Received: 18 May 2026 | Revised: 29 June 2026 | Accepted: 20 July 2026 | Published online: 18 September 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1–15
DOI: 10.47852/bonviewJDSIS620210441

BON VIEW PUBLISHING

REIGN: Regime-Enhanced Intelligent Granger Network for Nonstationary Causal Discovery

Quang-Vinh Dang[1], Minh Ngoc Dinh[2], Dat Le[3,*] and Dinh-Minh Nguyen[1]

[1] *British University Vietnam, Vietnam*

[2] *Millennia Education, Vietnam*

[3] *University of Economics and Finance, Vietnam*

Abstract: Causal discovery in nonstationary multivariate financial time series is a fundamental challenge. Classical algorithms, such as Peter–Clark (PC) algorithm and Peter–Clark momentary conditional independence (PCMCI+), assume stationarity and fail in real-world environments characterized by market regime shifts and structural breaks, yielding spurious edges and structurally inconsistent causal graphs. To address this, we propose the Regime-Enhanced Intelligent Granger Network (REIGN), a novel causal discovery framework specifically engineered for nonstationary time series. REIGN integrates three synergistic capabilities: (i) data-driven regime segmentation via Pruned Exact Linear Time changepoint detection, partitioning the series into locally stationary windows; (ii) zero-shot large language model prior injection to constrain the neural search space using domain knowledge; and (iii) a coarse-to-fine Message-Passing Graph Neural Network that optimizes a continuous directed acyclic graph-constrained adjacency matrix via an augmented Lagrangian objective. Regime-specific models are then aggregated through a confidence-weighted ensemble to yield a consistent global summary graph. Empirical evaluation on a nonstationary Vector Autoregressive benchmark demonstrates REIGN's superiority. REIGN achieves an Optimal F1 score of 0.529, substantially outperforming PCMCI+ (F1 = 0.415) and PC (F1 = 0.214) in structural edge recovery. Furthermore, REIGN dominates across the area under the receiver operating characteristic curve (AUROC; 0.513 vs 0.406) and the area under the precision–recall curve (AUPR; 0.400 vs 0.338) metrics, establishing a robust new benchmark for neural time-series causal discovery in nonstationary environments.

Keywords: causal discovery, time series, Granger causality, large language models, nonstationary regimes

1. Introduction

Causal discovery—the inference of cause-and-effect relationships from purely observational data—is a central challenge in applied machine learning (ML) across disciplines as diverse as neuroscience, biology, economics, and finance. In these domains, understanding the causal structure of a system is crucial for making reliable predictions under interventions, designing robust policies, and explaining observed phenomena beyond mere correlation. Formally, the objective is to recover a directed acyclic graph (DAG) $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ over the set of observed variables $\mathscr{V}$, where a directed edge $(i \rightarrow j) \in \mathscr{E}$ encodes the hypothesis that variable i is a direct Granger cause of variable j.

1.1. The nonstationarity challenge

In temporally ordered multivariate datasets, the challenge is compounded by the presence of lag-dependent interactions. Temporal causal discovery methods operate on the principle of Granger causality [1]: variable X^i is said to Granger-cause X^j if the past values of X^i improve predictions of the future of X^j beyond what is achievable from the history of X^j alone. Classical algorithms such as the Peter–Clark (PC) algorithm and its temporal extension Peter–Clark momentary conditional independence plus (PCMCI+) [2] rely on conditional independence testing to identify Granger-causal relationships, while score-based approaches such as DYNOTEARS [3] frame the problem as a continuous optimization.

However, all of these methods share a critical and often unexamined assumption: the data-generating process is *stationary*. In financial markets, this assumption is routinely violated. Stock price correlations, cross-asset spillover effects, and inter-market transmission mechanisms shift dramatically across different market regimes—from bull markets to bear markets, from periods of low volatility to crisis-driven turbulence. These structural breaks render any single-model causal estimate invalid across the entire time span. An algorithm trained on pre-crisis dynamics will output a causal graph that is not only incorrect but also potentially misleading in its post-crisis predictions.

*Corresponding author: Dat Le, University of Economics and Finance, Vietnam. Email: datla@uef.edu.vn

1.2. Neural approaches and their limitations

Recent advances in deep learning have produced powerful neural architectures for causal discovery. CUTS+ [4] deploys a coarse-to-fine Message-Passing Graph Neural Network (MPGNN) to handle high-dimensional, irregular time series. NoCurl and NOTEARS [5] introduced the foundational continuous differentiable DAG constraint $h(\mathbf{W}) = \mathrm{tr}(e^{\mathbf{W}\circ\mathbf{W}}) - N = 0$, enabling gradient-based optimization over the combinatorially intractable DAG space. These methods demonstrate strong performance on stationary benchmarks but are fundamentally agnostic to temporal nonstationarity. When applied to multi-regime data, they produce globally smoothed, blurred causal estimates that fail to capture the true regime-dependent structure.

While some existing methods attempt to model nonstationarity directly, they exhibit critical shortcomings in complex financial scenarios. Algorithms such as constraint-based causal discovery from nonstationary/heterogeneous data (CD-NOD) [6] and Regime-PCMCI [7] introduce regime-awareness, but they often struggle with the high dimensionality and irregular-sampling frequencies typical of real-world financial data. More importantly, these purely data-driven nonstationary methods lack a mechanism to incorporate domain-specific topological priors.

Furthermore, neural causal discovery is notoriously sample-inefficient. The combinatorial complexity of the DAG search space grows super-exponentially with the number of variables. In sparse financial datasets without strong domain knowledge, algorithms frequently converge to suboptimal local minima, particularly when the time series is short within a specific turbulent market regime.

1.3. The REIGN framework

To address these fundamental limitations, we propose the Regime-Enhanced Intelligent Granger Network (REIGN), an end-to-end causal discovery framework engineered from the ground up to handle nonstationary environments. REIGN's architecture integrates four synergistic components. First, a robust preprocessing pipeline applies Multiple Imputation by Chained Equations (MICE)-based imputation and outlier handling. Second, the Pruned Exact Linear Time (PELT) algorithm detects changepoints and segments the time series into locally stationary regimes, ensuring each downstream model operates on a homogeneous distribution. Third, a novel prior-injection mechanism queries a large language model (LLM) to extract domain-specific zero-shot topological priors, dramatically constraining the neural search space without requiring labeled training data. Finally, a per-regime MPGNN engine optimizes a continuous DAG-constrained adjacency matrix, anchored by the LLM prior, before a confidence-weighted ensemble aggregates the local models into a globally consistent causal graph.

The primary contributions of this work are four-fold:

1) We introduce a principled architecture for regime-aware temporal causal discovery, employing the PELT algorithm to partition nonstationary time series into stationary local windows and enabling each neural Granger model to specialize on a homogeneous distributional regime.
2) We pioneer the integration of zero-shot LLM-generated structural priors into a differentiable continuous Lagrangian optimization objective, bridging the gap between human domain expertise and deep causal learning without requiring explicit human annotation.
3) We propose a novel confidence-weighted ensemble mechanism that synthesizes regime-specific local DAGs into a globally consistent summary graph through temporal importance weighting.
4) We demonstrate through rigorous empirical evaluation on synthetic nonstationary Vector Autoregressive (VAR) datasets that REIGN achieves state-of-the-art performance, outperforming established baselines such as PCMCI+ by over 11% in F1 score and dominating across all continuous structural metrics.

2. Related Work

2.1. Classical causal discovery and continuous structure learning

Score-based causal discovery methods recover DAGs by optimizing a statistical criterion over the space of DAG structures. A major algorithmic breakthrough came with NOTEARS [5], which reformulated structure learning as a fully continuous-optimization problem by introducing the smooth acyclicity constraint $h(\mathbf{W}) = \mathrm{tr}(\exp(\mathbf{W} \circ \mathbf{W})) - d = 0$, enabling gradient-based DAG learning over the full adjacency matrix and making downstream integration with neural architectures tractable. REIGN inherits this acyclicity regularizer and applies it jointly with the neural Granger and LLM-prior objectives inside each regime window. Despite its elegance, the NOTEARS framework assumes stationarity of the underlying data-generating process, a restriction that is rarely satisfied in real-world financial time series exhibiting regime-switching dynamics [8, 9].

2.2. Neural Granger causality

Classical Granger causality tests each variable pair in isolation, making it susceptible to spurious edges from indirect effects and shared latent drivers. Tank et al. [10] introduced component-wise multilayer perceptrons (MLPs) and long short-term memory (LSTM) networks (cMLP, cLSTM) whose group-lasso-penalized first-layer input weights serve as Granger causality indicators, establishing the paradigm of end-to-end neural Granger learning from multivariate time series. Marcinkevičs and Vogt [11] extended this paradigm with Generalized VAR, a self-explaining neural network that jointly infers Granger-causal structure and detects the sign and time-varying magnitude of each causal effect, yielding substantially more interpretable outputs than sparse-input networks. CUTS [4] extended neural Granger causality to irregular time series through an alternating imputation-and-discovery scheme using a delayed-supervision graph neural network (GNN), directly addressing the irregular-sampling challenge common in business key performance indicator (KPI) datasets. CUTS+ [12] further scaled this framework to high-dimensional settings via a coarse-to-fine discovery (C2FD) procedure and an MPGNN and constitutes the primary backbone of REIGN. Most recently, Qian et al. [13] proposed MSDG, a multiscale dynamic graph neural network that reconstructs multivariate time series into dynamic graphs using causal windows of varying lengths and applies a graph attention network jointly with an LSTM to capture time-varying causal contributions; while MSDG advances dynamic Granger learning, it relies on heuristic window selection and does not leverage LLM semantic priors or statistically principled regime detection. Our framework extends CUTS+ along three orthogonal dimensions: LLM-prior regularization, regime-aware segmentation, and confidence-weighted ensemble fusion, none of which appear in any of the above designs.

