TRANSFORMERS ARE LIMITED CAUSAL LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive Transformers trained by next-step prediction underlie foundation models across modalities and appear to learn the structure of the worlds that generate their data. A growing body of work reads this structure as causal, through attention maps or Transformer-based causal-discovery estimators, yet when prediction recovers causal structure, and where it fails, remains unclear. In this work, we examine this question for time series from lagged structural causal models, distinguishing a variable's *predictive parents*, on which its conditional distribution depends, from its *structural parents*, which enter its generating mechanism. We show that under four identifying conditions the two sets coincide, so that structural parents can be read from the gradient of the log-density that a Transformer learns by next-step prediction, without any graph objective. Beyond these conditions, where causes are hidden, effects are instantaneous, or regimes are pooled, other past variables become predictive parents, and reading them as direct causes yields causal illusions that no model capacity can remove. Experiments with decoder-only Transformers bear out both halves: within the identifying conditions, the readout recovers structural parents across mechanisms, noise, dimensions, and lag horizons, including parents that act only on the shape of the distribution, while attention reflects this structure only in shallow models. Outside them, recovery declines as the number of hidden causes, instantaneous effects, or hidden regimes grows, and supplying the missing information through the regime identity or dedicated causal-discovery methods serves as a remedy. Taken together, our findings support viewing Transformers as causal learners within the identifying conditions and as limited causal learners beyond them, and they locate the limits of causal learning from prediction in the information that the observed history carries.

1 INTRODUCTION

Autoregressive Transformers have become a common interface for modeling sequential worlds. Text is predicted token by token, images and videos are generated as visual token sequences, audio is modeled through discrete acoustic codes, robot actions are emitted as sequence outputs, and time-series foundation models forecast future observations from long histories (Brown et al., 2020; Chen et al., 2020; Borsos et al., 2023; Zitkovich et al., 2023; Kim et al., 2024; Ansari et al., 2025; Das et al., 2024; Liu et al., 2025). At scale, next-step prediction also seems to give these models internal representations of the world that produces their data. A sequence model trained only on Othello moves encodes the board state (Li et al., 2022), a chess-playing language model tracks latent properties of the game (Karvonen, 2024), and large language models represent spatial and temporal variables linearly (Gurnee & Tegmark, 2023). These results suggest that prediction alone may lead a model to the structure that generates its observations.

A growing body of work reads this structure as causal structure. Attention weights of trained Transformers have been interpreted as causal graphs or conditional-independence relations (Rohekar et al., 2023; Zhang et al., 2023; Lu et al., 2023; Kong et al., 2024), and simplified Transformers trained by gradient descent have been shown to encode a latent causal graph in their attention (Nichani et al., 2024). Other work trains Transformers as amortized estimators that map observations to causal graphs (Ke et al., 2022; Wang & Kording, 2022; Dhir et al., 2024). Yet a model can predict well and still misrepresent the world that produced its data (Vafa et al., 2024; 2025). A model trained to predict a variable learns which past variables help predict it. We call these past variables its *predictive parents*. Causal discovery instead asks which past variables enter the mechanism that generates the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

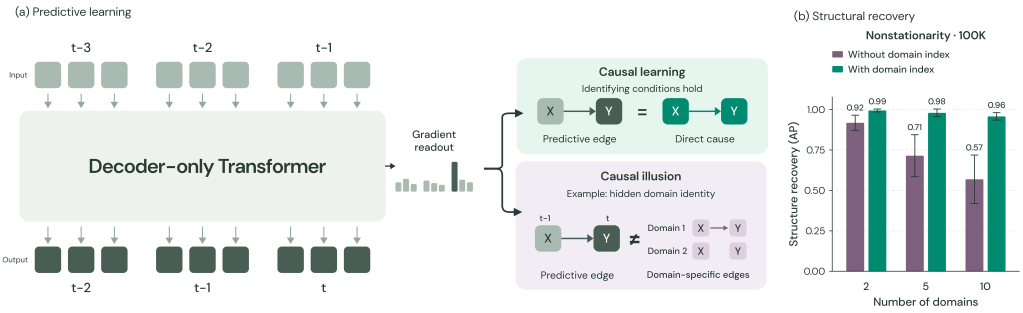


Figure 1: **Causal learning and causal illusion.** **(a)** A Transformer trained only for next-step prediction yields predictive parents through the gradient of its log-density, and within the identifying conditions these are the structural parents. Hidden domain identity is one example of missing information: pooling domains can yield predictive edges that are not direct causal edges in the current domain. **(b)** AP of the domain-specific structural parents in Extreme mixed systems with 100K samples, with the domain index hidden or supplied as an extra token (mean $\pm$ SD over three data seeds, after averaging optimizer seeds).

variable. We call these its *structural parents*. This raises the central question of this paper: *when are the predictive parents that a Transformer learns also structural parents?*

Lagged structural causal models make this question precise. The structural parents of $X_{i,t}$ are the past variables in its transition equation, and its predictive parents are the past variables that change the conditional distribution of $X_{i,t}$ given the observed history. We measure this change with the *score energy*, the expected squared derivative of the conditional log-density with respect to each candidate lagged input. Under four identifying conditions, namely conditional exogeneity, no instantaneous effects, lag-window coverage, and faithfulness, the two sets are identical (Theorem 1, Figure 1a), which ties Granger's predictive notion of causality (Granger, 1969) to structural parents.

This result connects structural causation directly to autoregressive training, the recipe behind modern foundation models. An autoregressive model is trained to maximize the likelihood of each next step given its history, so the object it learns is the conditional distribution on which the theorem is stated. Its causal mask, which restricts each prediction to the past, matches the temporal order that orients every candidate edge in a lagged model. Structural parents can therefore be read from a trained autoregressive model by differentiating its learned log-density with respect to each input, with no graph objective, sparsity penalty, or change to the standard training recipe. Within the identifying conditions, a model that learns the conditional distribution well is thus a causal learner, because its predictive parents are the structural parents. Beyond them, it is a limited causal learner, because its predictive parents can include variables that are not structural parents. A *causal illusion* arises when a genuine predictive dependence is read as a direct causal edge that is absent from the data-generating mechanism.

We test both halves with standard decoder-only Transformers. Each model is trained by next-step likelihood, using a mixture-of-Gaussians negative log-likelihood for continuous values, and we read predictive parents from the gradient of its log-density, which directly estimates the score energy of Theorem 1. Within the identifying conditions, we vary the mechanism, the noise type and variance, the dimension, and the lag horizon. Beyond them, we hide causes or the regime identity, or add instantaneous effects. We also apply the readout to benchmark data generated from real-world records, where these conditions cannot be verified. Across settings, we also vary the sample size, the readout, and the architecture, and compare the predictive parents that the Transformer recovers with the structural parents, using established temporal causal-discovery methods as references. Our findings are as follows.

- **Within the identifying conditions, Transformers recover structural parents, including parents that act only on the shape of the distribution.** Across nonlinear mechanisms, noise types, dimensions, and lag horizons, the gradient readout ranks structural parents almost perfectly, while

mean-based and linear methods miss parents that leave the conditional mean unchanged. Edge by edge, the score energy of the trained Transformer matches the score energy computed from the true data-generating process.

- **Attention reflects the structure in a one-layer Transformer and loses it as the Transformer gets deeper.** With one layer, attention ranks structural parents nearly as well as the gradient readout; with four layers, the gradient readout improves while attention degrades and does not recover with more data.
- **Beyond the identifying conditions, the Transformer's predictive parents stop being structural parents.** Each way of leaving these conditions produces a specific kind of spurious parent: a hidden cause shared by the history and target creates extra predictive parents, a hidden regime mixes the structural parents of different regimes, and instantaneous effects induce lagged edges. The mismatch grows across the tested counts of hidden causes, instantaneous edges, and regimes. On benchmarks built from air-quality, traffic, and medical records, where these conditions cannot be verified, the readout recovers structure comparably to established temporal causal-discovery methods.
- **Supplying the missing information offers a practical remedy.** Giving the model the regime identity, or combining the readout with a traditional discovery method built for the missing structure, such as L-PCMCI for hidden causes or PCMCI$^+$ for instantaneous effects, improves the recovery of structural parents in our experiments. In the ten-regime systems, providing the identity with 50K observations yields higher AP than using 1M observations without it.

2 PRELIMINARIES

Autoregressive learning. An autoregressive model factorizes a sequence in temporal order, $p(x_{1:T}) = \prod_t p(x_t \mid x_{<t})$, and learns each factor by predicting the next element from the elements before it. Modern Transformers learn in this way, whether the next element is a text token, a visual token, an acoustic code, or a patch of a time series. For a p-variate series $X_t = (X_{1,t}, \dots, X_{p,t})^\top$, the factor for variable i is the conditional law $p_i^*(\cdot \mid X_{t-L:t-1})$ of $X_{i,t}$ given the observed history of the last L steps. Under a likelihood objective, the population minimizer is exactly this conditional law. Which past variables this law depends on is set by the system that generates the data.

Multivariate dynamical systems. Let the series X_t be generated by a dynamical system in which each variable follows

$$X_{i,t} = f_i^{(r_t)}\big(\mathrm{Pa}^{(r_t)}(i,t),\, U_t,\, N_{i,t}\big), \qquad i = 1, \dots, p, \tag{1}$$

where $\mathrm{Pa}(i,t) \subseteq \{X_{j,t-\ell} : j \in [p],\ \ell \geq 0\} \setminus \{X_{i,t}\}$ denotes the set of observed parents, U_t denotes an unobserved process, r_t denotes the regime at time t, and $N_{i,t}$ are mutually independent noise variables satisfying $N_{i,t} \perp (X_{<t}, U_{\leq t})$. Parents at lag $\ell = 0$ represent instantaneous effects, the regime index represents nonstationarity, and a stationary system has a single regime. The observed history $X_{t-L:t-1}$ excludes U_t and r_t unless they are supplied explicitly.

Causal graph and structural parents. We write $j \xrightarrow{\ell} i$ if $X_{j,t-\ell} \in \mathrm{Pa}(i,t)$, that is, if $X_{j,t-\ell}$ is a direct cause of $X_{i,t}$, and within a regime these edges form the lagged causal graph $\mathcal{G}^*$. For each target i, the *structural parents* are its incoming lagged edges $\{(j,\ell) : j \xrightarrow{\ell} i,\ \ell \geq 1\}$, which we search for among the candidates $\{(j,\ell) : j \in [p],\ \ell \in [K]\}$ with $K \leq L$, including each variable's own lags. Instantaneous edges ($\ell = 0$) belong to $\mathcal{G}^*$ and shape the learned distribution, but they lie outside the candidate set because a next-step predictor conditions only on the past. The lagged causal structure is identifiable when the distribution of the observed series determines the structural parents of every target uniquely, even if systems with the same distribution differ in their functions f_i or noise laws (Spirtes et al., 2000; Peters et al., 2017). An autoregressive learner sees this distribution through the conditional laws $p_i^*(\cdot \mid X_{t-L:t-1})$, and Section 3 asks which past variables these laws depend on and under which conditions those variables are exactly the structural parents.

