DISTCUTS: DISTRIBUTION-AWARE CAUSAL DISCOVERY FROM INCOMPLETE MULTIVARIATE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering causal relationships from incomplete multivariate time series is challenging because missing observations obscure temporal dependencies. Existing approaches typically address this challenge by alternating between missing-value reconstruction and causal graph learning. However, their structure-learning objectives primarily assess whether a source variable's history improves point predictions of a target's future values. As a result, they may overlook causal edges whose effects manifest primarily in the target's conditional variability rather than its conditional mean, yielding incomplete causal structures. Omitting these edges can further impair missing-value reconstruction by excluding relevant causal information, with the resulting errors propagating into subsequent graph estimation. To address this limitation, we propose DISTCUTS, a distribution-aware framework that complements point prediction with history-conditioned residual distribution matching to capture both mean-driven and scale-driven dependencies. Specifically, DISTCUTS estimates history-dependent residual scales from graph-constrained source histories and uses these scales to parameterize conditional residual distributions. A maximum mean discrepancy (MMD) objective compares samples from these distributions with the corresponding observed residuals, encouraging the model to capture residual variability supported by the observations. As auxiliary components, we demean observed residuals to reduce sensitivity to predicted-mean offsets and use the learned residual scales to guide missing-value reconstruction. Experiments on synthetic and real-world datasets demonstrate that DISTCUTS outperforms the evaluated baselines.

1 INTRODUCTION

Discovering causal relationships from multivariate time series helps reveal the mechanisms underlying complex dynamical systems (Runge et al., 2023), supporting scientific reasoning and decision-making in healthcare, climate science, finance, and biological regulation. In practice, however, multivariate time series are rarely fully observed, as measurements could be missing or sampled at irregular intervals due to heterogeneous acquisition schedules, temporary sensor malfunctions, or recording failures (Che et al., 2018). These irregular observations obscure temporal dependencies and make reliable causal structure recovery more challenging.

Most existing time-series causal discovery methods assume regularly sampled and fully observed data (Runge et al., 2019; Tank et al., 2022). To handle irregular observations, recent neural approaches combine missing-value reconstruction with causal graph learning. For example, CUTS (Cheng et al., 2023) alternates between reconstructing the incomplete time series and estimating a Granger-causal graph (Granger, 1969), while CUTS+ (Cheng et al., 2024) extends this framework to high-dimensional systems. These methods identify a causal relation from incomplete observations mainly by examining whether the reconstructed history of one variable improves the prediction of another variable's future value. This value-oriented perspective is effective when causal effects manifest as changes in the predicted values, but it does not explicitly capture how a candidate cause alters the conditional predictive distribution of the target variable.

Ignoring distribution-driven relations has implications beyond missing a particular class of causal edges. In causal discovery from incomplete time series, graph estimation and missing-value reconstruction are inherently coupled: the learned graph determines which histories inform recon-

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.

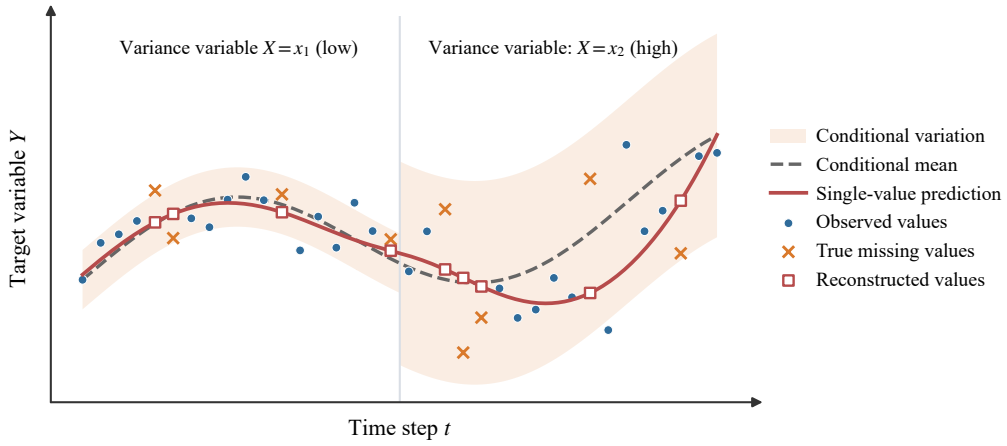


Figure 1: Illustration of value-oriented reconstruction in the presence of distribution-driven effects. The state of X changes from x_1 in the low-variance regime to x_2 in the high-variance regime, expanding the conditional variation of the target Y without changing its expected value. A single-value predictor can closely follow the expected mean when the conditional variation is small, but it cannot represent the increased uncertainty in the high-variance regime. Consequently, a value-oriented method may omit distribution-driven causal edges and fail to account for differences in reconstruction uncertainty across regimes.

struction, and the reconstructed sequence is reused to update the graph. If a variable modulates the conditional distribution of a target without changing its conditional mean, a value-oriented criterion may exclude it from the target's causal parent set. The reconstruction model then lacks information about the associated distributional regimes, potentially compromising reconstruction under high uncertainty (Fig. 1). Moreover, unexplained distributional variation may be misattributed to other variables, resulting in both missing and spurious causal edges. Capturing distribution-driven relations is therefore essential not only for recovering a more complete causal structure, but also for reliably reconstructing incomplete stochastic dynamics.

To address this limitation, we propose DISTCUTS, a distribution-aware framework for causal discovery from incomplete time series. The key idea is to complement value-based structure learning with history-dependent residual distribution matching, allowing source histories to inform causal discovery through their effects on residual scale. Specifically, DISTCUTS predicts history-dependent residual scales from source histories constrained by the causal graph to parameterize the conditional residual distribution. Samples drawn from this distribution are compared with the corresponding observed residuals through a maximum mean discrepancy (MMD)-based objective, which guides causal structure learning toward relations that explain history-dependent variations in residual scale. We also demean observed residuals across available minibatch entries to reduce sensitivity to predicted-mean offsets and use calibrated scale-aware blending to guide missing-value reconstruction. The updated sequence is then reused for the next round of graph learning, allowing DISTCUTS to jointly refine the causal structure and reconstruction while capturing dependencies beyond mean prediction.

Our contributions are threefold:

- We propose DISTCUTS, a distribution-aware framework that complements value-based causal structure learning with history-conditioned residual distribution matching, extending causal discovery from incomplete time series to distribution-driven dependencies.
- We develop an MMD-based objective that provides a distributional signal for discovering causal relations expressed through history-dependent changes in residual scale.
- We introduce residual demeaning and mean detachment to reduce sensitivity of distribution matching to predicted mean offsets, together with calibrated scale-aware blending that uses the learned residual scale to guide missing-value reconstruction.

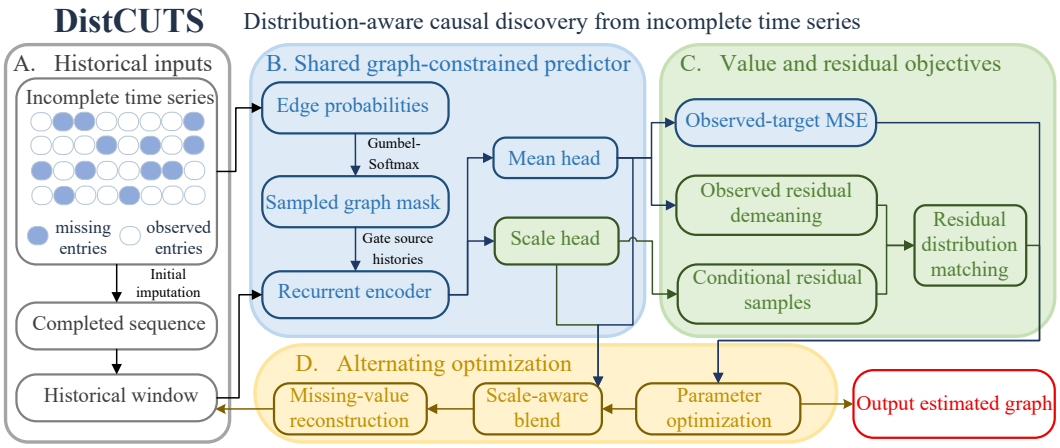


Figure 2: Overview of DISTCUTS. (A) Missing entries are initialized and iteratively reconstructed to form historical inputs. (B) A sampled graph mask constrains the shared recurrent encoder used by the mean and scale heads. (C) Observed-target prediction error and per-history residual distribution matching jointly guide learning. Observed residuals are demeaned, with gradients blocked through the predicted mean in the distributional branch. (D) Alternating predictor and graph updates are followed by a calibrated scale-aware blend of model predictions for missing-value reconstruction.

