Causal Discovery via Bayesian Optimization

Bao Duong, Sunil Gupta & Thin Nguyen

ICLR

A²I² APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Introduction

Problem

- Score-based causal discovery (SCD):

$$\mathcal{G}^* = \underset{\mathcal{G} \in \text{DAGs}}{\arg\max}\, S\left(\mathcal{D}, \mathcal{G}\right).$$

- Challenges:
    1. **Constraint**: the graphs must be acyclic.
    2. **Scalability**: high-dim & many trials.
    3. **Sample-efficiency**: score calculation can be expensive.

Introduction

Existing approaches

- Greedy search (e.g., GES)
  - **Slow exploration**: add/remove one edge at a time.
- Continuous optimization (e.g., NOTEARS, DAGMA, etc.)
  - **Lack exploration**: only follow the gradient direction.
- Reinforcement learning (e.g., RL-BIC, CORL, ALIAS, etc.)
  - **Inefficient exploration**: blindly explore random DAGs without pre-examining their potential.

Introduction

Motivation

- By modelling the explored DAG scores to detect promising exploration candidates, we may arrive at better solutions earlier

→ Bayesian Optimization (BO)

- Applying BO directly to SCD is hard:
  - SCD is usually **high-dim** and **constrained**, while BO works well **low-dim** and **unconstrained**.
  - BO **scales poorly** with #trials, while we may need **thousands or more trials** for SCD.
  - Acquisition function optimization in BO is itself a SCD problem, thus **requiring to be very efficient** to be practical.

→ we propose the **first BO-based SCD method** for sample-efficiency by making several innovations.

Introduction

Our work

- 4 innovations to specifically adapt BO to SCD:
    1. Low-rank unconstrained search space → addressing **acyclicity** & **dimensionality**.
    2. Replacing GPs with Dropout networks for surrogate modelling → addressing **scalability**.
    3. Indirect DAG score modelling → addressing **surrogate modelling accuracy**.
    4. Continual model training → addressing **scalability**.
- These enables **accurate** and **sample-efficient** SCD, as verified through extensive experiments and ablations.

Introduction

Our findings

- DrBO is highly accurate & sample-efficient compared with existing SOTAs.
    - **Accuracy:** SHD ≈ 0 for linear & nonlinear data, dense & large graphs, synthetic & real data.
    - **Sample-efficiency**: SHD ≈ 0 is reached earlier than other methods in both number of DAG evaluations & time.
- Ablations confirm that:
    - Lower rank = better sample-efficiency.
    - Dropout nets scale better than GPs.
    - Indirect DAG modelling = more accuracy.
    - Continual training = linear scalability.

Proposed method: DrBO

Low-dim unconstrained search space

- We turn the **constrained** optimization problem to an easier **unconstrained** problem with **low-dim search space**:

$$\mathcal{G}^* = \underset{\mathcal{G} \in \text{DAGs}}{\arg\max}\, S\left(\mathcal{D}, \mathcal{G}\right) \iff \mathbf{z}^* = \underset{\mathbf{z} \in \mathbb{R}^{d(1+k)}}{\arg\max}\, S\left(\mathcal{D}, \tau\left(\mathbf{z}\right)\right)$$

- The map τ turns **unconstrained continuous-value parameters** to a **DAG**:

$$\tau\left(\mathbf{p}, \mathbf{R}\right) := \underbrace{H\left(\text{grad}\left(\mathbf{p}\right)\right)}_{\text{ensures acyclitiy}} \odot \underbrace{H\left(\mathbf{R} \cdot \mathbf{R}^\top\right)}_{\text{low-rank connectivity}},$$

where $R \in \mathbb{R}^{d \times k}$ $(k \ll d)$ is an embedding matrix and z is concatenation of p and R.

- This is a low-rank adaptation of Vec2DAG (Duong et al., 2024).

→ Search dimensionality **scales linearly with d** and **allows generating more diverse DAGs**.

Duong, Bao et al. "ALIAS: DAG Learning with Efficient Unconstrained Policies." 2024.

Proposed method: DrBO

Acquisition function optimization

- Acquisition function optimization = SCD with acquisition function values as scores → sampling-based approach for efficiency:
  - Trust-region sampling: random $\{\boldsymbol{z}^{(j)}\}_{j=1}^{C}$ are generated from a hypercube centred at best solution so far $\boldsymbol{z}^*$.
  - Then, top-B candidates with highest acquisition function values are chosen.
- Larger C = higher-quality candidates → acquisition function evaluation must scale very well.

