DiSCO-SR: Residual-Guided Symbolic Selection under Predictive-Risk Constraints

Anonymous authors
Paper under double-blind review

Abstract

Predictively similar symbolic equations can imply different intervention responses. DiSCO-SR tests residual-guided equation selection within an explicit predictive-risk budget, separating equation choice from causal direction decisions. Separate data splits support fitting, selection, and direction decisions, with no refitting of the selected equations. We derive conditional guarantees for population risk and penalized finite-block residual dependence, and bound mean-intervention loss under correct orientation and bounded distribution shift. In independent confirmation on 1,920 synthetic cases from 240 mechanism clusters, the development-selected paired-stability rule achieves 95.21% direction accuracy and 0.1396 mean-intervention NRMSE, compared with 95.31% and 0.1385 for minimum-risk selection with the same direction rule. None of eight prespecified primary contrasts is significant after multiplicity correction. The stability rule changes 29.97% of directional equations, correcting four directions and corrupting six. Its certificate counterpart always returns the minimum-risk reference across 2,916 confirmation cases. Matched regression controls attain stronger point estimates, and a separate bounded-score study exhibits large errors on measured physical interventions. These results provide an auditable framework for constrained symbolic selection, but do not establish an incremental mechanistic benefit from residual guidance.

Keywords: Symbolic regression, causal direction, intervention prediction.

1 Introduction

An equation can predict observations accurately yet give the wrong response to an intervention. Symbolic regression (SR) makes this distinction tangible: both $Y = f(X)$ and $X = g(Y)$ can explain a scatterplot, but their executable mechanisms imply different responses when a variable is externally set. Under a causally sufficient additive-noise model (ANM), $Y = f^\star(X) + N$ with $N \perp\!\!\!\perp X$, a suitable forward regression admits independent residuals. The reverse generally does not, subject to identifiability exceptions (Hoyer et al., 2009, Peters et al., 2014). Predictive selection alone does not explicitly exploit this asymmetry.

We ask a narrower question than discovering a new causal asymmetry: *can residual evidence choose a useful directed equation among fitted symbolic candidates while controlling predictive loss?* A clean answer must separate equation selection from direction selection. A better direction rule can improve an entire pipeline without validating its equation selector. Likewise, lower residual dependence on selection observations is an optimization outcome, not independent evidence of a correct mechanism.

DiSCO-SR makes these distinctions executable. Disjoint samples fit the finite library, select equations under an observed-risk cap, and decide direction from residual evidence with a regression-error fallback. Independent queries evaluate the returned coefficients without refitting. The risk safeguard holds on selection observations. Its population interpretation needs additional assumptions, and neither version certifies causal truth or extrapolation.

We report three completed, separately frozen campaigns: the original expanded confirmation, a bounded-score study, and an independent paired-evidence confirmation. Their observations are not pooled. The paired-evidence study tests whether a selector should change the minimum-risk equation only when the estimated residual benefit is stable. Development freezes the rule before new confirmation. Every declared comparator remains in the report.

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.

The findings sharpen the contribution. Risk constraints provide an auditable tradeoff, but independently tested selection changes do not establish an incremental accuracy or intervention advantage. Matching the direction rule also removes the apparent advantage over strong regression representations. Our contribution is the constrained inference interface, conditional analysis of when it can help, and matched comparisons that separate equation selection from direction decisions. This distinction matters when a fitted equation is presented as a scientific mechanism rather than a predictor.

2 Positioning and Related Work

ANM and RESIT use residual independence for causal direction under appropriate assumptions (Hoyer et al., 2009, Peters et al., 2014). Least-squares independence regression already optimizes dependence within regression and performs model selection (Yamada & Sugiyama, 2010). Dependence-guided regression is therefore not new. Our focus is an explicit predictive-risk feasible set, a fixed-coefficient symbolic interface, and separate attribution to equation and direction. IGCI and RECI exploit other distributional asymmetries (Mooij et al., 2016, Bloebaum et al., 2018). Simple direction rules are essential controls, not dispensable competitors.

PySR, AI-Feynman, deep symbolic regression and sparse identification use different grammars and searches for equation discovery (Cranmer, 2023, Udrescu & Tegmark, 2020, Petersen et al., 2021, Brunton et al., 2016). Our primary backend selects sparse finite-basis models. Separate gplearn and PySR experiments compare selectors on shared fitted candidate sets. They test interface portability, not native solver superiority. MIIC-SR connects symbolic mechanisms to a graph. The concurrent SRIT preprint tests independence around genetic-programming expressions (Sella et al., 2025, Yan et al., 2026). CAM and nonlinear NOTEARS instead target multivariate graphs (Bühlmann et al., 2014, Zheng et al., 2020). Our primary estimand is a directed two-variable mechanism, not a graph-level guarantee.

3 Method and Testable Analysis

3.1 Problem, candidate class, and objective

Given i.i.d. observational pairs (X, Y), the output is a direction and one executable equation. A causal interpretation requires a two-node, causally sufficient ANM, an adequate retained representation, and directional identifiability. The algorithm cannot verify these assumptions from a score gap. Its computational task is finite-candidate selection under these assumptions, not discovery under arbitrary confounding or measurement error.

The original selector is the constrained rule in (4), followed by the gate below. For comparison, the matched fit- and joint-score baselines score the residual $r_f = Y - f(X)$ as follows:

$$F(f) = \widehat{R}(f) + \alpha C(f), \qquad J(f) = F(f) + \beta \widehat{D}(r_f, X), \qquad \widehat{R}(f) = n^{-1} \sum_i r_{f,i}^2. \tag{1}$$

Means and standard deviations are fitted on training rows only, so risk uses training-standardized target coordinates. The reference-score weights are $\alpha = 10^{-3}$ and $\beta = 0.35$. Risk preserves predictive adequacy. Complexity expresses a preference for a compact description. Dependence penalizes residual structure incompatible with the proposed ANM. These are design motivations, not a claim that all three terms improve every endpoint.

The primary dictionary for standardized input u is

$$\{1, u, u^2, u^3, \sin u, \tanh u, \log(1 + |u|), \sqrt{|u| + 10^{-12}}\}. \tag{2}$$

Training correlation screens terms. Candidate subsets contain singletons, ranked prefixes, pairs among the top six terms, and the full small dictionary. Each subset receives a ridge fit with penalty 10^{-6} and an unpenalized intercept. Independence changes the choice among these fits, not the continuous coefficient optimizer. Complexity sums active operator costs (1 for identity, 2 for square, 3 for other unary terms) plus $\max(1, \text{active nonconstant terms})$. Activity uses coefficient magnitude $> 10^{-4}$ in the standardized basis. This is a description-length proxy, not a validated measure of human interpretability.

The confirmed library retains every original candidate and adds analytic paths over powers 1–3 and shifted/scaled sine, cosine, tanh, log-absolute, and square-root-absolute atoms. Supports come from two training-residual greedy paths and single-atom alternatives. Each is fitted with ridge $10^{-3}, 10^{-2}, 10^{-1}$. These generic atoms do not include a complete test generator expression. Transformations, supports, and coefficients are frozen from fitting rows and exported at full precision.

Dependence is normalized HSIC with Gaussian kernels and median nonzero pairwise-distance bandwidths (Gretton et al., 2005, 2007):

$$\widehat{D} = \frac{\operatorname{tr}(K_c L_c)}{\sqrt{\operatorname{tr}(K_c K_c)\operatorname{tr}(L_c L_c)} + (n-1)^2 10^{-12}}. \tag{3}$$

K_c, L_c are centered Gram matrices on the rows being scored. For candidate j, $\widehat{D}_j$ evaluates this statistic. It is not an independence-test p value. Its magnitude depends on sample size, representation, and kernel choice. Distance correlation is an alternative in the frozen sensitivity grid (Székely et al., 2007).

For candidate j, let ℓ_{ji} be its squared error on selection row i, $\widehat{R}_j$ its empirical mean, and $b = \arg\min_j \widehat{R}_j$. With n_s selection rows, set $s_j = \operatorname{sd}_i(\ell_{ji} - \ell_{bi})/\sqrt{n_s}$ and write $C_j = C(f_j)$. The risk guard is

$$\mathcal{A} = \{j : \widehat{R}_j \le \widehat{R}_b + \min(\kappa s_j, \rho \widehat{R}_b)\}, \qquad \widehat{j} = \arg\min_{j \in \mathcal{A}} \{\widehat{D}_j + 10^{-3} C_j / n_s\}. \tag{4}$$

The development-locked values are $\kappa = 0.5$ and $\rho = 0.10$.

Proposition 1 (Observed-risk safeguard)**.** *$\mathcal{A}$ is nonempty and the returned candidate satisfies $\widehat{R}_{\widehat{j}} \le (1+\rho)\widehat{R}_b$ on the selection rows.*

Candidate b belongs to $\mathcal{A}$ because $s_b = 0$. Membership of $\widehat{j}$ and the relative cap prove the inequality. This directly supports predictive-risk control in the implemented selection step, but it is not a population confidence bound, an OOD guarantee, or a causal certificate. Theorem 1 transfers the cap under a simultaneous risk-estimation event for this fixed fitted library. Theorems 2 and 3 explain why multiple near-optimal candidates are needed for residual selection to matter. The collapse result fixes the fitted library, rather than a growing search.

3.2 Three information stages, two independent design choices

One random partition assigns 50% of rows to fitting, 25% to selection, and 25% to evaluation, identically in both directions. Fitting learns transformations, paths, and coefficients. Selection applies Equation 4 separately in each direction. Evaluation performs 199 residual permutations on at most 160 rows per direction. It uses residual orientation only when one p value is at least 0.10 and the other at most 0.005. Otherwise it uses a local cubic-ridge RECI rule fitted and scored on the 75% development rows. Nonrejection is not proof of independence. Gaps of magnitude at most 10^{-12} are ambiguous, and no stage receives a label or intervention query. Development uses synthetic truth for hyperparameter selection. The label-free claim applies to inference on each new pair.

The matched control replaces guard selection by minimum selection risk while keeping the expanded pool and the same direction gate. Additional ablations replace the gate by RECI or dependence, remove ridge paths or expanded atoms, or restore unconstrained J selection. The older JJ/FJ/FF factorial remains as a baseline: its letters denote joint or fit selection followed by joint or fit decision. Shared variants reuse exact candidates and split indices. Evaluation rows decide direction, so their winning statistic is not an unbiased post-selection estimate. Fresh query samples evaluate the returned coefficients. No full-data or test-dependent refit is performed.

Executable procedure. Fit both directional libraries, select one equation per direction, then apply the residual gate or RECI fallback. Return the chosen direction with unchanged coefficients and fitted root mean. The decision quarter is an algorithm input. Only independent queries measure performance.

3.3 POPULATION INTERPRETATION AND INTERVENTION CONSEQUENCES

Conditional on the fitting sample, suppose a simultaneous event gives $\max_j |\widehat{R}_j - R_j| \le \epsilon$ for the fixed fitted candidates. The observed cap then implies

$$R_{\widehat{j}} \le (1+\rho)\min_j R_j + (2+\rho)\epsilon. \tag{5}$$

Here R_j is population prediction risk, and $\widehat{R}_j$ is its selection-sample estimate. Theorem 1 proves the result and gives a bounded-loss specialization. Gaussian noise and polynomial predictions do not themselves satisfy that specialization. The conditional dependence oracle bound in Theorem 2 requires a simultaneous estimation event for all fitted candidates. It does not assert one for adaptive-bandwidth HSIC. Theorem 3 shows that, for a fixed finite library with a unique separated risk minimizer and finite loss variance, the standard-error guard eventually contains only that minimizer. These statements identify conditions for useful finite-sample choice rather than proving causal truth.

