MASK2CAUSE: TEMPORAL CAUSAL DISCOVERY BEYOND CAUSALITY IN MEAN

Anonymous authors
Paper under double-blind review

ABSTRACT

One approach to discovering causal relationships in multivariate time series is to ask whether a variable's history improves prediction of another variable. When this improvement is measured solely by reductions in optimal squared error, the criterion misses relationships that affect the target's conditional variance without changing its conditional mean. We propose Mask2Cause, an end-to-end Transformer framework for causal discovery through conditional mean and variance prediction. Each variable's history is embedded as a separate token, and a single learnable adjacency matrix, shared across layers and attention heads, controls information exchange between tokens. Gaussian negative log-likelihood and a sparsity penalty on this matrix jointly train the forecasting model and graph. Our population analysis connects the optimal log-loss reduction to causally conditioned directed information, providing an information-theoretic motivation for the training objective. We introduce Mixed Physics, a benchmark with separately controlled mean and variance dependencies, to evaluate recovery of mean-only and variance-only edges. In its pure-variance setting, Mask2Cause achieves 1.000 off-diagonal AUROC, compared with 0.521 for a corresponding squared-error model. Experiments on linear, chaotic, biological, and real-data-derived benchmarks demonstrate competitive graph recovery, while the inferred graphs support downstream tasks such as causal pruning for efficient forecasting and root-cause localization.

1 INTRODUCTION

Uncovering directed relationships within multivariate time series is a fundamental challenge in machine learning, with applications in forecasting, root cause analysis, and anomaly detection. Temporal causal discovery seeks to recover these relationships from observational data under assumptions on the underlying dynamics, using approaches such as conditional-independence testing (Runge et al., 2019), graph-score optimization (Pamfil et al., 2020), and restrictions on noise mechanisms (Hyvärinen et al., 2010; Peters et al., 2013). We adopt the predictive perspective of Granger causality (Granger, 1969): a time series Y Granger-causes another series X if including Y's history improves prediction of X beyond what is possible using the remaining observed histories, including X's own past. Such predictive dependence need not involve a change in the conditional mean: in heteroscedastic systems, a source may modulate the target's conditional variance. In financial markets, *volatility spillovers* describe how fluctuations in one asset's returns can help predict the volatility of another's (Diebold & Yilmaz, 2009), even when no corresponding predictive relationship is detected in mean returns (Gemici & Polat, 2021). A biological example arises in retinoic acid signaling, where changes in the binding protein Crabp2a alter temporal variability without changing mean retinoic acid levels (Sosnik et al., 2016). Yet a criterion based solely on reductions in optimal squared-error prediction cannot detect a source whose contribution is confined to conditional variance. Recovering these relationships therefore requires a discovery objective sensitive to predictive information beyond the mean.

To address this problem, we propose Mask2Cause (M2C), an end-to-end neural framework for discovering dependencies expressed through conditional mean, conditional variance, or both. We combine inverted variable embedding (Liu et al., 2024), which encodes each variable's history as a token, with adjacency-constrained masked attention, which gates attention according to learned parent relationships through a differentiable, sparsity-regularized adjacency matrix. Shared across encoder layers and attention heads, this matrix controls cross-variable information flow and is optimized

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.

directly alongside the forecasting weights. Its entries provide continuous edge scores, which are thresholded to obtain a stationary summary graph of lagged predictive dependencies. This design accommodates both squared-error prediction and heteroscedastic Gaussian likelihood prediction within this framework. In the latter, conditional mean and variance heads inform the same adjacency matrix, allowing either form of dependence to support an edge. Our population analysis motivates training objective with directed information (Quinn et al., 2015). We introduce Mixed Physics to evaluate mean-only and variance-only edge recovery through controlled comparisons, alongside standard discovery benchmarks and downstream applications.

Our contributions are fourfold: **(1) A shared graph-learning end-to-end architecture.** Unlike post-hoc frameworks, Mask2Cause integrates the causal structure directly into the forward pass. The structural adjacency matrix and the forecasting weights are optimized simultaneously, and the matrix that the sparsity penalty acts on is the causal graph itself, so no separate extraction step stands between training and the graph. This reduces the risk of the model overfitting to spurious correlations. **(2) A population-level analysis of the prediction objective.** We establish conditions under which a variance-only parent's history strictly reduces optimal Gaussian negative log-likelihood while leaving optimal squared error unchanged. **(3) A controlled benchmark for mean and variance dependence.** We introduce Mixed Physics, with separately controlled mean and variance parents, to evaluate graph recovery under combinations of the two dependence types. Corresponding squared-error and Gaussian-likelihood variants allow us to examine the effect of the prediction objective. **(4) An evaluation of recovery, efficiency, and downstream utility.** We assess graph recovery on linear, chaotic, biological, and real-data-derived benchmarks, examine architectural choices and computational costs, and evaluate the usefulness of inferred graphs for forecasting-model pruning and root cause localization.

2 Related Works

Classical Granger causal discovery commonly models linear dynamics through vector autoregression (Lozano et al., 2009; Lütkepohl et al., 2005). To accommodate nonlinear relationships, neural approaches introduced component-wise predictors, including SRU/eSRU, TCDF, and NGC (Khanna & Tan, 2020; Nauta et al., 2019; Tank et al., 2021), with more recent work employing Kolmogorov–Arnold networks (Liu et al., 2025a). Separate target predictors limit parameter sharing and can increase costs as dimensionality grows. Shared architectures address this limitation: CUTS+ uses message-passing networks for irregular time series (Cheng et al., 2024a), while JRNGC trains a joint predictor with Jacobian regularization (Zhou et al., 2024). Other approaches combine prediction with exogenous-variable modeling, as in AERCA (Han et al., 2025), or learn shared representations for dynamic discovery, as in UnCLe (Bi et al., 2026). Attention flexibly models temporal and cross-variable interactions, motivating CausalFormer, CSAM, and LLM-GC (Kong et al., 2024; Liu et al., 2025b; 2026). Graph extraction varies: CausalFormer uses relevance propagation, UnCLe temporal perturbations, and CSAM attention aggregation with permutation-based verification. Even with training-time structural constraints, separate attribution introduces choices that can affect how faithfully edges reflect learned dependencies. M2C shares forecasting transformations across targets and directly optimizes a sparsity-regularized adjacency that both gates attention across layers and heads and supplies edge scores. This couples graph estimation to prediction without separate attribution before thresholding. However, those relying on squared-error prediction lack a direct learning signal for variance-only dependencies.

Causality in variance has long been studied through residual-based tests (Cheung & Ng, 1996), multivariate GARCH models (Hafner & Herwartz, 2008), and AR-BEKK formulations combining mean and variance dynamics (Luo et al., 2011). These methods provide statistical inference under specified assumptions; complex nonlinear mean and variance interactions motivate neural alternatives. TS-LSNM models conditional location and scale using separate networks for each target, combining their weight-derived graph scores (Kikuchi & Shimizu, 2023). RHINO uses conditional normalizing flows and variational graph inference to capture history-dependent noise (Gong et al., 2022), while FANTOM additionally learns regime-specific graphs and regime boundaries (Rahmani & Frossard, 2026). M2C shares an attention-based encoder across targets and couples mean and variance prediction through one directly optimized summary adjacency, avoiding both per-target network replication and flow-based density estimation. Its Gaussian likelihood provides a simpler objective for learning dependencies expressed through either conditional moment, while Mixed Physics separately

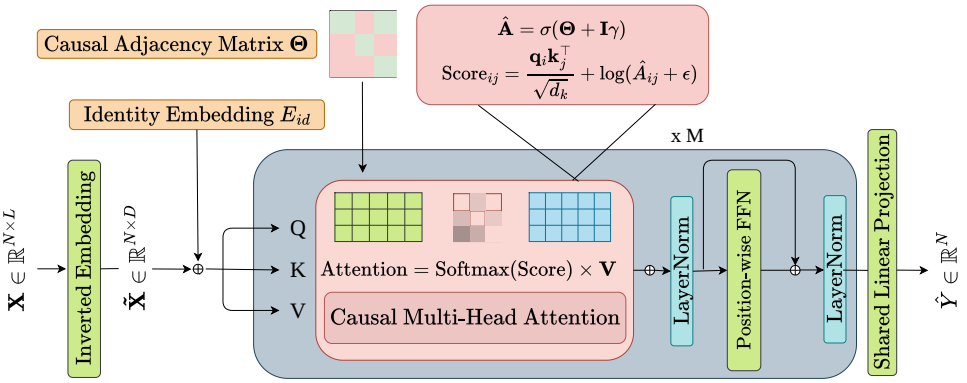


Figure 1: **The Mask2Cause Architecture.** The model maps a multivariate history into variable-specific tokens. A Transformer encoder, constrained by a learnable adjacency matrix, processes these tokens to predict the next-step state, discovering the causal graph through the forecasting objective.

evaluates mean-only, variance-only, and combined effects. To our knowledge, Mixed Physics is the first benchmark to do this. Appendix B provides detailed comparisons.

3 Mask2Cause Architecture

3.1 Problem Formulation and Assumptions

We observe $X \in \mathbb{R}^{T \times N}$ with state $x_t = [x_t^1, \ldots, x_t^N]^\top$. Let $x_{<t} = (x_1, \ldots, x_{t-1})$ denote its history and $x_{<t}^{-j}$ the history excluding variable j. We seek a stationary summary graph $A \in \{0,1\}^{N \times N}$, where $A_{ij} = 1$ denotes a predictive dependency $X^j \to X^i$ at one or more positive lags. The summary graph may contain feedback cycles; strict temporal precedence makes the time-unrolled graph acyclic. We learn this graph jointly with a forecasting model.

We ground the causal interpretation of the learned graph under the following assumptions commonly used in causal discovery, detailed in Appendix C: **(1) Strict Positivity:** The joint distribution is strictly positive. **(2) Causal Sufficiency:** There are no unobserved confounders. **(3) Strict Temporal Precedence:** Causal influences take time to propagate. **(4) Causal Faithfulness:** True causal effects do not perfectly cancel to produce statistical independencies. **(5) Stationarity and Finite Markov Order:** The process is stationary, with a time-invariant causal graph and Markov order at most L. Appendix D examines hidden confounding and switching graphs: confounding inflates scores among confounded children on VAR but has little effect on Lorenz-96; switching yields approximately the union of regime graphs, supporting screening. M2C approximates $x_{<t}$ with $x_{t-L:t-1}$.

M2C combines variable-history embeddings, adjacency-constrained attention, and shared prediction heads. We jointly optimize the continuous adjacency $\hat{A} \in [0,1]^{N \times N}$ and forecasting weights, then threshold it to obtain the summary graph (Appendix F).

3.2 Inverted Variable Embedding

Standard time-series Transformers embed each time step as a token, mixing variables and making their individual influences difficult to isolate Wu et al. (2021); Zhou et al. (2021). Instead, we project each variable's history window into a d-dimensional token:

$$\mathbf{e}_i = \text{Dropout}(\mathbf{W}_{emb} \mathbf{x}_{t-L:t-1}^i + \mathbf{b}_{emb}), \tag{1}$$

where $\mathbf{W}_{emb} \in \mathbb{R}^{d \times L}$ is learned. To distinguish between variables, we add a learnable variable identity embedding $\mathbf{E}_{id} \in \mathbb{R}^{N \times d}$ to the projected tokens:

$$\mathbf{Z}^{(0)} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]^\top + \mathbf{E}_{id}. \tag{2}$$

Each row of $\mathbf{Z}^{(0)} \in \mathbb{R}^{N \times d}$ encodes one variable's recent history and identity. Representing each variable's history as a token, produces an $N \times N$ attention matrix whose entries correspond to directed interactions between variables, allowing the learned adjacency to constrain information flow directly.

3.3 CAUSAL MASKED ENCODER

The core of our approach is the modification of the self-attention mechanism to enforce causal structure. In a standard Transformer, the attention score S_{ij} represents the relevance of token j to token i. We explicitly constrain these scores using a global, learnable adjacency parameter $\mathbf{\Theta} \in \mathbb{R}^{N \times N}$. A sigmoid yields the continuous adjacency, with diagonal forcing preserving access to each variable's own history:

$$\hat{\mathbf{A}} = \sigma(\mathbf{\Theta} + \mathbf{I} \cdot \gamma)\,. \tag{3}$$

Here $\mathbf{I}$ is the identity and a large constant γ (e.g., 100) makes diagonal entries approach 1. Although self-dependencies can be learned (Appendix O.4), this prior preserves the own-history access assumed by causally conditioned DI (Section 4). One adjacency is shared across all M layers and attention heads, representing a global graph rather than layer-specific dependencies.

Given token queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$, we add a logarithmic adjacency mask to attention scores:

$$\text{Score}_{ij} = \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_k}} + \log(\hat{A}_{ij} + \delta)\,, \tag{4}$$

where $\delta > 0$ ensures numerical stability. Near-unit gates leave attention essentially unchanged; near-zero gates suppress the source's contribution, with the positive floor making suppression approximate. Without this floor, a zero gate blocks attention. The head output is

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \hat{\mathbf{A}}) = \text{Softmax}(\text{Score})\mathbf{V}\,. \tag{5}$$

We stack M encoder blocks with masked attention, per-token layer normalization, and position-wise feed-forward networks implemented as 1×1 convolutions. These blocks learn nonlinear representations while sharing the graph constraint. Because normalization and feed-forward operations act independently on each token, masked attention is the sole route for cross-variable information flow.

3.4 PREDICTION AND OPTIMIZATION

The encoder output $\mathbf{Z}^{(M)} \in \mathbb{R}^{N \times d}$ feeds one of two prediction configurations:

1. Heteroscedastic Model (Variance-Aware). We model a conditional Gaussian $\mathcal{N}(\mu, \sigma^2)$ with time-varying moments predicted by two linear heads: $\hat{\mu}_t = \mathbf{Z}^{(M)}\mathbf{w}_\mu + \mathbf{b}_\mu$ and $\hat{\sigma}_t^2 = \text{Softplus}(\mathbf{Z}^{(M)}\mathbf{w}_\sigma + \mathbf{b}_\sigma) + \delta$. Negative log-likelihood (NLL) weights prediction errors by their estimated uncertainty:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{|\mathcal{B}|N} \sum_{t \in \mathcal{B}} \sum_{i=1}^{N} \left(\frac{1}{2} \log(\hat{\sigma}_t^i)^2 + \frac{(x_t^i - \hat{\mu}_t^i)^2}{2(\hat{\sigma}_t^i)^2} \right)\,, \tag{6}$$

where $\mathcal{B}$ contains the mini-batch time indices and N is the variable count.

2. Homoscedastic Model (Mean-Only). For additive noise with input-independent variance (e.g., standard VAR), we omit variance estimation and use one linear head, $\hat{\mathbf{x}}_t = \mathbf{Z}^{(M)}\mathbf{w}_{\text{out}} + \mathbf{b}_{\text{out}}$. With unit variance, NLL is equivalent, up to scaling and an additive constant, to mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{B}|N} \sum_{t \in \mathcal{B}} ||\mathbf{x}_t - \hat{\mathbf{x}}_t||_2^2\,. \tag{7}$$

Joint Optimization. We combine forecasting loss $\mathcal{L}_{\text{pred}} \in \{\mathcal{L}_{\text{NLL}}, \mathcal{L}_{\text{MSE}}\}$ with off-diagonal adjacency sparsity:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda \cdot \frac{1}{N(N-1)} \sum_{i \neq j} \hat{A}_{ij}\,. \tag{8}$$

Prediction and sparsity jointly optimize the same adjacency whose entries become the reported edge scores. In methods that penalize input weights or per-window coefficient tensors, graph extraction instead requires choices such as absolute values, aggregation over time and lags, and threshold selection. M2C avoids this intermediate attribution step: one graph parameter gates attention across layers and heads, receives the sparsity penalty, and is thresholded for evaluation.

4 THEORETICAL MOTIVATION

4.1 FROM FUNCTIONAL CAUSALITY TO DIRECTED INFORMATION

Neural Granger methods learn a predictor $g_i(\mathbf{x}_{<t})$ of x_t^i and identify noncausality through functional invariance (Tank et al., 2021): changing $\mathbf{x}_{<t}^j$ while holding the remaining histories fixed does not change the prediction. Under squared-error loss, the population-optimal predictor is the conditional mean. A source that changes only the target's conditional variance therefore leaves this prediction unchanged, even though it affects the distribution of possible outcomes (Appendix U.4 gives a concrete example).

To capture these dependencies, we broaden the prediction from a single value to a conditional distribution. Directed Information (DI) quantifies the reduction in uncertainty about a target's next value obtained from a source's past, beyond what the remaining histories already provide. The Directed Information Graph (DIG) contains an edge $X^j \to X^i$ if and only if the DI from X^j to X^i, causally conditioned on the other variables $\mathbf{X}^{-\{i,j\}}$, is strictly positive. For $j \neq i$, this quantity is the expected cumulative KL divergence between the true full and source-excluded conditional distributions:

$$I(X^j \to X^i \| \mathbf{X}^{-\{i,j\}}) = \sum_{t=1}^{T} \mathbb{E}\Big[D_{\mathrm{KL}}\Big(p(x_t^i \mid \mathbf{x}_{<t}) \,\|\, p(x_t^i \mid x_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}}) \Big) \Big]. \tag{9}$$

Under history-independent, zero-mean additive noise and finite first moments, functional noncausality and distributional noncausality coincide almost surely, making the functional criterion a special case of the DIG formulation (Appendix U.1). Under positivity and strict causality, the DIG also equals the unique minimal generative model graph (Quinn et al., 2015, Theorem 1): its parent sets are the smallest sets of histories preserving the full predictive distributions.

4.2 TRANSLATING DIRECTED INFORMATION INTO MODEL DESIGN

Proposition (Optimal log-loss reduction). Let $\mathcal{L}(p, x) = -\log p(x)$. For ideal predictors of the true full and source-excluded conditional densities, assuming finite expected losses,

$$\mathbb{E}\left[\sum_{t=1}^{T} \big(\mathcal{L}(p_{-j,t}, x_t^i) - \mathcal{L}(p_{\mathrm{full},t}, x_t^i) \big) \right] = I(X^j \to X^i \| \mathbf{X}^{-\{i,j\}}), \tag{10}$$

where $p_{\mathrm{full},t} = p(\cdot \mid \mathbf{x}_{<t})$ and $p_{-j,t} = p(\cdot \mid \mathbf{x}_{<t}^{-j})$ (Quinn et al., 2015, Proposition 1).

For a candidate parent set S, let $\mathcal{L}_S$ denote the cumulative negative log-likelihood of the true conditional densities $p(x_t^i \mid x_{<t}^i, \mathbf{x}_{<t}^S)$, and let $\mathcal{L}_{\mathrm{own}}$ denote the corresponding loss using only $p(x_t^i \mid x_{<t}^i)$. The same identity gives $I(\mathbf{X}^S \to X^i) = \mathbb{E}[\mathcal{L}_{\mathrm{own}} - \mathcal{L}_S]$: the information supplied by S equals the reduction in optimal expected log-loss obtained by adding its histories to the target's own past. The own-history term is constant with respect to S; maximizing the captured predictive information therefore amounts to minimizing negative log-likelihood. This motivates our combination of likelihood prediction and an L_1 penalty on the adjacency: prediction rewards informative histories, while the sparsity penalty discourages unnecessary connections.

M2C approximates the ideal density predictors with a parameterized network and finite windows. Variable-specific tokens and masked attention connect graph learning to source-history access. With $M = 1$, a zero gate suppresses a source. Consequently, the network's empirical increase in predictive loss points to a positive DI for the excluded parent. Furthermore, we establish in Appendix G that the approximation continues to hold in practice for multiple Transformer blocks ($M \geq 2$), allowing for the use of deep architectures that leverage the capacity of Transformers in capturing complex relations. By simultaneously imposing an L_1 penalty on the adjacency matrix, the optimization process is incentivized to retain only those parents X^j whose inclusion reduces this loss, or equivalently, those that contribute non negligible Directed Information, thereby incentivizing recovery of the DIG.

Gaussian NLL provides a prediction objective sensitive to both mean and variance dependencies. Under finite second moments, positive conditional variances and suitable integrability, minimizing its population expectation over unrestricted mean and variance predictors yields the true conditional

moments, even for non-Gaussian data. Consequently, including a source's history strictly reduces optimal Gaussian NLL if and only if it changes either conditional moment on a set of positive probability; optimal MSE decreases only when the conditional mean changes (Appendix U.2). When the true full-history conditional distribution is Gaussian, its mean and variance determine the entire distribution. In this case, a source has positive DI if and only if including its history strictly reduces optimal Gaussian NLL, even when the source-excluded conditional distribution is non-Gaussian (corollary in Appendix U.2).

Alternative likelihoods retain the log-loss motivation; Appendix I compares Gaussian and Student-t likelihoods empirically. Appendix U.6 shows how prediction errors affect the connection between fitted likelihood models and DI.

Diagonal Forcing as Structural Prior. Because the target's own history is already conditioned on in Eq. 9, its self-directed information under this definition is zero (Appendix U.3), without implying that its past lacks predictive value. Diagonal forcing preserves this own-history access and matches the self-dependencies ($W_{ii} \neq 0$) in standard benchmark graphs.

