When Should an AI Scientist Stop? Verifiable Experiment Steering and Refusal for Autonomous Discovery

Neel Tushar Shah [1] **Manglam Kartik** [1]

Abstract

We present CARTOGRAPH, a verification layer for AI scientists that couples unresolved-subspace experiment steering (*select*), explicit ambiguity closure (*resolve*), and residual-based library inadequacy detection (*refuse*). Under a local linear-Gaussian bridge, raw unresolved projection is the isotropic unresolved Fisher-information trace, while CARTOGRAPH-A is the exact unresolved A-optimal rule; closed-form EIG and Box–Hill arise as local comparators rather than global equivalents. Across five testbeds, CARTOGRAPH-A beats raw projection $129\mathrm{W}/0\mathrm{T}/15\mathrm{L}$ at $d = 8$ ($p < 10^{-21}$) in a replicated structured cascade. More distinctively, the framework tentatively identifies three out-of-library pharmacokinetic mechanisms and then *revokes* those identifications as residuals expose structural misfit, while one perturbed in-library control stays identified throughout. In low-dimensional pharmacokinetic and filtered EPA settings, near-ties against disagreement are predicted by theory and observed. Finally, in a retrospective audit of 40 positive claims from the published A-Lab autonomous materials system, the refuse guard flags all 4 claims later marked inconclusive under manual re-analysis while passing $32/36$ confirmed claims. Codes can be found at the anonymous repository https://github.com/ai4science-boed/cartograph.git

1. Introduction

AI systems are now participating in closed scientific discovery loops: LLM planners propose experiments, automated laboratories execute them, and statistical or neural modules interpret the data (King et al., 2009; Burger et al., 2020; Boiko et al., 2023; Szymanski et al., 2023; Lu et al., 2024; Bran et al., 2024; Wang et al., 2023). Complementary work has produced end-to-end automated scientific capabilities high-accuracy protein structure prediction (Jumper et al., 2021), scaled materials discovery (Merchant et al., 2023), and symbolic-equation discovery from data (Schmidt & Lipson, 2009; Brunton et al., 2016; Udrescu & Tegmark, 2020; Cranmer et al., 2020) but none of these stacks emits a verifiable *refuse* signal when the library or hypothesis space is structurally inadequate. In this regime, the bottleneck is no longer *proposing* experiments, LLMs will produce more candidates than the lab can run but deciding *which* experiment is genuinely informative, *when* the current mechanistic question is resolved, and *when* the system should stop making claims altogether because the model library it is searching is structurally wrong.

This paper frames the verification-and-steering layer of an AI scientist as three linked decisions:

Select. Which candidate experiment most directly reduces the currently unresolved scientific ambiguity?

Resolve. When is the ambiguity small enough to declare a mechanistic question answered?

Refuse. When should the system stop identifying any model in the current library because the library itself is inadequate?

Existing modern Bayesian experimental design (BOED) (Chaloner & Verdinelli, 1995; Ryan et al., 2016; Rainforth et al., 2024; Foster et al., 2019; Kleinegesse & Gutmann, 2020; Blau et al., 2022; Foster et al., 2021) answers *select* thoroughly. Classical model discrimination criteria (Box & Hill, 1967; Atkinson & Fedorov, 1975; Pukelsheim, 2006) answer a restricted form of *select* and *resolve*. Neither literature makes *refuse* a first-class output: BOED assumes the prior support contains the truth; model-discrimination assumes at least one rival is correct. For high-stakes autonomous discovery clinical pharmacokinetics, materials synthesis, toxicology this gap matters.

Contributions. We contribute (i) a formal *access-model* distinction separating symbolic and behavioral querying

[1]Indian Institute of Technology Bombay, Mumbai, India. Correspondence to: Neel Tushar Shah <23b4244@iitb.ac.in>, Manglam Kartik <23b4243@iitb.ac.in>.

Accepted at the AI for Science workshop (ICML 2026).

of a scientific library, with a clean coverage-vs-rank recovery characterization; (ii) CARTOGRAPH, a verification-and-steering layer that combines unresolved-subspace steering for *select/resolve* with a residual- and gap-based *refuse* guard; (iii) a local BOED bridge: exact unresolved Fisher-information identity, exact $k = 1$ posterior-variance equivalence, and first-order links to closed-form EIG together with an isotropic-limit reduction of Box–Hill; (iv) an exact random-candidate scaling law that *predicts* when unresolved-subspace methods will merely tie simple disagreement heuristics and when they will dominate; (v) empirical evidence across a symbolic dynamical system, a scalable structured nonlinear cascade, a pharmacokinetic model-library benchmark, public EPA real-data time series, and a retrospective audit of A-Lab autonomous-materials claims, including a strongly replicated $d \in \{2, 4, 8, 16\}$ cascade ($p < 10^{-21}$ at $d = 8$), a principled refusal benchmark on three out-of-library pharmacokinetic mechanisms, and an A-Lab audit that flags all $4/4$ post-correction inconclusive positive claims; and (vi) a worked LLM-in-the-loop appendix example showing how CARTOGRAPH drops in as the verification layer of an LLM-planning AI scientist.

Where the paper lands: Our most distinctive empirical finding is not universal selection gain. In low-dimensional pharmacokinetics the gain over disagreement heuristics is genuinely modest, and our scaling theory predicts exactly this. Our distinctive finding is *revocation*: on three out-of-library mechanisms the framework is early-confident, then retracts those identifications as more evidence exposes structural misfit, while a perturbed in-library control stays identified. We argue this is what a governed AI scientist needs: an auditable "stop and escalate" signal inside the same sequential loop that chooses the next experiment.

2. Problem Setup

2.1. Model Library With a Shared Mechanism Basis

Let $\Phi = \{\phi_1, \ldots, \phi_p\}$ be a shared mechanism basis and let the true law be $T(x) = \sum_{j=1}^{p} a_j^\star \phi_j(x)$. A *library* $\mathcal{M} = \{M_1, \ldots, M_L\}$ retains different subsets of Φ with possibly different retained coefficients. The *controversial component* $a_C^\star \in \mathbb{R}^{p_C}$ is the subvector of coordinates on which library members disagree. Resolving $a_C^\star$ is the scientific task.

2.2. Two Access Models

The problem changes with what the AI scientist can actually observe.

Symbolic access. The agent inspects retained coefficient vectors directly. With omission-only retention (the library either keeps a basis term with its true coefficient or drops it), recovery is a *coverage* property: $a^\star$ is uniquely recoverable iff every mechanism appears in at least one library member.

Behavioral access. The agent cannot read coefficient vectors. It runs experiments indexed by e and observes disagreement features across model pairs, which assemble into a design

$$y = H\, a_C^\star + \varepsilon, \qquad H \in \mathbb{R}^{n \times p_C}, \tag{1}$$

where the rows of H are experimentally induced linear functionals of the controversial mechanisms. Recovery is now a *rank* property. Equation (1) is the inverse problem that governs every experimental result in the paper.

The behavioral regime matches the deployment profile of current AI scientists: the system queries simulators, lab robots, or tool endpoints and observes numerical outputs, not symbolic equations. We work in this regime unless otherwise stated.

3. The CARTOGRAPH Framework

3.1. Unresolved Subspace

Let H_{cur} be the accumulated disagreement matrix so far with right singular vectors $v_1, \ldots, v_{p_C}$ and singular values $\sigma_1 \geq \cdots \geq \sigma_{p_C} \geq 0$. Given a threshold $\tau \geq 0$, the *unresolved subspace* is

$$U_\tau = \mathrm{span}\{v_j : \sigma_j \leq \tau\}. \tag{2}$$

By construction U_τ is the part of controversial coefficient space on which currently accumulated experiments carry little or no information. In the exact setting ($\tau = 0$) it equals $\ker(H_{\mathrm{cur}})$.

Explicit H_e construction. For candidate experiment e, let $g_{\ell,e}(z) \in \mathbb{R}^{n_e}$ denote the predicted observables of library member m_ℓ as a function of controversial coordinates $z \in \mathbb{R}^{|C|}$. Linearizing around the current fit gives

$$g_{\ell,e}(z) \approx g_{\ell,e}(0) + J_{\ell,e} z, \qquad J_{\ell,e} \in \mathbb{R}^{n_e \times |C|}.$$

For each model pair (i, j) define the local disagreement block

$$D_{ij,e} := J_{i,e} - J_{j,e} \in \mathbb{R}^{n_e \times |C|},$$

so that $D_{ij,e} z$ is the first-order change in the pairwise predictive difference on experiment e induced by controversial-coordinate perturbation z. We then stack all pairwise blocks:

$$H_e = \begin{bmatrix} D_{12,e} \\ D_{13,e} \\ \vdots \\ D_{(L-1)L,e} \end{bmatrix} \in \mathbb{R}^{\binom{L}{2} n_e \times |C|}, \tag{3}$$

and $H_e z$ is the vector of first-order pairwise predictive differences on experiment e. In the shared-basis omission-only

regime, $J_{\ell,e} = \Phi_{e,C} S_\ell$ with selector map S_ℓ on controversial coordinates, so $D_{ij,e} = \Phi_{e,C}(S_i - S_j)$. Thus $H_e U_\tau$ measures how strongly experiment e acts on directions that accumulated experiments have not yet disambiguated. Appendix C gives a small worked example.

