Causal Discovery Beyond Scaling: Mixed-Type DAG Learning with Native Missing-Data Inference

Anonymous Author
Anonymous Institution

Abstract

Scaling predictive optimization alone is not sufficient for causal discovery when data are mixed-type and partially observed. We study this setting with PredCoM, a predictive-coding DAG learner that combines mixed-node likelihoods, sparse acyclicity regularization, and native masked state inference in a single objective. Across ER/SF/WS synthetic benchmarks (RAW and CATMIX; complete, MCAR, and MAR), PredCoM is consistently competitive against NOTEARS, Peter Clark, DirectLiNGAM, LiM, mCMIkNN, and sortnregress. Missingness-rate and MAR-strength sweeps show that complete-case preprocessing deteriorates as retained-row fraction collapses, while masked training remains substantially stronger. A compute-budget ablation shows that increasing epochs does not rescue complete-case failures when data viability is near zero. The results identify a concrete boundary of scaling: in mixed-data causal discovery, optimization budget cannot substitute for explicit missingness-aware causal modeling.

Preliminary work. Under review by AISTATS 2026. Do not distribute.

1 Introduction

Recent progress in machine learning is often framed as a scaling story: larger models and longer optimization frequently improve predictive metrics. In causal discovery, this intuition is incomplete. When data are mixed-type and partially observed, preprocessing can dominate final graph quality, and poor missing-data handling can erase intervention-relevant structure even with substantial training budget.

We focus on a high-friction regime where this issue is explicit: continuous, binary, and categorical variables with MCAR/MAR masking. Complete-case row dropping can reduce usable samples exponentially with dimension. We compare such pipelines against PredCoM, a predictive-coding causal learner that handles missingness inside training via masked state inference rather than an external imputation/drop stage (Friston, 2005; Salvatori et al., 2023).

Main claim. In mixed and partially observed regimes, explicit causal structure learning with mask-aware optimization yields robustness that is not recovered by simply increasing optimization budget on reduced complete-case subsets.

2 Methodology

Let $x^{(n)} \in \mathbb{R}^d$ be sample n, $m^{(n)} \in \{0,1\}^d$ its observation mask, and $W \in \mathbb{R}^{d\times d}$ the weighted adjacency. PredCoM uses node drives

$$u_j^{(n)} = \sum_{i=1}^{d} W_{i,j} z_i^{(n)}, \qquad z_i^{(n)} = f(x_i^{(n)}),$$

with node-wise negative log-likelihoods

$$\begin{aligned} F_j^{\text{gauss}} &= \frac{(x_j - u_j)^2}{2\sigma_j^2}, \\ F_j^{\text{bern}} &= \log(1 + e^{u_j}) - x_j u_j, \\ F_j^{\text{cat}} &= -\sum_{c=1}^{C_j} \mathbf{1}\{x_j = c\} \log \operatorname{softmax}(\eta_j)_c. \end{aligned} \tag{1}$$

The training objective is

$$\mathcal{L}(W) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{d} m_j^{(n)} F_j + \lambda_1 \|W\|_1 + \lambda_h h_{\mathrm{dag}}(W), \tag{2}$$

where $h_{\mathrm{dag}}(W) = \mathrm{tr}(\exp(W \odot W)) - d$ is zero iff the graph is acyclic (Zheng et al., 2018).

Masked state inference and local gradients. Observed coordinates are clamped; missing coordinates are inferred by

$$x_{j,t+1}^{(n)} = x_{j,t}^{(n)} - \gamma\big(1 - m_j^{(n)}\big)\frac{\partial \mathcal{L}}{\partial x_{j,t}^{(n)}}. \tag{3}$$

For Gaussian/Bernoulli nodes, the likelihood gradient factorizes locally:

$$\frac{\partial \mathcal{L}}{\partial W_{i,j}} = \frac{1}{N}\sum_{n=1}^{N} m_j^{(n)}\, \varepsilon_j^{(n)}\, z_i^{(n)}, \tag{4}$$

with $\varepsilon_j = \partial F_j / \partial u_j$. Weights are updated by proximal gradient:

$$W^{t+1} = \mathrm{prox}_{\eta\lambda_1\|\cdot\|_1}\Big(W^t - \eta\big(\nabla_W \mathcal{L}_{\mathrm{lik}} + \lambda_h \nabla h_{\mathrm{dag}}\big)\Big). \tag{5}$$

Complete-case sample collapse. For a complete-case pipeline, the retained-row fraction is

$$\rho_{\mathrm{cc}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left\{\sum_{j=1}^{d} m_{n,j} = d\right\}. \tag{6}$$

Under MCAR with per-entry rate r, $\mathbb{E}[\rho_{\mathrm{cc}}] = (1-r)^d$. This predicts exponential loss of usable rows as d increases, even before optimization starts.