Proposition 2 (Wrong-direction intervention floor)**.** *Let $Y = f^\star(X) + N$, $N \perp\!\!\!\perp X$, $\mathbb{E}N = \mathbb{E}X = 0$, with finite second moments and no confounding. For positive normalizers s_f^2, s_X^2 and a cause-query distribution Q, any reversed two-node ANM has squared mean-intervention error $L^- = E_{\mathrm{do}}^2$ satisfying*

$$L^- \ge \frac{\mathrm{Var}_Q[f^\star(X)]}{2s_f^2}. \tag{6}$$

A forward model has $L^+ = \frac{1}{2}\{\mathbb{E}_Q[(\widehat{f} - f^\star)^2]/s_f^2 + (\widehat{\mu}_X - 0)^2/s_X^2\}$.

Under the reversed model, $do(X)$ leaves the predicted mean of root Y constant. Its squared approximation error is at least the query variance of $f^\star$. The other intervention error is nonnegative. The result also holds for the empirical query grid. It explains the potential cost of a wrong direction, not that the guard avoids it. Full proofs and a symmetric direction–equation attribution identity appear in Appendix A.3. These conditional analyses do not certify causal validity for an unbounded residual score or guarantee extrapolation accuracy.

3.4 PAIRED EVIDENCE WITH A MINIMUM-RISK FALLBACK

The follow-up uses the same fitted library and direction gate. Training fixes kernel widths. Selection rows form disjoint blocks with bounded dependence statistics $Z_{jt} \in [0,1]$. Let b minimize selection MSE and retain $\mathcal{A}_\rho = \{j : \widehat{R}_j \le (1+\rho)\widehat{R}_b\}$, with $\rho = .05$. For $c_j = 10^{-3}C_j/n_s$, the paired gain is the block average of $Z_{bt} - Z_{jt}$ plus $c_b - c_j$. The stability rule requires positive gain in at least a fraction τ of 399 shared block resamples. Otherwise it returns b. Development selects block size $q = 8$ and $\tau = .90$. This resampling frequency measures decision stability conditional on the observed blocks. It is not a confidence level.

A separate certificate rule switches only when a simultaneous empirical Bernstein lower bound on that gain is positive. It accounts for every ordered candidate pair in both directions, since the risk reference is adaptive. Theorem 5 guarantees improvement in expected penalized finite-block dependence conditional on fitting and i.i.d. selection blocks. It does not guarantee direction, intervention, or OOD improvement. Both rules retain their observed-risk cap. The certificate can be uninformative and return the risk minimizer on every task. Appendix A.2 gives the exact radii, tie rules, proof and power limitation. Theorem 7 additionally bounds mean-intervention loss under correct orientation and bounded distribution shift. Theorem 6 bounds finite-resampling decision error conditional on the observed blocks. It is not a causal confidence statement.

4 EXPERIMENTAL QUESTIONS AND PROTOCOL

Separate campaigns and statistical units. The original expanded campaign (V4) contains 384 development and 5,782 confirmation cases. Its 1,920 primary cases form 240 mechanism clusters. Its selected guard uses $\kappa = .5$ and $\rho = .10$. The bounded-score campaign (V5) contains 384 development, 400 calibration and 7,483 confirmation cases, including physical interventions. The focused paired-evidence campaign (V6) contains 384 development and 2,916 confirmation cases. All

are completed. Appendices A.1 and B.1 give their separate frozen designs. Earlier campaigns inform later designs, so only new within-campaign confirmation supports their prespecified comparisons.

In V6, nine development candidates compete by intervention NRMSE, subject to accuracy at least that of Risk+gate. Exact ties favor Risk+gate and then name. This selects the stability rule $(q, \tau) = (8, .90)$ before confirmation. Primary confirmation crosses 12 mechanism forms, 20 fresh parameter seeds, two centered cause laws, $n = 200/500$ and SNR 2/10: 1,920 cases in 240 clusters. Another 480 cases test transfer, 480 test six new forms, and 36 test certificate power at $n = 1,000/4,000$. New seeds separate stages. Previously studied mechanism families are not independent real-world domains. All 3,300 scientific cases and 58,500 method evaluations are included, with no failed or selectively omitted scientific cases.

Controls and fair attribution. Risk+gate is our minimum-selection-risk ablation, not a separately published method. Simple+gate and Joint+gate alter selection on the same candidates. The original guard, BIC heuristic, bounded upper-confidence selector, certificate, four stability settings and Risk+RECI remain visible. Seven regression representations—polynomial, spline, kernel ridge, random forest, extra trees, Gaussian process and SVR—receive the same residual gate and RECI fallback. This corrects the original comparison's direction-rule mismatch. Representation and fitting budgets still differ, so cross-model contrasts do not isolate a symbolic-selection effect. Native LSIR and native-budget solver superiority are not established.

Independent prediction and intervention queries. Oracle-direction ID error evaluates the true-direction equation on 512 fresh cause observations, normalized by their true-response standard deviation s_f. OOD queries extend beyond independent 1%/99% cause quantiles by $(s-1)$ times that interval's width, at $s = 1.25, 1.5, 2$. The same ID normalizer is retained. The OOD-1.5 endpoint is reported throughout. Oracle metrics do not depend on estimated direction. End-to-end error instead evaluates both intervention directions on $n_Q = 128$ central query points:

$$E_{\mathrm{do}} = \left[\frac{1}{2n_Q} \sum_{i=1}^{n_Q} \left\{ \frac{(\widehat{\mu}_{Y|do(x_i)} - f^\star(x_i))^2}{s_f^2} + \frac{(\widehat{\mu}_{X|do(y_i)} - 0)^2}{s_X^2} \right\} \right]^{1/2}. \tag{7}$$

These queries span 2–98% quantiles, s_X^2 is observed cause variance, and the centered cause has known mean zero. The model's root mean is estimated from fitting rows. Queries never select or refit coefficients. This synthetic endpoint concerns mean responses, not full interventional distributions. Physical responses use separately defined measured endpoints rather than substituting synthetic ground truth.

Inference and completeness. V6 compares the development-selected rule with four controls (risk, simple, joint and original guard) on accuracy and E_{do}. These eight prespecified two-sided cluster sign-flip tests receive Holm correction. Sign-flips require symmetric, exchangeable cluster effects. They are not treatment-randomization tests. The 20,000-draw cluster-bootstrap intervals are marginal and are not adjusted across the eight comparisons. Repeated conditions within a mechanism seed are paired, not independent replications. OOD, regression and transfer comparisons are descriptive. V4 and V5 retain their own twelve-test families. No favorable post-hoc subset or average rank redefines the principal claim.

5 Results: Selection, Direction and Generalization

5.1 Independent paired-evidence confirmation

Table 1 compares all 19 methods across three suites. The development-selected stability rule obtains 95.21% accuracy and 0.1396 intervention error. Risk+gate obtains 95.31% and 0.1385. Their accuracy difference is -0.104 percentage points, with marginal 95% cluster interval $[-0.417, 0.208]$. The intervention-error reduction is -0.00119, with interval $[-0.00407, 0.00164]$. Both adjusted p values equal 1.00. The development preference therefore does not reproduce as a resolved confirmation improvement. Failure to reject a difference does not establish equivalence, and we do not replace the selected configuration after seeing the confirmation table.

Table 1: Complete V6 comparison: 19 methods across all three principal suites (2,880 cases). Each group reports accuracy (%, higher is better), intervention NRMSE and OOD-1.5 (lower is better). All regression representations use the common gate. The selected rule was fixed using development only.

	Primary (1,920)			Transfer (480)			New forms (480)		
Method	Acc.	E_{do}	OOD	Acc.	E_{do}	OOD	Acc.	E_{do}	OOD
Risk+gate	95.31	0.138	3.13	88.33	0.206	2.38	90.42	0.193	3.86
Simple+gate	95.00	0.144	2.58	88.75	0.204	1.78	89.79	0.203	3.74
Joint+gate	95.26	0.142	2.21	88.75	0.202	1.64	89.38	0.207	3.32
Original guard	95.10	0.141	3.10	89.17	0.197	2.25	90.00	0.196	3.93
Block-UCB	95.05	0.145	3.09	88.33	0.209	2.34	89.58	0.204	3.64
Certificate	95.31	0.138	3.13	88.33	0.206	2.38	90.42	0.193	3.86
Risk+RECI	94.32	0.146	3.13	88.12	0.210	2.38	87.71	0.215	3.86
BIC+gate	94.74	0.153	1.83	88.96	0.208	1.55	88.75	0.222	2.22
Stability (4, 0.9)	95.26	0.139	3.11	88.75	0.203	2.47	89.79	0.199	3.97
Stability (4, 0.95)	95.31	0.139	3.15	88.54	0.204	2.48	90.00	0.197	4.03
Stability (8, 0.9) selected	95.21	0.140	2.96	88.33	0.206	2.26	89.58	0.202	3.98
Stability (8, 0.95)	95.31	0.139	3.00	88.54	0.204	2.23	89.79	0.199	4.03
Polynomial	95.00	0.145	215.42	88.75	0.202	193.68	89.79	0.209	421.87
Spline	95.26	0.142	2.91	88.96	0.205	2.22	91.46	0.177	1.86
Kernel ridge	95.47	0.136	3.08	88.96	0.198	2.01	90.21	0.185	1.78
Random forest	94.90	0.198	3.21	87.71	0.257	1.99	90.42	0.232	1.40
Extra trees	93.28	0.218	3.28	86.88	0.276	2.03	86.88	0.275	1.40
Gaussian process	95.52	0.133	2.60	88.75	0.197	1.47	90.83	0.177	1.29
SVR	95.00	0.130	2.92	88.96	0.189	1.91	89.17	0.184	1.46

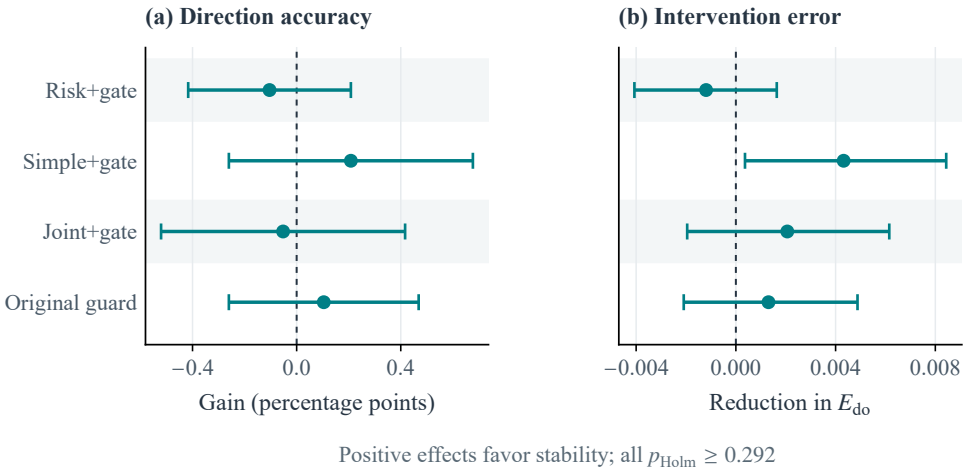

Figure 1: All eight prospective V6 primary comparisons of the selected stability rule against matched controls. Positive effects favor stability. Bars are marginal 95% cluster-bootstrap intervals. No contrast has a Holm-adjusted $p < .05$, including the Simple+gate error contrast whose marginal interval excludes zero.

