CAUSALARENA: BENCHMARKING CAUSAL DISCOVERY IN THE FOUNDATION MODEL ERA

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery aims to uncover causal structures from data and is fundamental to scientific reasoning and intervention-based decision making. Its evaluation relies heavily on structural causal models (SCMs), which specify a causal graph together with the mechanisms that generate data, yet existing studies differ substantially in graph families, mechanisms, and evaluation protocols. The emergence of causal discovery foundation models (CDFMs) further complicates evaluation: performance may reflect not only causal discovery ability, but also overlap between pretraining environments and test SCMs, making results on fixed synthetic benchmarks difficult to interpret. We introduce CausalArena, a unified and evolvable benchmark for causal discovery under a common protocol. Synthetic SCMs supply breadth over structures and mechanisms; semantic operational SCMs provide human-auditable, semantically grounded environments beyond standard synthetic generators; and formula-grounded SCMs test discovery under explicit scientific mechanisms. Public real-world datasets provide an additional external-validity check. Experiments across classical, neural, and pretrained methods reveal substantial ranking shifts across SCM families and protocols, showing that strong performance in one benchmark regime does not reliably transfer to others. These results highlight benchmark diversity and pretraining–evaluation overlap as central challenges for evaluating causal discovery in the foundation model era.

1 INTRODUCTION

Causal discovery aims to recover directed causal relationships from observational or interventional data and is fundamental to scientific reasoning and intervention-based decision making (Spirtes et al., 2000; Peters et al., 2017). Because the true causal graph is rarely known in real systems, empirical evaluation relies heavily on *structural causal models* (SCMs), which specify a causal graph together with the mechanisms that generate data. SCMs generate samples with known ground-truth structure and have therefore become a standard basis for comparing classical, neural, and pretrained causal discovery methods.

Yet causal discovery remains difficult to evaluate consistently. Different studies use different graph families, structural mechanisms, noise distributions, dimensions, sample sizes, intervention protocols, and thresholding rules. Performance can also depend strongly on the synthetic generator itself, as illustrated by generator artifacts such as varsortability (Reisach et al., 2021). Consequently, a strong result on one benchmark does not necessarily indicate robust causal discovery across other causal environments.

The emergence of causal discovery foundation models (CDFMs) makes this problem more fundamental. Methods such as AVICI (Lorch et al., 2022), SEA (Wu et al., 2025), Arrow (Thompson et al., 2026), CauScale (Peng et al., 2026), CDFM (Qiao et al., 2026), FoundCause (Blöbaum et al., 2026), and TabCausal (Li et al., 2026) are pretrained over large collections of synthetic or mixed causal environments. Their benchmark performance can therefore reflect not only causal structure-learning ability, but also overlap between pretraining and evaluation—from similar graph families and mechanisms to the same or closely related SCM generators. Moreover, once a fixed SCM benchmark becomes public, it can itself become future pretraining data. Thus, in the foundation-model era, the test distribution is no longer merely evaluative; it may become part of the training distribution.

Existing resources address different parts of this problem. Synthetic frameworks such as Bench-

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.

press (Rios et al., 2025) and CausalProfiler (Panayiotou et al., 2026) support controlled and scalable evaluation, while grounded resources such as Sachs (Sachs et al., 2005), CausalDynamics (Herdeanu et al., 2025), CausalBench (Chevalley et al., 2025), and Causal Chambers (Gamella et al., 2025) provide semantic, scientific, or physical context. However, these resources typically emphasize different graph distributions, domains, tasks, or protocols, making scores difficult to compare directly across methods and especially across pretrained models; Table 1 gives a compact comparison.

Table 1: Factual comparison of representative causal discovery evaluation resources and CausalArena. ✓ for general support, ○ for setting-dependent support, and – for absence.

Resource type	**Exec.**	**Broad**	**Ground.**	**Mech.**	**Proto.**	**Obs.+Int.**	**Tab. DAG**
Real/expert graphs	–	–	✓	○	–	○	✓
Method-specific synth.	✓	–	–	–	–	○	✓
Synthetic frameworks	✓	✓	–	–	✓	○	✓
Domain simulators	✓	–	✓	✓	–	○	○
Time-series benchmarks	○	–	✓	✓	–	○	–
Physical testbeds	–	–	✓	✓	–	✓	✓
Semantic/LLM graphs	–	–	✓	–	–	–	–
Foundation-model eval.	✓	✓	–	–	–	○	✓
CausalArena	✓	✓	✓	✓	✓	✓	✓

These limitations suggest four requirements for evaluating causal discovery in the foundation-model era. **Breadth** is needed because performance should not be determined by a single synthetic prior. **Freshness** is increasingly important because no fixed public SCM suite can remain unseen indefinitely; an evaluation framework should therefore support newly constructed causal environments as pretrained models evolve. **Grounding** complements arbitrary synthetic generators with causal systems whose variables, edges, and interventions have meaningful operational or scientific interpretations. Finally, **diagnosability** requires going beyond a single leaderboard score to identify which structures, mechanisms, and environments cause methods to succeed or fail.

Motivated by these requirements, we introduce CausalArena, a unified and evolvable benchmark for causal discovery. Its three SCM families play complementary roles. *Synthetic SCMs* provide breadth over graph structures, mechanisms, noise, dimensions, and intervention settings. *Semantic operational SCMs* ground causal structures in auditable real-world processes and provide a natural basis for introducing newly authored environments in future benchmark versions. *Formula-grounded scientific SCMs* anchor causal mechanisms in explicit scientific equations, making mechanism-level successes and failures easier to diagnose. All three families compile to a common executable SCM specification and share the same observational and interventional evaluation protocol; public real-world tables with published graphs serve as an additional external-validity check.

The current arena contains 1,200 executable SCM specifications—1,000 synthetic, 100 semantic operational, and 100 formula-grounded scientific SCMs—and evaluates representative classical, neural, and pretrained causal discovery methods. The current review package exposes a balanced half of each generated SCM family, while the remaining SCMs are reserved for held-out evaluation and future versioned updates. Beyond scores, we analyze performance by SCM family, graph and mechanism properties, sample size, intervention protocol, and semantic or scientific setting. The shared interface and protocol are fixed and standardized so that later rounds can introduce new SCMs, domains, and evaluation settings.

Our experiments yield several consistent conclusions. Rankings of classical, neural, and pretrained methods shift substantially across Synthetic, Semantic, Formula, and real-data settings: no method dominates every slice, and strong performance on synthetic SCMs does not reliably transfer to semantic, scientific, or real tables. Extra samples and interventions help methods unevenly, so a single pooled score is an incomplete summary. Beyond ranking methods, CausalArena aims to make such results diagnosable and, through a shared executable interface, to support successive evaluation rounds that can guide and re-test progress as new algorithms and foundation models appear.

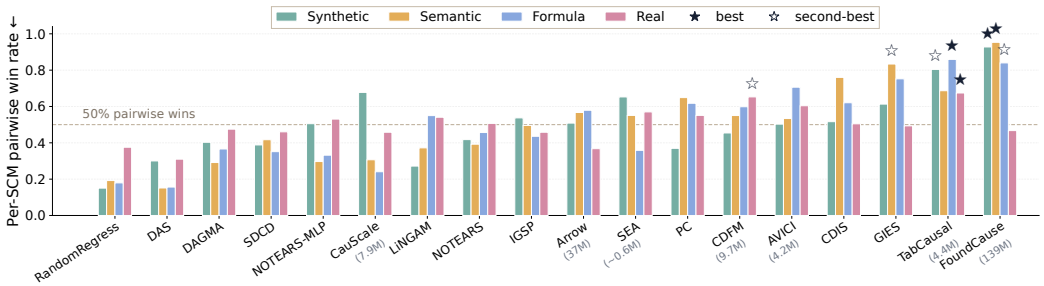


Figure 1: **Observation-only pairwise win profile across benchmark families.** For each SCM, method scores are first averaged over repeated runs. On each unit, every method pair is compared separately on F1 and SHD; higher F1 and lower SHD each count as a win, ties count as 0.5, and the reported win rate pools wins from both metrics. Each bar reports this pooled pairwise win fraction within one benchmark family. Values such as `139M` under pretrained method names denote parameter counts in millions for the evaluated checkpoints.

Figure 1 summarizes this ranking instability under a shared observation-only protocol. Methods are ordered by mean pairwise win rate across families, yet the family-wise bars show that the leader and second place often change from Synthetic to Semantic, Formula, and real data; a method that wins broadly on one family can sit mid-pack on another. The figure therefore motivates evaluating causal discovery under multiple SCM regimes. Our main contributions are:

- We formulate causal discovery evaluation in the foundation-model era around four requirements—breadth, freshness, grounding, and diagnosability.
- We introduce CausalArena, which combines broad synthetic SCMs, semantically grounded operational SCMs, and formula-grounded scientific SCMs whose interface can evolve by admitting new SCMs without changing how methods are compared.
- We conduct a broad evaluation of classical, neural, and pretrained causal discovery methods and provide family-, mechanism-, sample-size-, intervention-, and real-data analyses, revealing substantial ranking shifts and distinct failure profiles across evaluation environments.

2 Benchmark Formulation and Design

2.1 Task and Evaluation Formulation

We consider causal discovery over d observed variables $\mathbf{X} = (X_1, \dots, X_d)$ generated by an SCM $\mathcal{S}$. An SCM specifies a DAG $G = (V, E)$ together with structural assignments

$$X_j := f_j\big(\mathbf{X}_{\mathrm{Pa}_G(j)}, \epsilon_j\big)\,, \qquad j = 1, \dots, d, \tag{1}$$

where $\mathrm{Pa}_G(j)$ denotes the parents of X_j, f_j is its causal mechanism, and ϵ_j is an exogenous noise variable (Spirtes et al., 2000; Peters et al., 2017). Let $A_G \in \{0,1\}^{d\times d}$ denote the directed adjacency matrix of G. The task is to recover A_G from samples generated by $\mathcal{S}$. Here $\mathcal{S}$ supplies the controlled evaluation environment together with known ground-truth structure.

Each executable SCM induces an observational distribution $P_{\mathcal{S}}(\mathbf{X})$ and, after a do-style perturbation on intervention targets I, an interventional distribution $P_{\mathcal{S}}(\mathbf{X} \mid \mathrm{do}(I))$. We therefore construct two evaluation settings from the same SCM: an observation-only dataset $\mathcal{D}^{\mathrm{obs}}$ and an observation-plus-intervention dataset $\mathcal{D}^{\mathrm{obs+int}}$. A method $\mathcal{M}$ receives one of these datasets and returns

$$\widehat{A}_G = \mathcal{M}(\mathcal{D}), \tag{2}$$

which is scored against A_G with graph-recovery metrics such as F1 and SHD.

Classical and foundation-model pipelines differ in how $\mathcal{M}$ is obtained. A classical (or per-dataset neural) method is applied independently to each evaluation instance,

$$\psi_i \leftarrow \mathrm{Learn}(\mathcal{D}_i), \qquad \widehat{A}_i = \mathrm{Decode}(\psi_i; \mathcal{D}_i)\,, \tag{3}$$

so each evaluation dataset $\mathcal{D}_i$ induces its own fitted state ψ_i. A CDFM instead first learns reusable parameters from a collection of pretraining environments $\mathcal{E}_{\text{pre}}$,

$$\theta = \text{Pretrain}(\mathcal{E}_{\text{pre}}), \qquad \widehat{A}_i = \text{Decode}_\theta(\mathcal{D}_i), \tag{4}$$

as in amortized and pretrained causal discovery (Lorch et al., 2022; Qiao et al., 2026; Li et al., 2026). The distinction is whether structure-learning parameters are estimated separately for each evaluation dataset or amortized across many pretraining environments and reused at test time. For a benchmark $\mathcal{B} = \{(\mathcal{D}_i, A_i)\}_{i=1}^N$, a conventional aggregate score is

$$\text{Score}(\mathcal{M}; \mathcal{B}) = \text{Agg}_i \left[m\left(\widehat{A}_i, A_i\right) \right], \tag{5}$$

where m is a graph-recovery metric and Agg aggregates over instances and replicates.

2.2 Design Requirements

Concretely, three SCM families—synthetic, semantic operational, and formula-grounded scientific—are compiled into one executable specification. Classical, neural, and pretrained methods return directed adjacency estimates under those interfaces; scores are reported as a pooled $\text{Score}(\mathcal{M}; \mathcal{B})$ and broken down by family, mechanism, protocol, and domain factors. Public real-world tables with published graphs serve as an additional external-validity check alongside the three families. Figure 2 summarizes this organization.

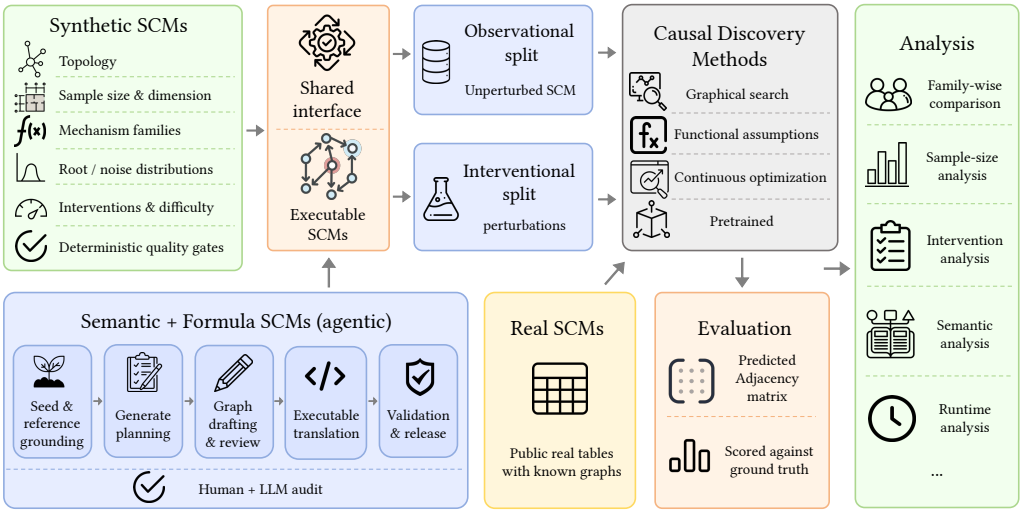


Figure 2: **Overview of the CausalArena framework.**

Let $\phi(\mathcal{S})$ denote construction factors of an SCM , and write slices of the arena as $\{\mathcal{B}_c\}$ indexed by factors in ϕ. We use this notation to state the four design criteria below.

Breadth. Because $\text{Score}(\mathcal{M}; \mathcal{B})$ depends on the chosen $\mathcal{B}$, the arena must support a diverse collection of slices $\{\mathcal{B}_c\}$—graph families, mechanisms, noise, dimensions, sample sizes, interventions, and difficulty—so that performance is assessed across many causal environments.

Freshness. Because CDFM predictions use shared parameters $\theta = \text{Pretrain}(\mathcal{E}_{\text{pre}})$, a fixed public $\mathcal{B}$ cannot remain outside $\mathcal{E}_{\text{pre}}$ indefinitely. Freshness requires that the evaluation pool can grow: if $\mathcal{E}_{\text{pre}}$ expands, the arena should admit $\mathcal{S}_{\text{new}}$. CausalArena therefore keeps the interface and protocol fixed and standardized, so that newly authored SCMs can be added as the arena evolves.

Grounding. Because anonymous variation of ϕ alone leaves open whether $\widehat{A}_G$ remains meaningful outside synthetic generators, some environments must attach operational or scientific meaning to variables, edges, and interventions. A recovered edge is then a binary entry in $\widehat{A}_G$ and an interpretable claim about a named process or equation.

Diagnosability. Because a pooled $\mathrm{Score}(\mathcal{M}; \mathcal{B})$ leaves success and failure unexplained, evaluation should retain $\phi(\mathcal{S})$ and report slice scores

$$\mathrm{Score}_c(\mathcal{M}) = \mathrm{Agg}_{i:\, c(i)=c} \left[m(\widehat{A}_i, A_i) \right], \tag{6}$$

so that rankings can be read as failure profiles across environments, together with a global order.

2.3 Evolvable Evaluation for Foundation Models

The dependence of Eq. (4) on $\mathcal{E}_{\mathrm{pre}}$ makes benchmark release fundamentally different for CDFMs. Publishing executable SCMs is valuable for reproducibility, inspection, and reuse, while those same SCMs can later enter pretraining corpora. A permanently hidden test would make construction and evaluation hard to inspect and reproduce, and public in-distribution / out-of-distribution splits alone leave the tension unresolved once SCMs have been released.

CausalArena therefore treats pretraining overlap as part of how a result is interpreted. Training disclosure clarifies what a reported score uses; the shared executable interface and common protocol allow newly constructed graphs, mechanisms, and scenarios to be added as the arena evolves. In this sense, *freshness* is enabled by evolving the evaluation pool under a fixed interface. A complementary pretraining-exposure OOD diagnostic is reported in Section E.2; release policy, contamination risks, and their limitations are discussed further in Section F.

3 Benchmark Implementation

To implement the four design criteria, CausalArena builds three complementary SCM families, detailed in Section B.1, and compiles them into one arena. Each SCM specification provides a DAG G, structural mechanisms $\{f_j\}$, exogenous or root distributions, intervention definitions, and the metadata required for generation and analysis. Once compiled, all families use the shared sampling and scoring interfaces in Section 3.5, with wrapper-level details in Section C.

The current arena contains 1,200 executable SCM specifications: 1,000 synthetic configurations, 100 semantic operational scenarios, and 100 formula-grounded scientific scenarios. The anonymous review package currently releases half of each family.

3.1 Synthetic SCMs: Controlled Breadth

The synthetic family is designed to systematically cover variations in the components of an SCM. We programmatically vary graph topology and dimension, structural mechanisms, root and noise distributions, dependencies among root variables, intervention settings, and difficulty factors. These factors yield different combinations of G, $\{f_j\}$, and noise distributions, while retaining exact ground-truth structure for evaluation.

The current pool contains 1,000 configurations over dimensions $\{10, 20, 30, 50, 100\}$, spanning 10 graph families, 16 mechanism families, 14 root-distribution families, 15 noise families, 4 root-dependency families, and 9 difficulty settings. The design includes high-dimensional configurations for stress testing and deliberately varies factors that can otherwise become implicit generator priors. Detailed family definitions and sampling rules are given in Section B.2.

Before release, deterministic quality gates check DAG validity, mechanism executability, numerical ranges, degenerate variables, and the intended coverage factors. These checks ensure that controlled variation in the generator corresponds to valid evaluation instances rather than numerical or implementation artifacts. Programmatic audits are detailed in Section B.3.

3.2 Semantic Operational SCMs: Grounded and Extensible Environments

The semantic family moves beyond unnamed synthetic variables by constructing SCMs around operational processes with interpretable variables, causal links, measurements, and interventions. Nodes are labeled by what they mean in that process, and directed edges carry reasons for why changing a parent can affect a child. Interventions change settable fields in the same executable SCM, and descendants are recomputed accordingly. Each specification also stores variable definitions,

measurement notes, intervention meanings, and edge reasons, so the causal structure is inspectable in ways that unnamed synthetic graphs are not, while retaining the same tabular input and directed adjacency target used by the other families.

The current semantic pool contains 100 scenarios from 10 operational domains: cybersecurity and IT operations, education, finance and credit, government services, healthcare delivery, housing and real estate, manufacturing, public health, urban transportation, and water and sanitation. Each domain contains 10 scenarios, with 17–25 variables and 25–67 directed edges per scenario.

Because semantic SCMs are authored through a reusable construction pipeline, new operational scenarios can later be added without changing the evaluation interface. This property provides a practical route toward the evolvability and future freshness discussed in Section 2.2; the current release itself does not assume that semantic knowledge is unseen by all pretrained models.

3.3 Formula-Grounded Scientific SCMs: Mechanistic Grounding

The formula-grounded family builds SCMs in which some mechanisms are named scientific or engineering equations, with units and valid ranges written down. Each scenario starts from one or more such equations together with the surrounding measurement and correction steps. Those equations define selected structural mechanisms f_j. Formula nodes are computed exactly from their parents. Randomness sits on inputs, instrument readings, or other non-formula quantities. After sampling, residual checks verify that rows that were not intervened still match the scientific relation.

The current pool contains 100 scenarios spanning 10 scientific and engineering domains: biology and ecology, astronomy, chemistry, mechanics and fluids, earth systems, electromagnetism, energy systems, materials and structures, optics and waves, and thermodynamics. Each domain contributes 10 scenarios, with 16–25 variables and 18–50 directed edges per scenario. Family-specific validation checks formula orientation, units, numerical validity ranges, and equation residuals.

3.4 Construction Workflow, Validation, and Diversity Control

Semantic and formula-grounded SCMs share a multi-pass construction pipeline. It starts from seed banks and grounding notes, performs structure planning before graph writing, expands the selected plan into a reviewed graph card, translates the graph card into an executable generator specification, and then applies deterministic validation, audit, and release gates. Human and LLM-assisted review focuses on variable necessity, causal-edge rationale, intervention meaning, measurement realism, and potential template or alias artifacts. The formula-grounded family additionally verifies equation orientation, units, valid ranges, and residual consistency.

The synthetic family does not require the same semantic authoring process; instead, deterministic programmatic checks validate graph structure, mechanisms, distributions, numerical behavior, and coverage constraints. Across all three families, diversity control removes redundant graphs, near-duplicate scenarios, repeated templates, unsupported edges, and low-information variables before inclusion in the evaluation pool.

3.5 Shared Protocol and Compared Methods

All families are scored under one shared protocol. For every executable SCM $\mathcal{S}_i$, CausalArena stores the ground-truth adjacency A_i and exports two matched tables from the same graph: an observation-only split $\mathcal{D}^{\mathrm{obs}}$ from the unperturbed SCM, and an observation-plus-intervention split $\mathcal{D}^{\mathrm{obs+int}}$ obtained by do-style perturbations with descendant recomputation. Intervention targets are recorded in masks shipped with the data. On the main leaderboard, the observation-only split uses $n_{\mathrm{obs}} = 1000$ rows; the mixed split uses $n_{\mathrm{obs}} = 800$ observational and $n_{\mathrm{int}} = 200$ interventional rows. Methods that do not consume intervention indicators are run only on supported splits, and unsupported cells are recorded as missing.

Every method wrapper receives standardized tabular inputs and returns a directed adjacency estimate. Each SCM is evaluated with multiple independent replicates (five in the current main protocol); replicate scores are averaged before aggregation over families, domains, dimensions, or the full arena. Primary metrics are directed-edge precision, recall, F1, and SHD relative to A_i; slice scores follow Eq. (6), so rankings can be read by family, mechanism, domain, sample size, or intervention

protocol as well as in pooled form. Thresholding, CPDAG/PAG conversion rules, runtime budgets, and wrapper defaults are specified in Section C.

Under this protocol the arena compares 18 methods with common wrappers: a random-order regression control (RandomRegress); graphical search (CDIS, GIES, IGSP, and PC); functional-assumption methods (DAS and LiNGAM); continuous optimization (DAGMA, NOTEARS, NOTEARS-MLP, and SDCD); and pretrained or amortized methods (Arrow, AVICI, CauScale, CDFM, FoundCause, SEA, and TabCausal). The next section reports what these methods do under the current release.

4 Results and Observations

We evaluate the methods and SCM pool from Section 3 under the shared protocol in Section C. The focus here is on observations: how rankings move across families and splits, how real tables and protocol choices change those rankings, and what that implies for foundation-model evaluation. Additional sample-size, intervention-protocol, tabular-foundation-model skeleton-probe, construction, runtime, and accuracy–cost experiments follow in Section D; factor-level, OOD, and Semantic/Formula analyses follow in Section E.

Diagnostic observations. The appendix analyses give several observations that help interpret the compact results below.

- **Sample-size saturation.** Most methods improve mainly from $n = 100$ to the low-thousands regime, after which the curves flatten. Additional observational rows therefore do not uniformly help pretrained wrappers, and sample-size gains should be read together with the fixed graph dimension and SCM mix used in that diagnostic.
- **Intervention-protocol sensitivity.** Methods differ substantially in how sensitive they are to the form of interventional data. Some remain relatively stable across intervention protocols, whereas others change markedly when the same sample budget is distributed across high-fraction, multi-target, or dense intervention regimes.
- **Generic TFM probes.** Generic tabular foundation models can be adapted to causal-discovery-style probes through feature attention, but the resulting skeletons are highly threshold-dependent and undirected. In their current form, these probes indicate that generic TFMs are not yet strong directed causal-discovery baselines.
- **Runtime and accuracy–cost trade-offs.** Runtime comparisons need to separate one-time model load from per-graph inference for pretrained wrappers. The accuracy–cost frontier shows that the fastest methods are not necessarily the most useful: cheap low-F1 wrappers and stronger, costlier methods occupy different parts of the practical frontier.
- **Factor preferences.** Factor-level analyses show clear method preferences across graph families, mechanisms, noise processes, and root-dependency patterns. A method that is strong on one construction factor can become weak on another, even under the same pooled benchmark.
- **Pretraining-support OOD.** The OOD analysis shows that the strongest pretrained methods do not owe their ranking only to documented pretraining support: FoundCause and TabCausal remain above the graphical-search reference on held-out factor categories, while methods with narrower or less balanced support show larger drops.
- **Scenario-level failure modes.** Semantic and Formula analyses make aggregate errors interpretable at the scenario level. They show not only which method ranks higher, but also what kinds of semantic processes or scientific mechanisms tend to trigger failures.

The main text therefore gives the aggregate rankings first, while the appendix analyses explain why those rankings change across benchmark conditions.

4.1 Main Results

Figure 1 previews cross-family ranking instability under the shared observation-only protocol. For each SCM or real dataset, we first average F1 and SHD *separately* over repeated runs, yielding one score per metric per unit. The intro figure then compares every method pair on F1 and on SHD independently (higher F1 wins; lower SHD wins; ties count as 0.5) and reports the pooled win fraction over both metric-wise comparisons within each family (full matrix in Figure 20). The tables

and plots below use the same per-unit, replicate-averaged scores, summarized as family-level means.

Table 2 compares all evaluated methods on Synthetic, Semantic, and Formula benchmarks. Figure 3 plots directed-edge precision against recall and F1 against SHD, with obs-only and obs+int shown together. Complete method-wise metrics are in Section G.1 for Synthetic, Section G.2 for Semantic, and Section G.3 for Formula.

Table 2: **Main benchmark results (F1 / SHD).** Cells report mean (standard deviation). Higher F1 and lower SHD are better; bold and underline mark the best and second-best F1 and SHD separately in each column.

Type	Method	Syn. obs	Syn. obs+int	Sem. obs	Sem. obs+int	Form. obs	Form. obs+int	Avg.
Control	RandomRegress	0.28 (0.07) / 218.0 (308.7)	–	0.25 (0.07) / 93.2 (69.0)	–	0.20 (0.06) / 108.3 (69.9)	–	0.24 (0.07) / 139.8 (149.2)
Graphical search	CDIS	0.38 (0.14) / 143.0 (351.2)	0.40 (0.13) / 140.1 (352.9)	0.40 (0.17) / 31.7 (9.0)	0.40 (0.16) / **31.9** (8.9)	0.23 (0.18) / 31.0 (6.8)	0.23 (0.18) / 30.9 (6.9)	0.34 (0.16) / 68.1 (122.6)
	GIES	0.48 (0.13) / 123.6 (174.1)	0.48 (0.11) / 134.0 (180.7)	0.55 (0.13) / 33.4 (15.2)	**0.57** (0.13) / 32.9 (15.4)	0.47 (0.12) / 36.7 (14.1)	0.52 (0.13) / 34.7 (14.0)	0.51 (0.13) / 65.9 (68.9)
	IGSP	0.41 (0.12) / 135.6 (191.9)	0.39 (0.12) / 132.8 (185.7)	0.31 (0.16) / 40.8 (13.8)	0.31 (0.16) / 39.9 (12.9)	0.21 (0.15) / 43.3 (16.9)	0.16 (0.16) / 40.7 (15.9)	0.30 (0.15) / 72.2 (72.8)
	PC	0.31 (0.12) / 115.4 (148.9)	–	0.33 (0.15) / 33.5 (9.4)	–	0.22 (0.16) / 31.1 (6.8)	–	0.29 (0.14) / 60.0 (55.0)
Functional assumptions	DAS	0.34 (0.11) / 142.6 (179.8)	–	0.21 (0.06) / 59.5 (15.4)	–	0.16 (0.06) / 64.8 (15.5)	–	0.23 (0.08) / 89.0 (70.2)
	LiNGAM	0.21 (0.12) / 109.6 (142.2)	–	0.20 (0.08) / 38.9 (9.3)	–	0.24 (0.08) / 34.7 (8.0)	–	0.22 (0.10) / 61.1 (53.2)
Continuous optimization	DAGMA	0.28 (0.12) / 107.2 (138.6)	–	0.18 (0.07) / 42.0 (9.1)	–	0.18 (0.07) / 39.7 (8.6)	–	0.21 (0.09) / 63.0 (52.1)
	NOTEARS	0.28 (0.11) / 104.2 (136.0)	–	0.20 (0.07) / 38.1 (8.5)	–	0.18 (0.08) / 35.2 (7.3)	–	0.22 (0.09) / 59.2 (50.6)
	NOTEARS-MLP	0.35 (0.10) / 107.0 (136.3)	–	0.22 (0.08) / 44.9 (10.7)	–	0.19 (0.06) / 44.1 (8.9)	–	0.26 (0.08) / 65.3 (52.0)
	SDCD	0.38 (0.09) / 135.5 (165.2)	0.38 (0.09) / 148.1 (174.1)	0.36 (0.09) / 51.2 (16.0)	0.36 (0.09) / 51.7 (15.8)	0.28 (0.08) / 61.5 (16.0)	0.29 (0.09) / 61.6 (16.3)	0.34 (0.09) / 84.9 (67.2)
Pretrained	Arrow	0.34 (0.13) / 106.0 (140.6)	–	0.31 (0.09) / 36.6 (9.5)	–	0.27 (0.08) / 35.9 (8.6)	–	0.31 (0.10) / 59.5 (52.9)
	AVICI	0.32 (0.15) / 104.5 (139.8)	0.35 (0.16) / 104.2 (139.8)	0.28 (0.13) / 35.5 (9.3)	0.30 (0.13) / 34.8 (9.3)	0.33 (0.12) / 31.8 (8.2)	0.41 (0.12) / 29.3 (8.0)	0.33 (0.14) / 56.7 (52.4)
	CauScale	0.41 (0.11) / 85.8 (107.1)	0.46 (0.10) / **84.6** (106.6)	0.25 (0.07) / 50.2 (19.2)	0.29 (0.08) / 48.6 (18.8)	0.21 (0.07) / 65.9 (23.4)	0.28 (0.09) / 59.6 (23.1)	0.32 (0.09) / 65.8 (49.7)
	CDFM	0.42 (0.14) / 136.4 (176.5)	–	0.43 (0.07) / 46.0 (12.4)	–	0.41 (0.08) / 45.8 (13.9)	–	0.42 (0.10) / 76.1 (67.6)
	FoundCause	**0.62** (0.12) / **80.8** (111.8)	–	**0.63** (0.11) / **26.8** (10.5)	–	**0.52** (0.12) / 33.4 (12.3)	–	**0.59** (0.12) / **47.0** (44.9)
	SEA	0.40 (0.13) / 99.2 (129.1)	–	0.33 (0.11) / 41.8 (17.9)	–	0.24 (0.09) / 52.8 (19.7)	–	0.32 (0.11) / 64.6 (55.6)
	TabCausal	0.50 (0.16) / 93.6 (127.3)	**0.50** (0.15) / 96.0 (128.4)	0.42 (0.13) / 36.1 (10.1)	0.45 (0.13) / 35.0 (10.2)	0.47 (0.13) / **29.5** (9.5)	**0.53** (0.14) / **26.6** (9.2)	0.48 (0.14) / 52.8 (49.1)

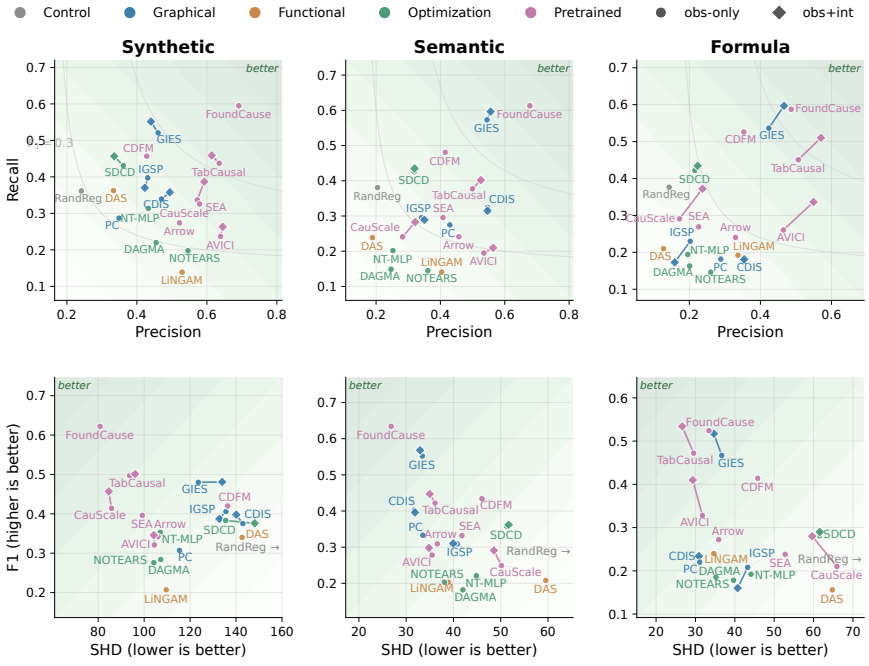


Figure 3: **Precision–recall and F1–SHD on the main benchmark.** Top row: directed-edge precision versus recall. Bottom row: F1 versus SHD. Circles are obs-only, diamonds are obs+int, and a line joins the two when both exist. Shaded corners markers indicate the preferred region.

