SAME CAUSAL SYSTEM, DIFFERENT GRAPHS: AUDITING PRETRAINED CAUSAL DISCOVERY ACROSS MEASUREMENT COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained causal discovery systems predict graph structure for new datasets, but aggregate accuracy does not establish whether their reported relationships persist across measurement representations. We introduce OrbitBench, a paired evaluation protocol that re-expresses the same observations through strictly increasing, invertible transformations of each variable while keeping the inference pipeline fixed. We evaluate three pretrained systems, with PC and GES as classical references, on linear and nonlinear synthetic tasks and measurements from one physical light-tunnel apparatus. On 16 controlled linear tasks, non-decreasing skeleton F1 can coexist with changed adjacencies; among the pretrained systems, Arrow remains near-best in both views while changing its skeleton in 63 of 64 task–view comparisons. Disagreement persists when predicted edge counts are fixed and extends to directed-path answers on fresh tasks. On 32 new nonlinear tasks, using two audit views, audit-assisted and stability-first selection reduce macroaverage held-out adjacency disappearance risk from 8.27% for score-only selection to 6.71% and 6.68%, respectively. These gains concern persistence within the tested transformation families, not established improvements in causal correctness. Our results motivate reporting relationship consistency alongside accuracy and model-choice summaries.

1 INTRODUCTION

Causal discovery, the inference of causal relationships from data, is an important tool for understanding complex systems, including cellular signaling networks (Sachs et al., 2005), Earth-system interactions (Runge et al., 2019), and controlled physical processes (Gamella et al., 2025). Classical methods include constraint-based procedures such as PC (Spirtes & Glymour, 1991) and score-based search such as GES (Chickering, 2002); subsequent work has introduced continuous optimization with NOTEARS (Zheng et al., 2018) and neural mechanism models with GraN-DAG (Lachapelle et al., 2020).

More recently, pretrained and amortized approaches to causal discovery have emerged, including AVICI (Lorch et al., 2022), Arrow (Thompson et al., 2026), and CDFM (Qiao et al., 2026). Whereas the earlier methods estimate structure separately for each dataset, these systems learn across simulated causal problems and reuse the learned parameters to predict graph structure for a new dataset (Lorch et al., 2022; Thompson et al., 2026). For practitioners, this makes it possible to apply the same inference model to multiple datasets without training a separate predictor for each one (Thompson et al., 2026).

Applying these reusable models requires assessing not only their accuracy but also the consistency of the relationships they report across measurement representations. Their graph predictions contain individual claims about relationships among variables, each of which can be examined across different representations of the same observations. Aggregate accuracy and model-choice summaries do not, by themselves, establish whether these individual claims persist. This matters in scientific applications: directed paths inform hypotheses about how changes might propagate through a system (Peters et al., 2017), and inferred signaling relationships can motivate targeted biological experiments (Sachs et al., 2005). We therefore ask: when the same observations are re-expressed

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.

through invertible, variable-wise transformations, does a fixed model still report the same causal relationships?

Prior work examines reparameterization sensitivity (Jørgensen & Hauberg, 2020) and robustness to assumption violations (Yi et al., 2025; Stein et al., 2026). We examine how often coordinate-dependent graph changes coexist with non-decreasing accuracy or retention of a near-best model, and whether these changes affect graph-derived answers. To investigate this question, OrbitBench pairs strictly increasing, invertible re-expressions of the same observations while holding the inference pipeline fixed. We evaluate AVICI, Arrow, and CDFM, with PC and GES as classical references, on synthetic tasks and measurements from one Causal Chambers light-tunnel apparatus (Gamella et al., 2025). On 16 controlled linear tasks, CDFM changes its skeleton without lowering F1 in 41 of 64 task–view comparisons; among the three pretrained systems, Arrow remains near-best in both views while changing its skeleton in 63 of 64 comparisons.

Although variable-wise invertible re-expressions preserve the causal graph, they need not preserve the functional-form or noise assumptions under which an estimator operates, and can change both the number of predicted edges and the variable pairs they connect (Jørgensen & Hauberg, 2020). To distinguish changes in edge count from changes in which pairs are selected, we compare learned-model predictions at fixed edge counts and find that the selected pairs still change on both linear and nonlinear synthetic tasks.

Inferred graphs are also used to answer structural questions that extend beyond individual adjacencies. For example, changing an edge need not change whether one variable can reach another through a directed path, so adjacency disagreement alone does not establish instability in path-based answers. We therefore compare path-based answers across measurement views, separately for predictions that are already acyclic and for outputs made acyclic by a fixed projection. On fresh linear-Gaussian tasks, the tested re-expressions change these path-based answers as well as individual adjacencies, showing that coordinate sensitivity can alter graph-derived conclusions without changing the underlying causal system.

Our contributions are threefold: (1) We introduce OrbitBench, a reusable paired evaluation protocol and implementation for assessing accuracy and relationship consistency across measurement coordinates. (2) We provide paired evidence that non-decreasing accuracy or retention of a near-best model can coexist with adjacency changes, and show that disagreement persists at fixed edge counts and extends to DAG-path answers. (3) On new nonlinear tasks, we demonstrate a practical use of coordinate audits: two audit views improve average held-out adjacency persistence relative to score-only selection within the tested transformation families.

2 Background and Related Work

2.1 Causal Discovery and Measurement Representations

Given a data table $X \in \mathbb{R}^{n \times d}$ with n observations of d measured variables, let $\widehat{G} = f(X)$ denote the graph predicted by a fixed inference pipeline f, including preprocessing and graph decoding. Learned systems may predict edge scores directly, as in AVICI (Lorch et al., 2022) and CDFM (Qiao et al., 2026), or combine skeleton scores with a variable ordering, as in Arrow (Thompson et al., 2026).

Although inference models differ in architecture, our comparison changes only the measurement representation supplied to a fixed inference system. We consider transformations of the form $T(X)_{ij} = t_j(X_{ij})$, where each t_j is strictly increasing and has an inverse on its image. Each transformation is applied consistently within its column, preserving the correspondence between observations and variables. Interpreted as a coordinate change of the same structural causal model, this operation preserves the target graph and population conditional independence relations (Jørgensen & Hauberg, 2020). Positive affine maps preserve linear and additive-noise model classes, whereas nonlinear monotone maps need not (Peters et al., 2014; Jørgensen & Hauberg, 2020). We therefore compare $f(X)$ and $f(T(X))$ to assess representation sensitivity, without presuming that every estimator has a theoretical invariance guarantee.

Graph outputs support different levels of structural comparison. An undirected skeleton records whether each pair of variables is adjacent, whereas a directed path records reachability through

a sequence of directed edges. Different DAGs can encode the same conditional independences, so observational data do not generally identify a unique DAG without further assumptions (Peters et al., 2014). We therefore interpret predicted edges and paths as model-implied structural claims, and evaluate paths only for DAG outputs. The evaluation separates accuracy relative to a reference graph, when available, from agreement across representations of the same observations.

