On Incorporating Prior Knowledge Extracted from Pre-trained Language Models into Causal Discovery

Anonymous ACL submission

Abstract

Pre-trained Language Models (PLMs) can reason about causality leveraging vast pre-trained knowledge and text descriptions of datasets, proving its effectiveness even when data is scarce. However, there are crucial limitations in the current PLM-based causal reasoning methods: i) PLM cannot utilize large datasets in prompt due to the limits of context length, and ii) the methods are not adept at comprehending the whole interconnected causal structures. On the other hand, data-driven causal discovery can discover the causal structure as a whole, although it works well only when the number of data observations is large enough. To overcome each other's limitations, we propose a new framework that integrates PLMs-based causal reasoning into data-driven causal discovery, which results in more improved and robust performance. Furthermore, our framework extends to the time-series data and exhibited superior performance.

1 Introduction

Causal discovery (Spirtes et al., 2000; Glymour et al., 2019) attempts to figure out the causal relations among the variables in a dataset, playing a core role in science and various applications (De La Fuente et al., 2004; Addo et al., 2021). Unfortunately, data are often scarce in real-world, thus causal discovery algorithms cannot accurately recover underlying causal structures. One approach to handle such data scarcity issue is using prior domain knowledge (Borboudakis et al., 2011; Kalainathan et al., 2018), e.g., by using an appropriate prior graph, the causal discovery algorithms can be guided by the prior when determining the direction of edges (Borboudakis and Tsamardinos, 2012; Sinha et al., 2021).

Recent breakthroughs in PLMs have demonstrated their potential for diverse reasoning tasks (Wei et al., 2022; OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023). Given the broad spectrum of text corpora utilized during pre-training, PLMs can address diverse tasks by employing specifically crafted task descriptions, including commonsense and numerical reasoning (Suzgun et al., 2022), code generation (Chen et al., 2021), and dialogue generation (Thoppilan et al., 2022).

Which of the following causal relationship is correct?
A. Changing $\{\alpha\}$ can directly change $\{\beta\}$.
B. Changing $\{\beta\}$ can directly change $\{\alpha\}$.
C. Both A and B are true.
D. None of the above. No direct relationship exists.
Let's think step-by-step to make sure that we have the right answer. Then provide your final answer within the tags, ⟨Answer⟩ A/B/C/D ⟨/Answer⟩

Figure 1: A multiple-choice template used in Kıcıman et al. (2023), to determine a pairwise causal relation.

Kıcıman et al. (2023) initiated *reasoning-based causal discovery*, harnessing such reasoning capability of PLMs. In particular, the authors designed a prompt template (Fig. 1), which queries whether one entity causes another entity, where the entities correspond to the column names of a tabular dataset. By recovering a causal structure via querying a causal relationship for every pair of variables, their method outperformed conventional causal discovery algorithms on benchmark datasets. This work showed the potential of utilizing the pre-trained knowledge of PLMs, at the same time, bypassing the issue of data scarcity.

However, PLM-based causal reasoning methods have inherent limitations compared to data-driven causal discovery. First, they cannot properly utilize large tabular data. Despite attempts to make use of tabular data with text, text-table multimodal models are limited to handling only small-scale tabular data (Wang et al., 2022; Dong et al., 2022; Liu et al., 2023; Lei et al., 2023; Li et al., 2023). Second, they mostly predict pairwise causal relations individually and cannot properly comprehend entire, interconnected causal structures.

Given that both PLM-based causal reasoning and data-driven causal discovery algorithms have their own strengths and weaknesses, we propose a novel framework that integrates the two approaches. In particular, we can harness the pre-trained knowledge of PLM to address data scarcity and utilize the patterns extracted from a dataset through causal discovery to gain a better understanding of the overall causal structure.

Moreover, we extended the application of PLM-based reasoning to our framework for addressing time-series datasets, which have numerous practical applications across various fields (Ding et al., 2006; Runge et al., 2019; Peters et al., 2013) but have not yet been addressed. We revealed that time-series causal discovery relying solely on PLMs is largely influenced by prompt design artifacts. We combined (i) the strengths of data-driven causal discovery, which is suitable for large datasets and capable of understanding the entire causal structure, with (ii) the effectiveness of PLM-based causal reasoning, which works well with small data, thereby outperforming both approaches.

Contributions We summarize our contributions. First, we demonstrate that PLM-based pairwise causal reasoning methods are not suitable for holistically eliciting a causal structure. Second, we propose a framework that integrates PLM-based causal reasoning with data-driven causal discovery, which compensates for one's weakness with the other's strength. Third, the proposed framework enhances the performance of existing causal discovery algorithms from static datasets to time-series datasets.

2 Preliminaries

In this section, we explain causal discovery algorithms and PLM-based causal reasoning.

2.1 Causal Discovery Algorithms

Causal discovery algorithms figure out a latent causal graph from a numeric dataset and are adept at effectively utilizing large tabular datasets. To begin with, we introduce notations. Given d variables and a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ with n observations, a causal graph can be expressed as a structural coefficients matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ under a linear assumption where $\mathbf{W}_{i,j}$ represents how much variable j would directly change to the change of variable i linearly.

First, DAG-GNN (Yu et al., 2019), learns a structural coefficient matrix through continuous optimization to approximate the distribution of causal graph of a dataset. Equipped with an encoder-decoder architecture, DAG-GNN is formulated as a variational autoencoder (Kingma and Welling, 2013), employing an acyclicity constraint and evidence lower bound. Zheng et al. (2018) proposed NOTEARS to solve combinatorial optimization as a continuous optimization, utilizing a DAG constraint. NOTEARS minimizes the following training objective, $L(\mathbf{W}) := \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1$, where the first term, fitting loss, is the Frobenius norm which indicates how well $\mathbf{W}$ fits the data, and the second term, sparsity loss, encourages a smaller number of edges, controlled by hyperparameter λ. NOTEARS minimizes the objective while ensuring the acyclicity of the learned graph (the acyclicity constraint is not shown here).

Time-series causal discovery aims to uncover temporal causal relationships, determining how variables influence each other across different time lags. DYNOTEARS (Pamfil et al., 2020) extends NOTEARS for time-series data, modelling time-lagged causal relations with a structural coefficient matrix called intra-slice $\mathbf{W}$, which represents contemporaneous causal relations, and a matrix called inter-slice $\mathbf{A} \in \mathbb{R}^{(T \times d) \times d}$, which represents time-lagged causal relations, where T is the maximum time lag. On the other hand, Sun et al. (2021) devised NTS-NOTEARS, which constructs weighted matrices with 1-dimensional CNNs for both intra-slice and inter-slice connections. It does not rely on the linear assumption. For readability, we simply refer the concatenation of $\mathbf{W}$ and $\mathbf{A}$ as $\mathbf{W}$, if no confusion can arise.

2.2 PLM-based Causal Reasoning

PLM-based causal reasoning on a static dataset Kıcıman et al. (2023) developed a multiple-choice prompt template (Fig. 1) for extracting pairwise causal relations through PLMs. By inserting the names of the variables into the prompt's α and β, PLM is guided to reason the existence and direction of causal relation between α and β. This process is repeated for all pairwise combinations of variables to build the causal graph.

Expansion to time-series data We expand the application of PLM-based causal reasoning to time-series datasets by proposing a prompt template (Fig. 2), which generalizes the multiple-choice prompt template (Fig. 1, Kıcıman et al. 2023). The prompt template (Fig. 2) inquires both time-lagged

Which of the following causal relationship is correct? For specific time step t,
A. Change $\{\alpha\}$ of time step t can directly change $\{\beta\}$ of time step t+1.
B. Change $\{\beta\}$ of time step t can directly change $\{\alpha\}$ of time step t+1.
C. Both A and B are true.
D. None of the above.

Let's think step-by-step to make sure that we have the right answer. Then provide your final answer within the tags, ⟨Answer⟩ A/B/C/D ⟨/Answer⟩

Which of the following causal relationship is correct? For specific time step,
A. Change $\{\alpha\}$ of time step can directly change $\{\beta\}$ of the same time step.
B. Change $\{\beta\}$ of time step can directly change $\{\alpha\}$ of the same time step.
C. Both A and B are true.
D. None of the above.

Let's think step-by-step to make sure that we have the right answer. Then provide your final answer within the tags, ⟨Answer⟩ A/B/C/D ⟨/Answer⟩

Figure 2: A multichoice template for the causal relation between two variables in time-series data. The upper prompt is for inter-slice, and the lower is for intra-slice.

	Method	NHD↓	NHD-R↓	SHD↓	#Edge	FDR↓	FPR↓	TPR↑
Arctic	GPT-4	0.23	0.42	19	32	0.28	0.09	0.47
	Revised	0.27	0.40	23	50	0.42	0.21	0.60
Sachs	GPT-4	0.14	0.45	18	21	0.47	0.09	0.57
	Revised	0.16	0.52	20	19	0.52	0.09	0.47

Table 1: Causal graph revision experiment using the Ban et al. (2023) revision prompt.

	Method	NHD↓	NHD-R↓	SHD↓	#Edge	FDR↓	FPR↓	TPR↑
Arctic Sea Ice	Pairwise	0.23	0.42	19	32	0.28	0.09	0.47
	LLM-complete	0.33	1.00	30	0	0.00	0.00	0.00
	LLM-cumulative	0.31	0.73	29	13	0.38	0.05	0.16
	LLM-ancestor	0.34	0.92	30	5	0.60	0.03	0.04
	GT-complete	0.33	1.00	30	0	0.00	0.00	0.00
	GT-cumulative	0.27	0.60	26	18	0.27	0.05	0.27
	GT-ancestor	0.31	0.81	28	6	0.17	0.01	0.10

Table 2: An ablation study to assess the effect of providing causal relations in prompts. Symbol ↓ indicates a lower-is-better metric. Full table is in Appendix G.

and contemporaneous causal relations for pairwise variables, $\{\alpha\}$ and $\{\beta\}$. We compared reasoning performance varying time units in the prompts (see Appendix B), and chose 'time step' as in Fig. 2 because there was no consistently meaningful difference across time unit.

Utilizing the prompt templates for static and time-series datasets, we can aggregate pairwise causal relations to construct a causal graph. The causal graph obtained by PLM is represented as a binary adjacency matrix, $\mathbf{K} \in \mathbb{R}^{d\times d}$, where $\mathbf{K}_{i,j}$ is 1 if i directly causes j and 0, otherwise. Since we do not enforce acyclicity, $\mathbf{K}$ might contain cycles. For time-series data, we similarly construct $\mathbf{K}$ through concatenating two adjacency matrices: one for intra-slice and the other for inter-slice.

3 Why Do We Need Causal Discovery for PLM-Based Causal Reasoning

We explain ablation studies about prompt templates to assess whether PLM can recognize the entire causal structure and to what extent PLM is affected by prompt artifacts when applied to time-series datasets. We first examine whether PLM recognizes the entire causal structure when determining pairwise causal relations, as data-driven causal discovery does by optimizing $\mathbf{W}$. Then, we explore how PLM's causal reasoning is affected by causally irrelevant artifacts of prompts, especially when applied to time-series datasets.

