiment types, all driven by code, effectively an "AWS for the bench." It is the reference implementation of software-defined wet-lab science.
- **Strateos** and **Arctoris** (both private) offer robotic, remotely-operated discovery platforms. **Culture Biosciences** (private) specializes the model to bioprocessing, renting cloud-connected bioreactors so teams can run and monitor fermentation experiments remotely.
- **Ginkgo Bioworks (DNA)** is the one public pure-play worth naming, and its cloud-lab effort is the layer's headline proof point: an AI agent set loose to run tens of thousands of experiments autonomously across its automated foundry. Ginkgo's own troubled history as a public company is a caution about platform economics, but the demonstration itself, one agent commanding a physical lab at scale, is the clearest evidence the full loop can close.

The autonomous labs are largely private or academic, and cluster, tellingly, in materials science:

- **Lila Sciences** (private, roughly $550 million raised, backed by Flagship Pioneering) is the best-funded commercial bet on "AI science factories" that run the scientific method autonomously across chemistry, materials, and biology.
- The **A-Lab** at Lawrence Berkeley National Laboratory is the canonical academic proof point, an autonomous facility that planned, synthesized, and characterized new inorganic materials with minimal human intervention. The **Acceleration Consortium** at the University of Toronto is the flagship academic program organizing the field. **Radical AI**

(private) is commercializing self-driving materials discovery. Note the pattern: the genuine autonomy is in materials, not yet in drugs.

The tech-enabled CROs are where public investors actually get exposure to "renting the loop":

- **Charles River Laboratories (CRL)** is the preclinical and safety-testing backbone of the industry, the physical wet-lab work of turning a candidate into something safe enough to put in a human. It is the most direct public exposure to the early-discovery half of the loop.
- **ICON (ICLR)** and **IQVIA (IQV)** dominate the clinical half, running trials, with IQVIA layering an enormous health-data-and-analytics business on top that is itself an AI asset. **Medpace (MEDP)** is the high-margin, small-biotech-focused clinical CRO. **Certara (CERT)**, met earlier in the Design layer, belongs here too as the biosimulation CRO whose models can substitute for physical work in a regulatory filing.

China is the unavoidable complication. **WuXi AppTec (2359.HK)** and **WuXi Biologics** are among the largest and most capable CRO/CDMO platforms in the world, deeply embedded in Western pharma supply chains. They now carry a specific regulatory overhang: the proposed U.S. BIOSECURE Act and the Department of Defense's 1260H list, which target certain Chinese biotech vendors. This cuts both ways and investors should hold both edges. A durable decoupling is a tailwind for Western CROs that absorb displaced work; a policy reversal or a US-China thaw unwinds part of that reallocation. The regulatory discount on the WuXi complex is not a simple bear or bull, it is a live, bidirectional catalyst.

How to Own It

The disciplined conclusion is that the most exciting version of this layer, the genuinely autonomous, full-loop-as-a-service lab, is mostly private, mostly early, and mostly proven in materials rather than medicine. Public exposure runs overwhelmingly through the tech-enabled CROs and, as the single pure-play, Ginkgo. That means the layer should be sized for what it is today, not what the pitch decks promise: a real and growing market for renting experimentation, wrapped in an "autonomous" narrative that is still, outside a few labs, more aspiration than product. The right posture is to own the incumbents whose data and scale compound regardless of how fast autonomy arrives, and to watch for the one signal that would change everything, a commercial cloud lab demonstrating genuine closed-loop, AI-chosen *drug* discovery, not just materials. Until that shows up, treat "self-driving lab" as a direction, not a destination.

The Discoverers — The Tenants Who Run the Loop to Find Drugs

At the top of the stack sit the companies everyone actually reads about: the AI-native drug hunters. They are the tenants of the machine this primer has described, assembling the models, the synthesis, the measurement, and the data into a single engine pointed at one goal, finding medicines faster and cheaper than the industry's dismal historical odds. They are also, for an investor, the most seductive and the most dangerous names in the entire field, because they wrap the binary risk of biotech inside the margins and narrative of software. Understanding why requires understanding how they actually make money, which is where we start.

Two Business Models Wearing the Same Clothes

Almost every company in this layer is one of two things, or an uneasy blend of both, and the distinction determines everything about how it should be valued.

The first archetype is the **platform** company. It builds a discovery engine and monetizes the engine itself, licensing the software, running discovery for many partners, collecting fees and milestones across a portfolio of programs it does not fully own. Revenue is recurring and diversified, closer to software or tooling than to biotech. A platform can be judged on adoption, partnerships, and the breadth of its pipeline, and it does not live or die on any single molecule.