Dimension-vs-sparsity note. In our generator, expected edge count scales as $e = 4d$, so global edge density scales as

$$p_d = \frac{e}{d(d-1)} \approx \frac{4}{d-1}.$$

Thus $d \in \{10, 15, 20\}$ corresponds to $p_d \approx \{0.44, 0.29, 0.21\}$: larger d means globally sparser graphs while local parent-set size stays around 4.

3 Experimental Design

We evaluate two questions: (i) mixed-data causal recovery under missingness, and (ii) whether optimization budget can compensate for complete-case data collapse.

Protocol. We use ER/SF/WS DAGs with RAW (continuous+binary) and CATMIX (continuous+binary+categorical) node families under complete/MCAR/MAR missingness. We report directed F1, SHD, SID (Peters and Bühlmann, 2015), and total-effect error. Baselines are NOTEARS, Peter Clark, DirectLiNGAM (Shimizu et al., 2011), LiM, mCMIkNN (Huegle et al., 2023), and sortnregress.

4 Empirical Results

All-method benchmark at $d = 20$. Figure 1 summarizes 18 benchmark conditions at $d = 20$. PredCoM achieves the highest mean F1 (0.711, rank 1/7), exceeds sortnregress in 15/18 conditions, and exceeds NOTEARS and mCMIkNN in 18/18 conditions. This supports that one mixed-type objective with explicit DAG regularization remains top-tier under partial observation.

Missingness-rate stress test. In the top panel, "Baseline" means the mean over non-PredCoM methods $\mathcal{M}$={NOTEARS, Peter Clark, DirectLiNGAM, LiM, mCMIkNN, sortnregress}, with

$$\Delta F1_{\mathrm{base}}(r) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Big(F1_{\mathrm{cc}}^{m}(r) - F1_{\mathrm{imp}}^{m}(r)\Big),$$

where "CC - Imputed" means complete-case minus imputed-input performance for the same method and condition. For PredCoM,

$$\Delta F1_{\mathrm{pc}}(r) = F1_{\mathrm{cc}}^{\mathrm{pc}}(r) - F1_{\mathrm{masked}}^{\mathrm{pc}}(r),$$

where "CC - Masked" compares complete-case against PredCoM's native masked training. Hence, more negative $\Delta F1$ is worse for complete-case preprocessing. The bottom panel shows $\rho_{\mathrm{cc}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\{\sum_j m_{n,j} = d\}$. As r increases, ρ_{cc} drops from roughly 0.38 to < 0.01, and complete-case degradation grows accordingly. The larger negative PredCoM curve does *not* mean PredCoM is weaker than baselines; it means complete-case harms PredCoM more relative to its stronger masked reference.

Compute budget versus data viability. Figure 3 identifies why the complete-case budget curves previously looked uninformative. In panel A, masked PredCoM improves with budget (MCAR: $0.689 \to 0.702$, MAR: $0.670 \to 0.687$, from 200 to 1600 epochs). In panel B, complete-case retention is extremely low ($\rho_{\mathrm{cc}} = 0.0346$ for MCAR and 0.0122 for MAR), and no condition yields finite complete-case F1 at any budget (annotated counts 0/3 per missingness at each epoch). Therefore, increasing optimization steps cannot repair the complete-case path in this regime because the failure is data viability, not optimizer convergence.

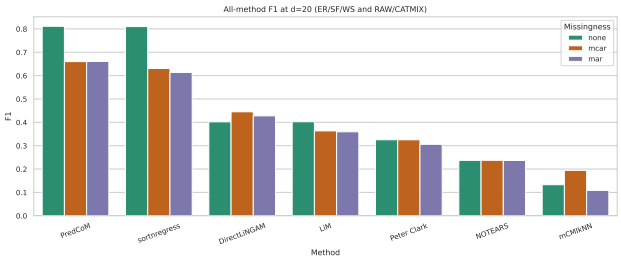

Figure 1: All-method F1 at $d = 20$ over ER/SF/WS, RAW/CATMIX, and complete/MCAR/MAR.