The main observation is that rankings depend strongly on the evaluation environment. FoundCause is strongest among the supported observation-only slices, GIES is competitive on semantic slices, and TabCausal is strong with interventional evidence. No method wins every slice or metric: F1 and SHD sometimes disagree, and the order changes across Synthetic, Semantic, and Formula.

Figure 3 makes the same pattern geometric. Most methods occupy a mid-precision, mid-recall band; FoundCause sits farther toward the high-F1 corner on observation-only slices, while RandomRegress is the high-SHD outlier (off-scale on the bottom row). The obs+int setting is a fixed-budget mixed-evidence protocol: it replaces 200 observational rows with interventional rows instead of adding

interventions on top of the same observational sample. Thus, moving from obs-only to obs+int changes both the information type and the sample composition, producing modest gains or trade-offs between F1 and SHD depending on the method. Semantic is more spread out than Formula, where several methods collapse toward similar precision–recall and F1–SHD values. Factor-level and semantic analyses in Section E explain these shifts.

4.2 Real-data Check

Figure 4 evaluates the same wrappers on real scored tabular datasets with public samples and published directed graphs. The observation-only sources are six CD-CSG datasets (Zhao et al., 2023): Abalone, Auto MPG, Cardiac Arrhythmia, Concrete Compressive Strength, Deutscher Wetterdienst, and Ozone. The interventional sources are Sachs flow-cytometry (Sachs et al., 2005); PetShop high traffic, low traffic, temporal traffic 1, and temporal traffic 2 (Hardt et al., 2024); and the Causal Chambers Light Tunnel and Wind Tunnel experiments (Gamella et al., 2025). The observation-only aggregate combines these six CD-CSG datasets with observation-only conversions of the seven interventional datasets, obtained by dropping intervention rows; the interventional aggregate uses the seven datasets with intervention rows.

Type	Method	Obs-only	Obs+int
Control	RandomRegress	0.20 (0.14) / 59.9	–
Graphical	CDIS	0.28 (0.22) / 99.5	0.14 (0.17) / 45.4
	GIES	0.26 (0.15) / 48.0	0.29 (0.13) / 118.7
	IGSP	0.18 (0.22) / 31.1	0.11 (0.12) / 59.1
Functional	PC	0.18 (0.14) / 28.9	–
	DAS	0.19 (0.23) / 61.9	–
	LiNGAM	0.13 (0.17) / **24.3**	–
Optim.	DAGMA	0.16 (0.14) / 31.8	–
	NOTEARS	0.16 (0.15) / 29.1	–
	NOTEARS-MLP	0.21 (0.19) / 55.0	–
Pretrained	SDCD	0.22 (0.17) / 44.5	0.27 (0.12) / 103.4
	Arrow	0.15 (0.13) / 70.5	–
	AVICI	0.22 (0.18) / 26.0	0.28 (0.12) / 42.6
	CauScale	0.08 (0.08) / 36.2	0.19 (0.07) / 62.1
	CDFM	**0.39** (0.14) / 38.5	–
	FoundCause	0.21 (0.23) / 98.3	–
	SEA	0.21 (0.14) / 31.3	–
	TabCausal	0.34 (0.24) / 28.4	**0.46** (0.11) / **38.4**

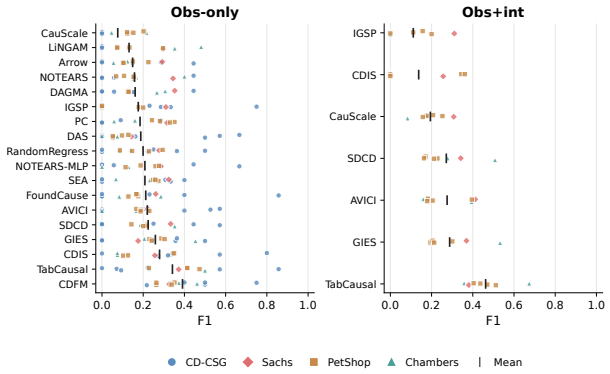


Figure 4: **Real-data check.** Left: compact F1 / SHD results, with cells reporting mean (std) over datasets (obs-only: 13; obs+int: 7). Right: per-dataset real-data F1; each marker is one scored table, and the tick is the method mean.

On these tables the generated-SCM ranking does not carry over: CDFM leads observation-only F1, TabCausal is second there and leads the observation-plus-intervention aggregate, while FoundCause is much weaker. The per-dataset markers show substantial variation within each method.

5 Conclusion

We introduced CausalArena, an evolvable benchmark for causal discovery in the foundation-model era. It combines broad synthetic SCMs, semantically grounded operational SCMs, and formula-grounded scientific SCMs under a shared observational and interventional protocol, supporting coverage, grounding, diagnosis, and extension with newly constructed causal environments. Experiments across classical, neural, and pretrained methods show that rankings vary substantially across SCM families, mechanisms, sample sizes, and intervention settings; strong performance in one regime does not necessarily transfer to others. For pretrained models, these results highlight the need to interpret scores in light of possible pretraining–evaluation overlap. By exposing where scores change and coupling reproducible public reference suites with an extensible SCM interface, CausalArena provides a foundation for continued evaluation as causal discovery methods and foundation models evolve.

Reproducibility Statement

We provide the benchmark construction details, evaluation protocols, method settings, runtime accounting, and complete result tables in the appendix and supplementary materials. The anonymous data and leaderboard package is available at `https://anonymous.4open.science/r/CausalArena-319B/`. The current review package includes a balanced half of each generated SCM family—500 of 1,000 synthetic SCMs, 50 of 100 semantic SCMs, and 50 of 100 formula-grounded SCMs—together with selected result files; the remaining SCMs are reserved for audited benchmark updates.

AI Use Statement

The authors used generative AI tools for language editing, code assistance, presentation-consistency checks, and the LLM-assisted Semantic/Formula SCM construction, quality assessment, and qualitative summaries described in Sections B and E. We did not use generative AI tools to produce unverified scientific claims or quantitative results. All AI-assisted content, code, benchmark artifacts, analyses, and final manuscript decisions were reviewed, validated, and controlled by the authors, who take full responsibility for the submitted work.

References

Nikolay Babakov, Ehud Reiter, and Alberto Bugarín. CausalGraphBench: a benchmark for evaluating language models capabilities of causal graph discovery. In *ACL Student Research Workshop*, pp. 240–258, 2025.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *NeurIPS*, 2022.

Patrick Blöbaum, Krishnakumar Balasubramanian, and Shiva Prasad Kasiviswanathan. FoundCause: Causal discovery with latent confounders from observational data, 2026. arXiv preprint arXiv:2606.17516.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *NeurIPS*, 2020.

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

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *ICLR*, 2024.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8(1):412, 2025.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

Haoyue Dai, Ignavier Ng, Jianle Sun, Zeyu Tang, Gongxu Luo, Xinshuai Dong, Peter Spirtes, and Kun Zhang. When selection meets intervention: Additional complexities in causal discovery. In *ICLR*, 2025.

Muhammad Hasan Ferdous, Emam Hossain, and Md Osman Gani. TimeGraph: Synthetic benchmark datasets for robust time-series causal discovery. In *KDD*, pp. 5425–5435, 2025.

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

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference, 2022. arXiv preprint arXiv:2202.02195.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Martin Roth, and Steffen Sonntag. causalAssembly: Generating realistic production data for benchmarking causal discovery. In *CLeaR*, pp. 609–642, 2024.

Léo Grinsztajn et al. TabPFN-3: Technical report, 2026. arXiv preprint arXiv:2605.13986.

Michaela Hardt, William Roy Orchard, Patrick Blöbaum, Elke Kirschbaum, and Shiva Kasiviswanathan. The PetShop dataset — finding causes of performance issues across microservices. In *CLeaR*, pp. 957–978, 2024.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. CausalDynamics: A large-scale benchmark for structural discovery of dynamical causal models. In *NeurIPS*, 2025.

Rebecca J. Herman, Jonas Wahl, Urmi Ninad, and Jakob Runge. Unitless unrestricted Markov-consistent SCM generation: Better benchmark datasets for causal discovery. In *CLeaR*, pp. 1506–1531, 2025.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NeurIPS*, 2008.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007.

Zi-Rong Li, Si-Yang Liu, Tian-Zuo Wang, and Han-Jia Ye. TabCausal: Pretraining across causal environments for tabular causal discovery, 2026. arXiv preprint arXiv:2605.31156.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *NeurIPS*, 2022.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *CLeaR*, pp. 752–771, 2023.

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

Achille Nazaret, Justin Hong, Elham Azizi, and David Blei. Stable differentiable causal discovery. In *ICML*, pp. 37413–37445, 2024.

Panayiotis Panayiotou, Audrey Poinsot, Alessandro Leite, Nicolas Chesneau, Marc Schoenauer, and Özgür Şimşek. CausalProfiler: Generating synthetic benchmarks for rigorous and transparent evaluation of causal machine learning. In *ICML*, 2026.

Bo Peng, Sirui Chen, Jiaguo Tian, Yu Qiao, and Chaochao Lu. CauScale: Neural causal discovery at scale. In *ICML*, 2026.

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, 2026. arXiv preprint arXiv:2607.11508.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model, 2026. arXiv preprint arXiv:2602.11139.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *NeurIPS*, 2021.

Felix L. Rios, Giusi Moffa, and Jack Kuipers. Benchpress: A versatile platform for structure learning in causal and probabilistic graphical models. *Journal of Statistical Software*, 114(12):1–43, 2025.

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.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

Muralikrishnna G. Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. NODAGS-Flow: Nonlinear cyclic causal structure learning. In *AISTATS*, pp. 6371–6387, 2023.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear Non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, second edition, 2000.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers - scaling up benchmarking of causal discovery for real-world time-series. In *ICLR*, 2025.

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

Guangya Wan, Yunsheng Lu, Yuqi Wu, Mengxuan Hu, and Sheng Li. Large language models for causal discovery: Current landscape and future directions. In *IJCAI*, pp. 10687–10695, 2025.

Yuhao Wang, Liam Solus, Karren D. Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *NeurIPS*, 2017.

Menghua Wu, Yujia Bao, Regina Barzilay, and Tommi Jaakkola. Sample, estimate, aggregate: A recipe for causal discovery foundation models. *Transactions on Machine Learning Research*, 2025.

Xingxuan Zhang et al. LimiX-2: A contextual mechanism network towards general structured-data intelligence, 2026a. arXiv preprint arXiv:2609.17488.

Yinghuan Zhang, Yufei Zhang, Parisa Kordjamshidi, and Zijun Cui. Bayesian network structure discovery using large language models. *Transactions on Machine Learning Research*, 2026b. Also available as arXiv:2511.00574.

Boxiang Zhao, Shuliang Wang, Lianhua Chi, Qi Li, Xiaojia Liu, and Jing Geng. Causal discovery via causal star graphs. *ACM Transactions on Knowledge Discovery from Data*, 17(7):98:1–98:24, 2023.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *NeurIPS*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *AISTATS*, 2020.

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, 2024. arXiv preprint arXiv:2406.04598.

APPENDIX

A RELATED WORK

A.1 CAUSAL DISCOVERY METHODS

Causal discovery aims to recover directed causal structure from observational or interventional data, typically represented by a directed acyclic graph (DAG) associated with an underlying SCM (Spirtes et al., 2000; Peters et al., 2017). Classical approaches differ substantially in their assumptions and learning principles. Constraint- and search-based methods infer structure from conditional-independence relations or graph scores, including PC (Spirtes et al., 2000; Kalisch & Bühlmann, 2007; Colombo & Maathuis, 2014), GIES (Hauser & Bühlmann, 2012), IGSP (Wang et al., 2017), and CDIS (Dai et al., 2025). Functional approaches exploit assumptions on causal mechanisms or noise, such as LiNGAM (Shimizu et al., 2006) and DAS (Montagna et al., 2023), which builds on additive-model structure learning (Bühlmann et al., 2014; Hoyer et al., 2008). Continuous-optimization methods instead formulate structure learning as differentiable optimization, including NOTEARS (Zheng et al., 2018), NOTEARS-MLP (Zheng et al., 2020), DAGMA (Bello et al., 2022), and SDCD (Nazaret et al., 2024). RandomRegress serves as a simple random-order regression control for sanity-checking scores against trivial baselines. Since these approaches rely on different structural, functional, and distributional assumptions, their relative performance can change substantially across data-generating conditions.

A.2 CAUSAL DISCOVERY FOUNDATION MODELS

More recently, amortized and pretrained structure learners have shifted causal discovery toward foundation-model-style inference. AVICI (Lorch et al., 2022) amortizes structure prediction over large collections of synthetic SCMs; SEA (Wu et al., 2025) aggregates classical discovery estimates from sampled variable subsets; Arrow (Thompson et al., 2026) factorizes graphs into skeletons and topological orders with an acyclicity guarantee; CauScale (Peng et al., 2026) targets neural discovery at large graph scales; CDFM (Qiao et al., 2026) treats unknown mechanisms as latents in a variational foundation-model recipe; FoundCause (Blöbaum et al., 2026) injects pairwise causal statistics and models latent confounding; and TabCausal (Li et al., 2026) pretrains across diverse tabular causal environments, including interventional settings. Unlike classical algorithms that are typically fitted or executed independently on each dataset, these models can encode priors learned from large synthetic or mixed causal environments. This changes the interpretation of benchmark performance: a test score may reflect not only causal structure-learning ability, but also similarity between the model's pretraining environments and the evaluation graphs, mechanisms, or generators. Comparisons among pretrained methods are therefore particularly sensitive to differences in their training distributions and evaluation protocols.

A.3 EVALUATION OF CAUSAL DISCOVERY

Synthetic SCMs are the dominant tool for scalable causal discovery evaluation because they generate samples together with known ground-truth graphs. Standard evaluations sample a DAG, assign structural mechanisms and noise distributions, and generate observational or interventional data (Spirtes et al., 2000; Peters et al., 2017). Many method papers consequently introduce their own synthetic evaluation suites, including those accompanying NOTEARS (Zheng et al., 2018), DAGMA (Bello et al., 2022), SDCD (Nazaret et al., 2024), DCDI (Brouillard et al., 2020), and NODAGS-Flow (Sethuraman et al., 2023). More general frameworks such as Benchpress (Rios et al., 2025), CausalProfiler (Panayiotou et al., 2026), and DECI's CSuite (Geffner et al., 2022) systematize generation and comparison across broader configurations. However, synthetic evaluation is itself sensitive to generator design: varsortability and related artifacts (Reisach et al., 2021), for example, demonstrate that methods may exploit properties of a particular data generator rather than recover causal structure in a more general sense. Complementary generator designs such as unitless unrestricted Markov-consistent (UUMC) SCM sampling (Herman et al., 2025) aim to reduce such nonphysical sortability patterns when constructing synthetic benchmarks.

A complementary line of work evaluates causal discovery on systems with semantic, scientific, or physical grounding. Classic resources include the Sachs protein-signaling data (Sachs et al.,

2005) and reference Bayesian networks in bnlearn (Scutari, 2010), while CauseEffectPairs (Mooij et al., 2016) remains a standard benchmark for bivariate cause–effect orientation, a narrower task than full multivariate DAG recovery. OCDB (Zhou et al., 2024) develops real-data evaluation with graph metrics designed for more comparable structure assessment. Domain-oriented resources provide richer generative mechanisms: CausalTime (Cheng et al., 2024) introduces realistically generated temporal data, CausalRivers (Stein et al., 2025) scales real-world hydrology time-series evaluation with constructed ground-truth graphs, CausalDynamics (Herdeanu et al., 2025) targets dynamical systems derived from scientific equations, CausalBench (Chevalley et al., 2025) focuses on perturbational biological networks, causalAssembly (Göbler et al., 2024) models industrial production processes, and Causal Chambers (Gamella et al., 2025) provides physical systems with controlled interventions. TimeGraph (Ferdous et al., 2025) further stresses nonstationarity, irregular sampling, missingness, and latent confounding in synthetic temporal settings. These settings improve realism and mechanistic grounding, but are typically specialized to particular domains, system classes, or data modalities.

Semantic information has also become increasingly relevant as language models are introduced into causal reasoning and graph construction. CausalGraphBench (Babakov et al., 2025) evaluates language-model-based graph discovery against curated causal structures, while PromptBN/ReActBN (Zhang et al., 2026b) studies language-assisted Bayesian-network construction in low- or no-data regimes; related work surveys the use of LLMs for causal extraction, prior injection, and graph refinement (Wan et al., 2025). Such resources evaluate a different source of causal information from purely numerical SCM benchmarks, since predictions can depend on semantic priors as well as statistical evidence.

B Benchmark Data Details

This section records the data construction details behind the current release.

Eqs. (3)–(5) make the Introduction's difficulties precise, and they constrain what a useful $\mathrm{Score}(\mathcal{M}; \mathcal{B})$ can mean. First, the score is defined only relative to a chosen $\mathcal{B}$: different papers use different graph families, mechanisms, noise processes, sample sizes, and intervention protocols, so a high score on one $\mathcal{B}$ need not transfer to another. Second, for a CDFM the score aggregates predictions $\widehat{A}_i = \mathrm{Decode}_\theta(\mathcal{D}_i)$ with $\theta = \mathrm{Pretrain}(\mathcal{E}_{\mathrm{pre}})$ as in Equations (4) and (5). Similarity between $\mathcal{E}_{\mathrm{pre}}$ and $\mathcal{B}$ may occur from an exact SCM or graph to a shared family or generator, so a high score may mix causal discovery ability with pretraining–evaluation affinity. Third, once a fixed public $\mathcal{B}$ is released, its SCMs can later enter $\mathcal{E}_{\mathrm{pre}}$, and the test distribution can overlap with later training. If $\mathcal{B}$ is narrow, the score tracks one generator prior; if it is fixed while $\mathcal{E}_{\mathrm{pre}}$ grows, the score can drift toward overlap; if it contains only anonymous tables, it says little about named processes or equations; if only a pooled value is reported, it hides where methods succeed or fail. CausalArena is therefore organized as a single evaluation arena whose benchmark $\mathcal{B}$ is built to answer these needs together.

These four requirements pull in different directions at once: breadth favors large regenerable grids; grounding favors named processes and equations; freshness favors newly authored environments over time; diagnosability favors explicit factors and slice reports. A single SCM family is unlikely to carry all four equally, which leads to the three complementary families next.

B.1 Three Complementary SCM Families

CausalArena assigns the requirements to three SCM families that do different jobs, then compiles them to one executable representation so that sampling, interventions, wrappers, and scoring stay comparable. The point of complementarity is coverage of roles that conflict if forced into one generator: scale and factor control on one side, operational and scientific meaning on the other, and an authoring path for new environments as models evolve.

Their construction procedures reflect their different roles in CausalArena. Synthetic SCMs are generated programmatically to provide controlled breadth over causal structures and data-generating mechanisms. Semantic operational SCMs are constructed around meaningful real-world processes to provide grounding and an extensible source of newly authored environments. Formula-grounded SCMs anchor parts of the causal system in explicit scientific equations, providing mechanistic

grounding and enabling more interpretable diagnosis. All SCMs pass family-specific validation and diversity checks before entering the evaluation pool.

Synthetic SCMs (breadth, and factor-level diagnosis). Breadth needs a regenerable factor grid: topology, mechanism and noise tags, dimension, interventions, and difficulty can be varied systematically and at large d. Anonymous synthetic SCMs are the practical carrier of that grid. Because $\phi(\mathcal{S})$ is explicit, the same family also supports diagnosability through slice scores in Eq. (6). What they leave open is meaning.

Semantic operational SCMs (operational grounding and freshness). Grounding needs variables and edges that mean something in a process—a policy setting, a queue length, a sensor reading, a counted outcome. Semantic SCMs supply that operational meaning and keep construction human-auditable. Because new scenarios can be authored through the same interface, they also provide a practical route to freshness as pretrained models evolve: new $\mathcal{S}_{\text{new}}$ can be added without changing the protocol. They do not replace the synthetic grid: operational scenarios are harder to enumerate at the same factor resolution and scale.

Formula-grounded scientific SCMs (mechanistic grounding and diagnosis). Grounding also needs scientific or engineering equations with units and validity ranges. Formula SCMs keep named identities exact given their parents and place noise on instruments and readouts. That structure strengthens mechanism-level diagnosability on equation-backed systems. Like semantic SCMs, they complement the synthetic family by supplying meaning that anonymous generators omit, at a cost in how densely ϕ can be swept.

Taken together, the synthetic family mainly carries *breadth* and factor-level *diagnosability*; the semantic family carries operational *grounding* and a path to *freshness*; the formula family strengthens mechanistic *grounding* and equation-level *diagnosability*. The shared protocol is what makes these roles jointly readable as one arena. Construction and validation details are in Section 3. The shared evaluation protocol and compared methods are specified there as well (Section 3.5).

B.2 Synthetic SCM Pool

The synthetic main benchmark contains 1,000 SCM configurations. It crosses 10 graph families with five dimensionalities $d \in \{10, 20, 30, 50, 100\}$ and 20 SCMs per graph-family/dimension pair. Each configuration is generated with both an observation-only split and an observation-plus-intervention split; generated files enter the release only after passing the audit rules in Section B.3.

Graph structure.

- **Erdos–Renyi:** random ordered graphs with edge density sampled from 0.05–0.25.
- **Scale-free:** preferential-attachment-style skeletons oriented by an acyclic order, with density 0.04–0.22.
- **Chain:** sparse path-like DAGs with density 0.03–0.16 and long directed propagation.
- **Tree:** branching sparse DAGs with density 0.03–0.16.
- **Layered:** layer-respecting DAGs with density 0.04–0.20.
- **Bipartite layered:** two-stage or multi-stage bipartite-like DAGs with density 0.04–0.20.
- **Small-world:** locally clustered DAGs with short directed paths, density 0.04–0.20.
- **Block-sparse layered:** block-structured layered DAGs with sparse cross-block edges, density 0.04–0.20.
- **Dense local-module:** modular DAGs with denser within-module connectivity, density 0.06–0.24.
- **Collider-rich:** DAGs that deliberately increase v-structure frequency, density 0.06–0.24.
- **Validity filters:** candidate graphs are resampled until they are DAGs with no isolated variables, sufficient graph depth, and at least $\lceil 0.8d \rceil$ directed edges. The optional in-degree cap is sampled from no explicit cap, 4, or 5 for $d \leq 20$, from 4, 5, or 6 for $20 < d \leq 50$, and from 5, 6, or 8 for $d = 100$.

Family mixing.

- **Mechanism/noise family count:** the number of active local mechanism and noise families increases with dimension: 2–4 at $d = 10$, 3–5 at $d = 20$, 4–6 at $d = 30$, 5–8 at $d = 50$, and 6–10 at $d = 100$.
- **Root-family count:** the number of active root families depends on the number of root variables: one family for a single root, 1–2 for at most three roots, 2–4 for at most eight roots, and 3–6 otherwise.
- **Local assignment:** after active families are selected, node-level mechanisms, root families, and noise families are assigned across the graph so that a single SCM can contain heterogeneous local structural equations and noise channels.

Mechanisms.

- **Linear additive:** parent weights are scaled by $0.8/\sqrt{k}$ and biases are sampled in $[-0.5, 0.5]$.
- **Polynomial additive:** powers 2–4 are used, with quadratic coefficients in $[-0.5, 0.5]$ and cubic coefficients in $[-0.15, 0.15]$.
- **Multiplicative parent product:** parent effects are multiplied after weight scaling by $0.8/\sqrt{k}$ and bias sampling in $[-0.4, 0.4]$.
- **Random-function product:** transformed parent responses are multiplied and rescaled with output scale in $[0.4, 1.6]$. Together with multiplicative parent product, this form is scored as one multiplicative family, so the analyses use 16 mechanism tags.
- **Additive interaction:** up to three parent pairs are sampled, with interaction weights scaled by $0.5/\sqrt{\#\text{pairs}}$.
- **Rational/ratio:** denominator-safe ratio transforms avoid near-zero denominators through bounded offsets and clipping.
- **Exponential/log:** exponential inputs are clipped to $[-3, 3]$ and paired with log or signed-log style transforms where needed.
- **Trigonometric/periodic:** periodic parent transforms use sine/cosine-style responses with randomized frequency and phase.
- **Piecewise/threshold:** cutoffs are sampled in $[-0.6, 0.6]$, low slopes in $[0.2, 1.0]$, high slopes in $[1.0, 2.5]$, and offsets in $[-1.5, 1.5]$.
- **Saturation/clipping:** saturating nonlinearities use response scales in $[0.8, 2.5]$.
- **Mixture/regime switching:** gates use scales in $[1, 4]$, cutoffs in $[-0.5, 0.5]$, and alternate response scales in $[0.5, 1.8]$.
- **Random neural:** one-hidden-layer neural mechanisms use 3–8 hidden units.
- **RFF/GP-style:** random-feature mechanisms use 64–256 features, length scales in $[0.25, 3.0]$, and output scales in $[0.4, 2.5]$.
- **Activation additive:** activation families include ReLU, leaky-ReLU, softplus, sigmoid, absolute value, sign, step, hard-tanh, ReLU6, ELU, SELU, SiLU, modulo, round, and rank transforms; modulo periods are in $[0.75, 3.0]$ and rounding steps in $[0.2, 1.0]$.
- **Tree-based:** shallow tree mechanisms use depth 2–4 with thresholds in $[-0.8, 0.8]$.
- **Discretization/assignment:** discretized mechanisms use 3–7 bins or clusters.
- **Aggregation nonlinearities:** max or log-sum-exp aggregation uses temperatures in $[0.3, 1.5]$.

Root distributions.

- **Uniform:** lower bounds are sampled in $[-2, 0]$ and upper bounds in $[0.5, 3.0]$.
- **Gaussian:** the mean is drawn from $\mathcal{N}(0, 1)$ and the standard deviation from $[0.5, 2.0]$.
- **Truncated Gaussian:** the mean is drawn from $\mathcal{N}(0, 0.8^2)$, the standard deviation from $[0.4, 1.4]$, and samples are clipped to mean ± 2.5 standard deviations.
- **Lognormal:** the log mean is sampled in $[-0.3, 0.7]$ and the log standard deviation in $[0.25, 0.9]$.

- **Gamma:** concentration is sampled in $[1, 5]$ and rate in $[0.5, 2]$.
- **Scaled beta:** shape parameters are sampled in $[0.8, 5]$, then values are mapped to a sampled interval with lower endpoint in $[-1, 0]$ and upper endpoint in $[0.8, 3.0]$.
- **Exponential:** the rate is sampled in $[0.4, 2.5]$.
- **Gumbel:** the location is sampled in $[-1, 1]$ and the scale in $[0.3, 1.8]$.
- **Bernoulli:** probabilities are sampled in $[0.15, 0.85]$.
- **Poisson:** rates are sampled in $[1, 12]$.
- **Ordinal categorical:** the number of ordered levels is sampled from $\{3, \ldots, 6\}$; category probabilities are produced by a softmax over i.i.d. normal logits, and samples take ordered values $0, \ldots, k-1$.
- **Mixture Gaussian:** the mixture probability is sampled in $[0.2, 0.8]$, the two means in $[-2, -0.2]$ and $[0.2, 2]$, and component standard deviations in $[0.3, 1.0]$.
- **Multinomial discrete:** the number of unordered categories is sampled from $\{2, \ldots, 10\}$; logits are sampled in $[-1.5, 1.5]$, converted to probabilities by softmax, and the sampled category index is standardized.
- **Zipf/power-law:** support size is sampled from $\{8, \ldots, 20\}$, probabilities are proportional to $v^{-\alpha}$ with $\alpha \in [2, 4]$, sampled values are capped at 10, and the result is centered.
- **Root dependency:** 800 of the 1,000 synthetic SCM configurations use independent standardized Gaussian root blocks. The remaining 200 use multi-root dependent blocks: a common-factor Gaussian with $\rho \in [0.25, 0.75]$ and random signs; a random-covariance Gaussian whose latent dimension is sampled from $1, \ldots, r$ for r root variables with covariance $LL^\top + 0.15I$; or a uniform-ball sampler that draws a random direction, radius $U^{1/r}$, scale in $[0.75, 2.0]$, and center coordinates in $[-0.5, 0.5]$ before column standardization.

