Learning to refine domain knowledge
for biological network inference

Peiwen Li
Shenzhen International Graduate School
Tsinghua University
lpw22{at}mails.tsinghua.edu.cn

Menghua Wu
Department of Computer Science
Massachusetts Institute of Technology
rmwu{at}mit.edu

Abstract

Perturbation experiments allow biologists to discover causal relationships between variables of interest, but the sparsity and high dimensionality of these data pose significant challenges for causal structure learning algorithms. Biological knowledge graphs can bootstrap the inference of causal structures in these situations, but since they compile vastly diverse information, they can bias predictions towards well-studied systems. Alternatively, amortized causal structure learning algorithms encode inductive biases through data simulation and train supervised models to recapitulate these synthetic graphs. However, realistically simulating biology is arguably even harder than understanding a specific system. In this work, we take inspiration from both strategies and propose an amortized algorithm for refining domain knowledge, based on data observations. On real and synthetic datasets, we show that our approach outperforms baselines in recovering ground truth causal graphs and identifying errors in the prior knowledge with limited interventional data.

1 Introduction

Large-scale perturbation experiments have the potential to uncover extensive causal relationships between biomolecules (Replogle et al., 2022), which may facilitate myriad applications in drug discovery, from disease understanding to mechanism of action elucidation (Schenone et al., 2013). Causal structure learning (discovery) algorithms are designed to extract these very relationships directly from data (Spirtes et al., 2001). Yet due to the high number of variables (genes), compounded with the low numbers of observations (cells) per setting (perturbation) (Nadig et al., 2024), these algorithms struggle to scale and perform robustly on such datasets. A key challenge is that causal discovery algorithms must not only infer the causal direction between variables, but also which variables are related in the first place. The latter can be alleviated in part by incorporating noisy priors regarding the data, e.g. by initializing the graph prediction using a biological knowledge graph (Ashburner et al., 2000). However, these graphs compile decades of discoveries from disparate experiments, rendering their relevance and correctness uncertain in individual cellular contexts. While the choice and quality of these priors generally does not impact consistency in the infinite data limit (Hauser & Bühlmann, 2012; Hägele et al., 2023), there are rarely sufficient data for these guarantees to hold in practice.

An orthogonal line of work aims to capture inductive biases that cannot be easily represented by individual graphs via amortized inference over synthetic data (Ke et al., 2022; Lorch et al., 2022). A simulator first generates pairs of "ground truth" causal graphs and datasets, following known rules regarding the domain of interest. For example, biological networks have been hypothesized to be scale-free (Barabási & Bonabeau, 2003), and transcription dynamics can be described through sets of differential equations (Chen et al., 1999). Once these data have been generated, a neural network is
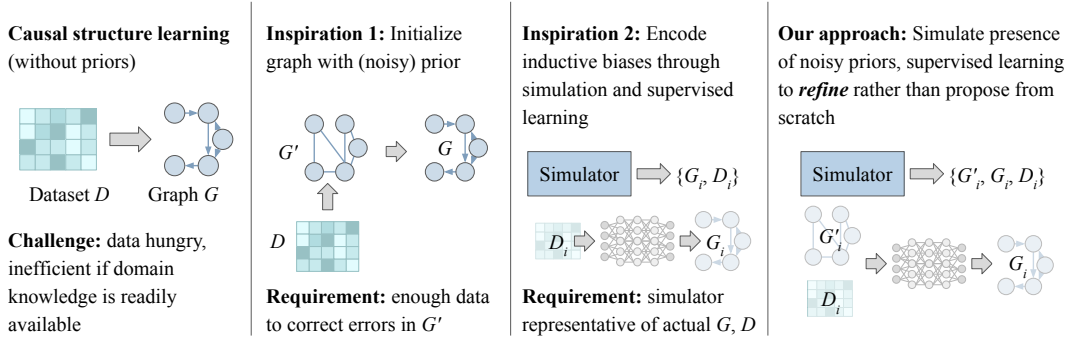
Causal structure learning (without priors)

Dataset D → Graph G

Challenge: data hungry, inefficient if domain knowledge is readily available

Inspiration 1: Initialize graph with (noisy) prior

G' → G

D

Requirement: enough data to correct errors in G'

Inspiration 2: Encode inductive biases through simulation and supervised learning

Simulator → $\{G_i, D_i\}$

D_i → G_i

Requirement: simulator representative of actual G, D

Our approach: Simulate presence of noisy priors, supervised learning to *refine* rather than propose from scratch