5 Experiments

Our experiments test whether M2C remains competitive on mean-driven benchmarks and recovers variance-driven dependencies missed by squared-error prediction (Section 4). We cover sparse linear systems, chaotic dynamics, gene regulatory networks and series generated from traffic, air-quality and medical recordings, including high-dimensional settings where standard configurations no longer saturate. Mixed Physics separates mean- and variance-driven edges with known ground truth.

5.1 Baselines and Metrics

We compare against component-wise, shared, attention-based and statistical baselines: cMLP and cLSTM Tank et al. (2021), TCDF Nauta et al. (2019), CUTSCheng et al. (2023), CUTS+ Cheng et al. (2024a), Causalformer Kong et al. (2024), SRU and eSRU Khanna & Tan (2020), PCMCI Runge et al. (2019), NGM Bellot et al. (2021), LCCM De Brouwer et al. (2020), and the more recent AERCA Han et al. (2025) and UnCLe Bi et al. (2026), as well as RHINO Gong et al. (2022). For KANGCI Liu et al. (2025a), LLM-GC Liu et al. (2026) and CSAM Liu et al. (2025b), we were unable to obtain a runnable implementation. Appendix R separately compares their published numbers with M2C evaluated on the reported configurations. Following standard practice Tank et al. (2021), we report AUROC: the probability that a true edge ranks above a non-edge, obtained by sweeping thresholds over continuous adjacency scores $\hat{A}_{ij} \in [0, 1]$. Appendix K reports AUPRC, SHD and F1.

Scoring convention. On every benchmark drawn from prior work we follow the scoring protocol established by its source paper, which is what allows the published baseline numbers to be quoted directly. Mixed Physics is our own benchmark and carries no such constraint, so we score it off-diagonal, removing self-loops from the metric and reporting only the inter-variable structure the benchmark was built to test.

5.2 Datasets

5.2.1 Standard Benchmarks

We use dataset implementations from the official SRU Khanna & Tan (2020) and CausalTime Cheng et al. (2024b) repositories.

We assess performance on synthetic systems including the Vector Autoregressive (VAR) processes to test sparse linear recovery, and the chaotic Lorenz-96 system Karimi & Paul (2010) (forcing constants $F = 10, 40$) to evaluate nonlinear continuous-time dynamics. To validate efficacy on realistic, high-dimensional proxies, we employ the DREAM3 gene regulatory networks Prill et al. (2010) and the CausalTime benchmark suite Cheng et al. (2024b). Both are governed by real-world dynamics. The DREAM3 topologies are sub-networks extracted from the experimentally derived *E. coli* and *S. cerevisiae* transcriptional maps, so the structure being recovered is a real regulatory network, and CausalTime generates its series from deep models fitted to real air-quality, traffic and medical recordings, so the dynamics inherit the statistical properties of the source data. These are

the standard means of obtaining real causal structure in domains where interventional ground truth is unobtainable. DREAM3, which features extremely short trajectories ($T = 21$), is purely used to evaluate how the model fares against baselines in highly data-scarce domains. We also evaluate lag-only ER type graphs from RHINO's synthetic benchmark with history-dependent noise (Gong et al., 2022), using shared datasets and off-diagonal summary-graph AUROC after aggregating lagged edges.

5.2.2 MIXED PHYSICS (HETEROSCEDASTIC)

Standard benchmarks (VAR, Lorenz) model interactions strictly via the conditional mean. To evaluate dependencies expressed through conditional variance, we introduce Mixed Physics, a synthetic benchmark with separately controlled mean and variance effects.

Generation Mechanism. We simulate a system where disjoint sets of cross-variable parent variables control the mean and the variance of the target. Mean self-dependencies are retained in every configuration. For variable i, the state evolves according to:

$$x_t^i = \underbrace{\sum_j \mathbf{W}_{ij}^{\mu} x_{t-L}^j}_{\text{Causality in Mean}} + \underbrace{\sqrt{\beta + \sum_k \mathbf{W}_{ik}^{\sigma} (x_{t-L}^k)^2}}_{\text{Causality in Variance}} \cdot \eta_t^i \tag{11}$$

where the innovations η_t^i are independent standard Gaussian variables across targets and time, independent of the past, L is a fixed time lag, and $\beta > 0$ is a baseline variance. The matrices $\mathbf{W}^{\mu}$ and $\mathbf{W}^{\sigma}$ represent the causal strengths for mean and volatility, respectively. A causal edge $j \to i$ is either a *Mean Edge* ($\mathbf{W}_{ij}^{\mu} \neq 0, \mathbf{W}_{ij}^{\sigma} = 0$) or a *Variance Edge* ($\mathbf{W}_{ij}^{\mu} = 0, \mathbf{W}_{ij}^{\sigma} > 0$). This distinction renders Variance-type edges invisible to standard MSE-based models. To analyze the sensitivity of our NLL objective to heteroscedastic dynamics, we vary the ratio of Variance-to-Mean edges across three regimes: 50:50, 75: 25, 100:0 (Pure Heteroscedastic control).

Second-moment stability. Unconstrained coefficients can produce diverging second moments, while artificially bounding the state can distort variance-driven dependencies. We therefore rescale the coefficients and verify stability without clipping the state. Let $\mathbf{S}_t = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top]$. Independence of the innovations gives the exact recursion

$$\mathbf{S}_t = \mathcal{T}(\mathbf{S}_{t-L}) + \beta \mathbf{I}, \qquad \mathcal{T}(\mathbf{S}) = \mathbf{W}^{\mu} \mathbf{S} (\mathbf{W}^{\mu})^\top + \mathrm{Diag}(\mathbf{W}^{\sigma}\, \mathrm{diag}(\mathbf{S}))\,, \tag{12}$$

where diag extracts a matrix's diagonal and Diag constructs a diagonal matrix. This expression retains cross moments between mean parents. If $\rho(\mathcal{T}) < 1$, where ρ denotes the spectral radius of the linear operator, its repeated applications decay and the accumulated innovation contributions converge. The second moments therefore approach the finite fixed point

$$\mathbf{S}_* = \beta \sum_{r=0}^{\infty} \mathcal{T}^r(\mathbf{I}), \qquad \mathbf{S}_* = \mathcal{T}(\mathbf{S}_*) + \beta \mathbf{I}. \tag{13}$$

Rescaling $\mathbf{W}^{\mu}$ by $\sqrt{k}$ and $\mathbf{W}^{\sigma}$ by k scales $\mathcal{T}$ by k without changing edge positions or mean/variance assignments. After rescaling, the generated configurations have spectral radii ≈ 0.9 satisfying the stability condition. Appendix J.5 provides the coefficient sampling, rescaling procedure, and detailed derivation.

Synthetic data provides known ground truth for variance-driven dependencies. In real data, financial connectedness indices and neural coherence measures provide estimated rather than known structure, so agreement with these measures would not by itself establish causal correctness. Appendix S provides evidence of forecasting utility on real world dataset ETTh1: pruning to M2C-NLL's graph improves forecasting over M2C-MSE's graph. Appendix J details all datasets and ground truths.

5.3 CAUSAL DISCOVERY RESULTS

On Lorenz-96 and VAR (Table 1), both M2C variants achieve near-perfect recovery. Several baselines also saturate at $N = 10$; at $N = 100$, M2C achieves 0.87–0.95 AUROC against 0.51–0.85 for the compared baselines (Appendix L). M2C also leads on Traffic and Medical in CausalTime (Table 2). On DREAM3 (Table 3), M2C-NLL ranks first on three of five networks despite trajectories of only

Table 1: **Average AUROC on Lorenz–96 and VAR systems.** Baselines are taken from Khanna & Tan (2020), except those marked with † which were reproduced using the official implementations.

Model	Lorenz–96				VAR	
	$F = 10$		$F = 40$		$T = 500$	$T = 1000$
	$T = 250$	$T = 500$	$T = 250$	$T = 500$		
cMLP(NGC)	0.93 ± 0.02	0.96 ± 0.03	0.85 ± 0.08	0.94 ± 0.03	0.94 ± 0.03	0.93 ± 0.02
cLSTM(NGC)	0.90 ± 0.02	0.95 ± 0.05	0.78 ± 0.09	0.90 ± 0.05	0.79 ± 0.12	0.80 ± 0.09
TCDF	0.70 ± 0.01	0.72 ± 0.04	0.62 ± 0.01	0.68 ± 0.04	0.77 ± 0.07	0.78 ± 0.04
SRU	0.84 ± 0.03	0.90 ± 0.02	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.82 ± 0.06	0.91 ± 0.04
eSRU	0.95 ± 0.02	0.98 ± 0.01	0.99 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.93 ± 0.05	0.98 ± 0.01
Causalformer†	0.68 ± 0.09	0.75 ± 0.04	0.85 ± 0.05	0.89 ± 0.10	0.66 ± 0.07	0.79 ± 0.07
CUTS+†	0.77 ± 0.06	0.84 ± 0.14	0.94 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.97 ± 0.02	0.99 ± 0.005
AERCA†	0.68 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.81 ± 0.04	0.97 ± 0.01	0.79 ± 0.01	0.91 ± 0.02
UnCLe†	0.92 ± 0.05	0.87 ± 0.02	0.95 ± 0.02	0.93 ± 0.01	0.83 ± 0.01	0.93 ± 0.05
M2C(MSE)	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.01	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$
M2C(NLL)	$\mathbf{0.99 \pm 0.02}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$

Table 2: **Average AUROC on CausalTime dataset.** Baselines are taken from Cheng et al. (2024b); †reproduced via official implementations.

Method	Traffic	AQI	Medical
PCMCI	0.54 ± 0.07	0.53 ± 0.07	0.70 ± 0.01
NGC	0.60 ± 0.01	0.72 ± 0.01	0.57 ± 0.01
TCDF	0.50 ± 0.00	0.41 ± 0.02	0.63 ± 0.04
CUTS	0.62 ± 0.02	0.60 ± 0.00	0.37 ± 0.03
CUTS+	0.62 ± 0.08	$\mathbf{0.89 \pm 0.02}$	0.82 ± 0.02
LCCM	0.55 ± 0.03	0.86 ± 0.07	0.80 ± 0.02
eSRU	0.60 ± 0.02	0.82 ± 0.03	0.76 ± 0.04
SCGL	0.59 ± 0.06	0.49 ± 0.05	0.50 ± 0.02
AERCA†	0.65 ± 0.01	0.50 ± 0.02	0.59 ± 0.01
UnCLe†	0.60 ± 0.01	0.66 ± 0.02	0.64 ± 0.04
M2C(MSE)	0.68 ± 0.04	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.90 \pm 0.05}$
M2C(NLL)	$\mathbf{0.71 \pm 0.02}$	0.84 ± 0.06	0.86 ± 0.03

Table 3: **AUROC on DREAM3 dataset.** Baselines are taken from Khanna & Tan (2020); †reproduced via official implementations.

Model	E.coli-1	E.coli-2	Yeast-1	Yeast-2	Yeast-3
cMLP(NGC)	0.644	0.568	0.585	0.506	0.528
cLSTM(NGC)	0.629	0.609	0.579	0.519	0.555
TCDF	0.614	0.647	0.581	0.556	**0.557**
SRU	0.657	0.666	0.617	0.575	0.550
eSRU	0.660	0.629	0.627	0.557	0.550
Causalformer †	0.420	0.261	**0.637**	0.483	0.484
CUTS+ †	0.585	0.563	0.547	0.532	0.509
AERCA †	0.668	0.616	0.621	0.567	0.465
UnCLe †	0.402	0.556	0.458	0.535	**0.558**
M2C(MSE)	0.643	0.672	**0.637**	0.563	0.530
M2C(NLL)	**0.672**	**0.703**	0.605	**0.578**	**0.557**

$T = 21$. Its AUROC ranges from 0.557 to 0.703, while several baselines fall below chance on individual networks.

Mixed Physics (Table 4) provides an idealized reference for the mean-prediction limitation in Section 4.1. Perfect ranking of mean edges and chance ranking of variance-only edges predicts AUROCs of 0.741, 0.630 and 0.500 from the respective edge counts (13/14, 7/20 and 0/27; derivation in Appendix U.5). M2C-MSE scores 0.764, 0.633 and 0.521, each within half a standard deviation of this reference. On the pure-variance Mixed Physics, M2C-NLL achieves 1.000 off-diagonal AUROC across five seeds. Given that Mixed Physics is a novel benchmark that separates mean- and variance-driven parents, we also investigate whether M2C can distinguish them. Appendix E shows that a two-graph extension achieves strong parent-type separation with near-perfect AUROC. Appendix T discusses objective selection; Appendix O reports sensitivity and hyperparameter studies.

On RHINO's synthetic benchmark (Table 4), M2C-NLL exceeds M2C-MSE by 0.110–0.322 AUROC and has the highest mean at $N = 5, 10, 40$; at $N = 20$, it scores 0.888 versus RHINO's 0.890.

Computational Analysis. Mask2Cause scales linearly with history length and quadratically with variable count, with lower computational costs than the compared baselines in large-system settings. At $N = 10, L = 5$, it requires 2.05M FLOPs and approximately 0.10M parameters (Appendix N).

5.4 Downstream Utility

M2C's graphs support forecasting and root cause localisation. On DREAM3, graph-guided pruning removes 59–99% of forecasting-model parameters while reducing MSE by 31–41% (Appendix P). On MSDS, M2C-NLL attains mean localisation accuracies of 0.446 at rank 1 and 0.995 within the top 5, supporting cloud microservice diagnosis (Appendix Q).

Table 4: **Off-diagonal AUROC on Mixed Physics and RHINO synthetic data.** Mixed Physics results use five seeds; the mean-based reference assumes perfect ranking of mean edges and chance ranking of variance-only edges. RHINO settings use lag-only graphs.

	Mixed Physics (variance:mean)			RHINO synthetic noise (History-dependent ER)			
Method	**50:50**	**75:25**	**100:0**	**N = 5**	**N = 10**	**N = 20**	**N = 40**
Mean-based reference	*0.741*	*0.630*	*0.500*	–	–	–	–
RHINO†	0.981 ± 0.000	0.982 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.798 ± 0.109	0.839 ± 0.060	$\mathbf{0.890 \pm 0.032}$	0.901 ± 0.024
TS-LSNM†	0.903 ± 0.002	0.863 ± 0.008	0.737 ± 0.003	0.687 ± 0.111	0.754 ± 0.076	0.752 ± 0.039	0.742 ± 0.049
M2C(MSE)	0.764 ± 0.047	0.633 ± 0.049	0.521 ± 0.047	0.674 ± 0.068	0.754 ± 0.076	0.633 ± 0.028	0.592 ± 0.039
M2C(NLL)	$\mathbf{0.998 \pm 0.003}$	$\mathbf{0.997 \pm 0.005}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.885 \pm 0.072}$	$\mathbf{0.864 \pm 0.045}$	0.888 ± 0.019	$\mathbf{0.914 \pm 0.016}$

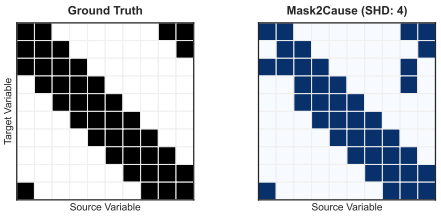


Figure 2: Ground truth and predicted causal graph for Lorenz-96 ($F = 40$, $T = 250$, AUROC = 0.98).

Table 5: Mean ablation AUROC: M2C-MSE on CausalTime and M2C-NLL on Mixed Physics.

Variant	**M2C-MSE CausalTime**	**M2C-NLL Mixed Physics**
Multiplicative Mask	0.74 ± 0.06	0.78 ± 0.03
Additive Bias Mask	0.73 ± 0.05	0.75 ± 0.01
Layer-wise Masks	0.66 ± 0.04	0.77 ± 0.04
Decoupled Heads	0.61 ± 0.05	0.86 ± 0.01
Residual Pred	0.78 ± 0.11	0.96 ± 0.01
Full Model	$\mathbf{0.82 \pm 0.09}$	$\mathbf{0.99 \pm 0.01}$

6 Ablation Study

We ablate masking, the global adjacency, shared projection heads and prediction targets in both M2C variants (Table 5). We retain variable tokenization because timestep attention lacks the $N \times N$ variable-pair structure required by $\hat{\mathbf{A}}$.

(1) **Alternative masking mechanisms:** We replace the log-barrier of Eq. 4 with two variants that keep a learnable $\hat{\mathbf{A}}$ and end-to-end extraction and differ only in how it enters attention: an *unnormalized multiplicative* gate $(\hat{\mathbf{A}} \odot \text{Softmax}(\mathbf{S}))\mathbf{V}$, and a *bounded additive bias* $\alpha \hat{A}_{ij}$ with α tuned per dataset. The multiplicative gate loses 0.08 AUROC on CausalTime and 0.21 on Mixed Physics, and the additive bias 0.09 and 0.24. A renormalized multiplicative gate is absent because it is algebraically identical to ours. Appendix H derives that equivalence and shows that the log gate approximately cancels sigmoid attenuation in the pre-softmax score derivative when $\delta \ll \hat{A}_{ij} \ll 1$.
(2) **Layer-wise Masks (No Global $\hat{\mathbf{A}}$):** Each layer learns its own mask; averaging them after training to obtain the graph costs 0.16 AUROC on CausalTime and 0.22 on Mixed Physics.
(3) **Decoupled Heads:** We replace the shared linear projection with a separate head per variable. The shared projection performs better, suggesting that learning common dynamics ("shared physics") reduces the parameter cost and overfitting risk of variable-specific functions.
(4) **Residual Prediction:** Predicting the raw state $\mathbf{x}_{t+1}$ rather than the increment $\Delta\mathbf{x}_{t+1}$ yields marginally better structural recovery.

7 Conclusion

Mask2Cause treats variables as tokens and constrains attention through a learned adjacency, recovering graphs in one shared model without component-wise networks or separate post-hoc extraction. Directed Information encompasses the mean-based criterion under history-independent additive noise; variance-only parents escape the population mean criterion while carrying positive DI. M2C detects these edges using log likelihood instead of squared error. M2C achieves competitive recovery across linear, chaotic, biological and real-data-derived benchmarks, with substantial computational savings in large-system settings. On the Mixed Physics benchmark we release, it recovers variance-driven structure that mean-based methods miss. The discovered graphs also transfer downstream: on DREAM3, graph-guided pruning removes 59–99% of forecasting-model parameters while reducing MSE by 31–41%, and on MSDS the graphs support accurate root cause localisation.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the checking of proofs and to provide feedback on research methodology and experiments. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, implement methods, clean and reformat dataset, or intepret results. Translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for discovering research topics and identifying gaps, brainstorming, sourcing and searching for information, improving readability, aid writing code, and identifying relevant literature. We have reviewed all AI-assisted work. The LLM-generated code was verified by the authors. The authors conducted the research analyses and interpreted the results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research was conducted in accordance with the ICLR Code of Ethics. The study evaluates a methodological contribution using synthetic and existing benchmark datasets. Potential downstream impacts and limitations of causal interpretation are discussed in Appendix V.

REPRODUCIBILITY STATEMENT

The models and newly introduced datasets and scripts to generate them are included in the supplementary material. Appendix J describes the datasets, and Appendix O details the evaluation protocols, model configurations, hyperparameter tuning, and baseline implementation settings. Appendix R distinguishes comparisons using published baseline results. The theoretical assumptions and proofs are provided in Appendices C and U, respectively.

REFERENCES

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. *arXiv preprint arXiv:2105.02522*, 2021.

Tingzhu Bi, Yicheng Pan, Xinrui Jiang, Huize Sun, Meng Ma, and Ping Wang. Uncle: Towards scalable dynamic causal discovery in non-linear temporal systems. *Advances in Neural Information Processing Systems*, 38:96158–96187, 2026.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024a.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations*, volume 2024, pp. 2923–2945, 2024b.

Yin-Wong Cheung and Lilian K Ng. A causality-in-variance test and its application to financial market prices. *Journal of econometrics*, 72(1-2):33–48, 1996.

Edward De Brouwer, Adam Arany, Jaak Simm, and Yves Moreau. Latent convergent cross mapping. In *International Conference on Learning Representations*, 2020.

Francis X Diebold and Kamil Yilmaz. Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, 119(534):158–171, 2009.

Eray Gemici and Müslüm Polat. Causality-in-mean and causality-in-variance among bitcoin, litecoin, and ethereum. *Studies in Economics and Finance*, 38(4):861–872, 2021.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Christian M Hafner and Helmut Herwartz. Testing for causality in variance using multivariate garch models. *Annales d'Economie et de Statistique*, pp. 215–241, 2008.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, volume 2025, pp. 45137–45153, 2025.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Alireza Karimi and Mark R Paul. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4), 2010.

Saurabh Khanna and Vincent Y. F. Tan. Economy statistical recurrent units for inferring nonlinear granger causality. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SyxV9ANFDH`.

Genta Kikuchi and Shohei Shimizu. Structure learning for groups of variables in nonlinear time-series data with location-scale noise. In *Causal Analysis Workshop Series*, pp. 20–39. PMLR, 2023.

Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Causalformer: An interpretable transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Bo Liu, Hongyan Li, and Shenda Hong. Llm-gc: Advancing granger causal discovery from time series with multimodel language modeling. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 387–395, 2026.