Structural noise.

- **Gaussian:** additive Gaussian noise uses scale $[0.05, 0.45]$.
- **Uniform:** additive bounded noise uses scale $[0.05, 0.45]$.
- **Laplace:** heavy-tailed additive noise uses scale $[0.03, 0.35]$.
- **Student-t:** degrees of freedom are sampled in 2.5–8 and scale in $[0.03, 0.30]$.
- **Multiplicative lognormal:** multiplicative noise uses $\sigma \in [0.03, 0.25]$.
- **Heteroskedastic:** the scale is $\text{base} + \text{slope} \cdot \sigma(|s|)$ with base in $[0.02, 0.15]$ and slope in $[0.02, 0.20]$.
- **RFF-heteroskedastic:** parent values pass through 64–256 random Fourier features with length scale $[0.25, 3.0]$ and output scale $[0.3, 1.8]$; a softplus transform converts this latent function into a sample-specific noise scale.
- **Target-R^2 Gaussian:** the noise variance is chosen from the empirical signal variance to target $R^2 \in [0.1, 0.9]$.
- **Mixture-outlier:** a base scale in $[0.02, 0.15]$ is mixed with rare outliers whose probability is $[0.01, 0.08]$ and scale is $[0.8, 2.5]$.
- **Quantized:** Gaussian-noisy values are rounded to a grid with step size $[0.05, 0.35]$.
- **Censored:** Gaussian-noisy values are clipped at their empirical 3rd and 97th percentiles.
- **Beta:** centered beta noise uses shape parameters in $[1, 10]$ and scale in $[0.05, 0.45]$.
- **Exponential:** centered one-sided disturbances use rate $[0.8, 4.0]$ and scale $[0.03, 0.35]$.
- **Gumbel:** centered extreme-value disturbances use scale $[0.03, 0.35]$.
- **No-noise option:** deterministic local equations are allowed, but the generator limits no-noise assignments so that at most 10% of nodes in a graph use this option.

Difficulty variants.

- **Schedule:** within each graph-family/dimension block, the generator cycles through one clean baseline case, two single-stressor cases, and one paired-stressor case.
- **Signal and noise stressors:** weak-signal cases multiply structural mechanism strength by 0.45; high-noise cases add extra Gaussian structural noise with scale 0.35 after the selected node noise.
- **Observation stressors:** feature warping applies signed-log, tanh, or signed-square-root transforms to about 55% of columns; outlier contamination perturbs about 2.5% of entries with heavy-tailed jumps; missingness/imputation affects about 5% of entries and fills them with column means plus small jitter; feature corruption permutes about 3% of entries within the same column.
- **Intervention stressors:** interventional distribution shift applies shift/scale perturbations to selected interventional columns; soft interventions blend the original structural value with the intervened value using softness 0.45.

Interventions.

- **Main split:** the obs+int split uses 800 observational rows and 200 interventional rows.
- **Target-pool size:** each SCM samples an intervention target pool whose size grows with dimension: 2–4 targets for $d = 10$, 3–6 for $d = 20$, 4–8 for $d = 30$, 5–12 for $d = 50$, and 8–20 for $d = 100$.
- **Row assignment:** every interventional row selects one target from this pool using balanced assignment, applies a hard do-resampling intervention, records the binary intervention mask, and recomputes all descendants.
- **Intervention value:** the do-resampling value is drawn uniformly from the target's empirical baseline 10th–90th percentile range. Intervention targets may include root or structural variables unless a sensitivity suite explicitly restricts targets to structural nodes.

B.3 Synthetic Audit and Protocol-sensitivity Suites

Synthetic SCMs are audited before release. Graph-level checks require an acyclic graph, no isolated variables, minimum depth 3 for chain/tree/small-world/scale-free families and 2 otherwise, and at least $\lceil 0.8d \rceil$ edges. Dataset-level checks require finite values, no constant or near-constant continuous variables, no discrete variable with a single category occupying more than 99% of rows, boundary mass at explicit clipping boundaries at most 10%, no near-duplicate variable pair above correlation 0.99, and intervention masks consistent with the target pool. For obs+int files, every recorded target must receive at least one affected row, target values must change relative to their observational mechanism, and descendants are regenerated under the intervened values.

The sample-size suite uses a fixed $d = 30$ representative subset of the synthetic pool, with 10 SCMs per graph family and $n \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$. In the main text we report the observation-only sample-size results to isolate observational sample count. The intervention-protocol suite also uses 100 SCMs at $d = 30$, but changes the intervention generator while keeping the same graph/mechanism pool:

- **Default resampling** uses the main protocol: 800 observational rows, 200 interventional rows, one target per interventional row, all sampled targets eligible, and hard do-resampling from the target's 10th–90th percentile range.
- **Fixed setpoint** keeps the $800/200$ split and one target per row, but sets targets to one of their 20th, 50th, or 80th percentile baseline values with jitter equal to 1% of the interquartile range.
- **Parameter shift** keeps the $800/200$ split, restricts targets to structural non-root nodes, and changes the target mechanism by scaling parent effects, shifting a bias/threshold by 0.3–0.8 interquartile ranges, or changing nonlinear strength by a factor sampled from 0.6–0.85 or 1.15–1.6.
- **High intervention fraction** uses 200 observational rows and 800 interventional rows with single-target hard resampling.
- **Multi-target per sample** keeps the $800/200$ split but samples 2–4 targets per interventional row and applies hard resampling to each selected target.

- **Dense mixed intervention** combines the harder settings: 200 observational rows, 800 interventional rows, 2–4 structural targets per interventional row, and a balanced mixture of hard resampling, fixed-setpoint interventions, and parameter shifts.

B.4 Agentic Construction Workflow and Quality Gates

Semantic and Formula SCMs are built with a multi-pass construction pipeline. The main goal is to avoid a common failure mode: a scenario may sound plausible in descriptions but still fail as an executable causal model. For the main construction batch and later repair passes, each scenario is saved as a sequence of records: the initial seed, grounding notes, selected structure plan, graph card, review notes, executable generator specification, validation report, revision log, and final manifest status. These records make it possible to trace why a variable or edge was included and how it is simulated. The construction ablation in Figure 9 uses four snapshots from this pipeline: seed only, after grounding, after structure planning, and after graph review.

Stage 1: seed bank and grounding notes. The pipeline starts from human-curated seeds. Semantic seeds follow a fixed 10-domain plan with 10 scenarios per domain. Each seed states the operational setting, the causal question, the measurement context, and plausible actions or interventions. Formula seeds state the scientific task family, candidate formula anchors or a formula-selection rule, the intended response variable, units, validity conditions, and experimental context. The first pass does not draw edges. Instead, it writes a grounding pack. For Semantic SCMs, this pack lists concrete records, sensors, actions, queues, delays, capacities, outcomes, and intervention handles. For Formula SCMs, it lists equations, constants, units, validity ranges, instruments, calibration or readout variables, and orientation constraints. This grounding pack is used to constrain the later graph.

Stage 2: structure planning before graph writing. The next pass decides the rough graph shape before detailed node and edge text is written. For each selected seed, the pipeline considers structurally different graph options and chooses one under batch-level diversity constraints. The selected plan records the intended node range, edge-density regime, root and sink structure, graph depth, recurring motifs, and likely intervention targets. The graph-design target is 15–25 observed variables with batch-average size close to 20; the Formula validator allows a wider 10–30-node safety range for justified cases, although the released Formula pool remains within 16–25 variables. The same plan also assigns each planned variable to a root sampler or a child mechanism family, together with a noise family, range policy, and possible intervention resampler. To avoid template-like graphs, the plan must contain several scenario-justified mechanism types and keeps generic bounded-sigmoid child mechanisms below 25% of children.

Stage 3: graph card, review, and repair. The selected plan is then expanded into a graph card with node definitions, directed edges, intervention handles, and local mechanism descriptions. Each node must state what it means, how it is measured, its scale and unit when applicable, its numerical range, its data source or sampling story, whether it can be intervened on, and how it is generated from its graph parents. Each edge must have a local causal explanation, rather than only saying that two variables are associated. The review pass checks for implausible causal directions, missing parents, vague latent placeholders, alias variables, unit-conversion duplicates, tiny-noise near copies, unsupported shortcuts, and low-information readout or status nodes. Formula graphs receive extra checks for equation correctness, unit compatibility, validity ranges, solve-for direction, response-variable orientation, invalid reverse directions, and whether calibration or readout nodes are scientifically justified. Blocking issues are sent back as graph-revision requests, and the revised graph is checked again.

Stage 4: executable generator specification. After graph approval, the graph card is translated into an executable generator specification. This specification maps each node to either a root sampler or a structured child operation, and maps each edge to a saved adjacency entry and parent contribution. A child operation must use exactly the parents declared in the graph: it cannot introduce hidden parents, ignore declared parents, depend on batch-level row statistics, or rely on prose-only mechanisms. The specification also records node-level noise, range handling, intervention resampling rules, generated-column names, masks, and metadata. Formula SCMs additionally store equation identifiers, inputs, outputs, constants, validity constraints, and residual checks. Formula outputs are recomputed from

parents and are not clipped simply to make them pass a range check. For mixed-interventional data, the generator resamples the targeted variables, records the intervention mask, and recomputes descendants.

Stage 5: deterministic validation and release gate. The final gate is run on the generated artifacts. It checks schema validity, acyclicity, unique node identifiers, valid edge references, finite generated values, nonconstant columns, compatible observation-only and mixed-interventional files, target changes under intervention, descendant recomputation, and agreement among adjacency matrices, generated columns, masks, graph metadata, and the executable specification. Continuous variables are revised when more than 10% of values lie exactly on a valid-range boundary without a declared boundary mechanism. Continuous or non-discrete near-copy pairs are revised when their absolute correlation reaches 0.90 without scientific justification. Binary, categorical, ordinal, and grid-valued variables are revised when one category dominates without a rare-event, pass-heavy, or threshold-based explanation. Formula SCMs must also pass residual checks for non-intervened formula rows within numerical tolerance. The final audit rejects duplicate support topologies, repeated coarse role patterns, slot-filled mechanism text, generic readout/flag/calibration padding, and excessive dominance of one root, mechanism, or noise family. Only scenarios that pass review, executable validation, audit, and the final manifest gate enter the benchmark package.

B.5 Semantic SCM Pool

Semantic scenario cards target operational DAGs with observable variables, actionable intervention handles, and scenario-specific measurement channels. The released semantic pool contains 100 operational scenarios, with 10 scenarios in each of 10 domains: cybersecurity and IT operations, education, finance and credit, government services, healthcare delivery, housing and real estate, manufacturing, public health, urban transportation, and water and sanitation. The realized release contains 17–25 variables (mean 22.89), 25–67 directed edges (mean 40.28), and 1–12 intervention targets (mean 3.82). Core metadata fields include variable definitions, units/ranges, observability, intervenability, parent lists, sampling policies, edge rationales, mechanism descriptions, monotonicity/sign tags, intervention resamplers, raw-value summaries, graph artifact paths, and generator-spec artifact paths. For example, the school-start-time scenario records that delaying the bell changes planned pickup schedules and circadian alignment, while grace-window policy affects whether an arrival is counted as tardy.

Figure 5 reports the released Semantic mix by mechanism and by noise. The largest mechanism shares are event/count processes (15.3%), thresholds (8.5%), stock-flow accumulation (8.3%), and ordinal rules (8.1%). On the noise side, 43.2% of nodes have no extra noise; the rest are led by normal (26.2%), then lognormal, Bernoulli, and count-style families.

Example. Figure 6 (top) shows the public-health scenario *Wildfire smoke and shelter behavior*: 20 variables and 26 edges. The SCM represents one household on a wildfire-smoke day. It separates three kinds of causal structure that are often collapsed in anonymous synthetic graphs. The first is an outdoor-warning chain: local outdoor PM2.5 is the true air condition near the house, it drives an official monitor reading, the monitor can trigger a shelter-in-place alert, and the alert changes perceived smoke risk. The second is a behavior and mitigation chain: perceived risk affects time spent outdoors and protective behavior, while a portable-filter intervention changes clean-air delivery inside the home. The third is a measurement chain: true indoor PM2.5 is not the same object as the consumer-sensor reading or the exceedance flag derived from that reading.

This example illustrates why semantic grounding makes graph errors more diagnostic. Missing the edge from portable filter to true indoor PM2.5 removes a real intervention handle; adding an edge from the consumer-sensor reading back to true indoor PM2.5 confuses measurement with the underlying air state; dropping the path from exposure dose to respiratory symptoms loses the health mechanism that the scenario is meant to evaluate. The executable card stores these variable meanings, intervention handles, measurement notes, and edge rationales together with the sampled tables, so graph recovery errors can be read as concrete claims about the smoke-response process.

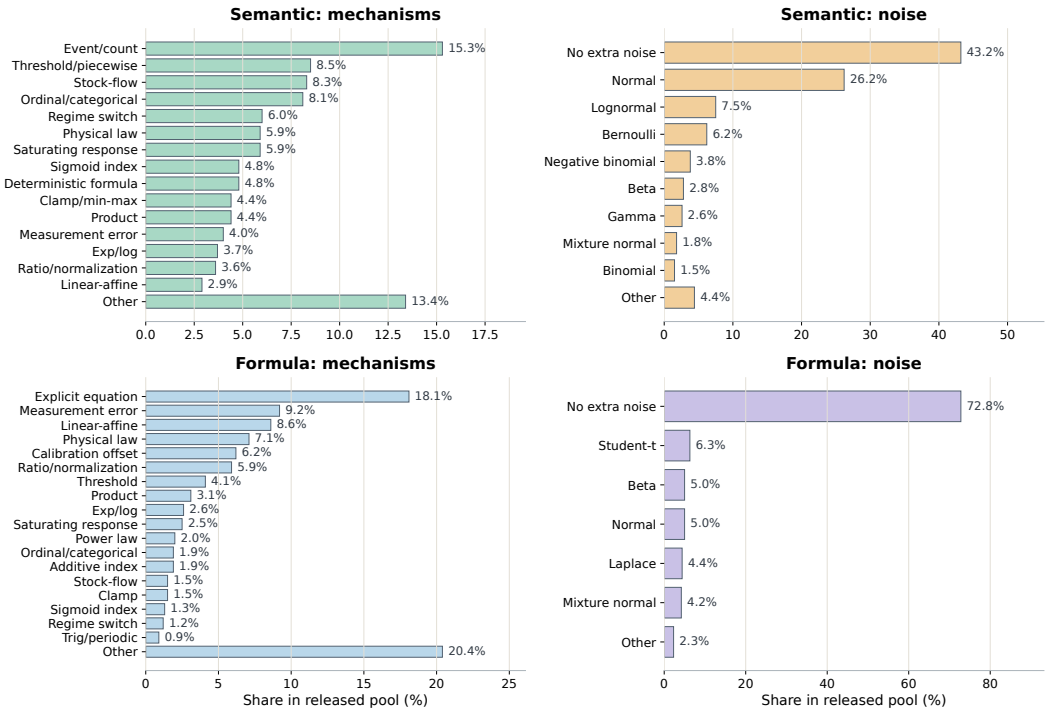


Figure 5: **Composition of the released Semantic and Formula pools.** Mechanism bars are edge shares; noise bars are node shares, from one representative file per released SCM. Remaining unlabeled families are grouped as Other. Formula nodes themselves have no extra noise, so that bar is large.

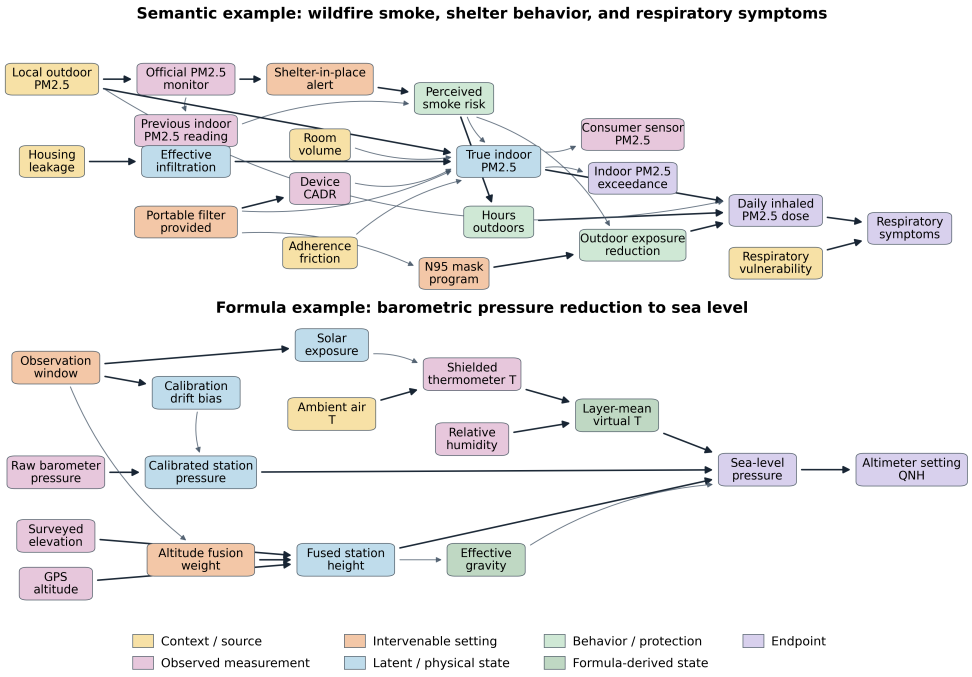


Figure 6: **Example Semantic and Formula SCMs (full released cards).** Node colors follow the legend, and arrows indicate directed causal edges. Top: wildfire smoke, shelter behavior, and respiratory symptoms (public health; 20 variables, 26 edges). Bottom: barometric pressure reduction to sea level (earth systems; 16 variables, 18 edges).

B.6 FORMULA-GROUNDED SCM POOL

Formula scenario cards target scientific DAGs whose key dependencies are grounded in explicit equations, causal-orientation contracts, units, validity ranges, and measurement/readout semantics. The released formula pool contains 100 equation-grounded scenarios, with 10 scenarios in each of 10 domains: biology and ecology, astronomy, chemistry, mechanics and fluids, earth systems, electromagnetism, energy systems, materials and structures, optics and waves, and thermodynamics. The realized release contains 16–25 variables (mean 22.68), 18–50 directed edges (mean 32.97), 1–17 intervention targets (mean 5.78), and explicit equation/residual metadata for formula-derived variables. Core metadata fields include variable definitions, units/ranges, constants, formula identifiers, equation inputs/outputs, solve-for variables, causal-orientation contracts, invalid reverse-orientation notes, measurement readouts, intervention resamplers, raw summaries, and residual checks on non-intervened rows. For example, the thin-lens scenario records the lens equation for target image distance, the Airy-disk diffraction relation, and the magnification equation, then adds actuator backlash, exposure, saturation, and quality-control readouts downstream.

Figure 5 also shows the Formula mix. The largest mechanism shares are explicit equations (18.1%), measurement-error edges (9.2%), linear-affine relations (8.6%), physical laws (7.1%), and calibration offsets (6.2%). Formula nodes are computed exactly from their parents, so they carry no extra noise; that is why "No extra noise" is 72.8% of nodes. The remaining noisy nodes are mostly measurement and readout terms: Student-t, beta, normal, Laplace, or mixture-normal. Both families use do-style intervention records with descendant recomputation. The difference is what the edges write down: a running process in Semantic, a named equation plus instruments in Formula.

Example. Figure 6 (bottom) shows the earth-systems scenario *Barometric pressure profile*: 16 variables and 18 edges. The SCM is built around a named physical calculation: reducing station pressure to sea-level pressure. Three upstream chains feed this calculation. A pressure chain turns a raw barometer reading and calibration drift into calibrated station pressure. A height chain fuses surveyed elevation and GPS altitude, with a settable fusion weight, and then determines effective gravity. A temperature chain uses the observation window to model solar exposure, converts ambient air temperature into a shielded thermometer reading, and combines that reading with relative humidity to obtain layer-mean virtual temperature.

These four quantities then enter the hypsometric reduction

$$\mathrm{SLP} = p_{\mathrm{stn}} \exp\left(\frac{g_{\mathrm{eff}}\, z}{R_d\, T_v}\right),$$

where p_{stn} is calibrated station pressure, z is fused height, g_{eff} is effective gravity, T_v is layer-mean virtual temperature, and R_d is the dry-air gas constant. The downstream QNH value is the sea-level pressure converted to hPa and rounded. This card therefore tests whether a method can recover both equation parents and instrument parents: omitting fused height or virtual temperature breaks the physical formula, while confusing raw and calibrated pressure collapses an instrument-correction step into the physical law. Interventions on the fusion weight or observation window recompute all downstream formula nodes, and residual checks ensure that non-intervened rows remain consistent with the stated equation.

C DETAILED EVALUATION PROTOCOL

This section records the evaluation conventions needed to interpret the reported results. File schemas, wrapper I/O details, and runnable commands are released with the code; here we summarize the protocol choices affecting comparability.

C.1 EVALUATION SETTINGS

Each executable SCM is evaluated under two exported splits from the same underlying graph. The *obs-only* split contains only unperturbed samples and uses $n_{\mathrm{obs}} = 1000$ in the main leaderboard. The *obs+int* split contains observational rows plus do-style interventional rows with descendant recomputation and uses $n_{\mathrm{obs}} = 800, n_{\mathrm{int}} = 200$ in the main leaderboard. Intervention targets are

recorded in the exported masks. Methods that do not consume intervention indicators are evaluated only on supported splits and are marked missing otherwise.

Each generated SCM benchmark scenario or synthetic configuration is evaluated with 5 random replicates, a fixed compromise between wall-clock cost and stable aggregate estimates. For a fixed method and split, we run the method on each replicate, score against the shared ground-truth adjacency, average replicate scores to a scenario/configuration-level value, and then aggregate over families, domains, dimensions, or the full arena. Failed, timed-out, or invalid runs are recorded explicitly and never silently removed.

C.2 METRICS

Given a predicted directed graph $\hat{G}$ and ground-truth adjacency G, we report directed-edge precision, recall, F1, and structural Hamming distance (SHD). Higher F1 and lower SHD are better. Compact tables report F1/SHD as mean(standard deviation); full F1 and SHD tables appear in Section G. For PC and GIES, remaining undirected CPDAG edges are randomly oriented with a fixed per-file seed before these DAG metrics are computed, so the scores are directed-graph scores, not equivalence-class scores. Appendix complete tables additionally report structural intervention distance (SID), the number of ordered pairs whose intervention distributions would be misidentified from $\hat{G}$; normalized SHD $\mathrm{nSHD} = \mathrm{SHD}/(d(d-1))$; and, when a method returns edge scores, AUROC and average precision (AP). Hard-graph methods without edge scores therefore have missing AUROC/AP entries. Runtime and failure status are reported separately from accuracy metrics.

C.3 RUNTIME, HARDWARE, AND TIMEOUT POLICY

GPU-based wrappers were evaluated on NVIDIA RTX 4090 GPUs when GPU inference was required. CPU baselines ran on AMD EPYC 7763 CPU cores with bounded process-level parallel workers. The current protocol uses a one-hour wall-time budget per graph row. Runs are executed in shards, with five replicate rows per shard, so the scheduler assigns a shard-level budget proportional to the number of rows plus a small startup guard. The main-text runtime ranking uses two protocols over the same synthetic mix: mean end-to-end wrapper time on the synthetic main benchmark for non-pretrained methods, and nested-k load/inference fits on a sample balanced across dimensions and graph families for pretrained methods.

Reduced configurations are used only when the full configuration has been verified to be impractical under the benchmark budget. In the real-data obs-int-as-obs conversion, PC timed out on four larger datasets (Wind Tunnel, Light Tunnel, PetShop temporal traffic 1, and PetShop temporal traffic 2) after a long best-effort budget; the reduced repair uses $\alpha = 0.001$, at most 200 sampled rows, constant-column dropping, and fixed-seed random orientation, with predictions mapped back to the original variable dimension. CDIS timed out or failed to complete on the same four datasets; its reduced repair caps the conditioning-set size at 1 and preserves the native PAG output. SEA required a preprocessing repair on the two Causal Chamber datasets because constant columns produced a smaller returned matrix; the repaired wrapper drops constant columns internally and maps predictions back to the original dimension.

C.4 METHOD SETTINGS

Unless noted below, methods are run once for the leaderboard with wrapper defaults and no hyperparameter search on benchmark labels. Library defaults not listed below are left unchanged by the wrapper. For pretrained or amortized methods, the checkpoint/source named below is fixed across generated-SCM and real-data runs. Methods that output equivalence-class objects preserve the native output when available, and directed-edge F1/SHD are computed after the deterministic conversion rules below.

- **RandomRegress.** A fixed-seed random-order regression baseline. It outputs a directed graph from nonzero regression support.
- **CDIS.** Uses observational and intervention groups under the wrapper default conditioning search. For directed DAG metrics, PAG endpoint uncertainty is converted to a directed graph using the

wrapper's liberal orientation rule. The reduced repair caps conditioning-set size at 1 only for cases verified to exceed the budget.

- **GIES.** Uses the interventional score with the BIC-style penalty left at the wrapper default. Undirected edges in the returned essential-graph/CPDAG-like object are randomly oriented with a fixed per-file seed for DAG metrics.
- **IGSP.** Uses Gaussian CI-style tests with $\alpha = \alpha_{\text{inv}} = 10^{-3}$, depth 4, and 5 runs by default. It reports a directed graph directly.
- **PC.** Uses stable PC with Fisher-z CI tests and $\alpha = 0.05$ by default. Undirected CPDAG edges are randomly oriented with a fixed per-file seed for DAG metrics. The reduced real-data repair uses $\alpha = 0.001$ and at most 200 rows after the full configuration times out.
- **DAS.** Uses the additive spline wrapper with pruning parameters $\eta_G = \eta_H = 0.001$, $\alpha = 0.05$, 10 splines, degree 3, and parent range 5–20.
- **LiNGAM.** Uses the standard DirectLiNGAM wrapper and reports the returned directed graph.
- **DAGMA.** Uses the wrapper's automatic linear/nonlinear selection with default optimization settings and default weight threshold 0.3.
- **NOTEARS.** Uses linear NOTEARS with $\lambda_1 = 0.1$ and fixed adjacency threshold 0.3.
- **NOTEARS-MLP.** Uses hidden size 10, $\lambda_1 = \lambda_2 = 0.01$, max iteration 100, and fixed adjacency threshold 0.3.
- **SDCD.** Uses the two-stage wrapper with the default stage-1/stage-2 schedule and model threshold 0.1; high-dimensional budget reductions are logged when invoked.
- **Arrow.** Uses the released ArrowFM base checkpoint `arrowfm-base.pt` and its official decoded binary output.
- **AVICI.** Uses the released AVICI `scm-v0` pretrained model (checkpoint `checkpoint_0300000.pkl`), and fixed threshold 0.5 for probabilistic adjacency outputs.
- **CauScale.** Uses the released checkpoint `synthetic/auprc=0.905_migrated.ckpt` and its official output logic.
- **CDFM.** Uses the released CDFM checkpoint `model.safetensors` and its official output logic.
- **FoundCause.** Uses the released checkpoint `checkpoint.pt`. The official CLI defaults to an adaptive GMM threshold, not a fixed 0.5 threshold. We scored both the default binary decision and a fixed 0.5 threshold on the saved edge probabilities, and report the latter because it gave higher average F1 and lower average SHD on most experiments. This is a documented wrapper rule, not per-dataset tuning.
- **SEA.** Uses the released obs-only FCI checkpoint `fci_synthetic/model_best_epoch=373_auprc=0.842.ckpt`, with 500 inference batches and FCI batch size 500. We do not run SEA as an interventional method: its interventional recipe uses GIES on sampled batches and is reported to need at least 250 observations per batch (Wu et al., 2025), while the main obs+int split has only $n_{\text{int}} = 200$ interventional rows in total, further split across several targets.
- **TabCausal.** Uses the released checkpoint `tabcausal-v2.pt` and its official output logic.

D Additional Experiments

D.1 Sample Size and Intervention Protocol

These experiments keep the synthetic SCM collection fixed and change sample size or intervention design. The sample-size suite fixes $d = 30$, uses 100 synthetic SCMs, and varies the observation-only sample size $n \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$. Figure 7a plots F1 against n for all evaluated methods. The corresponding F1/SHD table is in Section G.5.

Most of the F1 gain occurs between $n = 100$ and $n = 1{,}000$; beyond $n = 2{,}000$ the curves flatten. FoundCause remains highest throughout (0.43 at $n = 100$, 0.64 at $n = 1{,}000$, 0.66 at $n = 10{,}000$), and TabCausal has the next largest rise (0.33 to 0.57). These absolute levels should be read with the suite's fixed $d = 30$ SCM mix in mind: the curves isolate how F1 changes with n, while a method's height can also reflect affinity to that held-fixed dimension. Search and optimization methods such

as GIES, IGSP, CDIS, SDCD, and NOTEARS-MLP improve through $n = 1{,}000$ and then stall. NOTEARS, DAGMA, LiNGAM, and RandomRegress barely move. Arrow and CauScale are not monotone: both peak near $n = 1{,}000$–$2{,}000$ and then drop at $n = 10{,}000$ (Arrow $0.36 \to 0.33$, CauScale $0.41 \to 0.35$), so extra observational samples do not uniformly help pretrained wrappers.

The intervention-protocol suite also fixes $d = 30$ with 100 SCMs and compares the default resampling intervention with fixed-setpoint, parameter-shift, high-intervention-fraction, multi-target-per-sample, and dense mixed-intervention settings. Figure 7b plots the F1 change relative to Default; complete protocol-wise metrics are in Section G.6.

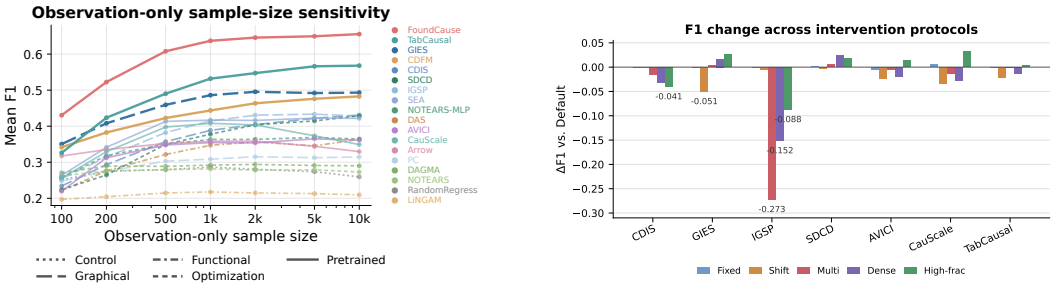


Figure 7: **Sample-size and intervention-protocol sensitivity.** (a) Mean F1 on the synthetic d=30 observation-only suite; line styles encode method class, and the right column lists full method names ordered by final F1. (b) F1 change relative to the default mixed protocol for each method that supports the intervention-protocol suite; negative values are drops.

