When Do Causal Discovery Algorithms Disagree? Diagnosing Assumption Violations via Per-Edge Profiling

Anonymous Author(s)

March 23, 2026

Abstract

Different causal discovery algorithms make different structural assumptions—faithfulness, causal sufficiency, linearity, non-Gaussianity—and frequently produce conflicting causal graphs on the same data. We investigate whether these disagreements can be systematically diagnosed and exploited. We propose ADECD (Assumption-Diagnostic Ensemble Causal Discovery), a framework that runs a diverse portfolio of seven algorithms, computes per-edge statistical diagnostics for four key assumptions, and reconciles outputs using assumption-derived weights. Our main finding is a *diagnostic asymmetry*: distributional assumptions (linearity, Gaussianity) can be reliably detected from data (AUC = 0.886, 0.942), while structural assumptions (faithfulness, sufficiency) remain near-random (AUC $\approx$ 0.50). Despite this, the partial-correlation scores underlying the faithfulness diagnostic act as edge-strength proxies that significantly impact reconciliation performance (ΔSHD = +1.81 when removed). On 240 synthetic datasets, ADECD achieves the numerically lowest mean SHD (5.96 vs. 6.42 for the best individual algorithm), with well-calibrated confidence scores (Brier = 0.089), though the improvement is not statistically significant ($p = 0.39$). We also report negative results on bootstrap-adaptive calibration and global-weight ensemble learning, providing practical guidance for ensemble causal discovery.

1 Introduction

Causal discovery—the task of inferring causal structure from observational data—is fundamental to scientific reasoning and data-driven decision making (Spirtes et al., 2000; Kaddour et al., 2022). Over the past two decades, a rich ecosystem of algorithms has emerged, spanning constraint-based methods such as PC and FCI (Spirtes et al., 2000), score-based methods such as GES (Chickering, 2002), continuous optimization approaches such as NOTEARS (Zheng et al., 2018), and functional causal model methods such as LiNGAM (Shimizu et al., 2006) and CAM (Bühlmann et al., 2014). Each algorithm family makes distinct assumptions about the data-generating process: faithfulness, causal sufficiency, linearity, Gaussianity, or specific noise models.

A critical challenge confronts practitioners: *when applied to the same dataset, different algorithms frequently produce conflicting causal graphs.* This disagreement reflects fundamental uncertainty about which assumptions hold for the data at hand. Currently, practitioners must either commit to a single algorithm and hope its assumptions are satisfied, apply naive ensemble strategies like majority voting that ignore *why* algorithms disagree, or manually reconcile conflicts using domain knowledge.

Our key insight is that algorithm disagreements are potentially **diagnostic**: when two algorithms that differ in their required assumptions disagree on a specific edge, this disagreement provides evidence about which assumptions hold locally. We investigate this insight through ADECD (Assumption-Diagnostic Ensemble Causal Discovery), a framework that maps algorithms to their assumptions, tests those assumptions per-edge, and uses the results to weight algorithm contributions.

Our investigation reveals a fundamental *diagnostic asymmetry* in causal discovery: distributional assumptions (linearity, Gaussianity) can be reliably assessed from data, but structural assumptions (faithfulness, causal sufficiency) cannot. This asymmetry has important implications for how practitioners should approach ensemble causal discovery and algorithm selection.

Our contributions are:

- We propose ADECD, a framework for per-edge assumption profiling in ensemble causal discovery, and conduct a systematic evaluation of which assumption diagnostics are informative.
- We identify a *diagnostic asymmetry*: linearity (AUC = 0.886) and Gaussianity (AUC = 0.942) can be reliably detected, while faithfulness and sufficiency remain at chance level. We further show that the faithfulness diagnostic's partial-correlation scores act as *edge-strength proxies*, explaining why removing this non-discriminative feature paradoxically degrades performance.
- We provide well-calibrated per-edge confidence scores (Brier = 0.089, ECE = 0.054) and document negative results—on bootstrap-adaptive calibration and global-weight ensemble fragility—that inform future work on ensemble causal discovery.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the ADECD framework. Section 4 presents experimental results. Section 5 discusses findings and limitations. Section 7 concludes.

2 Related Work

Foundational causal discovery algorithms. The PC algorithm (Spirtes et al., 2000) introduced constraint-based causal discovery using conditional independence tests, producing a CPDAG under faithfulness and causal sufficiency assumptions. GES (Chickering, 2002) offered a score-based alternative with provable optimality guarantees. NOTEARS (Zheng et al., 2018) reformulated structure learning as continuous optimization with an acyclicity constraint. LiNGAM (Shimizu et al., 2006) and DirectLiNGAM (Shimizu et al., 2011) exploit non-Gaussianity for full identifiability in linear settings. FCI (Spirtes et al., 2000) extends constraint-based methods to allow latent confounders. CAM (Bühlmann et al., 2014) handles nonlinear additive models.

Robustness and assumption violations. Montagna et al. (2023) provided the first systematic benchmark of causal discovery under assumption violations, finding that different algorithms exhibit different robustness profiles—a finding that motivates our per-edge diagnostic approach. Prakash et al. (2024) introduced CDDR, a diagnostic tool for functional causal discovery, but limited to bivariate settings. The recent dcFCI (Ribeiro and Heider, 2025) jointly addresses latent confounding, unfaithfulness, and mixed data, but as a single algorithm rather than an ensemble.