The second archetype is the **asset** company. It uses its AI to build its own proprietary drug pipeline and captures value the old-fashioned way: by owning drugs that, if approved, are worth billions, and if they fail, are worth nothing. This is biotech economics, binary, lumpy, and brutal. The AI is a means; the drug is the product.

The trouble is that most companies in this layer present as platforms while being valued, ultimately, on their assets. A brilliant discovery engine still has to produce a molecule that survives Phase 2, and as the earlier chapter on the drug pipeline made clear, roughly nine in ten candidates that reach the clinic fail, most of them in Phase 2, for lack of efficacy, after most of the money is spent. No model yet built has repealed that statistic. An AI-designed drug that binds its target perfectly still fails if the target turns out not to drive the disease. This is the hard truth that anchors the whole layer: a platform can be genuinely excellent and its lead asset can still die in the clinic, and the market will punish the stock as if the platform never worked.

The Data Flywheel, One More Time

The most durable edge a discoverer can have is not its model, models diffuse, as the Design chapter argued, but its proprietary data. A company that owns a way to generate experimental data no one else has, at a scale no one else can match, feeds a model that competitors cannot

replicate no matter how good their algorithms are. The measurement produces the data, the data trains the model, the model prioritizes the next measurements, and the loop compounds. This is why the strongest discoverers are the ones that own a piece of the measurement layer, not just the design layer. The company that owns the "hardware tested them" step owns the flywheel.

Deal Economics: How Partnerships Fund and Validate

Before a discoverer has an approved drug, which may be never, it survives on **partnerships** with large pharma. The structure is standard: an **upfront** payment when the deal is signed, a series of **milestone** payments as a program hits development and regulatory markers, and **royalties** on eventual sales. These deals do two things at once. They fund the discoverer with non-dilutive cash, and they serve as third-party validation, when a Lilly or a Novartis writes a large check and stakes a program on a startup's engine, it is a credible signal that the engine is real. The size and quality of a discoverer's partnerships are often a better read on its platform than any self-reported metric.

The Roster, Public and Private

The public pure-plays are where investors can actually participate, and they span the platform-to-asset spectrum:

- **Schrödinger (SDGR)**, met in the Design layer, is the most platform-like: a physics-and-ML software business with an owned pipeline bolted on as upside.
- **Recursion (RXRX)** is the highest-profile data-flywheel bet, an industrialized wet lab generating vast phenomics datasets (images of how cells respond to compounds) to train its models, strengthened by its merger with Exscientia and backed by NVIDIA. It is the purest listed expression of "platform, not a drug," and it trades with the volatility that implies.
- **AbCellera (ABCL)** is an antibody-discovery engine with a royalty-bearing model, optionality across many partnered programs rather than a bet on one asset, which makes it structurally more platform than asset. **Absci (ABSI)** uses generative AI to design antibodies and validate them in its own wet lab.
- **Tempus AI (TEM)** is a data-layer play more than a drug hunter, a multimodal clinical-and-molecular dataset that trains diagnostic and discovery models, monetized across many customers. In July 2026 it agreed to acquire **Personalis (PSNL)** for a $1.5B enterprise value, buying the ultrasensitive NeXT Personal MRD test to own the cancer-monitoring measurement outright, a data-layer name paying up for the instrument that generates the data, which is this primer's thesis compressed into one deal.

- Two 2026 IPOs extended the public set: **Eikon Therapeutics (EIKN)**, built on single-molecule imaging to watch drugs act on proteins in living cells (founded by a Nobel laureate), and **Generate:Biomedicines (GENB)**, whose Chroma generative model designs proteins across modalities.

The private leaders are, frustratingly, where the frontier and the best science mostly live:

- **Isomorphic Labs** is the flagship, the DeepMind spinout built on the AlphaFold lineage, which raised roughly $2.1 billion and signed target-rich partnerships with Eli Lilly, Novartis, and Johnson & Johnson. It is the clearest "AI biolab" tenant in existence, and it cannot be bought publicly. It also just got materially more serious: when DeepMind broke up the AlphaFold team in July 2026, part of that talent went to Isomorphic, and on August 5 Demis Hassabis stepped back from running Google DeepMind day to day, taking its chair and Alphabet's chief-scientist role, and said he would lean further into Isomorphic to work on curing disease. The man who won a Nobel for predicting protein structure has chosen the drug pipeline over the model lab. Read that as a statement about which layer he thinks compounds.
- **Terray Therapeutics** is the discoverer that owns its own measurement layer rather than renting it, which is this primer's thesis embodied in a single company. Its EMMI platform (Experimentation Meets Machine Intelligence) runs a proprietary, ultra-dense microarray that physically measures billions of small-molecule binding interactions — on the company's figures, on the order of a billion measurements a quarter — a dataset no competitor can buy, and feeds its COATI foundation model; its newer TerraSynth capability closes the loop on the make side with AI-designed synthesis. Because Terray owns the data, the instruments, the models, and the harness that runs the loop, its edge compounds instead of diffusing, unlike a design shop whose model leaks within months. NVIDIA-backed, it is pointed at an internal pipeline led by a brain-penetrant small-molecule inhibitor for multiple sclerosis, alongside partnered programs. (Disclosure: I am an angel investor.)
- **Xaira** (launched with around $1 billion), **insitro** (machine learning grounded in large in-house functional-genomics data), **Genesis Therapeutics** (deep learning for hard small-molecule targets), **Iambic** (physics-informed models, with a candidate already in Phase 1), **Chai Discovery**, and **Cradle** round out a private cohort collectively holding many billions in capital.