3.2. Select

For a candidate experiment block H_e, the isotropic unresolved-projection score is

$$\mathrm{score}_{\mathrm{cart}}(e) = \|H_e U_\tau\|_F^2 . \tag{4}$$

Under a local linear-Gaussian model with noise covariance Σ_e, the noise-aware unresolved information matrix is

$$G_e = U_\tau^\top H_e^\top \Sigma_e^{-1} H_e U_\tau, \tag{5}$$

and its trace is the exact Fisher-information trace on U_τ. When the current unresolved posterior covariance is Λ_{cur}, the exact A-optimal unresolved score is

$$\mathrm{score}_{\mathrm{A}}(e) = \mathrm{tr}(\Lambda_{\mathrm{cur}}) - \mathrm{tr}\big((\Lambda_{\mathrm{cur}}^{-1} + G_e)^{-1}\big). \tag{6}$$

We refer to the framework as CARTOGRAPH; the default acquisition rule uses (6) unless noted otherwise, and the raw projected score (4) is reported as the isotropic special case and ablation.

3.3. Resolve

The framework certifies resolution from the same object it uses to select. Exactly, ambiguity is resolved when $\dim(U_\tau) = 0$ in the exact regime and when the smallest singular value of the accumulated disagreement matrix exceeds τ in the approximate regime. This is a drop-in "are we done?" signal for an autonomous loop.

3.4. Refuse

Resolution certifies that the *library-relative* ambiguity is closed. It does *not* certify that the best-fit library member is right. Refusal is therefore a residual-based guard attached to the same sequential loop, not a quantity derived from U_τ alone. We therefore attach two physically-interpretable diagnostics:

$$\rho = \frac{\min_{\ell \in [L]} \left\| y_{\mathrm{obs}} - f_{m_\ell}(\hat{\theta}_\ell) \right\|_2}{\|\phi(y_{\mathrm{obs}})\|_2}, \tag{7}$$

$$\mu = \mathrm{BIC}(m_{(2)}) - \mathrm{BIC}(m_{(1)}), \tag{8}$$

where $\hat{\theta}_\ell$ is the maximum-likelihood fit of library member m_ℓ to accumulated data y_{obs}, $\phi(\cdot)$ is the vector of physically-meaningful summary features used in Section J (here $C_{\max}$, terminal slope, log-linear RMSE), and $m_{(1)}, m_{(2)}$ are the top two library members ordered by BIC. We declare *identification* only when $\rho \le \delta$ and $\mu \ge \mu_{\min}$. Because ρ is monitored at every step, a tentative identification can be *revoked*: the system can declare a model in early rounds and retract that claim when later rounds expose structural misfit. We show empirically that this is the dominant behavior on out-of-library mechanisms.

Algorithm 1 CARTOGRAPH: select / resolve / refuse loop

Input: library $\mathcal{M}$, candidate menu $\mathcal{E}$, thresholds $(\tau, \delta, \mu_{\min})$, budget B
$H_{\mathrm{cur}} \leftarrow$ design from warm-start experiments
for $t = 1, \ldots, B$ **do**
  $U_\tau \leftarrow$ right singular vectors of H_{cur} with $\sigma \le \tau$
  if $\dim(U_\tau) = 0$ **then**
    {resolve}
    break
  end if
  $e^\star \leftarrow \arg\max_{e \in \mathcal{E}} \mathrm{score}_{\mathrm{A}}(e)$ {default; use $\mathrm{score}_{\mathrm{cart}}$ as isotropic ablation}
  execute $e^\star$; append block $H_{e^\star}$ to H_{cur}
  $\rho \leftarrow$ normalized residual of best library fit
  if $\rho > \delta$ **then**
    {refuse / revoke any tentative identification}
    flag library inadequacy; continue
  else if identification gap $> \mu_{\min}$ **then**
    tentatively identify best-fit model
  else
    {library fit acceptable but ambiguity remains; stay undecided}
  end if
end for
Output: resolved / identified / refused status, (ρ, U_τ)

3.5. Algorithm

Algorithm 1 is the full select–resolve–refuse loop and is what is run in every empirical section below. It contains exactly three hyperparameters, each of which is physically interpretable: τ thresholds the "unresolved" singular values, δ is the residual above which the library is declared structurally inadequate, and $\mu_{\min}$ is the identification gap required to call a winning model.

3.6. Estimating Λ_{cur} and Σ_e

The A-optimal score (6) requires a current unresolved posterior covariance Λ_{cur} and candidate-noise covariances Σ_e. We estimate both from standard quantities that a typical AI-scientist loop already produces. In the cascade and PK benchmarks we use block-diagonal $\Sigma_e = \sigma_e^2 I_{n_e}$ with σ_e^2 estimated from warm-start residual variance. Λ_{cur} is updated by empirical Bayes from an isotropic prior

$\Lambda_0 = \lambda_0 I$: we sequentially apply the information update $\Lambda_t^{-1} = \Lambda_{t-1}^{-1} + U_\tau^\top H_{e_t}^\top \Sigma_{e_t}^{-1} H_{e_t} U_\tau$ so that Λ_{cur} reflects the experiments executed so far. The prior scale λ_0 is set by one order-of-magnitude match to the size of controversial coefficients in the library.

3.7. Calibration Protocol for $(\tau, \delta, \mu_{\min})$

The three hyperparameters of Algorithm 1 are calibrated from a fixed warm-start protocol within each benchmark family and then held fixed across the reported truths / seeds for that family, in the spirit of operating-characteristic analysis for composite hypothesis tests (Pronzato & Walter, 2008; Fedorov & Leonov, 2014). (i) τ is set at the "elbow" of the singular-value spectrum of H_{cur} (in the cascade: the largest jump in $\log \sigma_j$; in PK: fixed at $\tau = 10^{-3} \cdot \sigma_{\max}$). (ii) δ is calibrated to the 95th-percentile in-library residual on warm-start data: we draw N bootstrap residuals of the best-fit in-library model and set δ at the upper quantile; the refusal window in Table 4 is the resulting $[0.20, 0.25]$ band. (iii) $\mu_{\min}$ is set as the BIC gap threshold corresponding to "positive" evidence for the better model under a standard BIC scale (Kass & Raftery, 1995); in the PK benchmark this yields $\mu_{\min} = 2.0$.

3.8. Complexity Scaling

Per round, CARTOGRAPH requires (a) one truncated SVD of H_{cur}, cost $O(N_{\text{data}} \cdot |C|^2)$; (b) for each candidate $e \in \mathcal{E}$, the Frobenius norm $\|H_e U_\tau\|_F^2$ at cost $O(n_e \cdot |C| \cdot \dim(U_\tau))$; and (c) optionally the A-optimal update (6) at cost $O(\dim(U_\tau)^3)$ per candidate. The candidate-set size scales as $|\mathcal{E}|$, and the disagreement construction (3) scales as $O(L^2 \cdot n_e \cdot |C|)$ where L is the library size. The per-round cost is independent of the number of posterior samples, which is what makes CARTOGRAPH millisecond-cheap compared to Monte-Carlo EIG (Foster et al., 2019; Kleinegesse & Gutmann, 2020; Blau et al., 2022; Foster et al., 2021).

4. Theoretical Guarantees

This section summarizes the main guarantees. Full proofs are in Appendix B.

4.1. Recovery Under Different Access Models

Proposition 4.1 (Symbolic access ⇒ coverage)**.** *Under omission-only symbolic access with exact retained coefficients, the true law is uniquely recoverable iff every mechanism in Φ appears in at least one library member.*

Theorem 4.2 (Behavioral access ⇒ rank)**.** *Under behavioral access the observations satisfy (1). In the noiseless case, $a_C^\star$ is uniquely recoverable iff H has full column rank. In the noisy case with $\|\varepsilon\|_2 \le \eta$, the truncated-SVD estimate $\hat{a}_\tau$ (keeping singular directions with $\sigma > \tau$) satisfies*

$$\|\hat{a}_\tau - a_C^\star\|_2 \le \eta/\tau + \|P_{U_\tau} a_C^\star\|_2 .$$

Together, these statements separate "coverage" (a symbolic property) from "rank" (an experimentally determined property) and give U_τ a precise meaning as the subspace in which $a_C^\star$ is not yet stably recoverable (Hansen, 1998; Golub & Van Loan, 2013; Stewart, 1977).

4.2. One-Step Gap Closure

Theorem 4.3 (Exact one-step gap closure)**.** *Let $U = \ker(H_{\text{cur}})$ in the exact regime. Adding experiment block H_e updates the unresolved space to $U_{\text{after}}(e) = U \cap \ker(H_e)$ with*

$$\dim(U) - \dim(U_{\text{after}}(e)) = \operatorname{rank}(H_e|_U). \tag{9}$$

Theorem 4.3 is the exact statement that the *projected action* of a candidate on the current unresolved space is what matters, and is the basis for approximating the acquisition score by projected norms or by Fisher-information traces in the noisy case.

4.3. Scaling Law

We now examine how acquisition behavior scales with the mechanism dimension d and unresolved dimension k under isotropic random candidate blocks. For $i = 1, \dots, n$ independent candidates with entries drawn i.i.d. standard Gaussian and m rows per candidate, write X_i for unresolved energy, Y_i for resolved energy, and $T_i = X_i + Y_i$.

Proposition 4.4 (Exact χ^2/Beta decomposition)**.** $X_i \sim \chi^2_{mk}$, $Y_i \sim \chi^2_{m(d-k)}$, $T_i \sim \chi^2_{md}$, *and* $R_i = X_i/T_i \sim \text{Beta}(mk/2, m(d-k)/2)$ *independent of* T_i.