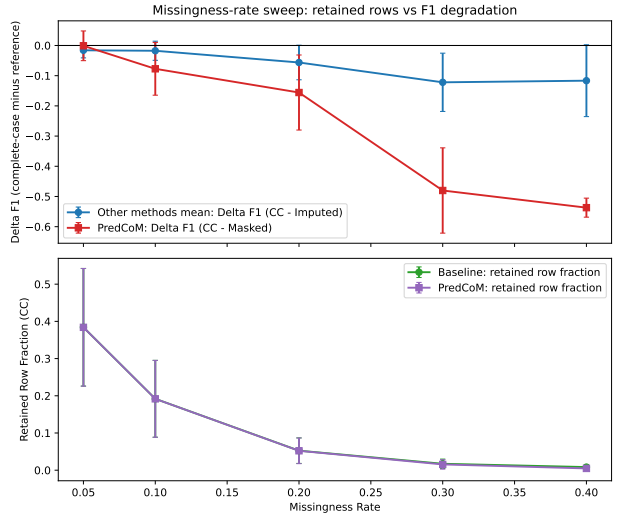

Figure 2: Missingness-rate sweep ($r \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$): top panel reports complete-case F1 deltas relative to each method's reference path; bottom panel reports complete-case retained-row fraction ρ_{cc}. Error bars show one standard deviation across aggregated conditions.

Quantitative synthesis. Across the $d = 20$ all-method benchmark (Figure 1), PredCoM attains mean F1 0.711 (rank 1/7), with condition-wise wins of 15/18 versus sortnregress and 18/18 versus NOTEARS and mCMIkNN. In the rate sweep (Figure 2) at $r = 0.40$, complete-case degradation is $\Delta F1_{\text{base}} = -0.117$ for the baseline mean and $\Delta F1_{\text{pc}} = -0.537$ for PredCoM relative to its masked reference, consistent with severe row-collapse at high r. In the budget ablation (Figure 3), masked gains from 200→1600 epochs are modest but positive (+0.0136 MCAR, +0.0170 MAR), while complete-case remains non-viable because retained-row fractions stay near zero.

Robustness matrix across topology/family/missingness. Figure 4 should be interpreted as 18 condition-specific slices (RAW/CATMIX × ER/SF/WS × none/MCAR/MAR), not as one

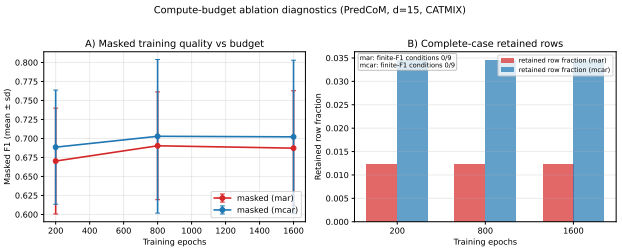

Figure 3: Compute-budget ablation ($d = 15$, CATMIX, MCAR/MAR, epochs $\{200, 800, 1600\}$): panel A shows masked F1 versus epochs; panel B shows complete-case viability through retained-row fractions, with finite-F1 run counts annotated above bars.

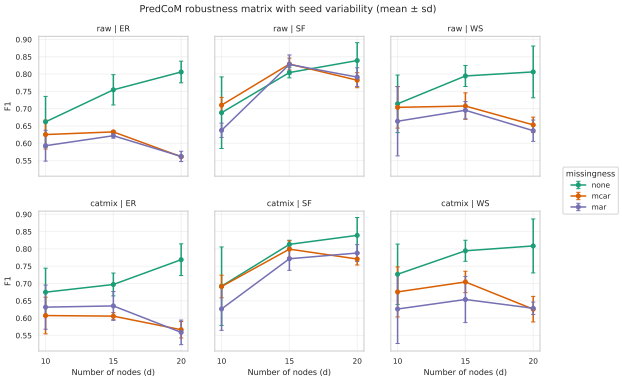

Figure 4: PredCoM F1 over ER/SF/WS, RAW/CATMIX, complete/MCAR/MAR, and node counts $d \in \{10, 15, 20\}$ (mean ± sd across seeds).

monotone scaling law in d. Non-monotonicity is therefore expected. Diagnostics in Appendix Table 2 show 10/18 non-monotone slices and 12/18 cases where $d = 15$ and $d = 20$ overlap within one standard deviation. Seed-level bootstrap comparisons (Appendix Table 3) further show 11/18 pointwise increases and 7/18 decreases for $d = 20 - d = 10$, with only 9/18 significantly positive and 2/18 significantly negative at the 95% bootstrap level. Hence the apparent upward trend is condition-dependent, not universal; it is compatible with (i) heterogeneity across topology/family/missingness and (ii) decreasing global density p_d as d grows under fixed mean degree.

Compute and real-data context. The synthetic findings are not explained by a hidden compute advantage. CPU-fair profiling in Appendix Table 8 shows PredCoM has low per-update cost (epoch-normalized) while remaining competitive in the highest-dimensional mixed setting. On fully observational real datasets, Appendix Table 9 shows that all methods return DAG hypotheses, but cycle-correction and sparsity behavior differ substantially across methods, reinforcing the

Table 1: Dimension-effect sign summary ($d = 20$ vs $d = 10$) for PredCoM F1 across the 18 condition slices of Figure 4.

