

# Causal Discovery in Multivariate Time Series through Mutual Information Featurization

From Statistical Tests to Pattern Recognition

---

Gian Marco Paldino, Gianluca Bontempi

Winter School on Causality and Explainable AI

October 20-24, 2025

Machine Learning Group, Computer Science Department, Université Libre de Bruxelles

(These slides are available)

# The Goal: Why Discover Causal Links in Time Series?

- **Climate Science:** Is rising CO2 *causing* global temperature anomalies?
- **Neuroscience:** How do brain regions *functionally influence* each other after a stimulus?
- **Economics:** Does a change in interest rates *cause* a change in inflation with a certain lag?



The bottom line: **correlation is not causation.**

To truly understand, predict, and **intervene**, we need a causal graph.

# Modeling Temporal Systems: Autoregressive Models

## Nonlinear Autoregressive (NAR) Models

We model each variable  $z_j$  at time  $t + 1$  as a function of the past values of itself and its causal parents, plus some noise.

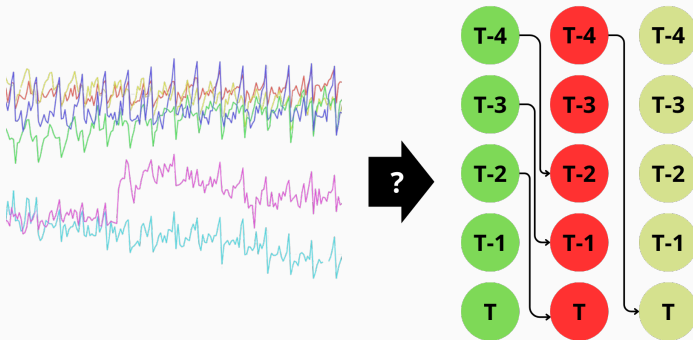
$$z_j^{(t+1)} = f_j(\text{past values of parents of } j) + \epsilon_j^{(t)}$$

For example, if  $z_i$  causes  $z_j$ :

$$z_j^{(t+1)} = f_j(z_j^{(t)}, z_j^{(t-1)}, \dots, z_i^{(t)}, z_i^{(t-1)}, \dots) + \epsilon_j^{(t)}$$

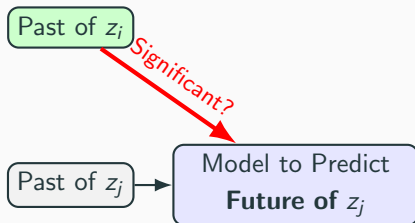
# How do we go from observations to graph? - I

We can “unroll” these equations over time to create a Directed Acyclic Graph (DAG) that represents the causal flow.



... so, how do we go from observations to DAG? How is the literature doing this?

# The Traditional Toolbox: Approach 1 - Granger Causality



Time series  $z_i$  “**Granger-causes**”  $z_j$  if adding the past of  $z_i$  to a model using the past of  $z_j$  significantly **improves the prediction** of  $z_j$ ’s future.

## Limitations

- (1) Improving predictions  $\neq$  Causing (eg. ice cream sales, shark attacks).
- (2) Mostly handles *linear* relationships (eg. VAR)

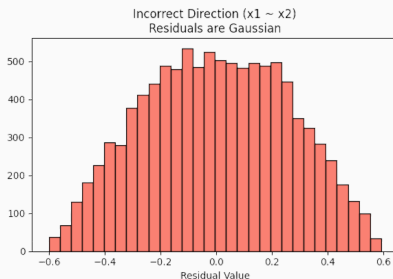
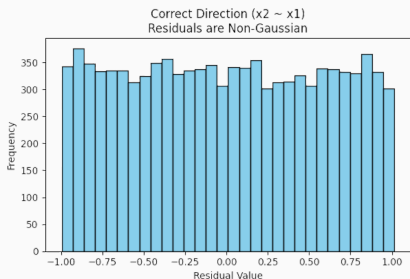
# The Traditional Toolbox: Approach 2 - Noise-Based

True Direction:  $x_1 \rightarrow x_2$

$$x_2 = f(x_1) + \text{Noise}_2$$

Wrong Direction:  $x_2 \rightarrow x_1$

$$x_1 = g(x_2) + \text{Noise}_1$$

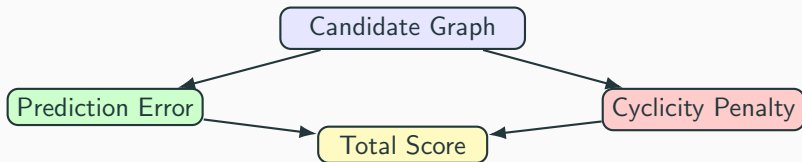


The noise of a true causal model should be statistically independent of the cause. Does not hold for the reverse direction.

## Limitations

Methods like VarLiNGAM [Hyvärinen et al., 2010] require the noise being non-Gaussian, and often assume linearity (VAR).

## The Traditional Toolbox: Approach 3 - Score-Based



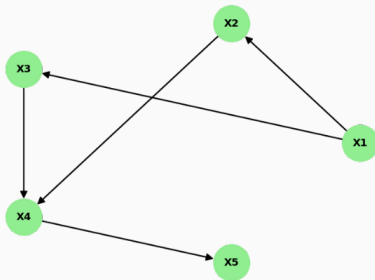
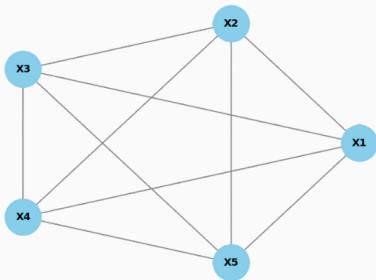
$$\underset{\mathbf{W}}{\text{minimize}} \quad \underbrace{\frac{1}{2n} \|\mathbf{X} - \mathbf{XW}\|_F^2}_{\text{Data Fit (Least Squares)}} + \underbrace{\lambda h(\mathbf{W})}_{\text{Acyclicity Constraint}}$$

Frame causal discovery as an optimization problem. Search for the graph structure that best explains the data by minimizing a score function.