2.2 Related Work on Representation Sensitivity and Robustness

Measurement representation is an established concern in causal discovery, with rank-based PC extending Gaussian methods to Gaussian copula models (Harris & Drton, 2013). Jørgensen & Hauberg (2020) assess causal decisions across random bijections, use resampling and reparameterization to quantify decision confidence, and develop a multivariate extension of their estimator. Reisach et al. (2021) show how scale-dependent variance patterns can drive performance on simulated benchmarks. Related work on benchmark generation seeks to reduce such artifacts by changing how structural causal models are sampled (Herman et al., 2025).

Recent benchmarks have made robustness to assumption violations an explicit evaluation objective (Yi et al., 2025; Stein et al., 2026). Yi et al. (2025) assess differentiable causal discovery under eight forms of misspecification, including scale variation, using Structural Hamming Distance and Structural Intervention Distance. For time-series causal discovery, TCD-Arena varies the severity of assumption violations within a modular testing framework (Stein et al., 2026). A complementary line examines how hyperparameter selection affects structure recovery and the choice between algorithms (Machlanski et al., 2024). OCDB complements these studies with a real-data benchmark and metrics for comparing causal structures and effects (Zhou et al., 2024). Causal Chambers provides controlled physical testbeds for evaluating causal methodology, including the light-tunnel system used in our study (Gamella et al., 2025).

Reparameterization-aware estimation and confidence assessment (Jørgensen & Hauberg, 2020) provide a foundation for treating measurement coordinates as an evaluation variable. OrbitBench focuses on how often a fixed multivariate predictor changes its reported relationships while accuracy does not decrease or a near-best method is retained. We assess these events on the same task–view pairs, complementing recovery-oriented evaluations under assumption violations (Yi et al., 2025; Stein et al., 2026). Fixed-edge-count comparisons and DAG-path analyses examine whether disagreement extends beyond changes in graph density and individual edges. A separate held-out study evaluates the persistence of selected adjacencies at new transformation parameters. This is an evaluation of cross-view agreement and its practical use for screening, not a new invariant causal estimator or a guarantee of causal correctness.

3 OrbitBench: Evaluation Protocol

3.1 Paired Measurement-Coordinate Views

To assess dependence on measurement coordinates, we compare predictions from different numerical representations of the same observations. For a reference table $X^{(0)} = X$, we construct views $X^{(v)} = T_v(X)$ by applying strictly increasing, invertible transformations separately to the columns, while retaining the same samples, variable identities, and row order. For each method, we pair the reference prediction $\widehat{G}^{(0)} = f(X^{(0)})$ with each transformed prediction $\widehat{G}^{(v)} = f(X^{(v)})$. The preprocessing and decoding rules are held fixed across views, and pretrained models are neither retrained nor fine-tuned.

We construct nonlinear views using two smooth transformation families, applied elementwise:

$$a_\lambda(x) = \frac{\operatorname{asinh}(\lambda x)}{\lambda}, \qquad c_\beta(x) = x + \beta x^3. \tag{1}$$

The asinh map compresses large magnitudes, whereas the cubic map expands them; both are strictly increasing and invertible for the positive parameters used here. Under the measurement-coordinate interpretation, these transformations preserve the target graph, but need not preserve an estimator's functional-form or noise assumptions.

For the fixed-view comparisons, we use $\lambda \in \{0.5, 1.5\}$ and $\beta \in \{0.05, 0.15\}$, with each parameter value shared across columns, yielding four nonlinear views that are each paired with the reference. The held-out screening experiment instead samples audit and test parameters separately within the same transformation families.

3.2 Evaluation Measures

We quantify adjacency disagreement between the reference and transformed views as

$$D_{\mathrm{adj}}(m, v) = \left| S_m^{(0)} \triangle S_m^{(v)} \right|, \tag{2}$$

where $S_m^{(v)}$ is the undirected skeleton edge set predicted by method m on view v, and $\triangle$ denotes symmetric difference. A larger value indicates changes in more variable pairs, while zero means that the predicted skeletons are identical. The comparison accommodates directed and partially directed outputs by using their skeletons, without comparing edge orientations. We report changed-pair counts and explicitly identify the denominator of normalized summaries.

Representation consistency and accuracy capture different properties: stable predictions can be wrong, and different graphs can achieve the same F1. When synthetic ground truth is available, we also report skeleton F1 and its paired change, defined as reference F1 minus transformed-view F1. We summarize model choice within each specified comparison set using the winner set, which contains all methods within 0.01 absolute F1 of the best score on that view. A winner-set flip occurs only when the reference and transformed winner sets have no member in common. A non-flip indicates a shared near-best method, not identical winner sets, an unchanged full ranking, or statistical equivalence of accuracy. We interpret these summaries alongside adjacency disagreement on the same task–view pairs wherever the quantities are jointly available.

To examine consequences beyond individual adjacencies, we also compare answers to directed-path queries. For DAG outputs, let $R_G(i, j)$ indicate whether a directed path connects i to j. The path-answer flip fraction is

$$D_{\mathrm{path}}(G_0, G_v) = \frac{1}{d(d-1)} \sum_{i \neq j} \mathbf{1}[R_{G_0}(i, j) \neq R_{G_v}(i, j)]\,. \tag{3}$$

Here, G_0 and G_v are paired graphs produced using the same decoding rule; the denominator counts all ordered pairs of distinct variables. Native-output comparisons require both graphs to be acyclic; otherwise this endpoint is undefined. Native and projected outputs are reported separately, as detailed in Section 3.3. The measure captures changes in model-implied reachability, not changes in estimated intervention effects or proof that either predicted path is correct.

3.3 Controls and Audit Design

To interpret differences across measurement coordinates, we check execution variability and numerical validity. Repeated inference on the reference input measures execution variability, while positive affine transformations, $t_j(x) = a_j x + b_j$ with $a_j > 0$, provide a coordinate control; graph agreement and score drift are assessed separately. In synthetic confirmation experiments, input-only checks verify finite values and preservation of within-column order and ties after numerical conversion and preprocessing. Eligible tasks are selected in fixed seed order within prespecified strata, without inspecting predictions, so these analyses concern the numerically admissible population.

To test whether predicted connections change even when edge count is held fixed, we compare learned-model skeletons at a common density across views. Each view retains the highest-scoring k variable pairs, with k set separately for each model and task to the reference prediction's skeleton edge count. For directed-path comparisons, we also construct DAGs by adding directed edges in descending score order while rejecting cycles, using the reference prediction's directed-edge count as a fixed budget. Both budgets come from predictions rather than ground truth; native and projected outputs are reported separately. Our implementation applies these score-based decoder comparisons to the three pretrained systems; PC and GES remain in native-output comparisons.