Against Simple+gate, the intervention-error reduction is 0.00432 with a marginal interval from 0.00037 to 0.00844, but adjusted $p = .292$. The other seven adjusted p values equal 1.00. All eight comparisons are reported in Appendix B.2. Stability improves mean OOD error descriptively from 3.132 to 2.956 relative to Risk+gate, while Joint+gate achieves 2.209. OOD was secondary in this campaign, so this smaller mean is not a prespecified primary success. The observed-risk constraint limits selection MSE, not extrapolation.

Figure 1 separates the two endpoints rather than combining their ranks. An accuracy change counts directions, whereas intervention error also reflects equation quality and root-mean estimation. The

Table 2: Secondary V6 suites, reported separately: 480 transfer, 480 new-form and 36 larger-sample cases. Certificate results equal Risk+gate exactly.

Suite	Method	Acc. (%)	E_{do}	OOD-1.5
Transfer	Risk+gate	88.33	0.2061	2.376
Transfer	Selected stability	88.33	0.2056	2.259
New forms	Risk+gate	90.42	0.1935	3.862
New forms	Selected stability	89.58	0.2015	3.976
Large sample	Risk+gate	91.67	0.1173	2.138
Large sample	Selected stability	91.67	0.1164	2.036

control whose point estimate is easiest to exceed need not be the strongest comparator. Keeping the four-control family fixed avoids turning the most favorable of eight uncertain effects into the headline result.

Matching direction rules changes the cross-model interpretation. GP+gate has the highest primary accuracy, 95.52%, and SVR+gate the smallest mean intervention error, 0.1301. Kernel ridge also reaches 95.47% and 0.1357. These are point estimates, not adjusted superiority tests against the selected rule. Nevertheless, they prevent attributing the earlier pipeline advantage over residual-direction regression to symbolic representation alone.

5.2 Activity, fallback and transfer

The selected stability rule switches 29.97% of the 3,840 primary directional equations. Its certificate counterpart makes no switch in any of the 2,916 confirmation cases and exactly matches Risk+gate's direction and intervention errors. At $n = 200/500$, all recorded certificate radii exceed one. The simultaneous bound cannot resolve the available paired signal. Even the 36 larger-sample cases yield no accepted switch. A mathematically valid bound can therefore be operationally conservative. Using the heuristic does not inherit the certificate's guarantee.

Table 2 separates the secondary suites. Transfer accuracy ties Risk+gate at 88.33%, with similar intervention error. On new forms, stability decreases accuracy from 90.42% to 89.58% and increases intervention error from 0.1935 to 0.2015. The 36 larger-sample cases show similar accuracy and small descriptive error differences, but are too narrow to support a broad power or robustness claim. The appendix reports all configurations, including rules not selected on development.

Library error versus selection error. The fresh-query audit evaluates every fitted candidate in the true direction only after inference has finished. Its minimum ID error is an unattainable oracle, not an input to selection. The primary average oracle error is 0.08462. Average excess selected error is 0.04428 for stability and 0.04431 for Risk+gate. These are paired averages within each fitted pool, not differences between independently optimized libraries. Frequent changes therefore leave average ID selection regret almost unchanged. Relative to Risk+gate, stability corrects four primary directions and corrupts six. Switch activity, residual score and intervention benefit measure different properties. Appendix C shows four fixed fitted equations inside and outside the observation range.

5.3 Historical attribution and physical validation

Table 3 isolates component changes on the completed 480-case V5 ablation suite. Removing the confidence radius lowers intervention error from 0.133 to 0.129, while removing residual selection gives 0.125. Dependence-only direction raises it to 0.208. Removing the expanded library raises OOD error from 2.989 to 5.671 despite improved direction accuracy. Components affect different endpoints, so these descriptive contrasts do not support a claim that every component helps. Appendix D.2 retains all 48 paired variants. Table 13 reports the complete representation–selection–direction factorial.

In the original expanded primary experiment, DiSCO-SR achieved 95.42% accuracy and 0.1390 intervention error, versus Risk+gate's 95.52% and 0.1379. The guard changed at least one directional equation in 1,597 of 1,920 cases, but changed direction in only 16, correcting seven and corrupting

Table 3: Completed matched component removals on 480 V5 ablation cases. ΔE_{do} is variant minus reference. A positive value means larger intervention error after the change. These descriptive contrasts are not V6 primary tests.

Variant	Acc. (%)	E_{do}	ΔE_{do}	OOD
V5 bounded reference	96.25	0.1335	+0.0000	2.989
No residual selection	97.08	0.1254	-0.0081	2.817
No confidence radius	96.46	0.1292	-0.0042	2.951
No complexity cost	96.25	0.1334	-0.0000	3.011
No analytic ridge	96.25	0.1333	-0.0001	3.206
Base library only	97.08	0.1271	-0.0064	5.671
RECI direction only	95.42	0.1391	+0.0057	2.989
Dependence direction only	87.08	0.2080	+0.0745	2.989

nine. Against the same equations with RECI, the gate corrected 21 and corrupted eight. These are different contrasts: the former isolates selection, the latter decision. None of the twelve primary tests established a favorable adjusted effect. Joint+gate had significantly lower OOD error. Appendix D.3 reports the complete primary test family.

The completed bounded-score campaign selects $q = 16$, $\rho = .025$ on development. Across 2,880 primary cases it reaches 95.87% accuracy and 0.137 intervention error. Risk+gate reaches 95.80% and 0.136. Again no favorable primary contrast survives Holm adjustment. Joint+gate has lower OOD error by 0.552, adjusted $p = .0006$. On 612 Tuebingen protocols, benchmark-weighted accuracy is 65.61% versus Risk+gate's 66.03% and IGCI's 69.41%. Repeated splits of 102 pairs are not new data sources, and these observational pairs provide no intervention ground truth.

Physical interventions expose a stronger limitation. Across 11 actuator tasks, with 165 repeated observational protocols, actuator-averaged normalized regime-mean response error is 78.35 for the bounded selector, 29.66 for Risk+gate, 15.32 for Joint+gate, and 0.715 for GP+gate. Lower is better. The bounded selector's oracle-direction error remains 79.08, so direction mistakes alone cannot explain the failure. These large errors invalidate a claim of reliable physical intervention prediction. Reused measurements and one hardware system do not provide independent physical replications. The all-method summary is retained in Appendix D.1.

Calibration and certificate informativeness. The V5 task-mixture calibration is separate from the within-pair residual gate. It uses 400 calibration tasks to choose a threshold for accepting a direction prediction. The selected threshold 0.05 accepts 386 calibration tasks with 21 errors and a simultaneous upper risk bound of 0.09753. On 400 fresh tasks, it accepts 391 (97.75% coverage), of which 365 are correct: the observed accepted-task error is 6.65%. This is evidence about the declared synthetic task mixture under exchangeability, not a guarantee that a particular physical pair satisfies an ANM. The threshold is never retuned on these 400 test tasks.

The separate V5 certificate check reports zero definite violations across 240 directional fits, but the dependence interval always spans the full unit range and the mean clipped-risk interval width is 8.803. Thus zero observed violations do not demonstrate a useful tight certificate. Monte Carlo reference errors are included in these intervals. It is not exact population coverage. These observations complement the V6 no-switch result: valid uncertainty control and useful selection power are separate requirements. Shrinking the radius after seeing these outcomes would require a new justified procedure and independent confirmation.

5.4 Computational cost

Direct residual-kernel scoring costs $O(Mn_s^2)$ arithmetic per direction with quadratic Gram-matrix memory. The gate uses 199 permutations on at most 160 rows. Paired stability additionally performs 399 block resamples. Prediction evaluates only the retained equation. These costs are distinct from symbolic-search budgets and cannot be inferred from a shared-library wall clock as if each method ran separately.

The complete V6 development and confirmation stages took 5 minutes 7 seconds and 2 hours 58 minutes 2 seconds, respectively, including their process and reporting overheads. Each used one worker and one numerical thread within a two-CPU, 4-GiB allocation, without a GPU. Peak batch memory was 878 MiB. The V5 campaign comprises 8,267 cases and 171,075 method evaluations. V6 comprises 3,300 cases and 58,500 evaluations. These counts describe scheduled scientific cases, not independent replications. Primary V6 inference uses 240 mechanism clusters rather than 1,920 conditions or 3,840 directional fits as its statistical units. The development-selected rule is fixed before confirmation.

6 Discussion and Limitations

The strongest supported property is an explicit selection-risk cap. A simultaneous risk-estimation event transfers it to population risk. A standard-error allowance alone does not supply that event. A bounded paired certificate targets expected penalized block dependence, not causal accuracy. The empirical studies show why these distinctions matter: the heuristic is active without a resolved primary gain, whereas the certificate is valid under its assumptions but inactive on every confirmation task.

Three limitations constrain use. First, a finite fitted library can exclude the mechanism. Residual selection cannot repair missing representation or unstable extrapolation. Second, strong synthetic direction accuracy does not verify causal sufficiency, additive noise or identifiability. In the original study the gate never abstained on 160 nonidentifiable controls or 180 assumption-violating noise cases. Third, transfer within specified mechanism families and one hardware system cannot establish broad scientific validity. Operator complexity is a proxy, not measured human interpretability.

These results favor evaluating Risk+gate and matched regression gates before adding residual-selection machinery. They do not prove that no residual selector can help: the intervals permit modest effects, the libraries and noise regimes are finite, and different applications may value different risk budgets. A stronger empirical claim would require new independently designed domains, calibrated assumption checks, and physical responses that remain accurate outside the observational range. Such evidence must be obtained prospectively rather than by discarding controls or selecting a favorable confirmation setting.

Why a risk cap cannot ensure extrapolation. A simple population example makes this limitation explicit. Let $X \sim \mathrm{Unif}[-1, 1]$ and $Y = X + N$, with independent, zero-mean noise of variance $\sigma^2 > 0$. Compare fixed candidates $f(x) = x$ and $g(x) = x + \epsilon x^3$. Because $\mathbb{E}[X^6] = 1/7$,

$$R(f) = \sigma^2, \qquad R(g) = \sigma^2 + \epsilon^2/7, \qquad |g(a) - f(a)| = |\epsilon|\,|a|^3. \tag{8}$$

Choosing $0 < |\epsilon| \le \sqrt{7\rho}\,\sigma$ satisfies the population relative-risk cap, yet the disagreement grows without bound as the query $|a|$ increases. Both equations are compact polynomial expressions. This example concerns what the cap alone implies. It does not claim that our selector necessarily chooses g. A residual criterion can add information, but a guarantee about bounded block dependence still needs an additional argument connecting that target to accuracy outside the observed cause distribution. The hardware errors and the Joint+gate OOD comparisons expose this gap empirically rather than contradicting the stated risk theorem.

Theorem 7 requires correct orientation, simultaneous risk control and a bounded query-to-observation density ratio. Its noise-dependent allowance persists even if the true mechanism is in the library. Theorem 6 controls resampling of observed blocks. More draws cannot repair a misspecified mechanism or unsupported intervention queries.

7 Conclusion

DiSCO-SR provides an executable, risk-constrained interface between symbolic prediction and residual evidence. Its conditional guarantees describe specific statistical targets. Completed matched studies delimit their empirical consequences. Independent paired selection does not establish incremental primary benefit, and physical response failures prevent a broad intervention claim. In this setting, Risk+gate remains the reference. Residual selection should be preferred only after independent tests establish a benefit at the intended intervention queries.

