

**AMORE
CONSULTING LIMITED**

INDUSTRY WHITE PAPER

AI Drug Discovery's Next Bottleneck: Building the Experimental Data Layer

AIDD industry, infrastructure, and value chain

October 6, 2026 | Authoritative English edition

From faster molecular design to reliable, reusable experimental evidence.

An industry and operating-model analysis. Company examples illustrate the value chain.

The defining industrial question in AI drug discovery is no longer whether models can generate a far larger number of molecular hypotheses. They can. The question is whether organizations can convert those hypotheses into **decision-grade experimental information** faster, more reproducibly, and at a lower total cost of failure than competing discovery systems. That conversion—not model output volume—is becoming the scarce capability.

The bottleneck is moving downstream because digital generation is compressing the cost of proposing structures, sequences, mutations, and assay conditions faster than it is compressing the physical work required to build, measure, interpret, reproduce, and ultimately manufacture them. A model may turn a target into 10,000 plausible designs in hours. Depending on the modality, the physical work includes DNA construction or chemical synthesis, expression and purification where relevant, assay controls, analytical quality, functional confirmation, developability, and traceable data capture. At the end of the chain, a human clinical program must establish safety and benefit. The economic prize therefore belongs less to a generic “AI layer” than to organizations that reliably transform a high volume of proposals into a compact, auditable set of experiments that change a program decision.

That framing changes how the value chain should be analyzed. A computational platform does not create drug value merely by producing more candidates. A synthesis provider does not create durable value merely by processing more oligos. An automated laboratory does not become a learning system merely by adding robotics. A data asset is not valuable merely because it is large. Each must be judged by the quality and provenance of the information it produces, its effect on the next decision, its transferability across programs, and its ability to survive the transition from discovery to preclinical development, chemistry, manufacturing, and controls (CMC), and clinical evidence.

1. The Constraint Has Shifted from Candidate Generation to Evidence Production

The most important unit in the next phase of AI drug discovery is not the candidate, the assay plate, or the model token. It is the **usable experimental result**: a result with known identity, context, controls, quality status, interpretation, and a defined relationship to the next decision. Usable results include reliable negatives. A properly documented nonbinder, expression failure, or unacceptable impurity profile can rule out an unproductive path. An invalid assay can also inform process improvement, but it cannot be treated as a valid biological negative. These outcomes must retain their distinct quality and failure labels rather than be collapsed into an undifferentiated zero.

This distinction explains why AI can simultaneously increase laboratory activity and reduce selected types of testing. Better ranking can eliminate poor candidates before synthesis. Better structure hypotheses can focus mutagenesis. Yet more ambitious target programs,

more variants per design cycle, and more decision points can increase total demand for construction, expression, screening, orthogonal assays, and data review. The direction of demand is conditional on budget elasticity, assay bottlenecks, capacity utilization, turnaround time, and whether a customer uses savings to stop work or to initiate another iteration.

The operating loop is best described as **Design-Make-Test-Learn (DMTL)**. Design proposes a constrained set of molecules or experiments. Make produces the relevant construct, material, or biological system. Test creates measurements under a versioned protocol. Learn joins the raw and processed observations to a model, a scientist's decision framework, or both, then selects the next action. The loop is broken when any output cannot be traced to a sample identity, protocol, instrument state, control set, and analytical method. In that case, a company may still be fast, but it is not accumulating a durable learning asset.

Active-learning research provides the right economic intuition: the objective is to select experiments expected to be most informative under resource constraints, not simply to maximize the number of experiments. [1] This does not authorize fewer controls or weaker independent replication. It means a well-designed system should compare information gained per dollar and per calendar day against a credible baseline. In biology, the baseline is often a combination of expert triage, standardized assays, and simple statistical models—not an imaginary world without human judgment.

The result is an industrial hierarchy. High-quality, high-throughput construction and assays are enabling assets. Automated execution is a productivity asset. Standardized, context-rich data capture is a compounding asset. An organization that can close the loop—with preserved negatives, batch metadata, and an experimental decision policy—has the potential to turn each project into an improvement in future selection. The gap between these categories will determine which suppliers command strategic relationships and which remain interchangeable capacity.

2. What Today's Foundation Models Know—and What They Do Not

The scientific frontier is real, but its industrial meaning depends on separating **structure**, **binding**, **function**, and **clinical value**. AlphaFold 3 (AF3) materially advanced complex-structure generation from protein, nucleic-acid, small-molecule, ion, and modified-residue inputs. Its reported confidence signals—such as pLDDT, PAE, and ipTM—are estimates about structural uncertainty, not direct forecasts of Kd, Ki, IC50, kon, koff, free energy, efficacy, or clinical probability. [2]

That distinction is more than semantic. In AF3's original evaluation, PoseBusters contained 428 protein-ligand structures deposited from 2021 onward. A separate recent-PDB assessment contained 8,856 complexes, and the filtered antibody-antigen subset contained

71 complexes, 166 interfaces, and 65 interface clusters. These are geometric comparisons with experimental structures—pocket RMSD, interface metrics, and related measures. They support the use of AF3 to propose poses, interfaces, and structural experiments. They do not establish that a high-confidence pose has a particular in-vitro binding constant or pharmacological effect. [2]

The model's limits also matter to laboratory design. Reported stereochemical violations in the PoseBusters setting, clashes, hallucinated interactions, performance degradation with shallow multiple-sequence alignments, and difficulty with large complexes or major induced conformational changes are operational reasons to retain multiple hypotheses and orthogonal experimental confirmation. Training/test independence is also task-specific: the PoseBusters analysis used an earlier structure-training cutoff, while other analyses rely on sequence-homology filtering. These are useful protections, not a universal guarantee of target-level novelty. [2]