### Limitations

NOTEARS [Pamfil et al., 2020] assumes a linear model structure (Least Squares) and requires the graph to be acyclic.

## The Traditional Toolbox: Approach 4 - Constraint-Based



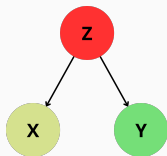
From a fully connected graph, we use Conditional Independence (CI) tests to remove links that are merely correlational.

### Limitations

Methods like PCMCi [Runge et al., 2019] heavily depend on the CI tests giving a clear “yes/no” answer.



# Conditional Independence and Mutual Information



Is  $\mathbf{X}$  independent of  $\mathbf{Y}$ ? No. ( $X \not\perp Y$ )

Is  $\mathbf{X}$  independent of  $\mathbf{Y}$  *given*  $\mathbf{Z}$ ? Yes. ( $X \perp Y \mid Z$ )

Another way to express this is *mutual information* (MI):

$$I(X; Y \mid Z) = 0 \iff X \perp Y \mid Z$$

$$I(z_i; z_j \mid z_k) = \mathbb{E}_{z_i, z_j, z_k} \left[ \log \frac{p(z_i, z_j \mid z_k)}{p(z_i \mid z_k) p(z_j \mid z_k)} \right]$$

MI quantifies the “amount of information” obtained about one RV by observing the other RV. It’s **non-parametric**, meaning it makes **no assumptions** about the shape of the data or the type of relationship.

## For later

We don’t know the distributions  $p(\cdot)$ , we’ll need to estimate  $I(\cdot)$

# Conditional Independence Tests

Test Name	Primary Data Type	Reference
Chi-squared ( $\chi^2$ ) Test	Categorical	[Pearson, 1900]
<b>G-test (Likelihood-Ratio)</b>	<b>Categorical</b>	<b>[Wilks, 1938]</b>
Cochran-Mantel-Haenszel	Stratified Categorical	[Cochran, 1954]
Partial Correlation	Continuous (Linear)	[Fisher, 1924]
Kernel CI Test (KCI)	Continuous / Mixed	[Zhang et al., 2011]
Generalized Covariance Measure	Continuous	[Shah & Peters, 2020]
Permutation-based CI Test	General (Non-parametric)	[Doran et al., 2014]

Example: the test statistic  $G$  is directly proportional to the mutual information:

$$G = 2 \cdot N \cdot I(\mathbf{z}_i; \mathbf{z}_j)$$

Where:

- $G$  is the G-statistic.
- $N$  is the total number of samples.
- $I(\mathbf{z}_i; \mathbf{z}_j)$  is the mutual information between the two variables.

# The Reality Leap

In the real world of complex time series, the measured CMI is **almost always a small, non-zero number**, even for non-causal pairs. Why?

- A **limited sample** introduces noise.
- **Hidden confounders** create residual information pathways.
- **Complex relationships** create “information leaks”.

## The idea

- What if the non-zero value of  $I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{z}_k)$  isn't just statistical noise, but a quantitative measure of residual information?
- What if these measures are fundamentally related to the **flow of information** through the system's temporal graph?

# Quantifying asymmetry in Information Flow

Can we somehow quantify asymmetry in the flow of information?

Something we can do is **counting open paths\***.

## D-separation

A path between two nodes  $X$  and  $Y$  is said to be **blocked** by a set of conditioning nodes  $\mathbf{S}$  if the path contains a chain or a fork where the middle node  $B$  is in  $\mathbf{S}$ , or if the path contains a collider where the middle node  $B$  is **not** in  $\mathbf{S}$ . We say that  $X$  and  $Y$  are **d-separated** if all paths between them are blocked.

## Open paths

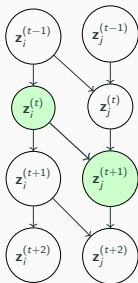
An open path represents a potential flow of information. Maybe a **quantitative asymmetry** ( $\Delta$ ) in path counts will manifest as a detectable asymmetry in the information flow.

---

\*To correctly count all open paths between two nodes, one must enumerate all simple paths in the undirected skeleton of the graph and then check each path individually against the d-separation rules on the original directed graph

# Visualizing the Asymmetry: From Obvious to Quantitative - I

We unroll a DAG to  $t - 20$ , and we count cause  $\rightarrow$  effect open paths.

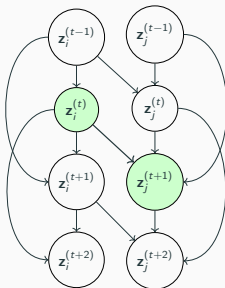


**Simple Case**

Forward ( $\rightarrow$ ) 1

Backward ( $\leftarrow$ ) 0

**Difference ( $\Delta$ ) 1**

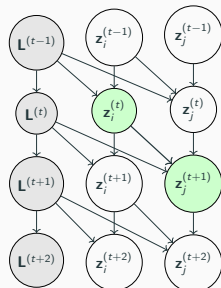


**Complex Case**

Forward ( $\rightarrow$ ) 370

Backward ( $\leftarrow$ ) 312

**Difference ( $\Delta$ ) 58**



**Confounded case**

Forward ( $\rightarrow$ ) 1186

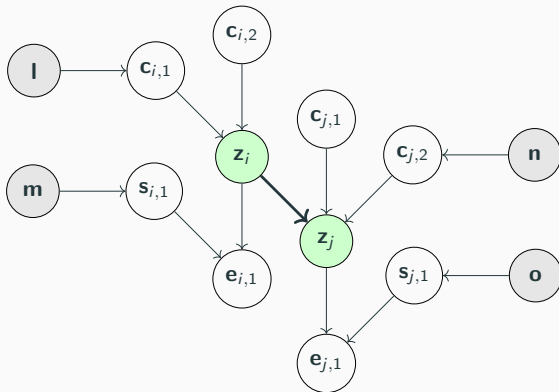
Backward ( $\leftarrow$ ) 898

**Difference ( $\Delta$ ) 288**

Is this asymmetry reflected in the flow of information?  
e.g. Could the distribution of CMI terms be asymmetric?

