CAURISK: WHEN TO TRUST CAUSAL DISCOVERY FOUNDATION MODELS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery foundation models can predict causal graphs for new datasets, but their confidence may not reflect graph accuracy when the underlying causal mechanisms change. We introduce CauRisk, a post-hoc framework that estimates graph-level prediction error under shifts in structural causal models (SCMs), without access to target ground-truth graphs at test time. Trained on synthetic source environments with known graphs, the risk estimator combines predictive uncertainty, stability across resampled datasets, and compatibility between the predicted graph and statistical patterns in the observed data. These diagnostics provide evidence about structural error without assuming that observational data can resolve non-identifiable causal directions. The discovery model remains frozen. CauRisk uses the estimated risk to decide which graph predictions to retain and which to defer. We specify an evaluation protocol that holds out entire causal mechanism and graph families, reserving target graphs exclusively for evaluation. The protocol evaluates CauRisk against diagnostic-specific and joint-feature baselines on failure detection, risk calibration, and risk–coverage trade-offs across two frozen discovery backbones. The results show that observable diagnostics can transfer useful graph-error rankings across SCM shifts, while numerical calibration and source-selected acceptance thresholds transfer less reliably.

1 INTRODUCTION

A predicted causal graph and a justified decision to rely on it are different outputs. Amortized discovery methods make the first increasingly accessible: AVICI learns structure inference from simulations, while Arrow uses large-scale synthetic pretraining for zero-shot discovery from tabular observations (Lorch et al., 2022; Thompson et al., 2026). The second remains a deployment question. When a pretrained model is applied to a new environment, how much structural error should be expected in its graph, and when should that graph be deferred for further assessment?

This question is especially consequential under shifts in structural causal models (SCMs). A model can assign confident edge scores to a wrong graph; resampling may reproduce the same error; and agreement with observational constraints cannot resolve every causal direction. These diagnostics describe different aspects of model behavior. Translating them into a judgment about the *whole graph* requires learning how observable evidence relates to structural error, and testing whether that relationship transfers across environments.

We study this relationship through *graph-level reliability estimation*. Synthetic source SCMs provide known graphs for supervising a risk estimator around a frozen discovery model. At test time, the estimator uses only observations and model outputs. Its target is the graph's structural error rather than an unfamiliarity score or an edgewise confidence summary. This separates reliability estimation from backbone adaptation and makes the resulting risk estimate directly relevant to retaining or deferring an entire graph.

CauRisk implements this perspective as a modular diagnostic framework. Uncertainty, resampling stability, and data–graph compatibility form distinct evidence views, each paired with structural metadata. Source-supervised experts map these views to a common total-error target. Mechanism-wise validation selects their combination, allowing unhelpful views to receive zero weight; disjoint source calibration then connects the combined score to an empirical deferral rule. This organization

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.

makes both the contribution of individual diagnostics and the transfer of the resulting decisions measurable.

A central finding is that *risk ranking, numerical calibration, and acceptance decisions exhibit different transfer behavior*. Across Arrow and AVICI, CauRisk detects graph loss above 0.5 with AUROC between 0.864 and 0.939 on same-family, mechanism-shift, and topology-shift tasks. These informative rankings coexist with shifts in numerical calibration and retained risk above the source budget under topology shift. Comparisons with joint-feature baselines show that the preferred diagnostic integration depends on the backbone and environment. The resulting picture connects graph-error prediction to the practical question of which graphs to accept.

Our contributions are threefold:

1. **Graph-level reliability estimation.** We formulate structural error as an explicit prediction target for frozen causal discovery models under SCM shift, distinguishing failure detection, numerical risk calibration, and selective graph acceptance.
2. **Post-hoc diagnostic risk estimation.** We introduce CauRisk, which combines diagnostic-specific risk experts through mechanism-wise source validation and source-only calibration. Its deployment interface requires neither target ground-truth graphs nor backbone updates.
3. **An analysis of reliability transfer.** We characterize the effects of observational ambiguity and distribution shift, and evaluate Arrow and AVICI on matched same-family, held-out-mechanism, and held-out-graph tasks. Across both backbones, the results distinguish transferable error rankings from the calibration and threshold behavior needed for selective acceptance.

2 Graph-Level Reliability from Observable Evidence

An environment e is an acyclic SCM on d observed variables,

$$X_j = f_{e,j}(X_{\mathrm{Pa}_{G_e}(j)}, \epsilon_j), \qquad j = 1, \dots, d, \tag{1}$$

with mutually independent exogenous noise variables. We consider continuous observations without latent confounding. A frozen discovery procedure F_θ, including preprocessing and graph decoding, maps discovery rows D_e^{disc} to a DAG $\widehat{G}_e$. Additional audit rows D_e^{audit} are used only for compatibility diagnostics. We evaluate this same graph, not a different graph refitted on the complete dataset.

The prediction target. An unordered variable pair has one of three states: no edge, a forward edge, or a backward edge. A disagreement costs one edit, including a reversal. The graph loss is

$$\ell_e = \frac{\mathrm{SHD}(\widehat{G}_e, G_e)}{d} = \frac{N_e^{\mathrm{extra}} + N_e^{\mathrm{missing}} + N_e^{\mathrm{reversal}}}{d}. \tag{2}$$

It is measured in edits per variable, not in probabilities. For fixed d, $0 \le \ell_e \le L_d = (d-1)/2$. A source-trained squared-error regressor targets $\mathbb{E}_{\Pi_{\mathrm{src}}}[\ell_e \mid \phi_e]$, where Π_{src} is the source environment distribution and ϕ_e contains observable diagnostics. Under SCM shift, this conditional expectation may change even if the diagnostics do not.

Uncertainty u. We summarize the frozen model's edge scores, probability entropies when probabilities are available, and directional margins. Predicted edges and nonedges are summarized separately. Arrow provides scores through a skeleton–order factorization, whereas AVICI predicts directed-edge probabilities; neither is treated as a calibrated posterior over whole DAGs. Each evidence array is summarized by its mean, standard deviation, and 10th, 50th, and 90th percentiles. Structural counts and missingness indicators accompany these summaries.

Stability s. For B bootstrap resamples of discovery rows, apply the same frozen model and decoder. A basic diagnostic is

$$s_e^{\mathrm{SHD}} = \frac{1}{Bd} \sum_{b=1}^{B} \mathrm{SHD}\left(\widehat{G}_e^{(b)}, \widehat{G}_e\right). \tag{3}$$

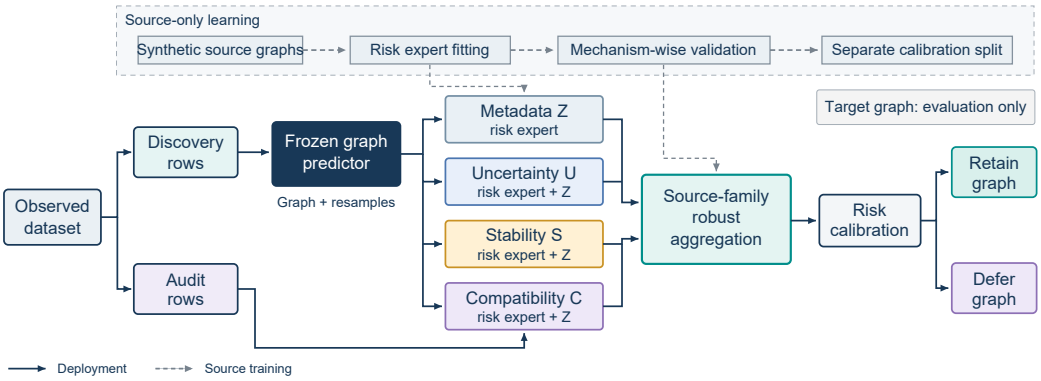


Figure 1: Source-supervised risk estimation around a frozen discovery model. Each diagnostic view predicts total graph error. Source mechanism groups determine aggregation weights; a disjoint source split determines calibration and the rejection threshold. The target graph is used only by the evaluator. Zero weights are allowed.

Other features summarize skeleton disagreement, orientation disagreement on common adjacencies, and edge-score variation. These features measure finite-sample sensitivity, not agreement with the unknown graph.