Boltz-2 illustrates the next step: attaching an affinity-oriented head to a broad structural representation. Its authors describe training that combines approximately 1.2 million continuous ChEMBL/BindingDB binder measurements across roughly 2,000 sequence-clustered targets and 600,000 compounds with PubChem, CeMM, MIDAS, and synthetic decoy data. The supervision mixes K_i , K_d , and IC_{50} -like measurements after normalization to $\log_{10}$ micromolar values. That design makes the output most useful as a ranking signal for general binding strength, not as a literal, assay-invariant affinity constant. [3]

The Boltz-2 paper evaluates two FEP+ subsets: the OpenFE set of 876 high-quality measurements and a separate focused four-target subset. The mean Pearson correlation of 0.66 refers to the four-target subset, not to all 876 measurements; training excludes proteins with at least 90% sequence identity to proteins in the FEP+ benchmark. The paper also reports an enrichment factor of 18.4 at the top-0.5% threshold on MF-PCBA and a zero-fine-tuning CASP16 evaluation of 140 protein–ligand pairs across two targets. Its eight internal blind tests are more sobering: only three exceeded Pearson r of 0.55. The TYK2 forward-looking exercise used absolute-binding-free-energy simulation after virtual screening, not an in-vitro test. As of the research date, the study remains a bioRxiv preprint and should be used as a strong technical signal rather than a settled peer-reviewed conclusion. [3]

The right deployment is therefore series ranking, virtual-screen prioritization, uncertainty stratification, and active-learning candidate selection. Wrong-pocket predictions, conformational errors, omitted water, ions, cofactors, multimeric partners, crop-level loss of long-range effects, and cross-assay heterogeneity can all propagate into an affinity score. A score cannot replace SPR or ITC, enzyme activity, cell potency, selectivity, ADME, toxicology, or developability testing.

ADiT offers a complementary route: an all-atom diffusion-transformer representation that removes MSA and template dependence and is trained at 12 million, 35 million, and 253 million parameters. Its formal splits are clearer than many benchmark narratives—for

example, protein–ligand LBA data split at 30% identity into 3,507/466/490 train/validation/test examples, and a HER2 antibody–antigen test set contains 419 examples. The paper also reports two antibody-maturation exercises covering 780 and 1,420 saturated single mutations, with seven wet-lab-validated affinity-enhancing mutation-ranking cases. [4] The evidence supports a promising shared representation across interaction types, especially when evolutionary alignment information is weak. It does not make the remaining proposed beneficial mutations experimentally confirmed, nor does it establish universal kinetics, functional activity, selectivity, or manufacturability prediction.

The practical implication is a disciplined division of labor. Structural models should generate several experimentally testable pose and interface hypotheses. Affinity models should rank alternatives and expose uncertainty. Experiments should determine which hypothesis survived contact with the real molecular system. The most valuable data layer will make it easy to record not only which prediction won, but also the assay definition, construct boundary, concentration regime, control behavior, and negative outcomes that explain why others did not.

3. Generative Protein Design Is Becoming an Experimental Funnel, Not a One-Shot Drug Factory

Generative antibody and protein design has crossed an important threshold: models can now be embedded in workflows that produce a limited number of expressible, binding molecules against multiple targets. But the evidence strongly favors a **funnel** interpretation over a claim that a model directly produces clinical antibodies.

In the RFantibody/RFdiffusion study, a structure generator was tuned for antibody design and paired with ProteinMPNN sequence completion and RoseTTAFold2 screening. The experiments centered on a fixed humanized VHH framework rather than de novo creation of a complete human IgG. Four high-throughput target arms screened 9,000 yeast-display designs each; lower-throughput *Escherichia coli* arms tested 95 designs each, with influenza HA examined through both approaches. Reported success rates across target/process combinations were 0% to 2%, and the text reports 28 SPR-confirmed binding VHHs. [5]

That denominator discipline matters. Nine thousand is a number of designs entering a display-selection arm, not a count of final soluble molecules. Ninety-five is a separate low-throughput design count. Twenty-eight is a total of confirmed VHH binders, not a universal numerator for every design generated in the work. Representative affinity values—78 nM for HA, 1.4 μ M for RSV site III, 5.5 μ M for RBD, and about 260–262 nM for TcdB—demonstrate real binding across different targets; none is a population-level hit rate. [5]

The study is industrially important because it moves beyond affinity alone. RBD and TcdB binders were evaluated through competition with known de novo binders at the intended epitope. A TcdB VHH produced a cell-protection signal in CSPG4-knockout cells with an EC50

of 460 nM and did not bind the roughly 70%-identical toxin TcsL. Some designs were assembled as scFvs, and selected cases received cryo-EM or structural support for the designed conformation. [5] Those are stronger steps in a discovery funnel than a computational docking score. They are still not evidence of animal efficacy, safety, pharmacokinetics, immunogenicity, or clinical benefit.

The remaining bottlenecks are informative. The work relies largely on purified antigens or display systems, uses one principal VHH framework, and reports low per-target/process success. HA experiments used an insect-cell-produced, monomeric antigen with less glycan shielding than native trimeric HA, a context chosen to match the design model rather than a replacement for native-antigen validation. The authors identify glycan/nonprotein representation, solubility, expression, aggregation, and immunogenicity as future needs. Affinity maturation that reaches low-nanomolar or subnanomolar values is a subsequent experimental evolution step, not the one-pass output of generation. [5]

OriginFlow, a 2025 preprint, provides an instructive contrast. It reports expression, solubility, and SPR testing for designs against PD-L1, SARS-CoV-2 RBD, and VEGF. The manuscript describes 17 selected designs entering soluble E. coli expression, yet also describes 11 of 12 VEGF designs as soluble and binding plus 8 of 8 across PD-L1/RBD. Its principal affinities are in the micromolar to 10^{-4} M range. The coexistence of 17 and 20 (12 plus 8) as sample descriptions means its headline success proportion should not be treated as a stable cross-target statistic. [6]

The broader lesson is not that generative design has failed; it is that **the endpoint must be defined before the funnel starts**. A credible program should report, separately, generated designs, computationally passing designs, constructed designs, expressed molecules, monomeric molecules, binders under a prespecified assay, epitope-correct binders, functional molecules, developable leads, and candidates that advance into formal preclinical packages. It should then repeat the test across independent target classes, antigen formats, and laboratories. That is the operational route by which generative success becomes a purchase decision rather than a publicity metric.

