DKCD: Domain Knowledge-Enhanced Causal Discovery from Unstructured Data

Anonymous ACL submission

Abstract

Causal discovery from unstructured data is a demanding yet underexplored task in several scientific domains. Recent methods leverage the general knowledge of large language models (LLMs) to identify causal factors from text and annotate them into structured data for constructing causal graphs. However, they remain limited in high-expertise domains, where two key challenges arise: (1) insufficient domain-specific knowledge impedes the identification of latent yet essential factors that are not explicitly stated in text, and (2) without domain-grounded reasoning, LLM-based annotation is unreliable, propagating errors into the final causal graph. To address these challenges, we introduce DKCD, a **D**omain **K**nowledge-enhanced **C**ausal **D**iscovery framework for extracting causal graphs from unstructured text in specialized domains. Specifically, DKCD is devised with a knowledge mining module to retrieve relevant knowledge. Through knowledge-guided reasoning, it effectively identifies overlooked latent factors and probes key causal clues to support more accurate data annotation, thereby improving the quality of causal discovery. Experimental results on two domain-specific datasets demonstrate that DKCD significantly improves performance in both causal factor identification and causal graph construction. The code and data are available at `https://anonymous.4open.science/r/DKCD-1E82`

1 Introduction

Exploring and uncovering causality from real-world data is fundamental across a wide range of scientific domains (Bunge, 2017; Illari et al., 2011), including healthcare (Shi and Norgeot, 2022; Yang et al., 2013), finance (Papana et al., 2017; Sokolov et al., 2025), and education (Morrison and van der Werf, 2016; Fancsali, 2014). In data-driven causal discovery, causal relationships are typically represented as causal graphs. Existing causal discovery algorithms, such as the Peter–Clark (PC) algorithm (Spirtes et al., 2000; Ramsey, 2016) and the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2013), are primarily designed for structured data with predefined schemas and fixed formats, commonly organized in tabular form (e.g., relational databases and spreadsheets). Thus, they face substantial limitations when applied to unstructured data (Malinsky and Danks, 2018; Yu et al., 2016). Meanwhile, the growing dominance of unstructured data, particularly natural-language text (Siddiqa et al., 2017; Eberendu et al., 2016; Azad et al., 2020), calls for methods that bridge unstructured text and statistical causal discovery.

With the emergence of large language models (LLMs), new opportunities have arisen for analyzing and discovering causal relationships from unstructured text (Wang et al., 2023; Wang, 2024). With strong capabilities in understanding, reasoning, and extracting knowledge from natural language, LLMs is able to well support automatic identification of causal factors, reducing reliance on manual domain expertise for data labelling (Ashwani et al., 2024; Dubey et al., 2024). Specifically, the COAT framework (Liu et al., 2024) is a representative approach that leverages LLMs to extract potential causal factors from unstructured text, annotate factor values, and then apply statistical causal discovery algorithms to construct causal graphs. Although COAT takes an important step forward, it suffers from critical limitations when applied to domains where causal factor identification requires substantial domain expertise, such as clinical text analysis, in which many causally relevant factors are implicit rather than explicitly stated in the text.

As shown in Figure 2, applying COAT to a diabetes-domain dataset reveals two critical limitations: **Challenge 1 (CH1):** It relies solely on LLMs' general knowledge to propose causal factors. Although LLMs are pre-trained on large-scale

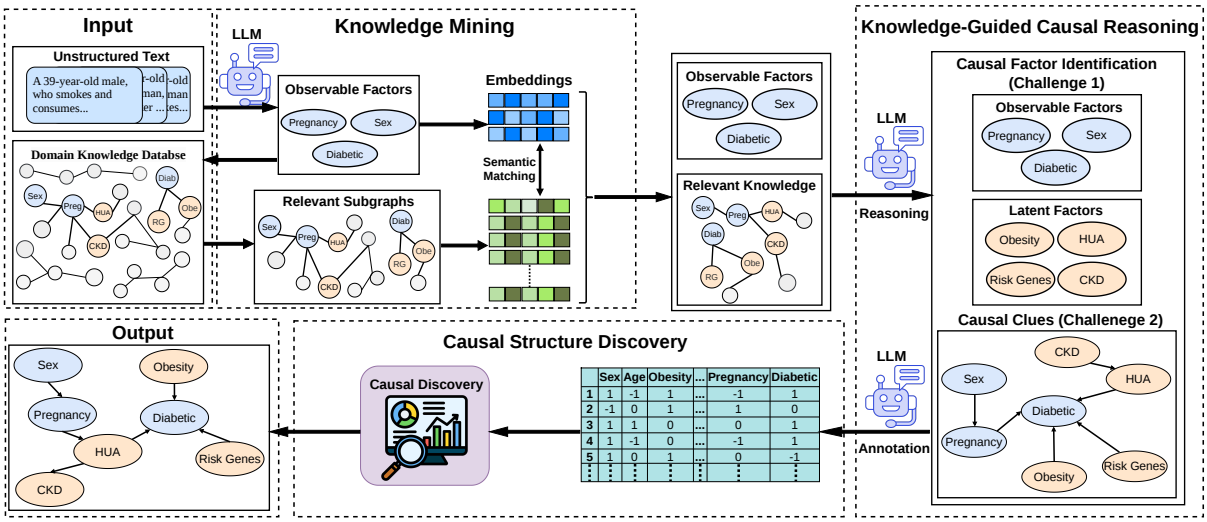


Figure 1: Overview of the DKCD framework. DKCD leverages domain knowledge graphs to guide latent factor discovery and enhances LLM-based causal reasoning by injecting external domain knowledge. The resulting knowledge-grounded factor scoring improves the robustness and accuracy of downstream causal graph construction.

corpora, their knowledge is general-purpose and often lacks sufficient domain specificity (Yu et al., 2024b; Shou et al., 2023). Consequently, they fail to identify domain-relevant *latent* factors that are not explicitly mentioned in the text but are essential for a complete causal model. As shown in Figure 2, COAT misses many latent factors (highlighted as red nodes), producing an incomplete set of causal factors that fundamentally limits the quality of the final causal graph. **Challenge 2 (CH2):** Beyond factor identification, constructing a causal graph requires accurately annotating each factor's value for each sample to form structured data, serving as the input to statistical causal discovery algorithms. However, LLMs' causal reasoning capabilities remain limited (Mirzadeh et al., 2024) and often produce inaccurate or inconsistent annotations without grounding in external domain knowledge (Burford et al., 2024; Zečević et al., 2023; Ban et al., 2025), which are further propagated into the inferred causal graphs, as evidenced by the substantial deviation from the ground-truth graph in Figure 2. Additionally, Jin et al. (Li et al., 2025a) extended COAT (Liu et al., 2024) by proposing the MLLM-CD framework for multimodal datasets, but it still faces the same challenges.

To address these challenges, we propose the DKCD (**D**omain **K**nowledge-enhanced **C**ausal **D**iscovery) framework, which leverages domain knowledge graphs to guide both factor discovery and factor annotation, thereby improving the quality of statistically inferred causal graphs from unstructured text. As illustrated in Figure 1, DKCD consists of three tightly connected components. (1) Knowledge Mining: It extracts observable factors from unstructured text using LLMs and retrieves relevant subgraphs from domain KGs. Semantic matching filters the retrieved knowledge to preserve the most relevant context, supporting subsequent causal reasoning. (2) Knowledge-Guided Causal Reasoning: Building on the observable factors and retrieved knowledge context, this module infers latent factors that are not explicitly stated in the text but are causally relevant in the domain, thereby directly addressing the incomplete factor identification challenge (**CH1**). In addition, the module generates *causal clues*, i.e., domain-grounded descriptions of plausible dependencies among factors. These clues provide structured guidance for scoring and annotation, helping reduce ambiguity and improve annotation reliability, thereby addressing the inaccurate factor annotation challenge (**CH2**). (3) Causal Structure Discovery: With a more complete factor set and more accurate annotations, this module applies a statistical causal discovery algorithm to infer the final causal graph. The main contributions of this work are summarized as follows:

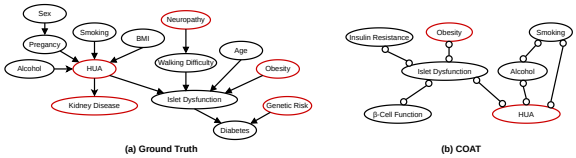


Figure 2: The COAT result on the diabetes dataset.

- We propose DKCD, a novel knowledge-guided framework that integrates domain knowledge graphs with LLM-based reasoning to enhance causal discovery from unstructured data with domain expertise in high-expertise domains.
- We design a knowledge-guided causal reasoning module that leverages retrieved domain knowledge to guide the discovery of latent causal factors and generate knowledge-grounded causal clues for improved annotation accuracy.
- We construct the first domain-specific benchmarks with real-world knowledge graphs for evaluation. Extensive experiments show that DKCD significantly outperforms existing methods regarding both causal factor identification and causal graph construction.

2 Related Work

Causal Discovery aims to identify causal relationships from data and typically represents them as causal graphs. Existing methods are commonly divided into three categories: constraint-based (Niu et al., 2024; Vowels et al., 2022), score-based (Vowels et al., 2022), and continuous optimization methods (Gong et al., 2024; Hyttinen et al., 2016). Constraint-based methods infer structure through conditional independence tests, with representative algorithms including FCI (Spirtes et al., 2013), PC (Spirtes et al., 2000; Ramsey, 2016), and PCMCI (Runge et al., 2019). Score-based methods evaluate candidate graphs using criteria such as the Bayesian Information Criterion (BIC) and the Bayesian Dirichlet equivalent uniform (BDeu) (Huang et al., 2018; Nogueira et al., 2022), while continuous optimization methods reformulate structure learning as differentiable optimization, such as NOTEARS (Zhou and Chen, 2022; Lopez et al., 2022). Although existing causal discovery algorithms perform well on structured data, they are difficult to apply directly to unstructured text. DKCD bridges this gap by enabling effective causal discovery on domain-specific unstructured data.

LLM for Causal Discovery has recently emerged as a promising direction (Gopalakrishnan et al., 2024; Wu et al., 2023; Tong et al., 2024). Prior studies use LLMs to extract causal factors (Liu et al., 2024), estimate causal edges (Ban et al., 2023; Darvariu et al., 2024), and support counterfactual reasoning from natural language data (Gendron et al., 2024; Liu et al., 2026). However, LLMs alone do not reliably perform causal reasoning, as their predictions are largely driven by pretraining data, which limits robustness in domain-specific settings.

KG-Guided LLMs for Causal Discovery introduce structured domain knowledge to improve LLM-based causal reasoning (Kim et al., 2024; Pan et al., 2024). Representative studies include RC2R (Yu et al., 2024a), RealTCD (Li et al., 2024), and biomedical KG-enhanced prompting or retrieval methods (Susanti and Färber, 2024; Lin et al., 2026). Despite these advances, LLMs still struggle to distinguish correlation from causation (Wu et al., 2024), especially in specialized domains with spurious correlations. This motivates combining statistical causal discovery, LLMs, and domain KGs, which is the core idea of DKCD.