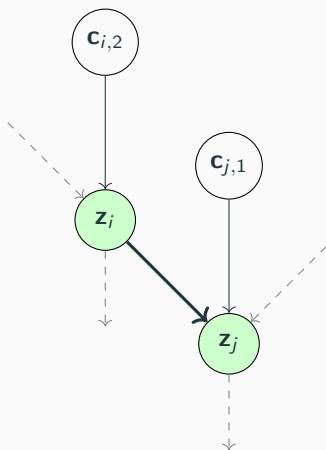
## Introducing Markov Blankets

We need to introduce the concept of Markov Blanket (MB).



The MB of  $z_i$  is a **set of variables** that renders  $z_i$  **conditionally independent** of all other variables in the system. It includes causes  $c_{i,j}$ , effects  $e_{i,j}$ , spouses  $s_{i,j}$ . **l, m, n, o** represent other parts of the graph.

# The Asymmetry of a Causal Link - I



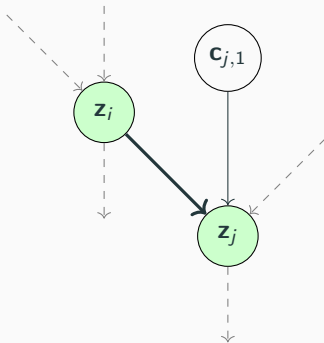
## The POV from the Cause ( $z_i$ )

- Consider the path:  
 $c_{i,2} \rightarrow z_i \rightarrow z_j$ . This is a **chain**.
- If we observe (condition on) the variable  $z_i$ , the path is **blocked**.
- Therefore, its parent  $c_{i,2}$  becomes **independent** of its effect  $z_j$ .
- $(c_{i,2} \perp z_j \mid z_i)$ , or  
 $(\mathbf{Pa}_i \perp z_j \mid z_i)$ , where  $\mathbf{Pa}$  is *parent*.

## The Asymmetry of a Causal Link - II

### The POV from the Effect ( $z_j$ )

- Consider the path:  
 $z_i \rightarrow z_j \leftarrow c_{j,1}$ . This is a **collider**.
- Conditioning on the collider  $z_j$  actually **opens** the path.
- Therefore, its parent  $z_i$  becomes **dependent** on its other parent  $c_{j,1}$ .
- $(c_{j,1} \not\perp z_i \mid z_j)$ , or  
 $(\mathbf{Pa}_j \not\perp z_i \mid z_j)$ , where  $\mathbf{Pa}$  is *parent*.





## The Asymmetry of a Causal Link - III

We have the following:

- $\mathbf{Pa}_i \perp \mathbf{z}_j \mid \mathbf{z}_i$ , or using MI,  $I(\mathbf{Pa}_i; \mathbf{z}_j \mid \mathbf{z}_i) = 0$
- $\mathbf{Pa}_j \not\perp \mathbf{z}_i \mid \mathbf{z}_j$ , or using MI,  $I(\mathbf{Pa}_j; \mathbf{z}_i \mid \mathbf{z}_j) > 0$

It's the exact same quantity measure by replacing  $i$  with  $j$ , and it's  $= 0$  in one case and  $> 0$  in the other case. **We found asymmetric measures.**

We call them *descriptors* [Bontempi et al., 2015]:

$$\begin{array}{ll} I(\mathbf{z}_i; \mathbf{c}_{j,k} \mid \mathbf{z}_j) > 0 & I(\mathbf{z}_j; \mathbf{c}_{i,k} \mid \mathbf{z}_i) = 0 \\ I(\mathbf{e}_{i,k}; \mathbf{c}_{j,k} \mid \mathbf{z}_j) > 0 & I(\mathbf{e}_{j,k}; \mathbf{c}_{i,k} \mid \mathbf{z}_i) = 0 \\ I(\mathbf{c}_{i,k}; \mathbf{c}_{j,k} \mid \mathbf{z}_j) > 0 & I(\mathbf{c}_{j,k}; \mathbf{c}_{i,k} \mid \mathbf{z}_i) = 0 \\ I(\mathbf{z}_i; \mathbf{c}_{j,k}) = 0 & I(\mathbf{z}_j; \mathbf{c}_{i,k}) > 0 \end{array}$$

So far **not useful**: they presuppose the ability to distinguish causes, effects, spouses, which is exactly the problem we want to solve.

# The Innovation: Population-Level Asymmetry

Instead of pre-selecting variables, we compute CMI terms for the entire Markov Blanket and compare the statistical distributions of the results.

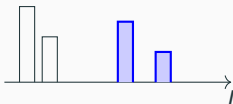
**True Direction:**  $\mathbf{z}_i \rightarrow \mathbf{z}_j$

We form the population  $D(i, j)$ :

$$\{I(\mathbf{z}_i; \mathbf{m}_{j,k} | \mathbf{z}_j)\}_{k=1, \dots, |\mathbf{M}_j|}$$

$\mathbf{m}_{j,k}$  is any member of  $\mathbf{z}_j$ 's MB.

- Because of colliders, some of these terms will be **greater than zero**.



Distribution is **skewed positive**

**Reverse Direction:**  $\mathbf{z}_j \rightarrow \mathbf{z}_i$

We form the population  $D(j, i)$ :

$$\{I(\mathbf{z}_j; \mathbf{m}_{i,k} | \mathbf{z}_i)\}_{k=1, \dots, |\mathbf{M}_i|}$$

$\mathbf{m}_{i,k}$  is any member of  $\mathbf{z}_i$ 's MB.

- Because of chains, many of these terms will be **close to zero**.



Distribution is **concentrated at zero**

## Conclusion:

We used nonzero CMI terms at a population-level to get **asymmetry**.

