Compressing the Validation Bottleneck: An Agentic Self-Driving Lab for Scientific Discovery

Kyunghoon Hur [1] **Chihun Lee** [2]

Abstract

Agentic AI-for-Science can automate ideation, planning, and analysis, but final validation still depends on real experiments. A self-driving lab (SDL) can execute those experiments, yet discovery still have bottlenecks when the agent spends too many rounds on low-value experiments or when each round requires high-cost experiment. We target these two physical bottlenecks with one agent. First, a prior-aware agentic DOE loop uses domain knowledge, past results, a surrogate, and a verifier to propose feasible and informative next experiments, reducing trials-to-target. Second, a cost-aware surrogate agent predicts high-cost, high-resolution measurements from low-cost, low-resolution measurements. It choose mesurement method between high or low-cost based on the uncertainty. We examine these directions in antibody process development and a metal-alloy, respectively. Together, these components aim to accelerate the SDL loop by combining with agent so that both the number of loops and the experiment budget burden can be reduced.

1. Agents Automate Everything but the Experiment

Agentic AI-for-Science can automate much of the research workflow apart from real-world experiments, spanning ideation, planning, experimental design, analysis, and writing, as illustrated by the Virtual Lab, the AI Scientist, and Coscientist (Swanson et al., 2025; Lu et al., 2024; Yamada et al., 2025; Boiko et al., 2023). However, lab validation remains a physical constraint: a hypothesis is confirmed only when it is tested in real-world experiments. To mitigate this, as those upstream stages accelerate, the self-driving lab (SDL) has emerged as a promising way to reduce the time and cost of validation by automating repeated experiments and measurements (MacLeod et al., 2020; Szymanski et al., 2023).

SDLs mainly focus on automating experiment execution, but a human still decides whether the proposed design of experiments (DOE) is feasible, useful, and scientifically informative. For an agentic SDL, that decision must be made by the agent itself, not by human intervention. We tackle two bottlenecks in this agent–bench loop: first, whether the agent can propose the next DOE by leveraging domain knowledge, feedback, and bench-side feasibility; and second, whether the agent can account for the time and resources needed to execute and measure each proposed experiment.

2. Bottleneck 1: The Agentic DOE Loop

In an SDL, when the agent and the lab side interact, the agent must propose a DOE and the lab must execute it under physical constraints, and the results must let the agent propose the experiment setup. The bottleneck is that this agent–bench interaction cannot always produce better results. Without combining domain priors, experimental feedback, and feasibility checks, the loop can run incorrectly while still spending rounds on low-value experiments.

Bayesian optimization (BO) formalizes sequential experimental design by updating a surrogate with each returned result and selecting the next DOE by balancing promising conditions against uncertain regions (Frazier, 2018; Balandat et al., 2020). However, vanilla BO usually encodes expert knowledge and laboratory constraints only indirectly, for example through fixed bounds, hand-designed variables, or manually curated search spaces. This motivates an agent-driven BO-DOE loop in which domain-specific expert knowledge, literature priors, and bench-side constraints actively shape each proposal before execution. The agent suggests candidate regions from these priors, ranks them using past results, and removes infeasible designs before execution.

For example, in antibody process development, the agent proposes a small-scale bioreactor DOE, the lab measures

[1]AX Research Division, Korea Electronics Technology Institute, Republic of Korea [2]Material Data Division, Korea Institute of Materials Science, Republic of Korea. Correspondence to: Kyunghoon Hur <kyunghoonhur@keti.re.kr>, Chihun Lee <chihunlee@kims.re.kr>.

Preprint. July 3, 2026.