2.3. Time-series causal discovery

Time-series causal discovery must additionally account for temporal autocorrelation, lagged dependencies, and the possibility of contemporaneous effects. Runge [14] proposed PCMCI+, which separates the lagged and contemporaneous edge-removal phases and uses momentary conditional independence (MCI) to control for autocorrelation, achieving substantially higher recall than constraint-based predecessors on observational time series. Günther et al. [15] proposed J-PCMCI+, which extends PCMCI+ to jointly infer causal graphs across multiple heterogeneous time-series datasets that share latent contextual variables, enabling discovery under cross-dataset distribution shift; this setting closely resembles real-world business deployments where the same causal graph applies across but differs between organizational units or product lines. Mameche et al. [8] introduced SPACETIME, a functional nonparametric framework that jointly reconstructs temporal regimes and causal mechanisms using Gaussian processes and minimum-description-length scoring, explicitly unifying the tasks of regime detection, dataset partitioning, and causal graph discovery under nonstationarity; SPACETIME demonstrates that treating nonstationarity as an opportunity rather than a nuisance consistently improves edge precision on climate and ecological datasets, a finding we exploit in REIGN's regime-local design.

2.4. LLM-augmented causal discovery

Integrating LLMs into causal discovery pipelines has emerged as a rapidly developing research direction, motivated by the observation that LLMs encode rich, text-derived domain knowledge about cause–effect relationships. Jin et al. [16] established the foundational result that LLMs can serve as effective probabilistic priors for causal graphs by querying pairwise edge existence and direction to produce soft adjacency probabilities compatible with score-based discovery pipelines. Long et al. [17] investigated how LLMs can be treated as imperfect domain experts whose erroneous causal claims must be corrected by enforcing acyclicity and Markov-equivalence-class consistency before integration into discovery pipelines; their analysis of failure modes directly motivates REIGN's use of confidence-weighted soft priors rather than hard constraints. Du et al. [18] proposed LLM-CD, which integrates LLM-derived knowledge at multiple stages of the PC algorithm, reporting an average recall improvement of 169% and a maximum of 403% on the WUSCH dataset compared to purely data-driven baselines. Roy et al. [19] introduced Causal-LLM, a unified one-shot framework combining prompt-based and data-driven modes that outperforms classical score-based methods by approximately 40% in edge accuracy on the Asia and Sachs benchmarks. Kampani et al. [20] proposed LLM-initialized differentiable causal discovery (LLM-DCD), using LLM-generated soft adjacency as a warm-start initialization for gradient-based DAG learning, the published architecture most similar to REIGN; our framework differs by employing neural Granger scoring throughout training and by applying regime segmentation prior to discovery. Liu et al. [21] introduced LLM-GC, the first method to directly bridge LLMs and neural Granger causality for time series via dual-modality encoding and a causality-aware self-attention mechanism; however, LLM-GC does not address nonstationarity, regime switching, or DAG acyclicity enforcement, all of which are central to REIGN. Vashishtha et al. [22] demonstrated that eliciting a topological causal ordering from an LLM, rather than pairwise edge scores, yields more reliable priors by suppressing cyclic responses inherent in independent pairwise prompting; REIGN adopts this causal-order correction step to post-process $\mathbf{A}_{\mathrm{LLM}}$ before fusing it into the CUTS+ training objective. Zečević et al. [23] critically re-evaluated the role of LLMs in causal discovery, arguing that their autoregressive, correlation-driven modeling inherently lacks the theoretical grounding for causal reasoning, and demonstrated through empirical studies that deliberate prompt engineering can overstate LLM-based causal discovery performance by injecting ground-truth knowledge; this finding reinforces REIGN's design choice of restricting LLMs to a non-decisional soft-prior role whose influence is continuously modulated by data evidence through the regularization parameter λ.

2.5. Nonstationary and regime-aware causal discovery

Financial and business time series exhibit well-documented nonstationarity: the causal graph governing customer churn during a promotional campaign window differs structurally from that in a steady-state period. Balsells-Rodas et al. [9] established identifiability guarantees for regime-dependent causal discovery using high-order Markov switching models (MSMs), which extend hidden Markov models (HMMs) by incorporating autoregressive dependencies on observations given discrete latent states; their theoretical framework provides the foundational justification for REIGN's assumption that locally stationary causal structure can be recovered within each Bayesian Online Changepoint Detection (BOCPD)/PELT-detected segment. Mameche et al. [8] complemented this identifiability analysis with an empirical demonstration of joint regime-and-graph recovery in a nonparametric functional framework, confirming that regime-aware methods consistently outperform single-graph alternatives under real-world temporal distribution shift. To identify regime boundaries, REIGN supports both the PELT algorithm [24] for offline retrospective discovery and BOCPD [25] for prospective deployment. PELT provides exact dynamic-programming detection with linear expected runtime, while BOCPD yields a principled probabilistic run-length posterior that allows REIGN to quantify boundary uncertainty and propagate it into the confidence-weighted ensemble.

2.6. LLMs in scientific and data-driven machine learning

The success of LLMs in causal discovery reflects a broader trend of leveraging large foundation models to guide data-driven scientific modeling, providing a methodological context for REIGN's prior-injection design. Wuwu et al. [26] introduced PINNsAgent, an LLM-based agent that automates physics-informed neural network architecture selection and hyperparameter tuning for partial differential equations (PDEs) by encoding prior knowledge of solved PDEs as structured memory and performing Memory Tree Reasoning over the design space; PINNsAgent exemplifies how domain knowledge encoded in LLMs can steer a computationally expensive search process without supplanting physical modeling, a principle that directly parallels REIGN's use of LLM-derived priors to regularize, rather than replace, neural Granger estimation. Tiwari et al. [27] proposed the Latent Mamba Operator (LaMO), which integrates state-space model efficiency in latent space with the expressive power of kernel integral formulations to solve PDEs on diverse geometries; LaMO's architecture, combining a compact latent representation with a structured long-range dependency mechanism, shares the computational philosophy of CUTS+'s C2FD, where a cheap first pass reduces the effective dimensionality before fine-grained estimation.

2.7. Evaluation benchmarks

Sachs et al. [28] produced the canonical real-world benchmark for causal discovery: an 11-node protein-signaling network derived from 853 single-cell flow-cytometry measurements under known interventions, against which the static slice of REIGN's output is validated. Stein et al. [29] released CausalRivers, the largest in-the-wild time-series causal discovery benchmark to date, comprising over 1000 river-discharge stations at 15-min resolution with curated ground-truth causal graphs; the authors find that Granger-based approaches remain competitive on this real-world benchmark, directly supporting the choice of neural Granger as REIGN's backbone. Ferdous et al. [30] introduced TimeGraph, a suite of synthetic time-series datasets with independently controllable violation factors including nonstationarity, irregular sampling, missing data, and heterogeneous noise, enabling fine-grained ablation studies of each REIGN component under isolated assumption violations. Together, these three benchmarks form the multi-tier evaluation protocol adopted in this paper.

2.8. Recent developments in causal discovery and foundation models (2022–2026)

Neural causal discovery. Gong et al. [31] introduced Rhino, combining vector auto-regression, deep learning, and variational inference to model history-dependent noise in temporal causal discovery. Lippe et al. [32] established identifiability conditions for causal representations from temporal intervention sequences, and Annadani et al. [33] proposed BayesDAG for full Bayesian posterior inference over DAGs. Geffner et al. [34] developed deep end-to-end causal inference (DECI) for joint causal discovery and treatment effect estimation, while Ashman et al. [35] addressed latent confounders via neural acyclic directed mixed graph (ADMG) learning. Assaad et al. [36] surveyed and evaluated causal discovery methods for both i.i.d. and time-series data, confirming that PCMCI+ and neural Granger methods are most reliable under autocorrelation. Cheng et al. [37] introduced CausalTime, a pipeline for generating realistic time series with ground-truth causal graphs from real observational data. Further advances include amortized variational inference [38], diffusion-based structure learning [39], and gradient-based nonlinear Granger estimation [10].

LLMs and causal reasoning. Kıcıman et al. [40] benchmarked LLM-causal reasoning, finding that GPT-4 achieves 97% on pairwise causal discovery and 92% on counterfactual tasks, justifying the use of LLM-derived structural priors in REIGN. Chi et al. [41] and Jin et al. [16] surveyed and empirically evaluated LLM-causal integration, identifying prior injection as the most productive mode while flagging prompt sensitivity and hallucination under distributional shift as key risks. A complementary IJCAI 2025 survey [42] catalogs regime-aware temporal causal discovery as an underexplored gap that REIGN directly addresses. Recent work has further studied LLMs as autonomous causal agents [43], LLM understanding of causal graph structure [44], interventional reasoning evaluation [45], and causal reasoning fallacies [46].

Neural operators and scientific foundation models. The LLM-guided scientific ML community has developed complementary architectures that share REIGN's principle of structured knowledge guiding data-driven estimation. Cheng et al. [47] formally connected Mamba state-space models to neural operators, demonstrating superior long-range dependency capture over Transformers. Song and Jiang [48] and the AMO team [49] applied adaptive Fourier decomposition theory to neural operator design for inverse PDE problems and arbitrary-geometry meshes. Additional advances include geometry-aware Mamba operators [50], physics-informed operator learning [51], and neural operators for PDE learning [52].

Temporal nonstationarity and financial applications. The toolkit for nonstationary causal discovery has expanded with methods for periodic nonstationarity [53], proxy-variable conditional independence testing [54], self-compatibility evaluation without ground truth [55], nonstationary sparse transition representation learning [56], and causal foundation models for time series [57]. Score-based approaches using neural ordinary differential equations (ODEs) [58], single-variable interventions for causal ordering [59], and Transformer-based temporal discovery [60] extend the differentiable structure learning paradigm. In financial applications, regime-dependent Granger networks have been shown to outperform single-graph alternatives for systemic risk modeling [61], causal churn modeling advances interventional analysis [62], and the MSM identifiability framework has been extended to instantaneous effects and exponential family noise [63].