Simulator → $\{G'_i, G_i, D_i\}$

G'_i

D_i

→ G_i

Figure 1: Drawing inspiration from amortized causal discovery algorithms, we learn how to refine graph priors, enabling robust graph predictions in low-data regimes.

trained to map the datasets to their associated causal graphs. Empirically, the resultant models are more robust to low-data situations in real settings, as they have ideally seen similar, synthetic datasets over the course of training (Wu et al., 2024). However, since these approaches incorporate inductive biases via simulation, it is imperative that the simulators accurately reflect the true data. In reality, it has been observed that truly scale-free graphs are the minority in biology (Broido & Clauset, 2019), and transcription dynamics can be highly heterogeneous and discontinuous, even within the same cell type (Lenstra et al., 2016).

In this work, we propose an amortized inference framework for *refining* noisy graph priors, to identify causal relationships in low-data regimes. We are motivated by the idea that while simulating biological data is hard, simulating the types of noise that occur may be easier. For example, gene-gene relationships may increase/decrease in strength, or appear/disappear based on the cell state, but the directionality of these relationships rarely changes (Belyaeva et al., 2021). Our model architecture is based on the supervised causal discovery model in Wu et al. (2024), in which datasets are featurized in terms of local causal graph estimates and summary statistics such as global correlation. During training, we augment these inputs with a corrupted, undirected graph, sampled at varying levels of noise. The neural network is forced to identify incorrect edges based on data, as well as orient edges where feasible.

We perform extensive experiments on both synthetic datasets and the real-world Sachs proteomics dataset (Sachs et al., 2005), evaluating the model's ability to 1) predict the ground truth graph and 2) detect errors in noisy priors, particularly under various down-sampling settings to assess performance in low-data regimes. Across all noisy prior conditions and sample sizes, OURS consistently achieves strong results in both causal structure learning and error detection in noisy priors. In contrast, baseline methods are significantly affected by the quality of prior knowledge and the amount of data. We conclude that high-quality graph priors provide strong starting points for inferring causal relationships, especially when limited data are available, and learning to denoise these priors is more data-efficient than using them to initialize graph predictions.

2 Background and related work

2.1 Biological network inference

Biological network inference is a classic systems biology problem, in which the goal is to uncover interactions between experimentally-quantifiable entities (e.g. genes, proteins) in the form of graphs (Albert, 2007; Huynh-Thu & Sanguinetti, 2018). For example, graphs of interest include gene regulatory networks (Badia-i Mompel et al., 2023), protein-protein interaction networks (Tsitsiridis et al., 2022), and metabolic pathways (Milacic et al., 2024). Early efforts towards biological network inference included the DREAM challenges (Marbach et al., 2012), which provided harmonized microarray data and were evaluated against known interactions at the time. Probabilistic graphical models are commonly used to infer gene regulatory networks in specific disease areas (Mao & Resat, 2004; Zhao & Duan, 2019; Dai et al., 2024). More recent works have also used graph neural networks to predict "missing" edges in these graphs (Feng et al., 2023). However, while biological network

inference algorithms have been applied to a variety of disease areas, many of these methods are typically engineered towards their respective datasets and do not share reproducible code.

2.2 Causal structure learning

Causality provides a formal framework for inferring data-generating mechanisms from experimental data. A causal graphical model is defined by a distribution P_X over a random variables X, associated with a directed acyclic graph $G = (V, E)$, where each node $i \in V$ corresponds to a random variable $X_i \in X$, and each edge $(i, j) \in E$ indicates a direct causal relationship from X_i to X_j (Spirtes et al., 2001). It is common to assume that the data distribution P_X is Markov to G, i.e. variables X_i are independent of all other $X_j \notin X_{\delta_i} \cup X_{\pi_i}$ (not descendants or parents), given its parents X_{π_i}. Causal graphical models introduce the concept of interventions on node i, by changing the conditional distribution $P(X_i \mid X_{\pi_i})$ to a new distribution $\tilde{P}(X_i \mid X_{\pi_i})$.