A complementary comparison examines how input canonicalization affects consistency and accuracy. Where included, Rank-Gaussian preprocessing serves as a separate canonicalization reference,

Table 1: Experimental groups and their roles.

Experimental group	Units	Setting and purpose
Controlled linear	16 tasks	Linear-Gaussian SCMs: eight Erdős–Rényi and eight scale-free graphs; joint comparison of structure, accuracy, and model choice.
Fresh-seed paths	16 tasks	New draws from the same linear-Gaussian generator; confirmation of path-answer changes.
Nonlinear confirmation	32 tasks	Nonlinear mechanisms, with 16 tasks per noise condition; structural and accuracy comparisons beyond the linear generator.
Held-out screening	32 tasks	New nonlinear tasks with separate audit/test parameters; evaluation of adjacency persistence after selection.
Physical measurements	57 environments	Light-tunnel measurements; within-environment structural consistency.

not an additional nonlinear test view. Order-preserving views can produce identical rank-based inputs, so we evaluate accuracy alongside consistency to distinguish coordinate agreement from correct graph recovery. The corresponding accuracy comparison is reported in Appendix E.2.

We then ask whether two audit views can help select predicted adjacencies that persist under additional coordinate changes. We score candidate adjacencies in the reference skeleton by $u(e) = r(e) - q(e)$, where $r(e)$ is the normalized midrank of the reference score among candidates and $q(e)$ is the fraction of the two audit views in which adjacency e disappears. The penalty weight is fixed at one across tasks and models. At the primary 80% retention level, rounded up, audit-assisted selection retains the candidates with the largest $u(e)$. We compare it with two rules retaining the same number of adjacencies: score-only selection ranks by decreasing reference score, while stability-first selection ranks by increasing audit disappearance fraction and then decreasing reference score. Exact rank normalization and tie-breaking rules are given in Appendix A.

The audit and test sets each contain one asinh and one cubic view. For each task, audit and test parameters are drawn independently within each family from log-uniform distributions, with $\lambda \in [0.5, 1.5]$ for asinh and $\beta \in [0.05, 0.15]$ for cubic; each sampled parameter is shared across columns. Retained sets are fixed before generating the held-out test predictions, without using ground truth. For nonempty candidate sets, held-out disappearance risk is the fraction of retained adjacencies absent from a test prediction, averaged over the two test views. This evaluates persistence at new parameter values within the tested families, not transfer to new families or a guarantee of causal correctness.

3.4 Models and Data

We evaluate measurement-coordinate sensitivity in three pretrained causal discovery systems: AVICI (Lorch et al., 2022), Arrow (Thompson et al., 2026), and CDFM (Qiao et al., 2026). PC (Spirtes & Glymour, 1991) and GES (Chickering, 2002) provide constraint-based and score-based classical references, respectively. The pretrained systems use fixed checkpoints without task-specific retraining or fine-tuning. All five methods enter the main structural comparisons; the fresh-seed path and held-out screening experiments use the three pretrained systems.

We use linear and nonlinear synthetic tasks and physical measurements to examine coordinate sensitivity under different mechanisms, noise conditions, and measurement settings. For each listed task or environment, participating methods receive matched coordinate views of 800 observations on 20 variables. Table 1 summarizes the experimental groups and their roles.

The nonlinear confirmation and screening tasks use AVICI's released random-Fourier-feature (RFF) generator under homoscedastic or heteroscedastic Gaussian noise, with fixed affine calibration estimated from observations disjoint from evaluation. These generator distributions are related to

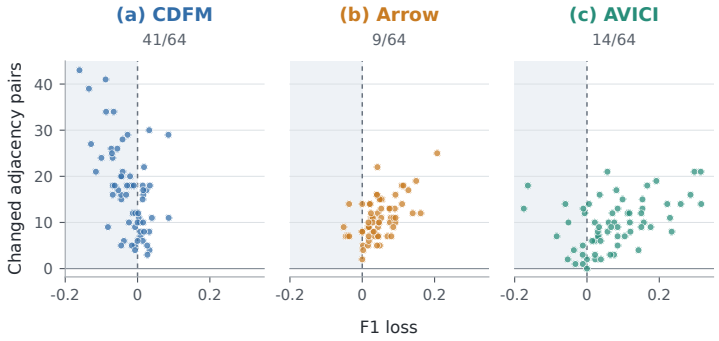


Figure 1: Paired skeleton F1 loss and adjacency disagreement on 16 controlled linear tasks. Each panel shows 64 comparisons: four nonlinear views per task. F1 loss is reference minus transformed F1. Shading and panel counts identify non-decreasing F1 with changed adjacencies. Counts are post-hoc descriptive summaries; comparisons within a task are not independent.

AVICI's training distributions. In the fresh-seed path experiment, the fixed-budget DAG decoder is the prespecified primary analysis; native outputs are reported separately. The physical environments come from a single Causal Chambers light-tunnel system (Gamella et al., 2025); comparisons are paired within environments and assess consistency, without accuracy or model-choice calculations.

Adjacency disagreement and F1 loss (reference minus transformed-view F1) are summarized by medians over four nonlinear views within each task, then across tasks within each group. Reference F1 in Table 2 is the across-task median; path- and winner-flip fractions use means. Screening risk reductions relative to score-only selection are averaged within model–noise groups, then equally across the six groups. Synthetic confidence intervals use whole-task bootstrap resampling within graph-family strata, preserving paired views and methods; screening additionally preserves noise strata and paired selection rules. Physical summaries remain descriptive. Appendices C and D detail implementation and resampling, respectively; Appendix E.1 reports matched signal-to-noise analyses. Task-level Robustness Cards separately report accuracy where available, representation consistency, and control diagnostics; Appendix F describes their implementation and reporting semantics.

4 Experiments

4.1 Accuracy and Model Choice versus Graph Consistency

Predicted adjacencies can change across measurement views even when skeleton F1 does not decrease. Figure 1 pairs F1 loss with adjacency disagreement for each of the 64 task–view comparisons on the 16 controlled linear tasks. In a descriptive paired analysis, F1 does not decrease while the predicted skeleton changes in 41, 9, and 14 comparisons for CDFM, Arrow, and AVICI, respectively. These comparisons show that non-decreasing accuracy does not ensure that the same adjacencies are reported.

Table 2 summarizes accuracy and adjacency disagreement, including the classical references. For these 20-variable graphs, the three pretrained systems change 10.0–16.5 variable pairs under the two-level median summary of Section 3.4. Median F1 losses are positive for Arrow and AVICI and negative for CDFM. Matched analyses of CDFM, Arrow, PC, and GES at additional local R^2 settings also show coordinate-dependent skeleton changes, with method-dependent trends across settings (Appendix E.1).

