

Next-Generation Enzyme Engineering at Ginkgo

Aarya Venkat, Tim Curran, Maria Fitzpatrick,
Nathan Schmidt, Kunal Mehta



September 2026



GINKGO
BIOWORKS



Introduction

Enzymes are attractive catalysts in industrial and pharmaceutical applications, from industrial detergents to active pharmaceutical ingredient (API) synthesis, and increasingly for defense, health, and supply-chain-relevant chemistries. Enzymes are particularly useful catalysts for reactions requiring very high stereo- and site-specificity and to create asymmetric structures. They can achieve in one step what often takes several with traditional catalysts, making biocatalysis simpler and shorter. And they are usually active in mild conditions (e.g. temperature, pH, solvents), making biocatalytic processes more sustainable than chemical manufacturing.

Successful application of enzymes requires engineering for activity and specificity, often on substrates and in conditions quite different from their native ones, and preserving that activity over long reaction times. The past two decades have seen engineered enzymes used in applications ranging from detergents to catalysts for API manufacturing. But doing this across enzymes and enzyme classes has been difficult, especially for complex synthetic chemistries which are often thermodynamically challenging and require biochemical cofactors. Thus, most commercialized enzymes today [degrade substrates like carbohydrates and lipids](#)—relatively simple chemistry—and *synthetic* applications largely come from just [three well-understood enzyme classes](#).

Optimizing enzymes is significantly more challenging than optimizing binding proteins like antibodies. Where antibody optimization primarily focuses on binding kinetics, balancing association (K_{on}) and dissociation (K_{off}) alongside developability, enzyme activity is a tradeoff between multiple parameters— K_{on} , K_{off} , K_{cat} , K_m —with no single optimum, e.g. improved catalysis often impairs stability. Epistasis, the property that the best amino acid at one position depends on those at others, creates a “rugged” fitness landscape for enzymes with multiple optima which is difficult to explore systematically. The design space is orders of magnitude larger: antibody binding is largely determined by 30 residues near the binding site; epistatic mutations in enzymes are often distributed over 100s to 1000s of residues. Finally, catalysis is dynamic and the sometimes complex conformational changes require integration of multiple approaches to model.

Advancing enzyme engineering is increasingly a matter of national priority. Engineered enzymes underpin resilient domestic manufacturing, secure supply chains, and the sustainable production of the chemicals, materials, and therapeutics on which government missions depend.¹ Ginkgo built and operates this platform in the United States, and the work described here emphasizes programs delivered for U.S. Government (USG) partners—and, by extension, the U.S.-based manufacturers who translate engineered enzymes into fielded products. Ginkgo seeks to advance enzyme engineering from a world of bespoke, arduous screening to one of predictable design. We believe this will require new infrastructure for automated, flexible, reproducible work in both computational design and experimental execution.

Automation frees scientists from managing low-level parameters to focus on solving the problem; flexibility respects the inherent diversity in biology and exposes the right tool for any job; and reproducibility provides the foundation needed for learning and prediction. Here, we describe the enzyme engineering platform built over a decade at Ginkgo to deliver this, with emphasis on work done for US Government partners, and outline our envisioned path to higher levels of autonomy and predictive power.

¹ [National Security Commission on Emerging Biotechnology, “Charting the Future of Biotechnology: Final Report” \(April 2025\)](#)—the bipartisan Congressional commission’s report framing the U.S. bioeconomy as a national-security priority and calling for scaling the domestic biomanufacturing base.

Value of Automation in Computational Enzyme Design

Until recently, computational protein design has had low success rates and a lack of generalizability. Predictive tools were highly specialized forcing designers to construct bespoke workflows for every new target, where breakthroughs on one target rarely translated to another.

The emergence of large-scale sequence and structural databases, new AI and physics-based models, and major improvements in computing has enabled protein design to become more predictive and reproducible. In spite of this, these workflows are still too fragmented for an individual designer to

execute alone, thus Ginkgo's platform resolves this bottleneck by automating these workflows at scale.

Two core pillars support our robust automated workflows for protein design: Kubernetes, a platform that orchestrates flexible compute, and Argo, a framework for defining modular, reusable workflows.

This level of automation allows scientists to design libraries leveraging a wide suite of prediction tools in days, and enables individual designers to manage up to ten programs in parallel, all without specialized computational training. We envision a platform that doesn't just focus on cross-functional collaboration, but cross-enablement, evolving the experimentalist into a computational designer.

Our enzyme development campaigns typically begin with a "discovery" phase to find the right starting point in nature or literature, followed by an "optimization" phase where multiple objectives are optimized at once. Across both phases we typically require 3-4 engineering cycles to deliver application-ready enzymes, with exceptions on either end: in several cases, we have delivered commercializable enzymes directly from the initial discovery screen; some challenging programs have gone up to ten cycles with continual—even accelerating—improvement throughout.

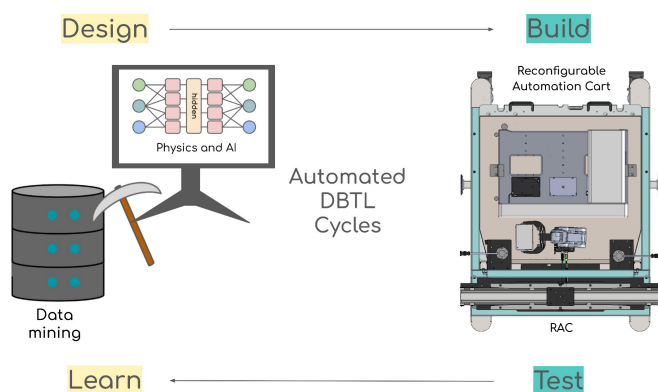


Figure 1: Schematic of an automated Design-Build-Test-Learn cycle. The workflow integrates AI- and physics-informed computational design and learning (left) with physical building and testing executed by a suite of Reconfigurable Automation Carts (RACs) (right).

Enzyme Discovery

The right starting point is crucial. By finding a natural enzyme with the desired activity, or close to it, we can avoid expensive and long iterative campaigns. This is especially valuable for developing enzymes for new-to-nature reactions or robustness in challenging industrial environments, both common requests. We invested in using natural protein diversity not just as training data for design models, but also as sources of variants for direct screening.