Meiliang Liu, Yunfang Xu, Zijin Li, Zhengye Si, Xiaoxiao Yang, Xinyue Yang, and Zhiwen Zhao. Kolmogorov-arnold networks for time series granger causality inference. *arXiv preprint arXiv:2501.08958*, 2025a.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Yusen Liu, Yong Wang, Yifan Yin, Tianqing Zhu, Xiufeng Liu, and Huan Huo. Causal discovery with inverted self-attention for multivariate time series. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 167–179. Springer, 2025b.

Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 577–586, 2009.

Qiang Luo, Tian Ge, and Jianfeng Feng. Granger causality with signal-dependent noise. *Neuroimage*, 57(4):1422–1429, 2011.

Helmut Lütkepohl et al. *New introduction to multiple time series analysis*, volume 1. springer Berlin, 2005.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Sasho Nedelkoski, Jasmin Bogatinovski, Ajay Kumar Mandapati, Soeren Becker, Jorge Cardoso, and Odej Kao. Multi-source distributed system data for ai-powered analytics. In *European conference on service-oriented and cloud computing*, pp. 161–176. Springer, 2020.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

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.

Christopher J Quinn, Negar Kiyavash, and Todd P Coleman. Directed information graphs. *IEEE Transactions on information theory*, 61(12):6887–6909, 2015.

Abdellah Rahmani and Pascal Frossard. Flow based approach for dynamic temporal causal models with non-gaussian or heteroscedastic noises. *Advances in Neural Information Processing Systems*, 38:8300–8349, 2026.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

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.

Julian Sosnik, Likun Zheng, Christopher V Rackauckas, Michelle Digman, Enrico Gratton, Qing Nie, and Thomas F Schilling. Noise modulation in retinoic acid signaling sharpens segmental boundaries of gene expression in the embryonic zebrafish hindbrain. *Elife*, 5:e14034, 2016.

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, 2021.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Wanqi Zhou, Shuanghao Bai, Shujian Yu, Qibin Zhao, and Badong Chen. Jacobian regularizer-based neural granger causality. *arXiv preprint arXiv:2405.08779*, 2024.

A NOTATION TABLE

Table 6 summarizes the primary mathematical notation used throughout the main text and appendices.

Table 6: Summary of notation

Symbol	Description
System and Data	
N	Number of variables in the multivariate system
T	Total number of observed time steps
X	The full multivariate time series matrix $\in \mathbb{R}^{T\times N}$
x_t	The state of all variables at time t, $x_t \in \mathbb{R}^N$
x_t^i	The scalar value of variable i at time t
L	Look-back window length (historical lag)
$x_{<t}$	Complete history of the system through time $t-1$
$x_{<t}^{-j}$	Complete history through time $t-1$, excluding variable j
$x_{t-L:t-1}$, $x_{t-L:t-1}^i$	The L most recent observations of the system and of variable i, respectively
Causal Graph & Information Theory	
G	The underlying directed Granger-causal graph
A	Ground-truth binary adjacency matrix $\in \{0,1\}^{N\times N}$; $A_{ij}=1$ denotes $j \to i$
$\hat{A}$	Learned continuous adjacency matrix (attention gates and reported edge scores) $\in [0,1]^{N\times N}$
Θ	Global learnable adjacency parameter matrix $\in \mathbb{R}^{N\times N}$
$I(X^j \to X^i \| X^{-\{i,j\}})$	Causally Conditioned Directed Information from X^j to X^i
$P(i)$	Predicted set of causal parents for target variable i
S	Candidate set of source variables for target i, excluding i; the target's own history is always included separately
$p_{\text{full},t}$	True conditional density of x_t^i given all variables' histories: $p_{\text{full},t}(u) = p(x_t^i = u \mid x_{<t})$
$p_{-j,t}$	True conditional density of x_t^i given every observed history except source j's: $p_{-j,t}(u) = p(x_t^i = u \mid x_{<t}^{-j})$. This still conditions on the target's own history ($j \neq i$)
$\mathcal{L}_{\text{own}}$	Cumulative log-loss of the ideal predictor using only target i's own past: $-\sum_{t=1}^{T} \log p(x_t^i \mid x_{<t}^i)$. Each density is the true conditional density evaluated at the observed target value
$\mathcal{L}_S$	Cumulative log-loss of the ideal predictor using target i's own past and the histories of sources in S: $-\sum_{t=1}^{T} \log p(x_t^i \mid x_{<t}^i, x_{<t}^S)$. Its expectation is the minimum expected cumulative log-loss with access to these histories
Architecture and Optimization	
d	Latent embedding dimension for variable tokens
d_k	Query/key dimension per attention head
$\mathbf{e}_i$	Projected history token for variable i, before adding its identity embedding
$\mathbf{W}_{\text{emb}}, \mathbf{b}_{\text{emb}}$	Shared history-projection weights and bias; $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{d\times L}$
E_{id}	Variable identity embedding matrix $\in \mathbb{R}^{N\times d}$
M	Number of Transformer encoder layers
$Z^{(m)}$	Latent token representation after the m-th encoder layer
Q, K, V	Query, Key, and Value matrices used in the masked attention mechanism
$S_{ij}, \text{Score}_{ij}$	Attention logit from source token j to target token i, before and after adding the adjacency mask
$\mathbf{I}, \sigma(\cdot)$	Identity matrix and elementwise sigmoid function
γ	Diagonal forcing constant for self-loops
δ	Positive numerical floor, used in the logarithmic attention mask and the predicted variance
λ	Sparsity penalty hyperparameter for the L_1 regularization

Symbol	Description
$\hat{\mu}_t^i$	Predicted conditional mean for variable i at time t
$(\hat{\sigma}_t^i)^2$	Predicted conditional variance for variable i at time t
$\mathbf{w}_\mu, \mathbf{b}_\mu;\ \mathbf{w}_\sigma, \mathbf{b}_\sigma$	Shared linear-head parameters for the conditional mean and the pre-Softplus variance output
$\mathbf{w}_{\text{out}}, \mathbf{b}_{\text{out}}$	Shared linear-head parameters for the mean-only prediction
$\mathcal{B}$	Set of target-time indices in the minibatch
$\mathcal{L}_{NLL}, \mathcal{L}_{MSE}$	Negative Log-Likelihood and Mean Squared Error objective functions
$\mathcal{L}_{\text{pred}}, \mathcal{L}$	Selected prediction loss (NLL or MSE), and the total objective including the adjacency sparsity penalty

B Detailed Review of Related Works

Causal Discovery Approaches. Causal discovery in time series is formalized through the following families of approaches:

- **Constraint-based approaches** depend upon conditional independence tests to construct a skeletal graph, which is subsequently oriented using temporal priority and faithfulness rules. Methods such as PCMCI (Runge et al. (2019)) reduce the conditioning sets through preliminary selection, but computational cost and statistical power remain sensitive to the number of variables, maximum lag, graph density, and conditioning dimension. Pairwise candidate links grow quadratically in the variable count; exhaustive searches over conditioning subsets can grow exponentially.
- **Score-based approaches** like DYNOTEARS (Pamfil et al. (2020)) formulate causal discovery as a continuous optimization problem by minimizing a regularized loss function that balances data likelihood against network complexity. The number of coefficients grows quadratically with the variable count and linearly with the maximum lag, while enforcing instantaneous acyclicity adds optimization overhead. These costs can become substantial in high dimensions; the linear parameterization also limits the nonlinear dependencies that DYNOTEARS can represent.
- **Noise-based approaches** achieve structural identifiability by exploiting statistical asymmetries in residual noise distributions under strict semi-parametric conditions, such as non-Gaussian linear or nonlinear additive noise models.
- **Granger causality** infers a causal link $X^p \to X^q$ if the inclusion of X^p's history strictly reduces the predictive error of an autoregressive model for X^q.

Linear Granger Models. These causal discovery models assumed linear dynamics and utilized Vector Autoregressive (VAR) models of order P (Lozano et al. (2009), Lütkepohl et al. (2005)). In this regime, the system dynamics for N variables are governed by:

$$x_t = \sum_{k=1}^{P} \mathbf{A}^{(k)} x_{t-k} + \epsilon_t \tag{14}$$

where $x_t \in \mathbb{R}^N$ is the observation vector at time t, $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$ are the lag-specific coefficient matrices, and ϵ_t is the noise term. The causal structure is determined by aggregating these coefficients into an adjacency matrix $\mathbf{A}$, where a variable X^j is identified as a Granger-cause of X^i if there exists at least one $k \in \{1, \ldots, P\}$ such that $A_{ij}^{(k)} \neq 0$. To handle high-dimensional data, these methods often employ group-sparse regularization to recover the underlying causal graph.

Non-linear Component-wise Models. These include models such as eSRU (Khanna & Tan (2020)), TCDF (Nauta et al. (2019)) and NGC (Tank et al. (2021)). They extract causal information by training a separate model for each variable and then inferring the causal structure from the learned model through various techniques. For instance Neural Granger Causality (cMLP) analyzes weights of the first hidden layer to interpret inter-variable influence. TCDF, on the other hand, utilizes Temporal Convolutional Networks and applies a post-hoc attention interpretation method to its

learned convolutional kernels to identify causal relationships. As mentioned before, these incur heavy computational overhead for high-dimensional data.

Non-linear Shared Weight Models. Models like CUTS+ (Cheng et al. (2024a)) utilize message-passing Graph Neural Networks (GNNs) for causal discovery in irregularly sampled time-series. Learning a large graph while performing graph-conditioned prediction creates computational challenges. CUTS+ addresses these scalability bottlenecks through shared message-passing weights and a Coarse-to-Fine discovery strategy that progressively refines the graph. Other shared-weight methods, such as JRNGC (Zhou et al. (2024)) and UnCLe (Bi et al. (2026)), utilize standard sequence forecasters rather than GNNs. However, they rely heavily on post-hoc analysis - such as evaluating the expected input-output Jacobian matrix over the full dataset (JRNGC) or measuring prediction error spikes after chronological data permutation (UnCLe) - to extract the causal graph after forecaster's convergence. This decoupling of representation learning from graph extraction causes the underlying models to optimize strictly for unconstrained prediction rather than structural sparsity, increasing the risk of overfitting to spurious correlations.

Transformer-based Models. CausalFormer (Kong et al. (2024)) and CSAM (Liu et al. (2025b)) perform post-hoc analysis in the form of Regression Relevance Propagation and statistical verification modules respectively. Hence, they inherit the risk of overfitting to spurious correlations as well. In CSAM the attention scores stay the primary object, and edges are read off them after training by a separate global causal algorithm and a verification module, whereas in Mask2Cause the adjacency matrix is itself a parameter that gates attention and receives the sparsity penalty, so the tensor we optimize is the one we report. Standard Transformer architectures face another issue - they tokenize the time-series data across time steps which mixes information across variables. This makes it challenging to isolate causal influences among variables. iTransformer (Liu et al. (2024)) provides a way to circumvent this by introducing inverted embedding, i.e. inverting the embedding axis to treat variables rather than time steps as tokens. This helps in preserving identifiability required for causal inference while also retaining the parameter efficiency of a Transformer architecture.

Recent Neural Granger Methods (2025–26). AERCA (Han et al. (2025)) treats anomalies as interventions on exogenous variables, jointly learning Granger structure and the normal-condition distribution of those exogenous variables, and reads its graph out of per-window coefficient tensors by taking an absolute value, a median over time, a maximum over lags and a quantile threshold after training. UnCLe (Bi et al. (2026)) pairs an Uncoupler and a Recoupler network and estimates dynamic causal influence from datapoint-wise prediction errors induced by temporal perturbations, which makes it one of the few methods in this comparison that targets a time-resolved rather than a single summary graph. KANGCI (Liu et al. (2025a)) replaces the MLP predictors of the component-wise family with Kolmogorov-Arnold Networks, extracting base weights from the KAN layers under sparsity and ridge penalties, and adds a time-reversed Granger criterion to suppress spurious links. LLM-GC (Liu et al. (2026)) is a multimodal approach that supplements the temporal signal with semantic priors retrieved from a large language model, which places it outside the purely observational setting the remaining methods share. All four estimate edges through quantities that depend on the data only through the conditional mean, which is the property our heteroscedastic experiments isolate.

Neural Discovery with Heteroscedastic Noise. TS-LSNM (Kikuchi & Shimizu, 2023) learns neural location and scale functions for each target and extracts lagged and contemporaneous relationships from their weights, with optional constraints on variable groups. M2C instead shares its attention encoder across targets, avoiding per-target network replication, and uses one sparsity-regularized adjacency to control both mean and variance prediction.

RHINO (Gong et al., 2022) combines nonlinear temporal mechanisms, conditional normalizing flows for history-dependent noise, and variational inference over lagged and contemporaneous graphs. M2C directly learns the stationary summary graph through an adjacency-constrained attention encoder. Its Gaussian mean and variance heads provide a straightforward likelihood objective without a conditional flow or variational graph posterior, and the same adjacency gates prediction and supplies the reported edge scores.

FANTOM (Rahmani & Frossard, 2026) uses conditional normalizing flows and variational Bayesian expectation-maximization to learn temporal causal models with non-Gaussian or heteroscedastic noise, including regime-specific mechanisms and graphs. For stationary summary-graph discovery, M2C uses a single shared encoder and adjacency without regime segmentation or alternating variational

inference. This design accommodates squared-error and Gaussian-likelihood training within the same architecture, enabling controlled comparisons of mean- and variance-driven discovery.

Table 7 gives a brief comparative analysis of the methodologies used in closely related works against **Mask2Cause**.

Table 7: Comparison of prediction models and graph-learning mechanisms in related temporal causal discovery methods.

Method	Parameter sharing	Graph learning and extraction	Prediction model
NGC (Tank et al., 2021)	Target-specific networks	Sparse input-weight groups; graph derived from their norms	MLP or LSTM mean predictor
CUTS+ (Cheng et al., 2024a)	Shared message-passing weights	Learned graph with coarse-to-fine discovery	Graph-based prediction for irregular series
JRNGC (Zhou et al., 2024)	Shared predictor	Jacobian regularization during training; Jacobian-based edge scores	Joint mean prediction
UnCLe (Bi et al., 2026)	Shared temporal networks	Sparse dependency matrices; aggregation for static graphs or temporal perturbation for dynamic graphs	Reconstruction and mean prediction
CausalFormer (Kong et al., 2024)	Joint Transformer	Regression relevance propagation through the trained model	Variable-history embeddings, causal convolutions and attention
CSAM (Liu et al., 2025b)	Joint attention model	Sparse attention, global graph extraction and causal verification	Inverted variable-history tokens
TS-LSNM (Kikuchi & Shimizu, 2023)	Target-specific location/scale networks	Weight-derived mean and variance adjacencies, combined across mechanisms	Neural location–scale model
RHINO (Gong et al., 2022)	Shared neural mechanisms	Variational inference over instantaneous and lag-specific edges	Nonlinear mechanisms with conditional-flow noise
FANTOM (Rahmani & Frossard, 2026)	Neural mechanisms within regimes	Variational graph inference with Bayesian EM for regime assignments	Regime-specific mechanisms with conditional-flow noise
M2C (ours)	**Shared attention encoder and heads**	**One learned adjacency gates attention, receives sparsity regularization, and supplies edge scores**	**Mean-only or Gaussian mean/variance prediction**

C Theoretical Assumptions for Causal Identifiability

Our framework relies on five standard assumptions for our model to discover discover the Ground Truth and for it to be unique.

1. Strict Positivity (Non-Determinism): We assume the true joint probability density of the system is strictly positive over the entire state space $\mathcal{X}$:

$$P(\mathbf{x}) > 0 \quad \forall \mathbf{x} \in \mathcal{X} \tag{15}$$

Necessity: This condition guarantees that no variable is a strictly deterministic, noiseless function of another. In purely deterministic systems, information becomes redundant, creating unresolvable causal symmetries. For example, suppose variable Z is a deterministic copy of X ($Z_t = X_{t-1}$), and Y depends on Z with some noise ($Y_t = Z_{t-1} + \epsilon_t$). Because Z_{t-1} is mathematically indistinguishable from X_{t-2}, substituting X for Z yields the exact same predictive loss. Consequently, an optimizer cannot distinguish between the causal chain $X \to Z \to Y$ and an alternative graph containing

the direct edge $X \to Y$. This symmetry renders the minimal generative causal graph non-unique. Positivity ensures that every variable possesses at least a marginal amount of independent noise, breaking these symmetries and ensuring that the minimal generative causal graph $\mathcal{G}$ is unique.

2. Causal Sufficiency: We assume there are no unobserved (hidden) confounding variables that simultaneously influence two or more observed variables in our system $\mathbf{X}$.

Necessity: If a hidden confounder exists, the network's optimizer will observe a spurious statistical correlation between the variables and hallucinate a direct causal edge to minimize the forecasting loss. Sufficiency ensures all inferred edges represent true causal mechanisms within the observed system.

3. Strict Temporal Precedence: We assume that causal influences strictly take time to propagate, precluding instantaneous (intra-step) causal effects. Each variable is generated from the system's past and its own innovation, with contemporaneous innovations mutually independent conditional on the complete past.

Necessity: Under this assumption, the state of variable i at time t is strictly determined by historical states and independent noise, rather than the concurrent states of other variables. This justifies our model's autoregressive design, where predictions for $\mathbf{x}_t^i$ are conditioned exclusively on the strictly past window $\mathbf{x}_{<t}$. The absence of instantaneous effects and the conditional independence of contemporaneous innovations imply the strictly causal factorization

$$p(x_t \mid x_{<t}) = \prod_{i=1}^{N} p(x_t^i \mid x_{<t}).$$

Under positivity and this factorization, the directed information graph equals the unique minimal generative model graph (Quinn et al., 2015, Theorem 1).

4. Causal Faithfulness: We assume the observed distribution is faithful to the time-unrolled causal graph: every conditional independence in the distribution is implied by the graph. *Necessity:* This excludes parameter cancellations that conceal structural dependencies, supporting the interpretation of predictive relationships in terms of the underlying causal mechanisms.

5. Stationarity and Finite Markov Order: We assume the process is strictly stationary, with time-invariant graph topology, and that the conditional transition probabilities satisfy a finite-order Markov property bounded by our look-back window L:

$$P(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}) = P(\mathbf{x}_t \mid \mathbf{x}_{t-L:t-1}) \tag{16}$$

Necessity: This guarantees that the complete causal footprint required to predict the next state is fully contained within the finite context window of the Transformer. Our model also assumes a stationary causal graph while trying to infer it.

D Behaviour Under Assumption Violations

Section 2.1 states the assumptions under which Mask2Cause is formulated. Causal sufficiency and a stationary graph are the two that real systems most often break, so this appendix characterises what happens when they do. Both suites are built as offshoots of our existing benchmarks so that the numbers stay comparable to the main results.

D.1 Hidden Confounding

Causal sufficiency is a classical assumption of the entire Granger family, so the question is how resilient M2C is when its assumption is violated. We add 10 observed variables plus one hidden AR(1) hub driving 4 of them. The model never sees the hub and is scored against the true 10×10 sub-graph. We measure whether it hallucinates edges specifically among the confounded children. The hub-to-children coupling strength was calibrated per system so the resulting hub-child correlation matched the same order of magnitude as the system's own true parent-child correlations.

The outcome is asymmetric across the datasets. On VAR, confounded children's false pairwise edges score 10–22% above other non-edges, which is the classical hidden-confounder signature, and AUROC degrades from 1.00 to roughly 0.80. On Lorenz-96 the effect is essentially absent (around 1%), plausibly because the nonlinear dynamics leave less of the hub's signature recoverable from pairwise structure.

Table 8: **Hidden confounding.** Inflation ratio compares the scores of false edges among confounded children against other non-edges.

Family	Objective	AUROC	Inflation ratio (children vs other non-edges)
VAR	MSE	0.795	1.22
VAR	NLL	0.822	1.10
Lorenz-96	MSE	0.981	1.01
Lorenz-96	NLL	0.992	1.01

D.2 Regime Switching

We take a fixed base graph and add one edge that is active only in the second half of the series. The learned $\hat{\mathbf{A}}$ is then scored against regime 1 and against regime 2 separately.

Table 9: **Regime switching.** An intermittently active edge is compared against edges that are true throughout.

Family	Objective	New edge score	Mean "always-on" true edge	Regime-1 / Regime-2 AUROC
VAR	MSE	0.277	0.291	0.986 / 0.985
VAR	NLL	0.708	0.755	0.973 / 0.967
Lorenz-96	MSE	0.210	0.256	0.943 / 0.930
Lorenz-96	NLL	0.709	0.748	0.993 / 0.980

The model outputs approximately the union of the two regimes. In all four configurations the intermittent edge reaches 82–95% of the mean score of permanently true edges, and AUROC remains high against both regimes. A practitioner therefore gets a superset of the true structure, which is the safe failure mode for screening. Recovering the switch points themselves is a different estimation problem. A possible mitigation is a sliding-window variant that re-estimates $\hat{\mathbf{A}}$ over overlapping segments, with change-point detection on $\|\hat{\mathbf{A}}_t - \hat{\mathbf{A}}_{t-1}\|$. We leave dynamic causal discovery to future work.