3 Domain Knowledge-Enhanced Causal Discovery

3.1 Problem Definition

We consider a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where each sample $\mathbf{x}_k \in \mathcal{X}$ represents a domain-specific textual description. Taking the healthcare domain as an example, patient health reports contain unstructured observations that can be used to identify potential causal factors (Li et al., 2025b; Ban et al., 2025). Let $\mathcal{V}_o^* = \{V_1, V_2, \ldots, V_d\}$ denote the set of ground-truth *observable causal factors* that can be directly extracted from the text. However, observable factors alone may not fully capture the causal mechanisms governing the domain process. Therefore, we introduce a set of ground-truth *latent causal factors* $\mathcal{V}_l^* = \{U_1, U_2, \ldots, U_{d'}\}$, which represent higher-level conceptual mechanisms that are not explicitly stated in the text but can be inferred with the assistance of domain knowledge. Let $\mathcal{V}^* = \mathcal{V}_o^* \cup \mathcal{V}_l^*$ denote the complete set of causal variables. We assume that the true causal relationships among these variables follow an unknown directed acyclic graph (DAG) $G^* = (\mathcal{V}^*, E^*)$, where E^* denotes the set of directed causal edges. To discover the true causal graph from unstructured data, three main tasks must be accomplished. First, observable factors are extracted from the text, and relevant domain knowledge is retrieved from the domain KG. Second, the retrieved knowledge is integrated with the observable factors to identify latent factors and generate causal clues. Third, the causal structure discovery module annotates each text sample based on the observable factors, latent factors, and causal clues, and converts the annotations into a structured

scoring table for statistical causal discovery.

Identifying Potential Causal Factors. This task involves identifying the relevant set of causal variables $\mathcal{V}$ from the unstructured dataset $\mathcal{D}$. Observable causal factors $\mathcal{V}_o$ can be extracted directly from the textual descriptions. However, latent factors $\mathcal{V}_l$, such as neuropathy, HUA, kidney disease, obesity, and genetic predisposition, these may not be explicitly mentioned in diabetic patient description datasets, yet are essential for a complete causal model. Identifying such latent factors requires domain-specific knowledge that goes beyond what is available in the text or in general LLM pre-training.

Annotating Factor Values. Given the identified factors $\mathcal{V}$, each text $\mathbf{x}_k$ must be annotated with a value v_{kj} for each variable $V_j \in \mathcal{V}$, converting $\mathcal{D}$ into a structured scoring table

$$\mathbf{S} = \{(v_1^k, v_2^k, \cdots, v_{|\mathcal{V}|}^k)\}_{k=1}^n, \tag{1}$$

where $|\mathcal{V}|$ denotes the cardinality of $\mathcal{V}$. This annotation step is critical because the quality of $\mathcal{S}$ directly determines the accuracy of downstream causal discovery. However, accurately scoring factors, especially latent ones, from unstructured text requires domain-grounded reasoning that LLMs lack when operating without external knowledge.

Discovering the Causal Structure. Given the structured data $\mathbf{S}$, a statistical causal discovery algorithm is applied to infer the causal graph G. The accuracy of the inferred graph depends on both the completeness of the factor set and the reliability of the annotations. Errors in either distort conditional independence tests and propagate to the graph.

The proposed DKCD framework aims to address these challenging tasks by integrating LLMs with domain knowledge graphs and statistical causal discovery algorithms for reliable causal structure discovery. The overall framework is illustrated in Figure 1 and consists of three key components. First, the knowledge mining module extracts observable causal factors from unstructured text using LLMs and retrieves relevant subgraphs from the domain knowledge graph to support the subsequent causal reasoning. Second, the knowledge-guided causal reasoning implements a knowledge-grounded reasoning to discover latent causal factors and derive causal clues to guide more reliable annotation. Third, the causal structure discovery module applies statistical causal discovery algorithms to infer causal relationships among the variables for constructing the final causal graph.

3.2 Knowledge Mining

As illustrated in Figure 1, this module takes a domain-specific textual description together with the domain KG, denoted by $\mathcal{G}_{KG}$, as the input. The objective of this module is not only to extract directly observable causal variables, but also to retrieve the most relevant contextual knowledge for subsequent knowledge-guided reasoning.

Observable Factor Proposal: Given a small set of textual samples $\mathcal{X}_s = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, where $\mathcal{X}_s \subset \mathcal{D}$ and each sample $\mathbf{x}_k \in \mathcal{X}_s$ denotes a textual description, the LLM identifies a set of observable causal factors $\mathcal{V}_o$ explicitly mentioned in the unstructured text. The LLM first identifies a set of observable causal factors $\mathcal{V}_o$ that are explicitly stated from the unstructured text, aiming to recover the ground-truth set $\mathcal{V}_o^*$ defined in the problem formulation. This step is formulated as a mapping

$$f_{\text{obs}} : \mathcal{X}_s \to \mathcal{V}_o. \tag{2}$$

These factors provide the initial semantic anchors for subsequent retrieval of relevant knowledge and latent factor mining.

Domain KGs Retrieval: Based on the identified observable factors $\mathcal{V}_o$, we retrieve relevant subgraphs from domain KGs, denoted by $\mathcal{G}_{KG}$. Specifically, the observable factors are used as queries to fuzzily match relevant entity nodes in $\mathcal{G}_{KG}$, and their neighboring entities and relations are collected to form contextual subgraphs

$$\mathcal{G}_{rel}^{(n)} \subseteq \mathcal{G}_{KG}, \tag{3}$$

where $\mathcal{G}_{rel}^{(n)}$ denotes the retrieved subgraph with n triples, where n is the number of retrieved triples. In the semantic matching stage, these triples are further transformed into natural language statements for relevance ranking. The retrieved subgraph contains structured domain concepts relevant to the observable factors. As an external knowledge source, the domain KG enriches the causal semantics of each textual description and supports subsequent latent factor inference.

Semantic Matching: To perform finer-grained retrieval aligned with semantic information, we further conduct semantic matching between observable factors and retrieved knowledge entities to preserve the most relevant knowledge context. Specifically, let $\mathcal{S}^{(k)} = \{s_1, s_2, \ldots, s_{n_k}\}$ denote the set of natural language statements verbalized

from the triples in the retrieved subgraph. For each observable factor $v_i \in \mathcal{V}_o$, we encode v_i and each statement $s_j \in \mathcal{S}^{(k)}$ into dense embeddings using the pretrained sentence-transformer model all-MiniLM-L6-v2 (Yin and Zhang, 2024; Vergou et al., 2023), denoted as $E(\cdot)$, and compute their semantic similarity using cosine similarity:

$$\mathrm{sim}(v_i, s_j) = \frac{E(v_i)^\top E(s_j)}{\|E(v_i)\| \|E(s_j)\|}. \tag{4}$$

We then select the top-r most relevant statements for each observable factor:

$$\mathcal{S}_{v_i}^{(k)} = \text{Top-}r\{s_j \in \mathcal{S}^{(k)} \mid \mathrm{sim}(v_i, s_j)\}, \tag{5}$$

where $\mathcal{S}_{v_i}^{(k)}$ denotes the set of matched knowledge statements associated with factor v_i. Further analysis of the parameter r is provided in Appendix B. The selected knowledge items are then used to construct the relevant knowledge context for each sample, providing more reliable support for latent factor discovery and causal clue inference.

3.3 Knowledge-Guided Causal Reasoning

This module leverages the retrieved domain knowledge to perform two key reasoning tasks that directly address the two identified challenges. First, it infers latent causal factors absent from the surface text, thereby addressing the problem of incomplete factor identification. Second, it generates knowledge-grounded causal clues that guide more reliable and accurate factor annotation.

Latent Factor Discovery: Given the observable factor set $\mathcal{V}_o$ and the matched knowledge context derived from $\mathcal{G}_{rel}^{(n)}$, the LLM performs knowledge-guided reasoning to infer a set of latent causal factors $\mathcal{V}_l$. These latent factors are not directly mentioned in the unstructured text, but are supported by the semantic relations and domain concepts provided by the retrieved knowledge. The complete causal factor set is therefore defined as

$$\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_l. \tag{6}$$

By incorporating domain knowledge to discover latent factors, this step extends causal factor discovery beyond surface-level textual mentions, leading to a more comprehensive factor set that better approximates the ground-truth factor set $\mathcal{V}^*$.

Causal Clue Generation: Beyond factor discovery, the LLM derives causal clues from the interaction between the identified factors and the retrieved domain knowledge through internal reasoning ability. These causal clues describe plausible causal dependencies and domain-informed influence patterns among factors. Formally, for each sample $\mathbf{x}_k$, the generated causal clues are denoted as

$$\mathcal{C}^{(k)} = L(\mathcal{V}_o, \mathcal{G}_{rel}^{(k)}), \tag{7}$$

where $L(\cdot)$ represents the knowledge-guided reasoning process implemented by the LLM. Note that the generated clues do not directly determine the final causal graph. Instead, they serve as domain-grounded guidance during the annotation step: by providing the LLM with explicit reasoning about how factors relate causally, the clues help produce more accurate and consistent factor values in the scoring table $\mathbf{S}$, thereby improving the reliability of the structured data fed into the downstream causal structure discovery module.

3.4 Causal Structure Discovery

With a more complete causal factor set $\mathcal{V}$ and more reliable annotations in the structured scoring table $\mathbf{S}^{(t)}$, the final step applies a statistical causal discovery algorithm $\mathcal{C}$ to infer the causal graph. Formally, the causal discovery process can be expressed as

$$G = \mathcal{C}\left(\mathbf{S}, \mathcal{V}\right), \tag{8}$$

where $\mathcal{C}$ denotes the causal discovery algorithm, and G denotes the inferred causal graph whose nodes correspond to the variables in $\mathcal{V}$ and whose edges represent the discovered causal relationships.

In this work, we instantiate $\mathcal{C}$ using the FCI algorithm (Spirtes et al., 2013) to perform the final causal structure discovery. The FCI algorithm is particularly suitable because it can infer causal structures in the presence of latent confounders without requiring all causal variables to be observed. This is important because, even with domain knowledge augmentation, it is often infeasible to identify every causal factor in specialized domains. Thus, FCI is well suited for inferring causal structures from the factor set discovered by DKCD, given its robustness to unobserved variables.

It is worth noting that traditional causal discovery algorithms, including FCI, require structured tabular data as input and are sensitive to noisy or incomplete inputs (Malinsky and Danks, 2018; Spirtes et al., 2013). The upstream modules of DKCD, including knowledge mining and knowledge-guided causal reasoning, are specifically designed to improve the quality of this input

by producing a more complete factor set and more accurate annotations, thereby enabling the statistical algorithm to recover a causal graph closer to the ground truth G^*.

4 Experiments

We evaluate the performance of the DKCD framework in both causal factor discovery and causal structure discovery on synthetic datasets from the diabetes and respiratory domains, using multiple LLMs, including GPT-4o (Hurst et al., 2024), Gemini 2.5 Flash (Comanici et al., 2025), Grok-3 (xAI, 2025), and LLaMA 3-70B (MetaAI, 2024). Due to space constraints, more detailed descriptions of the dataset construction process, prompt design, implementation details, evaluation metrics, and complete experimental results on both datasets, as well as further analyses of the ablation studies and parameter analysis, are provided in Appendix B.