3 When Next-Step Predictions Reveal Structural Parents

3.1 Predictive parents

Fix a target variable i and a candidate lagged input $X_{j,t-\ell}$ with $j \in [p]$ and $\ell \in [K]$. We consider a single regime in which the process is stationary, so that the joint law of $X_{i,t}$ and its history is the same at every t. The *score energy* of the candidate measures how strongly the conditional law of $X_{i,t}$ responds to it,

$$H^{\ell}_{j,i} = \mathbb{E}\Big[\big(\partial_{x_{j,t-\ell}} \log p^*_i(X_{i,t} \mid X_{t-L:t-1})\big)^2\Big]. \tag{2}$$

The derivative holds the value of $X_{i,t}$ fixed and measures how its conditional density changes when $X_{j,t-\ell}$ changes, and squaring before averaging keeps sensitivities that change sign across histories. Because the energy is defined on the whole conditional density, it registers a candidate whether the candidate moves the mean of $X_{i,t}$, its spread, or its shape. We call (j, ℓ) a *predictive parent* of i when $H^{\ell}_{j,i} > 0$. Write $X_{-(j,\ell)}$ for the history $X_{t-L:t-1}$ without $X_{j,t-\ell}$.

Lemma 1 (Predictive Parents as Conditional Dependencies)**.** *Under the regularity conditions of Appendix A.1, $H^{\ell}_{j,i} = 0$ if and only if $X_{i,t} \perp X_{j,t-\ell} \mid X_{-(j,\ell)}$.*

A predictive parent is therefore a past variable that still carries information about $X_{i,t}$ after the rest of the observed history is known (Lemma 3 in the appendix gives the proof).

3.2 When predictive parents are structural parents

Predictive parents are defined by the distribution of the observed series and structural parents by the mechanisms in Eq. (1), and the four identifying conditions below connect the two.

Assumptions

- A1 **Conditional exogeneity.** Given the structural parents of $X_{i,t}$ in the history, its unobserved inputs U_t are independent of the remaining history.
- A2 **No instantaneous effects.** Every parent of $X_{i,t}$ lies at a lag $\ell \geq 1$.
- A3 **Lag-window coverage.** The candidate lags $\ell \in [K]$, with $K \leq L$, contain every structural parent.
- A4 **Faithfulness.** Every conditional independence in the distribution of the system is implied by its causal graph, including the unobserved variables.

Identifiability of Structural Parents through Prediction

Theorem 1. *Consider a single stationary regime that satisfies A1–A4 and the regularity conditions of Appendix A.1. For every target i and every candidate (j, ℓ) with $\ell \in [K]$,*

$$H^{\ell}_{j,i} > 0 \quad \Longleftrightarrow \quad j \xrightarrow{\ell} i.$$

The predictive parents of every variable are therefore its structural parents, and the lagged causal graph $\mathcal{G}^*$ is identifiable from the conditional laws $p^*_i(\cdot \mid X_{t-L:t-1})$.

The proof in Appendix A.1 uses the conditions in two directions. A1 and A2, together with the independence of the noise, give $p^*_i(y \mid x_{t-L:t-1}) = p^*_i(y \mid x_{\mathrm{Pa}})$, where x_{Pa} collects the structural parents, so the rest of the history carries no information about $X_{i,t}$ once the structural parents are known, and every other candidate has zero energy by Lemma 1. A3 places each structural parent in the history, where its direct edge to $X_{i,t}$ stays open when the rest of the history is conditioned on, and A4 turns this open edge into a dependence with positive energy. Temporal order fixes the direction of every edge, so first-order derivatives of the conditional law suffice, whereas score-matching methods for i.i.d. data need second-order information of the joint density to recover an unknown causal order (Rolland et al., 2022; Montagna et al., 2023; Sanchez et al., 2022).

Causal sufficiency, which rules out unobserved inputs altogether, is the special case of A1 without U_t. A1 is weaker, because it admits unobserved inputs as long as the rest of the history carries no information about them once the structural parents are known (White & Lu, 2010). For example, an unobserved shock that hits several variables at the same time step and does not persist violates causal

sufficiency, yet it is independent of the entire history, so A1 holds and the predictive parents remain the structural parents. A3 can be met by enlarging the window, and A4 holds for almost all parameter values, since unfaithful parameters form a set of Lebesgue measure zero in linear Gaussian models (Meek, 2013) and a cancellation in nonlinear mechanisms must hold across all input values, which requires even more special parameters. We therefore treat A3 and A4 as defaults, and Section 3.4 studies what happens when A1 or A2 fails or when data from several regimes are pooled.

3.3 Reading predictive parents from autoregressive training

A model trained by next-step likelihood (Section 2) provides an approximation $q_{\theta,i}$ of the conditional laws p_i^* on which Theorem 1 is stated. We use a decoder-only Transformer whose input is the window $X_{t-L:t-1}$ flattened into one token per variable and lag, so that every candidate (j, ℓ) is a separate input with its own path to each target through self-attention. We add a node embedding and a time-step embedding that encode, respectively, which variable and which time step each token represents. Each output parameterizes a mixture of Gaussians for $q_{\theta,i}(\cdot \mid X_{t-L:t-1})$, a flexible density family that lets the score energy respond to every aspect of the conditional law. The model is trained with the negative log-likelihood without any explicit graph objective. We read predictive parents by applying Eq. (2) to the fitted density,

$$\widehat{H}^{\ell}_{j,i}(\theta) = \frac{1}{n} \sum_{r=1}^{n} \Big(\partial_{x_{j,t-\ell}} \log q_{\theta,i}\big(x_{i,t}^{(r)} \mid x_{t-L:t-1}^{(r)}\big) \Big)^2, \tag{3}$$

over n held-out target–history pairs, with the derivative taken by automatic differentiation through the whole network. The readout inherits Theorem 1 when the fitted log-density has the same input derivatives as the true one.

3.4 Beyond the identifying conditions

We now consider what happens when the identifying conditions are violated or when data from several regimes are pooled. We call an input of the mechanism of $X_{i,t}$ that is not contained in the observed history a *hidden parent*; examples include an unobserved input, a contemporaneous parent, a parent at a lag beyond the window, and the regime index of pooled data. A hidden parent alone leaves Theorem 1 intact (Section 3.2), and predictive and structural parents separate only when the history carries information about it, which Proposition 2 characterizes.

Proposition 2 (Predictive Parents under Hidden Parents)**.** *Suppose the full system, including all unobserved variables, is an acyclic structural causal model over time with independent noises that satisfies A4 and the regularity conditions. Then every structural parent of $X_{i,t}$ in the candidate set is a predictive parent of i, and any other candidate (j, ℓ) is a predictive parent of i if and only if $X_{j,t-\ell}$ is d-connected to a hidden parent of $X_{i,t}$ given $X_{-(j,\ell)}$.*

In other words, an extra predictive parent is a past variable that carries information about a hidden parent beyond what the rest of the history carries. To make Proposition 2 more intuitive, we consider three typical cases, which are also the settings studied in Section 4. (1) *Hidden causes*, which violate A1. A hidden cause that reaches both the history and the target, through persistence or effects at different times, makes its effects in the history predictive parents of the target as proxies of the hidden cause, and when such an effect also has observed causes, conditioning on it makes those causes predictive through explaining away. (2) *Instantaneous effects*, which violate A2. If $X_{k,t}$ is a contemporaneous parent of $X_{i,t}$, the past causes of $X_{k,t}$ become predictive parents of $X_{i,t}$. (3) *Hidden regimes*. When data from several regimes are pooled and the regime index is not observed, the conditional law of $X_{i,t}$ is a mixture over regimes, so the structural parents of every regime become predictive, together with the past variables that reveal the current regime. In all three cases, the extra predictive parents produce causal illusions when they are interpreted as direct causes.

In cases (1) and (2), these illusions persist with more data and larger models, since the lagged graph itself is not identifiable from the observed distribution. In fact, these are the well-known settings of causal discovery with latent confounders and contemporaneous effects, in which the graph can in general be recovered only up to an equivalence class (Malinsky & Spirtes, 2018; Gerhardus & Runge, 2020; Runge, 2020). On the other hand, if the missing information is made available, for example by measuring the hidden cause or by providing the regime index, the hidden parent becomes part of the history, and Theorem 1 applies again when no other hidden parent remains.

4 EMPIRICAL STUDIES OF TRANSFORMERS AS CAUSAL LEARNERS

Setup. For the main grid, we use the structural time-series simulator of Lawrence et al. (2020) and our own generator, varying one property at a time relative to a linear system with 10 variables, a maximum lag of 5, and three parents per variable in expectation (Appendix A.2.1). Except in the architecture ablations, every synthetic setting uses the same decoder-only Transformer with the same architecture and optimizer, without setting-specific tuning (Appendix A.2.2), and we read its predictive parents with Eq. (3) on held-out windows. We score the readout by average precision (AP) over all lagged candidates (j, ℓ), including each variable's own lags, and within each regime when the regime changes; the fraction of true edges provides an approximate chance reference. Established temporal causal-discovery methods serve as references, namely PCMCI with partial-correlation tests (Runge et al., 2017), DYNOTEARS (Pamfil et al., 2020), VAR-LiNGAM (Hyvärinen et al., 2010), NTS-NOTEARS (Sun et al., 2021), TCDF (Nauta et al., 2019), pairwise and multivariate Granger tests (Granger, 1969), and random scores, and the appendix gives the evaluation protocol.

4.1 WITHIN THE IDENTIFYING CONDITIONS

Five nonlinear mechanisms act on the conditional mean of $X_{i,t}$ (Interaction), on its variance (Scale), or only on the shape of its distribution at a fixed mean and variance (Shape and Transport), and Mixed assigns one of them to each variable (Appendix A.2.1).

The readout recovers structural parents within the identifying conditions. Across the five mechanisms and 50K to 1M samples, the readout reaches an AP of 0.986 to 1.000, where true edges make up 6.2% of the candidates; Figure 2a shows the 100K comparison. The parents of Scale, Shape, and Transport leave the conditional mean unchanged, and the readout registers them because the score energy covers the whole conditional density (Section 3.1), whereas the references, which test conditional means or linear dependence, stay near the fraction of true edges, at or below 0.11 on these three mechanisms and at or below 0.18 on all five. One architecture and one training configuration handle all five mechanisms, which reflects the expressivity of the Transformer as a conditional density estimator. In linear systems, references whose inductive bias matches the generator reach an AP near 1, and PCMCI with partial-correlation tests reaches 1.000 in every linear setting. The readout also reaches 1.000 with 10 variables and 5 lags, and it declines to 0.961 with 100 variables and to 0.898 at a maximum lag of 50, where true edges make up only 0.6% of the candidates (Figure 2b,c). This decline is consistent with the finite-sample cost of a model that assumes no functional form, since the candidate set grows while the graph stays sparse. Across noise types, the readout stays at approximately 0.998 or above, while DYNOTEARS falls from 0.91 under Gaussian noise to 0.81 in the large unequal-variance setting, consistent with Ng et al. (2024).

