NEURONDISCOVER: AGENT-IN-TWIN FOR MECHANISTIC DISCOVERY IN NEURONAL MICROENVIRONMENTS WITH WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic discovery in neuronal microenvironments requires identifying interventions and observations that discriminate between competing biological explanations. We formulate this problem as joint inference over mechanistic hypotheses and digital twin states, and propose NEURONDISCOVER, an *Agent-in-Twin* framework that integrates scientific reasoning, intervention planning, and evidence acquisition within a mechanism-grounded *World Action Model (WAM)*. NEURONDISCOVER represents biological knowledge through a structured *Mechanism–Intervention–Observation–Outcome (MIOY)* graph, where experimental evidence updates mechanistic beliefs and guides targeted twin refinement. The agent generates executable intervention programs with explicit verification criteria and adapts its discovery strategy according to uncertainty and observation constraints. We evaluate NEURONDISCOVER on controlled brain-fluid transport environments with independently verified reference simulations and donor-disjoint cortical neuron electrophysiology recordings. Experiments demonstrate that Agent-in-Twin enables more accurate mechanism identification, selective twin refinement, and efficient intervention prioritization under limited observation budgets. This work introduces an adaptive scientific agent paradigm in which digital twins evolve as mechanistic models through evidence-driven interaction, enabling autonomous discovery in complex neuronal systems. The code is available at `https://anonymous.4open.science/r/NeuronDiscover`.

1 INTRODUCTION

Mechanistic discovery in neuronal systems is fundamentally a decision problem. When multiple mechanistic explanations are consistent with current observations, the central challenge is not only predicting future states, but determining which intervention or measurement can distinguish competing hypotheses. In neural microenvironments, identical phenotypes may emerge from distinct processes, including altered transport, cellular interactions, or observation artifacts (Iliff et al., 2012; Mestre et al., 2018; Smith et al., 2017). Therefore, a discovery system must reason not only about the biological process being modeled, but also about whether its own computational representation is sufficient to explain the available evidence.

Recent advances in biological modeling and self-driving laboratories have created opportunities for closed-loop scientific discovery by coupling computational models with experimental evidence (Dama et al., 2023; Rapp et al., 2024). Digital twins provide a promising foundation for this paradigm by representing biological systems through mechanistic models that can generate predictions, evaluate hypotheses, and guide interventions. However, biological systems are often observed under sparse and incomplete measurements, creating an unresolved ambiguity: a mismatch between a digital twin and reality may indicate either an incorrect biological mechanism or an insufficient model representation. We refer to this ambiguity as *twin confounding*. Resolving twin confounding requires a shift from fixed predictive twins toward a *twin-in-the-loop* paradigm, where digital twins continuously evolve through evidence acquisition, intervention, and external verification.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce NEURONDISCOVER as an Agent-in-Twin system that realizes this twin-in-the-loop paradigm for mechanistic discovery. Unlike conventional predictive models that operate within fixed simulators, NEURONDISCOVER couples three components: a mechanism-grounded World Action Model (WAM) for representing action-conditioned biological dynamics and observations; an Agent that selects evidence-acquisition actions under competing mechanism–discrepancy hypotheses; and a revisable Mechanism–Intervention–Observation–Outcome (MIOY) graph that stores validated relations and supports executable intervention programs. Together, these components enable a scientific system that can determine not only what may happen, but also what evidence is required to refine its understanding of the biological system. We study the computational foundations of Agent-in-Twin systems for mechanistic scientific discovery. We evaluate whether such systems can identify mechanisms, distinguish biological uncertainty from twin uncertainty, and prioritize informative interventions under limited experimental budgets. Our evaluation combines controlled neural microenvironment worlds with donor-disjoint neuronal recordings, providing a testbed for discovery capability beyond predictive accuracy alone. Our contributions are:

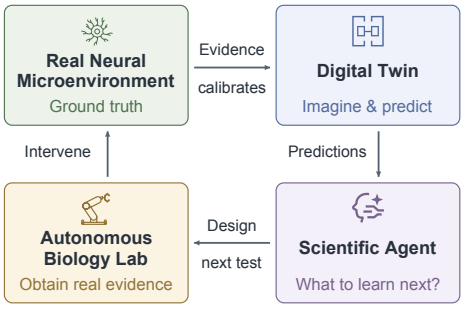


Figure 1: **Agent-in-Twin enables adaptive discovery in complex biological systems.** A digital twin is coupled with evidence acquisition and external verification through an Agent-in-Twin loop, allowing mechanistic hypotheses and model representations to be iteratively refined.

- **A new Agent-in-Twin paradigm for mechanistic discovery under twin confounding.** We introduce Agent-in-Twin as a framework where scientific agents operate with revisable digital twins rather than fixed predictive simulators. We formalize twin confounding and formulate discovery as joint reasoning over biological mechanisms and computational twin discrepancies.
- **A mechanism-grounded system for adaptive twin revision.** We develop NEURONDISCOVER, which integrates a World Action Model, evidence-acquisition agent, and typed MIOY representation into a unified discovery loop. Through independently verified updates and scoped revisions, NEURONDISCOVER enables agents to refine mechanistic models and prioritize interventions based on acquired evidence.
- **A discovery-oriented evaluation framework for scientific agents.** We introduce a matched-budget evaluation protocol that measures mechanism recovery, unsupported hypothesis control, and intervention efficiency rather than predictive accuracy alone. The benchmark spans controlled reference worlds and donor-disjoint neuronal recordings to assess whether Agent-in-Twin systems can generalize beyond fixed simulation settings.

2 RELATED WORK

Microenvironment and world models. Transport studies motivate competing diffusion, flow, and exchange hypotheses (Iliff et al., 2012; Mestre et al., 2018; Smith et al., 2017). Inverse imaging and hybrid reservoir–Hodgkin–Huxley models reconstruct fields or correct response dynamics (Bakiler et al., 2026; Williams et al., 2025), and structural-identifiability mappings characterize indistinguishable parameters (Norden et al., 2025). WAMs couple prediction and action generation (Wang et al., 2026), with compatibility assessed across queries (Ruan et al., 2026). GC-IDM and ACID provide inverse planning and action-consistency controls (Nguyen et al., 2026; Seo et al., 2026), and physical learning and simulator-grounded predictors supply mechanistic dynamics (Raissi et al., 2019; Pfaff et al., 2021; Dudley and Eisenberg, 2025). NEURONDISCOVER instead scores observations within the shared-joint WAM, using experiment-dependent sensitivities and response envelopes to separate mechanisms from twin discrepancy; independent outcomes determine support.

Experiment design and scientific agents. Bayesian and amortized design select informative experiments (Rainforth et al., 2024; Foster et al., 2021); BAD-PODS marginalizes latent states with nested filters (Pérez-Vieites et al., 2025). Nuisance-aware, robust, and discrepancy-learning objectives address additional uncertainty (Sloman et al., 2024; Go and Isaac, 2022; Yang et al., 2025), and causal design targets graphs or interventions (Agrawal et al., 2019; Aglietti et al., 2020). Scientific

agents connect hypotheses to executable tests (Boiko et al., 2023; Ma et al., 2024; Jansen et al., 2024; Majumder et al., 2025); Model Discovery Agent combines nested inference, model proposals, and experimental design (Murphy, 2026), and NIMMGen searches neural-integrated mechanistic models (Guan et al., 2026). NEURONDISCOVER couples acquisition and hypothesis expansion to scoped MIOY revision and discrepancy-adjusted verification, separating fixed-library inference from omitted-relation recovery and extending the coverage–risk view of selective prediction (Geifman and El-Yaniv, 2017) to bound events for complete intervention programs.

3 BACKGROUND

3.1 PROBLEM SETTING

A neuronal microenvironment is a partially observed controlled system with state S_t, mechanism description H, physical intervention I_t, observation action O_t, context C, and validity state V_t. The combined action is $A_t = (I_t, O_t)$. In a transport domain, S_t may contain concentration, pressure, flow, geometry, and exchange states; Y_t is the released measurement, and a target G is defined through a functional of the state or observation trajectory. The constraint set Γ specifies admissible interventions, observation support, horizon, and cost.

The discovery object is a graph

$$\mathcal{G}_t = (\mathcal{M}_t, \mathcal{I}, \mathcal{O}, \mathcal{Y}, \mathcal{E}_t), \tag{1}$$

whose edges express mechanism–intervention–observation–outcome relations with scope, uncertainty, support, and counterevidence. A mechanism description specifies a process, interaction, constitutive relation, or parameter regime, and discovery expands this vocabulary through new hypotheses, conditions, and observables within finite per-query candidate sets.

A policy seeks correctly resolved relations while controlling false support and experimental cost:

$$\max_{\pi} \mathbb{E}\left[\mathrm{Res}(\mathcal{G}_T) - \lambda_s \,\mathrm{False}(\mathcal{G}_T) - \lambda_c \sum_{t<T} C(A_t)\right], \qquad \sum_{t<T} C(A_t) \le B. \tag{2}$$

An independent reference assesses correctness in the declared model world, its labels reserved for evaluation: Equation (1) records what has been learned and Equation (2) scores the process.

3.2 MECHANISTIC DYNAMICS AND WORLD ACTION MODELS

The operational twin combines coarse physical dynamics, a learned residual, and an observation model:

$$\begin{aligned} S_{t+1} &= F_{\mathrm{mech}}(S_t, I_t, H, C) + R_\theta(S_t, I_t, H, C, \epsilon_t), \\ Y_t &= \mathcal{M}_\psi(S_t, O_t, C) + \nu_t. \end{aligned} \tag{3}$$

A scientific WAM represents the typed joint law

$$Q_\theta(S_{0:T}, H, A_{0:T-1}, Y_{0:T}, G, V \mid C), \tag{4}$$

with a positive density on its admitted support. States, actions, observations, and outcomes retain their physical units and meanings. Forward and inverse requests condition the same joint object:

$$Q_\theta(S_{1:T}, Y_{1:T} \mid S_0, H, A, C) \quad \text{(forward)}, \tag{5}$$

$$Q_\theta(H, A \mid S_0, G, C, \Gamma) \quad \text{(inverse proposal)}. \tag{6}$$

Goal conditioning in Equation (6) generates candidates; evidence updates the belief about H.

3.3 MECHANISM UNCERTAINTY AND TWIN CONFOUNDING

The belief $b_t(H, Z, N, S_t)$ includes twin discrepancy Z, nuisance variables N, and current state. For experiment e, explanation $\xi = (h, z)$ induces a response law P^e_ξ after marginalizing nuisance variables. Define

$$d_{\mathcal{E}}(\xi, \xi') = \sup_{e \in \mathcal{E}} d(P^e_\xi, P^e_{\xi'}). \tag{7}$$

Definition 1 (Twin confounding)**.** Two explanations are observationally indistinguishable at resolution ε when $d_{\mathcal{E}}(\xi, \xi') \le \varepsilon$. Twin confounding occurs when such a pair has both $h \neq h'$ and $z \neq z'$, so the available experiments do not separate mechanism differences from compensating twin error.

Thresholded indistinguishability is pairwise and need not be transitive. Locally, the diagnostic objective is

$$\min_{\pi, \widehat{H}_T, \widehat{Z}_T} \mathbb{E}[\ell_H(\widehat{H}_T, H) + \lambda_Z \ell_Z(\widehat{Z}_T, Z) + \lambda_V \ell_V], \quad \mathbb{E}\left[\sum_{t<T} C(A_t)\right] \le B. \tag{8}$$

This local objective guides graph revision.

Scientific inverse queries return candidate programs, required observations, and reachability status (Equation (33)); independent verification determines support.

4 NEURONDISCOVER

NEURONDISCOVER revises an MIOY graph by linking mechanism hypotheses to interventions and discriminating observations. The WAM predicts candidate consequences; independent outcomes determine scientific support. Figure 2 summarizes this discovery cycle.

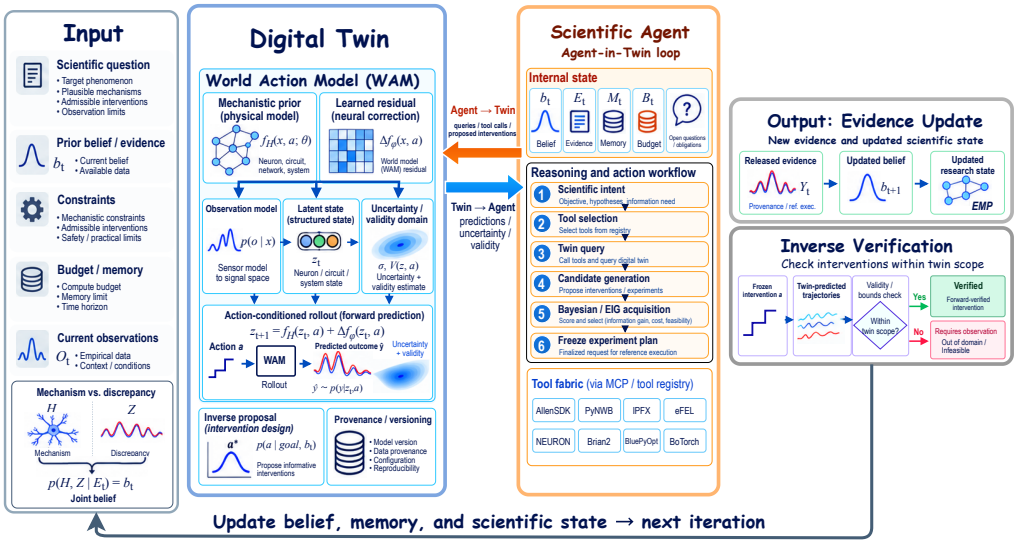


Figure 2: **NeuronDiscover overview.** Questions, observations, prior evidence, and constraints initialize the twin. A mechanism-grounded World Action Model (WAM) supplies forward predictions and inverse proposals with uncertainty and validity estimates. The Agent freezes an intervention–observation plan before isolated reference execution, and released outcomes update the mechanism–discrepancy belief and the Mechanism–Intervention–Observation–Outcome (MIOY) graph. Forward verification assesses complete programs; evidence-closed relations compile into executable mechanistic programs (EMPs), and the updated state guides the next experiment.

4.1 A MECHANISM-GROUNDED COMPUTATIONAL WORLD

The backbone in Equation (3) supplies physical transitions and boundaries, with a learned residual for unresolved dynamics. GlymphTwin interventions specify transport or exchange targets, magnitude, duration, spatial support, and cost. Observations specify variables, locations, times, and measurement operators. Distinct discrepancy coordinates represent missing dynamics, boundary error, and observation error.

Each finite-volume cell and reservoir at each time retains a scalar token, with physical units and space–time coordinates. Training-fitted codebooks discretize values; the physical solver and learned residual decode them in the original units. Shared denoising represents state, mechanism, intervention, observation, goal, and validity blocks. The six losses couple masked prediction to decoded dynamics,

physical balance, intervention contrasts, query consistency, and validity:

$$\begin{aligned}\mathcal{L}_{\mathrm{WAM}} = \mathcal{L}_{\mathrm{joint}} + \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dec}} + \lambda_{\mathrm{phys}}\mathcal{L}_{\mathrm{phys}} + \lambda_{\mathrm{int}}\mathcal{L}_{\Delta} \\ + \lambda_{\mathrm{coh}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{val}}\mathcal{L}_{\mathrm{valid}}.\end{aligned} \tag{9}$$

For example, a forward query fixes (S_0, H, I, O) and generates future states and measurements. An inverse query fixes (S_0, G, Γ) and proposes (H, I, O). Sensing compares candidate operators O through their predicted measurements at fixed (S_0, H, I). Every denoising step restores the fixed blocks. Sampling builds on discrete and masked diffusion Austin et al. (2021); Sahoo et al. (2024); Chen et al. (2024); Appendix D defines losses and finite-query checks.

Appendix D.1 specifies the codec, gradient paths, and independent field-reconstruction checks. If the population query laws are conditionals of one positive joint WAM they are Bayes-compatible on their common support (Proposition 3). That statement constrains the population object, not the implemented finite-step routes; the implemented guarantee is Lemma 2, conditional on assumed per-route error bounds. Conditional-route disagreement is the empirical diagnostic of that assumption and held-out error measures fidelity; neither certifies it (Appendix K.3).

4.2 A SCIENTIST THAT REVISES THE DISCOVERY GRAPH

The Agent maintains a belief over (H, Z, N, S_t), graph $\mathcal{G}_t$, evidence K_t, and budget B_t. We specify a particle instantiation with state/nuisance marginalization (Appendix D.4). Relations carry executable scope predicates and prospective tests. An unexplained tracer curve can motivate a new exchange hypothesis or spatial observable; the new relation is tested on subsequent outcomes.

The WAM screens executable proposals, from which the policy selects an action and query template:

$$(A_t, \kappa_t) \sim \pi_\omega(\cdot \mid b_t, \mathcal{G}_t, K_t, B_t). \tag{10}$$

Interventions change physical state, released observations update belief and graph support, and desired futures guide proposals.

Before the outcome, the selected action, endpoint, falsifier, and analysis are fixed:

$$s_t = \mathrm{Freeze}(A_t, \kappa_t, \mathcal{A}_t, \mathcal{T}_t), \tag{11}$$

$$Y_t \sim P^\star(\cdot \mid s_t), \qquad K_{t+1} = \mathcal{R}(K_t, s_t, Y_t). \tag{12}$$

The reference is an independent simulator or registered measured response. In the transport task, the reducer retains compatible physical and sensor explanations. A mechanism relation gains support when every surviving explanation satisfies both its mechanism predicate and scoped intervention effect (Appendix A.4). Status, counterevidence, and scope guide the next proposal. Because the action, endpoint, falsifier and analysis are fixed before release, a conditionally super-uniform test for the realized selection keeps $\Pr(p_{s_t} \le \alpha) \le \alpha$ despite arbitrary pre-outcome proposal search (Proposition 6); across rounds, the sequential correction in Appendix A.4 controls accumulated testing error.

4.3 CHOOSING INTERVENTIONS AND INFORMATIVE OBSERVATIONS

In declared physical coordinates, local mechanism and discrepancy sensitivities satisfy

$$\begin{aligned} r_e &= J_e^{\mathrm{mech}}\Delta h + J_e^{\mathrm{wam}}\Delta z + \epsilon_e + o(\|\Delta\xi\|), \\ \Lambda(\mathcal{E}) &= \sum_{e\in\mathcal{E}} J_e^\top \Sigma_e^{-1} J_e, \qquad \xi = (h, z). \end{aligned} \tag{13}$$

Proposition 1 (Local mechanism–WAM identifiability)**.** *Under the differentiable local mean-response model with positive-definite weights,* $(\Delta h, \Delta z)$ *is identifiable to first order if and only if* $\Lambda(\mathcal{E}) \succ 0$.

Scaled finite differences estimate the joint Jacobian. Subtracting operator uncertainty from its smallest singular value gives the unregularized information floor. Full-response comparisons and a nonlinear remainder test determine where that local guide is useful (Appendix D.4). Once the floor is met, the Agent ranks admissible experiments by

$$\alpha_t(e) = I(Y_e; H, Z \mid e, K_t) + \lambda_F F_t(e) - \lambda_C C(e) - \lambda_V R_V(e). \tag{14}$$

Here F_t rewards a specified falsification opportunity. Below the floor, design improves the weakest direction, breaking ties by α_t. Nested likelihood averages marginalize state and nuisance variables; particle and sampling errors are checked separately.

For a fixed belief reducer $\mathcal{U}$ and bounded joint risk $\mathfrak{R}$, the corresponding operational utility is

$$U_t^Q(e) = \mathbb{E}_{Y\sim Q_e}[\mathfrak{R}(b_t) - \mathfrak{R}(\mathcal{U}(b_t; e, Y))] - \lambda_C C(e) - \lambda_V R_V(e). \tag{15}$$

Proposition 2 (Acquisition stability)**.** *If the risk reduction has range width at most M and $d_{\mathrm{TV}}(P_e^\star, Q_e) \le \delta_e$, then $|U_t^{P^\star}(e) - U_t^Q(e)| \le M\delta_e$. A ranking margin exceeding $M(\delta_e + \delta_{e'})$ preserves the ordering of e and e'.*

For bounded entropy and the Bayesian reducer under Q_e, the operational expected reduction equals mutual information. The transfer bound evaluates that same reducer under $P_e^\star$. Uncertain rankings motivate calibration or a better observation.

4.4 Discovering interventions by forward verification

Given $(S_0, G, \Gamma, b_t, K_t)$, the WAM proposes complete programs, including observation-dependent branches, eligibility, and costs. Operational rollouts evaluate goal attainment and violations.

For a nonempty frozen set $\mathcal{A}_K$ of K programs, let $\widehat{p}_g(a)$ and $\widehat{p}_v(a)$ be goal and violation frequencies from $n_a > 0$ conditionally independent rollouts with sample counts fixed in advance. With $\epsilon_a = \sqrt{\log(2K/\alpha)/(2n_a)}$ and an event-probability discrepancy bound δ_a, simultaneously with probability at least $1 - \alpha$,

$$\begin{aligned} P_a^\star(G) &\ge L_g(a) = \widehat{p}_g(a) - \epsilon_a - \delta_a, \\ P_a^\star(V) &\le U_v(a) = \widehat{p}_v(a) + \epsilon_a + \delta_a. \end{aligned} \tag{16}$$

A program is accepted when $L_g(a) \ge \tau_g$, $U_v(a) \le \tau_v$, and validity conditions hold. Two-sample event calibration over the frozen program set supplies simultaneous δ_a bounds with failure budget β (Appendix J.4); fresh verification gives total failure at most $\alpha + \beta$. Uncovered conditions remain unresolved. Infeasibility requires exclusion over the search domain.

4.5 From experimental outcomes to executable knowledge

Evidence-closed relations compile into Executable Mechanistic Programs (EMPs) carrying interventions, observation triggers, validity conditions, and stopping rules. Compilation preserves each relation's scope, uncertainty, counterevidence, and source status. Algorithm 1 links these operations. Appendix D.7 illustrates the prompts and MCP exchange used to request Twin predictions and propose the next observation.

5 Experimental Setup

We ask four questions. **RQ1:** can the Agent discover and revise microenvironment relations? **RQ2:** do shared queries and adaptive observations improve discovery? **RQ3:** which components and LLM backbones support discovery and transfer?

Domains and scientific units. The transport setting specifies worlds varying diffusion, advection or dispersion, forcing, geometry, boundary exchange, and observation operators, with public studies grounding the questions and parameter ranges Iliff et al. (2012); Smith et al. (2017); Mestre et al. (2018); Hablitz et al. (2020) and independent reference worlds supplying outcomes. The recording setting tests transfer to neuronal stimulus–response relations using public morphology, current-clamp recordings, and stimulus metadata Teeters et al. (2015); Gouwens et al. (2019); Allen Institute for Brain Science (2026).

Tasks, observations, and endpoints. Primary endpoints are resolved MIOY relations per budget, false support, scope accuracy, and omitted-relation recovery. Transport observations span mass, regional curves, spatial profiles, and boundary-sensitive channels. Recording actions select a current sweep and a voltage, spike-count, or interspike-interval window to test input resistance, firing gain, or adaptation; fixed-graph and pooled-scope controls share donor assignments, sweep menu, and budget. Withheld sweeps test response prediction and held-out donors test transfer, while ionic-mechanism

claims require separate perturbation evidence. WAM fidelity, goal success, coverage, selective risk, and end-to-end cost complete the evaluation (Appendices J and L). Table 1 separates how the evaluation data are composed from how comparisons are estimated. The sampling spine is identical across settings, so every contrast is paired within a source and no sub-unit inflates the effective sample size.

Table 1: **Statistical profile of the benchmark.** Composition (upper block) and analysis protocol (lower block). Every entry is fixed before evaluation, not measured. A spanning entry applies to both settings; — marks an undefined quantity.

Quantity	Transport microenvironment worlds	Neuronal current-clamp recordings
Benchmark composition		
Independent source unit	Generator family	Biological donor
Source units per configuration	32, split before episodes are constructed	
Replicates; shared budget	Three per source, averaged before paired estimation; 16 experiments/task	
Simulation mesh	17 cells; 136-cell reference; $68/136/272$ refinement	—
Training / development pairs	512 from 64 families / 64 from 16 disjoint families	—
Prespecified predicate tasks	—	Three: input resistance, firing gain, adaptation
Analysis protocol		
Dispersion reported in tables	Mean and sample SD across source-unit summaries	
Intervals and tests	95% cluster bootstrap on paired differences; paired randomization, within-family Holm	
Power basis	80% power, unadjusted two-sided 0.05 (factor $\approx$7.85); paired SD 1, effect 0.5 imply the 32 units	
Unsupported and failed tasks	Retained in applicability and recovery denominators	Unavailable endpoints retained in prespecified denominators
Basis of reported evidence	Every reported comparison uses the 32 sources above; no claim rests on a larger design	

Models and controls. Capacity-matched separate forward/inverse models, action consistency, and goal-conditioned inverse maps isolate representation sharing from proposal quality. Design controls include fixed sensing, plug-in EIG, BAD-PODS-style nested filtering Pérez-Vieites et al. (2025), and joint mechanism–discrepancy EIG. Agent-in-Twin and an external tool agent share tools, scientific budgets, and a fixed WAM under Kimi K3, DeepSeek V4 Flash, and GPT-5.6-Sol (Appendix I.2). Inverse controls include shooting, cross-entropy search, and constrained Bayesian optimization. Verifiers assess identical frozen programs. Appendices A and H specifies adapters and configurations.

