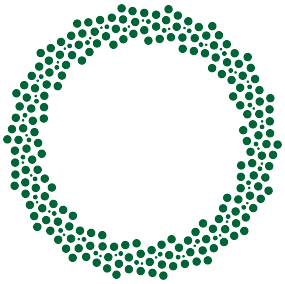
Open Access



Nucleus Developer Notes

Wrangled, Grounded, Constrained, & Mechanistic

What it took to successfully deploy an AI scientist in compositional biology

Maxwell Shapiro[1], **Margarita Khariton**[2], **Joseph Lozier**[1], **Ash Eldritch**[1], **Anton Jackson-Smith**[2], and **Scott C. Riggs**[1]

[1]Find What Matters, Inc., [2]b.next

Abstract

This Developer Note reports the third and final deliverable of our ARIA AI Scientist program, in which the pyruvate-acetate pathway (PAP) was deliberately chosen over more complex energy modules; with fewer components, it allowed the experimental loop run at an accelerated pace to test our AI co-scientist. We report on the experimental performance of IGOR (Iterative Guide and Orchestrated Research), an agentic AI co-scientist with a Bayesian Optimization backend, in a synthetic biology setting. Over five experimental rounds, we demonstrate how IGOR supported integration of PAP into a Cytosol cell-free protein synthesis system by offering reliable scientific insight throughout the experimental workflow, a feature that is often lacking when using purely large language model (LLM)-backed co-scientists. Across the rounds where a like-for-like comparison was possible, IGOR's designs improved yield by 12–56% over the standard PAP recipe, resolved four interacting small-molecule energy inputs into a single stoichiometric ridge, and identified the next rate-limiting factor for experimentation. The observed reliability was an emergent property of an empirically-grounded system equipped to strategically solicit scientific judgment from an engaged expert-in-the-loop.

Keywords AI co-scientist, Bayesian optimization, Cell-free protein synthesis, Cytosol, IGOR, Pyruvate-acetate pathway

Published Oct 26, 2026

Correspondence to
Scott C. Riggs
scr@findwhatmatters.ai

Open Access

Copyright © 2026 Shapiro *et al.*. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator.

Open Source

The materials and processes described in this article are distributed under the terms of the CERN Open Hardware License Permissive 2.0, which enables reusers to make and convey any component or device that apply the information made explicitly available under this license, so long as all notices are retained.

> **Note**
>
> This article was written by a **human author**.

1. Introduction

AI co-scientists are multi-agent systems in which LLM-backed agents search the literature, generate hypotheses, substantiate their hypotheses with scientific reasoning, and iteratively refine their proposals based upon receiving new information. Such systems promise to compress the time between a research question and an experimentally testable answer. Google's Co-Scientist [J. Gottweis *et al.* [1]], for example, orchestrates specialized Generation, Reflection, Ranking, Evolution, Proximity, and Meta-review agents in a tournament-style "generate, debate, evolve" loop, producing hypotheses that were later supported by laboratory experiments in drug repurposing for acute myeloid leukemia and epigenetic target discovery for liver fibrosis. However, novel hypothesis generation addresses only a small portion of the laboratory workflow. Laboratory researchers generally have a surplus of well-motivated hypotheses emerging from their scientific domain expertise, but what usually consumes their time and budgets is the process of setting up and executing their experiments (including

overcoming emergent experimental issues), analyzing and interpreting their results, and iterating until they achieve their scientific goals. What this looks like concretely in the lab: calibrating an instrument, identifying a formulation of biological reagents that yields a target protein, or diagnosing and rectifying sources of experimental variation.

How can an AI co-scientist help with overcoming these experimental challenges? A hint comes from observing what they have in common: 1) they all have quantifiable objectives (e.g., tuning instrument settings to match a reference standard, maximizing protein fluorescence, minimizing measurement noise, etc.); 2) they all require making and executing many interacting decisions (e.g., selecting material precursors, temperature profiles, reagent concentrations, time durations, and instrument settings) in order to achieve those quantifiable objectives; and 3) the size of the decision space (exponentially large in the number of decisions) and the per-experiment cost (large in time, money, and resources) preclude the viability of both trial-and-error and exhaustive search. These three properties define the problem class for which Bayesian optimization (a sequential model-based strategy for global optimization of black-box objective functions) is designed: finding the optimum of an objective that has no closed-form expression, can only be observed by running an (expensive) experiment, and is convolved with observation noise also of an unknown form. An AI co-scientist that incorporates Bayesian optimization reasoning into its generative process would thereby inherit its main benefit as the most experimentally efficient strategy for global optimization [E. Brochu, V. M. Cora, and N. de Freitas [2]].

Beyond sample efficiency, a Bayesian-optimization-backed AI co-scientist also enables reliability in situations where purely LLM-driven co-scientists are known to struggle. The outputs of LLM-driven co-scientists:

- inherit the weaknesses of their underlying models, including imperfect factuality and hallucination [A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang [3]];
- are subjected to the limitations of the literature they retrieve, including systematic underrepresentation of null results, contaminated reasoning from the presence of erroneous or irreproducible findings, and prevailing assumptions / blind spots within their corpus;
- tend to propagate their own internal biases because, quite often, other LLMs are used as evaluative filters on generative output (based upon characteristics like novelty and utility) [J. Gottweis *et al.* [1]]; and
- systematically deviate from experimental results when tested in the physical world [Rabanser *et al.* 2026].