Ginkgo's enzyme discovery capability integrates a large diverse genetic library with computational tools to search it and identify high-confidence candidates for screening. The library has close to six billion genes enriched for extreme and underexplored environments, ~60% of which are proprietary to Ginkgo. We first "search" the library for candidates homologous to human-selected

seeds. Next, we filter based on sequence (e.g. sequence-similarity networks built from homologs identified with hidden markov models), structure, and physics (e.g. molecular docking and molecular dynamics, which often reveal non-obvious sites for mutation) to rank candidates by predicted function. Finally, we further rank promising candidates using a collection of tools that predict biophysical properties (e.g. structure, thermostability, solubility, expression, post-translational modifications) and activity (e.g., binding affinity and catalytic parameters), and codon-optimize the selected sequences with host-specific usage tables.

This pipeline lets us consistently meet client goals with smaller library sizes. Delivering 1-10 positive hits now requires screening tens to hundreds, rather than thousands, of candidates (see Figure 3).

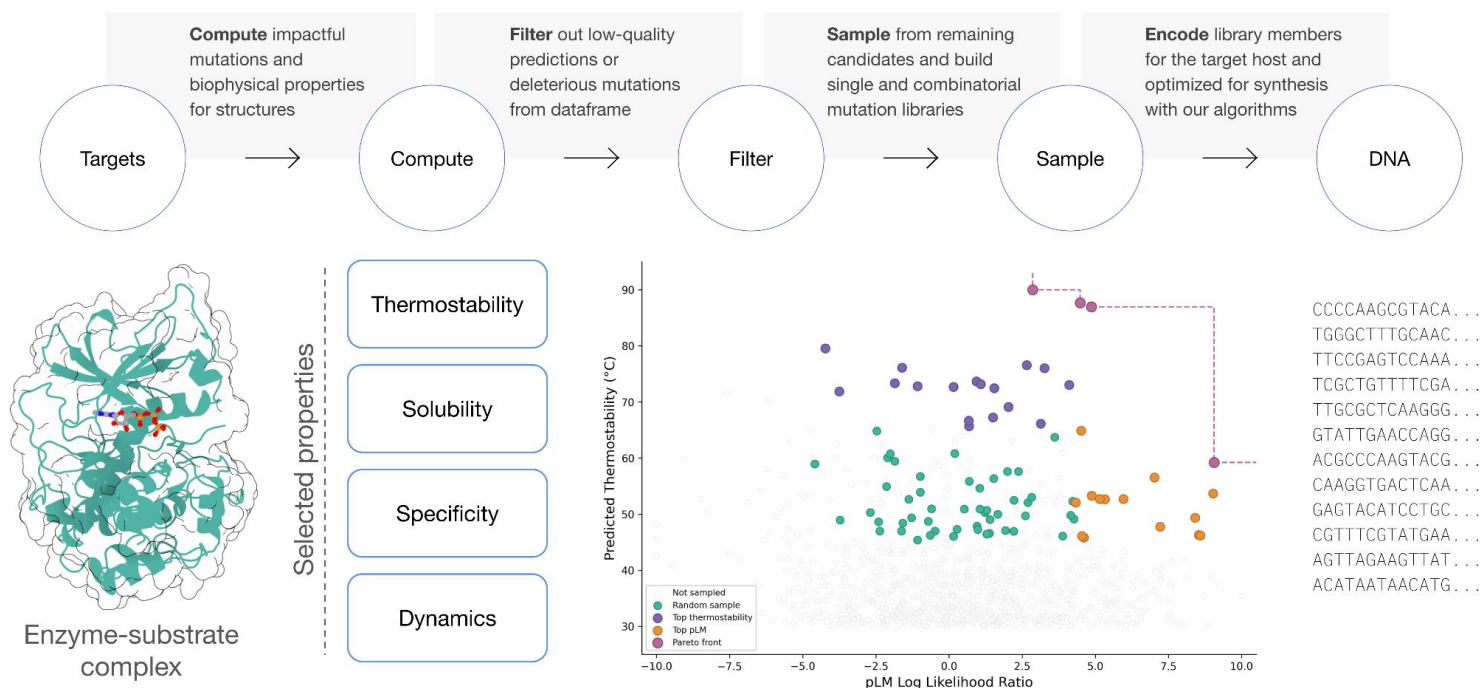


Figure 2: An automated pipeline for protein discovery and engineering. This pipeline takes in protein sequence or structure as inputs, generates predictions on biophysical properties and runs physics-based simulations. The data is automatically filtered and downsampled. For example, if trying to maximize protein validity and thermostability, a pareto sample can be applied to capture variants that fit these parameters. Protein experts analyze and sample the best variants predicted to reach enzyme function objectives.