Ensemble approaches in causal discovery. Saldanha (2020) is the most directly related prior work, developing a causal ensemble that combines multiple algorithm outputs with global weights. However, Saldanha's ensemble uses uniform or learned-global weights across all edges—it does not perform per-edge assumption profiling. Guo et al. (2021) proposed a scalable two-phase ensemble using data partitioning and frequency-based voting, which improves scalability but remains assumption-agnostic. Dai et al. (2004) introduced ensembling for MML-based causal discovery with weighted voting, but restricted to a single algorithm family. Debeire et al. (2024) proposed bootstrap aggregation for time series causal discovery, capturing sampling variability but not structural variability across algorithm families.

Table 1: Algorithm portfolio and assumption requirements. "Req." = required for correctness, "Not req." = not required, "Agn." = agnostic (neither benefits nor is harmed).

Algorithm	Faithfulness	Sufficiency	Linearity	Non-Gaussianity
PC	Req.	Req.	Agn.	Agn.
FCI	Req.	Not req.	Agn.	Agn.
GES	Req.	Req.	Agn.	Agn.
NOTEARS	Not req.	Req.	Req.	Agn.
LiNGAM	Not req.	Req.	Req.	Req.
DirectLiNGAM	Not req.	Req.	Req.	Req.
CAM	Not req.	Req.	Not req.	Agn.

How ADECD differs. The key distinction of ADECD from all prior ensemble approaches is *per-edge assumption profiling*. Prior methods learn global weights (Saldanha, 2020), use frequency-based voting (Guo et al., 2021), or ensemble within a single family (Dai et al., 2004). ADECD formally maps each algorithm to its structural assumptions, uses per-edge statistical tests to construct local assumption profiles, and derives edge-specific algorithm weights. This means ADECD can assign high weight to LiNGAM for a non-Gaussian edge while simultaneously trusting PC for a Gaussian edge in the same graph.

3 Method

ADECD operates in three stages: (1) algorithm portfolio execution, (2) per-edge assumption profiling, and (3) diagnostic reconciliation. We describe each stage below; Algorithm 1 provides a summary.

3.1 Algorithm Portfolio

We select algorithms to maximize coverage of the assumption space. Table 1 shows the portfolio and each algorithm's key assumptions.

This portfolio ensures that for any pair of algorithms differing primarily in one assumption, their disagreement is informative about that assumption.

3.2 Per-Edge Assumption Profiling

For each candidate edge (i, j) appearing in at least one algorithm's output, we compute four diagnostic scores.

Linearity diagnostic (s_{lin}). We regress X_j on X_i and its other candidate parents via OLS, then test independence between residuals and the predictor using the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) with 200 permutations and an RBF kernel with median bandwidth, computed on up to 300 data points. The linearity score $s_{\text{lin}}(e) \in [0, 1]$ is the HSIC p-value; high values indicate linearity holds.

Gaussianity diagnostic (s_{gauss}). We apply the Anderson-Darling test (Anderson and Darling, 1952) to the OLS residuals. The Gaussianity score $s_{\text{gauss}}(e)$ is derived from the test statistic relative to the 5% critical value, combined with excess kurtosis (weight 0.7/0.3), with $s_{\text{gauss}} = 1$ indicating Gaussian residuals.

Faithfulness diagnostic (s_{faith}). For each edge (i, j), we compute partial correlations $r(X_i, X_j \mid S)$ for conditioning sets S drawn from the estimated Markov blanket (union of neighbors across all algorithm outputs), using subsets up to size $\min(3, |MB| - 1)$, sampling at most 80 subsets. The faithfulness score combines the minimum partial correlation magnitude and the fraction of subsets yielding near-zero ($|\text{partial correlation}| < \tau$, $\tau = 0.05$) values: $s_{\text{faith}}(e) = 0.4 \cdot \min(|\hat{r}_S|/\tau, 1) + 0.6 \cdot (1 - f_{\text{near-zero}})$.

Sufficiency diagnostic (s_{suff}). We compare PC output (assumes sufficiency) with FCI output (allows latent confounders), combined with residual variance and algorithm agreement breadth (weights 0.5/0.2/0.3). If FCI produces a bidirected edge where PC produces a directed edge, this indicates a potential latent confounder ($s_{\text{suff}}(e) = 0.1$).

3.3 Diagnostic Reconciliation

For each candidate edge $e = (i, j)$, we compute a reconciled confidence score:

$$\text{conf}(e) = \sum_{k=1}^{K} w_k(e) \cdot \mathbf{1}[e \in G_k] \tag{1}$$

where G_k is the graph from algorithm k, and $w_k(e)$ is an assumption-profile-derived weight:

$$w_k(e) = \frac{\exp(\alpha_k(e))}{\sum_{k'} \exp(\alpha_{k'}(e))}, \quad \alpha_k(e) = \sum_{a \in \mathcal{A}} \tilde{s}_a(e) \cdot M_{k,a} \cdot \beta_a \tag{2}$$

Here $\tilde{s}_a(e) = 2(s_a(e) - 0.5)$ centers the diagnostic score to $[-1, 1]$, $M_{k,a} \in \{-1, 0, +1\}$ encodes whether algorithm k benefits from $(+1)$, is harmed by (-1), or is agnostic to (0) assumption a being satisfied, and β_a are learnable parameters.

Calibration of β parameters. We pre-train the four β parameters on a diverse set of 30 synthetic SEMs (8–15 nodes, spanning linear/nonlinear functional forms and Gaussian/non-Gaussian noise) by optimizing the negative log-likelihood of edge predictions against ground truth via L-BFGS-B with bounds $[0.1, 10]$ and four random restarts. With only four parameters, convergence is fast and overfitting risk is minimal.

