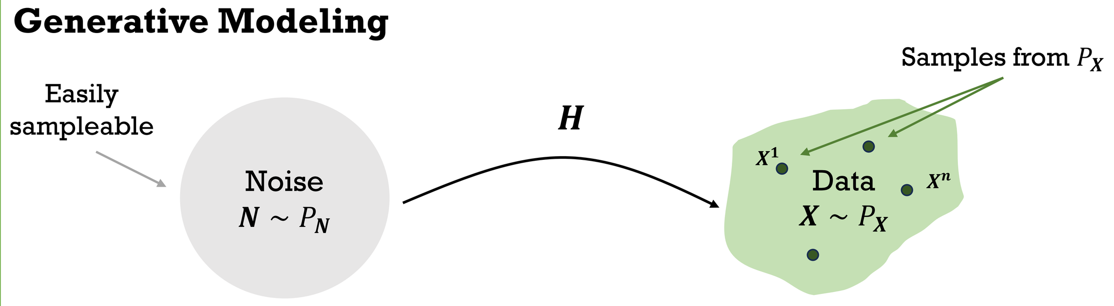
A Fixed-Point Approach for Causal Generative Modeling

Meyer Scetbon, Joel Jennings, Agrin Hilmkil, Cheng Zhang, Chao Ma

Why Causality?

Generative Modeling

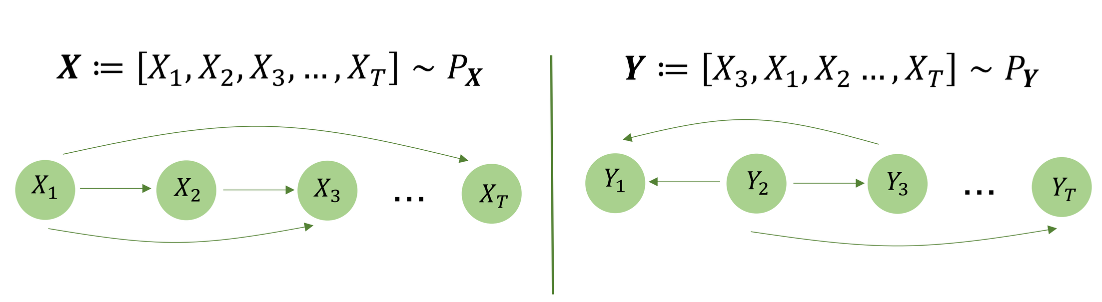


Goal: Learn a function $\boldsymbol{H}$ that maps P_N to P_X

Examples: VAEs, GANs, Normalizing Flows, Diffusions, …

Question: what if P_X admits an (unknown) structure?

Causal Generative Modeling



Goal: Learn a function $\boldsymbol{H}$ that maps P_N to P_X and that satisfies the *unknown structure* of P_X

Advantage: we can now simulate the **effects of changes** on the variables X_i w.r.t the structured generative process $\boldsymbol{H}$

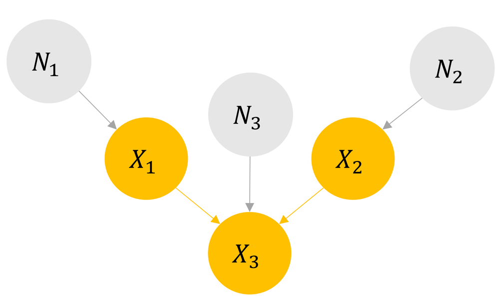
Causality offers a reliable tool for **decision-making** in generative modeling

Structural Causal Models

Parameterization of SCMs

- Noise distribution: P_N
- Directed Acyclic Graph: G
- System of Equations:

$X_i = H_i(\mathrm{PA}_G(\mathrm{X}_i), \mathrm{N}_i), \forall \mathrm{i} \in [|d|]$



Generate **observations**

- Sample $(N_1, N_2, N_3) \sim P_N$
- Solve $X_i = H_i(\mathrm{PA}_G(\mathrm{X}_i), \mathrm{N}_i)$