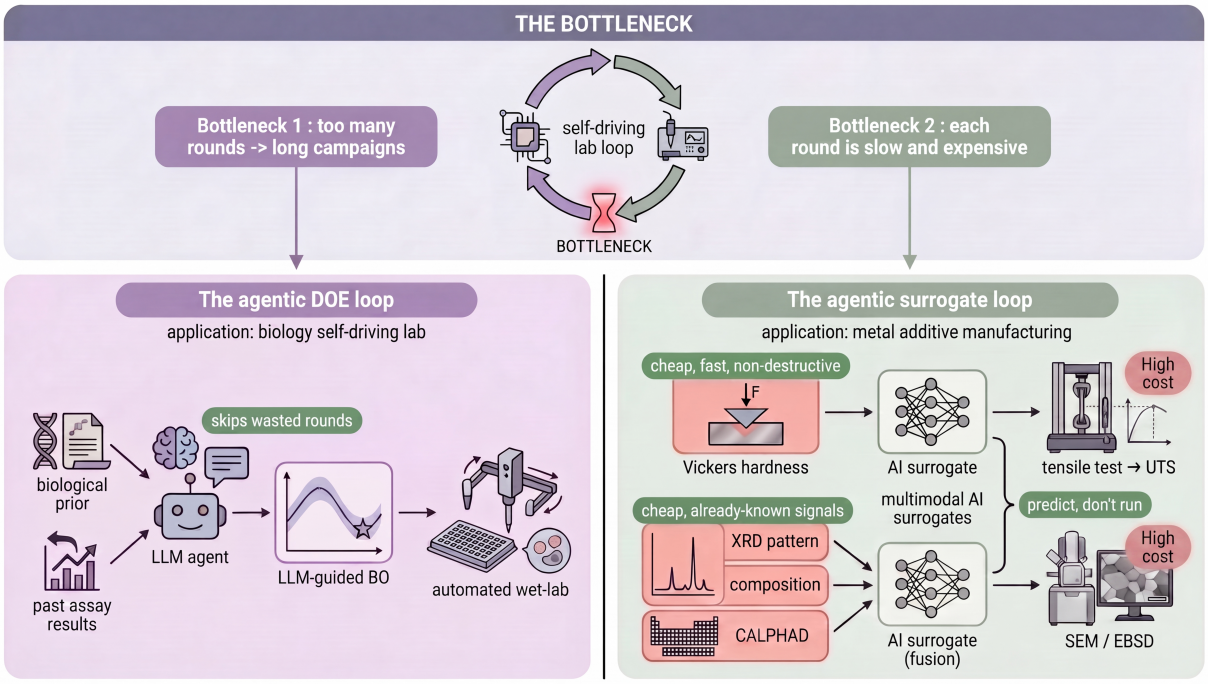

Figure 1. One agent attacks two bottlenecks of a domain-general method, shown in two application areas. Bottleneck 1 cuts the number of rounds. An LLM agent turns prior knowledge and past results into a Bayesian-optimization proposal, drawn here for antibody bioprocess optimization. Bottleneck 2 cuts the cost per round. Pre-trained surrogates predict an expensive measurement from cheap signals (hardness to tensile strength, and X-ray with composition and CALPHAD to phase fraction) for metal additive manufacturing, the lab we are building.

titer and product quality, and the next DOE updates feeding, pH, temperature, dissolved oxygen, and other culture conditions (Warr et al., 2011; Agarabi et al., 2015). These small-scale rounds are run before scale-up, because mini-bioreactors and scale-down systems screen conditions and predict larger manufacturing performance (Warr et al., 2011; Delouvroy et al., 2015). For a pharmaceutical company, finishing this optimization quickly lets a new antibody program move into scaled production sooner. In this setting, agentic DOE is expected to reduce trials-to-target by steering each round toward experimentally feasible and objective-relevant conditions.

We will compare against human-guided DOE, random search, grid search, and vanilla BO under the same experimental-loop budget, measuring trials-to-target, infeasible-proposal rate, and feedback sensitivity.

3. Bottleneck 2: The Agentic Measurement Loop

The AutoResearch concept introduced by Andrej Karpathy frames research as an iterative loop in which an agent proposes a change, executes it, evaluates the result, and keeps useful changes (**?**). This loop is straightforward in computational settings, where iterations can be repeated with low overhead. Moving the same idea to a wet-lab SDL is harder, since each loop consumes lab time, materials, and measurement capacity. If every candidate is sent to high-cost, high-resolution measurement, the loop becomes limited by measurement throughput rather than by the agent's ability to propose candidates (Kaufmann & Vecchio, 2024).

To reduce this cost, we propose a cost-aware surrogate agent that predicts selected high-cost, high-resolution measurement results from low-cost, low-resolution measurement results. When the surrogate uncertainty is low, the agent uses the prediction from the low-cost measurement result; when the uncertainty is high, the agent requests the high-cost measurement (Angelopoulos & Bates, 2021). This allows the SDL to evaluate more candidates under the same budget by spending high-cost measurements only where they are most informative (Alvi et al., 2026; Slautin et al., 2024). The surrogate is not used to guess an expensive measurement from weak inputs; it is used only when the low-cost measurements contain enough information to estimate the target reliably (Patel et al., 2025). When a low-cost measurement is not clear enough, the agent combines multiple inputs so their complementary information constrains the prediction.

