CONSISTENCY EVALUATION OF BENCHMARKS USED FOR CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

In graphical causal model, causal discovery aims to construct a causal graph based on numerical data and domain knowledge in plain text. However, the evaluation of causal discovery methods remains a challenge in the area as the progress of domain researches often makes benchmark causal graphs contain misaligned knowledge. This problem especially affects the evaluation of large language model (LLM) based causal discovery methods as they are sensitive to the new discoveries in the literature. This work is the first to systematically study the quality of benchmark causal graphs. Specifically, we design a pipeline that automatically retrieves relevant research papers from scientific databases, and prompts LLMs to check the consistency between the benchmark causal graphs and domain research papers. We evaluate 10 popular empirical benchmarks, for which our pipeline in total proceeds 35,340 domain papers. Our results show that popular benchmarks vary significantly in their consistency with domain research, with clear implications for causal discovery research.

1 INTRODUCTION

Causal discovery Glymour et al. (2019) aims to identify causal links between variables, serving as a fundamental research direction in both artificial intelligence Zheng et al. (2018); Roy et al. (2025) and application domains such as medical science Petersen et al. (2010), biology Sachs et al. (2005), and other fields Huang et al. (2021).

Our work considers the graphical model-based causal discovery problem for time-independent datasets. The target of this problem is to construct a causal graph, represented as a (mixed) directed graph, based on numerical data (statistical-based methods) or domain expertise. Numerous methods have been proposed to accomplish such tasks, e.g., PC Spirtes et al. (2001), GAS Chickering (2002), and more recent ones such as NoTears Zheng et al. (2018). These algorithms achieve impressive performance when evaluated against synthetic or real-world benchmarks. However, a crucial dilemma regarding evaluation benchmarks obscures the effectiveness of these methods in handling real-world challenges: A high-quality real-world benchmark dataset should contain a convincing ground-truth causal graph, but obtaining such a graph is challenging in many domains. As a result, many emerging causal discovery works rely on old empirical benchmarks, e.g., Spiegelhalter (1992), or semi-synthetic benchmarks, e.g., Tu et al. (2019), in their evaluation process. Gentzel et al. (2019) and Brouillard et al. (2025) reveal this issue in the causal discovery domain and advocate building and using new benchmarks.

Recent works, e.g., Zhang et al. (2024), reveal that some popular real-world benchmarks[1] contain out-of-date domain knowledge, potentially producing evaluation results that are inconsistent with real-world truth. This problem especially undermines the recently emerging stream of large language model (LLM)-based causal discovery methods. Unlike statistical-based methods that heavily rely on the alignment between the numerical data and the ground-truth causal graph, LLM-based methods explore knowledge in a "world model" Hao et al. (2023) to construct the causal graph. Therefore, using such practically non-aligned benchmarks for LLM-based methods especially drives the evaluation in an undesirable direction.

[1] A causal discovery benchmark typically consists of a causal graph and a corresponding numerical dataset. This work evaluates the consistency of benchmark causal graphs with domain knowledge; for simplicity, we use *benchmark* to refer to its causal graph unless otherwise specified.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A motivation example. *Figure 4 shows the causal graph of a popular causal discovery benchmark, Asia lau (1988), which is used in many LLM-based works Shen et al. (2025); Vashishtha et al. (2025); Jiralerspong et al. (2024); Takayama et al. (2024). These LLM-based methods achieve high performance against the Asia benchmark, usually fully or almost fully recovering the causal graph. However, domain research suggests that many of the benchmark's causal/conditional association relationships need to be updated. For instance, deep learning techniques can now reveal bronchitis via X-ray scans Chen et al. (2020); Ntiamoah et al. (2021); Nishino et al. (2014), contradicting the benchmark, e.g., X-ray and bronchitis are mutually independent conditioned on smoking. Therefore, high evaluation performance against such an out-of-date benchmark instead reveals problems with the method in practice.*

In this work, we intend to examine the quality of popular empirical causal discovery benchmarks. First, we review time-independent causal discovery works published in AI venues in recent years and summarise the usage of empirical benchmarks in these works. As the benchmark issue strongly influences LLM-based causal discovery methods, we also show the trend of this emerging stream based on metadata from arXiv. The number of LLM-based works has drastically increased since 2023, highlighting the necessity of systematically evaluating causal discovery benchmarks.

