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

**Chanhui Lee**[1*], **Juhyeon Kim**[2*], **Yongjun Jeong**[3], **Yoonseok Yum**[2], **Juhyun Lyu**[4], **Junghee Kim**[4], **Sangmin Lee**[4], **Sangjun Han**[4], **Hyeokjun Choe**[4], **Soyeon Park**[4], **Woohyung Lim**[4], **Kyunghoon Bae**[4], **Sungbin Lim**[5,6†], **Sanghack Lee**[2,7†]

[1]Department of Artificial Intelligence, Korea University
[2]Graduate School of Data Science, Seoul National University
[3]Department of Computer Science and Engineering, UNIST
[4]Data Intelligence Laboratory, LG AI Research
[5]Department of Statistics, Korea University
[6]LG AI Research
[7]SNU-LG AI Research Center

## Abstract

Pre-trained Language Models (PLMs) can reason about causality by leveraging vast pre-trained knowledge and text descriptions of datasets, proving their effectiveness even when data is scarce. However, there are crucial limitations in 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 sufficiently large enough. To overcome each other approaches' limitations, we propose a new framework that integrates PLMs-based causal reasoning into data-driven causal discovery, resulting in improved and robust performance. Furthermore, our framework extends to the time-series data and exhibits superior performance.

## 1 Introduction

Causal discovery [Spirtes et al., 2000] is to figure out the causal structure among the variables. Given that figuring causal structure greatly helps decision-making for effective resource utilization and risk management, recently, causal discovery has been actively applied in various fields [Nowack et al., 2020, Russell et al., 2023]. However, in real-world applications, data is often not enough to reveal the underlying causal structures. One approach to handle such data scarcity is using domain knowledge [Borboudakis et al., 2011, Kalainathan et al., 2018], e.g., by using an appropriate graph as prior knowledge, causal discovery can be effectively guided [Sinha et al., 2021].

Recent PLMs have demonstrated its prior knowledge over diverse reasoning tasks [Wei et al., 2022, OpenAI, 2023, Anil et al., 2023, Touvron et al., 2023]. By employing specifically crafted task descriptions, PLMs address *logical reasoning*, which aims to determine the truthfulness of a given statement, such as commonsense reasoning [Huang et al., 2019], arithmetic reasoning [Suzgun et al., 2022]. Based on that, Kıcıman et al. [2023] initiated *reasoning-based causal discovery*, which aims not only to determine the truthfulness of the individual hypotheses but also to comprehend the interconnections among hypotheses to discover a graphical causal structure. In *reasoning-based*

*Equal contributions. Chanhui Lee: chanhui-lee@korea.ac.kr, Juhyeon Kim: kimjh9474@snu.ac.kr,
†Corresponding Authors. Sungbin Lim: sungbin@korea.ac.kr, Sanghack Lee: sanghack@snu.ac.kr

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

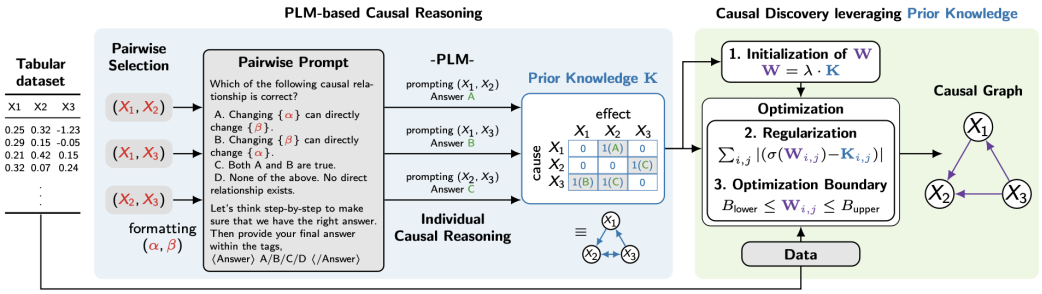



Figure 1: Overview of our framework. Given dataset, PLM-based causal reasoning returns an adjacency matrix as prior. Utilizing the prior, a causal discovery algorithm takes the dataset and returns a structural coefficient matrix, which is then mapped to a binary adjacency matrix.

*causal discovery*, one should consider causal structures like chain, fork, and collider so as not to suffer from false prediction of causal relation. For example, without noticing chain structure, one cannot differentiate between direct causal relation and indirect causal relation resulting in false predictions.

Kıcıman et al. [2023] used a prompt template (Figure 1) to determine pairwise causal relation of the variables. By aggregating individual predictions on pairwise causal relaionship, Kıcıman et al. [2023] reports superior causal discovery performance than data-driven causal discovery algorithms. Unlike question-answering works [Zellers et al., 2019, Sap et al., 2019, Huang et al., 2019], Kıcıman et al. [2023], which belongs to logical reasoning, *reasoning-based causal discovery* focuses on discovering the causal structure among variables as a graph, evaluated by graph evaluation metrics (*e.g.* Structural Hamming Distance).

However, PLM-based causal reasoning have inherent limitations compared to data-driven causal discovery. First, they cannot properly utilize large tabular data only limited to handling only small-scale tabular data [Liu et al., 2023, Lei et al., 2023, Li et al., 2023]. Second, in Kıcıman et al. [2023], PLM only predicts pairwise causal relations *individually*, and the entire causal structure is aggregated in an end-to-end manner, which lacks a comprehensive holistic understanding of causal structures. Third, attempts to leverage structural information upon Kıcıman et al. [2023] for a better holistic understanding of causal structure for PLM, rather than degraded performance. Though we tried various prompt schemas adapted to causal discovery in Table 3, such approaches actually degraded the causal reasoning performance, similar to the finding of Levy et al. [2024]. This suggests that leveraging structural information is not easily achievable via prompt engineering.

Given the pros and cons of both PLM-based causal reasoning and data-driven causal discovery algorithms, we propose a novel framework that integrates the two approaches. We combine (i) the strengths of data-driven causal discovery, which is suitable for learning large datasets to discover the entire causal structure, with (ii) the utility of PLM-based causal reasoning, which enables causal discovery even when data is scarce. Integration of the both methods mitigates the data scarcity problems of causal discovery algorithms and reduces PLM hallucination, which is hard to fix without leveraging observed data, more effectively than concurrent work [Abdulaal et al., 2024]. Moreover, we extended the application of PLM-based reasoning to address 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. During this, we revealed that time-series causal discovery relying solely on PLMs is largely influenced by prompt design artifacts.