Figure 7b shows that TabCausal and SDCD vary least across these protocols, while IGSP is most sensitive. Fixed-setpoint stays near Default for every method; the bars that stand out are IGSP under multi-target, dense mixed, and high-intervention-fraction interventions (ΔF1 -0.27, -0.15, and -0.09). The largest drops under multi-target and dense mixed interventions are plausible because these settings make the intervention regimes less cleanly attributable to one target. More generally, IGSP relies on conditional-independence and invariance tests and needs enough samples in both the observational distribution and each interventional regime, so increasing the intervention fraction can still hurt when it reduces observational rows and fragments interventional samples across many targets. GIES, which uses the same interventional tables, rises under dense mixed and high-intervention-fraction (ΔF1 $+0.02$ and $+0.03$) and drops mainly under parameter-shift. Complete protocol-wise metrics are in Section G.6.

D.2 Generic Tabular Foundation Models as Skeleton Probes

Motivated by recent use of tabular-foundation-model attention as a proxy for causal-skeleton recovery, including the attention-based diagnostic in LimiX-2 (Zhang et al., 2026a), we additionally test whether generic tabular foundation models expose usable causal-skeleton signal through feature-attention scores. For each Semantic and Formula observation-only instance, we run LimiX-2, TabICLv2 (Qu et al., 2026), and TabPFN-3 (Grinsztajn et al., 2026) by treating each variable as a target and extracting the last available feature-facing attention: LimiX-2 target-token to feature-group attention, TabICLv2 final row-interaction CLS-to-feature attention averaged over held-out rows, and TabPFN-3 final ColumnAggregator CLS-to-feature-group attention. The resulting directed scores are symmetrized and evaluated only as undirected skeletons. This protocol is therefore not comparable to the main directed-DAG leaderboard, but it probes whether supervised tabular foundation models contain recoverable causal-neighborhood information.

The curves show strong threshold sensitivity. Unlike causal-discovery methods that output edge probabilities or hard graphs, these attention scores are not calibrated edge probabilities. When the number of features is moderate or large, attention mass is spread across many candidate variables, so a fixed threshold such as 0.5 can yield nearly empty graphs. Conversely, very low thresholds put all three probes into a near-complete-graph regime: recall is almost one, predicted skeletons contain almost all variable pairs, and F1 collapses to a density-driven baseline. The informative region is therefore model- and benchmark-dependent. Even at their best fixed thresholds, these probes remain below most causal-discovery methods and do not produce edge directions. TabICLv2 is weaker

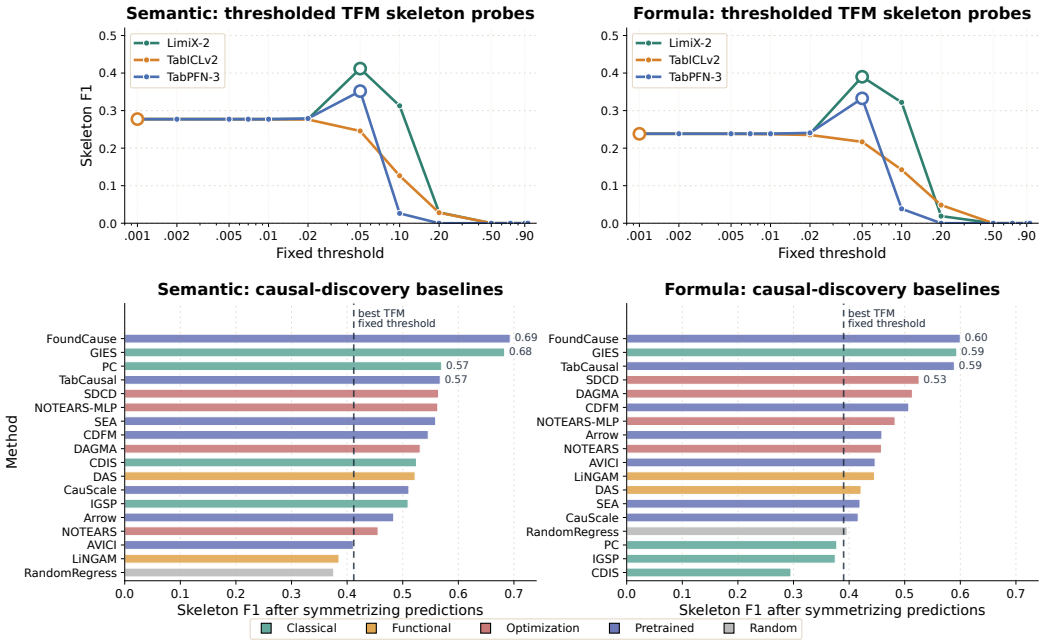


Figure 8: **Attention-based skeleton probes for generic tabular foundation models.** Top: skeleton F1 from thresholding final-layer feature or grouped-feature attention scores on Semantic and Formula observation-only benchmarks. Open circles mark each model's best fixed threshold in the plotted grid. Bottom: causal-discovery baselines evaluated under the same skeleton-F1 metric after symmetrizing their predictions; the dashed line marks the best fixed-threshold TFM score in the corresponding benchmark.

than the other two probes in this diagnostic; this may partly reflect its row-first design, where the extractable feature-facing signal comes from the final row-interaction block but still precedes the later in-context prediction blocks.

Thus, these generic tabular foundation models can be applied to this task, but their apparent skeleton-recovery ability depends heavily on the threshold used to convert attention scores into edges. Different fixed thresholds lead to very different graphs and scores, and the attention-based outputs only support undirected skeleton recovery rather than directed causal discovery. We therefore do not include them in the main directed-graph comparison.

D.3 Ablation Studies

Construction pipeline. Figure 9 compares successive stages of the agentic construction pipeline using an LLM-judge quality rubric on 20 sampled scenarios, 10 from Semantic and 10 from Formula. The full construction process is described in Section B. The reference pack adds external factual context; planning organizes variables and causal chains before graph drafting; graph review removes aliases, unsupported shortcuts, and low-information support nodes. In Figure 9 the average rises from 3.80 (seed only) to 3.87 after the reference pack, 4.37 after planning, and 4.48 after graph review. The reference pack mainly lifts semantic/mechanism diversity; planning is the largest step, raising causal plausibility, mechanism grounding, benchmark validity, and specifiability together; graph review adds a smaller further gain on the same criteria. Structural diversity is already high at the seed stage and stays flat. The score pattern supports keeping these stages.

D.4 Runtime and Resource

Pretrained and amortized wrappers pay a one-time load before they score a graph. On a short evaluation shard that load can dwarf the inference itself. We therefore time nested runs that score k graphs after one load, using a synthetic sample balanced across dimensions and graph families, and fit

$$T(k) = T_{\text{fixed}} + kT_{\text{graph}},$$

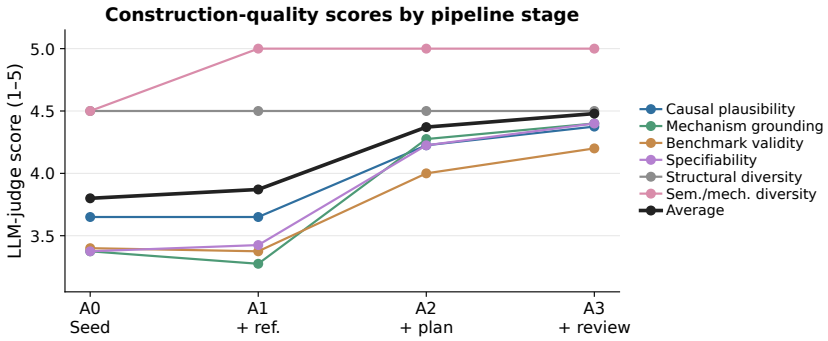


Figure 9: **Construction-pipeline ablation.** LLM-judge stage curves over 20 scenarios (10 Semantic, 10 Formula); the thicker line is the reported average. Criterion-level means with standard deviations are in Table 19.

where T_{fixed} is the load and T_{graph} is the extra seconds for one more graph once the model is already in memory. The nested k grid is larger at low d, where each extra graph is cheap and a short run would make the load/inference split unstable, and smaller at high d, where each extra graph is expensive: $k \in \{100, 200, 500, 1000\}$ at $d = 10$, $k \in \{50, 100, 200, 500\}$ at $d = 20, 30$, and $k \in \{20, 50, 100, 200\}$ at $d = 50, 100$.

Table 3: **Load and per-graph time for pretrained/amortized methods.** Fits use a synthetic sample balanced across dimensions $d \in \{10, 20, 30, 50, 100\}$ and graph families. T_{fixed} is the fitted one-time load, averaged over d. T_{graph} is the extra seconds for one more graph after the model is loaded, reported by d and as an unweighted mean over those d. Nested k is $\{100, 200, 500, 1000\}$ at d=10, $\{50, 100, 200, 500\}$ at d=20, 30, and $\{20, 50, 100, 200\}$ at d=50, 100: larger k at low d for a stable split, smaller k at high d because extra graphs cost more. Lower time is better. Bold is lowest in a time column and underline is second; R^2 columns are not ranked. All method–d fits have $R^2 > 0.97$, and all but one exceed 0.98.

		Per-graph seconds T_{graph}							
Method	T_{fixed}	d=10	d=20	d=30	d=50	d=100	**Mean**	**Mean** R^2	**Min** R^2
Arrow	14.47	**0.023**	**0.040**	**0.065**	**0.081**	**0.176**	**0.077**	0.9905	0.9837
AVICI	55.84	0.031	0.052	0.081	0.104	0.215	0.096	0.9933	0.9897
CDFM	10.02	0.068	0.098	0.108	0.249	0.472	0.199	0.9947	0.9922
TabCausal	**8.13**	0.140	0.171	0.170	0.224	0.346	0.210	0.9993	0.9979
CauScale	11.07	0.140	0.145	0.190	0.398	1.170	0.409	0.9902	0.9718
SEA	28.14	0.232	0.278	0.310	0.341	0.993	0.431	0.9951	0.9898
FoundCause	23.46	1.146	1.861	2.310	5.729	12.725	4.754	0.9995	0.9984

Table 3 and Figure 10 split that cost. All method–d fits have $R^2 > 0.97$, and all but CauScale at $d = 50$ exceed 0.98. After load, Arrow and AVICI are the cheapest extra graphs, CDFM and TabCausal are next, CauScale and SEA are slower, and FoundCause sits in a higher band. TabCausal also has the shortest mean load (8.13s); AVICI has the longest (55.84s). The later comparison still drops T_{fixed}. A pretrained wrapper is usually loaded once and then applied to many graphs—a leaderboard pool, a dimension or family shard, or a batch of related systems—so the one-time cost is T_{fixed}/k per graph and becomes small once k is large. The comparable quantity that remains is therefore T_{graph}.

Table 4 reports one comparable per-graph number for every method. At each d, the nested-k sample draws equally many graphs from each graph family, matching the synthetic-main mix, so the ranking is comparable; T_{fixed} is omitted because a large-k run amortizes the load.

Figure 11 uses the same comparable per-graph times. For pretrained methods the x-coordinate is T_{graph} with T_{fixed} omitted, because a large-scale causal-discovery run amortizes the load. The useful frontier is not simply the set of fastest methods. Arrow and AVICI are very cheap but have lower average F1; CDFM and TabCausal move the frontier upward at nearly the same cost scale; GIES remains the strongest low-cost graphical-search baseline; and FoundCause reaches the highest average F1 among its supported observation-only cells at a larger T_{graph}. Several classical or optimization methods are dominated in this view: they are slower than the frontier methods while also having

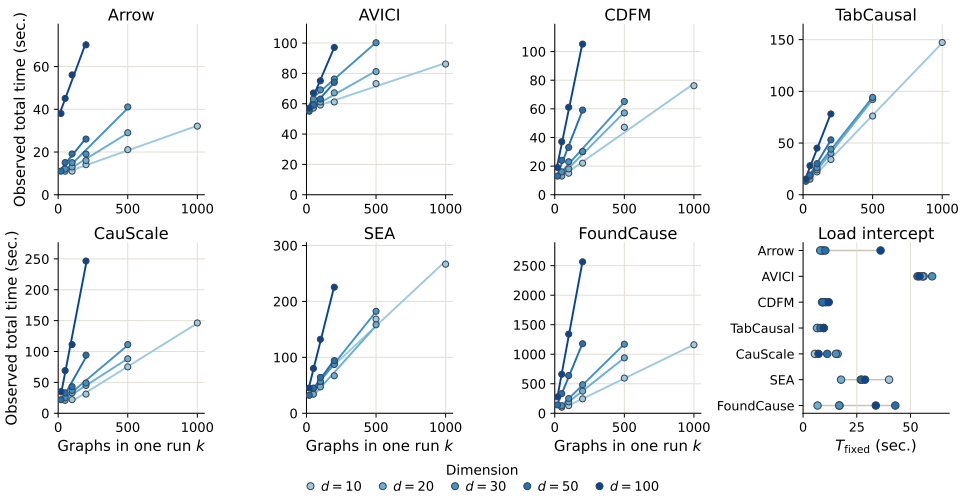


Figure 10: **Pretrained runtime fit quality.** Each line panel is one pretrained/amortized wrapper on its own vertical scale. Points are observed total wrapper time on a synthetic sample balanced across dimensions and graph families; a line is the fitted $T(k) = T_{\text{fixed}} + kT_{\text{graph}}$ within dimension d. Dots on the line mean the split is stable; a steeper line, on that panel's scale, costs more per extra graph. The last panel shows the fitted load T_{fixed}.

Table 4: **Per-graph runtime on the synthetic mix.** Non-pretrained entries are mean end-to-end wrapper time on the synthetic main benchmark, mean(standard deviation). † marks the unweighted mean of per-d T_{graph} from nested-k fits. At each d, the sample draws equally many graphs from each graph family, matching the synthetic-main mix, so the times are comparable; T_{fixed} is omitted because it amortizes at large k. Lower is better.

Rank	Method	Sec./graph	Rank	Method	Sec./graph	Rank	Method	Sec./graph
1	Arrow	0.077†	7	RandomRegress	0.82 (0.53)	13	NOTEARS	30.46 (57.50)
2	AVICI	0.096†	8	GIES	2.78 (1.19)	14	LiNGAM	36.17 (59.43)
3	CDFM	0.199†	9	FoundCause	4.754†	15	DAS	57.60 (62.02)
4	TabCausal	0.210†	10	DAGMA	5.33 (3.01)	16	CDIS	75.75 (253.70)
5	CauScale	0.409†	11	PC	20.36 (146.76)	17	NOTEARS-MLP	80.27 (99.35)
6	SEA	0.431†	12	IGSP	26.30 (131.27)	18	SDCD	192.64 (115.28)

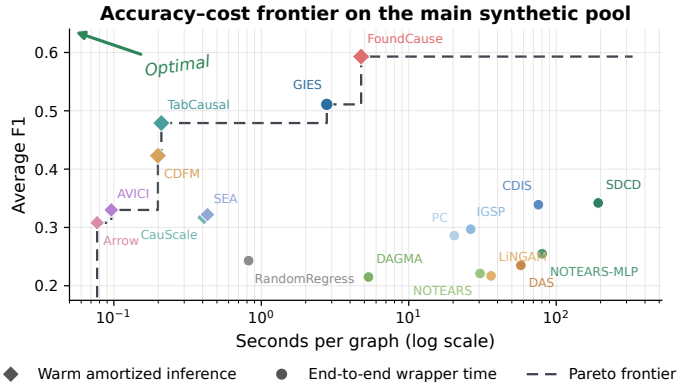


Figure 11: **Accuracy–cost frontier.** Average F1 from Table 2 against the comparable per-graph times in Table 4. Diamonds use pretrained T_{graph} from a sample balanced across synthetic dimensions and graph families, omitting T_{fixed}; circles use mean end-to-end time on the synthetic main benchmark. Average F1 follows the table Avg. column, which averages only the family–split cells available for each method. The dashed line is the Pareto frontier; *Optimal* points toward higher F1 and lower time.

lower average F1. Pretrained methods were not trained under the same documented SCM priors, so some scores may partly reflect overlap between a method's reported pretraining distribution and the evaluated slice. A later observation-only check in Section E.2 splits categories by each method's own audited prior and finds only modest ID-to-OOD shrinks in relative lead for that reference group; the coarse order is stable in a coverage-mix reading, with the limits of that diagnostic stated there.

E ANALYSIS

A released suite still needs to be checked for breadth and for whether scores can be traced to specific graph, mechanism, or domain properties. Following the diagnosability requirement in Section 2.2, this section reports synthetic factor analyses, a check of pretrained methods on categories outside their documented training support, and Semantic/Formula domain breakdowns.

E.1 SYNTHETIC BENCHMARK BEHAVIOR ANALYSIS

Interaction score. Aggregate leaderboards rank methods, but they do not say why the ranks differ. We analyze the synthetic obs-only benchmark with factor-specific interaction scores. For a method m and factor category c, let $F_{m,c}$ be the mean F1 on that category. We report

$$I_{m,c} = F_{m,c} - \bar{F}_{m,\cdot} - \bar{F}_{\cdot,c} + \bar{F}_{\cdot,\cdot}. \tag{7}$$

This double-centering removes each method's overall strength and each category's overall difficulty. A positive value means that a method is stronger on that category than expected from both marginal averages. For graph-level factors, categories are assigned per graph. For edge-level factors, mechanisms and noises are assigned to an edge by the child node's structural equation.

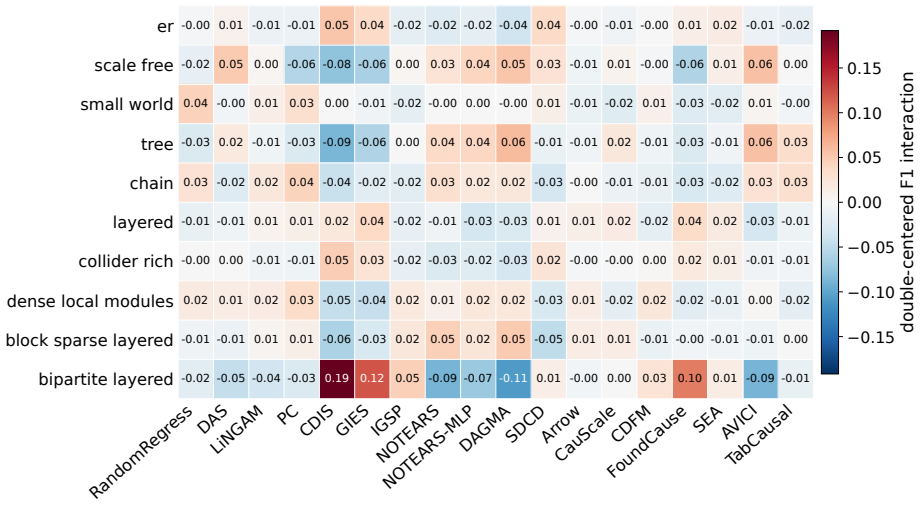


Figure 12: **Synthetic graph-family factor interactions.** Each cell is a double-centered F1 interaction on the observation-only synthetic benchmark.

Graph structure. Figure 12 shows that methods differ by graph topology even after category difficulty is removed. The clearest contrast is the bipartite-layered family: CDIS ($+0.192$), GIES ($+0.120$), and FoundCause ($+0.098$) are relatively stronger, while DAGMA (-0.108), NOTEARS (-0.089), and AVICI (-0.089) are relatively weaker. For CDIS/GIES, this is consistent with graphical-search behavior: bipartite and collider-rich structures create many separation and v-structure signals that conditional-independence or equivalence-class search can exploit. FoundCause is pretrained, but its pairwise-statistics pathway and triangular refinement module are also designed to reason over higher-order motifs such as chains and colliders, which plausibly explains why it benefits from the same structured signals. By contrast, continuous-optimization methods and AVICI are more favorable on sparse branching families such as trees or scale-free graphs, where recovering a weighted acyclic parent structure is closer to their optimization or amortized parent-selection bias.

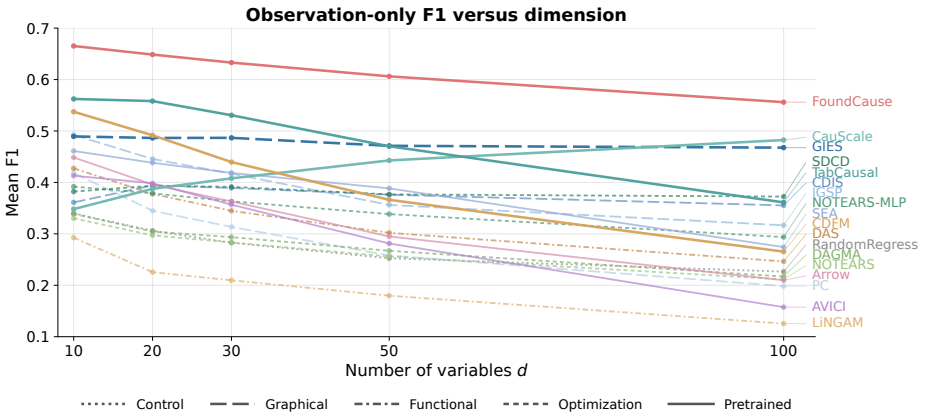


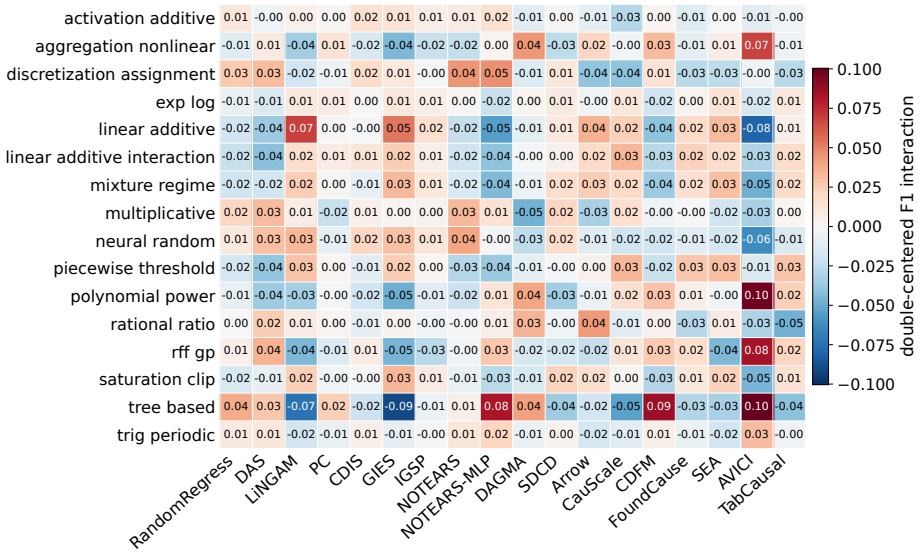
Figure 13: **Synthetic dimension scaling.** Mean observation-only F1 versus the number of variables d. Line styles follow method families, as in the sample-size plot. CauScale is the only method whose F1 rises with d.

Scale effects. Dimension is ordered, so Figure 13 reports it as curves. Almost every method loses F1 as d grows, but the slopes differ. FoundCause stays highest throughout (0.67 at $d = 10$, 0.56 at $d = 100$) while declining. CauScale is the exception that rises (0.35 to 0.48), matching its design goal and training setup for large graphs: its synthetic training includes graph sizes up to hundreds of nodes, and its reduction/tied-attention design is optimized for scaling. GIES, CDIS, and SDCD are comparatively flat. Arrow, CDFM, and AVICI drop sharply by $d = 100$ (to 0.21, 0.27, and 0.16); this likely reflects dense pairwise decoding, skeleton/order or edge-head uncertainty accumulating over many variable pairs, and training distributions whose strongest support is not centered on our high-dimensional heterogeneous synthetic pool. TabCausal also falls (0.56 to 0.36) but remains above those collapsing pretrained methods. Both TabCausal and FoundCause combine broad synthetic task variation with explicit pairwise or motif-level signals, which helps preserve relative performance as the variable set grows.

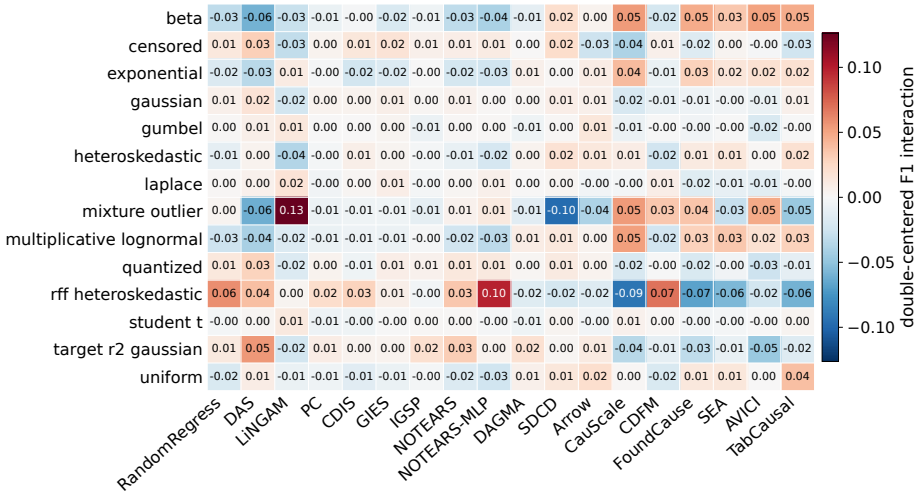
Mechanism and noise effects. Figure 14a shows a clear split between linear-identifiability assumptions and nonlinear mechanism coverage. On the linear-additive row, LiNGAM is the strongest positive method ($+0.069$), which is expected because this setting matches its linear non-Gaussian identification principle. NOTEARS is nominally a linear-SEM method, but its squared-loss objective is closer to a linear-Gaussian score and does not exploit non-Gaussian residual information for orientation; since most linear-additive cases in this pool are not Gaussian-noise linear SEMs, NOTEARS is not helped in the same way (-0.023). GIES follows LiNGAM closely ($+0.053$ on linear-additive, and a row-wise mechanism correlation of about 0.95 with LiNGAM), even though the algorithms are different. The common pattern is that both are most comfortable when causal effects are globally simple and separable, and both lose relative advantage when the edge signal is local, nonlinear, or non-smooth.

The tree-based row creates the strongest mechanism split. AVICI ($+0.100$), CDFM ($+0.086$), and NOTEARS-MLP ($+0.082$) are strongly positive, while GIES (-0.091), LiNGAM (-0.074), CauScale (-0.054), and TabCausal (-0.041) are negative after double-centering. This row corresponds to piecewise and threshold-like parent effects: the relevant parent set can be recoverable when a model has enough nonlinear function capacity or has seen similar nonlinear signatures during pretraining, but the same structure is hard for linear/non-Gaussian assumptions or score-based search to express as a stable global dependence pattern. The negative values for some pretrained models mark relative specialization: their overall accuracy may remain strong, but their advantage is less concentrated on this particular piecewise mechanism class. AVICI is especially mechanism-sensitive in this panel: its interaction variance across mechanism rows is the largest among the pretrained methods, with large positive values on tree-based, polynomial-power ($+0.097$), and RFF-GP ($+0.083$) rows but a large negative value on linear-additive (-0.081). We therefore read AVICI as less stable across mechanism families than broader or more motif-oriented pretrained methods such as TabCausal and FoundCause, without attributing this pattern to a single mechanism property.

The noise panel separates whether scores move more with the noise family or with the mechanism fam-



(a) Mechanism interactions (child-node mechanism family).



(b) Noise interactions (child-node noise family).

Figure 14: **Synthetic mechanism and noise interactions.** Each cell is a double-centered F1 interaction on the observation-only synthetic benchmark.

ily. Overall, noise families create weaker method separation than child mechanisms: the average row spread is smaller for noise than for mechanisms, although mixture-outlier and RFF-heteroskedastic noise are strong exceptions. Graphical-search methods are comparatively flat across noise rows, whereas pretrained methods show more visible noise-dependent swings than graphical search and most continuous-optimization methods. This sensitivity is less distinctive in the graph-family and mechanism panels, where variation is shared across several method classes. Noise families therefore show how each method uses residual shape, scale variation, and support constraints.

The largest contrasts come from noises that change more than marginal variance. Mixture-outlier noise injects sparse contamination, so methods that rely on smooth losses, ordering heuristics, or stable low-order associations can be pulled toward spurious supports; this is consistent with negative interactions for SDCD (-0.097) and DAS (-0.061). LiNGAM is the main exception ($+0.126$): in this slice, outliers amplify non-Gaussian residual cues, which its identification principle can sometimes exploit. RFF-heteroskedastic noise produces a different split because the noise scale changes with parent values, so the edge evidence is a conditional-variance signal. NOTEARS-MLP

is strongly positive ($+0.097$), plausibly because its nonlinear regression model can absorb associated nonlinear parent-dependent patterns, while CDFM is also positive ($+0.069$), consistent with explicit heteroskedastic coverage in its pretraining prior. Several pretrained methods are negative on this row: even a broad pretraining prior can remain sensitive to variance-driven evidence. RandomRegress also has visible tendencies across noise rows, for example on RFF-heteroskedastic noise ($+0.059$); because it is a fixed-random-order regression control, these tendencies reflect sensitivity of nonzero regression support to scale/noise artifacts.

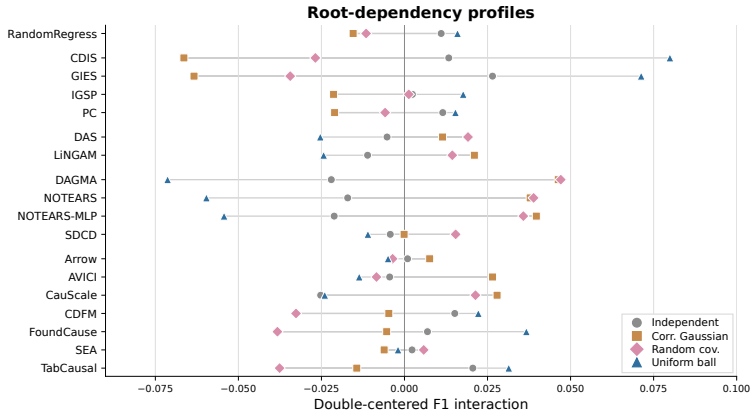


Figure 15: **Synthetic root-dependency interactions.** Each row is one method; markers are the four root-dependency categories, using double-centered F1. Wider rows are more root-sensitive. The uniform-ball versus common-factor (correlated-normal) reversal is the left–right swap of triangles and squares.