For example, the metal additive-manufacturing SDL can provide two cases. First, a quick hardness test provides

a cheap proxy for tensile strength within one alloy family (Pavlina & Van Tyne, 2008). This proxy is useful when hardness and strength change together, but it should not be trusted when tensile failure is dominated by hidden porosity that indentation cannot detect. Second, phase fractions can be estimated from a combination of X-ray diffraction (XRD) (Benrabah et al., 2025), elemental composition, and a CALPHAD phase-diagram calculation (Liu et al., 2024a), reducing reliance on slower electron-microscopy measurements such as SEM/EBSD. Each input alone may not be clear enough, but together they constrain the target quantity. Together, these cases illustrate the same principle: the agent uses low-cost, fast measurements when they provide enough information for the results, and reserves high-cost characterization for candidates where the surrogate remains uncertain.

Following this design, we are building a metal-AM SDL that links directed energy deposition (DED) printing, sample cutting, heat treatment, and X-ray, optical, and electron characterization into one loop. We will evaluate surrogate accuracy and uncertainty calibration on held-out measurements, on-target hits per characterization cost relative to measuring everything, and wall-clock time to reach a target specification.

4. Conclusion

The physical experiment is the rate-limiter for agentic scientific discovery. We focus on two places where an SDL still wastes time: too many experimental loops and a high cost for each loop. A prior-aware agentic DOE uses feedback to choose better next experiments, while a cost-aware surrogate agent uses low-cost measurements to avoid unnecessary high-cost measurements. The objective is "reach the target faster, with fewer experiments within the budget".

References

Agarabi, C. D., Schiel, J. E., Lute, S. C., Chavez, B. K., Boyne, M. T., Brorson, K. A., Khan, M., and Read, E. K. Bioreactor process parameter screening utilizing a Plackett–Burman design for a model monoclonal antibody. *Journal of Pharmaceutical Sciences*, 104(6):1919–1928, 2015. doi: 10.1002/jps.24420.

Alvi, S. M. A. A., Vela, B., Attari, V., Janssen, J., Perez, D., Allaire, D., and Arróyave, R. Deep Gaussian process-based cost-aware batch Bayesian optimization for complex materials design campaigns. *npj Computational Materials*, 12:105, 2026. doi: 10.1038/s41524-026-01981-7.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Benrabah, I.-E., Geandier, G., Nakonechna, O., Denand, B., Van Landeghem, H., Deschamps, A., and Allain, S. Y. P. Deep learning for real-time phase quantification from X-ray diffraction: Toward high-throughput steel microstructure mapping. *Advanced Engineering Materials*, 2025. doi: 10.1002/adem.202503172.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Delouvroy, F., Siriez, G., Tran, A.-V., Mukankurayija, L., Kochanowski, N., and Malphettes, L. ambr minibioreactor as a high-throughput tool for culture process development to accelerate transfer to stainless steel manufacturing scale. *BMC Proceedings*, 9(Suppl 9):P78, 2015. doi: 10.1186/1753-6561-9-S9-P78.

Forrester, A. I. J., Sóbester, A., and Keane, A. J. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A*, 463(2088):3251–3269, 2007. doi: 10.1098/rspa.2007.1900.

Frazier, P. I. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Gupta, R., Hartford, J., and Liu, B. LLMs for Bayesian optimization in scientific domains: Are we there yet? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2509.21403.

Han, D., Ai, Z., Cai, P., Xu, T., Li, Y., Zhang, S., et al. ChemBOMAS: Accelerated Bayesian optimization for scientific discovery in chemistry with LLM-enhanced multi-agent system. *arXiv preprint arXiv:2509.08736*, 2025.

Kaufmann, K. and Vecchio, K. S. Autonomous materials research and design: Characterization. *Current Opinion in Solid State and Materials Science*, 32:101192, 2024. doi: 10.1016/j.cossms.2024.101192.

Kennedy, M. C. and O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. doi: 10.1093/biomet/87.1.1.

Liu, S., Bocklund, B., Diffenderfer, J., Chaganti, S., Kailkhura, B., McCall, S. K., Gallagher, B., Perron, A., and McKeown, J. T. A comparative study of predicting high entropy alloy phase fractions with traditional

machine learning and deep neural networks. *npj Computational Materials*, 10:172, 2024a. doi: 10.1038/s41524-024-01335-1.

Liu, T., Astorga, N., Seedat, N., and van der Schaar, M. Large language models to enhance Bayesian optimization. In *International Conference on Learning Representations (ICLR)*, 2024b. arXiv:2402.03921.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6: 525–535, 2024.

MacLeod, B. P., Parlane, F. G. L., Morrissey, T. D., et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020.