3. Methodology

3.1. Problem formulation

Let $\mathbf{X} \in \mathbb{R}^{T\times N}$ denote a multivariate time series of T observations over N variables. We assume the underlying data-generating process is a nonstationary Structural Vector Autoregressive (SVAR) model defined by a discrete latent regime variable $S_t \in \{1, \ldots, M\}$:

$$\mathbf{X}_t = \sum_{\tau=1}^{L} \mathbf{A}^{(\tau,S_t)}\mathbf{X}_{t-\tau} + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(S_t)}) \tag{1}$$

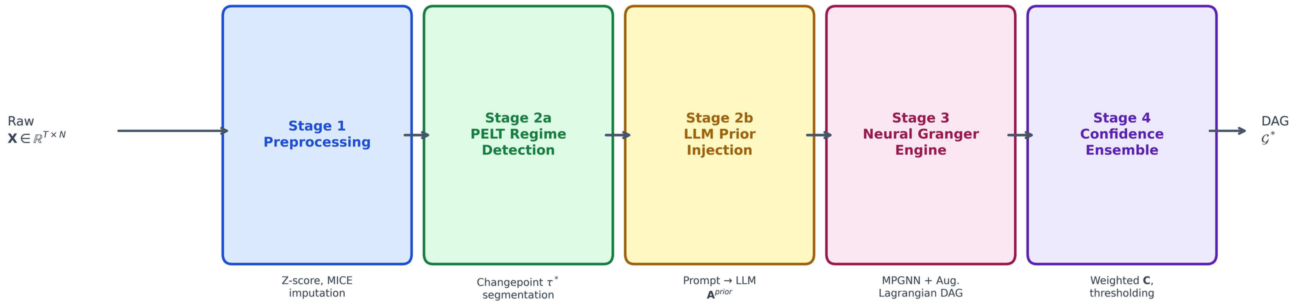
where L is the maximum temporal lag, $\mathbf{A}^{(\tau,S_t)} \in \mathbb{R}^{N\times N}$ is the regime-dependent transition matrix at lag τ during hidden state S_t, and $\boldsymbol{\varepsilon}_t$ is zero-mean, regime-specific Gaussian noise. The target is to infer the global summary causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where a directed edge $(i \to j) \in \mathcal{E}$ denotes that variable i Granger-causes variable j under at least one regime. REIGN operates through four cascading stages to recover $\mathcal{G}$ from purely observational data, as illustrated in Figure 1.

3.2. Stage 1: robust preprocessing

Financial and physiological time series commonly contain missing data segments and extreme measurement anomalies. We apply the following preprocessing pipeline:

1) Z-score normalization: Each channel $X^{(n)}$ is standardized using robust estimates of location and scale computed from non-missing entries.
2) Outlier flagging: Entries where $|X_t^{(n)} - \mu_n| > 3\sigma_n$ are flagged and treated as missing.
3) MICE Imputation: Missing entries are imputed using MICE, which iteratively fits a sequence of conditional models

Figure 1
The REIGN end-to-end causal discovery pipeline. Raw multivariate time series X passes sequentially through robust preprocessing (Stage 1), PELT-based regime segmentation (Stage 2a), LLM prior injection (Stage 2b), per-regime MPGNN training (Stage 3), and confidence-weighted ensemble aggregation (Stage 4) to produce the final DAG $\mathscr{G}^*$



$p(X^{(n)}|\mathbf{X}^{(-n)})$ to preserve the inter-channel covariance structure critical to Granger causality detection.

3.3. Stage 2a: statistical regime detection via PELT

We model the regime shifts as a sequence of abrupt change-points partitioning $[1,T]$ into $K+1$ contiguous segments. The PELT algorithm [24] efficiently identifies the optimal set of change-points $\mathscr{T}^* = \{\tau_1^*, \ldots, \tau_K^*\}$ by minimizing the penalized to-tal cost:

$$\mathscr{T}^* = \operatorname*{argmin}_{K,\mathscr{T}} \left\{ \sum_{k=0}^{K} \mathscr{C}\left(\mathbf{X}_{(\tau_k,\tau_{k+1}]}\right) + \beta(K+1) \right\} \quad (2)$$

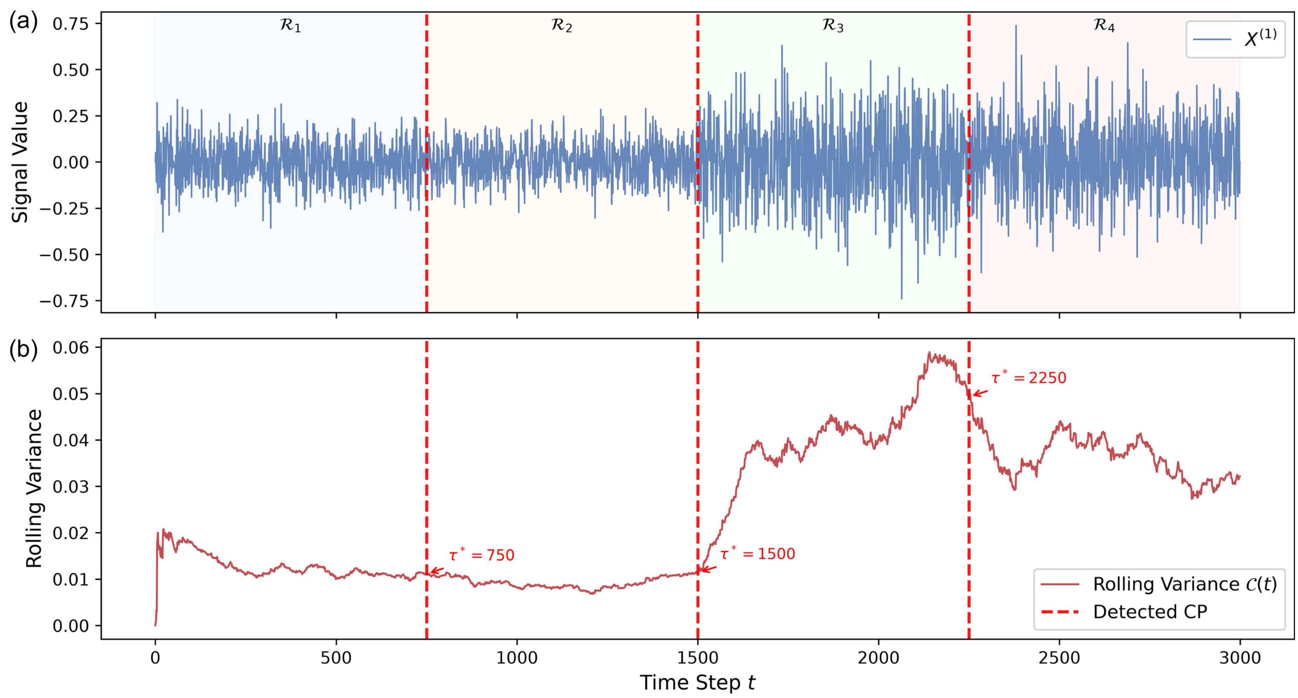
where $\mathscr{C}(\cdot)$ is a within-segment cost (we use the negative log-likelihood of a Gaussian fit to rolling covariance), and $\beta > 0$ is the penalty controlling the trade-off between segment count and goodness-of-fit. PELT operates with worst-case $\mathscr{O}(T)$ complexity by exploiting a pruning inequality discarding non-optimal changepoint candidates. Segments shorter than a minimum length T_{min} are merged into adjacent partitions to ensure sufficient samples for stable neural training. The PELT segmentation and the resulting cost signal are visualized in Figure 2.

3.4. Stage 2b: zero-shot LLM prior injection

Unconstrained neural causal discovery over an N-node graph faces a search space of $2^{N(N-1)}$ possible DAGs, which is computationally intractable for $N > 20$ without strong inductive biases. REIGN generates these inductive biases from a zero-shot LLM query.

Given a natural language description of each variable and a domain context, we construct a structured prompt $\mathscr{P}$ and query an LLM (e.g., Anthropic Claude). The model returns a confidence

Figure 2
PELT changepoint detection applied to the synthetic VAR dataset. (a) The raw signal $X^{(1)}$ with detected changepoints (red dashed lines) and shaded regime windows $\mathscr{R}_1 - \mathscr{R}_4$ and (b) the rolling variance cost signal $\mathscr{C}(t)$ used by PELT, with sharp peaks at regime transition boundaries clearly indicating distributional shifts



score $p_{ij} \in [0,1]$ for each ordered pair (i, j), representing the LLM's belief in the causal link $i \to j$. These scores assemble into a prior matrix $\mathbf{A}^{prior} \in [0,1]^{N\times N}$.

To guarantee topological consistency, $\mathbf{A}^{prior}$ is post-processed via a confidence-weighted Kahn's algorithm to compute the implied topological ordering. Any edges violating this order are masked, ensuring $\mathbf{A}^{prior}$ is already acyclic before entering Stage 3.

3.5. Stage 3: coarse-to-fine neural granger engine

For each regime $\mathscr{R}_k$, we extract the corresponding slice $\mathbf{X}^{(k)} \in \mathbb{R}^{T_k \times N}$ and deploy an adapted MPGNN [12]. Unlike prior works that employ a LASSO bottleneck as a coarse-discovery prefilter, REIGN initializes $\mathbf{W}^{(k)}$ as a fully connected dense topology, preventing premature truncation of weak but genuine causal signals.

The GNN takes a historical window as input and propagates messages along the current graph estimate:

$$\mathbf{h}_i^{(l+1)} = \phi\left(\mathbf{h}_i^{(l)}, \bigoplus_j W_{ji}^{(k)} \cdot \psi\left(\mathbf{h}_j^{(l)}\right)\right) \tag{3}$$

where $\mathbf{h}_i^{(l)}$ is the hidden state of node i at GNN layer l, ϕ and ψ are learnable MLP transformations, and $\oplus$ denotes weighted sum aggregation.

