SCORE-EM : ORDERING-BASED CAUSAL DISCOVERY WITH MISSING-AT-RANDOM DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from observational data is fundamental to machine learning and statistics, yet its application to real-world datasets is often complicated by the prevalence of missing values. Among ordering-based approaches, SCORE (Rolland et al., 2022) estimates a topological ordering using score-function identities under an additive-noise model, but cannot be directly applied to incomplete data. We introduce SCORE-EM, an EM-inspired framework for causal discovery under missing-at-random (MAR) mechanisms that jointly refines missing-value estimates and causal structure ordering. In the conditional-sampling step, missing entries are updated using Metropolis-adjusted Langevin algorithm (MALA) transitions targeting the conditional distribution induced by the current additive-noise model. In the structural-refinement step, the model is refitted and the causal ordering is updated using SCORE statistics aggregated across the resulting ensemble of completed datasets. Unlike structure-agnostic imputation, this procedure uses the evolving structural model to guide missing-value estimation, while the refined imputations provide increasingly informative data for structure learning. A subsequent multiple-imputation CAM pruning stage accounts for imputation uncertainty during edge selection. Empirical results on synthetic and real world data demonstrate that SCORE-EM improves causal ordering and graph recovery over competing missing-data baselines.

1 INTRODUCTION

Causal discovery, the task of inferring graph structures that model cause-and-effect mechanisms governing data generation, remains a cornerstone challenge in machine learning. Its utility spans diverse domains, including recommender systems (Xu et al., 2024), bioinformatics (Zhu et al., 2024), cyber-physical systems (Ghazi et al., 2025) and climatology (Nichol et al., 2025). Although randomized controlled trials serve as the gold standard for identifying causal links, interventional experiments are often impractical due to prohibitive costs, ethical constraints, or physical limitations. Consequently, considerable research has been shifted toward uncovering causal structures directly from observational data.

Most causal discovery methods assume that all variables are recorded for every sample. This assumption is frequently violated in real observational data, where missing values arise systematically and may depend on other observed variables. Under a missing-at-random (MAR) mechanism (Rubin, 1976), the probability that an entry is missing may depend on the observed portion of the data but not on the missing entries themselves. Consequently, the two default remedies — deleting incomplete samples or replacing missing entries by conditional means — are both inadequate. Complete-case analysis discards a non-representative subsample and therefore induces selection bias whenever the response indicator is associated with the variables of interest. Mean or regression imputation deflates conditional variance and attenuates the higher-order derivative information on which score-based procedures depend. In either case, the statistical quantities consumed by a causal discovery algorithm are distorted in a manner that is not correctable at the level of the final graph.

We study the problem of recovering a directed acyclic graph (DAG) from data generated by a nonlinear additive-noise model (ANM) and subsequently corrupted by MAR missingness. Among existing approaches designed for causal discovery with incomplete data, MissScore (Liu et al., 2025) extends SCORE-style ordering to the missing-data setting. MissScore estimates density-score quantities us-

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.

ing neural networks and, under MAR, additionally relies on estimated observation propensities. In contrast, SCORE-EM retains the kernel-based Stein estimator underlying SCORE and embeds the ordering procedure within an expectation–maximization framework. Given the current completed samples, the required score quantities are estimated using a regularized kernel solution, avoiding neural-network parameterization and stochastic neural-network optimization during the ordering stage. Moreover, under the MAR ignorability assumption, SCORE-EM does not require a separately estimated propensity model. These estimates are then used to construct SCORE's Hessian-based leaf statistic and iteratively recover a topological ordering.

MissDAG (Gao et al., 2022) also addresses causal discovery with missing data through an EM-style latent-variable framework and is therefore closely related to our approach. However, its M-step performs likelihood-based structure learning while enforcing a global DAG constraint. In contrast, SCORE-EM couples the missing-data estimation procedure specifically with SCORE's topological order recovery. This separation allows SCORE-EM to combine a latent-variable treatment of missingness with ordering-based causal discovery while avoiding both propensity-weighted neural score estimation and repeated global DAG-constrained optimization. Empirically, SCORE-EM achieves improved structural recovery and ordering accuracy relative to the considered missing-data baselines.

We build on SCORE (Rolland et al., 2022), which uses statistical score-function (Stein, 1972) identities to recover the ordering of variables in the causal system by iteratively identifying and removing leaf variables. SCORE declares as a leaf the variable with the smallest variation of this quantity across samples, since this variation is small for true leaves given that their behavior does not depend on the values of their descendants. SCORE estimates this quantity using a kernel-based statistical estimator, whose accuracy depends on subtle structural properties of the underlying data distribution. This is precisely the setting in which artificially filling in missing values tends to be most harmful.

Contribution. We introduce SCORE-EM, an EM-inspired alternating procedure motivated by expectation–maximization (Dempster et al., 1977) that couples Score matching based causal ordering with model-based conditional sampling of missing values. The method maintains an ensemble of M completed datasets, whose missing entries are repeatedly updated under the current structural model. This design is motivated by SCORE's leaf criterion, which depends on second derivatives of the log-density and can therefore be sensitive to imputation-induced distortions of its curvature. Point imputation, such as missForest (Stekhoven & Bühlmann, 2012), does not represent conditional uncertainty in missing entries and can distort the distributional features used for ordering. Multiple imputation (Rubin, 1988) represents this uncertainty, but its suitability for SCORE depends on whether the imputation model adequately captures the joint distribution and its relevant curvature. In a conventional impute-then-discover pipeline, the imputation model is not revised using the subsequently estimated causal structure. In contrast, SCORE-EM alternates between MALA transitions targeting $p(\mathbf{X}_{\mathrm{mis}} \mid \mathbf{X}_{\mathrm{obs}}; \hat{\theta}^{(t)})$, structural-model refitting, and ordering refinement using pooled SCORE statistics. This feedback allows the evolving structural model to guide missing-value estimation while retaining variability across completions.

2 Related Work

Causal discovery with complete data. Causal discovery aims to recover a directed acyclic graph from observational data, typically by imposing assumptions on the underlying data-generating process. We focus on additive noise models (Peters et al., 2014), in which each variable is generated as a nonlinear function of its causal parents plus an independent noise term. Under Gaussian noise and suitable differentiability assumptions, SCORE (Rolland et al., 2022) identifies leaf nodes using the variability of the diagonal entries of the log-density Hessian and recursively constructs a causal ordering. To improve scalability, DiffAN (Sanchez et al., 2023) employs a diffusion-based score model that enables more efficient Hessian estimation for datasets with larger sample sizes and dimensions. CAM (Bühlmann et al., 2014) learns nonlinear additive causal models by combining order-based structure search with regression-based edge pruning. Continuous optimization methods instead formulate DAG learning as a differentiable optimization problem. NOTEARS-MLP (Zheng et al., 2020) extends NOTEARS (Zheng et al., 2018) to nonlinear relationships using multilayer

perceptrons and a smooth acyclicity constraint, whereas DAGMA (Bello et al., 2022) introduces a log-determinant acyclicity characterization that supports efficient gradient-based structure learning.

Causal discovery with incomplete data. Existing approaches include extensions of the PC algorithm that use admissible observed samples while correcting biases in conditional-independence tests (Gain & Shpitser, 2018; Tu et al., 2019). While such constraint-based corrections extend naturally to the missing-data setting, they inherit the scalability and reliability limitations of CI-testing in high-dimensional or nonlinear regimes. We therefore restrict our empirical comparison to methods sharing our continuous-optimization/ordering-based formulation, which constitute the dominant paradigm for the settings considered in this work.

MissScore (Liu et al., 2025) extends denoising score matching to incomplete data by directly estimating first- and second-order density scores from partially observed samples. Under MAR, it uses inverse-probability weighting (IPW), with observation probabilities estimated using logistic regression, and applies the variance of the estimated Hessian diagonal to iteratively recover a SCORE-style topological ordering, followed by CAM pruning (Bühlmann et al., 2014). While this avoids explicit single imputation, its MAR correction depends on accurate estimation of the missingness propensities; small or misspecified probabilities can substantially increase the variance of the score estimates, a limitation also observed empirically by the authors.

Expectation Maximization (Dempster et al., 1977) based methods instead alternate between inferring missing values and learning the causal structure. Notable examples include MissDAG (Gao et al., 2022), which combines EM with continuous DAG optimization, such as NOTEARS (Zheng et al., 2020) and DAGMA (Bello et al., 2022), for identifiable additive noise models. Its E-step computes, or approximates using Monte Carlo EM, the expected complete-data likelihood, while the M-step applies a likelihood-based causal discovery method with an explicit DAG constraint. Although this directly couples missing-data inference with structure learning, the M-step requires repeatedly solving a generally non-convex DAG-constrained optimization problem and relies on an explicit likelihood specification for the structural equations and noise distributions. Moreover, Monte Carlo approximation of the latent-data posterior can become computationally demanding for nonlinear models.

MissOTM (Vo et al., 2024) approaches causal discovery with missing data through optimal transport. It formulates structure learning as a distribution-matching problem, seeking a causal model whose induced distribution minimizes the Wasserstein distance to the observed incomplete-data distribution. This formulation avoids relying on a separate imputation stage and can be combined with existing causal discovery methods designed for fully observed data.