Causal structure learning is the task of predicting causal graph G from dataset $D \sim P_X$. Classical *discrete optimization methods* operate over the combinatorial space of edge sets, and they make discrete changes to add/delete/orient edges. These include constraint-based algorithms, such PC and FCI for observational data (Spirtes et al., 1995), and JCI for mixed data (Mooij et al., 2020). While these algorithms can be initialized with an undirected graph as a prior, they cannot recover edges that are not present in this initial skeleton. There are also score-based methods that optimize a score, which represents the "goodness" of a particular graph, with respect to the data. These include GES (Chickering, 2002), GIES (Hauser & Bühlmann, 2012), CAM (Bühlmann et al., 2014), LiNGAM (Shimizu et al., 2006) and IGSP (Wang et al., 2017). Algorithms like GIES iterate between adding and deleting edges, so they can (in principle) identify edges that are missing from an initial estimate. However, due to the exponential space of potential graphs and the reliance on statistical power for discrete judgments, these classical approaches scale poorly with the number of variables and require copious data for reliable performance.

On the other hand, *continuous optimization methods* approach causal discovery through constrained continuous optimization over weighted adjacency matrices. Many of these approaches, exemplified by NoTears (Zheng et al., 2018), DCDI (Brouillard et al., 2020), and GranDAG (Lachapelle et al., 2020) train a generative model to capture the empirical data distribution, which is parameterized through the adjacency matrix. Several works have also been specifically designed to address challenges in biological problems. DCD-FG (Lopez et al., 2022) aims to scale to single-cell transcriptomics data and proposes a low-rank extension of DCDI. The hybrid IGSP algorithm (Wang et al., 2017) has also been applied to single-cell data. Prior knowledge can be used to initialize the graph parameters in these frameworks, but the same limitations apply with regards to data-efficiency.

2.3 Biological knowledge graphs for perturbations

Knowledge graphs have been indispensable to modeling biological perturbations. They are commonly used as undirected graphs, over which graph neural networks predict the cellular effects of unseen perturbation (Roohani et al., 2023; Bai et al., 2024) or infer perturbation targets for active learning (Huang et al., 2023) and target discovery (Gonzalez et al., 2024). This work focuses on an adjacent but distinct task: of inferring relationships between variables, rather their effects or identity as targets.

3 Methods

Let $D \sim P_X$ be a dataset containing M samples of N variables, and let $G = (V, E)$ be the causal graph that generated P_X. Let $G' = (V, E')$ be an undirected graph, where $E + E^\mathsf{T} \approx E'$ but $E + E^\mathsf{T} \neq E'$. Given D and E', the goal is to predict E.

3.1 Inference

When given a new dataset D and graph prior E', we summarize D in terms of local and global summary statistics, which are combined with E' as input to an attention-based neural network, trained to predict E (Figure 2A). We adapt the Sample, Estimate, Aggregate workflow (SEA, Wu et al. (2024)) to the task of refining noisy graph priors based on data observations as follows.

A) Inference procedure on real data   **B) Training procedure on simulated data**
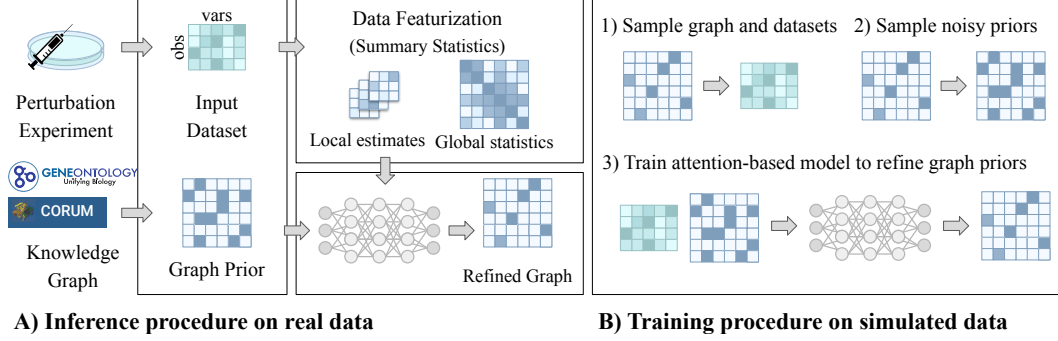
Figure 2: A) At inference, we use biological knowledge graphs as noisy graph priors, which we refine with perturbation data. B) We train an attention-based model to denoise simulated graph priors.

1. (Sample) We sub-sample batches from dataset D to focus on sets of variables that are likely to be related, using heuristics like correlation.
2. (Estimate) We compute pairwise correlation ρ between all variables as a summary statistic; and run classical causal discovery algorithms over smaller batches, sampled in (1).
3. (Aggregate) A neural network is provided the $N \times N$ noisy graph estimate E', the $N \times N$ summary statistic, and T estimates of size $k \times k$ (where $k = 5$ is small compared to N). The final output is a $N \times N$ matrix that represents the predicted causal graph, refined from the noisy prior with edges oriented.

