Sequential Causal Discovery with Noisy Language Model Priors

Anonymous authors
Paper under double-blind review

Abstract

Causal discovery from observational data typically assumes access to complete data and availability of perfect domain experts. In practice, data often arrive in batches, are subject to sampling bias, and expert knowledge is scarce. Language Models (LMs) offer a surrogate for expert knowledge but suffer from hallucinations, inconsistencies, and bias. We present a hybrid framework that bridges these gaps by adaptively integrating sequential batch data with LM-derived noisy, expert knowledge while accounting for both *data-induced* and *LM-induced* biases. We propose a representation shift from Directed Acyclic Graph (DAG) to Partial Ancestral Graph (PAG), that accommodates ambiguities within a coherent framework, allowing grounding the *global* LM knowledge in *local* observational data. To guide LM interactions, we use a sequential optimization scheme that adaptively queries the most informative edges. Across varied datasets and LMs, we outperform prior work in structural accuracy and extend to parameter estimation, showing robustness to LM noise.

1 Introduction

Inference of causal relations from observational data remains a challenge in applications across healthcare, economics, business, and scientific discovery (Sanchez et al., 2022; Tu et al., 2019; Sadeghi et al., 2023; Ebert-Uphoff & Deng, 2012). The challenge is addressed through a dual approach: applying causal learning algorithms to observational data while incorporating domain expertise to resolve structural uncertainties (Spirtes et al., 2000; Neapolitan et al., 2004; Spirtes & Zhang, 2016; Chickering, 2002). However, domain expertise can be a scarce resource (He & Geng, 2008; Choo & Shiragur, 2023; Mooij et al., 2016; Meek, 2013; Constantinou et al., 2023). Advanced Language Models (LMs) create opportunities to explore their potential as surrogate experts for causal discovery (Kiciman et al., 2024; Willig et al., 2022). LMs generate informative priors or contraints (Takayama et al., 2025; Long et al., 2022; Ban et al., 2023b), improving accuracy when combined with data-driven algorithms. Yet, LMs pose their own challenges: hallucination, inconsistency, or failure to capture context-specific nuances (Ji et al., 2023; Kiciman et al., 2024).

The challenges compound since in common applications, observational data arrive batch-wise at a cadence, instead of as a complete dataset. Examples include web and app metrics of all online firms, where, for example, data could arrive weekly. Privacy regulations and storage constraints may further restrict data access to a short look-back window. A given week's (batch) data may not be representative of the overall distribution, which we assume remains stationary. This kind of arrival introduces *data-induced* bias, since the non-random draw of a batch suffer from sample selection bias (hereafter, selection bias Spirtes et al. (1995)) that distorts causal discovery. Separately, use of LMs, including *large* ones, poses two problems: *(i)* As surrogates for domain expertise, LMs introduce an *LM-induced* bias—their responses in terms of informative causal priors are prone to hallucinations, contextual brittleness, and inconsistency (Ji et al., 2023; Kiciman et al., 2024). *(ii)* The *global* knowledge encoded in LMs may not align with domain-specific *local* patterns emblematic of batch-wise data, leading to potentially biased learning.

Inattention to the dual biases—data-induced and LM-induced—is a key gap in current approaches to causal discovery with LMs, which we address. First, we propose a change in representation shift from a Directed Acyclic Graph (DAG), which LM-augmented causal discovery methods currently use, to a Partial Ancestral

Table 1: **LMs are overly optimistic:** LM based DAG-Pairwise and DAG-triplet prompting methods achieve high recall with low precision across temperatures on two common causal discovery datasets. This limitation calls for explicitly modeling data and LM biases. SHD=Structural Hamming Distance.

Dataset	Temp.	Method	GPT-3.5$_{\text{turbo}}$			GPT-4o		
			SHD ↓	Precision ↑	Recall ↑	SHD ↓	Precision ↑	Recall ↑
Earthquake	0.0	Pairwise	3.0±0.0	0.57±0.00	1.0±0.0	2.0±0.0	0.67±0.00	1.0±0.0
		Triplet	2.1±0.3	0.66±0.03	1.0±0.0	4.8±0.4	0.46±0.02	1.0±0.0
	0.5	Pairwise	2.8±0.6	0.59±0.05	1.0±0.0	2.2±0.4	0.65±0.04	1.0±0.0
		Triplet	2.1±0.3	0.66±0.03	1.0±0.0	4.4±0.5	0.48±0.03	1.0±0.0
	1.0	Pairwise	3.2±0.6	0.56±0.05	1.0±0.0	1.8±0.4	0.69±0.05	1.0±0.0
		Triplet	2.0±0.6	0.67±0.07	1.0±0.0	6.2±0.8	0.39±0.03	1.0±0.0
Asia	0.0	Pairwise	25.2±0.4	0.24±0.00	1.0±0.0	11.2±0.4	0.42±0.01	1.0±0.0
		Triplet	24.5±0.5	0.25±0.00	1.0±0.0	19.0±0.6	0.30±0.01	1.0±0.0
	0.5	Pairwise	23.4±1.0	0.26±0.01	1.0±0.0	11.6±0.5	0.48±0.01	1.0±0.0
		Triplet	24.0±1.0	0.25±0.01	1.0±0.0	19.2±0.8	0.29±0.01	1.0±0.0
	1.0	Pairwise	23.2±1.5	0.25±0.02	0.9±0.1	11.8±0.4	0.40±0.01	1.0±0.0
		Triplet	23.3±1.0	0.26±0.01	1.0±0.0	26.2±1.7	0.23±0.01	1.0±0.0

Overly optimistic behavior of the LM experts lead to high recall and low precision.

✗ No *global* causal structure
✗ No grounding to *local* data
✗ Need heuristics to postprocess

Graph (PAG), to accommodate uncertainty in the causal structure arising from the dual biases. Second, we propose a novel Bayesian approach to causal structure discovery, where beliefs over causal structure are updated with new data-batch, while augmenting noisy LM-knowledge as priors In support of PAG, we show that popular methods of pairwise and triplet prompting are *overly-optimistic* (*cf.* Table 1) and generate unreliable causal structure in the form of a DAG.

We introduce NLPSCM (Noisy Languge Prior in Sequential Causal Modeling); see Fig. 1 for an overview of the framework. NLPSCM differs from existing methods that either rely solely on access to the complete observational data or treat LMs as primary discovery mechanism. In a novelty, the causal structure discovery itself is Bayesian-inspired in that the beliefs about causal structure from data are updated iteratively by information drawn from an LM, as data arrive in batches. That is, we adopt a *data-first* approach, where for each batch, a traditional causal discovery algorithm, such as FCI from Spirtes et al. (1995), constructs an initial PAG conditioned on the background knowledge, which is then iteratively refined through optimized LM queries that leverage the global knowledge while remaining grounded in observed data. To maximize performance under limited budget, LM interactions are framed as a *sequential optimization* problem, selecting the most important edges to query, while accumulating background knowledge over batches. Moreover, to complete the causal discovery process, the parameter (edge weights) estimation we propose is *also* Bayesian, which incorporates potentially noisy LM priors on latent confounders and causal relationships. Our method uncovers cross-sectional causal relationships within each batch, rather than modeling temporal dependencies. NLPSCM applies especially to situations where studying contemporaneous causal relations among metrics is separated from confounding temporal effects.

Contributions We summarize the contributions as follows: *(i)* We propose a representation shift from DAGs to PAGs, in a hybrid setup of batch, observational data and LM as noisy expert, that inherently captures uncertainty in causal structure learning. *(ii)* A Bayesian-inspired algorithm for *causal structure* discovery with sequential batch data treating LMs as noisy experts thus accounting for dual sources of bias. *(iii)* A *sequential optimization* strategy for selecting maximally informative LM edge queries under fixed LM budget constraints. *(iv)* A Bayesian *parameter estimation* algorithm that robustly integrates noisy LM priors with batched data. Taken together, these constitute NLPSCM—an end-to-end causal discovery framework that jointly addresses both *causal structure learning* and *parameter estimation*.

2 Literature Review

Traditional causal discovery Traditional causal discovery aims to recover the underlying causal structure from observational data by exploiting statistical dependencies, often formalized through graphical models such as DAGs and PAGs (Spirtes & Zhang, 2016; Pearl, 2009; Neapolitan et al., 2004). These approaches include constraint-based methods (*e.g.,* PC from Spirtes et al. (2000), FCI (Spirtes et al., 1995; Zhang, 2008), RFCI from Colombo et al. (2012)), score-based methods (*e.g.,* GES from Chickering (2002)), and functional causal models (*e.g.,* LiNGAM (Shimizu et al., 2006; 2011)), typically assuming causal sufficiency and faithfulness

(Spirtes et al., 2000; Zhang & Spirtes, 2012). They rely on conditional independence tests or likelihood-based scoring to causal relationships (Shah & Peters, 2020; Zhang et al., 2011; Peters et al., 2017; Glymour et al., 2019). However, these approaches often struggle under data scarcity, presence of latent confounders (Spirtes et al., 2000; Monti et al., 2020), or domain-specific constraints not captured by statistical patterns (Mooij et al., 2016; Peters et al., 2014). These limitations are addressed by hybrid methods incorporating external domain knowledge (Meek, 1995; Heckerman et al., 1995; Ogarrio et al., 2016) from human experts, refining causal graphs with new variables, modifying edge orientations, or resolving equivalence classes (Brouillard et al., 2020; Wang et al., 2017; Constantinou et al., 2023; Ban et al., 2023b). This helps restrict the search space and improves identifiability, particularly when data is limited, as shown by Wallace et al. (1996). The evolution from purely statistical methods to knowledge-augmented approaches fuels advanced machine learning techniques to enhance causal discovery (Glymour et al., 2019; Schölkopf et al., 2021). Also, as noted by Baldi & Shahbaba (2020), causal research distinguishes general *vs.* local settings and applies to diverse fields (Andrade & Zachariadis, 2016; Bilal & Känzig, 2024; Geist & Lambin, 2002; Kelly et al., 2011; McKinney et al., 2016; Mathers et al., 2009).