# Who Compares the Distributions?

We call these *populations of descriptors*:

$$D_1(i, j) = \left\{ I \left( \mathbf{z}_i; \mathbf{m}_{j, k_j} \mid \mathbf{z}_j \right), k_j = 1, \dots, K_j \right\}$$

$$D_1(j, i) = \left\{ I \left( \mathbf{z}_j; \mathbf{m}_{i, k_i} \mid \mathbf{z}_i \right), k_i = 1, \dots, K_i \right\}$$

$$D_2(i, j) = \left\{ I \left( \mathbf{m}_{i, k_i}; \mathbf{m}_{j, k_j} \mid \mathbf{z}_j \right), k_i = 1, \dots, K_i, k_j = 1, \dots, K_j \right\}$$

$$D_2(j, i) = \left\{ I \left( \mathbf{m}_{j, k_j}; \mathbf{m}_{i, k_i} \mid \mathbf{z}_i \right), k_i = 1, \dots, K_i, k_j = 1, \dots, K_j \right\}$$

$$D_3(i, j) = \left\{ I \left( \mathbf{z}_i; \mathbf{m}_{j, k_j} \right), k_j = 1, \dots, K_j \right\}$$

$$D_3(j, i) = \left\{ I \left( \mathbf{z}_j; \mathbf{m}_{i, k_i} \right), k_i = 1, \dots, K_i \right\}$$

Their computation will require *estimation of MBs*.

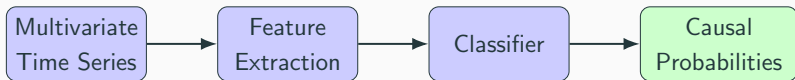
We end up with a vector, where  $\mathcal{Q}$  returns sample quantiles.

$$\mathbf{x} = [\mathcal{Q}(\hat{D}_1(i, j)), \mathcal{Q}(\hat{D}_1(j, i)), \mathcal{Q}(\hat{D}_2(i, j)), \mathcal{Q}(\hat{D}_2(j, i)), \mathcal{Q}(\hat{D}_3(i, j)), \mathcal{Q}(\hat{D}_3(j, i))]$$

## The Universality Hypothesis

The pattern that defines this 'causal footprint' is a **fundamental and transferable signal**. This means we could potentially **learn** the signature of causality **from diverse synthetic data** (e.g. using a classifier).

# Operationalizing the Hypothesis: The TD2C Framework



## Training

1. Generate diverse synthetic data
2. For every causal pair ( $z_i \rightarrow z_j$ ):
  - Extract *descriptors* vector  $\mathcal{X}_{train}$
  - Append label 1
3. For every noncausal pair ( $z_i \nrightarrow z_j$ ):
  - Extract *descriptors* vector  $\mathcal{X}_{train}$
  - Append label 0
4. Train a classifier

## Inference

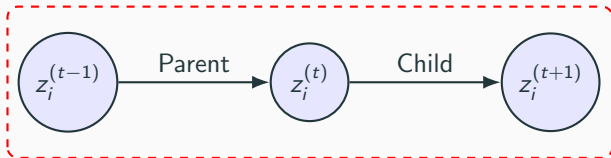
For every pair of interest:

1. Extract  $\mathcal{X}_{test}$
2. Predict  $\hat{y}$

# We still need estimates

## 1. MB Estimator

- Instead of a full MB search, we leverage the structure of time.
- *First-order Markov self-causality assumption.*



The defined Markov Blanket for  $z_i^{(t)}$

## 2. Non-Parametric Mutual Information (MI) Estimator

- K-Nearest Neighbor based KSG estimator [Kraskov et al., 2004].
- Allows to detect non-linear dependencies
- We used the decomposition  $I(X; Y | Z) = I(X, Z; Y) - I(Z; Y)$ .

# Building the “Causal Signature”: A Rich Vocabulary

## We can do feature engineering!

To capture the subtle causal signature, we engineer a diverse set of features for each potential link  $z_i \rightarrow z_j$ . Descriptors families:

- **Information-Theoretic:** what we've seen so far
- **Noise-Based:** inspired by the literature
- **Higher-Order Moments:** asymmetrical by definition
- **Linear Descriptors:** sometimes simplicity helps

## A novel Information Theoretical descriptor

We add a descriptor based on Transfer Entropy. It asks: “How much new information does the cause’s past provide about the effect’s present, given the effect’s own past?”

For a link  $z_i \rightarrow z_j$ , we compute:

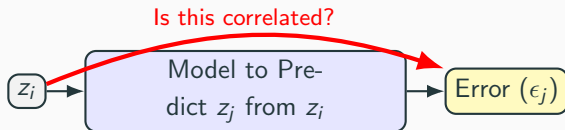
- **Forward TE:**  $I(z_i^{(t-1)}; z_j^{(t)} \mid z_j^{(t-1)})$
- **Backward TE:**  $I(z_j^{(t-1)}; z_i^{(t)} \mid z_i^{(t-1)})$
- $\Delta TE = \text{Forward TE} - \text{Backward TE}$

### In practice

We compute  $I(Z_i^{(t-1)}; Z_j^{(t)} \mid Z_j^{(t-k)})$  by iterating  $k$  through a fixed range from 1 to 15. If  $z_i \rightarrow z_j$  is a true causal link, we expect  $\Delta TE > 0$ .

## Error-Based Descriptors

For the true causal direction ( $z_i \rightarrow z_j$ ), the error in predicting the effect ( $z_j$ ) from the cause ( $z_i$ ) should be statistically independent of the cause itself. This independence often breaks in the reverse direction.



### Are they correlated?

A key descriptor is the **correlation between the prediction error and the cause**  $\text{Corr}(z_i, \epsilon_j)$ , where  $\epsilon_j = z_j - \text{model}(z_i)$ .

For a non-causal link, it is often **non-zero**.



## Distributional descriptors

Causal relationships can create subtle asymmetries in the *shape* of the joint data distribution, captured by higher-order statistical moments. For example, we can measure the interaction between the skewness of one variable and the values of another:

$$\text{HOC}_{3,1} = \mathbb{E}[(z_i - \mu_i)^3(z_j - \mu_j)]$$

---

### Wrapping up

By combining evidence from many different viewpoints, our model acts as a robust **meta-learner**, not over-relying on any single assumption. It learns *which type of signal to trust* for different dynamics. (we can see via **feature importance** that the model uses a diverse set of features from multiple families.)

# The Gauntlet: Experimental Setup

## Training Environment

- Trained **exclusively** on a rich, diverse library of 9 synthetic Nonlinear Autoregressive (NAR) processes.
- Multiple **noise distribution** (gaussian, uniform, and laplace) for assumptions and fairness.
- We derive a **classification threshold** ( $\tau = 0.309$ ) via Leave-One-Process-Out Cross-Validation.