3 PRELIMINARIES

3.1 ADDITIVE NOISE MODELS

We adopt the formalism of the structural causal model (SCM) (Peters et al., 2014) to characterize the causal relations among the variables under study. An SCM is a triple $\mathcal{M} = \langle \mathbf{Z}, \mathbf{X}, \mathcal{F} \rangle$, comprising a set of exogenous variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_d\}$, a set of endogenous variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$, and a collection of structural assignments $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$. Each assignment f_i determines the endogenous variable X_i from its causal parents $\mathrm{Pa}_i \subseteq \mathbf{X} \setminus \{X_i\}$ together with the associated exogenous variable Z_i, that is, $X_i = f_i(\mathrm{Pa}_i, Z_i)$.

Throughout this work we focus on a particular subclass of SCMs, namely the additive noise models (ANMs) (Peters et al., 2014), in which the exogenous term enters additively:

$$X_i \;=\; f_i(\mathrm{Pa}_i) + Z_i, \qquad i = 1, 2, \ldots, d, \tag{1}$$

where Z_i is interpreted as an additive noise variable. We assume throughout that Z_i is independent of every variable in Pa_i, and that the noise variables are mutually independent, i.e. $Z_i \perp\!\!\!\perp \mathbf{Z} \setminus \{Z_i\}$.

3.2 MISSING AT RANDOM

To consider the missing pattern of the data, we introduce $\boldsymbol{R} = (R_1, \ldots, R_d) \in \{0, 1\}^d$ as the corresponding missingness indicators for each variable X_j, namely, $R_j = 1$ if X_j is observed and

$R_j = 0$ otherwise. For a given missingness pattern, we partition $\boldsymbol{X}$ into observed and missing components denoted as $\boldsymbol{X}_{\text{obs}}$ and $\boldsymbol{X}_{\text{mis}}$. The missingness mechanism is said to be missing-at-random (MAR) if

$$p(\boldsymbol{R} \mid \boldsymbol{X}) = p(\boldsymbol{R} \mid \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}) = p(\boldsymbol{R} \mid \boldsymbol{X}_{\text{obs}}). \tag{2}$$

That is, conditional on the observed component $\boldsymbol{X}_{\text{obs}}$, the missingness does not depend on the unobserved values. Let θ parameterize the data model $p(\boldsymbol{X}; \theta)$ and let ψ parameterize the missingness mechanism. Under MAR and the standard parameter-distinctness condition between θ and ψ, the missingness mechanism is ignorable for likelihood-based inference (Rubin, 1976). Consequently, inference for θ can be based on the likelihood of observed data,

$$L(\theta, \psi \mid \boldsymbol{X}_{\text{obs}}, \mathbf{R}) = \int L(\theta \mid \boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}) p(\mathbf{R} \mid \boldsymbol{X}) d\boldsymbol{X}_{\text{mis}} = p(\mathbf{R} \mid \boldsymbol{X}_{\text{obs}}) L(\theta \mid \boldsymbol{X}_{\text{obs}}) \tag{3}$$

without explicitly modeling $p(\boldsymbol{R} \mid \boldsymbol{X}, \psi)$ given $p(\boldsymbol{X}_{\text{obs}}; \theta) = \int p(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}; \theta) d\boldsymbol{X}_{\text{mis}}$. This ignorability property lays the foundation for the Expectation-Maximization (EM) style (Dempster et al., 1977) treatment of missing values in this work.

3.3 SCORE and Causal Ordering

Rolland et al. (2022) show that the Stein score, defined as the gradient of the log-density, can be exploited to identify leaf nodes in an Additive Noise Model (ANM). For a nonlinear ANM with noise density p^{ϵ}, the j-th component of the score admits the analytical expression, where $\text{Ch}(X_j)$ denotes children of node j.

$$\begin{aligned}\nabla_{X_j} \log p(\boldsymbol{X}) &= \nabla_{X_j} \log \prod_{i=1}^{d} p\big(X_i \mid Pa_i\big) = \nabla_{X_j} \sum_{i=1}^{d} \log p\big(X_i \mid Pa_i\big) = \nabla_{X_j} \sum_{i=1}^{d} \log p^{\epsilon}(X_i - f_i) \\ &= \frac{\partial \log p^{\epsilon}(X_j - f_j)}{\partial X_j} - \sum_{i \in \text{Ch}(X_j)} \frac{\partial f_i}{\partial X_j} \frac{\partial \log p^{\epsilon}(X_i - f_i)}{\partial X_i}.\end{aligned} \tag{4}$$

Under the assumption that the second derivative of the noise log-density is constant, the diagonal entries of the Jacobian of the score, equivalently the Hessian of the log-density, provide a criterion for detecting leaf nodes. In particular, for dataset $\mathcal{D}$, node j is a leaf if

$$\text{Var}_{\mathcal{D}}\left[\boldsymbol{H}_{j,j}\big(\log p(\boldsymbol{X})\big)\right] = 0. \tag{5}$$

That is, the corresponding diagonal Hessian entry remains constant across samples in $\mathcal{D}$. This characterization enables the causal ordering to be recovered recursively by identifying a leaf node, removing it from the graph, and repeating the procedure on the remaining variables. Once the causal ordering is established, spurious edges are eliminated using CAM pruning (Bühlmann et al., 2014).

3.4 Problem Formulation: Causal Discovery Under Missing Data

For a dataset of n independently and identically distributed (i.i.d) observations denoted as $\{\boldsymbol{X}^{(m)}\}_{m=1}^{n}$, each $\boldsymbol{X}^{(m)} = \big(X_1^{(m)}, \ldots, X_d^{(m)}\big)^{\top}$ is considered as a d-dimensional random vector generated according to an underlying unknown directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Following the same notation, the corresponding complete data matrix is denoted by $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and the missingness indicators by $\boldsymbol{R} \in \{0,1\}^{n \times d}$, where $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ otherwise. Missing-at-random (MAR) mechanism is assumed throughout the study. Our objective is to recover a causal ordering $\pi = (\pi_1, \ldots, \pi_d)$ compatible with $\mathcal{G}$ and subsequently identify the directed edges of the DAG. Following the ordering-based formulation of SCORE, SCORE-EM first estimates the causal ordering π, followed by an edge-pruning step for the structure recovery as the candidate parents of X_{π_j} are restricted to variables appearing earlier in the order.

4 SCORE-EM

In this study, we introduce SCORE-EM, an EM-inspired alternating procedure that combines SCORE-based causal ordering with model-based sampling of missing values. The method alternates between approximate conditional sampling under a fitted additive noise model and updates of

the structural functions and causal ordering. The EM terminology reflects this conditional-sampling and model-refitting structure.

A key motivation for our EM-based formulation is that generic imputation can distort the distributional properties required by the leaf-identification criterion of SCORE. In particular, point imputation can artificially reduce residual variability and destroy the constant-curvature property of leaf nodes, while misspecified multiple imputation can induce mixture residual distributions whose Hessian diagonal has nonzero variance. These distortions may therefore affect the reliability of SCORE's ordering procedure. Our approach instead updates missing values using samples from the conditional distribution induced by the current structural model, thus more closely preserving the distributional assumptions underlying SCORE. A detailed illustrative example with three variables, showing these effects in contrast to imputing the data beforehand using MissForest (Stekhoven & Bühlmann, 2012) and multiple imputation, is provided in Appendix 13.

4.1 Ordering Phase

We estimate a causal ordering through alternating conditional sampling and structural refinement steps that explicitly account for missing values. Since the missing entries are unknown, we maintain an ensemble of M completed datasets, where the observed entries are fixed to their original values while the missing entries may differ across imputations.

The ordering procedure alternates between estimating a structural model and updating the missing values. We initialize the ensemble by mean-imputing the missing entries, refining them through two precision-matrix-based linear prediction steps, and adding Gaussian noise scaled by the observed residual standard deviation to generate multiple completed datasets. An initial causal ordering is then obtained by applying SCORE to this imputation ensemble. This ordering defines a provisional parent structure, based on which an additive noise model (ANM) is fitted. The collection of structural functions and noise variances is denoted by $\theta = (f_i, \sigma_i^2)_{i=1}^d$, which is estimated from the current ensemble of completed datasets and define the likelihood used to update the missing entries. Algorithm 1 shows the pseudocode for the ordering phase.

4.1.1 Conditional Sampling of Missing Values

Given the current structural parameters θ, the E-step updates the missing entries by sampling from the conditional distribution induced by the fitted ANM. For a completed dataset $\mathbf{X}^{(m)}$, the corresponding log-density is

$$-\frac{1}{2}\sum_{i=1}^{d}\left[\left(X_i^{(m)} - f_i(X_{\mathrm{pa}(i)}^{(m)})\right)^2/\sigma_i^2 + \log(2\pi\sigma_i^2)\right]. \tag{6}$$

We update the missing entries using Metropolis-adjusted Langevin dynamics (MALA) (Girolami & Calderhead, 2011). Let $\mathbf{X_{mis}}$ denote the current missing values. A Langevin proposal is

$$\mathbf{X}'_{\mathrm{mis}} = \mathbf{X}_{\mathrm{mis}} + \frac{\epsilon^2}{2}\nabla_{\boldsymbol{X}_{\mathrm{mis}}}\log p(\mathbf{X} \mid \theta) + \epsilon\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{7}$$

where ϵ is the step size and observed entries remain fixed. Because the gradient is computed under the fitted ANM, the proposal incorporates the learned structural dependencies when updating missing values; a Metropolis–Hastings correction yields samples approximating $p(\mathbf{X}_{\mathrm{mis}} \mid \mathbf{X}_{\mathrm{obs}}, \theta)$. Further details of the gradient and sampling procedure are provided in Appendix 9.1.