**Contributions** We summarize our contributions.

- We demonstrate that PLM-based pairwise causal reasoning methods are not suitable for holistically eliciting a causal structure.
- 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.
- The proposed framework outperforms baselines, demonstrating that our framework is more effective than concurrent work in mitigating PLM's hallucination.

## 2 PLM-based causal reasoning

We investigate PLM-based causal reasoning used in Kıcıman et al. [2023], and variations of Kıcıman et al. [2023] and their limitations.

Table 1: (a) Experiment to investigate revision prompt following Ban et al. [2023] causal-reasoning improves PLM-based causal reasoning. (b) Ablation study assessing the effect of quantity and quality of structural information on PLM-based causal reasoning.

(a)

| | Method | #Edge | FDR↓ | FPR↓ | TPR↑ |
|---|---|---|---|---|---|
| Arctic | GPT-4 | 32 | 0.28 | 0.09 | 0.47 |
| | Revised | 50 | 0.42 | 0.21 | 0.60 |
| Sachs | GPT-4 | 21 | 0.47 | 0.09 | 0.57 |
| | Revised | 19 | 0.52 | 0.09 | 0.47 |

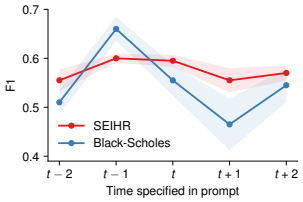
(b)

| | Method | #Edge | FDR↓ | FPR↓ | TPR↑ |
|---|---|---|---|---|---|
| Arctic Sea Ice | Pairwise | 32 | 0.28 | 0.09 | 0.47 |
| | PLM-complete | 0 | 0.00 | 0.00 | 0.00 |
| | PLM-cumulative | 13 | 0.38 | 0.05 | 0.16 |
| | PLM-ancestor | 5 | 0.60 | 0.03 | 0.04 |
| | GT-complete | 0 | 0.00 | 0.00 | 0.00 |
| | GT-cumulative | 18 | 0.27 | 0.05 | 0.27 |
| | GT-ancestor | 6 | 0.17 | 0.01 | 0.10 |

**PLM-based causal reasoning** Utilizing the prompt templates for static Figure 1 and time-series Figure 2 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.

**Difficulty in utilizing causal structural information** We examine whether PLM-based causal reasoning is adept at comprehending a causal structure, following Ban et al. [2023]. According to Ban et al. [2023], the PLM initially predicts causal relations for all pairwise variables, followed by an adjustment of these relations using a revision prompt. Our experiments, referenced in Table 1a, show only a slight alteration due to the revision prompt (detailed in Appendix A).

Additionally, we investigated the influence of both the quantity and quality of causal structural information. Regarding the impact of information quantity, we conducted experiments using three prompt schema: *complete prompting* which uses entire previous predictions similar to Ban et al. [2023], *cumulative* prompting, which predicts pairwise causal relationships based on previously accumulated causal relation predictions, whose information quantity bridges Kıcıman et al. [2023] and Ban et al. [2023], *ancestor* using only causally relative predictions. Table 1b shows all the revision prompt templates actually do not show notable improvements. Considering that the degraded performance is due to the causal structures predicted by PLM is not beneficial, we also experimented with ground truth (GT) structural information. Still, we observed there is not notable improvement over simple pairwise causal reasoning, indicating that structure-aware PLM-based reasoning is not easily achievable via prompt engineering only.




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 a time step can directly change $\{\beta\}$ of the same time step.
B. Change $\{\beta\}$ of a 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: Left) F1 of GPT-4 prediction (mean of 10 repetitions) on time-lagged causal relations. Right) Multiple choice templates (upper for time-lagged causal relation, lower for contemporaneous causal relation) for the causal relation between two variables in time-series data.

**Difficulty in expanding PLM reasoning to time series** As we know, there is no PLM-based causal reasoning work for time-series datasets, despite its practical importance in real-world applications. We expand the application of PLM-based causal reasoning to time-series datasets by a prompt template (Figure 2), which inquires both time-lagged and contemporaneous causal relations for pairwise variables, $\{\alpha\}$ and $\{\beta\}$. Through our experiments, we noted that performance is influenced by prompt-specific factors such as the time lag. To demonstrate this, we selected stationary temporal domains where the maximum time lag is 1 so that for a given pairwise variables, lexical choices to represent one time step difference should hold a fixed causal relation. For example, querying whether $\alpha_{t-1}$ is the cause of $\beta_t$ and $\alpha_t$ is the cause of $\beta_{t+1}$ should give the same result. The experiment in Figure 2 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 reveal the lack of capability to comprehend causal structures and understanding temporal consistency, suggesting that PLM alone does not strictly adhere to causal discovery.

## 3 Causal discovery with PLM-derived priors

To overcome the challenges of relying solely on PLM, we propose a framework which incorporates PLM-based causal reasoning into an optimization-based causal discovery algorithm, by utilizing a prior knowledge $\mathbf{K}$ extracted from PLM. Causal discovery algorithms figure out a causal graph effectively utilizing tabular datasets. 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.

The overall framework is depicted in Figure 1. Given static or time-series datasets as input, our framework performs PLM-based reasoning through specifically designed prompts (Figures 1 and 2). Then, by aggregating pairwise causal relations, we acquire a prior knowledge $\mathbf{K}$. The causal discovery algorithm's optimization process then makes use of the $\mathbf{K}$ in three ways (not exclusively): Graph initialization before data-driven optimization, optimization regularization and boundary optimization w.r.t. to the $\mathbf{K}$, both for robustness against anormaly in data. After the optimization of structural coefficient ends, we apply a thresholding return a binary adjacency matrix (i.e., a directed graph).

**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 $\lambda_{\text{init}}$ is introduced for adjustment of $\mathbf{K}$.

**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 $L_{\text{sim}}(\mathbf{W}) := \sum_{i,j} |(\sigma(\mathbf{W}_{i,j}) - \mathbf{K}_{i,j})|$, which can be viewed as $\ell_1$-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 $\sigma$, 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 $\lambda_{\text{sim}}$ is the hyperparameter.