2 Related Work

2.1 Causal Discovery from Incomplete Time Series

Causal discovery methods based on conditional-independence tests, autoregressive models, and neural predictors generally assume regularly sampled, fully observed data (Runge et al., 2019; Tank et al., 2022; Pamfil et al., 2020). Applying them after separate imputation is straightforward, but imputation without structural knowledge can distort dependencies and bias graph estimation (Tu et al., 2019). Approaches for incomplete static data jointly perform imputation and structure learning (Gao et al., 2022; Vo et al., 2024), but do not explicitly account for temporal dependencies. For incomplete time series, existing methods (Cheng et al., 2023; 2024; Li et al., 2026) couple temporal reconstruction with causal discovery, while they primarily learn relations through target-value prediction. DISTCUTS adds residual distribution matching to capture dependencies expressed through history-dependent residual scale.

2.2 Distribution-Aware Causal Discovery in Time Series

Temporal dependencies can affect conditional variance (Cheung & Ng, 1996; Hafner & Herwartz, 2006), quantiles (Jeong et al., 2012; Lee & Yang, 2014; Song & Taamouti, 2021), or full conditional distributions (Bouezmarni et al., 2012; Taamouti et al., 2014; Candelon & Tokpavi, 2016; Wang, 2026). Beyond statistical tests, RHINO (Gong et al., 2023) incorporates history-dependent noise into neural causal graph learning using conditional spline flows. These approaches primarily target fully observed series, while DISTCUTS combines residual distribution matching, graph-mask learning and missing-value reconstruction to address incomplete histories.

3 Methods

We propose DISTCUTS to learn temporal dependencies from incomplete time series by combining value prediction with history-conditioned residual distribution matching. A learnable graph determines which source histories contribute to each target's predictive mean and residual scale. The value objective captures dependencies useful for mean prediction, whereas the distributional objective provides an additional learning signal for dependencies expressed through residual-scale variation. We jointly use these objectives in an alternating procedure that updates the predictor,

refines the graph, and reconstructs missing values through a calibrated scale-aware blend of model predictions and temporal interpolation. The overall architecture is illustrated in Fig. 2.

3.1 Problem Formulation

Let $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}]^\top \in \mathbb{R}^{T\times d}$ denote a time series with T time steps and d variables, where $\mathbf{x}_t \in \mathbb{R}^d$. The observation mask $\mathbf{M} \in \{0,1\}^{T\times d}$ satisfies $M_{t,j} = 1$ if $X_{t,j}$ is observed and $M_{t,j} = 0$ otherwise. Irregular observations are represented as missing entries on a common discrete time grid.

To construct historical inputs in the presence of missing values, we maintain a completed sequence

$$\widetilde{X}_{t,j} = \begin{cases} X_{t,j}, & M_{t,j} = 1, \\ \widehat{X}_{t,j}, & M_{t,j} = 0, \end{cases} \tag{1}$$

where $\widehat{X}_{t,j}$ is the current reconstruction of a missing entry. Given a maximum lag L, the historical input for predicting time $t \geq L$ is $\mathbf{H}_t = [\widetilde{\mathbf{x}}_{t-L}, \ldots, \widetilde{\mathbf{x}}_{t-1}] \in \mathbb{R}^{d\times L}$. Thus, historical inputs may contain reconstructed values, while observed entries remain unchanged.

Our objective is to recover a directed summary graph $\mathbf{A} \in \{0,1\}^{d\times d}$. An edge $A_{i,j} = 1$ represents a temporal dependence in which the history of variable i provides additional predictive information about variable j through its conditional mean or residual scale, given the other source histories. Each edge gates the entire history window of its source, while the predictor models effects at individual lags. The graph is shared across time, and self-edges represent autoregressive dependence. Because all modeled effects run from past values to current targets, the summary graph need not be acyclic.

3.2 Graph-Constrained Mean and Scale Prediction

We parameterize the graph by a learnable edge-probability matrix $\Theta \in (0,1)^{d\times d}$, where $\Theta_{i,j}$ is the inclusion probability of edge $i \to j$. During training, binary graph masks are sampled according to these probabilities: $A_{i,j} \sim \mathrm{Bernoulli}(\Theta_{i,j})$. We use a straight-through Gumbel–Softmax estimator (Jang et al., 2017) to propagate gradients through the sampled masks. Each sampled mask determines the source histories available to the predictor and is shared by its mean and scale branches.

Following the graph-constrained recurrent prediction framework of Cheng et al. (2024), we compute

$$\begin{aligned} \mathbf{h}_{t,j} &= F_{\phi,j}(\mathbf{H}_t, \mathbf{A}), \\ \mu_{t,j} &= g_{\mu,j}(\mathbf{h}_{t,j}), \\ \sigma_{t,j} &= \mathrm{softplus}\big(g_{\sigma,j}(\mathbf{h}_{t,j})\big) + \sigma_{\min}. \end{aligned} \tag{2}$$

Here, $F_{\phi,j}$ is a recurrent encoder whose source-to-target messages are gated by $\mathbf{A}$, and $\mathbf{h}_{t,j}$ is its representation for target j at time t. The heads $g_{\mu,j}$ and $g_{\sigma,j}$ predict the conditional mean $\mu_{t,j}$ and residual scale $\sigma_{t,j}$, respectively. The transformation $\mathrm{softplus}(u) = \log(1 + \exp(u))$ and the fixed lower bound $\sigma_{\min} > 0$ ensure a positive scale. The parameter set ϕ includes the encoder and both output heads.

For a minibatch of target times $\mathcal{B} \subseteq \{L, \ldots, T-1\}$, we define

$$\mathcal{O}_{\mathcal{B}} = \{(t,j) : t \in \mathcal{B},\, M_{t,j} = 1\}, \qquad N_{\mathcal{B}} = |\mathcal{O}_{\mathcal{B}}|. \tag{3}$$

The value prediction objective is

$$\mathcal{L}_{\mathrm{val}} = \frac{1}{\max(1, N_{\mathcal{B}})} \sum_{(t,j)\in\mathcal{O}_{\mathcal{B}}} (X_{t,j} - \mu_{t,j})^2 . \tag{4}$$

This objective supervises the predictive mean using observed targets only. However, a source history may influence the target's residual scale without substantially changing its conditional mean. To capture such dependencies, we additionally train the scale branch through residual distribution matching.

3.3 History-Conditioned Residual Distribution Matching

The scale output $\sigma_{t,j}$ converts graph-constrained historical information into a history-dependent residual distribution. We compare samples from this distribution with the corresponding observed residual using an MMD-based objective (Gretton et al., 2012). Since both outputs use the same graph, this objective allows residual-scale variation to contribute to graph learning alongside mean prediction.

History-conditioned residual generation. For each history–target pair (t, j), we draw $S \geq 2$ independent innovations $\epsilon_{t,j}^{(s)} \sim p_\epsilon$, where p_ϵ is a fixed distribution with zero mean and unit variance. We generate residual samples as

$$r_{t,j}^{\mathrm{gen},(s)} = \sigma_{t,j} \sqrt{\frac{S}{S-1}} \left(\epsilon_{t,j}^{(s)} - \frac{1}{S} \sum_{s'=1}^{S} \epsilon_{t,j}^{(s')} \right). \tag{5}$$

Centering the innovations removes Monte Carlo mean fluctuations, and the factor $\sqrt{S/(S-1)}$ compensates for the resulting variance reduction in expectation. The history-dependent scale therefore controls the spread of the generated residuals.