D.3 On the Scope of the Stationary-Graph Setting

We do not claim to solve all variants of causal discovery, and we state these assumptions in Section 2.1 and their limitations in Appendix V.1. But stationary-graph discovery is not a toy restriction. Granger causality is classically defined for stationary processes, and almost all of the models we compare against report a single time-invariant causal graph, with UnCLe (Bi et al., 2026) being the notable exception. The setting also has ample real-world application. It is the operative regime for gene regulatory networks, where our DREAM3 ground truth is a real *E. coli* and yeast network with a fixed topology, and equally for industrial process monitoring, sensor and power networks, and climate teleconnections.

E Separating Mean and Variance Parents

Mixed Physics provides separate ground-truth labels for mean- and variance-driven dependencies. We investigate whether an extension of M2C-NLL can recover these parent types by learning two global adjacency matrices, $\widehat{A}^{\mu}$ and $\widehat{A}^{v}$, for conditional mean and variance prediction, respectively.

Two-graph architecture. The two prediction paths share variable-local embeddings and Transformer encoder weights, but maintain separate intermediate representations. Each path applies its own adjacency matrix throughout all encoder layers and attention heads, before its respective prediction head. Thus, cross-variable representations are computed separately for mean and variance prediction. This introduces 100 additional parameters for the ten-variable benchmark (26,122 versus 26,022 parameters with window length 3). We train both paths jointly using Gaussian NLL and a sparsity penalty equal to λ times the average of the two off-diagonal graph penalties. No parent-type labels or disjointness constraints are used during training.

Table 10: Parent recovery with two-graph M2C-NLL at the fixed final training epoch. Type accuracy is conditional on the true edge set. Mean-parent AUROC is undefined for 100:0 because no off-diagonal mean edges exist. (single seed run)

Setting	Mean AUROC	Variance AUROC	Union AUROC	Type accuracy
50:50	0.9690	0.9962	0.9806	25/27 (92.6%)
75:25	1.0000	0.9943	0.9971	27/27 (100.0%)
100:0	—	1.0000	1.0000	27/27 (100.0%)

Table 11: Binary recovery with separate density-matched thresholds. SHD is the number of false positives plus false negatives. Mean F1 is undefined for 100:0, where the true mean graph is empty.

Setting	Mean SHD	Mean F1	Variance SHD	Variance F1
50:50	6	0.7692	4	0.8571
75:25	0	1.0000	2	0.9500
100:0	0	—	0	1.0000

Experimental setup. The regenerated datasets store W_μ and W_σ, whose nonzero entries identify mean and variance parents in target-by-source orientation. The 50:50, 75:25, and 100:0 settings contain respectively 13/14, 7/20, and 0/27 off-diagonal mean/variance edges; the setting names refer to the variance-to-mean allocation ratio, subject to rounding. All datasets also contain ten mean self-loops. We run the model on hyper-parameters selected for the standard single graph M2C-NLL on Mixed Physics without additional tuning.

Parent recovery. We evaluate each learned adjacency matrix separately against its corresponding binary ground truth over all off-diagonal entries. Mean-parent AUROC therefore treats variance-only edges as negatives, and vice versa. These AUROCs use continuous scores without thresholding. For a separate parent-type diagnostic, we label each known true edge according to whichever adjacency matrix assigns it the larger score. This diagnostic is conditional on an edge existing and does not require that the edge first be discovered.

The model correctly labels all 13 mean edges and 12 of 14 variance edges in 50:50, and all seven mean edges and 20 variance edges in 75:25. In the variance-only 100:0 setting, all 27 true edges receive larger variance than mean scores, and variance-edge AUROC is 1.0000.

Separately thresholded graphs. For visualization and binary per-graph metrics, we threshold the matrices independently to match their respective ground-truth densities. An edge may appear in both predicted graphs.

Overall graph recovery. Using the elementwise maximum of the two matrices as a union score, the extension achieves AUROCs of 0.9806, 0.9971, and 1.0000 across the three settings. Matched single-graph controls achieve 0.9877, 0.9976, and 1.0000, respectively. The extension therefore provides additional parent-type information.

F THRESHOLDING

The primary output of the Mask2Cause architecture is a continuous adjacency probability matrix $\hat{\mathbf{A}} \in [0,1]^{N\times N}$. While continuous evaluation metrics such as AUROC and AUPRC comprehensively evaluate the model's ranking capability across all possible thresholds, downstream applications and discrete metrics (e.g., Structural Hamming Distance, F1) require a binarized causal graph.

Density-Matched Thresholding for Benchmarking. For standardized benchmark evaluations where the ground truth is accessible, we follow common practice by selecting a threshold that matches the edge density of the predicted graph to the known density of the ground-truth graph (expected edge density). For example, if the true VAR system possesses a known edge density of 30%, we threshold $\hat{\mathbf{A}}$ to retain exactly the top 30% of predicted edges.

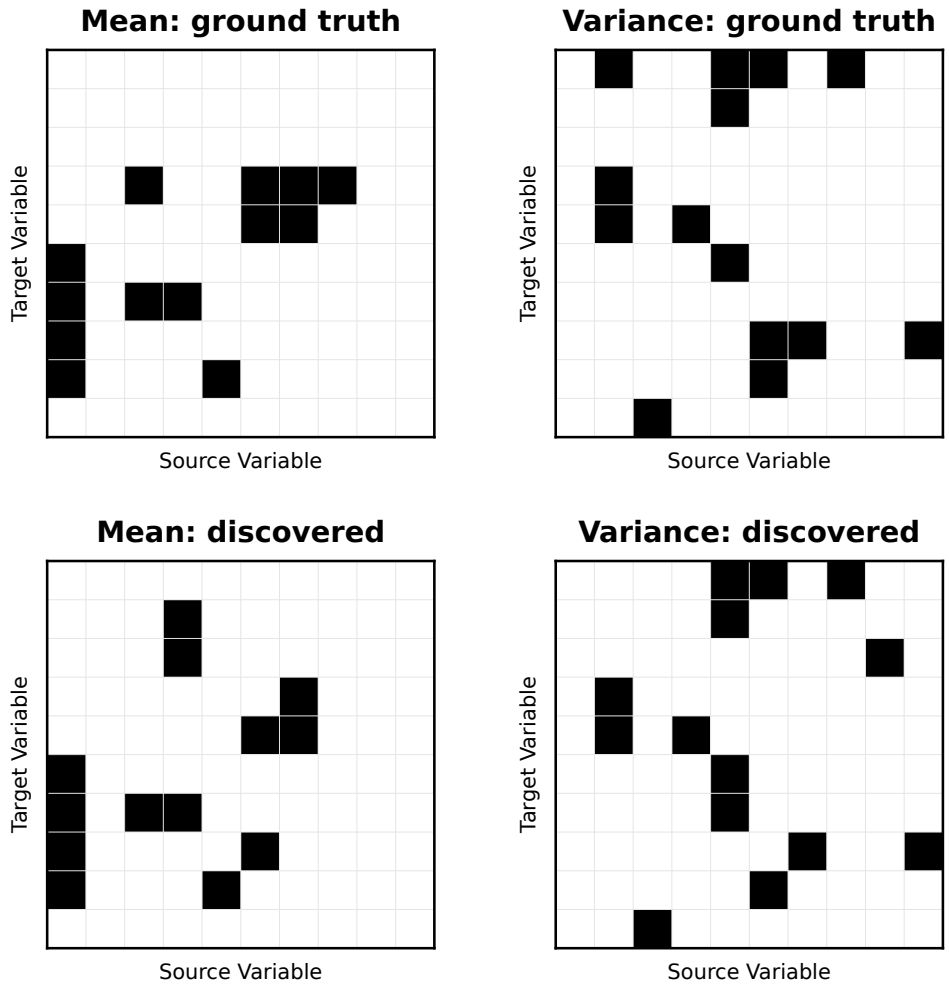


Figure 3: Mean and variance parent matrices for Mixed Physics 50:50. Rows denote targets and columns denote sources. Black cells indicate edges. Self-loops are excluded. The discovered matrices retain the top 13 mean scores and top 14 variance scores independently.

Unsupervised Thresholding for Real-World Application. In real-world settings where the expected density is unknown, a threshold must be inferred directly from the model's output. Mask2Cause is particularly well-suited for this scenario because it does not produce ambiguous, uniformly distributed scores. Instead, the combination of our log-gated attention mechanism and the L_1 sparsity penalty actively suppresses "lukewarm" connections.

True causal edges successfully resist the regularization penalty to maintain high activation weights, while non-causal edges are heavily penalized. Consequently, the optimized continuous matrix $\hat{\mathbf{A}}$ exhibits a stark, highly separable bimodal distribution. For example, the average weights for true causal edges versus non-causal edges naturally diverge to 0.514 vs. 0.164 on the Lorenz-96 system, and 0.491 vs. 0.208 on the VAR system.

Due to this strong bimodality, practitioners can reliably determine the binarization threshold via simple unsupervised 1D clustering. The clustering algorithm automatically identifies the optimal separation valley between the two score distributions, allowing for accurate discrete graph extraction without expected density.

How much is lost without ground-truth density. Table 12 quantifies the cost of discarding the density oracle. Every dataset is binarized by the unsupervised rule above, applied uniformly with no per-dataset tuning and no access to ground-truth density, alongside the density-matched thresholds

used elsewhere in the paper. AUROC is threshold-free and therefore identical under both rules, so only the discretization differs and we report SHD. The model is M2C-MSE.

Table 12: **Sensitivity to the thresholding rule.** SHD under density-matched thresholding against the unsupervised clustering rule of this appendix. AUROC is unaffected by the choice.

Dataset	AUROC	SHD (density-matched)	SHD (unsupervised)
Lorenz-96 ($F = 10, T = 500$)	0.997	2	2
Lorenz-96 ($F = 40, T = 500$)	1.000	0	3
Lorenz-96 ($F = 10, T = 250$)	0.995	6	10
Lorenz-96 ($F = 40, T = 250$)	0.999	2	4
VAR ($T = 1000$)	1.000	0	0
VAR ($T = 500$)	1.000	0	0
Medical	0.903	74	81
AQI	0.898	188	198
Traffic	0.681	54	57

Removing the density oracle never costs more than 4% of the candidate entries, and VAR is recovered exactly under both rules. This is the behaviour the bimodality argument predicts: the separation valley is wide enough that simple clustering locates it reliably.

G Robustness of Causal Masking in Deep Architectures ($M \geq 2$)

As introduced in Section 4, our optimization objective is motivated by the optimal log-loss identity in Eq. (10), which assumes the underlying predicting process is an ideal, perfectly minimizing estimator. To properly connect our model to this proposition, we must assume that setting $\hat{A}_{ij} = 0$ excludes variable j from the prediction of variable i. For a single-layer architecture ($M = 1$), this assumption effectively holds (for δ small in equation 4), as the mask explicitly blocks the only available pathway, guaranteeing that j cannot cause i in the model's prediction.

However, leveraging the full capacity of Transformer architectures requires stacking encoder layers ($M \geq 2$). In a multi-layer setting, a theoretical discrepancy arises between the ideal estimator assumption and the structural mask. Specifically, even if the direct edge is penalized and masked ($\hat{A}_{ij} = 0$), the computational graph of a deep Transformer could theoretically route the necessary information of j to i through an intermediate conduit variable k, provided that $\hat{A}_{ik} > 0$ and $\hat{A}_{kj} > 0$. This multi-hop capacity of the neural network threatens to violate the strict variable exclusion required by the theoretical proposition, potentially allowing the model to minimize the NLL for variable i while successfully evading the L_1 penalty on the direct edge $\hat{A}_{ij}$.

While this "multi-hop leakage" is mathematically possible in an unbounded optimization landscape, it does not manifest practically in Mask2Cause. Firstly, it is important to note that a potential "multi-hop leakage" is mostly a concern under specific configurations of the Ground Truth: namely, graphs containing Parallel Causal Pathways. If the ground truth contains a direct edge $j \to i$ alongside an alternative causal chain (e.g., a triangle $j \to k \to i$, or a longer path $j \to k_1 \cdots \to k_n \to i$), the model could theoretically evade the L_1 penalty on the direct edge by routing j's signal through the intermediaries. Even in such cases, the network is subject to architectural bottlenecks that make indirect routing practically unfeasible. After the usual hyperparameter tuning of the sparsity hyperparameter λ, we see that the massive NLL penalty incurred by attempting indirect routing automatically made to outweigh the L_1 savings of dropping the direct edge.

G.1 Architectural and Optimization Bottlenecks Suppressing Leakage

The architecture actively suppresses theoretical information leakage via two distinct mechanisms:

1. Representational Contention and Capacity Limits. To use variable k as a lossless conduit for j, the network must allocate specific attention heads and latent subspace dimensions within k's token strictly for j's signal. However, under the Directed Information framework, if j is a true direct causal parent of i (i.e., $I(X^j \to X^i | X^{-\{i,j\}}) > 0$), then j contains unique information in prediction of i that cannot be derived from k's history. Forcing k to embed this additional information means j and

k must compete for finite representational bandwidth, reducing the parameter capacity available to model k's own complex trajectory. Because the network jointly optimizes the forecasting loss for all variables simultaneously, the degradation in k's predictive precision causes a massive spike in the forecasting error (NLL/MSE). Consequently, the optimizer is incentivized to retain the true direct edge $\hat{A}_{ij}$, as the NLL degradation incurred by the conduit node drastically outweighs the $\lambda\|\hat{A}_{ij}\|_1$ savings.

Routing j's signal through k only becomes advantageous if the network possesses disproportionately high representational capacity and sparsity penalty (λ). As demonstrated in our hyperparameter sensitivity analysis O.4 on Lorenz-96 (a dataset that does contain Parallel Causal Pathways), the optimal space for λ and d_{model} is remarkably broad (indicating that standard hyperparameter tuning likely isolates the regime where indirect routing is heavily penalized).

2. Shared-Weight Disentanglement. Mask2Cause applies a universally shared Feed-Forward Network and final projection head across all variable tokens. This constraint in the architecture allows us to argue that node-specific routing strategies harder to learn, even when spare representational capacity is available. Let us assume that node k could partition its latent vector, dedicating one subspace to its own dynamics and an orthogonal "null space" to j's smuggled signal, thereby protecting its own prediction from interference. Such routing requires the shared transformations to preserve information useful for downstream targets while supporting the intermediate token's own prediction. Because these transformations are shared across all variables, a routing strategy must also remain compatible with their prediction tasks. Any resulting interference discourages indirect routing when its forecasting cost exceeds the sparsity savings from dropping the direct edge.

G.2 Empirical Validation

To empirically test whether the optimization landscape prefers direct causal edges over multi-hop routing, we conducted a masking ablation study. We locked the adjacency matrices and compared the baseline forecasting MSE of the true causal mask against an ablated mask.

For the Lorenz-96 system, the underlying ODE dictates that the true parents of node i are $i-2, i-1, i$, and $i+1$. We ablated the mask by explicitly removing the direct edge $i-2 \to i$, forcing the model to attempt to route the necessary information of $i-2$ exclusively through the intermediate parent $i-1$. The original VAR dataset does not have this potential weakness to test. So, we conducted a similar ablation on a custom VAR dataset parameterized such that $x_{i-2} \to x_i$ and $x_{i-1} \to x_i$.

Table 13: MSE degradation under forced multi-hop routing (ablated mask)

Dataset	**True Mask (MSE)**	**Ablated Mask (MSE)**	**Difference (Δ_{MSE})**
Lorenz-96, $F = 40, T = 500$	0.028651	0.076617	+0.047966
Lorenz-96, $F = 40, T = 250$	0.045148	0.100554	+0.055406
Lorenz-96, $F = 10, T = 500$	0.011014	0.016673	+0.005659
Lorenz-96, $F = 10, T = 250$	0.009012	0.013292	+0.004280
VAR, $T = 1000$	0.499738	0.503702	+0.003964
VAR, $T = 500$	0.521975	0.533000	+0.011025

In our baseline Lorenz-96 configuration, the effective L_1 sparsity penalty (λ) per node is 0.0002. Dropping the 10 direct edges across the system saves only 0.002 in the total L_1 penalty. As shown in Table 13, the MSE degradation (Δ_{MSE}) caused by forcing multi-hop routing entirely dwarfs the λ savings. The optimizer prefers direct edges.

Robustness to Encoder Depth. Because the architectural bottlenecks naturally regularize the flow of information, Mask2Cause remains remarkably robust even at extreme depths where multi-hop routing would theoretically be easiest. As shown in Table 14, scaling the number of encoder layers (M) up to 20 yields little to no degradation in structural recovery (in fact, we see an improvement in evaluation for Lorenz-96). We default to $M = 2$ in our main experiments solely to minimize parameter count, computational overhead, and the risk of overfitting in data-scarce regimes.

Table 14: Causal discovery performance vs. encoder depth (M) on Lorenz-96 ($F = 40, T = 500$) and VAR ($T = 500$). (single seed).

Dataset	Encoders (M)	AUROC	AUPRC	SHD
Lorenz-96	1	0.997	0.996	4
Lorenz-96	2	0.998	0.998	2
Lorenz-96	4	1.000	1.000	0
Lorenz-96	10	1.000	1.000	0
Lorenz-96	20	1.000	1.000	0
VAR	1	1.000	1.000	0
VAR	2	1.000	1.000	0
VAR	4	1.000	1.000	0
VAR	10	0.999	0.997	2
VAR	20	0.999	0.998	2

H Alternative Masking Mechanisms

This appendix expands the masking ablation of Section 5. All variants below retain a learnable adjacency matrix and end-to-end graph extraction, and differ only in how $\hat{\mathbf{A}}$ enters the attention computation. Everything else, including the encoder, the sparsity penalty and the diagonal forcing, is held fixed.

H.1 Which Alternatives Are Actually Distinct

The composition of Eq. 4 and the subsequent softmax is exactly equivalent to a multiplicative gate applied to the attention weights with renormalization:

$$\text{Softmax}(\mathbf{S} + \log(\hat{\mathbf{A}} + \delta)) \propto \exp(\mathbf{S}) \odot (\hat{\mathbf{A}} + \delta)\,. \tag{17}$$

A renormalized multiplicative mask is therefore not an ablation of our mechanism, since it is our mechanism written differently. Two alternatives remain genuinely distinct.

(A) Unnormalized multiplicative masking. We gate the attention weights without restoring the simplex constraint:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \hat{\mathbf{A}}) = \left(\hat{\mathbf{A}} \odot \text{Softmax}(\mathbf{S}^{\text{raw}})\right) \mathbf{V}\,. \tag{18}$$

(B) Bounded additive bias. We replace the logarithmic term of Eq. 4 with a linear penalty of tunable strength α, tuned per dataset with the best result reported:

$$\text{Score}_{ij} = \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_k}} + \alpha \hat{A}_{ij}\,. \tag{19}$$

Table 5 reports AUROC for both on CausalTime (M2C-MSE) and Mixed Physics (M2C-NLL).

H.2 Why the Alternatives Fall Short

(A) couples structure to signal magnitude. Once rows of the attention matrix no longer sum to one, the norm of a token's output scales with the number and strength of its retained parents. Layer normalization partially compensates for this, which dilutes the gating effect and, more damagingly, gives the optimizer an incentive to inflate $\hat{\mathbf{A}}$ purely to preserve activation scale, working directly against the L_1 penalty.

(B) makes the criterion depend on a tuned hyperparameter. The term $\alpha \hat{A}_{ij}$ acts as a bounded boost, not a penalty. A pruned edge receives no suppression, only the absence of a boost, and retains a fraction $e^{-\alpha}$ of the weight of a retained edge. The effective pruning floor is therefore a tuned quantity that also governs optimization stability.

The log gate preserves score sensitivity above the numerical floor. For $\hat{A}_{ij} \gg \delta$, the logarithm approximately cancels the sigmoid factor in the pre-softmax score derivative:

$$\frac{\partial \,\text{Score}_{ij}}{\partial \Theta_{ij}} = \frac{\sigma'(\Theta_{ij})}{\hat{A}_{ij} + \delta} = \frac{\hat{A}_{ij}(1 - \hat{A}_{ij})}{\hat{A}_{ij} + \delta} \approx 1 - \hat{A}_{ij}\,, \tag{20}$$

Thus, when $\delta \ll \hat{A}_{ij} \ll 1$, this score derivative is approximately 1. The direct multiplicative gate in (A) and additive bias in (B) instead retain the sigmoid factor $\hat{A}_{ij}(1 - \hat{A}_{ij})$ in their derivatives with respect to Θ_{ij}. This comparison concerns the local gating operations: the softmax and downstream network can still attenuate the loss gradient, so it does not guarantee recovery of suppressed edges. For fixed $\delta > 0$, the log-gated score derivative also tends to zero as $\hat{A}_{ij} \to 0$.

I Gaussian and Heavy-Tailed Likelihoods

Appendix V.1 notes that the Gaussian NLL objective can be swapped for a heavier-tailed likelihood when the data contains extreme outliers. This appendix demonstrates that substitution. The architecture and the μ and σ heads are unchanged, and only the loss differs between Gaussian NLL and Student-t NLL.

The data is a 30% dense random directed graph over 10 variables, where each edge is randomly assigned as either mean-causal or variance-causal. The setup follows Mixed Physics, except that the noise is changed from Gaussian to Student-t with 5 degrees of freedom, rescaled to unit variance so that it plays the same structural role as the original unit-variance Gaussian noise.

