DAG-PnP: Plug-and-Play Causal Discovery with Diffusion Priors

Nicolas Zilberstein[†] Ali Azizpour[†] Gonzalo Mateos[*] Santiago Segarra[†]

[†]Dept. of Electrical and Computer Engineering, Rice University [*]Dept. of Electrical and Computer Engineering, University of Rochester

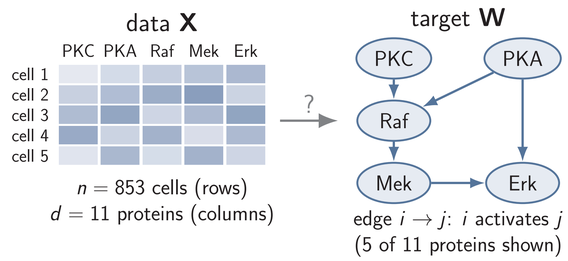
github.com/aliaaz99/dag-pnp

Causal Discovery as a Linear Inverse Problem

- **Why causal discovery?** Knowing *who drives whom* predicts the effect of interventions.
- **Inverse problem on graphs:** observe $\mathbf{X}$, recover the graph $\mathbf{W}$ via a linear SEM:

$$\mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E}, \qquad W_{ij} \neq 0 \;\Leftrightarrow\; i \to j$$

- **Example** (Sachs et al., 2005): protein signaling, measured in single cells.



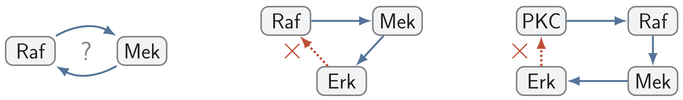
$n = 853$ cells (rows)
$d = 11$ proteins (columns)

edge $i \to j$: i activates j
(5 of 11 proteins shown)

- **Formulation:** data fit and acyclicity, plus a prior on the graph,

$$\min_{\mathbf{W}\in\mathcal{W}_s} \underbrace{\mathcal{L}_{\text{exp}}(\mathbf{W})}_{\text{data fit + acyclicity}} + \underbrace{\mathcal{R}_{\text{prior}}(\mathbf{W})}_{\text{prior}}$$

- **Why it is hard**
  - ⇒ **(a) Observational data:** passive measurements, no interventions.
  - ⇒ **(b) Direction is ambiguous:** one dependence, several causal explanations.
  - ⇒ **(c) Acyclicity:** combinatorial, global, with nonconvex surrogates.



(b) both directions fit (c) no cycles allowed (c) global: one edge, long cycle

Our approach in a nutshell

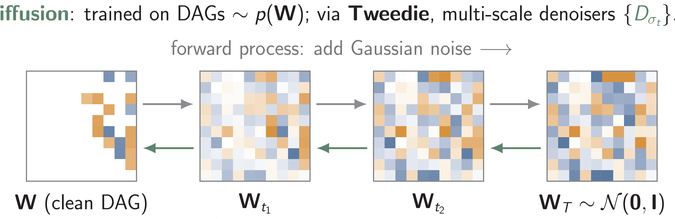
Plug-and-play: gradient steps on data fit + acyclicity, alternated with a **learned denoiser over DAGs** as an implicit prior.

Background: From Combinatorial Search to Learned Priors

- **Score-based search** over DAGs $\mathbb{D} \to$ **NOTEARS**: smooth h with zero level set $\mathbb{D}$:

$$\underbrace{\min_{\mathcal{G}(\mathbf{W})\in\mathbb{D}} \mathcal{S}(\mathcal{G}(\mathbf{W}); \mathbf{X})}_{\text{combinatorial: } |\mathbb{D}_8|\approx 7.8\times 10^{11}} \iff \underbrace{\min_{\mathbf{W}} \mathcal{S}(\mathbf{W}; \mathbf{X}) \;\text{ s.t. }\; h(\mathbf{W}) = 0}_{\text{continuous: first-order methods}}$$

  - ⇒ DAGMA: $h(\mathbf{W}) = -\log\det(s\mathbf{I} - \mathbf{W}\circ\mathbf{W}) + d\log s$ on $\mathcal{W}_s = \{\det(\cdot) > 0\}$.
  - ⇒ Regularization on $\mathbf{W}$ stays **hand-crafted** (e.g., ℓ_1 for sparsity).
- **Plug-and-play** (Venkatakrishnan et al., 2013): any **denoiser** can act as a regularizer.
- **Diffusion**: trained on DAGs $\sim p(\mathbf{W})$; via **Tweedie**, multi-scale denoisers $\{D_{\sigma_t}\}$.