All of these challenges contribute to unreliable output as judged by users; in one survey conducted by Anthropic of 80,508 people across 159 countries [Huang *et al.* 2026], unreliability (hallucinations, inaccuracies, and the resulting verification burden) was the most frequently cited concern, with greater concern expressed by people in higher-stakes professions. Rectifying these reliability concerns is nevertheless possible through access to a Bayesian optimization backend, where the corrective feedback ultimately derives from a quantitative model trained upon iteratively accumulated laboratory evidence that keeps the AI co-scientist grounded in physical reality.

In this work we experimentally deploy such an AI co-scientist to optimize a cell-free protein synthesis (CFPS) process. Using the Iterative Guide and Orchestrated Research (IGOR) platform from Find What Matters, we optimize the formulation for integrating a pyruvate-acetate pathway (PAP) into a cell-free system. The energy cost of sustaining the bioreaction is a major limiting factor in using Cytosol for CFPS, but some of this energy can be provided through the regeneration of ATP via PAP. Such a parallel pathway has been shown to increase protein yield by $\sim$77% when run in parallel with creatine phosphate in the PURE system [S. Yadav, A. J. P. Perkins, S. B. W. Liyanagedera, A. Bougas, and N. Laohakunakorn [4]]. Porting PAP

into Cytosol is both a route to energetically sustaining CFPS and a concrete high-dimensional design problem (given the many formulation components, mixing steps, and experimental settings involved) that is well suited to IGOR.

2. Methods

2.0.1. *Experimental Setup*

The full list of components used for Cytosol reactions is enumerated in Table 1. Reagents were assembled as either master mix solutions or individual stock reagents for dispensing via LabCraft [LabCraft 2026] automated liquid handler into a small-volume, black 384-well plate for multiplexed 2 µL reactions. Six components were dispensed independently for feature variation. Each component was dispensed via LabCraft at discrete droplets of volume 8–12 nL or 48–87 nL, depending on reagent and dispense mode (CV < 1% per droplet volume measurement). All reactions contained DNA plasmid with mScarlet as the reporter protein for Cytosol transcription/translation performance; the standard deGFP reporter recommended for Cytosol reactions cannot be used here as the flavin adenine dinucleotide (FAD) component of PAP reactions fluoresces at Ex/Em 450/515 nm.

Iterations of experimental rounds and their associated rationales are described in Table 2. Initial conditions (Round 0) for Cytosol reactions with and without added PAP-pathway components were performed to establish a baseline for the IGOR agent and model inputs. These were determined based on standard Cytosol reaction conditions [Nucleus, Assemble Base Cytosol 2026] and PAP protein and cofactor concentrations determined previously [S. Yadav, A. J. P. Perkins, S. B. W. Liyanagedera, A. Bougas, and N. Laohakunakorn [4]]. PAP reagents were added directly on top of base Cytosol with the exception of magnesium acetate (MgOAc) substitution with magnesium glutamate (MgGlu) in the small molecule solution (SMix); creatine kinase (CK) protein was present in the PAP system but this pathway remains inactive in conditions omitting creatine phosphate (CP).

Cytosol reaction performance was measured via mScarlet fluorescence time course readings at 37 °C on a BioTek Cytation3 plate reader for up to 15 h at 5 min intervals (Ex/Em 565/606 nm), with a 5 sec orbital shake at each interval. Each well was normalized against the no-template control to a LabCraft-dispensed 0–4 µM sulforhodamine 101 ladder and kinetic analysis performed with the Nucleus Cell Development Kit (CDK) pipeline [Nucleus, CDK 2026]; [Newman 2026]. The estimated parameters from nonlinear curve fits to fluorescence time course readings were used as supervised labels for training IGOR's surrogate model.

Table 1. Reaction conditions and tested ranges for PAP in Cytosol

Component	Reaction final concentration	Dispense mode
Cytosol small molecules (SMix)	1x	In master mix
Thiamine pyrophosphate (TPP)	2 mM	In master mix
Flavin adenine dinucleotide (FAD)	0.2 mM	In master mix
Creatine phosphate (CP)	SOP: 20 mM; Tested: 0–42 mM	In master mix and/or independent
Potassium phosphate buffer	SOP: 27.5 mM; Tested: 0–55 mM	Independent
Sodium pyruvate	SOP: 27.5 mM; Tested: 0–54 mM	Independent