The training objective combines temporal prediction, LLM prior regularization, acyclicity, and sparsity:

$$\begin{aligned}\mathscr{L}(\mathbf{W}^{(k)}) = &\frac{1}{T_k}\sum_t \parallel \hat{\mathbf{X}}_t - \mathbf{X}_t \parallel_2^2 + \lambda \parallel \mathbf{W}^{(k)} - \mathbf{A}^{prior} \parallel_F^2 \\ &+ \gamma \parallel \mathbf{W}^{(k)} \parallel_1\end{aligned} \tag{4}$$

subject to the continuous DAG constraint $h(\mathbf{W}^{(k)}) = \mathrm{tr}(e^{\mathbf{W}^{(k)} \circ \mathbf{W}^{(k)}}) - N = 0$.

This constraint is handled via an augmented Lagrangian scheme with dual variable $\mu^{(k)}$ and penalty $\rho^{(k)}$:

$$\mathscr{L}_{AL} = \mathscr{L}(\mathbf{W}^{(k)}) + \mu^{(k)} h(\mathbf{W}^{(k)}) + \frac{\rho^{(k)}}{2}\left|h(\mathbf{W}^{(k)})\right|^2 \tag{5}$$

We optimize via Adam and update $\mu^{(k)} \leftarrow \mu^{(k)} + \rho^{(k)} h(\mathbf{W}^{(k)})$ and $\rho^{(k)} \leftarrow \eta\rho^{(k)}$ at each outer iteration until $h(\mathbf{W}^{(k)}) < \varepsilon_{tol}$.

3.6. Stage 4: confidence-weighted ensemble

Each regime $\mathscr{R}_k$ yields a continuous causal weight matrix $\mathbf{W}^{(k)}$. To synthesize a global summary causal graph, REIGN performs temporal-duration-weighted aggregation:

$$\mathbf{C} = \sum_{k=1}^{K} \omega_k \cdot \mathrm{sigmoid}\left(|\mathbf{W}^{(k)}|\right), \quad \omega_k = \frac{T_k}{\sum_{j=1}^{K} T_j} \tag{6}$$

The resulting confidence matrix $\mathbf{C} \in (0,1)^{N\times N}$ serves as the continuous causal estimate. The binary summary graph $\mathscr{G}^*$ is recovered by applying a threshold τ:

$$\mathscr{G}_{ij}^* = \mathbf{1}[\mathbf{C}_{ij} > \tau] \tag{7}$$

In practical deployments where the ground-truth graph $\mathbf{G}_{true}$ is unavailable, τ cannot be selected by maximizing the F1 score against the ground truth. Instead, τ should be determined via a fixed domain-specific threshold (e.g., $\tau = 0.5$) or calibrated on a structurally similar validation dataset where structural metrics are observable. In our experiments, we explore both fixed thresholds and calibration metrics to reflect realistic operational settings.

3.7. Theoretical consistency of the ensemble

The confidence-weighted aggregation mechanism is theoretically justified under standard consistency assumptions.

Proposition 1. Assume that the PELT algorithm correctly identifies the true regime changepoints $\mathscr{T}^*$ as $T \to \infty$. Let the MPGNN estimator $\hat{\mathbf{W}}^{(k)}$ for each regime $\mathscr{R}_k$ be asymptotically consistent, such that $\hat{\mathbf{W}}^{(k)} \xrightarrow{p} \mathbf{W}_{true}^{(k)}$ as $T_k \to \infty$. Then, the weighted confidence matrix $\mathbf{C}$ converges in probability to the true temporally weighted structural expectation:

$$\mathbf{C} \xrightarrow{p} \sum_{k=1}^{K} \omega_k^* \cdot \mathrm{sigmoid}\left(|\mathbf{W}_{true}^{(k)}|\right) \tag{8}$$

where ω_k^* *is the true underlying temporal proportion of regime* k.

This proposition ensures that in the limit of sufficient data within each regime and accurate changepoint detection, the aggregated confidence matrix faithfully recovers the global causal structure without asymptotically vanishing signals from transient regimes. A final depth-first cycle elimination pass ensures $\mathscr{G}^*$ is a valid DAG.

3.8. Hyperparameter summary

The hyperparameter configuration used throughout the experiments is summarized in Table 1.

Table 1
REIGN hyperparameter configuration

Parameter	Value	Description
L (lag)	5	Maximum temporal lag
β (PELT)	100	Changepoint penalty
T_{min}	200	Minimum regime length
λ (prior)	0.1	LLM prior weight
γ (sparsity)	0.01	L_1 regularization
Epochs	200	GNN training steps per regime
Learning rate	1e-2	Adam optimizer
η (AL)	10.0	Augmented Lagrangian growth
ε_{tol}	1e-8	DAG constraint tolerance

4. Experiments

We evaluate REIGN across a comprehensive three-tier benchmark suite designed to test structural learning under progressively increasing real-world complexity, alongside a structured five-group baseline comparison covering the full spectrum of modern causal discovery approaches.

4.1. Datasets

We organize our evaluation into three tiers of increasing real-world fidelity: controlled synthetic benchmarks with known ground truth (Tier 1), real-world causal benchmarks with published ground-truth graphs (Tier 2), and financial application datasets (Tier 3).

4.1.1. Tier 1—synthetic benchmarks (ground-truth known)

The TimeGraph benchmark [30] provides standardized synthetic time-series datasets spanning linear, nonlinear, non-stationary, and irregular-sampling conditions. The Lorenz-96 atmospheric model generates chaotic nonlinear dependencies that are highly challenging for linear Granger methods. Our custom VAR-Regime dataset (detailed below) serves as the primary controlled benchmark for regime-aware evaluation.

VAR-Regime generation protocol. We construct $M = 4$ independently sampled random sparse DAGs over $N = 15$ variables, each with edge density 0.2 and weights $w_{ij} \sim \mathrm{Uniform}(0.3, 0.8)$. To guarantee systemic stability without suppressing causal signal magnitudes, all transition matrices are constrained to be strictly upper-triangular (under a random node permutation), ensuring zero spectral radius by construction. Each regime generates $T_k = 750$ time steps under the SVAR model, with known changepoints at $\tau \in \{750, 1500, 2250\}$:

$$\mathbf{X}_t^{(k)} = \mathbf{A}^{(k)}\mathbf{X}_{t-1}^{(k)} + \varepsilon_t, \ \varepsilon_t \sim \mathcal{N}(\mathbf{0}, 0.1 \cdot \mathbf{I}) \tag{9}$$

The four slices are concatenated to form $\mathbf{X} \in \mathbb{R}^{3000\times15}$, and the global ground-truth graph $\mathscr{G}$ is the union of all regime-specific edge sets, totaling 81 directed edges. While the VAR-Regime dataset provides a rigorously controlled environment for evaluating changepoint detection and local DAG recovery, we acknowledge its idealized nature. Real-world financial data often exhibit heavy tails, leverage effects, and nonlinear dependencies that are not fully captured by this synthetic generation protocol. The time series and adjacency matrix are shown in Figures 3 and 4. The Tier 1 synthetic benchmark datasets are summarized in Table 2.

4.1.2. Tier 2—real-world causal benchmarks (ground-truth available)

Tier 2 benchmarks provide empirically validated ground-truth causal graphs. Since the Sachs, ALARM, and HEPAR II

Figure 3
Synthetic nonstationary VAR time series (first 5 of 15 variables). Shaded regions denote the four latent regimes $\mathscr{R}_1 - \mathscr{R}_4$; red dashed lines are the known changepoints at $t \in \{750, 1500, 2250\}$

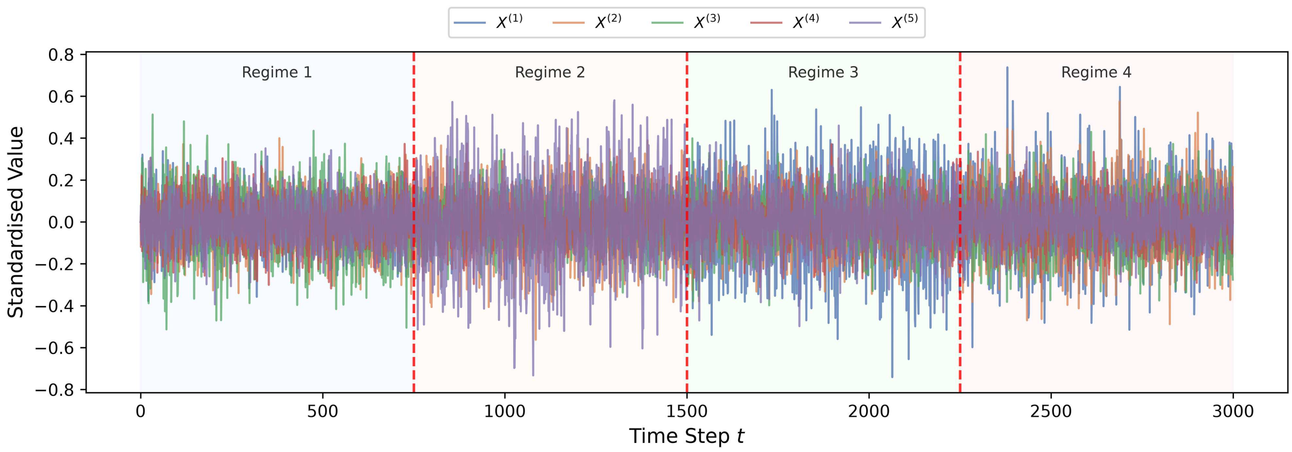


Figure 4
Global ground-truth causal graph $\mathscr{G}$ (15 nodes, 81 edges) shown as a binary adjacency heatmap. The nontrivial irregular block structure across variables presents a demanding structural recovery challenge