Statistic	Count
Point estimate > 0	11
Point estimate < 0	7
95% CI entirely > 0	9
95% CI entirely < 0	3
95% CI overlaps 0	6

practical need for explicit structural regularization.

Mechanism-strength sensitivity. The same collapse pattern appears when the MAR mechanism is hardened rather than when only the nominal missing rate is increased. Appendix Table 7 and Figure 7 show that as α grows from 0.5 to 4.0, complete-case retention remains in the low-fraction regime ($\rho_{\text{cc}} \approx 0.065 \rightarrow 0.024$), and complete-case deltas become more negative. This is important because it separates two effects: stronger missingness mechanisms reduce usable rows even when the target missing rate is fixed, and this reduction propagates directly to structure quality unless missingness is modeled in-training.

For a compact stress diagnostic, define a finite-difference sensitivity

$$S(\xi_1, \xi_2) = \frac{\Delta F1(\xi_2) - \Delta F1(\xi_1)}{\xi_2 - \xi_1},$$

with $\xi = r$ for rate sweeps and $\xi = \alpha$ for MAR-strength sweeps. Using the aggregate curves, $S_r(0.05, 0.40) \approx -0.287$ for the baseline mean and $S_r(0.05, 0.40) \approx -1.531$ for PredCoM complete-case controls. For MAR strength, $S_\alpha(0.5, 4.0) \approx -0.013$ (baseline mean) versus -0.058 (PredCoM complete-case controls). These slopes quantify how rapidly complete-case quality deteriorates under harder missingness.

Supplementary evidence map. Appendix Table 4 and Figure 5 report baseline complete-case vs imputed deltas by method. Appendix Table 5 and Figure 6 isolate PredCoM masked vs complete-case. Appendix Table 6 provides the full rate-sweep aggregates underlying Figure 2, and Appendix Table 7 with Figure 7 provides MAR-strength sweeps. Runtime and real-world diagnostics are summarized in Appendix Table 8, Figure 8, and Table 9.

5 Discussion and Limitations

This study isolates a practical failure mode in mixed-data causal discovery: optimization budget and model capacity are insufficient when preprocessing discards most rows. The rate sweep, MAR-strength sweep, and budget ablation all converge on the same mechanism-level interpretation: information loss from complete-case filtering dominates downstream graph quality once ρ_{cc} is small. In contrast, masked state inference preserves partially observed samples and yields stable structure quality across broader missingness regimes.

The main limitation is scope. Our benchmarks are observational, acyclic, and synthetic-heavy; they do not establish identifiability under latent confounding, feedback systems, or intervention shift. A second limitation is calibration: while structure metrics improve, downstream effect estimates still vary across topology and family, especially in hard MAR settings. A third limitation is variance: several condition-level differences are within one-standard-deviation overlap, suggesting that larger seed budgets or hierarchical uncertainty modeling would strengthen claims about fine-grained ranking gaps.

Two next directions are high-value for this line of work. First, evaluate explicit train-test mechanism shift (e.g., train under MCAR and test under stronger MAR) to connect structure recovery with out-of-distribution reliability. Second, extend the masked predictive-coding objective to partially cyclic or dynamic systems where acyclicity is relaxed but intervention semantics are retained. More broadly, these results suggest that progress in scalable causal learning requires coupling optimization scale with explicit structural objectives and missingness-aware inference, rather than treating missing data as a preprocessing afterthought.

References

Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.

Johannes Huegle, Christopher Hagedorn, and Rainer Schlosser. A KNN-based non-parametric conditional independence test for mixed data and application in causal discovery. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 541–558. Springer, 2023.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

Tommaso Salvatori, Luca Pinchetti, Amine M'Charrak, Beren Millidge, and Thomas Lukasiewicz. Causal inference via predictive coding. *arXiv preprint arXiv:2306.15479*, 2023.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, Kenneth Bollen, and Patrik Hoyer. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation

model. *Journal of Machine Learning Research*, 12 (Apr):1225–1248, 2011.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

A Additional Synthetic Evidence

Trend diagnostics for Figure 4. To avoid over-interpreting condition-level wiggles, Table 2 summarizes monotonicity diagnostics over the 18 panel-slices. The dominant regime is non-monotone (10/18), and 12/18 slices have overlapping $d = 15$ and $d = 20$ intervals within one standard deviation. This supports treating Figure 4 as a heterogeneous condition map rather than a single scaling trajectory.

Table 2: Trend diagnostics for PredCoM F1 curves in Figure 4.

Statistic	Count
Monotone increasing	7
Monotone decreasing	1
Non-monotone	10
d=15 vs d=20 overlap (1 s.d.)	12

Table 3: Dimension-effect sign summary ($d = 20$ vs $d = 10$) for PredCoM F1 across the 18 condition slices.