Generate **effects** of **changes**

- Intervene on $i: X_i \leftarrow a$
- Generate observations with the intervened system

Learning SCMs from observations is hard

- Computationally Challenging

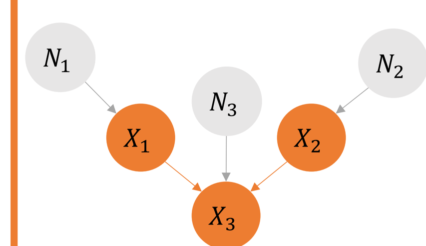
Finding the DAG from data is an **NP-hard combinatorial** problem

- Ill-posed Inverse Problem

The **uniqueness** of the SCM is **not** always **guaranteed**

Question: can we improve on both limitations?

From DAGs to Causal Order



A DAG (always) induces a causal order

Ex: $\pi = (X_1, X_2, X_3)$

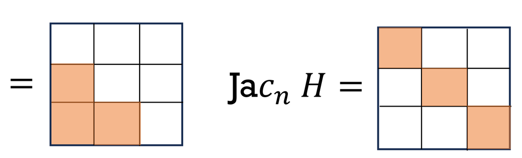
Equivalent Parameterization of SCMs

- Noise Distribution: P_N
- Topological Ordering: π
- Map: $\boldsymbol{H}: (\mathrm{x}, \mathrm{n}) \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ s.t.

No DAG required

Fixed-Point SCM: $\mathrm{X} = P_\pi^T \boldsymbol{H}(P_\pi \mathrm{X}, P_\pi \boldsymbol{N})$

$\mathrm{Jac}_x\, H =$ $\mathrm{Jac}_n\, H =$



Permutation matrix Associated to π

Partial Identifiability of Fixed-Point SCMs

Problem (Strong Identifiability): Given π and P_X, can we recover **uniquely** the **fixed-point SCM**?

Theorem (Monotonic Model)

If H **is monotonic increasing** w.r.t the noise, and P_N is fixed, then there exists a **unique SCM** generating P_X with order π and noise P_N

Problem (Weak Identifiability): Given π and P_X, can we recover **uniquely** the **interventional** and **counterfactual** distributions?

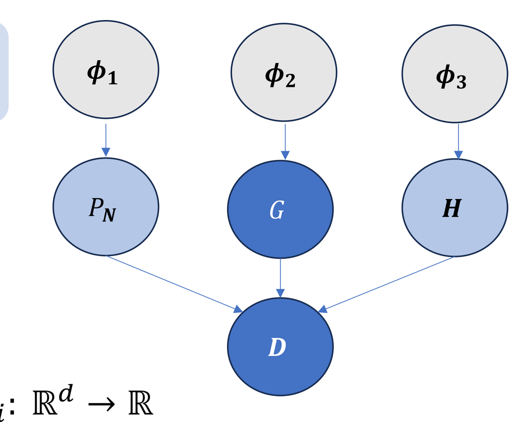
Theorem (Invertible Model)

- If H is **invertible** and C^1, then the **interventional** and **counterfactual** distributions are uniquely identifiable.
- Also, one can choose **arbitrarily** the **noise distribution** P_N, e.g. standard Gaussian.

Learning Fixed-Point SCMs

Recovering Causal Orders from Observations

Goal: Infer in a zero-shot manner the causal order of variables from observations



Dataset Generative Process

- P_N is a generated noise distribution on $\mathbb{R}^d$
- G is a generated DAG
- $\boldsymbol{H}$ is a generated sequence of d functions H_i: $\mathbb{R}^d \to \mathbb{R}$
- $D \in \mathbb{R}^{m \times d}$ is obtained by sampling m i.i.d samples $\boldsymbol{N}^{(k)} \sim P_N$

and then by solving for each k: $X_i^{(k)} = H_i\left(\mathrm{PA}_G\left(\mathrm{X}_\mathrm{i}^{(k)}\right), \mathrm{N}_i^{(k)}\right), \forall \mathrm{i} \in [|d|]$

Training: Given $(D_1, G_1), \dots, (D_n, G_n)$ i.i.d, we train a model $\mathcal{M}$ that predicts the leaves of the graphs in a sequential manner.