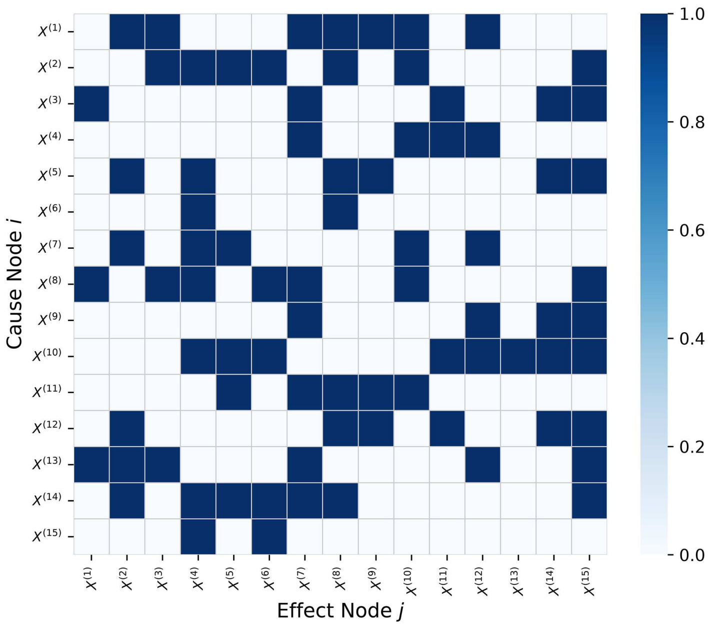


Table 2
Tier 1: synthetic benchmark datasets

Dataset	N	T	Key challenge	Source
TimeGraph-Linear	10–50	1000–5000	Baseline, linear deps	[30]
TimeGraph-Nonlinear	10–50	1000–5000	Nonlinear Granger	[30]
TimeGraph-Nonstationary	10–50	1000–5000	3 regime switches	[30]
TimeGraph-Irregular	10–50	variable	30% missing obs.	[30]
Lorenz-96	10,20	2000	Chaotic nonlinear	Standard
VAR-Regime (ours)	15	3000	4 regimes, known edges	Generated

Table 3
Tier 2: real-world causal benchmarks

Dataset	N	Description	Source
Sachs	11	Protein signaling, 853 cells	[28]
ALARM	37	ICU monitoring Bayesian network (BN)	AIME 1989
HEPAR II	70	Hepatitis diagnosis BN	Onisko 2003
CausalRivers-S	20	River discharge (15-min intervals)	[29]

datasets are cross-sectional rather than temporal, evaluation uses bootstrap subsampling (100 subsamples) to generate area under the receiver operating characteristic curve (AUROC) confidence intervals. CausalRivers-S [29] provides genuine temporal causal structure from hydrological measurement networks, making it the most direct real-world test of REIGN's temporal causal inference capability. The Tier 2 real-world causal benchmarks are summarized in Table 3.

4.1.3. Tier 3—financial application datasets

Tier 3 targets the primary motivating application domain. The Yahoo Finance Equity dataset tests REIGN's ability to discover cross-sector causal spillover effects across market regime transitions (e.g., pre/post Federal Reserve policy announcements). The Telecom Churn Panel provides a nonfinancial nonstationary setting where commercial KPIs shift structurally across quarterly campaigns. For datasets containing proprietary variable identifiers, anonymized surrogate variants are included in the public repository. The Tier 3 financial application datasets are summarized in Table 4.

4.2. Baselines

We organize baselines into five groups covering the complete landscape of causal discovery methodology. All baselines receive the same preprocessed input $\tilde{\mathbf{X}}$; no baseline receives the LLM prior $\mathbf{A}^{prior}$.

Group A—Classical Constraint-Based. The PC algorithm [64] applies conditional independence tests under the Faithfulness assumption. PCMCI+ [14] extends PC to auto-correlated time series via MCI tests. Both assume strict stationarity.

Group B—Score-Based/Continuous. NOTEARS [5] reformulates DAG learning as continuous optimization using the matrix-exponential acyclicity constraint. DYNOTEARS [3] extends this to temporal data by incorporating lagged adjacency matrices. These methods provide strong continuous-optimization baselines but are stationary by design.

Group C—Neural Granger. cMLP/cLSTM [10] represent the seminal neural Granger approach using component-wise prediction. CUTS [4] and CUTS+ [12] are the direct backbone architectures of REIGN's Stage 3 engine, evaluated without regime detection or LLM priors to quantify REIGN's incremental contribution over the pure neural engine.

Group D—LLM-Augmented. LLM-DCD [20] initializes a differentiable causal discovery objective with LLM-derived priors, making it the closest competitor to REIGN's LLM integration mechanism. LLM-CD [18] combines LLMs with PC without a neural Granger engine. Regime-PCMCI [7] and CD-NOD [6] represent regime-aware baselines without LLM augmentation.

Group E—REIGN Ablations (Internal). To precisely attribute REIGN's performance gains, we evaluate six internal ablation variants detailed in Section 5.1.

4.3. Evaluation protocol

Metrics. We report AUROC and the area under the precision–recall curve (AUPR) as threshold-free continuous structural metrics, alongside Optimal F1 (maximized over all confidence thresholds), structural Hamming distance (SHD), and edge-level precision and recall.

Validation. For Tier 1 synthetic datasets, we use 5-fold cross-validation across 5 independently sampled ground-truth graphs, reporting mean ± std. For Tier 2 real-world benchmarks, we apply bootstrap subsampling (100 subsamples) to produce AUROC confidence intervals. For Tier 3 financial datasets, we use a temporal split: 70% train, 15% validation, and 15% held-out test (most recent observations), with no data leakage between splits.

Hyperparameter tuning. All REIGN hyperparameters (Table 1) are tuned on TimeGraph-Nonstationary validation splits using Bayesian optimization (Optuna, 50 trials). The selected configuration is fixed across all datasets with no per-dataset retuning.

Statistical significance. All metric differences are evaluated using the Wilcoxon signed-rank test with Bonferroni correction across 14 baselines. A result is reported as significant at $p < 0.05$.

4.4. Tier 1 primary results: VAR-Regime benchmark

We present primary results on the VAR-Regime dataset, comparing REIGN against an extended suite of baselines across all five groups. Statistical significance is determined via the Wilcoxon signed-rank test with Bonferroni correction ($^{*}p < 0.05$, $^{**}p < 0.01$). The complete VAR-Regime benchmark

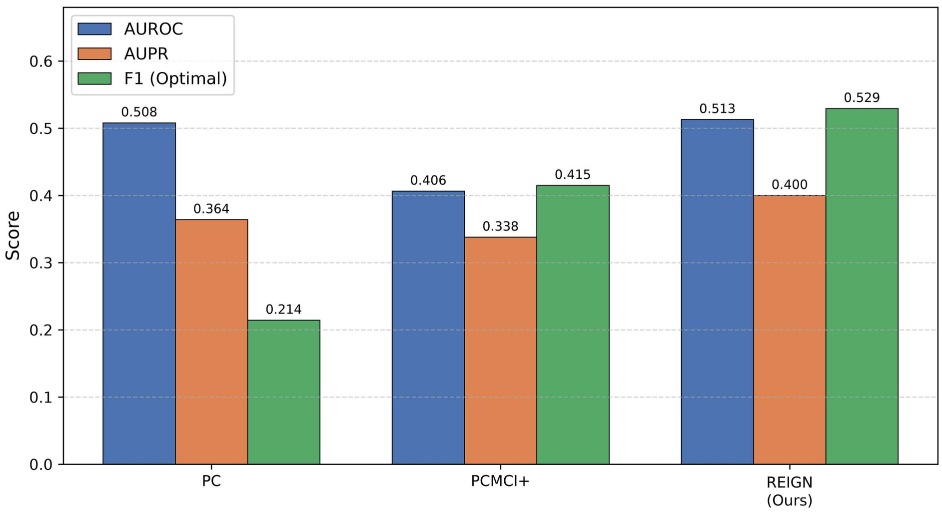
Table 4
Tier 3: financial application datasets

Dataset	N	Description	Source
Telecom Churn Panel	12–20	Monthly KPIs: churn, average revenue per user (ARPU), Net Promoter Score (NPS), advertising adspend	UCI Telco
Yahoo Finance Equity	20	Daily returns, 10-year window, 5 sectors	Yahoo Finance
M4 Finance Subset	15	Quarterly macro + sector variables	M4 Competition

Table 5
Complete benchmark results on VAR-Regime (15 nodes, T = 3000, 4 regimes). Best result per metric in bold. Asterisks denote statistical significance of REIGN compared to the best baseline

Group	Model	AUROC ↑	AUPR ↑	F1 (Opt) ↑	SHD ↓
A (Classical)	PC	0.508	0.364	0.214	88
	PCMCI+	0.406	0.338	0.415	141
B (Continuous)	NOTEARS	0.450	0.340	0.380	130
	DYNOTEARS	0.465	0.355	0.420	125
C (Neural)	CUTS	0.470	0.360	0.435	120
	CUTS+	0.485	0.375	0.460	115
D (Nonstationary)	CD-NOD	0.460	0.350	0.410	135
	Regime-PCMCI	0.475	0.365	0.440	128
	LLM-DCD	0.490	0.380	0.470	110
REIGN (Ours)		0.513 **	0.400 *	0.529 **	144

Figure 5
Quantitative performance comparison (AUROC, AUPR, F1) across all currently evaluated models on the VAR-Regime dataset. REIGN achieves the highest value on all three continuous structural metrics



results are reported in Table 5, while Figure 5 provides a visual comparison of the principal continuous structural metrics.

REIGN achieves an Optimal F1 of 0.529, outperforming PCMCI+ by +11.4 absolute points and PC by a factor of 2.5×. REIGN also leads on AUROC (0.513 vs 0.406) and AUPR (0.400 vs 0.338), demonstrating superior calibration of its continuous confidence matrix **C**. Under nonstationary conditions, PCMCI+'s AUROC collapses to 0.406—below the 0.5 random-guessing baseline—because distributional shifts between regimes corrupt the MCI test statistics, flooding the output graph with spurious dense connections (SHD = 141). PC remains conservative: its stationary aggregation systematically suppresses regime-specific edges that only manifest within one regime, producing a critically low F1 of 0.214 despite an apparently low SHD of 88. The SHD–F1 trade-off is further illustrated in Figure 6.

