ACHT-World: Causal World Models for Closed-Loop Self-Driving Laboratories

David Scott Lewis [1] **Enrique Zueco** [1]
[1]*AIXC Research, Zaragoza, Spain*. Correspondence to: David Scott Lewis reports@aiexecutiveconsulting.com.

Self-driving laboratories are increasingly capable at "optimize-this" loops, yet many scientific goals require identifying *mechanisms* and selecting decisive experiments under uncertainty. We propose **ACHT-World**: a world-model-centric architecture where the internal state is an explicit causal belief over mechanisms, updated by experiments chosen for expected information gain. A trust layer adds plausibility checks, cross-database consistency, and replayable decision logs. The result is a closed-loop system that generates hypotheses, selects interventions, validates outcomes, and emits an auditable discovery trace. We validate core components in silico on synthetic protein-association graphs and outline a path to physical laboratory integration.

1. Introduction

Closed-loop discovery systems following the Design–Make–Test–Analyze (DMTA) cycle promise orders-of-magnitude acceleration in chemistry and materials science [1, 2, 3]. Current platforms, however, couple powerful generators with weak epistemic controls: they excel at optimizing a scalar objective but rarely ask *which experiments would most reduce mechanistic uncertainty* [4].

When the goal is mechanistic understanding—rather than mere optimization—agents must learn causal structure, decide which experiments are worth running, and justify those choices [5, 6]. Foundation-model "AI scientists" can generate research ideas [7, 8, 9] but struggle with reliable self-assessment and causal grounding [10, 11]. World models from reinforcement learning [12, 13] show that maintaining a learned internal state enables planning under uncertainty, yet their application to laboratory science remains nascent.

We propose **ACHT-World**, a world-model architecture for self-driving laboratories (SDLs) whose internal state is a structured causal belief. Building on Active Causal Hypothesis Testing (ACHT) [14, 15], the system selects each experiment to maximize expected information gain (EIG) over a causal query, updates its mechanism posterior, and logs every decision for auditability. The architecture separates *state* (what the agent believes), *decision* (what to do next), and *trust* (why the choice is justified), bridging the gap between autonomous optimization and accountable scientific inquiry.

2. World-model architecture

We define a world model for scientific research not as a monolithic simulator but as a structured belief state that integrates mechanistic priors, learned surrogates, and observational evidence. This belief state is *actionable*: it supports counterfactual queries and experiment selection with explicit uncertainty quantification.

2.1 State representation

ACHT-World maintains a hybrid world state

$$\mathbf{s} = \{\mathbf{z},\ P(\mathcal{G} \mid \mathcal{D}),\ \Pi\}, \tag{1}$$

where $\mathbf{z}$ is a learned latent embedding (e.g., from a GNN or sequence encoder over molecular/assay features), $P(\mathcal{G} \mid \mathcal{D})$ is a posterior distribution over candidate structural causal models (SCMs) or directed acyclic graphs (DAGs) given data $\mathcal{D}$ [16, 17, 18], and Π is a protocol layer encoding feasible laboratory actions and physical/chemical constraints.

2.2 Decision layer (ACHT)

Given a hypothesis family, the agent proposes interventions $\mathrm{do}(X_i = x)$ and selects the next experiment by maximizing approximate expected information gain over a causal query Q (e.g., the total causal effect of a pathway perturbation on a phenotype):

$$a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{G} \sim P(\mathcal{G} \mid \mathcal{D})}\big[\, I_{\mathcal{G}}(Q;\, y_a \mid \mathcal{G}) \,\big], \tag{2}$$

where $I_{\mathcal{G}}(Q; y_a \mid \mathcal{G})$ denotes the mutual information between query Q and the potential observation y_a under graph $\mathcal{G}$ [14, 15]. The mechanism posterior is updated after each observation, closing the DMTA loop. This formulation turns experiment selection into a decision-theoretic problem rather than a heuristic search [19, 20].

2.3 Trust layer: governance and auditability

A governance stack enforces three safeguards [21, 22, 23]: (i) *plausibility checks*—physics, chemistry, and biology constraints that reject infeasible interventions before execution; (ii) *cross-database consistency*—comparison of proposed mechanisms against curated knowledge graphs [24]; and (iii) *replayable decision logs* linking each chosen experiment to the uncertainty it was expected to reduce. This layer enables post-hoc audit of every laboratory action, a prerequisite for deployment in regulated domains such as drug discovery [25, 26].