3.2 Training

At inference time, E' is provided (either a corrupted synthetic graph or a biological knowledge graph). During training, we sample graph priors by adding noise to the ground truth undirected graph. We compute E' as follows.

1. Sample noise level $p \sim \mathrm{Uniform}(0, 0.5)$, and binary mask $M \in \mathbf{1}^{N \times N}$ where

$$M_{i,j} = \mathbf{1}\{z_{i,j} \sim \mathrm{Uniform}(0, 1) < p\}. \tag{1}$$

2. Compute undirected graph $\tilde{E} = E + E^{\mathsf{T}}$.
3. Compute noisy prior E' where

$$E'_{i,j} = \begin{cases} \tilde{E}_{i,j} & M_{i,j} = 0 \\ 1 - \tilde{E}_{i,j} & \text{otherwise.} \end{cases} \tag{2}$$

We finetune all weights with the binary classification objective of predicting E, at the edge level. Thus, the objective both encourages the model to denoise E' and orient edges.

3.3 Implementation details

We adopt the same axial-attention architecture as SEA. Specifically:

1. The $k \times k \times T$ marginal estimates (local graph structures, inferred by standard causal discovery algorithms) are aligned by matching the same edges across estimates, and mapped to a $K \times T \times d$ *marginal* feature, where K is the number of unique edges.
2. E' is embedding using the same edge embeddings as marginal graphs (since they are in the same input space), and the result is added to that of the global statistic (since they are over the same nodes). The result is a $N \times N \times d$ *global* feature.
3. A series of 2D axial attention layers (Ho et al., 2020) attend over the rows and columns of both matrices. The final output is a $N \times N$ matrix, which is supervised by the (synthetic) ground truth E.

Our synthetic training set contains approximately 4000 datasets of size $N = 10, 20$ with linear additive and neural network (additive and non-additive) causal mechanisms. We use pretrained SEA weights with the GIES (Hauser & Bühlmann, 2012) estimation algorithm and inverse covariance as the global statistic. The introduction of E' does not add any new parameters, as we use the same embeddings as the existing edge estimates, which support undirected edges.

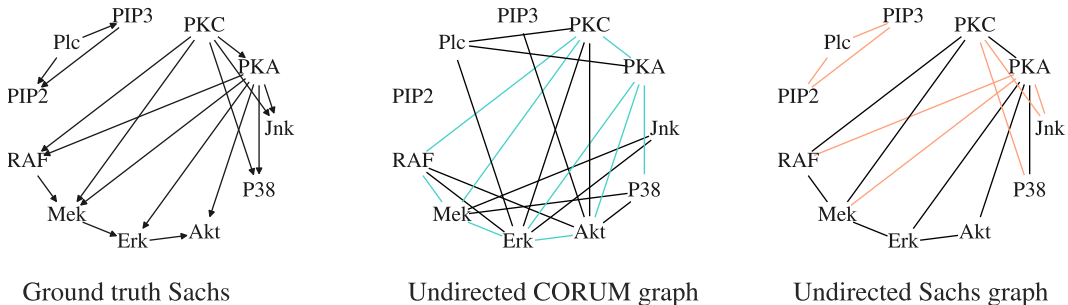| Ground truth Sachs | Undirected CORUM graph | Undirected Sachs graph |

Figure 3: Visualization of ground truth Sachs consensus graph (Sachs et al., 2005) and CORUM knowledge graph (Tsitsiridis et al., 2022). Blue: Undirected edges present in both CORUM and Sachs. Orange: Undirected edges present in Sachs but not CORUM. 9 of 17 undirected edges in Sachs are present on the CORUM graph; of the 38 pairs of nodes that have no relationship in Sachs, 26 also have no relationship in CORUM.

4 Experimental setup

We evaluate our approach against a variety of causal structure learning baselines on real and synthetic datasets. While it would be ideal to evaluate entirely on real applications, synthetic experiments allow us to systematically assess how performance changes based on the quality of our noisy priors and the availability of interventional data. In real settings, it is difficult to quantify the relevance of knowledge graphs to each cell line, disease type, or other factors, as these labels themselves are inherently approximations to the underlying biology.