China has produced two names worth knowing. **Insilico Medicine (2179.HK)** runs an end-to-end platform (Pharma.AI) and reached a milestone the company describes as a first: Rentosertib, which Insilico says is the first drug with both an AI-generated target and an AI-generated molecule to reach a Phase IIa clinical trial. Treat the "first" as the company's framing, but the underlying fact, an AI-originated candidate in mid-stage trials, is real evidence that the

full loop can produce a clinical candidate. **XtalPi (2228.HK)** pairs quantum-physics simulation with AI and lab robotics. Both carry the China regulatory overhang discussed elsewhere.

The incumbents are not spectators, and this is the most important point for a risk-conscious investor. **Eli Lilly (LLY)** is the most aggressive of the large pharmas, running federated AI programs on NVIDIA's stack (its TuneLab platform) and, as of February 2026, its own LillyPod supercomputer, a 1,016-GPU NVIDIA Blackwell DGX SuperPOD built to train proprietary models on data from more than a billion dollars of internal experiments, while its GLP-1 franchise generates the cash to fund it. Note what Lilly chose to build versus buy: it rented nothing on the data side and spent its own capital standing up LillyPod, because the GPUs are a purchase order and the billion-dollars-of-experiments dataset behind them is not. That combination, an AI program backed by one of the strongest cash engines in the industry, makes Lilly the least binary way to own the theme, because you are not betting the company on a single readout. **Novartis (NVS)** and **Johnson & Johnson (JNJ)** are renting the leading external engines (both are Isomorphic partners), and **Roche/Genentech (RHHBY)** runs deep internal efforts. The incumbents have what the startups lack: the data, the clinical infrastructure, the balance sheet to absorb failures, and the distribution to monetize a win.

The Read on the Tenants

The discoverers carry the field's asymmetric upside, a genuinely AI-designed blockbuster would reprice the whole category, and its genuine binary risk, since most of them are, underneath the software gloss, betting on drugs that will probably fail. Three practical conclusions follow. The AI-armed incumbent, Lilly above all, is the least-binary way to own the theme, because a diversified pharma with a cash engine survives the failures that would sink a single-asset startup. The public pure-plays (Recursion, AbCellera, Absci, Tempus) are the higher-beta "platform, not a drug" bets, appropriate in small size and with clear eyes about the clinical odds. And the best privates, Isomorphic, Xaira, Terray, are simply not investable on the public tape, which is the recurring lesson of this entire primer. When the most exciting tenants are private and the ones you can buy are binary, the durable trade is to own the layers every tenant has to rent from: the measurement, the data, the reagents, and the compute that get consumed whether the drug works or not. You would rather sell picks to all the miners than bet on which one strikes gold.

*That is the machine, end to end. Below the paywall: the investment framework that falls out of it, and **the AI-Science Basket**, a proposed twenty-name book with weights, tiered from the measurement core to the tenant optionality, the one-line case for each, the concentration risk the tiers hide, the bear case that could break the whole thesis (including the one that inverts it),*

and the specific catalysts I am watching. If you take one idea from the free section, take this: the money is not in guessing the drug, it is in owning the bench that every guess has to run on.

The rest of this primer is for paid subscribers

Part Three is the investment framework and **the AI-Science Basket**: twenty liquid, public names with tiers and weights, the one-line case for each, the concentration risk the tiers hide, the ranked bear case, and the catalysts being watched.

```
Paid subscribers: the password-protected full edition is linked in the post.
```

bepresearch.substack.com

Disclosure: *Author is long NVDA, LITE, CRDO, TSEM, LSCC, ALAB, WOLF, NOW, SMCI, ORCL (2027 LEAPS), and BE, and holds angel/private positions in Terray Therapeutics and VistaPath Bio. The AI-Science Basket above is a proposed research expression and may not reflect current positions. This piece is research and commentary, not investment advice. Do your own work.*