DECLARATIONS

AI USE STATEMENT

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

Ethics. The study uses synthetic and public benchmark data without new human-subject data collection. Observational direction scores are not sufficient evidence for clinical, biological, or policy interventions. Confounded and nonidentifiable controls are explicitly separated from identifiable-direction claims. Unfavorable results are retained.

Reproducibility. Sections 3–4 define the algorithm and primary endpoints. Appendices A, B.1 and E provide proofs, preprocessing, hyperparameters, statistical units and implementation details. The study retains the fixed configurations, seeds and case-level outcomes used for the reported comparisons.

REFERENCES

Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schoelkopf. Cause-effect inference by comparing regression errors. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 900–909. PMLR, 2018. URL https://proceedings.mlr.press/v84/bloebaum18a.html.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-AOS1260.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7:107–118, 2025. URL https://doi.org/10.1038/s42256-024-00964-x.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, pp. 63–77, 2005. doi: 10.1007/11564089_7.

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. URL https://doi.org/10.1080/01621459.1963.10500830.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. *International Conference on Learning Representations*, 2021.

Nadir Sella, Adam Perbost, and Louis Verny. MIIC-SR: From complex data to structural causal models. In *NeurIPS 2025 Workshop on Uncovering Causality in Science (CauScien)*, 2025. URL https://openreview.net/forum?id=vgcPkcSylO.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. doi: 10.1214/009053607000000505.

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.

Makoto Yamada and Masashi Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-gaussian noise. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 643–648, 2010. doi: 10.1609/aaai.v24i1.7655.

Bing Yan, Guangfei Yang, Jun Wang, and Dawei Lu. Causally constrained symbolic regression for mining interpretable functions from observational data. *SSRN Electronic Journal*, 2026. doi: 10.2139/ssrn.6879831. Preprint posted June 4, 2026.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3414–3425. PMLR, 2020. URL https://proceedings.mlr.press/v108/zheng20a.html.

A Theoretical Analysis and Guarantees

A.1 Risk constraints, finite libraries and bounded scores

Condition on fitting observations: all M functions, transforms and coefficients are fixed, while selection observations remain independent. The results do not establish causal validity under confounding or unrestricted extrapolation.

Theorem 1 (Transfer of the observed cap to population risk)**.** *Let R_j and $\widehat{R}_j$ be population and selection risks of candidate j. Suppose that, with probability at least $1-\delta$, $\max_{j\le M}|\widehat{R}_j - R_j| \le \epsilon$. For $b \in \arg\min_j \widehat{R}_j$, any selector satisfying $\widehat{R}_{\widehat{j}} \le (1+\rho)\widehat{R}_b$ obeys, on the same event,*

$$R_{\widehat{j}} \le (1+\rho)\min_{j\le M} R_j + (2+\rho)\epsilon. \tag{9}$$

In particular, if every candidate's loss lies in $[0, B]$ almost surely, one may take

$$\epsilon = B\sqrt{\frac{\log(2M/\delta)}{2n_s}}. \tag{10}$$

Here n_s is the number of independent selection observations. The loss bound B must hold on their population support, not merely on observed rows.

Proof. Let $j^\star$ minimize population risk. Uniform control and empirical optimality give

$$\begin{aligned} R_{\widehat{j}} &\le \widehat{R}_{\widehat{j}} + \epsilon \le (1+\rho)\widehat{R}_b + \epsilon \\ &\le (1+\rho)\widehat{R}_{j^\star} + \epsilon \le (1+\rho)R_{j^\star} + (2+\rho)\epsilon. \end{aligned} \tag{11}$$

For bounded losses, Hoeffding's inequality bounds each two-sided deviation by $2\exp(-2n_s\epsilon^2/B^2)$. A union bound over the fixed M candidates yields (10). Independence among candidates is unnecessary. □

Gaussian noise and polynomial extrapolation need not satisfy bounded loss. The explicit radius therefore applies only under its stated premise. A justified uniform bound for unbounded losses may replace it. The algorithm's paired standard-error allowance is not such a bound.

Theorem 2 (When residual selection has an opportunity to help)**.** *For a realized nonempty risk-feasible set $\mathcal{A}$, let $\widehat{j} \in \arg\min_{j\in\mathcal{A}}\{\widehat{D}_j + \lambda C_j\}$, with deterministic complexity C_j. Suppose an event gives uniform dependence-estimation control $\max_{j\le M} |\widehat{D}_j - D_j| \le \epsilon_D$. Then on that event,*

$$D_{\widehat{j}} + \lambda C_{\widehat{j}} \le \min_{j\in\mathcal{A}}(D_j + \lambda C_j) + 2\epsilon_D. \tag{12}$$

If an admissible candidate j_0 improves the population penalized dependence of the empirical risk minimizer b by $\Delta > 2\epsilon_D$, the selected candidate improves it by at least $\Delta - 2\epsilon_D$.

Proof. For every $j \in \mathcal{A}$, uniform control and selection optimality give $D_{\widehat{j}}+\lambda C_{\widehat{j}} \le \widehat{D}_{\widehat{j}}+\lambda C_{\widehat{j}}+\epsilon_D \le \widehat{D}_j + \lambda C_j + \epsilon_D \le D_j + \lambda C_j + 2\epsilon_D$. Taking the minimum proves (12). Substituting j_0 proves the final assertion. □

The simultaneous event covers the data-dependent feasible set, but the theorem does not derive an ϵ_D for adaptive-bandwidth normalized HSIC. Lower residual dependence alone does not establish causal direction.

Finite-library limits.

Theorem 3 (Fixed-library collapse of the standard-error guard)**.** *Fix a finite fitted library independently of an increasing sequence of i.i.d. selection observations. Assume the squared-loss variables have finite second moments and there is a unique population-risk minimizer $j^\star$ with positive risk gap to every other candidate. For fixed finite $\kappa \ge 0$ and $\rho \ge 0$, the guard using $\min(\kappa s_j, \rho\widehat{R}_b)$ satisfies*

$$\Pr(\mathcal{A} = \{j^\star\} \text{ eventually as } n_s \to \infty) = 1. \tag{13}$$

Proof. The strong law and finiteness of the library imply uniform almost-sure convergence of empirical risks. Thus $b = j^\star$ eventually, and every other empirical risk gap eventually exceeds half its positive population gap. The sample variance of each paired loss difference converges to a finite value, so $s_j \to 0$ almost surely. Its allowed excess risk is at most κs_j, eventually smaller than the corresponding empirical gap. Every nonoptimal candidate is therefore excluded, whereas b is always included. □

This statement fixes the training-fitted library while selection size increases. It does not automatically apply when training and candidate generation also change. It identifies the selector as a finite-sample choice among near-optimal predictions, and motivates the separately evaluated relative-risk-band alternative. No positive result for that alternative is assumed.

Bounded-score construction and target. Condition on fitting: candidate equations, coefficients, affine transforms and Gaussian-kernel widths are fixed. Widths are the median nonzero distances of fitting inputs and targets. The residual width is shared across candidates. Partition selection rows in their fixed random order into b complete blocks of q rows. Discard a final incomplete block only for dependence. All rows still contribute to risk. For centered input/residual Gram matrices, define

$$Z_{jt} = q^{-2}\operatorname{tr}(K_{ct}L_{cjt}), \quad \bar{D}_j = b^{-1}\sum_t Z_{jt}, \quad V_j = (b-1)^{-1}\sum_t (Z_{jt} - \bar{D}_j)^2. \tag{14}$$

The centered matrices are positive semidefinite with traces at most q. $\operatorname{tr}(AB) \le \operatorname{tr}(A)\operatorname{tr}(B)$ implies $0 \le Z_{jt} \le 1$. The target $D_j^{(q)} = \mathbb{E}[Z_{jt} \mid \text{fitting}]$ is expected finite-block dependence, not bias-free population HSIC. For $b \ge 2$, let

$$r_j = \sqrt{2V_j\log(8M/\delta)/b} + \frac{7\log(8M/\delta)}{3(b-1)}, \quad U_j = \bar{D}_j + r_j. \tag{15}$$

V5 minimizes $U_j + \lambda C_j/n_s$ inside $\mathcal{A}_\rho = \{j : \widehat{R}_j \le (1+\rho)\min_k \widehat{R}_k\}$. Its development choice is $q = 16$, $\rho = .025$, with $\lambda = 10^{-3}$. The common additive radius does not change rankings. The variance term can.

Theorem 4 (Simultaneous bounded-target certificate)**.** *For i.i.d. selection rows independent of fitting, with probability at least $1-\delta$, all candidates satisfy $|\bar{D}_j - D_j^{(q)}| \le r_j$ and $|\widehat{R}_j^{(B)} - R_j^{(B)}| \le s_j$. Here the risk loss is $\ell_j^{(B)} = \min\{(Y - f_j(X))^2, B\}$, and s_j is the radius in* (15) *using n_s observations and the variance of $\ell_j^{(B)}/B$, multiplied by B. With clipped endpoints $L_j^R = \max(0, \widehat{R}_j^{(B)} - s_j)$ and $U_j^R = \min(B, \widehat{R}_j^{(B)} + s_j)$, any selected candidate obeys*

$$R_{\widehat{j}}^{(B)} - \min_j R_j^{(B)} \le U_{\widehat{j}}^R - \min_j L_j^R. \tag{16}$$

Proof. Apply the one-sided empirical Bernstein inequality (Maurer & Pontil, 2009) to Z, $1-Z$, $\ell^{(B)}/B$ and $1 - \ell^{(B)}/B$, each at failure probability $\delta/(4M)$. A union bound covers the fitted library and hence the adaptive selection. Subtract the simultaneous lower bound on minimum risk from the selected upper bound. Independence between candidates is unnecessary. □

On the same event, upper-confidence optimality also gives

$$D_{\widehat{j}}^{(q)} + \lambda C_{\widehat{j}}/n_s \le \min_{j \in \mathcal{A}_\rho} \{D_j^{(q)} + \lambda C_j/n_s + 2r_j\}. \tag{17}$$

Indeed, bound the selected target by its upper score, compare that score with each feasible candidate, then bound its score by target plus $2r_j$. The risk audit uses $B = 25$ in training-standardized units. It certifies clipped risk, whereas selection caps unbounded empirical MSE. The theorem is per direction. Joint coverage requires another error-budget split. Neither target guarantees causal orientation or intervention accuracy.

A.2 Paired switching and resampling guarantees

Reuse the three splits and training-fixed widths in Appendix A.1. Let $b = \arg\min_j \widehat{R}_j$ be the empirical risk reference, $Z_{jt} \in [0,1]$ the block statistic, and $c_j = \lambda C_j/n_s$. The relative-risk band is $\mathcal{A}_\rho = \{j : \widehat{R}_j \le (1+\rho)\widehat{R}_b\}$. Define the paired estimated gain for each challenger:

$$T_{bjt} = Z_{bt} - Z_{jt}, \qquad \widehat{G}_{bj} = \overline{T}_{bj} + c_b - c_j. \tag{18}$$

Positive gain favors the challenger on penalized finite-block dependence. All candidates retain their training-fitted coefficients.

Stability heuristic. Resample the same block indices for all candidates 399 times. A challenger must belong to $\mathcal{A}_\rho$, have positive estimated gain, and have positive gain in at least a fraction τ of resamples. Choose the accepted challenger with the largest estimated gain, breaking ties by risk and index. If none qualifies, return b. Development compares $q \in \{4, 8\}$ and $\tau \in \{.90, .95\}$ with $\rho = .05$. These frequencies measure selection stability, not post-selection p values or calibrated causal probabilities. Their value requires independent validation.