Patel, D., Sharma, R., and Guo, Y. B. Computational, data-driven, and physics-informed machine learning approaches for microstructure modeling in metal additive manufacturing. *arXiv preprint arXiv:2505.01424*, 2025.

Pavlina, E. J. and Van Tyne, C. J. Correlation of yield strength and tensile strength with hardness for steels. *Journal of Materials Engineering and Performance*, 17 (6):888–893, 2008. doi: 10.1007/s11665-008-9225-5.

Ramos, M. C., Michtavy, S. S., Porosoff, M. D., and White, A. D. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341*, 2023.

Sabanza-Gil, V., Barbano, R., Pacheco Gutiérrez, D., Luterbacher, J. S., Hernández-Lobato, J. M., Schwaller, P., and Roch, L. Best practices for multi-fidelity Bayesian optimization in materials and molecular research. *Nature Computational Science*, 2025. doi: 10.1038/s43588-025-00822-9. arXiv:2410.00544.

Shang, X., Talbot, A., Li, E., Wen, H., Lyu, T., Zhang, J., and Zou, Y. Accurate inverse process optimization framework in laser directed energy deposition (AIDED). *Additive Manufacturing*, 102:104736, 2025. doi: 10.1016/j.addma.2025.104736.

Slautin, B. N., Pratiush, U., Ivanov, I. N., Liu, Y., Pant, R., Zhang, X., Takeuchi, I., Ziatdinov, M. A., and Kalinin, S. V. Co-orchestration of multiple instruments to uncover structure–property relationships in combinatorial libraries. *arXiv preprint arXiv:2402.02198*, 2024.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Szymanski, N. J., Rendy, B., Fei, Y., et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

Warr, S. R. C., Patel, J., Ho, R., and Newell, K. V. Use of micro bioreactor systems to streamline cell line evaluation and upstream process development for monoclonal antibody production. *BMC Proceedings*, 5(Suppl 8):P14, 2011. doi: 10.1186/1753-6561-5-S8-P14.

Whitman, S. E. and Latypov, M. I. Machine learning of microstructure–property relationships in materials leveraging microstructure representation from foundational vision transformers. *Acta Materialia*, 2025. doi: 10.1016/j.actamat.2025.121217. arXiv:2501.18637.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Yuan, X., Chen, Z., Zhang, J., Xiong, H., Ye, N., Li, Y., and Gu, Q. Unleashing LLMs in Bayesian optimization: A preference-guided framework for scientific discovery. In *International Conference on Learning Representations (ICLR)*, 2026.

A. Method Landscape and Positioning

This appendix organises the references cited in the main text by methodological family and the question each addresses for our framework. The goal is to make explicit (i) where the proposed agent sits in the current LLM + Bayesian optimization (BO) landscape, (ii) which line of work the surrogate characterization is built on, and (iii) which self-driving lab (SDL) and additive-manufacturing (AM) precedents anchor the case study.

A.1. LLM-augmented Bayesian optimization (Bottleneck 1)

Recent work has placed LLMs at different points in the BO loop. We compare them on four axes that matter for our framing, split across two tables to keep each readable: Table 1 covers *where the LLM enters the loop* and *how prior knowledge is represented*; Table 2 covers *how feedback is handled* and the *empirical domain* where each system has been validated.

Table 1. LLM placement and prior representation (§A.1, part 1 of 2). Last row is the agent proposed in this work.

Method	LLM role in loop	Prior representation
LLAMBO (Liu et al., 2024b)	Warm-start, sample proposer, surrogate augmentation	Natural-language conditioning
BO-ICL (Ramos et al., 2023)	LLM-as-regression-surrogate (in-context)	Prompt + experimental history
LGBO (Yuan et al., 2026)	Per-round region/point preference	Region/point + confidence
ChemBOMAS (Han et al., 2025)	Pre-loop search-space decomposition	Chemistry RAG + LLaMA fine-tune
Coscientist (Boiko et al., 2023)	Full plan-and-execute agent	Free-form chat reasoning
ChemCrow (M. Bran et al., 2024)	Tool-augmented LLM	Tool outputs + chat history
Gupta et al. (2025)	*Diagnostic only* (no new system)	—
This work	DOE proposer *and* domain selector	Physical-context block $\mathcal{P}$ + history $\mathcal{H}_t$ + literature

Table 2. Feedback handling and empirical setting (§A.1, part 2 of 2). "Random-label test" refers to the diagnostic introduced by Gupta et al. (2025): an LLM agent passes the test only if its selections change when feedback labels are randomly permuted.