We set `MAXLAGS` = 3 for the entire benchmark suite.

## Testing Environments

1. **Unseen Dynamics:** Evaluated on 9 completely new synthetic NAR processes not seen during training.
2. **Zero-Shot Generalization:** Applied directly to established, realistic benchmarks:
  - NetSim (Biological Networks)
  - DREAM3 (Gene Regulation)

## Our synthetic data (some)

$$Y_{t+1}[j] = -0.4 \frac{(3 - \bar{Y}_t[\mathcal{N}_j]^2)}{(1 + \bar{Y}_t[\mathcal{N}_j]^2)} + 0.6 \frac{3 - (\bar{Y}_{t-1}[\mathcal{N}_j] - 0.5)^3}{1 + (\bar{Y}_{t-1}[\mathcal{N}_j] - 0.5)^4} + W_{t+1}[j] \quad (1)$$

$$Y_{t+1}[j] = 1.5 \sin(\pi/2 \bar{Y}_{t-1}[\mathcal{N}_j]) - \sin(\pi/2 \bar{Y}_{t-2}[\mathcal{N}_j]) + W_{t+1}[j] \quad (2)$$

$$Y_{t+1}[j] = 2 \exp(-0.1 \bar{Y}_t[\mathcal{N}_j]^2) \bar{Y}_t[\mathcal{N}_j] - \exp(-0.1 \bar{Y}_{t-1}[\mathcal{N}_j]^2) \bar{Y}_{t-1}[\mathcal{N}_j] + W_{t+1}[j] \quad (3)$$

$$Y_{t+1}[j] = -2 \bar{Y}_t[\mathcal{N}_j] I(\bar{Y}_t[\mathcal{N}_j] < 0) + 0.4 \bar{Y}_t[\mathcal{N}_j] I(\bar{Y}_t[\mathcal{N}_j] < 0) + W_{t+1}[j] \quad (4)$$

$$Y_{t+1}[j] = 0.3 \bar{Y}_t[\mathcal{N}_j] + 0.6 \bar{Y}_{t-1}[\mathcal{N}_j] + \frac{(0.1 - 0.9 \bar{Y}_t[\mathcal{N}_j] + 0.8 \bar{Y}_{t-1}[\mathcal{N}_j])}{(1 + \exp(-10 \bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \quad (5)$$

$$Y_{t+1}[j] = \text{sign}(\bar{Y}_t[\mathcal{N}_j]) + W_{t+1}[j] \quad (6)$$

$$Y_{t+1}[j] = 0.8 \bar{Y}_t[\mathcal{N}_j] - \frac{0.8 \bar{Y}_t[\mathcal{N}_j]}{(1 + \exp(-10 \bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \quad (7)$$

$$Y_{t+1}[j] = 0.3 \bar{Y}_t[\mathcal{N}_j] + 0.6 \bar{Y}_{t-1}[\mathcal{N}_j] + \frac{(0.1 - 0.9 \bar{Y}_t[\mathcal{N}_j] + 0.8 \bar{Y}_{t-1}[\mathcal{N}_j])}{(1 + \exp(-10 \bar{Y}_t[\mathcal{N}_j]))} + W_{t+1}[j] \quad (8)$$

$$Y_{t+1}[j] = 0.38 \bar{Y}_t[\mathcal{N}_j] (1 - \bar{Y}_{t-1}[\mathcal{N}_j]) + W_{t+1}[j] \quad (9)$$

$$Y_{t+1}[j] = \begin{cases} 0.9 \bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] & \text{if } |\bar{Y}_t[\mathcal{N}_j]| < 1 \\ -0.3 \bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] & \text{otherwise} \end{cases} \quad (10)$$

$$Y_{t+1}[j] = 0.9 \cdot \bar{Y}_t[\mathcal{N}_j] + W_{t+1}[j] \quad (11)$$

$$Y_{t+1}[j] = 0.4 \cdot \bar{Y}_{t-1}[\mathcal{N}_j] + 0.6 \cdot \bar{Y}_{t-2}[\mathcal{N}_j] + W_{t+1}[j] \quad (12)$$

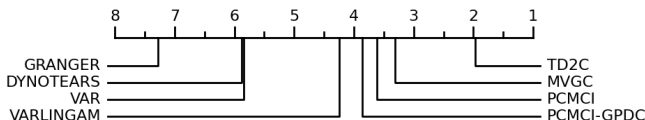
## Result 1: Dominating on Unseen Synthetic Dynamics

Causal discovery is a highly **imbalanced** problem (many more non-causal links than causal ones), we use F1-Score & Balanced Accuracy.

Method	Balanced Accuracy	F1-Score
<b>TD2C (ours)</b>	<b>0.822</b>	<b>0.631</b>
MVGC [Barrett et al., 2010]	0.741	0.489
PCMCi GPDC [Runge et al., 2019]	0.734	0.496
VARLINGAM [Hyvärinen et al., 2010]	0.712	0.450
DYNOTEARS [Pamfil et al., 2020]	0.533	0.104

Friedman/Nemenyi statistical test (right is better).

A **black bar** means statistically equal.



# Testing on Established, Realistic Benchmarks

## DREAM3 In Silico Challenge

- **Domain:** Systems Biology.
- **Goal:** Reverse-engineer gene regulatory networks (GRNs) from simulated gene expression data.
- **Data Type:** In silico (simulated) time-series and steady-state gene expression data from biological networks of e.g. E. coli.
- **Dynamics:** The data is generated using ordinary differential equations.

## NetSim

- **Domain:** Biological Network.
- **Goal:** To generate dynamic time-series data from known biological networks.
- **Data Type:** Simulated time-series data representing the activity levels of different nodes in a network over time.
- **Dynamics:** The simulator models the discrete-time dynamics of biological interactions.

## The Challenge

The underlying dynamics (based on biochemical kinetics) are fundamentally different from the NAR models used in training.