Certificate rule. For $M \ge 2$ candidates and $B \ge 2$ complete blocks, let $N = M(M-1)$, $L = \log(4N/\delta)$, and let V_{bj} be the sample variance of T_{bjt}. Define

$$r_{bj} = \sqrt{\frac{2V_{bj}L}{B}} + \frac{14L}{3(B-1)}, \qquad \underline{G}_{bj} = \widehat{G}_{bj} - r_{bj}. \tag{19}$$

The factor 14 accounts for the range $[-1, 1]$ of paired differences. The logarithm covers all ordered candidate pairs and both directions, since the reference is itself selected from these observations. The rule maximizes the positive lower bound over admissible challengers, or returns b when none exists. A singleton library returns its only candidate. Both switching rules retain the exact empirical MSE cap.

Theorem 5 (Paired switching certificate with an adaptive reference)**.** *Condition on fitting. Suppose selection rows are i.i.d. and independent of fitting, candidates and widths are fixed, and block statistics lie in* $[0,1]$*. With probability at least* $1-\delta$*, simultaneously for both directions, every accepted certificate switch satisfies*

$$D_{\widehat{j}}^{(q)} + c_{\widehat{j}} \le D_b^{(q)} + c_b - \underline{G}_{b\widehat{j}} < D_b^{(q)} + c_b. \tag{20}$$

If the rule falls back, the two penalized targets are equal.

Proof. For a fixed ordered pair, $(1-T_{ijt})/2 \in [0,1]$ is i.i.d. over blocks. Apply the one-sided empirical Bernstein inequality of Maurer & Pontil (2009) at failure probability $\delta/(2N)$. Undoing the affine transformation gives $\mathbb{E}T_{ijt} \ge \overline{T}_{ij} - r_{ij}$. A union bound over N ordered pairs in each direction provides simultaneous coverage. Each direction may have its own M and N. The event covers the adaptive reference, realized risk band and chosen challenger. Add the fixed cost difference and rearrange to obtain the result. □

This applies an existing inequality to paired selection. It is not a new causal identifiability result. It guarantees neither improved direction accuracy nor reduced intervention or extrapolation error.

Power limitation. Since $\overline{T}_{bj} \le 1$, no challenger can pass when the additive radius $14L/[3(B-1)]$ is at least $1 + \max_j(c_b - c_j)$. Small samples and large libraries can therefore force complete fallback to Risk+gate. Switch frequencies and uninformative radii reveal when the bound prevents a selection change. The stability heuristic is distinct: relaxing the certificate removes its guarantee.

Finite-resampling guarantees.

Theorem 6 (Monte Carlo stability of the switching decision)**.** *Condition on all fitted models and observed selection blocks. Let* $K \ge 1$ *be the number of challengers considered across the two directions, and let* π_j *be the probability that one block-bootstrap draw has positive estimated penalized gain for challenger* j*. For* L *independent bootstrap draws, let* $\widehat{\pi}_j$ *be their positive-gain frequency. Shared indices across challengers within each draw are allowed. For* $\delta \in (0,1)$*,*

$$\Pr\left\{ \max_{j \le K} |\widehat{\pi}_j - \pi_j| > \sqrt{\frac{\log(2K/\delta)}{2L}} \,\middle|\, \textit{observed blocks} \right\} \le \delta. \tag{21}$$

If every challenger that passes the deterministic risk and positive-gain checks is separated from threshold τ *by more than the displayed radius, the finite-*L *and infinite-resampling rules return the same equation, with conditional probability at least* $1-\delta$*.*

Proof. For a fixed challenger, the positive-gain indicators are i.i.d. Bernoulli variables of mean π_j conditional on the observed blocks. Hoeffding's inequality (Hoeffding, 1963) bounds the probability of an absolute frequency error exceeding u by $2\exp(-2Lu^2)$. A union bound over the K challengers gives $2K\exp(-2Lu^2)$. The radius in (21) makes this δ. No independence across challengers is needed. On that event, threshold decisions agree for all eligible challengers under the margin assumption. Their ranking uses the same observed gain, risk and index tie-breaks, so the accepted set and the returned equation agree, including the minimum-risk fallback. □

This controls conditional Monte Carlo error at rate $L^{-1/2}$, not sampling error in the blocks. It neither makes π_j a causal probability nor validates bootstrap population coverage. More resamples cannot supply a direction guarantee.

Proposition 3 (Variance of a paired block comparison)**.** *For i.i.d. paired block scores* (Z_{bt}, Z_{jt}) *with finite second moments, the variance of their average difference over* B *blocks is*

$$\mathrm{Var}(\overline{Z}_b - \overline{Z}_j) = \frac{\mathrm{Var}(Z_b) + \mathrm{Var}(Z_j) - 2\,\mathrm{Cov}(Z_b, Z_j)}{B}. \tag{22}$$

Compared with independent block samples from the same marginals, pairing does not increase this variance if and only if the covariance is nonnegative. The reduction is strict when the covariance is positive.

Proof. Expand $\mathrm{Var}(Z_b - Z_j)$ and use independence across blocks. With independent marginal samples the covariance term is zero. Subtracting the two variances proves the comparison. □

Pairing is not guaranteed to improve variance when the covariance is negative. This elementary identity explains a possible benefit of shared observations. It is not a claim that every empirical challenger has positive covariance.

A.3 Intervention consequences and transfer

Under the reversed ANM, the effect is a root with fitted mean c. Consequently

$$\mathbb{E}_Q(c - f^\star)^2 = \mathrm{Var}_Q(f^\star) + (c - \mathbb{E}_Q f^\star)^2. \tag{23}$$

The other intervention contribution is nonnegative, proving Proposition 2. With correct orientation, intervening on the effect leaves the fitted cause-root mean unchanged, which gives the stated forward loss. Both identities hold on a finite query grid. They concern squared loss, not the square of the mean NRMSE.

Let L_{de} evaluate the direction of pipeline d using the equations and root means of pipeline e, for $d, e \in \{F, J\}$. On the same query grid, define

$$\begin{aligned} A_{\mathrm{dir}} &= \tfrac{1}{2}[(L_{FF} - L_{JF}) + (L_{FJ} - L_{JJ})], \\ A_{\mathrm{eq}} &= \tfrac{1}{2}[(L_{FF} - L_{FJ}) + (L_{JF} - L_{JJ})]. \end{aligned} \tag{24}$$

Cancellation gives $A_{\mathrm{dir}} + A_{\mathrm{eq}} = L_{FF} - L_{JJ}$ exactly. This symmetrically allocates component interactions. It is algorithmic accounting, not causal mediation. Either contribution can be negative. For the historical JJ–FJ contrast, both use joint direction scoring. Only selection changes. Such accounting cannot turn an unfavorable confirmation comparison into evidence of mechanistic benefit.

Gate attribution. If a frozen gate changes a baseline only on event T, its accuracy gain is $\Pr(T, \text{correction}) - \Pr(T, \text{corruption})$. Outside T the correctness indicators coincide. Inside T their difference is $+1, -1$ or zero. With independent task clusters and mean within-cluster differences $Z_i \in [-1, 1]$, Hoeffding gives the lower bound $m^{-1}\sum_i \mathbb{E}Z_i \geq m^{-1}\sum_i Z_i - \sqrt{2\log(1/\delta)/m}$ with probability $1 - \delta$. Within-cluster independence is unnecessary. This is not calibration of a residual test and supplies no true direction label for nonidentifiable data.

Conditional transfer to mean interventions. Define the correctly oriented squared population intervention loss as $L_Q^+ = \mathbb{E}_Q[(f_{\widehat{j}} - f^\star)^2]/(2s_f^2) + (\widehat{\mu}_X - \mu_X)^2/(2s_X^2)$. Its square root corresponds to intervention NRMSE. All displayed risks and means are assumed finite.

Theorem 7 (Conditional transfer to mean-intervention loss)**.** *Condition on the fitting sample and its finite candidate library. Let $Y = f^\star(X) + N$, with $N \perp\!\!\!\perp X$, $\mathbb{E}N = 0$, and $\mathbb{E}N^2 = \sigma^2 < \infty$. Write P_X for the observational cause law, $R_j = \mathbb{E}[(Y - f_j(X))^2]$, and $R_\star = \min_j R_j$. Suppose the simultaneous selection-risk event $\max_j |\widehat{R}_j - R_j| \leq \varepsilon$ holds and the returned equation satisfies $\widehat{R}_{\widehat{j}} \leq (1+\rho)\min_j \widehat{R}_j$, with $\rho, \varepsilon \geq 0$. Let cause interventions follow $Q \ll P_X$ with density ratio $dQ/dP_X \leq W < \infty$ almost surely. If the returned direction is correct, its mean-intervention loss satisfies*

$$L_Q^+ \leq \frac{W\{(1+\rho)R_\star + (2+\rho)\varepsilon - \sigma^2\}}{2s_f^2} + \frac{(\widehat{\mu}_X - \mu_X)^2}{2s_X^2}. \tag{25}$$

Here $s_f, s_X > 0$ are the chosen normalizers, $\mu_X = \mathbb{E}X$, and $\widehat{\mu}_X$ is the returned root-mean estimate. All risks and normalizers must be expressed in the same response units.

Proof. Let $j_\star$ minimize population risk. On the stated event,

$$\begin{aligned} R_{\widehat{j}} &\leq \widehat{R}_{\widehat{j}} + \varepsilon \leq (1+\rho)\widehat{R}_{j_\star} + \varepsilon \\ &\leq (1+\rho)R_\star + (2+\rho)\varepsilon. \end{aligned} \tag{26}$$

Independence and the zero noise mean remove the cross term in the squared loss, giving $R_j = \mathbb{E}_{P_X}(f_j - f^\star)^2 + \sigma^2$. Absolute continuity and the density-ratio bound imply

$$\mathbb{E}_Q(f_{\widehat{j}} - f^\star)^2 = \mathbb{E}_{P_X}\left[\frac{dQ}{dP_X}(f_{\widehat{j}} - f^\star)^2\right] \le W(R_{\widehat{j}} - \sigma^2). \tag{27}$$

With correct orientation, intervening on the cause changes the effect mean to $f_{\widehat{j}}(x)$, whereas intervening on the effect leaves the cause mean at $\widehat{\mu}_X$. Substituting the preceding inequalities into the two normalized squared errors proves (25). □

If $f^\star$ belongs to the library, the first numerator is $W\{\rho\sigma^2 + (2+\rho)\varepsilon\}$: even then a positive risk budget permits mechanism error. Uniform control for unbounded losses and correct orientation remain assumptions. Outside-support queries may have no finite W (Equation 8). A discrete grid under continuous P_X also lacks this density ratio. The bound concerns a population query law, not every finite grid.

B Experimental Protocol and Independent Confirmation

B.1 Frozen campaigns, data and evaluation

V4, V5 and V6 use separate development/confirmation stages. Earlier results may inform a later design. They are not independent confirmation of it. Counts in Table 4 denote cases, not independent scientific replications. V6 development compares nine methods. Minimize mean intervention

Table 4: Campaign inventories and locked decisions. V5 calibration is separate from development and confirmation.