Compatibility c. For a predicted DAG, the local Markov property implies

$$X_i \perp X_j \mid X_{\mathrm{Pa}_{\widehat{G}_e}(i)}, \quad j \in \mathrm{ND}_{\widehat{G}_e}(i) \setminus \big(\mathrm{Pa}_{\widehat{G}_e}(i) \cup \{i\}\big). \tag{4}$$

We sample a bounded set of graph-implied queries and assess them on audit rows. Features include the fraction of small diagnostic p-values, log-p summaries, valid-query counts, and conditioning sizes. Two-fold out-of-sample kernel-ridge residuals and a randomized-feature permutation statistic provide approximate conditional-dependence diagnostics. Motivated by graph falsification (Eulig et al., 2025), we use these summaries as learned evidence about error rather than as a graph-level hypothesis test. Implementation details are given in Appendix B.

Metadata z. Sample size, variable count, predicted edge density, and predicted parent-set summaries help interpret the diagnostics. A dense graph can imply fewer testable constraints, so query counts must not be interpreted without structural context. Metadata-only prediction is a required control. In particular, none of these inputs includes the true edge count, target mechanism name, or target graph. Near-constant features are removed using source-training tasks only.

3 CauRisk: Post-hoc Diagnostic Risk Estimation

CauRisk maps observable diagnostics to a graph-error estimate and then to an acceptance decision. The diagnostic experts, aggregation weights, and calibration map are learned in separate source stages, allowing each stage's contribution to reliability transfer to be evaluated. Figure 1 shows the source-learning and deployment paths.

3.1 View-specific risk experts

Define four views $v_Z = z$, $v_U = (u, z)$, $v_S = (s, z)$, and $v_C = (c, z)$. Each expert predicts the same graph-level loss:

$$f_j \in \arg\min_f \frac{1}{|\mathcal{T}|} \sum_{e \in \mathcal{T}} \big(f(v_{j,e}) - \ell_e\big)^2, \qquad j \in \{Z, U, S, C\}. \tag{5}$$

The training set $\mathcal{T}$ contains balanced source SCM cells. We use ensembles of extremely randomized regression trees (Geurts et al., 2006). Leaf averages and convex aggregation keep predictions within the source-label range, which lies in $[0, L_d]$. The empirical implementation averages three seeds,

each with 128 trees and minimum leaf size five. These settings are fixed rather than selected from target results. All source validation and calibration tasks remain outside expert fitting.

The common total-error target allows each view to capture whichever structural failures it predicts in the source environments. No fixed correspondence is imposed between a diagnostic and an error component. The three counts in Equation (2) remain useful for failure analysis. Flat regressors on (u, s, c, z) and (u, s, z) provide direct controls for the effect of separating evidence views.

3.2 Mechanism-wise evidence aggregation

For weights w in the probability simplex, let

$$r_w(e) = \sum_j w_j f_j(v_{j,e}), \qquad w_j \geq 0, \quad \sum_j w_j = 1. \tag{6}$$

We evaluate weights on a disjoint source validation set, partitioned into mechanism groups $\mathcal{V}_g$. The selection rule is

$$\widehat{w} \in \arg\min_{w \in \mathcal{W}} \max_{g \leq K} \widehat{R}_g(w), \qquad \widehat{R}_g(w) = \frac{1}{|\mathcal{V}_g|} \sum_{e \in \mathcal{V}_g} \big(r_w(e) - \ell_e\big)^2. \tag{7}$$

Here $\mathcal{W}$ is the simplex grid with coordinates in $\{0, 0.1, \ldots, 1\}$; there are 286 four-view combinations. Ties are resolved by pooled validation MSE and then lexicographic weight order. The procedure minimizes prediction error rather than target AURC. It brings the worst-group perspective of Sagawa et al. (2020) to diagnostic weighting, using nonnegative prediction aggregation related to stacking (Breiman, 1996). Each source mechanism contributes to selection through its own prediction error, rather than only through a pooled average.

Source mechanism labels identify validation groups but are not inputs to the experts or the deployment rule. Validation uses source families represented in training; target mechanism and graph families are excluded from all source stages. Thus the criterion evaluates within-source family robustness, with transfer to held-out families assessed separately.

3.3 Calibration and deferral

A first source calibration partition fits a nonnegative isotonic map a to $r_{\widehat{w}}$. A second partition selects a threshold τ. The deployed score and decision are

$$S_e = a(r_{\widehat{w}}(e)), \qquad A_\tau(e) = \mathbb{I}[S_e \leq \tau]. \tag{8}$$

Candidate thresholds are score quantiles determined using only the map-fitting partition. We choose the highest-coverage candidate with at least ten accepted threshold-selection tasks and empirical mean loss at most α. If no candidate qualifies, the method defers all graphs. This rule enforces an empirical source-risk criterion; it differs from finite-sample risk-control procedures such as Angelopoulos et al. (2024).

For environment law Π, coverage is $C_\Pi(\tau) = \mathbb{P}_\Pi(A_\tau = 1)$ and selected risk is $\mathcal{R}_\Pi(\tau) = \mathbb{E}_\Pi[\ell \mid A_\tau = 1]$. Risk is undefined when coverage is zero. Raw scores determine ranking metrics; calibrated scores determine numerical calibration and fixed-threshold decisions. The source-fitted map and threshold remain fixed at test time.

4 Analysis of Graph-Level Reliability

The analysis distinguishes three questions: what observations can identify, what source-family validation can support, and what happens to selected risk under distribution shift. Together they specify the scope of a graph-reliability estimate without conflating it with causal identification.

Observational ambiguity under selective prediction. Observationally equivalent SCMs pose a limit even when a method is allowed to reject graphs. The following statement applies to any observation-based score and acceptance rule.

Proposition 1 (Selective ambiguity floor). *Let two SCMs with graphs G_0, G_1 on the same d variables induce identical observational distributions. Let a possibly randomized discovery procedure and an acceptance rule use only the observations and auxiliary randomness with the same law in both worlds. If their common acceptance probability is positive, then*

$$\max_{j\in\{0,1\}} \mathbb{E}_j\left[\frac{\mathrm{SHD}(\widehat{G}, G_j)}{d}\,\middle|\, A=1\right] \geq \frac{\mathrm{SHD}(G_0, G_1)}{2d}. \tag{9}$$

The proof couples the identical observable inputs and applies the triangle inequality conditionally on acceptance. For two observationally equivalent Gaussian models $X_1 \to X_2$ and $X_2 \to X_1$, the lower bound is $1/4$ edit per variable. Reducing coverage alone cannot provide a uniform guarantee below that bound across both worlds. This does not prohibit useful source-relative risk prediction or identification under additional functional, interventional, or background-knowledge assumptions.

Statistical control of source-family aggregation. Write $R_g(w) = \mathbb{E}_g[(r_w - \ell)^2/L_d^2]$ for normalized squared error in source family g. Conditioning on fitted experts, assume validation tasks are independent and represent the declared family mixtures.

Proposition 2 (Finite-grid validation bound). *Let n_g be the number of validation tasks in family g, and let $\widehat{w}$ minimize the largest empirical family error over finite $\mathcal{W}$. With probability at least $1-\delta$,*

$$\max_g R_g(\widehat{w}) \leq \min_{w\in\mathcal{W}} \max_g R_g(w) + 2\max_g \sqrt{\frac{\log(2K|\mathcal{W}|/\delta)}{2n_g}}. \tag{10}$$

The same right-hand side bounds the risk of $\widehat{w}$ under any mixture of these source family laws.

Bounded-loss concentration and a union bound yield the result. The bound relates validation sample size to the number of candidate mixtures and applies to the declared source families and their mixtures. At 32 validation tasks per mechanism, it serves as a finite-sample characterization; transfer beyond these families is assessed empirically.

Risk budgets depend on coverage and shift. Let P, Q be source and target joint distributions of the observable evidence and graph loss, with common fixed d.

