DIRECTED RESIDUAL INFORMATION: LEARNING FROM REDUNDANT HYPOTHESIS REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A new hypothesis can make different predictions without improving learning. It may repeat what existing hypotheses already predict, or vary in a way unrelated to the remaining error. We ask when adding a hypothesis helps a learner explain its target. For a linear learner with squared loss and a penalty on the new coefficient, the answer depends on the error left unexplained, the new prediction direction, and its agreement with that error. We call this relationship directed residual information. Classical regression identities make it exact for the stated objective; the contribution is an explicit account of these failure modes and of what can be known before evaluating a proposal. A controlled experiment separates useful from irrelevant directions while holding their novelty nearly fixed. On 11 public binary datasets, 2,086 generated proposals test whether the measured relationship predicts improvement on separate test data. The resulting score ranks improvement better than novelty alone by mean within-dataset Spearman .308 (95% CI $[.190, .431]$), but its difference from partial R^2 is unresolved ($-.010$ $[-.033, .004]$). Because the current representation is fixed within each split, this ranking test does not isolate the contribution of remaining target error. Frozen replay separately finds no support for one attempt to turn past measurements into a selection rule. The account explains evaluated proposals under stated objectives; choosing unseen proposals remains a prediction problem.

1 INTRODUCTION

Many learning systems improve by proposing and evaluating new hypotheses. Yet evaluating more hypotheses need not reduce prediction error. Two different expressions may make the same predictions. Even when their predictions differ, that difference may have little to do with the errors the learner still makes. Automated scientific learners make these failures concrete: they generate programs or equations, evaluate their predictions, and update what they try next (King et al., 2004; Romera-Paredes et al., 2024; Lu et al., 2024; Reddy & Shojaee, 2025). Hypotheses may also come from retrieval, language models, or human design.

These failures require different explanations. In the first, existing predictions already contain what the new hypothesis offers. In the second, the hypothesis adds something different but does not help explain the target. Here, a *proposal* means a hypothesis offered for evaluation. Novelty alone does not settle whether it is useful.

This paper asks a single question: *when does a new hypothesis actually improve learning?* Our answer proceeds in three steps:

Different syntax does not imply new behavior.
New behavior does not imply useful learning.
Useful learning requires new behavior that explains what remains unexplained.

Learning therefore depends not merely on whether a proposal is different, but on how it changes what the learner can explain about the target.

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.

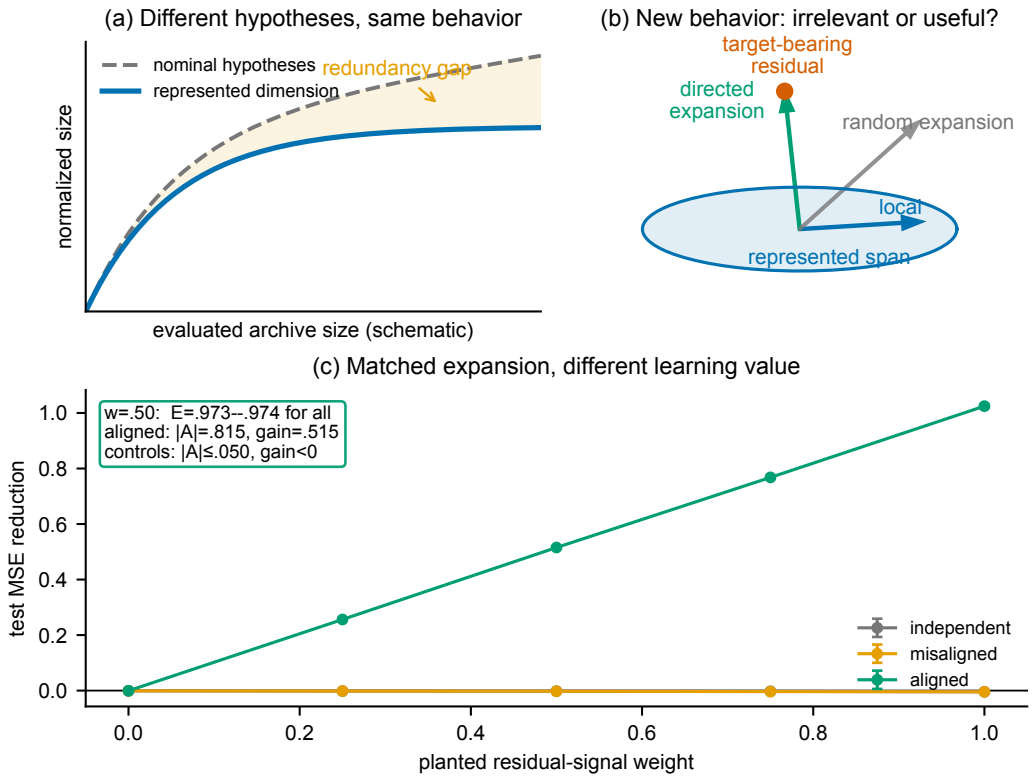


Figure 1: **Why novelty alone does not guarantee learning.** (a) Syntactically different hypotheses may induce redundant responses. (b) A genuinely new response direction may still be irrelevant to the target residual. (c) A controlled construction holds expansion nearly fixed while varying residual signal and alignment; only the aligned proposal improves held-out error. Training, calibration, and test samples are disjoint, with 300 repetitions per cell and normal 95% CIs. This is controlled gate separation, not prospective policy validation.

For a linear learner, the question becomes concrete. Combine the predictions of already evaluated hypotheses, then examine the target error that remains. A useful new prediction must add a direction that helps explain this error. We call this relationship *directed residual information*. The formal account separates remaining target error (B), a new prediction direction (E), and agreement with the residual error (A). Section 2 defines these three roles as learning gates, along with the response-space representation used to measure them.

A measurement can explain why an evaluated proposal helped without telling us which unseen proposal to try next. Its predictions must be available before we can measure their novelty and agreement with remaining error. Using those measurements to justify the earlier choice would leak future information into that choice. This timing restriction is the *observability boundary*: explanation after evaluation does not itself provide a selection policy.

The contribution connects these failure modes to measurable quantities and separates what they explain from what a future controller would need to predict or otherwise account for. The projection, partial-regression, ridge, and Gaussian relations are classical; Section 3 states their equivalences explicitly. The experiments test the resulting distinctions, not whether a renamed regression statistic outperforms its classical counterpart.

Figure 1 separates the two failures experimentally. Proposals add similarly new prediction directions, but only those aligned with the remaining target error improve held-out predictions. Independent training, calibration, and test samples separate fitting, measurement, and verification; Section 4 gives the results.

The central formal result is objective-specific. Under one-step ridge-regularized squared-loss learning, the exact improvement is

$$\frac{G_\lambda(z)}{\|y\|^2} = B_t \frac{E_t(z)}{E_t(z)+\lambda} A_t(z)^2, \tag{1}$$

with definitions given in Section 2. A rank/Gram identity states when capacity expands but contains no target, while a Gaussian noisy-channel identity supplies a conditional-exact information interpretation. Each result has its own assumptions; we do not claim a universal information product.

We separate evidence by role:

- **Exact-objective diagnostic.** Independent numerical paths check the stated asymmetric objective to machine precision; this is an implementation consistency test.
- **Controlled gate separation.** A sample-split construction holds proposal expansion nearly fixed while varying residual signal and alignment, then measures held-out predictive value.
- **Public Proposal-response Archive.** A fixed archive tests held-out discrimination after directions are generated and without a planted aligned direction.
- **Observability Analysis and Boundary Audit.** Quantity timing defines the boundary; frozen replay audits one attempted bridge to prospective control.
- **External Compatibility.** Finance and symbolic-regression studies test qualified compatibility across proposal sources.