Causal discovery extends to streaming data of networks, stock markets, and sensor systems. Causal Bayesian learning and causal discovery with progressively streaming features are well studied (Darvish Rouhani et al., 2018; Yu et al., 2010; You et al., 2023; Li et al., 2021; You et al., 2021; Yu et al., 2010). Instead, we focus on *contemporaneous* relationships among a fixed set of variables, which yield data across batches. In batch-wise experimentation in medicine and A/B testing, Bridgeford et al. (2021) examines associations between batches, with batch effects as causal effects, while Zhang & Yuan (2024) examines adaptive batch-wise intervention. We confine to sequential, batch-wise observational data.

LM-augmented causal structure discovery LM augmented causal discovery relies on LM's world knowledge and includes pairwise prompting (Willig et al., 2022; Long et al., 2022; Kiciman et al., 2024; Jin et al., 2024; Long et al., 2023), and triplet-based prompting incorporating voting proposed by Vashishtha et al. (2025). Hybrid frameworks integrate LM-generated insights in constraint-based methods or inform score-based approaches (Ban et al., 2023b; Takayama et al., 2025). Reliability of LM-derived constraints is sensitive to domain specificity and prompt framing (Kiciman et al., 2024; Ji et al., 2023). Ban et al. (2023a) proposes ILS-CSL involving iterative refinement of causal graphs by alternating between LM reasoning and statistical verification. Augmenting LM prompts with correlation matrices or statistical summaries is explored (Jiralerspong et al., 2024a; Susanti & Färber, 2025). Yet, LM-aided causal discovery shows inconsistent judgments across prompting strategies, difficulties with grounding, and biases. Recently, da Silva et al. (2025) proposed the BFS method. It is designed for full dataset while our method is anchored in sequential batched data. Additionally, its update is around information gain and does not perform edge-weight estimation.

Deviating from the prior art, we present a Bayesian framework for causal structure discovery, where data arrive sequentially in batches, and we recognize dual uncertainties—arising from limited observational data and noisy LM responses. We depart from DAG-centric discovery to adopt the PAG, a more robust representation for evolving, uncertain causal structures. We iteratively refine the PAG across batches, while framing LM queries as a sequential optimization problem.

Parameter estimation in SEMs While existing hybrid methods that combine observational data with language models (LMs) primarily focus on causal structure discovery, they typically stop short of estimating causal effect parameters, which we refer to as *parameter estimation*. In contrast, prior work on parameter estimation is predominantly framed within structural equation models (SEMs) and generally assumes either a known structure or purely data-driven priors.

Within the SEM framework, under assumptions of linearity and Gaussian noise (Bollen, 1989; Shimizu et al., 2006), classical approaches estimate parameters using maximum likelihood or two-stage least squares method, as described in Pearl (2009). Learning of SEM parameters extends to nonlinear or nonparametric settings using neural networks or Gaussian processes, enabling flexible modeling of complex dependencies while retaining causal interpretability (Zheng et al., 2020; Lachapelle et al., 2020), and helps in integrating expert knowledge with data-driven estimation(Peters et al., 2017; Schölkopf et al., 2021). However, the use of LM-derived priors and correlations for SEM parameter estimation remains largely unexplored.

To bridge this gap and provide an end-to-end causal discovery framework, we jointly perform causal structure discovery and SEM parameter estimation. The latter constitutes a key contribution of our work: we propose

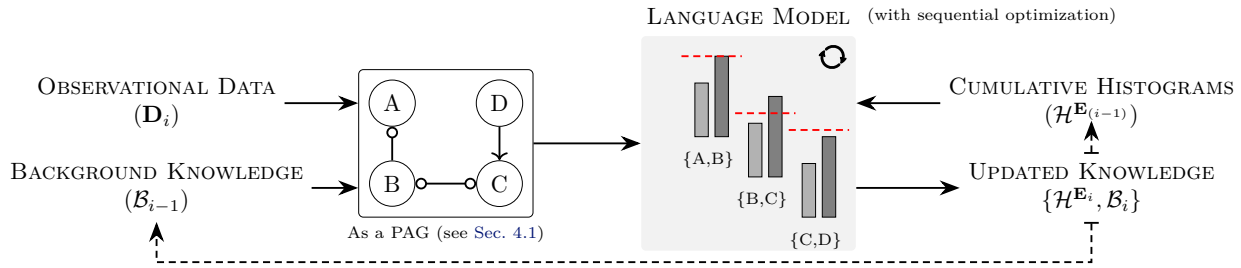


Figure 1: **Overview of the NLPSCM framework:** At each batch $\mathbf{D}_i$, observational data is combined with accumulated background knowledge $\mathcal{B}_{(i-1)}$ as prior to estimate a PAG structure. A language model is then queried—under sequential optimization—to produce beliefs over possible causal relations and update $\mathcal{B}_i$. The updated $\{\mathcal{H}^{\mathbf{E}_i}, \mathcal{B}_i\}$ are fed back into the next iteration.

a principled parameter estimation procedure that integrates LM-derived noisy priors into SEMs, yielding consistent estimators of causal strengths despite misspecification of priors.

3 Problem Setup: Sequential Causal Discovery with LMs

Traditional causal discovery methods uncover causal structure by exploiting statistical dependencies in observational data, typically assuming access to the complete dataset, and reliable domain knowledge. In contrast, we focus on the setting of sequential, batch-wise observational data. This setting introduces dual sources of bias: *(i)* potentially biased and limited batched observational data, and *(ii)* noisy LM responses. Below we introduce the notation and the problem setup.

Problem statement Given sequential batches of observational data subject to selection bias, we consider the problem of: *(i)* causal discovery in the presence of latent confounders, *(ii)* incorporating noisy priors provided by LMs, and *(iii)* parameter estimation of the inferred causal structure. We now define the problem formally.

Notation and setup We define batches of observational data $\mathcal{D} = \{\mathbf{D}_i\}_{i=0}^{N}$, where each ith batch is a sample from the same underlying *'true'* distribution, $\mathbf{D}_i \sim \mathbb{D}$. Each batch contains same set of observed variables, $\mathbf{V}^{\text{O}}$, with $\mathbf{D}_i \in \mathbb{R}^{n_i \times d}$, where n_i is the number of data points, varies by batch, and $d = |\mathbf{V}^{\text{O}}|$ is the number of observed variables. $\mathbb{R}$ denotes real numbers. Any categorical value for an observed variable is encoded as a numerical value. All notations are succinctly shown in Table A1.

For each batch $\mathbf{D}_i$, a causal graph $\mathcal{G}^{\text{D}_\text{i}}$=($\mathbf{V}^{\text{O}}, \mathbf{E}^{\text{D}_\text{i}}$) is inferred using a standard causal discovery algorithm, where $\mathcal{G}^{\text{D}_\text{i}}$ is a PAG with $\mathbf{V}^{\text{O}}$ nodes and $\mathbf{E}^{\text{D}_\text{i}}$ edges. We assume there exists a true causal graph $\mathcal{G}$=($\mathbf{V}, \mathbf{E}$), where $\mathbf{V}^{\text{O}} \subseteq \mathbf{V}$ and $\mathbf{V}$=$\{\mathbf{V}^{\text{O}}, \mathbf{V}^{\text{L}}\}$ represents all the variables, both observed and unobserved latent, and $\mathbf{E}$ represents the true causal relationships. We assume that each confounder may affect two observed variables.

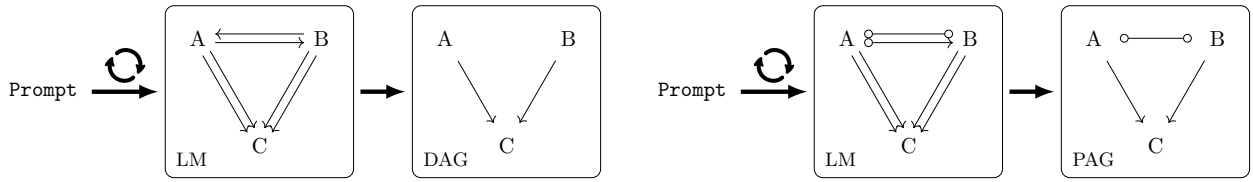
LM-augmented sequential causal discovery Traditional hybrid approaches rely on domain experts to narrow the gap between the inferred causal graph $\mathcal{G}^{\text{D}_\text{i}}$ and the true causal graph $\mathcal{G}$ by either introducing knowledge about unobserved variables, effectively reducing the set $\mathbf{V} \setminus \mathbf{V}^{\text{O}}$, and adding, removing, reorienting edges in the inferred graph, aiming to minimize the difference $\mathbf{E} \setminus \mathbf{E}^{\text{D}_\text{i}}$.