Proposition 3 (Shift amplification after deferral). *If $\mathrm{TV}(P,Q) \leq \varepsilon$, $P(A) > 0$, $Q(A) > 0$, and $\mathbb{E}_P[\ell \mid A] \leq \alpha$ for $0 \leq \alpha \leq L_d$, then*

$$\mathbb{E}_Q[\ell \mid A] \leq \min\left\{L_d,\, \alpha + \frac{L_d\varepsilon}{Q(A)}\right\}. \tag{11}$$

The $1/Q(A)$ factor shows how selection amplifies sensitivity to distribution shift. Here the shift concerns the joint law of evidence and graph loss, so a small observed-data discrepancy alone does not bound it. Full proofs and an exact indistinguishable-pair construction are in Appendix A.

5 Experimental Design

The evaluation follows four questions: (Q1) Can observable diagnostics rank structural error on new tasks? (Q2) How does diagnostic integration affect transfer? (Q3) Do numerical risk estimates remain calibrated? (Q4) Do source-selected acceptance rules retain their risk budgets? These questions separate the utility of risk estimation from the performance of any one estimator.

Backbones and tasks. We freeze the public Arrow base and AVICI `scm-v0` checkpoints. Arrow uses its native DAG decoder; AVICI edges above probability 0.5 are considered in descending score order and accepted if they preserve acyclicity. Both backbones receive identical datasets and row splits. Every task has $d = 10$ variables and $n = 500$ observations, split into 350 discovery and 150 audit rows. Each graph is predicted once, with $B = 4$ bootstrap predictions. Compatibility uses at most 16 graph-implied queries and 99 residual permutations. Risk estimators are fit separately for each backbone; within a backbone, all methods share prediction and diagnostic caches.

Table 1: Graph-risk AURC on matched tasks (lower is better). Each backbone uses 108 same-family, 100 mechanism-shift, and 100 topology-shift tasks. All learned estimators include metadata. Wide models match the four-expert library's tree count; oracle ordering uses evaluation labels only.

	Arrow			AVICI		
Risk estimator	Same-family	Mechanism	Topology	Same-family	Mechanism	Topology
Metadata	0.7350	0.4997	1.0989	0.4923	0.3585	0.7611
Uncertainty	0.6804	0.4891	0.8253	0.4266	0.3019	0.6907
Stability	0.6788	0.4715	0.8206	0.4283	0.3035	0.6656
Compatibility	0.7251	0.4878	1.0518	0.4656	0.3314	0.7004
U+S (trees)	0.6866	0.4610	0.7910	0.4287	0.2938	0.6700
All-feature (trees)	0.6802	0.4607	0.7901	0.4237	0.2903	0.6658
U+S (wide trees)	0.6853	0.4612	0.7929	0.4288	0.2942	0.6670
All-feature (wide trees)	0.6784	0.4603	0.7908	0.4248	0.2919	0.6647
CauRisk	0.6726	0.4648	0.8444	0.4271	0.3001	0.6659
Random order	0.9111	0.8330	1.0830	0.7444	0.6300	0.9730
Oracle order	0.5578	0.4162	0.6757	0.3535	0.2405	0.5598

Source SCMs cross three mechanisms (linear, tanh MLP, random Fourier features), two graph generators (random forward-edge sampling and preferential attachment), and two noise laws (Gaussian and Laplace). The 12 cells supply 25 training, eight validation, and eight calibration tasks each: 300, 96, and 96 tasks. Four-way diagnostic aggregation uses the three mechanism groups in the 96 validation tasks. Calibration is split into 48 map-fitting and 48 threshold-selection tasks.

Polynomial and post-nonlinear mechanisms, and modular and layered graph generators, are held out from risk fitting, validation, and calibration. This defines shift relative to the risk estimator rather than the backbone's pretraining distribution. Density ranges are controlled, with metadata controls accounting for variation in predicted graph structure (Reisach et al., 2021; Ormaniec et al., 2025).

Evaluation splits. The paired evaluation comprises 308 tasks: 108 same-family, 100 held-out-mechanism, and 100 held-out-topology tasks. Simulation seeds are disjoint from source and development tasks; family definitions are fixed from development. Each backbone's source-fitted risk models and calibration are frozen before target scoring, and test graphs are read by the evaluator only after scores are written. Appendix B records the split construction and provenance; Appendix C reports the separate development cohort.

Comparators. For each backbone, we fit the metadata expert Z, uncertainty expert U, stability expert S, and compatibility expert C; each diagnostic expert also receives Z. Two flat models concatenate U+S+Z and U+S+C+Z. All six share source samples, three seeds, trees per seed, and leaf regularization. Larger flat controls use 512 trees per seed, matching the full four-expert library's 1,536 trees. Aggregation ablations remove C, replace worst-family selection with pooled MSE, or use equal weights. The full and no-C grids contain 286 and 66 candidates. Eight source-fitted neural controls are also evaluated on Arrow (Appendix D).

Measurements. We report graph-risk MAE before and after calibration, failure AUROC for $\ell > \eta$, and risk–coverage curves. For ascending raw-risk order, with realized losses $\ell_{(i)}$,

$$\mathrm{AURC} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{k}\sum_{i=1}^{k}\ell_{(i)}. \tag{12}$$

Lower values are better. The oracle sorts true loss only for evaluation; expected random-order AURC equals the overall mean loss. Fixed source thresholds are evaluated without target-label adjustment at budgets $\alpha \in \{0.25, 0.5\}$. Paired 95% intervals use 2,000 task bootstrap draws within SCM cells, conditional on the fitted models and declared families, without multiplicity correction.

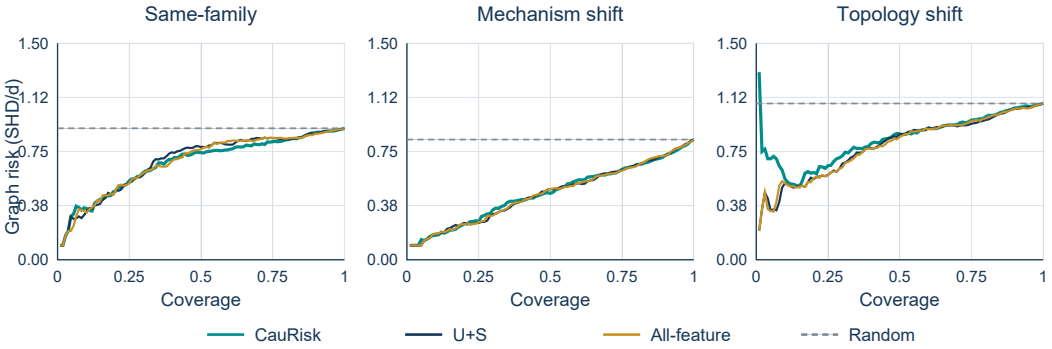


Figure 2: Risk–coverage curves for Arrow using raw risk scores. Lower curves are better; all methods converge to the same backbone error at full coverage. U+S and all-feature curves use the standard tree comparators. Dashed lines show expected random-order risk.

6 Results: From Risk Ranking to Acceptance Decisions

Q1. Observable diagnostics rank graph error across backbones. CauRisk separates lower- and higher-error graphs across all six backbone–test-group combinations (Table 1). Its AURC is 0.6726, 0.4648, and 0.8444 for Arrow, and 0.4271, 0.3001, and 0.6659 for AVICI, in same-family, mechanism-shift, and topology-shift order. Each is below the corresponding random-order reference. Failure AUROC for $\ell > 0.5$ ranges from 0.8659 to 0.9394 on Arrow and from 0.8636 to 0.9267 on AVICI. Thus graph-error information learned from source diagnostics remains useful under both mechanism and topology shifts. Comparisons between risk estimators use the same backbone predictions, so their differences reflect selection rather than changes to the predicted graphs.

Q2. The preferred integration depends on the backbone and shift. Source validation selects stability and compatibility weights of 0.7 and 0.3 for Arrow, and uncertainty, stability, and compatibility weights of 0.2, 0.6, and 0.2 for AVICI. These different combinations reflect how each backbone's diagnostics predict total graph error. For AVICI under topology shift, CauRisk improves over uncertainty alone by 0.0248 AURC (paired 95% interval for the difference: $[-0.0441, -0.0072]$) and performs similarly to joint-feature trees, at 0.6659 versus 0.6658. For Arrow, joint U+S trees obtain lower topology-shift AURC than CauRisk, at 0.7910 versus 0.8444; the paired difference is 0.0534 $[0.0145, 0.0849]$. All-feature trees also lead under AVICI mechanism shift, at 0.2903 versus 0.3001.