4. Virtual Cells Are Conditional Perturbation Systems, Not Replacements for Biology

“Virtual cell” is useful if it means a model that predicts a specified perturbation response under a stated cellular context. It becomes misleading if it implies a complete simulation that can replace biological experimentation. The strongest current evidence says that benchmark design, the quality of perturbation measurements, and baseline selection matter at least as much as model scale.

A Nature Methods benchmark examined double-perturbation prediction on Norman K562 CRISPRa data containing 100 single-gene and 124 double-gene perturbations with 19,264

RNA readouts. Training included all 100 singles and 62 doubles; the other 62 doubles were held out, with five random partitions. On the 1,000 most expressed genes, the tested models were outperformed by simply adding the single-perturbation log-fold changes. They also failed to improve genetic-interaction identification over a no-change baseline. In other K562 and RPE1 single-perturbation tests, deep models did not reliably exceed training-set mean or linear baselines. [7]

That does not mean foundation representations are useless. In the same analysis, scGPT and scFoundation gene embeddings paired with a linear decoder could meet or exceed their native decoders, and perturbation-data pretraining was more helpful than pretraining exclusively on observational atlases. The conclusion is more nuanced: representation quality may help, but “foundation model” is not a substitute for a well-specified causal prediction problem or a simple baseline.

An independent benchmark reached a similar directional finding: a mean predictor outperformed scGPT and scFoundation, while feature-informed conventional machine learning clearly exceeded scGPT in its comparisons. It also identified low perturbation-specific variance in existing Perturb-seq benchmarks, an important data-quality warning. [8] If most measured signal is shared baseline behavior or technical variation, a model can look plausible while contributing little incremental biological insight. Laboratory buyers should therefore ask about effect sizes, perturbation-specific variance, knockdown effectiveness, replicate structure, batch diversity, and pseudobulk as well as single-cell scoring.

The Arc Virtual Cell Challenge improves the test design by releasing H1 human embryonic stem-cell CRISPRi data with approximately 300,000 cells across 300 targets, more than 50,000 UMI per cell, and a median near 1,000 cells per perturbation. It reserves 100 perturbations for a sealed final test after 150 training and 50 validation perturbations and scores differential-expression sets, perturbation discrimination, and transcriptome MAE. [9] This is few-shot adaptation within a new cell background, not strict zero-shot prediction. It remains one cell type and single-gene perturbations, but it supplies a useful template: split by the condition that matters, use several metrics, and keep final experimental truth inaccessible until evaluation.

For industry, virtual-cell systems are best positioned to prioritize perturbations, identify mechanism hypotheses, and allocate limited Perturb-seq or high-content experiments. They should not yet be used to assert patient response, tissue-level safety, protein or epigenetic state, or clinical efficacy without those measurements. The data opportunity is not simply more single-cell sequencing. It is a context-rich perturbation dataset spanning cell type, state, donor or line, dose, time, delivery method, targeting efficacy, batch, and controls—organized so that a genuine unobserved-condition test is possible.

5. Closed-Loop Laboratories Must Learn from Instruments, Not Just Run Them

A laboratory becomes closed-loop only when its measurement changes the next experimental choice. A fixed robotic protocol is automation. A virtual screen followed by assay testing, without using the measurements to change the next selection round, remains open-loop. A true DMTL system connects Make/Test execution, instrument data, quality rules, and a decision algorithm that changes the next condition or candidate.

Three primary studies show both the promise and the narrowness of current evidence. Sheng and colleagues connected flow dispensing, automated cyclic voltammetry, deep-learning mechanism classification, and Bayesian optimization to investigate CoTPP/organic-halide electrochemistry. The system ran for approximately 50 hours without intervention and examined 2,520 parameter combinations; each second-stage inquiry generated six voltammograms. [10] It is a real experimental-feedback loop. Its learning claim is nonetheless bounded by a specialized electrochemical system, predefined mechanism classes, and no matched human-control cost or information-gain study.

Xu and colleagues' self-driving laboratory for PNIPAM lower critical solution temperature demonstrates a more frugal architecture: five parallel liquid-handling and optical modules, Peltier control, Gaussian-process surrogates, and expected-improvement selection. The two-salt experiment starts from 13 points and the three-salt experiment from seven. Reported five-module controls had a mean LCST of 30.59°C and a standard deviation of 0.15°C. [11] The work is a credible local optimization demonstration, but it is an arXiv preprint, depends on human preparation and interface choices, and validates on the same platform from which its data accumulate. It does not establish cross-laboratory generalization or a universal return-on-automation ratio.

A-Lab is a higher-throughput materials example. Across 17 days, it executed 353 inorganic-synthesis experiments against 57 human-specified targets, reporting 36 successes. Robot powder handling, heating, grinding, X-ray diffraction interpretation, and active learning were connected. [12] The authors also report by-products and four cases in which XRD could not exclude an ambiguity with an already known phase. A target phase appearing in an automated run is not equivalent to high purity, scalable manufacture, a new structure, or final device performance.