## Result 2: The Ultimate Test - Zero-Shot Generalization

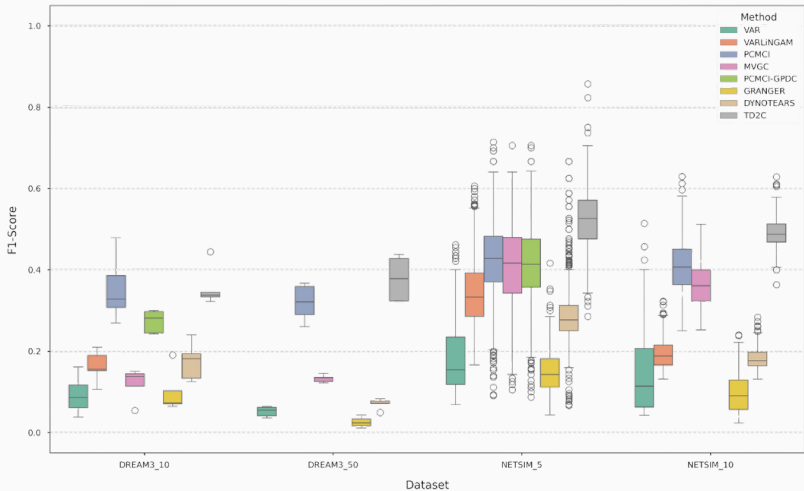


Figure 1: F1-Score distribution on realistic benchmark datasets.

# Conclusion & Key Takeaways

## 1. A New Paradigm

We successfully reframed causal discovery from a series of binary statistical tests to a supervised **pattern recognition** problem.

## 2. The TD2C Framework

Our framework learns the subtle, quantitative signatures of causality from a **rich, multi-faceted feature set**, synthesizing evidence from multiple methodological families.

## 3. Good Performance

TD2C demonstrates good zero-shot performance on realistic benchmarks, showing potential for generalizability.

# Future Work: The Path Ahead

## Addressing Scalability

- Currently  $O(n^2)$ .
- → infer the **entire DAG**.

## On top of Sota

- PC has theoretical guarantees
- → **post-hoc TD2C**.

## Prior knowledge

- Ground Truth is paramount
- → **include prior knowledge**.

## Hidden confounders

- An open research question
- → **hide some features**.

## Stronger validation

- Maybe our tests were lucky.
- → **extensive tests**.

## Tackling Non-Stationarity

- Real-world systems evolve.
- → **online TD2C**.

## Assumptions

- First-order might be wrong
- → **break assumptions**.

## Markov Blanket

- An essential step
- → **stronger MB estimation**.



# A First Step Towards Foundational Causal Models?

A foundational model (FM) for causality would be a large-scale model, pre-trained on a vast and diverse universe of causal systems, that learns the **fundamental, transferable signatures of causation itself**.

- We moved from single-use, dataset-specific statistical tests to a **trained, reusable model** of what causality "looks like" in data.
- Our (small) NAR library is a **diverse set of causal ground truths**.
- The learned concept was abstract enough for **unseen dynamics**.

But we're still *very far*:

- A true FM would need **billions** of diverse causal systems
- A FM would need to be pre-trained on **more general objectives**, full graph inference, estimating interventions, and counterfactuals.
- First next step: move **beyond RF** to GNNs or Transformers.

# Thank you for your attention.

Questions? Suggestions? Critiques?



Slides, code and data are available:  
<https://github.com/gmpal/TD2C-PP>

## References i



Barrett, A. B., Barnett, L., & Seth, A. K. (2010).  
**Multivariate Granger causality and generalized variance.**  
*Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*,  
81(4), 041907.






Bontempi, G., & Paldino, G. M. (2020).  
**Learning causal discovery from observational data.**  
In *International Conference on Discovery Science*.



Bontempi, G., & Meyer, P. E. (2015).  
**Dependency to causality: a machine learning approach.**  
*From dependency to causality (D2C) challenge*.



Granger, C. W. J. (1980).  
**Testing for causality: A personal viewpoint.**  
*Journal of Economic Dynamics and Control*, 2, 329–352.

-  Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008).  
**Nonlinear causal discovery with additive noise models.**  
*In Advances in neural information processing systems 21.*
-  Hyvärinen, A., Zhang, K., Shimizu, S., & Hoyer, P. O. (2010).  
**Estimation of a structural vector autoregression model using non-gaussianity.**  
*Journal of Machine Learning Research, 11*, 1709-1731.
-  Kraskov, A., Stögbauer, H., & Grassberger, P. (2004).  
**Estimating mutual information.**  
*Physical review E, 69*(6), 066138.



Pamfil, R., Sridhar, D., Bica, I., Cecen, Z., & van der Schaar, M. (2020).

**DYNOTEARS: Learning causal structure over time.**

In *International Conference on Machine Learning*.



Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019).

**Detecting and quantifying causal connections in a climate system.**

*Nature Communications*, 10(1), 1-12.



Cochran, W. G. (1954).

**Some methods for strengthening the common  $\chi^2$  tests.**

*Biometrics*, 10(4), 417-451.



Doran, G., Muandet, K., Zhang, K., & Schölkopf, B. (2014).

**A permutation-based kernel conditional independence test.**

*Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence.*



Fisher, R. A. (1924).

**The distribution of the partial correlation coefficient.**

*Metron*, 3, 329-332.



Mantel, N., & Haenszel, W. (1959).

**Statistical aspects of the analysis of data from retrospective studies of disease.**

*Journal of the National Cancer Institute*, 22(4), 719-748.

## References v



Pearson, K. (1900).

**On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.**

*Philosophical Magazine, Series 5, 50(302), 157-175.*



Shah, R. D., & Peters, J. (2020).

**The hardness of conditional independence testing and the generalised covariance measure.**

*The Annals of Statistics, 48(3), 1514-1538.*



Wilks, S. S. (1938).

**The large-sample distribution of the likelihood ratio for testing composite hypotheses.**

*The Annals of Mathematical Statistics, 9(1), 60-62.*



Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011).

**Kernel-based conditional independence test and application in causal discovery.**

*Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence.*



## Q0: The Source of Asymmetry: It's Structural, Not Random

The asymmetry is a direct result of how conditioning on a **cause** versus an **effect** interacts with their respective MBs.