A related question is whether retaining a near-best model also retains its reported relationships. Among the three pretrained systems, the winner sets are disjoint in only 1 of 64 comparisons (1.6%; 95% CI: 0.0–4.7%), using the 0.01 F1 tolerance defined in Section 3.2. In a descriptive paired breakdown, Arrow remains near-best in both views in all 63 non-flip comparisons, yet its predicted skeleton changes in each (Table 2). Including PC and GES increases the flip count to 9 of 64 (14.1%) in an expanded descriptive comparison, showing that model-choice persistence depends on the comparison set.

Table 2: Skeleton F1 and adjacency disagreement on 16 controlled linear tasks. Reference F1 is the across-task median; F1 loss and changed pairs are medians over four views, then tasks. Loss is reference minus transformed F1. Brackets give 95% bootstrap intervals for changed pairs only. The last column reports post-hoc descriptive joint counts over 64 task–view pairs.

Method	Reference F1	F1 loss	Changed pairs (95% CI)	F1 not lower & skeleton changed
CDFM	0.656	−0.014	16.5 [13.5, 18.5]	41/64
Arrow	0.894	0.040	10.5 [9.5, 11.5]	9/64
AVICI	0.617	0.066	10.0 [8.0, 11.5]	14/64
PC	0.925	0.000	0.75 [0.0, 1.0]	19/64
GES	0.857	−0.003	4.5 [3.0, 7.0]	33/64

Three-model comparison (CDFM, Arrow, AVICI): Arrow remains near-best in both views while its skeleton changes in **63/64** paired comparisons.

Table 3: Adjacency disagreement at fixed predicted edge counts. Each model–task pair uses its native reference skeleton's edge count for every view. Entries are medians over four nonlinear views, then 16 tasks per column, with pointwise 95% bootstrap confidence intervals.

Method	Linear	Nonlinear: homoscedastic	Nonlinear: heteroscedastic
CDFM	14.5 [12.0, 16.0]	9.5 [7.0, 13.0]	9.0 [8.0, 14.0]
Arrow	10.0 [8.0, 10.5]	9.0 [8.0, 9.5]	9.5 [7.0, 11.0]
AVICI	8.5 [8.0, 12.0]	3.5 [2.0, 4.5]	3.0 [2.5, 5.0]

Thus, on this panel, non-decreasing F1 or a retained near-best model does not ensure that the same adjacencies are reported. We next test whether these changes reflect only differences in predicted edge count and whether they extend to graph-derived path answers.

4.2 Does Graph Disagreement Persist at a Fixed Edge Count?

Changes in predicted edge count are one possible source of the observed graph disagreement. We therefore ask whether this disagreement persists when the edge count is held fixed. For each pretrained model and task, we retain the same number k of highest-scoring variable pairs in both views, with k set to the edge count of the native reference skeleton (Section 3.3). In separate checks of native outputs, repeated-input and positive-affine controls preserve skeletons for all five methods across the 48 synthetic tasks, although affine transformations do not preserve scores exactly.

Graph disagreement persists at fixed edge counts: on the 16 linear tasks, CDFM, Arrow, and AVICI change 14.5, 10.0, and 8.5 variable pairs, respectively, under the two-level median summary (Table 3). The pattern also occurs on the 32 nonlinear RFF tasks, comprising 16 tasks per noise condition. Across the two noise conditions, CDFM and Arrow change 9.0–9.5 pairs, whereas AVICI changes 3.0–3.5 pairs. Native-skeleton accuracy and disagreement for the same nonlinear tasks are reported in Appendix E.3.

With equal edge counts, any added adjacency must be accompanied by a removed one. For the tested pretrained systems, coordinate sensitivity therefore affects the selection of specific adjacencies, rather than only the density of the predicted graph.

4.3 Do Coordinate Changes Alter DAG-Path Answers?

Changes in individual adjacencies need not alter whether a directed path exists between two variables. We therefore ask whether measurement-coordinate sensitivity extends to these graph-derived answers. We evaluate the three pretrained systems on 16 fresh-seed, 20-variable tasks from the same linear-Gaussian generator. The prespecified primary analysis uses the fixed-budget DAG decoder of

Table 4: Path-answer flip fractions (%) on 16 fresh-seed linear-Gaussian tasks. Entries average over four nonlinear views and tasks, with 380 ordered variable pairs per comparison. Brackets give pointwise 95% bootstrap confidence intervals. The fixed-budget DAG analysis is the prespecified primary endpoint; native CDFM is undefined because its reference graphs contain cycles. Flips measure answer disagreement, not causal error.

Method	Fixed-budget DAG (primary)	Native output
CDFM	12.5 [10.3, 14.8]	Undefined (cyclic)
Arrow	15.7 [14.0, 17.4]	18.0 [16.4, 19.5]
AVICI	4.4 [3.9, 4.9]	4.3 [3.9, 4.7]

Section 3.3 and measures answer changes between paired views across all 380 ordered pairs of distinct variables.

Coordinate changes also alter DAG-path answers under the fixed-budget decoder, with mean flip fractions of 12.5% for CDFM, 15.7% for Arrow, and 4.4% for AVICI (Table 4). The lower 95% confidence limits for CDFM and Arrow exceed the prespecified 5% reference level for this experiment, whereas AVICI's mean flip fraction is nonzero but remains below this level.

Native outputs provide a separate check: Arrow and AVICI change 18.0% and 4.3% of path answers, respectively, so their changes are not limited to the standardized projection. CDFM's native endpoint is undefined on this panel because its reference graphs contain cycles. These findings concern model-implied reachability, not estimated intervention effects or evidence that either answer is correct. Within this fresh-seed panel, coordinate sensitivity therefore extends from individual adjacencies to graph-derived path answers.

4.4 Can Coordinate Audits Help Select More Stable Relationships?

We next ask whether two coordinate checks can help select predicted adjacencies that persist under further re-expressions. On 32 new nonlinear tasks, we compare audit-assisted, score-only, and stability-first selection at 80% retention of reference adjacencies (Section 3.3). Both audit-based rules use the same two audit views. Selected sets are determined without ground truth and fixed before the two held-out predictions are generated.

Both audit-based rules yield lower macroaverage held-out disappearance risks than score-only selection: 6.71% for audit-assisted and 6.68% for stability-first, compared with 8.27% for score-only (Table 5). The prespecified macroaverage reduction from score-only to audit-assisted selection is 1.56 percentage points (95% CI: 1.03–2.14). The interval supports a positive average gain, but does not establish that it exceeds the prespecified two-percentage-point reference. Audit-assisted selection has no demonstrated advantage over stability-first selection.

Greater persistence does not certify correctness: relative to score-only selection, the descriptive change in retained-adjacency precision is +0.41 percentage points (95% CI: -0.06–0.88). Audit and test parameters are drawn independently within the same asinh and cubic families. Within these tested families and parameter ranges, coordinate audits therefore provide useful information for selecting more persistent predicted relationships.