Appendix C reports the compatibility evidence; it does not serve as an independent gate test.

2 Directed Residual Information

Let $\mathcal{H}$ be a real Hilbert space of centered response vectors. At step t, the learner represents the closed subspace $\mathcal{S}_t = \mathrm{span}(\mathcal{A}_t)$ generated by evaluated hypotheses. The orthogonal projector onto $\mathcal{S}_t$ is P_t. This representation is empirical: two syntactically different hypotheses are equivalent for the present analysis if their response vectors coincide.

For target $y \neq 0$ and unit-norm proposal z, define the target and proposal residuals

$$y_{\perp,t} = (I - P_t)y, \qquad z_{\perp,t} = (I - P_t)z. \tag{2}$$

We use true orthogonal projection in the population statements. Ridge residuals and rolling predictive-novelty scores used in experiments are finite-sample proxies and are not assumed to equal these geometric objects.

Definition 1 (The three learning gates)**.** *The* unrepresented target energy*,* representation expansion*, and* residual alignment *are*

$$B_t = \frac{\|y_{\perp,t}\|^2}{\|y\|^2}, \qquad E_t(z) = \|z_{\perp,t}\|^2, \qquad A_t(z) = \frac{\langle y_{\perp,t}, z_{\perp,t}\rangle}{\|y_{\perp,t}\|\,\|z_{\perp,t}\|}, \tag{3}$$

with $A_t(z) = 0$ when either residual is zero.

B_t asks whether the current representation leaves any target information available. E_t asks whether the proposal adds a response direction. A_t asks whether the added direction is useful for the target. Each quantity answers a different question.

Equivalently, the three gates ask: *Is there anything left to learn? Does the proposal add a new response direction? Does that direction point toward what remains unexplained?*

Archive redundancy is an upstream diagnostic. For response correlation matrix $\Sigma_{\mathcal{A}}$ with eigenvalues λ_i and $p_i = \lambda_i / \sum_j \lambda_j$, we report the entropy effective rank (Roy & Vetterli, 2007)

$$r_{\mathrm{eff}}(\Sigma_{\mathcal{A}}) = \exp\left(-\sum_i p_i \log p_i\right). \tag{4}$$

A zero eigenvalue contributes $0 \log 0 = 0$ by continuity. A low r_{eff}/N means that the evaluated archive concentrates variance in few directions. It does not imply that the target has residual energy, that a particular proposal expands the span, or that the proposal aligns with the target. Conversely, $r_{\mathrm{eff}}/N = 1$ for a small archive does not exhaust a larger ambient space.

Principle 1 (Directed residual information)**.** *Redundancy diagnoses repeated use of represented directions; a learning opportunity exists only when target-relevant information remains outside the current representation. A proposal realizes that opportunity only by expanding the representation in a direction aligned with the target residual.*

The principle intentionally does not specify a generator. Local symbolic variation, retrieval, an LLM, program search, or a human scientist may produce z. Under the subspace model, linear recombination of represented responses remains in $\mathcal{S}_t$; other local operators need not. Any source is evaluated only after its proposal has been mapped into response space and the gates have been defined for a particular learning objective.

3 Objective-Specific Acquisition Identities

The knowledge claim is a typed, falsifiable account of generated-proposal acquisition, not a new algebraic identity. The results below formalize this account by specializing standard projection, residualized-regression, and Gaussian-information identities (Golub & Van Loan, 2013; Frisch & Waugh, 1933; Lovell, 1963; Hoerl & Kennard, 1970; Cover & Thomas, 2006). The finite-dimensional matrices are coordinate realizations of Section 2; proofs are in Appendix A.

Classical equivalence map. $z_\perp = (I - P_\mathcal{S})z$ is the Frisch–Waugh–Lovell residual. For centered targets, $B = 1 - R^2_\mathcal{S}$, scale-normalized $E = 1 - R^2_{z\sim\mathcal{S}}$ is candidate tolerance, and A^2 is partial R^2. In the unpenalized limit, $G_0(z)/\|y\|^2 = BA^2 = \Delta R^2$; the asymmetric penalty contributes $E/(E+\lambda)$. Residual-correlation forward selection and matching pursuit use the same one-step agreement with the current residual (Mallat & Zhang, 1993; Pati et al., 1993); Gaussian conditional mutual information is monotone in squared partial correlation. We claim none of these equivalences as new. We ask whether typing them as target opportunity, expansion, and alignment yields discriminative predictions for generated proposals.

3.1 Representation expansion is capacity, not learning

Proposition 1 (Exact representation expansion)**.** *Let the full-column-rank matrix X span $\mathcal{S}$ and let $r = (I - P_\mathcal{S})z$. Then*

$$\operatorname{rank}([X, z]) = \operatorname{rank}(X) + \mathbf{1}\{\|r\| > 0\}, \tag{5}$$

and

$$\det([X, z]^\top [X, z]) = \det(X^\top X)\,\|r\|^2. \tag{6}$$

Claim type: exact capacity identity. Proposition 1 is the Schur-complement/Gram identity in acquisition notation. It creates a nonzero higher-dimensional parallelotope exactly when $r \neq 0$, but numerical volume need not increase when $0 < \|r\| < 1$. The result contains no target y. Capacity expansion is therefore not evidence of learning, and entropy effective rank cannot replace the projector without additional assumptions.

3.2 Target-relevant learning gain

To make proposal scale operational, consider a one-step update that penalizes only the new coefficient. For $\lambda > 0$, define the improvement in this asymmetric penalized objective

$$G_\lambda(z) = \inf_{s\in\mathcal{S}} \|y - s\|^2 - \inf_{\substack{s\in\mathcal{S}\\ a\in\mathbb{R}}} \left\{\|y - s - az\|^2 + \lambda a^2\right\}. \tag{7}$$

Theorem 1 (Regularized residual information acquisition)**.** *Fixing t and suppressing its subscript, for unit-norm z and the quantities in Eq. 3,*

$$G_\lambda(z) = \frac{\langle y_\perp, z_\perp\rangle^2}{\|z_\perp\|^2 + \lambda} = \|y\|^2 B \frac{E(z)}{E(z) + \lambda} A(z)^2. \tag{8}$$

Consequently, $G_\lambda(z) > 0$ if and only if $B > 0$, $E(z) > 0$, and $A(z) \neq 0$.

Claim type: exact objective identity. Theorem 1 is the central formal statement. The three gates are individually necessary and jointly sufficient for positive gain under the asymmetric objective in Eq. 7. When $\lambda = 0$ and $E(z) > 0$, residual magnitude cancels by feature-rescaling invariance. Any claim that gain grows with expansion distance therefore requires regularization, observation noise, update constraints, or cost.

Section 4 checks this equality under the exact asymmetric objective. That numerical check is an implementation diagnostic; the held-out experiments ask a different predictive question.

Proof sketch. Residualizing both y and z against $\mathcal{S}$ reduces Eq. 7 to the one-dimensional quadratic $\|y_\perp - az_\perp\|^2 + \lambda a^2$. Its minimizer is $a^* = \langle y_\perp, z_\perp\rangle/(\|z_\perp\|^2 + \lambda)$; substitution and the definitions of B, E, A give Eq. 8. The complete argument is in Appendix A.

3.3 Information-theoretic interpretation

Let U denote the current representation and suppose a standardized external proposal is observed with noise. The Gaussian case makes conditional information explicit.