4.2 BEYOND THE IDENTIFYING CONDITIONS

We hide the regime identity in the nonstationary systems and, to make the extra predictive parents strong enough to outrank the structural parents, build two diagnostic systems with 50K observations. At fixed per-source strength, we activate 3, 5, or 10 independent hidden causes, or 3, 6, or 9 contemporaneous edges, retaining a zero-effect control with the same direct lagged graph and innovations. We supply the missing information by giving the model the regime identity as an extra token, and, when the hidden parent cannot be measured, by running bounded L-PCMCI (Gerhardus & Runge, 2020) or PCMCI$^+$ (Runge, 2020) on the full series with its lagged search restricted to the readout's top half of candidates per target (Appendix A.2.2).

Recovery falls as the number of hidden regimes grows. When the regime identity is hidden, the regime index becomes a hidden parent (case 3 in Section 3.4). At 100K samples, AP falls from 0.857 to 0.167 in linear systems and from 0.919 to 0.569 in nonlinear systems with regime-specific lags as the number of regimes rises from 2 to 10 (Figure 1b, and Figure 6 in the appendix).

Hidden causes and instantaneous effects lower recovery while the model still predicts accurately. In the diagnostic systems, AP falls from 1.000 in the zero-effect controls to 0.819, 0.663, and 0.319 with 3, 5, and 10 hidden causes, and to 0.924, 0.755, and 0.475 with 3, 6, and 9 contemporaneous edges (Figure 3a,b). The model still fits both systems closely: at the largest counts, its mean held-out

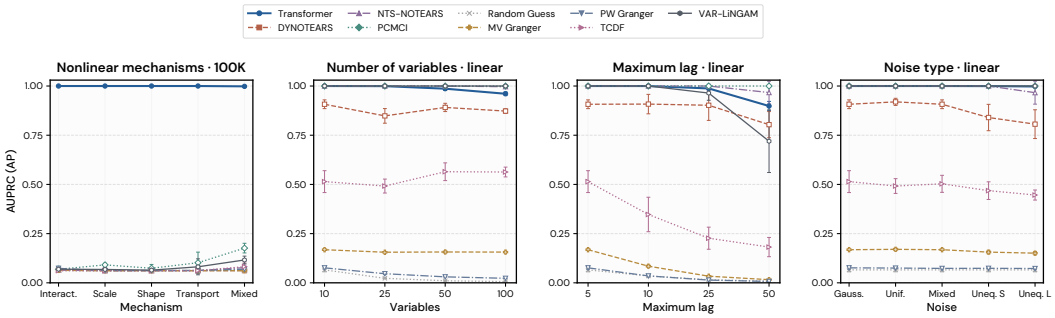


Figure 2: **Structural parents within the identifying conditions.** AP of the Transformer readout and of the reference methods across five nonlinear mechanisms (a), the number of variables (b), the maximum lag (c), and the noise type (d), with 100K samples. The readout reaches an AP of 0.898 or higher in every setting. The references stay near the fraction of true edges on the nonlinear mechanisms, and several of them reach an AP near 1 in the linear systems, whose functional form they assume. The Gaussian point in (d) is the linear reference system. Mean ± SD over three data seeds, with true edges making up 0.6% to 6.2% of the candidates.

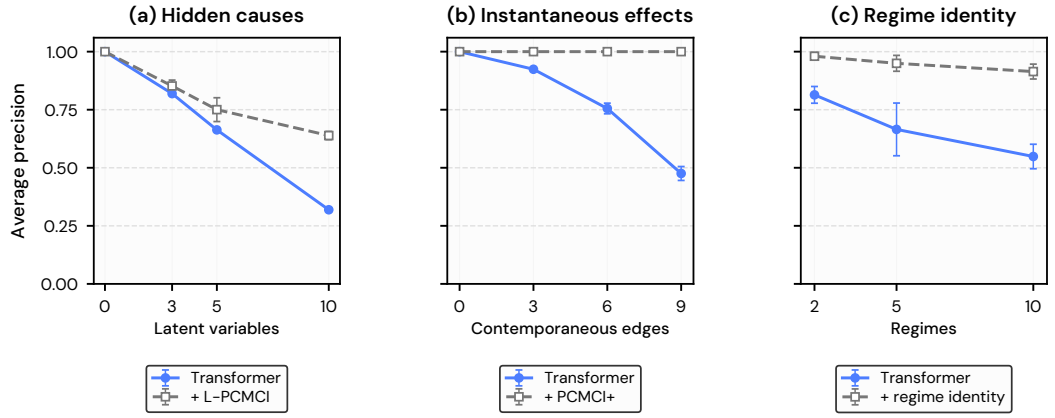


Figure 3: **Leaving the identifying conditions lowers recovery, and targeted refinement improves it.** (a,b) Increasing numbers of hidden causes and contemporaneous edges, for the readout and after L-PCMCI or PCMCI$^+$. (c) Increasing numbers of nonlinear regimes, with and without the regime identity. All panels use 50K observations. Mean ± SD over three data seeds, after averaging three optimizer seeds within each.

negative log-likelihood exceeds that of the true conditional law by only 0.010 and 0.001 nats per target in the hidden-cause and instantaneous systems, respectively, while the true laws themselves contain extra predictive parents. In the random graphs of our main grid, these effects are milder, with an AP of 0.971 under instantaneous effects and 0.984 with ten hidden processes at 100K samples. On benchmarks built from air-quality, traffic, and medical records (Cheng et al., 2023), where the identifying conditions cannot be checked, the readout reaches an AUPRC comparable to 14 established methods and at the upper end of their range on traffic (Table 1 in the appendix).

Supplying the missing information improves structural-parent recovery. At 50K samples, supplying the regime identity raises AP from 0.814, 0.665, and 0.549 to 0.980, 0.950, and 0.914 as the number of nonlinear regimes increases from 2 to 5 and 10 (Figure 3c), and in the linear family the identity raises AP from 0.303 to 0.999 at 1M samples. With 50K samples and the identity, both families already exceed the models without it at 1M samples, at 0.914 against 0.835 and 0.417 against 0.303, so supplied information closes a gap that more data leaves open. When the hidden parent cannot be supplied, PCMCI$^+$ conditions on the contemporaneous parent and restores an AP of 1.000 at every instantaneous-edge count, and bounded L-PCMCI raises AP to 0.852, 0.750, and 0.639

with 3, 5, and 10 hidden causes. The latent improvement is smaller with three hidden causes, and recovery remains incomplete because confounded adjacencies and some extra predictive relationships remain. These results agree with Section 3.4: once the regime identity enters the history, Theorem 1 applies within each regime, and the discovery-method refinements improve recovery without such a guarantee.

4.3 Readout, architecture, and data

From each trained Transformer, we also read attention rollout (Abnar & Zuidema, 2020), which multiplies the attention weights across layers to trace how much each input token contributes to the final position. Besides the four-layer Transformer used throughout, we train a one-layer Transformer and a one-layer causal convolutional network (CNN) with the same mixture head, objective, and readout.

Attention tracks the structure in a one-layer Transformer and loses it in a four-layer one. Within a checkpoint, the two readouts share the same learned density, so their difference reflects the readout alone. In the four-layer Transformer on Mixed at 100K samples, the gradient readout reaches an AP of 0.998 and attention rollout reaches 0.321, and in linear systems the two reach 1.000 and 0.375 (Figure 4). More data widens this difference on Mixed, since from 50K to 1M samples the gradient readout rises from 0.986 to 0.999 while attention rollout falls from 0.346 to 0.305. In the one-layer Transformer, attention rollout stays close to the gradient readout, at 0.889 against 0.944 on Mixed and 0.817 against 0.947 in linear systems. This pattern is consistent with Theorem 1, which places structural parents in the score energy, a measure of how the prediction itself responds to each input. Attention weights set how much each token contributes to another token's representation at a single layer, and in a deeper Transformer, value projections, residual streams, normalization, and MLP blocks mix the representations repeatedly, so a token can receive high attention and still have little effect on the prediction (Jain & Wallace, 2019).

The architecture sets how closely the model learns the conditional law. On Mixed at 100K samples, the gradient readout reaches an AP of 0.998 with the four-layer Transformer, 0.944 with the one-layer Transformer, and 0.922 with the CNN, and in linear systems the four-layer Transformer and the CNN both reach 1.000. The readout therefore works with any model that learns the conditional law, as Theorem 1 concerns the conditional law itself, and the four-layer Transformer comes closest to the structural parents on the nonlinear mechanisms. Under hidden regimes, depth helps less, since the four-layer model leads at two regimes (0.919 against 0.844) and trails the one-layer model at five and ten regimes (0.715 against 0.769 and 0.569 against 0.696), with a standard deviation across data seeds of up to 0.17. Between 50K and 1M samples on Mixed, the one-layer model rises from 0.916 to 0.966, so more data narrows the gap between the two depths within the identifying conditions.

5 Related Work

Temporal causal discovery. Temporal causal discovery uses time order together with structural assumptions to identify directed dependencies. Classical families include Granger-style predictive tests, structural time-series models such as TiMINo, conditional-independence methods such as PCMCI/PCMCI$^+$, continuous-optimization methods such as DYNOTEARS, and neural Granger approaches based on sparsity, Jacobians, or learned dependency scores (Granger, 1969; Peters et al., 2013; Runge et al., 2017; Runge, 2020; Pamfil et al., 2020; Tank et al., 2021; Nauta et al., 2019; Zhou et al., 2024). Score-based structure learning shows that derivatives of learned densities can reveal graph structure, through score matching and diffusion-based ordering (Rolland et al., 2022; Montagna et al., 2023; Sanchez et al., 2022). We place autoregressive training in this picture by asking when the predictive parents that it learns, which extend Granger's predictive notion to the whole conditional law, coincide with the structural parents that these methods target.

Attention and gradient attribution in Transformers. Prior work finds that attention weights often misrepresent what drives the outputs of a model, especially in deep models where token representations are repeatedly mixed across layers (Jain & Wallace, 2019; Serrano & Smith, 2019; Bastings & Filippova, 2020). Gradient-based attribution methods, including Integrated Gradients and Layer-wise Relevance Propagation, attribute predictions through the full computation graph

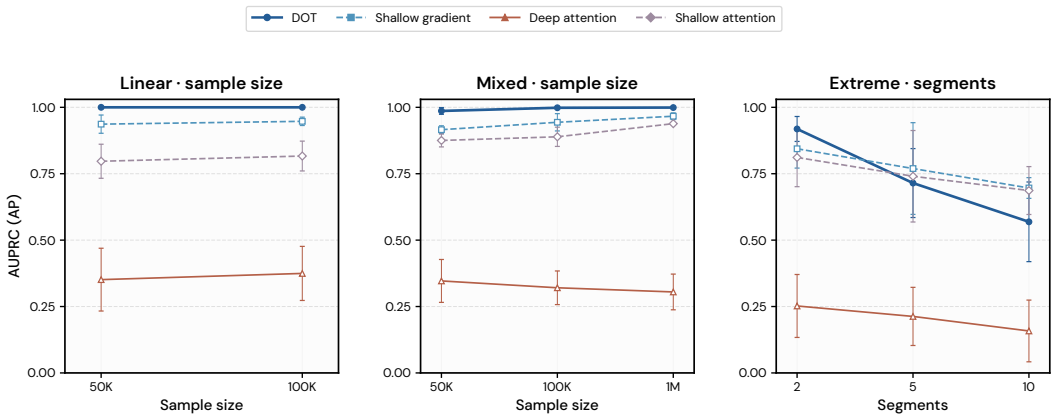


Figure 4: **Gradient readout and attention rollout from the same checkpoints.** In linear systems (left) and on Mixed (middle), attention rollout from the four-layer Transformer stays between 0.30 and 0.38 while its gradient readout reaches 0.986 to 1.000, and the one-layer Transformer keeps the two readouts close. Under hidden regimes (right), every readout falls as the number of regimes grows. DOT denotes the four-layer gradient readout. Mean $\pm$ SD over three data seeds.