Issue: limited capability to comprehend holistic causal structure To examine PLM's ability in comprehending a causal structure, we borrow the idea of Ban et al. (2023), who employed two phases of causal reasoning. First, in the reasoning phase, PLM predicts causal relations for pairwise variables, and then, in the following revision phase, PLM revises the whole causal relations via a revision prompt,

> Based on your explanation, check whether the following causal statements are correct, and give the reasons.
>
> $\{\alpha\}_1 \rightarrow \{\beta\}_1, \ldots, \{\alpha\}_i \rightarrow \{\alpha\}_i$

where the entire causal relations predicted in the reasoning phase are provided to be revised.

We investigated the effect of the revision prompt in static dataset (Arctic Sea Ice, Huang et al. 2021, on Earth science) with 10 repetitions and analyzed revised predictions. As depicted in Table 1, we can observe only a marginal effect of revision by prompt engineering.[1]

For in-depth investigations of structure-aware reasoning, we examined the effect of the quantity and quality of information provided. To verify the effect of the amount of information, we experimented with *cumulative prompting*, bridging Kıcıman et al. (2023) and Ban et al. (2023), which focuses only on pairwise causal relations (Kıcıman et al., 2023) at first and converges to the revision methodology (Ban et al., 2023) as the predicted causal relations accumulate. The result in Table 2 shows that PLM only repeats its predictions rather than revising the predictions considering the accumulated causal structure. Providing all pairwise

[1] SHD is the hamming distance between the estimated and true causal graphs (i.e., the number of wrongly predicted edges). NHD normalizes SHD by the size of adjacency matrix, and NHD ratio further normalizes NHD by baseline NHD (the worst case NHD with the same number of predicted edges). Considering correctly predicted edges as true positives, FDR, FPR, and TPR are computed. See Appendix C.1 for details.

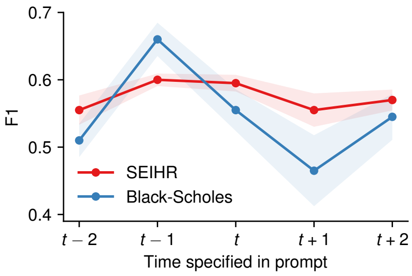

Figure 3: F1 of GPT-4 prediction (averaged over 10 repetitions) on time-lagged causal relations in two datasets. Shades represent 95% confidence interval.

relations at once in the prompt, referred to as *complete prompting*, also decreased the performance. The full prompts of *cumulative prompting* and *complete prompting* are in Appendix A.

We investigated that the low performance of the previous result was due to the low quality of its predictions in the reasoning phase. However, we still observed lower performance despite improved the quality of information in revision prompts. For this, we substituted PLM-predicted relations with the ground truth causal relations (rows starting with GT (Ground Truth) in Table 2). Similarly, there was no notable change in performance despite providing the true causal relations.

Given helpful information for the prediction, such as actual ground truth or causal ancestors, PLM's causal reasoning is expected to demonstrate better performance than that of vanilla pairwise reasoning. However, the experimental results indicate that whether predicted by PLM or known as ground truth, providing information on causal structure did not fulfill better performance than vanilla pairwise reasoning. This indicates that structure-aware PLM-based reasoning is not easily achievable via prompt engineering only.

Issue: prompt artifacts' influence in time-series While extending PLM's causal reasoning to time series, we discovered that the performance varies by prompt artifacts. To illustrate this, we experimented changing the way a one-step time-lagged causal relation in the prompt to explore the extent to which PLMs are influenced by the word choice in prompts rather than semantic meaning representing causality. We selected temporal domains where the maximum time lag is 1 and set the temporal causal relation unchanged, even if the start point is changed under the assumption that the causal structure remains unchanged over time. For example, querying whether α_{t-1} is the cause of β_t and α_t is the cause of β_{t+1} should give the same result. The experiment in Fig. 3 demonstrates that GPT-4's causal reasoning performance fluctuates based on specific numbers of time steps, even when all of them represent a one-step time lag.

These two experiments (i.e., on the lack of capability to comprehend causal structures and on being affected by prompt artifacts) suggest that PLM-based causal reasoning does not adhere strictly to the domain knowledge and that prompt engineering alone is insufficient to overcome these limitations. Therefore, it is desired to formally integrate PLM-based causal reasoning into time-series causal discovery algorithms, as we will do in the next section.

4 Causal Discovery with PLM-derived Priors

We now propose a causal discovery framework which incorporates PLM causal reasoning into an optimization-based causal discovery algorithm, by utilizing a prior knowledge $\mathbf{K}$ extracted from PLM. The overall framework is depicted in Fig. 4. Given static or time-series datasets as input, our framework performs PLM-based reasoning through specifically designed prompts (Figs. 1 and 2). Then, by aggregating pairwise causal relations, we acquire a prior $\mathbf{K}$. The causal discovery algorithm's optimization process then makes use of the prior $\mathbf{K}$ in three ways (not exclusively): The prior can be used as a starting point (Sec. 4.1); A regularization term is added to guide the learned structure reflects the given prior (Sec. 4.2) and; Boundaries are set for the structural coefficients based on the prior (Sec. 4.3). After the algorithm returns an estimated structural coefficient matrix, a threshold is applied to transform the structural coefficient matrix into a binary adjacency matrix (i.e., a directed graph).

4.1 Graph Initialization via Prior Knowledge

We suggest using $\mathbf{K}$ as an initial point for updating the edges. Typically, $\mathbf{W}$ is initialized as zero adjacency matrices (Zheng et al., 2018) without any prior. However, naively initializing the structural coefficient matrix can be sub-optimal by getting caught in local optima. Therefore, we devised initializing $\mathbf{W} = \lambda_{\text{init}}\mathbf{K}$ expecting that $\mathbf{K}$ of appropriate quality would help $\mathbf{W}$ avoid getting caught in local optima, where the scaling factor λ_{init} is introduced for adjustment of $\mathbf{K}$.

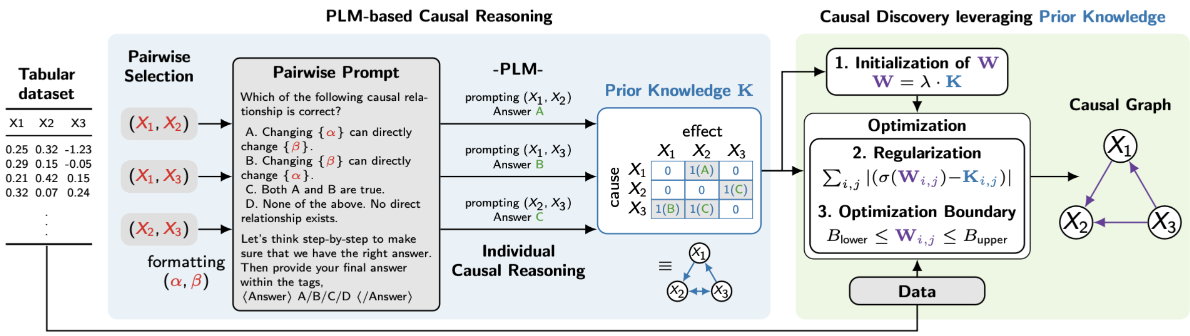

Figure 4: Overview of our framework. Given dataset, PLM-based causal reasoning returns an adjacency matrix as prior. Utilizing the prior, a causal discovery algorithm takes the dataset and returns a structural coefficient matrix, which is then mapped to a binary adjacency matrix.

4.2 Regularization with Prior Knowledge

We introduce a regularization term in the learning objective so that $\mathbf{W}$ reflects $\mathbf{K}$ throughout the optimization process, where the term is defined as

$$L_{\text{sim}}(\mathbf{W}) := \textstyle\sum_{i,j} |(\sigma(\mathbf{W}_{i,j}) - \mathbf{K}_{i,j})|,$$

which can be viewed as ℓ_1-regularization between $\mathbf{K}$ and the transformed, intermediate adjacency matrix $\mathbf{W}$. When regularizing $\mathbf{W}_{i,j}$ with binary $\mathbf{K}_{i,j}$ we applied a clamping function σ, which maps $\mathbf{W}_{i,j}$ between $[0, 1]$, to prevent large gradient flow from the regularization loss into $\mathbf{W}_{i,j}$. Then, our goal is to find an optimal matrix $\mathbf{W}_*$ which satisfies

$$\mathbf{W}_* = \arg\min_{\mathbf{W}} L(\mathbf{W}) + \lambda_{\text{sim}} L_{\text{sim}}(\mathbf{W}),$$

where λ_{sim} is the hyperparameter for scaling the regularization loss.

4.3 Setting Boundaries for Optimization

We now consider applying prior knowledge in setting each structural coefficient's boundary B as $B_{\text{lower}} \leq \mathbf{W}_{i,j} \leq B_{\text{upper}}$, where $B_{\text{lower}}, B_{\text{upper}} \in \mathbb{R}$, to be utilized during the optimization process. Sun et al. (2021) set B_{lower} larger than or equal to the threshold if edge (i, j) exists in the prior, and set $B_{\text{lower}} = B_{\text{upper}} = 0$ for $\mathbf{W}_{i,j}$ if prior knowledge indicates the absence of edge (i, j).

In our setting, the prior knowledge $\mathbf{K}$ is imperfect—we need to mitigate the risk of hallucination in prior knowledge. Therefore, we set a lower bound larger than 0 but smaller than the threshold for $\mathbf{W}_{i,j}$ if the corresponding edge presents in the prior, i.e., $\mathbf{K}_{i,j} = 1$. If there is no edge in the prior, i.e., $\mathbf{K}_{i,j} = 0$, we only set $B_{\text{lower}} = 0$. This modification prevents data-driven causal discovery from just following the prediction of $\mathbf{K}$ because the algorithm can now learn a structural coefficient $\mathbf{W}_{i,j}$ whose absolute value is smaller than the threshold. We implemented such boundary conditions for algorithms that employ L-BFGS (Byrd et al., 1995) (e.g., NOTEARS, and DYNOTEARS), replacing L-BFGS with L-BFGS-B (Zhu et al., 1997). Note that, when applying boundaries, we directly optimized the elements of the structural coefficient matrix $\mathbf{W}_{i,j}$ within $[B_{\text{lower}}, B_{\text{upper}}]$, without clamping, to ensure $\mathbf{W}_{i,j}$ within boundaries.

5 Experiments

We evaluate our proposed framework across the static and time-series datasets with various metrics.

5.1 Experimental Setup

We primarily investigated GPT-4 as our choice of PLM (OpenAI, 2023) since it outperformed other evaluated PLMs (detailed in Appendix D). We controlled the stochasticity of PLM by setting its temperature to 0.7 based on our experimental results over various temperature values. The prior knowledge $\mathbf{K}$ was determined based on the majority vote from 10 results (see Appendix C.2).

We employed NOTEARS, DAG-GNN, and CGNN for static datasets. For time-series, we employed a linear model, DYNOTEARS, and a nonlinear model, NTS-NOTEARS. The regularization method was not applied to NTS-NOTEARS since it is not straightforward to apply regularization over its architecture, i.e., convolution layers. The details on hyperparameters are in Appendix C.2, and the background and results of CGNN are illustrated in Appendices F and G.1.

	Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Arctic Sea Ice	GPT-4	0.23	0.42	19	32	0.28	0.09	0.47
	NOTEARS	0.31	0.63	26	23	0.43	0.10	0.27
	w/ random prior	0.44 (▲0.13)	0.60 (▼0.03)	37 (▲11)	56	0.63 (▲0.20)	0.37 (▲0.27)	0.43 (▲0.16)
	w/ GPT-4 prior	**0.22** (▼0.09)	**0.40** (▼0.23)	18 (▼8)	33	**0.27** (▼0.16)	0.09 (▼0.01)	**0.50** (▲0.23)
	DAG-GNN	0.31	0.76	27	12	0.41	**0.05**	0.14
	w/ random prior	0.41 (▲0.10)	0.64 (▼0.12)	37 (▲10)	44	0.62 (▲0.21)	0.29 (▲0.24)	0.33 (▲0.19)
	w/ GPT-4 prior	**0.22** (▼0.09)	**0.40** (▼0.36)	**17** (▼10)	33	**0.27** (▼0.14)	0.09 (▲0.04)	**0.50** (▲0.36)
	CMA	0.25	0.46	-	36	0.33*	0.13*	**0.50***
Sachs	GPT-4	0.14	0.45	18	21	0.47	0.09	0.57
	NOTEARS	0.22	0.65	22	22	0.68	0.14	0.36
	w/ random prior	0.27 (▲0.05)	0.82 (▲0.17)	28 (▲6)	21	0.83 (▲0.15)	0.17 (▲0.03)	0.18 (▼0.18)
	w/ GPT-4 prior	**0.10** (▼0.12)	0.41 (▼0.24)	**13** (▼9)	12	**0.25** (▼0.43)	**0.02** (▼0.12)	0.47 (▲0.11)
	DAG-GNN	0.18	0.68	19	13	0.61	0.07	0.26
	w/ random prior	0.27 (▲0.09)	0.81 (▲0.13)	29 (▲10)	21	0.83 (▲0.22)	0.17 (▲0.10)	0.20 (▼0.06)
	w/ GPT-4 prior	0.12 (▼0.06)	**0.36** (▼0.32)	15 (▼4)	22	0.40 (▼0.21)	0.08 (▲0.01)	**0.68** (▲0.42)

Table 3: Performances of NOTEARS and DAG-GNN on Arctic Sea Ice and Sachs datasets are indicated by red (improved) and blue (declined) arrows against the vanilla algorithm. For each algorithm, with and without GPT-4 prior, and random prior whose number of edges is the same with GPT-4 prior are investigated. $*$ indicates metrics that can be calculated via the true positive, precision, and recall reported in the CMA paper (Abdulaal et al., 2024).

5.2 Causal Discovery in Static Dataset

We report experimental results on two real-world datasets, Arctic Sea Ice (Huang et al., 2021) and Sachs (Sachs et al., 2005). Additional experiments on a physical-commonsense-based synthetic dataset are reported in Appendices E.1.3 and G.2.

Arctic Sea Ice Arctic Sea Ice dataset comprises 12 Earth science-related variables and only 486 observations, which is relatively small. Its causal graph, constructed by a meta-analysis of literature referred to in (Huang et al., 2021), contains 48 edges and includes cycles. This dataset presents two challenges for conventional causal discovery algorithms due to 1) a small sample size and 2) possible discrepancies between the causal relationships in the underlying data and the ground truth. We present the performance in Table 3.

GPT-4 shows better performance than data-driven causal discovery algorithms across metrics with big margins, in contrast, NOTEARS and DAG-GNN record NHD near 0.33, which is equivalent to NHD of an empty graph. The higher performance of PLM-based causal reasoning than data-driven causal discovery algorithms can be explained with the pre-train knowledge of the metadata. As PLM-based causal reasoning leverages the names of variables and related prior knowledge obtained in pre-training, it is not affected by the size of the dataset. Because the evaluation graph of Arctic Sea Ice dataset is constructed based on a meta-analysis of the literature, GPT-4 could have lots of chances to learn related prior knowledge.

Our proposed framework induces overall performance improvement with a big margin compared to causal discovery algorithms and even better or the same than GPT-4 across all metrics. Our framework also outperformed a recent work, Causal Modelling Agents (CMA) (Abdulaal et al., 2024), which likewise combines PLM and causal discovery, across all metrics except for TPR. Interestingly, when prior knowledge is incorporated, FDR decreases with little expense of FPR. This improvement is attributable to a well-constructed graph by PLM, and the revision by data-driven causal discovery with the support of data.

To better understand the effect of integration, we visualized the structural coefficients matrices as heatmaps (Figs. 5a to 5c). White circles denote false positives, and blue circles denote false negatives. The darker shades indicate the higher structural coefficients for the edges. In Fig. 5, our framework with DAG-GNN (Fig. 5c) resolves false positives and negatives by learning from the data compared to GPT-4 (Fig. 5a). We also observed that our model created edges where necessary, unlike the vanilla algorithm (Fig. 5b). Other heatmaps are in Appendix G. The effect of varying threshold values is depicted in Fig. 5d. We observed FDR and FPR of vanilla DAG-GNN and our framework with DAG-GNN, as increasing the threshold.

In addition, to investigate the effect of the quality

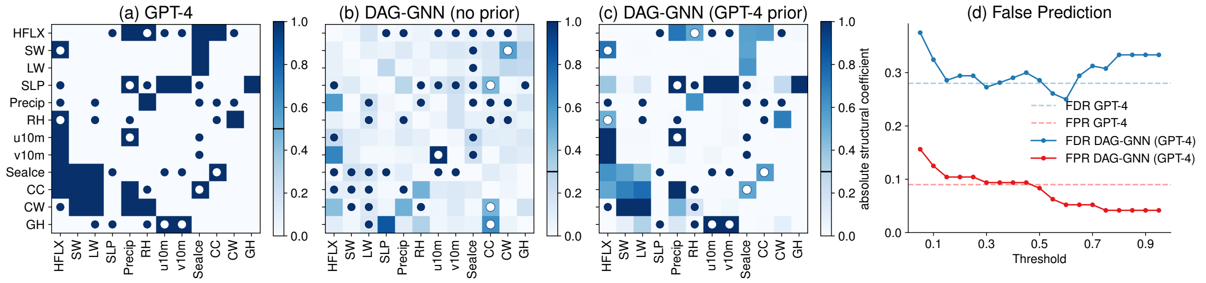

Figure 5: Predictions in Arctic Sea Ice of GPT-4, DAG-GNN, and DAG-GNN with GPT-4 prior. Dark circles are false negatives and white circles are false positives. A threshold is annotated as a line in a colorbar. (d) False prediction of DAG-GNN.

of a prior, we conducted an ablation analysis of our framework with a randomly sampled DAG of 43 edges as a prior where 43 is the number of edges predicted by GPT-4. Based on 20 repeated trials, the experimental results show that the performance improvement is not achieved by inadequate priors.

Sachs Sachs dataset (Sachs et al., 2005) is about protein signaling pathways and comprises 11 variables with 7,466 observations. Its causal graph consists of 19 edges and is acyclic (Ramsey and Andrews, 2018). Sachs dataset, in contrast to Arctic Sea Ice, is a wealth of data and exhibits strong alignment with the causal graph. For PLM prompting, we used full names instead of the abbreviations in the original data. We report experimental results in Table 3 where causal discovery algorithms exhibited different performance trends.

For DAG-GNN, we observed overall improvements except for the FPR. The reason vanilla DAG-GNN recorded a lower FPR without the PLM prior is that it predicted causal relations at roughly half the number of our framework. On the other hand, by increasing edge accurate predictions, our model improved performance except for FPR. Moreover, DAG-GNN with prior outperformed GPT-4 across all metrics. For NOTEARS, it gets even more consistent benefits than DAG-GNN, indicating that applying our framework improves performance across all metrics over vanilla NOTEARS. When compared to GPT-4, NOTEARS with prior outperform all metrics except for TPR, especially by big margins for FDR and FPR.

These results highlight the effectiveness of our framework. The overall improvement in FDR and FPR in every algorithm, compared to TPR, resulted in an overall increase in performance, as evidenced by NHD and SHD.

5.3 Causal Discovery in Time-series Datasets

For time-series, we simulated synthetic datasets regarding the well-known Partial Differential Equations (PDEs) with a maximum time lag of 1 where we adopted Black-Scholes (MacBeth and Merville, 1979) model in the finance domain and SEIHR (Niu et al., 2020) model in the epidemic domain. The reason why we used those synthetic datasets is that the conventional time-series dataset for causal discovery lacks the actual relationships among the variables for utilizing the pre-train knowledge of PLMs. The synthetic datasets via PDEs can offer rich, real-world semantic meanings annotated by domain experts, which provide PLMs enriched opportunities to learn necessary prior knowledge for causal reasoning at the same time while providing scalability in dataset size. Further detailed reasons for this selection are explained in Appendix E.2.

The overall process for creating these datasets involves 1) selecting a mathematical model with trustworthy, universally acceptable names and relationships, 2) generating time-series synthetic datasets, and 3) utilizing the models.

Black-Scholes Black-Scholes model is a probabilistic method to predict future stock prices, determining the current value of options (MacBeth and Merville, 1979). The PDEs of the model represents dynamics of future stock price, which acts as dependent variable. Based on the PDEs, we annotated the evaluation graph with 3 nodes and 5 edges. For each time step, we sampled observations with added noise.

Firstly, the overall performance of our framework with NTS-NOTEARS demonstrated a marked improvement compared to the vanilla NTS-NOTEARS and GPT-4 as detailed in Table 4. The prediction by our framework inferred the presence, absence, and direction of edges more accurately,

	Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Black-Scholes	GPT-4	0.11	0.40	4	5	0.40	0.06	0.60
	NTS-NOTEARS	0.22	0.67	8	7	0.71	0.16	0.40
	w/ random prior	0.22	0.80(▲0.13)	8	5	0.80(▲0.09)	0.13(▼0.03)	0.20(▼0.20)
	w/ GPT-4 prior	**0.06**(▼0.16)	**0.20**(▼0.47)	**2**(▼6)	5	**0.20**(▼0.51)	**0.03**(▼0.13)	**0.80**(▲0.40)
	DYNOTEARS	0.22	0.67	8	7	0.71	0.16	0.40
	w/ random prior	0.22	1.00(▲0.33)	8	3	1.00(▲0.29)	0.10(▼0.06)	0.00(▼0.40)
	w/ GPT-4 prior	0.08(▼0.14)	0.33(▼0.34)	3(▼5)	4	0.25(▼0.46)	**0.03**(▼0.13)	0.60(▲0.20)
SEIHR	GPT-4	0.09	0.33	9	14	0.36	0.06	**0.69**
	NTS-NOTEARS	0.11	0.44	11	12	0.42	0.06	0.54
	w/ random prior	0.16(▲0.05)	0.67(▲0.23)	16(▲5)	11	0.64(▲0.22)	0.08(▲0.02)	0.31(▼0.23)
	w/ GPT-4 prior	**0.07**(▼0.04)	**0.30**(▼0.14)	**7**(▼4)	10	**0.20**(▼0.22)	**0.02**(▼0.04)	0.62(▲0.08)
	DYNOTEARS	0.12	0.67	12	5	0.40	**0.02**	0.23
	w/ random prior	0.14(▲0.02)	0.70(▲0.03)	14(▲2)	7	0.57(▲0.17)	0.05(▲0.03)	0.23
	w/ GPT-4 prior	0.08(▼0.04)	0.33(▼0.34)	8(▼4)	11	0.27(▼0.13)	0.03(▲0.01)	0.62(▲0.39)

Table 4: Performances of NTS-NOTEARS, DYNOTEARS on Black-Scholes and SEIHR datasets are indicated by red (improved) and blue (declined) arrows against the vanilla algorithm. For each algorithm, with and without GPT-4 prior, and random prior whose number of edges is the same with GPT-4 prior are investigated.

which was evident across all metrics. Compared to GPT-4, our framework with NTS-NOTEARS also outperformed GPT-4 in all the metrics.