4.1 Experimental Setup

We construct two medical-domain datasets for evaluation: (1) the **Diabetic Patient Condition Descriptions Dataset (Diabetes Dataset)**, a synthetic dataset consisting of two components: 400 patient condition description samples with 14 high-level factors, and a diabetes-domain KG; and (2) the **Respiratory Patient Condition Descriptions Dataset (Respiratory Dataset)**, a synthetic dataset consisting of two components: 400 patient condition description samples with 8 high-level factors, and a respiratory-domain KG. Detailed implementation settings are provided in Appendix B.

Diabetes Dataset: Following the practice in work (Liu et al., 2024; Li et al., 2025a), we generate a medical text dataset containing clinical descriptions of diabetic patients. Each synthetic sample consists of a textual record describing patient symptoms, medical history, physiological indicators, lifestyle habits, and other clinical diagnostic information. This includes 9 observable factors (i.e., sex, pregnancy, alcohol, smoking, BMI, walking difficulty, age, diabetes, and islet dysfunction), and 5 latent factors (HUA, kidney disease, neuropathy, obesity, and genetic risk), as shown in Figure 3(a). Additionally, we construct a diabetes-domain knowledge graph using the DiaKG dataset (Chang et al., 2021), which contains annotated entities and relations for diabetes-related medical knowledge.

Respiratory Dataset: We also generate another medical text dataset containing clinical descriptions of patients with respiratory conditions, following prior work (Liu et al., 2024; Li et al., 2025a). Each synthetic sample includes 5 observable factors (i.e., tuberculosis, smoke, lung cancer, lung disease, and chest X-ray) and 3 latent factors (i.e., Asian travel, bronchitis, and dyspnea), as illustrated in Figure 10(a). Additionally, we construct a respiratory-domain knowledge graph based on previously published studies (Rahman et al., 2023; Stocks, 1959; Lauritzen and Spiegelhalter, 1988; Onozaki et al., 2015; Caplin and Festenstein, 1975). Specifically, GPT-4o (Hurst et al., 2024) is employed to extract entities and their relations from the literature to construct the knowledge graph.

Baselines: Zero-shot LLM (Antonucci et al., 2023; Du et al., 2025) refers to a setting where the causal graph is generated directly by the LLM without providing any examples or domain knowledge. COAT (Liu et al., 2024) is a state-of-the-art LLM-driven framework that constructs causal graphs from unstructured text by integrating LLM reasoning with traditional causal discovery algorithms. META (Liu et al., 2024) is a variant derived from the COAT ablation study, where contextual examples are not provided.

Metrics: Following common practice in the causal discovery literature, we evaluate the performance of factor identification and graph structure using widely adopted metrics (Zanga et al., 2022; Li et al., 2025a; Liu et al., 2024). For causal factor identification, we report: (1) Node Precision (NP), (2) Node Recall (NR), and (3) Node F1-score (NF). For causal structure discovery, we adopt: (1) Adjacency Precision (AP), (2) Adjacency Recall (AR) and (3) Adjacency F1-score (AF). To jointly assess both aspects, we also employ an extended Structural Hamming Distance (ESHD) (Peters and Bühlmann, 2015; Li et al., 2025a).

4.2 Analysis on the Datasets

The empirical results of causal discovery on the diabetes and respiratory datasets are presented in Table 1 and Figure 3, and Table 2 and Figure 10, respectively. The results demonstrate that DKCD consistently outperforms the baseline methods in most cases while maintaining stable performance across different LLM backbones.

Results on the Diabetes Dataset: From the perspective of causal factor identification, DKCD achieves the best overall performance across most LLM settings, obtaining higher NP, NR, and NF scores than the baseline methods, as shown in Ta-

Table 1: Causal factor identification and structure discovery performance on the diabetes dataset.

LLM	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
GPT-4o	Zero-shot LLM	0.26 ± 0.12	0.14 ± 0.07	0.15 ± 0.03	0.17 ± 0.29	0.03 ± 0.05	0.05 ± 0.08	36.67 ± 2.31
	META	0.57 ± 0.07	0.26 ± 0.11	0.35 ± 0.12	**1.00 ± 0.00**	0.13 ± 0.09	0.22 ± 0.13	32.00 ± 1.00
	COAT	0.61 ± 0.11	0.31 ± 0.09	0.41 ± 0.10	0.65 ± 0.04	0.18 ± 0.04	0.28 ± 0.05	30.67 ± 2.08
	DKCD	**0.84 ± 0.16**	**0.71 ± 0.08**	**0.77 ± 0.04**	0.50 ± 0.17	**0.33 ± 0.18**	**0.40 ± 0.18**	**25.33 ± 3.21**
Gemini 2.5 Flash	Zero-shot LLM	0.35 ± 0.07	0.17 ± 0.04	0.23 ± 0.06	0.50 ± 0.00	0.08 ± 0.00	0.14 ± 0.00	36.00 ± 3.61
	META	0.85 ± 0.13	0.85 ± 0.13	0.85 ± 0.13	0.41 ± 0.10	**0.59 ± 0.13**	0.48 ± 0.05	**29.67 ± 1.53**
	COAT	0.52 ± 0.25	0.24 ± 0.05	0.32 ± 0.09	0.37 ± 0.32	0.10 ± 0.12	0.16 ± 0.17	35.00 ± 7.00
	DKCD	**0.67 ± 0.03**	**0.62 ± 0.25**	**0.59 ± 0.21**	**0.86 ± 0.15**	**0.33 ± 0.16**	**0.45 ± 0.17**	**29.67 ± 1.53**
Grok-3	Zero-shot LLM	0.54 ± 0.27	0.16 ± 0.04	0.23 ± 0.06	0.17 ± 0.29	0.03 ± 0.05	0.05 ± 0.08	37.33 ± 2.08
	META	**0.78 ± 0.13**	0.41 ± 0.08	0.54 ± 0.09	0.81 ± 0.17	0.28 ± 0.05	0.41 ± 0.06	27.00 ± 4.58
	COAT	0.55 ± 0.04	0.26 ± 0.05	0.35 ± 0.05	**0.83 ± 0.29**	0.10 ± 0.04	0.18 ± 0.07	33.00 ± 0.00
	DKCD	0.71 ± 0.04	**0.64 ± 0.07**	**0.67 ± 0.05**	0.74 ± 0.14	**0.41 ± 0.16**	**0.44 ± 0.12**	30.33 ± 3.21
LLaMA 3-70B	Zero-shot LLM	0.56 ± 0.22	0.38 ± 0.18	0.45 ± 0.20	0.20 ± 0.09	0.29 ± 0.09	0.29 ± 0.09	29.67 ± 8.14
	META	**0.79 ± 0.10**	0.41 ± 0.04	0.53 ± 0.04	0.48 ± 0.14	0.20 ± 0.05	0.29 ± 0.07	39.00 ± 3.46
	COAT	0.74 ± 0.07	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.35	0.18 ± 0.05	0.27 ± 0.09	29.00 ± 3.61
	DKCD	0.71 ± 0.06	**0.57 ± 0.00**	**0.63 ± 0.07**	**0.65 ± 0.19**	**0.33 ± 0.12**	**0.44 ± 0.14**	**27.67 ± 3.51**

Table 2: Causal factor identification and structure discovery performance on the respiratory dataset.

LLM	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
GPT-4o	Zero-shot LLM	0.42 ± 0.07	0.38 ± 0.00	0.40 ± 0.03	0.44 ± 0.10	0.13 ± 0.00	0.20 ± 0.01	20.00 ± 5.20
	META	0.33 ± 0.06	0.25 ± 0.00	0.28 ± 0.02	**1.00 ± 0.00**	0.13 ± 0.00	0.23 ± 0.00	25.67 ± 1.53
	COAT	0.34 ± 0.06	0.25 ± 0.00	0.29 ± 0.02	0.67 ± 0.58	0.09 ± 0.08	0.15 ± 0.13	27.00 ± 4.00
	DKCD	**0.74 ± 0.13**	**0.75 ± 0.00**	0.74 ± 0.07	**0.87 ± 0.23**	**0.56 ± 0.22**	**0.68 ± 0.21**	**16.67 ± 2.31**
Gemini 2.5 Flash	Zero-shot LLM	0.39 ± 0.05	0.29 ± 0.08	0.33 ± 0.06	**0.44 ± 0.51**	0.09 ± 0.08	0.14 ± 0.12	23.00 ± 6.24
	META	0.21 ± 0.06	0.13 ± 0.00	0.16 ± 0.02	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	26.00 ± 4.36
	COAT	0.23 ± 0.06	0.21 ± 0.07	0.22 ± 0.07	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	30.00 ± 1.73
	DKCD	**0.84 ± 0.08**	**0.88 ± 0.13**	**0.86 ± 0.10**	0.41 ± 0.16	**0.42 ± 0.19**	**0.41 ± 0.17**	**17.33 ± 3.21**
Grok-3	Zero-shot LLM	0.51 ± 0.09	0.38 ± 0.00	0.41 ± 0.02	0.44 ± 0.10	0.13 ± 0.00	0.20 ± 0.01	22.33 ± 3.21
	META	0.37 ± 0.04	0.29 ± 0.08	0.33 ± 0.05	**1.00 ± 0.00**	0.13 ± 0.00	0.23 ± 0.00	25.67 ± 2.08
	COAT	0.32 ± 0.18	0.21 ± 0.07	0.25 ± 0.11	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	26.33 ± 6.03
	DKCD	**0.67 ± 0.14**	**0.63 ± 0.21**	**0.64 ± 0.18**	0.62 ± 0.13	**0.34 ± 0.08**	**0.43 ± 0.09**	**18.67 ± 1.15**
LLaMA 3-70B	Zero-shot LLM	0.32 ± 0.02	0.38 ± 0.00	0.34 ± 0.01	0.50 ± 0.00	0.13 ± 0.00	0.21 ± 0.00	27.00 ± 2.00
	META	0.35 ± 0.07	0.29 ± 0.08	0.32 ± 0.07	**1.00 ± 0.00**	0.13 ± 0.00	0.21 ± 0.00	25.67 ± 0.58
	COAT	0.53 ± 0.12	0.42 ± 0.07	0.47 ± 0.09	0.83 ± 0.29	0.25 ± 0.00	0.38 ± 0.04	23.33 ± 3.06
	DKCD	**0.73 ± 0.09**	**0.80 ± 0.14**	**0.69 ± 0.11**	0.81 ± 0.17	**0.50 ± 0.25**	**0.57 ± 0.18**	**17.33 ± 5.03**

ble 1. This indicates that the proposed DKCD framework can more accurately identify relevant causal factors, including both observable and latent factors, from unstructured clinical descriptions. In particular, compared with COAT (Liu et al., 2024), DKCD achieves stronger causal factor identification performance, indicating that incorporating domain KGs and latent factor reasoning provides more effective support for extracting complete causal factors from unstructured text. Additionally, DKCD demonstrates competitive performance regarding graph structures, with lower ESHD values and stable AR and AF scores across different LLM backbones, indicating that the generated causal structures are closer to the ground-truth causal graph. These results suggest that DKCD not only improves the identification of causal variables but also enhances the accuracy of the inferred causal relationships among factors.