4.1 Data preparation

Biological experiments The Sachs proteomics dataset (Sachs et al., 2005) is a common benchmark for causal structure learning approaches. In this work, we use the subset proposed by Wang et al. (2017), which contains 1755 observational samples and 4091 interventional samples, associated with a consensus graph of 11 nodes and 17 edges. We use both a corrupted version of the ground truth graph ("Synthetic KG") and the CORUM (comprehensive resource of mammalian protein complexes) knowledge graph (Tsitsiridis et al., 2022) as noisy priors. CORUM focuses on physical interactions between proteins, which are likely relevant for, but do not directly translate to quantifiable effects of perturbing certain proteins on others. Figure 3 depicts the CORUM graph alongside the ground truth.

Synthetic experiments The raw synthetic datasets were generated following DCDI (Brouillard et al., 2020): (1) sampling Erdős-Rényi graphs with $N = 10, 20$ nodes and $E = N$ expected edges; (2) sampling random instantiations of causal mechanisms (Linear, Neural Networks, Sigmoid with additive Gaussian noise, and Polynomial mechanisms); and (3) iteratively sampling observations in topological order. For each graph, we generated one observational regime and ten interventional regimes, with each regime consisting of 1000 samples, i.e., $1000 \times N$ data points, either entirely observational or following a set of single-node perfect interventions.

Down-sampling In perturbation experiments, observational data are generally easy to collect and abundant, while interventional data are costly and limited. For example, large-scale Perturb-seq datasets may include thousands of non-targeting control cells, but only a median of ~50 cells per perturbation (Nadig et al., 2024). To emulate this setting on all datasets, we sub-sample 50 or 100 examples from each interventional regime, while preserving all observational examples. This allows us to evaluate all models in realistic data settings.

4.2 Baselines

We compare our method to several state-of-the-art causal discovery algorithms, which are able to incorporate prior knowledge by initializing their graph parameters based on these undirected graphs. Specifically, the "vanilla" setting indicates that models do not consider prior knowledge, while the other settings specify the noise level ($10\%, 25\%$) that the undirected graphs were subject to (Section 3.2).

Table 1: Causal discovery and noise detection results on real proteomics dataset (Sachs et al., 2005), 50 samples per intervention. The vanilla setting is not evaluated on noise detection since there is no graph prior. The best results in each category are indicated in **bold**.

Prior Knowledge	Model	Graph		Noise	
		mAP ↑	SHD ↓	F1 ↑	Acc ↑
Vanilla	DCDI-G	0.14	23	—	—
	DCDI-DSF	0.14	29	—	—
	BACADI	0.15	20	—	—
	SEA	**0.20**	**17**	—	—
Synthetic KG (10% noise)	DCDI-G	0.14	21	0.55	0.73
	DCDI-DSF	0.15	18	0.53	0.75
	OURS	**0.25**	**17**	**0.73**	**0.80**
Synthetic KG (25% noise)	DCDI-G	0.19	19	0.67	0.76
	DCDI-DSF	0.15	23	0.35	0.60
	OURS	**0.24**	**17**	**0.78**	**0.80**
CORUM KG (34% noise)	DCDI-G	0.14	24	0.44	0.73
	DCDI-DSF	0.15	20	0.67	**0.82**
	OURS	**0.21**	**16**	**0.70**	0.78

- DCDI-G and DCDI-DSF (Brouillard et al., 2020) are two variations of DCDI, each employing a different density approximator. DCDI-G utilizes simple Gaussian distributions, while DCDI-DSF leverages the more expressive deep sigmoidal flows to represent non-linear causal relationships. To incorporate prior knowledge, we initialize the parameters of the Gumbel adjacency matrix such that the initial weighted adjacency reflects to the noisy graph. The "vanilla" DCDI initializes these parameters to a matrix of ones, i.e. a fully connected graph.

- BACADI (Hägele et al., 2023) is a fully Bayesian approach for inferring complete joint posterior over causal structure, parameters of causal mechanisms, and interventions in each experimental context.

- SEA (Wu et al., 2024) is the "vanilla" version of our model, which does not incorporate any prior knowledge and was not trained to denoise external information.

4.3 Metrics

We evaluate our models and all baselines on their ability to 1) predict the ground truth causal graph and 2) detect errors in the prior knowledge. To assess the quality of the predicted graphs, we report standard causal discovery metrics (discrete and continuous).

1. **Structural Hamming Distance** (SHD) measures the graph edit distance between the predicted DAG and the ground-truth DAG, as defined in Tsamardinos et al. (2006). Lower is better (0 is best). The discretization threshold is set to 0.5.
2. **Mean Average Precision** (mAP) calculates the area under the precision-recall curve for each edge and averages it across the entire graph. mAP ranges from 0 to 1 (best).