Overall performance of our framework with DYNOTEARS also was improved across all metrics than the vanilla model and GPT-4. Our model outperformed the vanilla DYNOTEARS across all metrics. Compared to GPT-4, our model showed significant improvement in overall metrics, especially FPR and FDR. Although the number of predicted edges is decreased, the reduction in FPR and FDR led to more accurate predictions, thus lowering SHD and NHD. This trend was consistently observed across different algorithms and datasets.

SEIHR SEIHR model estimates the transmission rate of an infectious disease (Niu et al., 2020). The dynamics of SEIHR is modeled using PDEs with 5 nodes and 13 edges.

In SEIHR dataset, we also found a consistent improvement in performance with our framework compared to the vanilla algorithms and GPT-4. In the case of NTS-NOTEARS, the integration of priors contributed to an overall performance increase, as seen in Table 4. SHD decreased by 4, and the performance enhancement in FPR and FDR was particularly notable compared to other metrics. When compared to GPT-4, there was an improvement in all metrics except TPR.

In DYNOTEARS, the overall performance improvement was significant, even with an increase in the number of predicted edges, compared to the vanilla algorithm. While the increase of 6 edges led to a slight rise in FPR, TPR saw a substantial increase, leading to an overall improvement compared to the vanilla DYNOTEARS. In contrast, when compared to GPT-4, there was no corresponding overall performance enhancement. We conjecture that the discrepancy may arise from the nonlinearity of the data, violating the linearity assumption of DYNOTEARS. Nonetheless, the ability of this approach to increase the number of edges while simultaneously enhancing precision, as opposed to vanilla DYNOTEARS, highlights the potential of our framework.

6 Conclusion

We proposed a novel framework that incorporates the prior knowledge extracted from PLMs into score-based causal discovery algorithms for both static and time-series datasets. The integration is achieved through graph initialization, regularization, and setting boundaries of structural coefficients, all leveraging the prior. This approach combines the strengths of both worlds: reducing the potential for false predictions of PLMs by applying data-driven structural learning and enhancing causal discovery performance by incorporating prior knowledge extracted from PLMs. We also demonstrated that solely relying on prompt engineering might diminish performance even when information is introduced to aid causal reasoning. This highlights the importance of combining data-driven causal discovery algorithms with PLM-based causal reasoning. We expect that our framework will open up new avenues for research and exploration in causal discovery.

7 Limitations

This paper has a few limitations. First, our assumption for time-series causal discovery is based on the premise that the latent causal structure does not change; therefore, performance may vary in cases where the causal structure changes. Second, the number of variables in our dataset was not large enough. Especially for time-series causal discovery, where variable names need to exist and have realistic relationships, we could not experiment with datasets that have an arbitrarily large number of variables.

8 Ethics Statement

We outline our ethics statement of the work as follows. (1) Our framework, based on a causal discovery algorithm, has less potential risks. We revealed our hyperparameters and other experimental settings in Sec. 5.1 and appendix C, and our experiments are based on repeated experiments. Moreover, while hallucinations within PLMs can lead to erroneous decision-making, the integration of causal discovery algorithms significantly minimizes such negative effects. Thereby, we propose that our work is robust to potential risks. (2) The static data used in the experiments are all open-source datasets and the time series datasets are newly created numeric data based on PDEs by us. Arctic Sea Ice and Sachs datasets are licensed under the Creative Commons Attribution-Share Alike License. Furthermore, we ensured that there is no data capable of identifying individuals. (3) The physical synthetic dataset in Appendix E.1.3, was annotated by human annotators using PIQA data (Bisk et al., 2019) to create ground truth graphs. We recruited student annotators with payment above the country's legal minimum wage. We announced to the annotators that the curated dataset and the annotations would be used for research purposes.

References

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. 2024. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*.

Peter Martey Addo, Christelle Manibialoa, and Florent McIsaac. 2021. Exploring nonlinearity on the co2 emissions, economic production and energy use nexus: a causal discovery approach. *Energy Reports*, 7:6196–6204.

Rohan Anil et al. 2023. PaLM 2 technical report.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *ArXiv*, abs/2306.16902.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Giorgos Borboudakis, Sofia Triantafillou, Vincenzo Lagani, and I. Tsamardinos. 2011. A constraint-based approach to incorporate prior knowledge in causal models. In *The European Symposium on Artificial Neural Networks*.

Giorgos Borboudakis and I. Tsamardinos. 2012. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. In *International Conference on Machine Learning*.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208.

Mark Chen et al. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.

Mingzhou Ding, Yonghong Chen, and Steven L Bressler. 2006. Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, pages 437–460.

Haoyu Dong, Zhoujun Cheng, Xinyi He, Mengyuan Zhou, Anda Zhou, Fan Zhou, Ao Liu, Shi Han, and Dongmei Zhang. 2022. Table pre-training: A survey on model architectures, pretraining objectives, and downstream tasks. *ArXiv*, abs/2201.09745.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. 2018. Learning functional causal models with generative neural networks. *Explainable and interpretable models in computer vision and machine learning*, pages 39–80.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. 2021. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in Big Data*.

Diviyan Kalainathan and Olivier Goudet. 2020. Causal discovery toolbox: Uncover causal relationships in python. *URL https://arxiv. org/abs*.

Diviyan Kalainathan, Olivier Goudet, Isabelle M Guyon, David Lopez-Paz, and Michèle Sebag. 2018. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv: Machine Learning*.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality.

Fangyu Lei, Xiang Lorraine Li, Yifan Wei, Shizhu He, Yiming Huang, Jun Zhao, and Kang Liu. 2023. S3hqa: A three-stage approach for multi-hop text-table hybrid question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2023. Table-GPT: Table-tuned GPT for diverse table tasks. *ArXiv*, abs/2310.09263.

Chuang Liu, Junzhuo Li, and Deyi Xiong. 2023. Tab-cqa: A tabular conversational question answering dataset on financial reports. In *Annual Meeting of the Association for Computational Linguistics*.

James D MacBeth and Larry J Merville. 1979. An empirical examination of the black-scholes call option pricing model. *The journal of finance*, 34(5):1173–1186.

Ruiwu Niu, Eric WM Wong, Yin-Chi Chan, Michaël Antonie Van Wyk, and Guanrong Chen. 2020. Modeling the COVID-19 pandemic using an seihr model with human migration. *IEEE Access*, 8:195503–195514.

OpenAI. 2023. GPT-4 technical report.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. DYNOTEARS: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26.

Joseph Ramsey and Bryan Andrews. 2018. FASK with interventional knowledge recovers edges from the Sachs model. *arXiv preprint arXiv:1805.03108*.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996.

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

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. 2021. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge. *arXiv preprint arXiv:2109.04286*.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Romal Thoppilan et al. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Dingzirui Wang, Longxu Dou, and Wanxiang Che. 2022. A survey on table-and-text hybridqa: Concepts, methods, challenges and future directions. *ArXiv*, abs/2212.13465.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23:550–560.

A Prompt Templates for Revision

Here, we describe the full text of the *cumulative prompting* Fig. 6 and *complete prompting* (Fig. 7). In both types of prompts, information about a causal structure is specified within ⟨Found Causal Relation⟩ . . . ⟨/Found Causal Relation⟩. In cumulative prompting, PLM performs causal reasoning over entire pairwise variables just once, and the predicted causal relations are accumulated. On the other hand, in complete prompting, PLM first performs causal reasoning over entire pairwise variables to draft an intermediate causal structure. Then, PLM repeats the causal reasoning again over the entire pairwise variables given the intermediate causal structure between ⟨Found Causal Relation⟩ . . . ⟨/Found Causal Relation⟩.

Here are previously found causal relations.

⟨Found Causal Relation⟩
Changing $\{\alpha\}$ can directly change $\{\beta\}$.
Changing $\{\gamma\}$ can directly change $\{\alpha\}$.
Changing $\{\alpha\}$ and changing $\{\delta\}$ have no direct causal relation.
⟨/Found Causal Relation⟩

Not only considering provided causal relationships but also incorporating your reasoning about the following question,

Which of the following causal relationship is correct?

A. Changing $\{\alpha\}$ can directly change $\{\epsilon\}$.
B. Changing $\{\epsilon\}$ can directly change $\{\alpha\}$.
C. Both A and B are true.
D. None of the above. No direct relationship exists.

Let's think step-by-step to make sure that we have the right answer. Then provide your final answer within the tags, ⟨Answer⟩ A/B/C/D ⟨/Answer⟩

Figure 6: A modified prompts from (Kıcıman et al., 2023) named "cumulative prompt" that uses cumulatively found relations from the previous prompts or ground-truth relationships.

B Prompt Engineering for Time-series Datasets

This section explains the ablation studies conducted to design prompts for time-series datasets. We conducted the ablation study where specific time units such as hour, day, month, and year were given instead of referring to it as a 'time step', as shown in Table 5. This approach somewhat yielded performance improvements in certain instances, though the effectiveness varied across different datasets. In

Here are previously found causal relations.

⟨Found Causal Relation⟩
Changing $\{\alpha\}$ can directly change $\{\beta\}$.
Changing $\{\alpha\}$ and changing $\{\gamma\}$ have no direct causal relation.
...
(relation between $\{\alpha\}$ and $\{\epsilon\}$ is not provided)
...
Changing $\{\delta\}$ can directly change $\{\alpha\}$.
⟨/Found Causal Relation⟩

Not only considering provided causal relationships but also incorporating your reasoning about the following question,

Which of the following causal relationship is correct?

A. Changing $\{\alpha\}$ can directly change $\{\epsilon\}$.
B. Changing $\{\epsilon\}$ can directly change $\{\alpha\}$.
C. Both A and B are true.
D. None of the above. No direct relationship exists.