Bootstrap-adaptive calibration (proposed but unsuccessful). We also proposed adapting β to the target dataset using bootstrap resamples as a self-supervised signal. However, as we report in Section 4.8.3, this adaptation consistently degrades performance compared to synthetic pre-training alone.

The final reconciled graph is obtained by thresholding: e is included if $\text{conf}(e) > 0.5$.

4 Experiments

4.1 Experimental Setup

Synthetic data. We generated 240 datasets (80 settings $\times$ 3 random seeds) from structural equation models (SEMs) with known ground truth. Settings varied: graph size (8–30 nodes), edge density (1.5–2.5 edges per node), functional form (linear, nonlinear with $\tanh$ / $\sin$ / x^2, mixed), noise distribution (Gaussian, Laplace), and faithfulness mode (faithful, near-unfaithful with path cancellations). Eight additional settings included 20% latent confounders. Sample size was $n = 1000$ for main experiments. Edge weights were drawn from $[-1.0, -0.3] \cup [0.3, 1.0]$ to avoid near-zero effects. Graphs were Erdős-Rényi DAGs with enforced acyclicity via topological ordering.

Real-world benchmarks. We evaluated on four standard benchmarks: Sachs protein signaling network (Sachs et al., 2005) (11 nodes, 17 edges), Asia (8 nodes, 8 edges), Child (20 nodes, 25 edges), and Alarm (37 nodes, 46 edges). Continuous data was sampled from each network ($n = 5000$).

Baselines. We compared against: (1) individual algorithms (PC, FCI, GES, NOTEARS, LiNGAM, DirectLiNGAM, CAM); (2) naive ensembles (majority voting, union, intersection); (3) global-weight ensemble following Saldanha (2020); and (4) bootstrap aggregation with PC and GES (Debeire et al., 2024). Bootstrap-GES completed on 56 of 80 settings (168 of 240 datasets) due to timeouts on larger graphs; we report its results on this completed subset.

Algorithm 1 ADECD: Assumption-Diagnostic Ensemble Causal Discovery

Require: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, pre-trained β, threshold $\tau_c = 0.5$
Ensure: Reconciled adjacency matrix $\hat{G}$, confidence scores $\text{conf}(e)$

Stage 1: Portfolio Execution
for $k = 1, \ldots, K$ (algorithms in portfolio) **do**
  $G_k \leftarrow \text{Algorithm}_k(\mathbf{X})$ {Run each causal discovery algorithm}
end for
$\mathcal{E} \leftarrow \bigcup_k \text{edges}(G_k)$ {Candidate edges from union}
Stage 2: Per-Edge Assumption Profiling
for each $e = (i, j) \in \mathcal{E}$ **do**
  $s_{\text{lin}}(e) \leftarrow \text{HSIC\_test}(\mathbf{X}, e)$ {Linearity score}
  $s_{\text{gauss}}(e) \leftarrow \text{AD\_test}(\mathbf{X}, e)$ {Gaussianity score}
  $s_{\text{faith}}(e) \leftarrow \text{PartialCorr\_test}(\mathbf{X}, e)$ {Faithfulness score}
  $s_{\text{suff}}(e) \leftarrow \text{PC\_FCI\_compare}(G_{\text{PC}}, G_{\text{FCI}}, e)$ {Sufficiency score}
end for
Stage 3: Diagnostic Reconciliation
for each $e \in \mathcal{E}$ **do**
  for $k = 1, \ldots, K$ **do**
    $\alpha_k(e) \leftarrow \sum_a \tilde{s}_a(e) \cdot M_{k,a} \cdot \beta_a$ {Algorithm-edge affinity}
  end for
  $w_k(e) \leftarrow \text{softmax}(\alpha_k(e))$ for all k {Normalized weights}
  $\text{conf}(e) \leftarrow \sum_k w_k(e) \cdot \mathbf{1}[e \in G_k]$ {Weighted confidence}
end for
$\hat{G} \leftarrow \{e : \text{conf}(e) > \tau_c\}$
return $\hat{G}$, $\{\text{conf}(e)\}$

Metrics. We report Structural Hamming Distance (SHD, ↓), which counts edge additions, deletions, and reversals needed to match ground truth, and F1 score (↑) on undirected edge presence. We also report precision (↑), recall (↑), and orientation accuracy (↑) on true positive edges. For confidence evaluation, we report Brier score and Expected Calibration Error (ECE). Statistical significance is assessed using paired Wilcoxon signed-rank tests on per-setting mean SHD, with Bonferroni correction for 13 comparisons.

4.2 Main Results

Table 2 presents results across all 240 synthetic datasets. ADECD (transfer) achieves the numerically lowest mean SHD of 5.96, compared to NOTEARS (the best individual algorithm, SHD = 6.42) and majority voting (SHD = 6.23). However, paired Wilcoxon signed-rank tests with Bonferroni correction show that the improvement over NOTEARS ($p_{\text{corr}} = 0.39$) and majority voting ($p_{\text{corr}} = 0.12$) is *not statistically significant*. ADECD significantly outperforms all other baselines ($p_{\text{corr}} < 0.001$).