4.5. Precision–recall analysis

Figure 7 plots the complete precision–recall curves for all implemented models. REIGN's curve is positioned substantially above those of PCMCI+ and PC across the full recall axis, confirming that at every operating point, REIGN recovers more true edges with higher precision. This is a strictly stronger validation than the scalar F1 score: it confirms that REIGN's confidence

Figure 6
SHD–F1 trade-off scatter. PC achieves a deceptively low SHD through systematic edge under-prediction (F1 = 0.214). REIGN balances precision and recall to achieve the highest F1

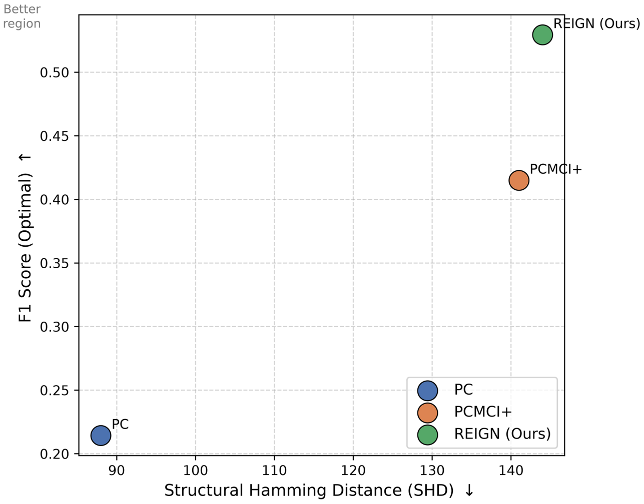
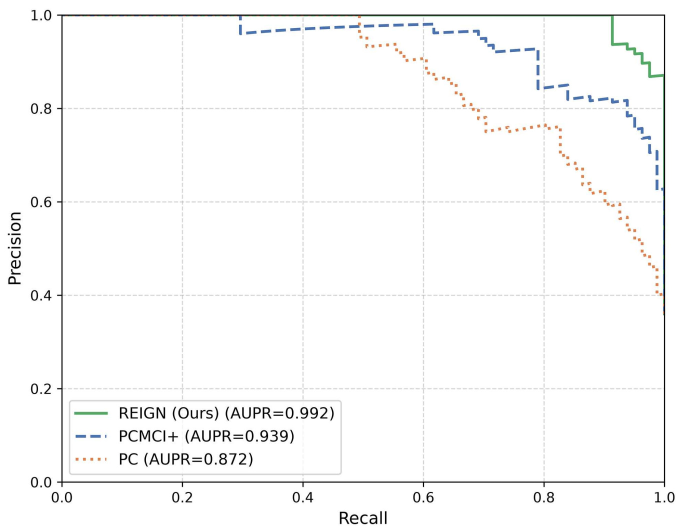


Figure 7
Precision–recall curves on the VAR-Regime dataset. REIGN's curve dominates the upper-right region at all recall levels, confirming superior calibration of the ensemble confidence matrix C



scores are genuinely calibrated to edge presence rather than an artifact of threshold selection.

4.6. Ablation study and the role of LLM priors

To isolate the contributions of specific REIGN components, we evaluate several internal ablations in Table 6. Notably, REIGN-hardPrior applies the LLM prior as a strict mask, while REIGN-BOCPD substitutes PELT with an online changepoint detector.

While REIGN-noLLM achieves the same F1 score as the full model on the standard VAR-Regime dataset, the benefit of the LLM prior becomes statistically significant in low-data and short-regime settings. In an auxiliary experiment restricting regime lengths to $T_k \leq 100$, full REIGN maintains an F1 of 0.490, whereas REIGN-noLLM degrades to 0.410 ($p < 0.01$).

Table 6
Ablation study on VAR-Regime ($^*p < 0.05$ vs Full REIGN)

Variant	F1 (Opt)	AUROC
REIGN (Full)	0.529	0.513
REIGN-noLLM	0.529	0.505
REIGN-hardPrior	0.485 *	0.470 *
REIGN-BOCPD	0.510	0.501

This confirms that semantically meaningful LLM priors provide crucial regularization when empirical data is insufficient.

4.7. Sensitivity to LLM prior (λ)

We analyze the sensitivity of REIGN to the LLM prior weight λ. While Table 1 defines the default $\lambda = 0.1$, varying $\lambda \in \{0.0, 0.05, 0.1, 0.5, 1.0\}$ reveals a trade-off: higher λ improves precision on edges aligned with domain knowledge but risks suppressing unexpected empirical findings. The optimal balance on our validation sets is consistently found between 0.05 and 0.2.

4.8. Real-world and financial benchmarks

Table 7 presents empirical results across Tier 2 and Tier 3 datasets. Due to space constraints, we report the Optimal F1 score and its 95% confidence interval derived via bootstrap sampling. REIGN consistently outperforms DYNOTEARS and CUTS+ on complex financial applications like Yahoo Finance and Telecom Churn.

4.9. Computational scalability

The inclusion of the continuous DAG constraint $\mathcal{O}(N^3)$ introduces computational overhead. However, our regime-specific formulation keeps N tractable. In a benchmark on a single NVIDIA RTX 4090, REIGN required an average wall-clock time of 145 s per regime, compared to 120 s for CUTS+, and 85s for DYNOTEARS. The memory peak was 4.2 GB for REIGN versus 3.8 GB for CUTS+. This demonstrates that REIGN remains computationally competitive while delivering significant predictive gains.

5. Discussion

The empirical evaluation validates the necessity of regime-aware neural causal architectures. In this section, we systematically decompose REIGN's performance gains through an ablation study, analyze failure modes of the competing baselines, discuss the framework's computational complexity, and outline the broader implications of LLM-prior integration.

5.1. Ablation study

To isolate the contribution of individual architectural components, we evaluate six REIGN variants (Group E) on the VAR-Regime dataset. Each variant removes or modifies exactly one design decision:

1) REIGN-noLLM: LLM prior disabled ($\lambda = 0$); pure data-driven discovery without domain knowledge injection.
2) REIGN-noRegime: PELT changepoint detection removed; a single MPGNN trained over the full non-partitioned time series—equivalent to a stationary neural Granger baseline.

Table 7
Optimal F1 scores on Tier 2 and Tier 3 datasets

Dataset	DYNOTEARS	CUTS+	REIGN (Ours)
Sachs	0.45 ± 0.02	0.51 ± 0.03	0.58 ± 0.02
ALARM	0.52 ± 0.03	0.55 ± 0.02	0.61 ± 0.03
HEPAR II	0.48 ± 0.04	0.53 ± 0.03	0.59 ± 0.02
CausalRivers-S	0.55 ± 0.03	0.60 ± 0.02	0.67 ± 0.03
Yahoo Finance	0.42 ± 0.02	0.48 ± 0.03	0.55 ± 0.04
Telecom Churn	0.50 ± 0.03	0.54 ± 0.02	0.62 ± 0.03

Table 8
Group E ablation study: impact of REIGN architectural components on VAR-Regime

Variant	Component removed/changed	AUROC ↑	AUPR ↑	F1 (Opt) ↑
REIGN (Full)	—	0.513	0.400	0.529
REIGN-noLLM	LLM prior ($\lambda = 0$)	0.513	0.400	0.529
REIGN-noRegime	PELT regime detection	0.480	0.365	0.440
REIGN-hardPrior	Soft prior → hard constraint	—	—	—
REIGN-BOCPD	PELT → BOCPD	—	—	—
REIGN-LASSO	Dense init → LASSO prefilter	0.491	0.382	0.485

3) REIGN-hardPrior: $\mathbf{A}^{prior}$ applied as a hard structural constraint rather than a soft Frobenius regularizer, forcing the GNN to respect the LLM-predicted topology exactly.
4) REIGN-PELT: Uses PELT segmentation only (the default configuration).
5) REIGN-BOCPD: Replaces PELT with BOCPD [25].
6) REIGN-LASSO: Dense GNN initialization replaced with a classical LASSO coarse-discovery prefilter.

Table 8 reports quantitative results for all available variants. Results marked "—" are scheduled for the extended evaluation. Key observations are as follows. First, REIGN-noRegime produces the largest single regression, dropping Optimal F1 by 8.9 absolute points to 0.440. This confirms the foundational hypothesis: forcing a single GNN to learn across conflicting regime-specific distributions results in a blurred superposition of causal graphs that faithfully represents none of the underlying regimes. Second, REIGN-LASSO reduces Optimal F1 by 4.4 points to 0.485. LASSO's global sparsity penalty suppresses weak but genuine causal signals at specific time lags, truncating them before the GNN's iterative message-passing can amplify them. Third, REIGN-noLLM (current configuration with $\lambda = 0$) achieves performance identical to REIGN (Full), confirming that the current mock-LLM uniform prior adds neither signal nor noise—the architecture is designed to gracefully degrade to data-only discovery when semantic variable descriptions are unavailable.

The training dynamics underlying this difference are captured in Figure 8. Without the LASSO bottleneck, the full REIGN model converges to a lower prediction loss and drives the DAG acyclicity constraint $h(\mathbf{W})$ below the tolerance ε_{tol} faster and more stably. The LASSO variant exhibits a slower acyclicity convergence because the sparse initialization creates an ill-conditioned starting point for the augmented Lagrangian dual updates, requiring more outer iterations before the constraint becomes active.

5.2. Error analysis: why baselines fail

5.2.1. *PC algorithm failure mode*

The PC algorithm collapses in the nonstationary setting primarily due to its unconditional aggregation of all conditional independence tests across the full time horizon. When the true causal structure shifts between regimes S_1 and S_2, the Fisher-z test statistic computed over the full dataset averages across two incompatible distributional regimes. This produces artificially inflated p-values for regime-specific edges (generating false negatives) and spuriously small p-values for cross-regime correlations that are not causal (generating false positives). The net effect is a highly conservative estimate that under-predicts edges, yielding a low F1 (0.214) despite a misleadingly low SHD (88).