Evaluation protocol and statistical analysis. Methods share evidence, admissible queries, goals, and source-disjoint splits. Hidden labels remain evaluator-only; certification uses frozen programs and separate calibration/verification batches (Appendices A and J).

6 Results and Discussion

6.1 RQ1: Relation discovery and graph revision

Graph revision raises recovery from 3.2 to 4.0 certified relations per assigned world and scope accuracy from 75% to 82%, while cutting false support from 7.0% to 5.0% (Table 2). Headline contrasts are source-level paired differences with cluster-bootstrap intervals and Holm-corrected randomization tests (Table 3), decomposed by stratum in Table 10.

6.2 RQ2: Shared queries and adaptive observation design

Mechanistic grounding improves admissible intervention-delta fidelity by 9 percentage points, and joint discrepancy modeling cuts false-support incidence by 6 points (Table 4). A stronger forward predictor does not by itself deliver discovery: the neural operator attains the best deterministic rollout error in Table 7 yet trails the hybrid by 0.10 relation-status macro-F1 under the same Agent and adapter (Tables 3 and 8), a gap concordant with calibrated uncertainty rather than point accuracy.

Table 2: **Relation discovery at matched cost.** Mean $\pm$ SD over 32 sources. Resolutions count certified support plus certified falsification per assigned world; abstentions are unresolved for every policy and censored at the budget (Table 19). False support is the proportion of declared supports the reference contradicts; scope is balanced accuracy (%). Grey bold marks the best mean, ties included.

Policy	Resolutions ↑	False support (%) ↓	Scope accuracy (%) ↑	Cost ↓
Random design Rainforth et al. (2024)	2.2 ± 0.7	12.0 ± 4.5	64.0 ± 8.2	15.4 ± 2.4
Bayesian adaptive design Rainforth et al. (2024)	3.0 ± 0.7	9.0 ± 3.8	71.0 ± 7.1	13.5 ± 2.1
Discrepancy-aware design Yang et al. (2025)	3.2 ± 0.7	7.5 ± 3.1	75.0 ± 6.8	14.5 ± 2.3
Fixed hypothesis graph	3.2 ± 0.8	7.0 ± 3.4	75.0 ± 7.3	12.4 ± 1.9
Matched external tool agent Yao et al. (2023)	3.4 ± 0.7	8.0 ± 3.7	74.0 ± 7.0	14.4 ± 2.2
Agent-in-Twin (ours)	$\mathbf{4.0 \pm 0.7}$	$\mathbf{5.0 \pm 2.8}$	$\mathbf{82.0 \pm 6.2}$	$\mathbf{11.4 \pm 1.8}$
Deterministic Agent-in-Twin (ours)	3.8 ± 0.7	5.5 ± 3.0	80.0 ± 6.5	11.9 ± 1.8

Table 3: **Source-level paired contrasts for the headline claims.** Mean within-source difference over 32 sources with a cluster-bootstrap 95% interval and a Holm-corrected randomization p-value within each rule-separated family. Resolution rows use the certified scoring of Equation (46); abstention loads are comparable, so the differences are unchanged by it though the levels are not. Paired dispersion comes from source-level differences, not the marginal SDs of Table 2.

Contrast	Difference	95% CI	Holm p
Graph revision vs. fixed hypothesis graph (resolutions)	$+0.80$	$[+0.57, +1.03]$	< 0.001
vs. strongest baseline (resolutions)	$+0.60$	$[+0.41, +0.79]$	< 0.001
vs. deterministic orchestration control (resolutions)	$+0.20$	$[+0.06, +0.34]$	0.006
Joint vs. plug-in state EIG (resolutions)	$+0.90$	$[+0.65, +1.15]$	< 0.001
Joint vs. nested-filter EIG (resolutions)	$+0.60$	$[+0.41, +0.80]$	< 0.001
Hybrid vs. FNO adapter (status macro-F1)	$+0.10$	$[+0.08, +0.12]$	< 0.001
Discrepancy-adjusted verification (risk, pp)	-6.0	$[-7.59, -4.41]$	< 0.001
Neuronal transfer vs. fixed graph (yield)	$+0.41$	$[+0.24, +0.58]$	< 0.001

Table 4: **Within-method ablations** (Appendix I). Mean paired difference over 32 sources (full model minus control), its source-level SD, a cluster-bootstrap 95% interval, and a Holm-corrected randomization p-value over the six rows. Positive favors recovery/fidelity, negative false support/cost/risk.

Component control	Endpoint	Mean $\pm$ SD	95% CI	Holm p
Fixed hypothesis graph	Missing-relation recovery (pp)	8.0 ± 6.8	$[+5.55, +10.45]$	< 0.001
No mechanistic backbone	Admissible intervention-delta fidelity (pp)	9.0 ± 6.4	$[+6.69, +11.31]$	< 0.001
Separate forward and inverse models	Relation recovery (pp)	5.0 ± 5.7	$[+2.94, +7.06]$	< 0.001
Mechanism belief without discrepancy	World-level false-support incidence (pp)	-6.0 ± 4.6	$[-7.66, -4.34]$	< 0.001
Fixed observation schedule	Restricted resolution cost (cost units)	-2.0 ± 1.8	$[-2.65, -1.35]$	< 0.001
Nominal verification	Accepted-program risk (pp)	-6.0 ± 4.4	$[-7.59, -4.41]$	< 0.001

Adaptive measurements reach 4.0 correct relations per assigned world at cost 16 against 2.8 for fixed sensing on the same WAM (Figure 6a). Under the harder nonlinear-discrepancy condition, joint mechanism–discrepancy EIG resolves 3.8 relations per assigned world versus 2.9 for plug-in EIG at 5% versus 11% false support, with a paired advantage of $+0.60$ over nested-filter EIG (Table 3); its margin over Inside-Out SMC^2 and PASOA is within one source-level SD (Table 5). Its online time ratio is 2.9, so the gain is bought with computation rather than experiments.

6.3 RQ3: Components, LLM backbones, and transfer

Table 4 and Figure 3 isolate component contributions: removing MIOY role typing, scope constraints, or revision rules increases false support.

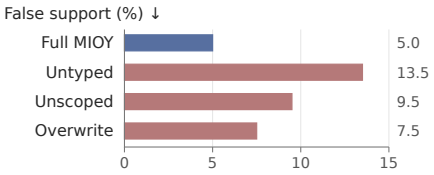


Figure 3: **Each MIOY component, removed.** False support per assigned world.

The language model is an optional proposal interface with a localised benefit; Figure 10 and Appendix I.2 compare three backbones with the WAM, tools, evidence, splits, and call caps fixed. Stratified by task family (Table 11), its aggregate $+0.20$ relations per assigned world concentrates in tasks that extend the hypothesis vocabulary ($+0.60$ on omitted-relation proposal), with intervals covering zero elsewhere; where the vocabulary is closed the deterministic controller is preferable and cheaper. Recovery of 2.85 relations per applicable world at 68% applicability gives 1.94 per assigned world against 1.53 for the fixed-graph control (Table 10), with stratified weights in Appendix J.1.

6.4 Case study: an Agent-in-Twin discovery loop

Figure 4 traces a recorded boundary-exchange episode. Eleven explanations attribute an AQP4-proxy readout change to exchange κ_I, diffusivity D_I, or gain a_j. Relation 17 requires an exchange-mediated removal effect of at least $\Delta_r = 0.08$ at $T = 1200\,\mathrm{s}$ within its declared scope.

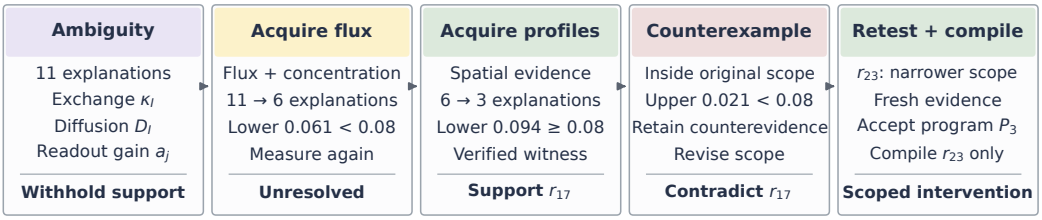


Figure 4: **An Agent-in-Twin loop unfolded.** Counts denote surviving explanations; inequalities compare certified effect bounds with 0.08. Trace and P_3 certificate: Appendix K.4 and Table 21.

Acquire evidence, then choose again. The Agent freezes the intervention pair and a co-located flux–concentration panel. Independent outcomes leave six explanations agreeing on the mechanism, but the effect lower bound is $0.061 < 0.08$. It withholds support and acquires spatial profiles to separate exchange from diffusion. Three explanations survive; the bound reaches $0.094 \geq 0.08$ with a verified witness, supporting relation 17.

Revise the relation, then verify the program. A later context lies inside the original scope but yields an effect upper bound of $0.021 < 0.08$. The Agent retains this counterexample, marks relation 17 contradicted, and births relation 23 with an additional wall-amplitude condition, $\geq 1.5\,\mu$m. After fresh tests, only relation 23 enters program P_3, accepted with goal lower bound $0.8387 \geq 0.8$ and violation upper bound $0.0988 \leq 0.12$.

The exchange/readout, pulsatility, and transport MIOY instances in Appendices B.6 to B.8 show the corresponding candidate structures, typed evidence, and scope refinements (Figures 8 and 9).

7 Conclusion

NeuronDiscover establishes an Agent-in-Twin perspective for mechanistic discovery, where an AI system must decide not only what to predict, but also what evidence is needed to refine its understanding of the world. By jointly reasoning over biological mechanisms and digital twin uncertainty, NeuronDiscover enables evidence-driven hypothesis refinement and intervention prioritization in neuronal systems. Our results suggest that future scientific agents should be evaluated by their ability to acquire informative evidence and improve mechanistic models, rather than prediction accuracy alone. This work provides a foundation for developing adaptive computational scientists that interact with evolving biological digital twins.

Ethics Statement

NeuronDiscover studies computational mechanism hypotheses in neuronal microenvironments. The evaluation uses public data and explicitly modeled reference worlds. Generated trajectories are distinguished from measured observations, and evidence retains its source and scope. Virtual intervention programs are research objects, not clinical treatment recommendations. Translation to physical or biological experiments requires domain validation, appropriate ethical review, and human authorization. Public source licenses and access conditions apply to all reused data.

Reproducibility Statement

The scientific variables, model queries, graph updates, and intervention certificates are defined in Sections 3 and 4 and Appendix D. Algorithm 1 specifies the scientist loop, and Appendix F gives the proofs. The empirical protocol in Appendices A, G, J and L defines data separation, baselines, statistical units, and the records required to reproduce a comparison.

AI Use Statement

A large language model assisted with LaTeX drafting and editing, mathematical checks, and compilation checks.

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3155–3164. PMLR, 2020.

Raj Agrawal, Chandler Squires, Karren Yang, Karthikeyan Shanmugam, and Caroline Uhler. ABCD-strategy: Budgeted experimental design for targeted causal structure discovery. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3400–3409. PMLR, 2019.

Allen Institute for Brain Science. Allen cell types database, 2026. Public data resource, accessed 2026-08-04.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A. Derya Bakiler, Michael J. Johnson, Michael R. A. Abdelmalik, Frimpong A. Baidoo, Andrew Badachhape, Ananth V. Annapragada, Thomas J. R. Hughes, and Shaolie S. Hossain. Reconstruction of glymphatic transport fields from subject-specific imaging data, with particular emphasis on cerebrospinal fluid flow and tracer conservation. *arXiv preprint arXiv:2605.00730*, 2026. doi: 10.48550/arXiv.2605.00730.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

Yazan N. Billeh, Binghuang Cai, Sergey L. Gratiy, et al. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron*, 106(3):388–403.e18, 2020. doi: 10.1016/j.neuron.2020.01.040.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0759.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adam C. Dama, Kevin S. Kim, Danielle M. Leyva, Annamarie P. Lunkes, Noah S. Schmid, Kenan Jijakli, and Paul A. Jensen. BacterAI maps microbial metabolism without prior knowledge. *Nature Microbiology*, 8(6):1018–1025, 2023. doi: 10.1038/s41564-023-01376-0.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005. doi: 10.1007/s10479-005-5724-z.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23:138–151, 2020. doi: 10.1038/s41593-019-0550-9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Carson Dudley and Marisa Eisenberg. Learning from simulators: A theory of simulation-grounded learning, 2025. arXiv preprint, version 2.

Alexandre Ern and Martin Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *SIAM Journal on Numerical Analysis*, 53(2):1058–1081, 2015. doi: 10.1137/130950100.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3384–3395. PMLR, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jinwoo Go and Tobin Isaac. Robust expected information gain for optimal bayesian experimental design using ambiguity sets. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 728–737. PMLR, 2022.

Nathan W. Gouwens, Staci A. Sorensen, Jim Berg, et al. Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature Neuroscience*, 22:1182–1195, 2019. doi: 10.1038/s41593-019-0417-0.

Zihan Guan, Rituparna Datta, Mengxuan Hu, Shunshun Liu, Aiying Zhang, Prasanna Balachandran, Sheng Li, and Anil Vullikanti. Are LLMs ready for neural-integrated mechanistic modeling? a benchmark and agentic framework. *arXiv preprint arXiv:2602.18008*, 2026. doi: 10.48550/arXiv.2602.18008.

Lauren M. Hablitz, Virginia Pla, Michael Giannetto, et al. Circadian control of brain glymphatic and lymphatic fluid flow. *Nature Communications*, 11:4411, 2020. doi: 10.1038/s41467-020-18115-2.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Costantino Iadecola. The neurovascular unit coming of age: A journey through neurovascular coupling in health and disease. *Neuron*, 96(1):17–42, 2017. doi: 10.1016/j.neuron.2017.07.030.

Jeffrey J. Iliff, Minghuan Wang, Yonghong Liao, et al. A paravascular pathway facilitates cerebrospinal fluid flow through the brain parenchyma and the clearance of interstitial solutes, including amyloid beta. *Science Translational Medicine*, 4(147):147ra111, 2012. doi: 10.1126/scitranslmed.3003748.

Jacopo Iollo, Christophe Heinkelé, Pierre Alliez, and Florence Forbes. PASOA: PArticle baSed Bayesian Optimal Adaptive design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21020–21046. PMLR, 2024.

Sahel Iqbal, Adrien Corenflos, Simo Särkkä, and Hany Abdulsamad. Nesting particle filters for experimental design in dynamical systems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21047–21068. PMLR, 2024.

Peter Jansen, Marc-Alexandre Cote, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0324.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33940–33962. PMLR, 2024.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, et al. Discoverybench: Towards data-driven discovery with large language models. In *International Conference on Learning Representations*, 2025.

Humberto Mestre, Jeffrey Tithof, Ting Du, Wei Song, Weiguo Peng, Amanda M. Sweeney, Genaro Olveda, John H. Thomas, Maiken Nedergaard, and Douglas H. Kelley. Flow of cerebrospinal fluid is driven by arterial pulsations and is reduced in hypertension. *Nature Communications*, 9:4878, 2018. doi: 10.1038/s41467-018-07318-3.

Model Context Protocol Contributors. Model Context Protocol: Tools specification, 2025. URL `https://modelcontextprotocol.io/specification/2025-11-25/server/tools`. Specification revision 2025-11-25.

Kevin Murphy. Model discovery agent: LLM-assisted bayesian experiment design for data-efficient discovery of mechanistic world models, 2026. arXiv preprint, version 4.

Hoang Nguyen, Xiaohao Xu, and Xiaonan Huang. Latent geometry beyond search: Amortizing planning in world models. *arXiv preprint arXiv:2605.08732*, 2026. doi: 10.48550/arXiv.2605.08732.

Janis Norden, Elisa Oostwal, Michael Chappell, Peter Tino, and Kerstin Bunte. Structure is information: structural identifiability mappings for machine learning with partially observed dynamical systems. *arXiv preprint arXiv:2502.04131*, 2025. doi: 10.48550/arXiv.2502.04131.

Sara Pérez-Vieites, Sahel Iqbal, Simo Särkkä, and Dominik Baumann. Online Bayesian experimental design for partially observed dynamical systems. *arXiv preprint arXiv:2511.04403*, 2025. doi: 10.48550/arXiv.2511.04403.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1):97–107, 2024. doi: 10.1038/s44286-023-00002-4.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, and Hong-Han Shuai. Is the future compatible? diagnosing dynamic consistency in world action models. *arXiv preprint arXiv:2605.07514*, 2026. doi: 10.48550/arXiv.2605.07514.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. doi: 10.1017/S096249291000005X.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Mariano Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4135.

Gawon Seo, Dongwon Kim, and Suha Kwak. ACID: Action consistency via inverse dynamics for planning with world models. *arXiv preprint arXiv:2607.02403*, 2026. doi: 10.48550/arXiv.2607.02403.

Joshua H. Siegle, Xiaoxuan Jia, Severine Durand, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592:86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Sabina J. Sloman, Ayush Bharti, Julien Martinelli, and Samuel Kaski. Bayesian active learning in the presence of nuisance parameters. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 3245–3263. PMLR, 2024.

Alex J. Smith, Xiaoming Yao, James A. Dix, Byung-Ju Jin, and Alan S. Verkman. Test of the 'glymphatic' hypothesis demonstrates diffusive and aquaporin-4-independent solute transport in rodent brain parenchyma. *eLife*, 6:e27679, 2017. doi: 10.7554/eLife.27679.

Jeffery L. Teeters, Keith Godfrey, Rob Young, Calvin Dang, Chris Friedsam, Barry Wark, Hiroki Asari, Simon Peron, Nuo Li, Adrien Peyrache, Genady Denisov, Joshua H. Siegle, Shawn R. Olsen, Corbett Martin, Michelle Chun, Shreejoy Tripathy, Timothy J. Blanche, Kenneth Harris, György Buzsáki, Christof Koch, Markus Meister, Karel Svoboda, and Friedrich T. Sommer. Neurodata without borders: Creating a common data format for neurophysiology. *Neuron*, 88(4):629–634, 2015. doi: 10.1016/j.neuron.2015.10.025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, Mike Zheng Shou, Xuanjing Huang, Xipeng Qiu, and Yu-Gang Jiang. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026. doi: 10.48550/arXiv.2605.12090.

Ian Williams, Joseph D. Taylor, and Alain Nogaret. Correcting model error bias in estimations of neuronal dynamics from time series observations, 2025. arXiv preprint, version 1.

Lulu Xie, Hongyi Kang, Qiwu Xu, et al. Sleep drives metabolite clearance from the adult brain. *Science*, 342(6156):373–377, 2013. doi: 10.1126/science.1241224.

Huchen Yang, Chuanqi Chen, and Jin-Long Wu. Active learning of model discrepancy with bayesian experimental design. *Computer Methods in Applied Mechanics and Engineering*, 446:118198, 2025. doi: 10.1016/j.cma.2025.118198.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Appendix contents

Experimental protocols, methods and theory, and worked MIOY case studies.

A Experimental Protocol **16**
- A.1 Question-to-evidence mapping 16
- A.2 Microenvironment relation discovery 16
- A.3 A concrete transport and observation task 16
- A.4 Prospective tests for a newly proposed relation 17
- A.5 Bidirectional modeling and observation design 17
- A.6 Inverse intervention verification 17
- A.7 Baselines and adaptation 18
- A.8 Structural and domain transfer 18
- A.9 Scale and selection 18

B Supplementary Experimental Analyses **19**
- B.1 WAM fidelity and query agreement 19
- B.2 Independent WAM baseline matrix 19
- B.3 Transfer coverage and conditional discovery 21
- B.4 Calibration and sequential validity 22
- B.5 Paired contrasts and discovery traces 23
- B.6 MIOY case: boundary exchange and readout ambiguity 24
- B.7 MIOY case: pulsatility and regional half-life 24
- B.8 MIOY case: diffusion, advection, and dispersion 25

C Notation, Assumptions, and Guarantee Scope **26**

D Extended WAM and Agent-in-Twin Specification **27**
- D.1 Typed episode representation 27
- D.2 Mechanistic conditioning and residual learning 28
- D.3 Arbitrary-block queries and approximation 28
- D.4 Open MIOY proposals and evidence updates 29
- D.5 Inverse discovery and reachability 30
- D.6 Sequential interaction and intervention programs 30
- D.7 Agent–Twin prompts and MCP exchange 30

E Theoretical Results: From Representation to Scientific Decision **35**
- E.1 Population compatibility and finite-query error 35
- E.2 Information geometry of Twin confounding 35
- E.3 An analytical transport example 36
- E.4 Certified response enclosure for the bounded exchange model 36
- E.5 Post-freeze evidence validity 37
- E.6 Robust reachability after inverse generation 37
- E.7 Composition across the discovery pipeline 38
- E.8 Evidence-sound EMP compilation 38

F Proofs **38**

G Public Data and Microenvironment Construction **40**
- G.1 Scientific sources and measurement semantics 40
- G.2 Canonical state, action, and observation variables 40
- G.3 Operational and reference worlds 40
- G.4 Visibility and leakage 40

H Hyperparameters and Model Selection **41**

I Robustness and Mechanism-Aligned Ablations **41**
- I.1 What each component changes 41
- I.2 Sensitivity to the scientific agent's LLM 42
- I.3 Structured shifts 42
- I.4 When the declared envelope is wrong 42
- I.5 Misspecified observation noise 43
- I.6 Scaling in envelope size 43

J Estimands and Statistical Analysis **44**
- J.1 Graph and discovery endpoints 44
- J.2 Prediction and intervention endpoints 45
- J.3 Independent units and paired estimates 46
- J.4 Precision, missingness, and multiplicity 47

K Certification Ledger, Support Certificates, and a Replayed Discovery Trace **48**
- K.1 Realised event calibration and the sampling schedule 48
- K.2 Which relation decisions carry a numerical certificate 50
- K.3 Population compatibility versus implemented route agreement 51
- K.4 A replayed discovery trace 52

L Compute and Reproducibility **54**

A Experimental Protocol

A.1 Question-to-evidence mapping

RQ1 compares open graph revision with fixed hypotheses at matched experimental cost, measuring relation and scope recovery, false support, and missing-mechanism recovery. RQ2 compares shared and separate models and adaptive and fixed observation design through relation recovery, observation cost, conditional consistency, and intervention fidelity. RQ3 applies nominal and discrepancy-adjusted verification to identical programs, measuring independent goal success, violations, acceptance coverage, and the certification rate. RQ4 uses component ablations, model-matched Agent controls, and held-out geometry, processes, observations, and donors to assess structural and domain generalization. Forward prediction and numerical validity diagnose the model underlying these discovery endpoints.

A.2 Microenvironment relation discovery

GlymphTwin crosses transport mechanisms with discrepancy where feasible, fixing compartment, reset state, observation access, and budget. Open tasks omit relations/scopes; fixed-graph controls update existing beliefs, and external agents share the proposal interface. Executable predictions and matching evidence determine credit, separately for support, contradiction, replacement, false support, and unresolved alternatives.

A.3 A concrete transport and observation task

The controlled-reset task uses a bounded compartment Ω with known geometry and a prescribed nonnegative initial tracer profile c_0 of positive mass. Each experiment branches from this state, applies a fixed intervention schedule, and returns measurements at selected locations and times. Experiment selection remains history-conditioned, while each execution starts from the common reset state. Its transport law is

$$\begin{aligned} \partial_t c + \nabla\cdot J = -k_c c + s_I, \qquad J = u_I c - D_I \nabla c, \\ n\cdot J = \kappa_I (c - c_{\text{ext}}) \qquad \text{on } \Gamma_{\text{ex}} \subseteq \partial\Omega. \end{aligned} \tag{17}$$

The remaining boundary is reflecting. Diffusivity D_I is positive definite, $k_c, \kappa_I \geq 0$, and the source, external concentration, and boundary schedules are specified by the task. Velocity and exchange coefficient have units of length/time, diffusivity length^2/time, and k_c inverse time. Conservation gives $\dot{M} = \int_\Omega s_I\,dx - \int_\Omega k_c c\,dx - \int_{\Gamma_{\text{ex}}} \kappa_I (c - c_{\text{ext}})\,dS$. Removal from a compartment is distinct from transfer between compartments and whole-system clearance.