Theorem 2 (Conditional information from a noisy proposal)**.** *Let scalar Y, Z and vector U be jointly Gaussian, with $\mathrm{Var}(Z) = 1$. Define $e_Y = Y - \mathbb{E}[Y \mid U]$ and $e_Z = Z - \mathbb{E}[Z \mid U]$, with $\mathrm{Var}(e_Y) > 0$, $E = \mathrm{Var}(e_Z) > 0$, and $A = \mathrm{Corr}(e_Y, e_Z)$. If the learner observes $W = Z + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, is independent of (Y, U, Z), then*

$$I(Y; W \mid U) = -\frac{1}{2}\log\left(1 - A^2\frac{E}{E + \sigma^2}\right). \tag{9}$$

Claim type: conditional exact identity. Expansion determines how much standardized proposal variance survives conditioning on the current representation; alignment determines what fraction is target-bearing. Observation noise makes proposal scale explicit, which is why the theorem fixes $\mathrm{Var}(Z) = 1$. Equations 8 and 9 express related residual mechanisms under different objectives. They do not establish a universal product law beyond their stated assumptions (Chechik et al., 2005).

4 Evidence by Role

Core evidence map.
Exact-objective diagnostic → verifies the factorization under the stated objective.
Controlled Gate Separation → isolates target alignment under matched expansion.
Public Proposal-response Archive → tests held-out discrimination without a planted direction.
Observability Analysis + Boundary Audit → defines the timing boundary and audits one direct policy bridge.
External Compatibility is reported in Appendix C and does not serve as an independent gate test. No level substitutes for another.

4.1 Exact-objective implementation diagnostic

Theorem 1 concerns the asymmetric objective in Eq. 7, not ordinary ridge with all coefficients penalized. We therefore compute the nested-objective difference and closed-form factorization by independent linear-algebra paths. Across 5,625 combinations of B, E, A, λ and random rotations, the maximum absolute discrepancy is 6.7×10^{-16}. This is an implementation consistency check, not evidence of generalization; Appendix B gives the grid and code hash.

4.2 Controlled gate separation

The ambient response is $h \sim \mathcal{N}(0, I_{20})$, and the current representation is $U = (h_1, \ldots, h_8)$. We generate

$$y = \sqrt{1 - w}\, h_1 + \sqrt{w}\, h_9 + 0.5\epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1), \tag{10}$$

where $w \in \{0, .25, .50, .75, 1\}$ varies planted signal outside the represented population subspace. The aligned proposal is $z = h_9$; a fixed misaligned proposal is h_{10}; and a randomized independent proposal is a unit-norm random combination of $(h_{10}, \ldots, h_{20})$, with its direction fixed across the

three samples within a repetition. All proposal types are outside $\mathrm{span}(U)$ in the population. Each cell has 300 repetitions. Ridge fitting uses 300 training observations, features standardized with training-sample statistics, and penalty 10. The gates B, E, A are estimated on 300 independent calibration observations, and final MSE is measured on 1,000 further observations. No target outcomes are reused across these stages.

Calibration estimates are marked with hats because they are finite-sample, scale-normalized proxies: $\widehat{E} = \|\widehat{z}_\perp\|^2/\|z\|^2$ and $\widehat{A}$ is the calibration residual correlation. Figure 1c shows that aligned-proposal test gain increases from -0.0007 at $w = 0$ to 1.0246 at $w = 1$. At $w = .50$, all proposals have $\widehat{E} \in [.973, .974]$, but the aligned proposal has $|\widehat{A}| = .815$ and reduces held-out MSE by 0.5154 (95% CI $[.5114, .5193]$). The two controls have $|\widehat{A}| \leq .050$ and gains near $-.002$. The held-out experiment asks whether predictive value varies with alignment when estimated expansion is matched. Within this construction it verifies the predicted ordering; it does not identify a prospective proposal rule.

4.3 Public proposal-response archive

The controlled study selects the aligned direction by construction. We therefore reconstruct a fixed archive of 2,086 deterministic response proposals on 11 public binary OpenML datasets, each with two frozen splits (Vanschoren et al., 2014). The current representation contains up to 10 raw features selected using training data only; proposals are raw, square, absolute-value, pairwise-product, and absolute-difference transforms. Proposal identities and train/validation/test responses exist before gate scoring. Gates are estimated on validation data, while an asymmetric update with mean-loss proposal penalty .1 is fitted on training data and evaluated by relative MSE reduction on the disjoint test split. The independent unit is the dataset; split-level and proposal-level results are averaged within dataset before 10,000-replicate cluster bootstrap inference.

Accounting for remaining target error predicts held-out improvement better than measuring novelty alone. Mean within-dataset Spearman is $.518$ (95% CI $[.321, .712]$) for the three-gate score and $.210$ $[.071, .366]$ for expansion; their paired difference is $.308$ $[.190, .431]$. Selecting the top 10% yields mean relative test-MSE reductions of 5.39% and $.91\%$, respectively, a 4.49-percentage-point difference $[1.27, 8.24]$ (positive on 9/11 datasets; exact sign-flip $p = .0195$).

The closest classical comparator, partial $R^2 = \widehat{A}^2$, is equally important (Figure 2). Its rank association is $.528$ $[.343, .717]$; the three-gate difference is unresolved ($-.010$ $[-.033, .004]$). Its top-decile gain is 5.54%. A post-hoc paired comparison on the frozen results gives a three-gate minus partial-R^2 difference of $-.14$ percentage points (95% CI $[-1.10, .54]$). Neither comparison establishes superiority or equivalence.

The remaining-error quantity $\widehat{B}$ is constant across proposals within each split, so it cannot affect their ranking. Thus this experiment tests the benefit of residual alignment over expansion alone, not the separate contribution of every gate. Appendix B details the classical score naming and the excluded 26-class sensitivity task.

4.4 External compatibility checks

Appendix C reports an A-share-calibrated semi-synthetic geometry sweep, a licensed unequal-budget market application, and a cached 158-task symbolic-regression subset. These studies test compatibility across proposal sources, but lack realized proposal geometry, equal budgets, or complete archive provenance. They therefore do not support the exact identity, public-archive discrimination, or generator dominance.

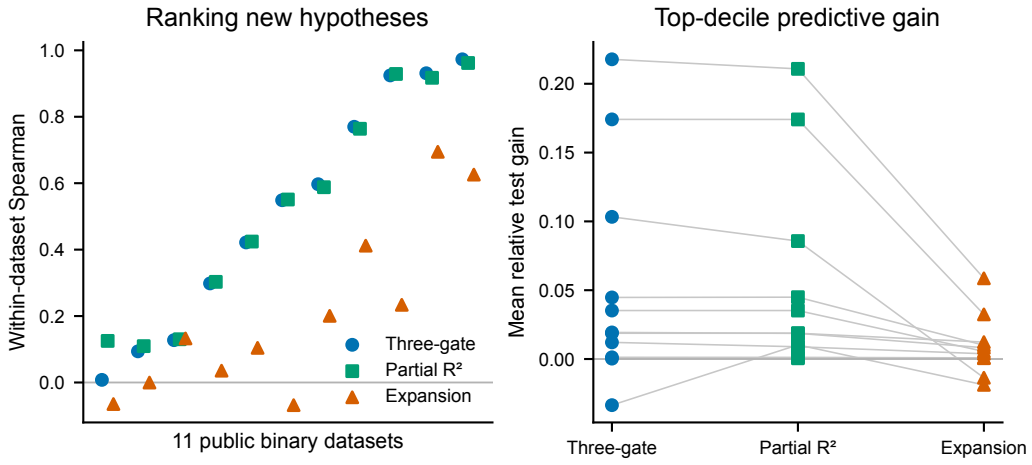


Figure 2: **Residual alignment helps discriminate useful proposals; the classical baseline remains competitive.** Left: rank association with disjoint test gain, including partial R^2 alongside the three-gate and expansion scores. Right: mean test gain for each method's top 10%. Points average two splits per dataset ($n = 11$); gray lines pair datasets. This observational test measures held-out discrimination, not prospective controller value.