Observed residual demeaning. To reduce the influence of mean-prediction errors on distribution matching, we detach the predicted mean and center the observed residuals within each minibatch. For variable j, let $\mathcal{O}_j = \{t \in \mathcal{B} : M_{t,j} = 1\}$ denote its observed target indices. We define

$$\begin{aligned} r_{t,j} &= X_{t,j} - \mathrm{sg}(\mu_{t,j}), \quad t \in \mathcal{O}_j, \\ \bar{r}_j &= \frac{1}{|\mathcal{O}_j|} \sum_{u \in \mathcal{O}_j} r_{u,j}, \\ \rho_{t,j} &= r_{t,j} - \bar{r}_j, \qquad\qquad t \in \mathcal{O}_j. \end{aligned} \tag{6}$$

Here, sg denotes stop-gradient to prevent the distributional loss from directly updating the predicted mean. Centering removes the minibatch-level residual offset, reducing sensitivity to shared mean-prediction bias. The distributional loss can still update the shared encoder and causal graph through the predicted scale.

Residual matching objective. We use a target-specific Gaussian kernel

$$k_j(a, b) = \exp\left(-\frac{(a-b)^2}{2\gamma_j} \right), \tag{7}$$

where $a, b \in \mathbb{R}$ are residual values and $\gamma_j > 0$ is a bandwidth parameter. We estimate γ_j from the median positive pairwise squared distance between observed residuals of variable j in the minibatch, apply a positive numerical safeguard, and detach it from gradient computation.

For each observed pair $(t, j) \in \mathcal{O}_\mathcal{B}$, the matching score is

$$\begin{aligned} \ell_{t,j}^{\mathrm{dist}} = \frac{1}{S^2} \sum_{s=1}^{S} \sum_{s'=1}^{S} k_j \left(r_{t,j}^{\mathrm{gen},(s)}, r_{t,j}^{\mathrm{gen},(s')} \right) \\ - \frac{2}{S} \sum_{s=1}^{S} k_j \left(r_{t,j}^{\mathrm{gen},(s)}, \rho_{t,j} \right) + 1. \end{aligned} \tag{8}$$

This score is the squared empirical MMD between the generated residual sample set and the single observed residual $\rho_{t,j}$. The first term includes all generated sample pairs, including identical indices, and the constant term follows from $k_j(\rho_{t,j}, \rho_{t,j}) = 1$. We average the scores over observed targets:

$$\mathcal{L}_{\mathrm{dist}} = \frac{1}{\max(1, N_\mathcal{B})} \sum_{(t,j) \in \mathcal{O}_\mathcal{B}} \ell_{t,j}^{\mathrm{dist}}. \tag{9}$$

Matching is performed separately for each history–target pair before averaging across observations. As a result, the objective evaluates residual samples generated under the corresponding historical input, rather than pooling all generated residuals into a single marginal distribution. Its gradients propagate through $\sigma_{t,j}$ to the shared encoder and graph mask, providing an additional signal for source histories that help explain residual-scale variation.

3.4 Alternating Optimization and Reconstruction

The predictor requires completed histories, while reconstruction depends on the predictor and its graph. We address this coupling through alternating refinement. Before training, we standardize each variable using its observed entries and initialize missing values to zero in standardized space. Each refinement round updates the predictor and graph with the completed sequence held fixed, followed by a missing-value update.

Predictor and graph updates. With graph parameters Θ fixed, we optimize predictor parameters ϕ using

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{val}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}, \tag{10}$$

where $\lambda_{\text{dist}} \geq 0$ controls the contribution of residual distribution matching. We then fix ϕ and optimize Θ using

$$\mathcal{L}_{\text{graph}} = \mathcal{L}_{\text{val}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \frac{\lambda_{\text{sp}}}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\Theta_{i,j}, \tag{11}$$

where $\lambda_{\text{sp}} \geq 0$ weights the sparsity penalty. The first two terms favor edges useful for explaining the predictive mean or residual scale, while the final term penalizes the average edge inclusion probability. During the graph update, gradients pass through the fixed predictor to the graph parameters via the straight-through mask estimator.

Scale-aware missing-value reconstruction. Let $n = 0, \ldots, N-1$ index refinement rounds, and let $I_{t,j}$ denote the temporal interpolation obtained from observed entries of variable j. During predictor optimization, we collect both the predicted mean $\mu_{t,j}^{(n)}$ and residual scale $\sigma_{t,j}^{(n)}$. A variable-wise factor $c_j^{(n)} > 0$ calibrates predicted scales against observed prediction errors, while $v_j^{\text{int},(n)} > 0$ denotes an estimated interpolation-error variance. We form the scale-aware blend

$$\begin{aligned}
\widetilde{\sigma}_{t,j}^{(n)} &= c_j^{(n)}\sigma_{t,j}^{(n)}, \\
w_{t,j}^{(n)} &= \frac{v_j^{\text{int},(n)}}{v_j^{\text{int},(n)} + \big(\widetilde{\sigma}_{t,j}^{(n)}\big)^2}, \\
b_{t,j}^{(n)} &= w_{t,j}^{(n)}\mu_{t,j}^{(n)} + \big(1 - w_{t,j}^{(n)}\big)I_{t,j}.
\end{aligned} \tag{12}$$

A lower calibrated predictive variance therefore assigns more weight to the model prediction; a less reliable prediction places more weight on interpolation. For missing entries with $t \geq L$, we update

$$\widehat{X}_{t,j}^{(n+1)} = (1 - \eta_n)\widehat{X}_{t,j}^{(n)} + \eta_n b_{t,j}^{(n)}, \qquad M_{t,j} = 0, \tag{13}$$

where $\eta_n \in [0, 1]$ is the reconstruction update rate. We adopt a delayed, gradual reconstruction schedule: the initial completion remains unchanged during the first 20 training epochs. Starting at epoch 21, missing entries are refreshed before each epoch's optimization using the scale-aware blends collected in the preceding epoch, with an update rate of $\eta_n = 0.1$. Thus, each refresh retains 90% of the current estimate and incorporates 10% of the cached blended estimate. This warm-up avoids feeding early, insufficiently trained predictions back into the historical inputs, while gradual updates limit abrupt changes in the completed sequence.

The updated estimates are substituted into Eq. 1 to form $\widetilde{\mathbf{X}}^{(n+1)}$, preserving all observed entries. This sequence supplies historical inputs for the next round. Scale-aware blending is a reconstruction rule rather than an additional training loss. The learned residual scale thus influences subsequent graph learning through reconstructed histories, in addition to its direct contribution through the distributional objective. After the final refinement round, the estimated summary graph could be obtained from the learned edge probabilities. The entire training procedure of DistCUTS is provided in Appendix A.1.

4 Experiments

We evaluate whether residual distribution matching improves causal discovery from incomplete time series, whether these improvements extend to distribution-dependent edges, and how graph learning

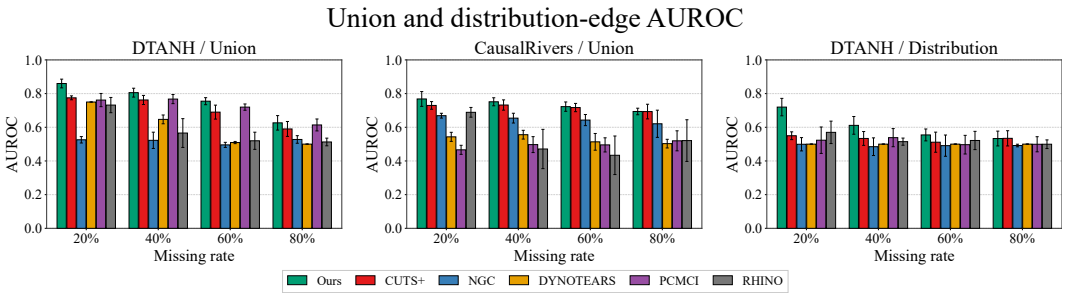


Figure 3: Graph recovery across four missing rates. Left: overall AUROC on DTANH, evaluated against the union of mean-driven and distribution-dependent edges. Middle: AUROC on CausalRivers, evaluated against the hydrological reference graph. Right: distribution-edge AUROC on DTANH, with mean-only edges excluded from the negative set. Bars report means over five seeds.

interacts with missing-value reconstruction. The ablation study follows to distinguish the contributions of proposed components in DISTCUTS. Implementation details and hyperparameter analysis for DISTCUTS are provided in Appendices A.3 and A.4, respectively.