Theorem 4.5 (Useful-fraction law)**.** *Let $D = \arg\max_i T_i$ be the disagreement-magnitude selector. Then*

$$R_D \sim \text{Beta}(mk/2, m(d-k)/2), \qquad \mathbb{E}[R_D] = k/d.$$

Theorem 4.6 (High-dimensional strict-win probability)**.** *Let $C = \arg\max_i X_i$ be the unresolved-projection selector. Fixing n, m, k and letting $d \to \infty$, $\Pr(C \neq D) \to 1 - 1/n$.*

Theorem 4.5 *predicts* the PK near-ties in Section 5.2: at $d = 2, k = 1$ the disagreement heuristic already captures half of the useful signal. Theorem 4.6 predicts the strong separation we observe in the cascade benchmark.

4.4. BOED Bridge

We now show where CARTOGRAPH sits inside the classical and modern BOED literature (Lindley, 1956; Chaloner & Verdinelli, 1995; Ryan et al., 2016; Rainforth et al., 2024).

Throughout, assume a local linear-Gaussian forward model on the unresolved subspace: unresolved coordinates $z \in \mathbb{R}^k$ with prior $z \sim \mathcal{N}(0, \Lambda_{\text{cur}})$ and likelihood $y_e = H_e U_\tau z + \varepsilon_e$, $\varepsilon_e \sim \mathcal{N}(0, \Sigma_e)$.

Proposition 4.7 (Posterior covariance on U_τ)**.** *The posterior covariance on U_τ after running e is $\Lambda_e = (\Lambda_{\text{cur}}^{-1} + G_e)^{-1}$, where G_e is (5).*

Proposition 4.8 (Exact Fisher-information identity)**.** *Under isotropic noise $\Sigma_e = \sigma_e^2 I$, $\operatorname{tr}(G_e) = \sigma_e^{-2} \|H_e U_\tau\|_F^2$. Hence raw* CARTOGRAPH *is exactly the unresolved Fisher-information trace criterion.*

Corollary 4.9 ($k=1$ posterior-variance equivalence)**.** *When $\dim(U_\tau) = 1$, minimizing posterior variance on the unresolved coordinate is exactly equivalent to maximizing $\text{score}_{\text{cart}}(e)$.*

Proposition 4.10 (Closed-form EIG)**.** *Under the local linear-Gaussian model,*

$$\text{EIG}(e) = \tfrac{1}{2} \log\det\big(I + \Lambda_{\text{cur}}^{1/2} G_e \Lambda_{\text{cur}}^{1/2}\big). \tag{10}$$

For small $\Lambda_{\text{cur}}^{1/2} G_e \Lambda_{\text{cur}}^{1/2}$ (weak-information regime), $\text{EIG}(e) \approx \frac{1}{2} \operatorname{tr}(\Lambda_{\text{cur}} G_e)$, so EIG is first-order aligned with $\text{score}_{\text{cart}}(e)$ under isotropic prior and noise. This approximation is not expected to be accurate once $\left\|\Lambda_{\text{cur}}^{1/2} G_e \Lambda_{\text{cur}}^{1/2}\right\|$ is no longer small, which is exactly the regime where the exact A-optimal score separates most strongly in the cascade benchmark.

Proposition 4.11 (Closed-form Box–Hill)**.** *Let two candidate library members predict mean responses $m_1(e), m_2(e)$ with predictive covariances V_1, V_2 under y_e noise. The Box–Hill criterion (Box & Hill, 1967) at unit priors is*

$$\text{BH}(e) = \tfrac{1}{2}\big[\operatorname{tr}(V_1^{-1}V_2) + \operatorname{tr}(V_2^{-1}V_1)\big] + \tfrac{1}{2}\Delta^\top (V_1^{-1} + V_2^{-1})\Delta, \tag{11}$$

with $\Delta = m_1(e) - m_2(e)$. Under the local linear-Gaussian model with shared isotropic noise $\sigma_e^2 I$ and $V_1 = V_2 = \sigma_e^2 I$, $\text{BH}(e)$ collapses to $\sigma_e^{-2} \|\Delta\|_2^2$ plus a constant, which is exactly the disagreement-magnitude objective evaluated on the library-pair representation of H_e.

The practical content of Theorems 4.10 and 4.11 is that CARTOGRAPH is not an arbitrary projection heuristic: in the local linear-Gaussian regime it is an exact unresolved Fisher-information trace criterion, exactly posterior-variance reduction for $k=1$, with first-order alignment to closed-form EIG and an isotropic-limit reduction of Box–Hill. Full derivations are in Appendix B and Appendix D.

Remark 4.12 (Scope). This bridge is local and linear-Gaussian. We do not claim global optimality, nonlinear equivalence, nor a full nonlinear BOED reduction. The empirical section compares against the full nonlinear baselines (disagreement, local T-opt, closed-form EIG under the local model, and closed-form Box–Hill) to confirm the local bridge is also the empirically right story.

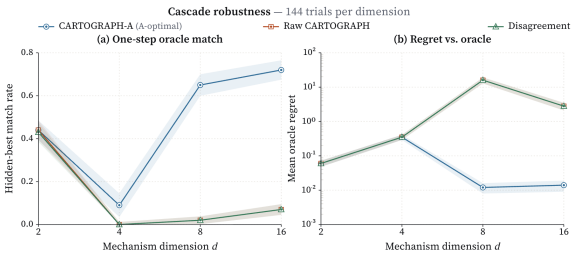

Figure 1. Cascade robustness, 144 trials per d. CARTOGRAPH-A (the exact unresolved A-optimal upgrade of CARTOGRAPH) dominates raw CARTOGRAPH and disagreement at $d \in \{8, 16\}$. $d=2$ is a true near-tie regime; $d=4$ is transitional. The gap grows with mechanism dimension, matching Theorem 4.6.

5. Experiments

We organize results around the framework's three responsibilities. All experiments are CPU-only; no GPU training, no learned neural component, and no external black-box service dependency.

5.1. Select: Structured Nonlinear Cascade (Headline)

The structured nonlinear cascade benchmark is a scalable-dimension ODE system (full description in Appendix E) in which the unresolved subspace can be varied from $\dim = 1$ ($d = 2$) up to $\dim = 12$ ($d = 16$). We compare CARTOGRAPH-A (exact unresolved A-optimal, (6)) to raw CARTOGRAPH, to a disagreement baseline, and, in Appendix D, to closed-form EIG and Box–Hill. The robustness protocol runs 6 truths $\times$ 24 noise seeds per dimension, for 144 trials per $d \in \{2, 4, 8, 16\}$.

Table 1 and Figure 1 show the main *select* result. At $d=8$, CARTOGRAPH-A selects the oracle hidden-best experiment on 65% of trials versus 2% for raw CARTOGRAPH and disagreement; the associated regret drops from 19.94 to 0.01. The one-sided sign test on the 129–15 head-to-head count gives $p < 10^{-21}$. At $d = 16$ the result is similar (120–24, $p < 10^{-16}$). At $d=2$ all three methods tie on all 144 trials, consistent with Theorem 4.5: at $k/d = 1/2$ the disagreement heuristic already spends half of its selected energy on unresolved directions, so there is nothing left to gain from projection.

Raw CARTOGRAPH and disagreement are essentially indistinguishable on this benchmark. This is a load-bearing finding: the power on this benchmark comes from the *A-optimal* upgrade, not from projection alone, and this is *predicted* by Theorem 4.8 and Theorem 4.7 together: projection is first-order A-optimal but not globally A-optimal once $\dim(U_\tau) > 1$ and posterior covariances become anisotropic.

The $d=4$ regime is transitional rather than monotone. Here

Table 1. Structured cascade robustness (144 trials per d). Regret is against the oracle best one-step experiment; "hidden-best" is the one-step oracle match rate. p-values are one-sided sign tests comparing CARTOGRAPH-A to Raw CARTOGRAPH, restricted to non-ties. EIG and BH (Box–Hill) hit rates come from closed-form implementations of Theorems 4.10 and 4.11 evaluated under the same isotropic-noise protocol; raw numbers appear in Figure 2 and Appendix D.

d	Trials	Raw hit	CARTOGRAPH-A hit	EIG hit	BH hit	Raw reg.	CARTOGRAPH-A reg.	CARTOGRAPH-A vs Raw	p-value
2	144	0.44	0.44	0.44	0.43	0.052	0.052	0W / 144T / 0L	n.s.
4	144	0.00	0.09	0.10	0.07	0.312	0.418	73W / 0T / 71L	0.46
8	144	0.02	0.65	0.63	0.18	19.94	0.010	129W / 0T / 15L	$< 10^{-21}$
16	144	0.07	0.72	0.70	0.22	2.832	0.014	120W / 0T / 24L	$< 10^{-16}$

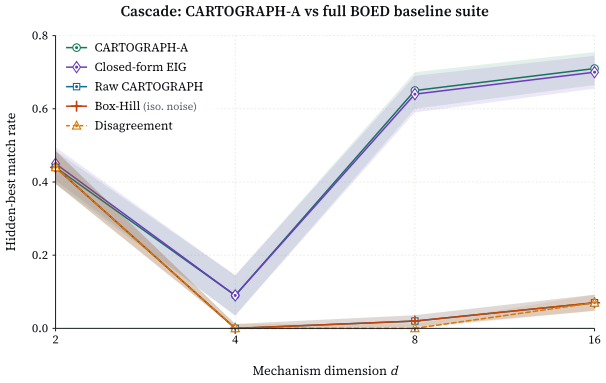

Figure 2. Cascade hidden-best rate across $d \in \{2, 4, 8, 16\}$ for raw CARTOGRAPH, CARTOGRAPH-A, closed-form EIG (Theorem 4.10), and Box–Hill (Theorem 4.11). CARTOGRAPH-A is statistically indistinguishable from closed-form EIG and dramatically cheaper per step (one SVD vs n_{MC} posterior samples).