Results of Enzyme Discovery campaigns with Diverse Objectives

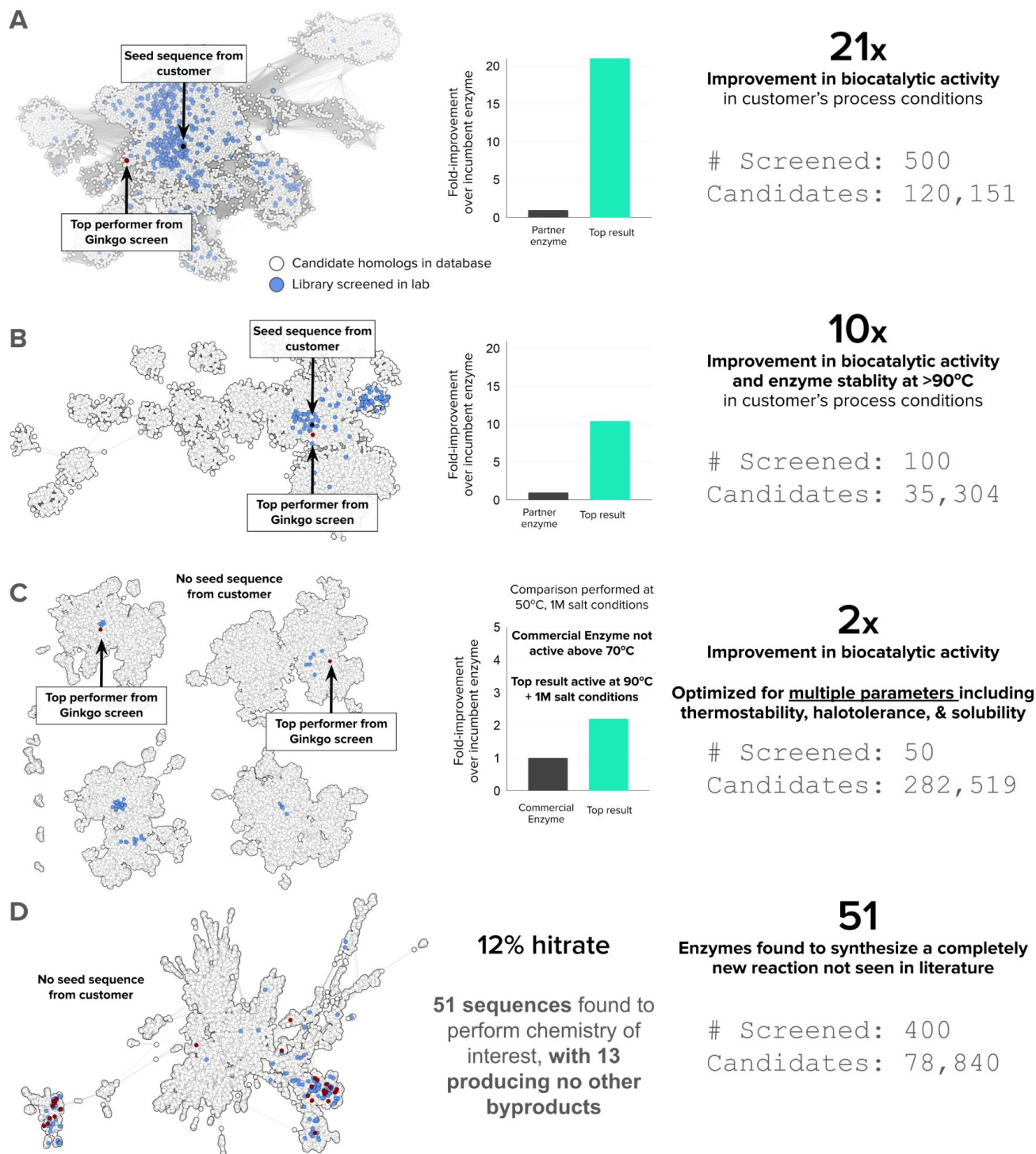
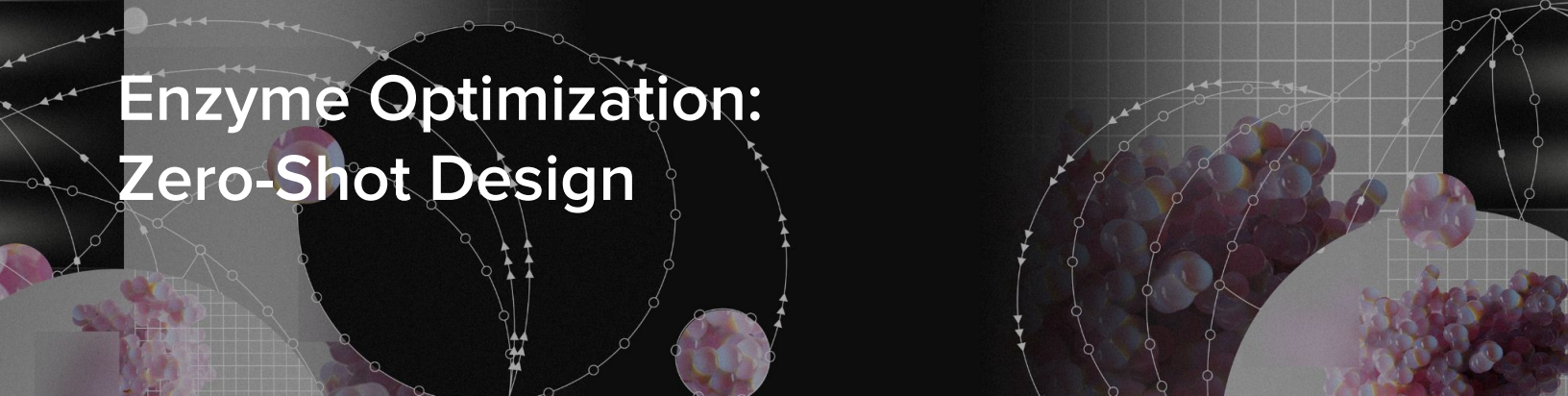


Figure 3: Four programs with small- and medium-sized libraries, focused on discovering enzymes with improved activity (a), activity and thermostability (b), activity, thermostability, salt tolerance, and solubility (c), and finally discovering an enzyme to catalyze a previously unseen chemical reaction (d).



Enzyme Optimization: Zero-Shot Design

Once we have a starting point, we can optimize for client targets. For an enzyme, this is typically a combination of productivity, specificity, thermostability, expression, and/or solubility. Because sequence-structure and sequence-function relationships in enzymes are complex, it is difficult to serially optimize enzymes for one trait at a time. Recognizing the interdependence of all these parameters, particularly with enzymes, we built our design toolset to simultaneously optimize all desired parameters together in each round of engineering.

The first round, without a detailed mutational dataset, is called a “zero-shot design”. To generate diversity, we combine whole-gene and site-directed approaches. AI models, including Protein Language Models (PLMs) which have learned the “grammar” of protein sequence, filter mutations that are computationally predicted to be deleterious or non-viable. Akin to large language models (LLMs), PLMs tokenize protein sequences, generating embeddings, or mathematical representations of protein sequences, that allow us to capture insights into sequence, structure, and function, and infer the effects of mutations.