**Setting boundaries for optimization** We now consider applying prior knowledge for setting each structural coefficient's boundary $B$ as $B_{\text{lower}} \leq \mathbf{W}_{i,j} \leq B_{\text{upper}}$, to be utilized during the optimization process. Sun et al. [2021] set $B_{\text{lower}}$ 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 $(i, j)$. When extracting $\mathbf{K}$ from PLM, we need to mitigate the risk of hallucination in prior knowledge. Therefore, we set $0 < B_{\text{lower}} < \mathbf{W}_{i,j}\ \forall i, j\ \mathbf{K}_{i,j} = 1$, and if $\mathbf{K}_{i,j} = 0$, we 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].

Table 2: The performances of NOTEARS and DAG-GNN on static datasets (upper two) and NTS-NOTEARS and DYNOTEARS on time-series datasets (lower two) are indicated with differences against the vanilla algorithm. * indicates metrics that can be calculated via the true positive, precision, and recall reported in the CMA paper [Abdulaal et al., 2024].

| | 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 | 0.17 | 0.60 | 20 | 16 | 0.56 | 0.08 | 0.36 |
| | w/ random prior | 0.27 (+0.10) | 0.82 (+0.22) | 28 (+8) | 21 | 0.83 (+0.27) | 0.17 (+0.09) | 0.18 (-0.18) |
| | w/ GPT-4 prior | 0.11 (-0.06) | 0.35 (-0.25) | 14 (-6) | 19 | **0.36** (-0.20) | **0.06** (-0.02) | 0.63 (+0.27) |
| | DAG-GNN | 0.19 | 0.62 | 22 | 18 | 0.61 | 0.10 | 0.36 |
| | w/ random prior | 0.27 (+0.08) | 0.81 (+0.19) | 29 (+7) | 21 | 0.83 (+0.22) | 0.17 (+0.07) | 0.20 (-0.16) |
| | w/ GPT-4 prior | **0.10** (-0.09) | **0.30** (-0.32) | **13** (-9) | 24 | 0.37 (-0.24) | 0.08 (-0.02) | **0.78** (+0.42) |
| 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 (+0.00) | 0.80 (+0.13) | 8 (+0) | 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 (+0.00) | 1.00 (+0.33) | 8 (+0) | 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) |

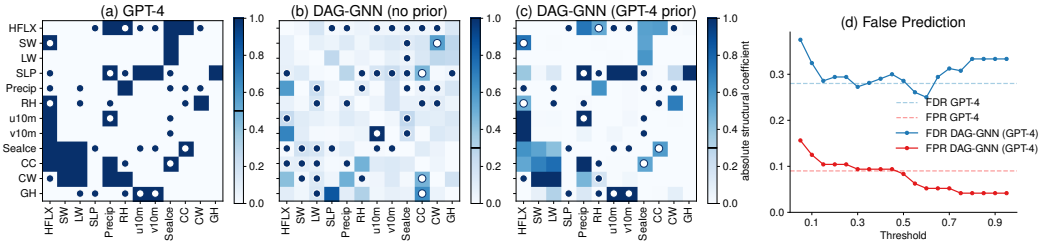



Figure 3: Predictions in Arctic Sea Ice of (a) GPT-4, (b) DAG-GNN, and (c) DAG-GNN with GPT-4 prior. Dark circles are false negatives and white circles are false positives. A threshold is annotated as a line in a colorbar. (d) False prediction of DAG-GNN.

## 4 Experiments

We evaluate our proposed framework across the static datasets and time-series datasets: Black-Scholes, SEIHR (detailed in Appendix C). We used prior knowledge from GPT-4 [OpenAI, 2023] determined with the majority vote from 10 results. For static datasets, we employed NOTEARS, DAG-GNN, and for time-series, adopted, DYNOTEARS and NTS-NOTEARS (detailed in Appendix B).

### 4.1 Causal discovery on static datasets: Arctic Sea Ice & Sachs

**Arctic Sea Ice** Arctic Sea Ice dataset comprises 12 Earth science-related variables and only 486 observations. The corresponding causal graph is constructed via a meta-analysis of literature referred to in Huang et al. [2021], containing 48 edges with 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. As shown in Table 2, the small sample size of this dataset poses challenges for data-driven causal discovery algorithms. On the other hand, PLM-based causal reasoning and our framework shows better performance than conventional causal discovery algorithms, which can be explained with the usage of prior knowledge related to metadata. As PLM-based causal reasoning leverages the names of variables and related knowledge obtained in pre-training, it is not restricted by the small size of the dataset.

When evaluating PLM-based causal reasoning against our framework, the absence of data usage results in a missing opportunity to correct the hallucinations of the PLM through observations, leading to a higher false discoveries of GPT-4. Our framework also outperformed a recent work, Causal Modelling Agents (CMA) [Abdulaal et al., 2024], which likewise combines PLM (GPT-4) and causal discovery across all metrics except for TPR. To better understand the effect of integration, we visualized the structural coefficients matrices (Figs. 3a to 3c). We observed that GPT-4 alone predicts lots of hallucinations with full confidence, figured out as false positives and negatives in Figure 3a. The effect of varying threshold values is depicted in Fig. 3d.

Our framework also outperformed a recent work, Causal Modelling Agents (CMA) Abdulaal et al. [2024], which likewise combines PLM (GPT-4) and causal discovery across all metrics except for TPR. While both CMA and our framework use the outcomes of PLMs as priors for causal discovery, CMA utilizes the PLM prior knowledge lowering the likelihood of fitting data. In contrast, our framework leverages PLM prior knowledge not only for optimization, but to provide effective initial structural coefficient (graph initialization) and robustness to anamaly data (boundary for optimization), whose effectiveness is supported by superior performances on diverse datasets Given that our framework and CMA use the same PLM (GPT-4), the better performance in false positives and negatives reveals that our framework reduces the hallucination of PLM more effectively than CMA.

**Sachs** Sachs dataset [Sachs et al., 2005] is comprises 11 variables about protein signaling pathways with 7,466 observations, which is large in the causal discovery community. We report experimental results in Table 2 where causal discovery algorithms exhibited different performance trends Similar to the trend in Arctic Sea Ice dataset, though causal discovery algorithms are provided with plenty of observation, integrating prior knowledge of GPT-4 still enhanced overall performance improvements over causal disocery algoirthms, in addition to the GPT-4 alone.