Increasing flat-model capacity preserves this pattern. The no-C and pooled-selection ablations have paired intervals overlapping zero relative to CauRisk in all three groups for each backbone (Appendices D and E). Joint diagnostics provide useful graph-risk information, while the incremental contribution of compatibility and the relative performance of aggregation rules vary across settings.

Q3. Numerical calibration behaves differently from ranking. Source isotonic calibration changes Arrow's MAE from 0.2683 to 0.2894 on same-family tasks, from 0.3346 to 0.3241 under mechanism shift, and from 0.3251 to 0.3083 under topology shift. The corresponding AVICI values are 0.2526 to 0.2622, 0.2596 to 0.2799, and 0.3058 to 0.2941. Both backbones overestimate risk under mechanism shift and underestimate it under topology shift: calibrated mean biases are $+0.1763/-0.1118$ for Arrow and $+0.1206/-0.0249$ for AVICI. Strong failure detection thus coexists with systematic changes in the meaning of a numerical risk estimate.

Q4. Selection gains and risk-budget transfer. At $\alpha = 0.5$, Arrow retains 23/108 same-family, 25/100 mechanism-shift, and 18/100 topology-shift graphs, with empirical risks 0.5087, 0.2640, and 0.6111. AVICI retains 47/108, 56/100, and 32/100 graphs, with risks 0.4447, 0.3179, and 0.5500. All six retained risks are below the respective full-coverage errors, but the topology-shift point estimates exceed the source budget for both backbones. Arrow's topology scores remain within the source calibration range, so score-range overlap alone does not establish threshold transfer. At $\alpha = 0.25$, neither backbone's CauRisk fit admits a source threshold and both defer all graphs.

These outcomes connect the empirical separation between ranking and acceptance to the coverage dependence in Proposition 3.

7 Related Work

Discovery and graph assessment. AVICI and Arrow amortize causal structure inference through synthetic training (Lorch et al., 2022; Thompson et al., 2026). We retain their distinction between observational evidence and learned inductive bias, while treating the frozen model's whole-graph error as a separate target. Bootstrap confidence has also been studied in time-series causal discovery (Debeire et al., 2024). Permutation-based graph falsification evaluates graph-implied independences against a graph reference distribution (Eulig et al., 2025); compatibility is one feature view here, not a replacement for its inferential procedure. Synthetic SCM design can introduce exploitable artifacts (Reisach et al., 2021; Ormaniec et al., 2025), motivating controlled task shapes and explicit metadata comparisons.

Predictive reliability. Calibration and its degradation under dataset shift are well established (Guo et al., 2017; Ovadia et al., 2019). Selective prediction studies risk as predictions are rejected (Geifman & El-Yaniv, 2019); our prediction unit is an entire causal graph. CauRisk uses established ensemble tools (Breiman, 1996; Geurts et al., 2006) with mechanism-wise robustness motivated by group DRO (Sagawa et al., 2020). Whereas conformal risk control studies calibration procedures with finite-sample guarantees (Angelopoulos et al., 2024), we examine the transfer of graph-error estimates and empirical source thresholds across SCM families. Propositions 1–3 connect this setting to causal ambiguity, source validation, and shift-sensitive selection.

8 Discussion and Conclusion

From structural predictions to reliability-aware use. CauRisk treats the quality of a predicted causal graph as a separate learning target. Its source-supervised diagnostic layer connects a frozen model's outputs to graph-level risk estimates and deferral decisions. The paired Arrow and AVICI experiments show that observable diagnostics carry useful information about structural error across the tested SCM shifts. They also identify an important deployment distinction: failure detection can transfer while numerical calibration and source-selected risk budgets change. Reliability assessment therefore requires evaluating the decision induced by a score, in addition to the score itself.

Scope and open directions. The study examines reliability transfer using two frozen discovery backbones, controlled task dimensions, and explicit SCM shifts. Joint-feature regressors and view-specific aggregation offer different trade-offs across these settings, motivating integration rules that adapt to the evidence available in each environment. The present evidence covers continuous synthetic tasks at $d = 10$ and $n = 500$, with family definitions fixed from development and 96 source calibration tasks. Broader task scales, scientific datasets, latent confounding, and feedback provide natural settings for testing how far the observed reliability patterns extend.

Graph-level reliability depends on both the information in a risk score and the decisions made from it. CauRisk provides a practical framework for measuring this relationship: learn error from observable diagnostics, evaluate its transfer across environments, and assess the risk of the graphs that remain accepted.

Reproducibility Statement

Appendix B specifies the SCM generators, diagnostic construction, model settings, calibration partitions, metric conventions, and artifact provenance. Arrow and AVICI share task inputs, while each has its own source-fitted risk estimators, calibration maps, and thresholds. Source-fitted artifacts are frozen before their respective target scoring runs. The accompanying evidence includes per-task losses and scores, and the experimental code provides protocol files and a label-free prediction interface. Appendix A provides complete proofs.

Ethics Statement

An estimated graph error is not authorization to intervene in a real system. Incorrect graphs may support misleading explanations, and deferring a graph leaves the need for additional evidence unresolved. The study uses synthetic observational datasets and existing pretrained software, with no new human-subject data collection. Scientific use requires domain-specific assessment beyond the synthetic reliability experiments.

AI Use Statement

Generative AI tools assisted with literature retrieval and discovery, research ideation and experimental implementation, manuscript drafting and editing, and the development and checking of mathematical arguments. An image-generation tool supplied a layout reference; all paper figures were subsequently redrawn as editable vectors, and all plotted values come from recorded experimental outputs. The authors manually reviewed and verified all AI-assisted code, experiments, mathematical claims, proofs, references, and manuscript content, and take responsibility for the final work.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996. doi: 10.1023/A:1018046112532. URL https://doi.org/10.1023/A:1018046112532.

Kevin Debeire, Andreas Gerhardus, Jakob Runge, and Veronika Eyring. Bootstrap aggregation and confidence measures to improve time series causal discovery. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 979–1007, 2024. URL https://proceedings.mlr.press/v236/debeire24a.html.

Elias Eulig, Atalanti A. Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26778–26786, 2025. doi: 10.1609/aaai.v39i25.34881. URL https://ojs.aaai.org/index.php/AAAI/article/view/34881.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.12934.

Weronika Ormaniec, Scott Sussex, Lars Lorch, Bernhard Schölkopf, and Andreas Krause. Standardizing structural causal models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ac3ec0493bec5ec5f15834f36149e422-Abstract-Conference.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.02530.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2102.13647.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Ryan Thompson, He Zhao, Daniel M. Steinberg, and Edwin V. Bonilla. Arrow: A foundation model for causal discovery. *arXiv preprint arXiv:2605.07204*, 2026. URL https://arxiv.org/abs/2605.07204.

A Proofs and an Indistinguishable Pair

A.1 Proof of Proposition 1

Couple the observational samples and the auxiliary random seeds identically in the two worlds. The discovery output $\widehat{G}$ and acceptance event A are then the same random objects under the common observable law. For every realization, the graph-edit metric satisfies

$$\mathrm{SHD}(\widehat{G}, G_0) + \mathrm{SHD}(\widehat{G}, G_1) \geq \mathrm{SHD}(G_0, G_1). \tag{13}$$

Condition on the event $A = 1$, which has the same positive probability in both worlds, and divide by d. The sum of the two retained risks is at least $\mathrm{SHD}(G_0, G_1)/d$. Their maximum is therefore at least half this value. Under our convention, SHD is Hamming distance on unordered-pair states and satisfies the triangle inequality. The argument does not require the fitted model or rejection rule to be deterministic.