4.1.2 Structural Model Refitting and Ordering Refinement

After updating the missing values, we recompute SCORE's leaf statistics separately within each imputed dataset and average them across the ensemble, selecting the variable with the smallest pooled statistic as the next leaf. The resulting ordering defines a shared parent structure, after which the ANM is refitted across imputations and used in the subsequent conditional-sampling step; additional details on edge aggregation and model fitting are provided in Appendix 9.4.

Stopping criteria and relation to EM. We terminate SCORE-EM when the pooled SCORE statistics, imputation chains, and ensemble-averaged complete-data log-likelihood stabilize, or when

Algorithm 1 Ordering Phase of SCORE-EM

Require: Incomplete data X, imputations M, maximum sweeps T
Ensure: Estimated ordering $\widehat{\pi}$

```
Initialize completion ensemble X^(0) with observed entries fixed
Estimate π^(0) using pooled SCORE statistics
Construct shared parent sets P^(0) consistent with π^(0) and fit ANM θ^(0)
for t = 1, ..., T do
    Conditional sampling: Obtain X^(t) by applying MALA to missing entries under the current ANM, keeping observed entries fixed
    Model refitting: Set π^(t) ← π^(t−1), P^(t) ← P^(t−1), and refit θ^(t) on X^(t)
    Ordering proposal: Estimate π̃^(t) using pooled SCORE statistics
    if π̃^(t) ≠ π^(t) then
        Construct candidate parent sets P̃^(t) and fit candidate ANM θ̃^(t)
        Evaluate candidate and incumbent complete-data log-likelihoods on X^(t)
        if the candidate passes the likelihood safeguard then
            (π^(t), P^(t), θ^(t)) ← (π̃^(t), P̃^(t), θ̃^(t))
        end if
    end if
    if the prescribed stopping diagnostics are satisfied then
        break
    end if
end for
return π̂ obtained using pooled SCORE statistics
```

the maximum number of sweeps is reached; detailed diagnostics are provided in Appendix 9.2. Unlike conventional EM, SCORE-EM does not optimize a common likelihood-based objective: its structural update is driven by SCORE's Hessian-based leaf-identification criterion, while the fitted ANM provides the conditional distribution used to update missing values. A detailed discussion contrasting SCORE-EM with conventional EM can be found in Appendix 9.3.

4.2 CAM Pruning with Multiple Imputation

The final stage of SCORE (Rolland et al., 2022) recovers a sparse DAG by pruning candidate edges that are consistent with the estimated causal ordering. Standard CAM pruning (Bühlmann et al., 2014), however, requires a fully observed dataset and therefore cannot be applied directly under MAR missingness. We address this limitation by introducing a multiple-imputation-based adaptation of CAM pruning. Given the recovered order, we perform CAM pruning across multiple stochastic completions. More details can be found in Appendix 10.

5 Experiments

5.1 Synthetic Data

Our data generation process follows the setup introduced by Sanchez et al. (2023). We represent the causal mechanisms f_i as functions sampled from a Gaussian Process (GP) with a radial basis function (RBF) kernel and unit bandwidth. Data are produced under additive noise models, with noise scales drawn uniformly from $[0.4, 0.8]$. Prior work has shown that assuming Gaussian noise yields identifiable causal models (Peters et al., 2014).

We construct the ground-truth causal graphs using two common random graph families: Erdős–Rényi (ER) and Scale-Free (SF). For each node count d, we adjust graph density by using two target expected edge totals, d and $4d$. We label each synthetic dataset as $[d]$[graph type][degree], where d is the number of nodes, graph type is ER or SF, and degree denotes the expected number of edges. The number of nodes has been varied in $d \in \{10, 20, 30\}$. For every setting, we generate datasets with 1000 samples.

To inject MAR missingness, we impose missing-value rates of 30% and 50% using a standard MAR mechanism. The missingness is generated via a logistic model following the same protocol as Liu et al. (2025). We first choose a random subset of variables to be fully observed; the remaining variables then have missing entries that depend only on this fully observed subset. Concretely, we sample random logistic coefficients and rescale the resulting probabilities so that the affected variables attain the desired overall missingness level.

5.2 Real World Data

For a real world causal discovery problem with known ground truth, we ran experiments on the widely used Sachs protein signalling dataset (Sachs et al., 2005). We introduced missing values using the known causal DAG, processing variables in topological order and keeping root variables fully observed. For each non-root variable X_j, we randomly selected one causal parent $X_{k(j)}$. In samples where that parent was observed, we masked X_j with probability

$$P(R_j = 0 \mid X_{k(j)}) = \sigma\Big(\alpha_j + \beta_j \widetilde{X}_{k(j)}\Big),$$

where $R_j = 0$ denotes missingness, $\widetilde{X}_{k(j)}$ is the standardized parent value, and β_j is drawn from a zero-mean Gaussian distribution. We calibrated α_j to approximately achieve the specified marginal missing rate for X_j. If the selected parent was missing in a sample, X_j was left observed, so masking decisions depended only on observed values.

5.3 Metrics

We evaluate the performance of the considered causal discovery methods using three metrics: Structural Hamming Distance (SHD), F1 score, and order divergence (Rolland et al., 2022). SHD measures the number of edge additions, deletions, or reversals required to transform the estimated graph into the ground-truth DAG, with lower values indicating better structural recovery. The F1 score summarizes the accuracy of edge identification by balancing precision and recall, with higher values being better. Order divergence measures the discrepancy between the estimated causal ordering and the ground-truth ordering.

5.4 Baselines

We evaluate SCORE-EM relative to five baseline methods. We consider the following methods: MissScore (Liu et al., 2025), MissDAG (Gao et al., 2022), MissOTM (Vo et al., 2024), MissForest + SCORE (MissForest imputation (Stekhoven & Bühlmann, 2012) followed by SCORE ordering, then CAM pruning), and MI+SCORE(multiple imputation (Rubin, 1988) followed by SCORE ordering, then CAM pruning). The implementation details of SCORE-EM and baselines are described in Appendix 8. The source code for the SCORE-EM implementation is available through this anonymous repository: https://anonymous.4open.science/r/scorem-iclr-3704/

For reference, we additionally report the performance of SCORE applied to the corresponding fully observed datasets, which serves as an oracle benchmark for assessing the degradation induced by missingness. We do not compare against constraint-based approaches such as Gain & Shpitser (2018) and Tu et al. (2019), as these differ fundamentally in identifiability assumptions and are known to scale poorly with dimensionality.

6 Results

6.1 Synthetic Data

Table 1 presents the associated results for $d = 20$ with a 30% missingness rate. Additional results for $d = 20$ under 50% missingness, as well as for $d \in \{10, 30\}$ for 30% and 50% missingness are provided in the Appendix. SCORE-EM achieves the highest mean F1 across all eight graph and missingness settings among methods using incomplete data, while consistently obtaining the lowest mean order divergence among the corresponding ordering-based baselines. At 30% missingness, it also achieves the lowest mean SHD in every graph setting, with particularly strong F1 scores of

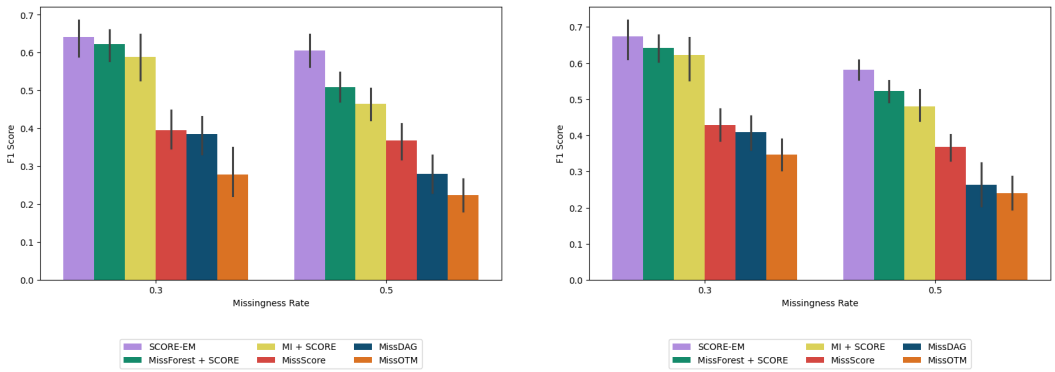
Table 1: Performance summary across graph topologies ($d = 20$, missing rate=0.3). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference. Order divergence is reported for ordering based methods only