Campaign	Development	Calibration	Confirmation	Locked rule
V4	384	–	5,782	$\kappa = .5, \rho = .10$
V5	384	400	7,483	$q = 16, \rho = .025$
V6	384	–	2,916	$q = 8, \tau = .90, \rho = .05$

NRMSE subject to accuracy at least Risk+gate's. Exact ties favor Risk+gate, then method name. This selects Stability $(8, .90)$, with 95.833% development accuracy and error 0.138252 versus Risk+gate's 0.138861 at the same accuracy. The selected rule is fixed before the independent confirmation cases are scored.

Mechanisms and randomization. V6 primary confirmation crosses 12 forms and 20 fresh parameter seeds with two centered cause laws, $n = 200/500$, and SNR 2/10: 1,920 cases in 240 mechanism clusters. Transfer and new-form suites each contain 480 cases. 36 additional cases use $n = 1,000/4,000$ for certificate power. V4 generator coefficients have signed uniform magnitudes $[.35, 1.1]$, input scale $[.7, 2.1]$ and shift $[-.5, .5]$. Centered unit-variance uniform and normal causes form the primary grid. A separate 16,384-sample reference fixes signal scale for SNR. Parameter, observation, swap, split and query streams have distinct SHA-256-derived seeds. Exact campaign functions and case keys remain in the frozen protocol sources. Sharing a form is not a new scientific domain. New forms include sine–tanh products, rational sine, log-square, exponential-tanh, asinh-cubic and double Gaussian bumps.

Queries and units. Section 4 defines the independent ID, OOD and two-direction intervention queries. All retain fitting-row coefficients and fixed ID normalizers. Queries score mean responses. They never choose an equation or orientation. Oracle errors exclude estimated-direction effects.

Inference. V6 uses eight primary tests. V4/V5 each use twelve. Holm applies within each family, while 20,000-draw cluster-bootstrap intervals are marginal. Sign-flips require symmetric, exchangeable cluster effects. Repeated conditions are paired. Transfer, V6 OOD and post-decision oracles remain descriptive. No favorable subset redefines success.

External observations and physical responses. The 102 scalar Tuebingen pairs use caps 500/1,000, three splits per cap and benchmark weights. Quality flags are excluded. These 612 protocols contain no intervention truth. Causal Chamber (Gamella et al., 2025) has 11 actuator tasks and 165 protocols from one light tunnel. Held-out regimes shift one actuator. Sensor errors use training standard deviations and equal actuator weights. Reused measurements are not independent physical domains.

B.2 Full primary tests and confirmation comparisons

Table 1 retains all 19 V6 methods for the primary, transfer and new-form suites. Table 5 gives the complete prespecified primary family. Positive effects favor stability. All errors are lower-is-better. Marginal intervals must not override adjusted tests.

Table 5: All eight V6 primary contrasts. Positive effects favor the development-selected stability rule. Intervals are marginal 95% mechanism-cluster bootstrap intervals. Tests are two-sided cluster sign-flips with Holm adjustment.

Control	Endpoint	Effect	95% interval	p_{Holm}
Risk+gate	Accuracy (pp)	-0.1042	[-0.4167, 0.2083]	1.000
Risk+gate	E_{do} reduction	-0.0012	[-0.0041, 0.0016]	1.000
Simple+gate	Accuracy (pp)	0.2083	[-0.2604, 0.6771]	1.000
Simple+gate	E_{do} reduction	0.0043	[0.0004, 0.0084]	0.292
Joint+gate	Accuracy (pp)	-0.0521	[-0.5208, 0.4167]	1.000
Joint+gate	E_{do} reduction	0.0021	[-0.0020, 0.0062]	1.000
Original guard+gate	Accuracy (pp)	0.1042	[-0.2604, 0.4688]	1.000
Original guard+gate	E_{do} reduction	0.0013	[-0.0021, 0.0049]	1.000

Switches and oracle diagnostics. Stability switches 1,151/3,840 primary directional fits (29.97%), correcting four directions and corrupting six versus Risk+gate. The shared-pool oracle ID error is 0.084622. Mean excess is 0.044279 for stability and 0.044311 for risk. These post-decision diagnostics cannot select or repair the pipeline. The certificate returns the risk reference in all 2,916 cases. Every radius exceeds one in the three principal suites and in half of larger-sample fits. None accepts a switch. Table 6 retains all larger-sample methods.

Table 6: Completed V6 large sample power confirmation (36 cases). All scheduled methods are retained. Accuracy is higher-is-better. All errors are lower-is-better. The main rule was selected on development only.

Method	Acc. (%)	E_{do}	ID	OOD-1.5
Risk+gate	91.67	0.1173	0.0856	2.138
Simple+gate	91.67	0.1181	0.0909	2.183
Joint+gate	94.44	0.0996	0.0981	2.187
Original guard+gate	91.67	0.1175	0.0928	2.244
Certificate+gate	91.67	0.1173	0.0856	2.138
Stability $q = 4$, $\tau = 0.9$	91.67	0.1169	0.0851	2.068
Stability $q = 4$, $\tau = 0.95$	91.67	0.1166	0.0844	2.054
Stability $q = 8$, $\tau = 0.9$ (selected)	91.67	0.1164	0.0833	2.036
Stability $q = 8$, $\tau = 0.95$	91.67	0.1163	0.0829	2.039

Completeness. All 384 development and 2,916 confirmation cases completed: 3,456 and 55,044 method evaluations, respectively. Three engineering cases are excluded. Raw identifiers, complete method lists, status counts and aggregate means were checked after transfer. The confirmation retains its development lock.

C Representative Symbolic Solutions

Figure 2 reconstructs four existing V6 cases from their saved full-precision coefficients. For each of four fixed mechanism families, we take the $n = 500$, SNR 10, uniform-cause case closest to the median stability ID error among its 20 seeds, breaking ties by case ID. Selection does not filter for correct direction, switching or improvement over risk. Axes are aligned to the known generator cause and effect for interpretation.

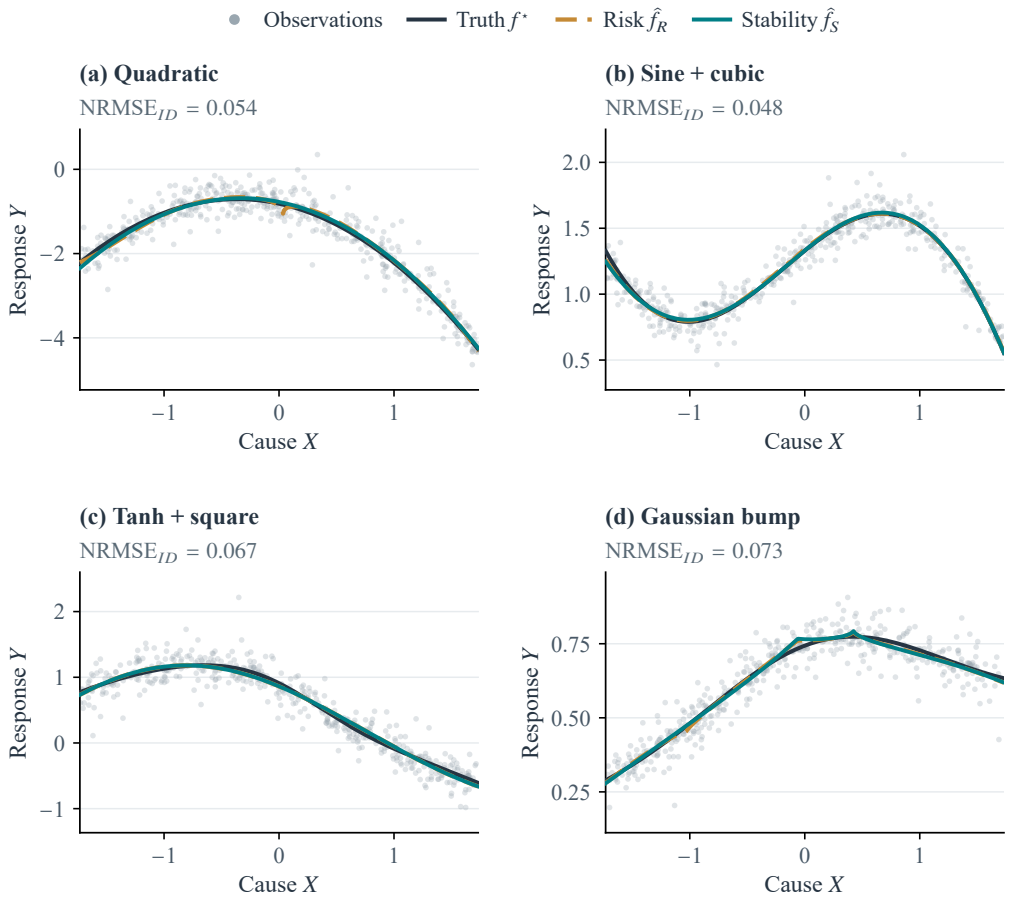

Figure 2: Selected equations within the observational support. Grey points are the 500 observations for each fixed synthetic case. Curves use the fitted equations without refitting. Labels report oracle-direction ID NRMSE for the stability equation. The known mechanism is black, minimum-risk selection gold and paired-stability selection teal. In panel (c), the two fitted equations coincide. These four descriptive cases do not replace the complete confirmatory comparison.

An explicit returned equation. For panel (a), the stored stability solution has the following compact form (coefficients rounded here to six decimals, with full precision in the plots):

$$u = (x - 0.035055)/0.931005,$$
$$\widehat{f}_S(x) = -1.521502 + 0.994939\left(0.732120 - 0.566838u - 0.732120u^2\right). \tag{28}$$

Equation (28) includes the fitted input and response transforms, not a refit in raw coordinates. The other three selected solutions combine sine, tanh, polynomial, log-absolute or square-root-absolute terms. All selected predictors and transforms are retained with the plotting data.

Behavior outside the observation range. Figure 3 extends exactly the same equations to the fixed OOD-1.5 query range. No coefficient or case is reselected for this display. The shaded interval is the support of the uniform observational cause. The known mechanism remains available because these are synthetic cases.