5 Observability Boundary

Proposal-specific expansion and alignment become available only after a proposal has been generated and evaluated. This timing creates an observability boundary: realized gates may explain an evaluated proposal and update the next archive state, but they cannot justify the action that produced that proposal. A prospective controller would need to predict—or otherwise account for—proposal-dependent expansion and alignment using leakage-safe information available before evaluation. This is a measurement and control boundary, not a second learning principle or a universal impossibility theorem. Appendix D visualizes the timing.

For auditability, Appendix D defines the discovery environment $\mathcal{E} = (G, V, U, c, R, \mu_0)$ and standard trajectory value $J(\pi; \mathcal{E})$. Shared response rules reduce obvious confounding by holding proposal access, verification, accounting, update timing, stopping, and scoring fixed (Russell & Wefald, 1991; Thomas & Brunskill, 2016; Majumder et al., 2025; Chen et al., 2025). They are validation infrastructure, not a third learning principle.

5.1 Frozen-replay boundary audit

We audit one attempted bridge across the observability boundary through frozen replay on 12 OpenML datasets (Vanschoren et al., 2014). Every policy receives the same frozen proposal supply, verification rules, cost models, budgets, stopping contract, and scoring rule. The endpoint is oracle-normalized utility–cost AUC. Repeated splits, orders, budgets, and costs are averaged within dataset before cluster inference; the independent unit is the dataset ($n = 12$), not the 124,992 nested trajectories. Proposal streams and terminal scores existed before the analysis freeze; the registry and analysis plan were then fixed.

The geometry-based CAHS stress test chooses between local and non-local sources using cost-normalized UCB scores built only from lagged archive compression, recent realized escape, recent predictive novelty, and cost; it never observes the same proposal's expansion, alignment, or terminal value before acting. Leave-one-dataset-out calibration uses the other 11 datasets and selects fixed 25% local/75% external in all 12 folds.

The frozen preregistration encoded a negative-sign machine rule for the trajectory-average effective-rank-ratio slope, while its intended diminishing-return interpretation treated lower rank ratio (greater compression) as lower utility. These directions are inconsistent. We preserve the frozen machine

status, but do not count it as substantive support: the final registered estimate is $-.3423$ (10,000-bootstrap 95% CI $[-.5988, -.0300]$). A separate post-hoc temporal audit of pre-action rank and next-action utility is inconclusive: $+.0017$ ($[-.0013, +.0046]$). Full CAHS differs from its no-rank ablation by only $+.0009$ ($[-.0009, +.0027]$) and loses to fixed 25/75 by $-.0135$ ($[-.0212, -.0057]$), with 10 losses in 12 dataset contrasts (exact sign-permutation $p = .009277$, Holm-adjusted $p = .027832$). Appendix D gives the endpoint, controls, design, archived 1,000-bootstrap artifact, final 10,000-bootstrap correction, distance diagnostics, stopping behavior, and asymmetric-cost audit.

This negative audit concerns the tested policy bridge, not the exact identity, controlled gate separation, or public-archive result. Effective rank remains a descriptive archive statistic, and realized expansion and alignment remain post-proposal quantities. A prospective controller would need to predict—or otherwise account for—the proposal-dependent gates before evaluation, together with cost and uncertainty.

6 Related Work

Residualized gain and greedy pursuit. Residualized regression is the closest classical account of our linear quantities (Frisch & Waugh, 1933; Lovell, 1963; Hoerl & Kennard, 1970). As made explicit in Section 3, B is unexplained target variation, scale-normalized E is candidate tolerance, A^2 is partial R^2, and $BA^2 = \Delta R^2$ without regularization. Matching and orthogonal matching pursuit similarly choose atoms from a fixed redundant dictionary by agreement with the current residual (Mallat & Zhang, 1993; Pati et al., 1993). Our generated proposal need not belong to a pre-enumerated dictionary: its identity and response may become available only after a source action.

Conditional relevance and representation. Information-theoretic feature selection already evaluates candidate relevance conditional on selected features (Brown et al., 2012; Wollstadt et al., 2023); our Gaussian theorem is a noisy scalar CMI specialization, not a new information criterion. Information Bottleneck separates compression from relevance (Tishby et al., 1999; Chechik et al., 2005), and concept models expose incomplete concept banks and operational informativeness (Yuksekgonul et al., 2023; Liu et al., 2024). We type these roles in the response span of evaluated hypotheses and test their held-out discrimination.

Prospective acquisition and value of information. Value of information is ex ante (Howard, 1966). Active feature-value acquisition ranks known costly measurements using predicted value, including modern generative-surrogate estimates (Saar-Tsechansky et al., 2009; Li & Oliva, 2021); active learning and entropy-search methods likewise value future observations before acquisition (MacKay, 1992; Hennig & Schuler, 2012; Hernández-Lobato et al., 2014). These methods anticipate the need to combine predicted value, uncertainty, and cost. Our observability boundary concerns a different interface: a source action may first generate an unseen hypothesis whose response, expansion, and alignment are not yet observed.

Scientific discovery systems. AI Feynman, symbolic regression, FunSearch, and LLM-SR instantiate different proposal mechanisms (Udrescu & Tegmark, 2020; La Cava et al., 2021; Romera-Paredes et al., 2024; Shojaee et al., 2025). DiscoveryBench and ScienceAgentBench define executable units and shared evaluation rules (Majumder et al., 2025; Chen et al., 2025). Our analysis is source-agnostic: an LLM is one generator, while the theoretical object is the verified response vector. Novelty search rewards behavioral novelty, while quality diversity combines quality and diversity (Lehman & Stanley, 2011; Pugh et al., 2016); our objective specifically values residual-aligned expansion.

These traditions study individual parts of the problem; we do not claim that generated proposals or any one component is introduced here. The contribution begins when proposal generation, response redundancy, target-relative value, and action-time observability jointly define one acquisition problem with falsifiable consequences.

7 Scope, Limitations, and Implications

Claim boundary. This is an objective-specific explanatory account, not a universal learning law, a uniquely minimal decomposition, a new regression identity, or an optimal controller.

Scope of the identities. Theorem 1 is exact for a linear subspace and one-step ridge-regularized squared loss. Theorem 2 assumes joint Gaussianity and independent additive noise. Nonlinear representation updates, non-Gaussian targets, multi-step feedback, and updates that change earlier features require new analysis. Without regularization or noise, expansion magnitude is scale-invariant once nonzero; Eq. 8 is not an unconditional product law.

Measurement. The accepted archive is endogenous to the generator, verifier, admission rule, deduplication, and finite-sample response matrix. Compression and mutual-information estimates can change under representation and estimator choices (Butakov et al., 2024). The controlled construction separates gates under a designed target direction. The public archive removes the planted aligned direction, but its current representation is selected using training targets; it remains observational, transformation-specific, and limited to a squared-loss surrogate on public binary tasks.