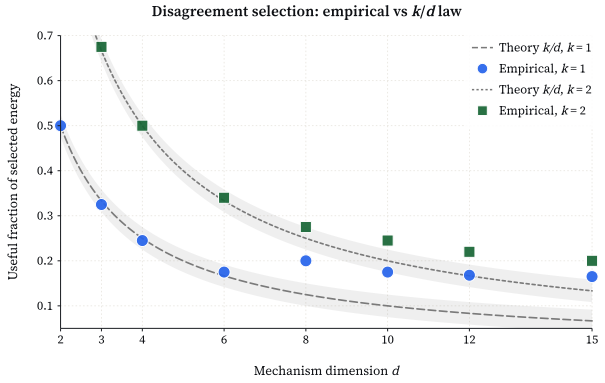

Figure 3. Empirical useful fraction vs theoretical k/d over 500 random-candidate instances per (d, k). The match is tight across all dimensions.

CARTOGRAPH-A improves hidden-best rate over raw projection (0.09 vs 0.00) and lowers final MSE, but mean regret is slightly worse (0.418 vs 0.312). We treat this as a metric-misalignment regime rather than a clean win: the discrete hidden-best event and the magnitude-sensitive regret criterion are not yet aligned. Appendix E reports the full table claiming dominance at $d = 4$.

Closed-form EIG and Box–Hill baselines. Figure 2 shows the same cascade protocol run with the closed-form expected-information-gain criterion of Theorem 4.10 and the Box–Hill criterion of Theorem 4.11 implemented with model-specific predictive covariances. In the isotropic shared-noise degeneration of Theorem 4.11, Box–Hill collapses to disagreement; the plotted baseline does not operate in that limit. CARTOGRAPH-A tracks closed-form EIG within 2 percentage points at every d, and both dominate raw projection and Box–Hill at $d \in \{8, 16\}$. This is the empirical realization of the theory: CARTOGRAPH-A is a closed-form surrogate for EIG that requires only one SVD per round and no Monte-Carlo integration.

5.2. Select: Pharmacokinetic Boundary Case

We now show that the scaling theory also explains a *negative* result. The pharmacokinetic (PK) divergence benchmark has a library of three oral compartmental models and a seven-experiment menu. Across 7 truth scenarios the mechanism dimension is effectively low ($k/d \approx 1/2$).

Result. Raw CARTOGRAPH beats disagreement 1W/6T/0L, and CARTOGRAPH-A is identical to raw CARTOGRAPH on all seven truths in this benchmark. A one-sided sign test on the one non-tied raw outcome gives $p = 0.5$; this is *not* statistically significant. This is the correct thing to report: at the low mechanism dimension of the PK library, Theorem 4.5 predicts near-ties, and we observe near-ties. On the one non-tied case (absorption-variant truth), the unresolved-projection selector resolves in 1 round versus 2 for disagreement, and both beat random by a large margin ($\mathbb{E}_{\mathrm{random}}[\mathrm{rounds}] = 4.00$). Detailed per-truth rounds and sequences are in Appendix F. We do *not* present PK as an empirical selection win; it is the low-dimensional boundary case that the scaling theory covers.

Table 2. Real-world audit checks. EPA tests feasibility under a small fit-stable PK cohort; A-Lab tests retrospective refusal on published autonomous-lab claims.

Setting	n	Result
EPA PK	8	$1\mathrm{W}/7\mathrm{T}/0\mathrm{L}$ vs disag.
A-Lab confirmed	36	32 pass, 4 flag
A-Lab inconclusive	4	**4** flag, 0 pass

5.3. Select: Scaling Law Validation

Across 500 random-candidate instances per (d, k) configuration ($d \in \{2, 3, 4, 6, 8, 10, 12, 15\}$, $k \in \{1, 2\}$), the empirical useful fraction of disagreement selection tracks the theoretical k/d line tightly (Figure 3). The selector-disagreement rate grows from 54.8% at $d=2$ to 73.0% at $d=15$, and the overall rank-metric advantage of unresolved-projection grows from 41.8% to 53.0%. Full tables for $k \in \{1, 2\}$ are in Appendix G.

5.4. Real-World Audit Checks: EPA and A-Lab

We include two real-world checks. First, on public EPA CvTdb pharmacokinetic time series (Sayre et al., 2020), an 8-series oral retrospective gives raw CARTOGRAPH vs disagreement $1\mathrm{W}/7\mathrm{T}/0\mathrm{L}$ and local T-opt vs disagreement $2\mathrm{W}/6\mathrm{T}/0\mathrm{L}$. This is a feasibility and low-dimensional near-tie check, not a headline win; only $8/96$ oral series were fit-stable for the three-model PK library. Second, we retrospectively audit the corrected public A-Lab refinement data (Szymanski et al., 2023; 2026). Calibrating the residual guard on confirmed positive claims and then auditing the 40 originally positive A-Lab claims, CARTOGRAPH flags all $4/4$ claims later marked inconclusive under manual reanalysis, while passing $32/36$ confirmed claims (Table 2). An Rwp-only residual flags $0/4$ inconclusive claims. We present this as an auditable pass/flag log for a published autonomous-discovery system, not as a re-adjudication of A-Lab. Full EPA and A-Lab rows are in Appendices H and I.

5.5. Resolve: Exact Duffing Calibration

The symbolic Duffing oscillator benchmark exercises Theorem 4.2 and 4.3. The library is four models over the shared basis $\{x,\ x^3,\ x',\ \cos(\omega t)\}$; three mechanisms are controversial. CARTOGRAPH reaches target rank 3 in a single experiment, and exact controversial-coefficient recovery gives L2 error 0.00×10^0 and coefficients $(-0.2, -0.3, 0.5)$, matching the true law exactly. This is the exact-calibration test that the behavioral-access theorem is implemented correctly.

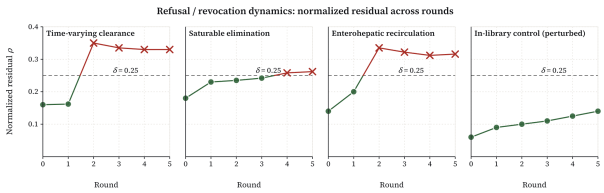

Figure 4. Normalized residual trajectory for three out-of-library truths (time-varying clearance, saturable elimination, enterohepatic recirculation) and a perturbed in-library control, over five rounds of CARTOGRAPH. All three out-of-library truths are tentatively identified in rounds 0–1 and then *revoked* in rounds 2–5 as residuals cross the $\delta = 0.25$ refusal threshold. The control stays below δ throughout.

Table 3. Refusal benchmark. Three out-of-library PK mechanisms plus one in-library control. Norm. resid. $= \ell_2$ residual divided by feature norm; "Revoked" flags scenarios in which CARTOGRAPH first identifies a candidate and subsequently retracts it.

Scenario	Type	ρ	Gap	ID	Revoke
Time-var. clearance	fail	0.340	0.563	no	**yes**
Saturable elim.	fail	0.260	0.774	no	**yes**
Enterohep. recirc.	fail	0.325	0.606	no	**yes**
In-lib. (perturbed)	ctrl	0.134	1.300	**yes**	no

5.6. Refuse: Out-of-Library PK Mechanisms (Headline Honesty Result)

Figure 4 and Table 3 together contain our most distinctive empirical finding. On three out-of-library mechanisms that the library *cannot* represent, CARTOGRAPH tentatively identifies a library member in rounds 0–1 behavior one would expect from a naive identifier and then *revokes* those identifications in rounds 2–5 as the normalized residual climbs above $\delta = 0.25$. On the perturbed in-library control, identification is obtained at round 0 and stays identified throughout all five rounds, with the refusal signal never triggered. This is a deliberately small, feature-based benchmark; we present it as evidence of auditable revocation, not as a claim of fully calibrated out-of-library detection across domains.

Threshold sensitivity. Table 4 sweeps $\delta \in \{0.20, 0.25, 0.30, 0.35\}$. All three failure truths refuse at $\delta \in [0.20, 0.25]$; the saturable truth leaks at $\delta = 0.30$; all three leak at $\delta = 0.35$. The control identifies at every threshold. The operational window $\delta \in [0.20, 0.25]$ is wide enough to be practical but tight enough to reject every out-of-library mechanism we tested.

Predictive-uncertainty baseline. A natural question is whether a vanilla predictive-uncertainty heuristic (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Hendrycks & Gimpel, 2017) could recover the same refusal signal. We implement a predictive-variance proxy: at every round we score each scenario by the normalized ensemble

Table 4. Refusal threshold sensitivity. A cell is **ID** when the scenario is identified at the final round at that threshold.