Proposed method: DrBO

Surrogate Modelling with Dropout Networks

- GPs scale cubically with number of datapoints, both in training and sampling.
- Dropout nets = approximate Bayesian inference (Gal & Ghahramani, 2016).

$$\mathrm{DropoutNN}\left(\mathbf{x}\right) := \mathbf{W}_2^\top \left(\mathrm{BatchNorm}\left(\mathrm{ReLU}\left(\frac{1}{1-p}\left(\left(1-\mathbf{m}\right)\circ\left(\mathbf{W}_1^\top\mathbf{x}+\mathbf{b}_1\right)\right)\right)\right)\right)+b_2.$$

- A forward pass $y \sim DropoutNN(x) \approx$ sampling from $P(y|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y})$ = Thompson sampling as acquisition function.

→ **constant-time acquisition function evaluation**.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning." ICML, 2016.

Proposed method: DrBO

Indirect Surrogate Modelling

- Naïve approach: train a network predicting $S(D, G)$ directly from G.
- However, partial scores are not well exploited. E.g.:

$$S_{\text{BIC-EV}}(\mathcal{D}, \mathcal{G}) := -nd \ln \frac{\sum_{i=1}^{d} \mathrm{MSE}_i\left(\mathrm{pa}_i^{\mathcal{G}}\right)}{d} - |\mathcal{G}| \ln n.$$

→ we use the evaluation data $\left\{\left(\mathrm{pa}_i^{\mathcal{G}^{(j)}}, \mathrm{MSE}_i\left(\mathrm{pa}_i^{\mathcal{G}^{(j)}}\right)\right)\right\}$ to train **separate dropout networks**, then combine the predictions:

$$\hat{S}_{\text{BIC-EV}}(\mathcal{D}, \mathcal{G}) := -nd \ln \frac{\sum_{i=1}^{d} \widehat{\mathrm{MSE}_i}\left(\mathrm{pa}_i^{\mathcal{G}}\right)}{d} - |\mathcal{G}| \ln n.$$

- Now all information is fully exploited → accurate score estimates.

Proposed method: DrBO

Continual Model Training

- Retraining the neural nets every BO iteration is costly, which prevents scaling to many trials.

→ we instead train them continually: each iteration apply several gradient steps on the **new data combined with a random batch of past data**.

→ **constant-time model update.**

Proposed method: DrBO

Overall Algorithm

Algorithm 1 The **DrBO** method for causal discovery.

Require: Dataset $\mathcal{D} = \left\{\mathbf{x}^{(j)} \in \mathbb{R}^d\right\}_{j=1}^{n}$ of d nodes and n observations, score function $S(\mathcal{D}, \cdot)$, DAG rank k, batch size B, no. of preliminary candidates C, and total no. of evaluations T.

Ensure: A DAG $\hat{\mathcal{G}}$ that maximizes $S(\mathcal{D}, \mathcal{G})$.

```
Initialize empty experience $\mathcal{H} := \emptyset$ and node-wise dropout neural nets: $\{\text{DropoutNN}_i\}_{i=1}^{d}$.
while $|\mathcal{H}| < T$ do
    Generate random DAGs: $\left\{\mathcal{G}^{(j)} := \tau\left(\mathbf{z}^{(j)}\right)\right\}_{j=1}^{C}$ where $\mathbf{z} \in [-1, 1]^{d(1+k)}$.   ▷ Secs. 4.1 & 4.2.
    Sample local scores: $\left\{\left\{l_i^{(j)} \sim \text{DropoutNN}_i\left(\text{pa}_i^{\mathcal{G}^{(j)}}\right)\right\}_{i=1}^{d}\right\}_{j=1}^{C}$.   ▷ Sec. 4.3.
    Combine local scores: $\left\{\text{AF}^{(j)} := \text{Combine}\left(l_1^{(j)}, \ldots, l_d^{(j)}\right)\right\}_{j=1}^{C}$.   ▷ Sec. 4.4.
    Select top $B$ candidates with highest AF values: $j_1, \ldots, j_B := \operatorname{argtop}_{B\, j=1,\ldots,C} \text{AF}^{(j)}$.   ▷ Sec. 4.2.
    Evaluate these candidates and update experience: $\mathcal{H} := \mathcal{H} \cup \left\{\left(\mathcal{G}^{(j)}, S\left(\mathcal{D}, \mathcal{G}^{(j)}\right)\right)\right\}_{j=j_1,\ldots,j_B}$.
    Update the neural nets on new $\mathcal{H}$.   ▷ Sec. 4.5.
end while
Get highest-scoring DAG so far: $\hat{\mathcal{G}} := \arg\max_{\mathcal{G} \in \mathcal{H}} S(\mathcal{D}, \mathcal{G})$.
Prune $\hat{\mathcal{G}}$ if needed.   ▷ Sec. 4.6.
```