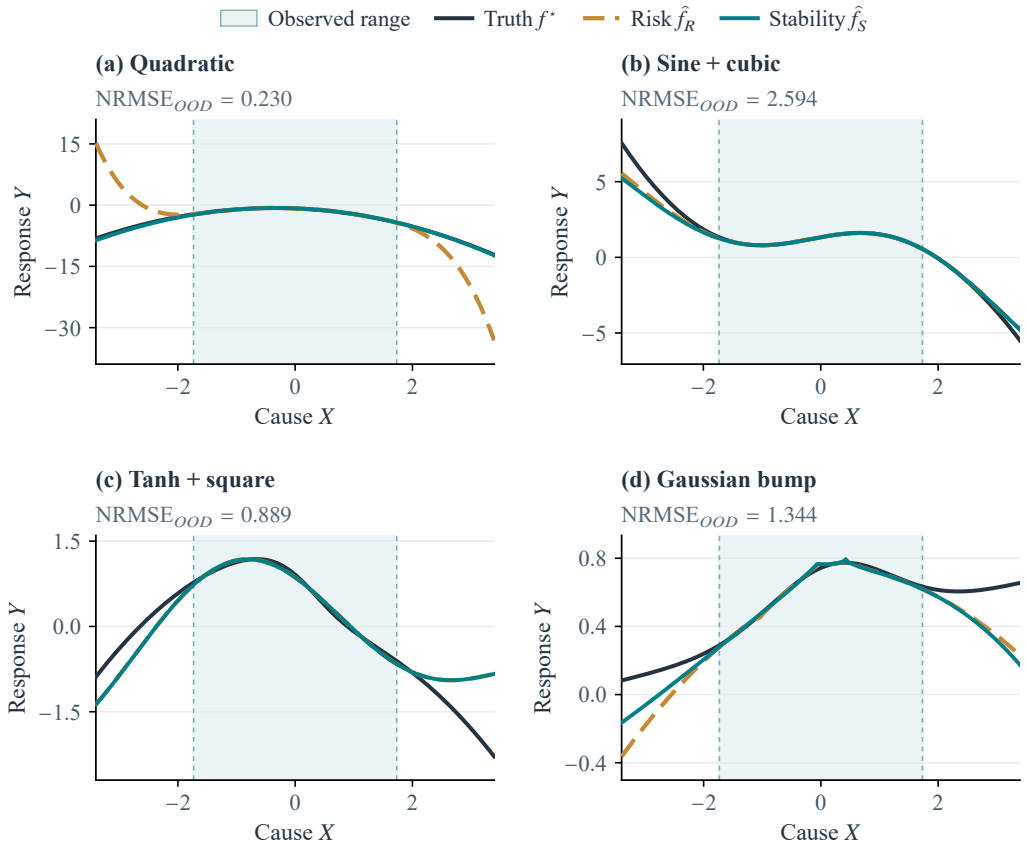

Figure 3: Extrapolation of the saved solutions, with the same panel order and colors as Figure 2. Full curves and ordinary linear axes are retained. The grey-green band marks the observation support. Labels give stability OOD-1.5 NRMSE on the independent outside-range query grid, using the fixed ID normalizer. Similar observed-range fits can imply different extrapolation, including deterioration under stability in panel (b).

Table 7: Case identities and oracle errors for the displayed equations. R denotes minimum risk. S denotes paired stability. Values are rounded only for display. These are four illustrations from the existing confirmation.

Panel / family	Seed	ID (R)	ID (S)	OOD (R)	OOD (S)
(a) Quadratic	76200013	.0670	.0543	9.2150	.2297
(b) Sine + cubic	76200018	.0438	.0484	2.2150	2.5936
(c) Tanh + square	76200013	.0674	.0674	.8890	.8890
(d) Gaussian bump	76200003	.0660	.0731	1.5412	1.3438

Table 7 includes favorable and unfavorable changes. These four cases illustrate how equations with similar in-range fit can diverge on outside-range queries. They provide no new training, independence certificate or superiority test.

D Additional Results and Component Analyses

D.1 Bounded-score confirmation and physical validation

The V5 primary suite contains 2,880 cases and 360 mechanism clusters. The bounded selector uses the frozen setting $(q, \rho) = (16, .025)$. Risk+gate minimizes selection risk. Regression labels specify minimum-risk fitting followed by dependence, RECI or the common gate. All 26 methods and all 12 primary tests appear in Tables 8 and 9. These are separate from V6.

Table 8: All V5 primary methods: higher accuracy and lower errors are preferable.

Method	Acc. (%)	E_{do}	ID	OOD-1.5
Block mean	95.868	0.135	0.131	2.890
Risk+gate	95.799	0.136	0.132	3.248
Risk+RECI	95.451	0.138	0.132	3.248
Bounded selector	95.868	0.137	0.134	2.987
GP+dependence	87.778	0.200	0.128	2.802
GP+gate	96.250	0.127	0.128	2.802
GP+RECI	95.451	0.133	0.128	2.802
Bounded+dependence	86.076	0.220	0.134	2.987
Bounded+RECI	95.451	0.139	0.134	2.987
IGCI	93.819	0.153	0.132	3.248
Joint+gate	95.972	0.137	0.135	2.434
Original guard+gate	95.729	0.136	0.132	3.210
Polynomial+dependence	86.424	0.219	0.905	237.844
Polynomial+gate	95.799	0.138	0.905	237.844
Polynomial+RECI	95.451	0.140	0.905	237.844
Kernel ridge+dependence	87.222	0.208	0.140	3.272
Kernel ridge+gate	96.042	0.131	0.140	3.272
Kernel ridge+RECI	95.451	0.136	0.140	3.272
Relative-cap HSIC	95.764	0.136	0.133	3.045
Simple+gate	95.799	0.137	0.133	2.773
Spline+dependence	86.771	0.214	0.149	3.019
Spline+gate	95.868	0.137	0.149	3.019
Spline+RECI	95.451	0.140	0.149	3.019
SVR+dependence	86.875	0.204	0.122	3.086
SVR+gate	96.076	0.120	0.122	3.086
SVR+RECI	95.451	0.125	0.122	3.086

Accuracy effects in the test table are proportions, not percentage points. None of the favorable contrasts survives Holm correction. The resolved contrast favors Joint+gate on OOD error. It cannot be reported as a selector win.

External pairs and physical intervention failures. V5 external accuracy is benchmark-weighted over 612 Tuebingen protocols. Physical errors average the 11 actuator tasks equally, with 165 reused observational protocols (Table 10). Neither source provides independent new systems for every row or partition. These physical tests evaluate V5, not V6.

Endpoint interpretation. The hardware endpoint compares predicted and measured regime means after one actuator distribution shifts. Reverse models predict the fitted sensor-root mean. The oracle column forces the known actuator direction. Its large errors show that orientation mistakes do not explain the failure. The bounded selector's mean error 78.3513 and oracle error 79.0840 contrast with GP+gate's 0.7149 for both. Polynomial extrapolation fails still more severely. These observations are retained as failures, not censored or imputed.

Table 9: All 12 fixed primary comparisons. Positive effects favor the selected specification. Intervals are marginal, tests are Holm adjusted.

Comparator	Endpoint	Effect	95% interval	p_{Holm}
Risk+gate	correct	0.0007	[-0.0021,0.0035]	1.0000
Risk+gate	E_{do}	-0.0006	[-0.0031,0.0019]	1.0000
Risk+gate	OOD-1.5	0.2612	[0.0287,0.5059]	0.3355
Original guard+gate	correct	0.0014	[-0.0021,0.0049]	1.0000
Original guard+gate	E_{do}	-0.0011	[-0.0043,0.0022]	1.0000
Original guard+gate	OOD-1.5	0.2233	[-0.0071,0.4613]	0.6440
Joint+gate	correct	-0.0010	[-0.0042,0.0021]	1.0000
Joint+gate	E_{do}	-0.0000	[-0.0029,0.0028]	1.0000
Joint+gate	OOD-1.5	-0.5524	[-0.7651,-0.3363]	0.0006
Bounded+RECI	correct	0.0042	[-0.0003,0.0087]	0.8982
Bounded+RECI	E_{do}	0.0025	[-0.0010,0.0060]	1.0000
Bounded+RECI	OOD-1.5	0.0000	[0.0000,0.0000]	1.0000

Table 10: Complete V5 external and physical summaries. External accuracy is in percent. All physical errors are lower-is-better. A dash means the method was not in that physical suite.

Method	External acc.	Mean error	Binned RMSE	Oracle error
Block mean	65.174	93.2967	105.8454	94.2632
Risk+gate	66.034	29.6568	33.8098	29.6468
Risk+RECI	67.869	29.6468	33.7998	29.6468
Bounded selector	65.607	78.3513	89.0961	79.0840
GP+dependence	52.810	0.7394	0.8158	0.7149
GP+gate	65.752	0.7149	0.7969	0.7149
GP+RECI	67.869	0.7149	0.7969	0.7149
Bounded+dependence	52.520	63.8589	72.6447	79.0840
Bounded+RECI	67.869	79.0840	89.9060	79.0840
IGCI	69.405	–	–	–
Joint+gate	67.567	15.3207	17.4523	15.3157
Original guard+gate	65.406	43.3382	49.3545	43.5670
Polynomial+dependence	51.932	131795.1206	191353.8622	270830.9802
Polynomial+gate	66.596	270823.7408	392940.6956	270830.9802
Polynomial+RECI	67.869	270830.9802	392949.4884	270830.9802
Kernel ridge+dependence	50.815	0.9303	1.0487	1.1397
Kernel ridge+gate	66.838	1.1379	1.2999	1.1397
Kernel ridge+RECI	67.869	1.1397	1.3123	1.1397
Relative-cap HSIC	65.250	66.8942	75.9536	67.1259
Simple+gate	67.123	23.3658	26.5784	23.3608
Spline+dependence	52.051	1.8813	2.0018	2.5078
Spline+gate	66.477	2.5023	2.6555	2.5078
Spline+RECI	67.869	2.5078	2.6615	2.5078
SVR+dependence	53.034	0.9345	1.0233	1.0496
SVR+gate	67.126	1.0545	1.1533	1.0496
SVR+RECI	67.869	1.0496	1.1504	1.0496

Two-node ANM identification is not asserted for this multivariable physical system. Training uses only subsamples of the independently randomized reference settings. Intervention measurements enter scoring only. Both published strengths are evaluated for colors and polarizers, and the published single strength for auxiliary LEDs.

No interchange of evidence. Tuebingen accuracy supplies observational direction evidence and no intervention ground truth. Its repeated partitions are not 612 independent sources. Conversely, measured hardware response errors do not validate the synthetic ANM assumptions. Synthetic direction accuracy, observational benchmark accuracy and measured response accuracy answer different questions.

D.2 Complete paired component sensitivity

All 48 variants in Tables 11 and 12 use the same completed 480-case V5 ablation suite. The reference is V5 bounded selection, not V6 stability. Positive ΔE_{do} means larger error. These are descriptive paired means. The confirmation does not select a replacement default.

Table 11: Complete component sensitivity, part 1 of 2. All 24 settings are shown.

Variant	Acc. (%)	E_{do}	ΔE_{do}	OOD
AIC+gate	96.25	0.1376	+0.0042	1.801
Base library only	97.08	0.1271	-0.0064	5.671
BIC+gate	96.04	0.1441	+0.0106	1.826
No confidence radius	96.46	0.1292	-0.0042	2.951
Block size $q = 16$	96.25	0.1335	+0.0000	2.989
Block size $q = 32$	97.08	0.1272	-0.0063	3.019
Block size $q = 4$	96.46	0.1311	-0.0024	2.885
Block size $q = 8$	96.04	0.1341	+0.0006	2.957
Candidate cap 16	96.88	0.1258	-0.0077	6.567
Candidate cap 32	97.08	0.1271	-0.0064	5.671
Candidate cap 64	96.25	0.1385	+0.0051	4.867
Complexity $\lambda = 0$	96.25	0.1334	-0.0000	3.011
Complexity $\lambda = 0.001$	96.25	0.1335	+0.0000	2.989
Complexity $\lambda = 0.01$	96.46	0.1317	-0.0018	2.569
Complexity $\lambda = 0.1$	96.04	0.1366	+0.0032	2.195
Confidence $\delta = 0.01$	96.25	0.1335	+0.0000	3.000
Confidence $\delta = 0.05$	96.25	0.1335	+0.0000	2.989
Confidence $\delta = 0.1$	96.25	0.1332	-0.0003	2.989
Confidence $\delta = 0.2$	96.25	0.1331	-0.0004	2.989
No residual selection	97.08	0.1254	-0.0081	2.817
Risk+RECI	95.42	0.1379	+0.0044	2.817
V5 bounded reference	96.25	0.1335	+0.0000	2.989
Gate thresholds $(0.005, 0.1)$	96.25	0.1335	+0.0000	2.989
Gate thresholds $(0.005, 0.2)$	96.25	0.1340	+0.0006	2.989

What the removals change. Removing the confidence radius retains the risk band, bounded block statistic and direction gate. Removing complexity alters the candidate penalty. Restricting the library changes which equations can be represented. The latter therefore tests representation, not only the residual criterion. The base-library row improves direction accuracy while worsening OOD error, illustrating why the three endpoints cannot be replaced by a single ranking.