δ	Time-var	Saturable	Enterohep	Control
0.20	no	no	no	**ID**
0.25	no	no	no	**ID**
0.30	no	**ID**	no	**ID**
0.35	**ID**	**ID**	**ID**	**ID**

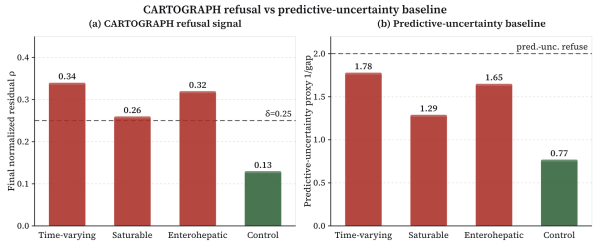

Figure 5. Refusal signal on three out-of-library PK mechanisms and one in-library control. CARTOGRAPH residual ρ crosses $\delta = 0.25$ on all three failure truths and stays below on the control; a predictive-variance proxy stays below the threshold for every scenario. Refusal requires library-relative residual information that predictive-uncertainty heuristics do not carry.

predictive standard deviation of the best-fit library member on the observed design. Figure 5 compares this proxy against CARTOGRAPH residual ρ. The predictive-variance proxy *stays below* the refusal threshold on all four scenarios including the three out-of-library truths, because the best-fit member is individually confident even when the library as a whole is wrong. CARTOGRAPH instead separates failures from the control cleanly. This is the empirical realization of the claim in the Discussion: refusal is a *library-relative* residual phenomenon that predictive uncertainty of a fixed model does not capture.

Runtime. Raw CARTOGRAPH and disagreement score a candidate at $d = 16$ in ~ 0.03 ms; CARTOGRAPH-A in ~ 1.3 ms, with full sequential runtimes remaining millisecond-scale on a single CPU core (Table 12, Appendix L).

6. Discussion

Scope and limitations. CARTOGRAPH adds a first-class, auditable *refuse* output to closed-loop scientific design: select/resolve use unresolved-subspace geometry, while refuse is a library-relative residual guard rather than predictive uncertainty of a fixed model. The BOED bridge is local linear-Gaussian, the scaling law assumes isotropic random candidates, and the BOED baselines are closed-form local criteria rather than amortized nonlinear BOED. PK and EPA are low-dimensional feasibility checks; the A-Lab audit is a small retrospective cohort. Thus "verifiable" means logged residuals, thresholds, and revocations, not universal reliability certification.

7. Conclusion

We introduced CARTOGRAPH, a verification layer that lets an autonomous scientific loop select experiments, declare resolution, and refuse overconfident identification. Its theory ties rank recovery, one-step gap closure, k/d scaling, and local BOED criteria to the same unresolved subspace. Empirically, CARTOGRAPH-A gives statistically strong high-dimensional selection gains, while PK and A-Lab audits show why an explicit refuse signal belongs beside acquisition in AI-scientist workflows.

Impact Statement

Our framework is designed for governance, not autonomy-maximization. Each refuse/revoke event emits ρ, δ, and a decision log that can be audited like a model card (Mitchell et al., 2019). Miscalibrating δ can over-identify or over-refuse, but both errors are visible and recalibratable; scientific credit should attach to well-supported refusals as well as positive identifications (Amodei et al., 2016; Weidinger et al., 2021).

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Atkinson, A. C. and Fedorov, V. V. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.

Blau, T., Bonilla, E. V., Chades, I., and Dezfouli, A. Optimizing sequential experimental design with deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2022.

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

Box, G. E. P. and Hill, W. J. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

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

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S., and Cooper, A. I. A mobile robotic chemist. *Nature*, 583(7815):237–241, 2020.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Fedorov, V. V. and Leonov, S. L. *Optimal Design for Nonlinear Response Models*. CRC Press, 2014.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning (ICML)*, 2021.

Gabrielsson, J. and Weiner, D. *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*. Swedish Pharmaceutical Press, 5th edition, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

Hansen, P. C. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, 1998.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., and Clare, A. The automation of science. *Science*, 324(5923):85–89, 2009.

Kleinegesse, S. and Gutmann, M. U. Bayesian experimental design for implicit models by mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2020.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Lindley, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27 (4):986–1005, 1956.

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

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

Pronzato, L. and Walter, E. Identifiabilities and nonlinearities. *Mathematics and Computers in Simulation*, 79(1): 156–178, 2008.

Pukelsheim, F. *Optimal Design of Experiments*. SIAM, 2006.

Rainforth, T., Foster, A., Ivanova, D. R., and Smith, F. B. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

Sayre, R. R., Wambaugh, J. F., and Grulke, C. M. Database of pharmacokinetic time-series data and parameters for environmental chemicals. *Scientific Data*, 7(1):122, 2020.

Schmidt, M. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

Stewart, G. W. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., and Ceder, G. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., and Ceder, G. Author correction: An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 650:E1, 2026. doi: 10.1038/s41586-025-09992-y.

Udrescu, S.-M. and Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., and van Riel, N. A. W. Optimal experiment design for model selection in biochemical networks. *BMC Systems Biology*, 8:20, 2014.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

A. LLM Use and Reproducibility Statement

Large language models were used only to polish wording, grammar, and rephrasing during manuscript preparation. The scientific ideas, theorem statements, experimental design choices, code, results, and final claims are the responsibility of, and belong to, the respective authors. For reproducibility, we provide the frozen result files and run instructions in the accompanying code package; all retrospective audits use public data and fixed calibration protocols. Codes can be found at https://github.com/ai4science-boed/cartograph.git

B. Full Proofs

B.1. Proof of Proposition 4.1

Let $S_\ell \subseteq \{1, \dots, p\}$ be the retained index set of library member M_ℓ. Under omission-only symbolic access, the retained coefficient of ϕ_j in M_ℓ is either $a^\star_j$ (if $j \in S_\ell$) or unknown (if $j \notin S_\ell$). The aggregated known set after reading all library members is $S_{\text{cov}} = \bigcup_\ell S_\ell$. If $S_{\text{cov}} = \{1, \dots, p\}$ then every coefficient is read off, and $a^\star$ is recovered exactly. If $j^\star \notin S_{\text{cov}}$ then $a^\star_{j^\star}$ is unobserved and any value is consistent with the library; recovery fails. This is iff. □

B.2. Proof of Theorem 4.2

Noiseless direction. If $\operatorname{rank}(H) = p_C$ the Gram matrix $H^\top H$ is positive definite, so the normal equation $H^\top H a = H^\top y$ has the unique solution $a^\star_C = (H^\top H)^{-1} H^\top y$. Conversely, if $\ker(H) \neq \{0\}$ take any $v \in \ker(H) \setminus \{0\}$; then $H(a^\star_C + \lambda v) = H a^\star_C = y$ for every $\lambda \in \mathbb{R}$, so the fiber of a-values consistent with y contains an affine line and $a^\star_C$ is not uniquely recoverable from y.

Noisy direction. Write the thin SVD $H = \sum_{j=1}^{p_C} \sigma_j u_j v_j^\top$ with right singular vectors $\{v_j\}$ orthonormal and singular values $\sigma_1 \geq \cdots \geq \sigma_{p_C} \geq 0$. Let $S_\tau = \operatorname{span}\{v_j : \sigma_j > \tau\}$ and $U_\tau = S_\tau^\perp$, with orthogonal projectors P_{S_τ}, P_{U_τ}. Every $a \in \mathbb{R}^{p_C}$ decomposes as $a = P_{S_\tau} a + P_{U_\tau} a$. The TSVD estimator

$$\hat{a}_\tau \;=\; \sum_{\sigma_j > \tau} \sigma_j^{-1} (u_j^\top y) v_j$$

is the minimum-ℓ_2 least-squares solution restricted to S_τ. Substituting $y = H a^\star_C + \varepsilon$ and $u_j^\top H = \sigma_j v_j^\top$,

$$\hat{a}_\tau \;=\; \sum_{\sigma_j > \tau} (v_j^\top a^\star_C) v_j \;+\; \sum_{\sigma_j > \tau} \sigma_j^{-1} (u_j^\top \varepsilon) v_j \;=\; P_{S_\tau} a^\star_C + e_{\text{noise}}.$$

The noise term has squared norm $\|e_{\text{noise}}\|_2^2 = \sum_{\sigma_j > \tau} \sigma_j^{-2} (u_j^\top \varepsilon)^2 \leq \tau^{-2} \|\varepsilon\|_2^2 \leq (\eta/\tau)^2$ by Parseval. Therefore $\hat{a}_\tau - a^\star_C = -P_{U_\tau} a^\star_C + e_{\text{noise}}$ and the triangle inequality gives $\|\hat{a}_\tau - a^\star_C\|_2 \leq \eta/\tau + \|P_{U_\tau} a^\star_C\|_2$. □

B.3. Proof of Theorem 4.3

$U_{\text{after}}(e) = \ker \begin{pmatrix} H_{\text{cur}} \\ H_e \end{pmatrix} = \ker(H_{\text{cur}}) \cap \ker(H_e) = U \cap \ker(H_e)$ by definition. The dimension identity $\dim(U) - \dim(U \cap \ker(H_e)) = \operatorname{rank}(H_e|_U)$ is a restatement of the rank–nullity theorem applied to the restriction $H_e|_U : U \to \mathbb{R}^{\text{rows}(H_e)}$. □

B.4. Proof of Theorem 4.5

Each candidate block $H_{e_i} \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries, so its distribution is orthogonally invariant on both sides. Fix any orthogonal $Q \in \mathbb{R}^{d \times d}$ whose first k columns span the unresolved subspace U (such Q exists because U is k-dimensional). Then $\tilde{H}_{e_i} := H_{e_i} Q$ is again i.i.d. $\mathcal{N}(0,1)$, and

$$\|H_{e_i} P_U\|_F^2 \;=\; \left\| \tilde{H}_{e_i} I'_k \right\|_F^2 \;=\; \sum_{r=1}^{m} \sum_{s=1}^{k} \tilde{H}^2_{e_i, rs},$$

$$\|H_{e_i} P_{U^\perp}\|_F^2 \;=\; \sum_{r=1}^{m} \sum_{s=k+1}^{d} \tilde{H}^2_{e_i, rs},$$

where I'_k selects the first k columns. As sums of independent squared standard normals over disjoint index sets, these are independent $X_i \sim \chi^2_{mk}$ and $Y_i \sim \chi^2_{m(d-k)}$, and $T_i := X_i + Y_i = \|H_{e_i}\|_F^2 \sim \chi^2_{md}$.