Results on the Respiratory Dataset: A similar trend can be observed on the respiratory dataset. In terms of causal factor identification, DKCD achieves the highest NP, NR, and NF scores across most LLM configurations, with quantitative results presented in Table 2. These results indicate that DKCD has a strong ability to extract meaningful causal factors from unstructured patient descriptions of respiratory conditions. In terms of causal graph structure generation, DKCD also achieves competitive performance, with lower ESHD values and higher AR and AF scores in most cases. These results indicate that DKCD can effectively recover the underlying causal structure among respiratory-related factors. Overall, the results further confirm that integrating domain knowledge with causal discovery algorithms significantly improves both causal factor identification and causal graph construction from unstructured text.

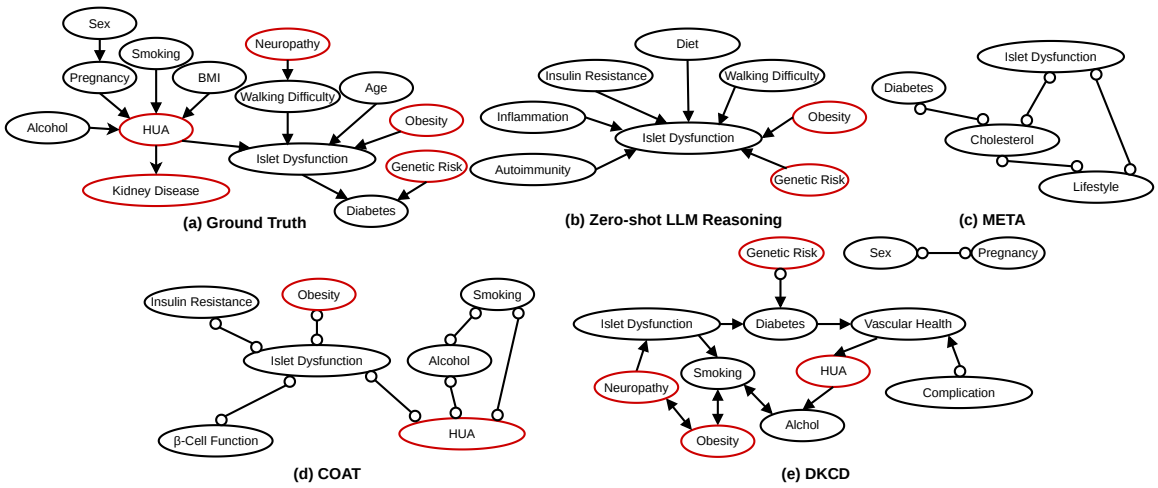


Figure 3: Ground-truth and discovered causal graphs on the diabetes dataset. Latent factors are marked in red.

4.3 Ablation Study

We conduct an ablation study to evaluate the effectiveness of key designs in the proposed DKCD framework. Specifically, we compare the full DKCD model with three variants. (1) w/o CC removes the causal clues in the knowledge-guided causal reasoning module, and performs causal structure discovery only based on the less reliable factor annotations. (2) w/o LF removes the support of the latent factors in the knowledge-guided causal reasoning module, causing the annotation process in the causal structure discovery module to rely solely on observable factors and lacking causal clues derived from latent factors. (3) w/o Both removes both causal clues and latent factor identification, simulating the case without introducing knowledge graphs.

The experimental results on the Diabetes and Respiratory datasets are shown in Table 3. Firstly, removing causal clues from the knowledge-guided causal reasoning module degrades performance, suggesting that causal clues help refine the causal structure during the causal discovery process. Additionally, we observe that removing latent factor identification results in a substantial decline in performance, particularly in the NF metric, indicating that domain-knowledge graph support is important for discovering meaningful, accurate causal factors. Removing both components (w/o Both) results in the most severe degradation, especially for the generated causal graph structure, demonstrating that the two designs are complementary and their combination is essential for effective causal discovery in high-expertise domains. Overall, the full DKCD framework achieves the best or competitive performance across both datasets, validating the effectiveness of the proposed design.

Table 3: Ablation study of the DKCD framework on the Diabetes and Respiratory datasets.

Dataset	Variant	NF↑	AF↑	ESHD↓
Diabetes dataset	w/o CC	**0.77** ± **0.04**	0.38 ± 0.15	26.67 ± 2.52
	w/o LF	0.33 ± 0.14	0.00 ± 0.00	26.67 ± 2.52
	w/o Both	0.54 ± 0.31	0.10 ± 0.09	34.00 ± 1.73
	DKCD	**0.77** ± **0.04**	**0.40** ± **0.18**	**25.33** ± **3.21**
Respiratory dataset	w/o CC	**0.74** ± **0.07**	0.62 ± 0.10	17.00 ± 2.65
	w/o LF	0.45 ± 0.11	0.32 ± 0.16	22.33 ± 4.51
	w/o Both	0.33 ± 0.15	0.14 ± 0.12	25.33 ± 1.53
	DKCD	**0.74** ± **0.07**	**0.68** ± **0.21**	**16.67** ± **3.21**

5 Conclusions

This paper presents DKCD, a novel knowledge-enhanced framework for causal discovery from unstructured text in high-expertise domains. Specifically, DKCD consists of three main components. First, it is devised with a knowledge mining module to retrieve relevant KG subgraphs. Then, it performs knowledge-guided causal reasoning to discover latent causal factors beyond surface-level factors and generates domain-grounded causal clues for more accurate factor annotation. The resulting higher-quality structured data is then fed into the statistical causal discovery algorithm for final graph construction. Extensive experiments on two domain-specific datasets with four LLM backbones demonstrate that DKCD significantly outperforms existing methods in both causal factor identification and causal graph construction. In future work, we will further explore DKCD's performance across a wider range of domain-specific datasets.

Limitations

Despite integrating LLMs with domain knowledge graphs to extract observable and latent causal factors, hallucinations in LLMs may still undermine the accuracy of causal factor discovery. In future work, we plan to investigate more effective statistical causal discovery methods to better constrain LLM outputs and mitigate hallucination issues.

In addition, our experiments are conducted on domain-specific synthetic datasets. This is because obtaining real-world datasets with reliable ground-truth causal graphs and corresponding domain knowledge graphs is challenging and often requires extensive expert annotation. As a result, the current evaluation may not fully reflect DKCD's performance in real-world settings. In future work, we will further evaluate DKCD on more diverse real-world domain-specific datasets.

Ethical Considerations

In this research work focuses on methodological research for causal graph construction from unstructured textual data. The datasets used in our experiments are either synthetic or derived from publicly available sources, and do not contain personally identifiable information. Additionally, our framework is intended to assist researchers in discovering potential causal structure and should not be used as a substitute for professional judgment, particularly in sensitive domains such as healthcare.Future applications should carefully consider issues related to data privacy, the potential biases inherent in language models, and the responsible generation, interpretation, and use of automatically inferred causal relationships in practical settings.

References

Alessandro Antonucci, Gregorio Piqué, and Marco Zaffalon. 2023. Zero-shot causal graph extrapolation from text via llms. *arXiv preprint arXiv:2312.14670*.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.

Poopak Azad, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. 2020. The role of structured and unstructured data managing mechanisms in the internet of things. *Cluster Computing*, 23(2):1185–1198.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. 2025. Integrating large language model for improved causal discovery. *IEEE Transactions on Artificial Intelligence*.

Mario Bunge. 2017. *Causality and modern science*. Routledge.

Kelly G. Burford, Nathan G. Itzkowitz, Anthony G. Ortega, Julien O. Teitler, and Andrew G. Rundle. 2024. Use of generative ai to identify helmet status among patients with micromobility-related injuries from unstructured clinical notes. *Jama Network Open*, 7(8):e2425981. 10.1001/jamanetworkopen.2024.25981.

Maxwell Caplin and Freda Festenstein. 1975. Relation between lung cancer, chronic bronchitis, and airways obstruction. *British Medical Journal*, 3(5985):678–680.

Dejie Chang, Mosha Chen, Chaozhen Liu, Liping Liu, Dongdong Li, Wei Li, Fei Kong, Bangchang Liu, Xiaobin Luo, Ji Qi, and 1 others. 2021. Diakg: An annotated diabetes dataset for medical knowledge graph construction. In *China Conference on Knowledge Graph and Semantic Computing*, pages 308–314. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. 2024. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 543–554.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Adanma Cecilia Eberendu and 1 others. 2016. Unstructured data: an overview of the data of big data. *International Journal of Computer Trends and Technology*, 38(1):46–50.

Stephen Fancsali. 2014. Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Educational Data Mining 2014*.

Gaël Gendron, Jože M Rožanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models. *arXiv preprint arXiv:2410.06392*.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and Yongjun Xu. 2024. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4):1–38.

Seethalakshmi Gopalakrishnan, Luciana Garbayo, and 1 others. 2024. Causality extraction from medical text using large language models (llms). *CoRR abs/2407.10020*.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. 2016. Causal discovery from subsampled time series data by constraint optimization. In *Conference on Probabilistic Graphical Models*, pages 216–227. PMLR.

Phyllis McKay Illari, Federica Russo, and Jon Williamson. 2011. *Causality in the Sciences*. Oxford University Press.

Yejin Kim, Eojin Kang, Juae Kim, and H Howie Huang. 2024. Causal reasoning in large language models: A knowledge graph approach. *arXiv preprint arXiv:2410.11588*.

Steffen L Lauritzen and David J Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

Jin Li, Shoujin Wang, Qi Zhang, Feng Liu, Tongliang Liu, Longbing Cao, Shui Yu, and Fang Chen. 2025a. Revealing multimodal causality with large language models. *arXiv preprint arXiv:2509.17784*.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. 2025b. Can large language models help experimental design for causal discovery? *arXiv preprint arXiv:2503.01139*.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. 2024. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4669–4677.

Damon Lin, Marzieh Mussavi Rizi, Conor O'Neill, Jeffrey C Lotz, Paul Anderson, and Abel Torres Espin. 2026. Knowledge augmented causal discovery through large language models and knowledge graphs: application in chronic low back pain. *MedRxiv*, pages 2026–02.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. Discovery of the hidden world with large language models. *Advances in Neural Information Processing Systems*, 37:102307–102365.

Zhonghua Liu, Wei Shao, Xia Cao, and 1 others. 2026. Crest: A causal framework for mitigating shortcut learning in language models through counterfactual reasoning. *Information Processing & Management*, 63(2):104418.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. 2022. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303.

Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Ai.meta.com/blog/meta-llama-3/.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, volume 2324, pages 141–147. AISEL.

Iman Mirzadeh, Keivan Alizadeh, and 1 others. 2024. Gsm-symbolic: understanding the limitations of mathematical reasoning in large language models. *CoRR abs/2410.05229*.

Keith Morrison and Greetje van der Werf. 2016. Searching for causality in educational research.

Wenjin Niu, Zijun Gao, Liyan Song, and Lingbo Li. 2024. Comprehensive review and empirical evaluation of causal discovery algorithms for numerical data. *arXiv preprint arXiv:2407.13054*.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1449.

Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. 2012. Hamming distance metric learning. *Advances in Neural Information Processing Systems*, 25.

Ikushi Onozaki, Irwin Law, Charalambos Sismanidis, Matteo Zignol, Philippe Glaziou, and Katherine Floyd. 2015. National tuberculosis prevalence surveys in asia, 1990–2012: an overview of results and lessons learned. *Tropical Medicine & International Health*, 20(9):1128–1145.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. 2017. Financial networks based on granger causality: a case study. *Physica A: Statistical Mechanics and its Applications*, 482:65–73.

Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799.

Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer.

Humairat H Rahman, Danielle Niemann, and Stuart H Munson-McGee. 2023. Association between asthma, chronic bronchitis, emphysema, chronic obstructive pulmonary disease, and lung cancer in the us population. *Environmental Science and Pollution Research*, 30(8):20147–20158.

Joseph Ramsey. 2016. Improving accuracy and scalability of the pc algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*.

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

Jingpu Shi and Beau Norgeot. 2022. Learning causal effects from observational data in healthcare: a review and summary. *Frontiers in Medicine*, 9:864882.

Xiao Shou, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, Oktie Hassanzadeh, and Kristin P Bennett. 2023. Pairwise causality guided transformers for event sequences. *Advances in Neural Information Processing Systems*, 36:46520–46533.

Aisha Siddiqa, Ahmad Karim, and Abdullah Gani. 2017. Big data storage technologies: a survey. *Frontiers of Information Technology & Electronic Engineering*, 18(8):1040–1070.

Alik Sokolov, Fabrizzio Sabelli, Behzad Azadie Faraz, Wuding Li, and Luis Seco. 2025. Toward automating causal discovery in financial markets and beyond. In *Transactions of ADIA Lab: Interdisciplinary Advances in Data and Computational Science*, pages 359–392. World Scientific.

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

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. 2013. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*.

Percy Stocks. 1959. Cancer and bronchitis mortality in relation to atmospheric deposit and smoke. *British Medical Journal*, 1(5114):74.

Yuni Susanti and Michael Färber. 2024. Knowledge graph structure as prompt: Improving small language models capabilities for knowledge-based causal discovery. In *International Semantic Web Conference*, pages 87–106. Springer.

Song Tong, Kai Mao, and 1 others. 2024. Automating psychological hypothesis generation with ai: large language models meet causal graph. *CoRR abs/2402.14424*.

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

Elena Vergou, Ioanna Pagouni, Marios Nanos, and Katia Lida Kermanidis. 2023. Readability classification with wikipedia data and all-minilm embeddings. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 369–380. Springer.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

Shijie Wu, Ozan Irsoy, and 1 others. 2023. Bloomberggpt: a large language model for finance. *CoRR abs/2303.17564*.

xAI. 2025. Grok 3 beta — the age of reasoning agents. X.ai/news/grok-3.

Jing Yang, Ning An, Gil Alterovitz, Lian Li, and Aiguo Wang. 2013. Causal discovery based on healthcare information. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 71–73. IEEE.

Chen Yin and Zixuan Zhang. 2024. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence*, pages 677–684. Atlantis Press.

Guanyuan Yu, Xv Wang, Qing Li, and Yu Zhao. 2024a. Fusing llms and kgs for formal causal reasoning behind financial risk contagion. *arXiv preprint arXiv:2407.17190*.

Kui Yu, Jiuyong Li, and Lin Liu. 2016. A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*.

Xuanqing Yu, Wangtao Sun, Jingwei Li, Kang Liu, Chengbao Liu, and Jie Tan. 2024b. Onsep: a novel online neural-symbolic framework for event prediction based on large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6335–6350.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8.

Wenxiu Zhou and QingCai Chen. 2022. A survey on causal discovery. In *China Conference on Knowledge Graph and Semantic Computing*, pages 123–135. Springer.

Xiaohan Zou. 2020. A survey on application of knowledge graph. In *Journal of Physics: Conference Series*, volume 1487, page 012016. IOP Publishing.

A Algorithm

Algorithm 1: The DKCD Framework

Input: $\mathcal{D}$, $\mathcal{X}_s$, $\mathcal{G}_{KG}$, f_{obs}, M, L, $\mathcal{C}$.
Output: $\mathcal{V}$, $\mathbf{S}$, G.

1. Propose observable factors: $\mathcal{V}_o \leftarrow f_{\text{obs}}(\mathcal{X}_s)$.
2. Initialize $\mathcal{V}_l \leftarrow \emptyset$, $\mathbf{S} \leftarrow [\,]$.
3. For each $\mathbf{x}_k \in \mathcal{D}$:
   (a) Retrieve subgraph: $\mathcal{G}_{rel}^{(n)} \subseteq \mathcal{G}_{KG}$.
   (b) Verbalize subgraph: $\mathcal{S}^{(k)} = \{s_1, \ldots, s_{n_k}\}$.
   (c) Semantic matching: $\mathcal{S}_{v_i}^{(k)} \leftarrow M(v_i, \mathcal{S}^{(k)})$, $\forall v_i \in \mathcal{V}_o$.
   (d) Build matched context: $\widetilde{\mathcal{S}}^{(k)} = \{\mathcal{S}_{v_i}^{(k)} \mid v_i \in \mathcal{V}_o\}$.
   (e) Discover latent factors: $\mathcal{V}_l^{(k)} \leftarrow L(\mathcal{V}_o, \widetilde{\mathcal{S}}^{(k)})$.
   (f) Generate causal clues: $\mathcal{C}^{(k)} \leftarrow L(\mathcal{V}_o, \mathcal{G}_{rel}^{(k)})$.
   (g) Annotate factors: $\mathcal{V}^{(k)} \leftarrow \mathcal{V}_o \cup \mathcal{V}_l^{(k)}$, $\mathbf{s}^{(k)} \leftarrow (v_1^k, \ldots, v_{|\mathcal{V}^{(k)}|}^k)$.
   (h) Update: $\mathcal{V}_l \leftarrow \mathcal{V}_l \cup \mathcal{V}_l^{(k)}$, $\mathbf{S} \leftarrow \mathbf{S} \cup \{\mathbf{s}^{(k)}\}$.
4. Merge factors: $\mathcal{V} \leftarrow \mathcal{V}_o \cup \mathcal{V}_l$.
5. Discover causal graph: $G \leftarrow \mathcal{C}(\mathbf{S}, \mathcal{V})$.
6. Return $\mathcal{V}, \mathbf{S}, G$.

B More Details about Experiments

We construct two medical-domain datasets: the Diabetic Patient Condition Descriptions Dataset (Diabetes Dataset) and the Respiratory Patient Condition Descriptions Dataset (Respiratory Dataset). Each dataset contains 400 patient condition descriptions, a set of high-level factors, and a corresponding domain knowledge graph.

B.1 Dataset Construction Process

B.1.1 Diabetes Dataset

Diabetic Patient Condition Descriptions Generation: We construct a dataset containing diabetic patient condition descriptions following the methodology proposed in (Liu et al., 2024). As shown in Figure 4, we first define a ground truth causal graph, from which a structured scoring table is derived. The FCI algorithm is then applied to the scoring table to reconstruct the causal graph. If the reconstructed graph is consistent with the

ground-truth structure, it is used to generate the corresponding descriptions of diabetic patient conditions.

Diabetes-Domain KG Construction: We construct the diabetes-domain KG using Neo4j (Miller, 2013), a graph database technology, based on the diabetes dataset provided in (Chang et al., 2021).

B.1.2 Respiratory Dataset

Respiratory Patient Condition Descriptions Generation: We construct a dataset containing respiratory patient condition descriptions following the methodology proposed in (Liu et al., 2024). First, a ground truth causal graph is defined, as illustrated in Figure 5, from which a structured scoring table is derived. The FCI algorithm is then applied to the scoring table to reconstruct the causal graph. If the reconstructed graph is consistent with the ground truth structure, it is then used as the basis for generating the corresponding respiratory patient condition descriptions.

Respiratory-Domain KG Construction: Since suitable respiratory-domain knowledge graphs are difficult to obtain, we collect a set of publications in the respiratory domain (Rahman et al., 2023; Stocks, 1959; Lauritzen and Spiegelhalter, 1988; Onozaki et al., 2015; Caplin and Festenstein, 1975). Based on these publications, we employ an LLM (GPT-4o) to extract relational triples in the form of (entity, relation, entity), which serve as the basic structure for KG construction (Pujara et al., 2013; Zou, 2020). The extracted triples are then using Neo4j (Miller, 2013) graph database technology to construct the respiratory-domain KG.

B.2 Details on Prompts

In this section, we provide examples of prompts used in our experiments, including:

- Generating diabetic patient condition descriptions for the diabetes dataset (Prompt B.1).
- Generating respiratory patient condition descriptions for the respiratory dataset (Prompt B.2).
- An example of observable causal factor mining (Prompt B.3).
- An example of knowledge-guided latent causal factor mining (Prompt B.4).
- An example of causal clues generation for causal structure discovery (Prompt B.5).
- An example of the text evaluation prompt (Prompt B.6).

Prompt B.1: Generating Diabetic Patient Condition Descriptions for the Diabetes Dataset

Task:
Please write a clinical description for a patient based on the given diabetic evaluation results.
Evaluation Results: {Diabetic Factors}
Requirement:
- Integrate all provided factors into a clinically plausible patient condition description.
- Embed natural clinical clues that imply the following latent factors without explicitly labeling them:

  - Neuropathy: numbness, tingling, burning sensation, or reduced sensation in the feet.
  - Obesity: elevated body weight, increased BMI, or weight-related mobility limitations.
  - HUA (Hyperuricemia): gout attacks, elevated serum urate, or urate-lowering medication.
  - ChronicKidneyDisease: albuminuria, reduced eGFR, kidney impairment, or long-standing metabolic disease.
  - RiskGenes: family history of diabetes or genetic predisposition to metabolic disorders.

- Single paragraph; No quotation marks; The description should be complete.
- Modern clinical English.
- 150–250 words.
- Only output the patient condition description directly without any other format or content.

Prompt B.2: Generating Respiratory Patient Condition Descriptions for the Respiratory Dataset

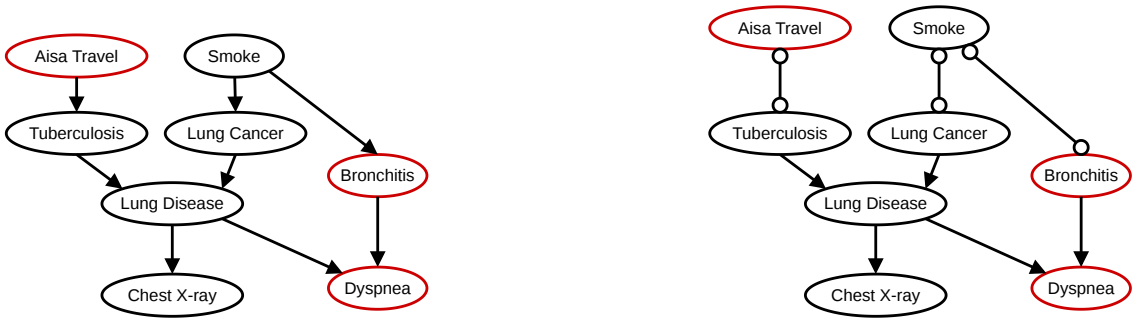
Task:
Please write a clinical description for a patient based on the given respiratory evaluation results.
Evaluation Results: {Respiratory Factors}
Requirement:
- Integrate all provided factors into a clinically plausible patient condition description.
- Embed natural clinical clues that imply the following latent factors without explicitly labeling them:

  - Visit_to_Asia: travel or exposure history related to TB risk.
  - Bronchitis: cough, sputum production, wheezing, or bronchitic symptoms.
  - Dyspnea: degree of shortness of breath or breathing difficulty.