Component	Reaction final concentration	Dispense mode
Magnesium glutamate	SOP: 12.5 mM; Tested: 4–25 mM	In master mix (4 mM floor) and independent
Cytosol proteins (PMix)	SOP: 1.8 µg/µL; Tested: 1.44–2.16 µg/µL (0.8x–1.2x)	In protein master mix
Pyruvate oxidase (Pox5)	SOP: 3.03 µM; Tested: 0–3.64 µM (2.42–3.64 µM in PAP)	In protein master mix
Acetate kinase (AckA)	SOP: 13.89 µM; Tested: 0–16.67 µM (11.11–16.67 µM in PAP)	In protein master mix
Catalase (KatE)	SOP: 2.38 µM; Tested: 0–2.86 µM (1.90–2.86 µM in PAP)	In protein master mix
Cytosol ribosomes	1.8 µM	In master mix
Cytosol tRNAs	3.5 µg/µL	In master mix
RNase inhibitor	SOP: 2000 U/mL; Tested: 0–1000 U/mL	In master mix
pOpen-mScarlet or pET21a-mScarlet DNA	SOP: 3–10 nM; Tested: 0–4 nM	Independent
Water	Remaining volume	In master mix and/or independent

Table 2. Experiment overview and rationale

Round	# Runs	Plexity	Unique conditions	Design basis	Features varied
R0	4	56, 48, 38, 54 (196 total)	12, 6, 4, 10 (32 total)	Initial sampling (literature) + method dev (expert)	CP/CK v. PAP v. combined pathway, protein addition order, DNA plasmid, [DNA], [MgGlu]
R1	1	84	27	Structured DOE + IGOR literature-guided	[CP], [MgGlu], [pyruvate], [Pi]
R2	1	56	10	IGOR literature/model-guided + model GES + expert + sample space anchors	[CP], [MgGlu], [pyruvate], [Pi]
R3	1	56	12	IGOR literature/model-guided + model GES iteration	[CP], [MgGlu], [pyruvate], [Pi], [DNA]
R4	1	60	16	IGOR literature/model-guided + model GES iteration + expert sub-studies	[CP], [MgGlu], [pyruvate], [Pi], [total protein]

2.0.2. *IGOR Setup*

IGOR populated its internal corpus of relevant literature with sources that were either autonomously retrieved or explicitly human-provided; the full list is included here [S. Berhanu, T. Ueda, and Y. Kuruma [5]]; [A. Doerr *et al.* [6]]; [R. Koivunen, R. Bollström, and P. Gane [7]]; [B. Lavickova and S. J. Maerkl [8]]; [J. Li, L. Gu, J. Aach, and G. M. Church [9]]; [S. Orbay, A. Ozcelik, J. Lata, M. Kaynak, M. Wu, and T. J. Huang [10]]; [A. Pandi *et al.* [11]]; [S. Yadav, A. J. P. Perkins, S. B. W. Liyanagedera, A. Bougas, and N. Laohakunakorn [4]]. Raw data and other supplementary information was provided to IGOR, and the platform's data wrangling capabilities were used to fully curate experimental data; an exportable CSV containing the curated data over all rounds can be found here, with the compiled data here. The laboratory scientist directly interfaced with IGOR, using the platform dynamically (including through data visualization, evaluation of quantitative analytics, and natural language chat) to plan experiments, troubleshoot ostensible problems with the experimental setup, and biologically interpret iterations of experimental results.

3. Results

Characterizing the performance of AI co-scientists is a known challenge; at present, no universally agreed-upon benchmarks exist [J. Gottweis *et al.* [1]]; [S. Kapoor *et al.* [12]]; [M. Mohammadi, Y. Li, J. Lo, and W. Yip [13]], and randomized control studies (involving parallel optimization campaigns orchestrated without AI) are cost prohibitive. In light of this, and also in lieu of an exhaustive chronology of all of IGOR's intermediate outputs, here we report on our distilled observations of IGOR's empirical performance over five rounds. Our aims are to clarify (via concrete examples) how IGOR supported experimental execution and to illuminate generalizable lessons for successful AI co-scientist deployment.

3.0.1. *Observation 1: IGOR generated scientific insight and facilitated quantitative interpretation*

- Example 1 – Magnesium-phosphate coupling: Both measured and modeled yield evinced non-monotonic dependence on magnesium glutamate, peaking for an intermediate value of around 13 mM with around 17 mM of creatine phosphate (and holding remaining components at the fixed reference conditions of `r1016_E17`, the highest yield reaction in Round 0). This phenomenology matched the scientist's mechanistic expectation of a chelation ridge, in which magnesium is sequestered by phosphorylated species at low concentrations and inhibits translation at high concentrations, leaving a productive window between the two regimes. IGOR's two-dimensional contour visualization Figure 1 displayed to the scientist that these mechanisms had been learned via data alone; being able to corroborate this scientifically valid behavior without explicit intervention in model training nor supplementary prompting increased the scientist's confidence in IGOR's other insights.