NLPSCM's hybrid approach extends the prior art in two directions: *(i)* using an LM as a noisy expert to improve the causal structure $\mathcal{G}^{\text{D}_\text{i}}$, obtained via known causal discovery algorithm, where *(ii)* data arrive sequentially in batches. The LM (noisy e**X**pert), represented X^i, helps in reducing the Markov equivalence class to yield causal structure ($\mathcal{G}^{\text{X}_\text{i}}$). Formally,

$$f_{\text{CD}} : \mathbf{D}_i \rightarrow \mathcal{G}^{\text{D}_\text{i}} \; ; \; f_{\text{LM}} : \mathcal{G}^{\text{D}_\text{i}} \rightarrow \mathcal{G}^{\text{X}_\text{i}} \, , \tag{1}$$

where $\mathcal{G}^{\text{D}_\text{i}} = (\mathbf{V}^{\text{O}}, \mathbf{E}^{\text{D}_\text{i}})$ and $\mathcal{G}^{\text{X}_\text{i}} = (\mathbf{V}^{\text{X}_\text{i}}, \mathbf{E}^{\text{X}_\text{i}})$. The aim is to reduce the discrepancy between the inferred graph and the true causal structure through LM expertise,

$$\mathcal{G} \setminus \mathcal{G}^{\text{X}_\text{i}} \leq \mathcal{G} \setminus \mathcal{G}^{\text{D}_\text{i}} \; ; \; \mathbf{V} \setminus \mathbf{V}^{\text{X}_\text{i}} \leq \mathbf{V} \setminus \mathbf{V}^{\text{O}} \; ; \; \mathbf{E} \setminus \mathbf{E}^{\text{X}_\text{i}} \leq \mathbf{E} \setminus \mathbf{E}^{\text{D}_\text{i}} \, . \tag{2}$$



(a) Conventional DAG-LM (b) PAG-LM in NLPSCM

Figure 2: **PAG-LM** streamlines how LMs compose the graph and allows for ambiguities to be indicated in the structure. *(a)* DAG is constructed by iterative prompting (🔄) leading to ambiguities (*e.g.,* $\leftrightarrows$) requiring heuristics that cannot be represented. *(b)* NLPSCM represents the causal structure as a PAG that implicitly allow ambiguities to be represented providing a richer representation (*e.g.,* $\circ\!\!-\!\!\circ$)

The $\mathbf{V}^{\mathrm{X_i}} \setminus \mathbf{V}^{\mathrm{O}}$ is the set of *LM-suggested* variables while $\mathbf{E}^{\mathrm{X_i}} \setminus \mathbf{E}^{\mathrm{D_i}}$ is the set of *LM-suggested* edges. To integrate the LM's noisy responses and address the inherent bias in batch data, our Bayesian causal discovery framework explicitly handles both *data-induced* and *LM-induced* biases.

Parameter estimation Once the LM-augmented causal structure $\mathcal{G}^{\mathrm{X_i}}$ is obtained, we focus on estimating the parameters of the structure equation model (SEM) *i.e.* the edge weights $\boldsymbol{\theta}$ and the noise variance σ^2; $\boldsymbol{\phi} = \{\boldsymbol{\theta}, \sigma^2\}$. A straightforward method to learn the parameters $\boldsymbol{\phi}$ is Maximum Likelihood Estimation (MLE), $\nabla_{\boldsymbol{\phi}} \log p(\mathbf{D}_i \,|\, \mathcal{G}^{\mathrm{X_i}}, \boldsymbol{\phi})$. However, a critical limitation appears in the presence of unobserved *expert-suggested* variables ($\mathbf{V}^{\mathrm{L}}$) in the augmented causal structure, as the likelihood becomes intractable. We address this by proposing a Bayesian parameter estimation algorithm that incorporates the *expert-suggested* information about the unobserved (latent) variable(s), $\mathbf{V}^{\mathrm{L}}$.

4 Sequential Causal Structure Discovery with PAGs

The promise of LMs as proxies for domain expert in causal structure discovery, as studied in prior art, typically queried via pairwise or triplet prompts, faces key limitations: *(i)* LMs may provide responses regardless of causal relations among other variables, resulting in inconsistent or cyclic causal graphs; *(ii)* LMs may hallucinate; *(iii)* LMs may be overly optimistic, predicting spurious causal relationships with high recall but low precision (*cf.* Table 1). While prior art addresses *(i)* and *(ii)* through heuristics or auxiliary models to enforce acyclicity and consistency, they add complexity and potentially degrading performance. Crucially, *(ii)* and *(iii)* remain largely unexplored in the setting of sequential batch data where two distinct sources of bias emerge: *LM-induced* bias and *data-induced* bias.

4.1 PAG to Incorporate Uncertainty

Prompting methods (pairwise or triplet) constrain the response format to—*'causal', 'non-causal',* or *'unknown'*—which prevent LMs from expressing uncertainty, thus exacerbating bias and inconsistency. To address this, we propose a representational shift from DAG to PAG when using LMs as proxy for experts. PAGs encode uncertainty and structural ambiguity in a principled manner, accommodating latent confounding and partial orientation (*cf.* Fig. 2). Formally, we expand the limited edge set of DAG, $\mathbf{E}^{\mathrm{DAG}}$=$\{\rightarrow, \leftarrow, \cdot\}$, to the richer set for PAG, $\mathbf{E}^{\mathrm{PAG}}$=$\{\rightarrow, \leftarrow, \leftrightarrow, \circ\!\!\rightarrow, \leftarrow\!\!\circ, \circ\!\!-\!\!\circ, \cdot, -\}$ [1], where $\cdot$ represents no causal relation. The expansion allows LM to select

Table 2: **DAG to PAG:** Structural Intervention Distance (SID) between DAG-Pairwise (Voting) and PAG-Pairwise depicts benefit of PAG to represent inherent causal uncertainty.

Dataset	Temp.	Method	SID ↓
Earthquake	0.0	Pairwise (Voting)	$(1.0, 1.0)\pm(0.0, 0.0)$
		PAG-Pairwise	$(0.0, 0.0)\pm(0.0, 0.0)$
	0.5	Pairwise (Voting)	$(1.4, 1.4)\pm(1.2, 1.2)$
		PAG-Pairwise	$(0.0, 0.0)\pm(0.0, 0.0)$
	1.0	Pairwise (Voting)	$(1.6, 1.6)\pm(1.5, 1.5)$
		PAG-Pairwise	$(0.8, 0.8)\pm(1.6, 1.6)$
Asia	0.0	Pairwise (Voting)	$(6.4, 6.4)\pm(0.8, 0.8)$
		PAG-Pairwise	$(3.8, 3.8)\pm(0.9, 0.9)$
	0.5	Pairwise (Voting)	$(5.2, 5.2)\pm(1.7, 1.7)$
		PAG-Pairwise	$(3.5, 3.5)\pm(0.8, 0.8)$
	1.0	Pairwise (Voting)	$(5.4, 5.4)\pm(1.8, 1.8)$
		PAG-Pairwise	$(4.0, 4.0)\pm(1.9, 1.9)$

[1] To capture selection bias, ambiguous edges ($\circ\!\!-\!\!\circ, \circ\!\!\rightarrow$) and undirected edges '$-$' can be used. Not all causal discovery algorithms output undirected edges (*e.g.,* https://causal-learn.readthedocs.io/en/latest/search_methods_index/Constraint-basedcausaldiscoverymethods/FCI.html), but they output ambiguous edges, which we use as proxy for selection bias.

Table 3: **Semantic Entropy in NLPSCM.** We report the mean Shannon entropy of empirical histograms over PAG edge predictions (mean ± std over 5 runs) across sequential batches for User Level Data–I and User Level Data–II. The observed reduction in entropy indicates decreasing predictive uncertainty over PAG edge types as structural context is progressively incorporated.

Dataset	Batch-1	Batch-2	Batch-3	Batch-4	Batch-5	Batch-6	Batch-7
User Level Data - I	0.48 ± 0.06	0.50 ± 0.05	0.36 ± 0.08	0.17 ± 0.13	0.10 ± 0.06	0.12 ± 0.06	0.09 ± 0.05
User Level Data - II	0.31 ± 0.25	0.27 ± 0.22	0.18 ± 0.18	0.12 ± 0.19	0.07 ± 0.15	0.00 ± 0.00	0.00 ± 0.00

from more options and improve causal discovery (*cf.* Table 2). In PAG-pairwise, the LM is queried per variable pair to select from $\mathbf{E}^{\text{PAG}}$ (see Sec. F.2 for the full prompt).

The representational shift to PAG is a necessary pivot and starting point for addressing dual sources of bias—*LM-induced* and *data-induced*. The LM's error prone response constitutes a prior and motivates a novel Bayesian causal structure discovery framework in the sequential batch setting, whereby we integrate causal predictions from observational data and LMs in a sequential, iterative manner. Building on this key observation, we introduce the causal structure learning algorithm of NLPSCM and then on the inferred structure introduce the parameter estimation algorithm.

4.2 LM-augmented Causal Structure Discovery

As a principled way to incorporate the dual sources of bias, we adopt a Bayesian-inspired formulation. Formally, given data batch $\mathbf{D}_i$, cumulative background knowledge $\mathcal{B}_{(i-1)}$ up to batch $(i-1)$, the posterior over $\mathcal{G}_i$ is obtained via Bayes' rules,

$$\underbrace{p(\mathcal{G}_i \mid \mathbf{D}_i, \mathcal{B}_{(i-1)})}_{\text{Posterior}} \propto \underbrace{p(\mathbf{D}_i \mid \mathcal{G}_i)}_{\text{Likelihood}} \underbrace{p(\mathcal{G}_i \mid \mathcal{B}_{(i-1)})}_{\text{Prior}} . \tag{3}$$

The prior $p(\mathcal{G}_i \mid \mathcal{B}_{(i-1)})$ is iteratively shaped by background knowledge, which accumulates across batches. Intuitively, the prior encodes the edge type between node pairs in the causal structure. Once the posterior is obtained, an LM is queried to update and obtain background knowledge $\mathcal{B}_i$ using Eq. (5) and Eq. (9), which shapes the prior for the next batch. We note that this formulation serves as a conceptual motivation for the design of NLPSCM; the implementation maintains uncertainty at the edge level (Eqs. (4) and (5)) rather than computing a full posterior over graph space. Next we discuss the formulation.