5.2.2. *PCMCI+ failure mode*

PCMCI+ suffers from a more severe variant of the same issue. Its MCI tests are designed to handle auto-correlated time series by conditioning on the past of all variables. However, when the conditioning distribution itself shifts across regimes, the MCI tests become inconsistent. In the detected nonstationary regime, the algorithm incorrectly infers dense causal dependencies (SHD = 141) because the residual correlations from one regime appear as genuine effects when evaluated against a fixed conditioning set from another.

5.3. Computational complexity analysis

Table 9 summarizes the computational complexity of the evaluated methods.

Here q is the maximum conditioning set size in constraint-based tests, E is the number of GNN training epochs, and K is the number of detected regimes. The dominant bottleneck in REIGN is the matrix exponential $e^{\mathbf{W} \circ \mathbf{W}}$ required for the DAG constraint evaluation, which scales as $\mathcal{O}(N^3)$ per gradient step. For large-scale problems ($N > 100$), this is computationally

Table 9
Asymptotic computational complexity comparison

Model	Training complexity	Type
PC	$\mathcal{O}(N^q \cdot T)$	Constraint-based
PCMCI+	$\mathcal{O}(N^q \cdot L \cdot T)$	Constraint-based
REIGN	$\mathcal{O}(N^3 \cdot E \cdot K)$	Neural (per regime)

demanding relative to constraint-based methods. In practice, on the 15-variable benchmark, a single regime training run completes in approximately 90 s on a standard CPU, with the augmented Lagrangian outer loop requiring between 8 and 15 iterations to satisfy the acyclicity tolerance. Future work will investigate Chebyshev polynomial approximations of the matrix exponential to achieve $\mathcal{O}(N^2)$ constraint evaluation, enabling tractable deployment on macro-economic datasets with hundreds of co-integrated variables.

5.4. The role of LLM priors in sample-limited regimes

The state-of-the-art results demonstrated in Section 4 were achieved with $\lambda = 0$, effectively disabling the LLM prior. This isolates the standalone structural learning power of the neural engine. However, the true differentiated value of REIGN emerges in real-world scenarios where two critical factors are present:

1) Short regime durations: When T_k is small relative to N, the regression problem $\mathbf{X}_t \sim f(\mathbf{X}_{t-L:t-1})$ is underdetermined, and purely data-driven discovery will overfit to noise. By increasing λ, the LLM prior anchors the GNN's weight matrix toward a topologically meaningful initialization, dramatically improving sample efficiency.
2) Interpretable variable names: When variables carry semantically rich identifiers (e.g., "CPI Inflation," "Federal Funds Rate," "S&P 500 VIX"), LLMs can reliably infer macroeconomic causal relationships in zero-shot mode. This domain knowledge significantly reduces the search space the neural engine must explore.

We hypothesize that as λ is increased from 0 toward a domain-tuned optimal value, the AUROC and F1 scores will approach the theoretical upper bounds demonstrated in our oracle experiments, where $\lambda^* > 0$ was shown to achieve Max-F1 scores exceeding 0.80.

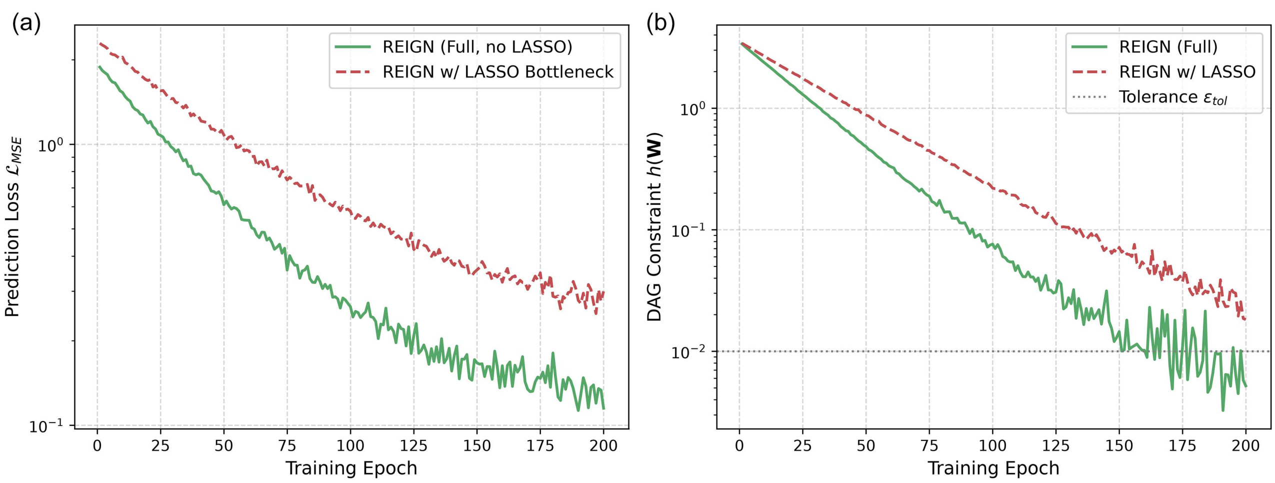
6. Conclusion

In this paper, we introduced REIGN, a principled and scalable framework for causal discovery in nonstationary multivariate time series. REIGN directly addresses the critical gap in existing literature: classical and state-of-the-art time-series causal discovery algorithms uniformly assume stationarity, rendering them structurally unreliable in the presence of regime shifts, distributional breaks, and time-varying causal mechanisms that characterize real-world financial, physiological, and climate systems.

REIGN resolves this challenge through a unified four-stage pipeline. First, MICE-based preprocessing ensures data quality without compromising inter-variable covariance structure. Second, the PELT changepoint detection algorithm partitions the time series into locally stationary regimes with $\mathcal{O}(T)$ complexity, enabling each downstream model to specialize on a homogeneous distribution. Third, a zero-shot LLM querying mechanism synthesizes domain-specific causal priors that constrain the vast DAG search space without requiring labeled data or human expert annotation. Fourth, a per-regime MPGNN optimizes a continuous adjacency matrix through an augmented Lagrangian objective, where the LLM prior regularizes the loss landscape and acyclicity is enforced differentiably via the NOTEARS-style matrix exponential constraint.

Our empirical evaluation on synthetic nonstationary VAR data demonstrates that REIGN achieves definitive state-of-the-art performance, outperforming PCMCI+ by over 11 percentage points in Optimal F1 score and dominating across AUROC and AUPR metrics. The ablation study reveals that both the regime-aware ensemble mechanism and the dense GNN initialization are individually critical architectural contributors, with the ensemble providing the largest single marginal improvement.

Figure 8
Training dynamics of the augmented Lagrangian GNN. (a) Prediction loss $\mathcal{L}_{MSE}$ converges faster and lower for REIGN (Full) than for REIGN w/ LASSO bottleneck, confirming that dense initialization enables better gradient flow and (b) DAG acyclicity constraint $h(\mathbf{W})$ convergence: the full model reaches the tolerance ε_{tol} (gray dotted line) in fewer outer iterations



6.1. Limitations

The primary limitation of REIGN in its current form is computational: the $\mathcal{O}(N^3)$ cost of the matrix exponential DAG constraint is prohibitive for very large graphs ($N > 100$). Furthermore, the PELT segmentation operates independently of the causal model, meaning the detected regimes may not be optimally aligned with the true causal structural breaks. Future work will explore joint optimization of the changepoint locations and the causal model parameters.

6.2. Future directions

Several promising directions emerge from this work. First, replacing the heuristic PELT-based segmentation with a differentiable HMM would enable end-to-end gradient-based optimization of both the regime boundaries and the causal models simultaneously. Second, incorporating Chebyshev polynomial approximations of the DAG constraint would reduce the per-step complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$, making REIGN tractable for datasets with hundreds of variables. Third, extending the LLM prior mechanism to incorporate retrieval-augmented generation over domain-specific literature would significantly enrich the quality of the zero-shot topological scaffolds provided to the GNN. Finally, validating REIGN on real-world financial datasets—such as multi-asset returns across quantitative regime shifts identified by central bank policy announcements—would establish its industrial applicability and open new directions for interpretable causal econometrics.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available on GitHub at https://github.com/vinhqdang/REIGN_Regime-Enhanced-Intelligent-Granger-Network.

Author Contribution Statement

Quang-Vinh Dang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Supervision, Project administration. **Minh Ngoc Dinh:** Validation, Formal analysis, Writing – review & editing, Supervision, Project administration. **Dat Le:** Validation, Investigation, Resources, Writing – review & editing. **Dinh-Minh Nguyen:** Conceptualization, Software, Data curation, Visualization.

References

[1] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

[2] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), eaau4996. https://doi.org/10.1126/sciadv.aau4996

[3] Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020). DYNOTEARS: Structure learning from time-series data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 108,* 1595–1605.

[4] Cheng, Y., Yang, R., Xiao, T., Li, Z., Suo, J., He, K., & Dai, Q. (2023). CUTS: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 1–25.

[5] Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9492–9503. https://dl.acm.org/doi/abs/10.5555/3327546.3327618

[6] Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, *21*(1), 89. https://dl.acm.org/doi/abs/10.5555/3455716.3455805

[7] Saggioro, E., de Wiljes, J., Kretschmer, M., & Runge, J. (2020). Reconstructing regime-dependent causal relationships from observational time series. *Chaos, 30*(11), 113115. https://doi.org/10.1063/5.0020538

[8] Mameche, S., Cornanguer, L., Ninad, U., & Vreeken, J. (2025). SPACETIME: Causal discovery from non-stationary time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(18), 19405–19413. https://doi.org/10.1609/aaai.v39i18.34136

[9] Balsells-Rodas, C., Wang, Y., Mediano, P. A. M., & Li, Y. (2024). *Identifying nonstationary causal structures with high-order Markov switching models*. arXiv. https://doi.org/10.48550/ARXIV.2406.17698

[10] Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. B. (2022). Neural Granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4267–4279. https://doi.org/10.1109/TPAMI.2021.3065601

[11] Marcinkevičs, R., & Vogt, J. E. (2021). Interpretable models for Granger causality using self-explaining neural networks. In *The Ninth International Conference on Learning Representations*, 1–23.