Table 15: **Gaussian versus Student-t likelihood** on heavy-tailed data and on a Gaussian control.

Data	Loss	AUROC	F1
Heavy-tailed 50:50	Gaussian NLL	0.641 ± 0.043	0.495
Heavy-tailed 50:50	**Student-t NLL**	$\mathbf{0.906 \pm 0.012}$	$\mathbf{0.815}$
Gaussian control 50:50	**Gaussian NLL**	$\mathbf{0.998 \pm 0.003}$	$\mathbf{0.970}$
Gaussian control 50:50	Student-t NLL	0.668 ± 0.039	0.539

The trade-off is symmetric. Under genuine outliers, Student-t NLL beats Gaussian NLL, and under correctly specified Gaussian noise the Gaussian loss retains the expected edge. This supplies the practical selection rule referred to in Appendix T: switch to the heavier-tailed likelihood when the data shows high kurtosis.

J Datasets

To evaluate the efficacy of our approach in identifying causal structures, we utilize the following standard and synthetic benchmarks with corresponding ground truth causal graphs:

- Synthetic : Vector Autoregressive (VAR) (Linear), Lorenz-96 (Non-Linear)
- Real-World : Causal Time Cheng et al. (2024b), DREAM3
- Heteroscedastic : Mixed Physics

J.1 Vector Autoregressive Dataset (Linear)

To evaluate the model's performance on linear systems with explicit multi-step memory, we simulate sparse Vector Autoregressive (VAR) processes. A VAR process of order K, denoted as VAR(K), evolves according to the linear recurrence:

$$\mathbf{x}_t = \sum_{k=1}^{K} B^{(k)} \mathbf{x}_{t-k} + \eta_t \tag{21}$$

where $\mathbf{x}_t \in \mathbb{R}^p$ is the state vector at time t, $B^{(k)} \in \mathbb{R}^{p \times p}$ is the transition matrix representing dependencies at lag k, and $\eta_t \sim \mathcal{N}(0, 0.1^2 I)$ is isotropic Gaussian noise.

The **Ground Truth Causal Graph** $A \in \{0,1\}^{p \times p}$ is defined by the existence of a dependency at *any* lag. Specifically, $A_{ij} = 1$ (variable j causes variable i) if and only if the coefficient for $j \to i$ is non-zero in at least one lag matrix:

$$A_{ij} = \mathbb{I}\left(\sum_{k=1}^{K} \left|B_{ij}^{(k)}\right| > 0\right) \tag{22}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This allows us to test if the model can aggregate causal signals distributed across multiple past time lags. Figure 4 shows the ground truth causal graph for a 10 variable system ($p = 10$). We source the dataset from Khanna & Tan (2020), accessible at `https://github.com/sakhanna/SRU_for_GCI/tree/master` (MIT License).

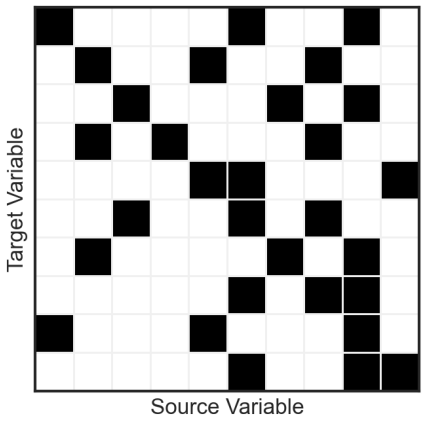


Figure 4: Ground truth for p=10 VAR system

J.2 Lorenz-96 Dataset (Nonlinear)

The Lorenz-96 model is a continuous-time dynamic system often used to simulate complex atmospheric physics Karimi & Paul (2010). It consists of p variables $x_1, \ldots, x_p$ arranged in a cyclic dependency structure. The evolution of variable x_i is governed by the system of ordinary differential equations (ODEs):

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F \tag{23}$$

where indices are taken modulo p. The forcing constant F determines the level of nonlinearity and chaos in the system. Following the benchmarks used in eSRU Khanna & Tan (2020), we have $p = 10$ and evaluate for $F = 10$ and $F = 40$ (a more turbulent regime). The high forcing constant in $F = 40$ regime amplifies the quadratic interaction terms, making nonlinear dependencies on parent variables significantly more dominant relative to linear decay.

The **Ground Truth Causal Graph** $A \in \{0,1\}^{p \times p}$ is defined as the sparse, static adjacency matrix $A \in \{0,1\}^{p \times p}$ derived from the ODE mechanism: $A_{ij} = 1$ if and only if $j \in \{i-2, i-1, i, i+1\} \pmod p$. This corresponds to a strict **Lag-1** causal dependency. Figure 5 shows the ground truth causal graph for a 10 variable system ($p = 10$). We source the dataset from Khanna & Tan (2020), accessible at `https://github.com/sakhanna/SRU_for_GCI/tree/master` (MIT License).

J.3 CausalTime Datasets

To rigorously evaluate performance on realistic, high-dimensional data, we utilize the CausalTime benchmark suite Cheng et al. (2024b). Unlike purely synthetic datasets (VAR, Lorenz) or real-world datasets with unverified ground truths, CausalTime employs a Real-to-Generated pipeline - It fits a deep generative model to real observational data and then generates new samples from the fitted model. This ensures the data retains realistic statistical properties while adhering to a known, mathematically precise causal graph.

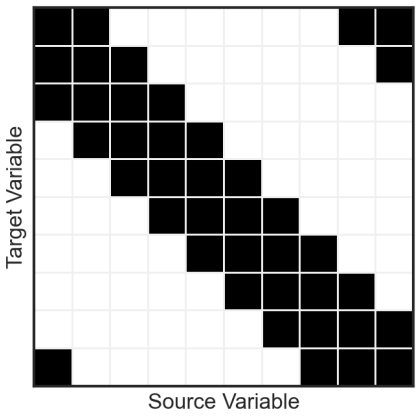


Figure 5: Ground truth causal graphs for p=10 Lorenz-96 system

The underlying generation process models the time series using a Non-Linear Autoregressive (NAR) framework implemented via Causally Disentangled Neural Networks (CDNNs) Cheng et al. (2024b). The value of variable i at time t is generated as:

$$\hat{x}_{t,i} = f_{\theta_i}(\text{Parents}(x_{t,i})) + \hat{x}^R_{t-1,i} + \eta_{t,i} \tag{24}$$

where f_{θ_i} represents the causal mechanism, $\hat{x}^R_{t-1,i}$ is a residual term capturing non-causal dynamics, and $\eta_{t,i}$ is noise modeled by a Normalizing Flow. This decomposition allows for the rigorous definition of a ground truth graph based solely on the inputs to f_{θ_i}, while the residual term ensures the trajectory complexity matches real-world recordings.

Domains and Configuration: We evaluate our method on three distinct domains provided by the benchmark:

- **Traffic ($p = 20$):** Based on traffic speed data from sensors in the San Francisco Bay Area. The causal structure is sparse and informed by the physical road network topology.
- **AQI ($p = 36$):** Derived from PM2.5 air quality measurements across 36 monitoring stations in China. The underlying graph reflects geographic proximity, where causal links exist between stations within a 40 km radius.
- **Medical ($p = 20$):** Constructed from the MIMIC-IV database, tracking vital signs and clinical events for ICU patients. Unlike the spatial datasets, this domain lacks a geometric prior, relying on the complex physiological interactions captured by the CDNN to define the causal structure.

For all three datasets, the **Ground Truth Causal Graph** is the binary adjacency matrix A used to constrain the generation process. We utilize the standard benchmark configuration consisting of 500 generated samples (each of length $T = 40$) for each domain.

Figure 6 shows the ground truth causal graph for Traffic, AQI and Medical datasets. We source the dataset from Cheng et al. (2024b), accessible at `https://www.causaltime.cc/` (MIT License).

J.4 DREAM3 Datasets

To evaluate performance on systems governed by continuous-time dynamics, we utilize the DREAM3 *InSilico* benchmark Prill et al. (2010). This suite provides a standardized platform for causal discovery in networks where interactions are defined by biologically-inspired structural equations.

The trajectories are generated using the *GeneNetWeaver* (GNW) simulator. The dynamics of the i-th variable x_i are modeled by a system of ordinary differential equations (ODEs):

$$\frac{dx_i}{dt} = m_i \cdot f_i(x_{\text{Parents}(i)}) - \lambda_i x_i \tag{25}$$

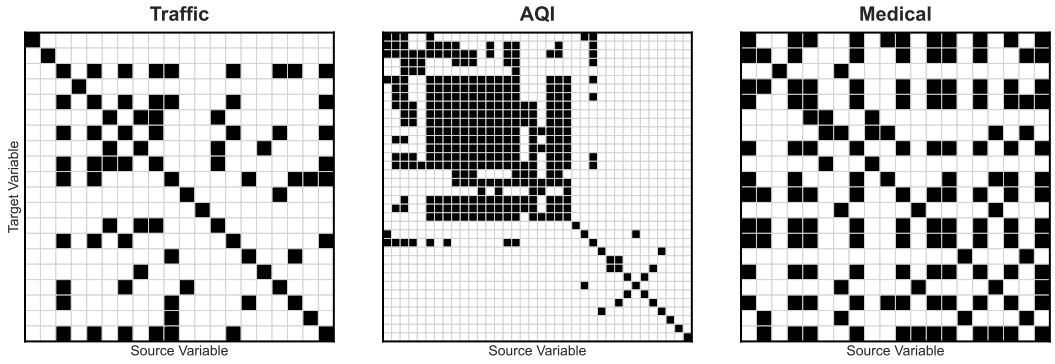


Figure 6: Ground truth causal graphs for CausalTime datasets

where $f_i(\cdot)$ is a non-linear input function (modeled via Hill-cube kinetics) representing the causal influence of parent nodes, m_i is the maximum production rate, and λ_i is the degradation rate. This formulation ensures that the ground truth graph is mathematically defined by the non-zero partial derivatives of the production term with respect to the input variables. **Domains and Configuration:** The benchmark utilized consists of five distinct networks in the 100-node ($p = 100$) category:

- **Ecoli1 & Ecoli2:** Sub-networks derived from the *Escherichia coli* transcriptional map.
- **Yeast1, Yeast2, & Yeast3:** Sub-networks derived from the *Saccharomyces cerevisiae* (Yeast) regulatory map.

For all networks, the **Ground Truth Causal Graph** is the binary adjacency matrix used to define the ODE couplings. The evaluation utilizes time-series data consisting of 46 independent trajectories, each containing $T = 21$ time points. Figure 7 illustrates the high sparsity and topological structure of these causal matrices. We source the dataset from Khanna & Tan (2020), accessible at `https://github.com/sakhanna/SRU_for_GCI/tree/master` (MIT License).

Note: 46 independent trajectories of 21 time points each is an extremely scarce dataset for learning a Causal Graph containing a 100 nodes. We use this purely as a stress test to see how our model would perform against baselines in such a data scarce regime.

J.5 Mixed Physics (Heteroscedastic)

We generate Mixed Physics using a heteroscedastic process with disjoint mean and variance parent sets:

$$x_t^i = \underbrace{\sum_{j=1}^{N} W_{ij}^{\mu} x_{t-L}^j}_{\text{Conditional mean}} + \underbrace{\left(\beta + \sum_{k=1}^{N} W_{ik}^{\sigma} (x_{t-L}^k)^2\right)^{1/2}}_{\text{Conditional standard deviation}} \eta_t^i. \tag{26}$$

The innovations η_t^i are independent standard Gaussian variables across targets and time, independent of the past. We use $L = 1$ and $\beta = 0.1$. The matrices $\mathbf{W}^{\mu}$ and $\mathbf{W}^{\sigma} \geq 0$ have disjoint supports and parameterize conditional mean and variance effects, respectively. Mean self-dependencies are retained in every configuration; variance edges are exclusively off-diagonal.

Configurations. Each dataset contains $N = 10$ variables, with 27 of the $N(N-1) = 90$ possible off-diagonal edges present, giving an inter-variable graph density of 30%. We evaluate three approximate variance-to-mean edge ratios:

- **50:50:** 14 variance edges and 13 mean edges.
- **75:25:** 20 variance edges and 7 mean edges.
- **100:0:** 27 variance edges and no off-diagonal mean edges.

These counts exclude the ten mean self-loops. Thus, the 100:0 configuration isolates variance-driven cross-variable dependencies while preserving autoregressive conditional means. The off-diagonal

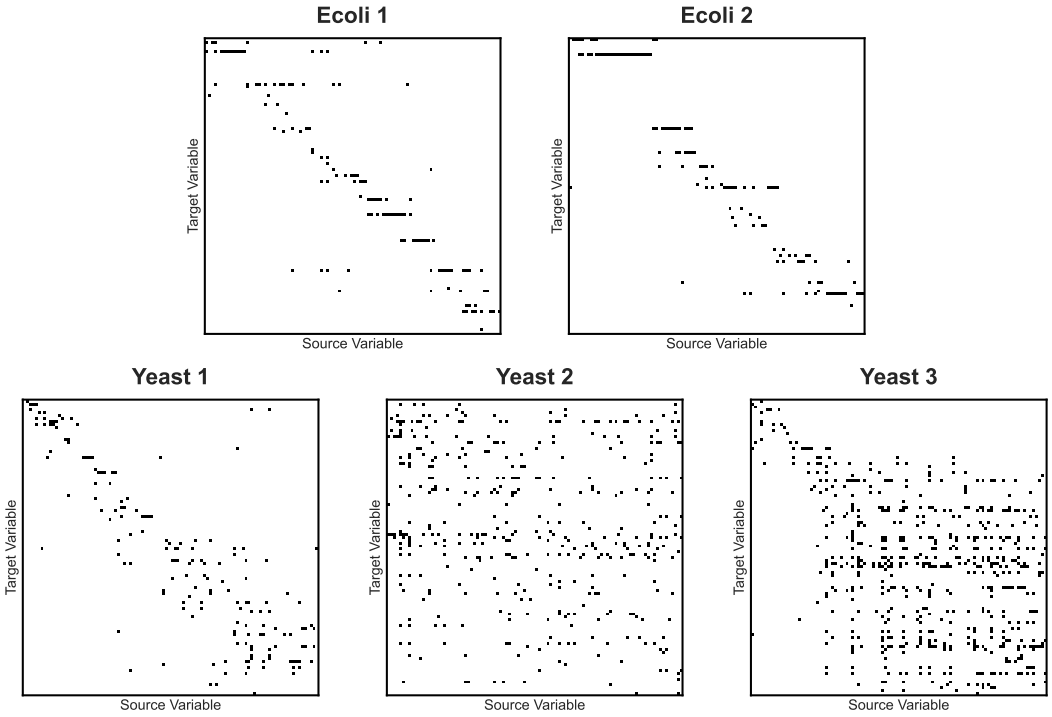


Figure 7: Ground truth causal graphs for DREAM3 networks

counts determine the idealized mean-based AUROC references reported in Section 5. Figure 8 shows the graph structures.

Coefficient sampling and scaling. For each configuration, we randomly partition its off-diagonal edges into mean and variance edges using the specified counts. Before rescaling, mean self-coefficients are sampled uniformly from $[0.3, 0.7]$. Off-diagonal mean coefficients have uniformly sampled magnitudes in $[0.5, 1.0]$ and independently sampled signs. Variance coefficients are sampled uniformly from $[0.5, 1.5]$. All other coefficients are zero.

Writing the sampled matrices as $\mathbf{B}_0$ and $\mathbf{C}_0$, we set

$$\mathbf{M}_0 = \mathbf{B}_0 \circ \mathbf{B}_0 + \mathbf{C}_0, \qquad k = \frac{0.9}{\rho(\mathbf{M}_0)}, \qquad \mathbf{W}^{\mu} = \sqrt{k}\,\mathbf{B}_0, \qquad \mathbf{W}^{\sigma} = k\,\mathbf{C}_0, \tag{27}$$

where $\circ$ denotes elementwise multiplication and ρ denotes spectral radius. This sets the overall coefficient scale while preserving edge positions and edge-type assignments. We verify second-moment stability of the resulting matrices using the full operator derived below.

Second-moment recursion. Let $\mathbf{B} = \mathbf{W}^{\mu}$, $\mathbf{C} = \mathbf{W}^{\sigma}$, and $\mathbf{S}_t = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^{\top}]$. Conditional on the past, the mean is $\mathbf{B}\mathbf{x}_{t-1}$ and the covariance is

$$\mathrm{Diag}(\beta \mathbf{1} + \mathbf{C}(\mathbf{x}_{t-1} \circ \mathbf{x}_{t-1}))\,.$$

Here diag extracts a matrix's diagonal and Diag constructs a diagonal matrix from a vector. Taking expectations gives the exact recursion

$$\begin{aligned} \mathbf{S}_t &= \mathcal{T}(\mathbf{S}_{t-1}) + \beta \mathbf{I}, \\ \mathcal{T}(\mathbf{S}) &= \mathbf{B}\mathbf{S}\mathbf{B}^{\top} + \mathrm{Diag}(\mathbf{C}\,\mathrm{diag}(\mathbf{S}))\,. \end{aligned} \tag{28}$$

The term $\mathbf{B}\mathbf{S}\mathbf{B}^{\top}$ retains cross moments between mean parents, so this expression does not assume that the variables are independent.

Stability condition and verification. The operator $\mathcal{T}$ is linear and can be represented by an $N^2 \times N^2$ matrix $\mathbf{K}$ satisfying

$$\mathrm{vec}(\mathcal{T}(\mathbf{S})) = \mathbf{K}\,\mathrm{vec}(\mathbf{S}).$$

We define $\rho(\mathcal{T}) = \rho(\mathbf{K})$. Iterating Eq. (28) yields

$$\mathbf{S}_t = \mathcal{T}^t(\mathbf{S}_0) + \beta \sum_{r=0}^{t-1} \mathcal{T}^r(\mathbf{I}). \tag{29}$$

If $\rho(\mathcal{T}) < 1$, the initialization term vanishes and the series converges. Hence the second moments approach the finite stationary solution

$$\mathbf{S}_* = \beta \sum_{r=0}^{\infty} \mathcal{T}^r(\mathbf{I}), \qquad \mathbf{S}_* = \mathcal{T}(\mathbf{S}_*) + \beta \mathbf{I}. \tag{30}$$

The coefficient rescaling above scales the full operator by k, since both terms in $\mathcal{T}$ scale by k. Evaluating this operator for the released configurations gives spectral radii 0.8864, 0.9230, and 0.9000 for 50:50, 75:25, and 100:0, respectively. All three therefore satisfy the second-moment stability condition.

Simulation. We initialize the state at zero, discard $2{,}000$ burn-in steps, and retain the following $5{,}000$ observations. The released configurations use coefficient sampling seed 0 and innovation seed 1000. No clipping or bounding of the state is applied.

Robustness to variable scaling. Simulated benchmarks can be gamed when marginal variance increases along the causal order, since variance-ranking alone then recovers much of the graph (Reisach et al., 2021). Two properties rule this out here. First, every loader in our pipeline standardises each variable before embedding, so multiplying variable k by any constant $c > 0$ scales its mean and standard deviation identically and leaves the model input unchanged. Rescaling each variable of the 100:0 data by a random constant drawn from $[0.1, 10]$ leaves off-diagonal AUROC identical at every seed. Second, the benchmark is far from the regime the concern applies to. Varsortability is 0.667, 0.667 and 0.630 across the three ratios, against the value near 0.94 reported for simulated additive-noise DAGs, and neither figure can account for an off-diagonal AUROC of 0.99. Standardisation is also the remedy prescribed in (Reisach et al., 2021), and it is applied by default rather than as a corrective.

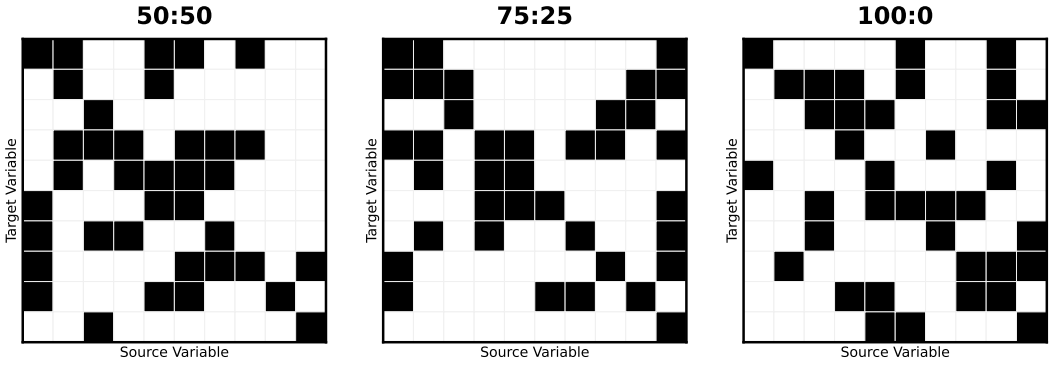


Figure 8: Ground truth causal graphs for Mixed Physics system

J.6 RHINO ERDŐS–RÉNYI SYNTHETIC BENCHMARK

Data source and generation. We use the Erdős–Rényi (ER) synthetic benchmark introduced by Gong et al. (2022). Each dataset is generated in four steps:

- Lag-specific directed graphs are sampled as ER random graphs with $2N$ edges per lag.
- Structural causal functions are parameterised as two-hidden-layer MLPs (64 units, ReLU).
- In the history-dependent noise setting, lagged-parent values are averaged, passed through a rational-quadratic spline, and used to scale an independent Gaussian innovation (noise level 0.5).
- Trajectories are simulated from random initial states, and an initial burn-in period is discarded.