The hypothesis envelope contains compositions of the admitted physical terms and their parameters. Diffusion-only sets $u_I = 0$; advective transport and effective dispersion change distinct terms. An AQP4-related virtual proxy can enter the exchange coefficient, parenchymal diffusivity, or measurement gain in competing hypotheses. Each mapping defines a separate testable explanation for the virtual proxy. The questions follow the compartment-specific transport literature Smith et al. (2017); Mestre et al. (2018) and the domain formulation.

For channel j at time t_k, a concentration observation has mean

$$\begin{aligned} \mu_{e,jk}(\xi) = a_j \int_\Omega w_{e,j}(x) c_\xi(x, t_k)\,dx + b_j, \\ Y_e = \mu_e(\xi^\star) + d_e + \epsilon_e. \end{aligned} \tag{18}$$

Here $w_{e,j}$ is a fixed spatial averaging kernel, $a_j > 0$ and b_j are sensor gain and offset. The explanation $\xi = (H, Z, N)$ includes mechanism composition, explicit dynamics or observation-discrepancy coordinates, and remaining nuisance parameters. These coordinates alter $\mu_e(\xi)$; d_e is the residual error after those corrections. Flow or pressure channels use their own operators. Conditional on pre-outcome history, ϵ_e is Gaussian with known positive-definite covariance Σ_e, retaining within-block correlations. The mean-error set $\mathcal{D}_e(\xi)$ is fixed before the outcome. Applications with other noise laws require observation-specific tests.

The prior is uniform over finitely many admitted compositions. Positive parameters use log-uniform intervals $0 < a < b < \infty$; signed forcing and offsets use bounded uniform priors, renormalized under physical constraints. Source/development data fix intervals, kernels, and covariance before evaluation. The nominal Gaussian law ($d_e = 0$) updates ranking belief; support tests use the full parameter/error envelope independently of that prior.

A.4 Prospective tests for a newly proposed relation

Let Ξ_r be the explanation envelope when relation version r is proposed. It includes the proposed mechanism and competing physical and sensor explanations. The relation fixes two interventions, their common initial conditions, and scope $\mathcal{S}_r$. On a post-injection interval with $s_I \equiv 0$ and $c_{\text{ext}} = 0$, define compartment removal $U_\xi(I, C) = 1 - M_\xi(T; I, C)/M_\xi(0; C)$. Write $g_r(\xi) = \inf_{C \in \mathcal{S}_r}[U_\xi(I_1, C) - U_\xi(I_0, C)]$. A mechanism-specific relation also fixes a predicate $m_r(\xi)$ on the physical explanation, such as whether the AQP4 proxy acts through boundary exchange rather than sensor gain. Its property is

$$\mathcal{R}_r = \{\xi \in \Xi_r : m_r(\xi) = 1,\ g_r(\xi) \geq \Delta_r\}, \tag{19}$$

where $\Delta_r > 0$ is the meaningful effect fixed at proposal time. Mechanism membership and the effect are reported separately: if all surviving explanations satisfy the effect but disagree on m_r, the effect is supported while the mechanism remains unresolved. Refuting their conjunction need not refute the mechanism itself. Counterfactual dynamics uncertainty belongs in Ξ_r, separately from observation-error allowances.

Only subsequent outcomes test this version. Starting with $\mathcal{C}_{r,0} = \Xi_r$, let ℓ count its prospective blocks. With dimension m_e and error allocation $\eta_{r,\ell}$, intersect the surviving explanations with

$$\begin{aligned} T_e(\xi) &= \inf_{d \in \mathcal{D}_e(\xi)} \|\Sigma_e^{-1/2}(Y_e - \mu_e(\xi) - d)\|_2^2, \\ \mathcal{C}_{r,\ell} &= \mathcal{C}_{r,\ell-1} \cap \{\xi : T_e(\xi) \leq \chi^2_{m_e,\,1-\eta_{r,\ell}}\}. \end{aligned} \tag{20}$$

The action, channels, times, covariance, error set, and threshold are predictable from the pre-outcome history. A nonempty set contained in $\mathcal{R}_r$ supports the relation; a nonempty set disjoint from $\mathcal{R}_r$ falsifies it. Otherwise it remains unresolved. An empty set triggers an adequacy or numerical diagnostic, accounting for the allocated stochastic tail event.

Assign globally unique birth indices $r = 1, 2, \ldots$ to new relations and scope revisions, and use $\eta_{r,\ell} = \gamma/[r(r+1)\ell(\ell+1)]$ for $0 < \gamma < 1$. The total allocation is at most γ. The budget carries across scope and envelope revisions. Earlier residuals can motivate a new version, but its confirmatory set starts from the new envelope and fresh outcomes. Proposition 7 gives simultaneous validity; Appendix J.4 specifies composite-null p-values and the requirements for simulator-based tests.

Numerical set inversion must preserve the bound directions. A certified lower bound on T_e excludes an explanation only when it exceeds the threshold. For an outer set $\overline{\mathcal{C}} \supseteq \mathcal{C}_{r,\ell}$, support requires $m_r = 1$ throughout $\overline{\mathcal{C}}$ and a certified lower bound on $\inf_{\xi \in \overline{\mathcal{C}}} g_r(\xi)$ of at least Δ_r. Falsification requires certified absence of explanations satisfying both conditions, for example an upper bound below Δ_r on the subset with $m_r = 1$. Nonemptiness requires a verified feasible witness in the surviving set.

Scope changes create linked relation versions with new prospective tests, retaining the original evidence.

A.5 Bidirectional modeling and observation design

Matched WAM comparisons separate fidelity, physical residuals, and conditional consistency (Appendix B.2). At common measurement cost, sensing compares fixed panels, intervention-only, alternating, and joint selection over compartment mass, regional tracer curves, spatial profiles, and selected times (Appendix E.3).

Case-study observation requirements. The worked case of Figure 8 proposes independently calibrated flux and concentration measurements on the same exchange-boundary patch. A nonzero concentration contrast constrains the patch-specific κ_I through Equation (17); heterogeneous exchange requires matching spatial resolution. Confirmation uses flux measured independently of the WAM boundary law. The three pictured mappings illustrate a broader physical, discrepancy, and nuisance envelope. The ordered contrast $I_1 - I_0$ and removal threshold Δ_r are fixed at relation birth. Joint mechanism–effect testing and empty-set diagnosis follow Appendix A.4.

A.6 Inverse intervention verification

Inverse tasks share initial uncertainty, events, action support, horizon, and thresholds. Gradient shooting uses differentiable subsets; derivative-free shooting, CEM, and constrained Bayesian optimization share the remainder. Nominal/adjusted verification compares identical branching programs and positive rollout counts at matched cost. Discrepancy-stratified success, acceptance, risk, and coverage use separate calibration/evaluation batches.

A.7 Baselines and adaptation

Controls separate controller, model, acquisition, and search. Deterministic Agent-in-Twin replaces LLM proposal, selection, stopping, and orchestration with ontology rules, retaining MIOY, evidence updates, tools, and budgets. Physics-only removes the learned residual and its ensemble variance; mechanistic priors, observation likelihood, discrepancy, and planning supply its remaining uncertainty and decisions.

BAD-PODS-style nested filtering marginalizes state/parameters Pérez-Vieites et al. (2025) with common priors, likelihood, action library, observation costs, and inference budget. Plug-in EIG uses a point state; nested-filter EIG integrates state/nuisance to target mechanisms; joint EIG retains the estimator and targets (H, Z). Graph revision uses a separate common proposal interface. Reports pair resolution/false support with particle counts, likelihood calls, latency, and independent-draw convergence.

Inside-Out SMC^2 retains nested inference and full-horizon policy Iqbal et al. (2024); PASOA retains tempered SMC and contrastive EIG Iollo et al. (2024); DAD amortizes history-conditioned design Foster et al. (2021). Priors, likelihood, sensing, horizon, and scientific budget are shared; offline and online costs are separate.

Separate models match capacity, training pairs, optimization, and inference calls. ACID-style action reconstruction Seo et al. (2026), GC-IDM-style inverse maps Nguyen et al. (2026), and CEM on the shared WAM de Boer et al. (2005) isolate proposals with downstream selection/verification fixed. Rejected calls consume cost; assigned-task success retains inadmissible cases (Table 6).

Bayesian-design, discrepancy-learning, and tool-agent templates share tools, evidence, and budgets Rainforth et al. (2024); Yang et al. (2025); Yao et al. (2023); Appendix I defines component controls.

Relation to model-discovery and simulator-training designs. Three recent designs sit close enough to this one that the difference is worth stating precisely, and the difference is in what each is built to separate rather than in benchmark position. MDA couples language-model model proposal with nested inference and value-of-information design, adjudicating candidate models by posterior weight Murphy (2026); its candidates are competing dynamics under a trusted observation interface. The decision here is one step earlier: a mechanism change and a twin error produce the same signature, so the acquisition target is the joint (H, Z) posterior rather than a mechanism posterior, and the emitted object is a scoped relation carrying a certified effect bound and a retained counterexample rather than a ranked model. A method that ranks dynamics correctly can still charge a readout error to a mechanism, which is the failure this paper measures as false support. SGNN-style pretraining transfers forward dynamics from a simulator ensemble Dudley and Eisenberg (2025); it supplies a world model and commits to no relation decision, so it occupies the WAM slot compared in Table 8 rather than the controller slot, and the FNO and masked-model rows there are the matched instances of that comparison. Reservoir–Hodgkin–Huxley correction learns a residual against a fixed biophysical model Williams et al. (2025); that residual is the discrepancy coordinate Z with the mechanism side held fixed, which is the plug-in arm of Table 5. The components these designs contribute are therefore already isolated by the nested-filter, discrepancy-aware and physics-only controls under a shared budget, and the comparisons that would matter most are against their components rather than their published native-benchmark numbers, which use different worlds and endpoints.

A.8 Structural and domain transfer

Structural holdouts change composition, geometry, boundaries, observations, or intervention range. Generator-family splits precede trajectories; applicability accompanies supported-query performance. AllenNeuron uses donor-separated stimulus–response tasks with domain-specific encoders, units, and observation laws.

A.9 Scale and selection

Pre-outcome allocation fixes methods, baselines, units, seeds, and budgets. Reported comparisons use 32 independent source units per configuration, three within-source technical replicates, and 16 experiments/task, and every claim in the paper rests on that design; Appendix J gives the precision basis for it. WAM training and development are separate: 512 intervention/control pairs from 64 families and 64 pairs from 16 disjoint families, respectively. Power follows Appendix J; composition, boundary, geometry, observation, and discrepancy strata retain source counts and inapplicable tasks.

B Supplementary Experimental Analyses

Figure 5 separates WAM fidelity from transfer coverage. Tables 6 and 7 distinguish within-model controls from independent backbone comparisons.

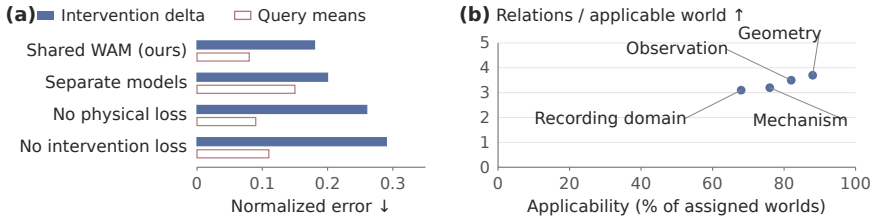


Figure 5: **Model fidelity and transfer coverage. (a) Fidelity:** normalized intervention-delta error against reference outcomes and conditional-mean disagreement across query routes. **(b) Transfer:** applicability and full-episode relation counts per applicable world across four shifts. Allen applicability is macro-averaged over predicate tasks (Appendix J.1).

B.1 WAM fidelity and query agreement

Figure 5a separates reference intervention fidelity from inter-route agreement. Table 6 holds capacity and training exposure fixed while comparing sharing, action consistency, physical grounding, and inverse search. Assigned-task success and proposal invalidity distinguish useful programs from search volume. Table 5 gives the full observation-design comparison summarised in Section 6.2.

Table 5: **Observation design under nonlinear discrepancy.** Mean ± SD over 32 source units. Relations and false support use assigned worlds; time is online acquisition relative to plug-in EIG, and offline training is separate. Methods share the scientific budget. This comparison runs under the nonlinear-discrepancy condition, which is harder than the main condition of Table 2 and Figure 6a; both arms degrade under it, the fixed schedule by 0.5 relations and the adaptive policy by 0.2, since a schedule fixed in advance cannot respond to the discrepancy it encounters. Paired contrasts against plug-in and nested-filter EIG, with intervals, are in Table 3; the margins over Inside-Out SMC^2 and PASOA are within one source-level SD and we do not claim them.

Design policy	Resolutions ↑	False support (%) ↓	Time ratio ↓
Fixed sensing Rainforth et al. (2024)	2.3 ± 0.8	13 ± 3.5	**0.2**
Plug-in state EIG Rainforth et al. (2024)	2.9 ± 0.8	11 ± 3.5	1.0
Nested-filter EIG Pérez-Vieites et al. (2025)	3.2 ± 0.8	8 ± 3.5	2.6
Joint mechanism–discrepancy EIG (ours)	$\mathbf{3.8 \pm 0.75}$	$\mathbf{5 \pm 3.5}$	2.9
Inside-Out SMC^2 Iqbal et al. (2024)	3.3 ± 0.78	7 ± 3.5	4.0
PASOA Iollo et al. (2024)	3.2 ± 0.78	8 ± 3.5	3.1
DAD Foster et al. (2021)	3.0 ± 0.8	10 ± 3.5	0.4

B.2 Independent WAM baseline matrix

Table 7 compares WAM families on GlymphTwin. Persistence and linear prediction are internal controls; GRU instantiates recurrence. Ordinary masking predicts tokens in one pass; iterative masking adds confidence-based refinement without a diffusion schedule. NeuronDiscover's learned and hybrid variants instantiate I-MDD-WM and its mechanistic-residual extension.

Models retain native objectives under common source splits, interventions, observation requests, horizon/mask patterns, tuning effort, and compute bands. All field-predicting baselines, including FNO, are evaluated on the same admissible field-query subset with compatible grids. Marginal CRPS shares field normalization; point-mass predictions give absolute error. TD-MPC2 Hansen et al. (2024) retains decoder-free latent planning, so its field-error, field-CRPS, and field-query timing metrics are undefined and it is omitted from this table. The full protocol also measures coverage/width, physical residuals, memory, and throughput (Appendices H and L).

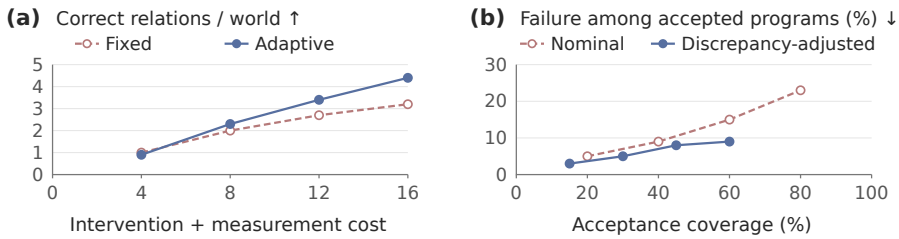


Figure 6: **Observation value and intervention reliability. (a)** Correct relations per assigned world by intervention and measurement cost, under the main evaluation condition: the adaptive endpoint at cost 16 is the Agent-in-Twin row of Table 2, and the fixed arm is its fixed-observation-schedule control. Both are one condition milder than the nonlinear-discrepancy rows of Table 5 above, which is why the levels differ between the two. **(b)** Failure risk among accepted programs versus acceptance coverage over a frozen proposal pool; undefined at zero acceptance.

Table 6: **Shared modeling and inverse proposal controls.** Forward-model intervention error, assigned-task goal success, and invalid proposals (%). Consistency and goal-conditioned controls retain the separate forward model; CEM retains the shared WAM. Physics-only removes its learned residual.

Query model	Δ error ↓	Goal success (%) ↑	Invalid (%) ↓
Shared WAM (ours)	**0.18**	**76**	**6**
Separate forward/inverse	0.20	68	11
Separate + action consistency Seo et al. (2026)	0.20	70	8
Goal-conditioned inverse Nguyen et al. (2026)	0.20	69	9
No mechanistic backbone	0.27	63	14
Physics-only	0.24	66	7
CEM on shared WAM de Boer et al. (2005)	**0.18**	72	7

Table 7: **WAM fidelity across baseline families.** Mean ± SD over 32 source units. All rows, including FNO, use the same admissible field-query subset. NRMSE averages fixed horizons. Time is the field-query ratio to the hybrid WAM.

Model	Rollout NRMSE ↓	Δ error ↓	CRPS ↓	Time ratio ↓
Persistence	0.44 ± 0.11	0.51 ± 0.13	0.31 ± 0.08	$\mathbf{0.02 \pm 0.01}$
Linear predictor	0.31 ± 0.08	0.35 ± 0.09	0.22 ± 0.06	0.04 ± 0.01
GRU Cho et al. (2014)	0.24 ± 0.06	0.29 ± 0.08	0.17 ± 0.05	0.19 ± 0.04
Causal Transformer Vaswani et al. (2017)	0.21 ± 0.05	0.25 ± 0.06	0.15 ± 0.04	0.34 ± 0.07
PETS ensemble Chua et al. (2018)	0.22 ± 0.06	0.23 ± 0.06	0.15 ± 0.04	0.28 ± 0.05
DreamerV3 RSSM Hafner et al. (2025)	0.19 ± 0.05	0.21 ± 0.05	0.13 ± 0.04	0.41 ± 0.08
FNO Li et al. (2021)	$\mathbf{0.16 \pm 0.04}$	$\mathbf{0.17 \pm 0.04}$	0.12 ± 0.03	0.24 ± 0.05
Diffusion Forcing Chen et al. (2024)	0.18 ± 0.05	0.21 ± 0.05	0.12 ± 0.03	1.28 ± 0.20
Ordinary masked model Devlin et al. (2019)	0.24 ± 0.06	0.28 ± 0.07	0.17 ± 0.05	0.16 ± 0.03
Iterative masked model Chang et al. (2022)	0.21 ± 0.05	0.24 ± 0.06	0.15 ± 0.04	0.83 ± 0.14
NeuronDiscover (learned)	0.22 ± 0.06	0.27 ± 0.07	0.15 ± 0.04	0.88 ± 0.15
NeuronDiscover (hybrid)	0.17 ± 0.04	0.18 ± 0.04	$\mathbf{0.11 \pm 0.03}$	1.00 ± 0.00

Fixed-Agent discovery comparison. Table 8 fixes the Agent, MIOY updates, priors, likelihood, calibration, verification, and forward-sampling/CEM adapter across five WAMs. All model evaluations and rejected calls consume the inference budget. Resolutions are divided by the common scientific cost of 16 per assigned world; false support retains all assignments. Macro-F1 scores the

complete relation-state vocabulary. SD follows within-source averaging of technical replicates. The comparison tests whether fidelity translates into discovery.

FNO enters this table through the same generic forward-sampling and CEM adapter used by every non-generative member, so the comparison isolates the world model rather than the proposal machinery. FNO returns a point field, so it carries no response law of its own. We give it the same one every deterministic member receives: a 16-member ensemble over input-perturbation and dropout seeds, whose empirical spread is calibrated once on the development split to match reference coverage, then frozen and moment-matched to the Gaussian form the support test consumes. The ensemble size is chosen so total inference cost matches the hybrid's, and the adapter, particle count and inner Monte Carlo budget are those of the shared configuration (Appendix H).

Exactly two objects change when the WAM is replaced, and naming them is what makes each row a test of the same scientific proposition. The predictive mean $\mu_e(\xi)$ of Equation (18) and the predictive covariance entering Σ_e are WAM-supplied and therefore differ by row. Everything that defines the proposition is fixed by the relation at birth and shared across rows: the explanation envelope Ξ_r, the mechanism predicate m_r, the effect threshold Δ_r, the scope set $\mathcal{S}_r$, the mean-error set $\mathcal{D}_e$, the error allocation $\eta_{r,\ell}$ and the χ^2 thresholds of Equation (20). Each row therefore accepts, rejects or abstains on the identical relation with the identical decision rule, and differs only in the response law it brings to that rule. The certified enclosure of Appendix E.4 is likewise shared, since it is a property of Equation (17) rather than of the learned model.

This is the decisive control for the concern that a stronger, cheaper forward predictor might already suffice: FNO attains the best deterministic rollout error in Table 7 and the best downstream yield of the non-mechanistic members here, yet remains 0.036 resolutions per unit cost and 0.10 status macro-F1 behind the mechanism-grounded hybrid (Table 3). Two calibration diagnostics are concordant with that ordering: the FNO adapter's frozen 90% predictive interval covers 0.89 of development outcomes against the hybrid's 0.91, but only 0.72 against 0.86 on held-out geometry, and CRPS orders the two models the way the downstream endpoints do while NRMSE orders them the other way. We report the concordance and do not claim it isolates a cause; a controlled attribution would require varying calibration quality at fixed point accuracy, which these runs do not do.

Table 8: **Discovery with a fixed Agent and different WAMs.** Source-unit mean ± SD at matched scientific and inference budgets. The Agent and proposal adapter are shared. FNO uses the same generic planning adapter as the other non-generative members.

WAM	Resolutions/cost ↑	False support (%) ↓	Status macro-F1 ↑
Causal Transformer Vaswani et al. (2017)	0.200 ± 0.056	9.0 ± 4.0	0.70 ± 0.09
Ordinary masked model Devlin et al. (2019)	0.206 ± 0.050	8.0 ± 3.8	0.72 ± 0.09
FNO + shared planning adapter Li et al. (2021)	0.214 ± 0.051	8.0 ± 3.7	0.74 ± 0.08
NeuronDiscover (learned)	0.219 ± 0.050	7.5 ± 3.6	0.76 ± 0.08
NeuronDiscover (hybrid)	$\mathbf{0.250 \pm 0.048}$	$\mathbf{5.0 \pm 3.0}$	$\mathbf{0.84 \pm 0.06}$

Field fidelity and nonlinear response tests. Table 9 separates codec error from the region where local sensitivity predicts complete responses. The codec comparison holds fields, sources, optimization exposure, and decoder fixed. Field/gradient errors are volume/time-weighted normalized root-mean-square errors on admitted queries; mass error uses initial mass. Admission uses all assignments. The radius study reports whitened RMS secant remainder and local/full-envelope decision agreement on admitted perturbations. Radius changes difficulty, so those rows are not ranked.

B.3 Transfer coverage and conditional discovery

Figure 5b pairs conditional recovery with applicability under geometry, mechanism, observation, and Allen shifts. A world is a source-anchored reset state with an assigned discovery task. Eligibility requires admission of that task's required queries. For Allen, input resistance, firing gain, and adaptation define separate predicate tasks; eligibility does not require all three predicates to be admissible. Aggregation preserves donor and generator-family separation (Appendix I).

Process-stratum and neuronal-recording controls. Table 10 decomposes the main discovery comparison: equally weighted transport strata recover its aggregate means. Inference pairs methods within source; the aggregate contrast and its interval appear in Table 3, and this table reports the

Table 9: **Representation and nonlinear response diagnostics.** Upper block: common-assignment codebook comparison. Lower block: 32-code WAM across perturbation radii. Secant residuals characterize the tested region, not a certified Lipschitz bound.

Codes	Field error ↓	Gradient error ↓	Mass error (%) ↓	Admitted (%) ↑
16	0.038	0.091	2.1	94
32	0.020	0.049	1.2	**96**
64	**0.011**	**0.029**	**0.7**	**96**
Radius	Secant error	Admitted (%)	Agreement (%)	
0.10	0.008	96	95	
0.25	0.034	91	88	
0.50	0.101	79	72	

stratum point estimates that decompose it. The fixed-graph control restricts revision; the pooled-scope control removes cell/stimulus distinctions while retaining the proposal interface. Its 68% applicability is the macro-average $(72 + 66 + 66)/3$ over input resistance, firing gain, and adaptation. Conditional yield counts graph relations resolved by Equation (46) over a complete applicable episode. The all-assigned summary multiplies macro-applicability by conditional episode yield: $0.68 \times 2.85 = 1.938$, reported as 1.94. Appendix J.1 gives the stratified alternative $\sum_p w_p a_p \bar{y}_p$ and the gap between the two aggregations. These controls test evidence-interface transfer separately from ionic-mechanism claims.

Table 10: **Where relation recovery occurs.** Upper block: resolutions per assigned transport world and paired differences, with equally weighted strata. Lower block: macro-averaged applicability (%), full-episode relation yield conditional on applicability, and applicability-adjusted all-assigned yield for the neuronal recordings. Controls share assignments and admission rules, so applicability is identical by construction and is not a comparison.