We treat error (noise) detection as an edge-level binary classification task. Since graph priors are symmetric, we symmetrize predicted graphs and omit the diagonal for this analysis. A "positive" label is an edge that is flipped in the noisy prior, while a "positive" prediction is an edge whose presence in the prediction differs from that in the noisy prior. We binarize with the standard threshold of 0.5.

1. **Acc** is the un-weighted accuracy of identifying errors in the undirected graph.
2. **F1 Score** is the harmonic mean of precision and recall, which provides a unified measure that considers both false positives and negatives.

Table 2: Causal discovery results on synthetic datasets where $N = 10, E = 10$. Each setting is the mean over *five* distinct datasets. Best results for each data setting are denoted in **bold**. Each dataset contains 1000 observational samples and the specified number of samples per intervention. The symbol † indicates that OURS and SEA were not trained on this setting (for fair comparison to SEA).

Samples	Noise	Model	Linear		NN non-add.		Sigmoid†		Polynomial†	
			mAP ↑	SHD ↓	mAP ↑	SHD ↓	mAP ↑	SHD ↓	mAP ↑	SHD ↓
50	Vanilla	DCDI-G	0.27±0.15	12.7±4.22	0.44±0.12	9.1±3.56	0.25±0.15	8.9±3.88	0.27±0.10	8.5±1.36
		DCDI-DSF	0.26±0.09	17.6±5.08	0.16±0.04	27.6±4.34	0.13±0.05	26.8±5.23	0.11±0.03	28.7±3.82
		BACADI	0.42±0.23	11.8±4.49	0.39±0.14	14.3±4.50	0.39±0.18	11.0±4.15	0.28±0.11	12.4±2.62
		SEA	0.94±0.05	2.0±1.55	0.84±0.07	5.4±1.43	0.78±0.18	4.2±2.36	0.64±0.15	**6.2±2.48**
	$p = 0.10$	DCDI-G	0.45±0.14	4.8±1.17	0.69±0.13	**3.7±2.00**	0.36±0.25	6.7±3.23	0.35±0.14	6.5±2.33
		DCDI-DSF	0.41±0.16	7.0±1.84	0.42±0.11	7.8±2.09	0.28±0.18	9.6±3.44	0.18±0.11	11.1±3.70
		OURS	**0.95±0.05**	**1.7±1.19**	**0.88±0.06**	4.1±1.70	**0.83±0.17**	**3.6±2.65**	**0.67±0.12**	**6.2±2.44**
	$p = 0.25$	DCDI-G	0.36±0.13	6.1±1.45	0.58±0.20	6.0±3.69	0.34±0.18	6.9±1.92	0.33±0.12	6.8±1.72
		DCDI-DSF	0.38±0.17	8.5±4.15	0.23±0.05	16.8±2.64	0.18±0.09	15.4±5.61	0.13±0.03	15.4±4.10
		OURS	0.93±0.07	1.9±1.51	0.84±0.08	6.4±2.73	0.80±0.14	4.5±2.46	0.64±0.12	6.5±2.62
100	Vanilla	DCDI-G	0.48±0.15	4.3±1.27	0.56±0.12	5.9±2.21	0.34±0.19	7.1±3.36	0.39±0.16	6.6±2.11
		DCDI-DSF	0.22±0.05	21.2±4.85	0.19±0.05	25.9±3.36	0.17±0.09	24.4±5.71	0.10±0.03	28.3±4.27
		BACADI	0.36±0.23	11.8±3.68	0.40±0.11	14.5±3.93	0.38±0.15	11.1±3.81	0.27±0.12	13.3±3.38
		SEA	0.93±0.06	2.2±1.25	0.84±0.07	5.5±2.33	0.84±0.16	3.8±2.32	0.63±0.15	**6.1±2.66**
	$p = 0.10$	DCDI-G	0.48±0.18	4.2±1.40	0.63±0.17	4.5±2.66	0.39±0.22	6.4±2.94	0.38±0.15	6.3±1.49
		DCDI-DSF	0.41±0.21	7.4±3.83	0.45±0.21	8.6±4.29	0.23±0.14	10.1±3.08	0.21±0.10	9.5±1.91
		OURS	0.94±0.06	1.8±1.17	**0.90±0.05**	4.3±1.79	**0.85±0.15**	3.3±2.41	**0.69±0.17**	6.2±2.44
	$p = 0.25$	DCDI-G	0.45±0.17	4.9±1.37	0.66±0.14	4.5±3.11	0.31±0.24	7.8±4.42	0.34±0.20	6.8±1.66
		DCDI-DSF	0.31±0.17	11.1±4.64	0.27±0.07	14.0±3.52	0.19±0.11	17.0±4.38	0.14±0.04	14.9±3.62
		OURS	**0.95±0.05**	**1.5±0.92**	0.86±0.05	4.9±1.81	0.82±0.15	4.1±2.26	0.67±0.16	6.3±2.69