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.593 ± 0.132	17.600 ± 5.038	2.800 ± 1.619
		MissDAG	0.584 ± 0.145	10.900 ± 2.378	-
		MissForest + SCORE	0.639 ± 0.121	15.000 ± 3.887	2.300 ± 1.337
		MissOTM	0.557 ± 0.139	12.600 ± 3.565	-
		MissScore	0.372 ± 0.071	35.700 ± 13.149	4.600 ± 2.366
		SCORE-EM	$\mathbf{0.770 \pm 0.097}$	$\mathbf{10.100 \pm 3.957}$	$\mathbf{0.300 \pm 0.483}$
		SCORE (Fully Observed)	0.969 ± 0.032	1.150 ± 1.348	0.150 ± 0.366
	4	MI + SCORE	0.589 ± 0.106	49.000 ± 9.487	7.100 ± 5.567
		MissDAG	0.386 ± 0.086	60.900 ± 9.036	-
		MissForest + SCORE	0.621 ± 0.096	46.000 ± 9.707	5.900 ± 4.606
		MissOTM	0.278 ± 0.110	67.900 ± 7.838	-
		MissScore	0.395 ± 0.086	81.400 ± 11.365	19.300 ± 9.452
		SCORE-EM	$\mathbf{0.641 \pm 0.082}$	$\mathbf{44.200 \pm 8.702}$	$\mathbf{3.300 \pm 2.111}$
		SCORE (Fully Observed)	0.757 ± 0.105	29.700 ± 11.178	1.300 ± 1.218
SF	1	MI + SCORE	0.598 ± 0.084	18.000 ± 3.651	3.200 ± 1.687
		MissDAG	0.551 ± 0.129	11.900 ± 2.923	-
		MissForest + SCORE	0.681 ± 0.080	13.900 ± 3.843	2.000 ± 1.333
		MissOTM	0.546 ± 0.149	14.200 ± 5.138	-
		MissScore	0.370 ± 0.063	43.000 ± 12.401	3.400 ± 2.319
		SCORE-EM	$\mathbf{0.797 \pm 0.050}$	$\mathbf{8.600 \pm 2.547}$	$\mathbf{0.600 \pm 0.843}$
		SCORE (Fully Observed)	0.940 ± 0.059	1.900 ± 1.774	0.600 ± 0.821
	4	MI + SCORE	0.622 ± 0.102	37.700 ± 7.587	5.800 ± 3.225
		MissDAG	0.408 ± 0.084	48.800 ± 4.662	-
		MissForest + SCORE	0.654 ± 0.097	35.000 ± 7.409	4.500 ± 3.136
		MissOTM	0.346 ± 0.074	53.300 ± 4.398	-
		MissScore	0.428 ± 0.074	72.000 ± 9.707	13.600 ± 4.248
		SCORE-EM	$\mathbf{0.673 \pm 0.097}$	$\mathbf{33.700 \pm 7.454}$	$\mathbf{2.600 \pm 1.578}$
		SCORE (Fully Observed)	0.760 ± 0.106	24.400 ± 8.407	1.650 ± 1.137

0.770 and 0.797 on ER1 and SF1, respectively. These results highlight SCORE-EM's ability to recover accurate causal orderings and maintain strong graph recovery as missingness increases.

Figure 1 compares the F1 scores of the evaluated methods on the ER4 and SF4 graphs under 30% and 50% missingness for $d = 20$. SCORE-EM achieves the highest mean F1 in all four settings, obtaining approximately 0.64 and 0.60 on ER4 and 0.67 and 0.58 on SF4 at 30% and 50% missingness, respectively. Although performance generally decreases as the missingness rate increases, SCORE-EM maintains a clear advantage over the competing methods, particularly at 50% missingness. These results indicate that the proposed approach remains robust even under substantial data incompleteness.

Effect of EM-inspired ordering. Since both SCORE-EM and MI + SCORE employ multiple-imputation-based CAM pruning, their primary difference lies in the ordering stage. We therefore use MI + SCORE as an ablation baseline to assess the contribution of the proposed EM-inspired ordering procedure. The improvements of SCORE-EM in F1, SHD, and order divergence suggest that its gains cannot be attributed solely to multiple imputation during pruning. The comparison with MissForest + SCORE further supports the benefit of coupling missing-value estimation with structural refinement. MissForest + SCORE applies prediction-based imputation before causal ordering, without revising the imputations using the estimated structural model. In contrast, SCORE-EM alternates between MALA updates of missing entries under the fitted ANM, structural-model refitting, and SCORE-based ordering refinement. Its completion ensemble thus represents uncertainty in the missing entries under an evolving structural model. These results support the effectiveness of this coupling for structural recovery and ordering accuracy.



(a) ER4 (b) SF4

Figure 1: F1 scores on (a) ER4 and (b) SF4 graphs at 30% and 50% missingness. SCORE-EM consistently achieves the highest F1 across both graph topologies and missingness levels.

Table 2: Performance on the Sachs dataset under 30% and 50% missingness (mean $\pm$ standard deviation). Order divergence is reported for ordering based methods only.

	Missing Rate = 0.3			Missing Rate = 0.5		
Model	F1 ↑	OD ↓	SHD ↓	F1 ↑	OD ↓	SHD ↓
MissDAG (Gao et al., 2022)	0.212 ± 0.105	-	14.667 ± 1.155	0.192 ± 0.093	-	15.000 ± 1.000
MissForest (Stekhoven & Bühlmann, 2012) + SCORE	$\mathbf{0.424 \pm 0.042}$	8.000 ± 0.000	14.333 ± 2.517	0.298 ± 0.042	8.667 ± 1.155	17.333 ± 4.041
MissOTM (Vo et al., 2024)	0.300 ± 0.061	-	14.667 ± 1.528	0.219 ± 0.126	-	15.333 ± 1.528
MI+SCORE	0.405 ± 0.008	8.000 ± 0.000	14.333 ± 0.577	0.302 ± 0.045	9.333 ± 1.155	15.667 ± 0.577
MissScore (Liu et al., 2025)	0.241 ± 0.185	8.667 ± 4.509	18.667 ± 4.509	0.253 ± 0.079	9.667 ± 1.528	21.667 ± 4.509
SCORE-EM	$\mathbf{0.424 \pm 0.018}$	$\mathbf{7.333 \pm 0.577}$	$\mathbf{13.333 \pm 1.155}$	$\mathbf{0.352 \pm 0.057}$	$\mathbf{7.667 \pm 0.577}$	$\mathbf{14.667 \pm 1.528}$
SCORE (Fully Observed)	0.444	8.000	12.000	0.444	8.000	12.000

6.2 Real World Data

SCORE-EM shows stronger robustness and improved graph reconstruction performance across different degrees of missing data in the Sachs protein signaling dataset. Table 2 highlights this under moderate (30%) and severe (50%) data incompleteness. At 30% missingness, SCORE-EM attains the lowest order divergence (7.333) and lowest SHD (13.333) while matching the top F1 score (0.424) of MissForest+SCORE. Under severe missingness (50%), SCORE-EM's advantage becomes even more pronounced: it outperforms all baselines across every metric, achieving an F1 of 0.352, an OD of 7.667, and an SHD of 14.667. Compared to the strongest baseline, MissForest+SCORE (0.298 F1, 17.333 SHD), this corresponds to an approximate 18% gain in F1 and a 15% reduction in SHD, while also maintaining the closest proximity to the fully observed benchmark (SHD 12.000). These trends confirm that jointly refining missing value imputations and causal ordering yields far more resilient causal structure discovery than traditional two-stage pipeline approaches, especially under extreme missingness.

7 Conclusion

We introduced SCORE-EM, an EM-inspired extension of SCORE (Rolland et al., 2022) for causal discovery with MAR missing data. The method alternates between conditional sampling of missing values under a fitted structural model and refinement of the model and causal ordering, while a multiple-imputation CAM pruning stage accounts for imputation uncertainty during edge selection. Across synthetic experiments, SCORE-EM improves causal ordering and graph recovery over competing missing-data baselines, with particularly strong performance on denser graphs. These results support incorporating structural information into missing-data estimation rather than relying solely on structure-agnostic imputation before causal discovery. Future work includes scaling to higher-dimensional and denser settings, studying richer missingness mechanisms such as MNAR, extending the approach to non-additive and mixed-type models, and developing objective-based structural updates with convergence guarantees and finite-sample uncertainty quantification.

AI USE STATEMENT

In this work, we used generative AI tools to **design or provide feedback on research methodology or experiments, implement methods, and interpret results**. We did not use generative AI tools for the other tasks requiring disclosure, and any remaining required-disclosure tasks were not applicable to this work.

Additionally, we used generative AI tools to **create or edit software code, draft parts of the research paper, discover research topics or identify gaps, brainstorm ideas, source or search for information, edit the research paper to improve readability, and propose a title or keywords for the research paper**.

We reviewed all AI-assisted work and independently verified its appropriateness for this research. In particular, methodological suggestions and interpretations were evaluated by the authors, AI-assisted code was reviewed and tested for correctness, sourced information was checked against the relevant literature or original sources, and AI-assisted manuscript text was reviewed and revised before inclusion in the final paper.