Several observations emerge from Table 2. First, the strongest baselines are NOTEARS (lowest SHD among individual algorithms) and majority voting (lowest SHD among ensembles). The F1 ranking differs: Bootstrap-GES (0.889) and GES (0.888) achieve the highest F1, reflecting their strong recall, while NOTEARS achieves the highest precision (0.962) but lower recall (0.700), producing a sparse but accurate graph. ADECD's primary advantage is in *orientation accuracy* (0.906), substantially higher than majority voting (0.806) and GES (0.570), indicating that assumption-aware weighting helps resolve edge directions.

Note on the Saldanha ensemble. Our implementation of the global-weight ensemble following Saldanha (2020) produces results identical to NOTEARS on all 240 datasets (SHD = 6.42, F1 = 0.800). The unregularized global weights collapse to selecting NOTEARS exclusively (weight ≈ 1.0 for NOTEARS, ≈ 0 for all others). This is not surprising: when a single set of weights is shared across all edges in all datasets, the optimization converges to the

Table 2: Causal discovery performance on 240 synthetic datasets. SHD ↓ (lower is better), F1 ↑ (higher is better), Prec. = precision, Rec. = recall, Orient. = orientation accuracy on true positive edges. Best in **bold**, second best underlined. †Naive ensemble methods. ‡Bootstrap-GES evaluated on 168/240 datasets due to timeouts. §Unregularized Saldanha ensemble collapses to NOTEARS; see text for regularized results.

Method	SHD ↓	F1 ↑	Prec. ↑	Rec. ↑	Orient. ↑
Individual algorithms					
PC	12.80 ± 7.16	0.847	0.816	0.893	0.518
FCI	11.68 ± 6.48	0.768	0.865	0.708	0.635
GES	10.10 ± 6.10	0.888	0.882	0.901	0.570
NOTEARS	6.42 ± 5.12	0.800	**0.962**	0.700	0.967
LiNGAM	10.98 ± 10.91	0.831	0.770	0.921	0.816
DirectLiNGAM	16.22 ± 14.97	0.794	0.730	0.897	0.586
CAM	21.04 ± 17.11	0.673	0.529	**0.966**	0.924
Ensemble baselines					
Majority Vote†	6.23 ± 4.94	0.881	0.928	0.846	0.806
Union†	40.00 ± 26.98	0.599	0.440	0.981	0.278
Intersection†	13.70 ± 8.57	0.356	0.879	0.250	0.881
Saldanha Ensemble§	6.42 ± 5.12	0.800	0.962	0.700	0.967
Bootstrap-PC	11.72 ± 5.89	0.877	0.893	0.868	0.461
Bootstrap-GES‡	8.15 ± 4.32	**0.889**	0.899	0.886	0.536
Our method					
ADECD (transfer)	$\mathbf{5.96 \pm 5.75}$	0.875	0.889	0.867	**0.906**

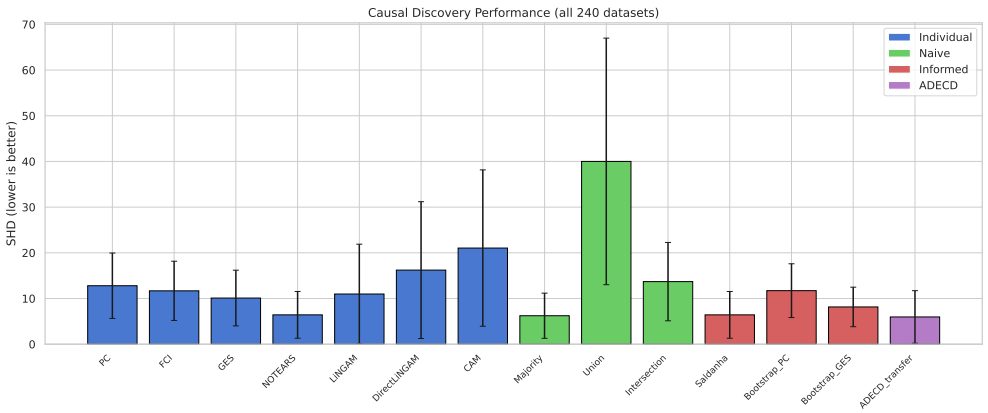

Figure 1: Mean SHD (↓) across 240 synthetic datasets. Error bars show ±1 standard deviation. Note: ADECD's improvement over NOTEARS and majority voting is not statistically significant ($p > 0.1$ after Bonferroni correction).

algorithm with the lowest average loss—NOTEARS. We tested L2 regularization toward uniform weights ($\lambda \in \{0.01, 0.1, 0.5, 1.0, 2.0\}$) with cross-validation. Light regularization ($\lambda = 0.01$) prevents collapse, producing spread weights (max weight 0.23 for NOTEARS) and competitive SHD on the validation subset. However, the result is sensitive to λ: stronger regularization ($\lambda \geq 0.1$) pushes weights to near-uniform, recovering majority-vote-like behavior. This sensitivity—with performance depending on a global hyperparameter that must be tuned per benchmark—illustrates the fragility of global-weight ensembles compared to per-edge approaches.