Experiments

Synthetic data

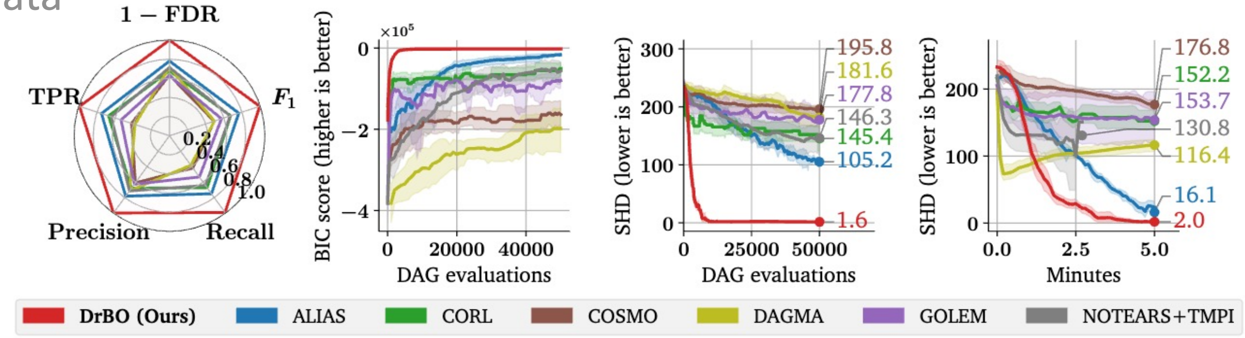


Figure 1. Linear-Gaussian data with dense graphs (30-node ER-8).

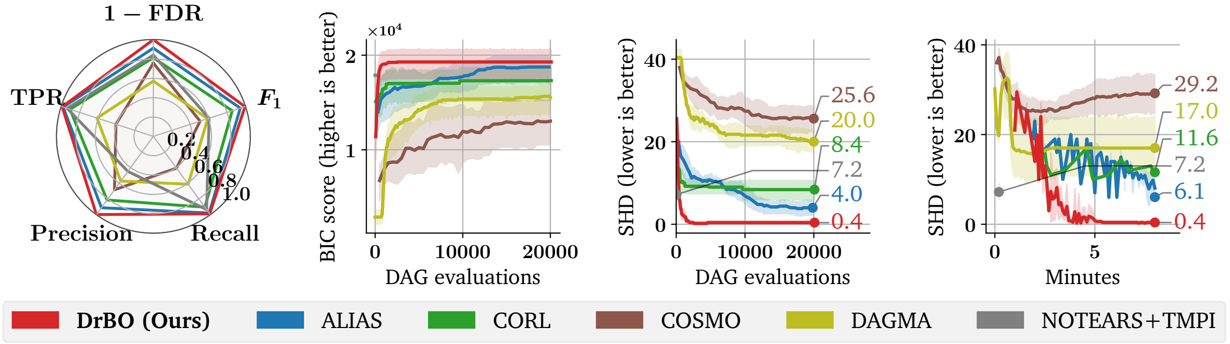


Figure 2. Non-linear data.