(Sundararajan et al., 2017; Bach et al., 2015; Achtibat et al., 2024). Transformer-based causal discovery methods often add specialized detectors, graph heads, or supervised or meta-learned graph prediction modules (Kong et al., 2024; Ke et al., 2022; Wang & Kording, 2022; Dhir et al., 2024). We study a Transformer trained only for next-step prediction, read the score energy of its learned density through the whole network, and compare this readout with attention rollout on the same checkpoints (Section 4.3).

Causal discovery beyond the clean lagged setting. Ideal causal discovery settings assume observed variables, lagged effects, sufficient context, and faithfulness. A large temporal-causal-discovery literature studies violations of these assumptions, with latent-confounder methods such as SVAR-FCI and L-PCMCI, contemporaneous-effect methods such as PCMCI$^+$, and heterogeneous or nonstationary discovery methods based on distribution shifts or sparse mechanism changes (Malinsky & Spirtes, 2018; Gerhardus & Runge, 2020; Runge, 2020; Huang et al., 2020; Perry et al., 2022). These are the settings in which predictive parents separate from structural parents (Section 3.4), and we combine L-PCMCI and PCMCI$^+$ with the readout as remedies (Section 4.2).

6 Conclusion

Under the identifying conditions of lagged structural causal models, the score energy of the conditional distribution that autoregressive training learns identifies the structural parents of each variable, so its predictive parents coincide with its structural parents. With decoder-only Transformers trained by next-step likelihood and a gradient readout, the recovered parents are comparable to those of established temporal causal-discovery methods across nonlinear, long-range, high-dimensional, and noisy settings, including parents that act only on the spread or shape of the distribution. Readout ablations locate this structure in output sensitivities, which attention tracks in a one-layer Transformer and loses in deeper ones. Beyond the identifying conditions, hidden causes, instantaneous effects, and hidden regimes turn other past variables into predictive parents, and supplying the missing information, through the regime identity or through L-PCMCI and PCMCI$^+$, restores much of the structure. These results support Transformers as causal learners within the identifying conditions and as limited causal learners beyond them.

AI USE STATEMENT

In this work, we used generative AI tools to assist with developing experimental code for the methods and evaluations reported in the paper, and with translating manuscript text. We did not use generative AI tools to develop the theoretical framework or mathematical proofs, design the experiments, generate or process datasets, or interpret the results. Additionally, we used these tools to prepare scientific figures and polish the manuscript's language for clarity, grammar, and readability. AI assistance was also used for LaTeX formatting and adaptation to the ICLR 2027 submission template.

The authors manually reviewed the AI-assisted code, figures, and text, and tested the experimental code. The authors take responsibility for the final content of this work, including the experimental implementation, the correspondence between figures and experimental results, the mathematical claims and their explanations, and the conclusions presented in the manuscript.

ETHICS STATEMENT

Inferred graphs from observational time series should not be treated as definitive causal knowledge without further validation, especially in high-stakes applications. Appendix A.4 discusses these risks and the need for domain expertise and, where possible, experimental or interventional evidence. Appendix A.2.4 documents the third-party assets, licenses, and terms of use relevant to our experiments.

REPRODUCIBILITY STATEMENT

The appendix provides the theoretical proofs in Appendix A.1. Appendix A.2 describes the data-generating processes and experimental settings, including model architecture and training details in Appendix A.2.2, compute resources in Appendix A.2.3, and third-party assets in Appendix A.2.4.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:218487351.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024.

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

Victor Akinwande and J Zico Kolter. Acceleratedlingam: Learning causal dags at the speed of gpus. *arXiv preprint arXiv:2403.03772*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*, 2020.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Anish Dhir, Matthew Ashman, James Requeima, and Mark van der Wilk. A meta-learning approach to bayesian causal discovery. *arXiv preprint arXiv:2412.16577*, 2024.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in neural information processing systems*, 33:12615–12625, 2020.

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

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, 90:208–225, 2017.

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.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Adam Karvonen. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*, 2024.

Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Anirudh Goyal, Jorg Bornschein, Melanie Rey, Theophane Weber, Matthew Botvinic, Michael Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. *arXiv preprint arXiv:2204.04875*, 2022.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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.

Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*, 2020.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025.

Ziyu Lu, Anika Tabassum, Shruti Kulkarni, Lu Mi, J Nathan Kutz, Eric Shea-Brown, and Seung-Hwan Lim. Attention for causal relationship discovery from biological neural dynamics. *arXiv preprint arXiv:2311.06928*, 2023.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, pp. 23–47. PMLR, 2018.

Christopher Meek. Strong completeness and faithfulness in bayesian networks. *arXiv preprint arXiv:1302.4973*, 2013.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pp. 752–771. PMLR, 2023.

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

Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, pp. 71–105. PMLR, 2024.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

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.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

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.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36:31450–31465, 2023.

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

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pp. 1388–1397. Pmlr, 2020.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Ashka Shah, Adela DePavia, Nathaniel Hudson, Ian Foster, and Rick Stevens. Causal discovery over high-dimensional structured hypothesis spaces with causal graph partitioning. *arXiv preprint arXiv:2406.06348*, 2024.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

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.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pp. 436–463, 2013.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Keyon Vafa, Peter G Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. *arXiv preprint arXiv:2507.06952*, 2025.

Xinyue Wang and Konrad Paul Kording. Learning domain-specific causal discovery from time series. *arXiv preprint arXiv:2209.05598*, 2022.

Halbert White and Xun Lu. Granger causality and dynamic structural systems. *Journal of Financial Econometrics*, 8(2):193–243, 2010.

Behrooz Zarebavani, Foad Jafarinejad, Matin Hashemi, and Saber Salehkaleybar. cupc: Cuda-based parallel pc algorithm for causal structure learning on gpu. *IEEE Transactions on Parallel and Distributed Systems*, 31(3):530–542, 2019.

Jiaqi Zhang, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between causal inference and attention. *arXiv preprint arXiv:2310.00809*, 2023.

Wenxia Zheng and Wenhe Liu. Symmetry-aware transformers for asymmetric causal discovery in financial time series. *Symmetry*, 17(10):1591, 2025.

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

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A APPENDIX

TABLE OF CONTENTS

A.1 Identifiability of the causal structure 15
A.1.1 Beyond the identifying conditions 17
A.2 Experiment setups 19
A.2.1 Data generation and simulation 19
A.2.2 Training details and model architecture 20
A.2.3 Compute resources 20
A.2.4 Existing assets, licenses, and terms of use 20
A.3 Limitations 22
A.4 Broader impacts 23
A.5 Discussions 24
A.5.1 Transformer-based causal discovery 24
A.5.2 Scalability in causal discovery 24
A.5.3 Structural causality and Granger causality 24
A.5.4 Relationships to previous theory results 24
A.6 Additional Experiments 26
A.6.1 Edge-level agreement between fitted and true score energy 26
A.6.2 Hidden regimes across families and sample sizes 26
A.6.3 Hidden regimes with small changes between regimes 27
A.6.4 Real-world benchmarks 27

A.1 IDENTIFIABILITY OF THE CAUSAL STRUCTURE

Definition 1 (Identifiability of the Lagged Causal Structure)**.** The lagged causal structure is identifiable if the structural parent set $S = \{j : X_j \in \mathrm{Pa}(Y_t)\}$ of every target can be uniquely recovered from the distribution of the observed series.

We formalize when the score energy of the conditional distribution recovers the structural parents. For a fixed target $Y := X_{i,t}$, let $X = (X_1, \ldots, X_d)$ collect the $d = pL$ history coordinates obtained by stacking all p variables over lags $1{:}L$, and write H_j for the score energy of coordinate X_j. This corresponds to $H^{\ell}_{j',i}$ in the main text, where j indexes the pair (j', ℓ). Within a single stationary regime, the target follows

$$Y_t = f_Y(\mathrm{Pa}(Y, t), U_t, N_{Y,t}),$$

where U_t collects the unobserved inputs of Y_t and $N_{Y,t}$ is independent noise satisfying $N_{Y,t} \perp (X, U_t)$. We write $S \subseteq \{1, \ldots, d\}$ for the index set of the structural parents of Y_t inside X and S^c for its complement.

Assumptions. We work under the following conditions.

- **A1, conditional exogeneity**: the unobserved inputs satisfy $U_t \perp X_{S^c} \mid X_S$. This holds automatically when U_t does not exist (causal sufficiency) or when it is independent of the observed past (White & Lu, 2010).
- **A2, no instantaneous effects**: all parents of Y_t lie at lags $\ell \geq 1$ (Peters et al., 2013; Runge et al., 2017).
- **A3, lag-window coverage**: the history X contains all structural parents of Y_t, so the chosen maximum lag is at least the causal horizon.
- **A4, faithfulness**: the distribution of the full system is faithful to its time-unrolled causal graph, including unobserved variables and, when present, instantaneous edges, so that no independence arises from measure-zero cancellations (Pearl, 2009; Spirtes et al., 2000; Peters et al., 2013).
- **Support and regularity**: the law of X admits a density supported on a rectangle $\Omega \subset \mathbb{R}^d$ (Evans, 2022; Adams & Fournier, 2003).
- **Conditional density and score regularity**: for $\mathbb{P}_X$-almost every $x \in \Omega$, the conditional law of Y given $X = x$ admits a density $p^*(\cdot \mid x)$ with respect to Lebesgue measure on $\mathbb{R}$ that is strictly positive on its support. The log-density

  $$\ell^*(y, x) \;:=\; \log p^*(y \mid x)$$

  belongs to $W^{1,2}_{\mathrm{loc}}(\mathbb{R} \times \Omega)$ as a function of (y, x), so that its weak partials $\partial_{x_j} \ell^*$ exist and are square-integrable under the joint law of (Y, X).

Conditional exogeneity is weaker than assuming causal sufficiency and independent noise at the same time. It permits unobserved inputs, provided they do not induce dependence between Y_t and the non-parents X_{S^c} beyond what is mediated by the parents X_S. The Causal Markov property $Y \perp X_{S^c} \mid X_S$ is not assumed separately, since it follows from conditional exogeneity and the data-generating process (Lemma 2). We assume plain faithfulness, which only requires every conditional independence in the distribution to be entailed by d-separation in the graph, and not strong faithfulness (Uhler et al., 2013). Unfaithful distributions form a collection of hypersurfaces in parameter space with Lebesgue measure zero (Meek, 2013; Spirtes et al., 2000), nonlinear mechanisms make exact cancellations even less likely, and any perturbation of the parameters restores faithfulness.