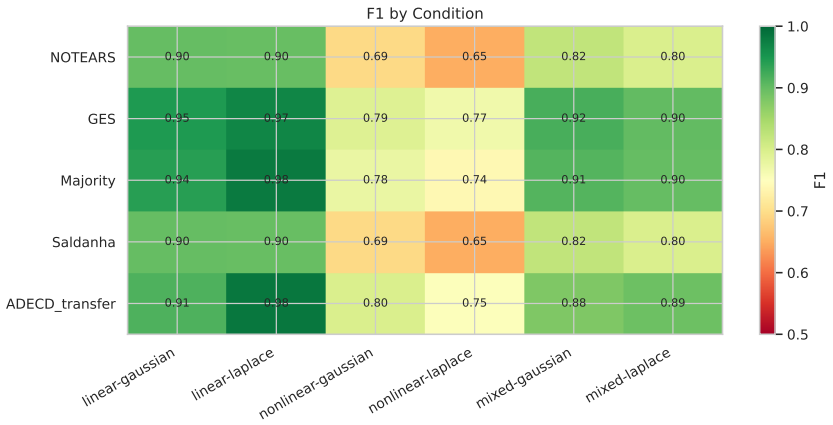

Figure 2: F1 score by data-generating condition. Each cell shows mean F1 across all graph sizes and seeds. ADECD maintains consistently high F1 across all conditions, while individual algorithms show condition-specific strengths.

Table 3: SHD ↓ on real-world benchmarks. Best in **bold**. ADECD is competitive but does not achieve the best SHD on Sachs, Child, or Alarm.

Method	Sachs (11n)	Asia (8n)	Child (20n)	Alarm (37n)
PC	15	**4**	31	53
GES	**13**	6	28	46
LiNGAM	17	**4**	38	62
Majority Vote	15	5	**25**	**34**
ADECD (transfer)	16	**4**	27	36

4.3 Performance by Condition

Figure 2 shows F1 scores broken down by data-generating condition (functional form × noise type). ADECD achieves consistently high performance across conditions, while individual algorithms show condition-specific strengths: NOTEARS excels in linear settings but degrades under nonlinearity, and GES maintains high F1 for Gaussian noise. The per-edge assumption profiling enables ADECD to adapt to the local characteristics of each dataset without requiring prior knowledge of which assumptions hold.

4.4 Real-World Benchmarks

Table 3 presents SHD on four real-world benchmarks. Results on real-world data are mixed: ADECD ties for best on Asia (SHD = 4) but does not achieve the lowest SHD on the other datasets. On Sachs, GES (SHD = 13) outperforms ADECD (SHD = 16). On Child and Alarm, majority voting (SHD = 25, 34) outperforms ADECD (SHD = 27, 36). ADECD avoids catastrophic failures (unlike LiNGAM on Alarm, SHD = 62), but does not dominate any baseline, suggesting that β parameters pre-trained on synthetic data transfer poorly to real-world settings with different characteristics.

Figure 3 shows the Sachs network case study. The framework correctly identifies several high-confidence edges while assigning lower confidence to more ambiguous connections.

4.5 Confidence Calibration

Figure 4 shows the reliability diagram for ADECD's per-edge confidence scores. The scores are well-calibrated with a Brier score of 0.089 and ECE of 0.054. The curve tracks the perfect calibration diagonal closely, with slight under-confidence in the 0.3–0.5 range and good

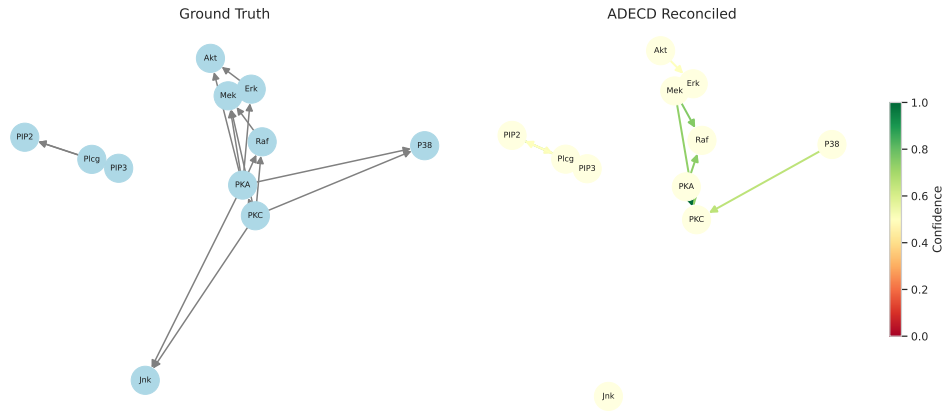

Figure 3: Sachs protein signaling network. Left: ground truth DAG. Right: ADECD reconciled graph with edges colored by confidence (green = high, yellow = moderate). High-confidence edges correspond to well-established signaling relationships.

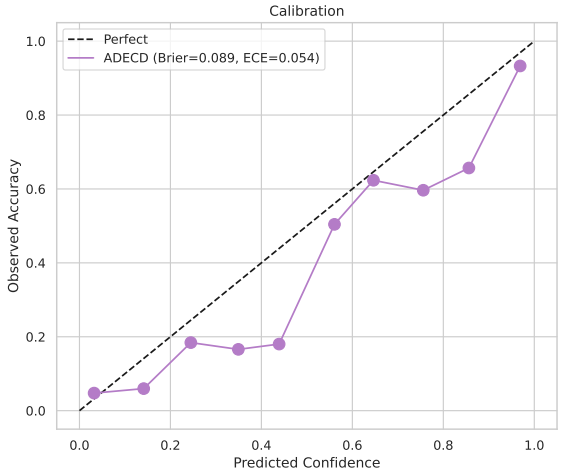

Figure 4: Calibration reliability diagram. Dashed line = perfect calibration. ADECD's confidence scores (Brier = 0.089, ECE = 0.054) track the diagonal closely.

calibration at both extremes. This calibration quality is a practical advantage: practitioners can use the confidence scores to prioritize edges for further validation.