Two other features of the two-dimensional contour visualizations are also noteworthy. First, the principal axes of the elliptical contours are noticeably rotated relative to the magnesium glutamate-creatine phosphate axes; this is direct evidence of a nonlinear interaction between the two components (i.e., the optimal magnesium concentration depends on creatine phosphate, and vice versa), which is emblematic of a typical class of multivariate interdependency in scientific processes. Estimating the magnitudes of these interactions requires a large number of experiments (i.e., exponential in the number of input dimensions) under traditional designs of experiment (e.g., factorial or one-factor-at-a-time) but substantially fewer experiments under multivariate approaches [Bergstra and Bengio 2012]. Second, the predicted yield landscape changes noticeably depending upon the specific values that are

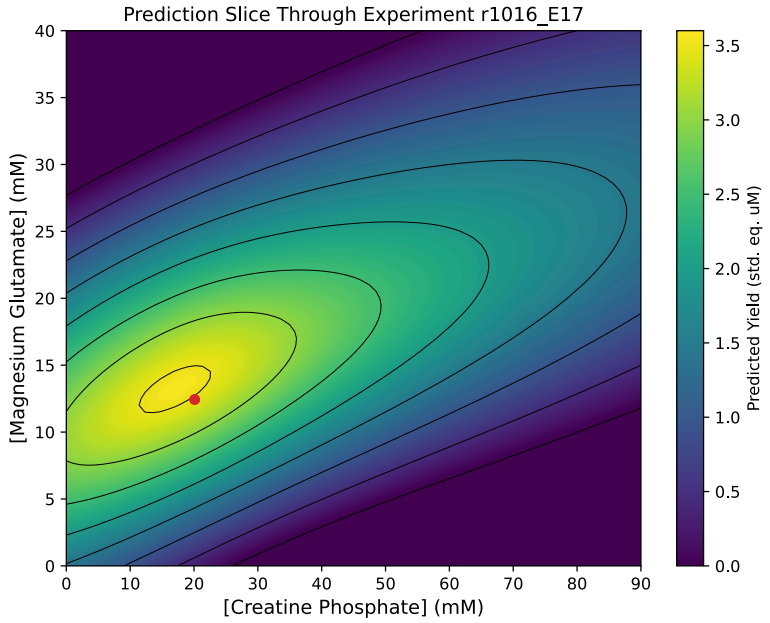


Figure 1. *Predicted yield as a function of magnesium glutamate and creatine phosphate (with other feature values fixed to those of experiment `r1016E17` (red dot), in the style of [S. Yadav, A. J. P. Perkins, S. B. W. Liyanagedera, A. Bougas, and N. Laohakunakorn [4]]). Model uncertainty is computed but not displayed.*

held fixed for the non-magnesium glutamate, non-creatine phosphate components Figure 2. This demonstrates how the magnesium-phosphate coupling itself interacts with the rest of

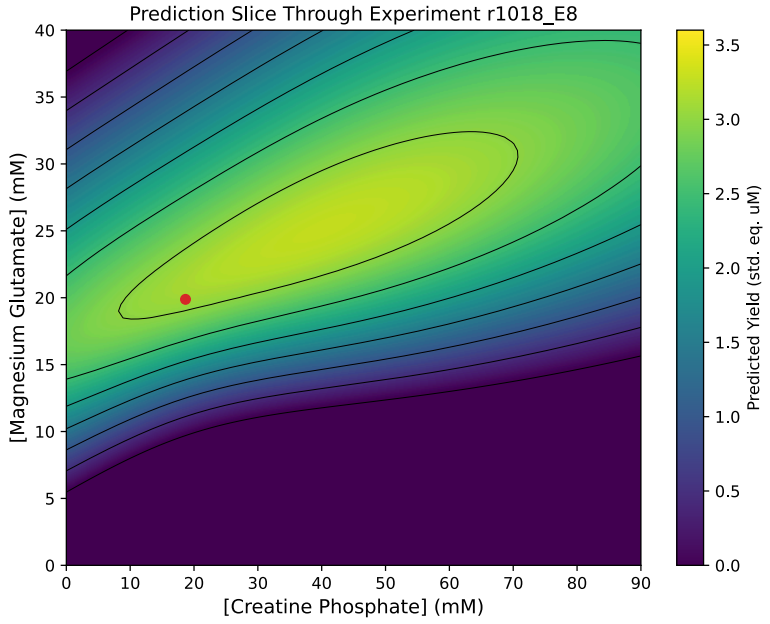


Figure 2. *As in Figure 1, but computed around experiment `r1018_E8`. Note the predicted changes in peak location, height, breadth and shape due to changes in the other parameters (chiefly increased pyruvate).*