⟵ learned denoisers D_{σ_t}: the structural prior used by DAG-PnP

Contributions

1. A **diffusion prior over weighted DAGs**, built on a direction-aware architecture.
2. A **principled, gradient-based method** that plugs in this **learned prior**.
3. A **barrier-aware step size**, with feasibility and local convergence guarantees.

DAG-PnP: MAP Estimation with a Learned Implicit Prior

- Causal discovery as **regularized (MAP-type) estimation**:

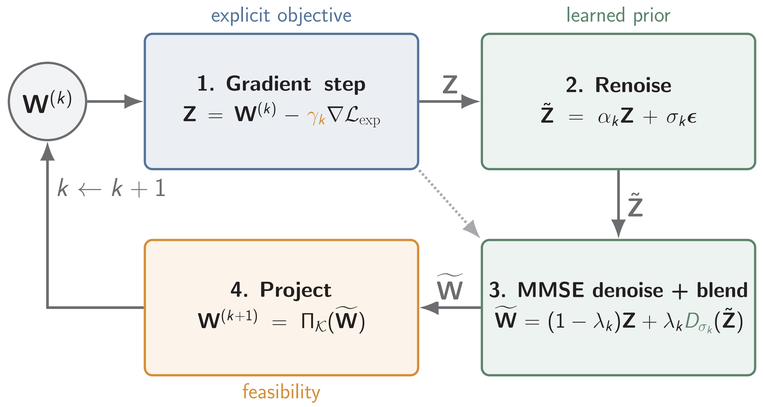
$$\min_{\mathbf{W}\in\mathcal{W}_s} \underbrace{\lambda_{\text{data}}\mathcal{L}_{\text{data}}(\mathbf{W}) + \lambda_{\text{dag}}\, h(\mathbf{W})}_{\text{Gaussian fit + acyclicity}} + \underbrace{\mathcal{R}_{\text{prior}}(\mathbf{W})}_{\text{role of } -\log p(\mathbf{W})}$$

- **Proximal gradient** alternates two steps:

$$\begin{cases} \mathbf{Z} = \mathbf{W}^{(k)} - \gamma_k \nabla\mathcal{L}_{\text{exp}}(\mathbf{W}^{(k)}) & \text{gradient step} \\ \mathbf{W}^{(k+1)} = \arg\min_{\mathbf{W}} \underbrace{\frac{1}{2\sigma^2}\|\mathbf{W} - \mathbf{Z}\|_F^2}_{\text{stay close to } \mathbf{Z}} + \underbrace{\mathcal{R}_{\text{prior}}(\mathbf{W})}_{\text{likely under the prior}} & \text{proximal step} \end{cases}$$

  - ⇒ The proximal step is **MAP denoising** of $\mathbf{Z}$: PnP replaces it with the learned $D_\sigma(\mathbf{Z})$.

Classical SP backbone (MAP, proximal denoising, Tweedie) + **a learned prior**



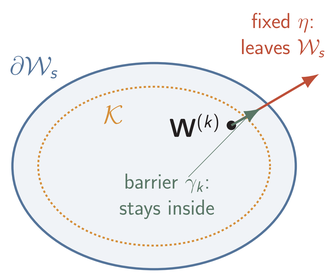
- **Noise schedule:** noise level and blend weight λ_k decrease: strong prior early, data fit late.
- **Projection** onto a safe set $\mathcal{K}$ (away from $\partial\mathcal{W}_s$): rescale so $\|\mathbf{W}\|_F \le W_{\max}$.

Barrier-Aware Step Size

- The gradient of h **explodes** near the boundary $\partial\mathcal{W}_s$, so a fixed step can **leave** the domain.
- Let $\delta(\mathbf{W})$ measure how close $\mathbf{W}$ is to $\partial\mathcal{W}_s$. Our interior-point-style rule:

$$\gamma_k = \min\left(\eta,\; c\,\frac{\delta(\mathbf{W}^{(k)})}{\|\nabla\mathcal{L}_{\text{exp}}(\mathbf{W}^{(k)})\|_F + \epsilon}\right)$$

  - ⇒ Far from $\partial\mathcal{W}_s$: base step η.
  - ⇒ Near $\partial\mathcal{W}_s$: step shrinks with δ.