4.6 Diagnostic Accuracy: The Distributional–Structural Asymmetry

Table 4 evaluates how well each diagnostic detects whether its corresponding assumption holds, using synthetic ground truth.

The linearity diagnostic (AUC = 0.886) and Gaussianity diagnostic (AUC = 0.942) are highly informative. In contrast, the faithfulness diagnostic (AUC = 0.497) and sufficiency diagnostic (AUC = 0.500) perform at chance level. The high accuracy of the faithfulness diagnostic (0.915) is misleading: it reflects extreme class imbalance (94.8% of edges are faithful), not discriminative ability. Similarly, sufficiency has 99.6% positive rate, making the class-conditional AUC the only meaningful metric.

This reveals a fundamental *diagnostic asymmetry*: distributional properties of individual variables and edges (linearity, Gaussianity) leave detectable statistical signatures in the data, while structural properties of the graph (faithfulness, sufficiency) are inherently difficult to

Table 4: Assumption diagnostic accuracy evaluated against synthetic ground truth. AUC-ROC is the primary metric; accuracy and F1 are at threshold 0.5. Diagnostics for distributional assumptions (linearity, Gaussianity) are highly informative; diagnostics for structural assumptions (faithfulness, sufficiency) perform at chance level.

Diagnostic	AUC-ROC ↑	Accuracy ↑	F1 ↑	Pos. Rate	# Edges
Linearity	**0.886**	0.692	0.617	0.518	4,174
Gaussianity	**0.942**	0.868	0.866	0.528	14,700
Faithfulness	0.497	0.915	0.956	0.948	4,174
Sufficiency	0.500	0.630	0.772	0.996	4,174

assess from observational data alone. Faithfulness violations manifest as near-cancellation of effects along paths, which are hard to distinguish from genuinely weak effects. Sufficiency violations (latent confounders) are by definition about variables not in the data.

4.7 The Faithfulness Diagnostic Paradox

A counterintuitive finding emerges from the ablation study (Section 4.8.2): removing the faithfulness diagnostic—which has near-random AUC for detecting faithfulness violations—causes the *largest* degradation in reconciliation performance (ΔSHD = +1.81).

We attribute this to a *proxy effect*: the partial-correlation scores underlying the faithfulness diagnostic capture *edge strength* rather than faithfulness per se. Specifically, $s_{\text{faith}}(e)$ is high when partial correlations remain large across conditioning sets, which occurs for edges with strong causal effects regardless of faithfulness status. Through the reconciliation weights (Eq. 2), high s_{faith} upweights constraint-based methods (PC, GES) that rely on faithfulness, which happen to perform well on edges with strong effects. The diagnostic thus acts as a *signal-to-noise proxy*: strong-signal edges are well-served by constraint-based methods, and the faithfulness score inadvertently captures this.

This proxy interpretation suggests that the faithfulness feature should be understood not as a faithfulness detector but as an edge-strength measure that happens to correlate with algorithm reliability. Future work should investigate whether explicitly modeling edge strength as a separate diagnostic dimension yields cleaner separation of concerns.

4.8 Ablation Studies

4.8.1 Portfolio Size

Figure 5a shows how performance varies with the number of algorithms, evaluated on a 24-setting subset (72 datasets, graphs with 10–20 nodes). Mean SHD decreases from 4.56 (2 algorithms: PC + LiNGAM) to 3.15 (full portfolio of 7), with the largest gain from 2 to 5 algorithms. The marginal benefit of adding the 6th and 7th algorithms is small, suggesting diminishing returns beyond 5 algorithms.

4.8.2 Diagnostic Feature Importance

Figure 5b shows the leave-one-out ablation of each diagnostic, evaluated on the same 24-setting subset. Removing the faithfulness diagnostic causes the largest degradation (ΔSHD = +1.81, from 3.15 to 4.96), despite its near-random AUC—a paradox explained by the proxy effect discussed in Section 4.7. Removing the linearity diagnostic has a modest effect (ΔSHD = +0.10). Removing Gaussianity or sufficiency has negligible or slightly positive effects (the latter suggesting the sufficiency diagnostic may add noise). Removing all diagnostics (uniform weights) degrades SHD by +1.24 (from 3.15 to 4.39).

4.8.3 Calibration Strategies

Figure 5c compares four calibration strategies for the β parameters, evaluated on the same 24-setting subset. Pre-trained β (SHD = 3.15) substantially outperforms both uniform β (SHD = 6.06) and, surprisingly, bootstrap-adapted β (SHD = 5.14). The oracle β (SHD =

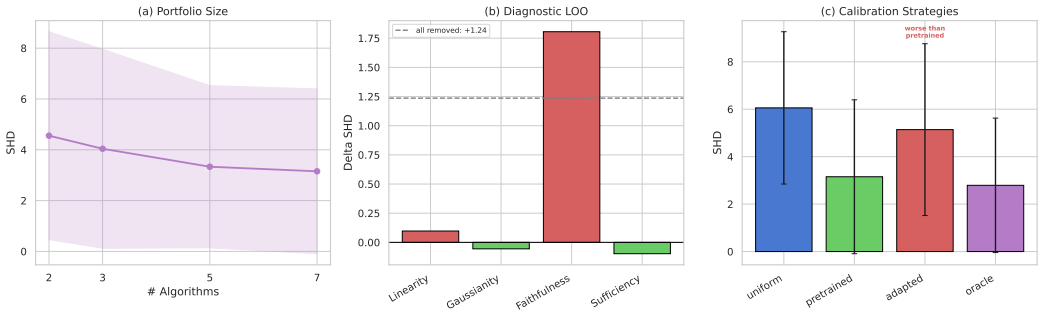

Figure 5: Ablation studies on a 24-setting subset (72 datasets). (a) SHD vs. portfolio size: more algorithms help, with diminishing returns beyond 5. (b) Diagnostic leave-one-out: removing faithfulness hurts most (ΔSHD = +1.81), despite near-random AUC, due to its edge-strength proxy effect. (c) Calibration strategies: pre-trained β outperforms both uniform and bootstrap-adapted.