External validity. The finance and symbolic-regression compatibility checks have unequal budgets, incomplete proposal geometry, dependent conditions, or incomplete provenance; they are not replications of the gate test. The fixed OpenML proposal family does not represent open-ended scientific search or long-run feedback. Replay outcomes existed before analysis freeze, so policy-time hiding is enforced but investigator-blind prospective confirmation is not. Dataset-cluster inference remains small-sample ($n = 11$ for the public archive; $n = 12$ for replay).

Implication. Redundancy can diagnose repeated represented directions, but it neither proves that target residual remains nor chooses a source. Under the studied interface, a prospective controller would need to represent or estimate the current opportunity B and predict—or otherwise account for—proposal-dependent E and A before evaluation, together with cost and uncertainty. We identify these design variables and their leakage constraints but do not provide that controller.

Ethics and data statement

The study uses no human subjects or personally identifiable information. Synthetic procedures, public benchmark identifiers, derived tables, and replay contracts are documented. Licensed Tushare Pro observations are not redistributed. The framework could affect automated allocation of scientific resources; users should not treat descriptive geometry as a decision rule without prospective validation.

8 Conclusion

New predictions help a learner when they explain target error that existing predictions leave unresolved. We formalize this account for a stated regularized linear objective and test its consequences with controlled and public-data studies. The public comparison supports residual alignment over novelty alone, not superiority to partial R^2 or an independent ranking contribution from every gate. The audit also has a concrete use: it exposes apparent policy wins created by unequal cost accounting (Appendix D). Under this interface, future controllers would need to predict or otherwise account for proposal-dependent value, cost, and uncertainty before evaluation. The account makes these requirements explicit; it does not solve the prospective selection problem.

AI use statement

OpenAI Codex was used to help refine the conceptual framing and hypotheses; suggest and critique formal statements, proof organization, validation design, and integrity checks; implement and revise verification scripts and figures; interpret frozen results; search and organize literature; and assist with drafting, language revision, LaTeX preparation, and citation formatting. It did not generate raw market observations or replace unavailable experiments with synthetic substitutes. The authors checked the mathematical derivations, inspected the cited primary sources and project artifacts, reran the shareable experiments and frozen statistical audits, and reviewed all AI-assisted text, code, figures, and interpretations. The authors take responsibility for the final content, including all claims and artifacts produced with AI assistance. DeepSeek-V4-Flash was used only as an experimental proposal generator in the cached symbolic-regression compatibility study; its outputs were treated as experimental data rather than as authoritative scientific or editorial text.

Reproducibility statement

The appendix gives complete proofs, diagnostic grids, all controlled-study cells, public-archive inference, transfer configurations, and frozen replay estimands. The anonymous supplement includes executable generators, split-hashed proposal responses, hash-pinned cached sources, public identifiers, replay envelopes, and offline verification scripts. Licensed raw market observations are not redistributed.

References

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012. URL https://www.jmlr.org/papers/v13/brown12a.html.

Ivan Butakov, Aleksandr Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, Alexey Frolov, and Kirill Andreev. Information bottleneck analysis of deep neural networks via lossy compression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b2b502c3629beadda06311386d2c6f73-Abstract-Conference.html.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6(6):165–188, 2005. URL https://www.jmlr.org/papers/v6/chechik05a.html.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. doi: 10.1002/047174882X.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. doi: 10.2307/1907330.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013. ISBN 9781421407944.

Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(57):1809–1837, 2012. URL https://jmlr.org/papers/v13/hennig12a.html.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/6488484c982e9af5c35689523ba1abfe-Abstract.html.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004. doi: 10.1038/nature02236. URL https://doi.org/10.1038/nature02236.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary symbolic regression methods and their relative performance. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c0c7c76d30bd3dcaefc96f40275bdc0a-Abstract-round1.html.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011. doi: 10.1162/EVCO_a_00025.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Ruizhe Liu, Qian Luo, and Yanchao Yang. InfoCon: Concept discovery with generative and discriminative informativeness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a3afe8afdc381e367da88ffd36803011-Abstract-Conference.html.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963. doi: 10.1080/01621459.1963.10480682.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992. doi: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d70af566e69f1dfb687791ecf955e28-Abstract-Conference.html.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. doi: 10.1109/78.258082.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, volume 1, pp. 40–44. IEEE, 1993. doi: 10.1109/ACSSC.1993.342465.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016. doi: 10.3389/frobt.2016.00040.

Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative AI: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28601–28609, 2025. doi: 10.1609/aaai.v39i27.35084. URL https://doi.org/10.1609/aaai.v39i27.35084.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007. doi: 10.5281/zenodo.40328.

Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1–3): 361–395, 1991. doi: 10.1016/0004-3702(91)90015-C. URL https://doi.org/10.1016/0004-3702(91)90015-C.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009. doi: 10.1287/mnsc.1080.0952.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/28df8e730c054c5331855fd4d5403ba9-Abstract-Conference.html.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148. PMLR, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377. University of Illinois, 1999. URL https://www.princeton.edu/~wbialek/our_papers/tishby+al_99.pdf.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *Journal of Machine Learning Research*, 24(131):1–44, 2023. URL https://www.jmlr.org/papers/v24/21-0482.html.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

A Proofs

A.1 Proof of Proposition 1

Write $z = P_{\mathcal{S}} z + r = Xb + r$ with $X^\top r = 0$. Elementary column operations replace the last column of $[X, z]$ by r without changing rank. Hence the rank increases exactly when $r \neq 0$. For the determinant, the block Gram matrix is

$$\begin{bmatrix} X^\top X & X^\top z \\ z^\top X & z^\top z \end{bmatrix}. \tag{11}$$

Because $X^\top X$ is invertible, its determinant equals $\det(X^\top X)$ times the Schur complement

$$z^\top z - z^\top X (X^\top X)^{-1} X^\top z = z^\top (I - P_{\mathcal{S}}) z = \|r\|^2. \tag{12}$$

A.2 Proof of Theorem 1

Orthogonality gives $\inf_{s \in \mathcal{S}} \|y - s\|^2 = \|y_\perp\|^2$. For fixed a, the best s absorbs the represented component of az, leaving

$$\|y_\perp - a z_\perp\|^2 + \lambda a^2. \tag{13}$$

This quadratic is minimized at

$$a^* = \frac{\langle y_\perp, z_\perp \rangle}{\|z_\perp\|^2 + \lambda}. \tag{14}$$

Substitution yields the first equality in Eq. 8. The second follows from $\|y_\perp\|^2 = \|y\|^2 B$, $\|z_\perp\|^2 = E$ for unit z, and $\langle y_\perp, z_\perp\rangle^2 = \|y_\perp\|^2\|z_\perp\|^2 A^2$. The denominator is positive because $\lambda > 0$; the gain is strictly positive exactly when the numerator is nonzero, equivalently when all three stated gates are open.

A.3 Proof of Theorem 2

Condition on U. Joint Gaussianity makes (e_Y, e_Z) bivariate Gaussian. Since $W = Z + \varepsilon$, the residual of W after conditioning on U is $e_Z + \varepsilon$. Its variance is $E + \sigma^2$, and

$$\mathrm{Corr}(e_Y, e_Z + \varepsilon)^2 = A^2 \frac{E}{E + \sigma^2}. \tag{15}$$

Conditional mutual information for a bivariate Gaussian pair with partial correlation ρ is $-\frac{1}{2}\log(1 - \rho^2)$, which gives Eq. 9.