Root dependency effects. Figure 15 isolates graph-level dependence among root variables. Independent-root markers sit near zero for most methods, as expected for the setting closest to the standard SCM assumption. Search methods such as CDIS and GIES have wide rows: uniform-ball markers (triangles) lie on the positive side and common-factor correlated-normal markers (squares) on the negative side. NOTEARS, NOTEARS-MLP, and DAGMA reverse that order. A plausible explanation is that smooth root covariance violates exogenous independence and can be mistaken by graphical-search procedures as root–root edges or v-structures, whereas global continuous objectives can sometimes absorb this covariance with fewer structural disruptions. Bounded uniform-ball roots favor CDIS/GIES and FoundCause, consistent with a setting where support constraints and sharper conditional changes provide useful search or motif cues but are harder to express with smooth additive objectives. Together, these views show that a synthetic score can be traced to graph structure, local mechanisms, noise processes, and root-dependency patterns.

E.2 Pretraining-Exposure OOD Score Analysis

The synthetic factor analysis above shows where methods are relatively strong or weak. For pretrained and foundation-style methods we add an observation-only check on categories they do not list as pretraining support. For each method and factor axis, we split categories by whether they appear in that method's documented prior, and compare the method to a fixed graphical-search reference (CDIS, GIES, IGSP, and PC) on the same held-out categories:

$$R^{\mathrm{OOD}}_{m,a} = \frac{F^{\mathrm{OOD}}_{m,a} - \bar{F}^{\mathrm{OOD}}_{\mathrm{ref},a}}{\bar{F}^{\mathrm{OOD}}_{\mathrm{ref},a}}. \tag{8}$$

The figure reports $100R^{\mathrm{OOD}}$ as a percent of the reference F1 (e.g. $+60\%$ means 60% above the reference mean), not an additive F1 increment. We use this reference group because these four methods form a coherent graphical-search family among the evaluated baselines. They still rely on standard causal-discovery assumptions, but they are comparatively less tied to a specific parametric mechanism family, neural pretraining simulator, or fixed variable ordering. Their average is a useful proxy for how hard a category is under common non-pretrained graph-search procedures.

All four axes use the same scenario-level F1 rows as the main benchmark. For graph family, each scenario has one category; for mechanism, noise, and root-source axes, a scenario contributes to every factor category present in its SCM, and category scores are macro-averaged before forming ID and OOD summaries. The overall OOD score is the unweighted mean of $R^{\mathrm{OOD}}_{m,a}$ over these four factor axes. A positive value means that the method stays above the four-method graphical-search reference on held-out categories, and larger values mean a larger relative lead. We also report an ID–OOD gap: the method's relative gain on supported categories minus its relative gain on held-out categories. A larger positive gap means weaker generalization under this split, because the lead over the reference is more concentrated on documented support. Smaller gaps mean more stable generalization: a gap of zero, or a small negative gap, means the OOD relative gain is not below the ID relative gain.

Taking the union of all pretrained methods' documented priors would leave too few held-out categories for a stable factor-level analysis, and that union would keep shrinking as new pretrained models appear. A per-method split asks a narrower question: given what a method documents as pretraining support, does its relative advantage remain outside that support?

The prior labels are audited from the corresponding papers or code. For the plot, source-level priors are mapped conservatively to CausalArena's factor tags and counted only when they have a clear counterpart in the current synthetic pool. Figure 16 visualizes this mapping as a documented prior-coverage profile over the four axes, with covered/total counts beside the radar. The mean column is the unweighted average of the four axis ratios, matching the four-axis averaging used for the OOD scores; we do not use radar-polygon area, which would mix axis order with the square-root radius. The full audited category mapping is released with the benchmark artifacts.

Method	Graph	Mech.	Noise	Root	Mean
Arrow	2/10	2/16	3/15	3/14	18%
AVICI	4/10	2/16	4/15	1/14	22%
CauScale	2/10	3/16	1/15	1/14	13%
CDFM	4/10	5/16	7/15	**5/14**	38%
FoundCause	4/10	**13/16**	6/15	**5/14**	**49%**
SEA	2/10	3/16	1/15	1/14	13%
TabCausal	**5/10**	8/16	**8/15**	4/14	45%

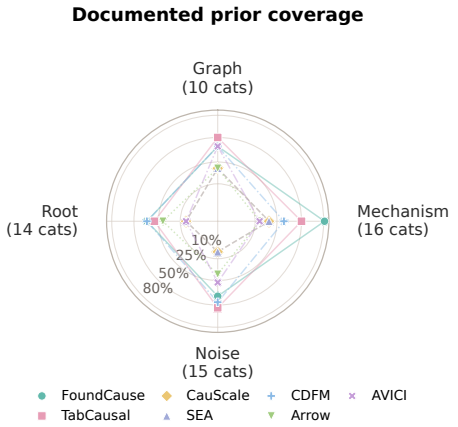


Figure 16: **Documented prior coverage for pretrained methods.** Source-level priors are mapped to CausalArena graph, mechanism, noise, and root-family tags. The radar shows axis-wise coverage ratios on a square-root radial scale. Table rows are alphabetical. Cells are covered/total tags on that axis; Mean is the unweighted average of the four ratios, reported as a percentage. Bold is the highest value in a column and underline is second; ties share the mark. CauScale and SEA have the same mapped coverage, so their radar markers overlap.

Figure 16 shows that the pretrained methods differ in documented support: FoundCause has the broadest mapped mechanism coverage ($13/16$), TabCausal is more balanced across graph, mechanism, and noise axes ($5/10$, $8/16$, and $8/15$), CDFM covers a relatively broad noise/root set, and AVICI has reasonable graph-family coverage but narrow mechanism and root coverage. CauScale and SEA share the same mapped counts. Those coverage differences are why the check is defined per method.

Figure 17 then asks whether a method's lead survives on categories it does not document as pretraining support. The top row is easiest to read as a location relative to the reference tick. FoundCause's ID and OOD bars both sit well to the right of the ticks on every axis, with OOD relative gains near $+60\%$. TabCausal remains above the ticks with smaller but still positive gains. CauScale is modestly above the reference, most clearly on mechanism and noise. CDFM is close to the ticks; SEA straddles them. Arrow and AVICI fall to the left of the ticks on several axes, so their OOD relative gains are negative. Holding out undocumented categories therefore changes how we read the scores more than it changes the coarse ranking: FoundCause and TabCausal stay first and second, but the bars show

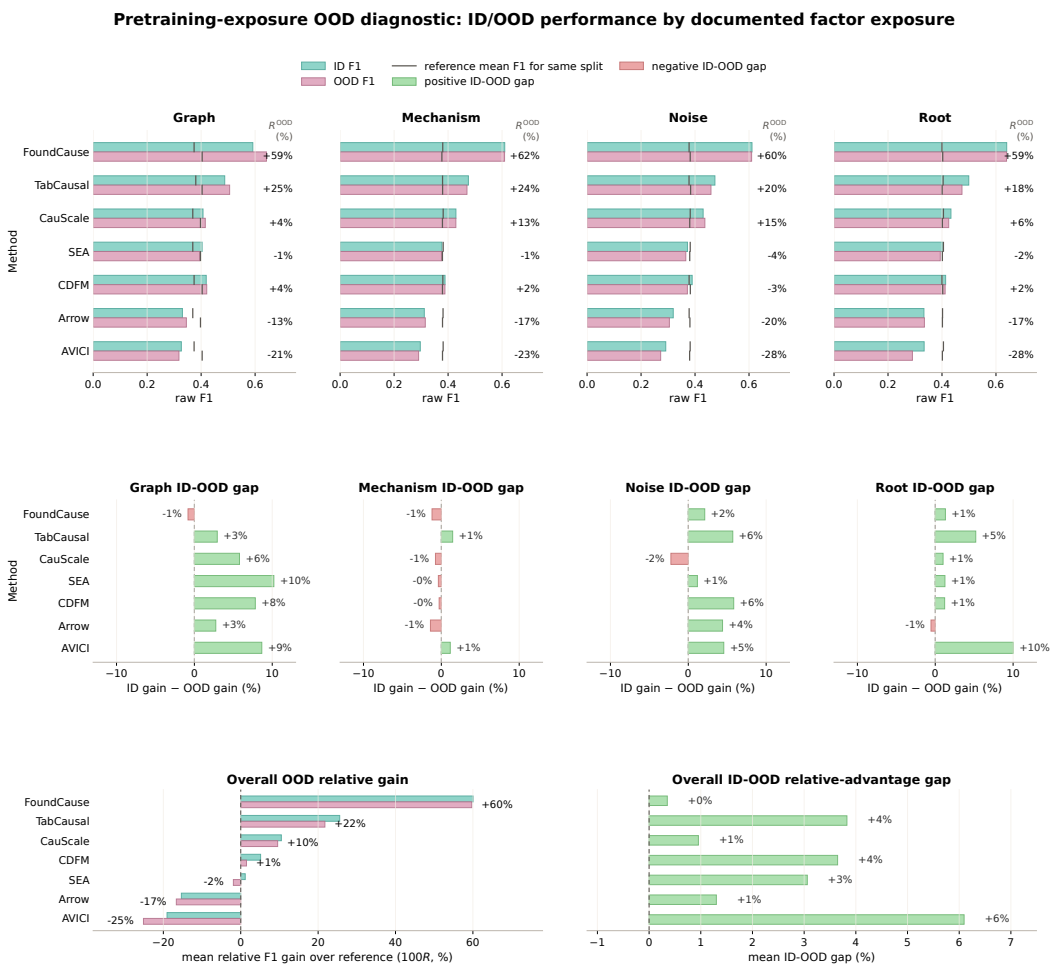


Figure 17: **Observation-only pretraining-exposure OOD diagnostics.** For each pretrained method we split synthetic obs-only categories into documented support (ID) and held-out categories (OOD), using that method's own audited prior. Columns are the four factor axes. *Top row:* raw F1 on ID (teal) and OOD (pink). The dark tick on each bar is the graphical-search reference mean (CDIS, GIES, IGSP, PC) on the same ID or OOD slice; a bar to the right of its tick is better than the reference. The number at the right of each OOD bar is the relative gain R^{OOD} from Equation (8), shown as a percent of the reference F1. *Middle row:* ID relative gain minus OOD relative gain, in the same percent units. A larger positive value means a larger drop from ID to OOD in that relative lead, so generalization is weaker; smaller values mean more stable generalization, and a gap of zero or a small negative value means OOD is not worse than ID. Green is a positive gap (advantage shrinks on OOD); red is a negative gap (advantage grows on OOD). *Bottom row:* unweighted means over the four axes. Left: overall R^{OOD}, how far the method sits above the four-baseline mean on held-out categories, again as a percent. Right: the mean of the middle-row gaps, how much larger the ID relative gain is than the OOD relative gain; smaller values indicate more stable generalization.

whether that order is confined to documented support.

The bottom-left panel summarizes the same OOD comparison: it is the unweighted mean of R^{OOD} over the four axes, i.e., how far each method sits above the four-baseline mean on held-out categories. FoundCause remains far above that reference (about $+60\%$). TabCausal is next ($+22\%$), with positive gains on all four axes. CauScale is also positive overall ($+10\%$), especially on mechanism and noise, while CDFM is only slightly positive ($+1\%$) and SEA is near the reference (-2%). Arrow and AVICI have negative overall OOD gains (-17% and -25%).

The middle row and the bottom-right panel measure generalization as a drop from ID to OOD in relative gain. A larger positive bar means weaker generalization: the method's lead over the reference

is more concentrated on documented support. Smaller values mean more stable generalization, and a gap of zero, or a small negative gap, means OOD is not worse than ID. Green is a positive ID–OOD gap. Under this reference group, 20 of 28 method–axis gaps are positive, and every overall mean gap is nonnegative or nearly zero. The bottom-right panel is the mean of those four middle-row gaps, so smaller values indicate more stable generalization. FoundCause is the exception that stays flat (overall gap $+0\%$), so its large lead is not concentrated on documented categories. TabCausal's advantage shrinks on held-out categories (overall gap $+4\%$) but remains substantial. CauScale's gap is small ($+1\%$). AVICI has the largest overall gap ($+6\%$), driven by graph ($+9\%$) and especially held-out root-family categories ($+10\%$), consistent with its narrow mapped root support ($1/14$). Arrow sits below the reference on most held-out axes; CDFM's positive OOD gains are smaller than those of FoundCause, TabCausal, and CauScale despite broader noise/root coverage. Broad documented pretraining support helps, but architecture, how coverage is balanced across factors, and how the model uses nonparametric graph signals still matter.

Those four-axis means also let us back out a crude full-ID / full-OOD F1 for the same average that Figure 11 reports. Let F be that average F1, c the unweighted mean documented coverage over the four axes, and

$$\delta = \frac{1 + \bar{R}^{\mathrm{ID}}}{1 + \bar{R}^{\mathrm{OOD}}} - 1 \tag{9}$$

the relative gap between the mean ID and mean OOD leads in Figure 17. Treating the reported score as a coverage mix,

$$F = c\,\hat{F}^{\mathrm{ID}} + (1 - c)\,\hat{F}^{\mathrm{OOD}}, \qquad \hat{F}^{\mathrm{ID}} = (1 + \delta)\,\hat{F}^{\mathrm{OOD}}, \tag{10}$$

gives $\hat{F}^{\mathrm{OOD}} = F/(1 + c\delta)$ and $\hat{F}^{\mathrm{ID}} = F(1 + \delta)/(1 + c\delta)$. This is only an estimate: c and δ are four-axis means, and F is the family–split average already used in Figure 11.

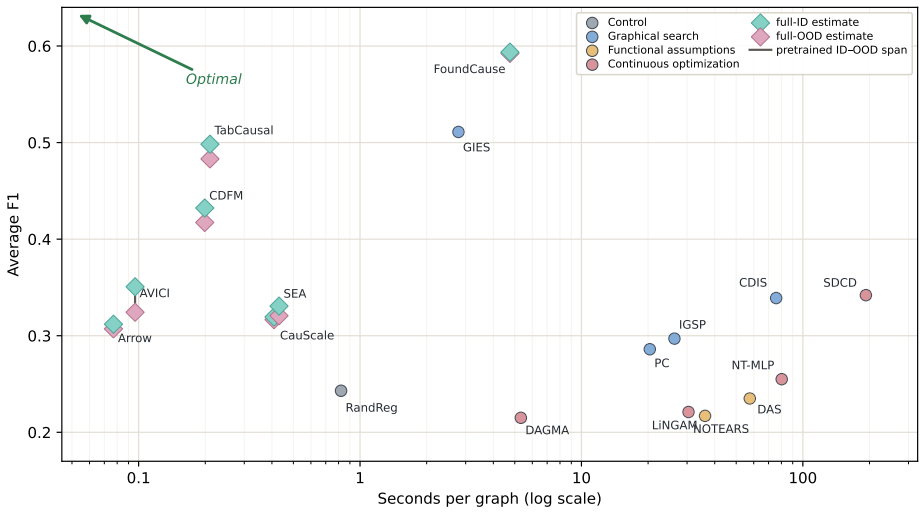


Figure 18: **Accuracy–cost view with full-ID / full-OOD estimates.** Same axes as Figure 11: average F1 against comparable per-graph time, with pretrained x equal to T_{graph}. There is no Pareto curve. Each pretrained method is a vertical pair: the teal diamond is $\hat{F}^{\mathrm{ID}}$ and the pink diamond is $\hat{F}^{\mathrm{OOD}}$ from Equation (10); the observed F lies on that segment. Circles are non-pretrained methods at their reported F1. *Optimal* points toward higher F1 and lower time.

Figure 18 plots the same ranking as a vertical F1 interval. The segments are short. FoundCause stays near 0.59 at either endpoint. TabCausal moves from about 0.48 (full OOD) to 0.50 (full ID) around its reported 0.49. CDFM is 0.42–0.43; AVICI has the longest relative span (0.32–0.35), consistent with its larger ID–OOD gap and narrower mapped support. CauScale, SEA, and Arrow barely move. The coarse pretrained order is therefore not an artifact of documented coverage: replacing the observed mix with either a fully ID or a fully OOD reading does not reorder FoundCause, TabCausal, and the cheaper lower-F1 group.

E.3 Semantic and Formula Benchmark Behavior Analysis

Semantic analysis protocol. The previous subsections use anonymous synthetic factors. Semantic and Formula SCMs let us look at operational processes and scientific formula families. We report self-centered F1 shifts: a method's F1 on a domain minus its own family average. Positive values mark domains where the method scores higher than its own family average; negative values mark domains where it scores lower. The plots show which semantic or scientific settings each method handles better or worse.

Example scenarios. Figure 6 and the examples in Sections 3.2 and 3.3 introduce two released SCMs: a wildfire smoke / shelter process and a barometric pressure reduction to sea level. We reuse those scenarios here when interpreting domain-level Semantic and Formula behavior: a missing or extra edge is then a claim about that smoke day (e.g., treating a consumer indoor reading as true indoor PM2.5, or dropping filter provision as a parent of indoor concentration) or a claim about the pressure reduction (e.g., writing the sea-level formula without fused height, or treating a thermometer reading as if it were already layer-mean virtual temperature). Released Semantic and Formula SCMs, including these two, pass several review and validation passes before they leave the construction pipeline (Section B.4).

We focus on observation-only behavior, restricting the comparison to methods evaluated on the same split. In Figure 19, the heatmaps show method-domain self-centered F1 shifts, and the side bars report each method's standard deviation across domains. A larger standard deviation means that the method is more domain-sensitive after subtracting its own average performance.

Domain-level patterns. After each method's own average is removed, some domains still sit consistently above or below that baseline. On Semantic, cybersecurity and IT operations is usually weak: many methods score lower there than on their own Semantic average. Education, healthcare delivery, manufacturing, and water and sanitation are the usual strong ones, though not for every method. The side bars repeat this by method. DAS, LiNGAM, and DAGMA hardly move across Semantic domains; PC, CDIS, AVICI, and TabCausal move more. On Formula, energy systems and thermodynamics tend to be easier. Astronomy, electromagnetism, and optics and waves more often invert a law, pass through a geometric identity, or put a sensor chain around the equation, and that is where several methods lose relative F1.

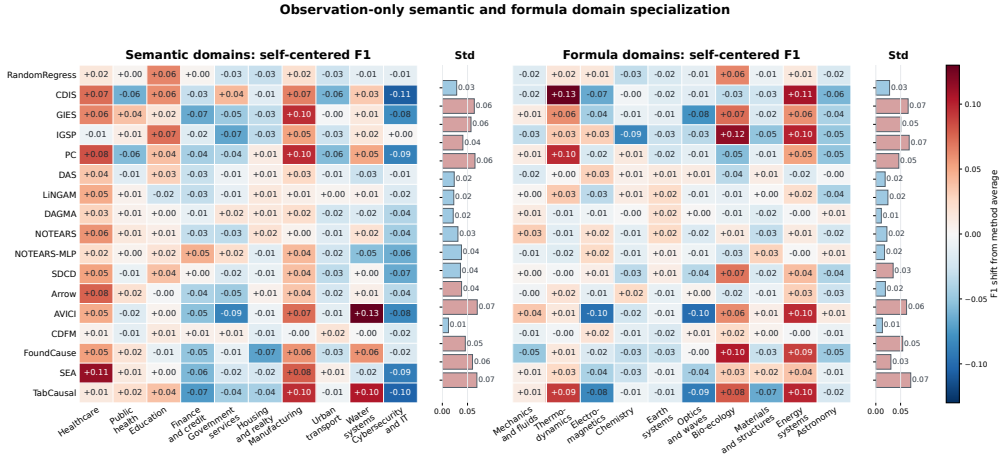


Figure 19: **Observation-only Semantic and Formula domain specialization.**

Method-level summaries. A domain heatmap only locates the gain or drop. It does not describe the processes or equations behind that domain. To read the Semantic and Formula benchmarks at that grain, we used an LLM agent to summarize each method's characteristic strengths and weaknesses from the observation-only results together with the scenario descriptions. Table 5 reports those summaries: how the method tends to behave on Semantic scenarios, how it tends to behave on Formula scenarios, and a short overall sketch.

Table 5: **Method-level behavior on the Semantic and Formula benchmarks.** An LLM agent summarized each method's characteristic strengths and weaknesses from the observation-only scores and the scenario descriptions. The table reports those qualitative directions; it does not list raw scores.

Method	Semantic benchmark direction	Formula benchmark direction	Overall profile
RandomRegress	Appears strong only when the variable order accidentally matches a simple operational sequence; collapses when the story requires distinguishing multiple paths, coordination, or behavioral response.	Shows the same order artifact on formula cases: it can match local support by chance, but fails on coupled equations, conservation relations, and measurement chains.	A sanity control, not a semantic reasoner; its large swings expose ordering and sparse-regression artifacts.
CDIS	Favors locally clear signal-to-decision structures, where an observed input flows into a screening or triage step and then to an outcome; weakens when demand, policy response, backlog, or containment propagates through the system.	Handles small directed formula chains better than coupled multiplicative systems or shared-parameter equations.	Works best when the causal story is locally separable; hidden demand, shared drivers, and cascading structure blur the conditional signals it relies on.
GIES	Prefers stable, sparse, stage-like workflows in which monitoring signals lead to interventions or outcomes; degrades under accumulated pressure, capacity limits, and cross-module propagation.	More reliable on static physical constraints than on readout chains, exponential/periodic transformations, or observation-heavy formula contexts.	Score/search behavior is strongest for stable staged mechanisms and less reliable when the observed variables are indirect traces of a latent process.
IGSP	Performs well on observable service workflows such as processing, maintenance, vaccination, or patching; weakens when rare positives, hidden eligibility, or surveillance logic determine the graph.	Handles some direct multiplicative or power-law chains, but struggles with thermodynamic, absorption, and exchange-process relations.	Benefits from explicit workflow/intervention alignment; sparse event evidence and scientific readout layers make its direction tests less decisive.
PC	Recovers local screening and inspection cascades better than adaptive, behavioral, or policy-response processes.	Handles low-coupling formula systems better than equations with shared parameters, multiple outputs, or derived readouts.	Conditional-independence search is useful when the semantic story decomposes into local tests, but brittle under adaptation, common drivers, and indirect measurements.
DAS	Shows mild advantages on anomaly-monitoring and escalation chains, but loses advantage in outreach, engagement, and priority-queue settings.	Works better on direct physical-quantity chains than on biological, load, or pump-like coupled systems.	Has no sharp semantic specialization; it is mainly helped when the ordering signal is simple and directly observable.
LiNGAM	Favors continuous, one-way trigger-to-action stories; weakens when capacity, adherence, follow-up management, or saturation controls the outcome.	More plausible on monotone scaling relations and weaker on nonlinear response, buffering, oscillation, or multi-input scientific systems.	Its linear non-Gaussian orientation prior can help on direct driving chains, but does not match capacity-limited or strongly nonlinear semantic processes.
DAGMA	Works comparatively well when a bottleneck or allocation lever explains a measurable downstream outcome; weakens on distributed monitoring, tracing, and coordination chains.	Fits smooth, compressible formula skeletons better than load, saturation, or shared-mechanism systems.	Continuous optimization favors centralized, smooth causal structure; diffuse coordination and non-smooth process logic are harder to represent.
NOTEARS	Favors direct risk-or-demand-to-action stories; degrades when the system involves propagation, containment, or multi-stage response.	Can recover some near-monotone physical chains, but loses reliability on complex multiplicative, thermal, and interference relations.	Linear DAG optimization is most useful when a semantic process can be compressed into a direct risk-to-action edge pattern.
NOTEARS-MLP	The nonlinear SEM helps on triage and resource-allocation chains, but not enough for saturation, alert fatigue, overflow, or backlog dynamics.	Captures closed-form nonlinear formulas better than linear NOTEARS, yet remains weak on control policies, threshold gates, and process feedback.	Extra function flexibility helps, but system-level semantics such as accumulation, saturation, and feedback remain the limiting factor.
SDCD	Performs better when an early warning signal leads to a resource intervention, and worse when capacity constraints or long-run accumulation modulate the effect.	Is relatively favorable on scale-law or power-law relations, and weaker on measurement readouts and geometric observation chains.	Most effective for monotone, observable intervention chains; less robust when a hidden state or measurement layer mediates the causal relation.
Arrow	Favors event-response stories: an incident is detected and then remediated or mitigated. It is weaker on planning, scheduling, and resource-prioritization tasks.	Handles some engineering and geometric relations, but weakens on biochemical response curves and buffered systems.	Better aligned with event-response semantics than with strategic allocation or delayed planning.

Method	Semantic benchmark direction	Formula benchmark direction	Overall profile
AVICI	Strongest when a concrete trigger or intervention changes a downstream metric; weaker in administrative coordination, workload, and prioritization cascades.	Covers several common physical-law structures, but its performance varies noticeably across formula types and readout contexts.	Its pretrained prior helps with local trigger-to-outcome chains, while multi-actor organization and instrumented scientific contexts remain harder.
CauScale	Works best on short, separable signal-to-action stories; weakens on drift, spillover, allocation competition, and equilibrium-like processes.	Very strong when calibrated physical inputs directly determine a law output, but much weaker once control, calibration, or instrument loops wrap the formula.	Especially good at compact causal chains; system-level equilibrium, propagation, and control layers are the main failure mode.
CDFM	Does not show a stable semantic specialization in this release; strong and weak residual cases are close after centering.	The formula readout is similarly flat, with no reliable direction that separates favorable and unfavorable scientific contexts.	The residual evidence does not support a clear method-specific semantic profile, so we avoid over-interpreting it.
FoundCause	Favors direct readiness, allocation, and response chains; weakens on drift, spillover, cascade, and feedback-like processes.	Handles some clearly structured formula chains, but struggles when shared physical parameters, vessels, instruments, or heat-exchange contexts mediate the relation.	Structural refinement appears helpful for direct response chains, but less sufficient for long-range propagation and shared-mechanism systems.
SEA	Favors short detection-to-intervention stories and quality/adherence-to-outcome chains; weakens under retention, allocation, cascade, and spillover dynamics.	Works better on classic single-law physics than on field-strength, wave/readout, or complex multiplicative relations.	Stronger on short protocol-like causal stories; slower accumulation and cross-system propagation remain difficult.
TabCausal	Favors clear policy or operational levers whose effects flow to downstream outcomes; weakens on load propagation, backlog, and cascade-heavy systems.	Strong on direct formula pipelines from physical inputs to derived quantities, but weaker when instrumentation, regime switches, or measurement chains surround the base law.	Its strongest semantic profile is clear intervention point plus directed downstream chain; multi-stage propagation and instrumented readout chains remain the main weaknesses.

Operational-semantic failure modes. The following two paragraphs paraphrase Table 5; they are LLM-assisted qualitative readings, not a separate human coding of errors. On Semantic cases, the summaries look better on a short visible chain and worse once backlog, drift, or a cascade takes over. RandomRegress looks strong only when the column order happens to match a simple sequence; it fails when the case needs several paths, coordination, or a behavioral response. PC, CDIS, IGSP, and GIES do better when a visible input goes into screening, inspection, or a staged workflow, and worse when demand, backlog, capacity, or an adaptive response moves through the rest of the system. IGSP also drops when rare events or hidden eligibility set the graph. DAS has only a mild edge on anomaly monitoring and escalation, and loses it on outreach and queues. LiNGAM prefers a one-way trigger-to-action chain, and drops when capacity, adherence, or saturation sets the outcome. DAGMA, NOTEARS, NOTEARS-MLP, and SDCD like one bottleneck, risk signal, triage step, or early warning that moves a measurable next action; they struggle with coordination, overflow, backlog, and long-run accumulation. Arrow fits detect-then-fix better than planning or scheduling. AVICI, CauScale, FoundCause, SEA, and TabCausal likewise prefer a short visible action that moves a downstream outcome, and they weaken on drift, spillover, workload, competing allocation, or cascade. CDFM is too flat in this release to assign a Semantic profile.

Scientific-formula failure modes. On Formula cases, they look better on a short equation chain, and worse once extra instruments sit around that equation. RandomRegress can match a few nearby variables by chance, then fail on coupled equations, conservation relations, and measurement chains. PC, CDIS, IGSP, and GIES handle small directed formulas better than shared parameters, extra outputs, or readout chains; IGSP is also weak on thermodynamic, absorption, and exchange relations. DAS prefers a direct physical-quantity chain over a biological, load, or pump-like coupled system. LiNGAM is more plausible on monotone scaling than on nonlinear response, buffering, oscillation, or several inputs. DAGMA and NOTEARS can fit a smooth or near-monotone formula, but not load, saturation, thermal, or interference structure. NOTEARS-MLP helps on a closed-form nonlinear formula, but not on control policies, threshold gates, or process feedback. SDCD likes power-law relations and not measurement or geometric readouts. CauScale is strong when calibrated physical inputs directly set a law output, and much weaker once control, calibration, or instrument loops wrap the law. FoundCause and TabCausal like a clear input-to-derived-quantity chain, and weaken when vessels, instruments, heat exchange, regime switches, or measurement chains sit around the

law. Arrow handles some engineering and geometric relations, not biochemical curves or buffered systems. SEA prefers a classic single-law physics case over field, wave, or complex product relations. AVICI covers several common physical-law forms, but the score still moves with the formula type and the readout. CDFM again has no reliable Formula direction.

F Discussion

F.1 Open and Evolvable Evaluation

The most direct validation of a causal discovery claim is often feedback in a real system: whether intervening on a recovered parent changes the intended downstream quantity. During method development, however, such closed-loop checks are usually unavailable at scale, so the field relies on benchmarks with known graphs. CausalArena is designed for that development setting. Using executable SCMs rather than only fixed sampled tables makes graph structure, mechanisms, interventions, and sampling procedures inspectable and reproducible, and it lets later releases add new SCMs while preserving the interfaces in Section 2.

For CDFMs, openness also changes how a score should be read. Once an SCM, its graph, or its generator becomes public, related environments may enter later pretraining. This does not make a public benchmark useless, but it does mean that a result should be interpreted together with the model's training context: performance on a familiar generator distribution and performance on newly introduced environments answer different questions. We therefore treat CausalArena as an *evolvable* arena. Versioned public releases provide reproducible reference evaluations, while the common SCM interface supports newly authored graphs, mechanisms, and scenarios as pretrained models evolve. This is how the freshness requirement in Section 2.2 is pursued without giving up transparency.

F.2 Interpreting Foundation-Model Results

To make pretrained comparisons easier to read, CausalArena ties results to explicit benchmark versions and release states. Leaderboard entries are encouraged to disclose whether public CausalArena SCMs, generated samples, scenario metadata, or closely related generators were used during development or pretraining. The current package distinguishes public and reserved subsets: 500 synthetic, 50 semantic, and 50 formula-grounded SCMs are public, while the remaining audited SCMs are reserved for held-out leaderboard evaluation and progressive release in later benchmark versions. Results in this paper use the complete 1,200-SCM pool and are labeled by this benchmark version.