Second, we investigate the alignment between the causal graphs of these popular empirical causal discovery benchmarks and the corresponding domain research. The local Markov property of causal graphs entails a series of conditional association relationships between variables (following the d-separation concept), which provide a natural connection to controlled experimental analyses in most domain research works. We design an LLM-equipped pipeline to verify the consistency between the benchmarks and domain research papers. In practice, we evaluate 10 empirical benchmarks that are the most frequently used among the 69 surveyed benchmarks, and we analyse 35,340 research papers to mine scientific evidence that supports or contradicts the benchmarks.

Our results show that these popular benchmarks vary in their alignment with the corresponding domain research. Some popular benchmarks are of high quality, e.g., the Sachs benchmark, whilst others present a high degree of misalignment with domain knowledge, e.g., the Child and Asia benchmarks. These findings can guide causal discovery works to better select benchmarks in their evaluation process.

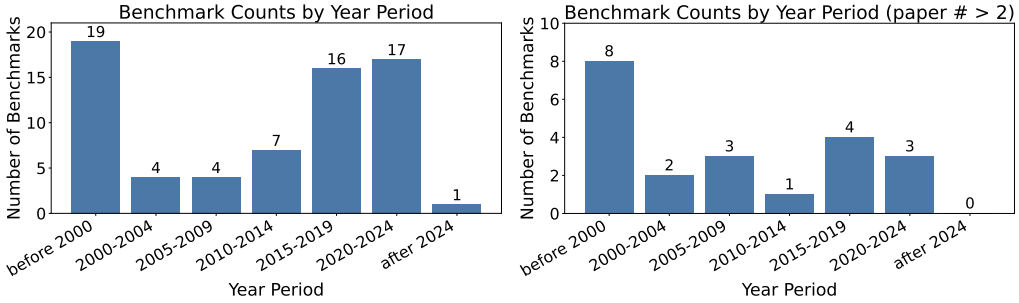
Related works. Evaluation benchmarks are basically divided into three types: purely synthetic, semi-synthetic (synthetic data based on real-world causal graphs), and real-world. Brouillard et al. (2025) provides a summary of the usage of the three types of benchmarks and reviews the real-world and semi-synthetic benchmarks. Prior to this survey, Gentzel et al. (2019) summarises evaluation methods and the usage of different types of benchmarks in the domain. They advocate using real-world benchmarks in evaluation processes and choosing interventional data rather than observational data. Similarly, Glymour et al. (2019) encourages using semi-synthetic data to better align with practice, compared with using purely synthetic data.

Our work focuses on benchmarks that are widely recognised as empirical benchmarks by the causal discovery community, as the community generally regards human judgment as playing an important role in constructing such benchmarks. These benchmarks are supposed to reflect real-world ground truth and are thus intended to be either semi-synthetic (e.g., Tu et al. (2019); lau (1988)) or real-world (e.g., Sachs et al. (2005)). We summarise the detailed usage of such benchmarks, particularly with an update on benchmark usage in LLM-based causal discovery methods, and evaluate their alignment with real-world knowledge.

To the best of our knowledge, we are the first to systematically evaluate empirical benchmarks from the perspective of alignment with domain research. Our survey updates benchmark usage in causal discovery, particularly the popularity of benchmarks and their usage in emerging LLM-based causal discovery works. We defer the related work on LLM-based causal discovery to Appendix B.

2 CAUSAL DISCOVERY BENCHMARK REVISIT

This section surveys empirical benchmarks used in causal discovery papers published at recent major AI conferences and characterises the trend of LLM-based causal discovery papers by mining the metadata retrieved from the arXiv repository.



(a) All benchmarks by year period.

(b) Frequently used benchmarks by year period.

Figure 1: Distribution of benchmarks across year periods.

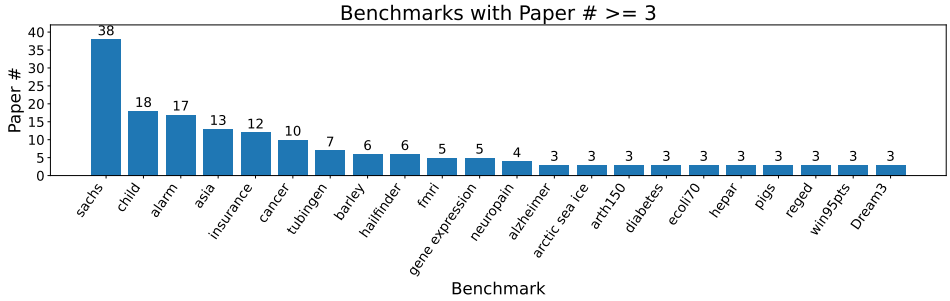


Figure 2: Number of papers that use frequently used benchmarks.