4.1 EXPERIMENTAL SETUP

Datasets and missingness. We use three synthetic datasets, CausalRivers (Stein et al., 2025), and the five DREAM3 Size-100 networks (Prill et al., 2010). Each synthetic dataset contains 32 variables and 2,000 retained time steps, with separately specified mean-driven and distribution-dependent edges. DLIN combines linear mean dynamics with state-dependent innovation scales; DTANH uses nonlinear mean dynamics and Laplace innovations; HETERO-StudentT combines heteroscedasticity with heavy-tailed innovations. CausalRivers comprises a fixed connected subgraph of ten East German river gauges, sampled every six hours over 3,000 time steps, with a hydrological reference graph. DREAM3 is an *in silico* gene-regulatory benchmark: Ecoli1, Ecoli2, and Yeast1–3 each contain 100 genes and 46 trajectories of 21 time points. We artificially hide observations completely at random at rates $r \in \{0.2, 0.4, 0.6, 0.8\}$ and repeat each condition with five seeds. Hidden values are excluded from training. The detailed data preparation is reported in Appendix A.2.

Baselines and metrics. We compare with CUTS+ (Cheng et al., 2024), NGC (Tank et al., 2022), DYNOTEARS (Pamfil et al., 2020), PCMCI (Runge et al., 2019), and RHINO (Gong et al., 2023). CUTS+ is the backbone baseline, whereas RHINO explicitly models conditional distributions. Graph recovery is measured by AUROC over directed cross-variable edges, excluding self-edges. For synthetic datasets, overall AUROC uses the union of mean-driven and distribution-dependent edges. Distribution-edge AUROC compares annotated distribution-dependent edges against non-edges, excluding mean-only edges from the negative set. For CausalRivers and DREAM3, AUROC uses the supplied reference graph. Missing value reconstruction is evaluated under RMSE on originally observed entries hidden for evaluation, excluding native missing entries.

4.2 MAIN RESULTS

We first examine overall graph recovery, then assess whether the improvement extends to distribution-dependent edges, and finally evaluate missing-value reconstruction. Figure 3 presents the main discovery results on DTANH and CausalRivers, covering a controlled nonlinear system and an observational benchmark. Results on the remaining datasets are provided in Appendix A.5.

Overall causal graph recovery. DISTCUTS achieves the highest mean AUROC at all four missing rates on both datasets. On DTANH, its consistent advantage over CUTS+ supports complementing value-based graph learning with residual distribution matching. Since the reference graph contains both mean-driven and distribution-dependent edges, the additional objective can provide information about dependencies that contribute little to point prediction. The distribution-edge evaluation below more directly examines this explanation. On CausalRivers, the advantage over CUTS+

progressively narrows as missingness increases, and their mean AUROCs are nearly identical at $r = 0.8$. This pattern suggests that the benefit of the additional learning signal depends on enough temporal information to estimate it. As observations become sparse, uncertainty in reconstructed histories may limit the usefulness of residual distribution matching for distinguishing candidate relations. Since the hydrological reference graph does not label edges by their effects on the conditional mean or scale, these results support overall recovery on an observational benchmark without establishing the mechanism underlying individual recovered edges.

Recovery of distribution-dependent edges. The DTANH results reveal a distinction between overall graph recovery and recovery of distribution-dependent edges. CUTS+ performs substantially better on the union graph than on the distribution-edge task, indicating that aggregate performance can conceal limited sensitivity to dependencies expressed through conditional variability. At $r = 0.2$, DISTCUTS raises distribution-edge AUROC from 0.5498 to 0.7199, providing more direct evidence that residual distribution matching captures information beyond target-value prediction. Among the baselines, RHINO achieves 0.5699 at $r = 0.2$, compared with 0.4992–0.5498 for the other baselines. This suggests that explicitly modeling conditional distributions can help identify distribution-dependent edges when observations are relatively abundant. However, this advantage is not sustained as missingness increases: RHINO no longer consistently outperforms the other baselines. This suggests that distribution modeling alone is insufficient under missingness, whereas DISTCUTS improves robustness by coupling it with iterative reconstruction and scale-aware blending.

4.3 MISSING-VALUE RECONSTRUCTION

Table 1 compares reconstruction performance with CUTS+, the backbone that also jointly learns a graph and reconstructs missing values. DISTCUTS achieves lower mean RMSE across all 18 settings reported in Table 1. CausalRivers exhibits the largest relative improvements, with RMSE reductions of 28.1% at $r = 0.2$ and 16.1% at $r = 0.6$. However, its performance advantage narrows as missingness increases, indicating that the additional distributional information does not fully compensate for the loss of observed temporal context. The remaining results are provided in Appendix A.5.

Table 1: Reconstruction RMSE (lower is better), averaged over five seeds. Bold indicates the lower RMSE in each comparison.

Method	DLIN	DTANH	HETERO-StudentT	Causal Rivers	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3
Missing rate $r = 0.2$									
CUTS+	0.2834	0.2813	0.2799	0.8005	0.0883	0.0883	0.0902	0.0863	0.0865
DISTCUTS	**0.2719**	**0.2703**	**0.2689**	**0.5752**	**0.0751**	**0.0752**	**0.0756**	**0.0734**	**0.0735**
Missing rate $r = 0.6$									
CUTS+	0.3003	0.3001	0.2971	0.8358	0.0947	0.0953	0.0964	0.0926	0.0930
DISTCUTS	**0.2876**	**0.2872**	**0.2851**	**0.7009**	**0.0829**	**0.0833**	**0.0837**	**0.0810**	**0.0810**

4.4 ABLATION STUDY

We evaluate four variants to examine the contributions of distribution learning, residual preparation, reconstruction feedback, and scale-aware reconstruction. *No MMD* removes the distributional objective; *No iteration* freezes the initial completion throughout training; *No demeaning* jointly disables observed-residual demeaning and allows MMD gradients to update the predicted mean; and *No scale* retains MMD but reconstructs missing values solely from the predicted mean, without using the predicted residual scale. All variants share data, masks, calibration partitions, and seeds. Table 2 summarizes results across four missing rates and five seeds. Synthetic distribution-edge AUROC is additionally averaged across the three synthetic datasets. CausalRivers AUROC and RMSE evaluate the corresponding effects on reference-graph recovery and reconstruction.

Table 2: Ablation results averaged over four missing rates and five seeds. Synthetic distribution-edge AUROC additionally averages DLIN, DTANH, and HETERO-StudentT.

Variant	Synthetic	CausalRivers	
	Dist. AUROC ↑	AUROC ↑	RMSE ↓
Full	**0.6092**	**0.7340**	**0.6786**
No MMD	0.4927	0.7133	0.7832
No iteration	0.5929	0.6822	0.7173
No demeaning	0.5860	0.6860	0.6976
No scale	0.6040	0.7330	0.7851

Residual distribution matching. Removing MMD reduces synthetic distribution-edge recovery to near-chance performance, revealing a limitation of value-based supervision: a source history may explain changes in a target's conditional variability without substantially improving its predicted mean. Residual distribution matching provides a complementary criterion for retaining such dependencies, which would otherwise receive little support from the value objective. The accompanying deterioration on CausalRivers supports the usefulness of this additional supervision beyond the synthetic settings. However, the reconstruction loss cannot be attributed solely to poorer graph recovery, since removing MMD also removes direct distributional supervision for the scale head used in reconstruction.

