# Compressing the Validation Bottleneck: An Agentic Self-Driving Lab for Metal Additive Manufacturing

**Anonymous Authors**[1]

## Abstract

Agentic AI-for-Science systems have walked the upstream research pipeline end-to-end – ideation, planning, experimental design, and feedback loops are no longer the long pole. The bottleneck has slid downstream onto the bench: each candidate must still be made and measured, and self-driving labs (SDLs) amortise human labour but not the underlying material, machine-time, and characterisation cost of *one* trial. We compress this bottleneck along two orthogonal cost axes with a single agent. **(i)** A prior-aware agent design-of-experiments (DOE) loop ingests physical context and trial history to cut trials-to-target below grid, random, and vanilla Bayesian optimisation; both prior and outcome are routed through a verifier and surrogate so feedback measurably affects selection. **(ii)** Cross-domain augmentation supplements scarce target-domain runs with a structurally aligned cheap surrogate; domain choice is itself an action of the same agent. We instantiate the framework for metal additive manufacturing with SLA resin printing as the surrogate, at orders-of-magnitude lower cost per trial.

## 1. Introduction: The Validation Bottleneck

Agentic AI-for-Science systems have, in two years, walked the upstream research pipeline end-to-end. The Virtual Lab orchestrates a multi-LLM team that designs SARS-CoV-2 nanobody binders with experimental validation at roughly one percent human intervention (Swanson et al., 2025); the AI Scientist proposes, runs, and writes up ML papers, with v2 producing the first AI-generated workshop manuscript to clear peer review (Lu et al., 2024; Yamada et al., 2025); Coscientist plans and executes Pd-catalysed cross-couplings from a chat prompt (Boiko et al., 2023); ChemCrow couples LLMs to expert chemistry tools for autonomous synthesis (M. Bran et al., 2024). With ideation, planning, and feedback no longer the long pole, the bottleneck has slid downstream onto the bench: each candidate must still be made and measured, and that step now governs the schedule.

Self-driving labs (SDLs) were the natural response, wiring agents to robotic execution (MacLeod et al., 2020; Szymanski et al., 2023); their leverage is real but concentrated on *human labour* and *handover overhead*. The underlying material, machine-time, and characterisation cost of *one* trial is unchanged. For high-value, low-throughput domains – metal additive manufacturing (AM), structural alloys, device-level characterisation – both the *number* of trials and the *cost per trial* remain prohibitive. We attack both with a single agent.

**Contributions.** **(i) Prior-aware Agent DoE (§2).** An LLM agent conditioned on physical priors and trial history proposes each next experiment, satisfying the target spec in fewer trials than grid, random, or vanilla Bayesian optimisation (BO) by ingesting unstructured natural-language priors – literature, hidden constraints, scientist intuition – that a probability-only surrogate cannot. **(ii) Cross-domain augmentation (§3).** Because per-trial cost stays high even when trial count drops, we co-train on a cheap, structurally aligned surrogate domain whose geometry and process axes overlap the target's; the same agent that selects the next experiment also chooses which domain $d$ to run it in. We instantiate the framework for metal AM (target) $\leftrightarrow$ SLA resin printing (surrogate), at orders-of-magnitude lower cost per trial.

## 2. Direction 1: Prior-Aware Agent DoE

Figure 1 shows the pipeline. At round $t$ the agent receives the trial history $\mathcal{H}_t$, a physical-context block $\mathcal{P}$ (allowed materials, machine envelopes, stability constraints, literature priors), and the target spec $\tau$, and returns a structured DOE $\mathbf{x}_t$ over composition, geometry, and process parameters with an information-gain estimate. The SDL (or a human, in the bootstrap phase) executes $\mathbf{x}_t$, returns measurement $y_t$, and the cycle repeats. The agent is hybrid: an LLM planner that emits candidate DOEs, a surrogate model (Gaussian