2.1 Benchmark usage in AI publications

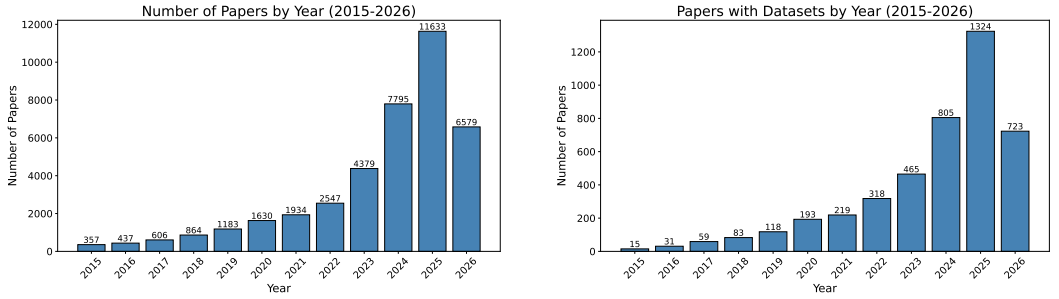
Paper selection. We consider causal discovery papers based on time-independent datasets from major AI venues, namely, NeurIPS, ICML, ICLR, AAAI, IJCAI, ACL, and EMNLP, published during 2021–2025, following the standard PRISMA guidelines (see details in Appendix C.1). Through a two-stage screening process, we include 86 papers in this survey section and identify 69 unique empirical benchmarks (see the full list in Appendix C.2).

Benchmark age in publications. We are interested in the distribution of the construction years of the 69 benchmarks, potentially revealing the timeliness of the evaluation process in the causal discovery domain. We review the construction year of each of the 69 benchmarks. Figure 1a presents the number of benchmarks constructed in each year period[2]. Among the 69 benchmarks, 19 were published before 2000, showing that the domain still uses many old benchmarks. Since 2000, new benchmarks have been increasingly constructed. A total of 34 new benchmarks have emerged since 2015, indicating that the community has become aware of the insufficient benchmark problem in recent years.

Figure 1b shows the number of frequently used benchmarks (i.e., those used in at least three publications) constructed in each year period. 21 out of the 69 benchmarks are frequently used, with Figure 2 showing the number of surveyed publications that select each benchmark in their method evaluation. The distribution of the frequently used benchmarks is relatively even over the years. Researchers still tend to use many of the old benchmarks, with 8 out of 21 constructed before 2000, while 7 out of 21 were constructed after 2014. The usage frequencies of these 21 benchmarks vary significantly. Sachs is the most popular benchmark, being used in 38 papers. All of the top 9 benchmarks, except the pairwise Tübingen dataset, are recorded in the bnlearn network repository Scutari et al. (2019), revealing its popularity in the domain. 15 of the 21 popular benchmarks are from the domains of medical science and biology.

LLM-based methods LLMs have become a tool for causal discovery in recent years. Among the 96 papers, 10 propose LLM-based causal discovery methods. These papers were published very

[2]Figure 1a excludes the Carpo benchmark, since the original publication source is missing.



(a) Number of papers per year found by searching "causal discovery" and "LLM".

(b) Number of papers per year found by searching "causal discovery", "benchmark", and "LLM".

Figure 3: Distribution of arXiv papers over years.

recently, with 9 of them published since 2025. That is, LLM-based methods account for 31% of the surveyed causal discovery papers published after 2024 (29 papers in total). There are 28 distinct benchmarks used in these 10 papers, indicating that these papers use diverse benchmarks (see details in Appendix C.3). However, the top 3 most frequently used benchmarks in LLM-based papers are Asia, Cencer, and Sachs, all of which are from the bnlearn small network repository.

2.2 Emerging LLM-based causal discovery works

arXiv search process. As new papers often appear on arXiv first, analysing their benchmark usage can reveal trends in this area. We perform a targeted bibliometric analysis of arXiv papers at the intersection of causal discovery and LLMs. Specifically, we form a search query using two keyword sets: one for causal discovery and one for LLMs. Then, we submit the query to the arXiv API to search for papers whose titles or abstracts contain keywords from both sets, with the aim of retrieving LLM-based causal discovery papers. Furthermore, we form another search query with an additional keyword set, "empirical dataset", aiming to retrieve LLM-based causal discovery papers that evaluate their methods with empirical benchmark datasets. We use these two search results as proxies for LLM-based causal discovery works and LLM-based causal discovery works with empirical benchmark evaluation.