**The View from the Cause ( $z_i$ )**  
**Reverse Direction:**  $\{I(z_i; \mathbf{m}_{i,k} | z_i)\}$

When we condition on the cause  $z_i$ :

- $\mathbf{Pa}_i \rightarrow z_i \rightarrow z_j$  (a **chain**).  
Conditioning on  $z_i$  **blocks**.
- $\mathbf{Ch}_i \leftarrow z_i \rightarrow z_j$  (a **fork**).  
Conditioning on  $z_i$  **blocks**.

Result: Many terms in this population become zero.

**The View from the Effect ( $z_j$ )**  
**True Direction:**  $\{I(z_i; \mathbf{m}_{j,k} | z_j)\}$

When we condition on the effect  $z_j$ :

- $z_i \rightarrow z_j \leftarrow \mathbf{Pa}'_j$  (a **collider**).  
Conditioning on the collider  $z_j$  **opens**.

Result: The presence of other causes for the effect *guarantees* some terms in this population will be greater than zero.

## Q1: How robust is the causal asymmetry?

The asymmetry we leverage is not an empirical artifact; it stems from the structural differences in how information flows through d-separation:

- **Chains** ( $A \rightarrow B \rightarrow C$ ): Conditioning **blocks** the path.
- **Colliders** ( $A \rightarrow B \leftarrow C$ ): Conditioning **opens** the path.

### Can it be inverted?

A true inversion is theoretically possible but practically unlikely.

It would require a "pathological" graph where numerous confounding paths create **more information flow in the non-causal direction** than the true causal link creates in the forward direction.

→ **It's unlikely in real world settings.**

## Q2: What if the Markov Blanket is wrong?

Our MB definition ( $\{z_i^{(t-1)}, z_i^{(t+1)}\}$ ) is a pragmatic choice, trading completeness for computational tractability and stability. It captures the most immediate and often strongest temporal influences.

**Performance is expected to degrade gracefully, not fail.**

Our framework is not solely reliant on the MB. Other feature families provide robustness. A key research direction is a **hybrid MB approach**, starting with the simple temporal neighbors and adaptively adding other highly relevant variables.

→ **It's unlikely, but we will try a hybrid MB.**

### Q3: How does TD2C handle unobserved confounders?

A confounder might create a strong spurious signal in one feature family. However, it is **less likely** to perfectly **mimic** the complex, asymmetric **signature** across all of our features. A true causal link tells a more consistent "story".

#### **No Method is Immune, But Our Approach is Resilient**

It is possible (although not done yet) to explicitly include systems with **latent confounders in our training data**. The model might learn to recognize the specific **statistical "footprint" of a confounded link**, which can differ from that of a direct causal connection.

→ **It's on our To Do List.**

## Q4: What about faithfulness violations?

Faithfulness violations (e.g., causal paths perfectly cancelling out) are most problematic for **constraint-based** methods, as they rely on finding perfect conditional independencies (i.e.,  $\text{CMI} = 0$ ).

### TD2C is a Pattern Recognition System, Not a CI Tester

- We do not test if CMI is zero. We analyze the **quantitative distribution of non-zero CMI values**.
- If a faithfulness violation zeroes out the signal in one specific information-theoretic descriptor, it is highly unlikely to simultaneously cancel the signals in our other feature families:
  - Asymmetries in prediction errors.
  - Asymmetries in higher-order statistical moments.
- Our model's strength is its ability to find a causal signature even when some individual signals are silent.

→ **However, this needs extensive assessment.**

## Q5: Why should training on NAR models generalize?

The hypothesis is not that NAR processes are universal. The hypothesis is that the **statistical manifestations of causality** are universal principles derived from probability and graph theory.

- Colliders inducing dependence is universal.
- The independence of residuals from a true cause is universal.
- Asymmetry in information flow is universal.

### **The NAR Library as a "Causality Flight Simulator"**

Our diverse set of 18 NAR processes (linear, non-linear, threshold-based, chaotic) is designed to be a rich training ground, exposing the model to these universal signatures in a wide variety of contexts.

→ **Generalizability needs to be tested more extensively.**

## Q6: Why were simple linear/error features most important?

The high rank of error-based and linear descriptors reveals a powerful insight into the model's logic:

- **They are strong, reliable first-pass filters.** Even in globally non-linear systems, local, time-lagged relationships often have a strong, detectable linear component. The model learns to trust these robust signals to solve the "easy" cases.

### But Simple Features Alone Are Not Enough

- The more complex, non-parametric features (Information Theory, Higher-Order Moments) are crucial for achieving state-of-the-art performance.

→ **The model's power comes from a synergistic combination.**

## Q7: How did you choose the MI estimator parameter 'k'?

The 'k' parameter in the KSG k-Nearest Neighbor estimator controls a fundamental trade-off:

- **Small 'k' (e.g., 1-3):** Low bias, high variance. Captures fine-grained local structure but can be sensitive to noise.
- **Large 'k':** High bias, low variance. More stable but oversmooths the probability distribution, potentially missing local dependencies.

### Our Approach: A Standard Default + Framework Robustness

- We chose **k=3**, a widely used default in the non-parametric statistics literature that is considered a good general-purpose balance.

→ **We did not perform exhaustive tuning. Systematic cross-validation to optimize 'k' is planned.**



## Q8: Why a Balanced Random Forest?

We selected the Balanced Random Forest classifier because it directly addresses three core issues in our task:

1. **Extreme Class Imbalance:** This is the most critical challenge. The algorithm is specifically designed to handle this by creating balanced bootstrap samples to train each tree, preventing the model from simply ignoring the rare "causal" class.
2. **High-Dimensional, Heterogeneous Features:** As an ensemble of trees, it is naturally robust to irrelevant or noisy features and can capture complex, non-linear interactions between features without manual engineering.
3. **Interpretability:** It provides a straightforward and reliable measure of feature importance, which was essential for us to understand the model's internal logic and validate our feature engineering choices.