4.5 Does Coordinate Sensitivity Extend to Physical Measurements?

Causal Chambers light-tunnel measurements also show coordinate-dependent adjacency disagreement (Gamella et al., 2025). After inspecting the initial environment's result, we designed a descriptive extension covering all 56 other environments meeting fixed input-eligibility criteria. In each environment, all five methods use re-expressions of the same 800 observations on 20 variables; we assess prediction consistency without reference-graph accuracy scoring.

Figure 2 shows how adjacency disagreement varies across methods and environments. Across the 56 extension environments, the two-level median numbers of changed adjacencies are 2.5 for CDFM, 6.0 for Arrow, and 4.5 for AVICI. PC and GES change 2.0 and 7.75 adjacencies, respectively, show-

Table 5: Held-out adjacency disappearance risk at 80% retention (rounded up) on 32 new nonlinear tasks. For each task, risk is the mean, over the two held-out views, of the fraction of retained adjacencies absent from that view's prediction. Results equally weight six model–noise groups, conditional on nonempty candidate sets (Section 3.4). Brackets give 95% paired bootstrap confidence intervals. Reductions are in percentage points (pp) relative to score-only selection, not accuracy gains.

Selection rule	Held-out disappearance risk (%)	Reduction vs. score-only (pp)
Score-only	8.27 [6.99, 9.61]	Reference
Audit-assisted	6.71 [5.65, 7.82]	1.56 [1.03, 2.14]
Stability-first	6.68 [5.61, 7.78]	1.59 [1.06, 2.18]

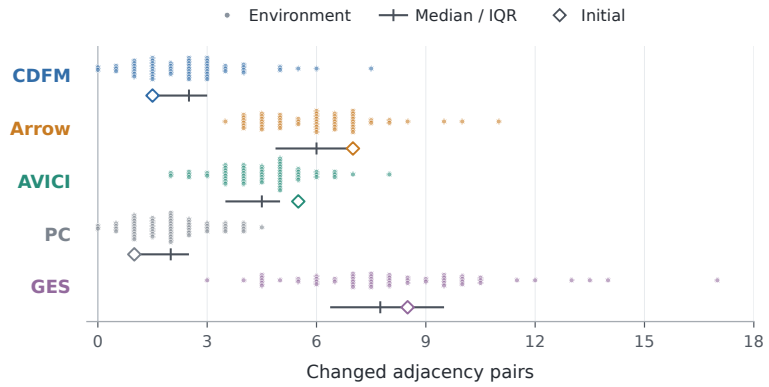


Figure 2: Native-skeleton disagreement in one Causal Chambers light-tunnel apparatus. Dots summarize 56 extension environments by their median changed-pair count over four nonlinear views. Bars and ticks show across-environment interquartile ranges and medians; open diamonds show the separate initial environment, excluded from these summaries. Vertical offsets prevent overlap. The extension is result-aware and descriptive; bars are not confidence intervals.

ing that the observed sensitivity is not confined to pretrained systems. Exact native-output summaries are reported in Appendix B (Table 6). For CDFM, Arrow, and AVICI, fixed-edge-count values are 2.0, 6.0, and 4.0, respectively, so fixing predicted edge counts does not eliminate disagreement.

Identity repeats and positive-affine controls preserve skeletons in all 285 method–environment combinations, although learned systems do not uniformly pass stricter continuous-score checks. This single-apparatus study extends the evidence to physical measurements, with smaller smooth-view disagreements for learned systems than on the linear synthetic panel.

5 Conclusion

Evaluating measurement-coordinate consistency addresses an evaluation gap between aggregate benchmark performance and the persistence of relationships reported by pretrained causal discovery systems. OrbitBench shows that non-decreasing accuracy or retention of a near-best model can coexist with adjacency changes under invertible re-expressions of the same observations. Disagreement persists at fixed predicted edge counts and extends to model-implied DAG-path answers, with its magnitude varying across methods and settings. Both audit-assisted and stability-first selection reduce average held-out adjacency disappearance risk relative to score-only selection within the tested transformation families. Our evaluation is limited to selected inference systems and transformation families, with physical measurements from a single apparatus. Future work should broaden model and physical-domain coverage, test generalization to unaudited transformation families, and determine whether coordinate-stable relationships improve downstream causal correctness, since persistence alone does not guarantee correctness.

AI USE STATEMENT

Generative AI tools were used solely for grammar and language polishing; all AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 3 specifies the paired coordinate-view protocol, evaluation measures, and controls. Appendices A and C provide the audit-selection rules, data-generation and preprocessing procedures, and native and controlled graph-decoding settings. Appendix D details aggregation, task-level bootstrap resampling, and the treatment of undefined endpoints. Appendix F describes the reference implementation, its input requirements, and the reporting semantics of the Robustness Cards.

REFERENCES

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(105):3365–3383, 2013.

Rebecca J. Herman, Jonas Wahl, Urmi Ninad, and Jakob Runge. Unitless unrestricted Markov-consistent SCM generation: Better benchmark datasets for causal discovery. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 1506–1531, 2025.

Martin Jørgensen and Søren Hauberg. Reparametrization invariance in non-parametric causal discovery. arXiv preprint arXiv:2008.05552, 2020.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

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, pp. 13104–13118, 2022.

Damian Machlanski, Spyridon Samothrakis, and Paul S. Clarke. Robustness of algorithms for causal structure learning to hyperparameter choice. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 703–739, 2024.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Jie Qiao, Ruichu Cai, Zijian Li, Weilin Chen, Pengfei Hua, Boyan Xu, Zhengming Chen, Zhifeng Hao, and Peng Cui. CDFM: Towards a general-purpose causal discovery foundation model. arXiv preprint arXiv:2607.11508, 2026.

Alexander 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.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10:2553, 2019.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

Gideon Stein, Niklas Penzel, Tristan Piater, and Joachim Denzler. TCD-Arena: Assessing robustness of time series causal discovery methods against assumption violations. In *International Conference on Learning Representations*, 2026.

Ryan Thompson, He Zhao, Daniel M. Steinberg, and Edwin V. Bonilla. Arrow: A foundation model for causal discovery. arXiv preprint arXiv:2605.07204, 2026.

Huiyang Yi, Yanyan He, Duxin Chen, Mingyu Kang, He Wang, and Wenwu Yu. The robustness of differentiable causal discovery in misspecified scenarios. In *International Conference on Learning Representations*, 2025.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Wei Zhou, Hong Huang, Guowen Zhang, Ruize Shi, Kehan Yin, Yuanyuan Lin, and Bang Liu. OCDB: Revisiting causal discovery with a comprehensive benchmark and evaluation framework. arXiv preprint arXiv:2406.04598, 2024.