These practices make pretraining–evaluation overlap easier to discuss; they are not a claim that overlap can be removed once and for all. As earlier versions become public, their SCMs may eventually enter future training corpora. Continued evaluation of new foundation models therefore benefits from periodically extending the SCM pool and reporting scores together with the benchmark version and training disclosure. The semantic construction pipeline is one practical route for adding newly authored operational environments in later releases.

F.3 Scope and Future Extensions

The current release focuses on multivariate tabular causal structure recovery with known directed graph ground truth. Synthetic SCMs emphasize controlled breadth, semantic SCMs emphasize operational grounding, and formula-grounded SCMs emphasize explicit scientific mechanisms. Public real-world datasets serve as an external-validity check; their published graphs should be read as reference structures, with the caveat that they are not uniquely verified causal ground truth.

The same executable interface can later support additional directions: new domains and SCM families, larger graphs, latent confounding, and neighboring causal tasks such as intervention-target reasoning, effect estimation, and counterfactual inference. Such extensions can reuse the current versioning and evaluation infrastructure while remaining comparable with earlier releases.

G Complete Results

This section reports the numerical result views behind the compact main-text tables and figures. Overall benchmark tables include F1, precision, recall, SHD, SID, normalized SHD (nSHD), AUROC, and AP when the corresponding method output supports the metric; hard-graph methods without edge scores therefore have missing AUROC/AP entries. nSHD is computed as $\mathrm{SHD}/(d(d-1))$. Values are reported as mean±standard deviation across heterogeneous graph instances or datasets. The reported standard deviation is a descriptive spread and can exceed the mean for nonnegative metrics such as SHD. Bold and underline mark the best and second-best values within each comparable block, separately for every metric. To keep the appendix readable, factor-level decompositions use compact F1 views; the released CSV reports retain the full precision/recall/SHD/SID columns for each slice.

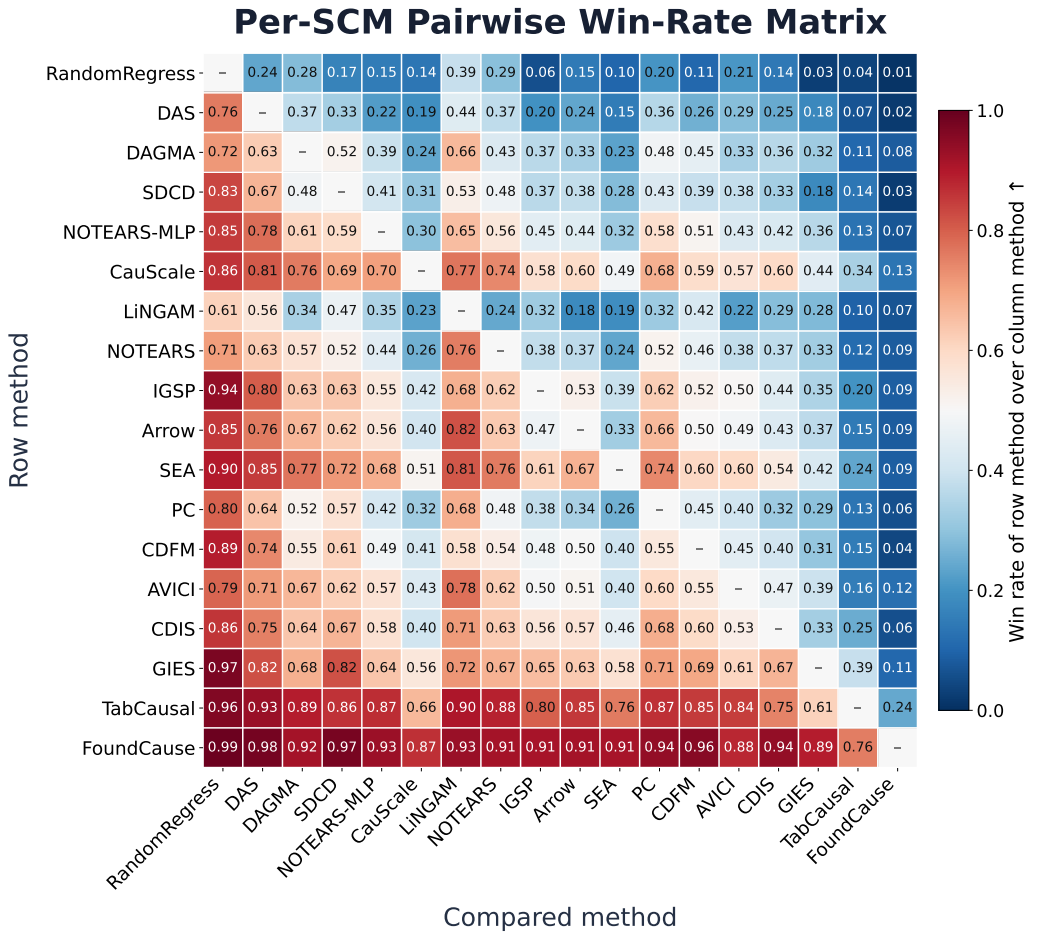


Figure 20: **Complete observation-only pairwise win-rate matrix.** Each cell reports how often the row method wins over the column method across per-SCM or per-dataset comparisons, after averaging repeated runs within the same unit. On each unit, F1 and SHD are compared separately (higher F1 wins, lower SHD wins, ties count as 0.5), and wins from both metrics are pooled into one rate. Methods are ordered by mean family-wise win rate, matching Figure 1. Red cells indicate that the row method wins more often; blue cells indicate the opposite.

G.1 Synthetic Benchmark

Table 6 reports the overall synthetic benchmark metrics. Tables 7 to 10 decompose the same benchmark by graph family, dimension, root-dependency mode, and difficulty component. Tables 11 and 12 give compact edge-level F1 views by mechanism and noise labels.

Table 6: **Complete synthetic benchmark metrics.** Values are mean±standard deviation over replicate-level rows.

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	RandomRegress	5k	0.28 ±0.10	0.24 ±0.11	0.36 ±0.11	218.0 ±310.0	0.103 ±0.054	1003.6 ±1488.4	0.50 ±0.06	0.13 ±0.07
Obs.	CDIS	5k	0.38 ±0.16	0.47 ±0.20	0.34 ±0.16	143.0 ±352.0	0.064 ±0.045	826.6 ±1357.4	–	–
Obs.	GIES	5k	0.48 ±0.15	0.46 ±0.17	0.52 ±0.14	123.6 ±174.9	0.063 ±0.039	798.1 ±1141.9	–	–
Obs.	IGSP	5k	0.41 ±0.15	0.43 ±0.18	0.40 ±0.14	135.6 ±192.3	0.063 ±0.034	783.3 ±1239.3	–	–
Obs.	PC	5k	0.31 ±0.14	0.35 ±0.15	0.29 ±0.15	115.4 ±149.0	0.062 ±0.034	990.2 ±1596.1	–	–
Obs.	DAS	5k	0.34 ±0.12	0.33 ±0.13	0.36 ±0.14	142.6 ±180.1	0.077 ±0.039	722.1 ±1128.5	–	–
Obs.	LiNGAM	5k	0.21 ±0.14	0.53 ±0.23	0.14 ±0.11	109.6 ±142.2	0.060 ±0.029	664.6 ±1139.3	–	–
Obs.	DAGMA	5k	0.28 ±0.14	0.45 ±0.17	0.22 ±0.13	107.2 ±138.6	0.059 ±0.031	697.4 ±1136.9	–	–
Obs.	NOTEARS	5k	0.28 ±0.14	0.55 ±0.18	0.20 ±0.12	104.2 ±136.0	0.057 ±0.029	669.1 ±1127.0	0.50 ±0.08	0.21 ±0.09
Obs.	NOTEARS-MLP	5k	0.35 ±0.12	0.43 ±0.14	0.31 ±0.14	107.0 ±136.4	0.060 ±0.033	742.4 ±1145.4	0.74 ±0.07	0.29 ±0.12
Obs.	SDCD	5k	0.38 ±0.11	0.36 ±0.13	0.43 ±0.13	135.5 ±165.8	0.076 ±0.042	787.9 ±1137.8	–	–
Obs.	Arrow	5k	0.34 ±0.15	0.52 ±0.16	0.27 ±0.15	106.0 ±140.7	0.056 ±0.031	661.5 ±1132.9	0.76 ±0.08	0.35 ±0.15
Obs.	AVICI	5k	0.32 ±0.17	0.64 ±0.18	0.24 ±0.15	104.5 ±139.7	0.054 ±0.026	652.3 ±1129.6	0.75 ±0.09	0.37 ±0.16
Obs.	CauScale	5k	0.41 ±0.13	0.57 ±0.18	0.34 ±0.12	85.8 ±107.2	0.054 ±0.033	605.2 ±983.2	0.84 ±0.06	0.44 ±0.12
Obs.	CDFM	5k	0.42 ±0.15	0.43 ±0.13	0.46 ±0.20	136.4 ±176.9	0.071 ±0.036	654.8 ±1072.2	0.83 ±0.08	0.43 ±0.18
Obs.	FoundCause	5k	**0.62 ±0.13**	**0.69 ±0.16**	**0.59 ±0.16**	**80.8 ±112.0**	**0.042 ±0.025**	**524.4 ±937.8**	**0.90 ±0.06**	**0.66 ±0.14**
Obs.	SEA	5k	0.40 ±0.15	0.58 ±0.18	0.33 ±0.15	99.2 ±129.2	0.053 ±0.029	646.9 ±1105.6	0.84 ±0.07	0.44 ±0.14
Obs.	TabCausal	5k	0.50 ±0.16	0.64 ±0.17	0.44 ±0.18	93.6 ±127.3	0.050 ±0.035	622.5 ±1067.0	0.83 ±0.09	0.51 ±0.18
Obs.+int	CDIS	5k	0.40 ±0.15	0.49 ±0.20	0.36 ±0.15	140.1 ±353.8	0.061 ±0.043	808.5 ±1330.8	–	–
Obs.+int	GIES	5k	0.48 ±0.14	0.44 ±0.15	**0.55 ±0.14**	134.0 ±181.4	0.067 ±0.038	859.7 ±1217.2	–	–
Obs.+int	IGSP	5k	0.39 ±0.14	0.42 ±0.18	0.37 ±0.14	132.8 ±186.1	0.062 ±0.033	780.7 ±1234.3	–	–
Obs.+int	SDCD	5k	0.38 ±0.11	0.34 ±0.12	0.46 ±0.13	148.1 ±174.9	0.084 ±0.043	808.8 ±1155.8	–	–
Obs.+int	AVICI	5k	0.35 ±0.17	**0.65 ±0.17**	0.26 ±0.17	104.2 ±139.7	0.053 ±0.026	651.3 ±1129.4	0.75 ±0.10	0.38 ±0.17
Obs.+int	CauScale	5k	0.46 ±0.12	0.59 ±0.16	0.39 ±0.12	**84.6 ±106.7**	0.053 ±0.032	**598.2 ±968.8**	**0.84 ±0.06**	0.48 ±0.12
Obs.+int	TabCausal	5k	**0.50 ±0.15**	0.61 ±0.15	0.46 ±0.19	96.0 ±128.4	**0.052 ±0.038**	622.9 ±1060.5	0.84 ±0.09	**0.52 ±0.18**

Table 7: **Compact synthetic graph-family breakdown.** Cells report Obs./Obs+int F1.

Method	Erdos–Renyi	Scale-free	Chain	Tree	Layered	Collider-rich	Bipartite layered	Dense local modules	Small-world	Block-sparse layered
RandomRegress	0.27/–	0.27/–	0.30/–	0.27/–	0.27/–	0.27/–	0.27/–	0.31/–	0.30/–	0.30/–
CDIS	0.42/0.42	0.30/0.32	0.33/0.36	0.31/0.35	0.40/0.42	0.41/0.42	0.58/**0.58**	0.34/0.39	0.34/0.36	0.34/0.38
GIES	0.50/**0.48**	0.42/0.46	0.45/0.50	0.44/0.48	0.52/**0.50**	0.49/0.47	0.61/0.51	0.45/0.46	0.44/**0.48**	0.47/0.47
IGSP	0.37/0.36	0.41/0.39	0.38/0.37	0.43/0.40	0.39/0.37	0.38/0.36	0.46/0.44	0.44/0.41	0.36/0.35	0.45/0.43
PC	0.28/–	0.25/–	0.34/–	0.30/–	0.32/–	0.28/–	0.29/–	0.35/–	0.31/–	0.34/–
DAS	0.34/–	0.39/–	0.31/–	0.38/–	0.33/–	0.33/–	0.30/–	0.37/–	0.30/–	0.35/–
LiNGAM	0.18/–	0.21/–	0.22/–	0.22/–	0.21/–	0.19/–	0.18/–	0.23/–	0.19/–	0.23/–
DAGMA	0.23/–	0.34/–	0.30/–	0.37/–	0.26/–	0.24/–	0.19/–	0.32/–	0.25/–	0.36/–
NOTEARS	0.24/–	0.30/–	0.30/–	0.33/–	0.27/–	0.23/–	0.20/–	0.30/–	0.24/–	0.34/–
NOTEARS-MLP	0.32/–	0.39/–	0.37/–	0.41/–	0.32/–	0.32/–	0.29/–	0.39/–	0.32/–	0.40/–
SDCD	0.41/0.40	0.41/0.41	0.34/0.35	0.39/0.38	0.39/0.39	0.39/0.40	0.41/0.38	0.37/0.35	0.36/0.37	0.36/0.34
Arrow	0.33/–	0.33/–	0.33/–	0.36/–	0.35/–	0.33/–	0.35/–	0.37/–	0.30/–	0.38/–
AVICI	0.30/0.31	0.38/0.41	0.34/0.37	0.40/0.43	0.29/0.31	0.30/0.31	0.24/0.24	0.34/0.37	0.30/0.33	0.33/0.36
CauScale	0.39/0.43	0.42/0.46	0.40/0.45	0.46/0.49	0.43/0.46	0.40/0.44	0.42/0.47	0.41/0.47	0.36/0.40	0.45/0.49
CDFM	0.41/–	0.42/–	0.40/–	0.43/–	0.40/–	0.41/–	0.46/–	0.45/–	0.40/–	0.43/–
FoundCause	**0.62**/–	**0.56**/–	**0.59**/–	**0.61**/–	**0.66**/–	**0.63**/–	**0.73**/–	**0.61**/–	**0.57**/–	**0.64**/–
SEA	0.40/–	0.41/–	0.37/–	0.41/–	0.41/–	0.39/–	0.42/–	0.40/–	0.35/–	0.41/–
TabCausal	0.47/0.48	0.51/**0.53**	0.52/**0.53**	0.55/**0.56**	0.47/0.48	0.48/**0.48**	0.47/0.44	0.52/**0.52**	0.46/0.48	0.52/**0.52**

Table 8: **Compact synthetic dimension breakdown.** Cells report Obs./Obs+int F1.

Method	10	20	30	50	100
RandomRegress	0.34/–	0.31/–	0.28/–	0.25/–	0.23/–
CDIS	0.36/0.44	0.39/0.42	0.39/0.40	0.38/0.38	0.36/0.35
GIES	0.49/0.55	0.49/0.49	0.49/0.48	0.47/0.45	0.47/0.44
IGSP	0.49/0.47	0.45/0.42	0.42/0.40	0.36/0.34	0.32/0.30
PC	0.42/–	0.34/–	0.31/–	0.26/–	0.20/–
DAS	0.43/–	0.38/–	0.34/–	0.30/–	0.25/–
LiNGAM	0.29/–	0.23/–	0.21/–	0.18/–	0.13/–
DAGMA	0.34/–	0.30/–	0.29/–	0.27/–	0.22/–
NOTEARS	0.33/–	0.30/–	0.28/–	0.26/–	0.21/–
NOTEARS-MLP	0.39/–	0.38/–	0.36/–	0.34/–	0.29/–
SDCD	0.38/0.40	0.39/0.39	0.39/0.38	0.38/0.36	0.37/0.36
Arrow	0.45/–	0.40/–	0.36/–	0.29/–	0.21/–
AVICI	0.41/0.46	0.40/0.43	0.36/0.38	0.28/0.29	0.16/0.16
CauScale	0.35/0.43	0.39/0.44	0.41/0.46	0.44/0.47	0.48/**0.49**
CDFM	0.54/–	0.49/–	0.44/–	0.37/–	0.27/–
FoundCause	**0.67**/–	**0.65**/–	**0.63**/–	**0.61**/–	**0.56**/–
SEA	0.46/–	0.44/–	0.42/–	0.39/–	0.27/–
TabCausal	0.56/**0.58**	0.56/**0.56**	0.53/**0.54**	0.47/**0.47**	0.36/0.35

Table 9: **Compact synthetic root-dependency breakdown.** Cells report Obs./Obs+int F1.

Method	correlated normal	independent roots	random covariance normal	uniform ball
RandomRegress	0.23/–	0.29/–	0.22/–	0.30/–
CDIS	0.27/0.29	0.38/0.41	0.30/0.30	0.45/0.47
GIES	0.37/0.40	0.49/0.49	0.39/0.40	0.54/**0.51**
IGSP	0.35/0.34	0.41/0.39	0.37/0.35	0.43/0.40
PC	0.25/–	0.31/–	0.26/–	0.32/–
DAS	0.33/–	0.34/–	0.33/–	0.32/–
LiNGAM	0.21/–	0.21/–	0.19/–	0.20/–
DAGMA	0.32/–	0.28/–	0.31/–	0.24/–
NOTEARS	0.30/–	0.28/–	0.29/–	0.24/–
NOTEARS-MLP	0.38/–	0.35/–	0.37/–	0.32/–
SDCD	0.36/0.35	0.39/0.38	0.36/0.35	0.38/0.38
Arrow	0.32/–	0.35/–	0.30/–	0.34/–
AVICI	0.32/0.34	0.33/0.35	0.28/0.29	0.32/0.34
CauScale	0.43/**0.46**	0.41/0.46	0.42/**0.45**	0.42/0.46
CDFM	0.38/–	0.43/–	0.34/–	0.44/–
FoundCause	**0.58**/–	**0.63**/–	**0.54**/–	**0.66**/–
SEA	0.36/–	0.40/–	0.36/–	0.40/–
TabCausal	0.43/0.46	0.51/**0.51**	0.42/0.43	0.51/0.50

Table 10: **Compact synthetic difficulty-component breakdown.** Cells report Obs./Obs+int F1; multi-component cases contribute to each active component.

Method	Baseline	Distribution shift	Feature corruption	Feature warping	High noise	Missingness	Outliers	Soft interventions	Weak signal
RandomRegress	0.28/–	0.28/–	0.27/–	0.27/–	0.30/–	0.27/–	0.26/–	0.28/–	0.30/–
CDIS	0.40/0.42	0.39/0.43	0.35/0.37	0.37/0.40	0.39/0.40	0.36/0.38	0.31/0.32	0.38/0.41	0.38/0.41
GIES	0.51/0.52	0.50/0.46	0.45/0.46	0.46/**0.48**	0.52/**0.53**	0.45/0.46	0.36/0.37	0.49/0.45	0.50/**0.51**
IGSP	0.42/0.41	0.43/0.41	0.38/0.35	0.39/0.38	0.46/0.43	0.39/0.37	0.30/0.29	0.41/0.40	0.43/0.41
PC	0.33/–	0.32/–	0.31/–	0.30/–	0.31/–	0.31/–	0.23/–	0.31/–	0.31/–
DAS	0.34/–	0.35/–	0.32/–	0.33/–	0.41/–	0.32/–	0.30/–	0.34/–	0.35/–
LiNGAM	0.26/–	0.23/–	0.18/–	0.21/–	0.16/–	0.20/–	0.06/–	0.22/–	0.22/–
DAGMA	0.31/–	0.30/–	0.28/–	0.29/–	0.30/–	0.29/–	0.16/–	0.29/–	0.27/–
NOTEARS	0.30/–	0.29/–	0.27/–	0.28/–	0.28/–	0.29/–	0.14/–	0.28/–	0.26/–
NOTEARS-MLP	0.36/–	0.36/–	0.34/–	0.35/–	0.39/–	0.35/–	0.32/–	0.36/–	0.35/–
SDCD	0.40/0.40	0.40/0.37	0.36/0.36	0.39/0.39	0.42/0.41	0.36/0.37	0.28/0.29	0.40/0.38	0.40/0.38
Arrow	0.36/–	0.38/–	0.34/–	0.32/–	0.38/–	0.34/–	0.24/–	0.35/–	0.35/–
AVICI	0.35/0.38	0.34/0.36	0.31/0.34	0.29/0.32	0.29/0.30	0.32/0.35	0.31/0.34	0.34/0.34	0.28/0.30
CauScale	0.44/0.49	0.44/0.47	0.39/0.44	0.40/0.46	0.45/0.49	0.40/0.44	0.30/0.38	0.43/0.44	0.41/0.45
CDFM	0.46/–	0.45/–	0.40/–	0.41/–	0.42/–	0.41/–	0.27/–	0.44/–	0.43/–
FoundCause	**0.66**/–	**0.65**/–	**0.59**/–	**0.59**/–	**0.64**/–	**0.60**/–	**0.51**/–	**0.64**/–	**0.62**/–
SEA	0.42/–	0.43/–	0.39/–	0.40/–	0.43/–	0.39/–	0.24/–	0.41/–	0.40/–
TabCausal	0.54/**0.54**	0.53/**0.51**	0.48/**0.49**	0.43/0.45	0.51/0.52	0.49/**0.50**	0.39/**0.42**	0.52/**0.50**	0.48/0.47

Table 11: **Compact synthetic edge-level mechanism breakdown.** Cells report Obs./Obs+int edge-F1 from aggregated TP/FP/FN counts.

Method	activation additive	aggregation nonlin.	discrete assign.	exp log	linear additive	linear interact.	mixture regime	mult.	neural random	piecewise threshold
RandomRegress	0.06/–	0.10/–	0.04/–	0.08/–	0.08/–	0.08/–	0.07/–	0.04/–	0.05/–	0.08/–
CDIS	0.34/0.34	0.32/0.32	0.32/0.32	0.37/0.36	0.39/0.39	0.40/0.39	0.38/0.38	0.34/0.35	0.37/0.36	0.37/0.36
GIES	0.48/**0.47**	0.43/0.42	0.45/**0.44**	0.52/0.51	0.59/**0.58**	0.55/0.53	0.56/0.55	0.47/0.47	0.51/**0.50**	0.54/0.54
IGSP	0.34/0.32	0.32/0.31	0.31/0.29	0.38/0.36	0.42/0.40	0.39/0.38	0.40/0.39	0.34/0.33	0.36/0.34	0.39/0.38
PC	0.23/–	0.25/–	0.20/–	0.27/–	0.30/–	0.29/–	0.29/–	0.21/–	0.23/–	0.28/–
DAS	0.28/–	0.30/–	0.29/–	0.30/–	0.31/–	0.29/–	0.32/–	0.31/–	0.33/–	0.29/–
LiNGAM	0.07/–	0.12/–	0.04/–	0.09/–	0.13/–	0.12/–	0.11/–	0.05/–	0.07/–	0.11/–
DAGMA	0.22/–	0.28/–	0.19/–	0.27/–	0.29/–	0.29/–	0.28/–	0.19/–	0.22/–	0.28/–
NOTEARS	0.06/–	0.12/–	0.04/–	0.08/–	0.09/–	0.09/–	0.08/–	0.03/–	0.04/–	0.09/–
NOTEARS-MLP	0.27/–	0.28/–	0.28/–	0.27/–	0.24/–	0.26/–	0.25/–	0.23/–	0.24/–	0.25/–
SDCD	0.37/0.37	0.39/0.39	0.36/0.35	0.43/0.42	0.42/0.42	0.46/0.45	0.44/0.44	0.38/0.37	0.39/0.38	0.42/0.42
Arrow	0.22/–	0.26/–	0.16/–	0.26/–	0.33/–	0.30/–	0.31/–	0.20/–	0.23/–	0.28/–
AVICI	0.20/0.21	0.29/0.30	0.18/0.19	0.23/0.24	0.20/0.21	0.24/0.25	0.22/0.23	0.18/0.19	0.16/0.18	0.25/0.26
CauScale	0.42/0.44	0.46/0.48	0.39/0.40	0.50/**0.52**	0.54/0.56	0.54/**0.57**	0.53/**0.56**	0.47/**0.49**	0.45/0.47	0.54/**0.56**
CDFM	0.21/–	0.25/–	0.20/–	0.23/–	0.23/–	0.24/–	0.23/–	0.21/–	0.21/–	0.24/–
FoundCause	**0.55**/–	**0.57**/–	**0.52**/–	**0.61**/–	**0.65**/–	**0.65**/–	**0.64**/–	**0.57**/–	**0.57**/–	**0.65**/–
SEA	0.33/–	0.34/–	0.28/–	0.38/–	0.42/–	0.40/–	0.41/–	0.31/–	0.32/–	0.41/–
TabCausal	0.38/0.39	0.48/**0.49**	0.36/0.35	0.45/0.45	0.43/0.44	0.46/0.47	0.44/0.45	0.38/0.37	0.38/0.37	0.46/0.47

Method	polynomial power	rational ratio	rff gp	saturation clip	tree based	trig periodic
RandomRegress	0.07/–	0.07/–	0.05/–	0.07/–	0.03/–	0.05/–
CDIS	0.31/0.31	0.32/0.33	0.32/0.31	0.37/0.36	0.21/0.21	0.33/0.33
GIES	0.42/0.41	0.47/**0.46**	0.40/0.39	0.54/**0.53**	0.27/0.29	0.45/0.44
IGSP	0.32/0.31	0.33/0.32	0.29/0.28	0.39/0.37	0.22/0.22	0.33/0.31
PC	0.23/–	0.23/–	0.19/–	0.27/–	0.15/–	0.22/–
DAS	0.25/–	0.31/–	0.30/–	0.32/–	0.20/–	0.29/–
LiNGAM	0.07/–	0.10/–	0.05/–	0.10/–	0.02/–	0.06/–
DAGMA	0.27/–	0.27/–	0.19/–	0.26/–	0.17/–	0.22/–
NOTEARS	0.09/–	0.08/–	0.04/–	0.07/–	0.04/–	0.05/–
NOTEARS-MLP	0.27/–	0.24/–	0.26/–	0.25/–	0.25/–	0.30/–
SDCD	0.37/0.36	0.39/0.39	0.35/0.35	0.43/0.42	0.28/0.28	0.39/0.39
Arrow	0.22/–	0.27/–	0.18/–	0.29/–	0.10/–	0.20/–
AVICI	0.31/0.31	0.18/0.20	0.27/0.28	0.21/0.22	0.21/0.22	0.24/0.25
CauScale	0.47/**0.49**	0.44/0.46	0.44/**0.46**	0.50/0.52	0.30/0.31	0.44/**0.45**
CDFM	0.24/–	0.21/–	0.21/–	0.22/–	0.19/–	0.21/–
FoundCause	**0.57**/–	**0.54**/–	**0.56**/–	**0.62**/–	**0.44**/–	**0.56**/–
SEA	0.32/–	0.34/–	0.26/–	0.39/–	0.19/–	0.30/–
TabCausal	0.48/0.48	0.37/0.37	0.46/0.46	0.42/0.42	0.34/**0.35**	0.41/0.41

Table 12: **Compact synthetic edge-level noise breakdown.** Cells report Obs./Obs+int edge-F1 from aggregated TP/FP/FN counts.

Method	beta	censored	expo-nential	gaussian	gumbel	heterosk.	laplace	mixture outlier	mult. lognormal	quantized
RandomRegress	0.09/–	0.06/–	0.08/–	0.06/–	0.06/–	0.08/–	0.06/–	0.05/–	0.09/–	0.06/–
CDIS	0.39/0.40	0.34/0.34	0.35/0.35	0.34/0.34	0.34/0.33	0.37/0.37	0.33/0.33	0.32/0.32	0.37/0.37	0.32/0.33
GIES	0.51/0.50	0.49/**0.48**	0.49/0.48	0.48/**0.47**	0.48/**0.47**	0.51/0.50	0.48/0.47	0.46/0.45	0.51/0.50	0.48/**0.47**
IGSP	0.39/0.39	0.34/0.32	0.38/0.37	0.35/0.33	0.34/0.32	0.37/0.35	0.34/0.32	0.32/0.30	0.38/0.38	0.34/0.32
PC	0.28/–	0.23/–	0.27/–	0.24/–	0.24/–	0.26/–	0.23/–	0.22/–	0.27/–	0.23/–
DAS	0.28/–	0.31/–	0.29/–	0.31/–	0.30/–	0.32/–	0.29/–	0.22/–	0.29/–	0.31/–
LiNGAM	0.11/–	0.06/–	0.11/–	0.07/–	0.08/–	0.09/–	0.08/–	0.09/–	0.11/–	0.07/–
DAGMA	0.29/–	0.23/–	0.28/–	0.24/–	0.24/–	0.27/–	0.24/–	0.22/–	0.29/–	0.24/–
NOTEARS	0.09/–	0.06/–	0.09/–	0.07/–	0.06/–	0.08/–	0.06/–	0.05/–	0.09/–	0.06/–
NOTEARS-MLP	0.31/–	0.26/–	0.30/–	0.27/–	0.25/–	0.29/–	0.24/–	0.20/–	0.30/–	0.26/–
SDCD	0.51/0.51	0.38/0.38	0.46/0.46	0.38/0.38	0.38/0.38	0.43/0.43	0.37/0.36	0.31/0.32	0.49/0.49	0.36/0.37
Arrow	0.30/–	0.20/–	0.28/–	0.24/–	0.25/–	0.28/–	0.23/–	0.19/–	0.28/–	0.23/–
AVICI	0.33/0.34	0.21/0.22	0.27/0.29	0.21/0.22	0.20/0.21	0.25/0.26	0.20/0.21	0.26/0.27	0.28/0.30	0.19/0.20
CauScale	0.57/**0.60**	0.41/0.44	0.54/**0.56**	0.44/0.47	0.46/0.47	0.50/**0.52**	0.45/**0.47**	0.50/**0.53**	0.56/**0.58**	0.43/0.46
CDFM	0.25/–	0.22/–	0.24/–	0.21/–	0.22/–	0.22/–	0.22/–	0.24/–	0.24/–	0.21/–
FoundCause	**0.68**/–	**0.55**/–	**0.64**/–	**0.57**/–	**0.57**/–	**0.61**/–	**0.56**/–	**0.60**/–	**0.65**/–	**0.55**/–
SEA	0.42/–	0.33/–	0.39/–	0.33/–	0.34/–	0.38/–	0.32/–	0.30/–	0.41/–	0.32/–
TabCausal	0.54/0.54	0.37/0.37	0.49/0.49	0.43/0.42	0.41/0.40	0.46/0.47	0.39/0.39	0.36/0.40	0.49/0.50	0.39/0.39

Method	RFF heterosk.	student t	target r2 gaussian	uniform
RandomRegress	0.02/–	0.06/–	0.06/–	0.08/–
CDIS	0.25/0.25	0.34/0.35	0.34/0.33	0.35/0.36
GIES	0.37/**0.36**	0.48/0.48	0.48/**0.46**	0.50/0.49
IGSP	0.22/0.20	0.35/0.33	0.36/0.33	0.37/0.36
PC	0.14/–	0.24/–	0.24/–	0.26/–
DAS	0.21/–	0.30/–	0.34/–	0.33/–
LiNGAM	0.03/–	0.08/–	0.07/–	0.10/–
DAGMA	0.11/–	0.24/–	0.26/–	0.28/–
NOTEARS	0.02/–	0.07/–	0.06/–	0.08/–
NOTEARS-MLP	0.13/–	0.26/–	0.28/–	0.30/–
SDCD	0.23/0.21	0.39/0.39	0.36/0.34	0.44/0.44
Arrow	0.11/–	0.24/–	0.24/–	0.29/–
AVICI	0.08/0.08	0.23/0.23	0.17/0.17	0.25/0.27
CauScale	0.25/0.26	0.47/**0.50**	0.42/0.43	0.49/**0.51**
CDFM	0.17/–	0.22/–	0.20/–	0.23/–
FoundCause	**0.40**/–	**0.58**/–	**0.54**/–	**0.62**/–
SEA	0.16/–	0.34/–	0.32/–	0.38/–
TabCausal	0.24/0.23	0.42/0.42	0.40/0.37	0.51/0.50

G.2 SEMANTIC BENCHMARK

Table 13 reports the full Semantic benchmark metrics. Domain-level behavior is analyzed in Section E.3.