Method	Feedback handling	Empirical domain
LLAMBO (Liu et al., 2024b)	In-context history (no posterior update on the LLM)	Hyperparameter tuning
BO-ICL (Ramos et al., 2023)	Re-prompt each round with full history	Catalysis (RWGS, OCM benchmarks)
LGBO (Yuan et al., 2026)	Mean-shift on GP from LLM region (forward-only)	Physics, chemistry, biology, materials
ChemBOMAS (Han et al., 2025)	Tree refinement (no closed loop on outcomes)	Buchwald and Suzuki couplings + wet-lab
Coscientist (Boiko et al., 2023)	Natural-language summaries fed to next prompt	Pd-catalysed cross-couplings
ChemCrow (M. Bran et al., 2024)	Tool-result ingestion via chat	Synthesis, retrosynthesis, repellent design
Gupta et al. (2025)	Reveals LLM agents are insensitive to feedback (random-label test)	Gene perturbation, molecules
This work	Surrogate-mediated; verifier-gated; both prior and outcome routed through the GP/neural surrogate, not the LLM alone	Antibody bioprocess (case study)

Two observations frame our position. First, LGBO (Yuan et al., 2026) is the closest published system: it integrates the LLM in every round (rather than once at warm-start), and gives the only theoretical guarantee in the table (bounded performance loss when the LLM is misaligned). It is, however, restricted to a single target domain and emits region/point preferences only. Second, Gupta et al. (2025) have shown empirically that off-the-shelf LLM experimental-design agents (BioDiscoveryAgent, LLAMBO) frequently *fail* to use feedback: replacing true outcomes with permuted labels does not change selections. Our framework does not bypass this failure mode by trusting the LLM more — we route both prior and feedback through the GP/neural surrogate and a feasibility verifier so feedback effects are measurable.

A.2. Surrogate Characterization and Multi-fidelity Modelling (Bottleneck 2)

Treating a cheap, fast measurement as a low-fidelity view of an expensive one is the multi-fidelity recipe. Co-kriging fuses the two through a learned relationship (Kennedy & O'Hagan, 2000), multi-fidelity optimization adds adaptive sampling between fidelity levels (Forrester et al., 2007), and recent guidance characterizes when a cheap source is worth using (Sabanza-Gil et al., 2025).

Our contribution is to make the cheap source a measurement modality rather than a coarser simulation, and to keep its use

honest. The closest work is microViT (Whitman & Latypov, 2025), which predicts a property from a frozen foundation backbone with no task-specific training. We place such a surrogate inside the agent loop and gate it by calibrated uncertainty, so the agent decides per candidate when a real measurement is worth its cost. This is different from co-orchestration (Slautin et al., 2024), which schedules the next instrument rather than replacing it, and from low-cost hardware laboratories, which cut the price of the machine rather than the cost of the measurement.

A.3. Self-driving Labs and Characterization Precedents (Case Study)

Closed-loop self-driving labs were popularised in chemistry and materials science (MacLeod et al., 2020; Szymanski et al., 2023; M. Bran et al., 2024). For metal AM, AIDED (Shang et al., 2025) couples a genetic algorithm with machine-learning surrogates to invert process parameters in laser directed-energy deposition in about an hour per round. It is a Bottleneck 1 system that tightens the on-rig loop on the expensive metal target alone. On the characterization side, deep models now read an X-ray diffraction pattern into phase fractions about two orders of magnitude faster than Rietveld refinement (Benrabah et al., 2025), and phase fractions also follow from composition and temperature through CALPHAD-trained surrogates (Liu et al., 2024a). Bottleneck 2 places these single-shot predictors under one uncertainty-gated agent, so a cheap prediction is trusted only when its calibrated uncertainty is low, and a real measurement is triggered otherwise.

A.4. Agentic Upstream Stack (Motivation)

The motivating works in the introduction — the Virtual Lab (Swanson et al., 2025), the AI Scientist (Lu et al., 2024; Yamada et al., 2025), Coscientist (Boiko et al., 2023), and ChemCrow (M. Bran et al., 2024) — are not direct competitors to the framework proposed here. Each demonstrates that a sufficiently capable agent can run the upstream pipeline (ideation, planning, paper writing, or chemistry-tool orchestration) at or near human productivity. They are cited because they collectively establish that the rate-limiter has shifted: the physical experiment, not the agent turn, now dominates the schedule for high-value low-throughput science. Bottleneck 1 and Bottleneck 2 are responses to that shift, not extensions of those systems.