A Audit Implementation Details

For each model–task pair, the candidate set consists of the m undirected adjacencies in the reference prediction. For $m > 1$, let $R(e)$ be the ascending rank of the reference score among candidates, with tied scores assigned their average rank. The normalized midrank is $r(e) = (R(e) - 1)/(m - 1)$. For $m = 1$, we set $r(e) = 0$. The audit disappearance fraction $q(e)$ is the number of the two audit predictions that omit adjacency e, divided by two.

Audit-assisted selection orders candidates by decreasing $u(e) = r(e) - q(e)$, breaking ties by decreasing reference score. Score-only selection orders candidates by decreasing reference score. Stability-first selection orders candidates by increasing $q(e)$ and then decreasing reference score. For all three rules, remaining ties are resolved lexicographically by variable-index pairs (i, j) with $i < j$. Each rule retains the first $\lceil 0.8m \rceil$ candidates.

If $m = 0$, the retained set is empty and disappearance risk is undefined rather than zero; such cases are excluded from risk averages. For nonempty candidate sets, risk is computed separately for each of the two held-out views as the fraction of retained adjacencies absent from that view's prediction, and then averaged over the two views.

B Physical Measurement Results

We report numerical summaries of the native-output adjacency disagreement shown in Figure 2. The initial environment is reported separately from the 56 extension environments, all from the same Causal Chambers light-tunnel apparatus. The extension was designed after inspecting the initial result and is descriptive.

Each task or physical environment contains 800 observations on 20 variables. Participating methods receive matched coordinate views, with observation and variable identities held fixed. The four smooth views follow Section 3.1, using asinh strengths 0.5 and 1.5 and cubic coefficients 0.05 and 0.15. The positive-affine control cycles through column scales $(0.5, 1, 2, 4)$ and shifts $(-2, -0.5, 1, 3)$; the identity control repeats inference on unchanged inputs. Rank-Gaussian analyses, where reported, use average ranks and map rank R among n observations to $\Phi^{-1}((R-0.5)/n)$, where Φ is the standard-normal cumulative distribution function. They are reported separately from the four smooth views.

Table 6: Changed unordered variable pairs in native predicted skeletons under the four smooth nonlinear views. The initial-environment column reports the median over views. For the extension, we first take the median over views within each environment, then report the median and interquartile range across the 56 environments. Brackets denote interquartile ranges, not confidence intervals. The initial environment is excluded from the extension summary.

Method	Initial environment	56 extension environments
CDFM	1.50	2.50 [1.50, 3.00]
Arrow	7.00	6.00 [4.88, 7.00]
AVICI	5.50	4.50 [3.50, 5.00]
PC	1.00	2.00 [1.00, 2.50]
GES	8.50	7.75 [6.38, 9.50]

C Additional Experimental and Implementation Details

C.1 Models, Preprocessing, and Native Outputs

We evaluate CDFM, Arrow, and AVICI using fixed released checkpoints without task-specific training or fine-tuning. We use the released CDFM checkpoint, Arrow's base checkpoint, and AVICI's `scm-v0` checkpoint, with PC and GES implemented through `causal-learn`. Each method retains its original preprocessing and native graph-construction procedure (Table 7). All inference is performed on CPU; pretrained-model inputs use `float32`, whereas PC and GES receive `float64` arrays.

Table 7: Preprocessing and native graph-construction settings.

Method	Settings
CDFM	Official columnwise standardization, applied once; the released predictor's automatic probability-threshold rule.
Arrow	Official columnwise standardization; edge probabilities above 0.5 combined with the predicted node ordering.
AVICI	Official continuous-data standardization, without intervention indicators; directed probabilities above 0.5.
PC	Stable PC with the Fisher-z conditional-independence test and significance level 0.05.
GES	The released `local_score_BIC` score and otherwise default search settings.

For skeleton comparisons, a variable pair is adjacent if either direction is present, and self-loops are excluded. PC and GES endpoint matrices are converted to skeletons without completing their partially directed outputs into DAGs. Native-output comparisons therefore preserve each method's own graph-construction rule.

C.2 Synthetic Tasks and Input Eligibility

The controlled linear panel contains eight Erdős–Rényi and eight scale-free tasks, each with 36 edges. The former use uniformly sampled fixed-edge-count skeletons, and the latter use Barabási–Albert attachment parameter 2; both are oriented by a random topological order. Edge magnitudes are uniform on $[0.75, 1.25]$, with independent random signs. An independent set of 8,192 calibration observations determines non-root noise scales for target local $R^2 = 0.5$ and the sequential affine normalization; root variables have standard-normal noise. The 800 evaluation observations do not enter calibration. The path panel uses 16 fresh-seed tasks from the same generator.

The nonlinear structural and screening panels use separate sets of 32 observational tasks from AVICI 1.0.7's released `rff-gauss.yaml` and `rff-gauss-heterosked.yaml` generators. Each panel contains 16 tasks per noise condition: four ScaleFree tasks and three each from ScaleFree-Transpose, WattsStrogatz, SBM, and GRG. Within each graph family, configurations are sampled uniformly from the released alternatives, retaining their mechanism-parameter distributions. The

first 8,192 generated observations determine column means and population standard deviations, which are applied to the remaining 800 evaluation observations. These generators are related to AVICI's training distributions; the panel is therefore not an out-of-distribution evaluation for every pretrained system.

Input eligibility is assessed without model predictions, using finite-value checks and within-column order and equality relations across views. Checks compare transformed and reference inputs at matching numerical layers, including `float64`, `float32`, and model-specific preprocessing. Failed cases are not repaired by jitter or rounding. After full scans of fixed candidate pools, tasks are selected in ascending candidate-index order. The controlled linear panel selects eight tasks per graph family from 256 candidates, requiring eligibility jointly at $R^2 \in \{0.25, 0.5, 0.75\}$; the path panel selects eight per family from 64 candidates at $R^2 = 0.5$. Each nonlinear panel uses 64 candidates per graph-family–noise stratum to fill the stated quotas.

For held-out screening, audit and test views each contain one asinh and one cubic transformation. Before input screening, their parameters are drawn independently from log-uniform distributions over $[0.5, 1.5]$ and $[0.05, 0.15]$, respectively, with each draw shared across columns. Ranking rules and the primary retention level of 80% are fixed before confirmation outputs are inspected. Selections use only reference and audit predictions and are sealed before test predictions are evaluated; Appendix A specifies the ranking rules.

C.3 Physical Measurements

We use the Causal Chambers dataset `lt_interventions_standard_v1`, taking the same fixed set of 20 variables and the first 800 rows in the released order within each environment, without imputation or row reordering. Eligibility requires all selected columns, at least 800 rows, finite selected values, and positive range in every column. The initial `uniform_reference` environment is reported separately from all 56 additional eligible environments. Unlike the synthetic panels, physical coordinate calibration uses the fixed 800 reference rows within each environment. The extension was designed after inspecting the initial result and is descriptive. Comparisons remain within environments from one apparatus, without reference-graph accuracy scoring.