### 4.2 Causal discovery on time-series datasets: Black-Scholes & SEIHR

For time-series causal discovery, we evaluated using synthetically generated datasets following Partial Differential Equations (PDEs) of Black-Scholes Black and Scholes [1973] model in the finance domain and SEIHR Niu et al. [2020] model in the epidemic domain. In addition, to provide real-life motivation and insight into how our framework works on a dataset of many variables, we report the results of the S&P 100 dataset. The result is consistent with previous research Pamfil et al. [2020], Wang et al. [2020], demonstrating the validity of our framework (detailed in Appendix E)

**Black-Scholes** Black-Scholes is a probabilistic model to predict stock prices, determining the current value of options [Black and Scholes, 1973]. Based on the PDEs, we annotated the evaluation graph with 3 nodes and 5 edges. The overall performance of our framework demonstrated a marked improvement compared to the vanilla algorithms and GPT-4 as in Table 2. Compared to GPT-4, our model showed significant improvement in overall metrics, especially in FPR and FDR.

**SEIHR** SEIHR model estimates the transmission rate of an infectious disease [Niu et al., 2020]. The dynamics of SEIHR are modeled using PDEs with 5 nodes and 13 edges. As reported in Table 2, our framework demonstrated a consistent improvement in performance compared to the vanilla algorithms and GPT-4. In contrast, when compared to GPT-4, there was no corresponding overall performance enhancement in DYNOTEARS. We conjecture that the discrepancy may arise from the nonlinearity of the data, violating the linearity assumption of DYNOTEARS.

## 5 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 for parameter optimization, to combines the strengths of both worlds. By integration, the proposed methods ourperformed baseline causal discovery algorithms and PLM through divers static, time-series datasets. 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.

## Acknowledgements

This work was supported by LG AI Research. This work was also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00410082), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. NRF-2022M3J6A1063595), and Korea University Grant(K2407521). This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City. This research was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center)(IITP-2024-RS-2024-00436857, 20%) grant funded by the Korea government(Ministry of Science and ICT).

## References

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

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

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *ArXiv*, abs/2306.16902, 2023. URL https://api.semanticscholar.org/CorpusID:259287331.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:208290939.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

Giorgos Borboudakis, Sofia Triantafillou, Vincenzo Lagani, and I. Tsamardinos. A constraint-based approach to incorporate prior knowledge in causal models. In *The European Symposium on Artificial Neural Networks*, 2011. URL https://api.semanticscholar.org/CorpusID:8926360.

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

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

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

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*, 2019.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in Big Data*, 2021. URL https://www.frontiersin.org/articles/10.3389/fdata.2021.642182.

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

Diviyan Kalainathan, Olivier Goudet, Isabelle M Guyon, David Lopez-Paz, and Michèle Sebag. Sam: Structural agnostic model, causal discovery and penalized adversarial learning. *arXiv: Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:88523786.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

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

Fangyu Lei, Xiang Lorraine Li, Yifan Wei, Shizhu He, Yiming Huang, Jun Zhao, and Kang Liu. S3hqa: A three-stage approach for multi-hop text-table hybrid question answering. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258823218.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table-tuned GPT for diverse table tasks. *ArXiv*, abs/2310.09263, 2023. URL https://api.semanticscholar.org/CorpusID:264127877.

Chuang Liu, Junzhuo Li, and Deyi Xiong. Tab-cqa: A tabular conversational question answering dataset on financial reports. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259370781.

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

Peer Johannes Nowack, Jakob Runge, Veronika Eyring, and Joanna D. Haigh. Causal networks for climate model evaluation and constrained projections. *Nature Communications*, 11, 2020. URL https://api.semanticscholar.org/CorpusID:212718883.

OpenAI. GPT-4 technical report, 2023.

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

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

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

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

Clark D. Russell, Nazir I. Lone, and John Kenneth Baillie. Comorbidities, multimorbidity and covid-19. *Nature Medicine*, 29:334–343, 2023. URL https://api.semanticscholar.org/CorpusID:256940540.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 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, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.33. URL https://aclanthology.org/2021.blackboxnlp-1.33.

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

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

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. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:252917648.

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

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

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

Yudong Wang, Zhiyuan Pan, Chongfeng Wu, and Wenfeng Wu. Industry equi-correlation: A powerful predictor of stock returns. *Journal of Empirical Finance*, 59:1–24, 2020.

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. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://api.semanticscholar.org/CorpusID:249674500.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:128358697.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23:550–560, 1997. URL https://api.semanticscholar.org/CorpusID:207228122.

Table 3: An ablation study to assess the effect of providing causal relations in prompts. Symbol ↓ indicates a lower-is-better metric.

| | 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 |
| | PLM-complete | 0.33 | 1.00 | 30 | 0 | 0.00 | 0.00 | 0.00 |
| | PLM-cumulative | 0.31 | 0.73 | 29 | 13 | 0.38 | 0.05 | 0.16 |
| | PLM-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 |

# A Prompt Templates and results for Revision

Here, we describe the full text of the *cumulative prompting* Figure 4 and *complete prompting* (Figure 5). 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 4: A modified prompts from [Kıcıman et al., 2023] named "cumulative prompt" that uses cumulatively found relations from the previous prompts or ground-truth relationships.

**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 4. This approach somewhat yielded performance improvements in certain instances, though the effectiveness varied across different datasets. In 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.

# B Experimental Details

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

## B.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

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 5: 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.

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

| | 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** |

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 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:

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

### B.2 Causal discovery algorithms

DAG-GNN [Yu et al., 2019] learns a structural coefficient matrix through continuous optimization to approximate the distribution of causal graphs 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. NOTEARS [Zheng et al., 2018] formulates causal discovery as a continuous optimization minimizing the following training objective: $L(\mathbf{W}) := \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_1$. 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 $\lambda$. NOTEARS minimizes the objective while ensuring the acyclicity of the learned graph.

**Time-series causal dsiscovery** 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, modeling time-lagged causal relations

Table 5: Performances of GPT-4 under varing temperatures on Arctic Sea Ice and Sachs dataset.

| | 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 |

with a structural coefficient matrix called intra-slice $\mathbf{W}$ for contemporaneous causal relations, and a matrix called inter-slice $\mathbf{A} \in \mathbb{R}^{(T \times d) \times d}$ for 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. For readability, we simply refer the concatenation of $\mathbf{W}$ and $\mathbf{A}$ as $\mathbf{W}$.