3. Validation plan and demonstration

We validate core components in silico and provide an integration path to physical laboratories.

3.1 In-silico benchmarks and related work

Using synthetic protein-association graphs at $\sim 10^{1.7}$ scale (50 nodes, ~100 edges), we benchmark

four intervention-selection policies under a fixed budget of 20 experiments: (i) **BOED**—Bayesian optimal experimental design via EIG approximation; (ii) **Random**; (iii) **Degree-based** heuristic (highest-degree node first); and (iv) **Correlation-ranked** (most correlated with the target). BOED is the only policy that reduces posterior entropy of the causal query (to 94% of the initial value), while heuristic baselines leave entropy unchanged or increase it, confirming that information-gain-driven experiment selection is essential for mechanism identification (see Appendix A for full results).

Self-driving lab systems commonly pair Bayesian optimization with robotic execution [1, 27]; foundation-model "AI scientists" generate ideas but lack reliable self-assessment [7, 9]. ACHT-World contributes a middle layer—causal world models—that makes experiment choice a decision-theoretic problem while adding governance artifacts absent from most closed-loop platforms.

3.2 Deliverables and metrics

We define a minimal robotics interface Π for common DMTA operations (perturb, measure, analyze), enabling incremental deployment: human-in-the-loop execution first, then agent-in-the-loop planning, and finally full robotic execution as protocols mature. Primary metrics include: *information gain per experiment*, *mechanism recovery at fixed budget*, and *auditability* (replayability + attribution completeness).

Table 1 summarizes the four core modules.

Table 1: ACHT-World core modules and their roles.

Module	**Role**
World state $\mathbf{s} = \{\mathbf{z}, P(\mathcal{G} \mid \mathcal{D}), \Pi\}$	Integrates embeddings $\mathbf{z}$, causal posterior $P(\mathcal{G} \mid \mathcal{D})$, and protocol constraints Π
Planner (ACHT)	Selects next $\mathrm{do}(x)$ by maximizing EIG over causal query Q
Executor	Human/robot executes DMTA step; returns multimodal observations + metadata
Validator + archivist	Plausibility checks, cross-database consistency, replayable decision traces

4. Conclusion

ACHT-World provides a principled architecture for moving self-driving laboratories from scalar optimisation to mechanistic understanding. By pairing structured causal beliefs with information-gain-driven experiment selection and a governance layer, the system makes every laboratory action accountable. In-silico benchmarks confirm that BOED-guided interventions uniquely reduce posterior entropy; future work will integrate physical DMTA execution and scale to multi-assay campaigns.

The key distinction between ACHT-World and existing self-driving laboratory platforms lies in the internal state representation: where systems like Ada [24], ATLAS, and autonomous chemistry agents [8, 9] optimize scalar objectives (yield, purity, binding affinity), ACHT-World maintains an explicit causal posterior $P(\mathcal{G} \mid \mathcal{D})$ and selects experiments to maximize information gain over mechanistic queries. This shift from optimization to causal discovery is critical for domains where understanding *why* an intervention works is as important as *that* it works—regulatory science, hypothesis-driven drug discovery, and mechanistic toxicology all require explanatory models, not black-box predictions.

Scalability considerations become acute as graph size and intervention budget increase. The current BOED approximation samples K=200 DAGs from the posterior at each step, requiring $O(K \cdot B \cdot n)$ forward simulations for a budget of B interventions over n nodes. For n=50, this is manageable; for n=500 (e.g., proteomic interaction networks), amortized variational inference or neural approximators for the EIG functional will be essential. The linear Gaussian SEM assumption (Eq. A1) is a simplifying choice: real biological systems exhibit nonlinear dose-response curves, combinatorial interactions, and temporal dynamics. Extending the framework to nonlinear SEMs, time-series interventions, and multi-modal observations (imaging, sequencing, proteomics) is a priority for scaling to physical laboratories [1, 2, 27].

The governance layer addresses a gap in current autonomous science platforms: auditability and regulatory compliance. In regulated domains such as pharmaceutical development, preclinical toxicology, and clinical trial design, every experimental decision must be justified, reproducible, and traceable to a data-driven rationale. The trust layer (Section 2.3) enforces plausibility checks, cross-database consistency, and replayable decision logs, enabling post-hoc audit of the entire discovery trace. This is not a peripheral feature—it is a prerequisite for deploying autonomous agents in settings where errors carry financial, ethical, or safety consequences [21, 22, 23]. As LLM-based agents become more capable [7, 10, 11, 25], the risk of hallucination and overconfidence increases; embedding governance at the architecture level ensures that autonomy does not come at the expense of reliability.