B Extended Experimental Details

B.1 Exact-objective diagnostic

The diagnostic constructs centered orthonormal response coordinates so that $\|y\| = \|z\| = 1$ and the population gates equal a prescribed grid: $B \in \{.1, .5, .9\}$, $E \in \{.05, .25, .5, .75, .95\}$, $A \in \{0, .25, .5, .75, 1\}$, and $\lambda \in \{.1, 1, 10\}$. For each of 25 random rotations, one implementation solves the block penalized least-squares problem directly, while a separate path evaluates Eq. 8. Across all 5,625 instances, the maximum absolute error is 6.6614×10^{-16} and the mean absolute error is 5.3313×10^{-17}. The seed is 20270731. Source and complete rows are in `recompute_exact_gate_diagnostics.py` and `exact_gate_identity.csv`.

A separate 2,000-world continuous construction draws $B, E \sim \mathrm{Unif}(.02, .98)$ and $A \sim \mathrm{Unif}(-1, 1)$, then uses independent samples of 300/300/1,000 for fitting, gate calibration, and testing. Under ordinary ridge with penalty 10, Spearman association with held-out gain is .901 for the scale-normalized three-gate score, .979 for BA^2, and .027 for expansion alone. Because this update penalizes all coefficients and differs from Eq. 7, it is a qualitative stress test. Here the expansion factor does not improve on the classical incremental-R^2 score BA^2, distinct from partial $R^2 = A^2$ across worlds with different B.

B.2 Controlled generator, timing, and configuration

For each repetition and residual-signal weight w, Eq. 10 is sampled independently for training ($n = 300$), calibration ($n = 300$), and final testing ($n = 1000$). A base ridge model uses $U = (h_1, \ldots, h_8)$; each augmented model retains all base features and adds one proposal. The randomized negative control draws one unit vector over $(h_{10}, \ldots, h_{20})$ per repetition and holds it fixed across all three samples. Features are standardized using training statistics only, the ridge penalty is 10, and test gain is base MSE minus augmented MSE. On calibration data, y and z are centered and separately residualized on centered U by least squares. We estimate $B = \|y_\perp\|^2/\|y\|^2$, scale-normalized $E = \|z_\perp\|^2/\|z\|^2$, and $|A|$. Reported 95% intervals are mean ± 1.96 Monte Carlo standard errors across 300 repetitions. The fixed seed is 20270730. The executable source is `experiments/recompute_principle_validation.py`; its CSV and JSON outputs retain every displayed cell and configuration field.

B.3 Complete controlled summary

B.4 Public proposal-response archive

The archive reuses 24 frozen train/validation/test task splits from the public OpenML replay manifest. Eleven binary datasets (22 splits) form the primary analysis. Letter recognition (OpenML 6) has a 26-class target encoded numerically by the shared loader; its 240 candidate evaluations are excluded from primary inference and retained as a declared sensitivity artifact. Within each split, up to 10 raw features are chosen by train-only absolute Spearman association and standardized using

w	Proposal	B	E	$\lvert A\rvert$	Test gain [95% CI]
0.00	Independent	.195	.974	.046	$-.0010$ [$-.0011, -.0008$]
0.00	Misaligned	.195	.972	.048	$-.0008$ [$-.0010, -.0006$]
0.00	Aligned	.195	.973	.048	$-.0007$ [$-.0009, -.0006$]
0.25	Independent	.389	.974	.049	$-.0017$ [$-.0021, -.0014$]
0.25	Misaligned	.389	.973	.049	$-.0016$ [$-.0020, -.0013$]
0.25	Aligned	.389	.973	.707	$+.2561$ [$+.2539, +.2584$]
0.50	Independent	.585	.973	.050	$-.0020$ [$-.0024, -.0016$]
0.50	Misaligned	.585	.973	.045	$-.0023$ [$-.0029, -.0018$]
0.50	Aligned	.585	.974	.815	$+.5154$ [$+.5114, +.5193$]
0.75	Independent	.781	.972	.049	$-.0026$ [$-.0032, -.0020$]
0.75	Misaligned	.781	.972	.050	$-.0036$ [$-.0043, -.0029$]
0.75	Aligned	.781	.973	.866	$+.7679$ [$+.7630, +.7728$]
1.00	Independent	.972	.974	.047	$-.0044$ [$-.0052, -.0035$]
1.00	Misaligned	.972	.974	.044	$-.0044$ [$-.0053, -.0035$]
1.00	Aligned	.972	.973	.893	$+1.0246$ [$+1.0181, +1.0311$]

Table 1: Complete controlled validation (15 cells). Values are Monte Carlo means over 300 repetitions. All proposal families remain matched in expansion; only the aligned proposal acquires substantial residual alignment when $w > 0$. Small negative control gains are finite-sample estimation costs, not negative population gain.

training statistics. The current representation is their span. A deterministic generator then produces raw, square, absolute-value, pairwise-product, and absolute-difference responses. Training targets determine the represented raw features; neither validation gates nor test outcomes change proposal inclusion.

For every proposal, B, E, A are estimated on validation data. The proposal is standardized using training statistics, and the augmented fit minimizes mean squared loss with penalty .1 only on its new coefficient; represented coefficients remain unpenalized. Relative test gain is base minus augmented test MSE, divided by base test MSE. The score is $\widehat{B}\widehat{E}(\widehat{E}+.1)^{-1}\widehat{A}^2$. Candidate-level rank correlations and top-decile gains are computed within each task, averaged over the two splits inside dataset, and then averaged over the 11 datasets. Percentile intervals resample datasets 10,000 times. The paired top-decile contrast also uses the exact 2^{11} sign-flip distribution.

Classical baseline and additional paired audit. Partial R^2 is $\widehat{A}^2$ (the archived `alignment_score`). The historical column `partial_r2_score` stores $\widehat{B}\widehat{A}^2$, the unpenalized incremental-R^2 score. Because $\widehat{B} > 0$ is constant within each split, both produce identical rankings and top-decile selections in this archive; neither supplies an independent ranking test of B. We retain the historical files unchanged and verify this identity using round-trip float parsing. The additional post-hoc audit uses seed 20260910 and 10,000 dataset-bootstrap replicates. Three-gate minus partial-R^2 top-decile gain is $-.001434$ with interval $[-.011012, .005394]$ (exact paired sign-flip $p = .9531$). These are relative-gain units; multiplying by 100 gives percentage points. The script `recompute_classical_comparison.py` produces Figure 2 and hash-bound results without refitting models or accessing new data.

The serialized archive contains every train/validation/test response matrix, candidate ID, task identity, analysis-scope flag, and split-specific SHA-256 digest. It comprises 2,086 primary and 240 sensitivity candidate evaluations. The response manifest, candidate-level table, task- and dataset-level summaries, exact sign-flip result, and figure-generation path are included in the anonymous supplement. The archive is observational and transformation-specific; it demonstrates held-out discrimination, not randomized causal evidence or prospective controller value.

Dataset	ρ(gates)	$\rho(E)$	ρ(partial R^2)	Top-10% Δ
Banknote	.924	.234	.929	+.159
Blood transfusion	.597	.201	.588	+.006
EEG eye state	.008	$-.065$	.125	$-.015$
ILPD	.094	.000	.109	.000
Nomao	.973	.626	.962	+.035
Phoneme	.549	$-.068$	.551	+.030
QSAR biodegradation	.298	.036	.303	+.008
Steel plates	.422	.104	.424	+.117
Adult	.931	.694	.917	+.011
KR-vs-KP	.770	.412	.764	+.142
Credit-G	.127	.133	.130	.000
Mean	.518	.210	.528	+.045

Table 2: Dataset-level public-archive results. Correlations rank proposals by held-out relative MSE gain. "Top-10% Δ" is gate-score minus expansion-only mean gain at the same selection budget. Values average the two frozen splits within each dataset.