### B.3 Setup

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 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 5, 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 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—$t$, $\lambda_{\text{sim}}$, thresholds in Table 6. $\lambda_{\text{init}}$ is the scaling factor for graph initialization and $\lambda_{\text{sim}}$ is that for prior graph similarity regularization. As we mentioned in the Experimental setup, hyperparameters of the 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 graph 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

Table 6: Hyperparameters of Arctic Sea Ice and Sachs

| | Method | Prior | $\lambda_{\text{init}}$ | $\lambda_{\text{sim}}$ | 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 |

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[3].

### B.4 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 7. We tested GPT4, GPT4 turbo [OpenAI, 2023], PaLM 2 [Anil et al., 2023], Claude[4], 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.

Table 7: 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)

| | 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 |

## C 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.

[3] `https://github.com/quantumblacklabs/causalnex`

[4] `https://www.anthropic.com/product`

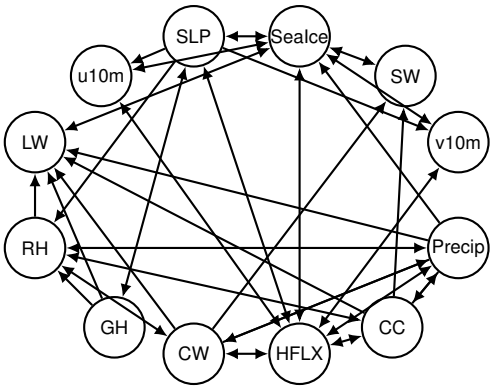



Figure 6: Arctic Sea Ice ground truth graph

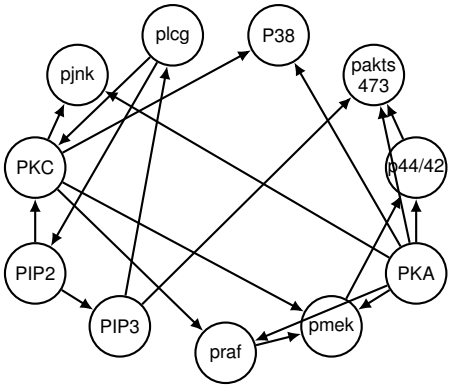



Figure 7: Sachs ground truth graph

### C.1 Static Datasets

#### C.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 (Figure 6), 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 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.

#### C.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 (Figure 7) 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.

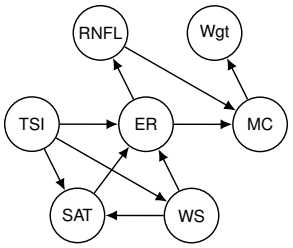



Figure 8: Physical knowledge-based synthetic graph with size 7. The components of the graph are Rainfall (RNFL), Total Solar Irradiance (TSI), Surface Air Temperature (SAT), Wind Speed (WS), Evaporation Rate (ER), Moisture Content of object (MC), and Weight of object (Wgt).

### C.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 Figure 8.

**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 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.

## C.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.

### C.2.1 Synthetic data generation

To the best of our knowledge, the usual benchmark datasets for time-series causal discovery lack real-world semantic meanings, which is essential to utilize PLMs, or ground-truth causal graphs for evaluation. Correspondingly, we simulated synthetic datasets referring to the well-established PDEs for Black-Scholes [Black and Scholes, 1973] model and SEIHR [Niu et al., 2020] model, with a 1 time-lag. The overall process for creating synthetic time-series datasets involves 1) selecting a mathematical model with trustworthy, universally acceptable semantic meaning and relationships, 2) generating synthetic time-series observation from PDE with Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.05)$, and 3) extracting causal relations to compose an evaluation causal graph. By creating a dataset from each domain's well-established PDE, we can utilize not only high-quality real-world semantic meanings annotated by domain experts but also robust ground-truth causal graphs for evaluation. Further detailed reasons for this selection are explained in the appendix.

### C.2.2 Black-Scholes

Black-Scholes model is a probabilistic model of predicting future stock prices, determining the current value of options [Black and Scholes, 1973]. 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 ($\sigma$). 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.

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

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 $\sigma$ 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 Figure 9. Figure 10 is the ground truth graph of Black-Scholes model.

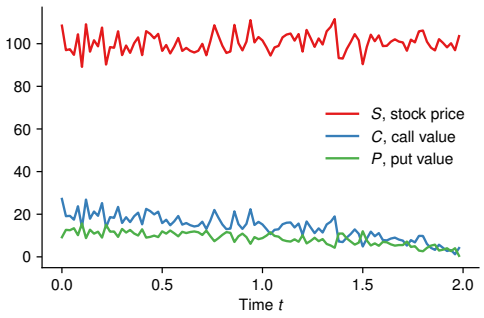



Figure 9: Data sampled from Black-Scholes model.

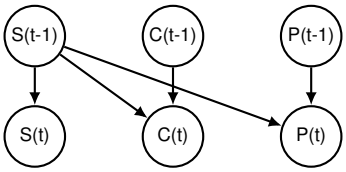



Figure 10: Black-Scholes model as a window causal graph.

### C.2.3 SEIHR

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

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

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: $\eta$ represents the transmission rate of individuals who have been exposed to the disease. $\alpha$ signifies the transmission rate, primarily applicable to infected individuals showing symptoms. $\beta$ stands as the reciprocal of the mean latent period. $\gamma$ represents the rate at which infected individuals require hospitalization. $\omega_E$, $\omega_I$, and $\omega_H$ are all denoting recovery rates. Specifically, $\omega_E$ stands for the rate at which non-hospitalized exposed individuals recover, while $\omega_I$ represents the recovery rate for non-hospitalized infected individuals. Lastly, $\omega_H$ 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 $\eta$ at 0.35, $\alpha$ at 0.46, $\beta$ at 0.14 and all $\omega$ at 0.1 over 180 days. Figure 11 shows the result of the setting and Figure 12 is a ground truth graph of SEIHR model.

## D 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.

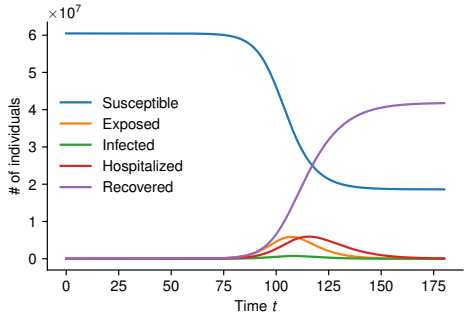



Figure 11: Data sampled from SEIHR model.