For an explicit pair, fix $0 < |\rho| < 1$ and take independent standard normal noises. Consider

$$M_0: \quad X_1 = \epsilon_1, \qquad X_2 = \rho X_1 + \sqrt{1-\rho^2}\epsilon_2, \tag{14}$$

$$M_1: \quad X_2 = \widetilde{\epsilon}_2, \qquad X_1 = \rho X_2 + \sqrt{1-\rho^2}\widetilde{\epsilon}_1. \tag{15}$$

Both induce a centered Gaussian with unit marginal variances and covariance ρ. Their graphs differ by one reversal, so $d = 2$ and the lower bound is $1/4$. If a discovery model always outputs $X_1 \to X_2$, its loss is zero in M_0 and $1/2$ in M_1, independently of any observation-based selection. A procedure that randomly chooses the two orientations with equal probability attains retained risk $1/4$ in both worlds when it accepts independently, establishing tightness of the minimax floor for this pair.

The same construction also limits risk estimation itself. For any observable estimate T, with losses $L_j = \mathrm{SHD}(\widehat{G}, G_j)/d$, the pointwise inequality

$$|T - L_0| + |T - L_1| \geq |L_0 - L_1| \tag{16}$$

implies $\max_j \mathbb{E}_j|T - L_j| \geq \frac{1}{2}\mathbb{E}|L_0 - L_1|$ under the common observational law. The bound depends on the loss difference between the two worlds and vanishes when the predicted graph has equal loss in both.

A.2 Proof of Proposition 2

Condition on all training data and fitted experts. After division by L_d, their predictions, their convex combinations, and the graph loss lie in $[0, 1]$. Thus $Z_e(w) = (r_w(e) - \ell_e)^2/L_d^2$ lies in $[0, 1]$. For any family g and weight vector w, Hoeffding's inequality gives

$$\mathbb{P}\Big(\Big|\widehat{R}_g^{\mathrm{norm}}(w) - R_g(w)\Big| > \epsilon_g\Big) \leq 2\exp(-2n_g\epsilon_g^2). \tag{17}$$

The bound also applies to independent, non-identically distributed tasks when R_g denotes the corresponding balanced cell-average expectation. Choose $\epsilon_g = \sqrt{\log(2K|\mathcal{W}|/\delta)/(2n_g)}$ and take a union bound over all families and grid points. With probability at least $1 - \delta$, every deviation is bounded simultaneously. Let $\epsilon = \max_g \epsilon_g$ and let w^* minimize $\max_g R_g(w)$ over $\mathcal{W}$. On this event,

$$\max_g R_g(\widehat{w}) \leq \max_g \widehat{R}_g^{\mathrm{norm}}(\widehat{w}) + \epsilon \tag{18}$$

$$\leq \max_g \widehat{R}_g^{\mathrm{norm}}(w^*) + \epsilon \tag{19}$$

$$\leq \max_g R_g(w^*) + 2\epsilon. \tag{20}$$

Multiplication of the empirical objective by the constant L_d^{-2} does not change its minimizer. For a mixture $Q = \sum_g \pi_g P_g$, its normalized squared risk is $\sum_g \pi_g R_g(\widehat{w}) \leq \max_g R_g(\widehat{w})$, proving the mixture statement. No such domination is asserted for a family law outside this convex hull.

A.3 Proof of Proposition 3

Assume $0 \le \alpha \le L_d$; larger budgets give a trivial statement. Define $h = (\ell - \alpha)\mathbb{I}[A]$, whose range is contained in $[-\alpha, L_d - \alpha]$ and has length L_d. For total variation defined as $\sup_B |P(B) - Q(B)|$, bounded-function duality yields

$$\mathbb{E}_Q h \le \mathbb{E}_P h + L_d \,\mathrm{TV}(P, Q). \tag{21}$$

The source assumption gives $\mathbb{E}_P h = P(A)(\mathbb{E}_P[\ell \mid A] - \alpha) \le 0$. Hence $Q(A)(\mathbb{E}_Q[\ell \mid A] - \alpha) \le L_d \varepsilon$. Divide by the positive target coverage and intersect with the trivial upper bound L_d. This is a population comparison; empirical source-threshold selection does not establish its source premise exactly, and observed covariate shift does not by itself bound joint shift involving graph loss.

B Implementation and Evaluation Details

B.1 SCM generation

Each unique task seed drives graph, mechanism, noise, and variable-permutation generation. Offsets of one, two, and three initialize separate generators for the row split, bootstrap resampling, and diagnostic queries. Graphs are constructed in a forward order and then jointly permuted with their observations. For the random forward-edge family, the edge count is sampled from $\{d-2, \ldots, 2d\}$, subject to the number of possible pairs, and that many pairs are sampled without replacement. This fixed-count construction is not an independent-edge Erdős–Rényi draw. The preferential-attachment family uses an undirected Barabási–Albert graph with one or two links per new node, oriented by construction order.

Modular graphs give an eightfold sampling weight to pairs in the same consecutive three-node block; other pairs have unit weight. Layered graphs allow forward edges only across such blocks. The number of eligible pairs caps the sampled edge count. These generators change structural motifs and can also change realized density; they are not exactly density-matched counterfactuals.

For a node with p parents, signed coefficients have magnitudes uniformly sampled between 0.5 and 1.3. The linear signal is the weighted parent sum divided by $\sqrt{p}$. MLP signals use 12 tanh hidden units with randomized weights and biases. Random Fourier signals use 24 cosine features with randomized frequencies and phases. Polynomial signals combine a linear term with centered squared-parent terms. The post-nonlinear mechanism applies $\mathrm{arcsinh}$ after adding noise to a linear parent signal. Noise is Gaussian or variance-standardized Laplace, with task noise scale sampled between 0.7 and 1.3. Observed variables receive independent positive scale factors after generation. The experiment therefore does not eliminate every scale-related shortcut; metadata and generator limitations remain explicit.

The risk-training mechanism families are linear, MLP, and random Fourier features. Polynomial and post-nonlinear tasks are held out from source risk fitting, validation, and calibration. Modular and layered graphs are held out in the graph-family experiment. Same-family tests use new SCM draws from the source generator cells. All tasks have 500 observations and ten variables.

B.2 Frozen backbones and evidence budget

The Arrow implementation is fixed at code revision `8487278a372f`, with the public base checkpoint revision `49a256f80fd3`. Full revision identifiers and the checkpoint SHA256 are recorded in the machine-readable evidence manifest. The checkpoint parameters have gradients disabled. The native decoder supplies acyclic predictions and edge scores, and each bootstrap uses the same preprocessing and decoder. The risk estimator never modifies these graphs.

AVICI uses the public `scm-v0` checkpoint from `larslorch/avici`, revision `3bcf141a2d01`, with AVICI 1.0.7, JAX 0.4.35, and jaxlib 0.4.34 under Python 3.12. Inference uses float32 and the highest matrix-multiplication precision setting. A fixed decoder sorts directed edges with probability strictly above 0.5 in descending order, resolves ties by row-major index, and skips edges that would create a cycle. The same rule applies to the original and bootstrap predictions. Before cycle removal, five source and four target task predictions contain cycles; each loses one edge. All graph losses and compatibility queries refer to the decoded DAG. Complete checkpoint identifiers and hashes are included in the evidence files.

The discovery/audit partition contains 350/150 rows. Four resamples are drawn from discovery rows only. Uncertainty and stability summaries, metadata, and compatibility produce 74 raw feature columns. Constant-column filtering is fit only on the 300 source-training tasks. No test-family labels, true graph statistics, or graph-error counts enter a feature view. Empty arrays have explicit missingness indicators. A graph with no valid audit queries is not treated as perfectly verified.

Compatibility queries are local Markov implications of the predicted graph. Residual regression uses two-fold cross-fitting within audit observations. Random-feature dependence checks use 99 permutations and a nominal small-p cutoff of 0.05. Their finite-sample conditional validity is not established by this implementation. Query count, conditioning dimension, and invalid-test summaries accompany the diagnostic values to expose variation in test opportunity and difficulty.

B.3 Risk estimators and decisions

Standard tree-based experts and flat comparators use three independent seeds, 128 extremely randomized trees per seed, minimum leaf size five, all available features considered at each split, and no bootstrap sampling of training tasks. Tree cut points are randomized. Larger flat controls use 512 trees per seed with otherwise identical settings. Experts average total-error labels directly. Neural controls compare total-error and three-component prediction, using source-fitted weights, calibration maps, and thresholds held fixed throughout evaluation.