We also extract benchmark usage information contained in the abstracts through rule-based matching against an extended benchmark inventory constructed by LLMs (details shown in Appendix C.4).

Summary. Figures 3a and 3b show the yearly numbers of papers on LLM-based causal discovery and LLM-based causal discovery with empirical dataset-based evaluation, respectively. The two numbers show a similar trend: both steadily increase from 2015 to 2025, with the increases becoming more pronounced since 2022. This trend is strongly driven by the rapid advancement of LLMs in recent years. In 2025, there are up to 11,633 LLM-based causal discovery papers, among which up to 1,324 papers contain evaluations based on empirical benchmarks.

In Appendix C.4, the arXiv search results and our surveyed benchmarks overlap on four benchmarks: Link, Cancer, Water, and Child. Three of them are recorded in the bnlearn repository, showing its popularity.

3 Domain knowledge consistency evaluation

3.1 Preliminary: graphical causal models

Causal graph. A causal graph is a directed acyclic graph (DAG), denoted as $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{V}$ is the set of variables and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed edges. For each pair of variables $V_i, V_j \in \mathbf{V}$, (V_i, V_j) being present in $\mathbf{E}$ denotes the existence of an directed edge from V_i to V_j, showing that V_i is a *direct cause* of V_j. An ordered sequence of distinct variables $(V_{i1}, V_{i2}, \cdots, V_{ik})$ from $\mathbf{V}$ forms a path if each adjacent pair $V_{ij}, V_{i(j+1)}$ is adjacent in $\mathcal{G}$, where $1 \leq j \leq k-1$, and the path is a directed path if $(V_{ij}, V_{i(j+1)})$ is in $\mathbf{E}$. By definition, a DAG does not contain any cycle, i.e.,

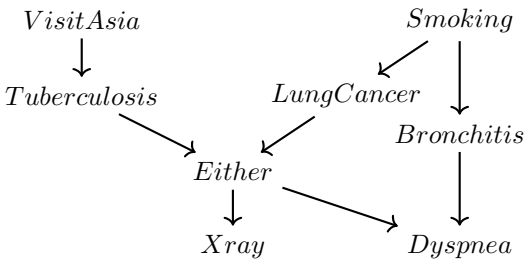


Figure 4: The causal graph of the Asia benchmark dataset: a semi-synthetic benchmark dataset.

a path with the first and the last variables coinciding. For each variable $V_i \in \mathbf{V}$, let $pa_{\mathcal{G}}(V_i)$ denote the parents of V_i in $\mathcal{G}$, i.e., $pa_{\mathcal{G}}(V_i) = \{V \in \mathbf{V} \mid (V, V_i) \in \mathbf{E}\}$, and let $de_{\mathcal{G}}(V_i)$ (resp. $nd_{\mathcal{G}}(V_i)$) denote the descendants (resp. non-descendants) of V_i in $\mathcal{G}$, i.e., all variables that are reachable (resp. unreachable) through a directed path from V_i. For a triple $V_i, V_j, V_k \in \mathbf{V}$ such that $V_i \to V_j \leftarrow V_k$ and V_i and V_k are not adjacent in $\mathcal{G}$, we call V_j a collider.

A ubiquitous assumption in our work is the local Markov property. That is, for each variable $V \in \mathbf{V}$, it holds that: $V \perp nd_{\mathcal{G}}(V) \mid pa_{\mathcal{G}}(V)$, where $\perp$ denotes statistical independence.

d-separation. The local Markov property entails a series of conditional association relationships between variables. The definition of *d-separation* characterises such conditional association relationships reflected in a causal graph.

Definition 3.1 (d-separation) *For a pair of variables $V_i, V_j \in \mathbf{V}$, a variable set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ d-separates V_i and V_j if the following hold:*
(1) $\mathbf{Z}$ contains at least one variable on each collider-free path between V_i and V_j; and
(2) If a path between V_i and V_j contains a collider, the collider and its descendants are not contained in $\mathbf{Z}$, or the path is blocked by a non-collider node on the path.