These studies show why the enterprise moat is often not the robot. It is the **protocol compiler and evidence model** around the robot: a versioned method, calibration status, consumable lot, sample map, reference material, raw sensor stream, QC gates, exception rules, and safe stop/approval path. The biopharmaceutical self-driving-laboratory literature identifies sensor compatibility, heterogeneous data formats, and multi-instrument integration as persistent constraints. [13] Vendor announcements about real-time quality

control or model-assisted protocol workflows demonstrate momentum, but they are not substitutes for disclosed end-to-end error rates or validated GxP deployment. [14]

A laboratory should thus be valued by its marginal reduction in uncertainty under a specified decision, not by its apparent autonomy. The next model iteration should be able to distinguish “not tested,” “not expressed,” “assay failed,” “reliable negative,” and “positive with caveats.” Without that distinction, automation can produce a larger database while amplifying a systematic experimental bias.

6. Each Drug Modality Creates a Different Experimental Demand Curve

A common analytical error is to treat all AI-generated candidates as if they required the same downstream infrastructure. They do not. A small molecule, an antibody, a designed enzyme, a nucleic-acid therapeutic, and a cell or gene therapy generate different failure modes, different evidence needs, and different transitions from discovery material to development material.

For small molecules, a useful discovery package commonly requires identity and purity confirmation, target engagement or functional assays, selectivity, solubility, permeability, metabolic stability, exposure relationships, and early safety signals. A structural or affinity model can prioritize compounds, but it does not guarantee synthetic accessibility, impurity control, off-target pharmacology, reactive metabolite behavior, central-nervous-system exposure, or organ risk. The best data layer connects each molecular structure to the actual batch, analytical result, assay conditions, and result status rather than treating a calculated score as a property of the molecule itself.

For antibodies, proteins, and other biologics, the experimental chain is shaped by sequence construction, expression yield, folding, affinity, epitope, cross-reactivity, thermal and colloidal stability, aggregation, glycosylation, formulation, and immune risk. In this domain, a K_d is one component of a broader developability decision. The RFdiffusion result is compelling precisely because it includes target diversity, epitope competition, a functional cell signal, and selected structural corroboration; its limits show why no one of those elements can be skipped. [5]

Nucleic-acid medicines and cell or gene therapies add delivery, editing efficiency, off-target effects, tissue specificity, durability, immunogenicity, and sometimes vector or cell-product attributes. These programs can create demand for sequencing, single-cell methods, proteomics, spatial biology, and high-content imaging, but only where the biological question and assay design require those modalities. More AI design iterations do not mechanically translate into more use of every instrument or consumable category.

The same principle applies to disease biology. A perturbation atlas may be extraordinarily valuable for target validation in a heterogeneous disease setting and irrelevant to a well-characterized biochemical lead-optimization decision. A high-throughput binding assay may triage a protein design library yet say little about agonism, antagonism, degradation, signaling bias, or tissue-specific function. The commercial consequence is that service providers should sell **decision packages**, not undifferentiated throughput. A decision package states the modality, intended decision, assay hierarchy, turnaround, data fields, failure handling, and any required orthogonal confirmation.

This is also why Discovery, Preclinical, CMC, and Clinical must not be conflated. Discovery asks whether a target and candidate merit further work. Preclinical asks whether a selected candidate has sufficient pharmacology, exposure, safety, and translational support to enter humans. CMC asks whether the material has a controlled identity, quality, purity, potency, stability, and traceable production history. Clinical development asks whether a product is safe and beneficial in people. Faster motion in one stage is valuable only if it improves the probability-adjusted path through the later stages.

7. CMC Is a Separate Bottleneck Between a Discovery Sample and a Clinical Material

The CMC handoff is where a discovery organization confronts a hard physical truth: a sample that supports an early assay does not automatically support dosing in people. For an investigational new drug application, FDA expects sufficient CMC information to assure identity, quality, purity, and strength. Early clinical development can use a phase-appropriate approach to stability and analytical validation; it does not waive the underlying obligation to establish an appropriate control strategy. [15]

For small molecules, this means understanding a repeatable synthetic route, starting-material boundaries, impurity formation and purge, residual solvents or metals where relevant, solid form or salt selection, packaging, transport, and stability. EMA's quality documentation framework and ICH Q11 emphasize process understanding, material attributes, impurities, and control strategies rather than a generic checklist applied identically to every molecule. [16] [17] An AI-generated scaffold may make a synthesis planning system look successful while creating a practical CMC burden through unstable intermediates, a difficult impurity profile, a nonrobust crystallization, or a form that does not hold through distribution.

For biologics, sequence is still less than the product. Cell bank, host, expression history, culture process, purification, formulation, glycan distribution, aggregation, process-related impurities, viral safety, and potency can all change what reaches a patient. EMA explicitly distinguishes the quality documentation expected for biological investigational medicinal products from small-molecule considerations. [18] A transiently expressed discovery

antibody and a stable production-cell-line product may share an amino-acid sequence while differing materially in product profile and control burden.

Comparability is therefore an industrial capability, not an administrative afterthought. ICH Q5E does not require every analytical value to be identical after a manufacturing change; it requires evidence that the change has not adversely affected safety, quality, or efficacy, with the depth of analytical, functional, nonclinical, or clinical bridging appropriate to the case. [19] When a program transfers from a discovery laboratory to a CDMO, method transfer, bridging lots, deviations, out-of-specification investigation, change control, and retained product knowledge can become the limiting steps.