Let's think step-by-step to make sure that we have the right answer. Then provide your final answer within the tags, ⟨Answer⟩ A/B/C/D ⟨/Answer⟩

Figure 7: A modified prompt from (Kıcıman et al., 2023) named "complete prompt" that uses all causal relations (except the relation to be queried) found from previous reasoning attempt or ground-truth relationships.

detail, for SEIHR model, using "day" and "hour" as the time unit yielded effective results, while in the case of Black-Scholes model, characterizing the interval as a 'time step' was more effective. Although the specific training corpora of PLM (GPT-4) is unknown, we guess that there were likely many predictions about the day-to-day variation in patient numbers since SEIHR model is based on COVID-19. For Black-Scholes model, the term "time step t" is frequently used in economics, supporting this assumption.

C Experimental Details

In this section, we illustrated the definitions of the metrics and the experimental setup for reproducibility.

C.1 Metrics

We introduce metrics employed in the experiments. Structural Hamming Distance (SHD) is the sum of the number of missing edges (false negative), extra edges (false positive), and reversed edges (Tsamardinos et al., 2006). Normalized Hamming Distance (NHD) is a metric that normalizes Hamming distance by dividing the distance by its matrix size. This yields values between 0 and 1, with lower values indicating greater similarity to the causal graph. NHD ratio is an NHD divided by the baseline NHD, which is the worst case NHD for the same number of edges. With NHD ratio, we can figure out how much the estimated adjacency matrix is improved compared to the worst case.

False Discovery Rate (FDR), False Positive Rate (FPR), and True Positive Rate (TPR) are derived

	Time Unit Naming	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Black-Scholes	Time-step	**0.11**	**0.40**	**4**	5	**0.40**	**0.06**	**0.60**
	Hours	0.14	0.56	5	4	0.50	**0.06**	0.40
	Day	0.14	0.56	5	4	0.50	**0.06**	0.40
	Month	0.14	0.56	5	4	0.50	**0.06**	0.40
	Year	0.14	0.56	5	4	0.50	**0.06**	0.40
SEIHR	Time-step	0.09	0.33	9	14	0.36	0.06	0.69
	Hours	**0.07**	**0.26**	**7**	14	**0.29**	**0.05**	**0.77**
	Day	**0.07**	**0.26**	**7**	14	**0.29**	**0.05**	**0.77**
	Month	0.11	0.38	11	16	0.44	0.08	0.69
	Year	0.11	0.35	11	18	0.44	0.09	**0.77**

Table 5: Ablation study for time-series datasets varying time unit specified in prompt, all with GPT-4.

from the four outcomes of a confusion matrix: False Positive, False Negative, True Positive, and True Negative and these metrics collectively evaluate the errors in classification:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.$$

C.2 Setup

The experimental setup, such as hyperparameter and model architecture, is as follows: First, GPT-4 prior is selected from start time t and time lag 1 and determined by majority voting over 10 repetitions. To leverage PLMs as causal reasoning agents, we should consider their randomness, which is usually controlled by temperature or top-p values in nucleus sampling. However, we found that just picking a deterministic result by setting the temperature near zero does not give the best performance. To handle the randomness of PLMs in causal reasoning and, at the same time, choose the best among diverse reasoning results, we collected 10 independent causal reasoning results for each dataset with varying temperatures. Given the result in Table 6, we chose temperature 0.7.

Second, in the experiment by Ban et al. (2023), further refining the Ban 2023 revision prompt, we utilized a modified version to revise a GPT-4 prior graph that we got from pairwise prompting. After 10 repetitions, we obtained a revised graph through majority voting and measured the performance by comparing the resulting revised graph with both the ground truth graph and the GPT-4 prior graph.

Third, we detail the hyperparameter of NOTEARS and DYNOTEARS—t, λ_{sim}, thresholds in Table 7. λ_{init} is the scaling factor for graph initialization and λ_{sim} is that for prior similarity regularization. As we mentioned in the Experimental setup of Sec. 5, hyperparameters of baseline were tuned to reproduce baseline experiments, and that of our experiments were selected by fine-tuning. For NTS-NOTEARS and DYNOTEARS, we experimented with two boundary settings for L-BFGS-B optimization, which is specified in the parentheses. The specific boundary setting is as follows. (NTS-NOTEARS, BS) : (0.4, 3.0), (NTS-NOTEARS, SEIHR) : (1.05, 3.0), (DYNOTEARS, BS) : (0.5, 3.0), (DYNOTEARS, SEIHR) : (0.8, 3.0).

Fourth, the model architecture and other setups are as follows. For DAG-GNN, we used the Adam optimizer and two layers each for the encoder and the decoder. We allocated 64 hidden nodes in each layer for Arctic Sea Ice model and 128 hidden nodes in each layer for Sachs model, with a uniform batch size set at 100 for DAG-GNN. For CGNN, we employed an average of $\mathbf{K}$ instead of using vanilla $\mathbf{K}$ to prevent CGNN being captured in a local minimum originated from the discrete value of $\mathbf{W}$. Moreover, CGNN does not use prior regularization in contrast to NOTEARS and DAG-GNN. The reason is that CGNN does not use explicit modeling of the structural coefficient matrix, which is essential in prior regularization.

Though the experiments are feasible on CPUs, our experiments were primarily conducted using NVIDIA RTX A6000 and Tesla V100-SXM2-32GB GPUs. Without repetition, individual training of algorithms can be conducted within an hour. All the baseline algorithms including DAG-GNN, NOTEARS, NTS-NOTEARS, DYNOTEARS have trainable parameters fewer than 10k. The baseline code was referenced from (Kalainathan and Goudet, 2020; Yu et al., 2019), CausalNex[2].

[2] https://github.com/quantumblacklabs/causalnex

	Temp.	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Arctic Sea Ice	0.01	0.27	0.43	23	41	0.39	0.15	0.52
	0.05	0.28	0.48	24	36	0.39	0.15	0.46
	0.10	0.26	0.44	24	37	0.35	0.14	0.50
	0.50	0.24	0.44	19	30	0.27	0.08	0.46
	0.70	0.23	0.42	19	32	0.28	0.09	0.47
	1.00	0.29	0.57	26	26	0.38	0.10	0.33
Sachs	0.01	0.20	0.57	22	23	0.61	0.14	0.47
	0.05	0.18	0.52	22	23	0.57	0.13	0.53
	0.10	0.17	0.54	21	20	0.55	0.11	0.47
	0.50	0.17	0.51	20	22	0.55	0.12	0.53
	0.70	0.17	0.50	20	21	0.52	0.11	0.53
	1.00	0.18	0.58	21	19	0.58	0.11	0.42

Table 6: Performances of GPT-4 under varing temperatures on Arctic Sea Ice and Sachs dataset.

	Method	Prior	λ_{init}	λ_{sim}	Threshold
Arctic Sea Ice	NOTEARS	GPT-4	0.55	0.6	0.2
	CGNN	GPT-4	1	-	0.99
	DAG-GNN	GPT-4	0.5	0.9	0.3
	NOTEARS	None	-	-	0.1
	CGNN	None	-	-	-
	DAG-GNN	None	-	-	0.3
Sachs	NOTEARS	GPT-4	0.3	0.4	0.2
	CGNN	GPT-4	1	-	0.65
	DAG-GNN	GPT-4	0.5	0.7	0.3
	NOTEARS	None	-	-	0.09
	CGNN	None	-	-	-
	DAG-GNN	None	-	-	0.3

Table 7: Hyperparameters of Arctic Sea Ice and Sachs

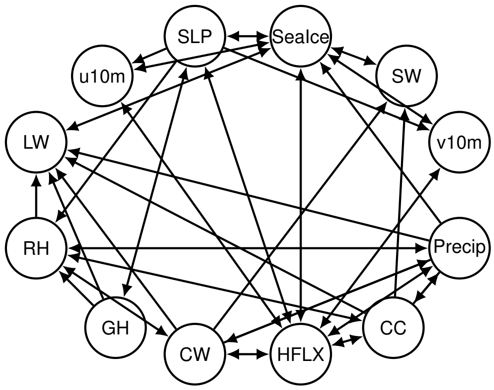

Figure 8: Arctic Sea Ice ground truth graph

D Comparison of Causal Reasoning Performance across PLMs

We chose GPT-4 as the baseline PLM for causal reasoning in our framework, based on comparative experiments conducted with recent PLMs on static datasets, detailed in Table 8. We tested GPT4, GPT4 turbo (OpenAI, 2023), PaLM 2 (Anil et al., 2023), Claude[3], and Gemini Pro (Team et al., 2023). GPT-4 and GPT-4 turbo recorded the best performance on both datasets, except for PaLM 2's exceptionally low FDR and FPR due to merely fewer edge predictions. Despite fluctuations in performance between GPT-4 and GPT-4 turbo, GPT-4 generally outperformed GPT-4 turbo on both dataset.

E Datasets

We explain the details of the datasets. For static datasets, we describe the characteristics of each dataset, the ground truth graphs, and the generation process of the physical commonsense-based static dataset. For time-series datasets, we illustrate the reason why we used the datasets based on PDEs instead of existing datasets, descriptions of the PDEs for each model, and the generation process based on PDEs.

E.1 Static Datasets

E.1.1 Arctic Sea Ice

Arctic Sea Ice dataset (Huang et al., 2021) consists of 12 Earth science-related variables and only 486 instances. Its causal graph (Fig. 8), constructed by a meta-analysis of literature referred to in (Huang et al., 2021), contains 48 edges without acyclicity. This dataset presents two challenges for conventional causal discovery algorithms due to 1) a small sample size and 2) possible discrepancies between the causal relationships in the underlying data and the ground truth because the causal graph of Arctic Sea Ice is annotated based on a literature review, without a comprehensive examination of alignment among the sources.

We infer that PLMs are not affected since each causal relation in the ground truth is based on published papers. Thus, PLMs could have learned

[3]https://www.anthropic.com/product

	Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Arctic Sea Ice	GPT4	**0.23**	0.42	**19**	32	0.28	0.09	0.47
	GPT4 turbo	0.26	**0.34**	26	62	0.41	0.27	0.75
	PaLM 2	0.27	0.71	22	8	**0.00**	**0.00**	0.16
	*Claude	0.40	0.41	38	92	0.55	0.53	**0.85**
	Gemini Pro	0.27	0.37	24	57	0.42	0.25	0.68
Sachs	GPT4	**0.14**	**0.45**	**18**	21	**0.47**	0.09	0.57
	GPT4 turbo	0.15	0.54	19	16	0.50	0.07	0.42
	PaLM 2	0.19	1.00	22	5	1.00	**0.04**	0.00
	*Claude	0.33	0.54	33	55	0.69	0.37	**0.89**
	Gemini Pro	0.35	0.64	35	48	0.75	0.35	0.63

Table 8: Performances of GPT-4, GPT-4 turbo,PaLM 2, Claude and Gemini Pro priors on the **Arctic Sea Ice and Sachs dataset**. Priors are determined by majority voting over 10 repetitions.(* only 1 time for Claude)

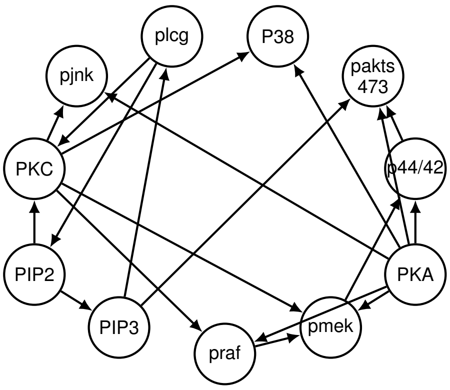

Figure 9: Sachs ground truth graph

related knowledge. This implies that the annotated causal graph could be misaligned with the ground truth in the data generation process in nature (e.g., cyclic). The two challenges mentioned previously contribute to the difficulties faced by traditional causal discovery algorithms in producing accurate predictions.