If $\mathbf{Z}$ d-separates V_i and V_j, then it holds that: $V_i \perp V_j \mid \mathbf{Z}$.

In a DAG causal graph, d-separation specifies all conditional association relationships for any triple V_i, V_j and $\mathbf{Z} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$: if $\mathbf{Z}$ d-separates V_i and V_j, the pair of variables are mutually independent conditioned on $\mathbf{Z}$; otherwise, they are correlated. We denote the conditional association relationship of the triple as $V_i - V_j|\mathbf{Z}$. Note that $\mathbf{Z}$ can be empty.

Example 3.1 *Figure 4 shows the causal graph of the Asia benchmark dataset from the BNLearn repository Scutari et al. (2019). Observe that variables $Xray$ and $Bronchitis$ are d-separated by $Smoking$, and therefore, we have $Xray \perp Bronchitis \mid Smoking$.*

On the other hand, if two variables are adjacent, e.g., $Smoking$ and $LungCancer$, they cannot be d-separated by any variable set, and therefore, they remain correlated, even when conditioned on any other variable set.

Our benchmark evaluation target. The main objective of this work is to evaluate the consistency of popular benchmarks used in causal discovery publications with the corresponding domain research. A typical original domain research work yields a conclusion by designing a controlled experiment, i.e., defining the target variable pair (e.g., studying whether variable V_i has any influence on the corresponding observation of variable V_j). A convincing controlled experiment usually specifies a set of covariates to avoid bias. With data collection and statistical/machine learning analysis, the conclusion reflects the association relationship between the target variable pair conditioned on the covariates. This naturally connects with the d-separation concept, providing an interface to verify the consistency between a domain research work and a causal graph.

Specifically, for a causal graph, we systematically retrieve research papers from scientific databases that are relevant to the conditional association relationships (determined by the d-separation criteria) reflected in the causal graph, and then evaluate the proportion of the retrieved papers that contain conclusions inconsistent with the causal graph's conditional association relationships. This inconsistency rate measures the misalignment between the causal graph and the domain research.

3.2 METHODOLOGY

We design an LLM-equipped pipeline to evaluate the consistency of causal graphs with domain research. The pipeline is composed of three components: (1) domain research paper retrieval, (2) conditional association relationship extraction, and (3) consistency verification. We mainly utilise LLMs to verify the consistency between paper conclusions and the causal graph's conditional association relationships in component (3). Given a causal graph $\mathcal{G}$, for *each pair* of variables $V_i, V_j \in \mathbf{V}$, the pipeline executes as follows.

Domain research paper retrieval. The first component constructs a research paper pool in which all papers study the relationship between V_i and V_j, by searching for and downloading papers from a scientific database and extracting text from each paper. Basically, we send a query "V_i and V_j" to search for relevant papers in a database, download the accessible papers among the search results, and extract the text from the downloaded files to make the papers ready for LLM reading.

A simple search query "V_i and V_j" is not sufficient for efficient retrieval due to the following facts: (i) Inconsistent term usage is common across various research domains, e.g., the term "Smoking" in Example 3.1 may be expressed as "Tobacco use" or "Cigarette consumption". (ii) Many databases use keyword matching against the full paper in their search engines, and the search query may retrieve a large number of irrelevant papers. Hence, we employ the *query expansion* Carpineto & Romano (2012) technique to form a search query and add an extra filtering process to screen out irrelevant papers.

The query expansion finds a set of synonyms for each of V_i and V_j, i.e., $\{V_i^1, \cdots, V_i^{ki}\}$ and $\{V_j^1, \cdots, V_j^{kj}\}$, and then forms a query:

$$(V_i^1 \text{ OR } V_i^2 \text{ OR } \cdots \text{ OR } V_i^{ki}) \text{ AND } (V_j^1 \text{ OR } V_j^2 \text{ OR } \cdots \text{ OR } V_j^{kj}).$$

From the returned search results, the filter screens in papers whose abstracts contain terms from both sets $\{V_i^1, \cdots, V_i^{ki}\}$ and $\{V_j^1, \cdots, V_j^{kj}\}$. Then, we download the PDF files of all open-access papers by interacting with the database's API and convert the downloaded files to JSON files. In brief, we expect that the retrieved papers study the relationship between V_i and V_j by analysing a dataset collected from experiments that control for a set of covariates $\mathbf{Z}$. Such content reveals the association relationship between V_i and V_j conditioned on $\mathbf{Z}$, i.e., $V_i - V_j|\mathbf{Z}$.