C.4 Controlled Graph Decoding

Fixed-edge-count skeletons retain the highest-scoring unordered pairs, using the native reference skeleton size as the budget for both views. Pair scores are the maximum of the two directed probabilities for CDFM and AVICI, and Arrow's edge-existence probabilities for Arrow. Ties are resolved lexicographically by variable indices.

The path control instead ranks directed scores and greedily adds edges that do not create a cycle, until the native reference directed-edge count is reached. Both views use the same budget, procedure, and lexicographic tie-breaking. Arrow uses directed probabilities here, rather than its undirected edge-existence scores. These controls post-process saved scores without changing model weights or native predictions.

D Aggregation and Statistical Analysis

D.1 Units and Point Estimates

The task, rather than an individual coordinate view, is the resampling unit in synthetic analyses. In Table 2, reference F1 is the median of the 16 reference-view F1 scores. F1 loss is computed as reference minus transformed F1 for each task–view pair, then summarized by a median over four smooth views within each task and a median across tasks. Changed-pair counts and fixed-edge-count disagreements use the same two-level median within each reported group. Negative F1 loss indicates improved accuracy; fractional changed-pair summaries arise from medians of integer counts. Fractions of possible unordered pairs use denominator $\binom{20}{2} = 190$.

For model choice, the near-best set includes methods within 0.01 of the highest F1 in each view. The main analysis compares the three pretrained systems and records a flip when the reference and transformed near-best sets are disjoint. Flip indicators are averaged over views and then tasks. Joint

counts over 64 task–view comparisons, such as non-decreasing F1 with a changed skeleton, are post-hoc descriptive summaries of 16 tasks, not 64 independent trials.

Path-answer flips are fractions over all $20 \times 19 = 380$ ordered pairs of distinct variables, averaged over four smooth views and then the 16 fresh-seed tasks. They measure changes in graph-implied reachability, not measured changes in interventional effects.

D.2 Bootstrap Intervals

Synthetic 95% confidence intervals use 10,000 bootstrap replicates, sampling whole tasks with replacement within graph-family strata. Each linear-panel replicate contains eight Erdős–Rényi and eight scale-free draws. Nonlinear analyses additionally preserve noise condition and the original graph-family counts within each condition. All paired views of a sampled task remain together, and paired rule comparisons use the same sampled tasks. Each replicate recomputes the complete endpoint-specific aggregation. Interval endpoints are the 2.5th and 97.5th percentiles, using linear quantile interpolation.

The intervals shown in Table 2 apply only to changed-pair counts, not to its F1 columns. Unless explicitly stated otherwise, intervals are pointwise, not simultaneous across models and endpoints. They are conditional on the input-eligibility protocol and do not propagate uncertainty from constructing or accepting candidate pools.

The fresh-seed path protocol prespecifies the fixed-budget DAG decoder as primary, requiring a lower 95% interval endpoint above a 5% flip rate for at least two of the three pretrained systems. Screening uses the equal-six-group risk reduction at 80% retention as its primary endpoint, with a two-percentage-point improvement as a separate practical reference. An interval above zero does not by itself establish improvement beyond this reference or in every model–noise group.

D.3 Held-Out Screening Summaries

At 80% retention, task-level disappearance risk is the fraction of retained reference adjacencies absent from a test graph, averaged over the two held-out views. Competing rules retain the same number of edges, as specified in Appendix A. Risk reduction is score-only risk minus the evaluated rule's risk, computed within each task. We average these differences within each model–noise group and then equally across the six groups. Edges are not pooled across tasks, so denser predictions do not receive extra weight.

The paired bootstrap samples tasks within graph-family–noise strata, carrying all methods and rules together before recomputing the six-group average. One AVICI task under homoscedastic noise has an empty reference candidate set, making its screening risk and precision undefined rather than zero. That group has 15 defined tasks; the other five each have 16. Undefined entries are omitted from their group means in both point estimates and bootstrap replicates; all reported replicates retain defined observations in every group. The estimand is conditional on nonempty candidate sets, with equal group weights.

D.4 Undefined Endpoints and Descriptive Reporting

Native DAG-path answers require both graphs in a comparison to be acyclic. CDFM's native reference graphs are cyclic on all 16 path tasks, so its native path endpoint is undefined, not a zero flip rate. The fixed-budget DAG endpoint is evaluated separately.

For physical measurements, we take the median over four views within each environment, then report the median and interquartile range across the 56 additional environments, keeping the initial environment separate. These interquartile ranges describe variation across observed environments; they are not confidence intervals. Identity repeats and positive-affine controls are reported separately and excluded from smooth-view aggregates.

Table 8: Native-skeleton changes across signal-to-noise settings. Entries are medians over the four smooth views within each topology, followed by a median across the 16 topology blocks. These descriptive changed-pair counts are not counts of errors.

Method	$R^2 = 0.25$	$R^2 = 0.50$	$R^2 = 0.75$
CDFM	22.25	16.50	12.25
Arrow	13.75	10.50	8.75
PC	0.25	0.75	1.25
GES	2.50	4.50	8.25

Table 9: Accuracy changes and the paired signal-to-noise contrast. The first three numerical columns report two-level median F1 loss, defined as reference minus transformed F1. The final column reports the median paired high-minus-low difference with its pointwise 95% bootstrap interval. Negative F1 loss indicates improved accuracy.

Method	F1 loss			Paired difference
	$R^2 = 0.25$	$R^2 = 0.50$	$R^2 = 0.75$	[95% CI]
CDFM	-0.0248	-0.0138	0.0012	0.0249 [-0.0130, 0.0361]
Arrow	0.0562	0.0404	0.0480	0.0021 [-0.0372, 0.0153]
PC	0.0000	0.0000	0.0061	0.0045 [0.0000, 0.0098]
GES	-0.0005	-0.0032	0.0052	0.0071 [-0.0044, 0.0171]

E Additional Results

E.1 Sensitivity across Signal-to-Noise Settings

To examine whether coordinate sensitivity is confined to the setting used in the main linear panel, we additionally evaluate target local R^2 values of 0.25 and 0.75 alongside the reference setting of 0.50. The three settings share the same 16 topology blocks, edge weights, and underlying Gaussian noise draws, with noise scales and affine normalization calibrated separately at each level. This extension includes CDFM, Arrow, PC, and GES, with 800 observations on 20 variables per configuration. Local R^2 is a controlled signal-to-noise parameter, not an assumed universal ordering of task difficulty across methods.