Given the inherent stochasticity of LMs, its response can be viewed as a *sample* from an implicit distribution over the edge types ($\mathbf{E}^{\text{PAG}}$). This allows explicit modeling of uncertainty in LM responses, $f_{\text{LM}}^{(i)}(A, B, \text{Pr}) \sim p(E_{AB} \mid A, B)$, where A, B are the two nodes, Pr is the prompt, and $E_{AB} \in \mathbf{E}^{\text{PAG}}$. Treating LM responses as noisy observations rather than ground truth, addresses the challenge of LM hallucination. As more batches are processed, the prompt Pr becomes more informative, thereby decreasing uncertainty in LM responses. Formally, we treat LM as a *black-box causal edge sampler* and aggregate multiple LM samples into empirical histograms that are updated iteratively over batches,

$$\mathcal{H}^{\mathbf{E}_i}(A, B)[E_{AB}] = \mathcal{H}^{\mathbf{E}_{i-1}}(A, B)[E_{AB}] + \mathbb{I}[f_{\text{LM}}^{(i)}(A, B) = E_{AB}] , \tag{4}$$

where $\mathcal{H}^{\mathbf{E}_i}(A, B)$ represents the cumulative histogram up to batch i, and E_{AB} represents a type of causal relation. These histograms define an approximate posterior distribution over edge types, capturing the LM's evolving beliefs about causal relationships.

To determine when accumulated LM evidence is sufficient to promote an edge to background knowledge, we define a dynamic threshold that balances distributional uncertainty and sampling uncertainty (motivated by the explore-exploit trade-off formalized in Sec. 4.3),

$$\tau_i^e = \alpha \times E_i^e \times T_i^e + (1-\alpha)\sqrt{T_i^e \left(1 - \frac{T_i^e}{T_i}\right)}, \quad \text{s.t.} \quad E_i^e = -\sum_j \frac{\mathcal{H}_{j,e}^{\mathbf{E}_i}}{T_i^e} \times \log\left(\frac{\mathcal{H}_{j,e}^{\mathbf{E}_i}}{T_i^e}\right). \tag{5}$$

Algorithm 1 LM-augmented structure learning

Require: $\mathbf{D}_i, \mathcal{H}^{\mathbf{E}_{(i-1)}}, \mathcal{H}^{\mathrm{L}_{(i-1)}}, \mathcal{B}_{(i-1)}, m^{\mathbf{E}}, m^{\mathrm{L}},$
Ensure: $\mathcal{H}^{\mathbf{E}_i}, \mathcal{H}^{\mathrm{L}_i}, \mathcal{B}_i$
1: **Initial causal structure**
2: $f_{\mathrm{CD}} : \mathbf{D}_i \times \mathcal{B}_{(i-1)} \rightarrow \mathcal{G}^{\mathrm{D_i}}$
3: **Expert-guided causal structure refinement**
4: $f_{\mathrm{LM}} : \mathcal{G}^{\mathrm{D_i}} \times \mathcal{H}^{\mathbf{E}_{(i-1)}} \times \mathcal{B}_{(i-1)} \times m^{\mathbf{E}} \rightarrow (\mathcal{H}^{\mathbf{E}_i}, \mathcal{B}_i)$
5: **Expert-suggested latent confounder**
6: **for** $A \leftrightarrow B$ in $\mathcal{B}_i$ **do**
7: $\quad f_{\mathrm{LM}} : \mathcal{H}^{\mathbf{E}_i} \times \mathcal{H}^{\mathrm{L}_{(i-1)}} \times A \times B \times m^{\mathrm{L}} \rightarrow \mathcal{H}^{\mathrm{L}_i}$
8: **end for**
9: **return** $\mathcal{H}^{\mathbf{E}_i}, \mathcal{H}^{\mathrm{L}_i}, \mathcal{B}_i$

Notation:
$m^{\mathbf{E}}, m^{\mathrm{L}}$: LM budget for edges and confounder
$\mathcal{H}^{\mathbf{E}}, \mathcal{H}^{\mathrm{L}}$: histograms for edges and confounders
$\mathcal{I}^{\mathrm{P}}, \mathcal{I}^{\rho}$: Prompt for prior and corelation

Algorithm 2 Bayesian Parameter Estimation

Require: $\mathbf{D}_i$, η, $\mathcal{G}^{\mathrm{X_i}}, \mathcal{I}^{\mathrm{P}}, \mathcal{I}^{\rho}$
Ensure: $\boldsymbol{\phi}$
1: **Initialization**
2: Warm-start for $\boldsymbol{\phi}^{\mathrm{O}}$
$\quad \boldsymbol{\phi}^{\mathrm{O}} \in \arg\max_{\boldsymbol{\phi}^{\mathrm{O}}} p(\mathbf{D}_i \,|\, \mathcal{G}^{\mathrm{X_i}}_{-\mathbf{V}^{\mathrm{L}}}, \boldsymbol{\phi}^{\mathrm{O}})$
3: Get prior over $\mathbf{V}^{\mathrm{L}}$
$\quad f_{\mathrm{LM}} : \mathcal{G}^{\mathrm{X_i}} \times \mathbf{V}^{\mathrm{L}} \times \mathcal{I}^{\mathrm{P}} \rightarrow \mathcal{N}(\boldsymbol{m}_p, \boldsymbol{S}_p)$
4: Get correlation $\rho(\mathbf{V}^{\mathrm{L}}, \mathbf{V}^{\mathrm{O}})$
$\quad f_{\mathrm{LM}} : \mathcal{G}^{\mathrm{X_i}} \times \mathcal{I}^{\rho} \rightarrow \rho(\mathbf{V}^{\mathrm{L}}, \mathbf{V}^{\mathrm{O}})$
5: Initialize $\boldsymbol{\theta}^{\mathrm{L}}$ to $\rho(\mathbf{V}^{\mathrm{L}}, \mathbf{V}^{\mathrm{O}})$ or randomly
6: **Iterative Optimization**
7: **while** not converged **do**
8: $\quad$ Compute posterior over $\mathbf{V}^{\mathrm{L}}$ using Eq. (10)
9: $\quad$ Optimize $\boldsymbol{\phi}$ variables using Eq. (11)
10: **end while**
11: **return** $\boldsymbol{\phi}$

Here, for batch i and edge e, τ_i^e denotes the threshold, E_i^e the posterior entropy, T_i^e the number of LM interactions, $T_i = \sum_e T_i^e$ the total interactions, and $\mathcal{H}_{j,e}^{\mathbf{E}_i}$ the frequency of bin j in e's histogram.

Intuitively, the first term in τ_i^e of Eq. (5) accounts for uncertainty in the histogram edge distribution, while the second term handles the sampling uncertainty that decreases as more batches of data arrive. The hyperparameter α balances between these two terms. Fig. 4 showcases the efficiency of the proposed dynamic background threshold (*cf.* Eq. (5)), with the additional details discussed in Sec. 6.1. The pseudo-code of the algorithm is outlined in Alg. 1.

A key property of our sequential framework is that the prompt Pr becomes progressively more informative across batches as $\mathcal{G}^{\mathrm{X_i}}$ expands, providing increasingly rich structural context to the LM. This corresponds to conditioning on accumulated causal constraints, and therefore the model's predictive uncertainty is expected to decrease over iterations. We empirically validate this behavior by reporting the mean histogram entropy across batches for the two real world datasets User Level Data–I and User Level Data–II in Table 3, and observe a consistent reduction as the graph is incrementally refined.

From an uncertainty modeling perspective, our histogram-based formulation is closely related to recent work on semantic entropy in language models (Farquhar et al., 2024; Nikitin et al., 2024; Kuhn et al., 2023). However, in our setting, NLPSCM restricts the LM output space to the finite set of PAG edge types $\mathbf{E}^{\mathrm{PAG}}$, yielding purely categorical predictions. Consequently, semantic entropy can be computed directly as the Shannon entropy of empirical histograms over sampled edge predictions, without requiring embedding-based clustering. While conceptually aligned with prior formulations of semantic entropy, this arises naturally in our setting from the discrete decision space induced by PAG edge selection.

4.3 LM Interaction: Sequential Optimization

Given the stochastic LM responses and cumulative histogram-based estimates of edge uncertainty, we next address how to allocate LM queries efficiently across candidate edges. We model LM interactions f_{LM} as a *sequential optimization* problem under a limited budget. At each batch i, up to m^{L} calls to the LM are allowed. The objective is to strategically allocate these calls to refine the edge distribution $\mathcal{H}^{\mathbf{E}_i}$ and expand the set of background knowledge $\mathcal{B}_i$.

In the edge refinement setting, each query corresponds to selecting a candidate edge e and querying f_{LM} to reduce uncertainty about its type. The LM's response is treated as a noisy sample from the underlying distribution over edge types. This induces a natural trade-off: we must *explore* uncertain edges to improve estimates and *exploit* promising edges that are likely to yield useful and increasing background knowledge.

Table 4: **NLPSCM improves causal discovery:** We experiment with six datasets–number of observed variables ranges 5 (small) to 37 (large)– using two paradigms: *Only-Data* and *Data-LM*. We evaluate with 5 metrics: *Modified SHD, SID, Precision, Recall, F1*. All methods use GPT-3.5$_{\text{turbo}}$ as an LM with temperature 1. We report mean and standard deviation over 5 runs and perform significance test with $\alpha = 0.05$.