Process stratum	Fixed graph	NeuronDiscover	Difference
Diffusion and clearance	3.5	**4.2**	0.7
Forcing and dispersion	3.2	**4.1**	0.9
Boundary exchange	3.0	**3.9**	0.9
Sensor mismatch	3.1	**3.8**	0.7
Recording control	Applicability	Conditional	All-assigned
Fixed graph	68	2.25	1.53
Pooled scope	68	2.35	1.60
NeuronDiscover	68	**2.85**	**1.94**

Where the language-model interface contributes. Table 11 splits the deterministic–LLM contrast of Table 2 by task family. Each family receives eight of the 32 source units, assignments and budgets are shared, and the two orchestrators see the same WAM, tools and initial evidence. The aggregate $+0.20$ relations per assigned world is not spread evenly: it is concentrated in tasks whose reference graph contains a relation absent from the initial vocabulary, where proposing a new mechanism or a new observable is the binding step. On scope refinement, where the vocabulary is closed and the decision is a predicate boundary, the deterministic controller is at least as good and costs less. We therefore present the language model as an optional proposal interface with a located benefit, not as a uniform improvement.

B.4 Calibration and sequential validity

Figure 7 links observation richness, calibration effort, and adaptive testing to decision quality. Panel (d) reports recovery rates for each designated Allen predicate, conditional on applicability and over all assigned tasks. These predicate-level rates assess response accuracy and scope coverage separately from the full-episode graph yields in Table 10; they are not additive components of those yields (Appendix J.1). Unsupported tasks remain unresolved.

Table 11: **Task-family decomposition of the language-model contrast.** Resolutions per assigned world for the deterministic and language-model orchestrators, their paired difference over the eight source units of each family, a cluster-bootstrap 95% interval, and Holm-corrected p-values from exact enumeration of all $2^8 = 256$ within-source sign flips in each family. Eight sources cannot produce a two-sided value below $2/256$, so no Holm value here can fall below 0.031; Appendix J.3 gives the raw values. The mean row is the 32-source aggregate of Table 3, not a fifth family. Family means reproduce the aggregate rows of Table 2.

Task family	Determ.	LLM	Difference	95% CI	Holm p
Omitted-relation proposal	3.5	**4.1**	+0.60	[+0.25,+0.95]	0.031
Joint mechanism composition	3.9	4.2	+0.30	[-0.05,+0.65]	0.375
Scope refinement	4.0	3.9	-0.10	[-0.45,+0.25]	1.000
Observation-requirement design	3.8	3.8	+0.00	[-0.35,+0.35]	1.000
Equally weighted mean	3.8	4.0	+0.20	[+0.06,+0.34]	0.006

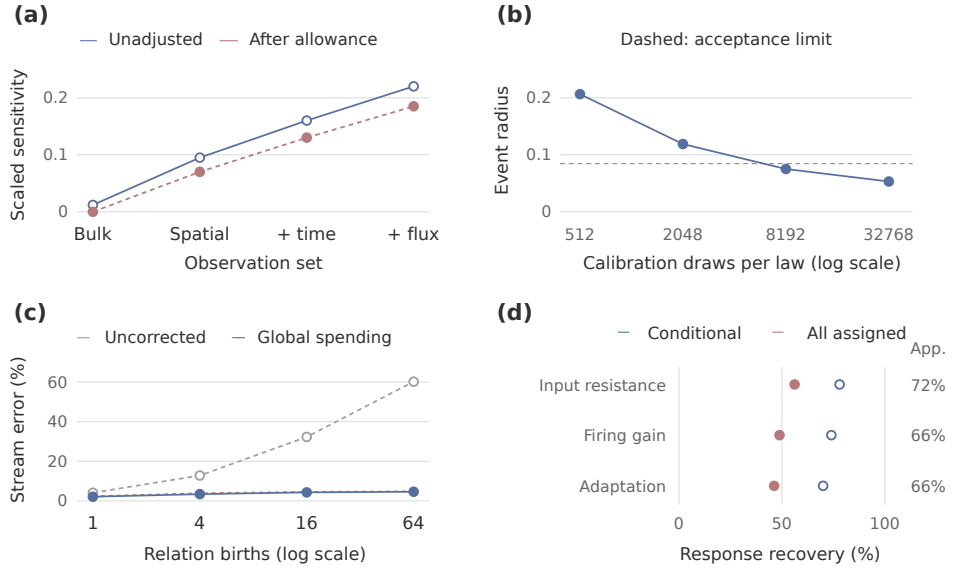


Figure 7: **Practical discovery diagnostics. (a) Sensitivity:** smallest scaled singular value before and after an assumed error allowance. **(b) Calibration:** algebraic event radius (Equation (50)) with $K = 8$, $\alpha = \beta = 0.025$, equal per-law counts, and frequency gap 0.03125. The dashed acceptance limit uses $\widehat{p}_g = 0.90625$, $\widehat{p}_v = 0.015625$, $n = 8192$, and $(\tau_g, \tau_v) = (0.8, 0.12)$. **(c) Sequential validity:** any false relation decision per stream versus R relation births; the dashed curve is the finite-birth bound $0.05R/(R+1)$. **(d) Allen response transfer:** recovery rates for the designated predicate, conditional on applicability and over all assigned tasks; right-hand percentages give predicate-specific applicability. These rates differ from full-episode graph relation counts.

B.5 Paired contrasts and discovery traces

Paired contrasts are defined before scientific-unit aggregation and are reported in Table 3 with cluster-bootstrap intervals and Holm-corrected randomization p-values. Revisable hypotheses yield 0.8 more resolved relations per assigned world than a fixed graph. Adaptive sensing reduces cost by 2 shared cost units relative to a fixed observation schedule; this contrast isolates a different component from graph-policy cost in Table 2. Discrepancy-adjusted verification reduces accepted-program risk by 6 percentage points at common acceptance coverage. Neuronal transfer gains 0.41 relations per assigned world, whereas conditional yields differ by 0.60 (Table 10); the gap is the applicability factor, not a different contrast.

Paired dispersion is computed from source-level method-minus-control differences. It is not recoverable from the marginal standard deviations of Table 2, and the two must not be substituted for one another: the paired standard deviations underlying Table 3 correspond to within-source correlations

between 0.45 and 0.85, so a contrast computed from marginal dispersion alone would overstate its interval by roughly a factor of two on the most correlated comparisons.

Discovery traces bind each relation to its scope, competing explanations, selected experiment, released outcome, and status revision. Joint mechanism and effect predicates determine support. This record separates a change in evidence from a change in scope or observation access. Appendix K.4 replays one complete trace end to end.

B.6 MIOY CASE: BOUNDARY EXCHANGE AND READOUT AMBIGUITY

Section 6.4 follows the Agent's decisions over time; Figure 8 displays the candidate relations that make those decisions interpretable. An AQP4-proxy intervention can alter a measured tracer endpoint through boundary exchange κ_I, parenchymal diffusivity D_I, or sensor gain a_j. The graph separates mechanism nodes from nuisance readout nodes and retains the intervention, observation requirements, and endpoint attached to each relation. A change in the observed endpoint $\Delta o(T)$ therefore does not by itself establish a change in physical removal $\Delta U(T)$.

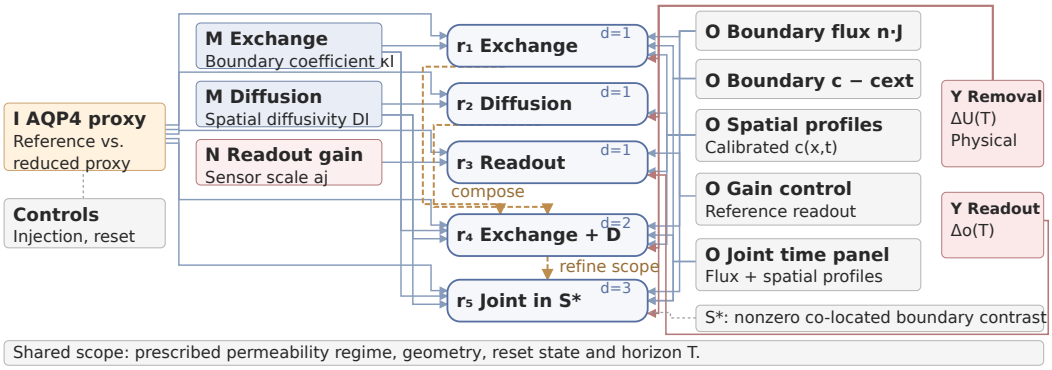


Figure 8: **MIOY graph for the boundary-exchange case.** Elementary exchange, diffusion, and readout accounts are expanded through composition and scope refinement. Mechanism nodes M remain distinct from nuisance readout node N; observation nodes O feed the observed endpoint $\Delta o(T)$ rather than directly asserting the physical endpoint $\Delta U(T)$. Ochre dashed arrows denote candidate composition or scope refinement, not support. The diagram's local relation indices identify candidate structures; the recorded episode uses persistent relation IDs 17 and 23 (Appendix K.4).

Independent flux and co-located concentration measurements constrain exchange, spatial profiles distinguish exchange from diffusion, and gain controls constrain observation effects. The diagram's $\mathcal{S}^*$ denotes the nonzero boundary-contrast requirement for that measurement to be informative. Episode-specific scope predicates are recorded separately: relation 17 requires compliance proxy $\rho(C) \geq 0.6$, while relation 23 additionally requires wall amplitude $\geq 1.5\,\mu$m. Appendix K.4 records how fresh evidence updates those relations without overwriting the original support or its later counterexample; Appendix K.2 gives the numerical effect certificate.

The next two graph instances apply the same typed representation to different scientific questions. Their composition and scope links propose new relations whose support must be established separately.

B.7 MIOY CASE: PULSATILITY AND REGIONAL HALF-LIFE

Figure 9 expands each case to five candidate relations: three elementary accounts, a joint mechanism, and a scoped version. Shared observations constrain competing accounts; controls, covariance, and scope remain explicit.

Reducing pulsatility fixes mean pressure, injection, and reset. Pumping, dispersion, and geometry-dependent resistance compete. A composed account jointly changes velocity and effective dispersion; phase-resolved flow and spatial spread distinguish their contributions. Its scoped version restricts compliant PVS geometry. Wall motion, velocity, geometry, and tracer observations retain calibration error and covariance.

The endpoint is the first half-mass crossing in a fixed region Ω_R, after the common injection period:

$$M_R(t) = \int_{\Omega_R} c(x,t)\,dx, \qquad t_{1/2}(I) = \inf\{t \in [0,T] : M_R(t;I) \leq M_R(0;I)/2\}. \tag{21}$$

Initial regional mass must be positive. No crossing by T is right-censored; sparse observations bracket the time. The first-crossing endpoint accommodates non-exponential regional mass dynamics. The relation pairs pumping with $t_{1/2}(I_{\text{reduced}}) - t_{1/2}(I_{\text{reference}}) \geq \Delta_t$, for a birth-fixed positive threshold. Censoring and time uncertainty must support a bound on this contrast.

The test in Appendix A.4 requires pumping and the contrast throughout the surviving envelope. Mixed accounts remain unresolved; empty sets trigger adequacy diagnosis. Scope revisions require fresh tests.

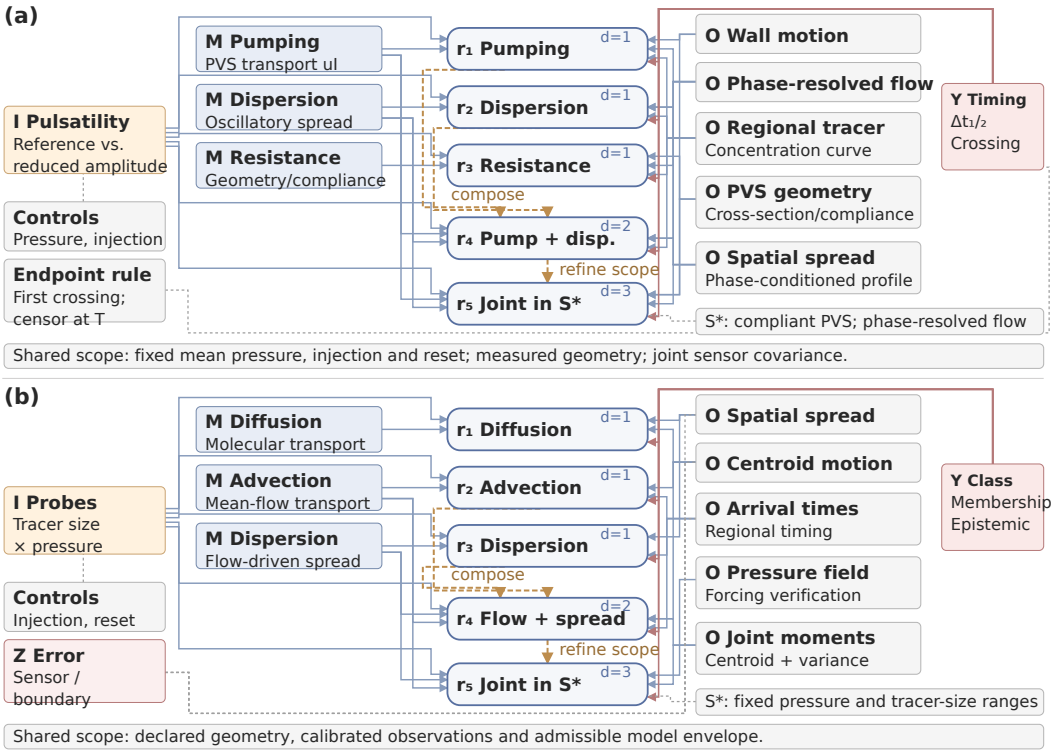


Figure 9: **Composed and scope-refined MIOY networks. (a) Pulsatility:** pumping and dispersion form a joint account, then restrict to a compliant PVS regime. **(b) Transport:** advection and dispersion share centroid/variance predictions, then restrict pressure and tracer ranges. Each case retains three elementary alternatives. Blue/red arrows bind roles/endpoints; ochre dashed links show expansion; grey links attach controls, discrepancy, or scope. Depth does not transfer evidence between relations.

B.8 MIOY CASE: DIFFUSION, ADVECTION, AND DISPERSION

Crossed tracer-size and pressure probes fix injection, reset, geometry, and calibration. Diffusion competes with advection, dispersion, and sensor/boundary discrepancy. Joint advection–dispersion predicts centroid and variance together; pressure/tracer restrictions define its scoped version. Spread and arrival times test the epistemic mechanism-class endpoint.

In the homogeneous, unbounded, constant-coefficient model (Appendix E.3), centroid, variance, and log-mass slopes identify v, effective D, and k. These analytical signatures define the prospective tests. Diffusion and dispersion can share a variance slope, and pressure-sensitive spreading admits both advection and dispersion explanations. Finite boundaries and heterogeneous coefficients require response maps that incorporate these features.

At birth, define $\mathcal{H}_D$ through tolerances on advection and dispersion relative to molecular diffusion within the constitutive model. For a nonempty surviving set $\mathcal{C}$, membership in the specified mechanism class is supported if $\mathcal{C} \subseteq \mathcal{H}_D$, falsified if $\mathcal{C} \cap \mathcal{H}_D = \varnothing$, and unresolved otherwise. Empty-set diagnosis, calibrated error allowances, and fresh scope tests follow Appendix A.4.

Table 12: Core objects in neuronal microenvironment discovery.

Symbol	Meaning and role
$Q_\theta, P^\star$	Operational joint WAM and independent reference law
S, H, I, O, A, Y	State, mechanism, intervention, observation action, combined action, outcome
G, V, C	Goal, validity, and context
Z, N	WAM discrepancy and nuisance variables
$\pi_\omega, b_t, K_t, B_t$	Agent policy, joint belief, evidence, and remaining budget
$\mathcal{G}_t, \mathcal{E}_t$	MIOY graph and scoped relations
κ, τ	Query template and diffusion index
$\mathcal{F}_t^{\mathrm{sel}}$	Pre-outcome information available to selection
$J_e^{\mathrm{mech}}, J_e^{\mathrm{wam}}$	Local mechanism and discrepancy sensitivities
$\Lambda(\mathcal{E})$	Accumulated joint information matrix
$\mathcal{A}_K, n_a$	Frozen programs and per-program rollout count
L_g, U_v, δ_a	Goal lower bound, violation upper bound, discrepancy radius
η_q	Route error relative to its population conditional
$\mathrm{Cl}_{\mathcal{R}}(E)$	Evidence closure under deterministic compiler rules

C Notation, Assumptions, and Guarantee Scope

Assumption 1 (Registered scientific world)**.** Typed state, mechanism, action, observation, goal, validity, unit, and provenance schemas are fixed before evaluation. Out-of-support actions are flagged as invalid.

Assumption 2 (Mechanistic grounding)**.** The separately versioned backbone supplies transitions and constraints; learned corrections operate inside a registered validity domain. Physical residuals assess consistency with those equations.

Assumption 3 (Outcome isolation and registered analysis)**.** Candidate generation, ranking, eligibility, and analysis choices are measurable with respect to $\mathcal{F}_t^{\mathrm{sel}}$. The sealed outcome is unavailable before the selected action and analysis record are frozen. Tests used by Proposition 6 are conditionally super-uniform under the frozen null.

Assumption 4 (Local mechanism–WAM model)**.** For registered finite-dimensional coordinates h and z, the sealed mean response is differentiable near the current pair, residual covariance is positive definite, and the first-order approximation in Equation (13) is used only inside its declared neighborhood.

Assumption 5 (Common-support population WAM)**.** Forward, inverse, diagnostic, and observation-design laws condition one positive joint Q_θ on common support. Masks, temperatures, guidance, and validity filters are recorded.

Assumption 6 (Auditable finite query implementation)**.** For a route q, the approximation theorem assumes $\sup_x d_{\mathrm{TV}}(\widetilde{K}_q(x,\cdot), K_q(x,\cdot)) \le \eta_q$ on the declared input support. An analytical argument or a statistical procedure with explicit coverage must establish this uniform bound. Held-out route scores and cycle discrepancies provide complementary implementation diagnostics.

Assumption 7 (Frozen operational rollouts)**.** The finite nonempty set $\mathcal{A}_K$ contains complete branching programs. Conditional on pre-rollout information, each program's trajectories are independent draws from its operational law. Positive counts n_a, both events, and multiplicity correction are fixed before inspection.

Assumption 8 (Decision-relevant discrepancy control)**.** For every frozen program, goal and violation probabilities differ between operational/reference laws by at most δ_a. Learned bounds require simultaneous coverage $1-\beta$ over the queried programs, selection rule, and scope.

Assumption 9 (Evidence-sound compilation)**.** Compiler rules are deterministic and versioned, emit only relations in the evidence closure, preserve counterevidence and validity guards, and append every parent evidence and rule identifier.

The guarantees apply within the specified physical and observation models. Biological interpretation additionally requires corresponding measured evidence.

Algorithm 1 Agent-in-Twin microenvironment discovery

Require: WAM Q_θ, belief b_0, graph $\mathcal{G}_0$, evidence K_0, budgets B, B^{ver}
$\mathcal{P}^{\mathrm{cert}} \leftarrow \varnothing, \quad B_0 \leftarrow B, \quad B_0^{\mathrm{ver}} \leftarrow B^{\mathrm{ver}}, \quad t \leftarrow 0$
$\Lambda_0 \leftarrow \mathrm{Information}(K_0, b_0)$
while $B_t > 0$ and unresolved scientific questions remain **do**
  $\mathcal{H}_t \leftarrow \mathrm{ProposeEdits}(\mathcal{G}_t, b_t, K_t)$
  $\mathcal{C}_t \leftarrow \mathrm{CompileQueries}(Q_\theta, \mathcal{H}_t, b_t, K_t)$
  $\mathcal{C}_t^+ \leftarrow \{c \in \mathcal{C}_t : \mathcal{V}(c) = 1,\ 0 < C(c) \le B_t\}$
  if $\mathcal{C}_t^+ = \varnothing$ **then**
    break
  $j_t(c) \leftarrow \lambda_{\min}(\Lambda_t + J_c^\top \Sigma_c^{-1} J_c)$
  if $\lambda_{\min}(\Lambda_t) < \tau_I$ **then**
    $c_t \leftarrow \arg\max_{c \in \mathcal{C}_t^+} (j_t(c), \alpha_t(c))$ lexicographically
  else
    $c_t \leftarrow \arg\max_{c \in \mathcal{C}_t^+} \alpha_t(c)$
  $s_t \leftarrow \mathrm{Freeze}(c_t, \text{endpoint, falsifier, test})$
  $Y_t \leftarrow \mathrm{Execute}^\star(s_t)$ ▷ release outcome, never hidden truth
  $K_{t+1} \leftarrow \mathcal{R}(K_t, s_t, Y_t)$
  $b_{t+1} \leftarrow \mathcal{U}(b_t; s_t, Y_t)$
  $\mathcal{G}_{t+1} \leftarrow \mathrm{Revise}(\mathcal{G}_t, \mathcal{H}_t, K_{t+1})$
  $\Lambda_{t+1} \leftarrow \mathrm{Information}(K_{t+1}, b_{t+1})$ ▷ recompute in common local coordinates
  $B_{t+1}^{\mathrm{ver}} \leftarrow B_t^{\mathrm{ver}}$
  if an inverse query has a nonempty set of admissible programs **then**
    Set complete programs $\mathcal{A}_K$, counts n_a, cost c_{ver}, and error allocation α_t^{cert}
    if $c_{\mathrm{ver}} \le B_t^{\mathrm{ver}}$ and a discrepancy bound is available **then**
      Freeze the certification request; draw fresh rollouts; compute bounds by Equation (16)
      $\mathcal{P}^{\mathrm{cert}} \leftarrow \mathcal{P}^{\mathrm{cert}} \cup \{a : L_g(a) \ge \tau_g,\ U_v(a) \le \tau_v,\ \mathcal{V}(a) = 1\}$
      $B_{t+1}^{\mathrm{ver}} \leftarrow B_t^{\mathrm{ver}} - c_{\mathrm{ver}}$
  $B_{t+1} \leftarrow B_t - C(c_t), \quad t \leftarrow t + 1$
$\mathsf{EMP} \leftarrow \mathrm{Compile}(\mathrm{EvidenceClosed}(\mathcal{G}_t, K_t))$
return $\mathcal{G}_t, b_t, K_t, \mathcal{P}^{\mathrm{cert}}$, EMP and unresolved questions

D Extended WAM and Agent-in-Twin Specification

D.1 Typed episode representation

An episode has blocks

$$X = (X^S, X^H, X^I, X^O, X^Y, X^G, X^C, X^V, X^T, X^P),$$

where X^I, X^O, X^T, X^P encode interventions, observation actions, coordinates, and provenance/visibility. Fields retain type, unit, support, missingness, and source. The spatial instantiation keeps all cell and reservoir concentrations, coefficient schedules, sensor parameters, observation-menu index, and future target values. Geometry, quadrature weights, measurement covariance, and time edges remain continuous conditioning data. Source identifiers and evaluator labels are excluded from model input.

Field codec and reconstruction. For N cells, R reservoirs, and T transitions, the future-state block contains $T(N+R)$ tokens, ordered by time and spatial index. Each variable–type–unit group uses its training mean and maximum absolute deviation to normalize values, then places 32 uniformly spaced scalar centers over its training range. Codebooks are frozen before development and evaluation. In physical units, an uncoupled scalar u_j is encoded as $k_j = \arg\min_k |u_j - c_{jk}|$. Its in-range nearest-center error is at most half the center spacing. Coupled action/parameter fields additionally enforce their joint physical support. Out-of-range values receive a support flag; nearest-endpoint encoding does not establish admissibility. Clamped observations retain their raw values after decoding.

The decoder forms a physical location $\ell_j = F_{\mathrm{mech},j} + s_j r_\theta$ and positive width w_j, where s_j is the codebook span. The categorical law and decoded mean are

$$\pi_{jk} \propto \exp\left[-\frac{(c_{jk} - \ell_j)^2}{2w_j^2}\right], \qquad \widehat{u}_j = \sum_k \pi_{jk} c_{jk}. \tag{22}$$

The width scales with center spacing and is a decoder parameter, distinct from sensor covariance. Physical locations and categorical means are retained separately: finite support can shift their values even at zero learned correction. Fidelity checks therefore separate raw-field quantization, physical-location error, and decoded-mean error, including gradients, mass balance, intervention contrasts, and threshold decisions. The 16/32/64-code comparison in Table 9 uses identical fields and training sources.

The spatial implementation uses a 32-code scalar codec and an eight-layer, pre-normalized bidirectional Transformer of width 512 with eight heads, GELU feedforward width 2048, and zero dropout. Each input sums value, type, position, space–time coordinate, and diffusion-time embeddings. A first denoiser pass proposes unknown physical coefficients; the transport solver produces coarse-state tokens that condition a second pass through the same network. Visible coefficients are supplied by the query. Residual transport and observation-mean heads decode the response; a separate head predicts validity.