E.1.2 Sachs

Sachs dataset (Sachs et al., 2005) consists of protein signaling pathways and comprises 11 variables with 7,466 observations. Its associated causal graph (Fig. 9) has a DAG structure with 19 edges (Ramsey and Andrews, 2018). Sachs dataset, in contrast to Arctic Sea Ice dataset, is a wealth of data and exhibits strong alignment with the causal graph. We replaced abbreviations of Sachs' original names of the variables with their full names for making the prior graph by PLMs.

E.1.3 Physical Commonsense-Based Synthetic Dataset

In this section, we explain why we created a physical commonsense-based synthetic dataset and how to construct it for evaluating causal discovery algorithms and the causal reasoning ability of PLM.

Reason for constructing physical commonsense-based synthetic dataset To evaluate the reasoning ability of PLM, we chose to construct a knowledge base within a specific domain. Because causal reasoning focuses on logical relations between variables, the annotated content based on the selected domain should contain clear ground truth. For this reason, domains where consensus on the ground truth is challenging, such as social or cultural domains, are unsuitable, so we decided to construct knowledge based on physics.

We utilized PIQA (Bisk et al., 2019) which is the QA dataset of physical commonsense to select the proper physical event that has indisputable causal relationships. We removed text that is ambiguous or described too specifically from our knowledge base. We selectively annotated entities that describe phase transition which refers to phenomena where a matter's phase, such as solid, liquid, or gas, transit to another phase. For example, the increase in 'surface air temperature' causes a change in the evaporation rate of water, transferring the object from the liquid phase to the gas phase.

Using this strategy to annotate PIQA dataset, we gathered the nodes of a causal graph whose nodes are entities involved in the phase transition. Then, human annotators evaluated the causal relationships among the nodes, to construct causal graph in Fig. 10.

Generation process of physical commonsense-based synthetic dataset To generate a synthetic dataset based on a physical commonsense-based causal graph, we selected seven nodes that represent the evaporation of water such that collected nodes and edges satisfy the DAG constraint. Given the causal graph, we added subgraphs of five and

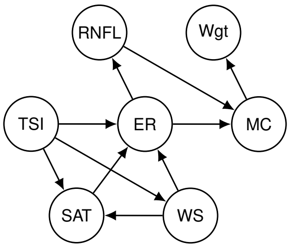

Figure 10: Physical knowledge-based synthetic graph with size 7. The components of the graph are Rainfall (RNFL), Total Solar Irradiance (TSI), Surface Air Temperature (SAT), Wind Speed (WS), Evaporation Rate (ER), Moisture Content of object (MC), and Weight of object (Wgt).

three nodes from the predefined graph by ensuring that causal relations were preserved even when nodes were removed. Removing nodes, we add additional edges from ancestor to descendant whenever the removed node connects the ancestor and descendant so that the chain relation holds. Using the constructed 3, 5, and 7 nodes graphs, we assumed a linear Structural Equation Model between variables and Gaussian noise of $\epsilon \sim \mathcal{N}(0, 0.5)$ within a given causal graph and generate 5000 data points.

E.2 Time-series Datasets

Numerous studies prefer synthetic datasets for time-series causal discovery due to scalability in dataset size and error-free evaluation. However, purely synthetic data lacks the semantic meaning found in written text, preventing using PLMs' causal reasoning. On the other hand, the synthetic datasets via PDE can offer real-world semantic meanings annotated by domain experts, which provide PLMs enriched opportunities to learn necessary prior knowledge for causal reasoning.

Our framework necessitates specific dataset conditions to effectively utilize Pre-trained Language Models (PLM). 1) The variable should be aligned with the consensus in the domain so that a solid ground truth holds; 2) The text descriptions based on the consensus are represented in various web-based sources so that PLMs can learn prior knowledge during the pre-training. However, several well-known datasets used in time-series causal discovery fail to fulfill both criteria, usually meeting only one of these conditions. Real datasets often come with meaningful variable names but lack a universally agreed-upon ground truth. For example, figuring out a consensus on the ground truth for the S&P 100 dataset is challenging hindering PLMs from learning the actual relationships.

E.2.1 Black-Scholes

Black-Scholes model is a probabilistic model of predicting future stock prices, determining the current value of options (MacBeth and Merville, 1979). This model accounts for various factors, including the price of the call options (C), the price of the put options (P), the current stock price (S), the strike price (K) of the option contract, the time remaining until the option's maturity (T), the prevailing risk-free interest rate (r), and the expected stock price volatility (σ). Normal distribution of d_1 and d_2 represent the sensitivity of the option price to changes in the price of the underlying asset and the probability when the underlying asset's price exceeds the strike price at maturity, i.e., the probability that a European call option will be exercised.

$$
\begin{aligned}
C &= SN(d_1) - Ke^{-r(T-t)}N(d_2) \qquad (1)\\
P &= Ke^{-r(T-t)}N(-d_2) - SN(-d_1)\\
d_1 &= \frac{\ln(\frac{S}{K}) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{(T-t)}}\\
d_2 &= \frac{\ln(\frac{S}{K}) + (r - \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{(T-t)}}.
\end{aligned}
$$

This equation estimates the expected option value of the stocks based on the stochastic path of the stock price (Eq. (1)).

We synthetically generated data for C, P, and S as the same as the equation assuming a hypothetical company's stock price as the basis for S. The assumption about S is grounded in the core principle of the model, that $\log S$ follows a normal distribution. K and T are constant values, while S has been modified to mimic realistic stock price fluctuations by adding Gaussian noise of $\epsilon \sim \mathcal{N}(0, 0.05)$. We set the random number at 1, the interest rate at 0.05, and the initial values for S and K at 100, with σ established at 0.3. The data was generated for a total of 100 steps.

Subsequently, for each time point with the added noise, we applied these values of S, to the model equation, generating values for C and P as shown in Fig. 11. Fig. 12 is the ground truth graph of Black-Scholes model.

E.2.2 SEIHR

SEIHR model estimates the transmission rate during the spread of an infectious disease (Niu et al., 2020) as follows:

$$\dot{S} = -(\eta E + \alpha I)S/N$$

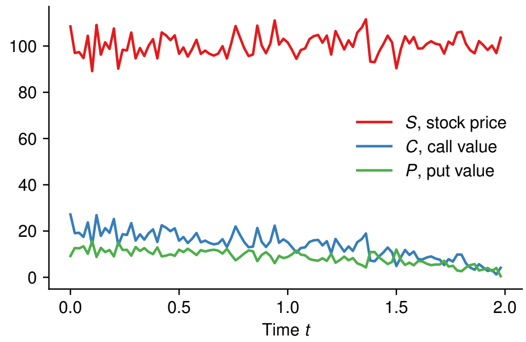

Figure 11: Data sampled from Black-Scholes model.

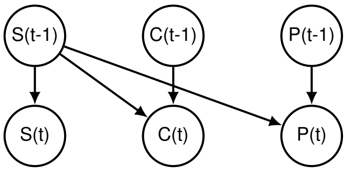

Figure 12: Black-Scholes model as a window causal graph.

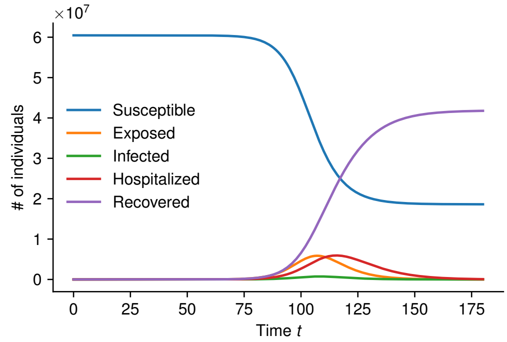

Figure 13: Data sampled from SEIHR model.

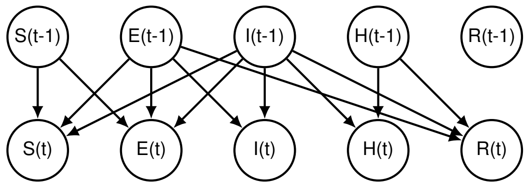

Figure 14: Ground truth graph of SEIHR model as a window causal graph.

$$
\begin{aligned}
\dot{E} &= (\eta E + \alpha I) S/N - (\beta + \omega_E) E \\
\dot{I} &= \beta E - (\gamma + \omega_I) I \\
\dot{H} &= \gamma I - \omega_H H \\
\dot{R} &= \omega_E E + \omega_I I + \omega_H H,
\end{aligned}
$$

where variables are susceptible individuals (S), those exposed to the disease (E), infected individuals (I), individuals receiving treatment (H), and individuals who have recovered (R). Other constants are as follows: η represents the transmission rate of individuals who have been exposed to the disease. α signifies the transmission rate, primarily applicable to infected individuals showing symptoms. β stands as the reciprocal of the mean latent period. γ represents the rate at which infected individuals require hospitalization. ω_E, ω_I, and ω_H are all denoting recovery rates. Specifically, ω_E stands for the rate at which non-hospitalized exposed individuals recover, while ω_I represents the recovery rate for non-hospitalized infected individuals. Lastly, ω_H corresponds to the rate at which hospitalized individuals recover.

We used the Italian region model in Niu et al. (2020). The reason for choosing the Italian region is that it presented a case where transmission dynamics were observable, offering a believable context for transmission events. The Italian region parameters assume a total population of 60,461,828, with an initial infected count of 1, and hyperparameters set as η at 0.35, α at 0.46, β at 0.14 and all ω at 0.1 over 180 days. Fig. 13 shows the result of the setting and Fig. 14 is a ground truth graph of SEIHR model.

F Preliminary of CGNN

CGNN is a differentiable generative model for score-based causal discovery (Goudet et al., 2018). We selected CGNN as a representative method to show the effectiveness of graph initialization because CGNN optimizes a skeleton graph derived from either the data or a prior graph. A skeleton graph is refined via a greedy procedure by reversing, adding, or removing edges.

G Additional Experimental Results of Section 5

We also conducted other experimental results and figures. Regardless of the choice of algorithm or dataset, we observed that our method reduced false positives and false negatives, resulting in a higher performance. Fig. 15 illustrates heatmaps for NOTEARS in Arctic Sea Ice dataset and Fig. 16 depicts heatmaps for NOTEARS and DAG-GNN for Sachs dataset. Figs. 17 and 18 are heatmaps for NTS-NOTEARS and DYNOTEARS representing both inter-slice and intra slice on Black-Scholes and SEIHR dataset. Table 9 and Table 10 details the result of CGNN on Arctic Sea Ice and Sachs dataset. Fig. 19 and Fig. 20 each shows SHD, FDR, TPR and FPR, NHD, NHD ratio of NOTEARS and CGNN on physical knowledge-based synthetic datasets whose sizes are 3, 5, and 7 nodes.