Because mutating even just three positions across a protein sequence expands the combinatorial space into the millions, we rank candidates for screening, similar to the approach described for Protein Discovery, integrating AI models to annotate sequence and structure; and physics-based approaches including molecular modeling, molecular dynamics, and quantum chemistry to predict the effect of mutations on protein-substrate interactions and stability. We find this combination

of AI with mechanistic models is essential for successful multi-parameter optimization.

A greater challenge for AI models is predicting biophysical properties like thermostability and solubility (termed “protein quality”). Here, data on the environment that the underlying sequences come from help us predict useful mutations and features - e.g., we know proteins from hot environments are more thermostable, so comparing their “grammar” to similar proteins from neutral environments with PLMs tells us what makes a protein thermostable.

However, the true frontier is enzyme kinetics. Transition-state stabilization and other aspects of catalysis are inherently dynamic, and cannot be predicted from sequence or structure alone. This is further complicated because kinetics and reaction mechanism data need to be context-specific to be useful for prediction, and are not broadly available other than for a few well-studied enzymes.

To design improved enzymes in this context, [AI models can sometimes identify marginal activity improvements through stabilizing mutations](#), but not the order-of-magnitude improvements across parameters typically needed to deliver commercial enzymes. Therefore, we supplement AI predictions with a suite of mechanistic models that allows us to move beyond obvious active-site residues and identify critical secondary and tertiary shell positions that subtly and allosterically influence catalysis. Combining AI, physical, and biochemical modelling consistently improves library quality and hit sizes, bringing us closer to predictive design (Figure 4).

Combining Physics with ML to Select Positions for Mutation Increases Hit Rate

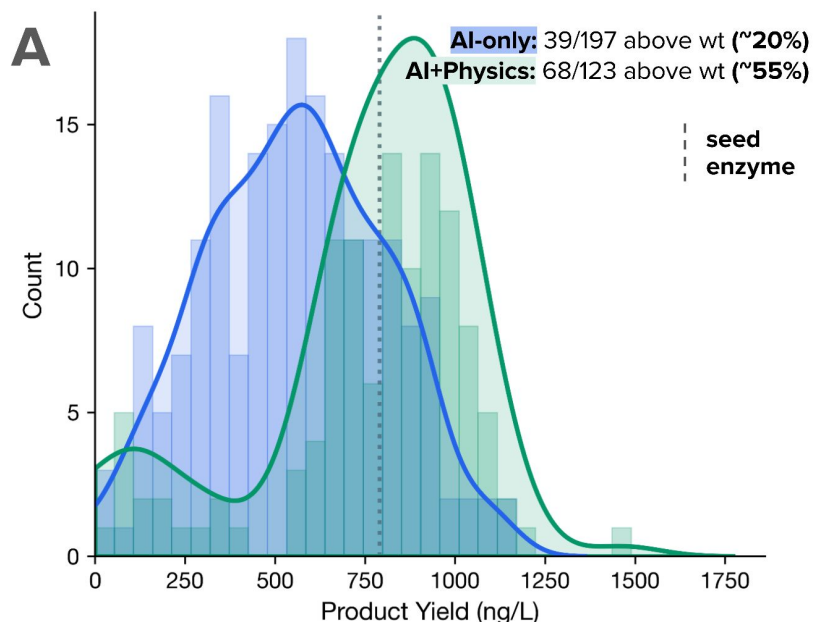
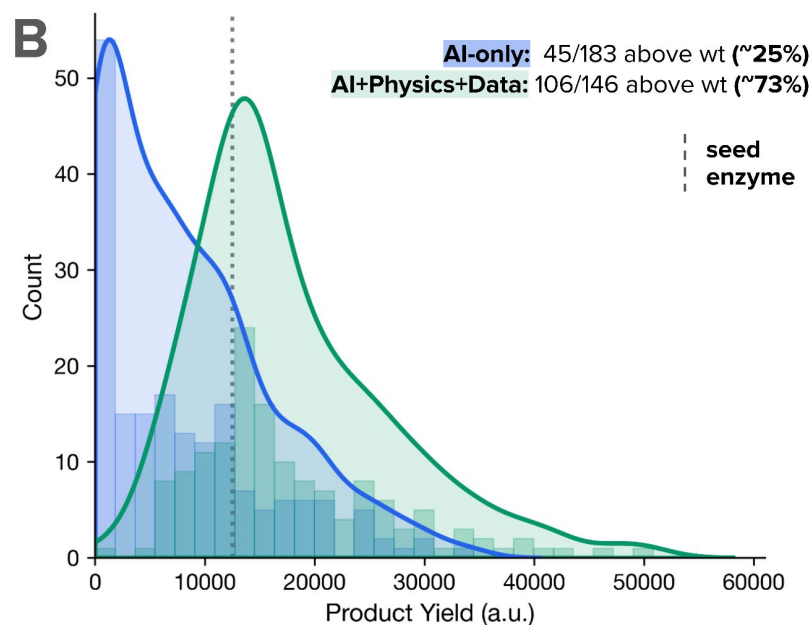


Figure 4. Enzymes predicted with AI express well enough to be measured for activity, but the models alone compared to the wild-type or seed enzyme only have a small percentage which improve over wild-type. Adding information from molecular dynamics simulations, data from protein electrostatics, as well as predictions from ML consistently drive our hit-rate higher in terms of product yield, up to 3-fold higher, even with fewer candidates sampled.



Enzyme Optimization: Iterative ML-guided Design

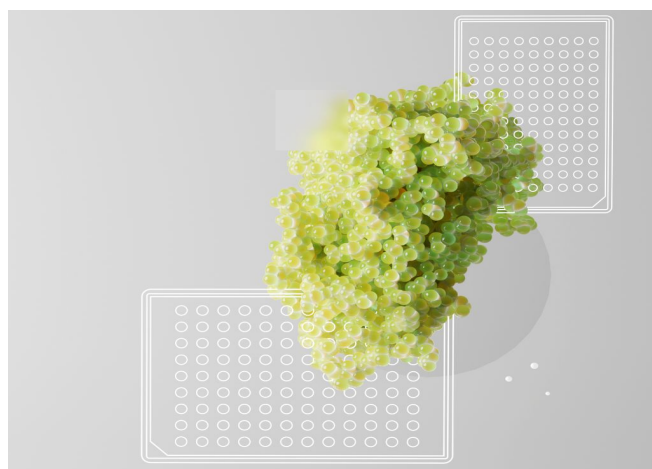
Once we have assay data, we join it to data on physical and chemical properties, structural information, and protein-language embeddings of sequences. To enable comparison across experiments, we normalize the dataset to correct issues like plate or batch variation, again leveraging our platform for automated, parameterized workflows. We also flag and label outliers and failed measurements (e.g. variants that failed to express). This “feature-engineered” dataset serves as the foundation for building predictive ML models for the next round of design.