Theoretical Guarantees

Lemma 1 (Feasibility; informal)

With a small enough safety factor c, the barrier-aware gradient step **never leaves** the DAGMA domain $\mathcal{W}_s$.

Theorem 1 (Local convergence; informal)

If the gradient-plus-denoiser map is locally contractive, the iterates **converge linearly** to a unique fixed point $\mathbf{W}^\star$ that balances data fit, acyclicity, and the learned prior.

- Classical PnP-style analysis; **formal statements and proofs in the paper**.

Numerical Experiments

- **Setup:** ER4/SF4 linear SEMs, **low-data** $n = d$, 10 seeds. Prior trained at $d = 20$.
- **Baselines:** NOTEARS, DAGMA (= DAG-PnP without denoiser), Greedy-BIC.
- **SHD/F1** computed after thresholding $|\hat{W}_{ij}|$ at its mean (off-diagonal).

Synthetic SEMs. **Bold**: best. Underline: second best.

		ER4				SF4				
d	Method	SHD↓	h↓	Fit↓	F1↑	SHD↓	h↓	Fit↓	F1↑	Time (s)
20	Greedy-BIC	**89.3**	**0.000**	8.57	0.244	**88.5**	**0.000**	8.41	0.206	1.4
	NOTEARS	142.8	0.239	1.09	0.359	145.1	0.216	1.19	0.326	0.6
	DAGMA	146.5	0.585	**0.77**	0.291	147.1	0.535	**0.87**	0.269	4.1
	DAG-PnP	130.2	0.176	1.43	**0.391**	135.2	0.150	1.57	**0.345**	6.1
50	Greedy-BIC	**255.5**	**0.000**	23.96	**0.268**	**256.1**	**0.000**	23.03	**0.242**	101.0
	NOTEARS	888.6	0.215	3.41	0.230	895.1	0.209	3.66	0.207	1.6
	DAGMA	931.6	0.487	**2.25**	0.213	921.5	0.499	**2.43**	0.199	1.8
	DAG-PnP	820.2	0.213	3.61	0.246	806.4	0.195	3.93	0.229	6.0

- Best SHD and h among continuous methods; best F1 overall at $d = 20$.
- Higher Fit: the prior trades data fit for plausible structure.
- Greedy-BIC wins SHD but is $\sim 72\times$ slower from $d = 20$ to $d = 50$.

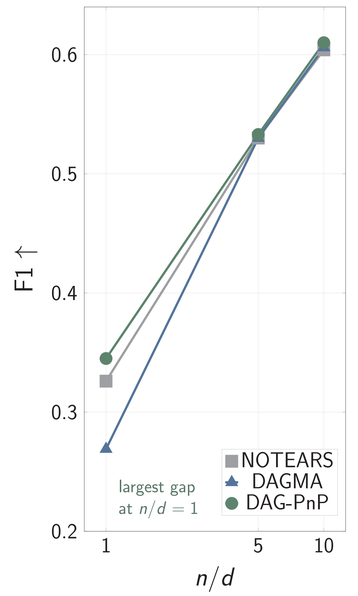
Ablation ($d = 20$, SF4).

Config.	SHD↓	h↓	Fit↓	F1↑
Fixed, no proj.	162.9	5.30	**0.002**	0.300
Barrier, no proj.	161.7	5.39	**0.002**	0.298
Fixed, proj.	139.0	0.23	1.33	0.326
Barrier, proj.	**135.2**	**0.15**	1.57	**0.345**

Sachs ($d = 11$, $n = 853$).

Method	SHD↓	h↓	Fit↓	F1↑
Greedy-BIC	**23.0**	**0.000**	1.91	0.26
NOTEARS	44.9	0.002	**0.98**	0.31
DAGMA	29.4	0.015	2.19	0.30
DAG-PnP	26.7	0.019	5.44	**0.45**

F1 vs. n/d ($d = 20$, SF4).



- **Ablation:** projection keeps iterates feasible; the barrier step adds a further gain.
- **Sachs:** single ground truth, so the prior is the synthetic ER4 one; best F1, higher fit error.
- **Plot:** the prior helps most when data are scarce.

Takeaways

- A **learned prior over DAGs** helps continuous methods most with **scarce data**.
- **Projection** keeps iterates feasible; the **barrier step** adds a further gain.
- **Limits**: linear SEMs, local theory. **Next**: nonlinear SEMs, global convergence.