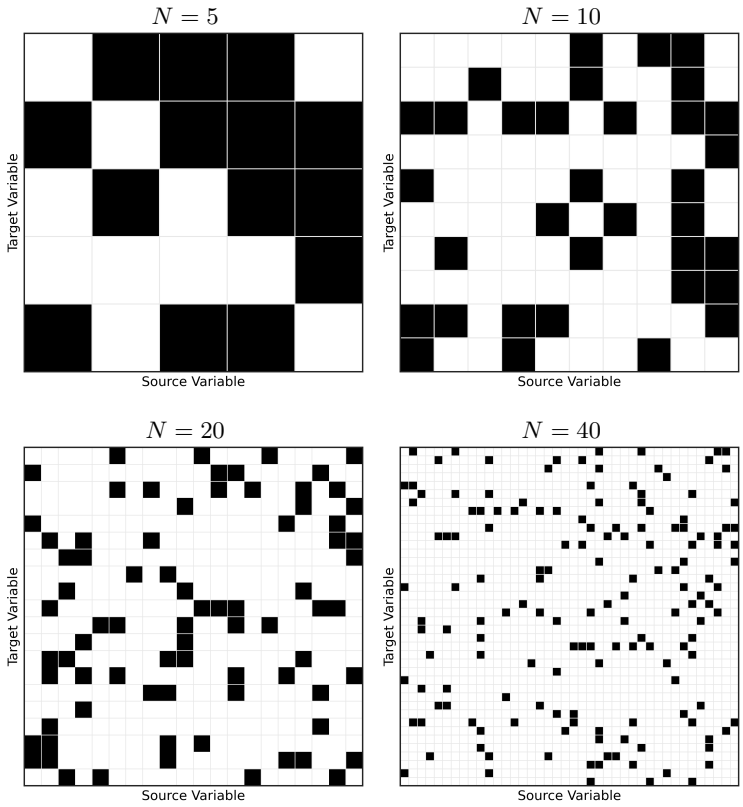


Figure 9: Ground-truth summary graphs for the Rhino ER benchmark (generator seed 1). Rows represent target variables and columns represent source variables. A black cell denotes an edge at either lag.

All data are generated using the official Causica implementation at commit `b3c79a0`.[1]

Evaluation and ground truth. From the full benchmark, we select the 20 ER datasets with no instantaneous effects and history-dependent noise, spanning $N \in \{5, 10, 20, 40\}$ variables and five generator seeds. We retain only the observational training trajectories and do not concatenate across trajectory boundaries. We set the temporal order to $L = 2$. We evaluate against a *summary graph*: an edge $j \to i$ is present whenever j causes i at lag 1, lag 2, or both:

$$A_{ij} = \mathbb{I}\Big(A_{ij}^{(1)} + A_{ij}^{(2)} > 0\Big). \tag{31}$$

Figure 9 shows representative ground-truth summary graphs for generator seed 1.

K Extended Evaluation Metrics: AUPRC, SHD, and F1

While AUROC serves as a standard measure for causal discovery, the Area Under the Precision-Recall Curve (AUPRC) provides a highly sensitive metric for evaluating false positives. A high AUPRC indicates that true edges are concentrated near the top of the ranking, making it informative for sparse graphs.

While AUROC and AUPRC are calculated by sweeping a continuous threshold across $\hat{\mathbf{A}}$, evaluating Structural Hamming Distance (SHD) and Precision/Recall requires a discrete graph. To construct this discrete graph, we performed thresholding as described in the Appendix F

As demonstrated in Table 16, Mask2Cause exhibits exceptionally strong performance across all metrics. On the synthetic benchmarks (VAR, Lorenz-96), the AUPRC nearly perfectly matches the

[1] `https://github.com/microsoft/causica/tree/b3c79a01f30f44ed36c582ffe2b4522058d82a73`

AUROC (frequently achieving a flawless 1.000), indicating near-perfect edge ranking. Furthermore, the model maintains robust precision-recall scores even on the highly complex, realistic proxies within the CausalTime suite (Medical, AQI, Traffic).

Table 16: Extended evaluation metrics for Mask2Cause (MSE & NLL) across synthetic and real-world benchmarks on a single seed

Dataset Configuration	**AUROC**	**AUPRC**	**SHD**	**F1**
Lorenz-96 ($F = 10$), $T = 500$, MSE	0.9967	0.9956	2	0.975
Lorenz-96 ($F = 40$), $T = 500$, MSE	1.0000	1.0000	0	1.000
Lorenz-96 ($F = 10$), $T = 250$, MSE	0.9950	0.9929	6	0.925
Lorenz-96 ($F = 40$), $T = 250$, MSE	0.9987	0.9981	2	0.975
VAR, $T = 1000$, MSE	1.0000	1.0000	0	1.000
VAR, $T = 500$, MSE	1.0000	1.0000	0	1.000
Medical, MSE	0.9030	0.8837	74	0.758
AQI, MSE	0.8980	0.7462	188	0.735
Traffic, MSE	0.6810	0.5355	54	0.671
Mixed Physics (50:50), NLL	0.9976	0.9943	2	0.9630
Mixed Physics (75:25), NLL	1.0000	1.0000	0	1.0000
Mixed Physics (100:0), NLL	1.0000	1.0000	0	1.0000
RHINO synthetic (5), NLL	0.9405	0.9735	2	0.9286
RHINO synthetic (10), NLL	0.9228	0.9116	12	0.8235
RHINO synthetic (20), NLL	0.8907	0.7803	50	0.6795
RHINO synthetic (40), NLL	0.9030	0.7139	102	0.6710

L Scalability on High-Dimensional Systems

In the main text experiments (Section 5), we evaluated Mask2Cause on synthetic systems (VAR and Lorenz-96) with $N = 10$ variables. This dimension was selected to mirror the established experimental protocols of our primary baselines (e.g., eSRU Khanna & Tan (2020) and CUTS Cheng et al. (2023)), ensuring a fair and direct comparison. As shown in Table 1, Mask2Cause achieves near-perfect recovery in this regime.

To rigorously evaluate the model's structural recovery capabilities at scale and explicitly address the ceiling effects observed at smaller dimensions, we extended our evaluation of the Mask2Cause(MSE) variant to $N = 20$ and $N = 100$ systems (with $T = 1000$).

As detailed in Table 17, the metric ceiling effect persists at $N = 20$, where the model still easily recovers the underlying dynamics. However, increasing N from 10 to 100 increases the number of candidate causal links quadratically from 100 to 10,000. This explicitly stresses the adjacency-constrained encoder. Despite this massive increase in structural complexity, Mask2Cause remains highly robust, maintaining an AUROC between 0.87 and 0.95 across both linear and chaotic continuous-time dynamics. Table 18 presents results of evaluation of other competitive baselines (CUTS+, cMLP, cLSTM) on high-dimensional VAR and Lorenz-96 datasets.

M Convergence Behaviour with Sample Size

We prove no consistency or identifiability result for the learned adjacency matrix. To our knowledge no neural Granger method provides finite-sample recovery guarantees under regularization and thresholding, and we do not claim one. What we can offer is empirical evidence on how recovery behaves as the sample size grows.

We sweep the trajectory length on Lorenz-96 and VAR at $N = 100$. The high dimension is deliberate, since at $N = 10$ the benchmark saturates and any trend would be masked by the ceiling.

Table 17: **Mask2Cause scaling performance.** Evaluation of the MSE variant on high-dimensional VAR and Lorenz-96 systems ($T = 1000$).

Dataset	**System Size (N)**	**AUROC**	**AUPRC**
Lorenz-96 ($F = 10$)	10	1.000 ± 0.000	1.000 ± 0.000
Lorenz-96 ($F = 40$)	10	1.000 ± 0.000	1.000 ± 0.000
VAR	10	1.000 ± 0.000	1.000 ± 0.000
Lorenz-96 ($F = 10$)	20	1.000 ± 0.000	1.000 ± 0.000
Lorenz-96 ($F = 40$)	20	1.000 ± 0.000	0.998 ± 0.001
VAR	20	0.997 ± 0.001	0.987 ± 0.004
Lorenz-96 ($F = 10$)	100	0.952 ± 0.064	0.871 ± 0.092
Lorenz-96 ($F = 40$)	100	0.883 ± 0.049	0.772 ± 0.084
VAR	100	0.867 ± 0.078	0.680 ± 0.095

Table 18: **Baseline scaling performance.** Evaluation of competitive baselines on high-dimensional VAR and Lorenz-96 systems ($T = 1000$) (mean $\pm$ std)

		LORENZ-96 (F=10)		**LORENZ-96 (F=40)**		**VAR**	
Metric	**Model**	$N = 20$	$N = 100$	$N = 20$	$N = 100$	$N = 20$	$N = 100$
AUROC	CUTS+	0.837 ± 0.013	0.781 ± 0.012	0.941 ± 0.026	0.826 ± 0.041	0.880 ± 0.015	0.601 ± 0.122
	cMLP	0.982 ± 0.007	0.850 ± 0.042	0.992 ± 0.004	0.798 ± 0.006	0.540 ± 0.025	0.506 ± 0.004
	cLSTM	0.719 ± 0.019	0.522 ± 0.021	0.698 ± 0.003	0.533 ± 0.011	0.641 ± 0.016	0.508 ± 0.006
AUPRC	CUTS+	0.625 ± 0.054	0.392 ± 0.113	0.875 ± 0.036	0.560 ± 0.142	0.760 ± 0.039	0.397 ± 0.144
	cMLP	0.939 ± 0.020	0.326 ± 0.085	0.974 ± 0.007	0.201 ± 0.003	0.277 ± 0.007	0.292 ± 0.001
	cLSTM	0.377 ± 0.076	0.033 ± 0.001	0.307 ± 0.042	0.033 ± 0.000	0.347 ± 0.016	0.295 ± 0.006

Table 19: **Recovery against trajectory length.** AUROC at $N = 100$, where the search space holds 10^4 candidate links.

Dataset	$T = 500$	$T = 1000$	$T = 2000$	$T = 5000$
Lorenz-96 ($F = 10$)	0.828 ± 0.083	0.952 ± 0.064	0.961 ± 0.045	0.978 ± 0.024
Lorenz-96 ($F = 40$)	0.810 ± 0.072	0.883 ± 0.049	0.923 ± 0.033	0.955 ± 0.042
VAR	0.791 ± 0.089	0.867 ± 0.078	0.917 ± 0.041	0.932 ± 0.040

Recovery improves monotonically with the number of samples and shows no sign of degradation or oscillation, which is consistent with convergence toward the true graph. Appendix V.1 states the limitation.

N Computational Complexity Analysis

We evaluate the efficiency of Mask2Cause using hardware-independent metrics: Parameter Count and Floating Point Operations (FLOPs). We conduct a single-sample inference sweep (Batch Size=1) to quantify scaling behavior with respect to history length L and system size N.

Scaling w.r.t Sequence Length (L). While standard temporal Transformers scale quadratically with sequence length ($O(L^2)$), our *Inverted Variable Embedding* strategy decouples the computational cost from the temporal window. Figure 10 (Left) confirms that Mask2Cause scales **linearly with history length**. Remarkably, increasing L from 5 to 2000 results in a negligible increase in computational cost (2.05M to 4.60M FLOPs). This efficiency arises because the sequence dimension is projected into a fixed latent vector d_{model} at the input layer; subsequent attention layers operate solely on the variable dimension, independent of L.

Scalability w.r.t System Size (N). Self-attention over variables implies $O(N^2)$ complexity. This is reflected in Figure 10 (Right) where we evaluate the computational cost for system sizes up to $N = 2000$. This behavior occurs because the quadratic attention term ($N^2 \cdot d_{model}$) dominates the position-wise Feed-Forward Networks (FFN) and linear projections, which scale as $O(N \cdot d_{model}^2)$.

Resource Footprint. Table 20 demonstrates the efficiency of our architecture while exposing fundamental scaling bottlenecks in baseline models. Because cLSTM and CUTS+ are recurrent architectures and employ temporal weight-sharing, their parameter counts remain constant as the sequence length L varies. In contrast, cMLP flattens the history window, causing a huge increase in parameter count as L scales from 5 to 2000. While all models inevitably confront an $\mathcal{O}(N^2)$ complexity bound regarding variable count N, their failure modes vary. Neural Granger methods (cMLP and cLSTM) become intractable at scale due to excessive parameter growth. CUTS+ remains parameter-efficient via spatial weight-sharing, but its graph-based message-passing mechanism incurs severe computational overhead, evident in its massive FLOP count at $N = 2000, L = 2000$. RHINO is compact at $N = 10, L = 5$, requiring 0.0323M parameters and 1.80M FLOPs. However, at $N = 2000, L = 5$, it grows to 70.6584M parameters and 8275.23M FLOPs, compared with 4.1005M parameters and 2511.10M FLOPs for Mask2Cause (NLL). Its history-length cost also rises sharply: at $N = 10, L = 2000$, RHINO requires 555.20M FLOPs versus 4.60M for Mask2Cause. In a separate single-GPU runtime comparison on the history-dependent RHINO ER dataset ($N = 10$, seed 1), complete runs excluding tuning took 42.04 minutes for RHINO, compared with 3.88 minutes for Mask2Cause (NLL), 3.67 minutes for Mask2Cause (MSE), 7.28 minutes for CUTS+, 12.88 minutes for cMLP, and 19.66 minutes for cLSTM. Mask2Cause circumvents these limits: it achieves parameter efficiency through shared Transformer weights and computational efficiency via global attention mechanisms. On the standard Lorenz-96 configuration ($N = 10, L = 5$), Mask2Cause requires only **0.10M parameters** and **2.05M FLOPs**, with this relative scaling advantage compounding rapidly at higher values of N and L. Figures 10 and 11 support this analysis, confirming the superior scalability of the proposed framework.

N.1 Hardware and Compute Resources

All experiments, including hyperparameter tuning and baseline evaluations, were conducted on a standard workstation equipped with an Intel Core Ultra 9 285H processor (32 GB of RAM) and two RTX A5000 GPU.

O Implementation Details and Hyperparameter Analysis

O.1 Data Sources and Evaluation Setup

We utilize standard dataset implementations sourced directly from the official repository of the eSRU Khanna & Tan (2020) framework (for VAR, Lorenz-96, and DREAM3) and the official CausalTime Cheng et al. (2024b) website.

Configuration: $N = 10$, $L = 5$

Model	Params (M)	FLOPs (M)
cMLP (NGC)	0.0333	0.13
cLSTM (NGC)	0.1952	2.00
CUTS+	0.1623	16.37
RHINO	0.0323	1.80
M2C (MSE)	0.1005	2.05
M2C (NLL)	0.1006	2.05

Configuration: $N = 2000$, $L = 5$

Model	Params (M)	FLOPs (M)
cMLP (NGC)	1280.2580	5120.51
cLSTM (NGC)	1057.9220	10589.44
CUTS+	25.2530	499939.84
RHINO	70.6584	8275.23
M2C (MSE)	4.1004	2510.85
M2C (NLL)	4.1005	2511.10

Configuration: $N = 10$, $L = 2000$

Model	Params (M)	FLOPs (M)
cMLP (NGC)	12.8013	51.20
cLSTM (NGC)	0.1952	801.28
CUTS+	0.1623	6549.68
RHINO	1.0298	555.20
M2C (MSE)	0.2282	4.60
M2C (NLL)	0.2283	4.60

Configuration: $N = 2000$, $L = 2000$

Model	Params (M)	FLOPs (M)
cMLP (NGC)	512000.2580	2048000.51
cLSTM (NGC)	1057.9220	4235776.00
CUTS+	25.2530	199975936.00
RHINO	24090.4584	2834472.03
M2C (MSE)	4.2281	3021.57
M2C (NLL)	4.2282	3021.82

Table 20: Complexity analysis: parameters and FLOPs across different configurations.

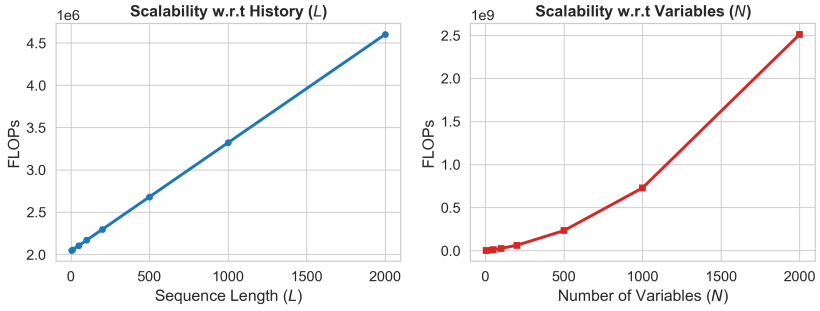


Figure 10: **Complexity scaling (MSE).** (Left) FLOPs vs. sequence length L. The model is virtually insensitive to the look-back window; increasing history from $L = 5$ to $L = 2000$ only increases cost from 2.05M to 4.60M FLOPs. (Right) FLOPs vs. variables N. The cost grows significantly with system size, spanning from 1.01M ($N = 5$) to nearly 2510.85M ($N = 2000$), driven by the variable-wise projections.

To ensure fair comparison and reproducibility, we adopt specific evaluation strategies tailored to the nature of each benchmark:

1. Synthetic Generative Systems (VAR, Lorenz-96) We strictly separate hyperparameter tuning from final evaluation to prevent leakage. For each physical configuration (e.g., forcing constant F), we use the available 6 independent dataset realizations using distinct random seeds. Seed 0 is used exclusively as a *Calibration Set* for hyperparameter tuning. Seeds 1–5 are reserved as *Test Sets*. We report the mean and standard deviation of the AUROC computed across the 5 independent Test Sets.

2. CausalTime (Static Real-World Proxies) For the fixed CausalTime datasets, where new samples cannot be generated, we employ a chronological split, using the first 20% of the data for tuning. To account for the variance inherent in neural network initialization, we train the model 5 times on the same dataset using different random seeds for weight initialization. Results are reported as the mean and standard deviation across these 5 runs.

3. DREAM3 (Gene Regulatory Networks) Consistent with the evaluation protocol of the baselines we compare against (which utilize the single fixed dataset provided by the challenge), we do not

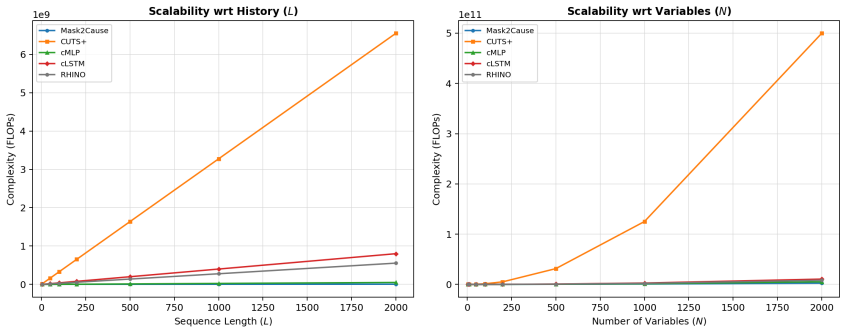


Figure 11: **Comparison of complexity scaling.** (Left) FLOPs vs. sequence length L (with $N = 10$ fixed). Mask2Cause demonstrates superior scalability with respect to sequence length; increasing history from $L = 5$ to $L = 2000$ only increases cost from 2.05M to 4.60M FLOPs. In contrast, baselines exhibit significantly steeper linear scaling: cMLP (0.13M to 51.20M), cLSTM (2.00M to 801.28M), CUTS+ (16.37M to 6549.68M), and RHINO (1.80M to 555.20M). (Right) FLOPs vs. variables N (with $L = 5$ fixed). While all models exhibit $O(N^2)$ quadratic scaling, Mask2Cause remains the most computationally efficient at large dimensions. Its cost grows from 1.01M ($N = 5$) to 2510.85M ($N = 2000$), substantially outperforming cMLP (5120.51M), cLSTM (10589.44M), CUTS+ (499939.84M), and RHINO (8275.23M).

perform multi-seed averaging for this benchmark. We use the first 20% of the data for tuning. We report the final AUROC from the single best model found after tuning on the validation set, matching the reporting style of the cited baselines.

4. Mixed Physics. Each variance-to-mean regime uses a fixed dataset. We report mean AUROC and standard deviation across five seeded model runs on that same dataset. The seeds vary model initialization and other seed-dependent training randomness; the selected hyperparameters remain fixed across runs. Thus, the reported variability reflects model-run variability, not independently generated datasets.

Baseline Configurations. For baseline methods that we ran locally (rather than quoting from published literature), we strictly utilized the hyperparameter configurations specified for each respective benchmark in their original publications. If that is not available (for the case of mixed physics), we tune the hyper-parameters ourselves following the guidelines established by the respective papers.