Conditional association extraction. Algorithm 1 (pseudocode in Appendix D.1) deterministically extracts the conditional association relationships from the causal graph $\mathcal{G}$. Intuitively, given $V_i, V_j \in \mathbf{V}$, Algorithm 1 checks whether each variable set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ d-separates V_i and V_j (Definition 3.1) and stores all variable sets that d-separate V_i and V_j. Therefore, it extracts the conditional association relationships from $\mathcal{G}$, denoted as $V_i \overset{\mathcal{G}}{-} V_j|\mathbf{Z}$.

Consistency verification. With the paper contents and the extracted $V_i \overset{\mathcal{G}}{-} V_j|\mathbf{Z}$ prepared, we employ LLMs to verify whether each paper's conclusion is consistent with each $V_i \overset{\mathcal{G}}{-} V_j|\mathbf{Z}$. For a paper studying the relationship between V_i and V_j, the LLM verifier extracts the conditional association relationship from the paper, denoted as $V_i \overset{\text{paper}}{-} V_j|\mathbf{Z}$, and verifies whether $V_i \overset{\mathcal{G}}{-} V_j|\mathbf{Z}$ and $V_i \overset{\text{paper}}{-} V_j|\mathbf{Z}$ are consistent. Specifically, the LLM-based conditional association relationship extraction first checks: (1) whether the paper contains statistical analysis results that reveal the association relationship between V_i and V_j; and (2) whether the dataset is collected from experiments that control for a set of covariates $\mathbf{Z}$. Then, it extracts $V_i \overset{\text{paper}}{-} V_j|\mathbf{Z}$ and verifies the consistency.

In practice, the LLM verifier returns "Unknown" if: (1) the paper does not contain data analysis revealing the relationship between V_i and V_j; or (2) $\mathbf{Z} \notin 2^{\mathbf{V}\setminus\{V_i,V_j\}}$. It returns "Consistent/Inconsistent" if $V_i \overset{\mathcal{G}}{-} V_j|\mathbf{Z}$ and $V_i \overset{\text{paper}}{-} V_j|\mathbf{Z}$ are consistent/inconsistent.

3.3 PRACTICAL IMPLEMENTATION

Among the 69 benchmarks (recall the benchmark survey in Section 2.1), we only evaluate the popular benchmarks, i.e., those used in at least 3 causal discovery papers. There are 21 popular benchmarks in total, and we exclude 11 benchmarks: (i) the pairwise benchmark (Tübingen) due to its

benchmarks	# nodes	# edges	# node pairs	# retrieved papers	publication year
sachs	11	17	55	4197	2005
child	20	25	190	4073	1992
alarm	37	46	200	12736	1989
asia	8	8	28	1413	1988
insurance	27	52	200	3408	1997
cancer	5	4	10	319	2010
fmri	6	7	15	2367	2015
alzheimer	11	19	55	5616	2010
arctic sea ice	12	40	66	784	2021
diabetes	4	4	6	427	1991

Table 1: Details of selected benchmark datasets.

simple conditional association type; (ii) the large benchmarks (namely, Hailfinder, Gene expression, Neuropain, Arth150, Hepar2, Pigs, Reged, Barley, WIN95PTS, and Ecoli70) due to our limited computational capacity.[3]. Table 1 lists the details of the evaluated benchmarks.

An n-variable benchmark has $\binom{n}{2}$ combinations of variable pairs. For each variable pair, we limit the database search engine to return only up to 500 of the most relevant papers and download the open-access papers among the search results. Among the 10 evaluated benchmarks, some have more than 200 combinations of variable pairs, and we randomly sample 200 variable pairs for evaluation. Therefore, the pipeline processes up to 100000 full papers for each benchmark[4].

Most of the benchmark causal graphs are from medical science and biology. For such causal graphs, we use PubMed[5] as the scientific database to search for papers, and we download papers that are available in the open-access database PMC. For causal graphs from other domains, we use OpenAlex to search for papers and download papers accessible through the institution's subscription databases. The downloaded PDF files are converted into JSON files using the Mistral OCR API. Throughout the entire process, our pipeline processes 35,340 full papers (see details in Table 1).

For variables from the medical science and biology, we expand the search query by finding 15 synonyms from both of the two resources: (1) the UMLS system and (2) ChatGPT. For variables from other domains, we only use ChatGPT to find 15 synonyms.