Table 8 summarizes changes in native predicted skeletons across the three settings. As local R^2 increases, the two-level median changed-pair count decreases from 22.25 to 12.25 for CDFM and from 13.75 to 8.75 for Arrow. PC and GES show the opposite descriptive trend, increasing from 0.25 to 1.25 and from 2.50 to 8.25, respectively. Coordinate sensitivity is therefore observed beyond the reference setting, but its magnitude and response to the generator setting are method-dependent.

The accuracy summaries do not track these structural changes uniformly. For example, at $R^2 = 0.25$, CDFM has a changed-pair summary of 22.25 but a median F1 loss of -0.0248, indicating a net improvement in the accuracy summary. Table 9 reports F1 loss at each setting and the prespecified high-minus-low contrast. For this contrast, we subtract the four-view median F1 loss at $R^2 = 0.25$ from that at $R^2 = 0.75$ within each topology, then take the median across topologies. The contrast is therefore not the difference between the two displayed marginal medians. Confidence intervals use the 10,000 graph-family-stratified bootstrap draws described in Appendix D, carrying all three settings together for each sampled topology.

All four paired F1-loss intervals include zero, so the prespecified directional increase from the low- to high-R^2 setting is not established. This does not imply that the generator setting has no effect: structural disagreement varies across settings, and the observed direction differs by method. The extension supports coordinate sensitivity across the tested signal-to-noise settings without establishing a common directional increase in its net F1 cost. Because the same topology blocks are reused, these results are a matched sensitivity analysis rather than an independent replication.

Table 10: Accuracy under Rank-Gaussian canonicalization on 16 controlled linear tasks. The F1 columns are medians across tasks. Paired cost is the median of reference F1 minus canonicalized F1 computed within each task; negative values indicate improvement. It is not the difference between the displayed F1 medians. Entries are descriptive point estimates.

Method	Reference F1	Rank-Gaussian F1	Paired cost
CDFM	0.6563	0.6596	0.0082
Arrow	0.8937	0.9210	−0.0268
AVICI	0.6168	0.6001	0.0208
PC	0.9254	0.9254	0.0000
GES	0.8573	0.8504	0.0000

E.2 Rank-Gaussian Canonicalization

To separate coordinate consistency obtained by preprocessing from accurate graph recovery, we examine Rank-Gaussian canonicalization on the 16 controlled linear tasks. Applying the average-rank transform of Appendix C to the reference, positive-affine, and four smooth views yields identical canonical inputs, as verified on the retained data. A single canonical prediction per method and task is reused across these identical inputs. Consequently, zero disagreement between canonicalized views follows by construction; it is not evidence from independent repeated inference runs or a guarantee of correct graph recovery.

Table 10 reports reference and canonicalized skeleton F1, together with the paired accuracy cost. The median paired cost is positive for CDFM and AVICI, negative for Arrow, and zero for PC and GES. These summaries show that the accuracy response is method-dependent, rather than uniformly improved by canonicalization. A zero median cost does not imply unchanged accuracy on every task. Rank-Gaussian therefore serves as a consistency reference whose accuracy must be evaluated separately, not as a universal repair.

E.3 Native-Graph Accuracy on Nonlinear Tasks

To complement the fixed-edge-count comparisons in Section 4.2, Table 11 reports native-skeleton accuracy and disagreement on the same 32 nonlinear confirmation tasks. Results are separated into 16 tasks with homoscedastic noise and 16 with heteroscedastic noise, using the generators and eligibility procedure described in Appendix C. All five methods are included, and their native output rules are unchanged. This is a fuller report of the existing nonlinear panel, not a new replication or the held-out screening experiment.

Every method has a nonzero median changed-pair count in both noise conditions, whereas median F1 losses range from -0.0050 to 0.0563. For example, GES changes a median of 13.5 pairs under heteroscedastic noise while its median F1 loss is 0.0025. These group summaries complement, rather than replace, the task–view co-occurrence analysis in Section 4.1. The pattern supports reporting structural consistency alongside accuracy beyond the linear generator, without requiring a uniform accuracy decline across methods. Because these generators are related to AVICI's training distributions, this panel is not an out-of-distribution evaluation for every pretrained system.

F Reference Implementation and Reproducibility

OrbitBench separates model inference from evaluation. Its evaluator accepts a task manifest specifying coordinate views and reference-graph availability, together with a prediction bundle containing a graph for each view and optional edge scores and repeated-input outputs. It validates these inputs and produces a machine-readable Robustness Card with a readable summary. Accuracy, graph consistency, and control diagnostics are reported separately; unsupported quantities remain unavailable rather than being recorded as zero.

The reference collection contains 16 controlled synthetic tasks and 57 measurement environments from one light-tunnel apparatus, with outputs for five methods and 365 method–task cards. These materials reuse the corresponding predictions analyzed in the paper and do not constitute addi-

Table 11: Native-skeleton accuracy and disagreement on nonlinear tasks. Each noise condition contains 16 tasks. Reference F1 is the median across tasks; changed pairs and F1 loss use the two-level median over four smooth views and then tasks. F1 loss is reference minus transformed F1. Brackets give pointwise 95% graph-family-stratified bootstrap intervals for F1 loss only, using 10,000 replicates. Changed pairs are adjacency differences, not necessarily errors.

Method	Reference F1	Changed pairs	F1 loss [95% CI]
Homoscedastic noise ($n = 16$)			
CDFM	0.6979	10.25	0.0053 [−0.0080, 0.0292]
Arrow	0.6033	9.00	0.0249 [0.0000, 0.0479]
AVICI	0.7974	4.50	0.0159 [0.0052, 0.0214]
PC	0.6068	5.25	0.0102 [−0.0079, 0.0224]
GES	0.6667	13.00	0.0134 [−0.0026, 0.0226]
Heteroscedastic noise ($n = 16$)			
CDFM	0.7067	10.25	0.0216 [0.0071, 0.0264]
Arrow	0.5520	10.25	0.0563 [0.0232, 0.0723]
AVICI	0.8114	3.25	0.0076 [−0.0149, 0.0242]
PC	0.5132	5.00	−0.0050 [−0.0119, 0.0044]
GES	0.5833	13.50	0.0025 [−0.0076, 0.0182]

tional experiments. Saved prediction bundles allow card-level evaluation without loading pretrained checkpoints. The collection is not a complete archive of every experiment in the paper, and its full-view card summaries do not replace the four-smooth-view aggregation and task-level resampling specified in Appendix D.

For a concrete example, the CDFM card for the first Erdős–Rényi task records a reference F1 of 0.6598 and a strong-cubic-view F1 of 0.7750, alongside 21 changed adjacency pairs out of 190 possible pairs. Thus, improved accuracy and altered relationships remain visible as separate fields. The same card distinguishes exact positive-affine skeleton agreement from a small score drift exceeding the configured tolerance, and marks native DAG-path quantities as undefined because the predicted graphs contain cycles. This illustrates the card's reporting semantics, rather than providing additional independent evidence.