5 Results

5.1 Real experiments

Table 1 illustrates our results on the Sachs proteomics dataset with varying graph priors. Incorporating synthetic KGs with 10% and 25% noise consistently improves performance in predicting the ground truth graph for OURS on mAP, but the effect is less evident on SHD. When provided with the CORUM KG, which contains a much higher noise level (34%), the advantage of leveraging prior knowledge diminishes. In terms of identifying noisy edges, OURS maintains robust performance across all prior knowledge settings, while the baselines are less consistent.

5.2 Synthetic experiments

Given the limited availability and challenges of evaluating real-world datasets – particularly in biological contexts where no standard evaluation exists – we perform a comprehensive comparison across synthetic datasets under various settings to assess the performance of each model (Table 2, Figure 4). We make the following observations.

Graph priors are useful in low-data scenarios, even at higher levels of noise. Incorporating graph priors improves performance for both our model and the DCDI baselines. For DCDI-G and DCDI-DSF, the performance with a prior noisy graph containing 25% noise is significantly better than that of the vanilla setting, particularly in the down-sampling scenario with only 50 samples per interventional regime. Our model is more sensitive to noise (perhaps because the SEA baseline starts at a high level), but a graph prior with 10% is consistently helpful across both data settings.

Graph priors are particularly helpful on hard settings. On easy settings (e.g. linear), the SEA baseline already achieves near-perfect mAP and SHD scores, as the assumptions of SEA's summary statistics (inverse covariance) match the linear Gaussian setting exactly. Correspondingly, our approach's improvement is less pronounced on linear Gaussian. However, there is slight mismatch
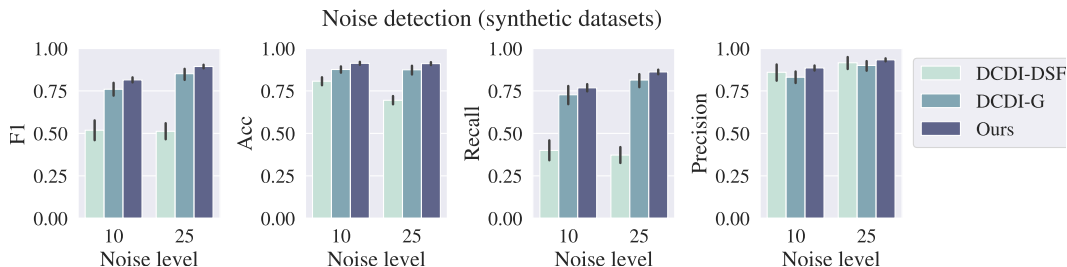
Figure 4: Noise detection on synthetic datasets. OURS outperforms DCDI at identifying errors in the graph prior, for both noise levels.
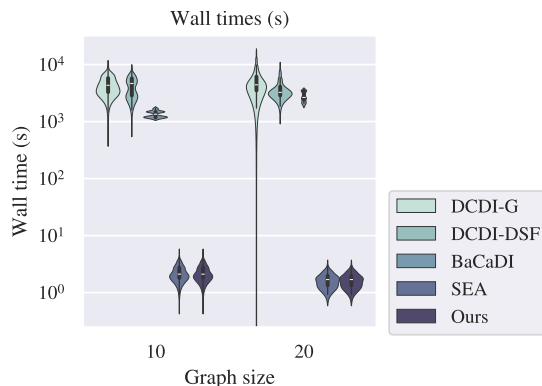

Figure 5: Runtime analysis of all algorithms. Amortized inference approaches (SEA, OURS) are orders of magnitude faster.

between SEA's assumptions and the remaining three settings; and as expected, providing a graph prior on those cases leads to larger improvements.

Additional observations   OURS outperforms DCDI at noise detection across noise levels (Figure 4). As an amortized inference method, OURS also achieves runtimes that are orders of magnitude faster than DCDI (Figure 5). The inclusion of a graph prior does not negatively impact runtime, compared to SEA.