The most consequential implication for AIDD is portfolio selection. A company that generates abundant design options should use manufacturing feasibility as an early, explicit selection objective—not wait until a lead becomes politically difficult to replace. For small molecules, this can mean route robustness and solid-form risk. For biologics, it can mean expression, product quality, viscosity, aggregation, glycan sensitivity, and stable-cell-line plausibility. CMC-aware design will not eliminate late surprises, but it can keep a discovery organization from optimizing itself into a manufacturing dead end.

8. The Data Product Is an Auditable Evidence System

Laboratory data become a product when another authorized scientist, model, customer, auditor, or manufacturing partner can determine what was measured, how it was measured, which version of the method applied, and whether the result is fit for a stated use. That is a higher standard than exporting a spreadsheet of endpoints.

A useful data product has four connected layers. The **record layer** preserves raw signals, sample identifiers, reagent and consumable lots, instrument and firmware state, timestamps, and chain of custody. The **protocol layer** versions the procedure, reference materials, controls, plate map, dilution logic, acceptance limits, and deviations. The **quality layer** keeps pre- and post-correction outputs, QC flags, detection limits, retries, invalid runs, and human overrides. The **decision layer** maps the result to assay interpretation, model version, uncertainty, and the action it informed. The NCATS Assay Guidance Manual program is a strong public reminder that experimental design, rigor, analytical technique, and data analysis jointly determine preclinical translation quality. [20]

FAIR principles are often misunderstood in commercial settings. They do not require a company to publish proprietary data. Properly applied, they support persistent identifiers, standardized metadata, interoperable formats, controlled access, and reuse under defined permissions. [21] High-throughput chemistry infrastructure research similarly argues that failed reactions, negative outcomes, branch decisions, batch identifiers, and instrument records are part of the record needed for reproducible computational learning. [22]

Instrument interoperability and biological semantic interoperability must be separated. SiLA 2 can standardize service descriptions, self-description, discovery, and information structures so that laboratory systems integrate with less bespoke engineering. [23] A common interface does not mean that two binding assays, cell models, or endpoints are biologically comparable. The first problem is about connecting instruments. The second is about defining the meaning, units, controls, and limitations of measurements. A provider that solves only the first may improve operating efficiency without creating cross-program training data.

This distinction affects pricing. A validation-oriented CRO typically charges for samples, plates, method development, repeat runs, and reports. A durable data-service offering may charge for an authorized data domain, curation, API access, version maintenance, and updates. The latter only earns a premium if it has rights to supply those features and can demonstrate that a version is reusable. A provider cannot presume that a client's failed experiments may be resold, pooled into a model, or retained indefinitely simply because those records are technically valuable.

Platform examples show the direction of travel. Lilly TuneLab sets minimum experimental contributions—1,000 experimental readings for small-molecule participants or 125 antibody sequences tied to an attribute or endpoint—and uses federated-learning-style parameter exchange rather than raw-data exchange. [24] Amazon Bio Discovery describes a workflow in which candidate ranking, external testing, result return, and customer-private model refinement are connected. [25] Both demonstrate that data is being operationalized as a collaborative asset. Neither public description substitutes for the contract provisions governing derived features, model updates, telemetry, retention, deletion, access after exit, or subcontractor use.

9. Build, Buy, or Hybrid Is a Design Choice, Not an Ideological One

The best operating model follows the location of iteration value. Building internal capability is justified when experiments are frequent, strategically differentiating, deeply coupled to scientists' decisions, dependent on proprietary samples, or too sensitive to queueing and data leakage. Buying external capacity is attractive for standardized assays, specialized equipment, variable peaks, independent replication, and capabilities that would otherwise sit underutilized. A hybrid model often retains the high-frequency or strategic decision experiments while outsourcing peak demand, specialized tests, and deliberate replication.

A build decision carries more than capital expenditure. It commits an organization to utilization risk, staffing, calibration, maintenance, laboratory information management, EHS, supply continuity, sample logistics, method governance, and quality documentation. The reward is tight feedback and control over scarce samples and data. A buy decision turns

some fixed costs into variable spend and can access standardized know-how; it may introduce queueing, context loss, contract complexity, and weaker control over the experimental loop. Neither model is inherently superior across modalities or company stages.

The hybrid option has an additional scientific advantage: independent replication. An internal result can be sent to an external laboratory under a blinded or prespecified plan, while external high-volume output can be checked internally on reference samples. Cross-plate anchors, randomized controls, reference materials, batch-aware analysis, and preservation of both adjusted and unadjusted outcomes guard against an organization learning its own laboratory artifact at scale.

Supplier evaluation should consequently move beyond headline throughput. Buyers should ask for median and distributional turnaround, not just best-case time. They should distinguish calendar time from instrument time, and peak capacity from sustained capacity. They should seek documentation on expression rate, assay validity rate, resubmission rate, failure categories, test/retest concordance, control performance, data-export structure, and the process for deviations. The highest-value supplier is often the one that can reduce the time from a candidate request to an auditable decision—not the one that advertises the largest number of theoretical operations per day.

This is also a reason to avoid treating every automation announcement as equivalent. Vendor capability announcements involving AI-assisted workflow design or live QC can be useful sourcing signals. [14] They become operational evidence only when an organization can show the performance of a bounded workflow: accepted samples, invalid runs, reruns, sustained throughput, data completeness, decision turnaround, and safe recovery from exceptions.

10. Commercial Signals Favor Providers That Convert Sequence into Trustworthy Data

Commercial adoption is beginning to appear, but it must be read with the same denominator discipline as scientific benchmarks. Twist Bioscience is a useful example of a provider positioned between sequence construction, antibody/protein discovery, and characterization. Its fiscal 2025 Form 10-K states that it can supply antibodies and high-quality characterization data to AI-enabled antibody-discovery customers for algorithm development, improvement, and developability screening. [26] Its public product materials advertise more than 6,000 IgG antibodies per week, with complete characterization data as fast as 15 days. [27] That is a platform capability statement, not a measure of utilization, paid orders, or universal workflow timing.