This data scale, typically hundreds to thousands of samples per round, is where our strategy diverges from traditional AI companies. AI models excel at generalization but require millions of data points, function as a “black box”, and [can struggle to optimize complex features like enzyme kinetics](#). In this domain, specificity beats generalization. We employ supervised models like gradient-boosted trees, random forests, and simple neural networks, which excel at identifying subtle patterns in focused datasets, are robust to the variability inherent in biological datasets, and have defined features and parameters allowing explanation of their predictions.

Training these models on our curated, focused datasets allows us to accurately predict and, crucially, explain, mutations that will yield further improvements. These ML models predict new beneficial combinations of (previously-seen) mutations based on experimental data; we also

supplement these variants by exploring previously untouched mutational space, re-running “Zero shot”-like optimization campaigns as described above on high-performing variants. This allows us to find pockets of epistatic space that are specific to these variants, not just the wild-type sequence, producing new data used to further refine our ML models as we progressively diverge from the natural sequence.

Our enzyme design philosophy is that of a toolkit. No single approach works for all tasks. We use generalized AI models to predict protein validity and quality. Specialized ML models and physics-based tools predict epistasis, specificity, and activity. Integrating this array of tools into a common, automated framework lets us choose the right tool for any job and design proteins with higher probabilities of success, faster R&D times, with lower library sizes and thus lower DNA and assay costs.