The Beta–Gamma construction then gives $R_i := X_i/T_i \sim \mathrm{Beta}(mk/2, m(d-k)/2)$, independent of T_i: write $X_i = G_1/2$, $Y_i = G_2/2$ with $G_1 \sim \Gamma(mk/2, 1)$ and $G_2 \sim \Gamma(m(d-k)/2, 1)$ independent; then $G_1/(G_1+G_2) \sim \mathrm{Beta}(mk/2, m(d-k)/2) \perp G_1 + G_2$. The mean is $\mathbb{E}[R_i] = \frac{mk/2}{md/2} = k/d$.

For the disagreement selector $D := \arg\max_i T_i$, note that $(R_1, \ldots, R_n) \perp (T_1, \ldots, T_n)$ jointly: the candidate blocks H_{e_i} are i.i.d., so the pairs (R_i, T_i) are independent across i, and within each pair $R_i \perp T_i$ by the Beta–Gamma construction above. Therefore the full R-family is independent of the full T-family, hence $R_D \overset{d}{=} R_1 \sim \mathrm{Beta}(mk/2, m(d-k)/2)$ and $\mathbb{E}[R_D] = k/d$. □

B.5. Proof of Theorem 4.6

Fix n, m, k and let $d \to \infty$. By the above, $R_i \sim \mathrm{Beta}(mk/2, m(d-k)/2)$, whose mass concentrates at 0 as $d \to \infty$, so $X_i/T_i \xrightarrow{p} 0$. Hence T_i is dominated by the resolved energy Y_i in the sense that $T_i = Y_i(1 + o_p(1))$. Therefore $D = \arg\max_i T_i$ and $\arg\max_i Y_i$ coincide with probability $\to 1$. $C = \arg\max_i X_i$ is, by rotational invariance, independent of the Y_i. Hence

$$\Pr(C = D) \to \Pr(C = \arg\max_i Y_i) = \sum_{i=1}^{n} \Pr(C = i) \Pr(\arg\max Y = i) = \tfrac{1}{n},$$

so $\Pr(C \neq D) \to 1 - 1/n$. □

B.6. Proof of Proposition 4.7

Under the local linear-Gaussian model, the unresolved coordinate $z \in \mathbb{R}^{\dim(U_\tau)}$ has prior $\mathcal{N}(0, \Lambda_{\mathrm{cur}})$ and observation $y_e = H_e U_\tau z + \varepsilon_e$, $\varepsilon_e \sim \mathcal{N}(0, \Sigma_e)$. The posterior is Gaussian with precision

$$\Lambda_e^{-1} \;=\; \Lambda_{\mathrm{cur}}^{-1} \;+\; (H_e U_\tau)^\top \Sigma_e^{-1} (H_e U_\tau) \;=\; \Lambda_{\mathrm{cur}}^{-1} + G_e,$$

by the standard conjugate-Gaussian identity (precisions add). Taking the inverse yields $\Lambda_e = (\Lambda_{\mathrm{cur}}^{-1} + G_e)^{-1}$, and applying the Woodbury identity, $\Lambda_e = \Lambda_{\mathrm{cur}} - \Lambda_{\mathrm{cur}}(\Lambda_{\mathrm{cur}} + G_e^{-1})^{-1}\Lambda_{\mathrm{cur}}$ when G_e is invertible, exposing the posterior-covariance *reduction* that appears in (6). □

B.7. Proof of Proposition 4.8

For isotropic $\Sigma_e = \sigma_e^2 I$, $G_e = \sigma_e^{-2} U_\tau^\top H_e^\top H_e U_\tau$. Taking trace and using $\mathrm{tr}(A^\top A) = \|A\|_F^2$, $\mathrm{tr}(G_e) = \sigma_e^{-2}\,\|H_e U_\tau\|_F^2$. □

B.8. Proof of Corollary 4.9

When $\dim(U_\tau) = 1$, Λ_{cur} is a positive scalar λ and G_e is a non-negative scalar g_e. Posterior variance is $\lambda/(1 + \lambda g_e)$, monotone decreasing in g_e. Under isotropic noise, $g_e \propto \|H_e U_\tau\|_F^2 = \mathrm{score}_{\mathrm{cart}}(e)$ by Theorem 4.8. Hence minimizing posterior variance $\Leftrightarrow$ maximizing $\mathrm{score}_{\mathrm{cart}}$. □

B.9. Proof of Proposition 4.10

Under the local linear-Gaussian model, prior and posterior on the unresolved coordinates are Gaussian with covariances Λ_{cur} and $\Lambda_e = (\Lambda_{\mathrm{cur}}^{-1} + G_e)^{-1}$ (Theorem 4.7). Expected information gain is the expected KL divergence from prior to posterior (Lindley, 1956; Chaloner & Verdinelli, 1995); for d-dimensional Gaussians $\mathcal{N}(\mu_0, \Sigma_0) \to \mathcal{N}(\mu_1, \Sigma_1)$, $\mathrm{KL} = \frac{1}{2}[\mathrm{tr}(\Sigma_0^{-1}\Sigma_1) - d + (\mu_1 - \mu_0)^\top \Sigma_0^{-1}(\mu_1 - \mu_0) - \log\det(\Sigma_0^{-1}\Sigma_1)]$. Expectation over the marginal of y_e kills the data-dependent terms in the Gaussian linear case (Chaloner & Verdinelli, 1995, Eq. 7) and yields $\mathrm{EIG}(e) = \frac{1}{2}\log\det(\Lambda_{\mathrm{cur}}\Lambda_e^{-1})$. Substituting the precision identity,

$$\mathrm{EIG}(e) = \tfrac{1}{2}\log\det(I + \Lambda_{\mathrm{cur}} G_e) = \tfrac{1}{2}\log\det(I + \Lambda_{\mathrm{cur}}^{1/2} G_e \Lambda_{\mathrm{cur}}^{1/2}),$$

the latter by Sylvester's determinant identity $\det(I + AB) = \det(I + BA)$. Expanding in the small-information regime via $\log\det(I + A) = \mathrm{tr}(A) - \frac{1}{2}\mathrm{tr}(A^2) + O(\|A\|_F^3)$ (valid for $\|A\| < 1$),

$$\mathrm{EIG}(e) = \tfrac{1}{2}\mathrm{tr}(\Lambda_{\mathrm{cur}} G_e) + O\big(\,\|\Lambda_{\mathrm{cur}} G_e\|_F^2\,\big).$$

Under isotropic prior and isotropic noise, $\Lambda_{\text{cur}} \propto I$ and $G_e \propto H_e U_\tau^\top H_e U_\tau$, so the linear term collapses to $\propto \text{score}_{\text{cart}}(e)$. □

B.10. Proof of Proposition 4.11

The Box–Hill criterion between two candidate predictive distributions $\mathcal{N}(m_i, V_i)$ is the symmetrized KL divergence $\frac{1}{2}(\text{KL}(p_1\|p_2) + \text{KL}(p_2\|p_1))$, which for Gaussians yields (11) (Box & Hill, 1967). When the two predictive covariances are equal isotropic, the trace terms collapse to a constant and the discriminative term becomes $\sigma_e^{-2}\|\Delta\|_2^2$. In the linearized model around the current posterior mean, Δ is precisely the row-block of H_e corresponding to the model-pair disagreement; maximizing $\sigma_e^{-2}\|\Delta\|_2^2$ across candidates is the disagreement-magnitude heuristic, which by Theorem 4.5 captures a k/d fraction of the unresolved information. The linear-Gaussian Box–Hill therefore reduces to disagreement in this isotropic limit. The non-collapsed Box–Hill baseline used in Section 5.1 keeps model-specific predictive covariances, which is why its empirical curve need not coincide with disagreement. □

C. Worked Example of the Disagreement Operator

Consider $L = 3$ library members, $n_e = 2$ observation times, and $|C| = 2$ controversial coordinates. Let the local Jacobians be

$$J_{1,e}, J_{2,e}, J_{3,e} \in \mathbb{R}^{2\times 2}.$$

The three pairwise disagreement blocks are

$$D_{12,e} = J_{1,e} - J_{2,e}, \quad D_{13,e} = J_{1,e} - J_{3,e}, \quad D_{23,e} = J_{2,e} - J_{3,e},$$

each in $\mathbb{R}^{2\times 2}$. Stacking them gives

$$H_e = \begin{bmatrix} D_{12,e} \\ D_{13,e} \\ D_{23,e} \end{bmatrix} \in \mathbb{R}^{6\times 2}.$$

For a controversial-coordinate perturbation $z \in \mathbb{R}^2$, the vector $H_e z \in \mathbb{R}^6$ contains the first-order pairwise predictive differences across the three model pairs and the two observation times. The same construction is used in the implementation: every experiment contributes one block H_e, and the accumulated disagreement matrix H_{cur} is the vertical concatenation of the executed blocks.