The stronger commercial signal is the company's stated order composition: its fiscal 2026 first-quarter materials said that more than \$25 million of fiscal 2025 order growth of \$66

million came from AI drug discovery. [28] These are booked orders, not automatically recognized revenue. They do not disclose an AI-only revenue line, cancellation rates, order duration, customer concentration, gross margin, or cash collection. Its fiscal 2026 third-quarter release reported total revenue of \$118.376 million and DNA Synthesis and Protein Solutions revenue of \$56.6 million, up 39% year over year; that segment supports AI-related workflows but is not an AI-only segment. [29]

GenScript illustrates a different route: an integrated Gene-to-Protein service model. The company reported that its AI-enabled drug-discovery business doubled year over year in the first half of 2026, while group revenue was about \$404.2 million, Life Science Group revenue about \$319.0 million, and Gene-to-Protein represented approximately 66% of Life Science Group revenue. [30] The facts show growth and a potentially relevant operating footprint. They do not permit an inference of AIDD revenue, margin, customer count, or conversion rate because the business is not separately reported in that way.

GenScript's AIDD materials advertise industrial-scale screening of more than 4,000 designs per day and a rapid sequence-to-data service as fast as four calendar days for VHHs and miniproteins, using 96/384-well cell-free expression and affinity ranking. They also describe a 7–10 day developability package across a broader set of formats and separate 14-day and 15-business-day examples. [31] The commercial lesson is precise: a fast, narrow sequence-to-data package can be valuable, but it is not a universal sequence-to-binding guarantee and is not the same as full developability characterization. The useful customer question is what data fields, assay depth, and repeatability accompany each turnaround promise.

There are also strategic signals from AI organizations entering or partnering with life-science laboratories. Anthropic disclosed a life-sciences research team and a Bay Area BSL-1/BSL-2 molecular-biology laboratory operated by human scientists; its protein-design work involved external wet-lab testing. [32] In one publicly described external validation, 1,320 designs produced 354 experimentally confirmed binders, while an expression rate near 95% did not equate to binding, function, or drug status. [33] Adapticv describes that validation as open-loop: using outcomes to guide the next design round is a prospective learning workflow, not evidence that these results have already retrained Claude. These examples point to rising demand for wet-lab interfaces. They do not, by themselves, establish a general clinical pipeline or a sector-wide demand forecast.

OpenAI's GPT-Rosalind offers a complementary route: a life-sciences reasoning model series designed for evidence synthesis, sequence-to-function interpretation, experimental planning, data analysis, and scientific tool use. OpenAI's September 2026 update expanded availability to eligible organizations globally through trusted access. [34] The industrial significance is workflow integration: a governed research assistant can connect biological questions, specialist tools, and laboratory decisions. A research benchmark or an access announcement does not, by itself, demonstrate a proprietary drug pipeline or improved clinical success.

Other company strategies show that the commercial model can extend from software toward partnered discovery and owned programs. Anew Labs publicly describes its platform and early target programs. [35] Reuters reported its separation from ByteDance alongside continued majority ownership by ByteDance, citing people familiar with the transaction. [36] Isomorphic Labs has stated that financing was intended to advance therapeutic programs into clinical development, and subsequent reporting moved its first-human-trial expectation to late 2026. [37] [38] Both cases underline a basic industry truth: a platform, a collaboration, a financing event, and a clinical study are different facts with different evidence thresholds.

11. Biosafety, Data Rights, and Supplier Governance

Become Part of Deliverable Capacity

Scaling design requests changes the compliance workload. Synthetic-nucleic-acid suppliers and buyers must contend with customer verification, sequence screening, review of legitimate purpose, escalation, and recordkeeping. The 2023 HHS screening guidance is a recommended baseline framework for providers, users, and benchtop-device manufacturers, not a single worldwide law. [39] The 2024 OSTP framework makes screening a condition of a core set of U.S. government-funded life-science procurement; a May 2025 executive order directed revision or replacement, and the status of any successor must be evaluated as it becomes available rather than assumed in advance. [40] [41]

The International Gene Synthesis Consortium's Harmonized Screening Protocol is a member-governed framework covering customer and sequence screening, manual escalation, and records. It is influential operational guidance, not an international treaty or a guarantee that every provider follows identical procedures. [42] The industrial implication is straightforward: safety review is part of usable capacity. A provider that can make designs quickly but cannot identify the requesting entity, preserve approval records, and manage escalations may not be able to deliver responsibly at the speed implied by an order queue.

Data rights need equally explicit architecture. Contracts should distinguish background intellectual property; customer sequences and raw results; derived features and model updates; supplier process know-how; external CRO and cloud-subprocessor access; retention periods; reuse permission; deletion; audit rights; and cross-border transfers. Ownership of a raw assay file is not automatically equivalent to permission to train a cross-customer model. Likewise, federated learning may reduce raw-data movement without eliminating risks related to model inversion, membership inference, retention, or post-termination access.

Governance should be embedded in the LIMS and commercial workflow, not bolted on at the end. Each order should connect identity, sequence or construct version, approval status, actual recipient, physical shipment, assay result, and allowed data use. This protects customers and providers while making the data more valuable for legitimate future learning.

It also creates a defensible boundary between a fast-turn service and an enterprise-grade experimental data product.