Results of Enzyme Optimization Campaigns with Diverse Objectives

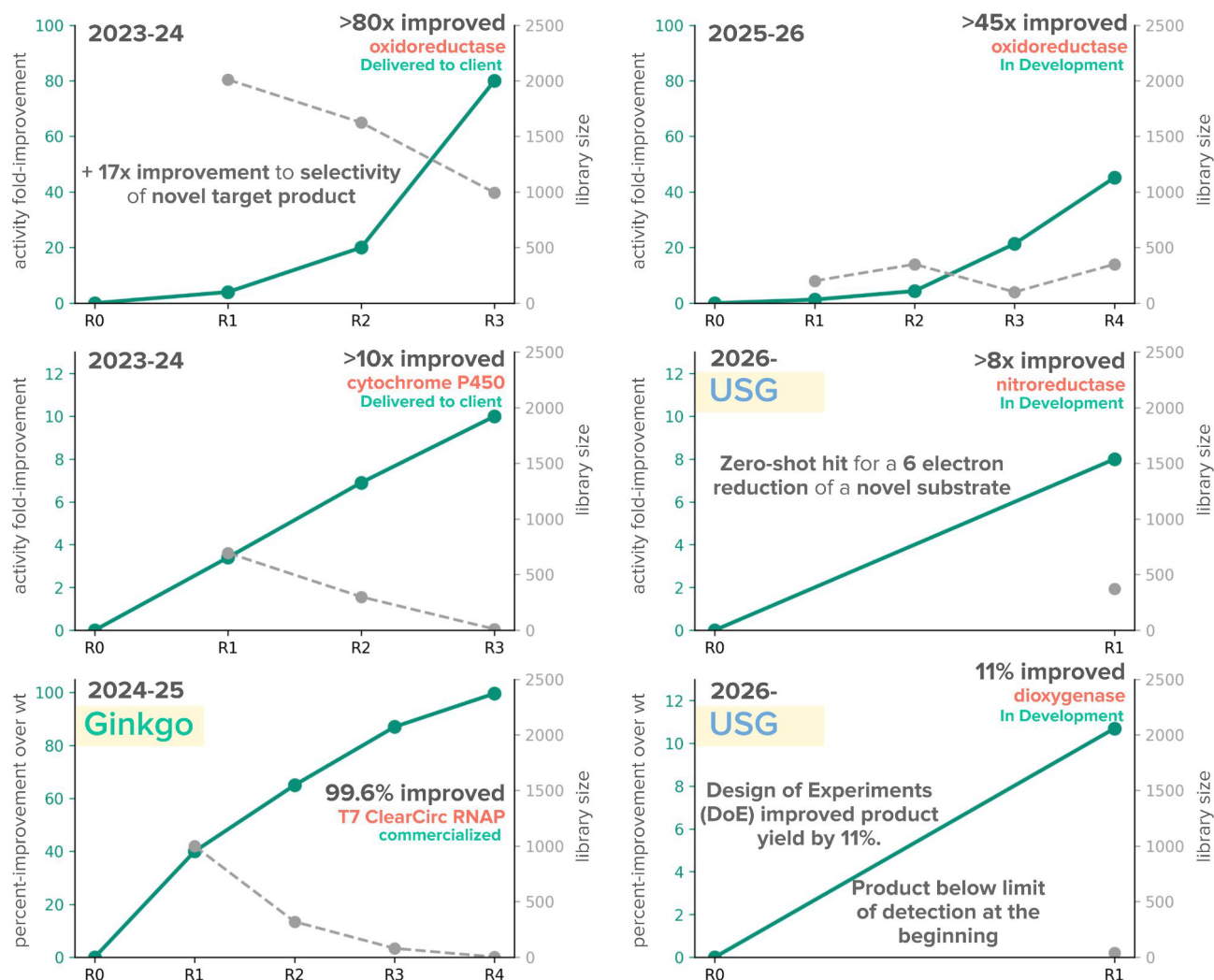


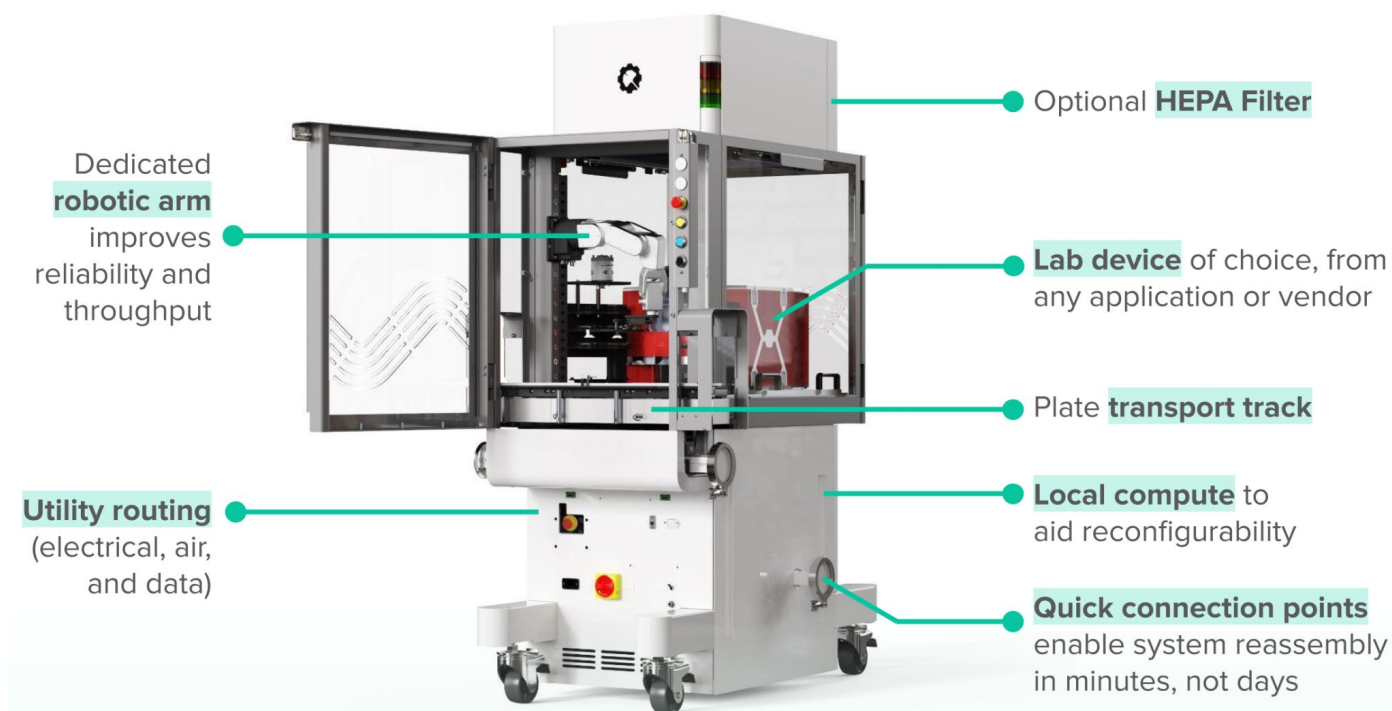
Figure 5: ML-guided protein design. In more recent campaigns, we maintain high fold-improvements per cycle with significantly smaller libraries. The 99.6% improved RNA Polymerase was Ginkgo Bioworks' first commercial enzyme product, which reduces the production of dsRNA which causes immunogenic effects. The 8x improved nitroreductase and 11% improved dioxygenase are enzymes currently in development for USG partners and demonstrate additional unique capabilities from powerful zero-shot prowess in spite of complex enzyme mechanisms, to high-throughput assay optimization aided by automation, leading to cumulatively improved downstream engineering.

Lab Automation: Reconfigurable Automation Carts

Training ML models to design successful libraries, with fast delivery, requires consistent, high-quality datasets. Biological data is typically highly variable and expensive; we believe rigorous automation that nevertheless retains the flexibility needed to address a diversity of applications is required to deliver these datasets at the scale needed to broadly commercialize enzymes.

Automation like this did not exist for biological data generation, so we built it at Ginkgo. Reconfigurable Automation Carts (RACs) and Catalyst software are our enabling technologies for automated, flexible, and reproducible data generation. They are the result of almost twenty years of work on automating biology R&D at Ginkgo and Zymergen.

Reconfigurable Automation Carts: The machines. Modular RACs allow a flexible automated laboratory to be built from disparate pieces of specialized equipment unified by a transport system. Akin to containerized software tools, each RAC encloses one or more laboratory instruments and provides communication, observability, and transport. A dedicated six-axis industrial robotic arm moves samples within each enclosure and to/from carriers that shuttle them to any other RAC in the system. Systems of multiple RACs compile many capabilities and redundancy permitting execution of complex workflows; e.g. liquid handlers, incubators, centrifuges, and analytical equipment. RACs are connected through standardized interfaces permitting rapid setup, reconfiguration, and data sharing.



Lab Automation: Catalyst Software

Automation Control Software: The brains.

[Catalyst ACS](#) (Automation Control Software) integrates protocol definition, execution, and device and data management with a common interface. Its modular architecture matches the flexibility of the physical systems that it controls. Observability of each instrument provides real-time visibility into the actions and health of each subsystem, and integrates all data into a unified structure.

Workflows are dynamically scheduled to interleave protocols providing efficient utilization of equipment. New protocol runs can be added to the queue at any time and are worked into the existing schedule. A modern interface allows users of all experience levels to build, execute, and monitor their workflows. Catalyst's "digital twin" of the physical system facilitates onboarding by providing a no-risk sandbox to simulate, troubleshoot, and plan workflows without disrupting ongoing work.



Figure 6: Ginkgo's Catalyst Software lets users build, launch, and monitor complex protocols. Protocols are built from operations (middle), which are composed of lower-level recipes (top) providing control at all levels of abstraction from individual devices to system. Launch forms collect payload and parameter information to customize each run. In-progress and scheduled runs are monitored with a real-time web-based UI (bottom) with each line representing a distinct protocol and the colored bars representing time on distinct devices.

Value of Flexible, Integrated Laboratory Automation

Flexible, integrated automation frees scientists to focus on biological hypotheses. The system is not limited to a single workflow and the scientist can run any experiment for which the instruments are available, accommodating the typical changes from one experiment to the next in R&D, like switching enzyme expression systems or purification methods. Overhead of sharing equipment between workflows and teams is reduced.