The 96 source validation tasks contain 32 tasks for each of the three mechanism groups, balanced across graph/noise cells. These validation tasks select only aggregation weights; they do not fit tree leaves. The four-view grid contains $\binom{13}{3} = 286$ points and the three-view no-C grid contains $\binom{12}{2} = 66$ points. The primary criterion is largest mechanism-wise MSE, followed by pooled MSE and a deterministic lexicographic tie rule. Pooled aggregation instead minimizes overall MSE. Equal aggregation assigns weight $1/4$ to each expert. A zero weight is a legitimate outcome, not replaced by a small positive constant.

Source calibration tasks with even within-cell replicate indices fit the isotonic map; odd indices select the threshold. Each partition contains 48 tasks. The threshold grid is the unique set of calibrated map-partition score quantiles at increments of 0.05. An admissible threshold must retain at least ten threshold-partition tasks and meet the empirical risk budget. A method with no admissible threshold defers all target graphs; it is assigned zero coverage and no numerical selected-risk value.

B.4 Task protocol and provenance

The Arrow aggregation protocol, source-selected weights, isotonic maps, and thresholds were serialized and hashed before evaluation-task collection; its larger flat and neural controls were fixed before evaluation on these task labels. The source seed schedule begins at 20260924. The evaluation schedule begins at 83000117 and increments by 1009 per task; all 308 seeds are distinct from those in the original 800 source and development tasks. Family definitions are shared with development, while SCM draws and observations are independent.

The cross-backbone design uses these same 492 source and 308 evaluation datasets for AVICI, with identical observations, row partitions, and ground-truth graphs. AVICI risk models are trained using their own source diagnostics and graph losses under the same fitting rules. They are frozen before AVICI target scoring. Thus the experiment compares reliability patterns across separately fitted backbone-specific estimators on matched tasks, rather than transferring one risk estimator's parameters between backbones. Task-level input equality, decoded graph losses, and recorded risk scores are checked against the released evidence.

Target scoring accepts only feature dictionaries and the frozen risk artifact. It writes all scores before reading graph-error labels. Ground-truth adjacencies are stored separately for evaluation and integrity checks. The 308 development tasks informed model design and are excluded from Table 1. The evidence records separate source fitting, validation, calibration, development, and evaluation partitions.

B.5 METRICS AND CONFIDENCE INTERVALS

Stable sorting by raw risk and then task order resolves score ties. Risk at a fixed coverage c uses $k = \lceil cN \rceil$ retained tasks; AURC averages all N nonzero coverages. Graph error remains normalized by d, and reversal costs one edit. Isotonic calibration can introduce ties, so raw scores rather than calibrated ties determine ranking metrics. Failure AUROC and average precision use thresholds $\eta \in \{0.25, 0.5, 1.0\}$; they are undefined if only one class is present.

Calibration summaries report raw and calibrated MAE, mean prediction bias, source-defined reliability bins, and the fraction of raw target scores outside the source map range. Accepted-risk reporting includes the number retained, coverage, mean graph loss, and excess over the declared budget. The source rule is applied unchanged within target subgroups; no subgroup-specific target tuning is performed.

Bootstrap comparisons resample task IDs within mechanism/graph/noise cells and preserve the empirical mixture. Every comparator uses the same sampled IDs within a contrast. The 95% percentile intervals are descriptive and unadjusted for multiple comparisons. Three expert seeds are averaged into one deployed predictor; the intervals condition on those fitted predictors rather than treating their seeds as independent test samples.

C DEVELOPMENT RESULTS AND MODEL CHOICE

Table 2 compares total-error and component-error prediction on 308 development tasks with Arrow. The neural models use two hidden ReLU layers, nonnegative outputs, three seeds, source-only model selection, and the same cached diagnostic inputs. Single-output models fit total graph error. Three-output models predict extra, missing, and reversed-edge errors using a component-sum squared loss plus a weighted total-error loss. Hyperparameters are selected using source-mechanism holdout refits.

Table 2: Arrow development-task AURC, lower is better. This cohort is separate from the paired evaluation in Table 1. Every diagnostic model includes metadata.

Neural estimator	Same-family	Mechanism	Topology
U+S, single output	0.6782	0.4805	0.7768
U+S, three outputs	0.6824	0.5119	0.7780
U+S+C, single output	0.6922	0.4833	0.7732
U+S+C, three outputs	0.6781	0.5343	0.7894

Compatibility improves the single-output model's topology-shift AURC, while same-family and mechanism-shift AURC increase. The three-output all-feature model has higher mechanism-shift AURC than its U+S single-output comparator, with a paired difference of 0.0539 and an exploratory 95% interval $[0.0188, 0.0933]$. Predicting error components separately therefore changes transfer behavior without a consistent ranking advantage in this comparison.

Under mechanism shift, mean missing-edge error is 0.116 edits per variable, while the all-feature three-output model predicts 0.273. A source-selected budget of 0.5 retains 27% of topology-shift tasks at empirical mean loss 0.633, with loss 0.725 in the modular subgroup. These component and threshold measurements distinguish numerical error estimation from graph-risk ranking.

D ARROW RELIABILITY MEASUREMENTS

The Arrow evaluator reports separate same-family, mechanism-shift, and topology-shift groups. Mechanism results are partitioned into polynomial and post-nonlinear tasks; topology results are partitioned into modular and layered tasks. Every subgroup uses the same source-fitted risk models and thresholds as its parent group.

Table 3 includes all twenty estimators, including eight neural controls. Tables 4–6 separate absolute risk error from failure detection. Table 7 gives selected paired contrasts; the machine-readable

Table 3: AURC for every frozen estimator on fresh tasks; lower is better. Neural three-output models predict component errors and sum them. Random and oracle references are in Table 1.

Risk estimator	Same-family	Mechanism	Topology
Metadata	0.7350	0.4997	1.0989
Uncertainty	0.6804	0.4891	0.8253
Stability	0.6788	0.4715	0.8206
Compatibility	0.7251	0.4878	1.0518
U+S (trees)	0.6866	0.4610	0.7910
All-feature (trees)	0.6802	0.4607	0.7901
U+S (wide trees)	0.6853	0.4612	0.7929
All-feature (wide trees)	0.6784	0.4603	0.7908
MLP: metadata	0.7571	0.4674	1.0944
MLP: uncertainty	0.7228	0.4913	0.8799
MLP: stability	0.6831	0.4698	0.9209
MLP: compatibility	0.7585	0.4698	1.0524
MLP: U+S	0.6723	0.4643	0.8019
MLP: U+S, three outputs	0.6686	0.4929	0.7950
MLP: all-feature	0.6907	0.4684	0.8077
MLP: all-feature, three outputs	0.6650	0.5084	0.8118
Aggregation without C	0.6677	0.4638	0.8447
Pooled aggregation	0.6768	0.4688	0.8394
Equal aggregation	0.6637	0.4639	0.8682
CauRisk	0.6726	0.4648	0.8444

Table 4: Same-family: numerical risk error and failure detection. MAE and mean prediction bias use edits per variable; AUROC detects loss above 0.5. Lower MAE and higher AUROC are better. Bias is calibrated prediction minus realized loss.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.3017	0.3113	0.0430	0.7168
Uncertainty	0.2655	0.3157	-0.0382	0.7989
Stability	0.2677	0.2859	-0.0183	0.8562
Compatibility	0.2980	0.3092	0.0159	0.7744
U+S (trees)	0.2696	0.2998	-0.0011	0.8313
All-feature (trees)	0.2671	0.2883	-0.0291	0.8449
U+S (wide trees)	0.2689	0.3044	-0.0091	0.8335
All-feature (wide trees)	0.2664	0.2927	-0.0348	0.8449
MLP: metadata	0.3332	0.3147	-0.0094	0.6798
MLP: uncertainty	0.3003	0.3231	-0.0567	0.7432
MLP: stability	0.3333	0.3000	0.0162	0.8483
MLP: compatibility	0.3270	0.3308	-0.0193	0.6830
MLP: U+S	0.2694	0.2836	-0.0023	0.8608
MLP: U+S, three outputs	0.2551	0.2945	-0.0798	0.8665
MLP: all-feature	0.2947	0.3173	-0.0644	0.8102
MLP: all-feature, three outputs	0.2504	0.3044	-0.1004	0.8631
Aggregation without C	0.2689	0.2873	-0.0171	0.8585
Pooled aggregation	0.2670	0.2872	-0.0173	0.8602
Equal aggregation	0.2714	0.2736	-0.0073	0.8460
CauRisk	0.2683	0.2894	-0.0279	0.8659

evidence includes every specified comparison. Table 8 reports the source rule within each shift subtype. Group labels determine reporting only, not prediction inputs.