Statistic	Count
Point estimate > 0	11
Point estimate < 0	7
95% CI entirely > 0	9
95% CI entirely < 0	3
95% CI overlaps 0	6

Baseline complete-case vs imputed. Table 4 and Figure 5 quantify how complete-case controls degrade relative to imputed baselines. Most method-level deltas are negative, confirming that row dropping reduces usable statistical signal even when the downstream learner itself is unchanged.

Table 4: Baseline complete-case vs imputed (aggregate).

Method	$F1_{\text{imp}}$	$F1_{\text{cc}}$	$\Delta F1_{\text{cc-imp}}$	ρ_{cc}
Peter Clark	0.452	0.289	-0.163	5.2%
sortnregress	0.791	0.736	-0.056	5.2%
DirectLiNGAM	0.447	0.400	-0.047	5.3%
NOTEARS	0.254	0.240	-0.014	5.3%
LiM	0.521	0.520	-0.002	5.2%

PredCoM masked vs complete-case. Table 5 and Figure 6 isolate PredCoM's missing-data mechanism. The central comparison is within-method: masked inference retains partially observed rows and consistently outperforms the complete-case control when retention collapses.

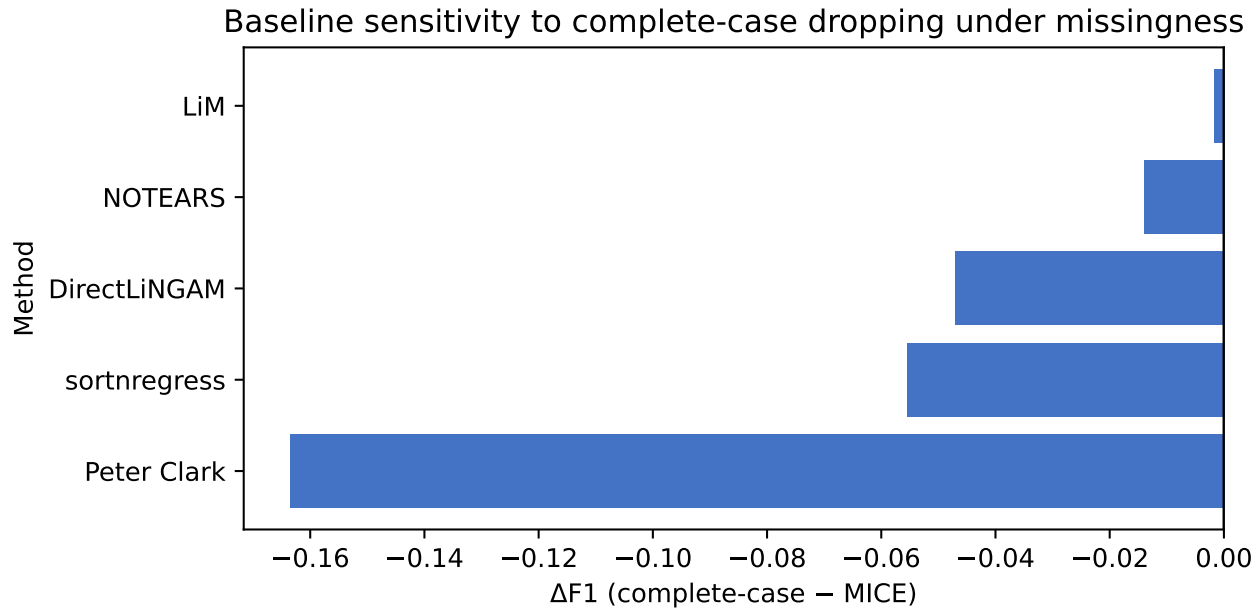

Figure 5: Baseline ΔF1 (complete-case minus imputed).

Table 5: PredCoM masked vs complete-case (aggregate by d and missingness).

d	Missingness	$F1_{\text{masked}}$	$F1_{\text{cc}}$	$\Delta F1_{\text{cc}-\text{masked}}$	ρ_{cc}
10	MAR	0.629	0.407	-0.221	5.9%
10	MCAR	0.654	0.569	-0.086	10.3%
15	MAR	0.693	–	–	1.2%
15	MCAR	0.704	–	–	3.5%