C Qualified Transfer Evidence

C.1 A-share-calibrated semi-synthetic sweep

The five-condition sweep is not a real-response experiment. It derives a proxy factor-correlation matrix from IC co-scoring across market-cap segments in one 3,638-factor A-share archive; pairs with fewer than three co-observations use IC similarity, and eigenvalue clipping makes the matrix positive semidefinite. Five dependent factor subsets determine the spectra. For each condition, 100 Monte Carlo repetitions generate 500 Gaussian observations from the calibrated spectrum and plant signal in accessible and inaccessible directions. Structured, external-style, and combined searches evaluate 40, 10, and $40+12$ proposals, respectively, with seed 42 and acceptance threshold .3. Table 3 reports every point entering the exploratory Spearman association. Random non-local proposals evaluate 20 directions and have zero accepted yield in all five conditions, so they are omitted. The association between r_{eff}/N and combined-minus-structured yield is $\rho = -.72$ for $n = 5$ dependent conditions. Subset construction changes both archive geometry and proposal reachability, so this is an A-share-calibrated semi-synthetic stress test, not identification of a compression effect in realized market responses.

Condition	N	r_{eff}/N	Structured (40)	External (10)	Combined (40+12) (Δ)
All factors	3,638	.003	25.2	5.0	31.2 ($+6.0$)
High correlation	500	.003	38.8	9.1	49.8 ($+11.0$)
Random 50%	1,819	.006	23.3	5.2	29.8 ($+6.5$)
Low correlation	500	.025	21.9	0.6	22.7 ($+0.8$)
Decorrelated	100	.086	23.2	2.6	26.4 ($+3.1$)

Table 3: Complete A-share-calibrated semi-synthetic compression sweep. Yield is the mean accepted count across 100 simulations; Δ is combined minus structured. Conditions are dependent subsets of one market-calibrated proxy geometry, and proposal budgets differ.

A second application covers 5,647 stocks from 2010–2026. It excludes ST/*ST stocks, histories shorter than 60 trading days, and IPOs within 30 days; discovery trains on 2010–2018, validates on 2019–2021, and evaluates on 2022–2026. Table 4 exposes the unequal proposal budgets rather than treating the union as a scheduler comparison. The union contains the accepted proposals from both streams and therefore establishes proposer compatibility only. Raw Tushare Pro observations are licensed and are not redistributed.

C.2 Symbolic regression

Our LLM-SRBench analysis is a cached evaluation on the 158 complete task identities common to two frozen runs: 72 synthetic equations and 86 adversarial transformations. The loader targeted the benchmark's 240 rows, but the retained run artifacts contain only these 158 executable tasks. An exclusion-by-reason manifest for the other 82 rows was not preserved, so this is neither a complete

Stream	Proposed	Accepted	Rate	Best ICIR	Breadth
External	100	9	9%	.52	9
Structured	55	10	18%	.42	10
Union	155	19	12%	.52	19

Table 4: Date-split A-share proposer-compatibility summary. Unequal proposal budgets preclude an adaptive-policy comparison.

benchmark evaluation nor a reproducible claim about the omission mechanism. It is a paper-specific rerun, not a result reported by the benchmark authors.

Data generation used seed 42, 500 training points and 100 test points per task, inputs sampled uniformly on $[-3, 3]$ with a positivity heuristic for logarithms and square roots, and training-target standardization. GP used PySR with 10 iterations, five populations of 30, a 30-second cap, maximum expression size 25, and the listed arithmetic/trigonometric operators. The configured LLM provider identifier was `deepseek-v4-flash`, with temperature .3, 4,096 output tokens, three retries, and five candidates. The later hybrid rerun gives GP the same full budget, then adds up to five residual-directed candidates when training $R^2 < .99$. Each cached table contains one result per task and method; no repeated-run uncertainty is available.

GP and LLM-only rows come from `results_20260529_165504.csv` (SHA-256 `75a8c0bf...22e15c4`); the later full-budget hybrid rows come from `results_20260529_174956.csv` (SHA-256 `65952584...96f`). This run selection was made to compare the earlier baselines with the later hybrid implementation that restored the full GP budget. The reconstruction script verifies both hashes and requires exactly 158 complete common task identities. Combined search improves 25 of the 46 transformed-equation cases in which GP has negative R^2, but only five reach nonnegative R^2. Fresh non-fallback loading may require external dataset access; fresh proposal runs require provider credentials and may not reproduce a mutable provider response. Complete historical proposal spectra were not recovered.

		GP		LLM		Hybrid	
Split	n	Median	$R^2 \geq .5$	Median	$R^2 \geq .5$	Median	$R^2 \geq .5$
All	158	.875	.639	.120	.386	.838	.633
Synthetic	72	.980	.958	.757	.708	.984	.986
Transform	86	$-.011$	.372	-1.253	.116	$-.065$	.337

Table 5: Complete aggregate reconstruction of the specified LLM-SRBench subset. GP retains the better overall and transformed-equation medians. The supplement reports all 158 task identities and method scores.

Of the 46 transformed tasks with negative GP R^2, hybrid has higher R^2 on 25 and reaches nonnegative R^2 on only five. The anonymous supplement includes both hash-pinned source tables and the complete 158-row merge. These results support failure improvement and compatibility, not general dominance.

D Frozen Replay and Attribution Audit

D.1 Environment and estimand

We write a discovery environment as

$$\mathcal{E} = (G, V, U, c, R, \mu_0), \tag{16}$$

where G supplies proposals, V verifies them, U updates state, c charges generation and verification, R scores a trajectory, and μ_0 specifies initial conditions. Standard trajectory value is

$$J(\pi; \mathcal{E}) = \mathbb{E}_{H \sim (\pi, \mathcal{E})}[R(H)]. \tag{17}$$

A declared policy contrast compares $J(\pi_1; \mathcal{E}) - J(\pi_2; \mathcal{E})$. Holding $\mathcal{E}$ fixed is necessary for this estimand but does not by itself guarantee unbiased causal attribution.

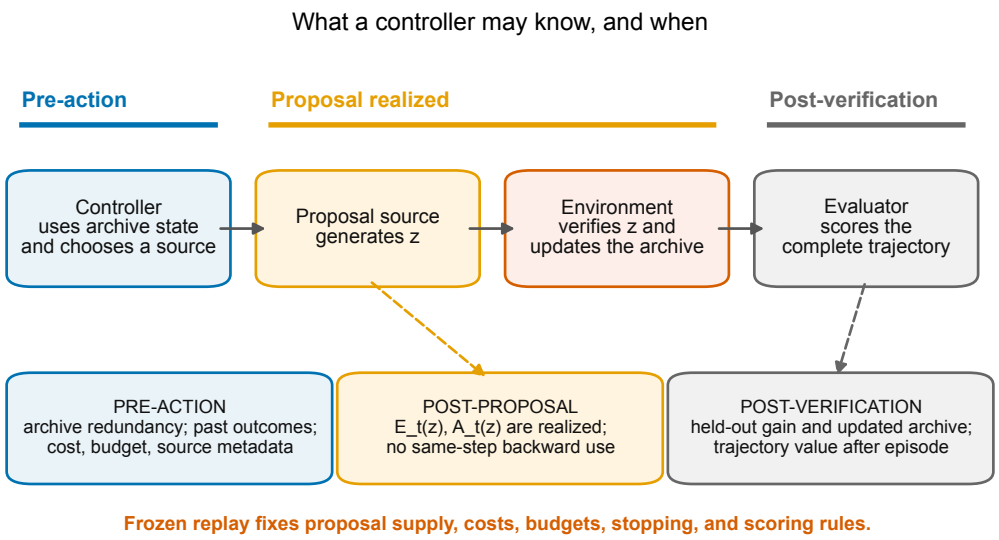


Figure 3: **Observability boundary.** Realized $E_t(z)$ and $A_t(z)$ may update the next archive state but cannot justify the action that produced z. Frozen replay fixes proposal supply, costs, budgets, stopping, and scoring while allowing policy-specific archives to diverge. The figure defines when quantities become available; replay audits one policy bridge but does not establish the boundary itself.