Table 13: **Complete Semantic benchmark metrics.** Values are mean±standard deviation over replicate-level rows.

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	RandomRegress	500	0.25 ±0.09	0.20 ±0.09	0.38 ±0.10	93.2 ±69.3	0.183 ±0.124	192.7 ±58.2	0.55 ±0.05	0.14 ±0.05
Obs.	CDIS	500	0.40 ±0.17	0.55 ±0.23	0.32 ±0.15	31.7 ±9.2	0.063 ±0.016	123.5 ±31.2	–	–
Obs.	GIES	500	0.55 ±0.14	0.54 ±0.16	0.57 ±0.14	33.4 ±15.7	0.066 ±0.028	107.9 ±53.4	–	–
Obs.	IGSP	500	0.31 ±0.18	0.34 ±0.22	0.30 ±0.16	40.8 ±14.4	0.081 ±0.025	148.0 ±48.5	–	–
Obs.	PC	500	0.33 ±0.16	0.43 ±0.21	0.27 ±0.14	33.5 ±9.7	0.067 ±0.017	139.5 ±36.6	–	–
Obs.	DAS	500	0.21 ±0.07	0.19 ±0.07	0.24 ±0.08	59.5 ±15.9	0.119 ±0.026	185.4 ±42.9	–	–
Obs.	LiNGAM	500	0.20 ±0.09	0.40 ±0.18	0.14 ±0.06	38.9 ±9.5	0.078 ±0.017	137.2 ±31.7	–	–

Continued on next page

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	DAGMA	500	0.18 ±0.07	0.25 ±0.10	0.15 ±0.06	42.0 ±9.2	0.084 ±0.014	166.8 ±43.1	–	–
Obs.	NOTEARS	500	0.20 ±0.08	0.36 ±0.14	0.14 ±0.06	38.1 ±8.5	0.076 ±0.014	146.1 ±35.4	0.59 ±0.06	0.20 ±0.06
Obs.	NOTEARS-MLP	500	0.22 ±0.09	0.25 ±0.11	0.20 ±0.08	44.9 ±10.9	0.089 ±0.017	179.3 ±46.8	0.69 ±0.06	0.18 ±0.07
Obs.	SDCD	500	0.36 ±0.10	0.32 ±0.10	0.43 ±0.12	51.2 ±16.4	0.102 ±0.028	150.2 ±45.2	–	–
Obs.	Arrow	500	0.31 ±0.10	0.46 ±0.14	0.24 ±0.09	36.6 ±9.6	0.073 ±0.017	134.0 ±33.5	0.69 ±0.08	0.30 ±0.09
Obs.	AVICI	500	0.28 ±0.14	0.54 ±0.20	0.19 ±0.11	35.5 ±9.3	0.071 ±0.016	125.4 ±31.9	0.72 ±0.08	0.32 ±0.12
Obs.	CauScale	500	0.25 ±0.08	0.28 ±0.13	0.24 ±0.08	50.2 ±19.5	0.100 ±0.034	152.3 ±37.0	0.73 ±0.07	0.23 ±0.08
Obs.	CDFM	500	0.43 ±0.08	0.42 ±0.10	0.48 ±0.11	46.0 ±13.1	0.092 ±0.024	122.4 ±38.2	0.82 ±0.06	0.44 ±0.10
Obs.	FoundCause	500	**0.63 ±0.12**	**0.68 ±0.12**	**0.61 ±0.14**	**26.8 ±10.8**	**0.053 ±0.020**	**77.6 ±31.1**	**0.90 ±0.05**	**0.68 ±0.12**
Obs.	SEA	500	0.33 ±0.11	0.41 ±0.17	0.30 ±0.10	41.8 ±18.4	0.083 ±0.032	136.6 ±34.7	0.80 ±0.07	0.34 ±0.13
Obs.	TabCausal	500	0.42 ±0.13	0.50 ±0.15	0.38 ±0.13	36.1 ±10.1	0.073 ±0.021	128.0 ±35.8	0.76 ±0.09	0.41 ±0.14
Obs.+int	CDIS	500	0.40 ±0.18	0.55 ±0.24	0.31 ±0.15	**31.9 ±9.2**	**0.064 ±0.016**	124.2 ±31.7	–	–
Obs.+int	GIES	500	**0.57 ±0.14**	0.56 ±0.16	**0.60 ±0.15**	32.9 ±15.8	0.066 ±0.029	**100.0 ±49.0**	–	–
Obs.+int	IGSP	500	0.31 ±0.18	0.35 ±0.22	0.29 ±0.16	39.9 ±13.4	0.079 ±0.023	147.9 ±47.0	–	–
Obs.+int	SDCD	500	0.36 ±0.10	0.32 ±0.10	0.44 ±0.12	51.7 ±16.3	0.103 ±0.028	150.8 ±46.4	–	–
Obs.+int	AVICI	500	0.30 ±0.13	**0.56 ±0.18**	0.21 ±0.11	34.8 ±9.4	0.070 ±0.016	124.4 ±31.1	0.72 ±0.08	0.34 ±0.12
Obs.+int	CauScale	500	0.29 ±0.09	0.32 ±0.13	0.28 ±0.09	48.6 ±19.2	0.096 ±0.033	148.2 ±38.2	0.74 ±0.07	0.27 ±0.09
Obs.+int	TabCausal	500	0.45 ±0.13	0.53 ±0.14	0.40 ±0.13	35.0 ±10.2	0.070 ±0.021	123.9 ±37.7	**0.77 ±0.09**	**0.45 ±0.14**

G.3 Formula Benchmark

Table 14 reports the full Formula benchmark metrics. Scientific-domain behavior is analyzed in Section E.3.

Table 14: **Complete Formula benchmark metrics.** Values are mean±standard deviation over replicate-level rows.

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	RandomRegress	500	0.20 ±0.08	0.14 ±0.07	0.38 ±0.12	108.3 ±70.1	0.221 ±0.134	187.2 ±60.8	0.56 ±0.05	0.12 ±0.05
Obs.	CDIS	500	0.23 ±0.19	0.35 ±0.28	0.18 ±0.16	31.0 ±6.9	0.064 ±0.015	106.8 ±33.1	–	–
Obs.	GIES	500	0.47 ±0.14	0.42 ±0.15	0.54 ±0.15	36.7 ±14.8	0.076 ±0.031	95.0 ±51.6	–	–
Obs.	IGSP	500	0.21 ±0.16	0.20 ±0.17	0.23 ±0.17	43.3 ±17.3	0.089 ±0.035	137.9 ±49.0	–	–
Obs.	PC	500	0.22 ±0.18	0.29 ±0.23	0.18 ±0.15	31.1 ±7.0	0.064 ±0.015	111.9 ±32.9	–	–
Obs.	DAS	500	0.16 ±0.07	0.13 ±0.06	0.21 ±0.08	64.8 ±16.5	0.133 ±0.034	171.5 ±46.6	–	–
Obs.	LiNGAM	500	0.24 ±0.10	0.34 ±0.15	0.19 ±0.08	34.7 ±8.6	0.071 ±0.017	122.2 ±36.4	–	–
Obs.	DAGMA	500	0.18 ±0.07	0.20 ±0.09	0.16 ±0.07	39.7 ±8.7	0.082 ±0.018	163.5 ±50.4	–	–

Continued on next page

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	NOTEARS	500	0.19 ±0.09	0.26 ±0.12	0.15 ±0.07	35.2 ±7.5	0.073 ±0.017	139.9 ±44.9	0.60 ±0.06	0.16 ±0.06
Obs.	NOTEARS-MLP	500	0.19 ±0.07	0.20 ±0.08	0.19 ±0.08	44.1 ±9.1	0.091 ±0.020	177.4 ±57.9	0.69 ±0.06	0.15 ±0.05
Obs.	SDCD	500	0.28 ±0.09	0.22 ±0.07	0.42 ±0.13	61.5 ±16.3	0.127 ±0.036	141.8 ±50.2	–	–
Obs.	Arrow	500	0.27 ±0.09	0.33 ±0.12	0.24 ±0.09	35.9 ±8.9	0.074 ±0.020	118.4 ±35.4	0.73 ±0.08	0.25 ±0.08
Obs.	AVICI	500	0.33 ±0.12	0.46 ±0.16	0.26 ±0.11	31.8 ±8.4	0.066 ±0.018	102.4 ±32.2	0.78 ±0.07	0.34 ±0.12
Obs.	CauScale	500	0.21 ±0.08	0.17 ±0.08	0.29 ±0.11	65.9 ±24.1	0.134 ±0.043	142.1 ±44.2	0.70 ±0.08	0.17 ±0.07
Obs.	CDFM	500	0.41 ±0.08	0.35 ±0.09	0.53 ±0.11	45.8 ±14.5	0.094 ±0.031	96.0 ±34.6	0.83 ±0.05	0.39 ±0.10
Obs.	FoundCause	500	**0.52 ±0.12**	0.49 ±0.14	**0.59 ±0.14**	33.4 ±12.7	0.069 ±0.024	**70.4 ±36.3**	**0.88 ±0.06**	**0.56 ±0.15**
Obs.	SEA	500	0.24 ±0.10	0.23 ±0.11	0.27 ±0.11	52.8 ±20.6	0.108 ±0.038	123.8 ±39.6	0.75 ±0.06	0.21 ±0.09
Obs.	TabCausal	500	0.47 ±0.13	**0.51 ±0.14**	0.45 ±0.14	**29.5 ±9.5**	**0.061 ±0.021**	93.1 ±36.2	0.82 ±0.08	0.46 ±0.16
Obs.+int	CDIS	500	0.23 ±0.19	0.35 ±0.28	0.18 ±0.16	30.9 ±7.1	0.064 ±0.015	106.1 ±32.7	–	–
Obs.+int	GIES	500	0.52 ±0.14	0.47 ±0.15	**0.60 ±0.15**	34.7 ±14.7	0.072 ±0.031	**78.6 ±47.8**	–	–
Obs.+int	IGSP	500	0.16 ±0.16	0.16 ±0.17	0.17 ±0.17	40.7 ±16.3	0.083 ±0.032	130.7 ±47.2	–	–
Obs.+int	SDCD	500	0.29 ±0.09	0.22 ±0.08	0.43 ±0.13	61.6 ±16.8	0.127 ±0.036	140.5 ±52.0	–	–
Obs.+int	AVICI	500	0.41 ±0.13	0.55 ±0.16	0.34 ±0.12	29.3 ±8.2	0.060 ±0.017	93.1 ±32.9	0.81 ±0.07	0.42 ±0.13
Obs.+int	CauScale	500	0.28 ±0.10	0.24 ±0.11	0.37 ±0.11	59.6 ±23.9	0.121 ±0.042	129.3 ±44.9	0.74 ±0.07	0.26 ±0.11
Obs.+int	TabCausal	500	**0.53 ±0.14**	**0.57 ±0.14**	0.51 ±0.14	**26.6 ±9.2**	**0.055 ±0.020**	82.9 ±36.8	**0.83 ±0.08**	**0.54 ±0.16**

G.4 Real-data Benchmark

Table 15 reports the complete real-data method summary. Table 16 lists every method–dataset pair, including observation-only and observation-plus-intervention splits. The six observation-only sources are the CD-CSG datasets Abalone, Auto MPG, Cardiac Arrhythmia, Concrete Compressive Strength, Deutscher Wetterdienst, and Ozone. The seven interventional sources are Sachs flow-cytometry; PetShop high traffic, low traffic, temporal traffic 1, and temporal traffic 2; and the Causal Chambers Light Tunnel and Wind Tunnel experiments.

Table 15: **Complete real-data method metrics.** Obs.-only includes the six native observation-only datasets plus observation-only conversions of the seven interventional datasets.

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	RandomRegress	13	0.20 ±0.14	0.15 ±0.11	0.31 ±0.21	59.8 ±55.9	0.208 ±0.148	144.2 ±186.5	–	–
Obs.	CDIS	13	0.28 ±0.22	0.24 ±0.20	0.40 ±0.23	99.5 ±133.7	0.204 ±0.118	179.8 ±222.8	–	–
Obs.	GIES	13	0.26 ±0.15	0.21 ±0.15	0.39 ±0.22	48.0 ±43.5	0.201 ±0.153	164.5 ±210.2	–	–
Obs.	IGSP	13	0.18 ±0.22	0.15 ±0.21	0.23 ±0.27	31.1 ±23.5	0.175 ±0.165	65.9 ±71.8	–	–
Obs.	PC	13	0.18 ±0.14	0.23 ±0.17	0.18 ±0.15	28.8 ±21.0	0.183 ±0.159	63.8 ±55.2	–	–
Obs.	DAS	13	0.19 ±0.23	0.14 ±0.17	0.30 ±0.36	61.9 ±55.1	0.207 ±0.141	130.7 ±126.7	–	–
Obs.	LiNGAM	13	0.13 ±0.17	0.24 ±0.30	0.09 ±0.12	**24.3 ±17.8**	0.164 ±0.151	**50.5 ±49.0**	–	–

Continued on next page

Split	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Obs.	DAGMA	13	0.16 ±0.14	0.17 ±0.14	0.18 ±0.19	31.8 ±27.1	0.197 ±0.184	62.4 ±54.5	–	–
Obs.	NOTEARS	13	0.16 ±0.15	0.22 ±0.20	0.17 ±0.19	29.1 ±24.0	0.194 ±0.186	57.6 ±49.5	–	–
Obs.	NOTEARS-MLP	13	0.21 ±0.19	0.20 ±0.17	0.29 ±0.30	55.0 ±106.0	0.185 ±0.147	146.8 ±300.8	–	–
Obs.	SDCD	13	0.22 ±0.17	0.21 ±0.18	0.28 ±0.22	44.5 ±38.7	0.202 ±0.161	89.8 ±95.7	–	–
Obs.	Arrow	13	0.15 ±0.13	0.13 ±0.12	0.22 ±0.21	70.5 ±102.6	0.232 ±0.152	133.9 ±175.7	0.65 ±0.16	0.20 ±0.16
Obs.	AVICI	13	0.22 ±0.18	0.27 ±0.18	0.21 ±0.23	26.0 ±20.7	0.143 ±0.144	55.4 ±52.4	0.68 ±0.21	0.29 ±0.21
Obs.	CauScale	13	0.08 ±0.08	0.17 ±0.28	0.08 ±0.09	36.2 ±34.0	**0.134 ±0.098**	82.9 ±100.1	0.64 ±0.19	0.23 ±0.16
Obs.	CDFM	13	**0.39 ±0.14**	0.33 ±0.20	**0.64 ±0.27**	38.5 ±34.5	0.189 ±0.172	72.9 ±77.6	**0.78 ±0.15**	**0.40 ±0.25**
Obs.	FoundCause	13	0.21 ±0.23	0.21 ±0.28	0.31 ±0.22	98.3 ±147.0	0.181 ±0.114	239.8 ±295.2	0.76 ±0.13	0.37 ±0.22
Obs.	SEA	13	0.21 ±0.14	0.28 ±0.26	0.22 ±0.16	31.3 ±26.0	0.158 ±0.136	70.9 ±71.6	0.68 ±0.20	0.28 ±0.16
Obs.	TabCausal	13	0.34 ±0.24	**0.33 ±0.24**	0.40 ±0.30	28.4 ±19.3	0.182 ±0.180	61.8 ±60.3	0.70 ±0.17	0.38 ±0.24
Obs.+int	CDIS	7	0.14 ±0.17	0.13 ±0.16	0.15 ±0.19	45.4 ±10.5	0.034 ±0.007	94.9 ±50.5	–	–
Obs.+int	GIES	7	0.29 ±0.13	0.20 ±0.11	**0.54 ±0.10**	118.7 ±53.8	0.097 ±0.053	403.1 ±215.0	–	–
Obs.+int	IGSP	7	0.11 ±0.12	0.09 ±0.10	0.15 ±0.17	59.1 ±24.3	0.043 ±0.013	157.6 ±130.0	–	–
Obs.+int	SDCD	7	0.27 ±0.12	0.22 ±0.13	0.43 ±0.08	103.4 ±55.0	0.081 ±0.030	252.4 ±144.6	–	–
Obs.+int	AVICI	7	0.28 ±0.12	0.42 ±0.20	0.23 ±0.12	42.6 ±12.1	0.033 ±0.005	88.1 ±49.0	0.72 ±0.07	0.26 ±0.16
Obs.+int	CauScale	7	0.19 ±0.07	0.20 ±0.07	0.20 ±0.08	62.1 ±17.2	0.048 ±0.010	135.1 ±81.4	**0.74 ±0.05**	0.20 ±0.10
Obs.+int	TabCausal	7	**0.46 ±0.11**	**0.50 ±0.16**	0.45 ±0.09	**38.4 ±9.9**	**0.029 ±0.008**	**87.6 ±46.9**	0.74 ±0.08	**0.41 ±0.13**

Table 16: **Complete real-data dataset-level metrics.** Each row is one method on one scored real dataset.

Dataset	Split	Method	F1	Precision	Recall	SHD	SID	nSHD
Abalone	Obs.	RandomRegress	0.00	0.00	0.00	24.0	34.0	0.429
Abalone	Obs.	CDIS	0.27	0.20	0.43	16.0	26.0	0.286
Abalone	Obs.	GIES	0.36	0.24	0.71	18.0	20.0	0.321
Abalone	Obs.	IGSP	0.23	0.16	0.43	19.0	24.0	0.339
Abalone	Obs.	PC	0.09	0.07	0.14	18.0	27.0	0.321
Abalone	Obs.	DAS	0.00	0.00	0.00	23.0	30.0	0.411
Abalone	Obs.	LiNGAM	0.00	0.00	0.00	11.0	14.0	0.196
Abalone	Obs.	DAGMA	0.00	0.00	0.00	8.0	11.0	0.143
Abalone	Obs.	NOTEARS	0.00	0.00	0.00	**7.0**	**8.0**	**0.125**
Abalone	Obs.	NOTEARS-MLP	0.00	0.00	0.00	9.0	15.0	0.161
Abalone	Obs.	SDCD	0.00	0.00	0.00	18.0	31.0	0.321
Abalone	Obs.	Arrow	0.00	0.00	0.00	20.0	28.0	0.357
Abalone	Obs.	AVICI	**0.53**	**0.42**	0.71	9.0	12.0	0.161
Abalone	Obs.	CauScale	0.00	0.00	0.00	12.0	18.0	0.214
Abalone	Obs.	CDFM	0.50	0.33	**1.00**	11.0	24.0	0.196
Abalone	Obs.	FoundCause	0.00	0.00	0.00	11.0	11.0	0.196
Abalone	Obs.	SEA	0.00	0.00	0.00	15.0	23.0	0.268
Abalone	Obs.	TabCausal	0.07	0.05	0.14	22.0	29.0	0.393
Auto MPG	Obs.	RandomRegress	0.25	0.20	0.33	4.0	6.0	0.333
Auto MPG	Obs.	CDIS	0.57	**0.50**	0.67	**3.0**	**2.0**	**0.250**
Auto MPG	Obs.	GIES	0.50	0.40	0.67	4.0	4.0	0.333
Auto MPG	Obs.	IGSP	0.29	0.25	0.33	4.0	5.0	0.333
Auto MPG	Obs.	PC	0.00	0.00	0.00	5.0	7.0	0.417
Auto MPG	Obs.	DAS	0.00	0.00	0.00	5.0	9.0	0.417

Continued on next page

Dataset	Split	Method	F1	Precision	Recall	SHD	SID	nSHD
Auto MPG	Obs.	LiNGAM	0.00	0.00	0.00	4.0	6.0	0.333
Auto MPG	Obs.	DAGMA	0.00	0.00	0.00	6.0	9.0	0.500
Auto MPG	Obs.	NOTEARS	0.00	0.00	0.00	6.0	9.0	0.500
Auto MPG	Obs.	NOTEARS-MLP	**0.67**	**0.50**	**1.00**	**3.0**	3.0	**0.250**
Auto MPG	Obs.	SDCD	0.44	0.33	0.67	4.0	5.0	0.333
Auto MPG	Obs.	Arrow	0.44	0.33	0.67	4.0	5.0	0.333
Auto MPG	Obs.	AVICI	0.00	0.00	0.00	6.0	9.0	0.500
Auto MPG	Obs.	CauScale	0.00	0.00	0.00	**3.0**	3.0	**0.250**
Auto MPG	Obs.	CDFM	0.33	0.22	0.67	6.0	8.0	0.500
Auto MPG	Obs.	FoundCause	0.00	0.00	0.00	4.0	5.0	0.333
Auto MPG	Obs.	SEA	0.40	**0.50**	0.33	**3.0**	3.0	**0.250**
Auto MPG	Obs.	TabCausal	0.22	0.17	0.33	5.0	8.0	0.417
Arrhythmia	Obs.	RandomRegress	0.50	0.40	0.67	3.0	5.0	0.250
Arrhythmia	Obs.	CDIS	0.33	0.33	0.33	3.0	5.0	0.250
Arrhythmia	Obs.	GIES	0.00	0.00	0.00	5.0	9.0	0.417
Arrhythmia	Obs.	IGSP	0.00	0.00	0.00	5.0	7.0	0.417
Arrhythmia	Obs.	PC	0.29	0.25	0.33	4.0	7.0	0.333
Arrhythmia	Obs.	DAS	0.67	0.50	**1.00**	3.0	3.0	0.250
Arrhythmia	Obs.	LiNGAM	0.00	0.00	0.00	4.0	4.0	0.333
Arrhythmia	Obs.	DAGMA	0.00	0.00	0.00	4.0	6.0	0.333
Arrhythmia	Obs.	NOTEARS	0.00	0.00	0.00	4.0	6.0	0.333
Arrhythmia	Obs.	NOTEARS-MLP	0.00	0.00	0.00	3.0	3.0	0.250
Arrhythmia	Obs.	SDCD	0.57	0.50	0.67	3.0	4.0	0.250
Arrhythmia	Obs.	Arrow	0.00	0.00	0.00	4.0	4.0	0.333
Arrhythmia	Obs.	AVICI	0.40	0.50	0.33	3.0	2.0	0.250
Arrhythmia	Obs.	CauScale	0.00	0.00	0.00	3.0	3.0	0.250
Arrhythmia	Obs.	CDFM	0.75	0.60	**1.00**	2.0	2.0	0.167
Arrhythmia	Obs.	FoundCause	0.40	0.50	0.33	3.0	3.0	0.250
Arrhythmia	Obs.	SEA	0.00	0.00	0.00	4.0	5.0	0.333
Arrhythmia	Obs.	TabCausal	**0.86**	**0.75**	**1.00**	**1.0**	**1.0**	**0.083**
Concrete	Obs.	RandomRegress	0.22	0.14	0.50	26.0	35.0	0.361
Concrete	Obs.	CDIS	0.32	**0.24**	0.50	17.0	27.0	0.236
Concrete	Obs.	GIES	0.25	0.17	0.50	22.0	28.0	0.306
Concrete	Obs.	IGSP	**0.33**	0.23	**0.62**	20.0	22.0	0.278
Concrete	Obs.	PC	0.00	0.00	0.00	21.0	35.0	0.292
Concrete	Obs.	DAS	0.16	0.10	0.38	26.0	32.0	0.361
Concrete	Obs.	LiNGAM	0.00	0.00	0.00	14.0	20.0	0.194
Concrete	Obs.	DAGMA	0.06	0.04	0.12	26.0	32.0	0.361
Concrete	Obs.	NOTEARS	0.06	0.04	0.12	26.0	33.0	0.361
Concrete	Obs.	NOTEARS-MLP	0.06	0.04	0.12	27.0	33.0	0.375
Concrete	Obs.	SDCD	0.00	0.00	0.00	33.0	39.0	0.458
Concrete	Obs.	Arrow	0.15	0.11	0.25	22.0	32.0	0.306
Concrete	Obs.	AVICI	0.00	0.00	0.00	14.0	22.0	0.194
Concrete	Obs.	CauScale	0.00	0.00	0.00	**8.0**	**8.0**	**0.111**
Concrete	Obs.	CDFM	0.22	0.13	**0.62**	26.0	56.0	0.361
Concrete	Obs.	FoundCause	0.23	0.17	0.38	20.0	24.0	0.278
Concrete	Obs.	SEA	0.00	0.00	0.00	**8.0**	**8.0**	**0.111**
Concrete	Obs.	TabCausal	0.09	0.06	0.25	33.0	43.0	0.458
DWD	Obs.	RandomRegress	0.36	0.29	0.50	5.0	8.0	0.250
DWD	Obs.	CDIS	0.80	0.67	**1.00**	2.0	2.0	0.100
DWD	Obs.	GIES	0.36	0.29	0.50	5.0	8.0	0.250
DWD	Obs.	IGSP	0.75	0.75	0.75	2.0	3.0	0.100
DWD	Obs.	PC	0.00	0.00	0.00	6.0	10.0	0.300
DWD	Obs.	DAS	0.57	0.40	**1.00**	6.0	6.0	0.300
DWD	Obs.	LiNGAM	0.00	0.00	0.00	8.0	9.0	0.400
DWD	Obs.	DAGMA	0.15	0.11	0.25	9.0	11.0	0.450
DWD	Obs.	NOTEARS	0.15	0.11	0.25	9.0	11.0	0.450
DWD	Obs.	NOTEARS-MLP	0.29	0.20	0.50	8.0	11.0	0.400
DWD	Obs.	SDCD	0.00	0.00	0.00	8.0	11.0	0.400
DWD	Obs.	Arrow	0.29	0.20	0.50	8.0	10.0	0.400
DWD	Obs.	AVICI	0.20	0.17	0.25	5.0	11.0	0.250
DWD	Obs.	CauScale	0.00	0.00	0.00	5.0	6.0	0.250
DWD	Obs.	CDFM	0.40	0.25	**1.00**	7.0	12.0	0.350
DWD	Obs.	FoundCause	**0.86**	**1.00**	0.75	**1.0**	**1.0**	**0.050**
DWD	Obs.	SEA	0.31	0.22	0.50	7.0	9.0	0.350
DWD	Obs.	TabCausal	0.57	0.40	**1.00**	6.0	7.0	0.300
Ozone	Obs.	RandomRegress	0.00	0.00	0.00	5.0	7.0	0.417
Ozone	Obs.	CDIS	0.00	0.00	0.00	5.0	6.0	0.417
Ozone	Obs.	GIES	0.00	0.00	0.00	5.0	7.0	0.417
Ozone	Obs.	IGSP	0.00	0.00	0.00	5.0	8.0	0.417
Ozone	Obs.	PC	0.29	0.25	0.33	4.0	5.0	0.333
Ozone	Obs.	DAS	0.50	0.40	0.67	3.0	5.0	0.250
Ozone	Obs.	LiNGAM	0.00	0.00	0.00	4.0	4.0	0.333
Ozone	Obs.	DAGMA	0.44	0.33	0.67	4.0	5.0	0.333