We take responsibility for the final content of this work, including all text, claims, analyses, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To enable reproducibility, we release an anonymous repository that includes the SCORE-EM implementation, the synthetic data generation pipeline, and scripts with step-by-step instructions to reproduce the reported results. The real-world dataset used in this work is publicly available (Sachs et al., 2005). Repository link: https://anonymous.4open.science/r/scorem-iclr-3704/

Additional implementation details covering model setups, hyperparameter choices, and the experimental protocol necessary to replicate our findings are provided in Section 8 of the Appendix.

REFERENCES

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8226–8239. Curran Associates, Inc., 2022. doi: 10.52202/068431-0598. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/36e2967f87c3362e37cf988781a887ad-Paper-Conference.pdf.

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Alexander Gain and Ilya Shpitser. Structure learning under missing data. In Václav Kratochvíl and Milan Studený (eds.), *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pp. 121–132. PMLR, 11–14 Sep 2018. URL https://proceedings.mlr.press/v72/gain18a.html.

Erdun Gao, Ignavier Ng, Mingming Gong, Li Shen, Wei Huang, Tongliang Liu, Kun Zhang, and Howard Bondell. Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35:5024–5038, 2022.

Youness Ghazi, Mohamed Tabaa, Mohamed Ennaji, and Ghita Zaz. Causal-bayesian modeling for risk-aware resilience in industrial cyber-physical systems. In *2025 37th International Conference on Microelectronics (ICM)*, pp. 1–4. IEEE, 2025.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011.

Wenqin Liu, Haoze Hou, Erdun Gao, Biwei Huang, Qiuhong Ke, Howard Bondell, and Mingming Gong. MissScore: High-order score estimation in the presence of missing data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38664–38691. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25aa.html.

J Jake Nichol, Michael Weylandt, G Matthew Fricke, Melanie E Moses, Diana Bull, and Laura P Swiler. Space-time causal discovery in earth system science: A local stencil learning approach. *Journal of Geophysical Research: Machine Learning and Computation*, 2(3):e2024JH000546, 2025.

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

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. An overview of multiple imputation. In *Proceedings of the survey research methods section of the American statistical association*, volume 79, pp. 84, 1988.

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.

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Idusfje4-Wq.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pp. 583–603. University of California Press, 1972.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1762–1770. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/tu19a.html.

Vy Vo, He Zhao, Trung Le, Edwin V. Bonilla, and Dinh Phung. Optimal transport for structure learning under missing data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49605–49626. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/vo24b.html.

Shuyuan Xu, Da Xu, Evren Korpeoglu, Sushant Kumar, Stephen Guo, Kannan Achan, and Yongfeng Zhang. Causal structure learning for recommender system. *ACM Transactions on Recommender Systems*, 3(1):1–23, 2024.

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.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Yuehua Zhu, Panayiotis V Benos, and Maria Chikina. A hybrid constrained continuous optimization approach for optimal causal discovery from biological data. *Bioinformatics*, 40(Supplement_2): ii87–ii97, 2024.

8 Implementation Details: Baselines

The implementation details of SCORE-EM and the baselines used in this paper are as follows:

- MissScore (Liu et al., 2025): We utilized the original authors' source code available at https://github.com/lww28/MissScore. For all experiments, we adopt the same hyperparameter settings and model configurations as those reported by the original authors. Specifically, we use a perturbation scale of $\sigma = 0.1$ and jointly optimize the objectives. The learning rate is set to 10^{-3}. The first two hidden layers have a latent dimension of 128, while the remaining three layers use a latent dimension of 1024. The model is trained for 250 epochs with a batch size of 256.
- MissOTM (Vo et al., 2024): We adopt the nonlinear MLP setup provided in the original authors' codebase at https://github.com/isVy08/OTM. We run the DAGMA algorithm for $T = 4$ central-path stages, starting from $\mu = 1.0$ with a decay rate $\alpha = 0.1$. The log-determinant hyperparameters are fixed to $(1.0, 0.9, 0.8, 0.7)$, and we use Adam with a learning rate of 2×10^{-4}. For optimization, we perform $2{,}000$ warm-up iterations in each of the first three stages and $3{,}000$ iterations in the final stage.
- MissDAG (Gao et al., 2022): We relied on the authors' original implementation available at https://github.com/ErdunGAO/MissDAG for NOTEARS-MLP-MCEM and adopted the same hyperparameter settings as reported by Gao et al. (2022). The nonlinear structural equations are parameterized with the NOTEARS-MLP model, and each M-step solves the resulting constrained optimization via an augmented Lagrangian method. We initialize the dual variable and penalty parameter as $\alpha_0 = 0$ and $\rho_0 = 0$, respectively, and set $\gamma = 0.25$, $h_{\min} = 10^{-8}$, and $\rho_{\max} = 10^{16}$. For sparsity, we use an ℓ_1 regularization weight of $\lambda_1 = 0.03$, and we additionally apply an ℓ_2 penalty with $\lambda_2 = 0.01$ to the MLP parameters. We run 10 EM iterations and derive the final adjacency matrix using an edge threshold of 0.3.
- MissForest+SCORE: We utilized Scikit learn's Iterative Imputer with Random Forest Regressor.
- Multiple imputation + SCORE: The MI + SCORE baseline generates 20 completed datasets using iterative Extra-Trees imputation with bootstrap resampling and additive Gaussian residual noise, then estimates a causal ordering by averaging the estimated Hessian-diagonal variances across imputations at each SCORE leaf-removal step. The same ensemble is reused for spline-based CAM pruning, retaining an order-consistent edge when its partial F-test is significant at $\alpha = 10^{-3}$ in at least half of the imputations.

9 Implementation Details: SCORE-EM

Our implementation is based on SCORE's (Rolland et al., 2022) original authors' implementation found at https://github.com/paulrolland1307/SCORE.

Hyperparameters. We use four parallel imputations ($M = 4$) and run the EM procedure for at most $T = 10$ iterations. The initial completed-data ensemble is refined using two burn-in iterations, each consisting of four Stein–ULA sweeps with step size 0.15. During the subsequent E-steps, each imputation is updated using $L = 15$ MALA transitions. The MALA step size is initialized at 0.4 and adapted during the first EM iteration toward a target acceptance rate of 0.574, with the step size constrained to the interval $[10^{-3}, 2]$. For the auxiliary overdispersed chains used in the convergence diagnostic, we perform $L_{\mathrm{od}} = \max(4, \lfloor L/3 \rfloor) = 5$ MALA transitions.

Convergence is assessed using the relative Stein-Hessian drift together with between-imputation variability, split-$\widehat{R}$, and the change in cross-fitted log-likelihood. We use $\tau = 2 \times 10^{-2}$ and $c = 1.5$ in the drift criterion, a split-$\widehat{R}$ threshold of 1.10, and require all convergence conditions to hold for two consecutive EM iterations. The causal order is reconsidered whenever the accumulated drift exceeds $\rho = 0.15$. During these order updates, the temporary ANM used to guide the EM procedure employs an annealed parent-selection threshold of $(0.005, 0.02, 0.02)$. Candidate parent sets are represented using additive hinge bases with three knots per predictor and are restricted to at most 16 parents per node. The structural functions are estimated using two-fold cross-fitted kernel ridge

regression with regularization parameter 10^{-2} and at most 800 training samples per fit. Gaussian-kernel bandwidths are selected using the median-distance heuristic.

9.1 Conditional Sampling Details

MALA (Girolami & Calderhead, 2011) is particularly suitable for the conditional-sampling step because the differentiable ANM provides the gradient of the conditional log-density with respect to the missing entries, enabling gradient-informed proposals rather than uninformed random-walk updates. Moreover, MALA requires the target density only up to a normalizing constant and therefore avoids direct evaluation of the generally intractable conditional normalizing constant.

The fitted ANM parameters θ affect the MALA update through both the likelihood and its gradient. In particular, the gradient with respect to a missing variable X_j contains the contribution from its own structural equation,

$$-\frac{X_j - f_j(X_{\mathrm{pa}(j)})}{\sigma_j^2},$$

as well as contributions propagated through its children. For a child X_i such that $X_j \in \mathrm{pa}(i)$, the corresponding contribution is

$$\frac{X_i - f_i(X_{\mathrm{pa}(i)})}{\sigma_i^2}\frac{\partial f_i}{\partial X_j}.$$

Thus, changes to a missing variable affect the ANM likelihood through both its own structural residual and the residuals of its descendants. The fitted functions f_i therefore allow the sampler to exploit the current causal structure when proposing missing values.

Each Langevin proposal is subsequently evaluated using the standard Metropolis–Hastings acceptance probability, which corrects for both the change in the ANM likelihood and the asymmetry of the Langevin proposal distribution. Repeated MALA transitions yield an updated ensemble of completed datasets that approximates the conditional distribution of the missing entries under the current structural model.

9.2 Stopping Criteria

Stopping diagnostics. We use practical stopping diagnostics based on the stability of the pooled SCORE statistics, agreement between imputation chains, and stabilization of the ensemble-averaged complete-data log-likelihood. Changes in the SCORE statistics are assessed relative to the within-ensemble Monte Carlo variability, while chain agreement is monitored using a split-$\hat{R}$ diagnostic on the missing entries. Iterations terminate when the prescribed criteria are simultaneously satisfied or when the maximum number of sweeps is reached. These diagnostics assess numerical stability and provide a practical stopping rule; they do not establish convergence to a stationary point of a common objective.