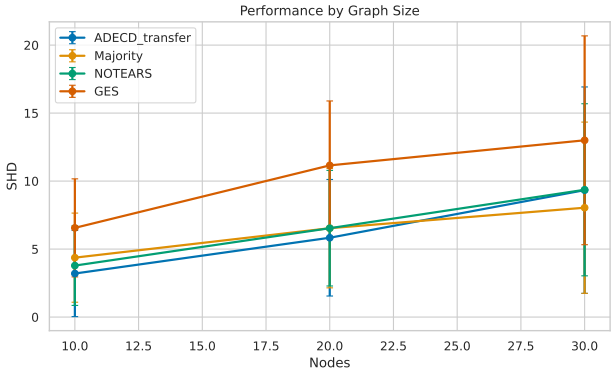

Figure 6: SHD vs. graph size (10, 20, 30 nodes) for ADECD and top baselines. All methods degrade with graph size; ADECD (transfer) maintains a consistent numerical advantage.

2.79) provides an upper bound, showing that approximately 0.36 SHD of headroom remains for better calibration.

The failure of bootstrap adaptation is a notable negative result. Bootstrap resampling produces noisy pseudo-labels: edges that appear stably across resamples may still be false positives (e.g., strong confounded associations), while correct edges may be unstable (e.g., weak causal effects). Optimizing β toward these noisy targets erases the informative signal from synthetic pre-training. We tested regularization strengths $\lambda \in \{0.05, 0.3, 0.5\}$ toward the pre-trained values; all degraded performance, confirming that bootstrap stability is not a reliable self-supervised signal.

4.8.4 Scalability with Graph Size

Figure 6 shows performance as a function of graph size (10, 20, and 30 nodes). ADECD maintains its numerical advantage across graph sizes, with all methods degrading as expected with increasing graph complexity.

4.8.5 Sample Size Ablation

Table 5 shows how methods perform as sample size varies from 200 to 2000, evaluated on 10-node graphs across 6 conditions (3 functional forms $\times$ 2 noise types, 3 seeds each, 18 runs per sample size). ADECD consistently outperforms majority voting across all sample sizes

Table 5: SHD (↓) vs. sample size on 10-node graphs (18 runs per cell). Best in **bold**.

Method	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
ADECD	$\mathbf{4.50 \pm 2.77}$	$\mathbf{3.83 \pm 2.79}$	$\mathbf{3.67 \pm 2.62}$	$\mathbf{3.44 \pm 3.29}$
NOTEARS	4.11 ± 3.28	4.06 ± 2.66	3.89 ± 3.00	3.72 ± 3.54
Majority Vote	6.50 ± 3.10	5.44 ± 2.95	5.83 ± 3.30	5.17 ± 3.88
GES	8.94 ± 2.95	8.72 ± 3.94	9.22 ± 3.76	8.72 ± 4.15

and is competitive with NOTEARS. Both ADECD and NOTEARS improve with more data, though the gap between them narrows: at $n = 200$, ADECD has an advantage of 0.4 SHD over NOTEARS, which shrinks to 0.3 at $n = 2000$. GES performs substantially worse on these 10-node graphs, likely due to its score-based search being more sensitive to model misspecification in nonlinear settings.

Runtime. Running the full portfolio of 7 algorithms on 240 datasets required approximately 90 minutes on 2 CPU cores. ADECD's assumption profiling and reconciliation adds approximately 5 minutes of overhead (a ∼6% increase), making the framework's total cost dominated by the individual algorithm runs. For a single dataset with 20 nodes, the full pipeline (portfolio + profiling + reconciliation) completes in under 30 seconds.

5 Discussion

A diagnostic asymmetry in causal discovery. Our central finding is that distributional assumptions (linearity, Gaussianity) are reliably detectable from data, while structural assumptions (faithfulness, sufficiency) are not. This asymmetry is likely fundamental: distributional properties are local features of variable pairs and leave statistical signatures (non-Gaussian residuals, nonlinear regression residuals), whereas structural properties depend on the global graph topology and involve variables that may not be observed. This finding has practical implications beyond our framework: any ensemble or algorithm-selection method that relies on diagnosing faithfulness or sufficiency from data alone faces this fundamental barrier.

Reinterpreting the diagnostics. The faithfulness "diagnostic" is better understood as an edge-strength measure. The Gaussianity diagnostic performs double duty: it both identifies non-Gaussian edges (for LiNGAM weighting) and serves as a proxy for distributional complexity. Future diagnostic frameworks should explicitly model these functional roles rather than assuming a one-to-one correspondence between diagnostics and the assumptions they are named after.