Experiments

Real data

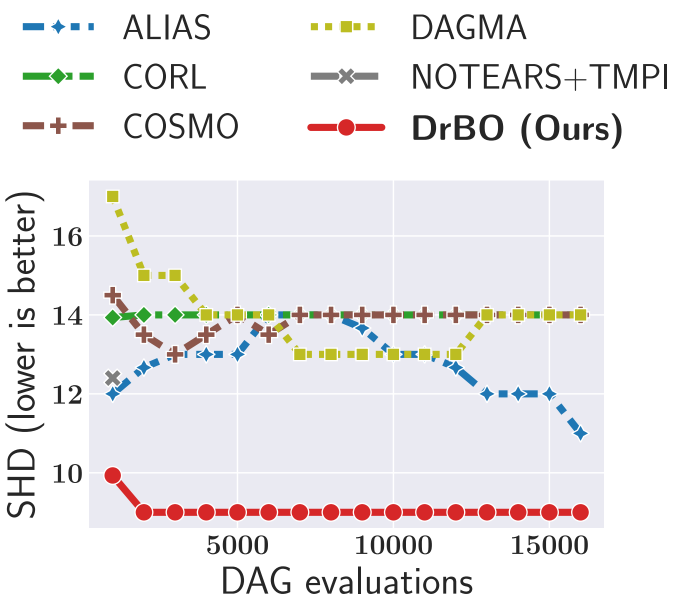
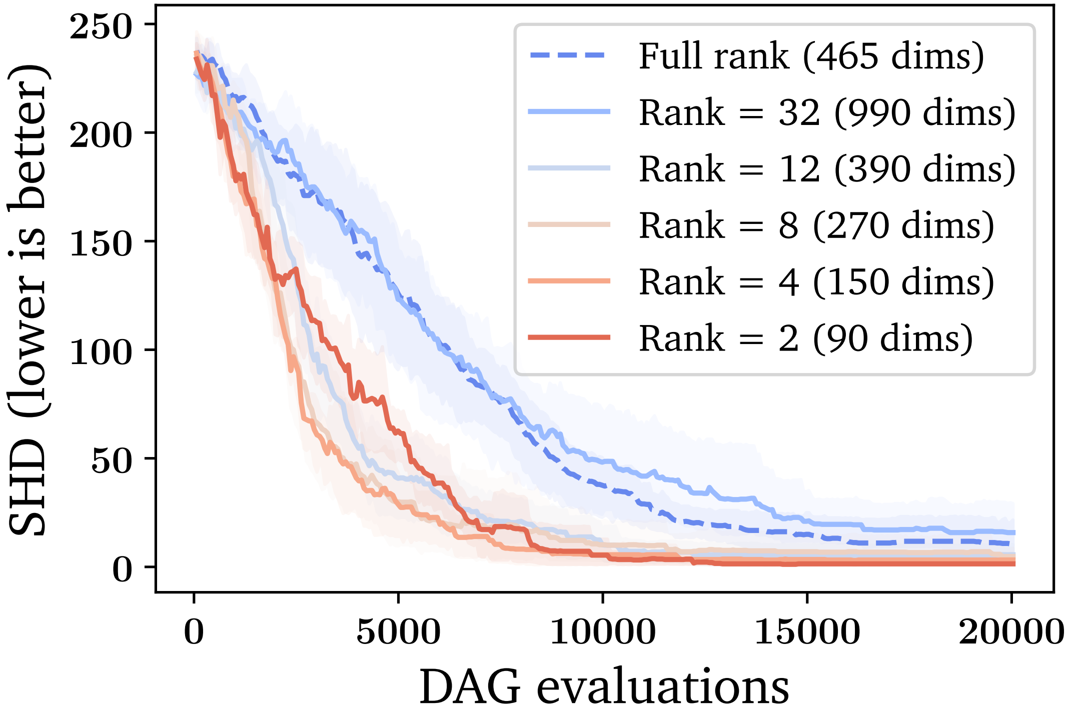