9.3 Relation to conventional EM

Conventional EM maximizes the observed-data likelihood by alternating between inference over the missing variables and maximization of the expected complete-data log-likelihood. SCORE-EM instead targets causal structure recovery under an additive-noise model. Its ordering update is driven by SCORE's Hessian-based leaf-identification criterion rather than solely by likelihood maximization. The fitted ANM provides a structure-aware conditional distribution for updating missing values, while the resulting completion ensemble improves the estimation of the SCORE statistics used for ordering. Consequently, SCORE-EM should not be interpreted as a general-purpose imputation procedure: its effectiveness relies on the assumed causal structure. Moreover, because the structural update is not derived from a common EM auxiliary objective, the likelihood is not guaranteed to increase monotonically at every update.

9.4 Structural Refinement Details

Rather than applying SCORE independently to each completed dataset and reconciling potentially different orderings, SCORE-EM performs ordering at the ensemble level. At each leaf-removal step,

the diagonal score-Jacobian variance is computed separately for each imputation. Let

$$V_j^{(m)}$$

denote the estimated leaf statistic for variable X_j in imputation m. We aggregate these statistics as

$$\bar{V}_j = \frac{1}{M} \sum_{m=1}^{M} V_j^{(m)},$$

and select the variable with the smallest pooled statistic as the next leaf. The selected variable is removed from all completed datasets simultaneously, yielding a single shared ordering.

Given the resulting ordering, candidate parents of each node are restricted to its predecessors. For each candidate edge, predictive importance is evaluated within every completed dataset by comparing the full additive regression model with the corresponding model after removing that predictor block. Importance scores are aggregated across imputations, and the shared parent set is obtained by thresholding the aggregated support.

Conditional on this shared structure, the ANM is refitted using cross-fitting. Kernel ridge regressions are fitted within each imputed dataset, while their predictions and gradients are averaged across imputations and residual variances are pooled. The resulting ensemble-averaged ANM is then used to define the conditional distribution employed by the next MALA update.

10 CAM Pruning with Multiple Imputation

Given the causal ordering estimated by SCORE-EM, the candidate parent set of each variable X_j is restricted to its predecessors in the ordering, denoted by $\mathrm{Pred}(j)$. To account for uncertainty in the missing values, we generate M stochastic completed datasets using iterative multiple imputation. By default, the conditional imputation models use Bayesian ridge regression with posterior sampling, producing multiple plausible realizations of the missing entries.

For each completed dataset m, we fit the additive model

$$X_j^{(m)} = \sum_{k \in \mathrm{Pred}(j)} f_{jk}^{(m)}\left(X_k^{(m)}\right) + \varepsilon_j^{(m)}, \tag{8}$$

where each component function $f_{jk}^{(m)}$ is represented using a cubic B-spline basis. Following CAM, the relevance of a candidate edge $X_k \to X_j$ is evaluated using a partial F-test that compares the full additive model with a reduced model excluding the spline block associated with X_k. Let $p_{kj}^{(m)}$ denote the resulting p-value for the candidate edge in imputation m.

We aggregate the evidence across imputations using stability voting. Specifically, the selection frequency of the edge $X_k \to X_j$ is defined as

$$s_{kj} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(p_{kj}^{(m)} < \alpha\right), \tag{9}$$

where α is the significance level used in each partial F-test. The edge is retained if

$$s_{kj} \geq \tau, \tag{10}$$

where τ is the stability threshold, set to 0.5 by default. Consequently, an edge is selected only when it is supported by at least half of the completed datasets.

11 SCORE-EM Computational Complexity

Let n denote the number of samples, d the number of variables, M the number of parallel imputations, T the number of EM iterations, and L the number of MALA transitions per E-step. Let $K \leq T + 2$ denote the number of full SCORE order estimations performed during the procedure.

A kernel-based SCORE ordering requires d successive leaf-removal operations. Each operation involves an $n \times n$ kernel solve, yielding

$$O(dn^3 + d^2n^2)$$

time for one completed dataset and

$$O\big(KM(dn^3 + d^2n^2)\big)$$

over the imputation ensemble.

The temporary ANM is fitted using kernel ridge regression. If s is the number of training samples used per fit and q is the maximum number of parents, this contributes

$$O\big(TMd(s^3 + qs^2)\big)\,.$$

Evaluating the ANM and its gradients during L MALA transitions costs

$$O(TMLndsq).$$

Thus, the ordering phase has overall complexity

$$O\big(KM(dn^3 + d^2n^2) + TMd(s^3 + qs^2) + TMLndsq\big)\,,$$

up to the lower-order convergence-diagnostic terms. In our implementation, $M = 4$, $T \leq 10$, $L = 15$, $q \leq 16$, and $s \leq 800$. The dominant cost is therefore typically the repeated kernel-based SCORE estimation, while the parent and sample caps bound the cost of the temporary ANM fits.

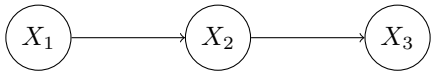
12 Wilcoxon Signed-Rank Tests

Table 3: Two-sided pairwise Wilcoxon signed-rank tests comparing SCORE-EM with the baseline methods. Δ denotes the median paired difference (SCORE-EM minus baseline). W/T/L denotes wins, ties, and losses for SCORE-EM, accounting for the direction of each metric. Holm correction is applied separately within each metric.

Metric	Baseline	N	Δ (median)	W/T/L	p	p_{Holm}	Significant
SHD	MissScore	480	-23	431/8/41	5.35×10^{-67}	2.14×10^{-66}	Yes
	MI + SCORE	480	-7	436/17/27	4.20×10^{-72}	2.10×10^{-71}	Yes
	MissOTM	480	-6	310/8/162	1.91×10^{-12}	3.83×10^{-12}	Yes
	MissForest + SCORE	480	-4	384/19/77	5.19×10^{-49}	1.56×10^{-48}	Yes
	MissDAG	480	-3	282/14/184	1.41×10^{-4}	1.41×10^{-4}	Yes
F1	MissOTM	480	0.3115	431/1/48	2.37×10^{-71}	7.12×10^{-71}	Yes
	MissDAG	480	0.2788	423/1/56	2.37×10^{-66}	4.74×10^{-66}	Yes
	MissScore	480	0.2414	453/1/26	3.46×10^{-76}	1.61×10^{-75}	Yes
	MI + SCORE	480	0.1055	451/5/24	3.23×10^{-76}	1.61×10^{-75}	Yes
	MissForest + SCORE	480	0.06931	411/5/64	4.00×10^{-63}	4.00×10^{-63}	Yes
OD	MissScore	480	-5	415/48/17	2.35×10^{-70}	7.04×10^{-70}	Yes
	MI + SCORE	480	-3	415/46/19	2.58×10^{-68}	5.16×10^{-68}	Yes
	MissForest + SCORE	480	-2	385/67/28	3.45×10^{-60}	3.45×10^{-60}	Yes

The paired Wilcoxon signed-rank tests in Table 3 show that SCORE-EM significantly outperforms all five evaluated baselines in both SHD and F1 after Holm correction within each metric. The median paired differences favor SCORE-EM in every comparison, with the largest reduction in SHD observed against MissScore ($\Delta = -23$) and the largest improvement in F1 against MissOTM ($\Delta = 0.3115$). The win/tie/loss counts further reflect the consistency of these improvements: SCORE-EM achieves lower SHD in 282–436 of the 480 paired evaluations and higher F1 in 411–453 evaluations, depending on the baseline. For order divergence, SCORE-EM also significantly outperforms all three ordering-based baselines for which this metric is available, with median paired reductions ranging from 2 to 5 and wins in 385–415 of the 480 evaluations. All reported comparisons remain statistically significant after Holm correction. Together, these results provide evidence that SCORE-EM yields more accurate graph recovery and causal ordering across the evaluated experimental settings.

13 ILLUSTRATIVE EXAMPLE: POINT IMPUTATION DISTORTS SCORE'S LEAF CRITERION



Consider the three-variable chain

$$X_1 = \epsilon_1, \qquad X_2 = X_1^2 + \epsilon_2, \qquad X_3 = X_2^2 + \epsilon_3, \qquad \epsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

The true leaf is X_3. Under the complete-data distribution,

$$\frac{\partial^2 \log p(x)}{\partial x_3^2} = -1, \qquad \mathrm{Var}\left[\frac{\partial^2 \log p(X)}{\partial x_3^2}\right] = 0,$$

whereas the corresponding variances for X_1 and X_2 are positive. Suppose half of the entries of X_3 are missing completely at random.