D.2 Replay design and observability

The 12 OpenML dataset IDs are 31, 1464, 1462, 1480, 1489, 1494, 1504, 1590, 3, 6, 1471, and 1486. Their names, sample counts, selected-feature counts, and two terminal split seeds are given in the supplement's `public_task_manifest.csv`. Split seeds are 20260610–20260621 and 20260723–20260734 in the same dataset order. Proposal streams and source-level test scores were generated before the analysis freeze. Policies cannot access terminal outcomes until reporting; this tests policy-time outcome hiding, not investigator-blind prospective confirmation.

Each of the 24 dataset–split tasks is replayed under the original proposal order plus 30 fixed permutations, budgets 8, 12, and 16, and provider-call weights 0, 1, 4, and 8. Verification cost is one; a provider call returns eight candidates, so the per-candidate non-local cost is $1 + w_c/8$. The Cartesian design has $24 \times 31 \times 3 \times 4 \times 14 = 124{,}992$ nested trajectories. The 14 registered policies are local-only, jump-only, fixed 25/75, fixed 50/50, fixed 75/25, random gate, recent yield, cost-aware UCB, admission sequential, four CAHS ablations (no rank, no proposal-value/distance, no failures, and no stop), and full CAHS.

For accumulated costs $0 = c_0 \le \cdots \le c_K$ and best verified utilities u_k, the raw trajectory endpoint is trapezoidal area divided by initial budget C,

$$\mathrm{AUC}_C = \frac{1}{C} \sum_{k=1}^{K} (c_k - c_{k-1}) \frac{u_{k-1} + u_k}{2}. \tag{18}$$

It is then normalized by the frozen dataset–split oracle maximum. Leave-one-dataset-out calibration uses the other 11 datasets to select each policy configuration and the strongest eligible non-geometric baseline; fixed 25/75 is selected in all 12 folds. The registered paired comparisons average repeated splits, orders, budgets, and costs inside dataset before 10,000-replicate dataset-cluster bootstrap inference. The policy grid, selected-policy hashes, exact endpoint code, and all envelopes are included in the supplement.

Full CAHS selects between local and non-local actions from cost-normalized UCB scores. Its non-local score adds lagged archive compression, recent realized escape, and recent predictive-novelty terms and subtracts non-local cost. These inputs summarize earlier evaluated proposals; no expansion,

alignment, or terminal value from the proposal being selected is exposed to the same decision. The policy is retained as a stress test, not proposed as a successful controller.

D.3 Registered direction and bootstrap audit

The frozen H1a regression uses trajectory-average `effective_rank_ratio` as predictor and normalized terminal utility–cost AUC as outcome, controlling for archive size, recent yield, proposal family, and total cost. The preregistration's machine rule requires a negative slope and same-direction clustered interval. Because lower effective-rank ratio denotes greater redundancy, that rule makes greater compression correspond to higher AUC; it is directionally inconsistent with the contemporaneous verbal hypothesis that greater compression predicts lower utility. We therefore retain the frozen machine-labeled status while declining to interpret it as support for the intended diminishing-return bridge.

The first exported comparison artifact used 1,000 rather than the registered 10,000 cluster-bootstrap replicates. It remains in the supplement as an immutable audit artifact: H1a was $-.3423$ with interval $[-.6068, -.0516]$. The corrected registered rerun uses the same estimator and 10,000 replicates, yielding $[-.5988, -.0300]$; point estimates and qualitative conclusions are unchanged. The post-hoc pre-action analysis is reported separately and does not amend H1a.

Question or contrast	Estimate	95% CI
Registered mean rank → trajectory AUC	-0.3423	$[-0.5988, -0.0300]$
Pre-action rank → next marginal utility	$+0.0017$	$[-0.0013, +0.0046]$
Distance → utility per cost	-0.4249	$[-0.6518, -0.2924]$
Held-out value, controlling distance	$+0.6490$	$[+0.1033, +1.4603]$
CAHS minus no-rank ablation	$+0.0009$	$[-0.0009, +0.0027]$
CAHS minus fixed 25/75	-0.0135	$[-0.0212, -0.0057]$

Table 6: Frozen replay. The registered rank endpoint is trajectory-level; the temporally aligned next-utility audit is post hoc. Policy differences are normalized utility–cost AUC with 10,000-replicate dataset-cluster inference.

The design contains 124,992 nested trajectories, but the independent unit is the dataset ($n = 12$). Full CAHS never stops early in 8,928 eligible matched episodes. The -0.0135 comparison in Table 6 is the registered LODO policy contrast averaged over all three budgets and four shared cost environments. The separate -0.0211 estimate (95% CI $[-0.0282, -0.0135]$) fixes every policy to the primary provider-weight-four environment $\mathcal{E}_4$ for a post-hoc accounting audit. Thus the two estimates have different, explicitly declared estimands and are not conflicting replications. Granting only CAHS free provider access while retaining primary cost for fixed 25/75 creates the invalid mixed-environment contrast

$$\widetilde{\Delta} = J(\text{CAHS}; \mathcal{E}_0) - J(\pi_{25/75}; \mathcal{E}_4) = +0.0401 \quad (95\% \text{ CI } [+.0311, +.0497]). \tag{19}$$

Across 182 ordered one-policy cost privileges, 75 reverse the winner. This audit demonstrates accounting sensitivity and is outside the registered policy-comparison family.

Quantity	Pre-action?	Permitted role
Archive effective rank	Yes	Descriptive state
Proposal expansion	No	Realized mechanism diagnostic
Residual alignment proxy	No	Realized mechanism diagnostic
Held-out incremental value	No	Verification outcome
Trajectory value	After episode	Policy comparison

Table 7: Observability contract. A prospective controller would need to predict—or otherwise account for—proposal-specific expansion, alignment, and verification value before they become observable.

E MANDATORY STATEMENTS

E.1 DATA AVAILABILITY

The anonymous supplement includes the synthetic generator and all cells, exact-objective and continuous diagnostics, the 2,326-response OpenML archive with split hashes, both hash-pinned LLM-SRBench sources and their 158-task reconstruction, fixed replay configuration, anonymous trajectory envelopes and decision states, offline reproduction scripts, and a SHA-256 manifest. LLM-SRBench and OpenML remain publicly accessible through their cited sources. Raw Tushare Pro market observations are licensed and cannot be redistributed; the supplement contains permissible derived summaries and documents the no-fallback boundary.

E.2 AUTHOR CONTRIBUTIONS

All authors contributed to conceptualization, methodology, investigation, validation, writing, and review. Specific CRediT assignments will be disclosed in the non-anonymous version.

E.3 CONFLICT OF INTEREST

The authors declare no known conflict of interest affecting this work.

E.4 FUNDING

Funding information is withheld during double-blind review and will be completed in the camera-ready record.