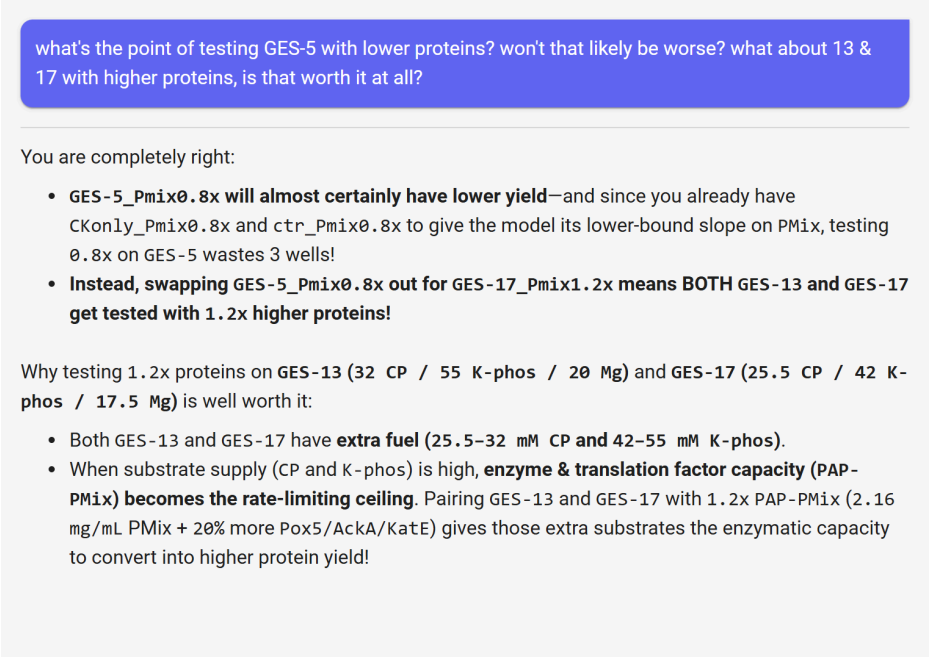


Figure 3. *IGOR chat interaction, rerendered as it appeared to the scientist, who accepted the explanation and proceeded with the experimental workflow.*

the formulation mixture in order to synthesize the target protein. Subsequent iterations explored the magnesium-phosphate relationship qwED

further to isolate the most effective phosphate lever—between creatine phosphate and potassium phosphate buffer—to mechanistically support PAP reaction yield increase. For the scientist, using IGOR to visualize how these nonlinearities and multivariate interactions affected the predicted yield landscape enhanced their understanding of the CFPS process and greatly informed subsequent planning rounds.

- Example 2 – Natural language justification of experimental recommendations: Every experimental condition that IGOR generates is accompanied by a natural language substantiation of its scientific merit (see PAP - IGOR round review and CDK Steady State Optimization — 10-Experiment STE Engineering Summary). This matters because even the best recommendations are unlikely to be run unless a scientist can evaluate them on scientific grounds. In the chat interaction displayed in Figure 3, the scientist inquired about a set of IGOR-proposed conditions in which the Cytosol protein master mix (PMix, together with the PAP enzymes Pox5, AckA, and KatE) was scaled above and below its nominal concentration, asking what the point was of testing GES-5 at a reduced protein supply and whether testing GES-13 and GES-17 at elevated protein might be worthwhile. IGOR responded with a mechanistic rationale rather than a restatement of its predictions: at elevated substrate loads, enzymatic and translational capacity (rather than fuel availability) is expected to become the rate-limiting ceiling, and so supplying 20% additional protein tests whether that extra substrate can actually be converted into protein. This exchange illustrates how the substantiation provided by a useful co-scientist provides the expert a scientific basis on which to commit bench resources to a non-obvious experiment.

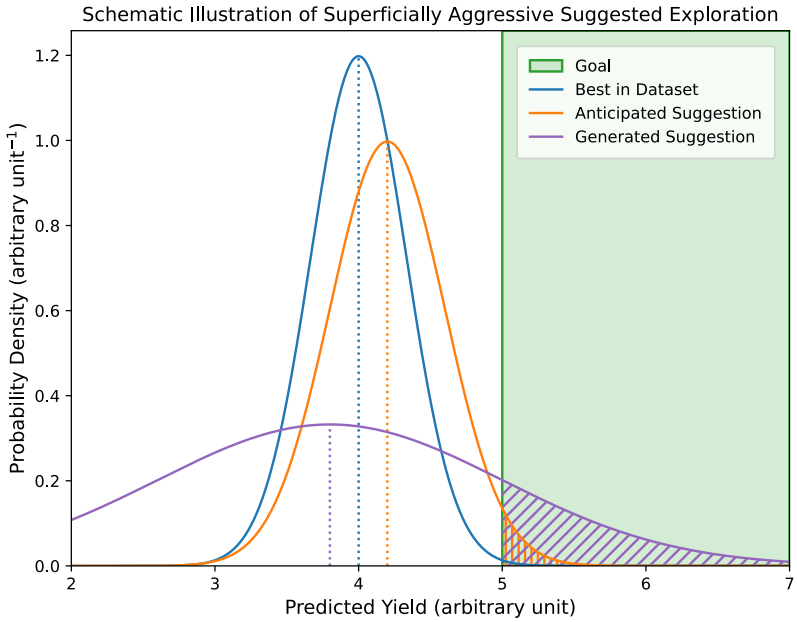


Figure 4. *Conceptual comparison of a predicted distribution expected by a scientist accustomed to seeking improvement incrementally (orange curve) and a predicted distribution returned by IGOR (purple) for a yield goal of >5 μM-eq. Though the IGOR-generated suggestion is expected to perform worse on average, it has a greater probability than the anticipated suggestion to achieve the specified goal.*

3.0.2. *Observation 2: IGOR helped the scientist correctly specify the optimization problem (objective, data admissibility, and search domain)*