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

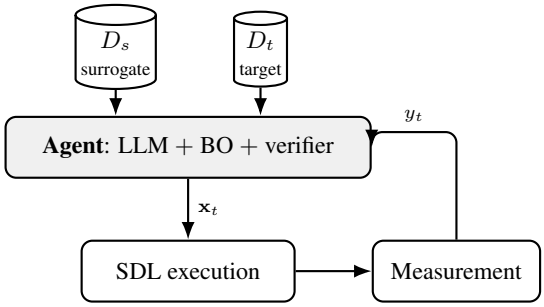


*Figure 1.* Agentic SDL with cross-domain augmentation. The agent unifies Direction 1 (prior-aware DoE on $D_t$) and Direction 2 (joint training on $D_s \cup D_t$); the domain choice $d \in \{s, t\}$ is itself an agent action. Outcomes feed back into the trial history and the joint dataset.

process or neural BO (Balandat et al., 2020; Frazier, 2018)) that scores them, and a verifier that gates against safety and feasibility constraints.

**Why an agent over pure BO.** Vanilla BO is black-box on hand-engineered features and cannot ingest *unstructured natural-language* priors – literature, hidden constraints, scientist intuition – without re-engineering the kernel or the feature space. The LLM contributes physics-aware reasoning a probability-only loop cannot, e.g. flagging a scan strategy that violates a thermal budget or melts an existing support. Recent LLM-augmented BO works – LLAMBO (Liu et al., 2024), BO with in-context learning (Ramos et al., 2023), LGBO (Yuan et al., 2026), and ChemBOMAS (Han et al., 2025) – form the proper comparison frontier. Gupta et al. (2025) have warned that off-the-shelf LLM agents often *fail* to use feedback (random-label sensitivity); we therefore route both prior and feedback through the verifier and the surrogate, not the LLM alone. The same agent also chooses *which domain* $d \in \{s, t\}$ to run the next trial in, foreshadowing §3.

We compare against grid, random, and vanilla BO with the same surrogate at a fixed trial budget; the primary metric is trials-to-target.

## 3. Direction 2: Cross-Domain Surrogate Transfer

Even an optimal agent loop converts each trial into information at maximum efficiency, but the absolute per-trial cost in metal AM (powder, machine time, post-processing) stays high. We amortise learning across a paired *surrogate* domain $D_s$ that we *hypothesise* satisfies, and the case study tests: **(H1)** shared decision variables (geometry, scan/print path, layer-level process parameters); **(H2)** a transfer-friendly outcome surface, where qualitative trends in $D_s$ predict the sign and rough magnitude of effects in $D_t$; **(H3)** per-trial cost orders of magnitude below $D_t$.

We train the agent's surrogate on $D_s \cup D_t$ via a multi-task / multi-fidelity objective in the spirit of co-kriging (Kennedy & O'Hagan, 2000; Forrester et al., 2007) and modern domain adaptation (Hospedales et al., 2021), with $D_t$ samples reweighted to anchor the target manifold; the model $f_\theta(\mathbf{x}, d)$ takes a domain indicator $d \in \{s, t\}$ and shares low-level features. Effective $|D_t|$ is augmented by the structurally aligned $|D_s|$, which in turn shortens the loop in Section 2. Recent SDL work in metal AM has demonstrated rapid surrogate-driven parameter optimisation in laser directed-energy deposition (Shang et al., 2025) and transfer-learning-based multi-fidelity surrogates for melt-pool modelling (Huang et al., 2022); our framing makes the cheap-surrogate domain a first-class citizen in the agent's policy rather than an offline warm-start.

## 4. Case Study and Evaluation Protocol

**Target / surrogate.** Laser powder-bed fusion (or directed-energy deposition) of a target alloy, optimising a spec $\tau$ over geometric fidelity, surface quality, and a target mechanical property (e.g., yield strength). The surrogate is an SLA resin printer covering the same envelope of part geometries and analogous process axes (layer thickness, exposure/scan, support strategy); per-trial cost is hours of resin and machine time vs. days and kilograms of powder for the metal target.