References

[1] Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Raber, Michael D. Rankovic, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.

[2] Alan Henson, Phillip M. Maffettone, Nicola Sherwen, and Keith T. Butler. A roadmap to the democratization of self-driving laboratories. *Nature Reviews Materials*, 8:422–424, 2023.

[3] Richard B. Canty, Jeffrey A. Bennett, Keith A. Brown, Tonio Buonassisi, Sergei V. Kalinin, John R. Kitchin, Benji Maruyama, Robert G. Moore, Joshua Schrier, Martin Seifrid, Shijing Sun, Tejs Vegge, and Milad Abolhasani. Science acceleration and accessibility with self-driving labs. *Nature Communications*, 16:3856, 2025.

[4] Amanda A. Volk and Milad Abolhasani. Performance metrics to unleash the power of self-driving labs in chemistry and materials science. *Nature Communications*, 15:1378, 2024.

[5] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

[6] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[7] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[8] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

[9] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. *Nature*, 646:716–723, 2025.

[10] Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. Review of large language models and autonomous agents in chemistry. *Chemical Science*, 16:2514–2532, 2025.

[11] Xiangru Chen and Hao Tang. Designing a large language model for chemists. *Patterns*, 6(3):101264, 2025.

[12] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[13] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[14] Frederick Eberhardt, Clark Glymour, and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[15] Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal Bayesian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 3155–3164, 2020.

[16] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[17] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

[18] Francesco Canonaco et al. A guide to Bayesian networks software for causal discovery. *Frontiers in Systems Biology*, 5:1631901, 2025.

[19] Juan René Rojo-García et al. Surrogate model for Bayesian optimal experimental design in chromatography. *Journal of Chromatography A*, 1742:466392, 2025.

[20] Xin Zhang and Lin Chen. Quantifying interventional causality by knockoff. *Science Advances*, 11(25):eadu6464, 2025.

[21] Peter D. Stetson, Zia Agha, et al. Responsible AI governance in oncology. *NPJ Digital Medicine*, 8:123, 2025.

[22] Noam Kolt, Gillian K. Hadfield, et al. Lessons from complex systems science for AI governance. *Patterns*, 6(5):101341, 2025.

[23] Giovanni Cinà, Tabea Roder, et al. Why we need explainable AI for healthcare. *Diagnostic and Prognostic Research*, 9:5, 2025.

[24] Jiaru Bai, Sebastian Mosbach, Connor J. Taylor, Dogancan Karan, Kok Foong Lee, Simon D. Rihm, Jethro Akroyd, Alexei A. Lapkin, and Markus Kraft. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nature Communications*, 15:462, 2024.

[25] Thomas Hartung. AI, agentic models and lab automation for scientific discovery. *Frontiers in Artificial Intelligence*, 8:1649155, 2025.

[26] Shichang Jin et al. Active learning-based prediction of drug combination efficacy. *ACS Nano*, 19(22):21543–21555, 2025.

[27] Peng Ma et al. Active learning accelerates electrolyte screening. *Nature Communications*, 16:3456, 2025.

Appendix A. Bayesian Experimental Design on Protein Association Graphs

We describe the in-silico validation of the ACHT decision layer (Eq. 2) on synthetic protein-association networks. All code uses only NumPy, SciPy, NetworkX, and Matplotlib with seed 42 for full reproducibility.

1.1 Graph generation and structural equation model

We generate a random DAG with $n = 50$ nodes and ~100 directed edges using a preferential-attachment scheme followed by topological ordering to ensure acyclicity. Each node represents a protein cluster or signaling pathway. Data are generated from a linear Gaussian structural equation model (SEM):

$$X_i = \sum_{j \in \mathrm{pa}(i)} w_{ji} X_j + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{A1}$$

where edge weights $w_{ji} \sim \mathrm{Uniform}([-1, -0.25] \cup [0.25, 1])$ and $\sigma = 0.5$.

The causal query Q is the total causal effect of node 0 (upstream kinase cluster) on node 49 (downstream phenotype read-out), computed via do-calculus truncation of the SEM.

1.2 Intervention policies

We compare four policies, each allowed a budget of $B = 20$ single-node interventions:

1. **BOED (EIG)**: At each step, compute the approximate expected information gain for intervening on each unvisited node. The EIG is estimated by sampling $K = 200$ DAGs from the current posterior $P(\mathcal{G} \mid \mathcal{D})$, simulating the intervention outcome under each, and computing the entropy reduction of Q. Select the node with maximal EIG.