Loss: $\mathcal{M}$ is learned by minimizing d-TOE where at each step we remove a leaf.

Algorithm 1 d-TOE($\mathcal{M}$, $(\mathcal{D}_{\mathrm{tr}}, \mathcal{G}_{\mathrm{tr}})$)

```
Input: M, (D_tr, G_tr)
Initialize d-TOE = 0.
for q = 1 to d do
    p ← M(D_tr),    y ← L(G_tr)
    d-TOE ← d-TOE + BN(p, y)
    ℓ̂ ← argmax_i [p]_i,    ℓ ← B(y, ℓ̂)
    D_tr ← R_1(D_tr, ℓ),    G_tr ← R_2(G_tr, ℓ)
end for
Return d-TOE
```

Learning SCMs on the Ordered Variables

Goal: Learn the fixed-point SCM associated to a single dataset given the causal order

Proposed Architecture

- Causal Embedding: $\mathcal{E}: \boldsymbol{X} \in \mathbb{R}^d \to [X_1 * \theta_1, \dots X_d * \theta_d] \in \mathbb{R}^{d \times D}$
- DAG Attention: $\mathrm{DA}(\boldsymbol{Q}, \boldsymbol{K}) = \dfrac{\exp((\boldsymbol{QK}^T - \mathrm{M})/\sqrt{D})}{\mathfrak{I}\left(\exp\left(\frac{\boldsymbol{QK}^T - \mathrm{M}}{\sqrt{D}}\right) 1_d\right)}$, where $\mathrm{M} = \begin{bmatrix} +\infty & +\infty & +\infty \\ 0 & +\infty & +\infty \\ 0 & 0 & +\infty \end{bmatrix}$

and $[\mathfrak{I}(v)]_i = v_i$ if $v_i \geq 1$, and $[\mathfrak{I}(v)]_i = 1$ otherwise

- Causal Encoder $\mathcal{C}_L$: $\boldsymbol{h}_{\ell+1} = h(DA(\boldsymbol{h}_\ell, \boldsymbol{X})\boldsymbol{X} + \boldsymbol{h}_\ell) \in \mathbb{R}^{d \times D}$
- Causal Decoder: $\mathcal{F}: \boldsymbol{X} \in \mathbb{R}^{d \times D} \to [\langle X_1, \omega_1 \rangle, \dots, \langle X_d, \omega_d \rangle] \in \mathbb{R}^d$

Model: $\mathcal{T}: \boldsymbol{X} \in \mathbb{R}^d \to \mathcal{F} \circ \mathcal{C}_L(\mathcal{E}(\boldsymbol{X})) \in \mathbb{R}^d$ satisfies the simple structure of fixed-point SCM.