Continued on next page

Dataset	Split	Method	F1	Precision	Recall	SHD	SID	nSHD
Ozone	Obs.	NOTEARS	0.44	0.33	0.67	4.0	5.0	0.333
Ozone	Obs.	NOTEARS-MLP	0.44	0.33	0.67	4.0	5.0	0.333
Ozone	Obs.	SDCD	0.25	0.20	0.33	4.0	7.0	0.333
Ozone	Obs.	Arrow	0.00	0.00	0.00	5.0	8.0	0.417
Ozone	Obs.	AVICI	**0.57**	**0.50**	0.67	**2.0**	4.0	**0.167**
Ozone	Obs.	CauScale	0.00	0.00	0.00	3.0	**3.0**	0.250
Ozone	Obs.	CDFM	0.50	0.33	**1.00**	5.0	9.0	0.417
Ozone	Obs.	FoundCause	0.00	0.00	0.00	3.0	4.0	0.250
Ozone	Obs.	SEA	0.33	0.33	0.33	4.0	4.0	0.333
Ozone	Obs.	TabCausal	0.00	0.00	0.00	4.0	6.0	0.333
Sachs	Obs.	RandomRegress	0.28	0.20	**0.45**	39.0	40.0	–
Sachs	Obs.	CDIS	0.26	0.26	0.25	26.0	34.0	–
Sachs	Obs.	GIES	0.18	0.14	0.25	37.0	78.0	–
Sachs	Obs.	IGSP	0.31	0.24	**0.45**	34.0	54.0	–
Sachs	Obs.	PC	0.32	0.33	0.30	24.0	77.0	–
Sachs	Obs.	DAS	0.14	0.11	0.20	38.0	64.0	–
Sachs	Obs.	LiNGAM	0.30	**0.57**	0.20	**18.0**	40.0	–
Sachs	Obs.	DAGMA	0.35	0.43	0.30	19.0	45.0	–
Sachs	Obs.	NOTEARS	0.34	0.56	0.25	**18.0**	44.0	–
Sachs	Obs.	NOTEARS-MLP	0.29	0.27	0.30	24.0	69.0	–
Sachs	Obs.	SDCD	0.33	0.32	0.35	24.0	62.0	–
Sachs	Obs.	Arrow	0.29	0.29	0.30	26.0	44.0	–
Sachs	Obs.	AVICI	0.22	0.43	0.15	19.0	44.0	–
Sachs	Obs.	CauScale	0.13	0.20	0.10	21.0	47.0	–
Sachs	Obs.	CDFM	0.33	0.28	0.40	25.0	48.0	–
Sachs	Obs.	FoundCause	0.26	0.23	0.30	28.0	56.0	–
Sachs	Obs.	SEA	0.32	0.35	0.30	22.0	59.0	–
Sachs	Obs.	TabCausal	**0.37**	0.35	0.40	22.0	**32.0**	–
Sachs	Obs.+int	CDIS	0.26	0.26	0.25	26.0	**34.0**	–
Sachs	Obs.+int	GIES	0.37	0.27	**0.60**	35.0	**34.0**	–
Sachs	Obs.+int	IGSP	0.31	0.24	0.45	35.0	62.0	–
Sachs	Obs.+int	SDCD	0.34	0.33	0.35	24.0	42.0	–
Sachs	Obs.+int	AVICI	**0.41**	**0.50**	0.35	**16.0**	**34.0**	–
Sachs	Obs.+int	CauScale	0.31	0.32	0.30	24.0	**34.0**	–
Sachs	Obs.+int	TabCausal	0.38	0.36	0.40	21.0	51.0	–
PetShop HT	Obs.	RandomRegress	0.29	0.21	0.48	89.0	208.0	0.060
PetShop HT	Obs.	CDIS	0.35	0.33	0.38	51.0	152.0	0.034
PetShop HT	Obs.	GIES	0.29	0.22	0.40	74.0	361.0	0.050
PetShop HT	Obs.	IGSP	0.18	0.15	0.21	70.0	200.0	0.047
PetShop HT	Obs.	PC	0.35	0.33	0.38	51.0	150.0	0.034
PetShop HT	Obs.	DAS	0.10	0.07	0.14	94.0	286.0	0.063
PetShop HT	Obs.	LiNGAM	0.30	**0.67**	0.19	**36.0**	**99.0**	**0.024**
PetShop HT	Obs.	DAGMA	0.13	0.19	0.10	43.0	127.0	0.029
PetShop HT	Obs.	NOTEARS	0.11	0.21	0.07	41.0	105.0	0.028
PetShop HT	Obs.	NOTEARS-MLP	0.26	0.29	0.24	49.0	170.0	0.033
PetShop HT	Obs.	SDCD	0.14	0.11	0.21	90.0	190.0	0.061
PetShop HT	Obs.	Arrow	0.15	0.16	0.14	58.0	147.0	0.039
PetShop HT	Obs.	AVICI	0.17	0.28	0.12	44.0	105.0	0.030
PetShop HT	Obs.	CauScale	0.20	0.16	0.26	76.0	199.0	0.051
PetShop HT	Obs.	CDFM	0.35	0.29	0.45	62.0	**99.0**	0.042
PetShop HT	Obs.	FoundCause	0.18	0.11	0.45	163.0	637.0	0.110
PetShop HT	Obs.	SEA	0.26	0.29	0.24	49.0	132.0	0.033
PetShop HT	Obs.	TabCausal	**0.47**	0.43	**0.52**	45.0	122.0	0.030
PetShop HT	Obs.+int	CDIS	0.34	0.31	0.38	53.0	159.0	0.036
PetShop HT	Obs.+int	GIES	0.19	0.13	0.43	142.0	528.0	0.096
PetShop HT	Obs.+int	IGSP	0.16	0.13	0.19	70.0	211.0	0.047
PetShop HT	Obs.+int	SDCD	0.17	0.11	0.43	162.0	398.0	0.109
PetShop HT	Obs.+int	AVICI	0.40	0.39	0.40	47.0	103.0	0.032
PetShop HT	Obs.+int	CauScale	0.25	0.23	0.29	66.0	139.0	0.045
PetShop HT	Obs.+int	TabCausal	**0.51**	**0.50**	**0.52**	**40.0**	**84.0**	**0.027**
PetShop LT	Obs.	RandomRegress	0.23	0.16	0.37	97.0	230.0	0.059
PetShop LT	Obs.	CDIS	**0.35**	0.33	0.37	54.0	144.0	0.033
PetShop LT	Obs.	GIES	0.30	0.23	0.44	78.0	344.0	0.048
PetShop LT	Obs.	IGSP	0.20	0.19	0.21	60.0	209.0	0.037
PetShop LT	Obs.	PC	0.33	0.31	0.35	54.0	148.0	0.033
PetShop LT	Obs.	DAS	0.13	0.10	0.19	96.0	271.0	0.059
PetShop LT	Obs.	LiNGAM	0.13	0.24	0.09	44.0	**128.0**	0.027
PetShop LT	Obs.	DAGMA	0.12	0.19	0.09	46.0	135.0	0.028
PetShop LT	Obs.	NOTEARS	0.07	0.13	0.05	45.0	129.0	0.027
PetShop LT	Obs.	NOTEARS-MLP	0.11	0.15	0.09	49.0	151.0	0.030
PetShop LT	Obs.	SDCD	0.22	0.17	0.30	80.0	182.0	0.049
PetShop LT	Obs.	Arrow	0.23	0.29	0.19	47.0	140.0	0.029
PetShop LT	Obs.	AVICI	0.23	**0.39**	0.16	**42.0**	145.0	**0.026**
PetShop LT	Obs.	CauScale	0.15	0.13	0.19	77.0	214.0	0.047

Continued on next page

Dataset	Split	Method	F1	Precision	Recall	SHD	SID	nSHD
PetShop LT	Obs.	CDFM	0.34	0.24	**0.58**	82.0	228.0	0.050
PetShop LT	Obs.	FoundCause	0.13	0.08	0.28	146.0	517.0	0.089
PetShop LT	Obs.	SEA	0.25	0.22	0.28	66.0	174.0	0.040
PetShop LT	Obs.	TabCausal	0.23	0.25	0.21	51.0	144.0	0.031
PetShop LT	Obs.+int	CDIS	0.36	0.33	0.40	54.0	**146.0**	0.033
PetShop LT	Obs.+int	GIES	0.30	0.21	**0.53**	102.0	462.0	0.062
PetShop LT	Obs.+int	IGSP	0.20	0.19	0.21	60.0	214.0	0.037
PetShop LT	Obs.+int	SDCD	0.16	0.10	0.44	183.0	450.0	0.112
PetShop LT	Obs.+int	AVICI	0.21	0.23	0.19	53.0	153.0	0.032
PetShop LT	Obs.+int	CauScale	0.18	0.17	0.19	65.0	188.0	0.040
PetShop LT	Obs.+int	TabCausal	**0.40**	**0.37**	0.44	**50.0**	159.0	**0.030**
PetShop T1	Obs.	RandomRegress	0.15	0.09	0.33	144.0	458.0	0.088
PetShop T1	Obs.	CDIS	0.10	0.06	0.30	210.0	538.0	0.128
PetShop T1	Obs.	GIES	0.23	0.15	**0.47**	122.0	552.0	0.074
PetShop T1	Obs.	IGSP	0.00	0.00	0.00	43.0	**113.0**	0.026
PetShop T1	Obs.	PC	0.28	0.43	0.21	42.0	115.0	0.026
PetShop T1	Obs.	DAS	0.05	0.04	0.09	125.0	289.0	0.076
PetShop T1	Obs.	LiNGAM	0.07	0.18	0.05	44.0	121.0	0.027
PetShop T1	Obs.	DAGMA	0.13	0.21	0.09	44.0	131.0	0.027
PetShop T1	Obs.	NOTEARS	0.15	**0.44**	0.09	**40.0**	114.0	**0.024**
PetShop T1	Obs.	NOTEARS-MLP	0.19	0.29	0.14	43.0	138.0	0.026
PetShop T1	Obs.	SDCD	0.21	0.15	0.35	98.0	251.0	0.060
PetShop T1	Obs.	Arrow	0.10	0.09	0.12	74.0	207.0	0.045
PetShop T1	Obs.	AVICI	0.19	0.29	0.14	44.0	125.0	0.027
PetShop T1	Obs.	CauScale	0.12	0.10	0.16	82.0	240.0	0.050
PetShop T1	Obs.	CDFM	0.26	0.19	0.44	94.0	186.0	0.057
PetShop T1	Obs.	FoundCause	0.17	0.11	0.42	160.0	557.0	0.098
PetShop T1	Obs.	SEA	0.27	0.25	0.30	57.0	192.0	0.035
PetShop T1	Obs.	TabCausal	**0.41**	0.41	0.42	44.0	156.0	0.027
PetShop T1	Obs.+int	CDIS	0.00	0.00	0.00	43.0	**113.0**	0.026
PetShop T1	Obs.+int	GIES	0.21	0.13	**0.47**	145.0	510.0	0.088
PetShop T1	Obs.+int	IGSP	0.00	0.00	0.00	43.0	**113.0**	0.026
PetShop T1	Obs.+int	SDCD	0.23	0.16	0.37	96.0	279.0	0.059
PetShop T1	Obs.+int	AVICI	0.18	0.26	0.14	47.0	118.0	0.029
PetShop T1	Obs.+int	CauScale	0.16	0.15	0.16	67.0	156.0	0.041
PetShop T1	Obs.+int	TabCausal	**0.47**	**0.48**	**0.47**	**42.0**	119.0	**0.026**
PetShop T2	Obs.	RandomRegress	0.09	0.05	0.21	166.0	587.0	0.101
PetShop T2	Obs.	CDIS	0.13	0.08	0.37	207.0	593.0	0.126
PetShop T2	Obs.	GIES	0.24	0.16	**0.49**	121.0	552.0	0.074
PetShop T2	Obs.	IGSP	0.00	0.00	0.00	43.0	**113.0**	0.026
PetShop T2	Obs.	PC	0.24	0.35	0.19	43.0	124.0	0.026
PetShop T2	Obs.	DAS	0.05	0.04	0.09	125.0	289.0	0.076
PetShop T2	Obs.	LiNGAM	0.07	0.18	0.05	44.0	121.0	0.027
PetShop T2	Obs.	DAGMA	0.13	0.21	0.09	44.0	131.0	0.027
PetShop T2	Obs.	NOTEARS	0.15	**0.44**	0.09	**40.0**	114.0	**0.024**
PetShop T2	Obs.	NOTEARS-MLP	0.28	0.41	0.21	41.0	125.0	0.025
PetShop T2	Obs.	SDCD	0.20	0.14	0.35	107.0	261.0	0.065
PetShop T2	Obs.	Arrow	0.10	0.09	0.12	74.0	207.0	0.045
PetShop T2	Obs.	AVICI	0.19	0.29	0.14	44.0	125.0	0.027
PetShop T2	Obs.	CauScale	0.12	0.10	0.16	83.0	244.0	0.051
PetShop T2	Obs.	CDFM	0.26	0.19	0.44	94.0	186.0	0.057
PetShop T2	Obs.	FoundCause	0.17	0.10	0.42	165.0	542.0	0.101
PetShop T2	Obs.	SEA	0.26	0.26	0.26	51.0	159.0	0.031
PetShop T2	Obs.	TabCausal	**0.41**	0.41	0.42	44.0	156.0	0.027
PetShop T2	Obs.+int	CDIS	0.00	0.00	0.00	**43.0**	**113.0**	**0.026**
PetShop T2	Obs.+int	GIES	0.21	0.14	**0.47**	139.0	665.0	0.085
PetShop T2	Obs.+int	IGSP	0.11	0.08	0.19	107.0	404.0	0.065
PetShop T2	Obs.+int	SDCD	0.21	0.15	0.35	100.0	277.0	0.061
PetShop T2	Obs.+int	AVICI	0.18	0.25	0.14	47.0	128.0	0.029
PetShop T2	Obs.+int	CauScale	0.21	0.17	0.26	72.0	274.0	0.044
PetShop T2	Obs.+int	TabCausal	**0.44**	**0.43**	0.44	44.0	121.0	0.027
Light Tunnel	Obs.	RandomRegress	0.14	0.14	0.14	83.0	122.0	0.059
Light Tunnel	Obs.	CDIS	0.08	0.04	0.30	400.0	432.0	0.284
Light Tunnel	Obs.	GIES	0.45	0.55	0.39	52.0	62.0	0.037
Light Tunnel	Obs.	IGSP	0.00	0.00	0.00	57.0	57.0	0.041
Light Tunnel	Obs.	PC	0.06	0.18	0.04	63.0	81.0	0.045
Light Tunnel	Obs.	DAS	0.07	0.06	0.11	140.0	197.0	0.100
Light Tunnel	Obs.	LiNGAM	0.48	0.86	0.33	41.0	**39.0**	0.029
Light Tunnel	Obs.	DAGMA	0.31	0.28	0.33	84.0	85.0	0.060
Light Tunnel	Obs.	NOTEARS	0.40	0.47	0.35	57.0	64.0	0.041
Light Tunnel	Obs.	NOTEARS-MLP	0.13	0.07	**0.54**	402.0	1127.0	0.286
Light Tunnel	Obs.	SDCD	0.35	0.54	0.26	49.0	62.0	0.035
Light Tunnel	Obs.	Arrow	0.12	0.07	0.44	349.0	627.0	0.248
Light Tunnel	Obs.	AVICI	0.00	0.00	0.00	57.0	63.0	0.041

Continued on next page

Dataset	Split	Method	F1	Precision	Recall	SHD	SID	nSHD
Light Tunnel	Obs.	CauScale	0.22	0.47	0.14	56.0	52.0	0.040
Light Tunnel	Obs.	CDFM	0.46	0.86	0.32	41.0	42.0	0.029
Light Tunnel	Obs.	FoundCause	0.08	0.05	0.44	530.0	714.0	0.377
Light Tunnel	Obs.	SEA	0.07	**1.00**	0.04	55.0	55.0	0.039
Light Tunnel	Obs.	TabCausal	**0.50**	0.78	0.37	**40.0**	41.0	**0.028**
Light Tunnel	Obs.+int	CDIS	0.00	0.00	0.00	57.0	57.0	0.041
Light Tunnel	Obs.+int	GIES	0.53	0.42	**0.72**	71.0	199.0	0.050
Light Tunnel	Obs.+int	IGSP	0.00	0.00	0.00	57.0	57.0	0.041
Light Tunnel	Obs.+int	SDCD	0.51	0.46	0.56	59.0	123.0	0.042
Light Tunnel	Obs.+int	AVICI	0.39	0.79	0.26	46.0	40.0	0.033
Light Tunnel	Obs.+int	CauScale	0.18	0.24	0.14	75.0	90.0	0.053
Light Tunnel	Obs.+int	TabCausal	**0.67**	**0.84**	0.56	**29.0**	**37.0**	**0.021**
Wind Tunnel	Obs.	RandomRegress	0.09	0.07	0.12	93.0	134.0	0.094
Wind Tunnel	Obs.	CDIS	0.07	0.04	0.31	299.0	376.0	0.301
Wind Tunnel	Obs.	GIES	0.21	0.17	0.26	81.0	114.0	0.082
Wind Tunnel	Obs.	IGSP	0.00	0.00	0.00	42.0	42.0	0.042
Wind Tunnel	Obs.	PC	0.16	0.50	0.10	**40.0**	44.0	**0.040**
Wind Tunnel	Obs.	DAS	0.00	0.00	0.00	121.0	218.0	0.122
Wind Tunnel	Obs.	LiNGAM	0.35	0.46	0.29	44.0	51.0	0.044
Wind Tunnel	Obs.	DAGMA	0.27	0.22	**0.33**	76.0	83.0	0.077
Wind Tunnel	Obs.	NOTEARS	0.17	0.14	0.21	81.0	107.0	0.082
Wind Tunnel	Obs.	NOTEARS-MLP	0.00	0.00	0.00	53.0	58.0	0.053
Wind Tunnel	Obs.	SDCD	0.19	0.21	0.17	60.0	62.0	0.060
Wind Tunnel	Obs.	Arrow	0.06	0.03	0.17	226.0	282.0	0.228
Wind Tunnel	Obs.	AVICI	0.17	0.28	0.12	49.0	53.0	0.049
Wind Tunnel	Obs.	CauScale	0.05	**1.00**	0.02	41.0	**41.0**	0.041
Wind Tunnel	Obs.	CDFM	**0.37**	0.42	**0.33**	45.0	48.0	0.045
Wind Tunnel	Obs.	FoundCause	0.29	0.43	0.21	44.0	46.0	0.044
Wind Tunnel	Obs.	SEA	0.23	0.23	0.24	66.0	99.0	0.067
Wind Tunnel	Obs.	TabCausal	0.23	0.29	0.19	52.0	58.0	0.052
Wind Tunnel	Obs.+int	CDIS	0.00	0.00	0.00	**42.0**	42.0	**0.042**
Wind Tunnel	Obs.+int	GIES	0.20	0.12	**0.60**	197.0	424.0	0.199
Wind Tunnel	Obs.+int	IGSP	0.00	0.00	0.00	**42.0**	42.0	**0.042**
Wind Tunnel	Obs.+int	SDCD	0.28	0.20	0.48	100.0	198.0	0.101
Wind Tunnel	Obs.+int	AVICI	0.16	**0.50**	0.10	**42.0**	**41.0**	**0.042**
Wind Tunnel	Obs.+int	CauScale	0.08	0.10	0.07	66.0	65.0	0.067
Wind Tunnel	Obs.+int	TabCausal	**0.36**	0.48	0.29	43.0	42.0	0.043

G.5 SAMPLE-SIZE SCALING

Table 17 reports the numerical F1/SHD values for the observation-only sample-size scaling suite, which fixes $d = 30$ and varies the number of samples.

Table 17: **Complete sample-size sensitivity results.** Cells report F1/SHD for the observation-only $d = 30$ synthetic suite.

Method	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
RandomRegress	0.25/76.6	0.28/76.8	0.28/84.1	0.29/88.8	0.28/99.2	0.27/113.0	0.26/125.7
CDIS	0.23/51.1	0.29/49.4	0.36/48.4	0.39/50.8	0.40/54.2	0.42/60.0	0.43/61.7
GIES	0.35/61.8	0.41/56.2	0.46/54.2	0.49/54.6	0.50/57.9	0.49/65.3	0.49/70.6
IGSP	0.25/52.3	0.31/51.8	0.38/52.2	0.41/54.3	0.43/58.9	0.43/68.8	0.43/76.8
PC	0.25/50.6	0.27/50.2	0.30/50.8	0.31/52.5	0.32/54.3	0.31/57.9	0.31/60.1
DAS	0.22/61.5	0.28/61.5	0.32/65.4	0.35/67.5	0.36/71.0	0.35/104.7	0.36/87.1
LiNGAM	0.20/51.3	0.20/50.5	0.21/49.7	0.22/49.5	0.21/49.4	0.21/49.5	0.21/49.5
DAGMA	0.27/53.0	0.29/50.1	0.29/49.6	0.29/49.1	0.29/48.8	0.29/49.0	0.29/49.0
NOTEARS	0.26/49.2	0.28/48.2	0.28/47.7	0.28/47.4	0.28/47.5	0.28/47.3	0.27/47.5
NOTEARS-MLP	0.26/78.2	0.32/57.9	0.35/51.5	0.36/49.9	0.36/49.9	0.37/49.4	0.36/49.6
SDCD	0.22/118.1	0.26/100.0	0.35/75.3	0.38/68.7	0.40/62.8	0.41/59.2	0.43/54.2
Arrow	0.32/**47.5**	0.34/46.5	0.35/46.3	0.36/46.6	0.36/47.5	0.34/50.0	0.33/53.5
AVICI	0.22/58.0	0.31/47.5	0.35/44.9	0.35/44.5	0.35/44.3	0.36/43.7	0.36/40.4
CauScale	0.26/66.5	0.33/49.7	0.40/44.6	0.41/44.6	0.40/47.0	0.37/52.6	0.35/56.3
CDFM	0.34/70.5	0.38/65.5	0.42/61.2	0.44/58.0	0.46/54.5	0.48/52.4	0.48/51.2
FoundCause	**0.43**/49.7	**0.52**/45.2	**0.61/39.0**	**0.64/35.6**	**0.65/34.7**	**0.65/34.1**	**0.66/33.7**
SEA	0.26/80.8	0.34/58.6	0.41/45.7	0.42/45.0	0.42/45.1	0.42/44.7	0.42/44.8
TabCausal	0.33/47.6	0.42/**43.9**	0.49/41.4	0.53/39.2	0.55/39.0	0.57/38.1	0.57/38.6

The sample-size suite fixes $d = 30$ and varies only the number of observational samples. Higher F1 and lower SHD are better; bold and underline mark the best and second-best F1 and SHD separately in each column.

G.6 INTERVENTION-PROTOCOL SENSITIVITY

Table 18 reports the full numerical values for the intervention-protocol sensitivity suite.

Table 18: **Complete intervention-protocol sensitivity metrics.** All rows use the mixed observational/interventional split.

Protocol	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Default	CDIS	500	0.40 ±0.12	0.48 ±0.16	0.35 ±0.10	48.1 ±19.4	0.055 ±0.022	237.9 ±117.7	–	–
Default	GIES	500	0.48 ±0.11	0.43 ±0.13	**0.56 ±0.11**	59.7 ±31.4	0.069 ±0.036	251.3 ±78.4	–	–
Default	IGSP	500	0.40 ±0.11	0.44 ±0.15	0.38 ±0.11	53.7 ±26.9	0.062 ±0.031	232.6 ±107.2	–	–
Default	SDCD	500	0.38 ±0.09	0.33 ±0.11	0.48 ±0.11	73.0 ±25.5	0.084 ±0.029	246.0 ±92.0	–	–
Default	AVICI	500	0.39 ±0.13	**0.66 ±0.15**	0.29 ±0.12	43.4 ±17.3	0.050 ±0.020	189.6 ±109.6	0.77 ±0.09	0.42 ±0.12
Default	CauScale	500	0.45 ±0.11	0.56 ±0.12	0.39 ±0.11	43.6 ±18.7	0.050 ±0.022	201.3 ±108.4	**0.86 ±0.05**	0.47 ±0.11
Default	TabCausal	500	**0.54 ±0.12**	0.62 ±0.14	0.49 ±0.13	**40.5 ±17.8**	**0.047 ±0.020**	**187.9 ±96.7**	0.85 ±0.08	**0.56 ±0.14**
Fixed setpoint	CDIS	500	0.40 ±0.12	0.48 ±0.16	0.35 ±0.10	48.3 ±19.6	0.056 ±0.022	237.7 ±119.2	–	–
Fixed setpoint	GIES	500	0.48 ±0.12	0.43 ±0.14	**0.56 ±0.12**	59.7 ±31.7	0.069 ±0.036	253.1 ±79.2	–	–
Fixed setpoint	IGSP	500	0.40 ±0.12	0.43 ±0.15	0.38 ±0.11	54.2 ±26.7	0.062 ±0.031	233.2 ±105.7	–	–
Fixed setpoint	SDCD	500	0.38 ±0.09	0.33 ±0.11	0.48 ±0.11	72.7 ±26.5	0.084 ±0.030	242.8 ±91.2	–	–
Fixed setpoint	AVICI	500	0.38 ±0.12	**0.65 ±0.15**	0.28 ±0.11	43.7 ±17.1	0.050 ±0.020	189.3 ±109.8	0.77 ±0.09	0.41 ±0.12
Fixed setpoint	CauScale	500	0.46 ±0.10	0.57 ±0.11	0.39 ±0.11	43.0 ±18.6	0.049 ±0.021	198.7 ±106.2	0.85 ±0.05	0.47 ±0.11
Fixed setpoint	TabCausal	500	**0.54 ±0.12**	0.61 ±0.14	0.49 ±0.13	**40.6 ±17.7**	**0.047 ±0.020**	**186.6 ±94.4**	**0.85 ±0.08**	**0.55 ±0.14**
Parameter shift	CDIS	500	0.40 ±0.12	0.48 ±0.16	0.35 ±0.10	48.3 ±19.6	0.055 ±0.023	240.0 ±120.0	–	–
Parameter shift	GIES	500	0.43 ±0.11	0.39 ±0.12	**0.50 ±0.10**	64.1 ±32.1	0.074 ±0.037	279.8 ±79.1	–	–
Parameter shift	IGSP	500	0.39 ±0.11	0.43 ±0.15	0.38 ±0.10	54.4 ±27.0	0.062 ±0.031	235.9 ±107.2	–	–
Parameter shift	SDCD	500	0.38 ±0.09	0.33 ±0.11	0.47 ±0.11	73.2 ±25.5	0.084 ±0.029	244.8 ±88.5	–	–
Parameter shift	AVICI	500	0.37 ±0.13	**0.62 ±0.15**	0.27 ±0.11	44.6 ±17.7	0.051 ±0.020	192.5 ±109.1	0.76 ±0.09	0.39 ±0.13
Parameter shift	CauScale	500	0.42 ±0.10	0.52 ±0.12	0.36 ±0.11	45.2 ±18.8	0.052 ±0.022	209.8 ±106.1	0.84 ±0.06	0.41 ±0.11
Parameter shift	TabCausal	500	**0.52 ±0.12**	0.59 ±0.15	0.47 ±0.13	**41.8 ±18.0**	**0.048 ±0.021**	**191.4 ±94.1**	**0.85 ±0.08**	**0.53 ±0.15**
Multi-target	CDIS	500	0.38 ±0.12	0.47 ±0.16	0.33 ±0.10	48.9 ±19.4	0.056 ±0.022	240.6 ±120.8	–	–
Multi-target	GIES	500	0.48 ±0.12	0.44 ±0.14	**0.56 ±0.11**	59.3 ±31.2	0.068 ±0.036	252.8 ±79.0	–	–
Multi-target	IGSP	500	0.13 ±0.19	0.14 ±0.22	0.12 ±0.19	54.6 ±21.7	0.063 ±0.025	210.2 ±114.3	–	–
Multi-target	SDCD	500	0.38 ±0.09	0.34 ±0.11	0.48 ±0.11	72.4 ±25.9	0.083 ±0.030	243.8 ±93.4	–	–
Multi-target	AVICI	500	0.39 ±0.12	**0.64 ±0.14**	0.28 ±0.11	43.7 ±17.2	0.050 ±0.020	189.9 ±109.7	0.77 ±0.08	0.41 ±0.12
Multi-target	CauScale	500	0.44 ±0.10	0.51 ±0.12	0.39 ±0.11	45.7 ±18.7	0.053 ±0.021	210.0 ±104.7	0.84 ±0.06	0.44 ±0.11

Continued on next page

Protocol	Method	N	F1	Prec.	Rec.	SHD	nSHD	SID	AUROC	AP
Multi-target	TabCausal	500	**0.54 ±0.12**	0.62 ±0.14	0.49 ±0.13	**40.6 ±17.6**	**0.047 ±0.020**	**186.3 ±97.5**	**0.85 ±0.08**	**0.55 ±0.14**
Dense mixed	CDIS	500	0.36 ±0.10	0.52 ±0.15	0.29 ±0.09	47.2 ±17.9	0.054 ±0.021	210.4 ±113.2	–	–
Dense mixed	GIES	500	0.50 ±0.11	0.45 ±0.13	**0.57 ±0.11**	56.3 ±28.0	0.065 ±0.032	251.3 ±76.9	–	–
Dense mixed	IGSP	500	0.25 ±0.16	0.35 ±0.22	0.20 ±0.13	51.8 ±21.0	0.060 ±0.024	214.3 ±112.1	–	–
Dense mixed	SDCD	500	0.40 ±0.09	0.36 ±0.12	0.49 ±0.11	68.3 ±23.6	0.079 ±0.027	232.1 ±88.9	–	–
Dense mixed	AVICI	500	0.37 ±0.12	**0.64 ±0.15**	0.27 ±0.11	44.3 ±17.1	0.051 ±0.020	190.3 ±111.1	0.77 ±0.08	0.40 ±0.11
Dense mixed	CauScale	500	0.42 ±0.10	0.50 ±0.13	0.38 ±0.10	46.7 ±18.8	0.054 ±0.022	211.4 ±104.3	0.82 ±0.06	0.41 ±0.11
Dense mixed	TabCausal	500	**0.53 ±0.11**	0.61 ±0.13	0.48 ±0.12	**41.0 ±16.9**	**0.047 ±0.019**	**186.9 ±98.2**	**0.85 ±0.07**	**0.53 ±0.13**
High fraction	CDIS	500	0.36 ±0.10	0.51 ±0.15	0.28 ±0.09	47.6 ±17.7	0.055 ±0.020	211.5 ±114.9	–	–
High fraction	GIES	500	0.51 ±0.12	0.46 ±0.14	**0.58 ±0.12**	56.7 ±29.9	0.065 ±0.034	238.2 ±82.5	–	–
High fraction	IGSP	500	0.31 ±0.10	0.43 ±0.15	0.25 ±0.09	52.0 ±22.4	0.060 ±0.026	224.1 ±109.5	–	–
High fraction	SDCD	500	0.40 ±0.10	0.35 ±0.12	0.49 ±0.10	70.3 ±25.5	0.081 ±0.029	237.3 ±93.0	–	–
High fraction	AVICI	500	0.40 ±0.13	**0.67 ±0.14**	0.30 ±0.12	42.9 ±17.2	0.049 ±0.020	187.2 ±109.4	0.77 ±0.09	0.43 ±0.12
High fraction	CauScale	500	0.48 ±0.10	0.60 ±0.12	0.41 ±0.11	40.8 ±17.6	0.047 ±0.020	193.3 ±107.6	**0.86 ±0.05**	0.51 ±0.11
High fraction	TabCausal	500	**0.54 ±0.11**	0.62 ±0.13	0.49 ±0.12	**40.0 ±16.8**	**0.046 ±0.019**	**181.7 ±97.0**	0.86 ±0.07	**0.56 ±0.13**

G.7 Construction Ablation

Table 19 repeats the criterion-level construction-quality scores with standard deviations for the same 20 sampled scenarios summarized in Figure 9.

Table 19: **Agentic construction ablation quality scores.**

Stage	Variant	Cases	Plaus.	Mech.	Bench.	Spec.	Single avg.	Struct. div.	Sem./mech. div.	Avg.
A0	Seed only	20	3.650 (0.462)	3.375 (0.626)	3.400 (0.447)	3.375 (0.535)	3.450 (0.492)	**4.500**	4.500	3.800
A1	+ reference pack	20	3.650 (0.328)	3.275 (0.499)	3.375 (0.358)	3.425 (0.520)	3.431 (0.398)	**4.500**	**5.000**	3.871
A2	+ planning	20	4.225 (0.302)	4.275 (0.302)	4.000 (0.363)	4.225 (0.302)	4.181 (0.291)	**4.500**	**5.000**	4.371
A3	+ graph review	20	**4.375 (0.275)**	**4.400 (0.262)**	**4.200 (0.377)**	**4.400 (0.262)**	**4.344 (0.262)**	**4.500**	**5.000**	**4.479**

Scores are on a 1–5 scale judged by an external LLM evaluator. Single-graph criteria are mean (standard deviation) over 20 cases. The final average also includes two batch-level diversity criteria. Higher is better; bold and underline mark the best and second-best values in each score column. Ties share the same mark.