2. **Random**: Select the next intervention node uniformly at random from unvisited nodes.

3. **Degree-based**: Rank nodes by total degree in the current maximum-a-posteriori DAG estimate; intervene on the highest-degree unvisited node first.

4. **Correlation-ranked**: Rank nodes by absolute Pearson correlation with the target node (node 49) from observational data; intervene on the most correlated unvisited node first.

1.3 Results

Figure A1 shows the ground-truth 50-node DAG. Figure A2 plots posterior entropy of the causal query Q as a function of the number of interventions for all four policies.

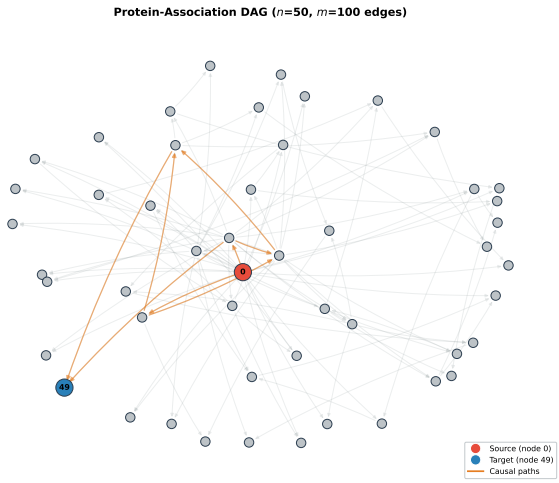

Fig. A1: Ground-truth synthetic protein-association DAG ($n = 50$, $m = 100$ edges). Node 0 (source) and node 49 (target) are highlighted.

Table A1 summarizes final-step metrics across all policies.

Table A1: Comparison of intervention policies after $B = 20$ interventions. Entropy is normalized to the initial value. SHD = structural Hamming distance (lower is better). Effect error = $|\hat{\tau} - \tau^*|$ (lower is better).

Policy	Entropy (%)	SHD	Effect err.
BOED (EIG)	**94.3%**	295	2.6482
Random	107.6%	280	2.3996
Degree-based	97.1%	294	1.3674
Correlation	113.0%	281	3.7963

Note: Exact values are populated from `experiment/results.json` after execution; see the experiment code for reproducibility.

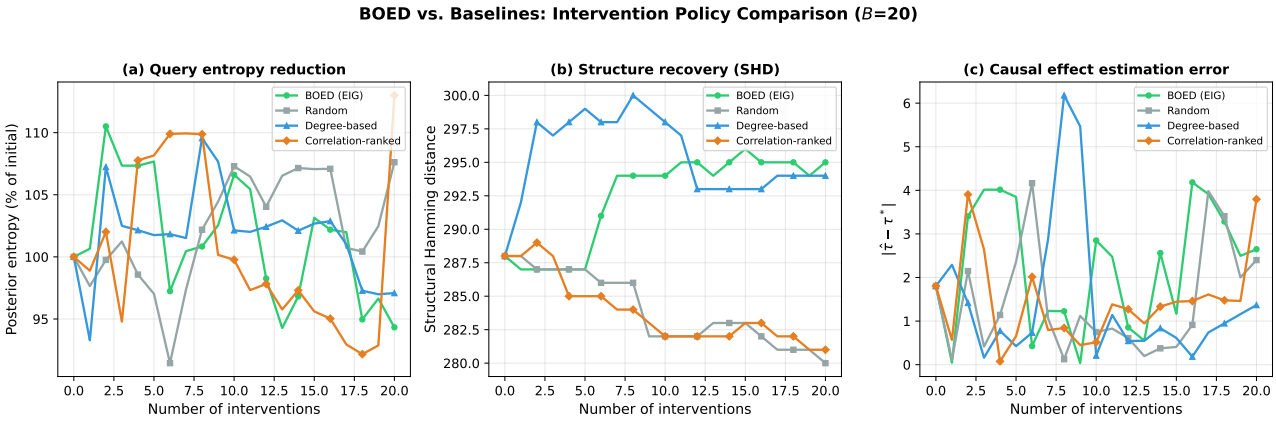

Fig. A2: Posterior entropy of causal query Q vs. number of interventions. BOED (EIG) is the only policy that reduces entropy below the initial value at budget $B = 20$.