Residual demeaning and mean detachment. Jointly removing these operations impairs graph recovery despite retaining MMD, suggesting that the effectiveness of distributional supervision depends on how its residual targets are constructed. Without demeaning, systematic residual offsets can enter the matching objective; without detachment, the objective can also be reduced by adjusting the predicted mean. Both effects can divert optimization from explaining history-dependent residual variation. The observed degradation is consistent with residual preparation making the distributional signal more informative for structure learning.

Iterative reconstruction. Freezing the initial completion produces the largest decline in CausalRivers graph recovery among the variants. This highlights the dependence of structure learning on the quality of the historical inputs: both value prediction and residual-scale estimation operate on completed histories, so retaining the initial fill can obscure the dependencies that either objective seeks to capture. Iterative reconstruction allows the evolving predictor and graph to revise these inputs, which can in turn support subsequent structure estimation. The simultaneous deterioration in graph recovery and reconstruction is consistent with this feedback mechanism.

Scale-aware reconstruction. Mean-only reconstruction preserves graph recovery relatively well but substantially worsens reconstruction accuracy. Because MMD remains active, the residual scale can still provide distributional supervision for graph learning even when it is excluded from missing-value reconstruction. This helps explain why the two evaluation metrics respond differently to this ablation: retaining scale-dependent dependencies does not by itself ensure that reconstruction fully exploits the learned information. The results therefore support distinct roles for residual scale modeling in structure learning and reconstruction.

5 Conclusion

We presented DistCUTS, a framework for causal discovery from incomplete time series that combines mean prediction with history-conditioned residual distribution matching. An MMD-based objective enables a shared graph to capture dependencies expressed through residual scale, while residual demeaning and mean detachment reduce direct interference from mean adjustment. Calibrated scale-aware reconstruction feeds updated histories back into graph learning. Experiments demonstrate improved distribution-edge recovery on synthetic data and gains in graph recovery and reconstruction on benchmark datasets. Ablations distinguish the contributions of distribution matching and reconstruction feedback. Future work will extend the residual model beyond scale variation and investigate robustness to nonrandom missingness.

AI USE STATEMENT

In this work, we did not use generative AI tools for any tasks requiring disclosure. Additionally, we used generative AI tools to assist with editing software code, analyzing existing literature, searching for information, editing the manuscript to improve readability, identifying relevant literature, formatting references, and proposing the title or keywords. We reviewed all AI-assisted work and verified the resulting code, literature-related information, references, and manuscript edits for correctness and consistency with the original sources and our research. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed information to support the reproducibility of DISTCUTS throughout the main paper and appendix. The problem formulation and graph-constrained mean and scale prediction are described in Sections 3.1 and 3.2. The residual distribution matching objective, including residual demeaning and mean detachment, is specified in Section 3.3. Alternating optimization and scale-aware reconstruction are presented in Section 3.4, with the complete training procedure provided in Appendix A.1. Experimental settings, datasets, baselines, and evaluation metrics are described in Section 4.1. Additional details on synthetic data generation and real-world datasets are provided in Appendix A.2, while preprocessing, calibration, training configurations, and evaluation protocols are documented in Appendix A.3. Hyperparameter and initialization sensitivity analyses are reported in Appendix A.4. Complete dataset-specific results across missing rates are presented in Appendix A.5. Anonymous source code is available at `https://anonymous.4open.science/r/distcuts_iclr27-BFBE` to facilitate reproduction of the reported results.

REFERENCES

Taoufik Bouezmarni, Jeroen V. K. Rombouts, and Abderrahim Taamouti. Nonparametric copula-based test for conditional independence with applications to granger causality. *Journal of Business & Economic Statistics*, 30(2):275–287, 2012. doi: 10.1080/07350015.2011.638831.

Bertrand Candelon and Sessi Tokpavi. A nonparametric test for granger causality in distribution with application to financial contagion. *Journal of Business & Economic Statistics*, 34(2):240–253, 2016. doi: 10.1080/07350015.2015.1026774.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS+: high-dimensional causal discovery from irregular time-series. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 11525–11533. AAAI Press, 2024.

Yin-Wong Cheung and Lilian K. Ng. A causality-in-variance test and its application to financial market prices. *Journal of Econometrics*, 72(1):33–48, 1996. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(94)01714-X.

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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5024–5038. Curran Associates, Inc., 2022. doi: 10.52202/068431-0363.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Christian M. Hafner and Helmut Herwartz. A lagrange multiplier test for causality in variance. *Economics Letters*, 93(1):137–141, 2006. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2006.04.008.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Kiho Jeong, Wolfgang K. Härdle, and Song Song. A consistent nonparametric test for causality inquantile. *Econometric Theory*, 28(4):861–887, 2012. doi: 10.1017/S0266466611000685.

Tae-Hwy Lee and Weiping Yang. Granger-causality in quantiles between financial markets: Using copula approach. *International Review of Financial Analysis*, 33:70–78, 2014. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2013.08.008.

Weihong Li, Baohong Li, Anpeng Wu, Zhihan Li, Ming Ma, Keting Yin, and Kun Kuang. Causal discovery for irregularly time series with consistency guarantees. In *Forty-third International Conference on Machine Learning*, 2026.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605, 2020.

Robert J. Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS ONE*, 5(2):e9202, 2010. doi: 10.1371/journal.pone.0009202.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023. doi: 10.1038/s43017-023-00431-y.

Xiaojun Song and Abderrahim Taamouti. Measuring granger causality in quantiles. *Journal of Business & Economic Statistics*, 39(4):937–952, 2021.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers - scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025.

Abderrahim Taamouti, Taoufik Bouezmarni, and Anouar El Ghouch. Nonparametric estimation and inference for conditional density based granger causality measures. *Journal of Econometrics*, 180 (2):251–264, 2014. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2014.03.001.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022.

Algorithm 1 Alternating training of DISTCUTS

Require: Series $\mathbf{X}$, observation mask $\mathbf{M}$, maximum lag L, refinement rounds N, distribution weight λ_{dist}, sparsity weight λ_{sp}, and update rates $\{\eta_n\}$.

Ensure: Edge probabilities Θ and completed sequence $\widetilde{\mathbf{X}}^{(N)}$.

Estimate variable-wise normalization statistics from available training observations only.
Initialize $\widetilde{\mathbf{X}}^{(0)}$ and construct the temporal interpolation I used by the blend separately; preserve all observed entries.
Initialize predictor parameters ϕ and edge probabilities Θ.
for $n = 0, \ldots, N-1$ **do**
  Form historical windows from fixed $\widetilde{\mathbf{X}}^{(n)}$.
  for each predictor minibatch **do**
    Draw a graph mask from Θ without updating Θ.
    Compute $\mu_{t,j}$ and $\sigma_{t,j}$ using Eq. 2.
    Construct demeaned residuals and generated residual samples using Eqs. 5 and 6.
    Update ϕ with Θ fixed by minimizing Eq. 10.
  end for
  for each graph minibatch **do**
    Draw a hard graph mask with straight-through gradients to Θ.
    Update Θ with ϕ fixed by minimizing Eq. 11.
  end for
  if $n + 1 < 20$ **then**
    Keep the completion unchanged: $\widetilde{\mathbf{X}}^{(n+1)} \leftarrow \widetilde{\mathbf{X}}^{(n)}$.
  else
    Estimate calibration factors and interpolation-error variances from observations withheld for calibration.
    Form the scale-aware blend using Eq. 12.
    Update missing entries with Eq. 13; copy observed entries unchanged into $\widetilde{\mathbf{X}}^{(n+1)}$.
  end if
end for
After training, calibrate predictive scales and interpolation error using the held-out calibration entries, then form the final reconstruction $\widetilde{\mathbf{X}}^{(N)}$.
return Θ and $\widetilde{\mathbf{X}}^{(N)}$.

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, 2019.

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, 2024.

Tao Wang. Testing distributional granger causality with entropic optimal transport. *Journal of Time Series Analysis*, 2026. doi: https://doi.org/10.1111/jtsa.70057.

A APPENDIX

A.1 TRAINING PROCEDURE

Algorithm 1 makes explicit the order in which the predictor, graph, and completed series are updated. The observed entries are never replaced. Distribution matching is evaluated at observed target entries, whereas reconstructed values may enter the lagged histories used in later refinement rounds. The same sampled graph gates the predictive-mean and residual-scale branches.