Score energy. Define the *score* of the conditional density with respect to x as

$$s(y, x) \;:=\; \nabla_x \ell^*(y, x), \qquad s_j(y, x) \;:=\; \partial_{x_j} \ell^*(y, x),$$

and the corresponding *score energy*

$$H_j \;:=\; \mathbb{E}\Big[\big(s_j(Y, X)\big)^2\Big], \quad j = 1, \ldots, d.$$

The score $s_j(Y, X)$ measures how sensitive the entire conditional law of Y is to perturbations of the coordinate X_j at the realized pair (Y, X), and H_j aggregates this sensitivity over the population.

Lemma 1 (Functional Independence under a Zero Weak Partial Derivative)**.** *Let $f \in W^{1,1}_{\mathrm{loc}}(\Omega)$ on a rectangle $\Omega \subset \mathbb{R}^d$. If $\partial_{x_j} f = 0$ holds almost everywhere on Ω, then there exists a measurable h with $f(x) = h(x_{-j})$ almost everywhere. Conversely, if f does not depend on x_j, then $\partial_{x_j} f = 0$ holds almost everywhere.*

Proof. Assume $\partial_{x_j} f = 0$ almost everywhere. Fix x_{-j}. For almost every line $t \mapsto (t, x_{-j})$, $f(t_2, x_{-j}) - f(t_1, x_{-j}) = \int_{t_1}^{t_2} \partial_{x_j} f(s, x_{-j})\, ds = 0$, so $f(t, x_{-j})$ is almost everywhere constant in t. Thus there is a measurable h with $f(x) = h(x_{-j})$ almost everywhere. Conversely, if f does not depend on x_j, then its weak partial $\partial_{x_j} f$ is 0 almost everywhere. □

Lemma 2 (Causal Markov Property under Conditional Exogeneity)**.** *Under the data-generating process $Y = f_Y(X_S, U_t, N_{Y,t})$ with $N_{Y,t} \perp (X, U_t)$, and conditional exogeneity $U_t \perp X_{S^c} \mid X_S$, the conditional distribution of Y given X depends only on X_S,*

$$p^*(y \mid x) = p^*(y \mid x_S).$$

In particular, $Y \perp X_{S^c} \mid X_S$.

Proof. By the data-generating process, $Y = f_Y(X_S, U_t, N_{Y,t})$. For any measurable set A,

$$P(Y \in A \mid X = x) = \int P(f_Y(x_S, U_t, n) \in A \mid X = x)\, dP_{N_{Y,t}}(n),$$

where we used $N_{Y,t} \perp (X, U_t)$. By conditional exogeneity, $P(U_t \mid X = x) = P(U_t \mid X_S = x_S)$, so the integrand depends on x only through x_S. Hence $P(Y \in A \mid X = x) = P(Y \in A \mid X_S = x_S)$. □

Lemma 3 (Zero Score Partial and Conditional Independence)**.** *Under the conditional density and score regularity assumption, the following are equivalent for each $j \in \{1, \ldots, d\}$.*

1. *$\partial_{x_j} \ell^*(y, x) = 0$ for almost all $(y, x) \in \mathbb{R} \times \Omega$.*
2. *There exists a measurable h such that $\ell^*(y, x) = h(y, x_{-j})$ almost everywhere.*
3. *The conditional density $p^*(y \mid x)$ does not depend on x_j, that is, $p^*(y \mid x) = \tilde{p}(y \mid x_{-j})$ almost everywhere for some measurable $\tilde{p}$.*
4. *$Y \perp X_j \mid X_{-j}$.*

In particular, $H_j = \mathbb{E}[(\partial_{x_j} \ell^*(Y, X))^2] = 0$ if and only if $Y \perp X_j \mid X_{-j}$.

Proof. $H_j = 0$ if and only if $\partial_{x_j} \ell^*(y, x) = 0$ almost everywhere. By Lemma 1, this holds if and only if $\ell^*(y, x) = h(y, x_{-j})$ for some measurable h, which is equivalent to $p^*(y \mid x) = \exp(h(y, x_{-j}))$ depending on x only through x_{-j}. This factorization of the conditional density is the standard characterization of $Y \perp X_j \mid X_{-j}$. □

Lemma 3 is Lemma 1 of the main text.

Theorem 2 (Score-Based Characterization of Lagged Parents (Restating Theorem 1))**.** *Under the data-generating process, A1–A4, and the conditional density and score regularity assumptions, the score energy H_j recovers the structural parent set,*

$$H_j = 0 \iff j \notin S.$$

Proof. **Claim 1:** $j \notin S \Rightarrow H_j = 0$. By A2 and A3, every parent of Y_t is a lagged variable inside X, so the data-generating process has the form $Y = f_Y(X_S, U_t, N_{Y,t})$ with $N_{Y,t} \perp (X, U_t)$. Lemma 2 and A1 give $p^*(y \mid x) = p^*(y \mid x_S)$. For $j \notin S$, the coordinates x_S are contained in x_{-j}, so the conditional density does not depend on x_j. Lemma 3 then gives $H_j = 0$. This step handles the unobserved inputs by integrating over them with A1.

Claim 2: $j \in S \Rightarrow H_j > 0$. Suppose $j \in S$. Then the causal graph contains the direct edge $X_j \to Y_t$. This single-edge path is not blocked by conditioning on X_{-j}, because it has no intermediate node and

neither endpoint is conditioned on. Thus the graph does not entail $Y_t \perp X_j \mid X_{-j}$. By faithfulness, this conditional independence cannot hold in the distribution, and Lemma 3 gives $H_j > 0$.

Applying both claims to every target identifies each structural parent set, and hence the lagged causal graph, from the conditional laws $p^*(\cdot \mid x)$. □

A.1.1 BEYOND THE IDENTIFYING CONDITIONS

Proof of Proposition 2. Let $\mathcal{G}$ denote the time-unrolled causal graph of the full system, including its unobserved variables, its contemporaneous edges, and its regime index, and write $Y = X_{i,t}$, $X_a = X_{j,t-\ell}$, and $Z = X_{-(j,\ell)}$. Since the full system is an acyclic structural causal model with independent noises, its distribution is Markov with respect to $\mathcal{G}$ (Pearl, 2009), and by A4 it is faithful to $\mathcal{G}$. Together with Lemma 3, this gives, for every candidate, $H^{\ell}_{j,i} > 0$ if and only if X_a and Y are d-connected given Z in $\mathcal{G}$. Every edge of $\mathcal{G}$ points forward in time or stays within a time step, so every descendant of Y lies at time t or later, whereas Z contains only variables at times up to $t-1$.

Structural parents. If (j, ℓ) is a structural parent, the edge $X_a \to Y$ is a path without intermediate nodes, so it is open given any conditioning set, and $H^{\ell}_{j,i} > 0$. This part uses neither A1 nor A2.

Other candidates. Let (j, ℓ) be a candidate outside the structural parents, and suppose first that a path π from X_a to Y is open given Z. If the last edge of π points out of Y, then, following π backward from Y, the edges point away from Y until a first collider, because X_a lies in the past and is not a descendant of Y. This collider is a descendant of Y, and neither it nor any of its descendants belongs to Z, so π would be blocked. Hence the last edge of π is $W \to Y$ for a parent W of Y. The parent W differs from X_a, since (j, ℓ) is not a structural parent, and W is a non-collider on π, so $W \notin Z$. Therefore W lies outside the observed history, that is, W is a hidden parent of Y, and the part of π from X_a to W is open given Z.

Conversely, suppose that a path π' from X_a to a hidden parent W of Y is open given Z. If π' passes through Y, then Y is a non-collider on π', since a collider at Y would require Y or one of its descendants to belong to Z, and the part of π' up to Y is an open path from X_a to Y. Otherwise, appending the edge $W \to Y$ to π' gives a path on which W is a non-collider outside Z, so this path is open given Z. In both cases X_a and Y are d-connected given Z, and (j, ℓ) is a predictive parent of i. □

Hidden-cause example. Let $U_t \sim \mathcal{N}(0, 1)$, $N_{k,t} \sim \mathcal{N}(0, s_k)$, all independent over k and t, and set $X_{1,t} = U_t + N_{1,t}$ and $X_{2,t} = U_{t-1} + N_{2,t}$. In the history of $X_{2,t}$, only $X_{1,t-1} = U_{t-1} + N_{1,t-1}$ shares a random component with $X_{2,t}$, and all other history coordinates are independent of $(X_{1,t-1}, X_{2,t})$. Hence $X_{2,t}$ given the history is Gaussian with mean $\mathbb{E}[U_{t-1} \mid X_{1,t-1}] = X_{1,t-1}/(1 + s_1)$ and constant variance $1 + s_2 - 1/(1 + s_1)$. Its score with respect to $x_{1,t-1}$ is $(1+s_1)^{-1}(X_{2,t} - \mu)/\sigma^2$, where μ and σ^2 are this mean and variance, so $H^1_{1,2} = (1+s_1)^{-2}/\sigma^2 > 0$ although $X_{1,t-1}$ is not a structural parent of $X_{2,t}$.

Instantaneous-effect example. Let $X_{1,t} = aX_{1,t-1} + N_{1,t}$ and $X_{2,t} = bX_{1,t} + N_{2,t}$ with $N_{k,t} \sim \mathcal{N}(0, s_k)$ independent. Substituting the first equation into the second gives $X_{2,t} = abX_{1,t-1} + bN_{1,t} + N_{2,t}$, where $bN_{1,t} + N_{2,t}$ is independent of the past. The conditional law of $X_{2,t}$ given the past is Gaussian with mean $abX_{1,t-1}$ and constant variance $b^2 s_1 + s_2$, so $H^1_{1,2} = (ab)^2/(b^2 s_1 + s_2) > 0$ whenever $ab \neq 0$, although $X_{2,t}$ has no lagged structural parent.

Hidden-regime mixture. Let the series be pooled over segments with regime index r_t, and consider histories that lie within one segment. By the law of total probability,

$$p_i(y \mid x_{t-L:t-1}) = \sum_r P(r_t = r \mid x_{t-L:t-1})\, p_i^{(r)}(y \mid x_{t-L:t-1}),$$

where $P(r_t = r \mid x_{t-L:t-1})$ is the probability of regime r given the history. Each component depends on the structural parents of its own regime, and the weights depend on every history coordinate that is informative about r_t. Unless these dependencies cancel exactly, the pooled law therefore depends on the union of the regime-specific structural parents and on the regime proxies.

The main text states that the lagged graph is not identifiable once A1 or A2 fails. The following two pairs of linear Gaussian systems make this concrete.