Table 5: Mechanism shift: numerical risk error and failure detection. MAE and mean prediction bias use edits per variable; AUROC detects loss above 0.5. Lower MAE and higher AUROC are better. Bias is calibrated prediction minus realized loss.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.3656	0.3491	0.2279	0.8746
Uncertainty	0.3481	0.3405	0.1527	0.9358
Stability	0.3284	0.3090	0.1507	0.9291
Compatibility	0.3565	0.3498	0.2315	0.9042
U+S (trees)	0.3164	0.2992	0.1363	0.9590
All-feature (trees)	0.3144	0.2922	0.1328	0.9577
U+S (wide trees)	0.3162	0.2953	0.1244	0.9608
All-feature (wide trees)	0.3125	0.2964	0.1340	0.9586
MLP: metadata	0.3616	0.3141	0.2059	0.9583
MLP: uncertainty	0.3673	0.3343	0.2207	0.9501
MLP: stability	0.3599	0.3539	0.2178	0.9416
MLP: compatibility	0.3592	0.3182	0.1979	0.9456
MLP: U+S	0.3097	0.3284	0.1651	0.9479
MLP: U+S, three outputs	0.2978	0.3174	0.0026	0.9403
MLP: all-feature	0.3351	0.3501	0.2224	0.9332
MLP: all-feature, three outputs	0.3293	0.3204	0.0058	0.9367
Aggregation without C	0.3393	0.3241	0.1884	0.9465
Pooled aggregation	0.3305	0.3160	0.1660	0.9336
Equal aggregation	0.3437	0.3347	0.2430	0.9603
CauRisk	0.3346	0.3241	0.1763	0.9394

Table 6: Topology shift: numerical risk error and failure detection. MAE and mean prediction bias use edits per variable; AUROC detects loss above 0.5. Lower MAE and higher AUROC are better. Bias is calibrated prediction minus realized loss.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.4121	0.4450	-0.0577	0.4741
Uncertainty	0.3322	0.3312	-0.0917	0.8871
Stability	0.3068	0.2965	-0.1091	0.8918
Compatibility	0.4011	0.4152	-0.1200	0.4580
U+S (trees)	0.3017	0.2912	-0.1048	0.9090
All-feature (trees)	0.2982	0.2802	-0.1168	0.9075
U+S (wide trees)	0.3005	0.2889	-0.1094	0.9106
All-feature (wide trees)	0.2989	0.2907	-0.1171	0.9043
MLP: metadata	0.4206	0.4387	-0.1020	0.5494
MLP: uncertainty	0.3785	0.3782	-0.1603	0.7969
MLP: stability	0.4384	0.3688	-0.1100	0.7812
MLP: compatibility	0.4129	0.4199	-0.1132	0.5780
MLP: U+S	0.2951	0.3009	-0.1205	0.8980
MLP: U+S, three outputs	0.2895	0.3019	-0.1432	0.8949
MLP: all-feature	0.3696	0.3437	-0.1459	0.8933
MLP: all-feature, three outputs	0.3109	0.3389	-0.1461	0.8447
Aggregation without C	0.3262	0.3096	-0.1112	0.8682
Pooled aggregation	0.3106	0.3044	-0.1129	0.8878
Equal aggregation	0.3514	0.3393	-0.1108	0.8337
CauRisk	0.3251	0.3083	-0.1118	0.8824

Table 7: Paired AURC differences: CauRisk minus comparator. Negative values favor CauRisk. Intervals are cell-stratified 95% percentile intervals from 2,000 bootstrap samples, without multiplicity correction.

Group	Comparator	Difference	95% interval
Same-family	U+S (trees)	-0.0140	[-0.0419, 0.0074]
Same-family	All-feature (trees)	-0.0076	[-0.0296, 0.0105]
Same-family	U+S (wide trees)	-0.0127	[-0.0389, 0.0085]
Same-family	All-feature (wide trees)	-0.0058	[-0.0254, 0.0105]
Same-family	MLP: U+S	0.0003	[-0.0305, 0.0302]
Same-family	MLP: all-feature	-0.0181	[-0.0506, 0.0148]
Same-family	Aggregation without C	0.0049	[-0.0064, 0.0178]
Same-family	Pooled aggregation	-0.0042	[-0.0149, 0.0050]
Mechanism	U+S (trees)	0.0038	[-0.0057, 0.0139]
Mechanism	All-feature (trees)	0.0041	[-0.0045, 0.0131]
Mechanism	U+S (wide trees)	0.0036	[-0.0054, 0.0137]
Mechanism	All-feature (wide trees)	0.0045	[-0.0041, 0.0138]
Mechanism	MLP: U+S	0.0005	[-0.0113, 0.0120]
Mechanism	MLP: all-feature	-0.0036	[-0.0184, 0.0099]
Mechanism	Aggregation without C	0.0009	[-0.0042, 0.0057]
Mechanism	Pooled aggregation	-0.0040	[-0.0102, 0.0009]
Topology	U+S (trees)	0.0534	[0.0145, 0.0849]
Topology	All-feature (trees)	0.0543	[0.0164, 0.0858]
Topology	U+S (wide trees)	0.0515	[0.0133, 0.0827]
Topology	All-feature (wide trees)	0.0536	[0.0152, 0.0847]
Topology	MLP: U+S	0.0425	[-0.0065, 0.0859]
Topology	MLP: all-feature	0.0368	[-0.0191, 0.0831]
Topology	Aggregation without C	-0.0003	[-0.0200, 0.0180]
Topology	Pooled aggregation	0.0051	[-0.0064, 0.0179]

Table 8: CauRisk with unchanged source-selected thresholds. Risk is undefined when no task is retained (shown as a dash). A positive excess indicates failure to meet the nominal source risk budget on target tasks.

Group	Budget	Retained	Coverage	Risk	Excess
Same-family	0.25	0	0.0000	–	–
Same-family	0.5	23	0.2130	0.5087	0.0087
Mechanism (all)	0.25	0	0.0000	–	–
Mechanism (all)	0.5	25	0.2500	0.2640	0.0000
Polynomial	0.25	0	0.0000	–	–
Polynomial	0.5	13	0.2600	0.2615	0.0000
Post-nonlinear	0.25	0	0.0000	–	–
Post-nonlinear	0.5	12	0.2400	0.2667	0.0000
Topology (all)	0.25	0	0.0000	–	–
Topology (all)	0.5	18	0.1800	0.6111	0.1111
Layered	0.25	0	0.0000	–	–
Layered	0.5	8	0.1600	0.6125	0.1125
Modular	0.25	0	0.0000	–	–
Modular	0.5	10	0.2000	0.6100	0.1100

E AVICI RELIABILITY MEASUREMENTS

Tables 9–14 report the twelve risk estimators on AVICI using the same task groups as Arrow. Each diagnostic expert includes structural metadata. Raw scores determine AURC and failure detection; source-calibrated scores determine MAE, bias, and acceptance at a fixed threshold.

Worst-family validation selects weights $(0, 0.2, 0.6, 0.2)$ for (Z, U, S, C), with worst-mechanism MSE 0.1209. Removing compatibility gives weights $(0, 0.4, 0.6)$ and MSE 0.1218. Pooled validation selects $(0, 0.6, 0.4, 0)$, with worst-mechanism MSE 0.1226. The corresponding Arrow weights are $(0, 0, 0.7, 0.3)$, $(0.3, 0.1, 0.6)$, and $(0, 0, 0.9, 0.1)$, with worst-mechanism MSE 0.0916, 0.0981, and 0.0995. Metadata remains an input to each diagnostic expert even when the standalone Z weight is zero.