Dataset	Approach	Method	Mod. SHD ↓	SID ↓	Precision ↑	Recall ↑	F1 Score ↑
Earthquake ($d = 5$)	Only-Data	FCI-Cumulative	2.00±0.00	(0.00, 5.00)±(0.00, 0.00)	**1.00±0.00**	0.50±0.00	0.67±0.00
		FCI-Vanilla	3.60±0.80	(8.20, 9.20)±(3.60, 1.60)	0.20±0.40	0.05±0.10	0.08±0.16
		FCI-Iterative	5.00±1.67	(12.20, 12.20)±(4.66, 4.66)	0.30±0.27	0.20±0.19	0.24±0.22
		FCI-Heuristics	3.60±0.80	(8.20, 9.20)±(3.60, 1.60)	0.20±0.40	0.05±0.10	0.08±0.16
	Data-LM	LLM-first	6.00±0.82	(15.00, 15.00)±(0.82, 0.82)	0.11±0.16	0.08±0.12	0.09±0.13
		ILS-CSL	2.38±0.96	(5.25, 6.50)±(0.83, 2.69)	**0.88±0.22**	0.44±0.21	0.56±0.21
		NLPSCM	**1.00±0.63**	**(2.20, 2.20)±(1.60, 1.60)**	**1.00±0.00**	**0.75±0.16**	**0.85±0.11**
Asia ($d = 8$)	Only-Data	FCI-Cumulative	7.00±0.00	(23.00, 49.00)±(0.00, 0.00)	0.00±0.00	0.00±0.00	0.00±0.00
		FCI-Vanilla	7.80±0.75	(30.00, 35.00)±(5.90, 2.45)	0.00±0.00	0.00±0.00	0.00±0.00
		FCI-Iterative	8.00±1.26	(33.00, 33.00)±(7.46, 7.46)	0.45±0.24	0.23±0.15	0.29±0.18
		FCI-Heuristics	7.80±0.75	(30.00, 35.00)±(5.90, 2.45)	0.00±0.00	0.00±0.00	0.00±0.00
	Data-LM	LLM-first	7.33±0.94	(27.67, 27.67)±(2.49, 2.49)	0.58±0.12	0.29±0.06	0.39±0.08
		ILS-CSL	6.50±0.50	(28.50, 28.50)±(3.20, 3.20)	**0.79±0.12**	0.28±0.11	0.40±0.12
		NLPSCM	**4.60±1.02**	**(13.60, 13.60)±(3.83, 3.83)**	**0.80±0.12**	**0.60±0.12**	**0.67±0.08**
User Level Data-I ($d = 9$)	Only-Data	FCI-Cumulative	15.00±2.77	(47.80, 47.80)±(4.62, 4.62)	0.72±0.18	0.37±0.09	0.47±0.09
		FCI-Vanilla	21.30±3.50	(61.60, 61.60)±(5.54, 5.54)	0.41±0.18	0.22±0.10	0.29±0.13
		FCI-Iterative	**5.60±1.20**	(23.40, 23.40)±(7.94, 7.94)	**0.93±0.04**	0.77±0.02	**0.84±0.03**
		FCI-Heuristics	21.30±3.50	(61.60, 61.60)±(5.54, 5.54)	0.41±0.18	0.22±0.10	0.29±0.13
	Data-LM	LLM-first	9.68±1.67	(38.12, 38.12)±(4.06, 4.06)	0.80±0.06	0.66±0.04	0.72±0.05
		ILS-CSL	9.20±1.72	(34.20, 34.20)±(3.49, 3.49)	0.82±0.05	0.66±0.05	0.73±0.05
		NLPSCM	**5.00±0.71**	**(13.75, 13.75)±(2.49, 2.49)**	**0.90±0.04**	**0.83±0.02**	**0.86±0.02**
User Level Data-II ($d = 8$)	Only-Data	FCI-Cumulative	19.80±2.04	**(40.00, 40.00)±(3.03, 3.03)**	0.15±0.07	0.10±0.04	0.12±0.05
		FCI-Vanilla	17.40±1.83	**(36.80, 39.40)±(2.04, 3.38)**	0.07±0.13	0.02±0.03	0.03±0.05
		FCI-Iterative	20.60±1.77	(42.80, 43.40)±(4.07, 4.22)	0.06±0.08	0.05±0.07	0.05±0.07
		FCI-Heuristics	17.40±1.83	**(36.80, 39.40)±(2.04, 3.38)**	0.07±0.13	0.02±0.03	0.03±0.05
	Data-LM	LLM-first	18.75±0.43	(44.00, 44.00)±(1.00, 1.00)	0.18±0.01	**0.17±0.00**	**0.18±0.00**
		ILS-CSL	**16.90±1.50**	(44.40, 47.80)±(6.05, 3.49)	0.16±0.03	0.10±0.04	0.12±0.03
		NLPSCM	**16.33±1.80**	**(40.17, 40.17)±(3.14, 3.14)**	**0.26±0.04**	**0.18±0.06**	**0.20±0.04**
Child ($d = 19$)	Only-Data	FCI-Cumulative	27.50±0.00	**(111.00, 131.00)±(0.00, 0.00)**	0.38±0.00	0.36±0.00	0.37±0.00
		FCI-Vanilla	28.00±1.48	(129.20, 133.20)±(10.46, 10.76)	0.38±0.04	0.26±0.05	0.31±0.04
		FCI-Iterative	32.10±1.16	(149.00, 164.40)±(7.16, 10.33)	0.27±0.03	0.26±0.04	0.26±0.03
		FCI-Heuristics	28.00±1.48	(129.20, 133.20)±(10.46, 10.76)	0.38±0.04	0.26±0.05	0.31±0.04
	Data - LLM	LLM-first	31.67±2.05	(172.00, 172.00)±(13.42, 13.42)	0.29±0.05	0.30±0.05	0.30±0.05
		ILS-CSL	32.00±2.10	(154.00, 154.00)±(26.53, 26.53)	0.28±0.05	0.28±0.05	0.28±0.05
		NLPSCM	**25.50±0.89**	**(103.40, 112.20)±(8.09, 7.05)**	**0.43±0.01**	**0.44±0.02**	**0.43±0.01**
Alarm ($d = 37$)	Only-Data	FCI-Cumulative	**45.00±0.00**	**(626.00, 626.00)±(0.00, 0.00)**	0.25±0.00	0.02±0.00	0.04±0.00
		FCI-Vanilla	49.50±1.61	(617.80, 699.20)±(29.23, 68.43)	0.00±0.00	0.00±0.00	0.00±0.00
		FCI-Iterative	52.40±6.21	**(612.20, 636.80)±(49.87, 43.83)**	**0.33±0.14**	0.12±0.05	0.17±0.07
		FCI-Heuristics	49.50±1.61	(617.80, 699.20)±(29.23, 68.43)	0.00±0.00	0.00±0.00	0.00±0.00
	Data - LLM	LLM-first	52.33±3.09	(673.33, 673.33)±(25.63, 25.63)	**0.33±0.08**	0.13±0.02	0.19±0.03
		ILS-CSL	51.20±1.60	(657.20, 657.20)±(24.07, 24.07)	0.34±0.05	0.10±0.01	0.15±0.02
		NLPSCM	50.90±1.32	**(589.80, 591.00)±(26.48, 25.59)**	**0.42±0.03**	**0.22±0.02**	**0.29±0.02**

Formally, we cast LM interactions as a sequential decision-making problem:

$$\textbf{Arms:} \quad \mathcal{A} = \{\text{All possible edges between variables, } \mathbf{E}^{\text{PAG}}\}, \tag{6}$$

$$\textbf{Reward:} \quad r_k(e) = \text{Information gain from querying edge } e \text{ at step } k, \tag{7}$$

$$\textbf{Policy:} \quad \pi : \mathcal{H}^{\mathbf{E}_i} \times \mathcal{G}^{\text{D}_\text{i}} \to e \quad \text{(Edge selection rule).} \tag{8}$$

The optimization objective is to maximize cumulative information gain over m^{L} LM calls, balancing both the expansion of background knowledge, and the uncertainty reduction in $\mathcal{H}^{\mathbf{E}_i}$. To implement π, since the true information gain $r_k(e)$ is not known before querying, we propose a scoring function that serves as a proxy for the expected reward, jointly accounting for epistemic uncertainty, proximity to background knowledge thresholds, and exploration,

$$S_i^e = w_1 E_i^e + w_2 \left(\frac{1}{TD_i^e}\right) + w_3 \sqrt{\frac{\log T_i}{T_i^e}}, \quad \text{s.t.} \quad TD_i^e = \tau_i - \max(\mathcal{H}^{\mathbf{E}_i}(e)), \tag{9}$$

where TD_i^e is the threshold distance from being included in background knowledge, E_i^e, T_i, T_i^e are as defined in Eq. (5), and w_1, w_2, w_3 are hyper-parameters controlling the trade-off. After querying, the LM's response

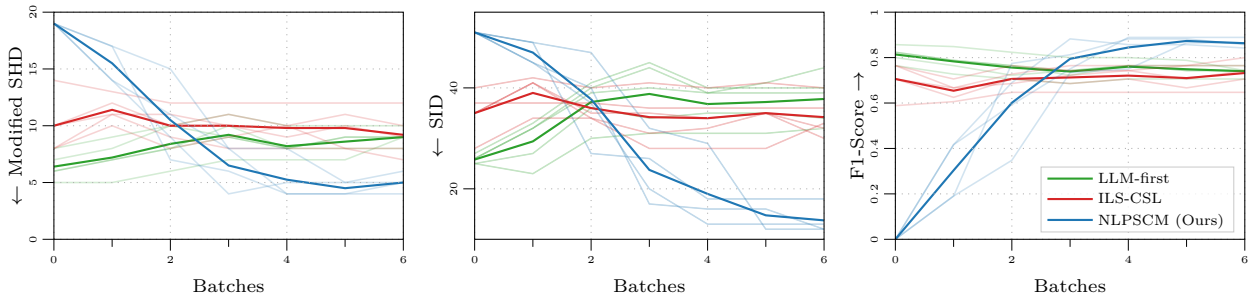


Figure 3: User Level Data - I: Performance evolution across batches for *Data-LM* methods. Left: Modified Structural Hamming Distance (↓), Middle: Structural Intervention Distance (↓), and Right: F1-Score (↑). NLPSCM consistently outperforms other approaches as data accumulation progresses.

updates the histogram $\mathcal{H}^{\mathbf{E}_i}(e)$, which constitutes the realized information gain. At each step, the edge $e^* = \arg\max_e S_i^e$ is selected for LM interaction. Fig. 4 showcases the effectiveness of the proposed selection score (*cf.* Eq. (9)), with details discussed in Sec. 6.1. This formulation generalizes to other expert-guided tasks (*e.g.,* confounder detection) by redefining the arms and reward function.