Figure. Causal Discovery performance on the Sachs dataset

Ablation studies

Lower rank = more sample-efficiency



Ablation studies

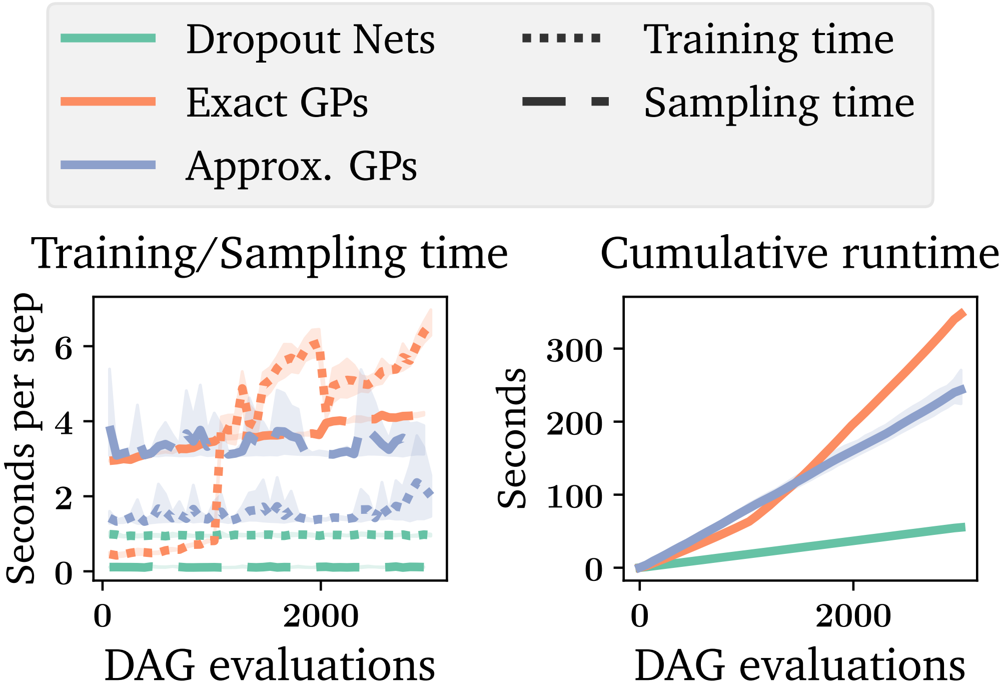
Lower rank = more diverse candidates

Table 5: **Effect of DAG Rank on Exploration Diversity.** We generate 1,000 DAGs with $d = 30$ nodes using $\mathcal{G} := \tau(\mathbf{z}), \mathbf{z} \in [-1, 1]^{d\cdot(1+k)}$ with different k. The numbers are mean $\pm$ std over 10 simulations.

Rank k in Eq. (4)	**Number of dimensions**	**Number of unique 30-node DAGs over 1,000 random DAGs**
2	90	926.7 ± 7.0
4	150	779.2 ± 12.7
8	270	493.5 ± 12.3
12	390	332.4 ± 10.8
32	990	90.7 ± 9.5
Full rank (Vec2DAG, Duong et al., 2024)	465	421.9 ± 13.8