G.1 Experimental Results of CGNN

CGNN showed a notable performance improvement by solely using graph initialization.

CGNN exhibited higher performance compared to random prior and GPT-4 in Arctic Sea Ice dataset. Vanilla CGNN failed to make any predictions, but with GPT-4 prior, it produced more accurate predictions than GPT-4 as detailed in Table 9. Using a random prior resulted in worse predictions than making no predictions at all, showing a decline in performance. However, it still slightly outperformed GPT-4.

In Sachs dataset, it also outperformed vanilla CGNN and performed similarly to GPT-4. The performance improved across all metrics compared to vanilla CGNN and to CGNN with a random prior, but there was no significant difference compared to GPT-4 as detailed in Table 10.

G.2 Experimental Results on Physical Synthetic Dataset

We report all results about the physical synthetic dataset in Table 11. Overall, we observed that the integration of PLM prior improves performance when the number of nodes is larger than three (except for TPR of CGNN on five node dataset). When the number of nodes is three, the causal graph of the dataset is too simple for NOTEARS so that it exactly predicted causal graphs of the dataset, resulting in no difference whether integrating PLM prior or not. If the number of nodes is larger than three, vanilla NOTEARS did not predict any edges, and integration of PLM prior brings out consistent performance enhancement over all metrics.

Similarly to NOTEARS, when the node size is smallest, CGNN showed no difference following the integration of PLM prior. However, except for TPR, CGNN performance is improved with a huge difference, more than that of NOTEARS. From the insights of (Goudet et al., 2018), which indicate that utilizing priors closer to the ground truth graph enhances the performance of CGNN, we interpret that PLM priors provide promising skeleton graphs.

Generally, the bigger the number of nodes gets, the harder the combinatorial problems are so SHD and TPR are getting worse as we can observe in Figs. 19 and 20. In contrast, our framework mitigated the decline in performance than conventional causal discovery algorithms and GPT-4. For the five and seven nodes datasets, NOTEARS shows enhancement of all the metrics concretely when integrated with PLM prior.

G.3 Full Results of Revision Prompt

In Table 12, we provide a full table including the results of Sachs Dataset. Similar to the result of the Arctic Sea Ice dataset, the simplest pairwise causal reasoning prompt recorded the best performance across all metrics, proving that mere prompt engineering is not effective in utilizing additional information of causal structure.

Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
GPT-4	0.23	0.42	19	32	0.28	0.09	0.47
CGNN(*)	0.33	0.33	48	0	-	-	-
w/ random prior	0.42 (▲0.09)	0.66 (▲0.33)	39 (▼9)	43	0.64	0.28	0.31
w/ GPT-4 prior	0.22 (▼0.11)	0.39 (▲0.06)	19 (▼29)	35	0.28	0.10	0.52

Table 9: Performances of CGNN on the **Arctic Sea Ice dataset**. With and without GPT-4 prior, and uniform random prior whose number of the edge is the same with GPT-4 prior are investigated.

Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
GPT-4	0.14	0.45	18	21	0.47	0.09	0.57
CGNN	0.26	0.84	30	19	0.84	0.15	0.15
w/ random prior	0.29 (▲0.03)	0.84	31 (▲1)	23	0.85 (▲0.01)	0.20 (▲0.05)	0.17 (▲0.02)
w/ GPT-4 prior	0.14 (▼0.12)	0.47 (▼0.37)	18 (▼12)	19	0.47 (▼0.37)	0.08 (▼0.07)	0.52 (▲0.37)

Table 10: Performances of CGNN on the **Sachs dataset**. With and without GPT-4 prior, and uniform random prior whose number of the edge is the same with GPT-4 prior are investigated.

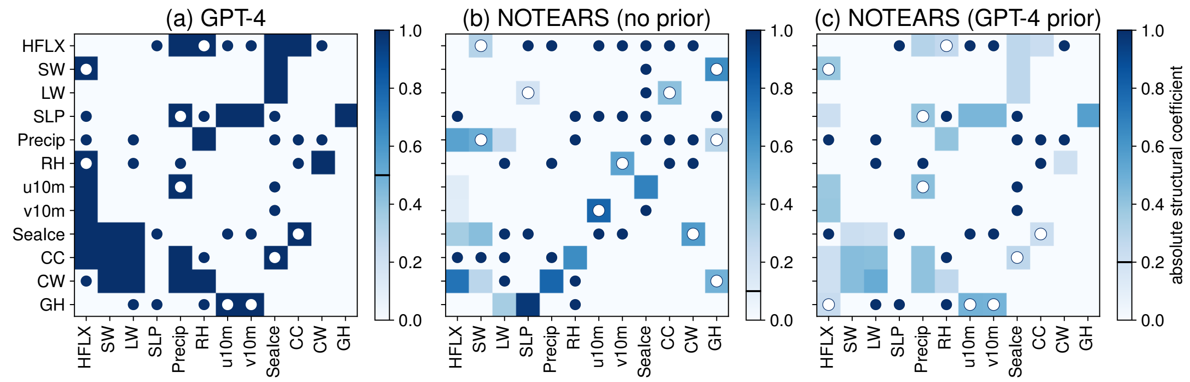

Figure 15: Heatmaps in Arctic Sea Ice dataset by a) GPT-4, b) NOTEARS, and c) NOTEARS with GPT-4 prior.

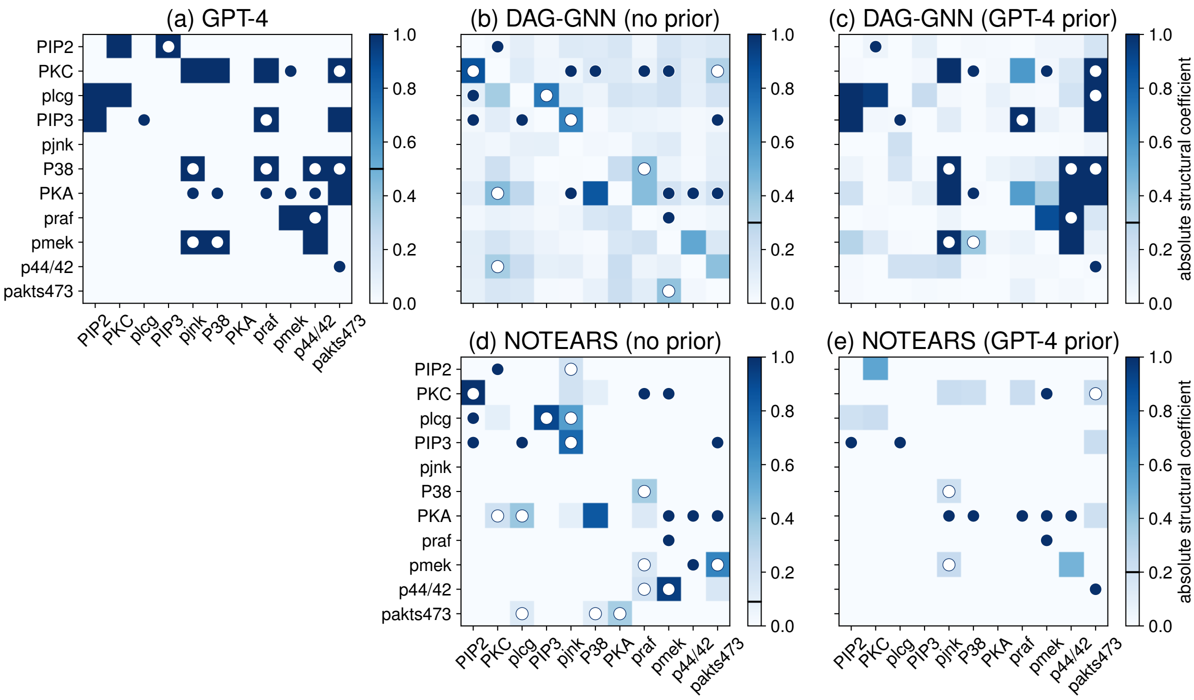

Figure 16: Heatmaps in Sachs dataset by a) GPT-4, b) NOTEARS, and c) NOTEARS with GPT-4 prior, d) DAG-GNN, and e) DAG-GNN with GPT-4 prior

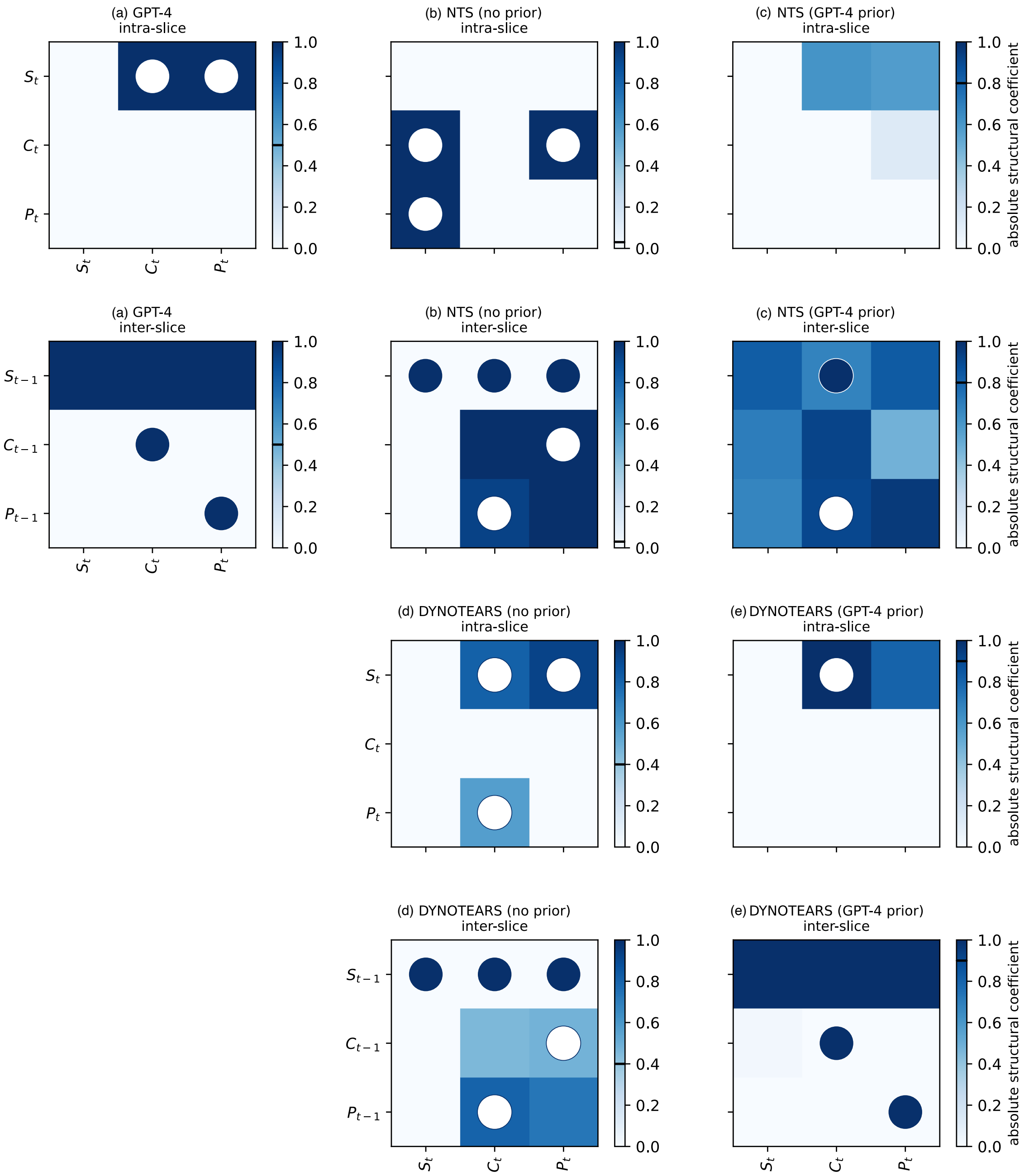

Figure 17: Heatmaps in Black-Scholes dataset by a) GPT-4, b) NTS-NOTEARS, c) NTS-NOTEARS with GPT-4 prior. d) DYNOTEARS, and e) DYNOTEARS with GPT-4 prior.