While the connection to multi-armed bandits and the derivation of regret bounds are appealing in this setup, LM's stochastic nature, inter-dependent arms, and implicit priors induced by causal structure constraints warrant caution. We discuss this in detail in Sec. D.

5 Bayesian Parameter Estimation

Once we obtain a causal structure $\mathcal{G}^{\mathrm{X_i}}$, we address the critical task of *parameter estimation* within the Structural Equation Model (SEM). The parameters $\boldsymbol{\phi} = \{\boldsymbol{\theta}, \sigma^2\}$ include edge weights (coefficients) and noise parameters. With $\mathcal{G}^{\mathrm{X_i}}$ potentially containing both observed and latent variables $\mathbf{V}^{\mathrm{X_i}}{=}\{\mathbf{V}^{\mathrm{O}}, \mathbf{V}^{\mathrm{L}}\}$, we represent observed variable edges as $\boldsymbol{\theta}^{\mathrm{O}}$ and latent confounder edges as $\boldsymbol{\theta}^{\mathrm{L}}$, giving $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{\mathrm{O}}, \boldsymbol{\theta}^{\mathrm{L}}\}$.

When no latent confounders exist ($\mathbf{V}^{\mathrm{L}}{=}\emptyset$), standard Maximum Likelihood Estimation (MLE) optimizes the parameters $\boldsymbol{\phi}{=}\{\boldsymbol{\theta}^{\mathrm{O}}, \sigma^2\}$ as $\boldsymbol{\phi} = \arg\max_{\boldsymbol{\phi}} \log p(\mathbf{D}_i \,|\, \mathcal{G}^{\mathrm{X_i}}, \boldsymbol{\phi})$ using a conventional gradient-based methods. We focus on the more important and challenging scenario where latent confounders exist ($\mathbf{V}^{\mathrm{L}}{\neq}\emptyset$).

With latent confounders, MLE is ill-posed and intractable. We instead employ an iterative Expectation–Maximization (EM) algorithm that incorporates LM-provided probability $p(\mathbf{V}^{\mathrm{L}})$ and correlation $\rho(\mathbf{V}^{\mathrm{O}}, \mathbf{V}^{\mathrm{L}})$ about latent confounders. Specifically, we propose the following EM steps:

- **E-step:** Compute conditional posterior of latent confounder(s) given $\mathbf{D}_i$ and SEM parameters $\boldsymbol{\phi}$,

$$p(\mathbf{V}^{\mathrm{L}} \,|\, \mathcal{G}^{\mathrm{X_i}}, \mathbf{D}_i, \boldsymbol{\phi}) \propto p(\mathbf{D}_i \,|\, \mathcal{G}^{\mathrm{X_i}}, \mathbf{V}^{\mathrm{L}}, \boldsymbol{\phi}) p(\mathbf{V}^{\mathrm{L}}). \tag{10}$$

- **M-step:** Update parameters by maximizing the expected log-likelihood, incorporating LM-provided regularization for latent confounder edges,

$$\boldsymbol{\phi} \in \arg\max_{\boldsymbol{\phi}} \mathbf{E}_{p(\mathbf{V}^{\mathrm{L}} \,|\, \mathcal{G}^{\mathrm{X_i}}, \mathbf{D}_i, \boldsymbol{\phi})}[\log p(\mathbf{D}_i \,|\, \mathcal{G}^{\mathrm{X_i}}, \mathbf{V}^{\mathrm{L}}, \boldsymbol{\phi})] - \lambda \|\boldsymbol{\theta}^{\mathrm{L}} - (\rho(\mathbf{V}^{\mathrm{O}}, \mathbf{V}^{\mathrm{L}}) \sigma_{\mathbf{V}^{\mathrm{O}}} \sigma_{\mathbf{V}^{\mathrm{L}}}^{-1})\|_2 \,. \tag{11}$$

Alg. 2 details the EM parameter estimation algorithm. In Sec. 6.3, we demonstrate the robustness and recovery capability of the proposed parameter estimation algorithm.

6 Experiments

To provide a robust empirical examination, we conduct experiments on four language models: GPT-3.5$_{\text{turbo}}$, GPT4.1$_{\text{nano}}$, Llama3.1$_{\text{8B-instruct}}$, Qwen3$_{\text{4B-instruct}}$ and six datasets: Earthquake from Korb & Nicholson (2010), Asia from Lauritzen & Spiegelhalter (1988), User Level Data - I, User Level Data - II from

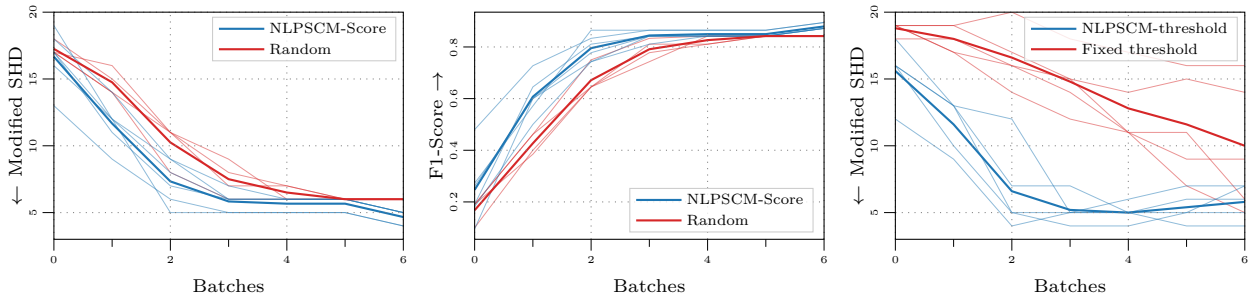


Figure 4: **Structure learning ablation:** The impact of two key components: *selection score* and *dynamic background threshold*. (Left, Middle) Modified SHD (↓) and F1-score (↑) on the User Level Data - I dataset, comparing NLPSCM—*selection score* against random selection. (Right) Modified SHD(↓) comparing NLPSCM— *dynamic threshold* with a conventional fixed threshold.

Google & Kaggle (2018), Child from Spiegelhalter et al. (1993), and Alarm from Beinlich et al. (1989). They range in number of observed variables (nodes) from small (5) to medium (19) to large (37). Earthquake, Asia, Child, and Alarm datasets are standard benchmarks, providing observational data and ground-truth causal structures. To simulate a streaming batch setting, each dataset is split into batches. For the two User Level Data, which contain only observational data, the underlying DAG is inferred using DirectLiNGAM, an algorithm proposed by Shimizu et al. (2011). We treat the DAG inferred from DirectLiNGAM as the ground truth for evaluation. Further details on the data sets and simulation process are provided in Sec. A.

Our evaluation metrics include a *modified* Structural Hamming Distance (Mod. SHD), which extends SHD to account for uncertain edges in PAG; Structural Intervention Distance (SID); and precision, recall, and F1-score for causal relations that are certain. Together, these metrics assess structural accuracy, interventional soundness, and edge-wise discovery performance (details in Sec. C).

6.1 Structure Learning

We evaluate NLPSCM against *Only-Data* and *Data-LM* baselines in Table 4. As shown in Table 1, *Only-LM* methods exhibit *overly optimistic* behavior, producing globally plausible but locally unreliable causal structures. This highlights the need for data-grounded post-processing. Table 4 compares NLPSCM with several baselines, including multiple FCI variants (cumulative, vanilla, iterative, heuristics), as well as *Data-LM* approaches (LM-first, and ILS-CSL proposed by Ban et al. (2023a)). Across all evaluation metrics, NLPSCM consistently outperforms baselines, which also holds for different LM temperatures (Table A3). While Table 4 reports metrics for the final batch, we also show performance evolution across batches, a crucial step in sequential settings (*cf.* Fig. 3). We justify the use of FCI over other causal discovery algorithms in Sec. B. We provide more experiment details in Sec. F.

Finally, going beyond GPT-3.5$_{\text{turbo}}$ (Table 4), results with recent LMs, GPT-4o and GPT-5 (Table A2) show good performance gains for NLPSCM across LMs. The much higher inference cost of GPT-4o and GPT-5, over GPT-3.5$_{\text{turbo}}$ constrain their use for large set of experiments.

Structure learning ablations We ablate two key components of NLPSCM: *(i)* selection score for sequential optimization, and *(ii)* dynamic background threshold. Results of ablations performed on User Level Data - I are shown in Fig. 4.

Effectiveness of proposed selection score (Eq. (9)) in guiding edge selection under a fixed budget of LM queries, is compared against a random-selection baseline. Fig. 4 (Left, Middle) shows that NLPSCM achieves significantly better Mod. SHD and F1-score across batches, demonstrating the benefit of a principled edge selection policy in sequential structure learning.