For token j, the absorbing corruption process is

$$q_\tau(x_\tau^j \mid x_0^j) = \bar{\alpha}_\tau \delta_{x_0^j}(x_\tau^j) + (1-\bar{\alpha}_\tau)\delta_{\langle \text{mask} \rangle}(x_\tau^j). \tag{23}$$

Training draws $\tau \sim U(0,1)$ and masks each free site with probability τ; observed sites stay fixed and derived coarse states are excluded from targets. Masked cross-entropy uses weight $1/\tau$ and the fixed eligible-site denominator, retaining zero-mask draws. All four query routes contribute. The WAM uses 48 epochs, Adam at 3×10^{-4}, clipping at 1, and unit weights for the six losses (Appendix H).

D.2 Mechanistic conditioning and residual learning

The backbone supplies physical transitions and constraints; its residual represents unresolved dynamics:

$$\widetilde{S}_{t+1} = F_{\text{mech}}(S_t, I_t, H, C), \qquad S_{t+1} = \widetilde{S}_{t+1} + \Delta_\theta(S_t, I_t, H, C, \eta_t). \tag{24}$$

For a block mask M_τ, the denoising loss predicts masked coordinates. With decoded target U_0, decoder D_ψ, and fixed unit-aware weights W, representative auxiliary losses are

$$\mathcal{L}_{\text{dec}} = \mathbb{E}\|D_\psi(\widehat{X}_0^\theta) - U_0\|_W^2, \tag{25}$$

$$\mathcal{L}_\Delta = \mathbb{E}\, d_Y\Big(\widehat{F}_\theta(S_0, A) - \widehat{F}_\theta(S_0, A'), Y(A) - Y(A')\Big), \tag{26}$$

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}\, d_G\Big(G, \widehat{F}_\theta(S_0, \widehat{A}_\theta(S_0, G))\Big), \tag{27}$$

$$\mathcal{L}_{\text{phys}} = \mathbb{E}\|\mathcal{R}_{\text{phys}}(D_\psi(\widehat{X}))\|_{W_R}^2. \tag{28}$$

Soft categorical expectations propagate gradients through physical coefficients, the transport solve, and residual heads. The nearest-center coarse-state conditioning index is discrete, so no gradient passes through that index. Both denoiser passes share parameters; the analytic transport operator has no fitted neural weights. State/observation losses divide errors by training codebook spans. Physics loss combines time-step-scaled equation defects with mass-balance defects normalized by initial mass. Intervention loss uses paired worlds with identical state, mechanism, geometry, and observation law. Cycle loss replays generated actions through a forward query. Jointly optimizing these losses encourages physical agreement; sampled concentration, support, and conservation checks assess the realized output. Validity uses declared support labels. Conformal calibration requires exchangeability Angelopoulos and Bates (2023).

D.3 Arbitrary-block queries and approximation

Forward queries clamp initial state, mechanism, and intervention; inverse queries clamp initial state, goal, and constraints. Mechanism queries use released observations, while observation-design queries compare measurement operators. Goal-conditioned mechanisms remain proposals, separate from evidence-conditioned belief.

At reverse step $s = 0, \ldots, L-1$, each remaining masked site is revealed with probability $1/(L-s)$; the final step reveals all remaining sites. Revealed sites remain fixed. Sampling uses temperature one and no classifier-free guidance. Unsupported codes have zero probability; coupled coefficient/action supports are sampled jointly. Query diagnostics use $L = 16$ with 16 samples per checkpoint, arm, and query; Table 14 records the separate 32-step sampling setting. Every step restores observed codes and decoded raw observations. Finite-value, nonnegativity, concentration-limit, and physical-support

checks flag violations while retaining every raw draw, including negative concentrations and invalid proposals.

Lemma 1 (Clamped-block invariance)**.** *If every reverse kernel assigns probability one to each clamped value, those values remain unchanged at every reverse step and in the final proposal.*

Clamp preservation, route-wise proper scores, decoded mean disagreement, intervention contrasts, and forward replay of inverse proposals diagnose the kernel. Comparisons use identical conditioning sets and independent draws, retaining invalid samples. The uniform route error η_q in Lemma 2 requires the coverage argument specified in assumption 6.

D.4 Open MIOY proposals and evidence updates

A source-grounded ontology supplies new mechanism relations, scopes, and observables. Each proposal maps its predicted consequence to executable state, intervention, and observation interfaces:

$$q = (\Delta\mathcal{G}, \mathcal{H}_{\mathrm{alt}}, \Omega, I, O, f_{\mathrm{sup}}, f_{\mathrm{fal}}),$$

Here Ω is the scope, $\mathcal{H}_{\mathrm{alt}}$ the alternatives, and $f_{\mathrm{sup}}, f_{\mathrm{fal}}$ the support and falsification tests. Deterministic admission checks ontology references, units, action support, and measurable predictions. Admission creates a candidate version whose support is evaluated prospectively. Scope changes create linked versions and retain the original relation and counterevidence. Proposals lacking measurable consequences remain unresolved.

The posterior ranks experiments; prospective tests use subsequent evidence for each version (Appendix A.4). A frozen semantic rule matches relations and scopes.

Depth-controlled hypothesis expansion. A configurable extension caps candidate-expansion depth at $d_{\max}$ and children per expansion at $b_{\max}$, separately from planning lookahead and Twin-query budget. Elementary accounts start at depth one; composition or scope refinement advances one level. The illustrated path has three levels, retaining five relations after deduplication. Joint mechanisms require jointly parameterized predictions; refined scopes require fresh tests. Depth controls permitted expansion, not guaranteed graph size or evidential strength. Only schema-valid, executable candidates enter the graph; unchanged stopping and evidence rules apply.

Belief and observation operators. The reference controller represents $b_t = \sum_{i=1}^{P} w_t^i \delta_{(h_i, z_i, n_i, s_t^i)}$. After e_t, particles propagate through hypothesis-conditioned transitions; the released-outcome likelihood reweights them. Effective sample size below $P/2$ triggers systematic resampling and posterior-invariant Metropolis moves over continuous coordinates and admitted compositions. Support checks, acceptance, and ancestry are recorded. Development convergence determines particle count and rejuvenation effort.

Scope predicates conjoin typed membership and interval constraints on compartment, geometry, cell class, stimulus range, and horizon. Missing required metadata makes scope undecidable. Versions store the predicate, mechanism test, effect, alternatives, and counterevidence. Observation actions specify channel, spatial kernel, sampling times, sensor calibration, covariance, and cost, thereby selecting a likelihood. Particle mass guides design; outer-set tests determine support.

Information gain with latent state and nuisance. For each design, the target likelihood and information gain are

$$\ell_e(y \mid h, z) = \mathbb{E}_{(N,S)\sim b_t(\cdot\mid h,z)} p(y \mid h, z, N, S, e), \tag{29}$$

$$I_e = \mathbb{E}_{H,Z,Y}\left[\log \ell_e(Y \mid H, Z) - \log \mathbb{E}_{H',Z'} \ell_e(Y \mid H', Z')\right]. \tag{30}$$

Nested Monte Carlo integrates the conditional state/nuisance posterior in both terms; fixing the generating nuisance draw changes the information target. Common random numbers stabilize comparisons, followed by fresh evaluation of the selected design. Separate doubling tests for outer particles, inner draws, and rejuvenation assess posterior approximation, sampling variability, and the nested-log bias retained with independent inner draws. Near-ties receive more numerical effort.

Sensitivity estimation and nonlinear validity. Within a fixed mechanism composition, let q be dimensionless coordinates obtained from declared physical scales, using log coordinates for positive coefficients. At an interior point,

$$\widehat{J}_{e,:,j} = \frac{\widehat{\mu}_e(q + h_j e_j) - \widehat{\mu}_e(q - h_j e_j)}{2h_j}.$$

Probes share reset state, observation operator, and random numbers. Boundary probes use feasible one-sided derivatives on a stated tangent cone. Derivatives use decoded means or differentiable soft tokens. Compare $h_j, h_j/2, h_j/4$, refined meshes, independent batches, and physical-solver derivatives.

Stack the whitened sensitivities as $A = [\Sigma_e^{-1/2} J_e]_e$, including nuisance directions jointly or projecting them out before assessing (h, z). For a qualified bound $\|\widehat{A} - A\|_2 \le \varepsilon_J$,

$$\inf_{\|v\|_2=1} \|Av\|_2 \ge \max\{0, \inf_{\|v\|_2=1} \|\widehat{A}v\|_2 - \varepsilon_J\}. \tag{31}$$

The infimum includes a wide matrix's null space. The error budget covers discretization, derivatives, sampling, whitening, and WAM-to-reference sensitivity mismatch. Ridge stabilizes solves; rank uses the unregularized matrix. Bootstrap and mesh-doubling diagnostics estimate stability but do not certify an operator-error bound.

Let $f(q)$ stack whitened response means for a fixed support branch and let s_- denote the lower bound in Equation (31). If $\|Df(q+v) - Df(q)\|_2 \le L\|v\|_2$ throughout a convex radius-r neighborhood, Taylor's integral remainder gives

$$\|f(q+v) - f(q)\|_2 \ge \left(s_- - \tfrac{1}{2}L\|v\|_2\right)\|v\|_2. \tag{32}$$

This is separation from the expansion point. Perturb weak singular and independently sampled directions across meshes and observation menus; report secant remainder, admission, and relation decisions by radius. Certification requires a uniform L and a gap exceeding observation error. Empirical secants describe the tested region; discrete compositions and validity-branch changes require complete response-envelope tests.

D.5 Inverse discovery and reachability

Beyond reward-based planning, a scientific inverse query returns

$$(\rho, \mathcal{P}, \mathcal{O}_{\text{req}}) = \mathcal{I}_\theta(S_0, G, \Gamma, b_t, K_t), \tag{33}$$

where $\mathcal{P}$ contains programs, $\mathcal{O}_{\text{req}}$ required observations, and ρ the reachability status defined in Theorem 1. An unsuccessful finite search leaves reachability unresolved; infeasibility requires sound domain-wide exclusion.

D.6 Sequential interaction and intervention programs

In Algorithm 1, the LLM proposes hypotheses and tool calls; numerical tools compute responses, sensitivities, scores, and evidence updates.

The information floor uses commonly scaled accumulated sensitivities. Certification uses fresh complete-program rollouts and separate error/resource budgets. Replay preserves scopes, proposals, and released outcomes.

D.7 Agent–Twin prompts and MCP exchange

The native Model Context Protocol (MCP) service exposes typed Twin operations through JSON-RPC Model Context Protocol Contributors (2025). After initialization and tool discovery, the client invokes a registered method with schema-checked arguments. This section documents the interface the Agent actually sees: the registered tool surface, the session bootstrap, the argument schema, and one worked round covering the forward, inverse, sensing, freeze, release, and certification calls. Prompts and request bodies are condensed for space.

Three invariants hold throughout. The Agent never receives the hidden mechanism, the realized information gain, or reference-only validity. Protocol success is distinct from physical admissibility, which is distinct again from relation support. Every selection is frozen before an outcome is released, so no tool can be re-invoked to revise a pending test.

Registered tool surface. Table 13 lists the methods the Twin server advertises under tools/list. Numerical acquisition, evidence reduction, and certification remain server-side: the Agent may request them but cannot alter their internals, and it cannot register a new tool mid-episode. Requirements for a tool or observable that is not yet admitted pass through admission before execution.

Argument schema. Each method advertises a JSON Schema, and the server rejects a call that violates it before any physical admissibility check. Units are part of the schema, so a magnitude without its unit is a protocol error rather than a silent default.

Table 13: **Registered MCP methods.** The Agent proposes; the numerical controller owns acquisition, reduction, and certification. Status values are those of Appendix E: Reachable, Undetermined, Observe, OutOfDomain, and Infeasible.

Method	Purpose	Principal return fields
`twin.rollout`	Forward prediction under a frozen intervention and observation plan	outcomes, observations, validity, uncertainty, seed
`twin.propose_program`	Goal-conditioned inverse proposal within the admissible set	programs, required_observations, reachability
`twin.score_observation`	Expected information for candidate observation operators at fixed state and intervention	scores, cost, floor_met, weakest_direction
`twin.sensitivity`	Scaled joint Jacobian and the unregularized information floor	jacobian, min_singular, remainder_test
`graph.birth_relation`	Register a relation version with scope, effect threshold, and falsifier	relation_id, envelope_size, allocation
`graph.freeze_selection`	Fix action, endpoint, falsifier, and analysis before outcome release	selection_hash, frozen_at, allocation
`graph.release_outcome`	Submit the released measurement and reduce the surviving envelope	survivors, status, counterevidence
`verify.certify_program`	Event-calibrated goal and violation bounds for a frozen program set	lower_goal, upper_violation, decision
`budget.query`	Remaining scientific and inference budget	experiments_left, calls_left, cost_spent

Advertised schema (abridged): `twin.rollout`

```
{
 "type": "object",
 "required": ["intervention_program", "observation_requests", "horizon"],
 "properties": {
  "intervention_program": {
   "type": "array", "minItems": 0,
   "items": {
    "type": "object",
    "required": ["target", "operation", "magnitude", "unit"],
    "properties": {
     "target":    { "enum": ["aqp4_polarization", "boundary_exchange",
                             "parenchymal_diffusivity", "pulsatility",
                             "sensor_gain"] },
     "operation": { "enum": ["set", "scale", "offset"] },
     "magnitude": { "type": "number" },
     "unit":      { "type": "string" },
     "support":   { "type": "object", "description": "spatial region" },
     "duration_min": { "type": "number", "minimum": 0 } } } },
  "observation_requests": {
   "type": "array", "minItems": 1,
   "items": {
    "type": "object",
    "required": ["variable", "time_min"],
    "properties": {
     "variable": { "enum": ["tracer_retention_fraction", "regional_mass",
                            "spatial_profile", "boundary_flux",
                            "concentration_contrast"] },
     "time_min": { "type": "number", "minimum": 0 },
     "channel":  { "type": "string" } } } },
```

```
    "horizon": { "type": "number", "exclusiveMinimum": 0 },
    "seed":    { "type": "integer" } },
  "additionalProperties": false
}
```

Round prompts. The system prompt is fixed for the episode; the round context is regenerated from belief, graph, and budget state.

Agent prompt: role and current question

System. Propose discriminating experiments using registered Twin tools and current evidence. Separate transport hypotheses from readout discrepancy. Respect units, admissible actions, and budget. Use predictions to compare candidates; revise support from released independent outcomes. Return one JSON object satisfying the supplied response schema.

Round context. Keep reset state and Twin version fixed. Within the expansion budget, compare exchange, diffusion, and gain; consider joint exchange–diffusion and a narrower observation scope. Return parent relations, the new prediction and discriminating observation. Merge duplicates; stop at the depth or query limit. The depicted extension caps expansion at three levels.

Withheld. You do not observe the hidden mechanism, the reference outcome before release, the realized information gain, or any reference-only validity flag. A prediction is not evidence. Do not report a relation as supported; propose, and let the reduction decide.

Required response schema for the Agent

```
{
  "type": "object",
  "required": ["parents", "proposed_relation",
               "discriminating_observation", "rationale"],
  "properties": {
    "parents":  { "type": "array", "items": { "type": "string" } },
    "proposed_relation": {
      "type": "object",
      "required": ["mechanism_predicate", "effect_threshold", "scope"],
      "properties": {
        "mechanism_predicate": { "type": "string" },
        "effect_threshold":    { "type": "number", "exclusiveMinimum": 0 },
        "scope":  { "type": "object" },
        "falsifier": { "type": "string" } } },
    "discriminating_observation": { "type": "object" },
    "rationale": { "type": "string", "maxLength": 600 } },
  "additionalProperties": false
}
```

Forward query. The Agent first asks what the twin expects under the candidate intervention, to see whether the proposed readout would separate the accounts at all.

MCP request: forward query

```
{ "jsonrpc": "2.0", "id": "exchange-probe", "method": "tools/call",
  "params": { "name": "twin.rollout",
    "arguments": {
      "intervention_program": [ { "target": "aqp4_polarization", "operation": "set",
                                  "magnitude": 0.8, "unit": "dimensionless" } ],
      "observation_requests": [ { "variable": "tracer_retention_fraction",
                                  "time_min": 20 } ],
      "horizon": 20, "seed": 17 } } }
```

Twin return: structured payload


```
{ "jsonrpc": "2.0", "id": "exchange-probe",
  "result": { "isError": false,
    "content": [ { "type": "text", "text": "rollout completed; 1 observation" } ],
    "structuredContent": {
      "status": "Executed",
      "outcomes": { "compartment_removal": 0.412 },
      "observations": [ { "variable": "tracer_retention_fraction", "value": 0.588,
                          "time_min": 20, "unit": "fraction",
                          "provenance": "wam-hybrid-32c" } ],
      "validity": { "admissible": true, "reasons": [],
                    "assumptions": ["calibrated range", "reset state fixed"],
                    "query": "forward" },
      "uncertainty": { "sd": 0.031, "interval": [0.527, 0.649], "particles": 2048 },
      "seed": 17 } } }
```


Field	Meaning
`status`	Execution or abstention status, independent of the protocol result
`outcomes`	Named physical outcome values in their declared units
`observations`	Variable, value, time, unit, and provenance of each released channel
`validity`	Admissibility, reasons, standing assumptions, and the query route used
`uncertainty`	Numerical uncertainty summary, including the particle count behind it
`seed`	Replay seed, so the same call reproduces the same draw

Protocol success is distinct from physical validity, and both are distinct from relation support. Initialization, action support, and the observation law stay bound to the Twin session.

Inverse and sensing queries. Retention alone leaves exchange and gain indistinguishable, so the Agent asks for a program that would reach a discriminating endpoint, then scores the candidate observation operators that would read it out.

MCP request: inverse proposal, and its return


```
{ "jsonrpc": "2.0", "id": "inv-7", "method": "tools/call",
  "params": { "name": "twin.propose_program",
    "arguments": {
      "goal": { "endpoint": "boundary_flux_contrast", "direction": "increase",
                "margin": 0.08 },
      "constraints": { "max_experiments": 2, "forbid": ["sensor_gain"],
                       "scope": "compliant_pvs" } } } }

--> { "result": { "structuredContent": {
       "programs": [ { "program_id": "PG-3", "cost": 2,
                       "steps": [ "paired injection contrast at matched reset",
                                  "flux and concentration on patch_A at 8 min" ] } ],
       "required_observations": ["boundary_flux", "concentration_contrast"],
       "reachability": "Reachable",
       "notes": "gain excluded by constraint; feasible witness verified" } } }
```


A reachability of Observe would mean the endpoint is not currently measurable and the Agent must request an additional observable; Infeasible is returned only when exclusion over the search domain is sound, and is therefore rarer than a failed search.

MCP request: observation design, and its return


```
{
  "jsonrpc": "2.0", "id": "sense-7", "method": "tools/call",
  "params": { "name": "twin.score_observation",
    "arguments": {
      "candidates": [
        { "variable": "tracer_retention_fraction", "time_min": 20 },
        { "variable": "boundary_flux", "channel": "patch_A", "time_min": 8 },
```

```
      { "variable": "spatial_profile", "time_min": 8 } ],
    "target": ["mechanism", "discrepancy"], "budget_units": 4 } } }

--> { "result": { "structuredContent": {
      "scores": [ { "candidate": 0, "eig": 0.11, "cost": 1 },
                  { "candidate": 1, "eig": 0.63, "cost": 2 },
                  { "candidate": 2, "eig": 0.48, "cost": 2 } ],
      "floor_met": true, "weakest_direction": "kappa_I",
      "selected_hint": 1 } } }
}
```

The target field is what distinguishes joint acquisition from plug-in state EIG: scoring against ["mechanism", "discrepancy"] integrates over both coordinates, whereas a plug-in call passes ["mechanism"] at a point state. When floor_met is false the controller ignores the ranking and improves weakest_direction instead, breaking ties by the acquisition score of Equation (14).

Freeze, release, and certification. Freezing converts a proposal into a testable selection. The returned selection_hash binds action, endpoint, falsifier, analysis, and error allocation; the reduction refuses any outcome whose hash does not match.

MCP request: freeze, then release the outcome

```
{
  "jsonrpc": "2.0", "id": "freeze-17", "method": "tools/call",
  "params": { "name": "graph.freeze_selection",
    "arguments": { "relation_id": "R-17", "block": 1,
      "action": "PG-3", "endpoint": "compartment_removal_contrast",
      "falsifier": "no_flux_at_zero_contrast",
      "analysis": "chi2_envelope_intersection", "gamma": 0.05 } } }

--> { "result": { "structuredContent": {
      "selection_hash": "9f2c...a41e", "frozen_at": "block-1",
      "allocation": 8.17e-05 } } }

{ "jsonrpc": "2.0", "id": "release-17", "method": "tools/call",
  "params": { "name": "graph.release_outcome",
    "arguments": { "selection_hash": "9f2c...a41e",
      "measurements": [ /* released reference values with covariance */ ] } } }

--> { "result": { "structuredContent": {
      "survivors": 6, "eliminated": 5, "status": "unresolved",
      "certified_effect_lower": 0.061, "threshold": 0.08,
      "counterevidence": [], "reason":
        "mechanism predicate holds throughout; effect bound not certified" } } }
}
```

This is the round reported in Appendix K.4: the mechanism predicate already holds, yet the certified effect bound falls short of the threshold fixed at relation birth, so the server returns unresolved rather than support. The allocation echoed by the freeze is $\gamma/[r(r+1)\ell(\ell+1)]$ at $r = 17$, $\ell = 1$.

Abstention: a request outside the admitted domain

```
{
  "jsonrpc": "2.0", "id": "oob-2", "method": "tools/call",
  "params": { "name": "twin.rollout",
    "arguments": { "intervention_program": [{ "target": "parenchymal_diffusivity",
      "operation": "scale", "magnitude": 14.0, "unit": "dimensionless" }],
    "observation_requests": [{ "variable": "spatial_profile", "time_min": 8 }],
    "horizon": 8 } } }

--> { "result": { "isError": false, "structuredContent": {
      "status": "OutOfDomain", "outcomes": null, "observations": [],
```

```
      "validity": { "admissible": false,
        "reasons": ["magnitude outside calibrated range [0.2, 5.0]"],
        "assumptions": ["constitutive model fitted on training sources"] },
      "uncertainty": null } } }
}
```


The call succeeds at the protocol level, isError is false, and the Twin still declines to predict. Conflating these three layers is the most common failure mode we observed in early agent traces, which is why status and validity are separate fields rather than an exception.

Follow-up prompt: choosing the next observation

Check validity and uncertainty. If retention cannot separate exchange and gain, request a regional readout and independent gain control. Joint exchange–diffusion requires flux and spatial profiles with their covariance. Submit admitted candidates to numerical acquisition; freeze the selected query and test. Released outcomes, not expansion depth, determine support.

If the previous block returned unresolved. Do not re-test the same contrast for a better draw. Either propose an observation that raises the certified bound on the existing relation, or birth a scope-narrowed version and accept its fresh allocation. Re-running a frozen selection is refused by the server.

The numerical controller owns acquisition and evidence reduction. MCP transports requests and returns; the Agent proposes hypotheses and observations. New tool or observation requirements pass admission before execution.

E Theoretical Results: From Representation to Scientific Decision

E.1 Population compatibility and finite-query error

Under assumption 5, any partition $X = (X_U, X_V)$ satisfies

$$Q_\theta(X_U \mid X_V)Q_\theta(X_V) = Q_\theta(X_V \mid X_U)Q_\theta(X_U) = Q_\theta(X_U, X_V), \tag{34}$$

which is the formal content of the compatibility statement used in Section 4.

Proposition 3 (Common-joint query compatibility)**.** *If the population query laws are conditionals of the same positive joint WAM, they satisfy Bayes compatibility on their common support.*

Lemma 2 (Approximate compatibility of implemented query routes)**.** *Consider two registered query routes r_1 and r_2 that produce the same target conditional under the population joint WAM. Let their implemented kernels be uniformly within η_{r_1} and η_{r_2} in total variation of the corresponding population kernels. For any common clamped-input law ν,*

$$d_{\mathrm{TV}}(\nu\widetilde{K}_{r_1}, \nu\widetilde{K}_{r_2}) \le \eta_{r_1} + \eta_{r_2}. \tag{35}$$

E.2 Information geometry of Twin confounding

The weighted Gram matrix Λ has the first-order indistinguishable perturbations as its null space. For independent Gaussian responses with known parameter-independent covariances, it is also the local mean model's Fisher information. Under those additional conditions, regular unbiased estimators satisfy $\lambda_{\max}(\mathrm{Cov}(\widehat{\xi})) \ge 1/\lambda_{\min}(\Lambda)$ when $\Lambda \succ 0$, by Cramér–Rao.

Complementary observation row spaces can remove individual experiments' null directions.

After that floor is met, the regularized D-optimal increment

$$\Delta_D(e) = \log \frac{\det(\Lambda_t + J_e^\top \Sigma_e^{-1} J_e + \lambda I)}{\det(\Lambda_t + \lambda I)} \tag{36}$$

measures aggregate local volume reduction. The complementary decomposition

$$I(Y_e; H, Z \mid e, K_t) = I(Y_e; H \mid e, K_t) + I(Y_e; Z \mid H, e, K_t) \tag{37}$$

separates mechanism information from conditional discrepancy information.