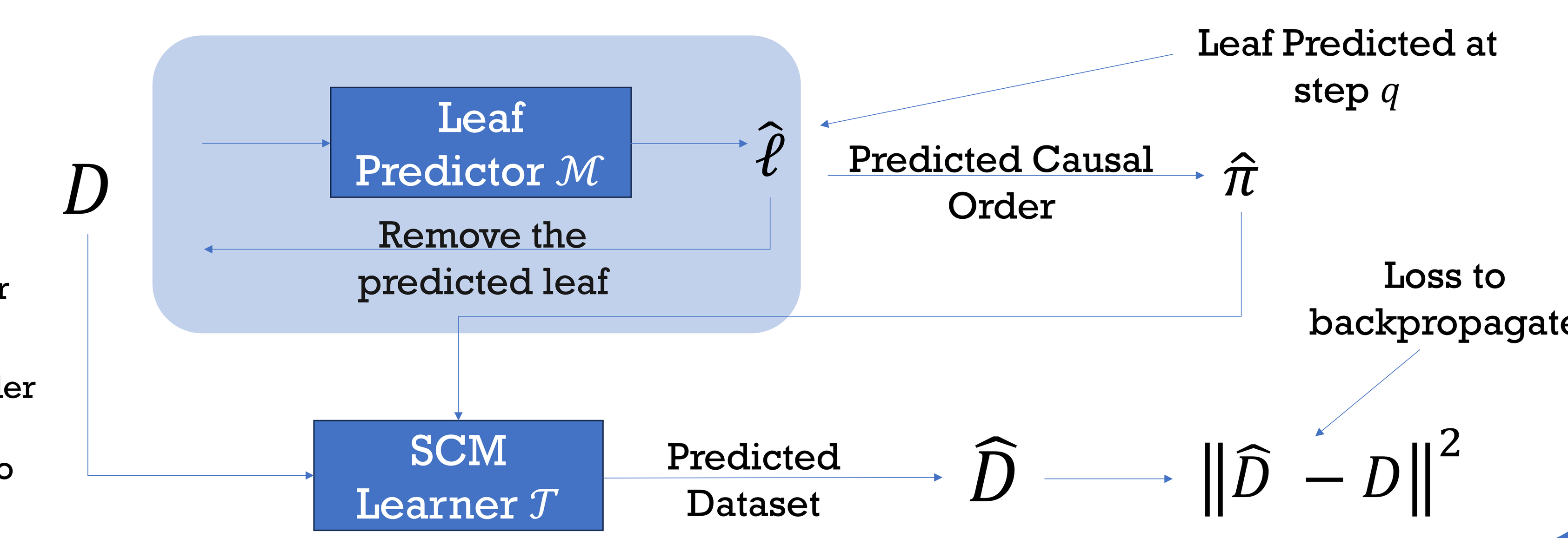
Training: minimize the MSE $\mathbb{E}_{\boldsymbol{Z} \sim P_{P_\pi X}} \|\mathcal{T}(\boldsymbol{Z}) - \boldsymbol{Z}\|^2$ where $P_{P_\pi X}$ is the causally ordered distribution of observations.

End-to-End Pipeline

Problem Set Up: Let $\mathrm{X}^{(1)}, \dots, \boldsymbol{X}^{(m)}$ i.i.d samples from P_X. Our goal is to learn an SCM generating P_X.

Training: We learn the SCM by training $\mathcal{T}$ on D ordered according to the causal order predicted by $\mathcal{M}$ on D.

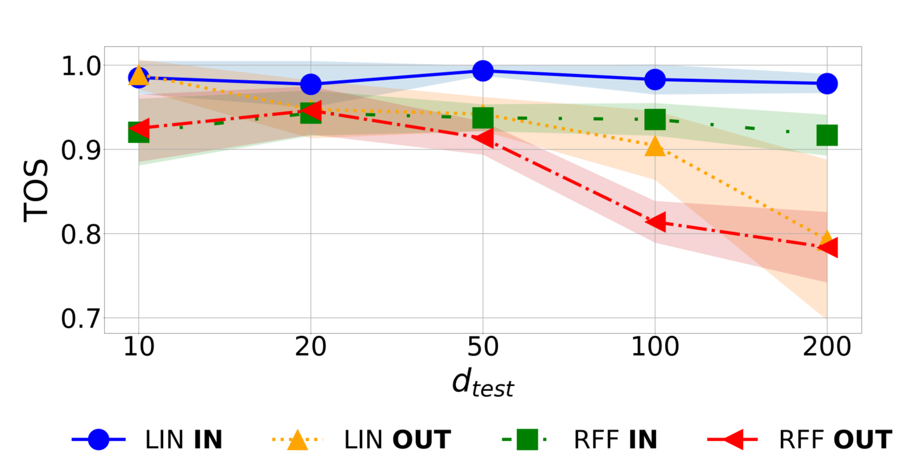
Inference: Once trained, we show that under ANM, and correctness of $\hat{\pi}$, $\hat{\mathcal{T}}$ learned converges (in the limit of infinite sample) to the true SCM generating the data.