6   Conclusion

In this work, we have presented an amortized inference algorithm for refining prior knowledge into data-dependent graphs. We demonstrated in synthetic and real settings that incorporating prior knowledge is particularly helpful in low-data settings, and that our approach is able to detect errors in these priors with high accuracy. However, we also observed that biological knowledge graphs contain high levels of noise in their connectivity alone, so it could be valuable to incorporate semantic information regarding the graphs and/or data for future work.

Acknowledgements

References

Réka Albert. Network inference, analysis, and modeling in systems biology. *The Plant Cell*, 19(11): 3327–3338, 2007.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather L. Butler, J. Michael Cherry, Allan Peter Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna E. Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, and Gerald M. Rubin. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

Pau Badia-i Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, 24(11):739–754, 2023.

Ding Bai, Caleb N Ellington, Shentong Mo, Le Song, and Eric P Xing. Attentionpert: accurately modeling multiplexed genetic perturbations with multi-scale effects. *Bioinformatics*, 40:i453–i461, 06 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae244.

Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific american*, 288(5):50–9, 2003.

Anastasiya Belyaeva, Chandler Squires, and Caroline Uhler. DCI: learning causal differences between gene regulatory networks. *Bioinformatics*, 37(18):3067–3069, 03 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab167.

Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1): 1017, 2019.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data, 2020.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014. doi: 10.1214/14-AOS1260.

Ting Chen, Hongyu L He, and George M Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, pp. 29–40, 1999.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, November 2002.

Haoyue Dai, Ignavier Ng, Gongxu Luo, Peter Spirtes, Petar Stojanov, and Kun Zhang. Gene regulatory network inference in the presence of dropouts: a causal view. *ArXiv*, abs/2403.15500, 2024.

Ke Feng, H. Jiang, Chaoyi Yin, and Huiyan Sun. Gene regulatory network inference based on causal discovery integrating with graph neural network. *Quant. Biol.*, 11:434–450, 2023.

Guadalupe Gonzalez, Isuru Herath, Kirill Veselkov, Michael Bronstein, and Marinka Zitnik. Combinatorial prediction of therapeutic perturbations using causally-inspired neural networks. *bioRxiv*, 2024. doi: 10.1101/2024.01.03.573985.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1): 2409–2464, 2012.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers, 2020.

Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. *bioRxiv*, 2023. doi: 10.1101/2023.12.12.571389.

Vân Anh Huynh-Thu and Guido Sanguinetti. Gene regulatory network inference: An introductory survey. *Methods in molecular biology*, 1883:1–23, 2018.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions, 2023.

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, 2022.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning, 2020.

Tineke L Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R Larson. Transcription dynamics in living cells. *Annual review of biophysics*, 45(1):25–47, 2016.

Romain Lopez, Jan-Christian Hütter, Jonathan K. Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems*, 2022.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning, 2022.

Linyong Mao and Haluk Resat. Probabilistic representation of gene regulatory networks. *Bioinformatics*, 20(14):2258–2269, 2004.

Daniel Marbach, James C. Costello, Robert Küffner, N. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R Allison, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9:796 – 804, 2012.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678, 2024.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Steven A McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide characterization of genetic perturbations. *bioRxiv*, 2024. doi: 10.1101/2024.07.03.601903.

J. M. Replogle, R. A. Saunders, A. N. Pogson, J. A. Hussmann, A. Lenail, A. Guna, L. Mascibroda, E. J. Wagner, K. Adelman, G. Lithwick-Yanai, N. Iremadze, F. Oberstrass, D. Lipson, J. L. Bonnar, M. Jost, T. M. Norman, and J. S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575, Jul 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 2023.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809.

Monica Schenone, Vlado Dancík, Bridget K. Wagner, and Paul A. Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology*, 9 4: 232–40, 2013.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvarinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pp. 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2001. doi: https://doi.org/10.7551/mitpress/1754.001.0001.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 51(D1):D539–D545, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1015.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Menghua Wu, Yujia Bao, Regina Barzilay, and Tommi Jaakkola. Sample, estimate, aggregate: A recipe for causal discovery foundation models. *arXiv 2402.01929*, 2024.

Haitao Zhao and Zhong-Hui Duan. Cancer genetic network inference using gaussian graphical models. *Bioinformatics and biology insights*, 13:1177932219839402, 2019.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018.