O.2 Architectural Configurations

Fixed Parameters. Through preliminary ablation on the Calibration Sets, we identified a globally robust architecture that performs well across diverse dynamical regimes. We fix the number of encoder layers to $M = 2$ and the number of attention heads to $H = 4$. Deeper architectures were observed to overfit the noise in sparse causal discovery tasks. The embedding dimension usually d_{model} set to 64 by default, but reduced to 32 or 16 for datasets with extremely short sequences to strictly constrain model capacity.

Diagonal Forcing. The diagonal forcing parameter controls the prior on self-loops. For autoregressive systems (VAR, Lorenz), a variable's immediate history is its strongest predictor. We force the diagonal logits to $+100$ to ensure the model always attends to itself, stabilizing the learning of cross-variable interactions. For the DREAM3 benchmark, the standard evaluation protocol explicitly excludes self-regulation edges from the ground truth. So, we remove the diagonal component from the predicted adjacency matrix before evaluating. Because of this we are no longer forced to keep the diagonals open, allowing us to include "Diag Force" for tuning, taking values {-100, 0, 100}.

O.3 Optimal Hyperparameters

We performed a grid search on the Calibration/Validation sets to determine the optimal learning rate ('lr'), sparsity penalty (λ), and training duration ('epochs'). The selected configurations for our Homoscedastic (MSE) and Heteroscedastic (NLL) models are detailed below.

Table 21: **Optimal hyperparameters for M2C-MSE**. CausalTime settings use seed 0; initial graph logits are $+1$ for Traffic and AQI and -1 for Medical.

Dataset	Learning Rate	Batch Size	Seq Len	d_{model}	λ	Diag Force	Epochs
VAR ($T = 500$)	0.001	32	3	64	0.01	+100	10
VAR ($T = 1000$)	0.001	32	3	64	0.01	+100	5
Lorenz-96 ($F = 10, T = 250$)	0.01	32	1	32	0.02	+100	200
Lorenz-96 ($F = 10, T = 500$)	0.001	32	1	64	0.02	+100	150
Lorenz-96 ($F = 40, T = 250$)	0.01	32	1	64	0.02	+100	150
Lorenz-96 ($F = 40, T = 500$)	0.001	32	1	64	0.02	+100	150
DREAM3 (E.coli-1)	0.001	32	5	32	0.5	-100	20
DREAM3 (E.coli-2)	0.001	32	5	32	0.005	+100	65
DREAM3 (Yeast-1)	0.001	16	5	64	0.0005	+100	35
DREAM3 (Yeast-2)	0.001	32	5	32	0.001	0	20
DREAM3 (Yeast-3)	0.001	16	5	64	0.5	-100	20
CausalTime (Traffic)	0.01	32	3	64	0.5	+100	17
CausalTime (AQI)	0.01	32	3	64	1.0	+100	16
CausalTime (Medical)	0.01	32	3	64	1.0	+100	35
MixedPhysics (50:50)	0.001	32	3	32	0.01	+100	35
MixedPhysics (75:25)	0.0001	64	3	16	0.01	+100	5
MixedPhysics (100:0)	0.001	16	3	16	0.001	+100	5
RHINO synthetic (5)	0.001	32	2	32	0.03	+100	35
RHINO synthetic (10)	0.001	32	2	32	0.003	+100	35
RHINO synthetic (20)	0.001	32	2	32	0.03	+100	35
RHINO synthetic (40)	0.001	32	2	32	0.03	+100	35

Table 22: **Optimal hyperparameters for M2C-NLL**. CausalTime settings use seed 0; and initial graph logits of $+1$.

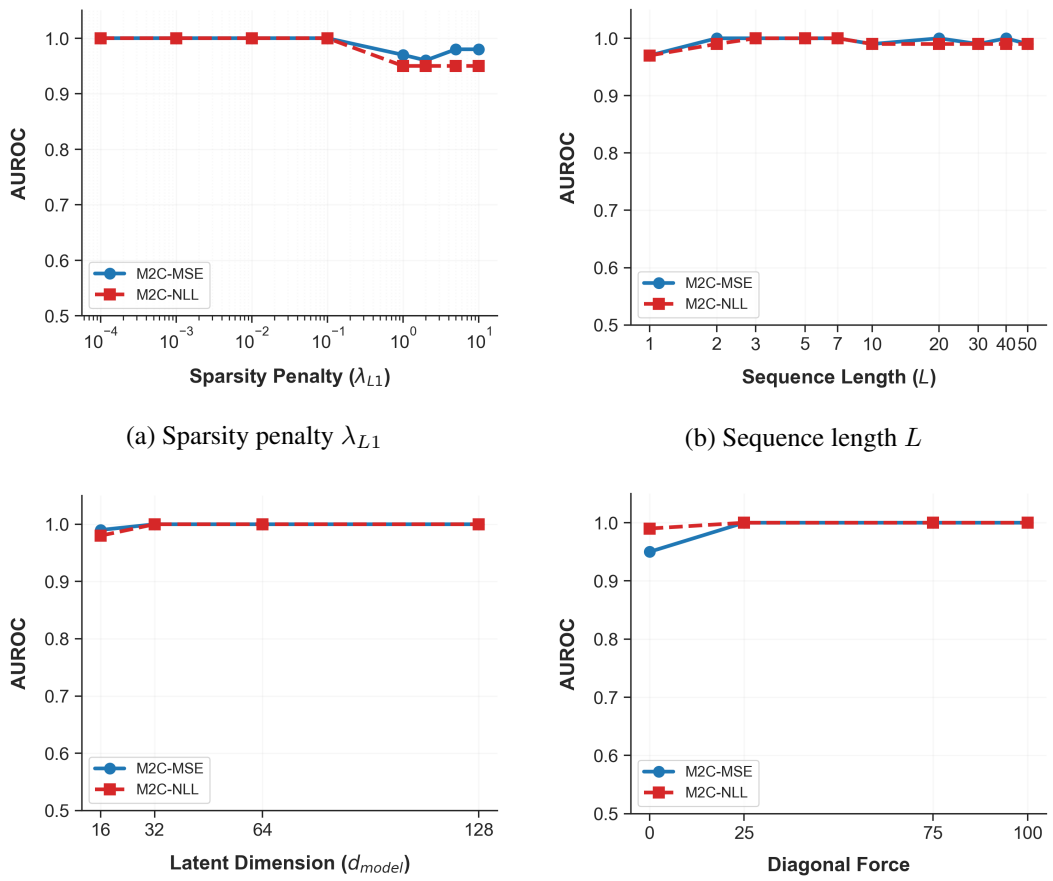
Dataset	Learning Rate	Batch Size	Seq Len	d_{model}	λ	Diag Force	Epochs
VAR ($T = 500$)	0.001	32	3	64	0.01	+100	10
VAR ($T = 1000$)	0.001	32	3	64	0.01	+100	5
Lorenz-96 ($F = 10, T = 250$)	0.01	32	1	32	0.02	+100	185
Lorenz-96 ($F = 10, T = 500$)	0.001	32	1	64	0.02	+100	140
Lorenz-96 ($F = 40, T = 250$)	0.01	32	1	64	0.02	+100	150
Lorenz-96 ($F = 40, T = 500$)	0.001	32	1	32	0.02	+100	30
DREAM3 (E.coli-1)	0.001	32	5	64	0.01	-100	20
DREAM3 (E.coli-2)	0.001	32	5	64	1.00	+100	3
DREAM3 (Yeast-1)	0.001	32	5	64	0.001	+100	35
DREAM3 (Yeast-2)	0.0001	32	5	64	0.001	+100	15
DREAM3 (Yeast-3)	0.01	16	1	16	0.1	-100	2
CausalTime (Traffic)	0.01	32	3	64	0.001	+100	11
CausalTime (AQI)	0.003	64	3	64	0.01	+100	20
CausalTime (Medical)	0.01	32	3	64	0.5	+100	15
MixedPhysics (50:50)	0.001	32	3	32	0.01	+100	75
MixedPhysics (75:25)	0.003	32	1	32	0.01	+100	35
MixedPhysics (100:0)	0.003	32	1	32	0.01	+100	20
RHINO synthetic (5)	0.001	128	5	32	0.0005	+100	50
RHINO synthetic (10)	0.001	32	2	32	0.003	+100	35
RHINO synthetic (20)	0.003	128	5	64	0.003	0	43
RHINO synthetic (40)	0.001	128	5	64	0.01	0	158

O.4 Sensitivity Analysis

To assess the robustness of Mask2Cause, we analyze the sensitivity of the AUROC metric to variations in four key hyperparameters: Sparsity Penalty (λ_{L1}), Sequence Length/Look back window (L), Latent Dimension (d_{model}), and Diagonal Forcing. We vary one parameter at a time while fixing the others to their optimal settings on the Lorenz-96 ($F = 10$) calibration set.

Observations:

- **Sparsity Penalty (λ):** As shown in Panel (a), the model is exceptionally robust to the regularization strength, maintaining perfect recovery (AUROC 1.00) for $\lambda \in [10^{-4}, 10^{-1}]$. A slight degradation (0.98) is observed only when $\lambda \geq 1$, where the penalty begins to over-prune weak but true causal signals.
- **Sequence Length (L):** Panel (b) highlights a critical property of the Inverted Variable Embedding. The true lag of the system is small ($K \approx 3$). Remarkably, even when the input window is extended to $L = 50$, the performance does not degrade ($0.99 - 1.00$). This confirms that the variable-specific tokenization effectively learns to attend to the relevant



(a) Sparsity penalty λ_{L1}

(b) Sequence length L

(c) Latent dimension d_{model}

(d) Diagonal forcing

Figure 12: **Hyperparameter sensitivity analysis.** The model exhibits high robustness across wide ranges of hyperparameters. Notably, (b) shows that performance remains near-perfect even as sequence length increases to $L = 50$, demonstrating the embedding's ability to prioritize relevant recent history.

immediate history (the first few time steps) while suppressing noise from the distant past, without requiring manual lag selection.

- **Latent Dimension (d_{model}):** As seen in Panel (c), the model is insensitive to the hidden dimension size, achieving AUROC ≥ 0.99 for $d \in [16, 128]$. This suggests that the causal graph structure is the primary information bottleneck, and even compact representations are sufficient to capture the dynamics.

- **Diagonal Forcing:** Panel (d) demonstrates that our model does not rely on diagonal forcing as a prerequisite for success. Crucially, even in the complete absence of this prior (Force = 0), the model maintains high recovery accuracy (AUROC ≥ 0.95 for MSE model, 0.99 for NLL model). This result confirms that diagonal forcing is not a "cheat code" or structural crutch; rather, the model is fully capable of autonomously learning autoregressive dependencies from scratch, with the prior serving merely to accelerate convergence in known autoregressive settings.

P FORECASTING USING CAUSAL PRUNING

P.1 FORECASTING ACCURACY AND PARAMETER REDUCTION

We use the causal map discovered by **Mask2Cause** to perform *Causal Pruning* at the input level and evaluate a forecaster's performance on a single-step-ahead objective. For any target variable $X_{i,t+1}$, the model's receptive field (input) is restricted to its predicted causal parents and its own historical value, formally defined as $P(i) = \{X_{j,t} : M_{j,i} = 1\} \cup \{X_{i,t}\}$. This decomposition of forecasting loss minimization into N univariate forecasting tasks ensures that the model isolates direct causal effects and discards spurious associations driven by conditionally independent variables. It also offers an added advantage of reducing the architectural complexity from $O(N)$ to $O(|P(i)|)$ per node. Model-specific implementation details follow below.

Empirical Results and Comparative Analysis. To validate our hypothesis that causal pruning can improve forecasting accuracy by avoiding the influence of spurious correlations and also improve resource optimization, we evaluate the parameter reduction achieved and the gain in the mean square error over the vanilla[2] versions of common forecasting models in Table 23. We observe that in the DREAM3 and most Lorenz datasets, causal pruning achieves a better accuracy while reducing the computational overhead significantly. In VAR dataset, the parameter reduction comes at the cost of a slight drop in accuracy for some models. Our findings show that causal pruning with **Mask2Cause** outperforms the **CUTS+** baseline for most datasets and models, indicating it discovers preciser matrices that provide a superior inductive bias (see Appendix P.3) . We also observe causal pruning performs competitively against alternative feature selection methods (see Appendix P.4). Appendix Q reports a second downstream task on real data, localising root causes in cloud microservice telemetry from the discovered graph.

Table 23: Model-wise comparison of parameter reduction (PR) and MSE reduction (MSE-R) in %

	DREAM3 ($N = 100$)					**Lorenz**		**VAR**	
Model / Metric	**Ecoli1**	**Ecoli2**	**Yeast1**	**Yeast2**	**Yeast3**	**F10**	**F40**	**T500**	**T1000**
ARIMAX									
PR (%)	99.0	99.0	99.0	99.0	98.9	60.69	58.50	72.20	70.50
MSE-R (%)	+40.01	+36.70	+36.35	+38.11	+41.15	+2.34	+1.20	-5.11	-5.62
MLP									
PR (%)	97.0	97.0	96.9	97.0	96.9	50.44	48.62	60.01	58.60
MSE-R (%)	+31.67	+33.81	+31.91	+33.74	+33.13	+38.48	+5.44	-4.41	-4.06
N-BEATS									
PR (%)	59.2	59.2	59.2	59.2	59.2	7.9	7.6	4.8	4.7
MSE-R (%)	+33.26	+35.25	+32.12	+34.10	+34.05	+12.93	-19.47	+10.72	+1.13
Linear Reg.									
PR (%)	98.0	98.0	97.9	98.0	97.9	54.8	53.2	65.6	64.0
MSE-R (%)	+36.97	+35.07	+34.20	+35.57	+38.17	+2.50	+2.39	-5.02	-5.28

P.2 MODEL-SPECIFIC PRUNING MECHANISMS

Causal Pruning offers a benefit of reducing architectural complexity from $O(N)$ to $O(|P(i)|)$ per node, with implementation strategies specific to each model class. For neural architectures such as MLP and N-BEATS, pruning is enforced by slicing the input tensors so that instead of feeding a full N-dimensional vector to the first hidden layer, the model only receives indices corresponding to $P(i)$. In the case of ARIMAX, the set $P(i)$ is treated as the collection of exogenous regressors, where masking the exogenous input matrix to include only causal parents allows the model to avoid estimating coefficients for non-causal variables. Similarly, Linear Regression (VAR) is formulated as

[2]The vanilla versions of forecasting models have access to history of all the variables for one-step-ahead prediction task.

a sparse Ordinary Least Squares (OLS) problem, ensuring that coefficients are estimated exclusively for variables within the causal neighborhood $P(i)$.

P.3 Comparative Analysis with CUTS+ Baseline

Table 24 compares the performance of different forecasting models for one-step-ahead prediction task by performing causal pruning using causal masks from **Mask2Cause** and **CUTS+** baseline. We observe that for DREAM3 and Lorenz96 datasets, Mask2Cause provides causal matrices that have a better inductive bias. For VAR dataset, we see a slight drop in accuracy which is compensated by lower computational complexity.

Table 24: Relative improvement of Mask2Cause over CUTS+ baseline (%) in MSE reduction across datasets

	DREAM3					Lorenz		VAR
Model	**Ecoli1**	**Ecoli2**	**Yeast1**	**Yeast2**	**Yeast3**	**F=10**	**F=40**	**T=500**
ARIMAX	+2.25	+4.16	+0.67	+2.57	+2.74	+11.40	-1.92	-0.38
MLP	+6.31	+9.70	+6.31	+8.02	+7.93	+20.66	+11.70	-0.52
N-BEATS	+8.19	+10.79	+6.36	+7.82	+10.34	+31.24	+20.59	-0.62
Linear Reg.	+2.14	+4.08	+0.82	+2.61	+2.93	+9.49	-1.33	-0.20

P.4 Comparison with other Feature Selection Methods

Table 25 provides details of various feature selection mechanisms (Causal Pruning, Mutual Information and Linear Granger Causality) evaluated on parameter reduction and gain in mean square error for common forecasting models. We observe that Causal Pruning, which is a zero-shot byproduct of the Mask2Cause framework, demonstrates competitive results against alternative feature selection methods.

Table 25: Model-wise comparison of parameter reduction (PR, %) and MSE reduction (MSE-R, %) across pruning methods. **Bold** indicates best performer, underline indicates second best.

	VAR				Lorenz ($F = 10$)				Lorenz ($F = 40$)			
Model / Method	**T=500**		**T=1000**		**T=250**		**T=500**		**T=250**		**T=500**	
	PR	MSE-R	PR	MSE-R	PR	MSE-R	PR	MSE-R	PR	MSE-R	PR	MSE-R
MLP												
M2C (Ours)	**60.01**	-4.41	58.60	-4.06	**50.84**	**+49.35**	**50.04**	**+27.61**	48.37	**+10.33**	48.87	**+0.55**
Mutual Info.	54.19	-1.74	**59.43**	-3.41	0.83	+4.36	1.16	+1.09	33.08	-3.21	46.71	-16.74
Linear Granger	20.45	**+0.85**	9.00	**+0.53**	26.46	+27.78	18.62	+22.25	**59.18**	-9.61	**51.70**	-10.52
ARIMAX												
M2C (Ours)	**72.20**	-5.11	70.50	-5.62	**61.17**	+2.41	**60.20**	**+2.27**	58.20	+1.13	58.80	**+1.27**
Mutual Info.	65.20	-2.88	**71.50**	-3.85	1.00	-0.24	1.40	-0.91	39.80	+3.01	56.20	+0.37
Linear Granger	24.60	**+0.45**	10.83	**+0.30**	31.83	**+2.42**	22.40	-0.33	**71.20**	**+3.18**	**62.20**	+1.02
N-BEATS												
M2C (Ours)	4.8	**+10.72**	4.7	**+1.13**	**7.9**	**+20.67**	**7.9**	+5.18	7.6	**-11.26**	7.6	**-27.67**
Mutual Info.	**8.47**	+5.68	**9.28**	-0.69	0.13	-7.82	0.18	-15.73	5.17	-58.71	7.30	-103.78
Linear Granger	3.19	+3.34	1.41	+0.92	4.13	+13.35	2.91	**+19.04**	**9.24**	-76.56	**8.08**	-92.93

Q Downstream Utility: Root Cause Localisation

This appendix adds a second downstream task on real data, following the protocol of AERCA (Han et al., 2025). We use MSDS (Nedelkoski et al., 2020), a real cloud microservice telemetry benchmark. No ground-truth causal graph exists for MSDS, so graph quality is assessed indirectly, by supplying the discovered adjacency matrix to Personalized PageRank and measuring root cause localisation accuracy at k (AC@k).

M2C-NLL localises the root cause more often at $k = 1$ and $k = 5$, and the two methods are tied at $k = 10$. The wide standard deviation on AERCA's AC@1 is worth noting, since it indicates that its top-ranked candidate is far less stable across runs than the aggregate figures suggest.

Table 26: **Root cause localisation on MSDS.** AC@k using the discovered causal matrix as input to Personalized PageRank.

Method	AC@1	AC@3	AC@5	AC@10
AERCA	0.381 ± 0.408	$\mathbf{0.908 \pm 0.062}$	0.974 ± 0.027	$\mathbf{1.000 \pm 0.000}$
M2C-NLL	$\mathbf{0.446 \pm 0.108}$	0.834 ± 0.046	$\mathbf{0.995 \pm 0.003}$	$\mathbf{1.000 \pm 0.000}$

R Comparison Under Published Protocols

The tables in Section 5 report methods we ran ourselves through a single scoring harness. Three recent methods could not be placed in those tables on the same footing, because we were unable to obtain a runnable implementation of any of them. We could not locate public code for CSAM (Liu et al., 2025b) or KANGCI (Liu et al., 2025a), and the only implementation of LLM-GC (Liu et al., 2026) we found was an anonymised review artifact that we could not access. LLM-GC would in any case need the semantic modality it distils from a language model, which our synthetic benchmarks do not carry.

For these three, we quote their published numbers unchanged and run M2C on the configurations they report. This keeps the comparison meaningful at the cost of not passing every method through one harness, and we report it separately so that the provenance of each number is unambiguous. Two further limits apply to CSAM: it reports only precision, recall and F1, so F1 is the only comparable metric, and it does not state the Lorenz-96 forcing constant, so we report M2C at both values we evaluate. Table 27 also lists JRNGC (Zhou et al., 2024), a shared-weight method distinct from the component-wise NGC family, because LLM-GC reports it under the same protocol.

Table 27: **Lorenz-96 configurations from the LLM-GC protocol.** Baseline numbers are as published; M2C is run on their configurations. Settings are denoted (N, T, F).

Setting	Metric	KANGCI	JRNGC	LLM-GC	M2C
$(20, 1000, 10)$	AUROC	0.862 ± 0.018	0.934 ± 0.018	0.979 ± 0.033	$\mathbf{0.989 \pm 0.003}$
	AUPRC	0.875 ± 0.009	0.946 ± 0.015	$\mathbf{0.984 \pm 0.018}$	0.964 ± 0.005
$(20, 500, 20)$	AUROC	0.775 ± 0.016	0.903 ± 0.020	0.943 ± 0.008	$\mathbf{0.998 \pm 0.001}$
	AUPRC	0.780 ± 0.014	0.925 ± 0.022	0.950 ± 0.015	$\mathbf{0.993 \pm 0.002}$
$(40, 1000, 20)$	AUROC	0.719 ± 0.015	0.907 ± 0.008	0.932 ± 0.005	$\mathbf{0.996 \pm 0.002}$
	AUPRC	0.766 ± 0.011	0.909 ± 0.017	0.940 ± 0.016	$\mathbf{0.968 \pm 0.004}$

M2C leads on five of the six entries, and is the only method whose AUROC stays above 0.98 in all three settings.