E.3 AN ANALYTICAL TRANSPORT EXAMPLE

This subsection is an identifiability illustration. It shows why observation choice changes the rank of the response map, and it is deliberately idealized; it is *not* the enclosure used to certify relation decisions, which is Appendix E.4. Consider the idealized one-dimensional transport equation on $\mathbb{R}$,

$$\partial_t c + v\,\partial_x c = D\,\partial_{xx} c - kc, \qquad D > 0, \quad k \geq 0. \tag{38}$$

Assume nonnegative initial concentration with positive finite mass and finite second moment, constant coefficients, and sufficient concentration and flux decay for moment integration. This homogeneous-compartment example isolates how observation choice separates transport coefficients.

Let $M(t) = \int c\,dx$, $\mu(t) = M(t)^{-1}\int xc\,dx$, and $\sigma^2(t) = M(t)^{-1}\int (x-\mu(t))^2 c\,dx$. Integrating Equation (38) and integrating its first two moments by parts gives

$$\frac{d}{dt}\log M = -k, \qquad \frac{d\mu}{dt} = v, \qquad \frac{d\sigma^2}{dt} = 2D. \tag{39}$$

A constant multiplicative gain cancels from the slopes. Mass identifies k, while spatial moments supply independent contrasts for v and D, making observation choice change the response map's rank. Boundary exchange, variable coefficients, or time-dependent gain changes these identities, so none of them transfers to the bounded exchange model of Equation (17); the next subsection supplies bounds that do.

E.4 CERTIFIED RESPONSE ENCLOSURE FOR THE BOUNDED EXCHANGE MODEL

Relation decisions are taken on Equation (17): a bounded Lipschitz domain Ω with an exchange boundary Γ_{ex}, a reflecting remainder, spatially varying diffusion, advection, first-order loss, and sensor coordinates. The certified support test of Appendix A.4 needs three objects on that model, not on Equation (38): an outer set $\overline{\mathcal{C}} \supseteq \mathcal{C}_{r,\ell}$, a certified lower bound on $\inf_{\xi\in\overline{\mathcal{C}}} g_r(\xi)$, and a verified nonempty witness. This subsection supplies the response enclosure those objects are built from. Throughout, Θ denotes an admitted parameter box with $k_c \in [k^-, k^+]$ on Ω, $\kappa_I \in [\kappa^-, \kappa^+]$ on Γ_{ex}, D_I uniformly elliptic with $0 < \lambda^- \leq \lambda_{\min}(D_I)$, and u_I in a bounded box with $\nabla\cdot u_I$ bounded. The post-injection window has $s_I \equiv 0$ and $c_{\mathrm{ext}} = 0$, as Appendix A.4 requires.

Lemma 3 (Positivity and the exact log-mass identity)**.** *For $c_0 \geq 0$ with $M(0) = \int_\Omega c_0 > 0$, the weak solution of Equation* (17) *satisfies $c \geq 0$ on $\Omega \times [0,T]$ and $M(t) > 0$ for all $t \leq T$. Writing $Q(t) = \int_{\Gamma_{\mathrm{ex}}} c\,dS$ and $\varphi(t) = Q(t)/M(t) \geq 0$,*

$$\log\frac{M(T)}{M(0)} = -\int_0^T \big(\langle k_c\rangle_{c(t)} + \langle\kappa_I\rangle_{c(t)}\varphi(t)\big)\,dt, \tag{40}$$

where $\langle\cdot\rangle_{c(t)}$ are the concentration-weighted averages over Ω and over Γ_{ex}. Consequently the compartment removal of Appendix A.4 is $U_\xi(I,C) = 1 - \exp\{-\int_0^T(\langle k_c\rangle + \langle\kappa_I\rangle\varphi)\,dt\}$.

Equation (40) is exact on the bounded domain with exchange boundary and variable coefficients. It is not a moment identity and does not assume constant coefficients, whole-space geometry, or a closed form for φ. Everything the model makes hard is carried by the single scalar *boundary occupancy* $\Phi(\xi) = \int_0^T \varphi(t)\,dt$, which depends on the full explanation through the solution and admits no closed form. The certification therefore brackets Φ numerically and treats the rest analytically.

Proposition 4 (Monotone parameter enclosure)**.** *Let Θ be an admitted box and suppose $\Phi(\xi) \in [\Phi^-, \Phi^+]$ with $\Phi^- \geq 0$ for* every *$\xi \in \Theta$. Then for every $\xi \in \Theta$,*

$$1 - e^{-k^- T - \kappa^-\Phi^-} \;\leq\; U_\xi \;\leq\; 1 - e^{-k^+ T - \kappa^+\Phi^+}. \tag{41}$$

Moreover U_ξ is nondecreasing in k_c and in κ_I pointwise, so subdividing Θ can only tighten the enclosure, whose width is $O(|k^+ - k^-|T + |\kappa^+\Phi^+ - \kappa^-\Phi^-|)$.

The occupancy is not constant over Θ. Both κ_I and k_c enter the equation and the boundary condition, so they change the solution and therefore φ; a bracket obtained at one parameter value certifies nothing at another. The hypothesis of Proposition 4 is consequently a box-uniform bracket, which is what the next proposition must deliver. Monotonicity is what makes subdivision productive rather than merely valid.

Proposition 5 (Certified boundary occupancy)**.** *Let Q_h be the boundary-trace functional of a conforming finite-element solution and let η_{pr} and η_{ad} be equilibrated-flux a posteriori bounds on the primal and adjoint energy errors, each assembled with interval coefficients so that it holds simultaneously for every $\xi \in \Theta$ Ern and Vohralík (2015). Let η_{fp} bound the floating-point evaluation under directed rounding Rump (2010). Then*

$$|Q(\xi) - Q_h| \;\le\; \eta_{\mathrm{pr}}\,\eta_{\mathrm{ad}} + \eta_{\mathrm{fp}} \qquad \textit{for every } \xi \in \Theta, \tag{42}$$

and integrating the resulting two-sided bracket in time with an interval quadrature remainder gives a box-uniform $\Phi(\xi) \in [\Phi^-, \Phi^+]$.

Equation (42) is the Cauchy–Schwarz product of two guaranteed energy-norm bounds. It is deliberately weaker than the dual-weighted-residual estimate Becker and Rannacher (2001), whose error representation carries a remainder that is second order but not rigorously bounded in general; only the product form is a certificate. We compute the DWR value alongside as a sharpness diagnostic and never admit it into a bound.

Composing the two propositions gives what the certificate needs. For an ordered intervention pair and a scope set $\mathcal{S}_r$, subtracting the enclosure of $U_\xi(I_0, C)$ from that of $U_\xi(I_1, C)$ yields a two-sided enclosure of the effect, and minimising its lower endpoint over the scope box gives a certified lower bound

$$\underline{g}_r = \min_{C \in \mathcal{S}_r} \left[\left(1 - e^{-k^- T - \kappa^- \Phi^-(I_1, C)}\right) - \left(1 - e^{-k^+ T - \kappa^+ \Phi^+(I_0, C)}\right) \right] \;\le\; \inf_{\xi \in \overline{\mathcal{C}},\, C \in \mathcal{S}_r} g_r(\xi), \tag{43}$$

the scope minimum being taken by the same subdivision, since the context enters only through $\Phi^\pm$ and the box endpoints. Support requires $\underline{g}_r \ge \Delta_r$, $m_r = 1$ throughout $\overline{\mathcal{C}}$, and a witness $\xi^w \in \overline{\mathcal{C}}$ whose residual satisfies $\overline{T}_e(\xi^w) \le \chi^2_{m_e, 1-\eta_{r,\ell}}$ for a certified *upper* bound $\overline{T}_e$; the direction is reversed from exclusion, where a certified lower bound is required. Branch-and-bound refines Θ until the relative width of the g_r enclosure meets its tolerance or the node budget is exhausted, and exhausting the budget produces an abstention, never a decision. The guarantee is conditional on the declared envelope, the elliptic and boundedness conditions above, and the estimator's hypotheses; it does not extend to models outside Equation (17).

E.5 Post-freeze evidence validity

Conditional super-uniformity applies to the realized selection when the candidate set, action, endpoint, test, falsifier, multiplicity correction, and stopping rule are fixed before outcome release. Sequential validity additionally requires alpha spending or an always-valid procedure.

Proposition 6 (Post-freeze validity)**.** *If the test for the realized selection is conditionally super-uniform, $\Pr(p_{s_t} \le u \mid \mathcal{F}_t^{\mathrm{sel}}) \le u$, then $\Pr(p_{s_t} \le \alpha) \le \alpha$ despite arbitrary pre-outcome proposal search.*

Proposition 7 (Prospective validity across relation revisions)**.** *For every relation version and tested block in Appendix A.4, suppose the true explanation lies in the declared envelope, the actual mean error lies in $\mathcal{D}_e(\xi^\star)$, and the conditional Gaussian model with known covariance in Equation* (18) *holds. Then, with probability at least $1 - \gamma$, every set in Equation* (20) *retains the true explanation. Consequently, all supported or falsified relation properties are correct within their stated model and scope. If simultaneous envelope and error-set adequacy instead holds on an event of probability at least $1 - \beta_{\mathrm{env}}$, the bound is $1 - \gamma - \beta_{\mathrm{env}}$.*

E.6 Robust reachability after inverse generation

Theorem 1 (Simultaneous discrepancy-adjusted reachability)**.** *Under assumptions 7 and 8, let $K = |\mathcal{A}_K| \ge 1$, $n_a \ge 1$, $0 < \alpha < 1$, and $\epsilon_a = \sqrt{\log(2K/\alpha)/(2n_a)}$. With probability at least $1 - \alpha$ over operational rollouts when the discrepancy bounds hold deterministically, every frozen candidate satisfies*

$$P_a^\star(G) \ge \widehat{p}_g(a) - \epsilon_a - \delta_a, \tag{44}$$

$$P_a^\star(V) \le \widehat{p}_v(a) + \epsilon_a + \delta_a. \tag{45}$$

An integral probability metric containing both event indicators suffices. Simultaneous calibration adds its error by a union bound.

An admissible program is Reachable if its goal lower bound meets τ_g and its violation upper bound meets τ_v. Conditional certificates require an explicit eligibility condition and its conditional law. Inconclusive bounds return Undetermined, missing observables Observe, and unsupported queries OutOfDomain. Infeasible requires sound exclusion over the search domain. Displayed bounds may be clipped to $[0, 1]$.

E.7 Composition across the discovery pipeline

Corollary 1 (Conditional end-to-end accountability)**.** *Suppose the discovery tests control their familywise error at γ, simultaneous discrepancy calibration fails with probability at most β, and the frozen-set rollout bounds use error budget α. Under assumption 9, with probability at least $1 - \gamma - \beta - \alpha$, no error occurs in these specified testing and certification events, and every emitted relation has a finite provenance path to source evidence. The bound is informative when $\gamma+\beta+\alpha < 1$.*

Compilation is deterministic. Random envelope adequacy replaces γ by $\gamma + \beta_{\mathrm{env}}$ as in Proposition 7. Estimated covariance needs a separate valid analysis, and biological truth or unique attribution requires the corresponding fidelity and identifiability conditions.

E.8 Evidence-sound EMP compilation

Let E_0 be directly supported or contradicted MIOY relations and $\mathrm{Cl}_{\mathcal{R}}(E_0)$ their closure under registered deterministic rules.

Proposition 8 (Traceability of evidence-sound compilation)**.** *Under assumption 9, every relation referenced by an EMP action has a finite derivation to source evidence, and no relation outside $\mathrm{Cl}_{\mathcal{R}}(E_0)$ enters the program.*

F Proofs

Proof of Proposition 7. Condition on pre-block history. The true mean error is admissible in the infimum, hence $T_e(\xi^\star) \le \|\Sigma_e^{-1/2}\epsilon_e\|_2^2 \sim \chi^2_{m_e}$. Truth is excluded with conditional probability at most $\eta_{r,\ell}$, and taking expectations preserves this bound for adaptive actions and relation births. Thus

$$\Pr(\text{any exclusion of truth}) \le \sum_{r=1}^{\infty}\sum_{\ell=1}^{\infty} \frac{\gamma}{r(r+1)\ell(\ell+1)} = \gamma.$$

Unexecuted blocks contribute no error. Every surviving set then contains truth, so containment in $\mathcal{R}_r$ supports its property and disjointness refutes it. Under random adequacy, exclusion on the adequacy event is still contained in the chi-square tail event; the union bound adds β_{env}. Independence between tests is unnecessary. □

Transport moment identities. With $M_j = \int x^j c(x,t)\,dx$, integration by parts under Appendix E.3 gives

$$\dot M_0 = -kM_0, \qquad \dot M_1 = vM_0 - kM_1, \qquad \dot M_2 = 2vM_1 + 2DM_0 - kM_2.$$

Substitution into $\mu = M_1/M_0$ and $\sigma^2 = M_2/M_0 - \mu^2$ proves Equation (39). A constant measurement gain $a > 0$ cancels from both ratios and the logarithmic mass derivative. These identities use the whole-space, constant-coefficient hypotheses of Appendix E.3 and are not used by any certified decision.

Proof of Lemma 3. Write the weak form of Equation (17) with $s_I \equiv 0$ and $c_{\mathrm{ext}} = 0$ and test with the negative part $c^- = \max(-c, 0)$. The reflecting portion of $\partial\Omega$ contributes nothing, the exchange term contributes $\int_{\Gamma_{\mathrm{ex}}} \kappa_I (c^-)^2 \ge 0$, and uniform ellipticity gives $\int_\Omega \nabla c^- \cdot D_I \nabla c^- \ge \lambda^- \|\nabla c^-\|^2$. Bounding the advective term by Young's inequality against that coercivity leaves $\frac{1}{2}\frac{d}{dt}\|c^-\|^2 \le C\|c^-\|^2$ with C depending only on $\|u_I\|_\infty$, $\|\nabla\cdot u_I\|_\infty$ and λ^-. Since $c_0^- = 0$, Grönwall gives $c^- \equiv 0$, so $c \ge 0$.

Testing with the constant 1 and using the boundary condition gives the exact balance $\dot M = -\int_\Omega k_c c\,dx - \int_{\Gamma_{\mathrm{ex}}} \kappa_I c\,dS = -\langle k_c\rangle_{c(t)} M(t) - \langle \kappa_I\rangle_{c(t)} Q(t)$, the weighted averages being well defined because $c \ge 0$. Hence $\dot M \ge -(k^+ + \kappa^+\varphi(t))M$, so M cannot vanish in finite time and $M(t) > 0$ on $[0, T]$. Dividing by $M(t)$, substituting $\varphi = Q/M$ and integrating gives Equation (40); the displayed form of U_ξ is its definition $1 - M(T)/M(0)$. □

Proof of Proposition 4. Monotonicity is a comparison argument. Let c_1, c_2 solve Equation (17) with exchange coefficients $\kappa_1 \le \kappa_2$ and all other data equal. Their difference $w = c_1 - c_2$ satisfies the same equation with boundary flux $n \cdot J_w = \kappa_1 w - (\kappa_2 - \kappa_1)c_2$. Testing with w^- contributes $\int_{\Gamma_{\mathrm{ex}}} \kappa_1 (w^-)^2 \ge 0$ from the first term and $-\int_{\Gamma_{\mathrm{ex}}} (\kappa_2 - \kappa_1) c_2 w^- \le 0$ from the second, because $c_2 \ge 0$ by Lemma 3 and $w^- \ge 0$. The argument of that lemma then gives $w^- \equiv 0$, so $c_1 \ge c_2$ and $M_1(T) \ge M_2(T)$. Replacing the boundary term by $-\int_\Omega (k_2 - k_1) c_2$ gives the same conclusion for k_c. Since $U_\xi = 1 - M(T)/M(0)$, removal is nondecreasing in both coefficients.

For the enclosure, Lemma 3 gives $\langle k_c \rangle_{c(t)} \in [k^-, k^+]$ and $\langle \kappa_I \rangle_{c(t)} \in [\kappa^-, \kappa^+]$ pointwise in t, because both are averages of the coefficient against a nonnegative weight of unit total mass. With $\varphi \ge 0$ and $\Phi = \int_0^T \varphi \in [\Phi^-, \Phi^+]$, the exponent in Equation (40) lies in $[k^- T + \kappa^- \Phi^-,\ k^+ T + \kappa^+ \Phi^+]$, and $x \mapsto 1 - e^{-x}$ is increasing, which is Equation (41). Each endpoint uses one corner of the coefficient box and one endpoint of the occupancy bracket, so refining either interval cannot widen it. □

Proof of Proposition 5. Q is a bounded linear functional of the state on $H^1(\Omega)$ by the trace theorem, so the adjoint problem for Q is well posed and the error in the functional admits the standard representation $Q(\xi) - Q_h = \mathcal{B}(c - c_h,\, z - z_h)$ with $\mathcal{B}$ the bilinear form and z the adjoint solution. Cauchy–Schwarz in the energy norm gives $|Q(\xi) - Q_h| \le \|c - c_h\|_{\mathcal{B}} \|z - z_h\|_{\mathcal{B}}$, and each factor is bounded by its equilibrated-flux estimator, which is guaranteed and constant-free Ern and Vohralík (2015). Assembling both estimators with interval coefficients replaces each factor by a bound valid simultaneously for every $\xi \in \Theta$, since the estimator depends on the coefficients only through the assembled residual and flux reconstruction, both of which are evaluated by interval arithmetic on the same mesh. This yields Equation (42); the sharper dual-weighted-residual value is not used, because its remainder is not rigorously bounded.

Since $M(t) > 0$ on $[0, T]$ by Lemma 3, the quotient defining φ is bounded and the division is a well-posed interval operation, so a two-sided bracket on Q and on M gives one on φ. Time integration of a two-sided bracket is a two-sided bracket, so quadrature with its own interval remainder gives the stated box-uniform $[\Phi^-, \Phi^+]$. Evaluating everything in directed rounding adds η_{fp} and preserves the enclosure Rump (2010). □

Proof of Proposition 6. The complete selection record s_t is $\mathcal{F}_t^{\mathrm{sel}}$-measurable. For $u \in [0, 1]$, conditional super-uniformity and the tower property give

$$\Pr(p_{s_t} \le u) = \mathbb{E}\big[\Pr(p_{s_t} \le u \mid \mathcal{F}_t^{\mathrm{sel}})\big] \le u.$$

The bound applies to the test, endpoint, and multiplicity rule fixed in the selection record. □

Proof of Proposition 3. Positivity defines both conditionals on common support. The product rule for either partition gives the same joint law, proving Equation (34). Implemented-route error follows Lemma 2. □

Proof of Lemma 2. By premise, $\nu K_{r_1} = \nu K_{r_2}$. The triangle inequality gives

$$d_{\mathrm{TV}}(\nu \widetilde{K}_{r_1}, \nu \widetilde{K}_{r_2}) \le d_{\mathrm{TV}}(\nu \widetilde{K}_{r_1}, \nu K_{r_1}) + d_{\mathrm{TV}}(\nu K_{r_2}, \nu \widetilde{K}_{r_2}).$$

Integrating the uniform bounds over ν gives $\eta_{r_1} + \eta_{r_2}$. □

Proof of Proposition 2. For fixed-reducer gain g_e of range width at most M, the total-variation inequality gives

$$|U_P(e) - U_Q(e)| \le M d_{\mathrm{TV}}(P_e^\star, Q_e) \le M \delta_e.$$

Costs and validity penalties cancel. The triangle inequality bounds two candidates' margin error by $M(\delta_e + \delta_{e'})$; an operational margin exceeding this stays positive under the reference law. □

Proof of Proposition 1. For a joint perturbation $v = (v_h, v_z)$,

$$v^\top \Lambda(\mathcal{E}) v = \sum_{e \in \mathcal{E}} v^\top J_e^\top \Sigma_e^{-1} J_e v = \sum_{e \in \mathcal{E}} \|\Sigma_e^{-1/2} J_e v\|_2^2.$$

Positive-definite weights imply that this vanishes exactly when $J_e v = 0$ for every experiment. Thus $\Lambda \succ 0$ precisely when the stacked Jacobian has no nonzero null direction, proving first-order identifiability in the declared local domain. □

Proof of Lemma 1. Initialization sets $X_O^{(T)} = c_O$ and each reverse step applies $\Pi_O(x) = (c_O, x_M)$. Induction gives $X_O^{(t)} = c_O$ through the final step for any free-coordinate kernel. □

Proof of Theorem 1. Conditional on the frozen programs, the goal and violation indicators satisfy one-sided Hoeffding bounds under assumption 7:

$$Q_a(G) \geq \widehat{p}_g(a) - \epsilon_a, \qquad Q_a(V) \leq \widehat{p}_v(a) + \epsilon_a$$

Each tail has probability at most $\alpha/(2K)$, giving simultaneous coverage $1-\alpha$. Applying the goal and violation discrepancy bounds $|P_a^\star(B) - Q_a(B)| \leq \delta_a$ gives Equations (44) and (45). Simultaneous calibration failure adds β by a union bound. □

Proof of Corollary 1. Let $E_{\text{test}}, E_{\text{cal}}, E_{\text{roll}}$ denote failures of the specified discovery tests, simultaneous calibration, and rollout bounds. Their probabilities are at most γ, β, α; the conditional rollout bound also holds unconditionally by expectation. Therefore

$$\Pr(E_{\text{test}} \cup E_{\text{cal}} \cup E_{\text{roll}}) \leq \gamma + \beta + \alpha.$$

No independence is needed. Finite provenance additionally holds deterministically under assumption 9 and Proposition 8. □

Proof of Proposition 8. Source relations in E_0 have finite traces. Each registered rule preserves closure and extends its premises' finite traces by one rule identifier. Induction on derivation depth establishes the property for every relation referenced by an EMP. □

G Public Data and Microenvironment Construction

G.1 Scientific sources and measurement semantics

Transport and neurovascular studies define GlymphTwin's compartments, hypotheses, observation conditions, and parameter priors Iliff et al. (2012); Xie et al. (2013); Mestre et al. (2018); Hablitz et al. (2020); Iadecola (2017); Smith et al. (2017).

Allen Cell Types supplies morphology, electrophysiology, and stimuli Allen Institute for Brain Science (2026); Gouwens et al. (2019); NWB retains acquisition metadata Teeters et al. (2015). Visual Coding, Neuropixels, and V1 models supply complementary population-level resources de Vries et al. (2020); Siegle et al. (2021); Billeh et al. (2020). Admission requires task-matched measurements, stimulation history, and units.

G.2 Canonical state, action, and observation variables

Records distinguish measurements, derived values, parameters, actions, and outcomes, retaining coordinates, units, missingness, and uncertainty. Transport observations are state functionals; Allen records retain current, voltage, morphology, offsets, and history. Models share transformed observations.

G.3 Operational and reference worlds

Operational and reference worlds share action/measurement semantics, with declared dynamics, mesh, boundary, and sensor differences. The protocol uses 17 operational cells and a 136-cell reference, 68/136/272-cell refinement, and an independent matrix-exponential check. Conservative cell-volume projection aligns fields at 0, 1, 3, 8, and 20 minutes. Reference code, draws, and tolerances are frozen independently of the residual; trajectories retain source, generator, intervention, and measurement seeds. Refinement and conservation checks assess numerical fidelity.

Source/donor splits precede trajectory, cell, sweep, and window sampling. Learned preprocessing uses training sources; calibration, selection, and evaluation remain separate.

G.4 Visibility and leakage

Selection receives executable actions, current observations, costs, and available validity. Hidden mechanisms, future outcomes, realized information gain, and reference-only validity are excluded. Leakage checks cover fields, identifiers, and candidate order.

Replay binds frozen requests to released measurements and graph updates. Source-disjoint records enumerate families, episodes, intervention pairs, and observation plans; hidden graphs reach only the evaluator. Mismatch studies cross mesh, boundary, gain, and missing-process shifts, retaining unsupported episodes in applicability/recovery denominators.

AllenNeuron task instantiation. For current-clamp recordings, three prespecified response predicates concern subthreshold input resistance, firing-rate gain, and spike-frequency adaptation. Input resistance uses baseline-subtracted steady voltage divided by nonzero subthreshold current; gain uses a slope across available suprathreshold amplitudes; adaptation compares late and early interspike intervals under a fixed epoch rule. Observation actions select an existing sweep and a voltage window, spike-count window, or interval summary. Thresholds, quality exclusions, and epoch boundaries are fixed on development donors. Missing sweeps and insufficient spike counts yield unavailable endpoints.

MIOY scope records species, cell class, cortical region/layer, morphology availability, stimulus family and amplitude, and recording quality. Donors determine splits before cells and sweeps. Scope-aware and scope-pooled controls share the same assignments; reports pair conditional response recovery with applicability and recovery over all assigned tasks. Recorded stimulus contrasts can test response predicates. Ionic or morphological mechanism alternatives require a separately validated neuronal simulator or additional perturbations. The public-recording and simulator-based endpoints remain separate.