- Single paragraph; No quotation marks; The description should be complete.
- Modern clinical English.
- 150–250 words.
- Only output the patient condition description directly without any other format or content.

(a) Ground Truth **(b) Reconstructed Causal Graph**

Figure 4: Ground truth causal graph and the causal graph reconstructed by the FCI algorithm from the generated structured data derived from the ground truth on the diabetes dataset. Latent factors are marked in red.



(a) Ground Truth **(b) Reconstructed Causal Graph**

Figure 5: Ground truth causal graph and the causal graph reconstructed by the FCI algorithm from the generated structured data derived from the ground truth on the respiratory dataset. Latent factors are marked in red.

Prompt B.3: An Example of Observable Causal Factor Mining

You are an expert clinical analyst specializing in medical record interpretation. You are ready to analyze contrasting patient condition descriptions and identify observable causal factors.
Task:
Patient Condition Descriptions: {Patient Condition Descriptions}
Based on these patient condition descriptions, analyze the underlying differences among causal factors that may contribute to the observed distinctions.
Requirements:
- Identify the key observable factors that differentiate the patient records.
- Each factor should focus on one concrete aspect without semantic overlap.
- Keep the criterion for value 0 fixed as: not exist; or not mentioned.
- Each factor is only allowed to take one value from {-1, 0, 1}.
Output Format:
Your output should contain the following three parts.
Part 1: Consideration
Explain your reasoning process when identifying causal factors from the patient records. Focus on how the textual evidence supports the abstraction of candidate factors.
Part 2: Factor Filtration
Evaluate each candidate factor and decide whether to retain it.
Selection causal factor criteria:

- Each factor should help distinguish different patient conditions.
- Each factor should represent a single concrete aspect.
- Semantic overlap between factors should be minimized.

Part 3: Final Output
Report the final factors using the following template:
{Factor Name}

- - 1: [Positive Criterion]
- - 0: [Not exist; or not mentioned]
- - -1: [Negative Criterion]

Prompt B.4: An Example of Knowledge-guided Latent Causal Factor Mining

Use the patient condition description, relevant domain knowledge, and observable causal factor criteria to identify clinically meaningful **latent causal factors**.
Patient Condition Descriptions: {Patient Condition Descriptions}
Relevant Domain Knowledge: {Relevant Domain Knowl-

edge}
Observable Causal Factor Criteria: {Observable Causal Factor Criteria}
Task:
Based on the patient condition description and the relevant domain knowledge, identify **latent causal factors**. These latent factors should capture deeper causal abstractions rather than directly repeating observable factors.
Requirements

- Latent factors should represent higher-level mechanisms or conceptual causes.
- They must be distinct from observable factors.
- Do not duplicate, paraphrase, or rename observable factors.
- Avoid semantic overlap with observable factors.

Output Format:
Part 1: Consideration
Briefly describe how you abstract latent factors from observable factors and domain knowledge.
Part 2: Factor Filtration
Explain why each candidate factor should be kept or discarded.
Part 3: Final Output
Report the final factors using the following template:
{Factor Name}

- - 1: [Positive Criterion]
- - 0: [Not exist; or not mentioned]
- - -1: [Negative Criterion]

Prompt B.5: An Example of Causal Clues Generation for Causal Structure Discovery

Task:
Using ONLY the factors finalized according to {Patient Condition Description} and {Causal Factor Criteria}, infer causal clues.
Patient Condition Description: {Patient Condition Description}
Causal Factor Criteria: {Causal Factor Criteria}
Relevant Domain Knowledge: {Relevant Domain Knowledge}
Requirements
- Infer causal clues only when there is clear textual evidence from the patient condition description and causal factor criteria.
- Each causal clue should suggest a potential directional relationship between two factors.
- If the relationship is uncertain or unsupported by the text, omit the edge rather than guessing.
Output Format
Part 1: Inferred Causal Clues
List the inferred causal clues in the following format: Source_Factor → Target_Factor
Part 2: Justification
Provide a brief justification for each clue based on the patient's condition description.

Prompt B.6: An Example of Text Evaluation Prompt

You are a highly capable assistant for analyzing, abstracting, and processing textual data.
Task:
Evaluate the following factors based on the provided patient condition description.
Patient Condition Descriptions: {Patient Condition Descriptions}
Each factor must be assigned exactly one value from {-1, 0, 1} according to its evaluation criteria.
Factor Criteria and Causal Clue (used as reference during scoring):
{Factor Criteria} {Causal Clue}
Requirements
- For each factor, select exactly one value from {-1, 0, 1}.
- If the factor is not mentioned in the text, return 0.
- The assigned value must strictly follow the given factor criteria.
Output Format
Return ONLY a single-line compact JSON object:
Rules:

- Output JSON only (no explanations, no markdown, no extra text).
- Keys must match exactly (including spaces and lowercase).
- Values must be integers from {-1, 0, 1}.

Example format: {json_example}

B.3 Implementation Details

We implement DKCD as follows. In the knowledge mining module, we first sample representative patient condition descriptions from different outcome groups. Following the COAT method, we empirically select up to 20 samples per group to form a diverse, balanced input set for the identification of observable factors. For relevant domain knowledge retrieval, we query the Neo4j domain knowledge database with the extracted observable factors to obtain candidate subgraphs, and then apply semantic matching using the pretrained sentence-transformer model all-MiniLM-L6-v2 (Yin and Zhang, 2024; Vergou et al., 2023) to rank retrieved knowledge sentences and retain the top-r most relevant contexts for each factor. Based on the parameter analysis in Table 5, we set $r = 20$ to balance relevant knowledge coverage and contextual noise. In the knowledge-guided causal reasoning module, the LLM takes representative patient condition descriptions, observable factor criteria, and retrieved knowledge contexts as input to infer latent causal factors and causal clues, and further annotates each description in the scoring dataset with a structured scoring vector in $\{-1, 0, 1\}$. In the causal structure discovery module, we apply the FCI algorithm im-

plemented in the causal-learn library (Zheng et al., 2024) to the annotated scoring table to infer the final causal graph structure. All experiments are conducted on a server with two Intel Xeon 6346 CPUs, 256GB RAM, and two NVIDIA A40 GPUs, with all LLMs accessed via API calls.

B.4 Evaluation Metrics

To evaluate the effectiveness of our method, we adopt a set of commonly used metrics for both causal factor identification and causal graph structure discovery, following prior work (Li et al., 2025a; Liu et al., 2024; Zanga et al., 2022).

Factor Identification: We evaluate the accuracy and completeness of the identified causal factors using precision (NP), recall (NR), and F1 score (NF). Since factor names may vary across different methods and models, we manually align semantically equivalent factors with the ground-truth names during evaluation to ensure consistency.

Causal Graph Structure Discovery: The quality of the discovered causal structure is evaluated using adjacency precision (AP), adjacency recall (AR), and their F1 score (AF). We further adopt the extended structural Hamming distance (ESHD) metric (Li et al., 2025a). Compared with the standard SHD metric (Tsamardinos et al., 2006; Norouzi et al., 2012), ESHD additionally accounts for missing and spurious factors and their associated edges.

B.5 Baselines

We provide a detailed list of the representative baselines used in our experiments, including:

- **Zero-shot LLM** (Antonucci et al., 2023; Du et al., 2025) relies solely on the reasoning capability of LLMs without providing any examples or domain knowledge.

- **META** (Liu et al., 2024) is a variant derived from the COAT ablation study, where contextual examples are not provided. It performs zero-shot causal factor proposal based only on the given context using LLMs.

- **COAT** (Liu et al., 2024) is a state-of-the-art LLM-driven framework for causal discovery from unstructured text. It integrates LLM reasoning with traditional causal discovery algorithms to construct causal graphs.

B.6 Full Experimental Results

The complete qualitative comparison of different methods across multiple LLM backbones, including GPT-4o, Gemini 2.5 Flash, Grok-3, and LLaMA 3-70B, on the Diabetes dataset is presented in Figures 6, 7, 8, and 9. These visual results offer a more comprehensive view of the causal graph structures produced by each method under different LLM backbones, thereby facilitating a detailed qualitative assessment of their structural discovery capabilities. Likewise, the corresponding qualitative comparisons on the Respiratory dataset are presented in Figures 10, 11, 12, and 13.

B.7 Full Ablation Study Results

To complement the findings in Section 4.4, we further present detailed ablation results on the Diabetes and Respiratory datasets using GPT-4o as a representative LLM backbone. These results offer a more fine-grained analysis of the role of each component in the proposed framework for causal factor identification and causal structure discovery. Specifically, they allow us to examine how removing individual components affects factor-level performance as well as the structural quality of the inferred causal graphs. The complete results are reported in Table 4.

B.8 Parameter Analysis Results

In this section, we analyze the impact of the parameter r on DKCD performance on the Diabetes dataset using GPT-4o. Here, r denotes the number of top-r relevant knowledge contexts selected for each factor after semantic matching. When r is small, insufficient knowledge is retrieved, leading to lower performance. As shown in Table 5, performance improves as r increases and reaches a relatively good level at around $r = 20$. In contrast, incorporating too much contextual knowledge can introduce excessive noise, thereby degrading DKCD performance. Therefore, we use $r = 20$ in the main experiments.

Table 4: Ablation study of the DKCD framework on the Diabetes and Respiratory datasets.

Dataset	Variant	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
Diabetes dataset	w/o CC	**0.84** ± **0.16**	**0.71** ± **0.08**	**0.77** ± **0.04**	0.49 ± 0.16	0.31 ± 0.13	0.38 ± 0.15	26.67 ± 2.52
	w/o LF	0.51 ± 0.17	0.24 ± 0.11	0.33 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	34.00 ± 1.73
	w/o Both	0.43 ± 0.14	0.24 ± 0.11	0.33 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	34.00 ± 1.73
	DKCD	**0.84** ± **0.16**	**0.71** ± **0.08**	**0.77** ± **0.04**	**0.50** ± **0.17**	**0.33** ± **0.18**	**0.40** ± **0.18**	**25.33** ± **3.21**
Respiratory dataset	w/o CC	**0.74** ± **0.13**	**0.75** ± **0.00**	**0.74** ± **0.07**	0.73 ± 0.14	0.54 ± 0.08	0.62 ± 0.10	17.00 ± 2.65
	w/o LF	0.50 ± 0.17	0.42 ± 0.07	0.45 ± 0.11	**1.00** ± **0.00**	0.21 ± 0.14	0.32 ± 0.16	22.33 ± 4.51
	w/o Both	0.38 ± 0.16	0.30 ± 0.14	0.33 ± 0.15	0.33 ± 0.29	0.09 ± 0.08	0.14 ± 0.12	25.33 ± 1.53
	DKCD	**0.74** ± **0.13**	**0.75** ± **0.00**	**0.74** ± **0.07**	0.87 ± 0.23	**0.56** ± **0.22**	**0.68** ± **0.21**	**16.67** ± **3.21**

Table 5: Impact of the top-r relevant knowledge contexts selected during semantic matching in the knowledge mining module on the performance of DKCD.