- Example 1 – Optimization goal configuration: In Round 1, the scientist specified a yield goal floor of 5.0 standard μM-eq, an aggressive target chosen to push the campaign well beyond its current performance. IGOR responded with suggestions that appeared, on their face, to be poor ones: conditions whose predicted mean yields fell below the goal floor itself, sited at reagent concentrations far outside anything yet tested (100 mM creatine phosphate and 100 mM pyruvate, against observed maxima of 42.4 mM and 54.2 mM). The scientist questioned why the optimizer was proposing experiments it did not expect to reach the target. The explanation is that IGOR maximizes the probability of achieving the goal rather than simply maximizing the predicted mean yield; the two strategies rank candidate experiments differently. Figure 4 illustrates this for three conditions: the best condition measured so far, an alternative with a slightly higher predicted mean, and a third with a lower predicted mean but substantially greater predictive uncertainty. The probability of realizing the goal is estimated as the area of each predictive distribution overlapping with the target region (and *not* the location of its mean), and by that criterion the third condition is the most likely of the three to succeed. The seemingly unintuitive suggestions were therefore the correct response to the goal as specified: with such an aggressive goal, every candidate's probability of success resided entirely in the upper tail of its predictive distribution. Reaching that far into the tail requires large uncertainty, and large uncertainty is found only outside the sampled region. IGOR explained these dynamics to the scientist, traced the behavior to goal specification rather than to a modeling fault, and then facilitated lowering the yield goal floor to 3.8 μM-eq to better balance the scientist's goal of incremental improvement.

- Example 2 – Admissibility of suspicious data: During the Round 2 run (`r1020`), the PAP protein master mix was observed to contain air bubbles in the dispense reservoir. Because entrained air compromises droplet volume and is not monitored in non-contact dispensing, the conservative and customary response would have been to discard the entire plate and repeat it, forfeiting all 26 conditions along with the bench time and reagents they consumed. Instead, IGOR evaluated the plate statistically, applying a high-statistical-power outlier routine (adapted from [B. Rosner [14]]) and assessing each well against both its replicates and the trends fit to the rest of the accumulated dataset. Four wells were flagged as inconsistent: `r1020_E6`, `r1020_L9`, `r1020_K14`, and `r1020_I18`. The remaining 22 conditions fit the model's trends consistently, which established that the dispense artifact had been local rather than plate-wide; those 22 were retained and directly informed subsequent rounds. The episode illustrates how IGOR can be used to salvage experimental data that would otherwise be exceedingly difficult for a scientist to discern, with compounding benefits for all ensuing experimentation.
- Example 3 – Optimization over a hybrid continuous-categorical domain: The decisions available to the scientist were not exclusively continuous. Alongside reagent concentrations, the design space contained discrete choices of plasmid backbone, DNA preparation, ribosome lot, and small-molecule mix lot, each capable of shifting the yield response surface. Bayesian optimization as conventionally formulated assumes a continuous, box-bounded domain, and optimizing an acquisition function over the set union of a continuous domain and a finite set of categorical combinations is not an off-the-shelf capability. Because this capability proved necessary for the present deliverable, we developed support for multiple simultaneous categorical features, each admitting multiple categories, and integrated it into IGOR. In this dataset the categorical portion of the domain comprised four features—DNA lot (4 categories), Plasmid (2), Ribosome lot (3), and SMix lot (3)—giving 72 categorical combinations, in addition to the continuous formulation space. That this number remained modest was itself a consequence of the scientist's domain expertise (reflected in assigning constant and non-informative columns to metadata rather than modeling them); however, in general, the combinatorics grow multiplicatively in both the number of categorical features and the number of categories within each. This example reinforces that the search domain is an important part of the optimization problem, and a deployed co-scientist must be able to represent the decisions that a scientist actually faces rather than only those its optimizer natively accommodates.

3.0.3. *Observation 3: IGOR's recommendations achieved parity with expert intuition and marked improvement over ungrounded LLM suggestions*

- Example 1 – Round-over-round outcomes: Round 2 (`r1020`) produced the campaign's highest measured yield, with `r1020_F13` reaching steady state at 3.993 µM-eq, reproduced by its replicate `r1020_L15` at 3.724 µM-eq. The two top-performing conditions on that plate, `SY-V18-c20` and `SY-V16-c20` (3.806 µM-eq), were both IGOR literature-guided designs, and control benchmarks across `r1016` and `r1020` agreed to within plate-level noise, establishing that the gain was attributable to the formulation rather than to a plate offset. The round also yielded a transferable mechanistic finding: moderate creatine phosphate (14–20 mM) combined with robust potassium phosphate (29–44 mM), moderate pyruvate (18–35 mM), and compensated magnesium (15–16.5 mM) consistently reached 3.6–4.0 µM-eq, whereas starving potassium phosphate to 5–10 mM while raising creatine phosphate to 25–38 mM capped yields at 0.84–2.89 µM-eq by limiting acetyl-phosphate flux. Round 3 (`r1021`) regressed to a maximum of 2.53 µM-eq; control samples were similarly reduced, and