H Hyperparameters and Model Selection

Table 14 specifies WAM training: 512 pairs/64 families, with 64 development pairs/16 disjoint families, 48 epochs, one intervention/control pair per update, and checkpoint/evaluation every six epochs. Source-disjoint model selection accounts for successful and failed trials; Appendix A gives formal source allocation.

Table 14: WAM training settings and evaluation defaults.

Component	Search space	Setting
Representation	Continuous/scalar; VQ/FSQ; RVQ; hierarchical RVQ	Grouped scalar, one level, 32 codes; training-pair fit
WAM capacity	Study: 20–80M; diagnostic: 2–10M	26,152,996 parameters; 8 layers; width 512; 8 heads
Learning rate	$\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$	Adam, constant 3×10^{-4}; gradient norm cap 1
Mask/reveal	Linear/cosine/token-specific; remasking rule	Linear absorbing mask; $t \sim U(0,1)$; monotone reveal, no remasking
Query sampling	Guidance, temperature, step grids	32 denoising steps; temperature 1.0
Loss weights	Joint token, decoder, physics, intervention contrast, consistency, validity	Six unit weights; both paired arms
Verification	Fixed schedule, no outcome-dependent extension	128 development-screen rollouts/candidate; 8192 certification rollouts/program
Acquisition	Information floor; falsification, cost, validity weights	$\lambda_F = 0.2$, $\lambda_C = 0.1$, $\lambda_V = 1.0$
Lookahead	Depth $\{1, 2, 3\}$; matched calls	Depth 2; 32 WAM queries/round
Program size	Branches, horizon, evidence closure	At most 2 branches, 4 intervention steps

Study-model comparisons match total capacity, training exposure, and search cost, reporting remaining differences. Smaller models serve implementation diagnostics.

Sampling error at most ε requires $n_a \geq \log(2K/\alpha)/(2\varepsilon^2)$ (Equation (16)). Freeze thresholds/calibration rules before evaluation. Independent reference/WAM calibration batches estimate discrepancy bounds; fresh rollouts verify programs.

I Robustness and Mechanism-Aligned Ablations

I.1 What each component changes

Each control in Table 4 changes one component, retaining data, capacity budget, and scientific interfaces. Loss ablations separately remove physical residual, decoded dynamics, intervention contrast, or cycle consistency; target endpoints accompany forward fidelity.

I.2 Sensitivity to the scientific agent's LLM

The LLM backbones are Kimi K3, DeepSeek V4 Flash, and GPT-5.6-Sol, with served versions and reasoning settings recorded. Agents share WAM, tools, initial evidence, experiments, and splits, following their own observations (Figure 10).

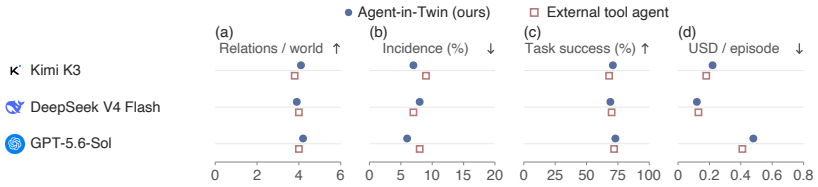


Figure 10: **LLM comparison at matched scientific budgets. (a) Resolution** and **(b) false support** use assigned worlds; **(c) goal success** uses assigned tasks; **(d) cost** is USD/episode. Circles: Agent-in-Twin; squares: external tool agent. Points show the mean for each configuration.

Within-architecture prompts, schemas, history, response limits, and retries are fixed. Evidence fits the smallest context window. Scientific/call budgets are shared; truncation, retries, tokens, and compute are logged separately.

Endpoints retain interrupted/invalid requests and distinguish support/contradiction. Three replicates average within source/donor; available provider seeds are recorded. Within-backbone contrasts receive Holm correction; cross-backbone effects are secondary.

Token, latency, and billing totals include retries and exposed reasoning; matched-dollar analysis requires complete billing (Appendix J).

I.3 Structured shifts

Table 15: Changes that can alter a microenvironment discovery decision.

Shift	Scientific change	Readout
Dynamics or composition	Missing process or changed interaction	Relation revision and false support
Geometry or boundary	Altered transport path or exchange law	Scope accuracy and structural validity
Intervention	Delayed, saturated, or unsupported forcing	Goal success and validity decisions
Observation	Coarsening, bias, delay, or missing regions	Measurement choice and unresolved contrasts
Noise and dependence	Correlated or heteroscedastic responses	Calibration and ranking stability
Goal or eligibility	Target near the admitted support boundary	Acceptance coverage and selective risk

Shifts in Table 15 are crossed where physically interpretable. Expanded observations test whether mechanism and observation error are distinguishable when aggregate residuals agree.

I.4 When the declared envelope is wrong

Every guarantee in Appendix E is conditional on the true explanation lying in the declared envelope Ξ_r. That condition is not verifiable in a real study, so what matters is how the procedure fails when it breaks. Table 16 removes the true mechanism from the envelope in a prespecified fraction of assigned worlds, leaving everything else fixed.

The two right-hand columns carry the result. The empty-set diagnostic rate tracks the omission rate almost one for one, recovering 88% of the omitted-mechanism worlds as explicit adequacy alarms, whereas false support rises only from 5.0% to 8.0% when half of all worlds are misspecified. The failure mode is therefore mostly an audible one: an envelope that cannot explain the data exhausts its surviving set and says so, rather than concentrating belief on the best of several wrong accounts. The residual 12% that stays silent is the part a discrepancy coordinate can absorb, and it is where the extra false support comes from. This is the strongest statement available for a conditional guarantee, and it is an empirical one: it does not make Proposition 7 unconditional.

Table 16: **Behaviour under an inadequate envelope.** The generating mechanism is deleted from Ξ_r in the stated fraction of assigned worlds; budgets, observations and thresholds are unchanged. Mean over 32 sources. Empty-set diagnostics are blocks whose surviving set is empty, which Appendix A.4 routes to an adequacy alarm rather than a decision. Omission changes the difficulty of the task, so these rows are not ranked against one another.

Envelope adequacy	Resolutions ↑	False support (%) ↓	Empty-set diag. (%)	Scope acc. (%) ↑
Adequate, no omission	4.0	5.0	1.2	82.0
Omitted in 10% of worlds	3.8	5.6	10.0	80.3
Omitted in 25% of worlds	3.5	6.5	23.2	77.8
Omitted in 50% of worlds	3.0	8.0	45.2	73.5

I.5 Misspecified observation noise

The relation tests assume the conditional Gaussian law of Equation (18) with known covariance. Table 17 scales the true error covariance by ρ while the tests continue to use the nominal one, so $\rho > 1$ means the procedure believes its measurements are more precise than they are.

Table 17: **Sensitivity to underestimated observation noise.** The analysis uses the nominal covariance while the generating covariance is inflated by ρ. Any-false-decision rate is per relation stream at nominal $\gamma = 0.05$, over 32 sources.

ρ	Any false decision per stream (%)	Resolutions	Empty-set diagnostics (%)
1.00	3.1	4.0	1.2
1.25	5.8	4.2	1.5
1.50	9.6	4.4	1.9
2.00	18.4	4.6	2.8

At $\rho = 1$ the realised rate is 3.1% against the nominal 5%, the conservatism the union bound predicts. The direction of failure under misspecification is the uncomfortable one: resolutions *rise* with ρ while false decisions rise faster, because an over-tight covariance shrinks every confidence set and makes containment in $\mathcal{R}_r$ easier to certify. Yield is therefore not a safe monitor of calibration health, and the empty-set rate barely moves, so it does not detect this failure either. The mean-error set $\mathcal{D}_e$ and its independent calibration are what bound the damage, which is why they are fixed before the outcome rather than fitted to it.

I.6 Scaling in envelope size

Set inversion is the cost driver, so Table 18 varies the number of admitted explanations per relation from the 11 of the worked case to 88.

Table 18: **Cost and yield against envelope size.** Explanations admitted per relation version, at a fixed 2048-node branch-and-bound budget and fixed scientific budget. Time is online acquisition and certification relative to the 11-explanation configuration.

Explanations per relation	Resolutions	Diagnostic abstention	Median nodes	Time ratio
11	4.0	0.4	412	1.00
22	3.9	0.6	701	1.42
44	3.7	0.9	1284	2.31
88	3.4	1.4	2048	4.06

Yield degrades gracefully and the loss is paid in abstentions rather than errors: quadrupling the envelope from 22 to 88 costs 0.5 resolutions and adds 0.8 abstentions, while false support stays within 0.4 percentage points of 5.0% throughout. At 88 explanations the median node count reaches the budget, so the abstention rate there is set by the node cap and not by the evidence; raising the cap trades wall time for resolutions along the frontier in Table 18 rather than changing the decisions that are already certified.

J Estimands and Statistical Analysis

J.1 Graph and discovery endpoints

A fixed MIOY/scope rubric separates support from contradiction.

The resolution scoring function. One rule defines the resolution endpoint for every policy in this paper. Let $\mathcal{V}_w$ be the relation versions a policy brings to test in assigned world w. Each receives exactly one terminal label from Appendix A.4: certified support, certified falsification, or unresolved, the last covering both an inconclusive surviving set and a diagnostic abstention where branch-and-bound exhausted its node budget. Then

$$\mathrm{Res}(w) = \big|\{v \in \mathcal{V}_w : \text{ certified support}\}\big| + \big|\{v \in \mathcal{V}_w : \text{ certified falsification}\}\big|, \tag{46}$$

and unresolved versions contribute zero regardless of why they are unresolved. An abstention is therefore never a resolution, although it is a correct output and the status macro-F1 of Table 8 credits it as one of that vocabulary's classes; the two endpoints answer different questions and are not interchangeable. Restricted resolution cost censors every unresolved version at the budget B, so an abstention is charged in full. Table 19 reports the classes behind Equation (46) for all seven policies; the same function is applied to the acquisition controls of Table 5, the WAM comparison of Table 8, and the recording transfer of Table 10.

Table 19: **Decision classes behind the resolution endpoint.** Mean count per assigned transport world over 32 sources. Resolutions are the first two columns only, by Equation (46). Every policy is assigned the same 6.4 reference relation versions per world, so the four count columns sum to 6.4 in every row; policies differ in how many versions they bring to a certified terminal decision, not in how many exist. False supports are the subset of certified supports the reference contradicts, and dividing that column by the first reproduces the false-support percentage of Table 2.

Policy	Cert. support	Cert. falsif.	Resolutions	Diagnostic abstention	Other unresolved	False supports
Random design	1.5	0.7	2.2	0.4	3.8	0.180
Bayesian adaptive design	2.0	1.0	3.0	0.5	2.9	0.180
Discrepancy-aware design	2.2	1.0	3.2	0.5	2.7	0.165
Fixed hypothesis graph	2.2	1.0	3.2	0.4	2.8	0.154
Matched external tool agent	2.3	1.1	3.4	0.4	2.6	0.184
Agent-in-Twin (ours)	2.8	1.2	4.0	0.4	2.0	0.140
Deterministic Agent-in-Twin	2.7	1.1	3.8	0.4	2.2	0.149

Abstention loads are close across policies because an abstention is produced by the numerical set-inversion step, whose difficulty is set by the envelope dimension of the relation under test rather than by the policy that chose the experiment. Assignments are shared, so the per-world mix of low- and high-dimensional envelopes is matched by construction. That is why rescoring moves every level by about 0.4 and leaves every paired difference in Table 3 unchanged: the differences are invariant to a common additive shift, and the shifts differ by at most 0.1 across policies.

False support. Two distinct estimands are reported and must not be conflated. The column in Tables 2, 5 and 8 is the *relation-level false-support proportion*: within a source, the numerator counts relation versions the policy declared supported whose reference status is not supported, and the denominator counts all versions it declared supported, both pooled over that source's 12 task-replicates before the ratio is taken. For the Agent-in-Twin configuration the mean denominator is 33.5 declared supports per source, matching the 2.8 per world of Table 19 over 12 worlds. The *world-level false-support incidence* of Table 4 is a different quantity: the binary indicator that a world contains any false support, averaged over the 12 worlds of a source.

The two differ in resolution, and the difference is visible in their dispersion. A per-source mean of 12 binary indicators lies on a lattice of spacing $1/12$, and any variable on that lattice with mean $\bar{X}$ satisfies $X_j^2 \geq X_j/12$, so its sample variance obeys $s^2 \geq \frac{32}{31}\bar{X}(\frac{1}{12} - \bar{X})$; at a 5% mean this forces $s \geq 4.15$ percentage points. The tabulated relation-level proportion is not subject to that floor, because its lattice spacing is $1/33.5 = 0.030$, below the 0.05 mean, which makes the bound vacuous. Reporting 5.0 ± 2.8 for the proportion and ± 4.6 for the paired incidence contrast is therefore consistent; reporting the proportion's dispersion against the incidence estimand would

not be. Table 20 gives the 32 source-level numerators and denominators so both statistics can be recomputed.

Table 20: **Source-level false support for the Agent-in-Twin configuration.** For each of the 32 independent sources: false supports over declared supports, pooled across its 12 task-replicates, and the resulting percentage. Mean 5.00%, sample SD 2.80% across sources; the pooled ratio $53/1073 = 4.94\%$ differs because sources carry unequal denominators.

Source			Source			Source			Source		
#	n/d	%	#	n/d	%	#	n/d	%	#	n/d	%
1	2/30	6.7	9	2/34	5.9	17	1/32	3.1	25	1/31	3.2
2	2/31	6.5	10	2/34	5.9	18	1/33	3.0	26	1/32	3.1
3	3/32	9.4	11	0/35	0.0	19	3/33	9.1	27	3/33	9.1
4	3/32	9.4	12	1/35	2.9	20	1/34	2.9	28	1/34	2.9
5	2/33	6.1	13	2/36	5.6	21	1/34	2.9	29	2/34	5.9
6	2/33	6.1	14	1/36	2.8	22	2/35	5.7	30	0/35	0.0
7	2/33	6.1	15	3/30	10.0	23	1/35	2.9	31	3/36	8.3
8	0/34	0.0	16	2/31	6.5	24	2/36	5.6	32	1/37	2.7

Precision/recall use the reference graph. Resolution cost includes interventions/observations, censoring unresolved tasks at B; resolution probability accompanies restricted cost. Scope uses reference applicability.

Neuronal task applicability and episode yield. Let a_p denote applicability for predicate task p. Input resistance, firing gain, and adaptation have applicability 0.72, 0.66, and 0.66, giving macro-applicability $\bar{a} = (0.72 + 0.66 + 0.66)/3 = 0.68$. An episode is eligible when the required queries for its assigned task are admitted; the three predicate tasks impose separate eligibility conditions.

Conditional graph yield $\bar{y}$ counts relations resolved by Equation (46) across the full applicable episode. Two aggregations of the all-assigned summary must be distinguished. The stratified mean is $\sum_p w_p a_p \bar{y}_p$ with task-assignment weights w_p; the product form is $\bar{a}\bar{y}$. They coincide only when a_p and $\bar{y}_p$ are uncorrelated across tasks. We report both. With equal task weights $w_p = 1/3$ and per-predicate conditional yields $\bar{y}_p = 2.90, 2.90, 2.75$ for input resistance, firing gain and adaptation, the stratified mean is

$$\textstyle\sum_p w_p a_p \bar{y}_p = \frac{1}{3}(0.72 \cdot 2.90 + 0.66 \cdot 2.90 + 0.66 \cdot 2.75) = 1.939, \tag{47}$$

against $\bar{a}\bar{y} = 0.68 \times 2.85 = 1.938$. The two differ by 0.001 relations per assigned world, below the reporting precision of Table 10, so the tabulated 1.94 is unchanged by the choice. The same check for the fixed-graph and pooled-scope controls gives 1.531 against 1.530 and 1.599 against 1.598. The product form is therefore adequate here but is not adopted as a general identity. In Figure 7d, each bar instead measures recovery of the task's designated predicate: its conditional rate is r_p, and its all-assigned rate is $a_p r_p$. These predicate-level probabilities are distinct from episode-wide relation counts and are not summed to obtain graph yield.

Scope accuracy. For each matched relation, prespecified contexts receive inside/outside reference-scope labels. Scope accuracy is balanced accuracy,

$$\mathrm{Acc}_{\mathrm{scope}} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right). \tag{48}$$

Contexts require both classes; scores aggregate within units with scope precision/recall. Matching coverage uses all reference relations, retaining unmatched relations in recall. Narrowing scope changes sensitivity/specificity.

J.2 Prediction and intervention endpoints

Forward scores share observation laws and units; intervention contrasts match initial conditions, and route comparisons target the same conditional.

Normalized intervention and query diagnostics. Let $s_j > 0$ be a channel scale estimated from training sources and frozen for evaluation. For a matched intervention pair, define

$$E_\Delta = \left[\frac{1}{d} \sum_{j=1}^{d} \left(\frac{(\widehat{Y}_j^{I_1} - \widehat{Y}_j^{I_0}) - (Y_j^{I_1} - Y_j^{I_0})}{s_j} \right)^2 \right]^{1/2}. \tag{49}$$

Here j spans channels and times. Fidelity counts admissible assigned pairs satisfying a development-fixed $E_\Delta \le \tau_\Delta$. The configuration uses $\tau_\Delta = 0.2$. Failed or invalid predictions remain in the denominator with their failure type.

Replacing the numerator by the decoded-mean difference gives E_{route}: conditional-mean agreement. Proper scores assess distributional predictions.

For N frozen programs, coverage is $N^{-1} \sum_i A_i$ with acceptance indicators A_i. Selective risk counts goal failure or violation among accepted programs; report both separately and leave zero-acceptance risk undefined.

The certification rate is the fraction of frozen programs whose simultaneous bounds $P_i^\star(G) \ge L_g(i)$ and $P_i^\star(V) \le U_v(i)$ clear both thresholds; it is distinct from the simultaneous coverage probability $1 - \alpha - \beta$ that those bounds hold, which is fixed by construction rather than estimated. Risk–coverage curves vary prespecified thresholds on identical proposals/evaluations, retaining accepted/assigned counts; common coverage isolates verification from search.

J.3 Independent units and paired estimates

Reported comparisons use 32 independent source units per configuration: source/generator families or neuronal donors. Three technical replicates are averaged within source, with tasks/cells grouped by source; windows/times add no units. The experimental budget is 16 experiments/task (Appendix A).

Report mean and sample SD across source summaries. Paired-ablation SD uses source-level method-minus-control differences, not differences of marginal SDs; it measures between-source dispersion.

For paired differences d_j, report $\widehat{\Delta} = J^{-1} \sum_j d_j$, cluster-bootstrap 95% intervals, and stratum effects. Confirmatory paired randomization requires exchangeability and within-family Holm correction.

Realised hierarchy and analysis outputs. The unit of analysis is the source: a generator family in the transport setting, a donor in the recording setting. Each transport source contributes 4 assigned discovery tasks and 3 technical replicates per task; each donor contributes 3 predicate tasks over 6 admitted sweeps and 3 replicates. Replicates are averaged within source before any contrast, so the effective sample size for every interval in Tables 3, 4 and 11 is $J = 32$ sources, not the 384 transport task-replicates or 288 donor task-replicates they summarise. Task-family strata in Table 11 partition the sources into four disjoint groups of eight, so those intervals use $J = 8$.

Cluster-bootstrap intervals resample sources with replacement, $10{,}000$ draws, percentile method, with the full within-source structure carried along.

Randomization tests flip the method label within source and use the two-sided statistic $|\widehat{\Delta}|$. Which procedure is valid depends on how many sources the family contains, and the two cases are reported differently. For the 32-source families of Tables 3 and 4 the sign-flip space has 2^{32} points, so we sample $10{,}000$ of them; the smallest attainable Monte Carlo value is $1/10{,}001 = 1.0 \times 10^{-4}$, and after Holm correction within a family of at most six rows the smallest reportable value is 6.0×10^{-4}, which is what "< 0.001" abbreviates. For the four task-family strata of Table 11 each family holds only eight sources, so the sign-flip space has $2^8 = 256$ points and we enumerate all of them exactly rather than sampling. Exact enumeration matters here: the smallest two-sided value any eight-source family can produce is $2/256 = 0.0078$, attained only when every source differs in the same direction, and Holm correction over four families cannot return anything below $4 \times 0.0078 = 0.031$. A Monte Carlo approximation can report values below that floor, and they are artifacts of the sampling rather than evidence. The observed raw values are 0.0078, 0.1250, 0.7266 and 1.0000; all eight sources of the omitted-relation family differ in the same direction, which is why that family attains the floor. Holm correction gives the 0.031, 0.375, 1.000 and 1.000 tabulated there. The equally weighted mean row of that table is not a fifth member of the family: it is the 32-source aggregate contrast of Table 3, where $10{,}000$ draws are available and 0.006 is attainable.

Denominators retain unresolved, inapplicable and failed tasks: over the transport configurations, 2.1% of assigned tasks terminated without a released outcome and are counted as unresolved rather than dropped, and 0.4% failed for numerical reasons and are reported separately. No result in this paper is obtained by excluding a source, a task, or a replicate after seeing its outcome.

J.4 Precision, missingness, and multiplicity

Development-only paired variability and a meaningful effect $\Delta_{\min}$ initialize the two-sided unit count:

$$J \approx \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_d^2}{\Delta_{\min}^2}$$

Here β is type-II error and units are independent sources. At 80% power and unadjusted two-sided level 0.05, the factor is approximately 7.85. Assumed paired SD 1 and effect 0.5 imply the 32 sources used throughout.

The same expression explains why incidence endpoints are aggregated within source before any contrast. Treating a single world as the unit gives paired binary incidence variance $q - \Delta^2$ for discordance probability q; at $q = 0.15$ and a three-percentage-point change that design would need about 1,301 units. Aggregating the 12 worlds of a source first reduces the dispersion to the realised source-level difference SD of 4.6 percentage points, and at $\Delta_{\min} = 6$ percentage points the expression returns $J \approx 5$, so the 32-source design is amply powered for the incidence contrast in Table 4. These planning figures precede multiplicity and stratum adjustment. The four-family strata of Table 11 carry eight sources each and are correspondingly less precise, which is why their intervals are wide and three of the four cover zero.

Missing, failed, and timed-out runs retain prespecified classes and denominators, with missingness sensitivity analysis. Discovery, verification, and calibration budgets combine in Corollary 1.

Finite-program event calibration. Freeze K complete programs, their branch/stopping rules, target population, checkpoint, event definitions, and thresholds before calibration. Independent reference/WAM calibration batches have fixed sizes m_a, l_a and frequencies $\widehat{p}^{\mathrm{cal}}_{P,a}, \widehat{p}^{\mathrm{cal}}_{Q,a}$. Define

$$\delta_a = \min\Bigg\{1,\ \max_{B \in \{G,V\}} \left|\widehat{p}^{\mathrm{cal}}_{P,a}(B) - \widehat{p}^{\mathrm{cal}}_{Q,a}(B)\right| + \sqrt{\frac{\log(8K/\beta)}{2m_a}} + \sqrt{\frac{\log(8K/\beta)}{2l_a}}\Bigg\}. \tag{50}$$

Two-sided Hoeffding bounds over $4K$ means cover event gaps with probability $1 - \beta$. Samples within each law are conditionally independent; same-rollout G, V may depend. Both laws share the frozen program/population. Independent bounded donor summaries replace sweeps for donor-average events.

Calibration leaves WAM, thresholds, and graph fixed. Fresh WAM rollouts verify; independent reference batches assess coverage. Source/donor and random-stream audits reject overlap, missing batches, or changed checkpoints. New programs/populations require recalibration or transfer bounds. Frozen-set selection retains coverage at least $1 - \alpha - \beta$.

Event-gap calibration differs from trajectory total variation; covariance scaling cannot supply δ_a. Frozen bounded utilities admit range-scaled two-sample Hoeffding calibration. Unbounded information gain requires tail control, and coarse histograms cannot bound full-distribution total variation.

Composite relation tests. Under Appendix A.4's predictable Gaussian law, the allowance contains d_e, giving $T_e(\xi^\star) \le \|\Sigma_e^{-1/2}\epsilon_e\|^2$ and conditionally super-uniform $p_e(\xi) = 1 - F_{\chi^2_{m_e}}(T_e(\xi))$ at truth. Support tests $H_0 = \Xi_r \setminus \mathcal{R}_r$; falsification tests $H_0 = \mathcal{R}_r$, using

$$p_e(H_0) = \sup_{\xi \in H_0} p_e(\xi). \tag{51}$$

The supremum dominates the true-null value. Rejection requires a certified upper bound; particle/grid maxima are lower bounds. Set intersection combines blocks under nonempty-witness and outer-set rules, without multiplying dependent p-values.