12. Value Capture Requires Information Quality, CMC Readiness, and Clinical Proof

A simplified operating model shows how changes in price, cost, and data quality affect value capture. Let **N** be accepted experimental requests, **p** the normalized unit price, **c** the direct variable cost per request, **F** allocated platform fixed cost, and **q** the usable-and-interpretable data rate. Here, **q** includes valid negative results; it is explicitly not a binding-hit rate. Define revenue as **R = Np**, total cost as **C = Nc + F**, and operating contribution as **R – Nc – F**. This is an operating contribution construct, not a reported gross-margin or EBITDA measure.

The following are hypothetical index-based scenarios, not company forecasts or observed results; monetary values are normalized to the same baseline unit-price scale. Usable-data units equal **Nq**. Three cases show why throughput alone is insufficient. In Scenario A, **N = 100**, **p = 1.00**, **c = 0.60**, **F = 20**, and **q = 0.80**. Revenue is 100, total cost is 80, operating contribution is 20, contribution margin is 20.0%, and usable data are 80 units. In Scenario B, **N = 150**, **p = 0.85**, **c = 0.50**, **F = 30**, and **q = 0.85**. Revenue is 127.5, total cost is 105, contribution is 22.5, contribution margin is 17.65%, and usable data are 127.5 units. In Scenario C, **N = 200**, **p = 0.70**, **c = 0.40**, **F = 45**, and **q = 0.60**. Revenue is 140, total cost is 125, contribution is 15, contribution margin is 10.71%, and usable data are 120 units.

The mechanism is the point. Scenario C doubles requests relative to Scenario A but produces less operating contribution and fewer usable data than Scenario B because price compression, higher fixed-cost absorption needs, and a lower **q** overwhelm volume. Where **p** exceeds **c**, the break-even volume is **N* = F / (p – c)**: 50 requests in A, about 85.7 in B, and 150 in C. Discounting or idle capacity can therefore block economic conversion even as a laboratory appears busier.

This framework also explains why broad pharmaceutical R&D budgets cannot be treated as an AIDD addressable market. Roche, Novartis, and IQVIA report research-and-development expenditures using different entities, accounting definitions, time periods, currencies, and population scopes. [43] [44] [45] Discovery experiments, outsourced laboratory spend, clinical development, CMC, and AI software are overlapping but nonidentical pools. The relevant commercial measures are order intake, accepted requests, queue time, completed experiments, usable-data yield, customer acceptance, revenue recognition, receivables collection, repeat behavior, and the contribution cost of each usable result.

Clinical proof remains the ultimate boundary. A randomized phase 2a study of the AI-discovered TNIK inhibitor rentosertib in idiopathic pulmonary fibrosis enrolled 71 people across 21 Chinese centers and reported a 12-week FVC change of +98.4 mL in the 60 mg once-daily group versus –20.3 mL with placebo. It also reported treatment-related adverse

events in 77.8%, treatment-related serious adverse events in 11.1%, and discontinuation for adverse events in 22.2% of the 60 mg group. [46] The study is an important early human signal, not proof that AI improves clinical success: the sample is small, follow-up is short, missingness and attrition matter, and a larger, longer, more representative program is required.

The frequently cited early analysis of AI-native pipelines reached 21 successes among 24 molecules completing Phase I and four successes among 10 completing Phase II. [47] Those small, public, completed-stage denominators are useful for hypothesis formation, but they are susceptible to selection, disclosure, survivorship, target, modality, and stopping-decision effects. They do not establish an unbiased industry-wide clinical success rate.

Regulatory evolution may change the composition of preclinical evidence, not eliminate the need for it. FDA's New Approach Methodologies work is framed around a stated context of use, human relevance, technical reliability, and fit for purpose. [48] Draft guidance addressing some monoclonal-antibody cases can reduce or omit six-month nonhuman-primate toxicology studies in particular situations; it is not a blanket rule for small molecules, vaccines, or complex biologics. [49] Other draft guidance encourages scientifically appropriate alternatives to animal testing. [50] Where nonhuman primates remain necessary, availability and regulated import logistics still matter; Charles River's acquisition disclosure and CDC's minimum 31-day quarantine requirements illustrate that these are operational constraints, not interchangeable assay slots. [51] [52]

The white paper's central thesis is therefore falsifiable. It will be supported if providers demonstrate, over time, that AI-associated requests convert into recognized revenue and cash; that usable-data rates, protocol reuse, customer acceptance, and decision-cycle time improve without a collapse in price or quality; that CMC feasibility is incorporated earlier in candidate selection; and that independently measured preclinical and clinical outcomes improve against appropriate historical and contemporaneous comparators. It will be weakened if order growth fails to convert, q declines as throughput rises, discounting forces contribution below the added fixed-cost burden, experimental results do not reproduce across laboratories, CMC eliminates a disproportionate share of generated leads, or later-stage human evidence fails to exceed credible baselines. Those are the operating observations that should replace both unqualified optimism and blanket skepticism.

Notes

The **English edition is the authoritative version** of this white paper. The Chinese companion edition follows its facts and judgments. In the event of any difference of interpretation, the English edition prevails. Background materials supplied by the commissioning party included texts attributed to **Freda Duan**, *AI Drug Discovery Is Becoming a Bottleneck Trade*, and **metavestor**, *The ultimate bio x AI thematic primer for investors*. They are treated as background context, not as independently cited primary sources. The report uses **Twist Bioscience**, the brand presented on the company's official

Chinese-language site. Quantitative statements unsupported by adequately identified primary or high-quality sources—including certain reported capacity and revenue figures—are not used for quantitative analysis.

Disclaimer: This paper is industry information and analysis only. It is not a substitute for clinical, regulatory, or legal judgment.