Proposition 3 (Observationally Equivalent Systems). *Let $s_1, s_2 > 0$, $|a| < 1$, and $ab \neq 0$.*

(a) The hidden-cause example above and the system $X_{1,t} = N'_{1,t}$, $X_{2,t} = \beta X_{1,t-1} + N'_{2,t}$, with $\beta = 1/(1+s_1)$, $N'_{1,t} \sim \mathcal{N}(0, 1+s_1)$, and $N'_{2,t} \sim \mathcal{N}(0, 1+s_2-\beta)$, induce the same distribution of the observed series. The second system satisfies A1–A4 and contains the structural edge $1 \xrightarrow{1} 2$, which the first system lacks.

(b) The instantaneous-effect example above induces the same distribution as a linear Gaussian system in which $X_{2,t}$ causes $X_{1,t}$ instantaneously and $X_{1,t-1}$ is a structural parent of $X_{2,t}$ with coefficient ab.

Proof of Proposition 3. **(a)** Both systems generate zero-mean Gaussian processes, so it suffices to match their covariance functions. In the hidden-cause system, $\mathrm{Var}(X_{1,t}) = 1 + s_1$, $\mathrm{Var}(X_{2,t}) = 1 + s_2$, $\mathrm{Cov}(X_{1,t-1}, X_{2,t}) = \mathrm{Var}(U_{t-1}) = 1$, and every other covariance between distinct coordinates of the observed process is zero, because distinct coordinates share no random component. In the second system, $\mathrm{Var}(X_{1,t}) = 1 + s_1$, $\mathrm{Cov}(X_{1,t-1}, X_{2,t}) = \beta(1+s_1) = 1$, and $\mathrm{Var}(X_{2,t}) = \beta^2(1+s_1) + 1 + s_2 - \beta = 1 + s_2$, using $\beta^2(1+s_1) = \beta$. All other covariances vanish because X_1 is independent over time and N'_2 is independent of everything else. The noise variance $1 + s_2 - \beta$ is positive since $\beta < 1$. The two covariance functions coincide, so the two observed processes have the same distribution. The second system has no unobserved inputs and no instantaneous edges, it uses one lag, and its linear Gaussian distribution with $\beta \neq 0$ is faithful to its graph, so it satisfies A1–A4.

(b) In the instantaneous-effect system, given the past, $(X_{1,t}, X_{2,t})$ is Gaussian with mean (ax, abx) for $x = X_{1,t-1}$ and covariance

$$\begin{pmatrix} s_1 & bs_1 \\ bs_1 & b^2 s_1 + s_2 \end{pmatrix}.$$

Factorizing this Gaussian in the order $X_{2,t}$, then $X_{1,t}$, gives the system

$$X_{2,t} = abX_{1,t-1} + E_{2,t}, \qquad X_{1,t} = cX_{2,t} + a(1-bc)X_{1,t-1} + E_{1,t},$$

with $c = bs_1/(b^2 s_1 + s_2)$, $E_{2,t} \sim \mathcal{N}(0, b^2 s_1 + s_2)$, and $E_{1,t} \sim \mathcal{N}(0, s_1 s_2/(b^2 s_1 + s_2))$. The residual $E_{1,t}$ of a Gaussian regression is independent of its regressor, so $E_{1,t}$, $E_{2,t}$, and the past are mutually independent. Both systems therefore have the same transition kernel from X_{t-1} to X_t, and with $|a| < 1$ the same stationary law, so their observed processes have the same distribution. In the second system, $X_{2,t}$ causes $X_{1,t}$ instantaneously and $X_{1,t-1}$ is a structural parent of $X_{2,t}$ with coefficient $ab \neq 0$, while in the first system $X_{2,t}$ has no lagged structural parent. □

A.2 Experiment setups

A.2.1 Data generation and simulation

Simulators. For the main grid, the stationary linear, high-dimensional, long-lag, noise, and linear-regime benchmarks come from the structural time-series simulator of Lawrence et al. (2020). The nonlinear mechanisms, the nonlinear regime family, and the main-grid systems with hidden causes or instantaneous effects come from our own generator, which discards a burn-in of 1,000 steps. We vary the property of interest relative to a linear reference system: the stationary and regime sample-size sweeps use 50K, 100K, and 1M samples; the high-dimensional, long-lag, noise, and latent-variable grids use 100K; and the instantaneous systems use 10K, 50K, and 100K. The count diagnostics below use 50K samples.

Variables and lags. Systems have N observed variables and maximum lag K, with $N = 10$ and $K = 5$ unless noted. We use $N \in \{10, 25, 50, 100\}$ to study high-dimensional systems and $K \in \{5, 10, 25, 50\}$ to study long-range systems, both with linear mechanisms.

Graph density. In the main grid, every variable has three parents in expectation over all lagged candidates, including its own lags, to obtain comparable sparsity across N and K. Our main-grid generator also requires at least two parents per variable; the count diagnostics use fixed graphs described below.

Nonlinear mechanisms. For Scale, Shape, and Transport, let g denote the normalized sum of the tanh-transformed parents of $X_{i,t}$, multiplied by the target's signed mechanism gain. We use five mechanisms. (1) Interaction, in which the parents enter the conditional mean through sinusoidal terms and cross-products between parents, with additive Gaussian noise. (2) Scale, $X_{i,t} = \exp(g/2)\,\varepsilon_{i,t}$, in which the parents set the conditional variance at zero mean. (3) Shape, in which $X_{i,t}$ comes from a bimodal logistic mixture with a probability that increases with g and from a standard Gaussian otherwise, both with zero mean and unit variance. (4) Transport, a standardized sinh–arcsinh transform of Gaussian noise whose skewness grows with g, at zero mean and unit variance. (5) Mixed, which assigns Interaction, Scale, Shape, or Transport to each variable.

Noise types. We consider Gaussian noise, which the linear reference system uses, uniform noise, and mixed noise. In the mixed-noise setting, each variable's noise family is drawn once from Gaussian, uniform, Laplace, and Student-t distributions with probabilities 0.4, 0.3, 0.2, and 0.1, and is held fixed over time. We also study unequal-variance noise with a small variance range from 0.5 to 5 and a large range from 0 to 10 (implemented with a positive lower bound of 10^{-8}).

Hidden regimes. We split each series into $S \in \{2, 5, 10\}$ contiguous regimes of equal length. In the linear family (Regular), every regime has maximum lag 5 and an independently sampled graph. In the nonlinear family (Extreme), every regime uses the Mixed mechanisms, draws its maximum lag from 1 to 5, edits the graph of the previous regime, and draws new signs and strengths for its mechanisms. The low-drift variants change only a small part of the graph between consecutive regimes and keep the mechanisms nearly unchanged.

Hidden causes. In the main grid, we add 3, 5, or 10 hidden autoregressive processes with coefficient magnitude 0.35 to a fixed linear system, and each process drives three observed variables at the next time step with loading magnitude 0.4. A paired system without hidden processes shares the observed mechanisms.

Instantaneous effects. In the main grid, we add five instantaneous edges with coefficients of magnitude 0.35 to 0.43 to linear systems with a fixed lagged graph, using 10K, 50K, and 100K samples.

Count diagnostics. Figure 3a uses 22 observed variables $W, V, M_1, \ldots, M_{10}, Y_1, \ldots, Y_{10}$, with $K = 2$. Here W, V, ten hidden sources H_j, and all innovations are independent standard Gaussian sequences, with $M_{j,t} = a_j W_{t-1} + b_j H_{j,t} + 0.2\varepsilon^M_{j,t}$ and $Y_{j,t} = 0.4V_{t-1} + b_j H_{j,t-1} + 0.2\varepsilon^Y_{j,t}$, where $a_j \in \{-1, 1\}$. Activating 0, 3, 5, or 10 sources sets their $b_j = 2$, leaving the other loadings zero

and the 20 direct observed edges fixed. These series are generated directly, with two initialization positions removed. Panel b uses 10 variables, $K = 5$, and lagged coefficient matrix $B_1 = 0.3I$, with all other lag matrices zero; a nested contemporaneous star has 0, 3, 6, or 9 edges of magnitude 1.5, after 1,000 burn-in steps. Both diagnostics share coefficients and innovations across counts within each data seed (242, 243, 244), and use optimizer seeds 0, 1, 2. They use windows of $K + 5$ lags and supervise the last six positions.

A.2.2 Training details and model architecture

For the synthetic experiments, we train autoregressive Transformers on node–lag tokens after per-variable z-score normalization with training statistics. We use an embedding dimension of 64, 4 attention heads, and either 1 ("shallow") or 4 ("deep") layers with pre-LayerNorm, residual connections, and a 2-layer ReLU feed-forward network, together with causal masking and node/time embeddings. Each output token parameterizes a five-component Gaussian mixture, and models are optimized with Adam (learning rate 10^{-3}, batch size 256) under the negative log-likelihood, with early stopping on the validation negative log-likelihood. We split every series, or every regime, chronologically into 80% training, 10% validation, and 10% test data. The reused linear benchmark families (stationary, high-dimensional, long-lag, noise, and linear-regime systems) use windows of K lags with all positions supervised. Newly generated nonlinear, latent-variable, and instantaneous-effect systems, including the count diagnostics and the paired linear control, use windows of $K + 5$ lags with the last six positions supervised. Validation evaluates the final position in the former protocol and all six supervised positions in the latter. The CNN applies one causal one-dimensional convolution of kernel width K, with ten channels per variable, followed by a ReLU and the same five-component mixture-head design. The readout uses 256 held-out windows. We use official implementations for PCMCI (Runge, 2020), VAR-LiNGAM (Hyvärinen et al., 2010), TCDF (Nauta et al., 2019), NTS-NOTEARS (Sun et al., 2021), and implementations for Granger causality from a collection repository of time-series causal discovery algorithms. The random baseline assigns uniform random scores to all candidates. Unless otherwise stated, we use default hyperparameters from each official implementation.

Evaluation protocol. For series with regime changes, we read the model on the held-out windows of each regime and score it against that regime's graph. The reference methods run on the full series and output one graph per series, which we score against the graph of every regime, and TCDF and the Granger tests output binary graphs whose AP reflects a single operating point. L-PCMCI and PCMCI$^+$ run on the full series with their lagged search restricted to the top half of the readout's candidates for each target, while all off-diagonal contemporaneous candidates remain allowed. Both use partial-correlation tests at $\alpha = 0.05$; L-PCMCI uses fixed search bounds `max_p_global` $= 3$ and `max_q_global` $= 20$. We compute AP from native $-\log_{10} p$ scores over the full lagged candidate set, clipping p-values at the smallest positive normal float64 value and assigning unsearched candidates score zero, without masking by the final graph. Synthetic results average three optimizer seeds for each of three data seeds, and error bars show the standard deviation across data seeds.

A.2.3 Compute resources

We implement the Transformer and CNN models in PyTorch and train them on NVIDIA H100 GPUs, and run the reference methods on CPUs. Every run contributing to the reported synthetic aggregates completed.

A.2.4 Existing assets, licenses, and terms of use

We use third-party assets only as cited simulators, benchmarks, pretrained models, or baseline implementations for research evaluation. We do not redistribute third-party datasets, model weights, or baseline source code with this manuscript. When code or data are released, the package should include the corresponding notices below and should link to the original source for assets whose redistribution terms are not explicit.