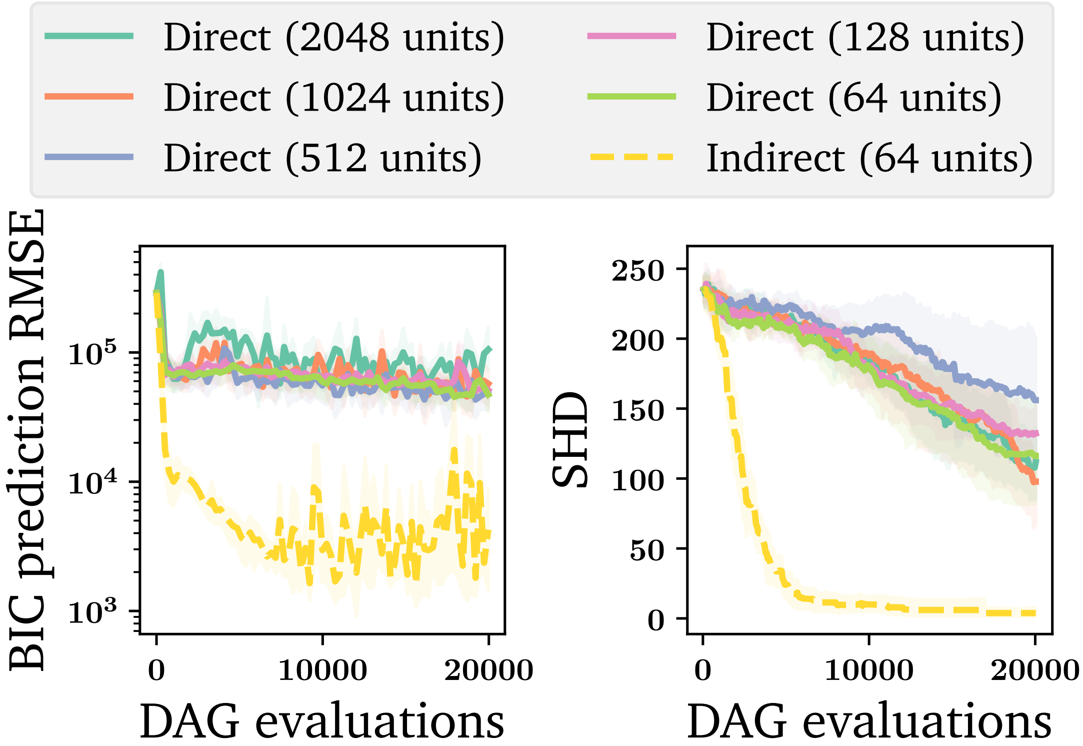
Ablation studies

Dropout Nets scale much better than GPs



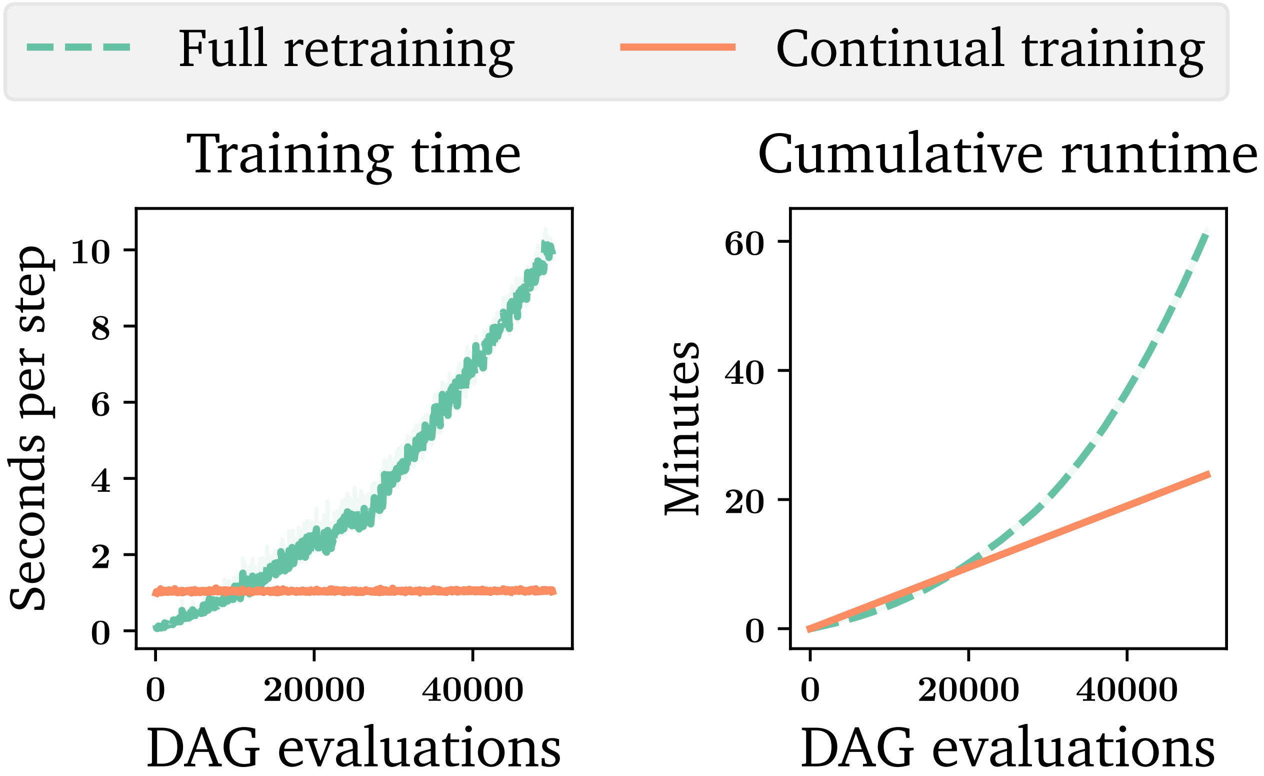
Ablation studies

Indirect DAG Modelling = more accuracy



Ablation studies

Continual Training = linear scalability



Key takeaways

- We propose to the use of Bayesian optimization for **sample-efficient score-based causal discovery**.
- 4 innovations to specifically adapt BO to SCD:
    1. Low-rank unconstrained search space.
    2. Replacing GPs with Dropout networks for surrogate modelling.
    3. Indirect DAG score modelling.
    4. Continual model training.
- These enables **accurate** and **sample-efficient** SCD, as verified through extensive experiments and ablations.