MissForest-style point imputation. Even if the regression function is learned perfectly, point imputation replaces missing values by

$$\widehat{X}_3 = \mathbb{E}[X_3 \mid X_1, X_2] = X_2^2.$$

Consequently, the completed residual $\widetilde{\epsilon}_3 = \widetilde{X}_3 - X_2^2$ follows

$$\widetilde{\epsilon}_3 = \begin{cases} \epsilon_3, & \text{with probability } 1/2, \\ 0, & \text{with probability } 1/2. \end{cases}$$

Half of the residuals are exactly zero, and their variance decreases from 1 to $1/2$. More fundamentally, the residual distribution is no longer Gaussian, so SCORE's constant-curvature leaf argument no longer applies. This represents the ideal conditional-mean behavior of numerical MissForest predictions.

Nonlinear multiple imputation. Suppose an imputer learns the correct nonlinear mean but overestimates the noise standard deviation:

$$\widehat{X}_3 = X_2^2 + 2Z, \qquad Z \sim \mathcal{N}(0,1).$$

The completed residual density is then

$$h(r) = \frac{1}{2}\phi_1(r) + \frac{1}{2}\phi_2(r), \qquad r = x_3 - x_2^2,$$

Here, ϕ_σ denotes the density of $\mathcal{N}(0, \sigma^2)$:

$$\phi_\sigma(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

Since X_1, X_2 remain unchanged, the completed joint density factorizes as

$$q(x) = p(x_1, x_2)h(r), \qquad r = x_3 - x_2^2.$$

The factor $p(x_1, x_2)$ does not depend on x_3, and $\partial r / \partial x_3 = 1$; therefore,

$$\frac{\partial^2 \log q(x)}{\partial x_3^2} = \frac{d^2}{dr^2} \log h(r) = \frac{h''(r)}{h(r)} - \left(\frac{h'(r)}{h(r)}\right)^2.$$

At $r = 0$, symmetry gives $h'(0) = 0$. Using $\phi_\sigma''(0) = -\phi_\sigma(0)/\sigma^2$ and $\phi_2(0) = \phi_1(0)/2$, we obtain

$$h(0) = \frac{3}{4}\phi_1(0), \qquad h''(0) = -\frac{1}{2}\phi_1(0) - \frac{1}{8}\phi_2(0) = -\frac{9}{16}\phi_1(0).$$

Consequently,

$$\left.\frac{\partial^2 \log q(x)}{\partial x_3^2}\right|_{r=0} = \frac{h''(0)}{h(0)} = -\frac{3}{4}.$$

In the tails, the wider Gaussian component dominates. More explicitly,

$$h(r) = \frac{1}{2}\phi_2(r)\left(1 + 2e^{-3r^2/8}\right),$$

so

$$\log h(r) = C - \frac{r^2}{8} + \log\left(1 + 2e^{-3r^2/8}\right),$$

where C is constant. The second derivative of the final term vanishes as $|r| \to \infty$, yielding

$$\lim_{|r|\to\infty} \frac{\partial^2 \log q(x)}{\partial x_3^2} = -\frac{1}{4}.$$

Thus, the leaf curvature is nonconstant and continuous over a residual distribution with positive density everywhere, implying

$$\mathrm{Var}_q\left[\frac{\partial^2 \log q(X)}{\partial x_3^2}\right] > 0.$$

This contrasts with the true Gaussian ANM, for which the leaf curvature is identically -1 and its variance is zero.

Scope. This example demonstrates distortion of SCORE's leaf criterion, rather than guaranteed ordering failure. The advantage of SCORE-EM is conditional on correct model estimation and sampling; correctly specified nonlinear multiple imputation would also preserve the criterion.

14 Additional Results

14.1 30% Missingness

Table 4: Performance summary across graph topologies ($d = 10$, missing rate=0.3). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference.

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.670 ± 0.116	6.400 ± 2.366	1.100 ± 0.994
		MissDAG	0.482 ± 0.284	5.900 ± 2.923	-
		MissForest + SCORE	0.725 ± 0.119	5.500 ± 2.369	0.600 ± 0.966
		MissOTM	0.422 ± 0.178	8.700 ± 3.561	-
		MissScore	0.448 ± 0.120	12.900 ± 4.358	1.600 ± 1.174
		SCORE-EM	**0.795 ± 0.144**	**3.700 ± 2.263**	**0.300 ± 0.483**
		SCORE (Fully Observed)	0.953 ± 0.086	0.700 ± 1.380	0.150 ± 0.366
	4	MI + SCORE	0.788 ± 0.069	12.500 ± 3.749	2.300 ± 1.160
		MissDAG	0.442 ± 0.055	25.400 ± 2.011	-
		MissForest + SCORE	0.808 ± 0.050	11.500 ± 2.550	1.700 ± 1.252
		MissOTM	0.347 ± 0.111	28.600 ± 3.658	-
		MissScore	0.643 ± 0.073	19.000 ± 3.399	5.800 ± 2.348
		SCORE-EM	**0.839 ± 0.046**	**10.300 ± 2.983**	**0.900 ± 0.738**
		SCORE (Fully Observed)	0.846 ± 0.057	9.600 ± 3.102	0.200 ± 0.410
SF	1	MI + SCORE	0.665 ± 0.146	6.000 ± 2.828	1.700 ± 1.160
		MissDAG	0.577 ± 0.197	5.100 ± 1.853	-
		MissForest + SCORE	0.722 ± 0.141	4.800 ± 2.573	1.300 ± 1.059
		MissOTM	0.470 ± 0.207	8.400 ± 3.777	-
		MissScore	0.548 ± 0.112	10.500 ± 5.104	0.900 ± 1.197
		SCORE-EM	**0.869 ± 0.107**	**2.600 ± 2.459**	**0.200 ± 0.422**
		SCORE (Fully Observed)	0.970 ± 0.057	0.450 ± 0.826	0.100 ± 0.308
	4	MI + SCORE	0.778 ± 0.096	8.900 ± 3.446	1.800 ± 1.033
		MissDAG	0.514 ± 0.130	16.000 ± 2.906	-
		MissForest + SCORE	0.772 ± 0.092	9.500 ± 2.915	1.500 ± 1.269
		MissOTM	0.435 ± 0.101	18.600 ± 1.897	-
		MissScore	0.574 ± 0.132	18.100 ± 4.458	4.000 ± 2.981
		SCORE-EM	**0.823 ± 0.076**	**7.500 ± 2.677**	**0.600 ± 0.843**
		SCORE (Fully Observed)	0.886 ± 0.085	4.900 ± 2.954	0.250 ± 0.444

Table 5: Performance summary across graph topologies ($d = 30$, missing rate=0.3). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference.

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.685 ± 0.072	21.200 ± 4.158	3.100 ± 2.025
		MissDAG	0.601 ± 0.070	18.300 ± 4.473	-
		MissForest + SCORE	0.683 ± 0.082	21.300 ± 5.250	3.200 ± 2.201
		MissOTM	0.543 ± 0.064	20.500 ± 5.759	-
		MissScore	0.300 ± 0.069	77.500 ± 19.609	9.600 ± 4.300
		SCORE-EM	**0.764 ± 0.050**	**15.500 ± 3.064**	**1.400 ± 1.647**
		SCORE (Fully Observed)	0.934 ± 0.043	3.750 ± 2.531	0.950 ± 1.191
	4	MI + SCORE	0.553 ± 0.082	77.400 ± 15.932	14.000 ± 4.643
		MissDAG	0.340 ± 0.088	92.800 ± 15.943	-
		MissForest + SCORE	0.583 ± 0.078	73.500 ± 15.785	11.700 ± 3.653
		MissOTM	0.221 ± 0.115	103.800 ± 16.838	-
		MissScore	0.321 ± 0.060	138.600 ± 11.462	36.600 ± 10.793
		SCORE-EM	**0.602 ± 0.070**	**70.500 ± 14.532**	**6.500 ± 3.028**
		SCORE (Fully Observed)	0.726 ± 0.098	48.800 ± 17.065	3.450 ± 2.139
SF	1	MI + SCORE	0.658 ± 0.084	22.400 ± 6.004	3.500 ± 1.958
		MissDAG	0.547 ± 0.064	18.700 ± 2.669	-
		MissForest + SCORE	0.703 ± 0.051	19.300 ± 3.974	2.200 ± 1.317
		MissOTM	0.576 ± 0.143	18.700 ± 5.498	-
		MissScore	0.273 ± 0.076	91.100 ± 20.910	5.700 ± 2.751
		SCORE-EM	**0.775 ± 0.064**	**14.900 ± 5.195**	**0.800 ± 1.033**
		SCORE (Fully Observed)	0.937 ± 0.062	3.200 ± 2.913	0.800 ± 2.215
	4	MI + SCORE	0.522 ± 0.094	74.700 ± 12.667	11.800 ± 4.237
		MissDAG	0.322 ± 0.051	85.800 ± 4.541	-
		MissForest + SCORE	0.567 ± 0.094	68.200 ± 12.656	10.700 ± 2.584
		MissOTM	0.257 ± 0.072	92.300 ± 5.355	-
		MissScore	0.339 ± 0.056	134.600 ± 16.304	27.800 ± 9.319
		SCORE-EM	**0.578 ± 0.098**	**67.000 ± 11.738**	**6.400 ± 2.319**
		SCORE (Fully Observed)	0.684 ± 0.084	50.000 ± 10.458	5.050 ± 2.502

14.2 50% MISSINGNESS

Table 6: Performance summary across graph topologies ($d = 10$, missing rate=0.5). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference.