We also assess the impact of dynamic background threshold (Eq. (5)) used to promote edges from the histogram $\mathcal{H}^{\mathbf{E}}$ into the background knowledge $\mathcal{B}$. Fig. 4 (Right) reports Mod. SHD over batches, highlighting

Table 5: **NLPSCM improves causal discovery across language model families:** We showcase results on the CHILD and USER LEVEL DATA-I datasets using three language models: *Llama3.1$_{8B\text{-}instruct}$*, *Qwen3$_{4B\text{-}instruct}$* and a GPT-4.1$_{\text{nano}}$. We report 5 metrics: *Modified SHD, SID, Precision, Recall, F1*, with mean and standard deviation over 5 runs and perform significance test with $\alpha = 0.05$.

Dataset	Model	Method	Mod. SHD ↓	SID ↓	Precision ↑	Recall ↑	F1 Score ↑
CHILD	Llama3.1$_{8B\text{-instruct}}$	LLM-first	45.80±1.94	(211.20, 211.20)±(15.25, 15.25)	0.13±0.01	0.19±0.02	0.16±0.01
		ILS-CSL	27.60±2.24	(146.40, 146.40)±(15.93, 15.93)	0.37±0.05	0.38±0.07	0.38±0.06
		NLPSCM	**24.60**±0.80	(**101.80**, **101.80**)±(8.11, 8.11)	**0.44**±0.01	**0.45**±0.02	**0.44**±0.01
	Qwen3$_{4B\text{-instruct}}$	LLM-first	35.80±2.04	(173.80, 173.80)±(14.13, 14.13)	0.27±0.02	**0.42**±0.02	0.34±0.01
		ILS-CSL	30.60±2.42	(196.80, 196.80)±(14.13, 14.13)	0.30±0.06	0.28±0.05	0.29±0.05
		NLPSCM	**21.20**±0.75	(**107.80**, **107.80**)±(13.42, 13.42)	**0.51**±0.01	0.32±0.03	**0.39**±0.02
	GPT4.1$_{\text{nano}}$	LLM-first	31.83±2.61	(203.50, 203.50)±(23.06, 23.06)	0.16±0.01	0.14±0.03	0.15±0.02
		ILS-CSL	31.60±2.06	(**147.40**, **147.40**)±(11.25, 11.25)	0.29±0.05	0.19±0.04	0.23±0.03
		BFS	34.75±7.80	(252.50, 252.50)±(23.36, 23.36)	0.18±0.08	0.12±0.04	0.14±0.03
		BFS$_{\text{corr}}$	32.30±4.16	(301.10, 301.10±(14.42, 14.42)	0.21±0.03	0.15±0.01	0.18±0.01
		NLPSCM	**27.90**±1.61	(**143.50**, **165.50**)±(12.23, 35.97)	**0.42**±0.06	**0.23**±0.04	**0.30**±0.03
USER LEVEL DATA-I	Llama3.1$_{8B\text{-instruct}}$	LLM-first	11.50±0.50	(42.50, 42.50)±(1.50, 1.50)	0.74±0.01	0.61±0.03	0.67±0.02
		ILS-CSL	9.75±1.64	(32.25, 32.25)±(3.56, 3.56)	0.78±0.05	0.66±0.03	0.72±0.04
		NLPSCM	**7.50**±0.45	(**25.80**, **25.40**)±(2.32, 4.22)	**0.92**±0.07	**0.69**±0.01	**0.79**±0.01
	Qwen3$_{4B\text{-instruct}}$	LLM-first	23.67±1.11	(64.33, 64.33)±(0.94, 0.94)	0.35±0.04	0.28±0.04	0.31±0.04
		ILS-CSL	20.60±4.18	(56.80, 56.80)±(7.36, 7.36)	0.45±0.13	0.36±0.10	0.40±0.12
		NLPSCM	**9.80**±2.04	(**26.60**, **26.60**)±(6.83, 6.83)	**0.91**±0.08	**0.54**±0.08	**0.67**±0.07
	GPT4.1$_{\text{nano}}$	LLM-first	19.42±1.14	(72.10, 72.10)±(5.36, 5.36)	0.76±0.07	0.32±0.03	0.45±0.03
		ILS-CSL	18.00±1.41	(54.00, 54.00)±(2.53, 2.53)	0.53±0.05	**0.42**±0.03	0.47±0.04
		BFS	36.60±8.96	(65.80, 65.80)±(2.17, 2.17)	0.13±0.06	0.29±0.14	0.18±0.06
		BFS$_{\text{corr}}$	24.00±6.24	(42.20, 42.20)±(6.53, 6.53)	0.17±0.12	0.28±0.11	0.21±0.11
		NLPSCM	**10.60**±2.13	(**34.80**, **35.20**)±(5.56, 5.84)	**0.97**±0.04	**0.43**±0.06	**0.60**±0.06

the advantages of a dynamic threshold over a conventional fixed one. The adaptive mechanism yields more stable and accurate graph recovery throughout the learning process.

6.2 Other LM Families and Memorization in LMs

We evaluate NLPSCM using three *recent* LMs across different families: Llama3.1$_{8B\text{-instruct}}$, Qwen3$_{4B\text{-instruct}}$, and GPT4.1$_{\text{nano}}$. The first two are open models. We evaluate on two datasets: CHILD and the real-world USER LEVEL DATA - I from Google & Kaggle (2018). Measuring performance on the 5 metrics over 5 runs, the results in Table 5 are consistent with those of the previously presented GPT3.5$_{\text{Turbo}}$ model, demonstrating robustness across LMs with different training data, and architectures.

We introduce an additional baseline which does causal discovery using LM and Pearson correlation in the prompt, namely BFS, as proposed by Jiralerspong et al. (2024b). A comparison of NLPSCM with BFS also serves an important purpose by throwing light on the issue of potential memorization of datasets seen in training by LMs. As noted by Jiralerspong et al. (2024b), their method relies on the LM's training knowledge (Sec. 5 of their paper), thereby recognizing the dependence on memorization. To make a fair comparison, we use a recent model GPT4.1$_{\text{nano}}$. Results reported in the bottom panel of Table 5 show that for both CHILD and USER LEVEL DATA - I, NLPSCM beats BFS handily. To elucidate about memorization, we compare the difference-of-differences in metric-values between BFS and NLPSCM. The difference is *smaller* for CHILD dataset, a standard causal discovery benchmark that is likely to be a part of the LM training data. In contrast, the difference is *much larger* for the USER LEVEL DATA-I dataset. The latter dataset or its causal graph is unlikely to have been seen by any LM due to the processing and construction of attributes we performed in this dataset, from the publicly available large data at Google & Kaggle (2018). That BFS performs worse highlights its reliance on memorized knowledge, whereas NLPSCM demonstrates greater robustness across datasets.

6.3 Parameter Estimation

The datasets used in Sec. 6.1 do not contain latent confounders in their causal structures. Consequently, parameter estimation reduces to standard maximum likelihood estimation, which NLPSCM (Sec. 5) replicates by design. To meaningfully evaluate NLPSCM in the presence of latent confounders, we consider a real-world dataset: the RED-WINE QUALITY from Cortez et al. (2009) (details in Sec. A).

Table 6: **LM-predicted priors for confounding variable:** LM suggest relevant Gaussian priors when the confounder is meaningful and default to $\mathcal{N}(0, 1)$ when uncertain.

Variables	GPT-3.5$_{\text{turbo}}$	GPT-4o
density$\leftarrow$ alcohol $\rightarrow$ quality	$\mathcal{N}(12.5, 2.5)$	$\mathcal{N}(10.5, 1.2)$
density $\leftarrow$ alcohol $\rightarrow$ volatile-acidity	$\mathcal{N}(12.5, 2.5)$	$\mathcal{N}(10.5, 1.2)$
cat $\leftarrow$ alcohol $\rightarrow$ mouse	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$
bed $\leftarrow$ alcohol $\rightarrow$ shopping	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$

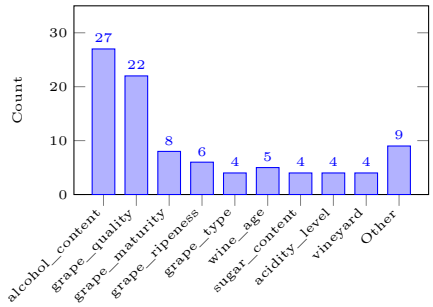


Figure 6: LM predicted confounders.

For the RED-WINE QUALITY, NLPSCM's structure learning algorithm predicts a latent confounder between variables *quality* and *density*. Applying DirectLiNGAM on the full dataset indicates that the true confounder is *alcohol content*, aligning with domain knowledge. Following Sec. 4.2, we query an LM to identify potential latent confounders using world knowledge. Fig. 6 shows a histogram of LM predictions, with *alcohol_content* emerging as the top candidate. We incorporate this LM-provided confounder into NLPSCM's parameter estimation pipeline and query the LM for its marginal distribution. Table 6 presents the LM-provided Gaussian distributions. Notably, when queried with obscure variables (*e.g.,* (cat, mouse)), LM often defaults to unit Gaussian $\mathcal{N}(0, 1)$.

To assess parameter recovery, we track evolution of $\boldsymbol{\theta}$ over sequential batches. As a performance metric, we compute the ℓ_2-norm error $\|\boldsymbol{\theta}^\star - \boldsymbol{\theta}\|_2$, where $\boldsymbol{\theta}^\star$ denotes the parameters obtained via MLE assuming the confounder (*alcohol_content*) is observed. Fig. 5 visualizes this convergence behavior.