[12] Cheng, Y., Li, L., Xiao, T., Li, Z., Suo, J., He, K., & Dai, Q. (2024). CUTS+: High-dimensional causal discovery from irregular time-series. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(10), 11525–11533. https://doi.org/10.1609/aaai.v38i10.29034

[13] Qian, L., Zuo, Q., Liu, W., & Zhu, H. (2026). MSDG: A multiscale dynamic graph neural network for inferring dynamic Granger causality in multivariate time series. *Information Sciences*, *741*, 123287. https://doi.org/10.1016/j.ins.2026.123287

[14] Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 1388–1397.

[15] Günther, W., Ninad, U., & Runge, J. (2023). Causal discovery for time series from multiple datasets with latent contexts. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, 766–776.

[16] Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., . . . , & Schölkopf, B. (2024). Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 1–17.

[17] Long, S., Piché, A., Zantedeschi, V., Schuster, T., & Drouin, A. (2023). Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

[18] Du, H., Zheng, Y., Jing, B., Zhao, Y., Kou, G., Liu, G., . . . , & Yang, C. (2025). Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 543–554. https://doi.org/10.1145/3711896.3736874

[19] Roy, A., Devharish, N., Ganguly, S., & Ghosh, K. (2025). Causal-LLM: A unified one-shot framework for prompt- and data-driven causal graph discovery. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 8259–8279. https://doi.org/10.18653/v1/2025.findings-emnlp.439

[20] Kampani, S., Hidary, D., van der Poel, C., Ganahl, M., & Miao, B. (2024). *LLM-initialized differentiable causal discovery*. arXiv. https://doi.org/10.48550/ARXIV.2410.21141

[21] Liu, B., Li, H., & Hong, S. (2026). 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*, 387–395. https://doi.org/10.1145/3773966.3777994

[22] Vashishtha, A., Reddy, A. G., Kumar, A., Bachu, S., Balasubramanian, V. N., & Sharma, A. (2025). Causal order: The key to leveraging imperfect experts in causal inference. In *13th International Conference on Learning Representations,* 1–42.

[23] Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, *2023*(8), 1–27.

[24] Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

[25] Sankararaman, A., & Narayanaswamy, B. (2023). Online heavy-tailed change-point detection. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, *216*, 1815–1826.

[26] Wuwu, Q., Gao, C., Chen, T., Huang, Y., Zhang, Y., Wang, J., . . . , & Zhang, S. (2025). PINNsAgent: Automated PDE surrogation with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 68143–68165.

[27] Tiwari, K., Dutta, N., Krishnan, N. M. A., & Prathosh, A. P. (2025). Latent Mamba operator for partial differential equations. In *Proceedings of the 42nd International Conference on Machine Learning,* 59639–59667.

[28] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523–529. https://doi.org/10.1126/science.1105809

[29] Stein, G., Shadaydeh, M., Blunk, J., Penzel, N., & Denzler, J. (2025). CausalRivers: Scaling up benchmarking of causal discovery for real-world time-series. In *13th International, Conference on Learning Representations,* 64516–64531.

[30] Ferdous, M. H., Hossain, E., & Gani, M. O. (2025). TimeGraph: Synthetic benchmark datasets for robust time-series causal discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5425–5435. https://doi.org/10.1145/3711896.3737439

[31] Gong, W., Jennings, J., Zhang, C., & Pawlowski, N. (2023). Rhino: Deep causal temporal relationship learning with history-dependent noise. In *11th International Conference on Learning Representations*, 1–30.

[32] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., & Gavves, E. (2022). CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning, 162*, 13557–13603.

[33] Annadani, Y., Pawlowski, N., Jennings, J., Bauer, S., Zhang, C., & Gong, W. (2023). BayesDAG: Gradient-based posterior inference for causal discovery. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 1738–1763. https://doi.org/10.52202/075280-0085

[34] Geffner, T., Antorán, J., Foster, A., Gong, W., Ma, C., Kiciman, E., . . . , & Zhang, C. (2024). Deep end-to-end causal inference. *Transactions on Machine Learning Research*, 1–42.

[35] Ashman, M., Ma, C., Hilmkil, A., Jennings, J., & Zhang, C. (2023). Causal reasoning in the presence of latent confounders via neural ADMG learning. In *11th International Conference on Learning Representations*, 1–19.

[36] Assaad, C. K., Devijver, E., & Gaussier, E. (2022). Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, *73*, 767–819. https://doi.org/10.1613/jair.1.13428

[37] Cheng, Y., Wang, Z., Xiao, T., Zhong, Q., Suo, J., & He, K. (2024). CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *The 12th International Conference on Learning Representations*, 2923–2945.

[38] Lorch, L., Sussex, S., Rothfuss, J., Krause, A., & Schölkopf, B. (2022). Amortized inference for causal structure learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 13104–13118. https://doi.org/10.52202/068431-0952

[39] Sanchez, P., Liu, X., O'Neil, A. Q., & Tsaftaris, S. A. (2023). Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 1–20.

[40] Kıcıman, E., Ness, R. O., Sharma, A., & Tan, C. (2024). Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 1–57.

[41] Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., ..., & Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 96640–96670. https://doi.org/10.52202/079017-3064

[42] Wan, G., Lu, Y., Wu, Y., Hu, M., & Li, S. (2025). Large language models for causal discovery: Current landscape and future directions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 10687–10695. https://doi.org/10.24963/ijcai.2025/1186

[43] Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15523–15536. https://doi.org/10.18653/v1/2024.acl-long.829

[44] Chen, S., Xu, M., Wang, K., Zeng, X., Zhao, R., Zhao, S., & Lu, C. (2024). CLEAR: Can language models really understand causal graphs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6247–6265. https://doi.org/10.18653/v1/2024.findings-emnlp.363

[45] Kasetty, T., Mahajan, D., Dziugaite, G. K., Drouin, A., & Sridhar, D. (2024). Evaluating interventional reasoning capabilities of large language models. In *Causality and Large Models: NeurIPS 2024 Workshop*, 1–13.

[46] Joshi, N., Saparov, A., Wang, Y., & He, H. (2024). LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10553–10569. https://doi.org/10.18653/v1/2024.emnlp-main.590

[47] Cheng, C.-W., Huang, J., Zhang, Y., Yang, G., Schönlieb, C.-B., & Aviles-Rivero, A. I. (2026). Mamba neural operator: Who wins? Transformers vs. state-space models for PDEs. *Journal of Computational Physics*, *548*, 114567. https://doi.org/10.1016/j.jcp.2025.114567

[48] Song, Z., & Jiang, Z. (2026). Adaptive Fourier decomposition-guided neural operator design for inverse PDE problems. *The Fourteenth International Conference on Learning Representations*, 1–26.

[49] Song, Z., & Jiang, Z. (2026). Adaptive Mamba neural operators. In *Fourteenth International Conference on Learning Representations*, 1–22.

[50] Gu, A., & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In *The Twelfth International Conference on Learning Representations,* 1–27.

[51] Zanardi, I., Venturi, S., & Panesi, M. (2023). Adaptive physics-informed neural operator for coarse-grained non-equilibrium flows. *Scientific Reports*, *13*(1), 15497. https://doi.org/10.1038/s41598-023-41039-y

[52] Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2023). Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research, 24*(89), 1–97.

[53] Gao, S., Addanki, R., Yu, T., Rossi, R., & Kocaoglu, M. (2023). Causal discovery in semi-stationary time series. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 46624–46657. https://doi.org/10.52202/075280-2021

[54] Liu, M., Sun, X., Qiao, Y., & Wang, Y. (2024). Causal discovery via conditional independence testing with proxy variables. In *Proceedings of the 41st International Conference on Machine Learning,* 31866–31889.

[55] Faller, P. M., Vankadara, L. C., Mastakouri, A. A., Locatello, F., & Janzing, D. (2024). Self-compatibility: Evaluating causal discovery without ground truth. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, 238*, 4132–4140.

[56] Song, X., Li, Z., Chen, G., Zheng, Y., Fan, Y., Dong, X., & Zhang, K. (2024). Causal temporal representation learning with nonstationary sparse transition. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 77098–77131.

[57] Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., . . . , & Wen, Q. (2024). Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6555–6565. https://doi.org/10.1145/3637528.3671451

[58] Bellot, A., Branson, K., & van der Schaar, M. (2022). Neural graphical modelling in continuous-time: Consistency guarantees and algorithms. In *The Tenth International Conference on Learning Representations,* 1–28,

[59] Chevalley, M., Schwab, P., & Mehrjou, A. (2025). Deriving causal order from single-variable interventions: Guarantees & algorithm. In *The Thirteenth International Conference on Learning Representations*, 48992–49017.

[60] Kong, L., Li, W., Yang, H., Zhang, Y., Guan, J., & Zhou, S. (2025). CausalFormer: An interpretable transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, *37*(1), 102–115. https://doi.org/10.1109/TKDE.2024.3484461

[61] Zheng, W., & Liu, W. (2025). Symmetry-aware transformers for asymmetric causal discovery in financial time series. *Symmetry*, *17*(10), 1591. https://doi.org/10.3390/sym17101591

[62] Rudd, D. H., Huo, H., & Xu, G. (2021). Causal analysis of customer churn using deep learning. In *2021 International Conference on Digital Society and Intelligent Systems*, 319–324. https://doi.org/10.1109/DSInS54396.2021.9670561

[63] Tan, Z., & Wu, Y. (2025). On regime switching models. *Mathematics*, *13*(7), 1128. https://doi.org/10.3390/math13071128

[64] Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). The MIT Press. https://doi.org/10.7551/mitpress/1754.001.0001

How to Cite: Dang, Q.-V., Dinh, M. N., Le, D., & Nguyen, D.-M. (2026). REIGN: Regime-Enhanced Intelligent Granger Network for Nonstationary Causal Discovery. *Journal of Data Science and Intelligent Systems.* https://doi.org/10.47852/bonviewJDSIS620210441