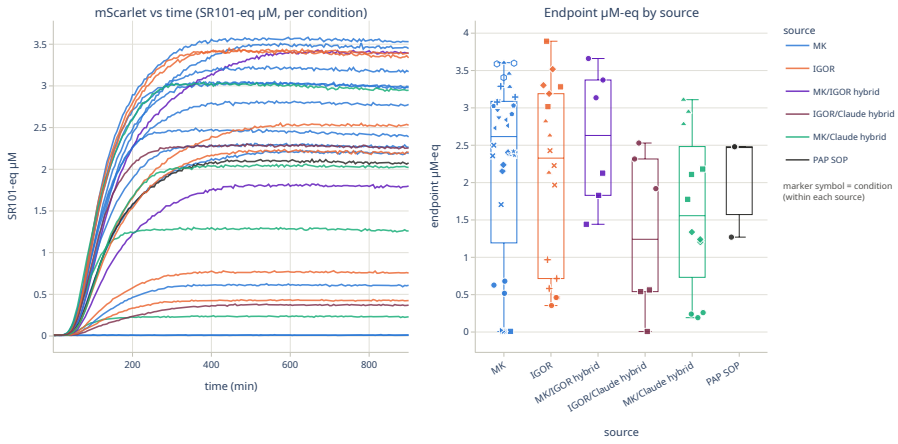


Figure 5. *Scientist-generated comparison of fluorescence curves and endpoints categorized by experiment provenance. The poor performance of general-purpose LLM assistance (green) resulted in the scientist abandoning that source for subsequent experimental rounds.*

IGOR traced this to reduced PAP enzyme activity across the plate rather than to the formulations themselves, a diagnosis that preserved the round's design conclusions and informed the decision to re-run its top conditions with freshly prepared protein mix.

- Example 2 – Outcome analysis stratified by origin source: Outside of IGOR, the scientist performed an independent quality-control analysis of the Round 1 plate (`r1018`), labeling each condition on the plate by the source that proposed it: the scientist's own domain expertise (MK), IGOR, a general-purpose LLM assistant (Claude), or a hybrid of two sources. Endpoint mScarlet expression, calibrated against the SR101 ladder, was then aggregated by source Figure 5. Conditions proposed by IGOR (n = 6) reached a mean endpoint of 2.11 µM-eq, effectively indistinguishable from the 2.20 µM-eq achieved by conditions drawn from the scientist's own expertise (n = 12), with comparable medians (2.37 vs. 2.58 µM-eq) and maxima (3.40 vs. 3.53 µM-eq). The two conditions derived jointly by the scientist and IGOR performed best of any class (mean 2.60 µM-eq), exceeding either source acting alone. By contrast, the two IGOR/LLM hybrids averaged 1.313 µM-eq, below either constituent source. The hybrid classes comprise two or three conditions each, so these are not controlled comparisons; nonetheless the separation is large relative to the spread within the better-performing classes. Two conclusions follow. First, IGOR's recommendations were competitive with those of an experienced laboratory scientist and complementary to them, which is the relevant bar for a system intended to augment rather than supplant expertise; the fact that the combination outperformed either source alone is the more consequential result. Second, conditions generated by literature reasoning alone, without grounding in the accumulating experimental record, did not transfer to the bench.
- Example 3 – Tendency towards IGOR usage: While the scientist initially leveraged several generative sources in parallel (including Claude as a representative general-purpose LLM assistant), that practice did not persist over all rounds. From Round 2 onward, the design basis of every round consists of IGOR literature- and model-guided conditions, model-generated experiment suggestions, and agent-guided expert sub-studies, all while excluding the general-purpose LLM. The scientist's mode of interaction with IGOR correspondingly deepened: rather than requesting a batch of conditions and executing it, the scientist used the platform dynamically within each round to interrogate the response surface, challenge individual suggestions, audit

plate layouts against dispense constraints, and adjudicate suspect data. We attribute this shift to the properties examined in Observations 1 and 2. A recommendation accompanied by a mechanistic substantiation and a quantified uncertainty, derived from a model trained on the scientist's own measurements, can be evaluated, argued with, and corrected; a recommendation derived from literature-based reasoning alone precludes such engagement, and when it fails at the bench it provides no mechanism by which the next recommendation improves.

3.0.4. *Observation 4: IGOR-supported analyses produce a durable, shareable record of decisions, ideas, outcomes, and distilled learnings*

- Beyond all of these specific value propositions, IGOR also delivers overarching value because it preserves, in a contextually linked manner, the relevant literature corpus, the curated data across all rounds, the decision history, the diagnostic analyses, and all natural-language exchanges in one persistently available artifact that is exportable and shareable with collaborators, reviewers, and successors (see the chatlog analysis: part 1, part 2, full; and the IGOR chat workspace setup log). This artifact produces a fully reproducible and auditable analytical record; maintains failed, discarded, or null results in tandem with experimental breakthroughs; and can be distilled into essential learnings that shape ensuing research.