The graph matrix in the implementation is indexed as $[\text{target}, \text{source}]$; thus its element $[j, i]$ is the score for $i \to j$. Throughout the manuscript, $\Theta_{i,j}$ denotes that same directed edge conceptually. All reported discovery metrics remove the diagonal, since autoregressive self-dependence is not the cross-variable discovery target.

A.2 Datasets and Preparation

A.2.1 Controlled distributional time series

For each synthetic process, $d = 32$ and $T = 2000$ retained observations follow a burn-in of 500 steps. Indices are circular: $m(j) = 1 + ((j-2) \bmod d)$ denotes the mean parent of target j, and $v(j) = 1 + ((j-4) \bmod d)$ denotes its distinct scale parent. The mean-parent and scale-parent cross-variable edge sets are therefore disjoint. An autoregressive self-edge and the $m(j) \to j$ edge belong to the mean graph; the $v(j) \to j$ edge belongs to the distribution graph. The union of the two graphs is the overall reference graph.

We use a common state-dependent scale function

$$q_j(t) = 0.12 \left[1 + 2.5 \ \text{sigmoid}\big(2X_{t-1,v(j)}\big)\right], \qquad \text{sigmoid}(u) = \frac{1}{1+e^{-u}}. \tag{14}$$

The innovation $\epsilon_{t,j}$ has zero mean and unit variance; all innovations are drawn independently across variables and time, conditional on the state-dependent scale. DLIN uses standardized Student-t innovations with five degrees of freedom and

$$X_{t,j} = 0.30X_{t-1,j} + 0.20X_{t-1,m(j)} + 0.05X_{t-2,m(j)} + q_j(t)\epsilon_{t,j}. \tag{15}$$

DTANH uses variance-one Laplace innovations and

$$X_{t,j} = 0.30X_{t-1,j} + 0.20\tanh\big(X_{t-1,m(j)}\big) + 0.05\sin\big(X_{t-2,m(j)}\big) + q_j(t)\epsilon_{t,j}. \tag{16}$$

HETERO-StudentT uses standardized Student-t innovations with five degrees of freedom and

$$X_{t,j} = 0.30X_{t-1,j} + 0.20\tanh\big(X_{t-1,m(j)}\big) + q_j(t)\epsilon_{t,j}. \tag{17}$$

For $Z \sim t_5$, the standardized innovation is $\epsilon = \sqrt{(5-2)/5}\, Z$; for a Laplace variable of unit scale, $\epsilon = Z/\sqrt{2}$. The simulated truth is stored separately as mean, distribution, and union matrices.

A.2.2 CausalRivers and DREAM3

The CausalRivers experiment fixes one connected ten-gauge subgraph in East Germany, resamples the series on a six-hour grid, and retains 3000 steps. The same gauges, time grid, and hydrological reference graph are reused across methods, missing rates, and random seeds. This reference graph does not label mean-driven and scale-driven edges separately, so it supports only overall reference-graph AUROC.

The DREAM3 evaluation uses the five Size-100 networks Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3. Each network contains 46 trajectories of 21 time points. Historical windows are formed within trajectories and no window crosses a trajectory boundary. Each network is scored against its own supplied directed reference graph. DREAM3 likewise does not provide the mean-versus-scale decomposition used for the synthetic distribution-edge metric.

A.3 Implementation and Evaluation Details

Missingness and calibration. Artificial MCAR masks are applied only to entries present in the source data. Rates $r \in \{0.2, 0.4, 0.6, 0.8\}$ are evaluated with five seeds per condition. The artificially hidden, originally observed entries are test targets and never enter training or calibration. A further 10% of the remaining observed entries are withheld from model fitting to estimate scale-calibration and interpolation-error quantities. Normalization statistics are estimated from training-available entries, not from test targets or native missing values.

Table 3: Principal saved settings for the Full experiments.

Hyperparameter	Value
History length L	2
Training epochs N	64
Batch size B	1024
Sparsity coefficient λ_{sp}	0.30
Predictor learning rate α_ϕ	$0.01 \to 0.001$
Graph learning rate α_Θ	$0.01 \to 0.001$
Gumbel–Softmax temperature τ	$1.0 \to 0.3$
MMD weight λ_{dist}	1.0
Residual draws S	32

Scale calibration. For variable j, let $\mathcal{C}_j$ be its calibration entries. A variable-wise factor can be estimated by

$$c_j = \sqrt{\frac{\sum_{t\in\mathcal{C}_j}(X_{t,j}-\mu_{t,j})^2}{\max\left(|\mathcal{C}_j|v_{\min}, \sum_{t\in\mathcal{C}_j}\sigma^2_{t,j}\right)}},$$
$$v_j^{\mathrm{int}} = \max\left(v_{\min}, \frac{1}{|\mathcal{C}_j|}\sum_{t\in\mathcal{C}_j}(X_{t,j}-I_{t,j})^2\right). \tag{18}$$

The implementation clips c_j to $[0.05, 20]$, uses $v_{\min} = 10^{-6}$, and falls back to the training-variable variance when too few calibration entries are available. These quantities are held fixed when forming the blend in Eq. 12; the blend itself is not an extra loss term. A small positive floor of 10^{-4} prevents zero predicted scales.

Training settings. The common predictor is a graph-gated recurrent model with one recurrent layer and a 32-unit hidden representation. Table 3 lists the principal settings saved with the Full runs. The MMD branch uses 32 Gaussian innovations per observed target, a distribution weight of 1.0 in $\mathcal{L}_{\mathrm{val}} + \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}}$, minibatch-wise observed-residual demeaning, and stop-gradient through the predicted mean. The Gaussian-kernel bandwidth is estimated from positive pairwise squared residual distances, detached from gradient computation, and bounded below by 10^{-6}. The learned graph is evaluated through its continuous edge scores.

Metrics. AUROC is computed over ordered, off-diagonal source–target pairs. For synthetic overall recovery, positives are the union of mean and distribution edges; for distribution-edge AUROC, positives are distribution edges and mean-only edges are excluded from the negative set. CausalRivers and DREAM3 use their supplied reference graphs. RMSE is calculated in the original value units only at artificial test targets that were observed in the source data. We report five-seed means. No best seed or missing rate is selected.

A.4 Hyperparameters and Initialization Sensitivity

We analyze one-factor changes around the Full configuration on DTANH at missing rate $r = 0.2$. The anchor uses MMD weight $\lambda_{\mathrm{dist}} = 1$, 32 generated residual samples, Gaussian predictive noise, a two-step history, 64 epochs, and the remaining settings in Table 3. It attains distribution-edge AUROC 0.7199, union-graph AUROC 0.8599, and RMSE 0.2703. Each non-anchor case changes only one named axis, so the comparisons below measure local sensitivity rather than interactions among hyperparameters.

Table 4 shows the complete MMD-weight sweep. Distribution-edge AUROC rises from 0.5676 at weight 0.1 to 0.7199 at weight 1.00, while union-graph AUROC rises from 0.7838 to 0.8599. Thus, in this controlled setting, the weight chiefly affects graph discrimination.

The tested levels of the most consequential remaining axes are presented below. 1) Changing the number of MMD samples from 8 to 48 produces a much smaller distribution-edge AUROC range (0.7034–0.7199) than changing the MMD weight. 2) Gaussian noise has the highest observed score

Table 4: DTANH one-factor MMD-weight sweep at $r = 0.2$.

λ_{dist}	Dist. AUROC ↑	Union AUROC ↑	RMSE ↓
0.10	0.5676	0.7838	0.2703
0.25	0.5969	0.7984	0.2703
0.50	0.6458	0.8229	0.2703
1.00	0.7199	0.8599	0.2703

among the tested predictive families, despite DTANH using Laplace innovations in its data generator. 3) The predictor and graph configuration also matter: two recurrent layers, a shared decoder, or a growing 16-group graph each substantially reduce distribution-edge recovery under the otherwise fixed anchor. These are one-factor observations, not an optimized joint configuration.