Automatic scheduling and transport ensures that operations are completed as planned. The consolidated equipment population minimizes laboratory space requirements while making it easier to monitor equipment health and perform maintenance. Batch sizes are no longer limited by scientists' ability to organize physical plates and

equipment. Iteration time shrinks as equipment is utilized around the clock.

In the rare case where a scientist needs to isolate a device for testing or method development, the modular architecture allows temporarily disconnecting it to be used in "walk-up mode."

We call Ginkgo's installation of this automation infrastructure "Nebula" (Figure 7). Nebula can run specific protocols repeatedly, similar to existing "workcell" systems. But increasingly, we use it to run a large variety of protocols from different scientists and projects. In sharp contrast to purpose-built workcells, the system can work in both modes simultaneously.

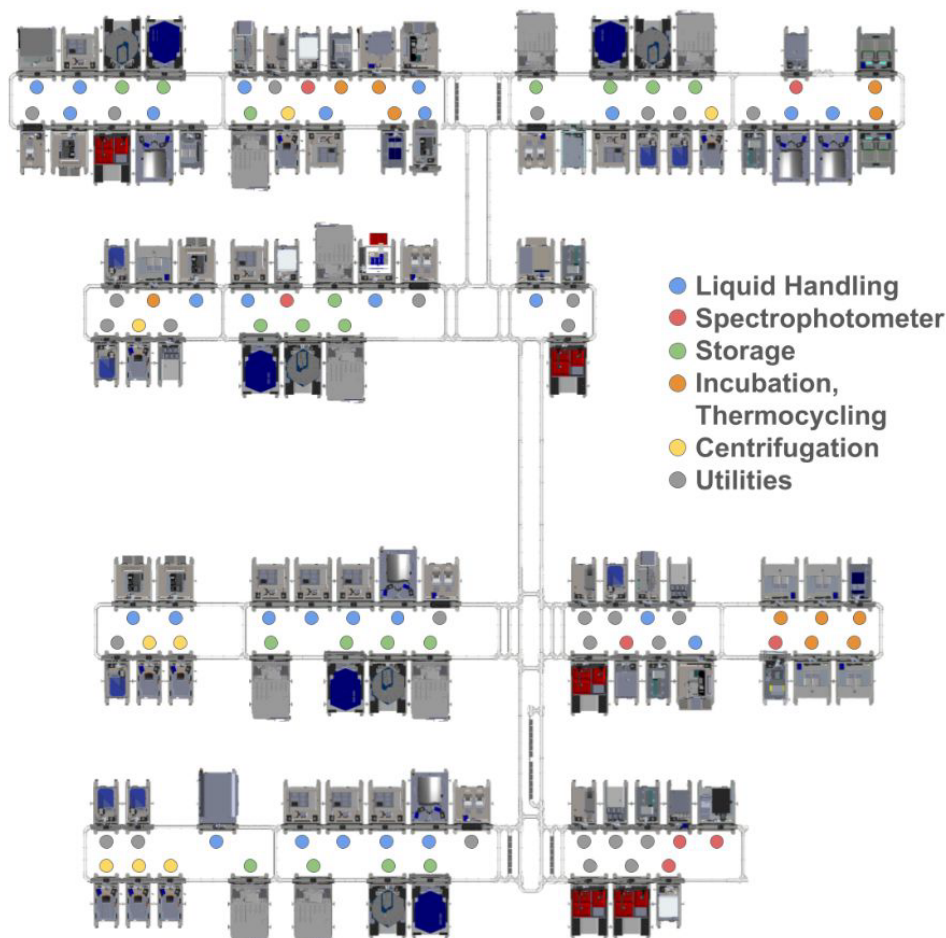
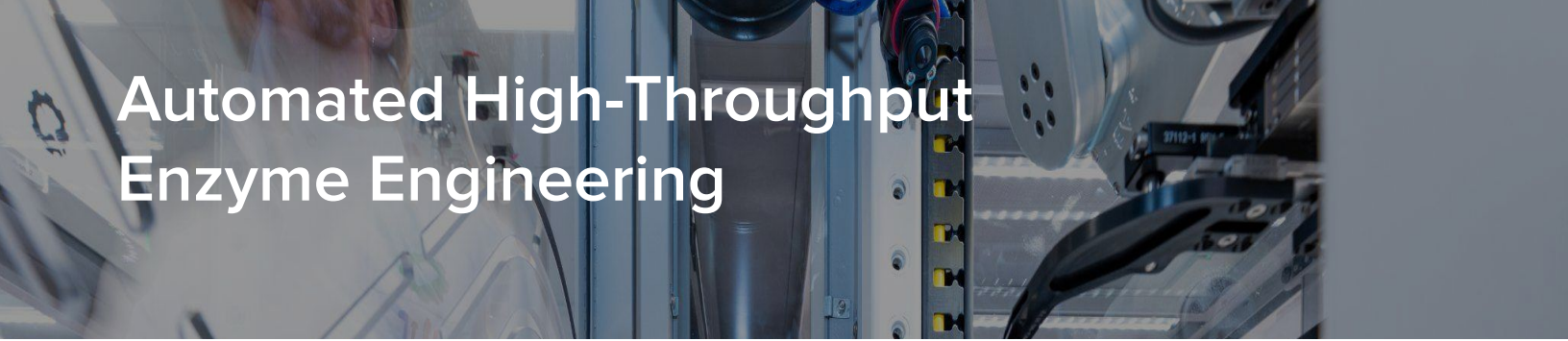


Figure 7: Nebula, Ginkgo's 100+ RAC system. Full integration of each device with RACs means that scientists can run workflows involving any sequence or combination of devices. Carriers move samples along tracks to any device on the system. Key functional categories include liquid handling tools covering nL-mL, spectrophotometers, temperature controlled storage, incubation, centrifugation, along with utilities including plate sealers/peelers, labelers, and waste.



Automated High-Throughput Enzyme Engineering

In 2025, Nebula became mature enough to begin to replace traditional “walk-up” automation and protocol-specific workcells. Here, we highlight key enzyme engineering protocols on Nebula that enable higher degrees of automation and ultimately autonomous enzyme development.