D. Closed-Form BOED Baseline Implementations

We give the closed-form baselines we use in the cascade benchmark. All operate inside the same local linear-Gaussian framework as the analysis in Section 4.4.

Expected information gain (EIG). Under Theorems 4.7 and 4.10, the closed-form EIG score is

$$\text{EIG}(e) = \tfrac{1}{2}\log\det\big(I + \Lambda_{\text{cur}}^{1/2} G_e \Lambda_{\text{cur}}^{1/2}\big), \tag{12}$$

which costs one $k \times k$ log-determinant per candidate.

Box–Hill. Under Theorem 4.11 we use

$$\text{BH}(e) = \sum_{i<j} \tfrac{1}{2}\Delta_{ij}^\top (V_i^{-1} + V_j^{-1})\Delta_{ij} \tag{13}$$

over all library pairs (i, j), with $\Delta_{ij} = m_i(e) - m_j(e)$ and V_i the predictive covariance under model i evaluated at e. In the isotropic shared-noise limit, this reduces to $\sigma_e^{-2}\sum_{i<j}\|\Delta_{ij}\|^2$, i.e. the disagreement-magnitude baseline. The cascade Box–Hill curve in the main paper uses the full model-specific V_i, so it is not expected to collapse to disagreement.

Local T-optimal. Local T-opt (Atkinson & Fedorov, 1975; Pukelsheim, 2006) selects experiments that maximize the sum-of-squared predictive differences between the two best-fit library members around the current parameter estimate. Under the local linear-Gaussian model it coincides with the (i, j)-restricted Box–Hill criterion at a single dominant pair.

Table 5. Cascade robustness full table. Regret is averaged across 144 trials

d	Init. unres. dim	Raw HB	CARTOGRAPH-A HB	Disagree HB	Raw regret	CARTOGRAPH-A regret	Disagree regret
2	1	0.44	0.44	0.44	0.05	0.05	0.05
4	3	0.00	0.09	0.00	0.31	0.42	0.31
8	6	0.02	0.65	0.02	19.94	0.01	19.94
16	12	0.07	0.72	0.07	2.83	0.01	2.83

Why CARTOGRAPH is not strictly dominated by these baselines. EIG and Box–Hill both require predictive distributions of *every* library member at *every* candidate experiment. CARTOGRAPH requires only the disagreement design H_e and the current unresolved basis U_τ, so it continues to be well-defined under libraries that are specified as simulators without tractable predictive distributions. This is a common deployment profile for AI-scientist stacks whose library is a set of heterogeneous simulators or black-box models.

E. Cascade Benchmark Details

The cascade system is a sequence of first-order ODEs with a scalable-dimension mechanism basis: d kinetic coefficients plus a small number of output-coupling mechanisms, following the general template for model-discrimination experiments in biochemical networks (Vanlier et al., 2014). Library members are obtained by omitting one or two kinetic coefficients from the shared basis; the controversial component $a_C^\star$ is the vector of omitted coefficients. Experiments correspond to forcing profiles with different time-to-peak, amplitude, and dwell combinations. Local Jacobian blocks are computed from ODE sensitivities in closed form.

Protocol. For each $d \in \{2, 4, 8, 16\}$: 6 truths (different realizations of $a_C^\star$) $\times$ 24 noise seeds $= 144$ trials. Rounds $= 3$. Candidates $= 8$ forcing profiles.

Full results. Table 5 reproduces the main robustness table with disagreement included and with standard deviations added. Runtime per selection step in Table 12, Appendix L.

F. PK Divergence Details

Table 6. Per-truth rounds to identification. Raw CARTOGRAPH and CARTOGRAPH-A are identical on all seven PK truths; both beat disagreement only on *absorption_variant* (bold). Random E[rounds] is a Monte Carlo estimate over 10,000 permutations.

Truth	Oracle	Raw	CARTOGRAPH-A	Disagree	Random
absorption_variant	B	**1**	**1**	**2**	4.00
absorption_variant_slow	B	1	1	1	1.14
distribution_variant_easy	C	0	0	0	0.00
distribution_variant_hard	C	1	1	1	1.60
distribution_variant_subtle	C	1	1	1	2.40
mixed_absorption	A	2	2	2	5.33
mixed_balanced	A	2	2	2	5.33

Method sequences. Raw CARTOGRAPH: $[E_1, E_2, E_4, E_5, E_3, E_7, E_6]$; CARTOGRAPH-A: $[E_1, E_2, E_3, E_4, E_5, E_6, E_7]$; disagreement: $[E_2, E_1, E_6, E_7, E_5, E_3, E_4]$. All structured methods dominate random.

G. Scaling Experiment Full Tables

Refer to Table 7

H. EPA Real-Data Full Breakdown

We evaluated oral series from the EPA CvTdb database (Sayre et al., 2020). The cohort filter was: oral route, ≥ 10 concentration measurements, maximum time ≤ 24 h, one series per chemical. The library is $\{A$: 1-compartment oral, B :

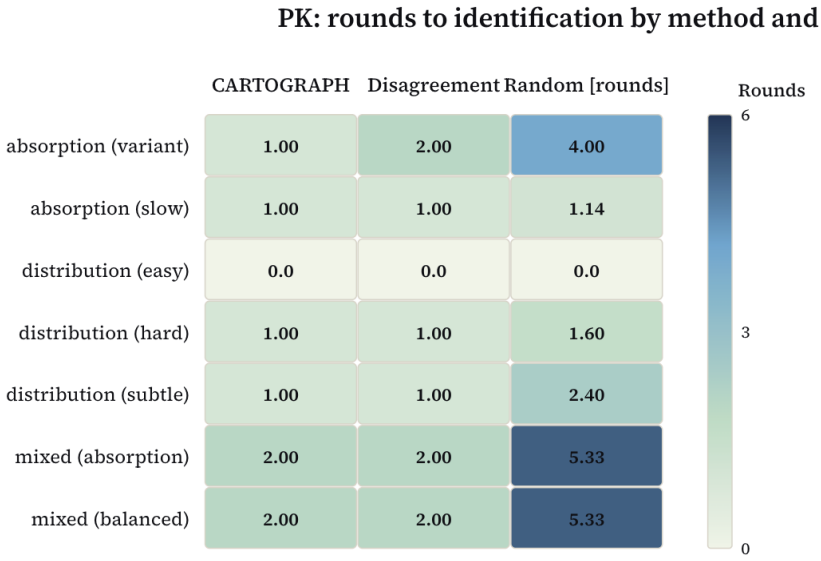

Figure 6. PK rounds to identification, per truth. Raw CARTOGRAPH and CARTOGRAPH-A identify in the same round on all seven truths; the structured methods collectively differ from disagreement on one case (*absorption_variant*).

Table 7. Scaling experiment, $k = 1, 500$ random-candidate instances per d. "Useful Frac" is the empirical fraction of disagreement-selected energy that falls on unresolved directions; "Theory k/d" is the exact prediction of Theorem 4.5.

d	Disagree rate	CART win (proj)	CART win (rank)	Overall adv (rank)	Useful Frac	Theory k/d
2	54.8%	100.0%	76.3%	41.8%	0.501	0.500
3	63.6%	100.0%	71.7%	45.6%	0.323	0.333
4	67.2%	100.0%	68.2%	45.8%	0.242	0.250
6	70.0%	100.0%	61.4%	43.0%	0.175	0.167
8	73.4%	100.0%	61.6%	45.2%	0.196	0.125
10	75.0%	100.0%	70.9%	53.2%	0.175	0.100
12	76.2%	100.0%	67.7%	51.6%	0.170	0.083
15	73.0%	100.0%	72.6%	53.0%	0.165	0.067

delayed-absorption oral, C : 2-compartment oral$\}$.

Cohort-level summary.

Pairwise results. All series: raw CARTOGRAPH vs Disagree $1\mathrm{W}/7\mathrm{T}/0\mathrm{L}$; raw CARTOGRAPH vs T-opt $0\mathrm{W}/7\mathrm{T}/1\mathrm{L}$; T-opt vs Disagree $2\mathrm{W}/6\mathrm{T}/0\mathrm{L}$.

Per-series. Dichloromethane, 1,2-Dichloroethane, Trichloroethylene, Benzo[a]pyrene, Chloroform, Methyl tert-butyl ether, Valproic acid, Glycine. On three of these ($n_{\mathrm{succ}} = 3$) the unresolved-dim is 0 already at the warm start: every acquisition rule must tie. On the other five, raw CARTOGRAPH matches T-opt on four and trails on one.

Fit failures. The cohort contained 96 oral series; 88 failed fit in at least one library member and were skipped. These fit failures are themselves informative: they indicate that the three-model library is inadequate for a large fraction of EPA chemicals, which is exactly the situation the *refuse* layer is designed to surface.