The remaining tested axes include batch size, decoder-state concatenation, predictor weight decay, final Gumbel temperature, supervision and fill schedules, the MMD bandwidth floor, and the scale and fusion-variance floors. Varying the bandwidth floor from 10^{-8} to 10^{-2} leaves mean distribution-edge AUROC unchanged at the reported precision, indicating limited sensitivity to this setting in the evaluated configuration. By contrast, a batch size of 256, a predictor width of 16, and a halved predictor learning rate yield 0.6823, 0.6330, and 0.6648, respectively. These contrasts indicate that training capacity and optimization choices can matter at least as much as numerical floors in this setting.

A.5 Full Results across Missing Rates

Figures 4–12 report the full results across all missing rates. Bars show means over five seeds, and error bars show dispersion across the five seeds. For each synthetic dataset, the panels show union-graph AUROC, distribution-edge AUROC, and RMSE; for CausalRivers and DREAM3, they show reference-graph AUROC and RMSE. The RMSE panels compare only CUTS+ and DistCUTS.

Distribution-dependent recovery across synthetic systems. Figures 4–6 show that the benefit of residual distribution matching extends across linear dynamics, nonlinear dynamics, and heavy-tailed innovations. The clearest improvements occur in distribution-edge recovery at lower missing rates, where the baselines often recover the union graph more effectively than the distribution-dependent subset. This distinction supports the motivation for complementing mean prediction: histories that explain conditional variability may provide little additional information about the target mean. The gains on DTANH and HETERO-StudentT are also informative because the MMD branch samples predictive residuals from Gaussian base noise, whereas the data-generating innovations are Laplace and Student-t, respectively. Useful scale-dependent information can therefore be learned even when the working innovation family differs from the data-generating distribution, although this does not imply recovery of its full shape.

Effect of increasing missingness. The synthetic results also reveal a common limitation: distribution-edge AUROC generally declines as missingness increases, and the advantage over CUTS+ narrows substantially at $r = 0.8$. Residual-scale learning relies on both observed targets and informative historical inputs. Increasing missingness reduces the available supervision while making those inputs more dependent on reconstructed values. The resulting degradation is consistent with uncertainty in the completed histories limiting the additional information provided by residual matching.

Recovery on CausalRivers. The reference-graph results in Figure 7 show that DistCUTS improves on CUTS+ most clearly when more observations are available, while their graph-recovery performance becomes similar at the highest missing rate. This pattern is consistent with the dependence of the proposed learning signal on informative temporal context. The reconstruction improvements reported in Table 1 complement this structural evaluation, but should not be attributed exclusively to better graph recovery: calibrated blending also directly changes how missing values are estimated. Moreover, the hydrological reference graph does not distinguish mean-driven from scale-driven effects. These results therefore assess recovery of the reference structure without establishing which recovered relations operate through conditional variability.

Graph recovery and reconstruction on DREAM3. Figures 8–12 show a more consistent advantage in reconstruction than in graph recovery. Across the five networks and four missing rates, DISTCUTS yields lower mean RMSE than CUTS+, whereas their AUROCs are often close and other baselines remain competitive on some networks. This difference suggests that improving missing-value estimates does not necessarily produce a comparable improvement in edge ranking. Scale-aware blending can directly improve missing-value estimates even when the resulting edge scores yield similar AUROC. The short trajectories also provide limited temporal context, which may constrain the structural benefit of the additional distributional objective. Because DREAM3 does not separately annotate scale-driven relations, its results support the utility of the combined framework for reconstruction, rather than directly verifying the recovery of distribution-dependent edges.

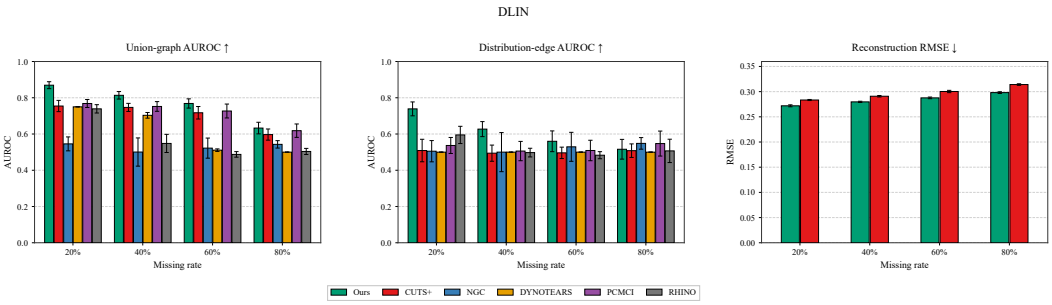


Figure 4: Full four-rate results on DLIN: union-graph AUROC, distribution-edge AUROC, and reconstruction RMSE. Each bar is a five-seed mean; error bars are the stored dispersion values.

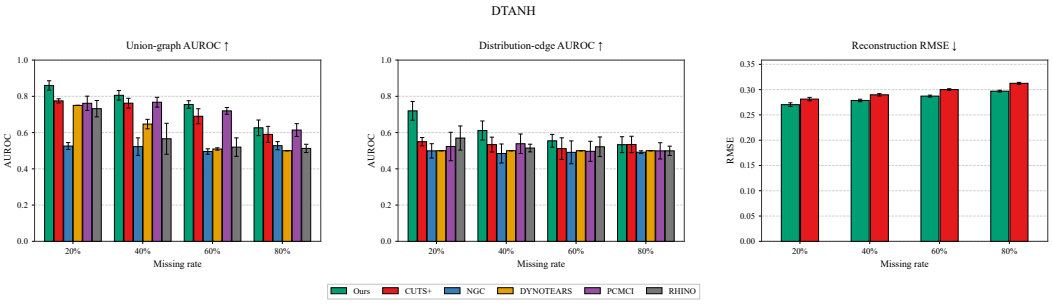


Figure 5: Full four-rate results on DTANH.

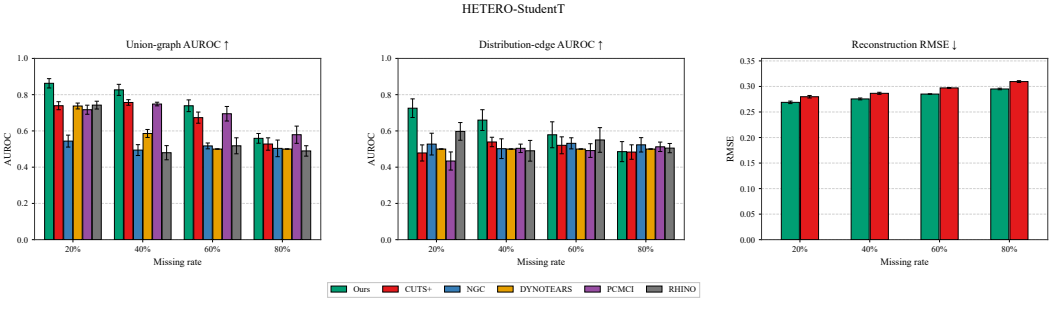


Figure 6: Full four-rate results on HETERO-StudentT.

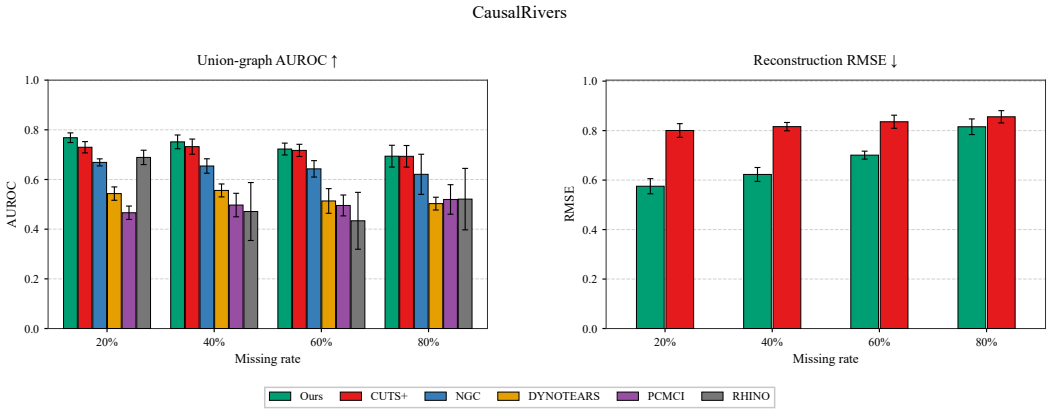


Figure 7: CausalRivers reference-graph AUROC and RMSE.