**Mapping and three-number protocol.** An aligned schema (geometry primitives, scan/exposure strategy, layer parameters) plus a domain indicator $d$ feeds both heads of $f_\theta$; outcomes project to a shared latent of geometric/structural quality with material-specific heads on top. The evaluation commits to three numbers, each individually falsifiable: **(i)** trials-to-target on $D_t$ versus grid, random, and BO baselines (Direction 1); **(ii)** sample-efficiency lift attributable to $D_s$, via an ablation that drops the surrogate domain from joint training (Direction 2); **(iii)** wall-clock target-hit time including all surrogate runs (the end-to-end claim). We are constructing the metal-AM SDL and the SLA surrogate rig now; the workshop submission presents the framework and protocol, with empirical results to follow.

## References

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.

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

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.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169, 2021.

Huang, X., Xie, T., Wang, Z., Chen, L., Zhou, Q., and Hu, Z. A transfer learning-based multi-fidelity point-cloud neural network approach for melt pool modeling in additive manufacturing. *ASME Journal of Risk and Uncertainty in Engineering Systems, Part B*, 8(1):011104, 2022. doi: 10.1115/1.4051749.

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

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.

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.

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.

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 optimisation (BO) landscape, (ii) which classical line of work the cross-domain transfer is built on, and (iii) which self-driving lab (SDL) and additive-manufacturing (AM) precedents anchor the case study.

## A.1. LLM-augmented Bayesian Optimisation (Direction 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., 2024) | 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., 2024) | 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 | Metal AM ↔ SLA (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 extra cost 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. Multi-fidelity and Cross-domain Surrogate Modelling (Direction 2)

The classical recipe for fusing cheap and expensive evaluations is co-kriging (Kennedy & O'Hagan, 2000), in which the high-fidelity Gaussian process is autoregressive on the low-fidelity one through a scaling factor and a bias process. Forrester et al. (2007) extended this to multi-fidelity surrogate-based optimisation with adaptive sampling between fidelity levels and demonstrated practical gains in aerodynamic and materials design. Modern domain-adaptation surveys (Hospedales et al., 2021) frame the same problem statistically: a shared backbone $f_\theta(\mathbf{x}, d)$ with domain indicator $d$, multi-task heads, and

re-weighting of the target-domain samples to anchor the joint manifold.

Our setup generalises the classical recipe in two ways. (i) The cheap and expensive sources are *different physical processes* — SLA resin photopolymerisation versus metal laser powder-bed fusion / directed-energy deposition — not two fidelities of the same simulator. The transfer is therefore not just a smoothness assumption between fidelities but a hypothesised structural alignment in geometry and process variables, made explicit as (H1)–(H3) in Section 3. (ii) The agent, rather than a fixed acquisition rule, decides at each round whether to spend the next trial on $D_s$ or $D_t$; domain choice is a first-class action in the policy.

### A.3. Self-driving Labs and Metal AM 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 specifically, two recent threads anchor opposite ends of our problem. AIDED (Shang et al., 2025) couples a genetic algorithm with machine-learning surrogates to invert process parameters in laser directed-energy deposition, producing optimal print geometries in $\sim$1 h per closed-loop round. It is a Direction-1 system: a tighter on-rig optimisation loop, but on the expensive metal target alone. MF-PointNN (Huang et al., 2022) uses transfer learning over point-cloud features to bridge low-fidelity finite-element melt-pool simulations and high-fidelity experiments; this is a Direction-2 system at the simulation $\rightarrow$ experiment boundary, on a single physical process. Neither thread combines the two axes. Our framework asks a single agent to do both: pick the next experiment *and* pick the domain it runs in.

### 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. Direction 1 and Direction 2 are responses to that shift, not extensions of those systems.