- **CDML-NeurIPS2020 simulator.** We use the structural time-series data-generating process of Lawrence et al. (2020) to generate synthetic benchmarks. The public repository is `https://github.com/causalens/cdml-neurips2020` and is listed under the

BSD-3-Clause license. We use the simulator to generate new synthetic datasets under our experimental settings and do not redistribute the original downloaded experiment datasets.

- **Tigramite/PCMCI.** We use Tigramite implementations of PCMCI/PCMCI$^+$ (Runge et al., 2017; Runge, 2020). The public repository is `https://github.com/jakobrunge/tigramite` and is listed under the GNU GPL-3.0 license. We use it as an external baseline/refinement package and do not redistribute its source code.
- **CausalNex/DYNOTEARS.** We use DYNOTEARS (Pamfil et al., 2020) via the CausalNex package. The package page `https://pypi.org/project/causalnex/` lists the license as Apache-2.0. We use it as an installed baseline implementation and do not redistribute its source code.
- **LiNGAM package.** We use the Python `lingam` package for VAR-LiNGAM-style baselines (Hyvärinen et al., 2010). The public repository is `https://github.com/cdt15/lingam` and is listed under the MIT license. We use it as an installed baseline implementation and do not redistribute its source code.
- **TCDF.** We use TCDF (Nauta et al., 2019) as a neural time-series causal-discovery baseline. The public repository is `https://github.com/M-Nauta/TCDF` and is listed under the GNU GPL-3.0 license. We use it as an external baseline and do not redistribute its source code.
- **NTS-NOTEARS.** We use the official NTS-NOTEARS implementation (Sun et al., 2021). The public repository is `https://github.com/xiangyu-sun-789/NTS-NOTEARS` and is listed under the Apache-2.0 license. We use it as an external baseline and do not redistribute its source code.
- **CausalTime benchmark.** The real-world feasibility experiment uses the CausalTime benchmark (Cheng et al., 2023). The official website is `https://www.causaltime.cc/`, and the dataset page is `https://www.causaltime.cc/dataset/`. We did not find an explicit license statement on the public dataset page or linked code directory at the time of writing; therefore, we use CausalTime only for evaluation, cite the original paper, do not redistribute the dataset, and direct users to the official download page and any terms maintained by the original authors.

A.3 LIMITATIONS

Theorem 1 characterizes the structural parents through the conditional laws in the population, and a trained model inherits it to the extent that the input derivatives of its log-density match those of the true conditional laws. We verify this match edge by edge on systems whose true score energy can be computed, and a finite-sample or optimization guarantee for the trained readout remains open. The identification target is the lagged parent set of each variable within a stationary regime, so the mechanisms, the effect sizes, and the contemporaneous edges are left to other methods. Our synthetic systems come from two generators, with 10 variables in most settings and up to 100 variables and 50 lags in linear systems, and the diagnostics for hidden causes and instantaneous effects use constructed systems with fixed graphs and three data seeds. Supplying the regime identity adds an input token to the model, so the paired comparison changes the available information and the input format together. We evaluate screened L-PCMCI and PCMCI$^+$ refinements empirically. The original methods target equivalence classes under their respective assumptions; identifiability of the screened, bounded combination is not established here. On the real-world benchmarks, the identifying conditions cannot be checked. We train each Transformer from scratch on one system, and pretrained forecasters with other objectives, such as point or quantile forecasts, remain to be examined with the same readout.

A.4 BROADER IMPACTS

This work connects autoregressive prediction with causal structure learning, offering a scalable approach to discovering temporal causal relationships. Potential positive applications include scientific discovery in climate, neuroscience, healthcare, engineering, and environmental time-series modeling, where identifying candidate causal dependencies can support better downstream analysis and inform interventions or policies. At the same time, inferred structures from observational time series should not be treated as definitive causal knowledge without further validation. Misuse or overconfidence in such graphs could lead to incorrect conclusions, especially in high-stakes domains such as medicine, infrastructure, and public policy. Practitioners should therefore validate findings with domain expertise and, where possible, experimental or interventional evidence. The paper also mitigates this risk by stating the identifiability assumptions and the failure modes under latent confounding, contemporaneous effects, and model misspecification.

A.5 DISCUSSIONS

A.5.1 TRANSFORMER-BASED CAUSAL DISCOVERY

Recently, the transformer architecture has been proposed as a powerful alternative for discovering causal structures from data. CausalFormer combines additional modules such as multi-kernel causal convolution and a decomposition-based detector to extract the global causal graph from time-series data (Kong et al., 2024). Symmetry-aware transformers preserve temporal asymmetry for causal discovery in financial time series (Zheng & Liu, 2025). Besides unsupervised causal discovery, transformers are applied to learn a broad range of predefined causal relationships with synthetic data (Ke et al., 2022) and domain-specific causal graphs with interventional data (Wang & Kording, 2022) in a supervised manner. BCNP (Bayesian Causal Neural Process) meta-learns Bayesian networks by training a transformer-like encoder-decoder architecture on synthetic data to learn priors and adjusting the posterior at test time (Dhir et al., 2024). These methods add graph-specific modules or learn from graph supervision, while we train a standard decoder-only Transformer only for next-step prediction and characterize when the predictive parents it learns are the structural parents.

A.5.2 SCALABILITY IN CAUSAL DISCOVERY

Scalability is crucial for causal discovery algorithms in real-world scenarios. Factors such as sample size, number of variables, graph density, functional forms, and data quality (e.g., latent confounders and missing values) all put pressure on scalability. Previous search methods tackle combinatorial challenges in large graphs via computational optimization, such as device acceleration (Zarebavani et al., 2019; Akinwande & Kolter, 2024) and algorithmic techniques like divide-and-conquer and dynamic programming (Shah et al., 2024; Ramsey et al., 2017; Hyttinen et al., 2017). In contrast, continuous-optimization and Neural Granger methods (Pamfil et al., 2020; Nauta et al., 2019; Tank et al., 2021) recover graphs at model convergence within a fixed optimization budget, and are not strictly limited by combinatorial search. However, these methods are trained on a homogeneous dataset and recover a corresponding population-level graph; they do not naturally accommodate heterogeneity and are often jointly optimized with heuristics that are sensitive to hyperparameters (e.g., ℓ_1 penalties). Autoregressive training offers another route to scale: across our synthetic settings, we keep the architecture and optimizer fixed except in the architecture ablations, with the two supervision protocols specified in Appendix A.2.2. The readout needs only the gradients of the trained model on held-out windows; the separately tuned CausalTime models are described in Appendix A.6.4. Whether pretrained forecasters with other objectives expose the same structure remains open (Appendix A.3).

A.5.3 STRUCTURAL CAUSALITY AND GRANGER CAUSALITY

Our framework is grounded in *structural causal models* (SCMs), where $X_j \to Y$ means X_j appears in the structural equation for Y—a mechanistic notion independent of statistical tests (Pearl, 2009; Peters et al., 2017). White & Lu (2010) linked Granger and structural causality: under conditional exogeneity ($U_t \perp X_{S^c} \mid X_S$), Granger noncausality equals *structural noncausality almost surely*. However, without faithfulness, edge cases exist where true causes are undetectable, falling short of exact identification. We strengthen this in three ways. First, by imposing faithfulness, we obtain an *exact* characterization: $H_j = 0 \iff X_j \notin S$ (Theorem 2). Second, we provide a concrete criterion, the score energy $H_j = \mathbb{E}[(\partial_{x_j} \log p(Y|X))^2]$, directly computable from learned densities. Third, this criterion is defined on the conditional law that autoregressive training learns, so a trained model gives direct access to it (Section 3.3). The logical direction runs from structural causality to predictive dependence. Under the identifying conditions, structural identification and distributional Granger causality coincide, and the guarantee derives from the structural causal model.

A.5.4 RELATIONSHIPS TO PREVIOUS THEORY RESULTS

Classical Granger causality (Granger, 1969) tests predictive improvement in linear VARs but lacks structural interpretation. Subsequent work pursued structural identification through diverse assumptions: LiNGAM (Hyvärinen et al., 2010) achieves full DAG identification via non-Gaussianity but requires linearity; PCMCI (Runge, 2020) performs conditional independence testing under Causal Markov, faithfulness, and causal sufficiency; DYNOTEARS (Pamfil et al., 2020) enables continuous

optimization for linear time-series DAGs; Neural Granger Causality (Tank et al., 2021) extends to nonlinear functions but establishes only sufficient conditions. The closest antecedent is White & Lu (2010), who proved under conditional exogeneity ($D^t \perp U^t \mid Y^{t-1}, X^t$) that Granger non-causality is equivalent to structural non-causality almost surely—that is, structural causality may exist on measure-zero sets without manifesting in conditional distributions. Our Theorem 1 strengthens this by imposing faithfulness to eliminate such pathological cases, yielding an exact characterization: $H_j = 0 \Leftrightarrow j \notin S$. Moreover, while White & Lu (2010) relies on conditional independence testing, we provide a computable criterion, score energy, that directly connects structural identification to the prediction objective, enabling practical extraction via gradient-based attribution in transformers. A separate line uses score functions for DAG identification in i.i.d. settings (Rolland et al., 2022; Montagna et al., 2023; Sanchez et al., 2022). These methods rely on the *variance* of Hessian diagonals to identify leaf nodes, a necessary step when topological order is unknown. In our time-series setting, assumption A2 (no instantaneous effects) provides this order for free, allowing us to use the simpler first-order score energy H_j directly. This distinction enables natural integration with autoregressive training: the conditional $p(X_t \mid X_{<t})$ is precisely what next-step prediction optimizes, avoiding the need for specialized machinery.

A.6 Additional Experiments

A.6.1 Edge-level agreement between fitted and true score energy

For three systems whose conditional densities are known in closed form, namely a paired linear control, Interaction, and Scale, we compute the true score energy of every candidate edge and compare it with the readout of the trained model on the same 256 held-out windows (Figure 5). Over 13,500 candidate edges from three data seeds and three optimizer seeds per system, the fitted energy of a structural parent has a median ratio of 0.966 to 0.991 to its true energy, and its rank correlation with the true energy is 0.90 to 0.99 across the three systems. The true energy of every other candidate is exactly zero, and the readout separates these candidates from the structural parents with an AP of 1.000 in 25 of 27 fits and at least 0.978 in the other two.