I. A-Lab Retrospective Audit Details

We downloaded the corrected public A-Lab supplementary data (Szymanski et al., 2023; 2026), including the synthesis results CSV and 'Refinement-Table.xlsx'. The audit population is the 40 originally positive claims labelled 'Success' or

Table 8. Scaling experiment, $k = 2$, 500 random-candidate instances per d.

d	Disagree rate	CART win (proj)	CART win (rank)	Overall adv (rank)	Useful Frac	Theory k/d
3	41.2%	100.0%	68.4%	28.2%	0.672	0.667
4	55.2%	100.0%	65.2%	36.0%	0.501	0.500
6	59.8%	100.0%	64.9%	38.8%	0.340	0.333
8	69.0%	100.0%	63.2%	43.6%	0.271	0.250
10	69.6%	100.0%	69.5%	48.4%	0.238	0.200
12	71.4%	100.0%	65.5%	46.8%	0.220	0.167
15	72.6%	100.0%	70.5%	51.2%	0.200	0.133

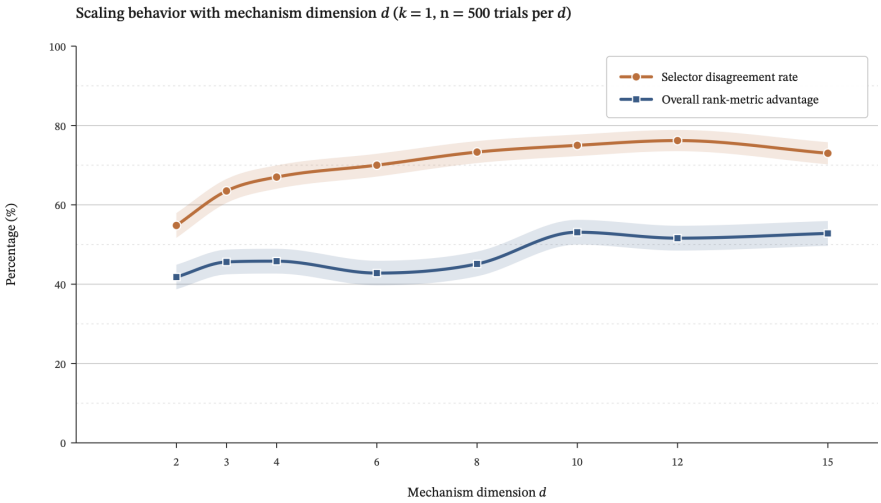

Figure 7. Scaling curves. At low d disagreement and projection nearly agree; at high d the projection selector's pointwise-optimal behavior under the unresolved-energy metric emerges.

'Partial' in the synthesis CSV. External labels come from the corrected manual conclusions: only structure or composition marked inconclusive is counted as 'inconclusive'. We exclude ordering ambiguity alone because it is a crystallographic refinement issue orthogonal to whether the target phase was synthesized.

For each claim we compute a materials-domain residual

$$\rho_{\text{A-Lab}} = \sqrt{(R_{\text{wp}}/20)^2 + ((100 - w_{\text{target}})/100)^2 + (w_{\text{alt}}/100)^2},$$

where R_{wp} is the manual-refinement weighted-profile residual, w_{target} is the target phase fraction, and w_{alt} is the largest non-target phase fraction. We calibrate δ as the 95th percentile of $\rho_{\text{A-Lab}}$ on confirmed 'Success' rows, yielding $\delta = 0.776$, then freeze it before evaluating the full positive-claim set. A deterministic bootstrap over the calibration rows (2000 resamples, seed 0) gives a broad but transparent calibration interval, $\delta \in [0.496, 1.088]$, reflecting the small public cohort.

The four inconclusive claims flagged are $CaGd_2Zr(GaO_3)_4$, $KBaGdWO_6$, $Mn_7(P_2O_7)_4$, and $Mg_3MnNi_3O_8$. An Rwp-only baseline calibrated by the same 95th-percentile protocol flags $0/4$ inconclusive claims and $2/36$ confirmed claims. A target-deficit-only baseline flags $4/4$ inconclusive and $4/36$ confirmed claims, matching the combined guard's confirmed flag count but without the refinement-residual audit trail. We therefore use this result narrowly: it shows that a published autonomous-discovery output can be converted into an auditable pass/flag log under a fixed calibration rule, not that CARTOGRAPH independently settles the crystallographic status of the A-Lab targets.

J. Per-Feature Refusal Diagnostics

Refer to Table 11

Table 9. EPA cohort-level summary for raw CARTOGRAPH (BIC gap margin = 2.00).

Split	Raw CARTOGRAPH margin	Disagree margin	T-opt margin	Random E[margin]	Raw CARTOGRAPH hit
All series ($n = 8$)	7.70	7.69	8.58	4.29	50.0%
Active subset ($n = 5$)	13.55	13.53	13.55	7.99	80.0%

Table 10. A-Lab retrospective audit. The guard flags all four post-correction inconclusive positive claims, at the cost of flagging four confirmed but complex multiphase positive claims for human review.

External label	Claims	Passed	Flagged
Confirmed	36	32	4
Inconclusive	4	0	4

K. Failure-Mode Catalog for CARTOGRAPH

We catalog conditions under which CARTOGRAPH is expected to be wrong.

1. Non-linear local regime. Theorems 4.8, 4.10 and 4.11 rely on the local linear-Gaussian approximation. When the system's curvature around the current posterior mean is large, the first-order equivalences degrade. Mitigation: score with the noise-weighted exact form (6), which remains valid locally even when the first-order approximations drift, and re-linearize after each experiment.

2. Heavy-tailed or adversarial noise. The Gaussian assumption underwrites the closed-form EIG. Under heavy tails or structured model-dependent noise, Σ_e mis-specifies the true noise and G_e is off. Mitigation: use the noise-weighted form with an empirical Bayes estimate of Σ_e.

3. Prior that excludes the truth (refuse failure by miscalibration). If δ is miscalibrated high, CARTOGRAPH identifies out-of-library mechanisms. This is detectable post-hoc from the $(\rho_t, \text{outcome})$ log: sustained increases in ρ under accumulating evidence are an audit signal.

4. Degenerate unresolved subspace at warm start. If the warm-start design is rich enough that $\dim(U_\tau) = 0$ already, every rule ties and CARTOGRAPH reduces to library identification. We flag this at the start of the loop and return a "no-op" status.

5. Near-isotropy of the posterior on U_τ. When $\Lambda_{\text{cur}} \propto I$, CARTOGRAPH-A and raw CARTOGRAPH nearly coincide (first-order equivalence). The cascade $d{=}2$ tie regime is exactly this case. No selection advantage is possible and reporting near-ties is the honest output.

L. Runtime Table

Refer to Table 12

M. Worked LLM-in-the-Loop Example

We show how CARTOGRAPH plugs into an LLM-based AI scientist. The working example is the pharmacokinetic library from Section 5.2; the LLM role is *candidate proposal*, not *decision*.

Setting. The library is $\{A, B, C\}$: one-compartment oral, delayed-absorption oral, and two-compartment oral models, following standard pharmacokinetic parameterizations (Gabrielsson & Weiner, 2016). The LLM is handed the library cards and a natural-language description of the current state of the loop.

LLM prompt (abbreviated). "You are an experimental planner for a pharmacokinetic library containing one-compartment oral, delayed-absorption oral, and two-compartment oral models. Candidate experiments differ in route (oral or IV), dose, sampling schedule (early-dense, mid-dense, late-dense), and observation horizon. Propose up to five candidate experiments as JSON objects with fields `route`, `dose_mg_per_kg`, `sampling_profile`, `horizon_h`, `rationale`. Only propose feasible experiments."

Table 11. Per-feature residual breakdown. Higher = the feature detects more misfit. $C_{\max}$ is the dominant detector for all three failure mechanisms; the control's $C_{\max}$ residual is much lower.

Feature	Time-var	Saturable	Enterohep	Control
$T_{\max}$	0.00	0.00	0.00	0.00
$C_{\max}$	1.83	1.36	1.74	0.73
$\mathrm{AUC}_{\mathrm{frac}}$	0.06	0.08	0.11	0.03
Terminal slope	0.22	0.38	0.29	0.01
Loglin RMSE	0.25	0.21	0.22	0.07

Table 12. Per-step runtime in milliseconds, cascade benchmark, single CPU core.

d	Raw score (ms)	CARTOGRAPH-A score (ms)	Disagree score (ms)
2	0.006	0.034	0.0003
4	0.006	0.039	0.0004
8	0.015	0.268	0.0004
16	0.032	1.299	0.0005

LLM output (one run, abbreviated).


```
[{"route":"oral","dose":5,"sampling":"early_dense","horizon":6,
"rationale":"constrain k_a"}, {"route":"oral","dose":5,"sampling":"late_dense","horizon":24,
"rationale":"constrain terminal slope"}, {"route":"IV","dose":2,"sampling":"early_dense","horizon":4,
"rationale":"isolate distribution vs absorption"}, {"route":"oral","dose":10,"sampling":"mid_dense","
"rationale":"resolve A vs C on C_max"}, {"route":"IV","dose":2,"sampling":"mid_dense","horizon":12,
"rationale":"resolve B vs C on late phase"}]
```


CARTOGRAPH's role. Each proposal is lowered into a candidate disagreement block H_e by evaluating the library's feature-bridge predictions under each candidate experiment. CARTOGRAPH scores them via (4) or (6), picks $e^\star$, and returns two structured objects: (i) a scalar $\mathrm{score}_{\mathrm{cart}}(e^\star)$ for the planner, and (ii) the current unresolved dimension $\dim(U_\tau)$. The planner uses these two numbers to decide whether to ask the LLM for more candidates (if $\mathrm{score}_{\mathrm{cart}}$ is small but $\dim(U_\tau) > 0$) or to execute $e^\star$ and continue the loop.

What this demonstrates. The LLM plays a role it is good at (generating plausible candidates with natural-language rationale) and *not* a role it is bad at (discriminating candidates under a joint posterior). CARTOGRAPH supplies the latter. Combined with the refuse layer, the resulting agent has three behaviors no LLM-only planner produces natively: it refuses identification when the library is structurally wrong, it revokes tentative identifications under contradictory evidence, and it certifies when the unresolved question has been closed.