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.404 ± 0.145	16.600 ± 3.534	2.400 ± 1.350
		MissDAG	0.296 ± 0.222	$\mathbf{7.400 \pm 2.951}$	-
		MissForest + SCORE	0.427 ± 0.149	14.800 ± 3.259	2.300 ± 1.337
		MissOTM	0.357 ± 0.129	10.300 ± 3.974	-
		MissScore	0.385 ± 0.163	14.700 ± 7.181	2.400 ± 2.366
		SCORE-EM	$\mathbf{0.577 \pm 0.131}$	11.500 ± 2.677	$\mathbf{0.500 \pm 0.707}$
		SCORE (Fully Observed)	0.953 ± 0.086	0.700 ± 1.380	0.150 ± 0.366
	4	MI + SCORE	0.683 ± 0.091	16.800 ± 3.521	7.000 ± 3.333
		MissDAG	0.243 ± 0.106	30.700 ± 2.869	-
		MissForest + SCORE	0.733 ± 0.055	14.800 ± 2.098	4.600 ± 2.221
		MissOTM	0.280 ± 0.075	30.700 ± 2.452	-
		MissScore	0.535 ± 0.137	22.900 ± 4.864	8.900 ± 3.071
		SCORE-EM	$\mathbf{0.805 \pm 0.036}$	$\mathbf{12.400 \pm 2.221}$	$\mathbf{2.200 \pm 1.476}$
		SCORE (Fully Observed)	0.846 ± 0.057	9.600 ± 3.102	0.200 ± 0.410
SF	1	MI + SCORE	0.416 ± 0.124	15.800 ± 3.938	2.600 ± 1.430
		MissDAG	0.378 ± 0.239	$\mathbf{6.600 \pm 1.776}$	-
		MissForest + SCORE	0.416 ± 0.118	16.700 ± 3.860	2.300 ± 1.567
		MissOTM	0.344 ± 0.117	10.800 ± 2.573	-
		MissScore	0.379 ± 0.151	14.400 ± 4.526	2.300 ± 1.703
		SCORE-EM	$\mathbf{0.602 \pm 0.103}$	11.000 ± 3.916	$\mathbf{0.500 \pm 0.527}$
		SCORE (Fully Observed)	0.970 ± 0.057	0.450 ± 0.826	0.100 ± 0.308
	4	MI + SCORE	0.615 ± 0.113	17.900 ± 5.343	3.800 ± 1.687
		MissDAG	0.293 ± 0.092	20.300 ± 1.703	-
		MissForest + SCORE	0.609 ± 0.083	17.800 ± 3.882	3.500 ± 0.850
		MissOTM	0.355 ± 0.103	20.800 ± 2.821	-
		MissScore	0.529 ± 0.093	20.200 ± 3.393	4.600 ± 2.171
		SCORE-EM	$\mathbf{0.697 \pm 0.079}$	$\mathbf{14.900 \pm 3.814}$	$\mathbf{1.200 \pm 1.229}$
		SCORE (Fully Observed)	0.886 ± 0.085	4.900 ± 2.954	0.250 ± 0.444

Table 7: Performance summary across graph topologies ($d = 20$, missing rate=0.5). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference.

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.411 ± 0.094	35.400 ± 7.074	4.200 ± 1.229
		MissDAG	0.472 ± 0.112	**13.000 ± 2.108**	-
		MissForest + SCORE	0.430 ± 0.129	33.600 ± 8.396	3.900 ± 1.524
		MissOTM	0.442 ± 0.128	16.000 ± 4.595	-
		MissScore	0.353 ± 0.057	39.800 ± 12.770	3.900 ± 2.025
		SCORE-EM	**0.547 ± 0.111**	28.000 ± 6.360	**0.400 ± 0.516**
		SCORE (Fully Observed)	0.969 ± 0.032	1.150 ± 1.348	0.150 ± 0.366
	4	MI + SCORE	0.465 ± 0.072	66.000 ± 6.377	17.600 ± 8.030
		MissDAG	0.280 ± 0.078	67.000 ± 8.472	-
		MissForest + SCORE	0.509 ± 0.094	61.800 ± 7.584	13.500 ± 7.764
		MissOTM	0.224 ± 0.072	71.900 ± 7.965	-
		MissScore	0.368 ± 0.084	81.900 ± 9.398	20.500 ± 6.329
		SCORE-EM	**0.606 ± 0.071**	**52.700 ± 9.604**	**4.300 ± 2.497**
		SCORE (Fully Observed)	0.757 ± 0.105	29.700 ± 11.178	1.300 ± 1.218
SF	1	MI + SCORE	0.387 ± 0.092	36.600 ± 8.847	4.500 ± 1.509
		MissDAG	0.413 ± 0.132	**14.500 ± 1.900**	-
		MissForest + SCORE	0.416 ± 0.068	34.700 ± 6.413	3.800 ± 1.229
		MissOTM	0.456 ± 0.166	16.600 ± 4.671	-
		MissScore	0.296 ± 0.061	49.000 ± 10.477	4.500 ± 2.718
		SCORE-EM	**0.542 ± 0.059**	25.400 ± 4.326	**1.600 ± 1.174**
		SCORE (Fully Observed)	0.940 ± 0.059	1.900 ± 1.774	0.600 ± 0.821
	4	MI + SCORE	0.481 ± 0.070	56.100 ± 6.118	10.700 ± 3.889
		MissDAG	0.264 ± 0.098	55.900 ± 4.228	-
		MissForest + SCORE	0.523 ± 0.072	52.500 ± 8.489	8.100 ± 2.234
		MissOTM	0.241 ± 0.078	59.700 ± 4.111	-
		MissScore	0.368 ± 0.063	77.100 ± 8.543	14.000 ± 5.055
		SCORE-EM	**0.581 ± 0.045**	**49.200 ± 4.780**	**3.900 ± 2.183**
		SCORE (Fully Observed)	0.760 ± 0.106	24.400 ± 8.407	1.650 ± 1.137

Table 8: Performance summary across graph topologies ($d = 30$, missing rate=0.5). **Bold** indicates the best performance among methods handling missing data; *SCORE (Fully Observed)* is shown as a Reference.

Graph	Degree	Model	F1 ↑	SHD ↓	Order Divergence ↓
ER	1	MI + SCORE	0.433 ± 0.063	47.400 ± 6.535	8.100 ± 2.923
		MissDAG	0.397 ± 0.121	**24.500 ± 5.720**	-
		MissForest + SCORE	0.457 ± 0.063	48.000 ± 6.272	6.200 ± 1.619
		MissOTM	0.369 ± 0.075	27.600 ± 7.516	-
		MissScore	0.246 ± 0.036	91.100 ± 24.973	10.000 ± 5.312
		SCORE-EM	**0.560 ± 0.082**	39.000 ± 9.006	**1.700 ± 1.418**
		SCORE (Fully Observed)	0.934 ± 0.043	3.750 ± 2.531	0.950 ± 1.191
	4	MI + SCORE	0.415 ± 0.069	106.600 ± 11.983	24.800 ± 9.343
		MissDAG	0.207 ± 0.068	103.700 ± 13.417	-
		MissForest + SCORE	0.436 ± 0.070	103.900 ± 16.093	22.300 ± 6.881
		MissOTM	0.163 ± 0.081	110.400 ± 15.064	-
		MissScore	0.253 ± 0.081	146.600 ± 11.673	47.600 ± 15.862
		SCORE-EM	**0.536 ± 0.054**	**89.000 ± 12.719**	**9.700 ± 3.335**
		SCORE (Fully Observed)	0.726 ± 0.098	48.800 ± 17.065	3.450 ± 2.139
SF	1	MI + SCORE	0.434 ± 0.090	48.400 ± 11.607	6.000 ± 2.000
		MissDAG	0.369 ± 0.133	**23.200 ± 3.120**	-
		MissForest + SCORE	0.462 ± 0.052	47.000 ± 6.128	4.600 ± 2.459
		MissOTM	0.453 ± 0.112	23.900 ± 4.864	-
		MissScore	0.214 ± 0.043	110.300 ± 19.805	7.300 ± 2.497
		SCORE-EM	**0.543 ± 0.068**	40.600 ± 7.919	**1.300 ± 1.337**
		SCORE (Fully Observed)	0.937 ± 0.062	3.200 ± 2.913	0.800 ± 2.215
	4	MI + SCORE	0.398 ± 0.057	98.700 ± 7.987	22.900 ± 7.549
		MissDAG	0.210 ± 0.069	95.000 ± 4.190	-
		MissForest + SCORE	0.435 ± 0.058	95.700 ± 8.381	18.900 ± 5.820
		MissOTM	0.196 ± 0.052	98.400 ± 4.477	-
		MissScore	0.255 ± 0.066	147.500 ± 20.129	39.900 ± 10.450
		SCORE-EM	**0.514 ± 0.082**	**85.700 ± 12.902**	**8.500 ± 2.461**
		SCORE (Fully Observed)	0.684 ± 0.084	50.000 ± 10.458	5.050 ± 2.502