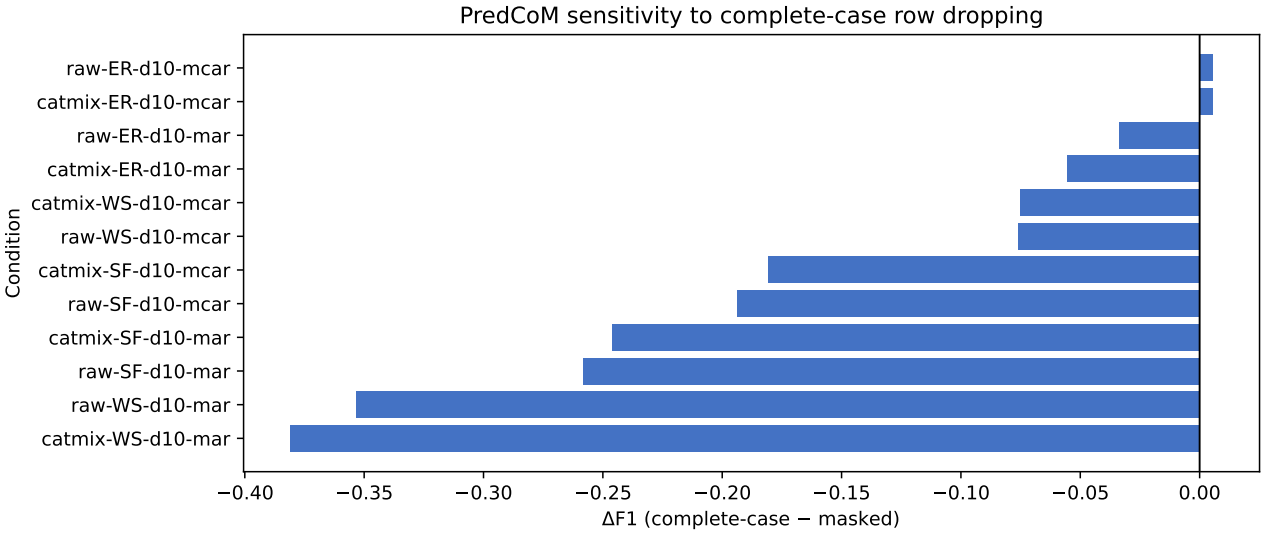

Figure 6: PredCoM ΔF1 (complete-case minus masked).

B Sweep Evidence for OOD-Style Missingness Stress

Rate sweep. Table 6 gives the full numerical breakdown for the rate sweep summarized in main-text Figure 2. At high rates ($r \geq 0.30$), ρ_{cc} reaches the low-10^{-2} regime and complete-case deltas become sharply negative.

Table 6: Missingness-rate sweep summary.

r	Baseline ρ_{cc}	Baseline $\Delta F1$	PredCoM ρ_{cc}	PredCoM $\Delta F1$
0.05	0.384 ± 0.158	-0.016 ± 0.025	0.384 ± 0.158	-0.001 ± 0.049
0.10	0.192 ± 0.103	-0.017 ± 0.032	0.192 ± 0.103	-0.077 ± 0.087
0.20	0.053 ± 0.034	-0.056 ± 0.057	0.052 ± 0.034	-0.156 ± 0.124
0.30	0.017 ± 0.013	-0.122 ± 0.096	0.016 ± 0.013	-0.480 ± 0.141
0.40	0.008 ± 0.007	-0.117 ± 0.119	0.005 ± 0.006	-0.537 ± 0.031

MAR-strength sweep. Table 7 and Figure 7 show larger penalties as mechanism hardness increases. Increasing α lowers complete-case retention and amplifies the gap between complete-case and missingness-aware paths.

Table 7: MAR-strength sweep summary.

α	Baseline ρ_{cc}	Baseline $\Delta F1$	PredCoM ρ_{cc}	PredCoM $\Delta F1$
0.5	0.065 ± 0.034	-0.034 ± 0.045	0.065 ± 0.034	-0.150 ± 0.093
1.0	0.052 ± 0.030	-0.046 ± 0.051	0.052 ± 0.030	-0.116 ± 0.093
2.0	0.037 ± 0.025	-0.065 ± 0.072	0.036 ± 0.026	-0.221 ± 0.138
3.0	0.031 ± 0.023	-0.057 ± 0.100	0.028 ± 0.024	-0.369 ± 0.125
4.0	0.029 ± 0.023	-0.078 ± 0.089	0.024 ± 0.023	-0.353 ± 0.115