References

- [1] PLoS Computational Biology | June 1, 2017 | A review of active learning approaches to experimental design for uncovering biological networks
- [2] Nature | May 8, 2024 | Accurate structure prediction of biomolecular interactions with AlphaFold 3
- [3] bioRxiv / PMC | June 18, 2025 | Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction
- [4] ICLR | April 23, 2026 | Towards All-Atom Foundation Models for Biomolecular Binding Affinity Prediction
- [5] Nature | November 5, 2025 | Atomically accurate de novo design of antibodies with RFdiffusion
- [6] bioRxiv | April 29, 2025 | Robust and Reliable de novo Protein Design: A Flow-Matching-Based Protein Generative Model Achieves Remarkably High Success Rates
- [7] Nature Methods | August 4, 2025 | Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines
- [8] BMC Genomics | April 14, 2025 | Benchmarking foundation cell models for post-perturbation RNA-seq prediction
- [9] Cell | June 26, 2025 | Virtual Cell Challenge: Toward a Turing test for the virtual cell
- [10] Nature Communications | March 30, 2024 | Autonomous closed-loop mechanistic investigation of molecular electrochemistry via automation
- [11] arXiv | September 2, 2025 | Self-driving Laboratory Optimizes the Lower Critical Solution Temperature of PNIPAM in Multi-salt Systems
- [12] Nature | November 29, 2023 | An autonomous laboratory for the accelerated synthesis of inorganic materials
- [13] Nature | 2026 | Towards self-driving laboratories in the biopharmaceutical industry
- [14] NVIDIA | January 12, 2026 | NVIDIA BioNeMo Platform adopted by life sciences leaders to accelerate AI-driven drug discovery
- [15] U.S. Food and Drug Administration | Updated June 22, 2026 | IND Applications for Clinical Investigations: Chemistry, Manufacturing, and Control (CMC) Information
- [16] European Medicines Agency | January 28, 2022 | Requirements to the chemical and pharmaceutical quality documentation concerning investigational medicinal products in clinical trials
- [17] International Council for Harmonisation | May 2012 | Q11 Development and Manufacture of Drug Substances
- [18] European Medicines Agency | January 28, 2022 | Requirements for quality documentation concerning biological investigational medicinal products in clinical trials
- [19] International Council for Harmonisation | November 2004 | Q5E Comparability of Biotechnological/Biological Products Subject to Changes in Their Manufacturing Process
- [20] Nature Reviews Drug Discovery | December 2022 | The NCATS Assay Guidance Manual programme: advancing the practice and rigour of preclinical translation
- [21] Frontiers in Drug Discovery | July 12, 2023 | A FAIR data management: what does it mean for drug discovery?
- [22] Digital Discovery | 2025 | A FAIR research data infrastructure for high-throughput digital chemistry
- [23] SiLA Organization | Accessed October 6, 2026 | Frequently Asked Questions
- [24] Eli Lilly | Accessed October 6, 2026 | Introducing Lilly TuneLab
- [25] Amazon Web Services | April 14, 2026 | Amazon Bio Discovery: AI-powered drug research
- [26] Twist Bioscience | November 17, 2025 | Form 10-K for fiscal year ended September 30, 2025
- [27] Twist Bioscience | Accessed October 6, 2026 | Official Chinese-language site

- [28] Twist Bioscience | February 2, 2026 | Fiscal 2026 First Quarter Financial Results
- [29] Twist Bioscience | August 3, 2026 | Fiscal Third Quarter 2026 Financial Results
- [30] GenScript Biotech | August 18, 2026 | GenScript Reports First-Half 2026 Results
- [31] GenScript | Accessed October 6, 2026 | AI Drug Discovery Solutions
- [32] Anthropic | September 23, 2026 | Claude discovers a novel enzyme system
- [33] Adaptiv Bio | August 20, 2026 | Benchmarking Claude's protein designs in the wet lab
- [34] OpenAI | April 16, 2026; updated September 11, 2026 | Introducing GPT-Rosalind for life sciences research
- [35] Anew Labs | Accessed October 6, 2026 | Anew Labs official site
- [36] Reuters | September 16, 2026 | ByteDance completes fundraising for AI drug unit after its spin-off
- [37] Isomorphic Labs / PR Newswire | March 31, 2025 | Isomorphic Labs announces funding to advance therapeutic programs into the clinic
- [38] Reuters | January 20, 2026 | Google-backed AI drug discovery startup Isomorphic Labs delays clinical trial timeline
- [39] U.S. Department of Health and Human Services | 2023 | HHS Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids
- [40] OSTP / HHS | April 29, 2024 | OSTP Framework for Nucleic Acid Synthesis Screening
- [41] The White House | May 5, 2025 | Executive Order 14292: Improving the Safety and Security of Biological Research
- [42] International Gene Synthesis Consortium | September 3, 2024 | Harmonized Screening Protocol v3.0
- [43] Roche | 2026 | Annual Report 2025: Innovating for a healthier future
- [44] IQVIA Institute | March 2025 | Global Trends in R&D 2025
- [45] Novartis | February 3, 2026 | Annual Report 2025
- [46] Nature Medicine / PMC | June 3, 2025 | A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial
- [47] Drug Discovery Today | 2024 | How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons
- [48] U.S. Food and Drug Administration | Accessed October 6, 2026 | New Approach Methodologies (NAMs)
- [49] U.S. Food and Drug Administration | December 2, 2025 | FDA Releases Draft Guidance on Reducing Testing on Non-Human Primates for Monoclonal Antibodies
- [50] U.S. Food and Drug Administration | March 18, 2026 | FDA Releases Draft Guidance on Alternatives to Animal Testing in Drug Development
- [51] Charles River Laboratories | 2026 First Quarter | Form 10-Q
- [52] U.S. Centers for Disease Control and Prevention | Accessed October 6, 2026 | Bringing a Nonhuman Primate into the U.S.