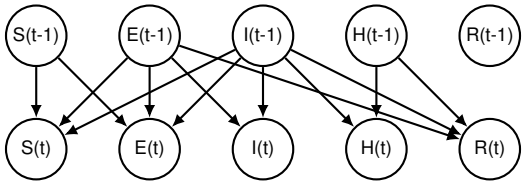



Figure 12: Ground truth graph of SEIHR model as a window causal graph.

# E Additional Experimental Results

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. Figure 13 illustrates heatmaps for NOTEARS in Arctic Sea Ice dataset and Figure 14 depicts heatmaps for NOTEARS and DAG-GNN for Sachs dataset. Figures 15 and 16 are heatmaps for NTS-NOTEARS and DYNOTEARS representing both inter-slice and intra slice on Black-Scholes and SEIHR dataset. Table 8 and Table 9 details the result of CGNN on Arctic Sea Ice and Sachs dataset. Figure 17 and Figure 18 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.

Table 8: 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 | 0.23 | 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 |

## E.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 8. 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 9.

Table 9: 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.

| 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) |

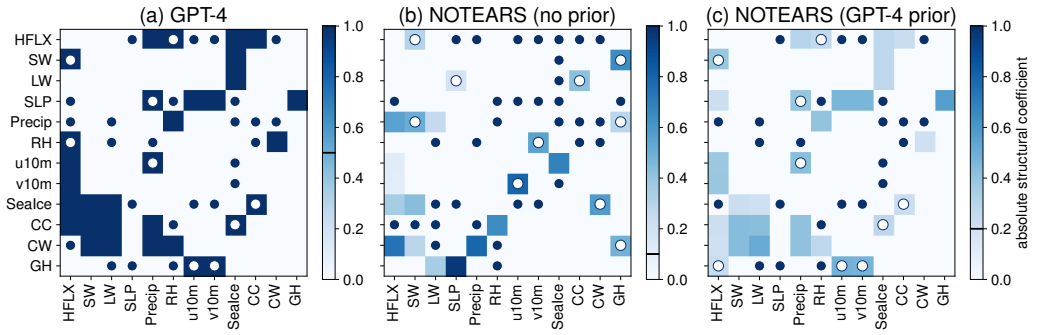



Figure 13: Heatmaps in Arctic Sea Ice dataset by a) GPT-4, b) NOTEARS, and c) NOTEARS with GPT-4 prior.

### E.2 Experimental Results on Physical Synthetic Dataset

We report all results about the physical synthetic dataset in Table 10. 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 Figures 17 and 18. 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.

### E.3 Full Results of Revision Prompt

In Table 11, 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.

### E.4 Experimental Results on S&P100 Dataset

For real-life motivation, we analyzed the S&P 100 data from 2020 to 2023. Although there are no graphs for evaluation, the analysis showed high correlations among companies in similar sectors,

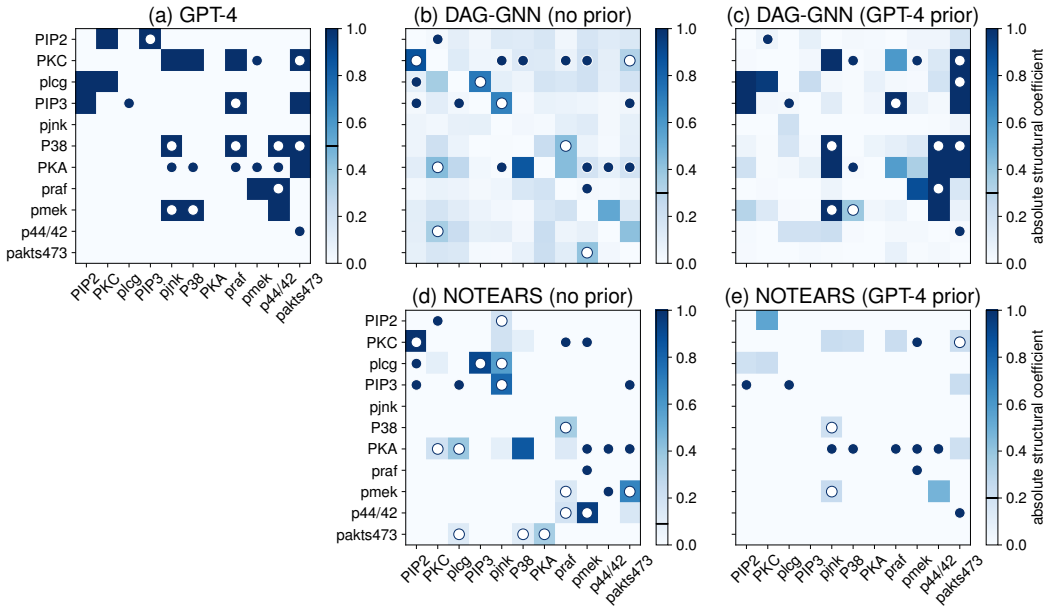



Figure 14: Heatmaps in Sachs dataset by a) GPT-4, b) NOTEARS, and c) NOTEARS with GPT-4 prior, d) DAG-GNN, and e) DAG-GNN with GPT-4 prior

Table 10: Performances of causal discovery algorithms on the Physical Knowledge Based Synthetic datasets.

| 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 |

consistent with previous research [Pamfil et al., 2020, Wang et al., 2020], demonstrating the validity of our framework with real-world data.