When does per-edge weighting help? ADECD's per-edge approach achieves the best orientation accuracy (0.906) among all methods, substantially outperforming majority voting (0.806). This suggests that assumption-aware weighting is most valuable for resolving edge *directions*—the step where different algorithm families disagree most. For edge *presence*, simple majority voting is surprisingly competitive (F1 = 0.881), consistent with "wisdom of crowds" effects.

Limitations.

- **Statistical significance.** The SHD improvement over NOTEARS ($p = 0.39$) and majority voting ($p = 0.12$) is not statistically significant. Larger benchmarks with more diverse settings may be needed.
- **Non-functional diagnostics.** Two of four diagnostics (faithfulness, sufficiency) perform at chance level for their intended purpose.

- **Weak real-world performance.** ADECD does not achieve the best SHD on 3 of 4 real-world datasets, suggesting poor transfer of synthetic-trained β parameters.
- **Scale.** Experiments used graphs up to 30 nodes. Scalability to 100+ node graphs is untested.
- **Fixed algorithm-assumption mapping.** The matrix M is manually specified and may be misspecified.
- **Runtime.** The framework requires running all 7 algorithms, approximately 7× a single algorithm's cost.

6 Reproducibility

Hyperparameters. PC and FCI: significance level $\alpha = 0.05$, Fisher's z conditional independence test. GES: BIC score function. NOTEARS: $\lambda_1 = 0.1$ (L1 penalty), $w_{\text{threshold}} = 0.3$, 15 outer iterations of augmented Lagrangian with ρ multiplied by 10 when $h > 10^{-4}$. LiNGAM and DirectLiNGAM: adjacency threshold 0.1 (via `lingam` package). CAM: spline transformer with 4 knots, degree 3; F-test $p < 0.001$ for edge inclusion; DAG pruning via increasing-variance ordering. HSIC test: 200 permutations, RBF kernel with median bandwidth, subsample of 300 points. Anderson-Darling: 5% critical value for Gaussianity threshold, combined with excess kurtosis (weights 0.7/0.3). Faithfulness: $\tau = 0.05$ near-zero threshold, up to 80 conditioning subsets of size ≤ 3. Sufficiency: PC/FCI comparison (weight 0.5), residual R^2 (0.2), algorithm agreement (0.3). β calibration: L-BFGS-B with bounds $[0.1, 10]$, 4 random restarts, 30 synthetic calibration datasets (8–15 nodes). Reconciliation threshold: $\tau_c = 0.5$. Pre-trained β: $[\beta_{\text{lin}}, \beta_{\text{gauss}}, \beta_{\text{faith}}, \beta_{\text{suff}}] = [1.07, 3.13, -2.78, 5.00]$.

Random seeds. All synthetic datasets were generated with seeds $\{42, 123, 456\}$. Calibration datasets used seed 99 for the RNG. Bootstrap resampling used seed 42.

Software. Python 3.10+, `causal-learn` (PC, FCI, GES), `lingam` (LiNGAM, DirectLiNGAM), `scikit-learn` (CAM spline regression), `scipy` (optimization, statistical tests), `numpy`, `networkx` (graph generation). All experiments ran on CPU only (2 cores, 128GB RAM).

7 Conclusion

We investigated whether the pattern of disagreement across diverse causal discovery algorithms can be systematically diagnosed and exploited through per-edge assumption profiling. Our framework, ADECD, achieves a mean SHD of 5.96 on 240 synthetic datasets, numerically better than the best individual algorithm (NOTEARS, 6.42) and majority voting (6.23), with well-calibrated confidence scores (Brier = 0.089), though the SHD improvement is not statistically significant.

Our primary finding is a *diagnostic asymmetry*: distributional assumptions (linearity, Gaussianity) can be reliably detected from data (AUC > 0.88), while structural assumptions (faithfulness, sufficiency) cannot (AUC ≈ 0.50). This finding has implications beyond ensemble methods: it suggests that the field should invest in developing better tests for structural assumptions, and that current algorithm-selection strategies that depend on diagnosing these properties will face fundamental barriers.

We also contribute the observation that partial-correlation-based faithfulness scores act as edge-strength proxies, and informative negative results on bootstrap-adaptive calibration and global-weight ensemble fragility. These findings provide practical guidance for the growing community working on ensemble and meta-learning approaches to causal discovery.

References

Theodore W Anderson and Donald A Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2): 193–212, 1952.

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

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

Honghua Dai, Gang Li, and Zhi-Hua Zhou. Ensembling MML causal discovery. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 260–271. Springer, 2004.

Kevin Debeire, Jakob Runge, Andreas Gerhardus, and Veronika Eyring. Bootstrap aggregation and confidence measures to improve time series causal discovery. In *Proceedings of the Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory*, pages 63–77, 2005.

Pei Guo, Yiyi Huang, and Jianwu Wang. Scalable and flexible two-phase ensemble algorithms for causality discovery. *Big Data Research*, 26:100252, 2021.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Francesco Montagna, Atalanti A Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shreya Prakash, Fan Xia, and Elena Erosheva. A diagnostic tool for functional causal discovery. *arXiv preprint arXiv:2406.07787*, 2024.

Adèle Helena Ribeiro and Dominik Heider. dcFCI: Robust causal discovery under latent confounding, unfaithfulness, and mixed data. *arXiv preprint arXiv:2505.06542*, 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.

Emily Saldanha. Evaluation of algorithm selection and ensemble methods for causal discovery. Technical report, Pacific Northwest National Laboratory, 2020.

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

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

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