We deploy the Qwen3-30B-A3B-Thinking-2507 model on a single H100 GPU with 80GB memory to verify the consistency between each paper content and the d-separation query.

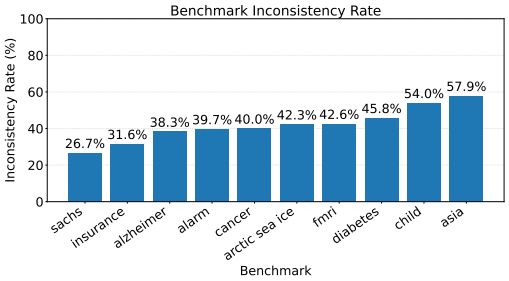
LLM consistency verification accuracy. To show the accuracy of the LLM-based consistency verifier, we randomly sample 100 retrieved papers and manually check whether each answer generated by the LLM consistency verifier is correct. For instance, for a "Consistent" answer generated by the LLM to the question "Is paper0's result consistent with the conditional association relationship $A \perp B \mid C$?", we manually read the paper and check whether the data analysis results of paper0 indicate that A and B are mutually independent in experiments controlling for covariate C. Our manual evaluation shows that the LLM consistency verifier achieves an accuracy of 0.9 based on the sampled 100 papers. See Appendix D.3 for details.

The main metric: inconsistency rate. Recall that investigating each paper returns one of three answers: "Unknown", "Consistent", or "Inconsistent". The answer "Unknown" basically indicates that the paper does not contain sufficient information to support a decision on the conditional as-

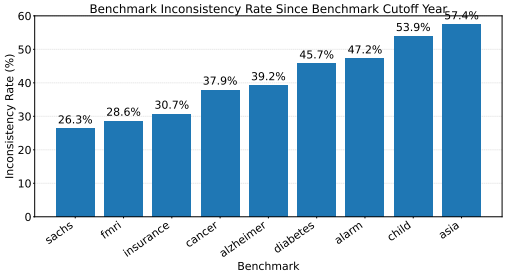
[3]The computational complexity of finding all conditional association queries for a given pair of variables is $O(2^n)$. We set a threshold of $n = 45$, beyond which the enumeration of conditional relationships becomes computationally costly. Moreover, due to the limited resource for deploying LLMs, we only sample a limited number of variable pairs for evaluation. When the graph is large, we are more likely to sample distant variable pairs that have weak connections in practice. This makes the evaluation process less meaningful.

[4]The actual number of papers is much smaller than 100000, since (1) some variable pairs have few search results due to their weak relationship, and (2) only a limited number of papers have open-access full content.

[5]Tools: (1) PubMed: `https://pubmed.ncbi.nlm.nih.gov/`, (2) OpenAlex: `https://openalex.org/`, (3) Mistral OCR: `https://mistral.ai/news/mistral-ocr/`, (4) UMLS: `https://www.nlm.nih.gov/research/umls/index.html`.



(a) Benchmarks' inconsistency rates.



(b) Benchmarks' inconsistency rates since the construction year.

Figure 5: Inconsistency rates.

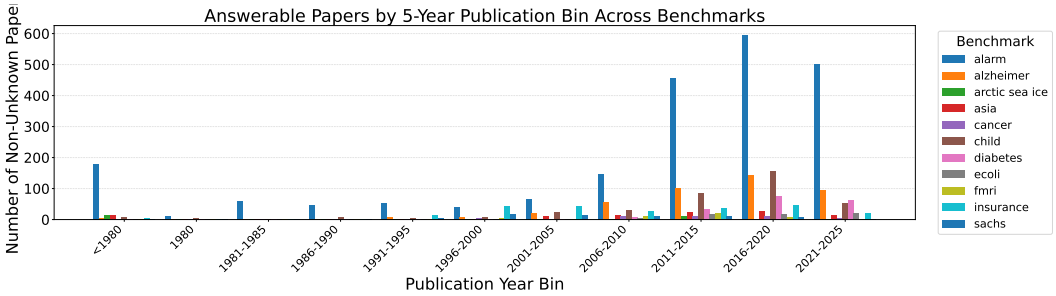


Figure 6: Benchmarks' relevant paper number per year period.