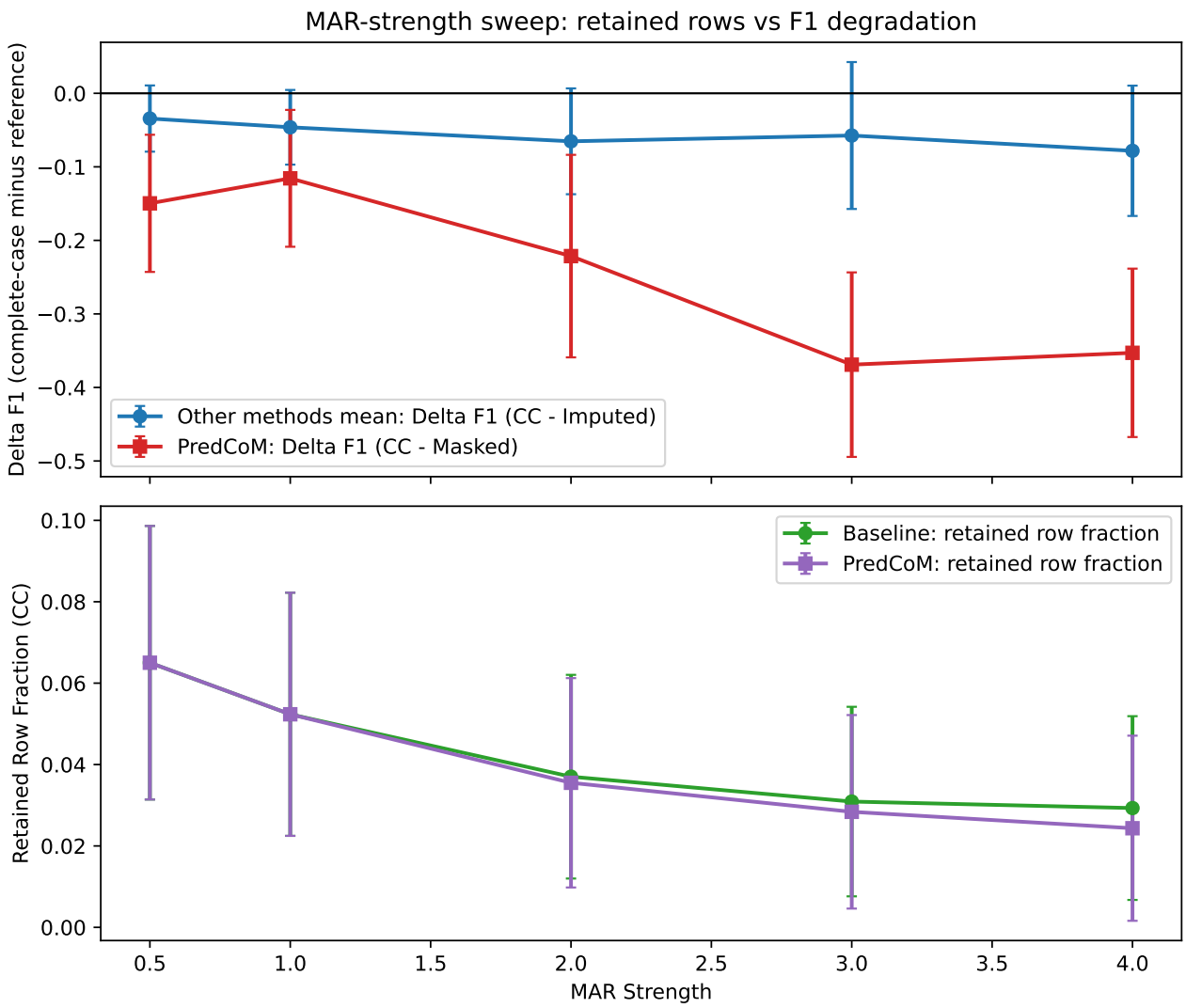

Figure 7: Retained-row fraction and complete-case F1 deltas across MAR strengths (mean ± sd).

C Runtime and Real-World Diagnostics

CPU-fair profiling. Table 8 and Figure 8 summarize runtime and memory under CPU-only execution with method-normalized optimization units. The spread confirms that predictive performance differences are not trivially explained by one method receiving disproportionate compute per update.

Real-world discovery outputs. Table 9 reports German Credit and Boston Housing diagnostics. These results are descriptive (no ground-truth DAG), but show that all compared methods return valid graph hypotheses on mixed observational data, with distinct sparsity and orientation patterns.

Table 8: CPU-fair runtime and memory summary.