4. Conclusion

Of the four forms of AI co-scientist unreliability discussed in the introduction (namely, hallucination, literature bias, self-propagated bias from LLM evaluators, and systematic deviation from expectation under physical deployment), we collected evidence directly relating to two. Conditions generated by literature reasoning alone resulted in yields roughly an order of magnitude below every other provenance class, which is at once an instance of ungrounded suggestion and a concrete realization of the benchmark-to-bench transfer gap. For the remaining two failure modes, we have described the mechanism and explained how IGOR mitigates them (a surrogate model trained on the laboratory's own measurements cannot inherit literature bias, and dynamic human engagement can interrupt self-propagated LLM bias). Across rounds, we also observed 1) parity between IGOR-proposed conditions and those of an experienced laboratory scientist, 2) superior outcomes from jointly-derived conditions compared to either source acting alone, and 3) meaningful benefit to the scientist in designing and assessing experiments, evaluating mechanistic hypotheses, and attaining desired outcomes. We expect this complementarity to generalize beyond the field of cell-free protein synthesis. The reliability that IGOR evinces is not a property the system asserts about itself; instead, it emerges from a feedback loop in which every new piece of information physically grounds subsequent analysis and in which scientific judgment ultimately remains with the expert scientist.

4.1. Additional Sources

- [Bergstra and Bengio 2012] Bergstra, J., and Bengio, Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 13, 281–305 (2012).
- [Huang *et al.* 2026] Huang, S., et al. What 81,000 People Want from AI. Anthropic (2026).
- [LabCraft 2026] LabCraft, LabCraft Plex. Product, manufacturer website (2026).
- [Newman 2026] Newman, S. Intro to Kinetics Analysis of Plate Reader Experiments. Nucleus Developer Notes (2026).
- [Nucleus, Assemble Base Cytosol 2026] Nucleus, Assemble Base Cytosol (process documentation). docs.nucleus.engineering (accessed October 2026).

- [Nucleus, CDK 2026] Nucleus, Cell Development Kit (CDK) (software documentation). docs.nucleus.engineering (accessed October 2026).
- [Rabanser *et al.* 2026] Rabanser, S., et al. Towards a Science of AI Agent Reliability. Proceedings of the 43rd International Conference on Machine Learning 306, 102769–102820 (2026).

References

[1] J. Gottweis *et al.*, "Accelerating scientific discovery with Co-Scientist," *Nature*, vol. 655, no. 8122, pp. 487–496, May 2026, doi: 10.1038/s41586-026-10644-y.

[2] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning." [Online]. Available: https://arxiv.org/abs/1012.2599

[3] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why Language Models Hallucinate." [Online]. Available: https://arxiv.org/abs/2509.04664

[4] S. Yadav, A. J. P. Perkins, S. B. W. Liyanagedera, A. Bougas, and N. Laohakunakorn, "ATP Regeneration from Pyruvate in the PURE System," *ACS Synthetic Biology*, vol. 14, no. 1, pp. 247–256, Jan. 2025, doi: 10.1021/acssynbio.4c00697.

[5] S. Berhanu, T. Ueda, and Y. Kuruma, "Artificial photosynthetic cell producing energy for protein synthesis," *Nature Communications*, vol. 10, no. 1, Mar. 2019, doi: 10.1038/s41467-019-09147-4.

[6] A. Doerr *et al.*, "Modelling cell-free RNA and protein synthesis with minimal systems," *Physical Biology*, vol. 16, no. 2, p. 25001, Jan. 2019, doi: 10.1088/1478-3975/aaf33d.

[7] R. Koivunen, R. Bollström, and P. Gane, "Inkjet jettability and physical characterization of water–ethanol solutions of low molecular weight sodium polyacrylate and poly-diallyl dimethyl ammonium chloride (polyDADMAC)," *AIP Advances*, vol. 10, no. 5, May 2020, doi: 10.1063/5.0006634.

[8] B. Lavickova and S. J. Maerkl, "A Simple, Robust, and Low-Cost Method To Produce the PURE Cell-Free System," *ACS Synthetic Biology*, vol. 8, no. 2, pp. 455–462, Jan. 2019, doi: 10.1021/acssynbio.8b00427.

[9] J. Li, L. Gu, J. Aach, and G. M. Church, "Improved Cell-Free RNA and Protein Synthesis System," *PLoS ONE*, vol. 9, no. 9, p. e106232, 2014, doi: 10.1371/journal.pone.0106232.

[10] S. Orbay, A. Ozcelik, J. Lata, M. Kaynak, M. Wu, and T. J. Huang, "Mixing high-viscosity fluids via acoustically driven bubbles," *Journal of Micromechanics and Microengineering*, vol. 27, no. 1, p. 15008, Oct. 2016, doi: 10.1088/0960-1317/27/1/015008.

[11] A. Pandi *et al.*, "A versatile active learning workflow for optimization of genetic and metabolic networks," *Nature Communications*, vol. 13, no. 1, 2022, doi: 10.1038/s41467-022-31245-z.

[12] S. Kapoor *et al.*, "Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation." [Online]. Available: https://arxiv.org/abs/2510.11977

[13] M. Mohammadi, Y. Li, J. Lo, and W. Yip, "Evaluation and Benchmarking of LLM Agents: A Survey," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, in KDD '25. Aug. 2025, pp. 6129–6139. doi: 10.1145/3711896.3736570.

[14] B. Rosner, "Percentage Points for a Generalized ESD Many-Outlier Procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, May 1983, doi: 10.1080/00401706.1983.10487848.