All-feature trees achieve lower point-estimate AURC than U+S trees in each AVICI group. Relative to CauRisk, their mechanism-shift difference is 0.0098 in favor of the flat model, with paired interval $[0.0012, 0.0184]$. The topology comparison is nearly equal, at 0.6658 versus 0.6659. These results complement the Arrow comparisons by separating the information in the joint diagnostic set from the performance of a particular aggregation rule.

Table 9: AVICI AURC for all twelve risk estimators; lower is better.

Estimator	Same-family	Mechanism	Topology
Metadata	0.4923	0.3585	0.7611
Uncertainty	0.4266	0.3019	0.6907
Stability	0.4283	0.3035	0.6656
Compatibility	0.4656	0.3314	0.7004
U+S (trees)	0.4287	0.2938	0.6700
All-feature (trees)	0.4237	0.2903	0.6658
U+S (wide trees)	0.4288	0.2942	0.6670
All-feature (wide trees)	0.4248	0.2919	0.6647
CauRisk	0.4271	0.3001	0.6659
Aggregation without C	0.4264	0.2989	0.6728
Pooled aggregation	0.4244	0.2985	0.6789
Equal aggregation	0.4358	0.3051	0.6796

Table 10: AVICI, same-family: numerical risk error and failure detection. Bias is calibrated prediction minus graph loss; AUROC detects loss above 0.5.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.3214	0.3313	0.0555	0.7819
Uncertainty	0.2431	0.2457	0.0284	0.8679
Stability	0.2507	0.2855	0.0475	0.8689
Compatibility	0.2985	0.3093	0.0419	0.8271
U+S (trees)	0.2452	0.2444	0.0396	0.8653
All-feature (trees)	0.2427	0.2384	0.0253	0.8726
U+S (wide trees)	0.2450	0.2439	0.0400	0.8668
All-feature (wide trees)	0.2424	0.2368	0.0432	0.8719
CauRisk	0.2526	0.2622	0.0547	0.8700
Aggregation without C	0.2433	0.2569	0.0359	0.8722
Pooled aggregation	0.2426	0.2457	0.0361	0.8755
Equal aggregation	0.2708	0.2566	0.0551	0.8613

Table 11: AVICI, mechanism shift: numerical risk error and failure detection. Bias is calibrated prediction minus graph loss; AUROC detects loss above 0.5.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.3082	0.3198	0.0021	0.8180
Uncertainty	0.2553	0.2521	0.0434	0.9151
Stability	0.2732	0.3017	0.1361	0.9251
Compatibility	0.2776	0.3046	0.0288	0.8764
U+S (trees)	0.2521	0.2728	0.1104	0.9340
All-feature (trees)	0.2509	0.2782	0.1251	0.9396
U+S (wide trees)	0.2518	0.2724	0.1047	0.9344
All-feature (wide trees)	0.2513	0.2757	0.1294	0.9372
CauRisk	0.2596	0.2799	0.1206	0.9267
Aggregation without C	0.2612	0.2699	0.1151	0.9287
Pooled aggregation	0.2580	0.2607	0.0995	0.9255
Equal aggregation	0.2603	0.2989	0.0992	0.9122

Table 12: AVICI, topology shift: numerical risk error and failure detection. Bias is calibrated prediction minus graph loss; AUROC detects loss above 0.5.

Estimator	Raw MAE	Cal. MAE	Bias	AUROC
Metadata	0.3566	0.3629	-0.0187	0.7509
Uncertainty	0.3307	0.3474	-0.0894	0.8240
Stability	0.3033	0.3274	-0.0390	0.8782
Compatibility	0.3184	0.3190	-0.0346	0.8100
U+S (trees)	0.3102	0.3124	-0.0574	0.8788
All-feature (trees)	0.3049	0.3005	-0.0601	0.8730
U+S (wide trees)	0.3089	0.3076	-0.0557	0.8730
All-feature (wide trees)	0.3038	0.2963	-0.0506	0.8718
CauRisk	0.3058	0.2941	-0.0249	0.8636
Aggregation without C	0.3110	0.3187	-0.0595	0.8636
Pooled aggregation	0.3165	0.3059	-0.0477	0.8537
Equal aggregation	0.3192	0.3094	-0.0012	0.8526

Table 13: AVICI CauRisk with source-selected thresholds. A dash denotes undefined risk at zero coverage. Excess is the positive part of empirical risk minus budget.

Group	Budget	Retained	Coverage	Risk	Excess
Same-family	0.25	0	0.0000	–	–
Same-family	0.5	47	0.4352	0.4447	0.0000
Mechanism	0.25	0	0.0000	–	–
Mechanism	0.5	56	0.5600	0.3179	0.0000
Topology	0.25	0	0.0000	–	–
Topology	0.5	32	0.3200	0.5500	0.0500

Table 14: AVICI paired AURC differences: CauRisk minus comparator. Negative values favor CauRisk. Intervals use 2,000 cell-stratified bootstrap samples, without multiplicity correction.

Group	Comparator	Difference	95% interval
Same-family	U+S (trees)	-0.0016	[-0.0133, 0.0097]
Same-family	All-feature (trees)	0.0034	[-0.0084, 0.0156]
Same-family	U+S (wide trees)	-0.0016	[-0.0129, 0.0093]
Same-family	All-feature (wide trees)	0.0023	[-0.0085, 0.0133]
Same-family	Aggregation without C	0.0007	[-0.0057, 0.0082]
Same-family	Pooled aggregation	0.0027	[-0.0066, 0.0131]
Same-family	Equal aggregation	-0.0087	[-0.0201, 0.0009]
Same-family	Uncertainty	0.0006	[-0.0141, 0.0159]
Same-family	Stability	-0.0011	[-0.0074, 0.0061]
Mechanism	U+S (trees)	0.0063	[-0.0032, 0.0150]
Mechanism	All-feature (trees)	0.0098	[0.0012, 0.0184]
Mechanism	U+S (wide trees)	0.0059	[-0.0031, 0.0146]
Mechanism	All-feature (wide trees)	0.0082	[-0.0002, 0.0164]
Mechanism	Aggregation without C	0.0012	[-0.0043, 0.0064]
Mechanism	Pooled aggregation	0.0017	[-0.0064, 0.0091]
Mechanism	Equal aggregation	-0.0050	[-0.0168, 0.0069]
Mechanism	Uncertainty	-0.0017	[-0.0151, 0.0102]
Mechanism	Stability	-0.0034	[-0.0139, 0.0053]
Topology	U+S (trees)	-0.0041	[-0.0177, 0.0126]
Topology	All-feature (trees)	0.0001	[-0.0110, 0.0117]
Topology	U+S (wide trees)	-0.0011	[-0.0140, 0.0123]
Topology	All-feature (wide trees)	0.0011	[-0.0106, 0.0129]
Topology	Aggregation without C	-0.0069	[-0.0158, 0.0054]
Topology	Pooled aggregation	-0.0130	[-0.0239, 0.0003]
Topology	Equal aggregation	-0.0137	[-0.0254, 0.0016]
Topology	Uncertainty	-0.0248	[-0.0441, -0.0072]
Topology	Stability	0.0003	[-0.0123, 0.0165]

F DEPLOYMENT PROCEDURE

Algorithm 1 CauRisk scoring without target graphs

Require: Observations D, frozen discovery model F_θ, source-fitted experts f_j, weights $\widehat{w}$, map a, threshold τ

Split D into discovery and audit observations using the fixed rule.
Predict $\widehat{G}$ and edge scores using F_θ on discovery rows.
Predict four bootstrap graphs using the same frozen procedure.
Extract uncertainty u, stability s, and metadata z.
Test sampled graph-implied queries on audit rows to form compatibility diagnostics c.
Compute $r = \sum_j \widehat{w}_j f_j(v_j)$ and calibrated risk $S = a(r)$.
if a source-admissible τ exists and $S \leq \tau$ **then**
    return $(\widehat{G}, S, \mathrm{retain})$.
else
    return $(\widehat{G}, S, \mathrm{defer})$.
end if