First, we collected the S&P 100 data from the past four years using Yahoo Financials[5], analyzing 1005 instances of the top 15 companies for clarity in visualization. We calculated log returns based

[5]https://pypi.org/project/yahoofinancials/

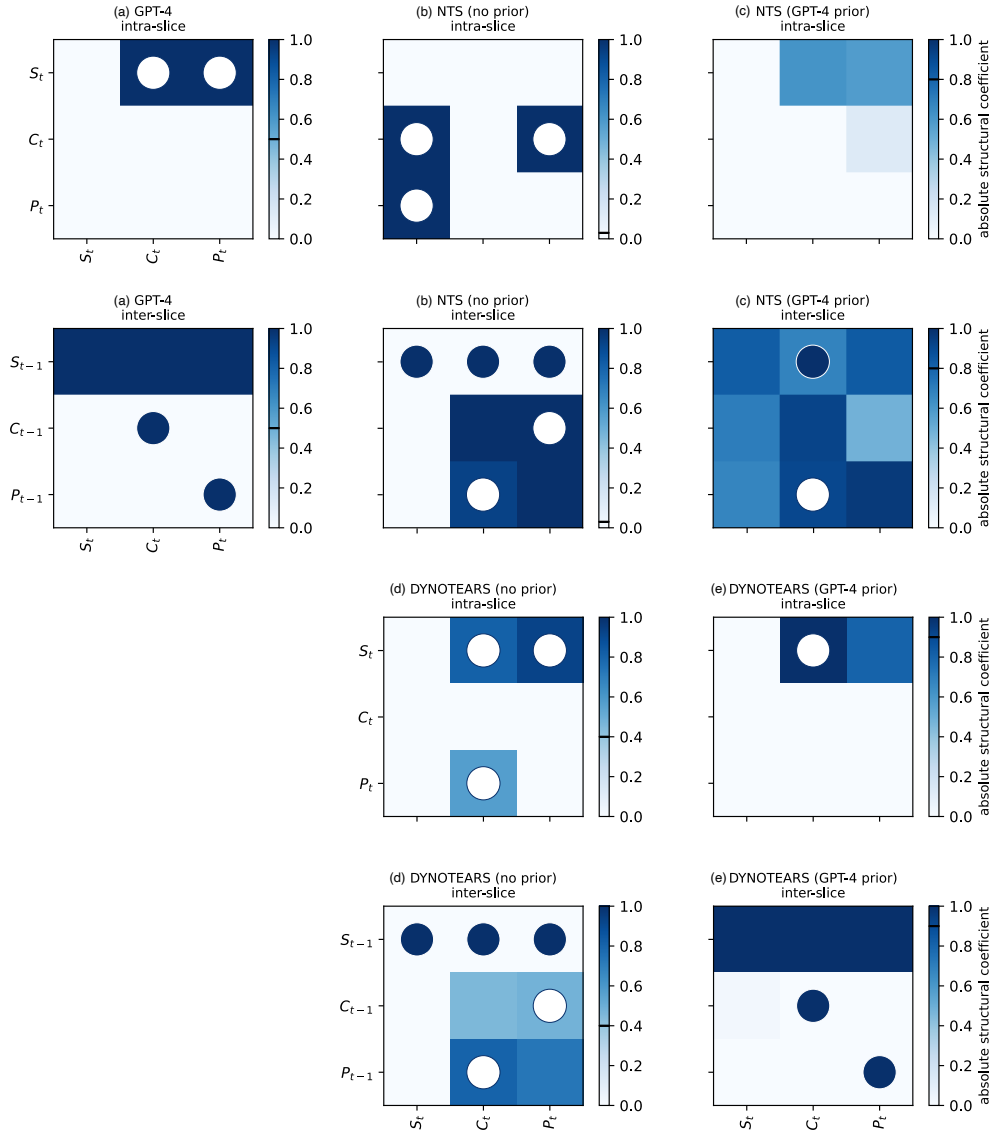



Figure 15: Heatmaps in Black-Scholes dataset by a) GPT-4, b) NTS-NOTEARS, c) NTS-NOTEARS with GPT-4 prior. d) DYNOTEARS, and e) DYNOTEARS with GPT-4 prior.

on daily end prices as $\log\left(\frac{\text{price}(t)}{\text{price}(t-1)}\right)$ to normalize values by converting them into ratios, avoiding bias towards higher values. We then normalized the log values.

The results are shown in the following Figure 19. In each heatmap, the closer to blue, the higher the coefficient. Applying NTS NOTEARS and DYNOTEARS in our framework revealed high correlations among companies in similar sectors. For example, JP Morgan, being a financial company, had no correlations with IT-based companies, retail companies, or communication service companies. On the other hand, companies like Google were related to Meta, NVIDIA, etc., and Microsoft showed similar correlations with companies like Apple and Amazon. Although classified as a retail company, Amazon showed similarities with IT due to its video streaming and cloud services.