Dataset	Method	Runtime / unit (s)	Runtime / unit / 1k (s)	Peak memory (MB)	Unit
CATMIX-ER, d=20, complete	DirectLiNGAM	0.724	0.362	1245.342	full run
CATMIX-ER, d=20, complete	LiM	1.448	0.724	1245.342	optimizer iter
CATMIX-ER, d=20, complete	NOTEARS	2.587	1.293	1243.706	optimizer iter
CATMIX-ER, d=20, complete	Peter Clark	3.085	1.542	1245.342	full run
CATMIX-ER, d=20, complete	PredCoM	0.043	0.022	1239.893	epoch
CATMIX-ER, d=20, complete	mCMIkNN	170.724	85.362	1245.342	full run
CATMIX-ER, d=20, complete	sortnregress	0.065	0.032	1245.342	full run
RAW-ER, d=20, complete	DirectLiNGAM	0.711	0.356	1244.546	full run
RAW-ER, d=20, complete	LiM	1.439	0.720	1244.546	optimizer iter
RAW-ER, d=20, complete	NOTEARS	2.555	1.277	1242.428	optimizer iter
RAW-ER, d=20, complete	Peter Clark	2.981	1.490	1243.854	full run
RAW-ER, d=20, complete	PredCoM	0.043	0.022	1238.643	epoch
RAW-ER, d=20, complete	mCMIkNN	171.465	85.732	1244.546	full run
RAW-ER, d=20, complete	sortnregress	0.082	0.041	1244.546	full run

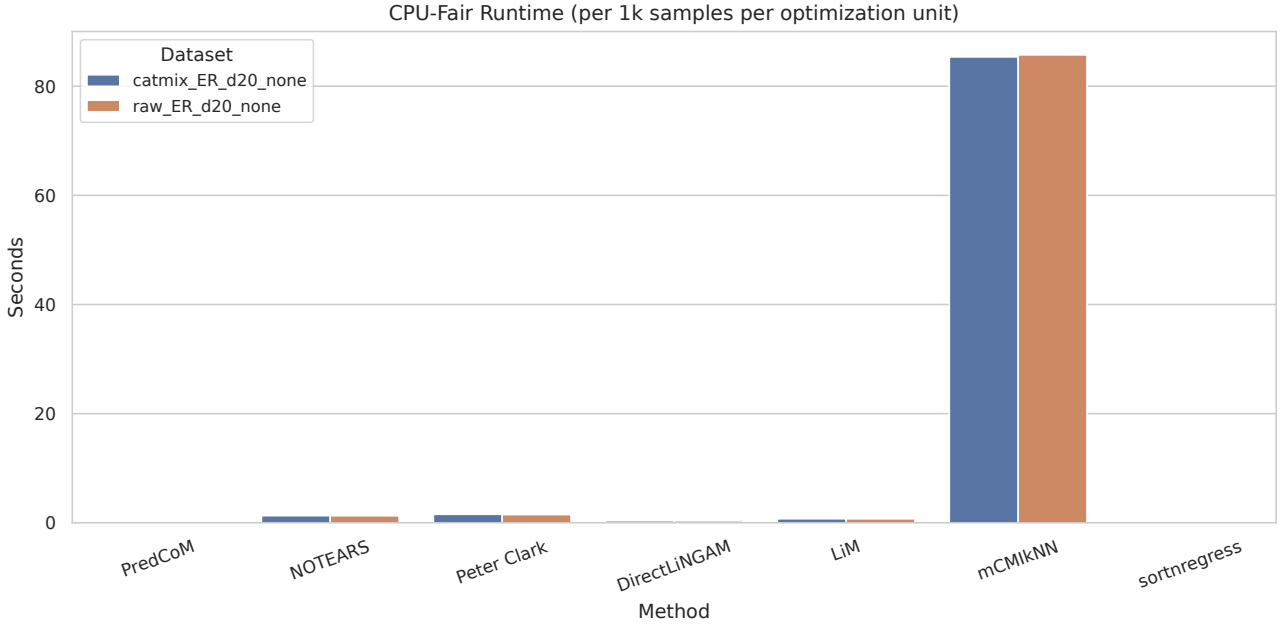

Figure 8: Runtime per optimization unit under CPU-fair execution.

Table 9: Real-world DAG diagnostics (German Credit and Boston Housing).

Dataset	Method	Runtime (s)	Edges (raw)	Edges (DAG)	Cycle in raw	Edges removed to DAG
Boston Housing	PredCoM	14.437	30.000	29.000	1.000	1.000
Boston Housing	NOTEARS	30.546	17.000	17.000	0.000	0.000
Boston Housing	Peter Clark	0.379	20.000	20.000	0.000	0.000
Boston Housing	DirectLiNGAM	0.113	8.000	8.000	0.000	0.000
Boston Housing	LiM	17.035	24.000	24.000	0.000	0.000
Boston Housing	sortnregress	0.020	27.000	27.000	0.000	0.000
German Credit	PredCoM	40.462	63.000	63.000	0.000	0.000
German Credit	NOTEARS	200.497	167.000	165.000	1.000	2.000
German Credit	Peter Clark	4.142	66.000	55.000	1.000	11.000
German Credit	DirectLiNGAM	0.724	26.000	26.000	0.000	0.000
German Credit	LiM	0.725	200.000	104.000	1.000	96.000
German Credit	sortnregress	0.058	24.000	24.000	0.000	0.000