Reference and rounding. The V5 reference is $q = 16$, $\rho = .025$ with upper-confidence selection. Its intervention error is 0.1335 at the displayed precision. Paired changes are calculated from the original case-level values before rounding. The difference between rounded displayed means can therefore differ in the last digit. A displayed signed zero is retained from the source table. Duplicated default settings are explicit named configurations, not independent replications.

Table 12 continues with the remaining 24 variants. No row is removed because its result is unfavorable, and the V6 development choice remains unchanged.

Remaining direction and sensitivity settings. This is the second half of the same 480-case V5 analysis. Candidate sharing, the reference and all endpoint definitions are unchanged. Positive error changes favor the reference. Negative changes favor the listed variant.

Table 12: Complete component sensitivity, part 2 of 2. These descriptive contrasts do not select a new default.

Variant	Acc. (%)	E_{do}	ΔE_{do}	OOD
Gate thresholds $(0.01, 0.1)$	95.83	0.1364	+0.0030	2.989
Gate thresholds $(0.025, 0.1)$	95.42	0.1384	+0.0050	2.989
Dependence direction only	87.08	0.2080	+0.0745	2.989
RECI direction only	95.42	0.1391	+0.0057	2.989
Joint+gate	96.46	0.1330	-0.0005	1.916
No complexity cost	96.25	0.1334	-0.0000	3.011
No analytic ridge	96.25	0.1333	-0.0001	3.206
Original guard+gate	96.88	0.1288	-0.0047	2.675
Permutation row cap 160	96.25	0.1335	+0.0000	2.989
Permutation row cap 320	96.25	0.1335	+0.0000	2.989
Permutation row cap 80	96.88	0.1299	-0.0036	2.989
Permutations 199	96.25	0.1335	+0.0000	2.989
Permutations 499	96.67	0.1308	-0.0027	2.989
Permutations 999	96.67	0.1308	-0.0027	2.989
Random+gate	96.67	0.1294	-0.0040	2.973
Risk cap $\rho = 0$	97.08	0.1254	-0.0081	2.817
Risk cap $\rho = 0.01$	96.25	0.1324	-0.0010	2.693
Risk cap $\rho = 0.025$	96.25	0.1335	+0.0000	2.989
Risk cap $\rho = 0.05$	96.46	0.1360	+0.0025	3.004
Risk cap $\rho = 0.1$	96.67	0.1372	+0.0038	3.463
Risk cap $\rho = 0.2$	96.25	0.1472	+0.0138	3.398
Relative-cap HSIC	96.67	0.1298	-0.0036	2.700
Simple+gate	96.46	0.1328	-0.0007	2.596
Wide risk cap	95.83	0.1553	+0.0218	3.344

Direction versus equation. Replacing the gate with dependence-only or RECI-only direction retains the selected directional equations. Oracle ID and OOD metrics therefore remain unchanged, while end-to-end intervention error can change substantially. The dependence-only row reaches 87.08% accuracy and error 0.2080, against 96.25% and 0.1335 for the reference. These are matched descriptive contrasts, not independent new tests of V6 paired stability.

Risk budgets and computation. A zero relative cap recovers the risk selector on these observations. Wider bands admit more equations, but neither observed-risk membership nor switch activity guarantees a better mechanism. Permutation counts and caps change the direction procedure's computation. Block size changes the finite-block dependence target. Thus different sensitivity rows need not optimize the same statistical object. Tables 11 and 12 report every scheduled setting.

The factorial in Table 13 independently crosses representation, selection and direction on its separate 240-case suite. It complements these removals without pooling their observations.

D.3 Factorial attribution and historical primary tests

The 240-case V5 factorial below crosses ten representations, two selectors and three direction rules. Within a representation the same fitted pool is shared. Direction changes retain equations, so oracle ID and OOD errors are unchanged. Fitting budgets differ across representations. The factorial and 480-case ablation suite use distinct observations and are not pooled.

Table 13: Completed 240-case V5 factorial: ten representations, two selectors and three direction rules (60 configurations). Each cell is accuracy (%) / intervention NRMSE. Fit minimizes risk. Block uses bounded upper-confidence selection. Both selectors share the same fitted representation.

Representation	Selector	Dependence	RECI	Gate
Linear	Fit	30.83 / 0.722	96.67 / 0.453	95.42 / 0.461
	Block	30.83 / 0.722	96.67 / 0.453	95.42 / 0.461
Polynomial	Fit	87.08 / 0.215	96.67 / 0.139	97.08 / 0.133
	Block	85.83 / 0.227	96.67 / 0.141	96.67 / 0.139
Spline	Fit	86.25 / 0.216	96.67 / 0.130	97.08 / 0.126
	Block	87.92 / 0.204	96.67 / 0.131	97.08 / 0.127
Kernel ridge	Fit	88.75 / 0.198	96.67 / 0.126	96.67 / 0.125
	Block	87.08 / 0.213	96.67 / 0.126	97.08 / 0.123
Random forest	Fit	85.83 / 0.282	96.67 / 0.192	96.25 / 0.195
	Block	85.42 / 0.285	96.67 / 0.193	96.25 / 0.195
Extra trees	Fit	89.17 / 0.252	96.67 / 0.199	96.25 / 0.203
	Block	87.50 / 0.264	96.67 / 0.199	96.25 / 0.203
Gaussian process	Fit	88.33 / 0.196	96.67 / 0.121	97.50 / 0.114
	Block	88.33 / 0.196	96.67 / 0.121	97.50 / 0.114
SVR	Fit	86.67 / 0.210	96.67 / 0.115	97.50 / 0.108
	Block	87.08 / 0.209	96.67 / 0.115	97.50 / 0.108
Huber polynomial	Fit	82.92 / 0.262	96.67 / 0.167	96.67 / 0.167
	Block	84.58 / 0.248	96.67 / 0.168	96.67 / 0.168
MLP	Fit	88.75 / 0.200	96.67 / 0.130	96.67 / 0.130
	Block	88.33 / 0.205	96.67 / 0.130	96.67 / 0.130

Original expanded campaign (V4). The V4 primary suite contains 1,920 cases in 240 clusters. All twelve prespecified comparisons appear in Table 14. Positive effects favor the original guard. Accuracy effects are proportions. Intervals are marginal 95% cluster intervals, while the tests receive Holm adjustment.

Table 14: All 12 fixed primary comparisons. Positive effects favor the selected specification. Intervals are marginal, tests are Holm adjusted.

Comparator	Endpoint	Effect	95% interval	p_{Holm}
Risk+gate	correct	-0.0010	[-0.0052,0.0031]	1.0000
Risk+gate	E_{do}	-0.0011	[-0.0048,0.0026]	1.0000
Risk+gate	OOD-1.5	-0.0973	[-0.3437,0.1477]	1.0000
Risk+RECI	correct	0.0068	[0.0005,0.0135]	0.5357
Risk+RECI	E_{do}	0.0045	[-0.0007,0.0098]	0.7161
Risk+RECI	OOD-1.5	-0.0973	[-0.3420,0.1522]	1.0000
Joint+gate	correct	0.0026	[-0.0016,0.0068]	1.0000
Joint+gate	E_{do}	0.0036	[-0.0001,0.0075]	0.5872
Joint+gate	OOD-1.5	-0.5625	[-0.8182,-0.3266]	0.0006
Guard+RECI	correct	0.0068	[0.0005,0.0135]	0.5365
Guard+RECI	E_{do}	0.0048	[-0.0001,0.0101]	0.5872
Guard+RECI	OOD-1.5	0.0000	[0.0000,0.0000]	1.0000

No favorable V4 primary effect is resolved after adjustment. The significant OOD contrast favors Joint+gate. The 1,597 changed-equation cases and seven corrections versus nine corruptions refer to selection against Risk+gate. Gate-versus-RECI counts are a different contrast.

E REPRODUCIBILITY AND CALIBRATION

Regression budgets. Table 15 lists the fixed search grids used in the expanded protocol.

Table 15: Regression hyperparameter grids. Width multipliers use the median distance.

Representation	Hyperparameter grid
Polynomial	Degrees $2, 3, 5, 9$, ridge $.001, .1, 1$.
Cubic splines	Knots $5, 9, 15$, ridge $.001, .1, 1$.
Random forest	128 trees, minimum leaf size $2, 5, 10, 20$.
Extra trees	128 trees, minimum leaf size $2, 5, 10, 20$.
Kernel ridge	Width multipliers $.5, 1, 2$.
SVR	Width multipliers $.5, 1, 2$, $C = 1, 10, 100$, $\epsilon = .05$.
Gaussian process	Scaled RBF plus white noise, initial widths $.5, 2$, one optimizer restart.

Each grid is fitted on training rows and selected on selection rows. Cross-model comparisons do not isolate a symbolic-selection effect.

Symbolic backends and versions. The expanded implementation records Python 3.11.11, NumPy 1.26.4, SciPy 1.14.1, scikit-learn 1.7.2, gplearn .4.2 and PySR 1.5.9, with a Julia manifest. The V4 gplearn archive uses population 300, 20 generations, tournament 15 and parsimony .001, capped at 64 candidates. PySR uses 40 iterations, four populations of 30, 100 cycles, maximum size 20 and a 90-second directional timeout. V5 separately compares 5/10/20 gplearn generations and 15/45/90-second PySR budgets. Archives and native selections remain separate. Realized archives are shared within each task. Time limits can affect cross-machine reproducibility. Verified native LSIR superiority remains unestablished.

Task-mixture calibration. V5 freezes three pipelines and 11 thresholds, giving 33 acceptance rules. For e errors among m accepted calibration tasks, its upper bound is

$$U(e, m) = \begin{cases} 1, & m = 0 \text{ or } e = m, \\ \mathrm{Beta}^{-1}(1 - \delta/33, e + 1, m - e), & \text{otherwise.} \end{cases} \tag{29}$$

Select greatest coverage with $U \leq .10$, ties favoring the smaller threshold. If none qualifies, abstain everywhere. Conditional binomial coverage and a union bound justify selection over fixed rules for i.i.d. tasks independent of development. They do not detect confounding on an arbitrary new pair. Threshold .05 accepts 386/400 calibration tasks with 21 errors and bound .09753. On 400 fresh tasks it accepts 391 with 26 errors (6.65%). No retuning uses this test set. The separate certificate audit has zero definite violations over 240 directional fits, but dependence intervals span $[0, 1]$ and mean clipped-risk width is 8.803. Monte Carlo uncertainty is included.

Execution completeness and timing correction. V5 contains 8,267 cases and 171,075 evaluations with zero failed cases. V6 contains 3,300 scientific cases and 58,500 evaluations. Development and confirmation took 5:07 and 2:58:02 with peak batch memory 898,772 KiB. One V4 timing record was invalid because basis removal depended on query values. All 150 timings were recomputed with a fitting-frozen basis. The scientific coefficients and observations were unchanged. The V4 inventory comprises 6,166 cases and 89,650 evaluations.

Planned validation of the paired selector. A separate component study is planned with 768 fresh-seed cases and 24 variants (18,432 evaluations), a fixed V6 reference and 32 primary tests. It varies the stability test, risk band, candidate library, regularization, resampling budget and direction rule. These planned evaluations have no results in this paper and contribute no evidence to the reported conclusions.