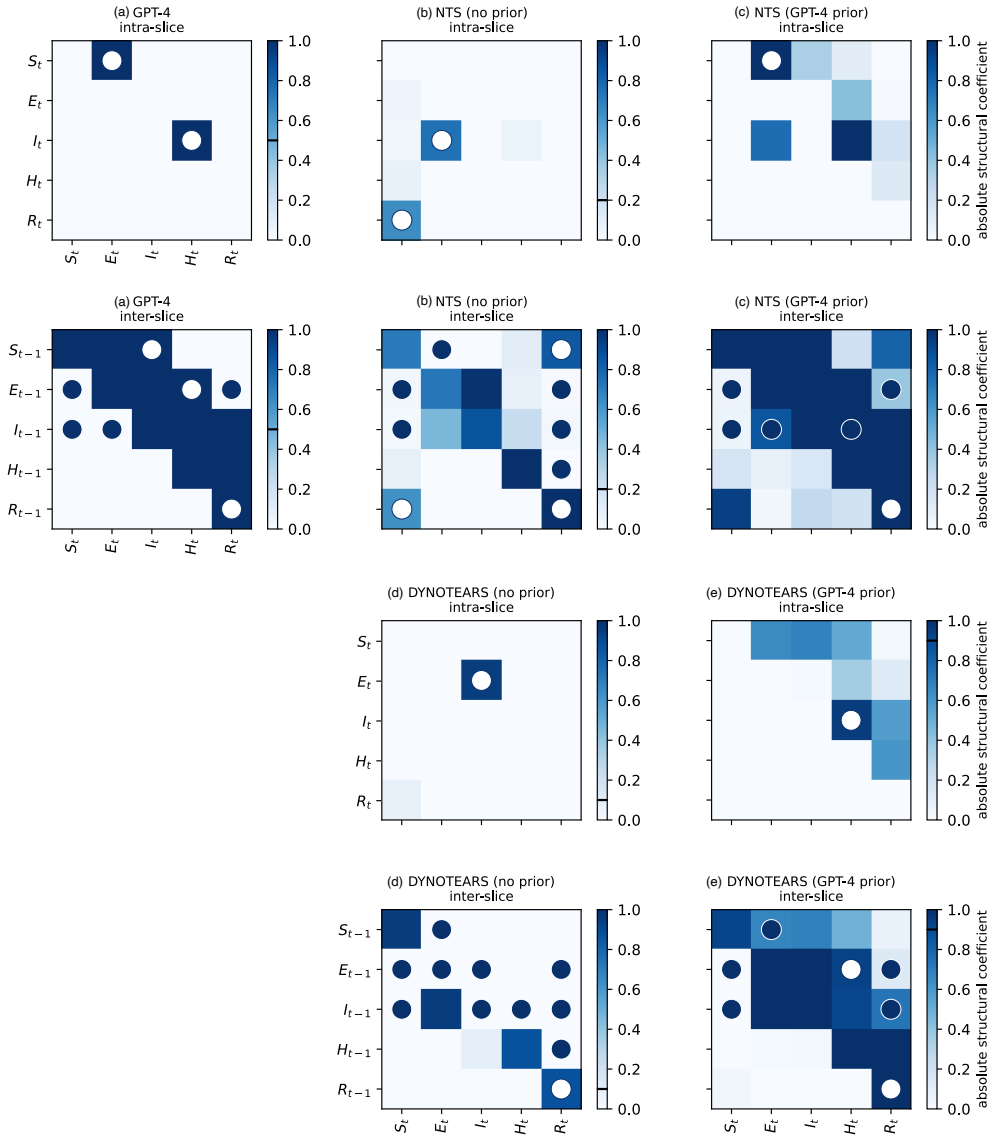



Figure 16: Heatmaps in SEIHR dataset by a) GPT-4, b) NTS-NOTEARS, c) NTS-NOTEARS with GPT-4 prior. d) DYNOTEARS, and e) DYNOTEARS with GPT-4 prior.

Table 11: An ablation study to assess overcoming pairwise prompts via providing the information of causal relations on prompt formats.

| | 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 |
| | PLM-complete | 0.33 | 1.00 | 30 | 0 | 0.00 | 0.00 | 0.00 |
| | PLM-cumulative | 0.31 | 0.73 | 29 | 13 | 0.38 | 0.05 | 0.16 |
| | PLM-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 |
| | PLM-complete | 0.15 | 1.00 | 19 | 0 | 0.00 | 0.00 | 0.00 |
| | PLM-cumulative | 0.15 | 0.82 | 19 | 4 | 0.50 | 0.01 | 0.10 |
| | PLM-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 |

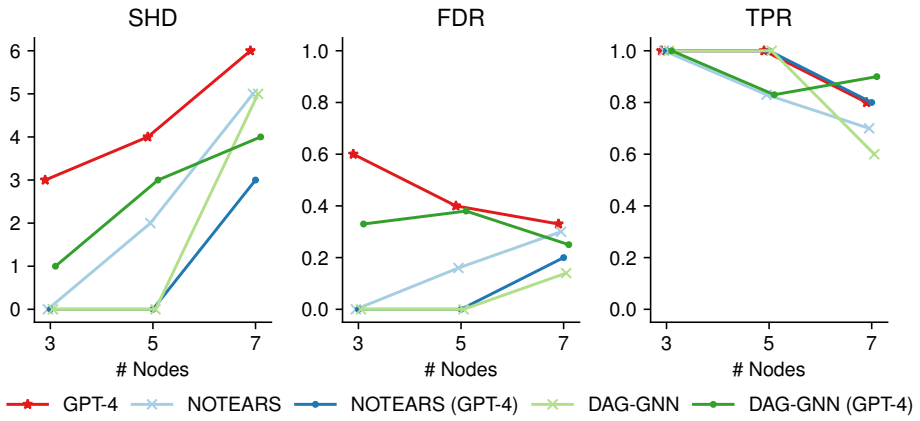



Figure 17: SHD, FDR, and TPR of NOTEARS and CGNN on the physical knowledge-based synthetic datasets with and without PLM prior.

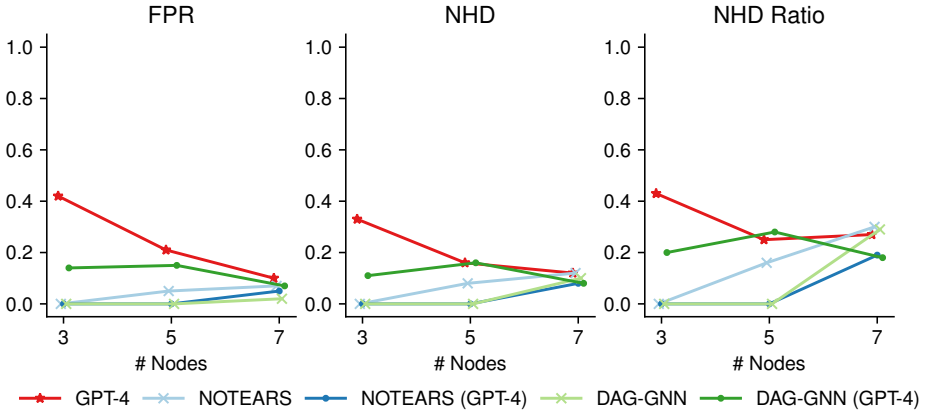
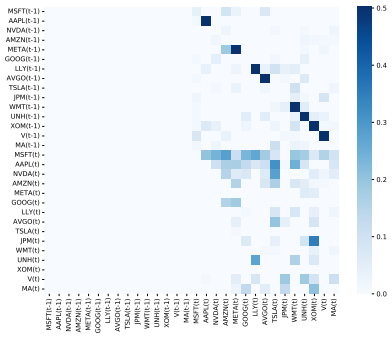


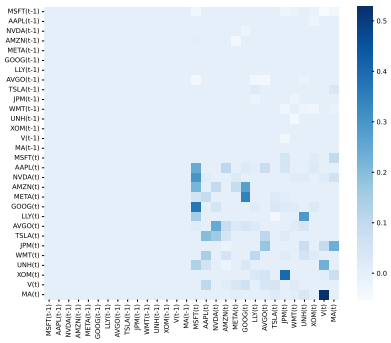

Figure 18: FPR, NHD, NHD Ratio of comparison on the physical knowledge-based synthetic datasets.

Table 12: Causal graph revision experiment using the Ban et al. [2023] revision prompt.

| | 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 |



(a) Heatmap of S&P 100 by NTS-NOTEARS



(b) Heatmap of S&P 100 by DYNOTEARS

Figure 19: Heatmaps of S&P 100 showing the correlation coefficients among companies.