Empirical Evaluations

- We train $\mathcal{M}$ on $\simeq 200k$ datasets with $n = 200$ samples and $d = 50$ dimensions.
- We evaluate the models on $\simeq 400$ test datasets of various sizes, newly sampled from either the distribution used during training (**in-distribution**), or from a variant with a substantial shift (**out-of-distribution**).

Evaluation of the Leaf Predictor $\mathcal{M}$



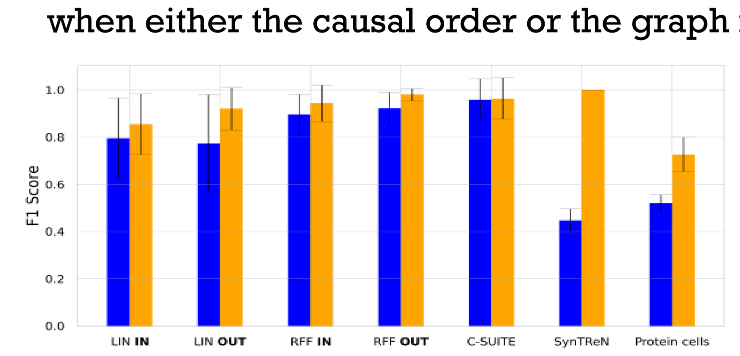
TOS measures the proportion of nodes correctly ranked in the causal order.

Obtain (at worst) an accuracy of 80% on problems with 200 nodes.

Evaluation of the SCM Learner $\mathcal{T}$

Datasets	True P	True $\mathcal{G}$
LIN **IN**	0.037 / 0.066 (0.057)	0.012 / 0.039 (0.060)
LIN **OUT**	0.065 / 0.11 (0.084)	0.017 / 0.034 (0.048)
RFF **IN**	0.065 / 0.10 (0.089)	0.033 / 0.059 (0.075)
RFF **OUT**	0.11 / 0.12 (0.088)	0.033 / 0.042 (0.040)
C-Suite	0.026 / 0.032 (0.030)	0.022 / 0.025 (0.022)

Counterfactual predictions errors on synthetic test datasets when either the causal order or the graph is known.



F1 score on various problems when either the causal order or the graph is known.

Benchmarking of End-to-End Pipeline

Datasets	LIN OUT	RFF OUT
PC	0.47 (0.14)	0.40 (0.12)
GES	0.56 (0.12)	0.37 (0.060)
GOLEM	0.73 (0.29)	0.31 (0.13)
DECI	0.36 (0.13)	0.74 (0.14)
GraN-DAG	0.29 (0.19)	0.50 (0.26)
DAG-GNN	0.61 (0.19)	0.44 (0.15)
DP-DAG	0.17 (0.074)	0.16 (0.067)
AVICI	0.73 (0.16)	0.74 (0.17)
FiP (Ours)	**0.76(0.20)**	**0.81(0.15)**

Comparison of the F1 score against various baselines on O.O.D datasets.

We obtain **SoTA** results for **causal discovery** and **causal inference tasks** on **O.O.D** test datasets.

Datasets	LIN OUT	RFF OUT
DECI	0.39 (0.29)	0.18 (0.12)
DoWhy - AVICI	0.20 (0.18)	0.16 (0.096)
FiP (Ours)	**0.13 (0.10)**	**0.13 (0.096)**
FiP w. $\mathcal{G}$	0.034 (0.048)	0.042 (0.040)
DoWhy w. $\mathcal{G}$	0.0017 (0.0017)	0.088 (0.072)

Comparison of the counterfactual predictions against various baselines on O.O.D datasets.

Dataset	Model	CF Error (RMSE)
TRIANGLE	Causal NF	0.13 (0.02)
	CAREFL	0.17 (0.03)
	VACA	4.19 (0.04)
	FiP	**0.094(0.021)**
SIMPSON	Causal NF	**0.12(0.02)**
	CAREFL	0.17 (0.04)
	VACA	1.50 (0.04)
	FiP	**0.12(0.0089)**

Comparison of the counterfactual predictions on synthetic datasets when the order is known.