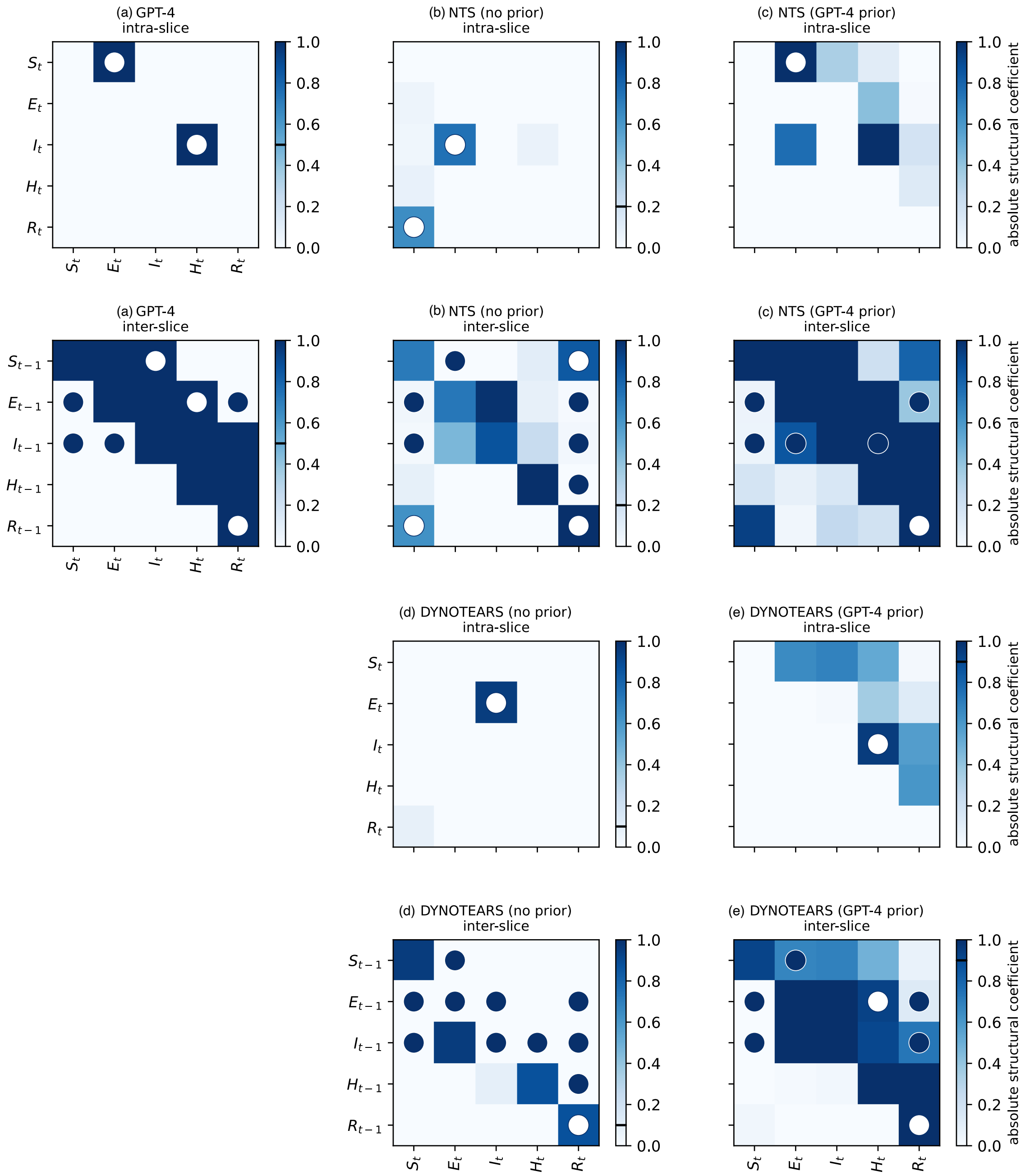

Figure 18: Heatmaps in SEIHR dataset by a) GPT-4, b) NTS-NOTEARS, c) NTS-NOTEARS with GPT-4 prior. d) DYNOTEARS, and e) DYNOTEARS with GPT-4 prior.

Dataset	Method	NHD	NHD Ratio	SHD	No. Edge	FDR	FPR	TPR
3 nodes	GPT-4	0.33	0.43	3	5	0.60	0.42	1.00
	NOTEARS	0.00	0.00	0	2	0.00	0.00	1.00
	NOTEARS (GPT-4 prior)	0.00	0.00	0	2	0.00	0.00	1.00
	CGNN	0.11	0.33	1	1	0.50	0.12	1.00
	CGNN (GPT-4 prior)	0.22	0.33	1	4	0.50	0.28	1.00
	DAG-GNN	0.00	0.00	0	2	0.00	0.00	1.00
	DAG-GNN (GPT-4 prior)	0.11	0.20	1	3	0.33	0.14	1.00
5 nodes	GPT-4	0.16	0.25	4	10	0.40	0.21	1.00
	NOTEARS	0.08	0.16	2	6	0.16	0.05	0.83
	NOTEARS (GPT-4 prior)	0.00	0.00	0	6	0.00	0.00	1.00
	CGNN	0.12	0.33	7	3	0.50	0.13	1.00
	CGNN (GPT-4 prior)	0.12	0.23	3	7	0.28	0.10	0.83
	DAG-GNN	0.00	0.00	0	6	0.00	0.00	1.00
	DAG-GNN (GPT-4 prior)	0.16	0.28	3	8	0.38	0.15	0.83
7 nodes	GPT-4	0.12	0.27	6	12	0.33	0.10	0.80
	NOTEARS	0.12	0.30	5	10	0.30	0.07	0.70
	NOTEARS (GPT-4 prior)	0.08	0.19	3	10	0.20	0.05	0.80
	CGNN	0.41	0.41	20	10	1.00	0.26	0.00
	CGNN (GPT-4 prior)	0.12	0.30	5	10	0.30	0.07	0.70
	DAG-GNN	0.10	0.29	5	7	0.14	0.02	0.60
	DAG-GNN (GPT-4 prior)	0.08	0.18	4	12	0.25	0.07	0.90

Table 11: Performances of causal discovery algorithms on the Physical Knowledge Based Synthetic datasets.

	Method	NHD (↓)	NHD Ratio (↓)	SHD (↓)	# Edges	FDR (↓)	FPR (↓)	TPR (↑)
Arctic Sea Ice	Pairwise	0.23	0.42	19	32	0.28	0.09	0.47
	LLM-complete	0.33	1.00	30	0	0.00	0.00	0.00
	LLM-cumulative	0.31	0.73	29	13	0.38	0.05	0.16
	LLM-ancestor	0.34	0.92	30	5	0.60	0.03	0.04
	GT-complete	0.33	1.00	30	0	0.00	0.00	0.00
	GT-cumulative	0.27	0.60	26	18	0.27	0.05	0.27
	GT-ancestor	0.31	0.81	28	6	0.17	0.01	0.10
Sachs	Pairwise	0.14	0.45	18	21	0.47	0.09	0.57
	LLM-complete	0.15	1.00	19	0	0.00	0.00	0.00
	LLM-cumulative	0.15	0.82	19	4	0.50	0.01	0.10
	LLM-ancestor	0.16	0.61	19	12	0.50	0.06	0.32
	GT-complete	0.14	0.90	18	1	0.00	0.00	0.05
	GT-cumulative	0.14	0.80	17	2	0.00	0.00	0.10
	GT-ancestor	0.17	0.91	20	3	0.67	0.02	0.05

Table 12: An ablation study to assess overcoming pairwise prompts via providing the information of causal relations on prompt formats.

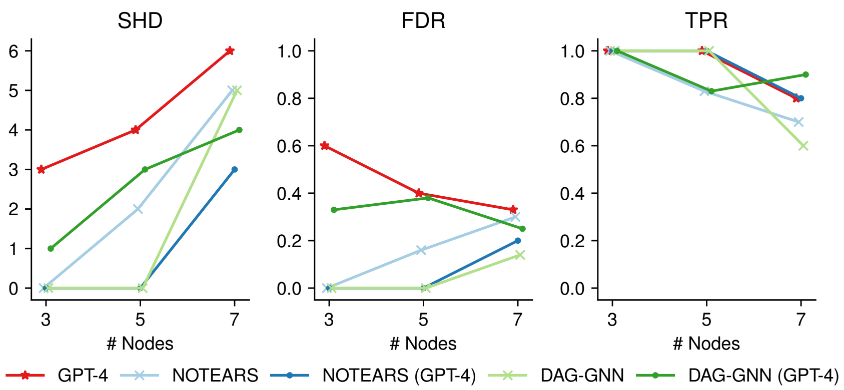

Figure 19: SHD, FDR, and TPR of NOTEARS and CGNN on the physical knowledge-based synthetic datasets with and without PLM prior.

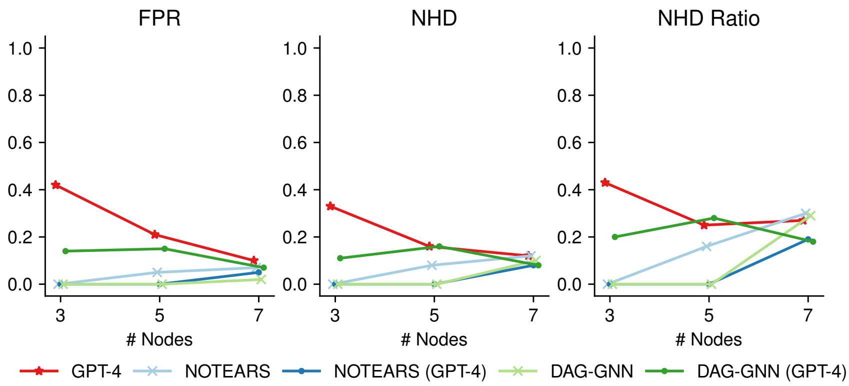

Figure 20: FPR, NHD, NHD Ratio of comparison on the physical knowledge-based synthetic datasets.