Dataset	r	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
Diabetes dataset	5	0.33 ± 0.17	0.14 ± 0.07	0.20 ± 0.10	0.02 ± 0.03	0.03 ± 0.05	0.05 ± 0.08	36.67 ± 2.08
	10	0.55 ± 0.15	0.26 ± 0.09	0.36 ±0.11	**0.83** ± **0.29**	0.13 ± 0.04	0.21 ± 0.06	32.00 ± 2.65
	20	**0.84** ± **0.16**	**0.71** ± **0.08**	**0.77** ± **0.04**	0.50 ± 0.17	**0.33** ± **0.18**	**0.40** ± **0.18**	**25.33** ± **3.21**
	50	0.61 ± 0.03	0.41 ± 0.15	0.48 ± 0.11	0.39 ± 0.24	0.16 ± 0.13	0.22 ± 0.17	33.00 ± 1.00

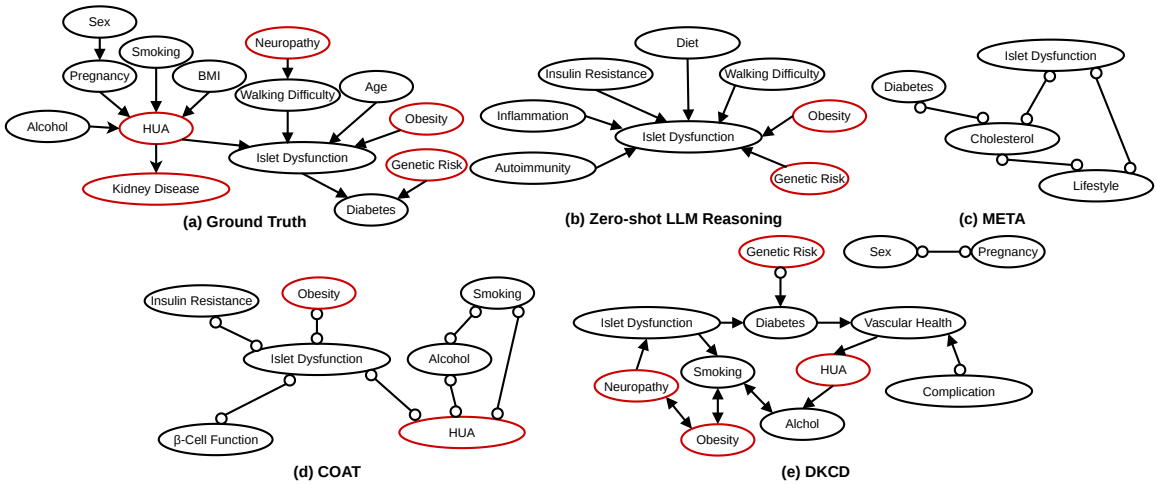


Figure 6: Causal graphs discovered with GPT-4o on the Diabetes dataset. Latent factors are marked in red.

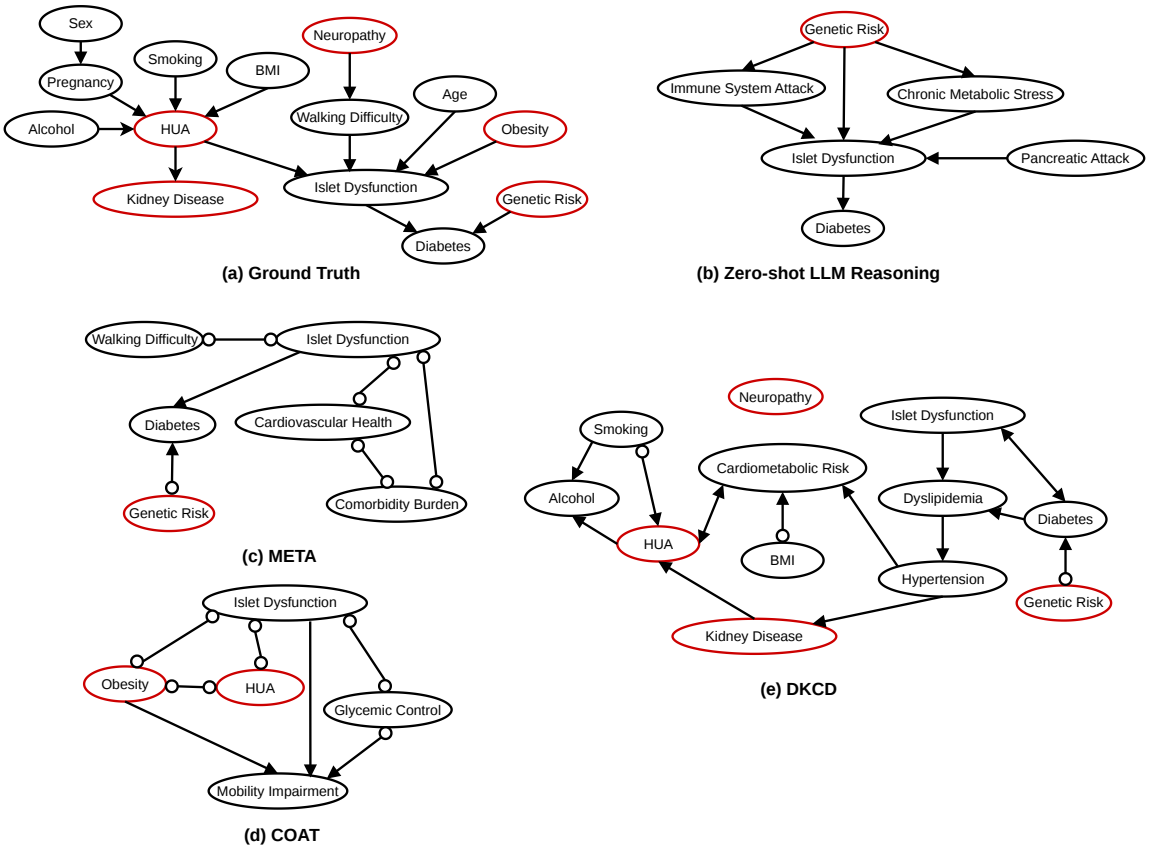


Figure 7: Causal graphs discovered with Gemini 2.5 Flash on the Diabetes dataset. Latent factors are marked in red.

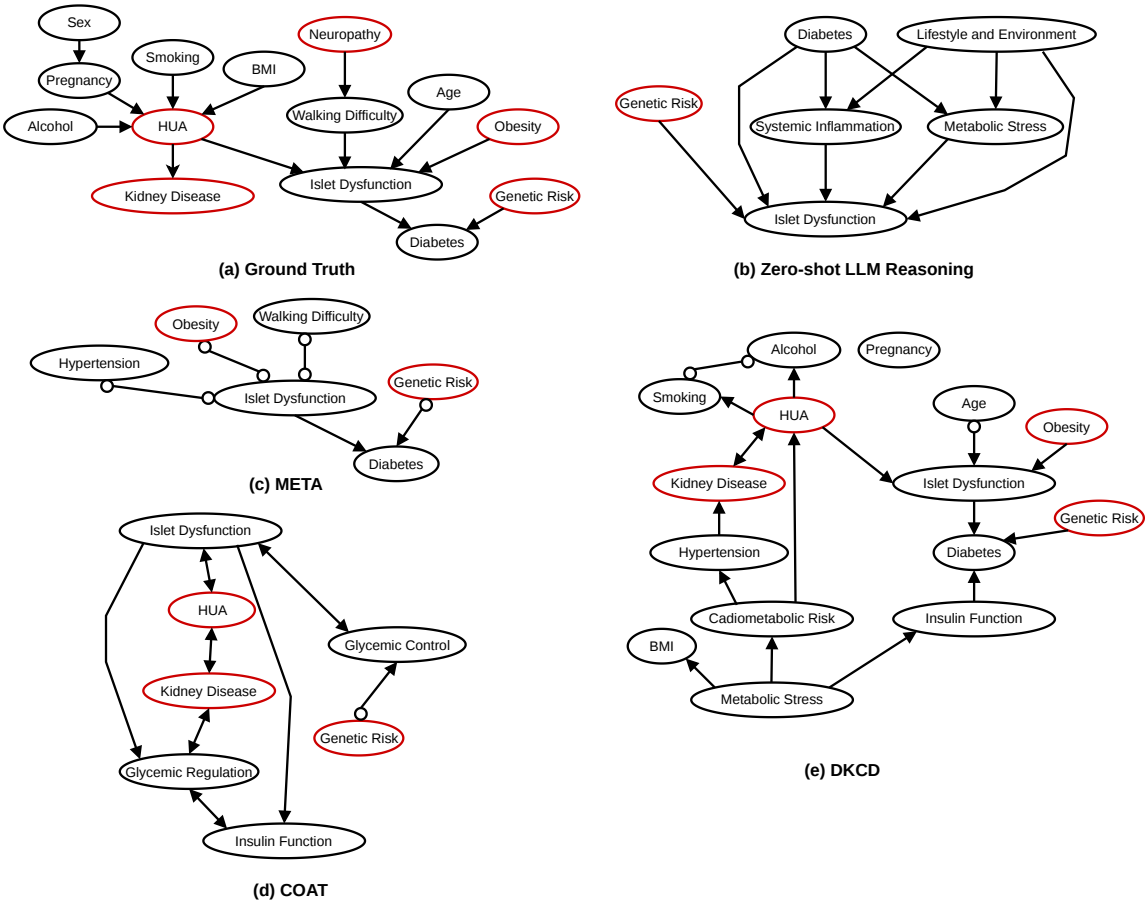


Figure 8: Causal graphs discovered with Grok-3 on the Diabetes dataset. Latent factors are marked in red.