Table 28: **CSAM comparison on Lorenz-96 ($N = 10$), F1.** The CSAM number is taken directly from the published paper.

Method	Samples	F1
CSAM (Liu et al., 2025b)	2048	0.769
M2C-MSE ($F = 10$)	2048	**1.000**
M2C-MSE ($F = 40$)	2048	**1.000**

S Heteroscedastic Structure in Real Data

The Mixed Physics benchmark is synthetic, and this appendix sets out why that is difficult to avoid and what indirect evidence real data can supply.

Why no validated real benchmark exists. We looked for a curated real-world time series carrying a known variance-driven causal graph and found none. Volatility spillover in finance and coherence modulation in neuroscience are the canonical examples, but in both fields the structure would have to

be taken from tools such as Diebold-Yilmaz connectedness indices or coherence measures, and those tools are themselves estimators of that structure. Scoring a discovery method against their output would measure how closely two estimators agree, not whether either one is correct. This is why we treat Mixed Physics as a contribution in its own right: a capability cannot be evaluated until a benchmark exists that isolates it.

Indirect evidence on ETTh1. During our experiments we observed that ETTh1, an electricity transformer temperature dataset, behaves as a heteroscedastic system would predict. We prune a one-step-ahead forecaster to the parents each method discovers, and compare against a random graph matched to the same edge density, which controls for capacity rather than structure.

Table 29: **Causal pruning on ETTh1.** One-step-ahead forecasting MSE after restricting each target to its discovered parents, against a random graph matched to the same edge density.

Pruning source	Edges	Param. reduction	MSE	MSE (density-matched random)
Vanilla (all edges)	42	0%	0.179	—
AERCA (Han et al., 2025)	3	72.6%	0.176	**0.174**
M2C-MSE	20	40.9%	**0.174**	0.182
M2C-NLL	19	42.8%	**0.169**	0.178

Both M2C graphs beat their density-matched random controls while AERCA's does not, which indicates that the discovered structure carries information beyond sparsity alone. The NLL variant's graph is the more useful of the two, and both improve on the unpruned model. On a purely homoscedastic system we would expect the NLL and MSE variants to be interchangeable, as they are on VAR and Lorenz-96, so the gap between them is itself the signal of interest.

This is suggestive evidence, not proof. ETTh1 has no causal ground truth and the differences are small. What it shows is consistent with variance-driven structure being present and exploitable in a real system, which is the most that can be claimed without a labelled benchmark.

T Practical Guidelines: Choosing Between Homoscedastic (MSE) and Heteroscedastic (NLL) Objectives

To contextualize the empirical results presented in the main text, it is important to note that the underlying data-generating processes for the VAR, Lorenz-96, CausalTime, and DREAM3 benchmarks are inherently homoscedastic. Therefore, the Mean Squared Error (MSE) objective is usually optimal for these systems.

The introduction of the Negative Log-Likelihood (NLL) objective—which jointly estimates a time-varying variance—carries a trade-off between representational capacity and parameter efficiency. On highly complex datasets with scarce samples, such as the Yeast1 network (which contains 100 genes and 166 regulatory edges, with up to seven regulators per gene), the variance estimation can destabilize, degrading the structural learning signal. Conversely, on simpler DREAM3 datasets, the NLL objective occasionally outperforms MSE. In these specific regimes, the dynamic variance acts as an adaptive regularizer, intelligently down-weighting noisy or highly volatile transitions during gradient descent.

When deploying Mask2Cause on a novel, real-world dataset where the ground truth is unknown, we recommend the following decision framework:

- **Data Scarcity:** If the dataset is highly constrained in the time dimension (T) relative to the number of variables (N), default to the **MSE** objective. The parameter efficiency and optimization stability provided by a fixed-variance assumption will generally outweigh the benefits of adaptive variance modeling.
- **Rolling Variance Check:** For appropriately sized datasets, practitioners should pre-compute the rolling variance of the target variables. If the variance appears stationary across time, default to **MSE**. However, if the time series exhibits distinct volatility clustering or heteroscedastic shocks (e.g., high-frequency financial econometrics, neurological seizure data), deploy the **NLL** objective.

U PROOFS

U.1 NEURAL GRANGER CAUSALITY IS A SPECIAL CASE OF DIRECTED INFORMATION

Theorem Let $x_t^i = g_i(\mathbf{x}_{<t}) + \epsilon_t^i$, where ϵ_t^i has a density, is independent of the full history, and has zero mean. Assume finite first moments. Then X^j is functionally non-causal for X^i almost surely (i.e., g_i admits a representation using only $\mathbf{x}_{<t}^{-j}$) if and only if $I(X^j \to X^i \mid \mathbf{X}^{-\{i,j\}}) = 0$.

Proof. The Causally Conditioned Directed Information (DI) from a source X^j to a target X^i is defined as the expected sum of Kullback-Leibler (KL) divergences between the full conditional distribution and the restricted conditional distribution:

$$I(X^j \to X^i \mid \mathbf{X}^{-\{i,j\}}) = \sum_{t=1}^{T} \mathbb{E}\Big[D_{\mathrm{KL}}\big(P(x_t^i \mid \mathbf{x}_{<t}) \,\big\|\, P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})\big)\Big] \tag{32}$$

Under the Additive Noise Model (ANM) assumption, the randomness in x_t^i arises solely from the noise term ϵ_t^i. Consequently, the conditional probability density function is equivalent to the noise density shifted by the deterministic prediction function g_i:

$$P(x_t^i \mid \mathbf{x}_{<t}) = P_\epsilon\left(x_t^i - g_i(\mathbf{x}_{<t})\right) \tag{33}$$

If X^j is functionally non-causal for X^i, then by definition, the function g_i does not depend on the history of X^j. Mathematically, this implies invariance:

$$g_i(\mathbf{x}_{<t}) = g_i(\mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}}) \tag{34}$$

Substituting this invariance into the density formulation:

$$\begin{aligned} P(x_t^i \mid \mathbf{x}_{<t}) &= P_\epsilon\left(x_t^i - g_i(\mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})\right) \\ &= P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}}) \end{aligned} \tag{35}$$

Since the full conditional distribution $P(x_t^i \mid \mathbf{x}_{<t})$ is identical to the restricted distribution $P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})$ at every time step t, the KL divergence term vanishes:

$$D_{\mathrm{KL}}\Big(P(x_t^i \mid \mathbf{x}_{<t}) \,\Big\|\, P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})\Big) = \mathbb{E}\left[\log \frac{\cancel{P(x_t^i \mid \mathbf{x}_{<t})}}{\cancel{P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})}}\right] = 0 \tag{36}$$

Thus, the Causally Conditioned Directed Information is exactly zero.

Conversely, if DI is zero, nonnegativity of KL divergence implies equality of the full and source-excluded conditional distributions almost surely at every evaluated time step. Their conditional means therefore agree. Since the noise is independent of the history and has zero mean,

$$g_i(\mathbf{x}_{<t}) = \mathbb{E}[x_t^i \mid \mathbf{x}_{<t}] = \mathbb{E}[x_t^i \mid \mathbf{x}_{<t}^{-j}] \quad \text{almost surely.}$$

The right-hand side depends only on the remaining histories, establishing functional noncausality almost surely. □

Remark (Failure of the Converse in Heteroscedastic Systems). It is crucial to note that, for general dynamical systems with finite first moments, zero DI implies invariance of the conditional mean, but the converse does not hold. If a variable X^j influences the *variance* (or higher-order moments) of X^i on a set of positive probability without affecting its conditional mean g_i, then $g_i(\mathbf{x}_{<t}) = g_i(\mathbf{x}_{<t}^{-j})$ holds, satisfying the condition for Functional Non-Causality (zero NGC). However, in such a case, the probability distributions would differ in shape (e.g., width), i.e., $P(x_t^i \mid \mathbf{x}_{<t}) \neq P(x_t^i \mid \mathbf{x}_{<t}^{-j})$. The Directed Information, which detects any distributional discrepancy, would correctly be non-zero ($I > 0$). This gap highlights precisely why standard functional methods fail on the Mixed Physics benchmark, while probabilistic objectives succeed.

U.2 Optimal Gaussian Prediction and Conditional Moments

Proposition. Fix a target x_t^i with finite second moment. Let (μ, v) and (μ_{-j}, v_{-j}) be its conditional mean and variance given $\mathbf{x}_{<t}$ and $\mathbf{x}_{<t}^{-j}$, respectively. Assume $v, v_{-j} > 0$ almost surely and $\mathbb{E}|\log v| + \mathbb{E}|\log v_{-j}| < \infty$. Among unrestricted measurable Gaussian predictors with positive variance, the population-optimal predictions equal these conditional moments. Excluding X^j strictly increases optimal Gaussian NLL if and only if either moment changes on a set of positive probability; it strictly increases optimal MSE if and only if the conditional mean changes on such a set.

Proof. For either choice of history, the conditional expected Gaussian NLL at predicted mean m and variance $s^2 > 0$, omitting the constant $\frac{1}{2}\log(2\pi)$, is

$$\frac{1}{2}\log s^2 + \frac{v + (\mu - m)^2}{2s^2}.$$

It is uniquely minimized by $m = \mu$ and $s^2 = v$; the same argument applies to the source-excluded moments. Evaluating both optimal predictors under the full history and then taking expectations gives

$$\begin{aligned}\mathcal{L}^*_{\mathrm{NLL},-j} - \mathcal{L}^*_{\mathrm{NLL,full}} &= \frac{1}{2}\mathbb{E}\left[\log\frac{v_{-j}}{v} + \frac{v + (\mu - \mu_{-j})^2}{v_{-j}} - 1\right] \\ &= \mathbb{E}[D_{\mathrm{KL}}(\mathcal{N}(\mu, v) \,\|\, \mathcal{N}(\mu_{-j}, v_{-j}))]\,.\end{aligned}$$

This is zero exactly when both moments agree almost surely, and strictly positive otherwise. For squared error, the conditional expectation minimizes risk and the orthogonality of its residual yields

$$\mathcal{L}^*_{\mathrm{MSE},-j} - \mathcal{L}^*_{\mathrm{MSE,full}} = \mathbb{E}[(\mu - \mu_{-j})^2].$$

Thus a variance-only parent improves optimal Gaussian prediction while leaving optimal squared error unchanged. The data need not be Gaussian: the Gaussian distributions in the KL expression are the optimal predictive models, not necessarily the true conditional densities. □

Corollary (Gaussian full conditionals and DI). Under the preceding moment and integrability assumptions at each evaluated time step, suppose additionally that $p(x_t^i \mid \mathbf{x}_{<t})$ is Gaussian almost surely. For $j \neq i$, the cumulative optimal Gaussian NLL increases upon excluding source j if and only if $I(X^j \to X^i \| \mathbf{X}^{-\{i,j\}}) > 0$. The source-excluded conditional density need not be Gaussian.

Proof. Zero DI means that the full and source-excluded conditional distributions agree almost surely at every evaluated time step. Their moments then agree, so the preceding proposition gives zero Gaussian NLL increase. Conversely, zero cumulative Gaussian NLL increase implies equality of the full and reduced moments at every step, since each step's increase is nonnegative. The full conditional law is therefore $\mathcal{N}(\mu_{-j}, v_{-j})$, whose parameters depend only on the source-excluded history. Averaging this law over the missing history leaves it unchanged, so it also equals the reduced conditional law. Every DI summand is consequently zero. Taking the contrapositive establishes the equivalence of strict positivity. This identifies the same population edges, not necessarily the same numerical loss reduction as DI, and does not assert recovery by a fitted network. □

U.3 Proof of Zero Self-Information under Causal Conditioning

Proposition. *For any variable X^i, the Causally Conditioned Directed Information from X^i to itself is exactly zero:*

$$I(X^i \to X^i \mid \mathbf{X}^{-\{i\}}) = 0 \tag{37}$$

Proof. Recall the definition of Causally Conditioned Directed Information from a source X^j to a target X^i, conditioned on the set of all other variables $\mathbf{X}^{-\{i,j\}}$:

$$I(X^j \to X^i \mid \mathbf{X}^{-\{i,j\}}) = \sum_{t=1}^{T} \mathbb{E}\Big[D_{\mathrm{KL}}\big(P(x_t^i \mid \mathbf{x}_{<t}) \,\big\|\, P(x_t^i \mid \mathbf{x}_{<t}^i, \mathbf{x}_{<t}^{-\{i,j\}})\big)\Big] \tag{38}$$

where $\mathbf{x}_{<t}$ denotes the full history of all variables in the system up to time $t-1$.

To evaluate the self-information flow, we set the source variable to be the target itself, i.e., let $j = i$. In this case, the exclusion set $\mathbf{X}^{-\{i,j\}}$ becomes $\mathbf{X}^{-\{i,i\}} = \mathbf{X}^{-\{i\}}$. The term inside the second probability distribution becomes:

$$\{\mathbf{x}^i_{<t}, \mathbf{x}^{-\{i\}}_{<t}\} \equiv \mathbf{x}_{<t} \tag{39}$$

This is simply the full system history $\mathbf{x}_{<t}$.

Substituting this back into the KL divergence term:

$$D_{\mathrm{KL}}\big(P(x^i_t \mid \mathbf{x}_{<t}) \,\|\, P(x^i_t \mid \mathbf{x}_{<t})\big) \tag{40}$$

Since the Kullback-Leibler divergence between two identical distributions is exactly zero ($D_{\mathrm{KL}}(P\|P) = 0$), every term in the summation vanishes. Therefore:

$$I(X^i \to X^i \mid \mathbf{X}^{-\{i\}}) = \sum_{t=1}^{T} 0 = 0 \tag{41}$$

□

U.4 A Concrete Variance-Only Dependency

Let (X_t) and (ϵ_t) be mutually independent sequences of independent standard Gaussian variables, and define

$$Y_t = \sqrt{\beta + aX^2_{t-1}}\,\epsilon_t, \qquad \beta > 0, \quad a > 0.$$

The source X_{t-1} is independent of $Y_{<t}$. Consequently,

$$\mathbb{E}[Y_t \mid X_{<t}, Y_{<t}] = \mathbb{E}[Y_t \mid Y_{<t}] = 0,$$

whereas

$$\mathrm{Var}(Y_t \mid X_{<t}, Y_{<t}) = \beta + aX^2_{t-1}, \qquad \mathrm{Var}(Y_t \mid Y_{<t}) = \beta + a.$$

Both optimal squared-error predictors therefore output zero and have expected loss $\beta+a$. Nevertheless, the full and reduced conditional distributions differ with positive probability, since their variances differ. Thus $I(X_{<t}; Y_t \mid Y_{<t}) > 0$, yielding a positive contribution to the strictly lagged DI from X to Y.

The optimal expected Gaussian NLL reduction is

$$\frac{1}{2}\left\{\log(\beta + a) - \mathbb{E}\big[\log(\beta + aX^2_{t-1})\big]\right\} > 0.$$

Strict positivity follows from strict concavity of the logarithm and the nonconstant conditional variance; $\beta > 0$ and finite second moments ensure the required integrability. This is a variance-only dependency that optimal Gaussian prediction detects and optimal squared-error prediction misses. It lies outside the history-independent additive-noise premise of Appendix U.1.

U.5 Idealized Mean-Based AUROC Reference

Proposition. Suppose a graph has m mean-dependent edges, v variance-only edges and at least one non-edge, with $m + v > 0$. Assume every mean-dependent edge scores above every non-edge, while a uniformly selected variance-only edge versus a uniformly selected non-edge has probability of winning plus half its probability of tying equal to $1/2$. Then the tie-adjusted AUROC is

$$\mathrm{AUROC}_{\mathrm{ref}} = \frac{m + \frac{1}{2}v}{m + v}.$$

Proof. AUROC is the probability that a uniformly selected positive edge scores above a uniformly selected non-edge, with ties contributing one half. The positive edge is mean-dependent with probability $m/(m + v)$ and variance-only with probability $v/(m + v)$. Conditioning on these two cases gives

$$\mathrm{AUROC}_{\mathrm{ref}} = \frac{m}{m + v} \cdot 1 + \frac{v}{m + v} \cdot \frac{1}{2}.$$

If the chance-ranking premise holds in expectation over random scores, the same calculation gives expected AUROC. □

The Mixed Physics configurations have $(m, v) = (13, 14), (7, 20), (0, 27)$, giving $20/27 \approx 0.741$, $17/27 \approx 0.630$ and $1/2 = 0.500$, respectively. These are reference values under the stated ranking assumptions, not universal bounds or guaranteed AUROCs for MSE-trained networks. The absence of a population MSE gain for a variance-only source does not itself determine the edge-score rankings of a fitted model.

U.6 Log-Loss Reduction for Fitted Density Predictors

The identity in Eq. (10) uses the true conditional densities. A neural density predictor may approximate these imperfectly. The following standard log-loss decomposition separates the information supplied by a source from the errors in predicting the two conditional distributions.

Proposition. For a fixed target i and source $j \neq i$, let $p_{\text{full},t}$ and $p_{-j,t}$ be the true full-history and source-excluded conditional densities. Let $q_{\text{full},t}$ and $q_{-j,t}$ be any fixed predictive densities using only their respective histories; in particular, $q_{-j,t}$ cannot access source j's past. Assume the expected absolute log-densities evaluated at x_t^i are finite. Define the cumulative expected modeling errors

$$\mathcal{E}_{\text{full}} = \sum_{t=1}^{T} \mathbb{E}\big[D_{\text{KL}}(p_{\text{full},t} \| q_{\text{full},t})\big], \qquad \mathcal{E}_{-j} = \sum_{t=1}^{T} \mathbb{E}\big[D_{\text{KL}}(p_{-j,t} \| q_{-j,t})\big].$$

The cumulative expected NLL reduction from using the full predictor is

$$\begin{aligned} \Delta_q &:= \mathbb{E}\left[\sum_{t=1}^{T} \log \frac{q_{\text{full},t}(x_t^i)}{q_{-j,t}(x_t^i)}\right] \\ &= I(X^j \to X^i \| \mathbf{X}^{-\{i,j\}}) + \mathcal{E}_{-j} - \mathcal{E}_{\text{full}}. \end{aligned} \tag{42}$$

Proof. For either history choice, conditional expectation gives

$$\mathbb{E}[-\log q_t(x_t^i)] = \mathbb{E}[-\log p_t(x_t^i)] + \mathbb{E}[D_{\text{KL}}(p_t \| q_t)].$$

Subtract the full-history identity from the source-excluded identity, sum over time, and apply Eq. (10), attributed to Quinn et al. (2015). □

Interpretation. Each predictor's expected loss equals the best achievable loss given its available histories plus its error in modeling the corresponding distribution. Their loss difference therefore reflects both the source's predictive information and the difference between their modeling errors. A reduced predictor can model its conditional distribution perfectly despite having less information: the information loss is already accounted for by DI. If both modeling errors are zero, Δ_q equals DI; if each is at most ε, then $|\Delta_q - I| \leq \varepsilon$, where I denotes the DI in Eq. (42).

V Limitations and Broader Impacts

V.1 Limitations

While Mask2Cause successfully captures both mean and variance-driven causality, the heteroscedastic NLL objective uses a Gaussian predictive density, although its population-optimal moments remain valid for non-Gaussian data under the conditions in Appendix U.2. Because Gaussian NLL penalizes errors quadratically, extreme outliers in heavy-tailed data force the network to artificially inflate the predicted variance, which can wash out the structural gradient (potentially leading to spurious edge detection). If heavy tails are detected in the dataset (e.g., via a high kurtosis statistic), the framework can however be trivially adjusted by swapping the Gaussian likelihood for a Student-t or Laplace distribution NLL. Like all methods grounded in the Granger Causality framework, Mask2Cause relies on the assumption of causal sufficiency. It assumes there are no unobserved latent confounders driving the observed variables. In the presence of hidden confounders, Granger-based methods may infer spurious direct edges. Extending this framework to detect or mitigate hidden confounding remains an important direction for future work. Our architecture assumes the causal graph is fixed. While the conditional mean and variance can change with the observed history even in a stationary process, the underlying structural adjacency matrix $\hat{\mathbf{A}}$ is assumed to be static across the evaluated time horizon T. Adapting the masking mechanism to handle non-stationary, regime-switching causal structures is a promising avenue for further research.

V.2 BROADER IMPACTS

Causal discovery from time series has profound implications across domains ranging from climate science to financial econometrics. On the positive side, robust causal graphs can improve the interpretability and safety of automated decision-making systems (e.g., medical interventions or traffic grid optimization) by distinguishing true mechanisms from spurious correlations. A potential negative impact arises if the models are applied to sensitive human-centric data without accounting for unobserved confounders, which could lead to flawed causal conclusions and biased policy interventions. To mitigate this, practitioners must validate the discovered graphs against domain expertise before deployment.