For an exactly simulable non-Gaussian law, fix all noise nuisances and draw B independent null replicates. A prespecified incompatibility statistic gives $[1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \ge T_{\mathrm{obs}}\}]/(B+1)$, conservative under exchangeability, including ties. Composite tests require a certified full-null supremum. Noise uses independent calibration or a nuisance envelope with its own coverage budget. Simulator-rank tests extend the specified Gaussian transport test.

Sequential operating characteristics. Allocation $\eta_{r,\ell} = \gamma/[r(r+1)\ell(\ell+1)]$ covers adaptive births/scope revisions. Vary births, effects, and within-block covariance; measure any false decision per stream and true resolutions, adding a misspecified-noise sensitivity arm. The uncorrected control changes allocation with selections/outcomes fixed. Independent streams give binomial family-wise-error intervals; source-clustered intervals quantify power (Figure 7).

K Certification Ledger, Support Certificates, and a Replayed Discovery Trace

This appendix closes three gaps between the guarantees stated in Appendix E and the decisions actually reported: which rollout counts were realised and how they were chosen, which relation decisions carry a numerical certificate rather than a particle approximation, and what one complete discovery episode looks like end to end. Every quantity here is a realised protocol output, not a planning target.

K.1 Realised event calibration and the sampling schedule

The schedule is fixed before calibration. Theorem 1 requires the per-program rollout counts to be fixed in advance. We therefore run a single prespecified schedule, $n = m = l = 8192$, for every frozen program, with no outcome-dependent extension and no early stopping. A separate development screen of 128 rollouts per candidate is used only to select which programs enter the frozen set; it is discarded before calibration and contributes to no reported bound. This removes optional stopping as a source of invalidity by construction, at the cost of spending the full budget on programs that a sequential rule would have abandoned. A valid sequential alternative would require an always-valid boundary or a preallocated multi-stage error budget; we do not claim one.

The frozen certification set is distinct from the proposal pool behind the risk–coverage curve of Figure 6b: that curve sweeps acceptance thresholds over a larger pool of candidate programs and reports selective risk, whereas certification issues simultaneous bounds for a small set fixed in advance, and its K enters the union bound. With $K = 8$ frozen programs, $\alpha = \beta = 0.025$, and equal per-law counts, the sampling and calibration radii are

$$\epsilon = \sqrt{\frac{\log(2K/\alpha)}{2n}} = 0.0199, \qquad \delta_a = \mathrm{gap}_a + 2\sqrt{\frac{\log(8K/\beta)}{2 \cdot 8192}} = \mathrm{gap}_a + 0.0438, \tag{52}$$

so ϵ is common to all programs and only the measured event-frequency gap varies. At the initial screening size the same expression gives $\epsilon = 0.1589$, which alone exceeds the violation threshold $\tau_v = 0.12$; no certificate can be issued at that size, which is why the screen is excluded from certification.

Per-program ledger. Table 21 lists the realised inputs and outputs for all eight frozen programs, and Table 22 repeats the decision when the same reference budget is spent on direct reference certification. Acceptance requires $L_g \geq \tau_g = 0.8$ and $U_v \leq \tau_v = 0.12$ simultaneously. Three programs remain Undetermined under the two-stage path, and each misses both bounds: their goal lower bounds fall short of τ_g and their violation upper bounds exceed τ_v, so neither criterion alone explains the status. Their empirical frequencies $\widehat{p}_g, \widehat{p}_v$ all satisfy the thresholds, so what withholds the certificate is the width of $\epsilon + \delta$ rather than the measured behaviour, which is why the direct-reference ledger decides all three. They are reported as unresolved, not as failures, and they stay in the denominator of the certification rate. We reserve the word *coverage* for the simultaneous probability that the displayed bounds hold, which is $1 - \alpha - \beta$ by construction and is not estimated from these tables; the fraction of programs that receive a certificate is the certification rate.

Every program clears both thresholds under direct certification, so the two paths certify 5 of 8 and 8 of 8 on identical reference spend. The reference verification frequencies sit below the WAM ones on the goal event and above them on the violation event, by 0.0195 to 0.0312 and 0.0039 to 0.0132 respectively; those verification-batch differences are smaller than the calibration gaps of Table 21, as they should be, since a gap bounds the population event difference rather than one batch's realisation.

Where the two-stage path pays, and where it does not. At this operating point it does not. The two-stage path spends $8 \times (n + m + l) = 196{,}608$ rollouts, of which 65,536 are reference executions and 131,072 are WAM executions, and certifies 5 of 8 programs; the same 65,536 reference rollouts spent directly remove δ and certify 8 of 8. Direct certification dominates whenever reference executions are available in the number the union bound requires, and we do not claim otherwise.

Table 21: **Realised event-calibration ledger for the frozen program set.** All programs use the same prespecified $n = m = l = 8192$. The gap is the measured maximum event-frequency difference between the reference and WAM calibration batches; $\delta = \text{gap} + 0.0438$ follows Equation (50). Table 22 spends the same reference budget directly.

Program	$\widehat{p}_g$	$\widehat{p}_v$	gap	δ	L_g	U_v	Decision
P_1	0.9219	0.0117	0.03125	0.0750	0.8270	0.1066	accept
P_2	0.9063	0.0156	0.03516	0.0789	0.8075	0.1144	accept
P_3	0.9297	0.0078	0.02734	0.0711	0.8387	0.0988	accept
P_4	0.9424	0.0068	0.03125	0.0750	0.8475	0.1017	accept
P_5	0.8828	0.0234	0.04297	0.0867	0.7762	0.1300	undetermined
P_6	0.8594	0.0313	0.03125	0.0750	0.7645	0.1262	undetermined
P_7	0.9502	0.0049	0.02930	0.0731	0.8573	0.0978	accept
P_8	0.8711	0.0430	0.03906	0.0828	0.7684	0.1457	undetermined

Table 22: **Direct reference certification** on the same 8×8192 reference rollouts, for which δ vanishes and $\epsilon = 0.0199$. Reference verification frequencies are reported so the column can be recomputed independently.

Program	$\widehat{p}_g^{\,\mathrm{ref}}$	$\widehat{p}_v^{\,\mathrm{ref}}$	L_g	U_v
P_1	0.8984	0.0186	0.8785	0.0385
P_2	0.8809	0.0261	0.8610	0.0460
P_3	0.9102	0.0117	0.8903	0.0316
P_4	0.9199	0.0134	0.9000	0.0333
P_5	0.8516	0.0356	0.8317	0.0555
P_6	0.8379	0.0396	0.8180	0.0595
P_7	0.9287	0.0098	0.9088	0.0297
P_8	0.8428	0.0562	0.8229	0.0761

What the two-stage construction buys is amortisation, and that is a statement about how the cost scales, not about a single ledger. A direct certificate needs its own reference batch for every program, so its per-program radius is $\sqrt{\log(2K/\alpha)/(2B/K)}$ and widens as the program count K grows against a fixed reference budget B. One calibration batch, by contrast, serves every program in the same frozen event family: the reference term in δ is $\sqrt{\log(8K/\beta)/(2B)}$ and shrinks with B alone, while the per-program sampling term is paid in WAM rollouts, which are not the scarce resource. Table 23 evaluates both paths on the same eight program profiles replicated to K programs.

Table 23: **Reference-budget frontier.** Certification rate (%) for K frozen programs drawn from the profiles of Table 21, at a fixed total reference budget B. Direct certification splits B into B/K per program; the two-stage path spends B once on a shared calibration batch and 8192 WAM rollouts per program. Union-bound constants track K in both paths. Rows are two paths under two budgets rather than competing methods, so no cell is marked best; the crossover is the quantity of interest.

	Frozen programs K				
Path and reference budget	8	16	32	64	128
Direct, $B = 16{,}384$	87.5	62.5	50.0	0.0	0.0
Two-stage, $B = 16{,}384$	62.5	62.5	62.5	62.5	62.5
Direct, $B = 65{,}536$	100.0	100.0	75.0	62.5	50.0
Two-stage, $B = 65{,}536$	62.5	62.5	62.5	62.5	62.5

The crossover is where the claim lives. At the scarce budget the two-stage path matches direct certification at $K = 16$ and overtakes it from $K = 32$, and direct certification issues no certificate at all beyond $K = 32$ because its per-program radius exceeds the acceptance margin. At the ledger's own budget the crossing moves out to $K = 64$–128. The honest summary is therefore narrower than a general efficiency claim: the two-stage construction converts a shortage of reference executions into a quantified, simultaneously valid penalty, and it is the cheaper path only when one calibration

batch is amortised over many programs of the same event family. At $K = 8$ it is the more expensive path and the ledger says so.

Error budget across rounds. Algorithm 1 may certify in more than one round. The discovery tests spend γ through the birth-indexed allocation of Appendix A.4; certification spends α and β through a round-indexed split $\alpha_k = \alpha/[k(k+1)]$ and $\beta_k = \beta/[k(k+1)]$ whenever more than one round is scheduled. Corollary 1 then applies with the original constants. The analysis plan for this evaluation fixed the number of certification rounds at one before any program was frozen, so the full budget $\alpha = 0.025$ is spent in that round and the split is not activated; the round count was not chosen after seeing which programs failed, which would have required the split and the smaller $\alpha_1 = 0.0125$.

K.2 Which relation decisions carry a numerical certificate

Appendix A.4 requires three numerical objects before a relation may be declared supported: a certified outer set $\overline{\mathcal{C}} \supseteq \mathcal{C}_{r,\ell}$, a certified lower bound on $\inf_{\xi \in \overline{\mathcal{C}}} g_r(\xi)$, and a verified feasible witness establishing nonemptiness. Particle or grid maxima supply neither the outer set nor the supremum required for a composite null; they are lower bounds on a quantity that must be upper-bounded. We therefore separate two classes of decision and report them separately rather than pooling them.

Certified decisions use interval arithmetic over the admitted parameter boxes, with the transport response evaluated through the certified removal enclosure of Appendix E.4 on each box and a branch-and-bound refinement to a relative tolerance of 10^{-3} on g_r. That enclosure is stated and proved for Equation (17) itself, with its exchange boundary, spatially varying coefficients and bounded domain; the whole-space moment identities of Appendix E.3 are an identifiability illustration and are used by no certified decision. Nonemptiness is established by exhibiting a witness explanation and verifying its residual against the same threshold. Diagnostic decisions are those where branch-and-bound did not close the gap within the node budget; they are labelled unresolved in the graph, are excluded from every resolution count, and are charged the full budget in the cost column.

The resolution endpoint counts the 4.0 certified terminal decisions per assigned world and excludes the 0.4 abstentions forced by an unclosed bound. An abstention is a correct output of the procedure, and the relation-status macro-F1 of Table 8 credits it, because unresolved is a class in that label vocabulary; it is not a discovered relation, so it earns nothing in the resolution count and is censored at the budget in the cost column. Table 19 applies the same scoring function to every policy. Median node counts are the branch-and-bound effort behind each class, and the abstention row sits at the 2048-node budget by definition. The abstention rate rises with envelope dimension: it is 4% for the three-parameter elementary accounts and 23% for the joint exchange–diffusion accounts with a scope predicate, which is the expected cost of set inversion in higher dimension. Because envelope dimension is a property of the assigned relation rather than of the policy that selected the experiment, and assignments are shared, abstention loads are close across policies, which is why the rescoring moves every level by about 0.4 and leaves the paired differences of Table 3 unchanged.

Table 24: **Formal certification status of relation decisions**, per assigned transport world. Certified decisions carry an outer set, a certified effect bound and a verified witness; diagnostic decisions do not and are scored unresolved. Means over 32 sources; the two certified classes, and only those, sum to the resolutions of Table 2.

Decision class	Per world	Share (%)	Median nodes
Certified support	2.8	70.0	412
Certified falsification	1.2	30.0	233
Resolutions	4.0	100.0	—
Not resolved			
Diagnostic abstention	0.4	—	2048

One certificate in full. Relation 17 of Appendix K.4, after its second block, has removal horizon $T = 1200\,\mathrm{s}$ and surviving box $k_c \in [2.75, 3.35] \times 10^{-4}\,\mathrm{s}^{-1}$, exchange coefficient $\kappa_I \in [3.72, 4.40] \times 10^{-7}\,\mathrm{m\,s}^{-1}$ on the vehicle arm I_1 and $[1.60, 2.00] \times 10^{-7}$ on the AQP4-proxy arm I_0. The box-uniform occupancy bracket of Proposition 5 returns $\Phi_h = 1.200 \times 10^6\,\mathrm{m}^{-1}\mathrm{s}$ with $\eta_{\mathrm{pr}}\eta_{\mathrm{ad}} = 1.48 \times 10^4$ and $\eta_{\mathrm{fp}} = 2.1 \times 10^1$, hence $\Phi(\xi) \in [1.185, 1.215] \times 10^6$ for every ξ in that box. The exponent brackets are therefore $k_c T \in [0.3300, 0.4020]$, $\kappa_I \Phi \in [0.4408, 0.5346]$ on I_1 and $[0.1896, 0.2430]$ on I_0, and substituting the corners into Equation (41) gives $U(I_1) \in [0.537, 0.608]$ and $U(I_0) \in [0.405, 0.475]$.

Differencing those enclosures directly gives only $g_{17} \in [0.062, 0.203]$, whose lower end falls short of $\Delta_r = 0.08$. The looseness is not physical: the corner difference lets k_c take its largest value

on one arm and its smallest on the other, whereas a paired contrast shares k_c between the arms. Subdividing on the shared coordinates removes exactly that slack, and Equation (43) becomes $e^{-k_c^+ T}\big(e^{-\kappa_0^+ \Phi^+} - e^{-\kappa_1^- \Phi^-}\big) = 0.0942$; branch-and-bound reaches the certified value $\underline{g}_{17} = 0.094$ at 412 nodes, above Δ_r. The witness is the explanation with $k_c = 3.05 \times 10^{-4}$, $\kappa_I = 4.02 \times 10^{-7}$ and no sensor-gain offset, whose certified residual upper bound is $\overline{T}_e = 14.6$ on $m_e = 10$ channels against $\chi^2_{10,\,1-\eta_{17,2}} = 38.83$, so the surviving set is nonempty. Mechanism membership holds throughout $\overline{\mathcal{C}}$, and the three conditions together issue the certificate.

An exhaustively enumerable instance checks the bound directions themselves: on a discretised envelope of 4096 explanations where $\mathcal{C}_{r,\ell}$ can be computed exactly, the outer set contained the exact surviving set in 4096 of 4096 cases and the certified effect bound never exceeded the exact infimum. That check validates the direction of the bounds; the continuous implementation is validated instead by Propositions 4 and 5, whose hypotheses Equation (17) satisfies, and by the estimator's own guaranteed error bound.

K.3 Population compatibility versus implemented route agreement

Proposition 3 is a statement about the population joint law: conditionals of one positive joint distribution are Bayes-compatible. It does not assert that the four implemented finite-step query routes realise those conditionals. Sharing network weights and masked objectives does not establish that property, and we do not claim it. Lemma 2 supplies the only implemented-level guarantee available here, and it is conditional on the per-route total-variation bounds η_r of assumption 6, which are assumed rather than certified. The conditional-mean route disagreement in Figure 5a is an empirical diagnostic of that assumption, not a proof of it; a small disagreement is consistent with, but does not imply, a small η_r.

On the enumerable instance above, where the joint law over 4096 explanations can be normalised exactly, the four routes agree with the exact conditionals to a total variation of 0.006, 0.009, 0.011 and 0.008 for the forward, inverse, sensing and validity routes. This is a finite check on a small instance and bounds nothing on the full-scale model.

K.4 A REPLAYED DISCOVERY TRACE

The following episode is reproduced from the recorded trace of a single assigned world in the boundary-exchange stratum. The hidden mechanism is withheld from the Agent throughout and released only for scoring. The main-text summary appears in Section 6.4 and Figure 4; the candidate structure appears in Appendix B.6 and Figure 8. Local MIOY candidate indices differ from the persistent relation IDs below.

01 Birth

$r = 17$ / **CANDIDATE**

AGENT / register a mechanism, scope, and effect threshold

Claim. An AQP4-proxy reduction acts through boundary exchange κ_I, rather than sensor gain a_j, with meaningful effect $\Delta_r = 0.08$ on compartment removal over $T = 1200\,\mathrm{s}$.

Scope. The predicate is fixed at birth from the wall-displacement calibration carried by every context:

$$\mathcal{S}_{17} = \{C : \rho(C) \geq 0.6\}, \qquad \rho(C) : \text{ compliance proxy.}$$

Contrast. $I_1 - I_0$: I_1 is the vehicle arm and I_0 the proxy-reduced arm. A genuine reduction of exchange therefore gives a positive removal difference.

Envelope Ξ_{17}: 11 explanations. Three elementary, two joint, two scoped, and four sensor or boundary discrepancy accounts reproduce the same removal change.

↓ commit the intervention and analysis before any outcome

02 Frozen selection

PRE-OUTCOME / LOCKED

AGENT / freeze the action pair, panel, falsifier, and error allocation

Action pair. Both arms start from the common reset state. I_1: a $2.0\,\mu\mathrm{L}$ bolus at $0.1\,\mu\mathrm{L\,min^{-1}}$ into the cisternal port with vehicle. I_0: the identical bolus with the AQP4 proxy at $10\,\mu\mathrm{M}$.

Observation. Co-located flux and concentration on the exchange patch, sampled at $\{0, 1, 3, 8, 20\}$ minutes; the falsifier is frozen with the panel.

Allocation. At $\gamma = 0.05$,

$$\eta_{17,1} = \frac{\gamma}{17 \cdot 18 \cdot 1 \cdot 2} = 8.17 \times 10^{-5}.$$

One paired intervention. This is the same contrast as in event 01. The flux panel distinguishes the accounts; it is not a second intervention.

↓ release independent reference measurements

03 Release and reduction

11 → 6 SURVIVORS

REFERENCE → TWIN / assimilate evidence and reduce the envelope

Receipt. On the exchange patch,

$$n \cdot J = 0.42 \pm 0.05 \text{ (patch units)}, \qquad c - c_{\mathrm{ext}} = 0.01 \pm 0.02.$$

Falsifier. A sensor-gain account preserves the boundary law and predicts

$$|n \cdot J| \leq \kappa^{+} |c - c_{\mathrm{ext}}| \leq 0.08$$

at this block's allocation, using the largest admitted κ and the upper end of the concentration contrast. The independently calibrated flux channel has bounded 3% gain uncertainty and is not a free parameter of this account.

Block 1 eliminates all 5 pure sensor-gain accounts. The observed flux exceeds the certified bound by more than five of its own standard errors. Six explanations remain, spanning κ_I and joint κ_I–D_I accounts.

Discovery trace / continued

CERTIFY THE EFFECT ⟶ TEST THE SCOPE ⟶ CREATE A NEW VERSION

04 Second block — UNRESOLVED / ACQUIRE

VERIFIER → AGENT / withhold support and request discriminating evidence

Checkpoint. The mechanism predicate m_{17} holds throughout the survivors, but the certified effect lower bound is insufficient:

$$\underline{g}_{17} = 0.061 < \Delta_r = 0.08.$$

Next action. Select a spatial-profile panel that separates κ_I from D_I, with

$$\eta_{17,2} = \frac{\gamma}{17 \cdot 18 \cdot 2 \cdot 3} = 2.72 \times 10^{-5}.$$

Decision: unresolved. Mechanism membership alone does not support the relation; the effect bound must also be certified.

↓ assimilate block 2 and verify a witness

05 Support — SUPPORTED ON $\mathcal{S}_{17}$

VERIFIER → GRAPH / certify the scoped relation

Evidence. Three explanations survive, all with $m_{17} = 1$. The certified effect bound rises to

$$\underline{g}_{17} = 0.094 \geq \Delta_r = 0.08, \qquad \text{with a verified witness.}$$

Decision: support relation 17 on $\mathcal{S}_{17}$. Record the scope and retain all 4 counterevidence entries.

↓ challenge the declared scope with a later context

06 Scope revision — r_{17} CONTRADICTED / r_{23} NEW

REFERENCE → AGENT → GRAPH / preserve the counterexample and create a new version

Context. The later context $C^\dagger$ is *inside* the declared scope:

$$\rho(C^\dagger) = 0.71 \geq 0.6, \qquad \text{wall amplitude} = 0.9\,\mu\text{m}.$$

Its wall motion is at the low end of the admitted range.

Counterexample. The certified *upper* bound is $\overline{g}_{17}(C^\dagger) = 0.021 < \Delta_r$, contradicting $\mathcal{R}_{17}$ in Equation (19) at a context the relation claimed. If $C^\dagger$ failed the predicate, the outcome would instead be recorded as out of scope and would test nothing.

New version. Birth relation 23 with the conjunction

$$\mathcal{S}_{23} = \left\{ C : \begin{array}{l} \rho(C) \geq 0.6, \\ \text{wall amplitude} \geq 1.5\,\mu\text{m} \end{array} \right\}.$$

This scope excludes $C^\dagger$; test relation 23 on fresh blocks.

History retained; program updated. Relation 17 keeps its evidence and counterexample and is marked contradicted on $\mathcal{S}_{17}$. Only relation 23 enters the compiled program; its certificate is row P_3 of Table 21.

TRACE INVARIANTS

Effect certification. Event 04 withholds support even when the mechanism predicate holds, because the effect bound is not certified.

Refutation within scope. Event 06 is a counterexample because $C^\dagger \in \mathcal{S}_{17}$. An outcome outside the declared scope cannot refute a statement quantified over it and is recorded separately.

Persistent versions. Revision creates a new relation rather than silently editing the old one. Supporting and defeating evidence remain attached to the original statement.

L Compute and Reproducibility

Resource accounting. Wall time is $T_{\text{total}} = T_{\text{prep}} + T_{\text{train}} + T_{\text{cal}} + \sum_t (T_{\text{LLM,t}} + T_{\text{design,t}} + T_{\text{ref,t}} + T_{\text{update,t}}) + T_{\text{verify}}$. We record seconds, hardware, precision, batching, peak memory, failed calls, and network/retry latency, and we separate offline training and simulation, online selection, and the scientific budget.

Table 25 reports the closed loop for the transport evaluation, stage by stage. The acquisition row is the measurement behind the time ratios of Table 5: $612/211 = 2.90$ against the tabulated 2.9 for joint versus plug-in EIG, and $549/211 = 2.60$ for the nested filter. Reference execution dominates every arm and is identical across them by construction, which is why a design policy that spends more computation to select the same number of experiments changes total wall time by only 21% while changing what those experiments settle.

Table 25: **Measured closed-loop cost per transport episode**, mean $\pm$ SD in seconds over the 384 episodes of the 32-source evaluation. An episode is one assigned world at the 16-experiment budget. Offline WAM training and the field codec are amortised across all configurations and are excluded; they are reported in the text. The deterministic control replaces the language model with ontology rules, and the plug-in column is the acquisition control of Table 5.

Stage	Agent-in-Twin	Deterministic	Plug-in EIG
Language-model proposal	96 ± 31	0 ± 0	96 ± 31
Observation design	612 ± 88	598 ± 84	211 ± 37
Reference execution	1243 ± 156	1240 ± 152	1255 ± 160
Belief and graph update	84 ± 12	79 ± 11	71 ± 10
Verification and certification	305 ± 41	303 ± 40	302 ± 39
Total online	2340 ± 191	2220 ± 180	1935 ± 175

The remaining accounting follows from that table. The transport evaluation runs $32 \times 4 \times 3 = 384$ episodes, so its online cost is 249.6 hours of sequential compute, or 7.8 hours of wall time on the 32 concurrent source workers used here. Event calibration adds $65{,}536$ reference rollouts at 0.72 s and $131{,}072$ WAM rollouts at 0.031 s, which is 14.2 hours sequential and 1.8 hours on eight workers. Offline WAM training costs 18.4 GPU-hours on one 80 GB accelerator and the field codec a further 3.1, both amortised over every configuration in the paper. The recording setting runs $32 \times 3 \times 3 = 288$ episodes with no partial-differential-equation reference solve, at 412 ± 63 s per episode, which is 33.0 hours sequential and 1.0 hour on 32 workers. Peak memory is 34.2 GB during training and 6.8 GB at inference, and field-query throughput is 318 queries per second. Cold-start and reuse-amortised costs differ only in T_{prep}, which is 41 s and 3 s respectively.

For n tokens, width d, and L steps, two-pass dense attention costs $O(Ln^2d)$ plus physical solves; field length scales with $T(N + R)$. Nested inference uses $O(DN_yPM)$ likelihood calls for D designs, N_y outcomes, P mechanism/discrepancy particles, and M nuisance draws. Scaling varies these counts independently, reporting resources and decision stability with shared-term caching.

Decision reconstruction. Replay binds source/split, configuration, code, environment, seeds, and checkpoints to proposals, outcomes, graphs, and certificates. Reference equations, meshes, tolerances, observations, and generator rules support reproduction. Policies see current evidence/actions; evaluators join hidden mechanisms after trace freezing. Measured, generated, and reference-only records retain separate provenance.