→ **Nonlinear, interpretable model, handling class imbalance**

## Q9: Was the experimental comparison fair?

We made specific choices for fairness when possible (eg. PCMCI w/ nonlinear CI tests), but also:

- We used the standard implementations of all benchmark methods with their recommended default parameters.
- Our decision threshold was tuned via cross-validation on the training set.

### Comparing Core Algorithms, Not Tuning Strategies

While it's possible that exhaustively tuning each of the 7 competing methods could alter their scores, our goal was to compare the **inherent power of the algorithms themselves**.

→ We went for default, mostly, trying to maintain fairness.

## Q10: How to interpret the High-Recall result?

In many scientific domains, the cost of missing a true discovery (a false negative) is far higher than the cost of investigating a false lead (a false positive).

- Our method's **statistically significant superiority in Recall** means it is an exceptional tool for discovery. It is less likely to miss novel, true causal relationships.

### Balancing Discovery with Practicality

TD2C is not just a high-recall method; it maintains good, competitive Precision. This is critical because it ensures the list of generated hypotheses is of high quality and not flooded with noise.

→ **TD2C can be used to generate a rich yet manageable set of high-confidence candidate causal links.**

## Q11: Sensitivity to the Lag-1 Ground Truth Assumption?

To go from static ground truth for DREAM3 and NETSIM, we made a Lag-1 Ground Truth Assumption. Creating a temporal ground truth from the static benchmark graphs required some decision to be made.

However:

- This assumption affects **all methods equally**, ensuring the **relative ranking** between them remains a valid measure of performance.

### TD2C Has Built-in Robustness to This Assumption

During inference on the benchmarks, we set 'max\_lag=3', allowing TD2C to search for and identify causal links beyond the assumed lag of 1.

→ **A more diverse testing scenario would be beneficial.**

## Q12: Future Work: Adapting Pairwise Features for a GNN

The goal is to move from isolated pairwise decisions to a single, context-aware graph inference. We envision a three-step process:

1. Use a fast, scalable method (e.g., lagged correlation) to create an initial, dense, **undirected graph**.
2. For each directed edge candidate ( $i \rightarrow j$  and  $j \rightarrow i$ ), compute our full 63-dimensional TD2C feature vector (initial **edge embedding**).
3. A Graph Attention Network (GAT) iteratively updates each edge embedding by learning to "**attend**" to the neighborhood.

### Adjacency matrices

Another option is use CNNs to learn full adjacency matrices and perform one-shot inference with matrix as output.

→ We want the final prediction for edge  $i \rightarrow j$  to be informed by the entire local graph structure.

## Q13: Future Work: Tackling Non-Stationarity

The current global model would likely fail. The solution is to adapt the framework to be local in time.

### Rolling-Window Analysis

- Apply the pre-trained, static TD2C model to sliding windows.
- Produce a temporal series of causal graphs.

### ”Change-Aware” Feature Engineering

- Engineer new descriptors that explicitly quantify change.
- Ex:  $\Delta MI = MI(\text{first half of window}) - MI(\text{second half})$ .

→ allow a single, more powerful model to learn the signatures of a changing causal link and potentially predict regime shifts.

## Q14: Is the output probability a measure of causal strength?

### An Important Distinction: Confidence, Not Effect Size

The model's output is the **probability that a causal link exists**, based on the learned patterns. It is a measure of the model's confidence.

- It is **not** a direct, calibrated measure of causal strength or effect size (e.g., the coefficient  $\beta$  in an equation  $Y = \beta X + \epsilon$ ).
- While a stronger causal link will likely produce a higher probability score, the relationship is not guaranteed to be linear or directly proportional.

→ **We want to investigate the relationship between probability score and causal strength**

## Q15: Data Volume at a Glance: Synthetic

### Synthetic Data Generation For Training:

- **3,240** Unique Time Series were generated.
- *Composition*: 9 NAR Processes  $\times$  3 Noise Types (*Gauss, Uniform, Laplace*)  $\times$  120 Instances each.
- *Dimensions*: 5 Variables  $\times$  250 Timesteps.
- *Total Candidate Links*: 243,000 pairs.

### For Testing:

- **1,080** Unique Time Series were generated.
- *Composition*: 9 different NAR Processes  $\times$  3 Noise Types  $\times$  40 Instances each.
- *Dimensions*: 5 Variables  $\times$  250 Timesteps.
- *Total Candidate Links*: 81,000 pairs.



## Q16: Data Volume at a Glance: Realistic

This data was **only used for evaluation** and never seen during training.

Dataset	Time Series	Vars	Timesteps
DREAM3-10	5	10	84
DREAM3-50	5	50	483
NETSIM-5	1,050	5	200
NETSIM-10	250	10	200

## Q17: Why The Kraskov (KSG) MI Estimator?

### The Challenge of Estimating MI

Estimating mutual information from data is hard, especially in high dimensions, as it requires estimating the full probability density functions.

### The KSG Estimator [Kraskov et al., 2004]

- It avoids direct density estimation.
- Instead, it is based on the distances to the  $k$ -th nearest neighbors for each point in the joint and marginal spaces.
- By comparing the average volume occupied by neighbors in the joint space vs. the marginal spaces, it can accurately estimate MI.
- It is non-parametric, meaning it makes no assumptions about the underlying data distribution (e.g., Gaussianity).
- This is crucial for capturing the non-linear relationships in our synthetic and real-world data.

## Q18: What did the model learn?

Rank	Feature Description
1	<i>(Error-Based)</i> Corr(Prediction Error, Cause)
2	<i>(Error-Based)</i> Partial Corr(Cause, Effect)
3	<i>(Info-Theoretic)</i> Mean MI from Effect's MB
4	<i>(Linear)</i> Regression Coefficient (Effect)
5	<i>(Info-Theoretic)</i> Mean MI from Cause's MB
...	...
8	<i>(Higher-Order)</i> Cross-moment (HOC 1,3)
9	<i>(Info-Theoretic)</i> Generalized TE Asymmetry

### Key Insight

The model does not rely on a single type of causal signal. It learns to synthesize evidence from a **diverse toolkit** of features spanning multiple theoretical families. This versatility is the foundation of its robust, generalizable performance.