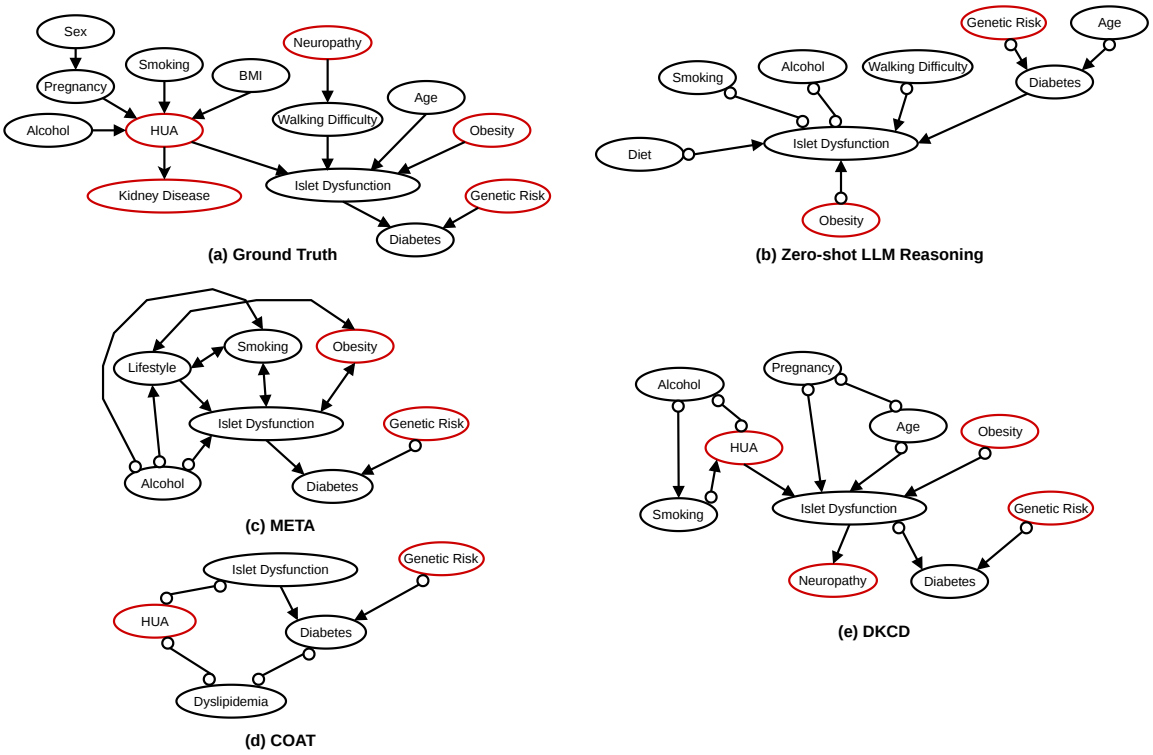


Figure 9: Causal graphs discovered with LLaMA 3-70B on the Diabetes dataset. Latent factors are marked in red.

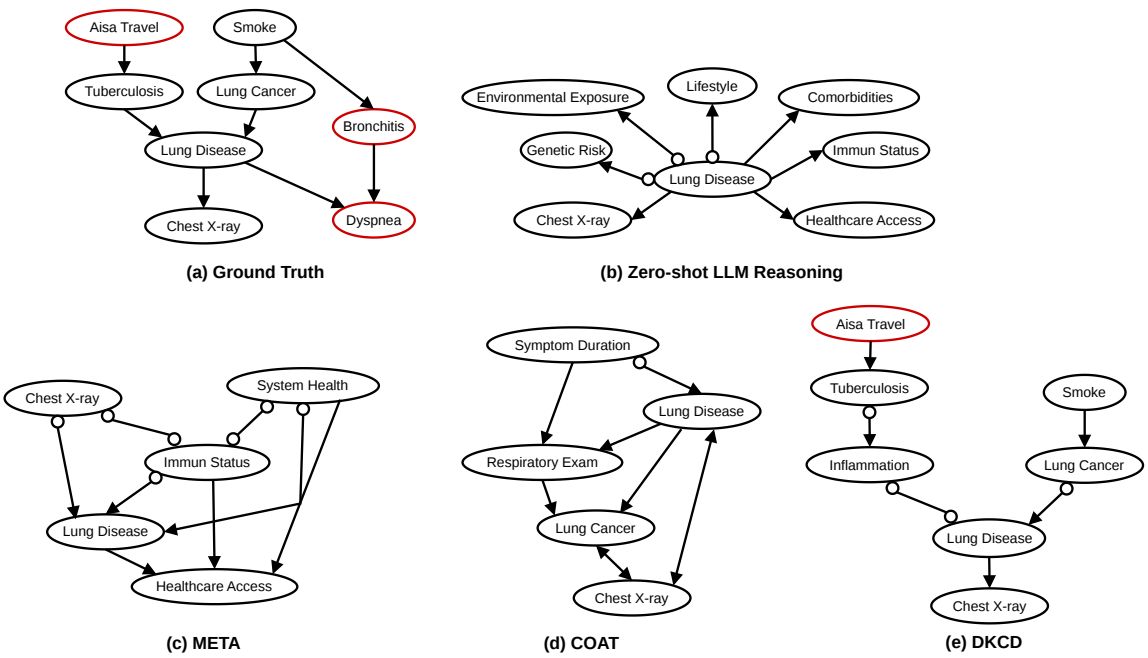


Figure 10: Causal graphs discovered with GPT-4o on the Respiratory dataset. Latent factors are marked in red.

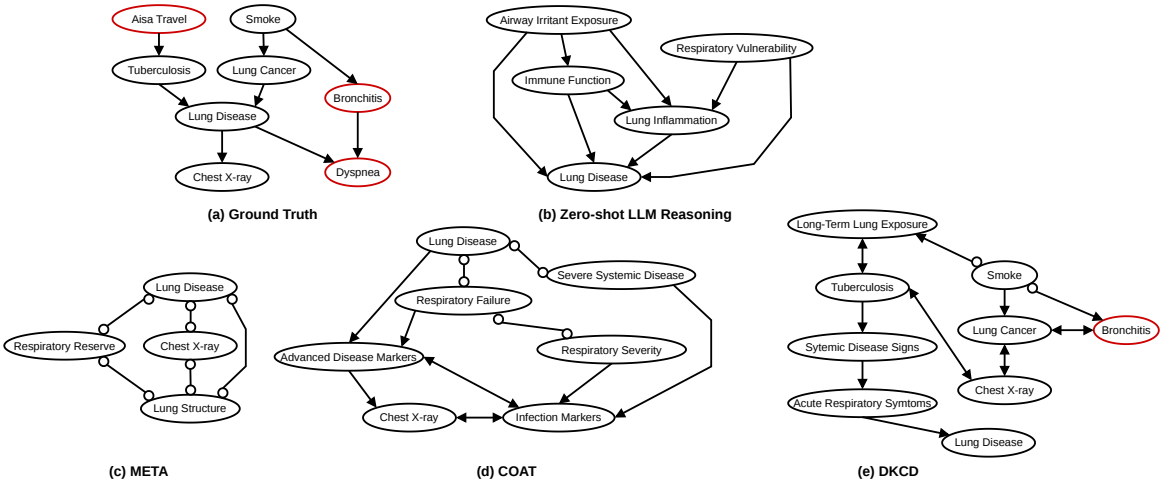


Figure 11: Causal graphs discovered with Gemini 2.5 Flash on the Respiratory dataset. Latent factors are marked in red.

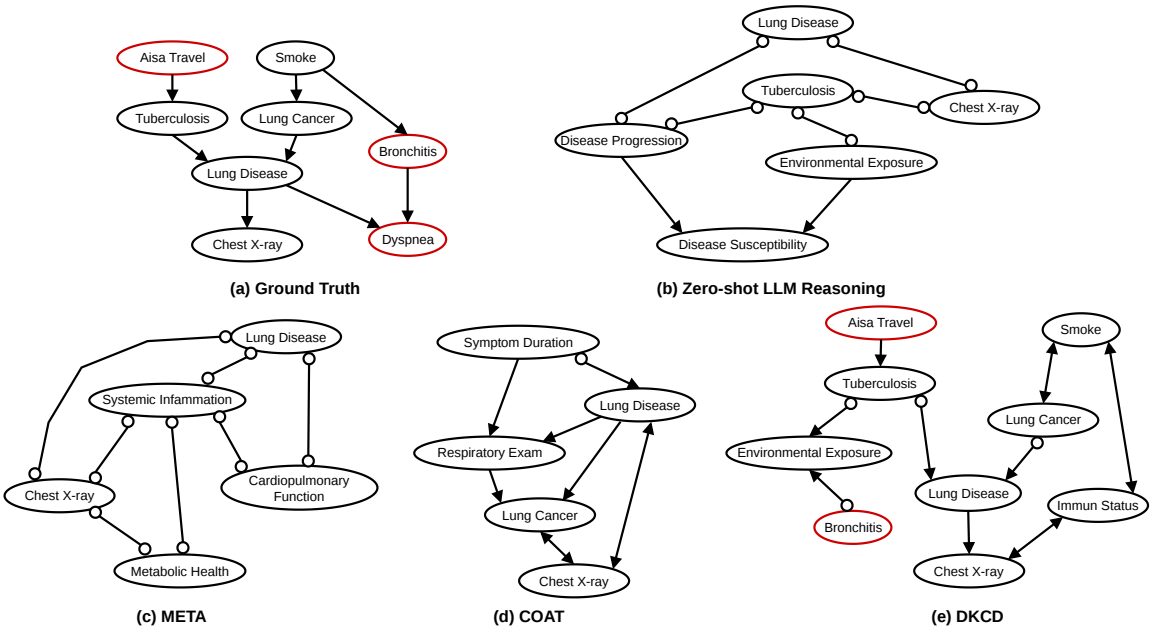


Figure 12: Causal graphs discovered with Grok-3 on the Respiratory dataset. Latent factors are marked in red.

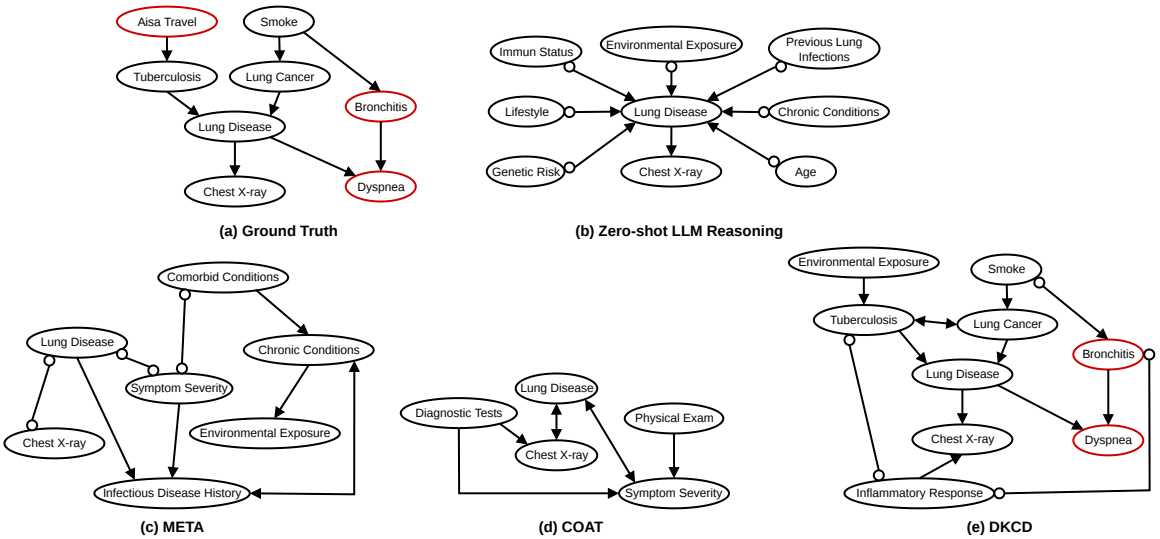


Figure 13: Causal graphs discovered with LLaMA 3-70B on the Respiratory dataset. Latent factors are marked in red.