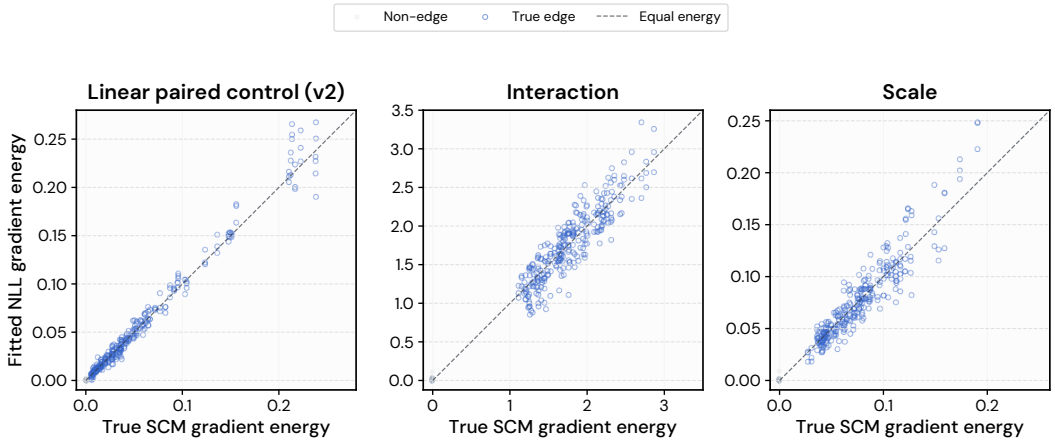


Figure 5: **Fitted and true score energy of every candidate edge.** Each point is one candidate edge in one fit, and the dashed line marks equal energy. Structural parents lie close to the line, and the other candidates stay near zero in both energies.

A.6.2 Hidden regimes across families and sample sizes

Figure 6 reports the Regular and Extreme hidden-regime settings of the main grid, with 2, 5, and 10 regimes and 50K, 100K, and 1M samples.

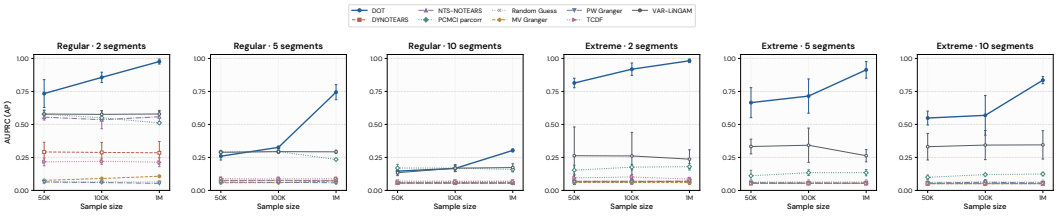


Figure 6: **Hidden regimes across families and sample sizes.** AP of the Transformer readout (DOT) and of the reference methods when the regime identity is hidden, for linear systems with a fixed maximum lag (Regular, left three panels) and nonlinear systems with regime-specific lags (Extreme, right three panels). The readout is scored within each regime, and each reference outputs one graph per series, which is scored against the graph of every regime. AP falls as the number of regimes grows and rises with the sample size, most strongly in the nonlinear family, where 10 regimes reach 0.835 at 1M samples against 0.303 in the linear family. The linear family uses the training protocol with all positions supervised (Appendix A.2.2). Mean $\pm$ SD over three data seeds, with true edges making up 5.2% to 6.0% of the candidates.

A.6.3 HIDDEN REGIMES WITH SMALL CHANGES BETWEEN REGIMES

In the hidden-regime systems of Section 4.2, consecutive regimes can differ in their graph, their maximum lag, and the sign and strength of their mechanisms. Real systems often change only part of their mechanisms from one regime to the next, so we also generate low-drift series, in which consecutive regimes differ in a small part of the graph and keep their mechanisms nearly unchanged (Figure 7). With ten hidden regimes and 100K samples, AP rises from 0.569 to 0.937 in the nonlinear family and from 0.167 to 0.599 in the linear family, and with 1M samples the low-drift series reach 0.945 and 0.886. This agrees with Proposition 2, since regimes that share most of their structural parents add few extra predictive parents when they are pooled.

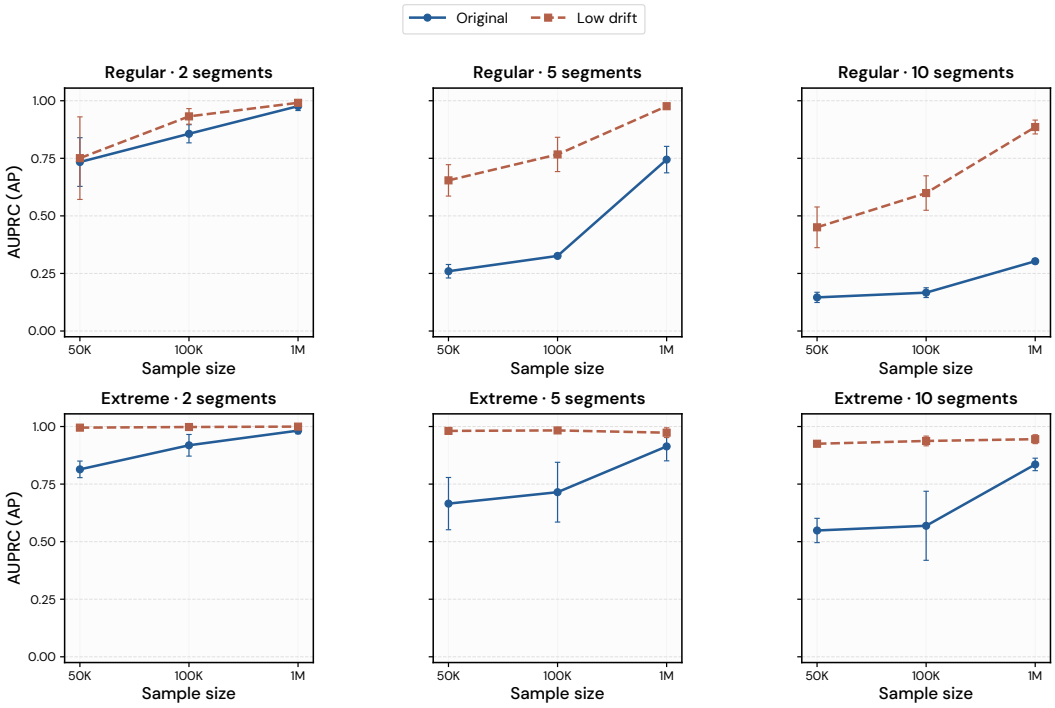


Figure 7: **Hidden regimes with small changes between regimes.** AP of the Transformer readout when the regime identity is hidden, for the regimes of Section 4.2 (Original) and for regimes that change little from one to the next (Low drift), in linear (Regular) and nonlinear (Extreme) systems with 2, 5, and 10 regimes. Mean $\pm$ SD over three data seeds.

A.6.4 REAL-WORLD BENCHMARKS

CausalTime provides air-quality, traffic, and medical series with annotated summary graphs (Cheng et al., 2023). We aggregate the readout over lags and score it against the summary graph by AUROC and trapezoidal AUPRC. On these data, we use a lighter Transformer with one layer, one attention head, and one token per time step, trained with a sparsity penalty on its input gradients, together with a two-layer MLP that uses the same readout. We use 15-step windows and the maximum score over the ten most recent lags for each summary edge. The 480 trajectories are split into 384/48/48 training/validation/test trajectories, with training-set standardization, batch size 128, and early stopping on validation NLL plus the penalty (patience 10). The Transformer supervises the last six positions and the MLP the final position; both read out 256 held-out windows. Table 1 compares both predictors with the 14 methods reported by Zhou et al. (2024). The Transformer performs comparably to the strongest of these methods on traffic and medical records, at the upper end of their range on traffic, and the MLP does so on air quality.

Table 1: Benchmark results on three datasets. The three highest values in each column are marked in **bold**, with an underline, and with a dagger (†), in that order. Baseline values are taken from Zhou et al. (2024); our rows report mean $\pm$ SD over optimization seeds 1–5. AUPRC is the trapezoidal area under the precision–recall curve.

Methods	AQI		Traffic		Medical	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GC	0.4538 ± 0.0377	0.6347 ± 0.0158	0.4191 ± 0.0310	0.2789 ± 0.0018	0.5737 ± 0.0338	0.4213 ± 0.0281
SVAR	0.6225 ± 0.0406	0.7903 ± 0.0175	0.6329 ± 0.0047	0.5845 ± 0.0021	0.7130 ± 0.0188	0.6774 ± 0.0358
N.NTS	0.5729 ± 0.0229	0.7100 ± 0.0228	0.6329 ± 0.0335	0.5770 ± 0.0542	0.5019 ± 0.0682	0.4567 ± 0.0162
PCMCI	0.5272 ± 0.0744	0.6734 ± 0.0372	0.5422 ± 0.0737	0.3474 ± 0.0581	0.6991 ± 0.0111	0.5082 ± 0.0177
Rhino	0.6700 ± 0.0983	0.7593 ± 0.0755	0.6274 ± 0.0185	0.3772 ± 0.0093	0.6520 ± 0.0212	0.4897 ± 0.0321
CUTS	0.6013 ± 0.0038	0.5096 ± 0.0362	0.6238 ± 0.0179	0.1525 ± 0.0226	0.3739 ± 0.0297	0.1537 ± 0.0039
CUTS+	$0.8928 \pm 0.0213^\dagger$	$0.7983 \pm 0.0875^\dagger$	0.6175 ± 0.0752	$\underline{0.6367 \pm 0.1197}$	$\mathbf{0.8202 \pm 0.0173}$	0.5481 ± 0.1349
NGC	0.7172 ± 0.0076	0.7177 ± 0.0069	0.6032 ± 0.0056	0.3583 ± 0.0495	0.5744 ± 0.0096	0.4637 ± 0.0121
NGM	0.6728 ± 0.0164	0.4786 ± 0.0196	0.4660 ± 0.0144	0.2826 ± 0.0098	0.5551 ± 0.0154	0.4697 ± 0.0166
LCCM	0.8565 ± 0.0653	$\mathbf{0.9260 \pm 0.0246}$	0.5545 ± 0.0254	0.5907 ± 0.0475	$\underline{0.8013 \pm 0.0218}$	$\mathbf{0.7554 \pm 0.0235}$
eSRU	0.8229 ± 0.0317	0.7223 ± 0.0317	0.5987 ± 0.0192	0.4886 ± 0.0338	$0.7559 \pm 0.0365^\dagger$	$\underline{0.7352 \pm 0.0600}$
SCGL	0.4915 ± 0.0476	0.3584 ± 0.0281	0.5927 ± 0.0553	0.4544 ± 0.0315	0.5019 ± 0.0224	0.4833 ± 0.0185
TCDF	0.4148 ± 0.0207	0.6527 ± 0.0087	0.5029 ± 0.0041	0.3637 ± 0.0048	0.6329 ± 0.0384	0.5544 ± 0.0313
JRNGC-F	$\mathbf{0.9279 \pm 0.0011}$	0.7828 ± 0.0020	$\underline{0.7294 \pm 0.0046}$	$0.5940 \pm 0.0067^\dagger$	0.7540 ± 0.0040	$0.7261 \pm 0.0016^\dagger$
DOT (Ours)	0.8040 ± 0.0084	0.6571 ± 0.0089	$\mathbf{0.7915 \pm 0.0172}$	$\mathbf{0.6446 \pm 0.0118}$	0.7554 ± 0.0085	0.6984 ± 0.0109
MLP (Ours)	$\underline{0.9190 \pm 0.0046}$	$\underline{0.8281 \pm 0.0066}$	$0.7283 \pm 0.0192^\dagger$	0.5642 ± 0.0369	0.6457 ± 0.0197	0.5852 ± 0.0378