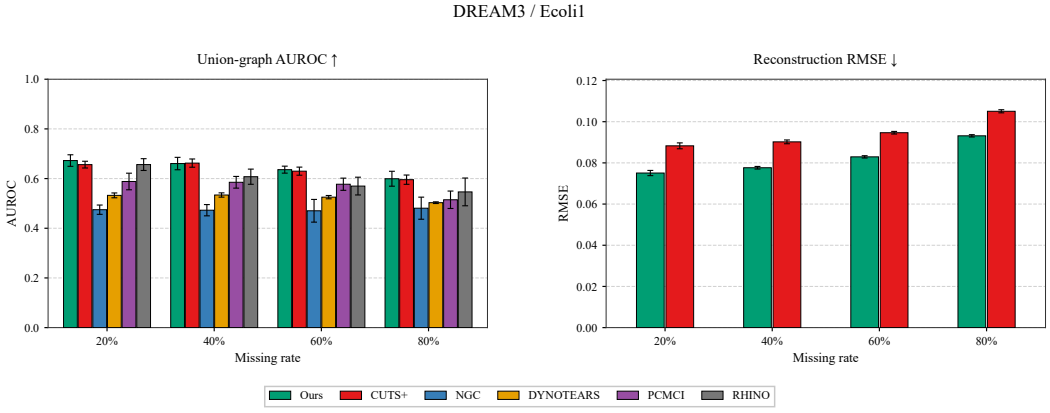


Figure 8: Ecoli1 reference-graph AUROC and reconstruction RMSE.

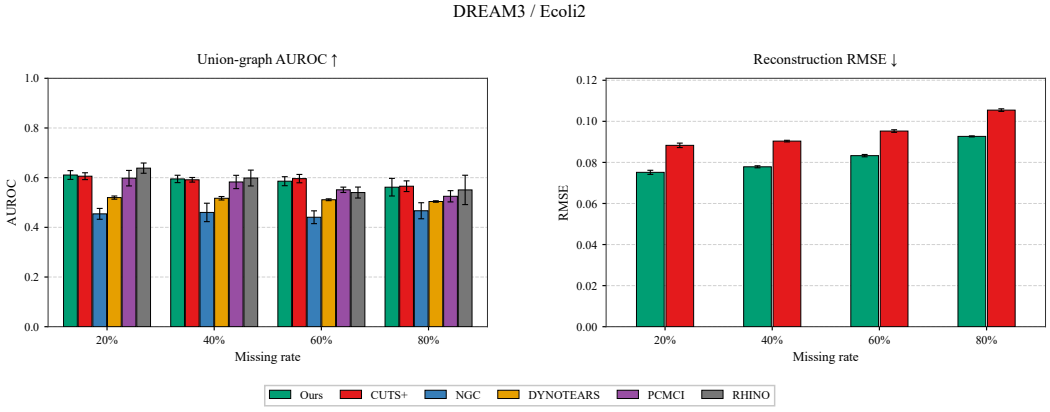


Figure 9: Ecoli2 reference-graph AUROC and reconstruction RMSE.

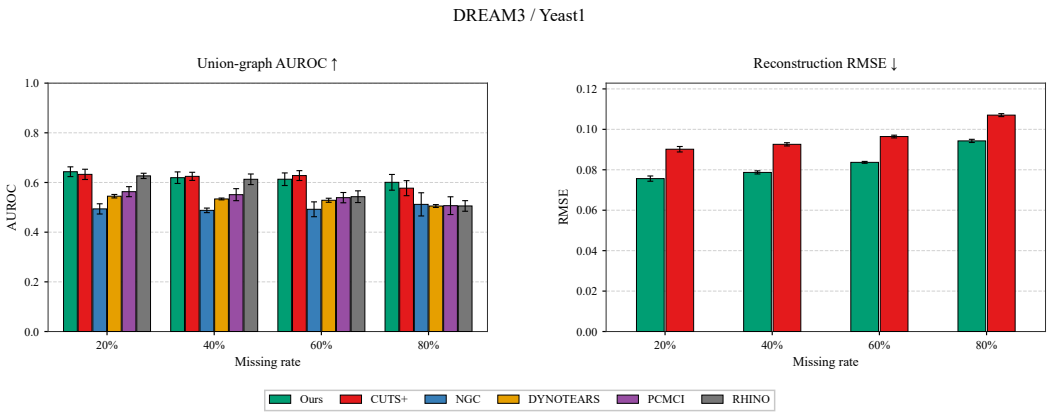


Figure 10: Yeast1 reference-graph AUROC and reconstruction RMSE.

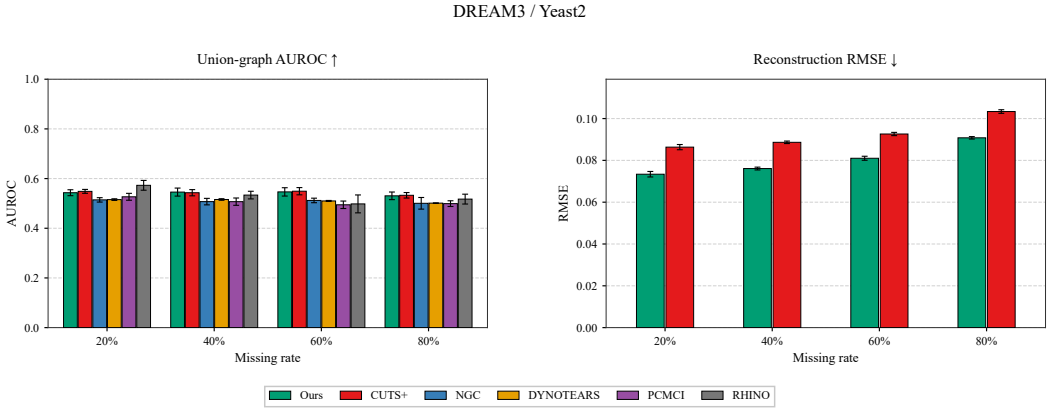


Figure 11: Yeast2 reference-graph AUROC and reconstruction RMSE.

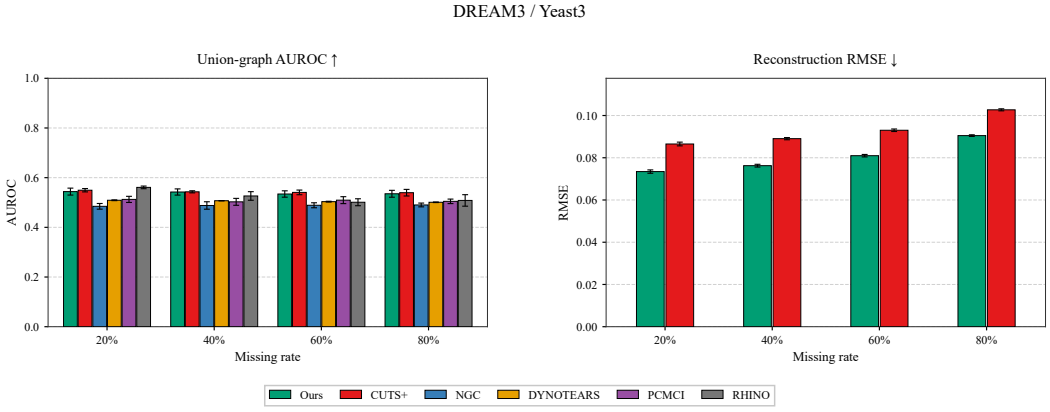


Figure 12: Yeast3 reference-graph AUROC and reconstruction RMSE.