Parameter estimation: robustness and recovery We demonstrate robustness of the proposed parameter estimation algorithm under misspecified or ill-informed priors. Based on domain knowledge and observational data, we estimate latent confounder *alcohol_content* to follow distribution $\mathcal{N}(11, 1.0)$. To stress-test NLPSCM, we experiment with three alternative priors: $\mathcal{N}(12.5, 2.5)$ (suggested by GPT-3.5$_{\text{turbo}}$), $\mathcal{N}(0, 1)$, and a severely misspecified prior $\mathcal{N}(50, 1.5)$. Fig. 5 illustrates the evolution of the learned parameters $\boldsymbol{\theta}$ across training batches for each prior. As expected and aligned with the Bayesian principle, convergence is slower when initialized with an inaccurate prior. Nevertheless, the model progressively refines its estimates as more data is processed, ultimately converging towards $\boldsymbol{\theta}^\star$. This demonstrates both the robustness and recovery capabilities of NLPSCM's Bayesian parameter estimation algorithm—even under poor initialization. Additionally, we incorporate LM-suggested Pearson correlation coefficients in the *M-step* objective to further guide estimation, Eq. (11) (see discussion and results in Sec. E.).

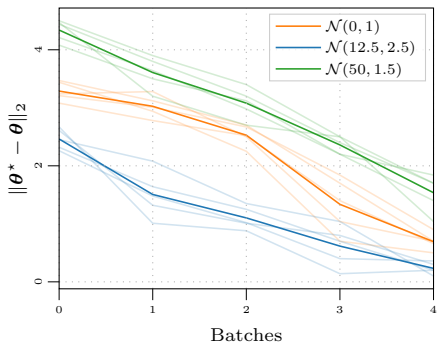


Figure 5: **Parameter Estimation:** Convergence of parameters and robustness to prior misspecification as more batches are processed.

7 Discussion and Conclusion

We present NLPSCM (Noisy Language Prior in Sequential Causal Modeling)—a Bayesian framework for causal structure discovery and parameter (edge weights) estimation in sequential, batch-wise data settings. By treating language models (LMs) as noisy surrogate experts, NLPSCM addresses the dual *LM-induced* and *data-induced* biases. A key contribution is the representation shift from DAGs to PAGs, allowing uncertainty and confounders to be modeled explicitly. Through LM interactions modeled in sequential optimization framework and EM style iterative parameter estimation algorithm, NLPSCM improves both structural accuracy and parameter recovery in hybrid *data-LM* pipelines. NLPSCM leverages global LM knowledge while staying grounded in local data, and offers a robust foundation for hybrid causal discovery in the presence of sequential, batched data.

Limitations and Future Work In the Bayesian formulation in NLPSCM, incorporating Dirichlet or hierarchical Bayesian priors over LM judgments could provide another way to capture LM uncertainty. Extending to fully probabilistic inference over graph structures, rather than edge-level belief tracking, is another promising direction. Future work may also explore adaptive calibration of LM responses or memory-based accumulation of observational data across batches, following Chang et al. (2023). We employ FCI for causal discovery due to its compatibility with our setup (Section 3). The current framework assumes that each latent confounder may affect two observed variables; extending to multi-variable confounding is a direction for future work. Systematically evaluating the robustness of NLPSCM for other causal discovery algorithms remains important future work. Additionally, extending our approach to incorporate interventional data and active learning strategies could improve sample efficiency and the quality of discovered causal structures. Systematic evaluation of LM accuracy in confounder identification across diverse domains also remains a useful future effort. Finally, a theoretical analysis of sequential optimization and a bandit-style framework remains an open problem, discussed preliminarily in Sec. D.

Broader Impact NLPSCM combines observational data with LM-derived knowledge to discover causal structures, which may inform downstream decisions in domains such as healthcare, policy, and business. As with any causal discovery method, the learned structures reflect the assumptions and limitations of the underlying algorithms, data quality, and in our case, the reliability of LM priors. We recommend that practitioners treat discovered structures as hypotheses to be validated through domain expertise or interventional studies before acting on them.

References

Philippe Andrade and Marios Zachariadis. Global versus local shocks in micro price dynamics. *Journal of International Economics*, 98:78–92, 2016.

Pierre Baldi and Babak Shahbaba. Bayesian causality. *The American Statistician*, 74(3):249–257, 2020.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *CoRR*, 2023a.

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 preprint arXiv:2306.16902*, 2023b.

Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings*. Springer, 1989.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Adrien Bilal and Diego R Känzig. The macroeconomic impact of climate change: Global vs. local temperature. Technical report, National Bureau of Economic Research, 2024.

Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 1989.

Eric W Bridgeford, Michael Powell, Gregory Kiar, Ross Lawrence, Brian Caffo, Michael Milham, and Joshua T Vogelstein. Batch effects are causal effects: applications in human connectomics. *bioRxiv*, 3, 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Paul Edmund Chang, Prakhar Verma, S. T. John, Arno Solin, and Mohammad Emtiyaz Khan. Memory-based dual Gaussian processes for sequential learning. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 2002.

Davin Choo and Kirankumar Shiragur. Subset verification and search algorithms for causal DAGs. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 2012.

Anthony C Constantinou, Zhigao Guo, and Neville K Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8), 2023.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 2009.

Tiago da Silva, Bruna Bazaluk, Eliezer de Souza da Silva, António Góis, Dominik Heider, Samuel Kaski, Diego Mesquita, and Adèle Helena Ribeiro. Expert-aided causal discovery of ancestral graphs, 2025.

Bita Darvish Rouhani, Mohammad Ghasemzadeh, and Farinaz Koushanfar. Causalearn: Automated framework for scalable streaming-based causal Bayesian learning using fpgas. In *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 1–10, 2018.

Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17), 2012.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 2024.

Helmut J Geist and Eric F Lambin. Proximate causes and underlying driving forces of tropical deforestation: Tropical forests are disappearing as the result of many pressures, both local and regional, acting in various combinations in different geographical locations. *BioScience*, 52(2):143–150, 2002.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10, 2019.

Google and Kaggle. Google analytics sample dataset, 2018. URL https://www.kaggle.com/datasets/bigquery/google-analytics-sample/data.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov), 2008.

David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20, 1995.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *International Conference on Learning Representations*, 2024.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024a.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models, 2024b.

Ryan P Kelly, MM Foley, WS Fisher, RA Feely, BS Halpern, GG Waldbusser, and MR Caldwell. Mitigating local causes of ocean acidification with existing laws. *Science*, 332(6033):1036–1037, 2011.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50, 1988.

Longzhu Li, Yaojin Lin, Hong Zhao, Jinkun Chen, and Shaozi Li. Causality-based online streaming feature selection. *Concurrency and Computation: Practice and Experience*, 33(20):e6347, 2021.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

Takashi Nicholas Maeda and Shohei Shimizu. RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2020.

Colin D Mathers, Ties Boerma, and Doris Ma Fat. Global and regional causes of death. *British medical bulletin*, 92(1):7–32, 2009.

Melissa A McKinney, Kylie Dean, Nigel E Hussey, Geremy Cliff, Sabine P Wintner, Sheldon FJ Dudley, M Philip Zungu, and Aaron T Fisk. Global versus local causes and health implications of high mercury concentrations in sharks from the east coast of south africa. *Science of the Total Environment*, 541:176–183, 2016.

Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995.

Christopher Meek. Causal inference and causal explanation with background knowledge. *arXiv preprint arXiv:1302.4972*, 2013.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of The Thirty-Fifth Uncertainty in Artificial Intelligence Conference*. PMLR, 2020.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17, 2016.

Richard E Neapolitan et al. *Learning Bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, 2004.

Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, Proceedings of Machine Learning Research, pp. 368–379. PMLR, 2016.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Agathe Sadeghi, Achintya Gopal, and Mohammad Fesanghary. Causal discovery in financial markets: A framework for nonstationary time-series data. *arXiv preprint arXiv:2312.17375*, 2023.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 2021.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48, 2020.

Shohei Shimizu. LiNGAM: Non–Gaussian methods for estimating causal structures. *Behaviormetrika*, 41, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12, 2011.

David J Spiegelhalter, A Philip Dawid, Steffen L Lauritzen, and Robert G Cowell. Bayesian analysis in expert systems. *Statistical science*, 1993.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied Informatics*. Springer, 2016.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

Yuni Susanti and Michael Färber. Can LLMs leverage observational data? Towards data-driven causal discovery with LLMs. *arXiv preprint arXiv:2504.10936*, 2025.

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. Integrating large language models in causal discovery: A statistical causal approach. *Transactions on Machine Learning Research*, 2025.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32, 2019.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *International Conference on Learning Representations*, 2025.

Chris Wallace, Kevin B Korb, and Honghua Dai. Causal discovery via MML. In *ICML*, volume 96, pp. 516–524. Citeseer, 1996.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

Dianlong You, Ruiqi Li, Shunpan Liang, Miaomiao Sun, Xinju Ou, Fuyong Yuan, Limin Shen, and Xindong Wu. Online causal feature selection for streaming features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(3):1563–1577, 2021.

Dianlong You, Siqi Dong, Shina Niu, Huigui Yan, Zhen Chen, Shunfu Jin, Di Wu, and Xindong Wu. Local causal structure learning for streaming features. *Information Sciences*, 647:119502, 2023.

Kui Yu, Xindong Wu, Hao Wang, and Wei Ding. Causal discovery from streaming features. In *2010 IEEE International Conference on Data Mining*, pp. 1163–1168. IEEE, 2010.

Yue Yu, Tian Gao, Naiyu Yin, and Qiang Ji. DAGs with no curl: An efficient DAG structure learning approach. In *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2021.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172, 2008.

Jiji Zhang and Peter L Spirtes. Strong faithfulness and uniform consistency in causal inference. *arXiv preprint arXiv:1212.2506*, 2012.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011.

Zezhong Zhang and Ted Yuan. Practical batch Bayesian sampling algorithms for online adaptive traffic experimentation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 471–480, 2024.

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

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.