DNA synthesis. We use multiple DNA synthesis routes for flexibility in library design. Beyond working with multiple commercial suppliers, we now run automated workflows leveraging “dial-out PCR” to assemble genes from pooled oligonucleotides on Nebula for particularly challenging constructs or where [faster production is useful](#).

Enzyme expression. To accelerate enzyme development, we increasingly use cell-free protein synthesis (CFPS) to express libraries without assembling plasmids or transforming strains. These features also make CFPS much easier to automate than in vivo expression. *E. coli*-based CFPS is a workhorse, and we have optimized our extract and expression conditions to enable expressing a [broad range of proteins](#). We are also developing wheat germ-based CFPS for proteins with glycosylation and other post-translational [modifications](#).

Producing enough protein to assay activity is also a key challenge. We tasked an LLM with improving protein yield from CFPS reactions. The LLM designed experiments, analyzed data, and generated new hypotheses; Nebula executed experiments without human intervention other than producing and loading reagents like cell-free extract. The “closed-loop autonomous lab” increased protein production (g/L titer) by 27% relative to [state-of-the-art](#).

For proteins that don’t express well in CFPS, we leverage Nebula’s flexibility to support in vivo expression. Delivering engineered DNA into cells has traditionally been the biggest barrier to automating in vivo expression, with most protocols giving low transformation rates that impede high-

throughput work. We automated yeast ([S. cerevisiae](#)) and mammalian ([human Expi293F](#)) transformations by systematically optimizing protocol parameters for the many sensitive operations involved.

Purification. Purifying enzymes before testing significantly increases data quality by reducing background noise and cross-reactivity. We believe this is especially critical in the ML era. But purification is often skipped because it is expensive and difficult to automate for high-throughput work.

We focused on magnetic bead-based purification because it is broadly applicable (e.g. for different affinity tags) and well-suited for automation. Nebula runs a [bead-based purification protocol](#) that delivered 60-90% pure material for 96 ~10 kDa His-tagged proteins from *E. coli* lysate. It is fully parameterized to provide flexibility across proteins, affinity tags, or expression systems (CFPS extracts, cell lysate from intracellular in vivo production, or fermentation broth for secreted in vivo production).

Activity assay and characterization. Nebula can run most spectrophotometric assays today. For analytical chemistry like chromatography and mass spectrometry, we currently prepare samples on Nebula and run them on autosampled systems offline, with full integration in progress.

Measuring protein yield enables determining specific activity and thus distinguishing low enzyme activity from lack of expression, but is often omitted for cost and logistics. We standardized on the HiBiT tag for high-throughput quantification, and Nebula runs batches of up to 10,000 samples [today](#).

Outlook for Autonomous Enzyme Engineering in Government and Commercial Programs

The pace of scientific discovery is limited by the complexities of experimental execution. Advanced lab automation is fundamentally decoupling researchers from these operational burdens. By deploying highly flexible, automated platforms across the entire Design-Build-Test-Learn (DBTL) cycle, from initial gene synthesis to final protein purification, we can shatter the traditional tradeoff between speed, cost, and quality.

This evolution toward fully [autonomous laboratories](#) will empower every scientist to act as a principal investigator, freely guiding experimental direction while automated systems reliably handle end-to-end execution. Ultimately, Ginkgo’s new standard for rapid, 24/7 workflows and ultra-precise biological design will drastically reduce required library sizes,

seamlessly bridging the final gap between a revolutionary idea and a tangible, manufactured product. This trajectory aligns directly with current U.S. Government priorities.

Recent defense authorizations direct the expansion of domestic biomanufacturing capacity for critical products and the development of biotechnologies for supply-chain resilience;² dedicated public-private infrastructure is building demonstration-scale biomanufacturing capacity across the country;³ and the reauthorized Small Business Innovation Research and Small Business Technology Transfer programs now offer scale-up awards designed to carry technologies across the “valley of death” toward production and procurement.⁴

Near-term roadmap for further automation of key workflows in automated high-throughput enzyme engineering	
Workflow	2027–2028 roadmap
Library design	Enable experimentalists to onboard and run new pipelines, enabled by tools like Claude Code. Automatically generate simulation data and turn the experimentalist into a computational designer.
Protein production	Autonomous, sequence-driven selection of optimal expression system per protein. Closed-loop optimization of expression conditions with no manual input.
Purification	Generalized purification across all expression systems (CFPS, in vivo, secreted) and affinity tags. AI-driven protocol selection per protein class. Yield data fed into closed-loop expression optimization.
Assay/test	Generalized sample prep, fully integrate complex analytics (e.g. Echo MS)

Ginkgo is poised to develop and transition technology through a variety of federal R&D programs including, as a small business under most federal size standards, through SBIR/STTR programs. Ginkgo provides a domestic platform that delivers engineered enzymes for government missions and equips the domestic manufacturers who turn that science into products, whether the end customer is a federal agency or a commercial market.

² FY2026 National Defense Authorization Act, biotechnology and biomanufacturing provisions (e.g., §§ 242–248), authorizing the Department of Defense to expand domestic biomanufacturing capacity for critical products and to develop and scale biotechnologies for defense supply-chain resilience; see also Defense Production Act Title III industrial-base authorities.

³ [BioMADE \(Bioindustrial Manufacturing and Design Ecosystem\)](#), the Department of Defense–sponsored Manufacturing Innovation Institute for domestic bioindustrial scale-up and commercialization—e.g., its 2025 investment in a demonstration-scale biomanufacturing facility.

⁴ Small Business Innovation and Economic Security Act (2026), reauthorizing the SBIR and STTR programs through FY2031 and adding Strategic Breakthrough (Phase II) awards to bridge the “valley of death” toward production, procurement, and commercialization.



We thank Jama Hagi-Yusuf, Peter Punt, Aaron Berman, Reshma Shetty, and Jason Kelly for valuable feedback. We additionally thank and acknowledge the work of the Solutions, Protein Engineering, and Automation teams at Ginkgo for their instrumental role in developing this platform.