sociation relationship query. Then, for each benchmark, let $\#_{\text{Con}}$ denote the number of papers that present results consistent with the corresponding d-separation query, and, in contrast, let $\#_{\text{InCon}}$ denote the number of papers that present results inconsistent with the corresponding d-separation query. The *inconsistency rate* of a benchmark is then defined as:

$$R_{\text{InCon}} = \frac{\#_{\text{InCon}}}{\#_{\text{InCon}} + \#_{\text{Con}}}.$$

4 Evaluation Results

4.1 Inconsistency Rates and Paper Distribution

Figure 5a shows that Sachs has the lowest inconsistency rate (26.70%), potentially indicating the best alignment with the domain research, and it is also the most frequently used benchmark (used in 38 surveyed papers). The second- and third-lowest benchmarks are Insurance and Alzheimer, with inconsistency rates of 31.6% and 38.3%, respectively. In contrast, Diabetes, Child, and Asia have the highest inconsistency rates (45.8%, 54.0%, and 57.9%, respectively). It is worth noting that all of them were constructed before 2000.

Among the 11 benchmarks, only Insurance and Arctic Sea Ice do not belong to the fields of medical science or biology. However, we do not find that the benchmark field is related to benchmarks' inconsistency rates.

We also use the construction year as a cutoff to measure the inconsistency rate. When only considering the consistency of papers published after each benchmark's construction year, the inconsistency rates vary significantly, as shown in Figure 5b. Fmri's consistency rate decreases from 42.6% to 28.6%, with its ranking improving to second. In contrast, Alarm's inconsistency rate increases from 39.7% to 47.2%. All benchmarks constructed before 2000, except for Insurance, are ranked at the buttom, i.e., Diabetes, Alarm, Child, and Asia.

Figures 6 and 7 show the numbers of retrieved relevant papers (i.e., papers that do not return "Unknown") and inconsistent papers over the years, respectively. Observe that for most of the bench-

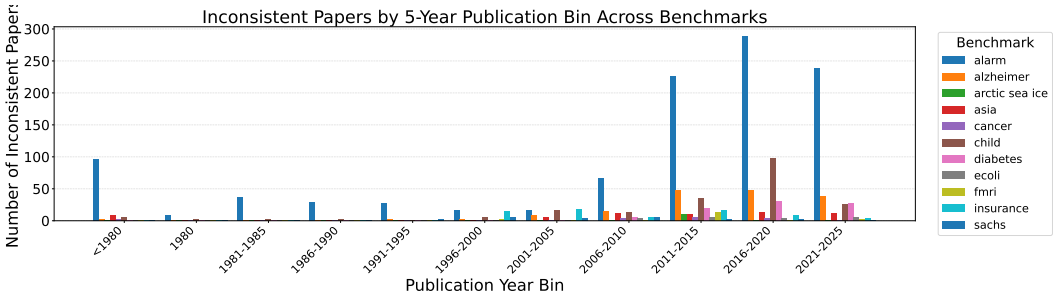


Figure 7: Benchmarks' inconsistent paper number per year period.

marks, both paper numbers start increasing from 2001 and reach their highest levels during 2011–2025. Interestingly, many of the benchmarks have a relatively lower number of inconsistent papers around their publication years compared with the numbers in the preceding and subsequent years. In particular, Alarm shows the most obvious trend (see the detailed inconsistent paper numbers in Appendix D.2).

4.2 Discussion

Benchmarks' validity over time. A benchmark often reflects the state-of-the-art understanding of a problem at the time it is constructed. The inconsistency rate may increase over time as the understanding of the problem progresses. Figure 5b clearly shows such a trend: dated benchmarks suffer from the domain misalignment problem. Our results indicate that new benchmarks provide a better measurement for causal discovery methods, especially for LLM-based methods. On the other hand, maintaining and updating dated benchmarks can help the development of better causal discovery methods.

Hypothesis on growing inconsistency. For most benchmarks, the number of inconsistent papers increases after the benchmarks' publication years, whilst the inconsistency rates do not show such a trend. We conjecture that the subsequent increase in the number of inconsistent papers is due to richer studies on more complex scenarios. More recently, high-quality datasets with higher dimensions and more advanced data handling techniques have become available. Therefore, researchers are able to construct more complex causal models, including variables that were previously treated as latent confounders. Such studies may not invalidate the previous benchmarks, but they consider additional factors missing from previous studies, which may lead to the inconsistency.

Benchmark construction process matters. The Sachs benchmark achieves the lowest inconsistency rates in both Figures 5a and 5b. Though it was constructed over 20 years ago, it retains the highest quality among the evaluated benchmarks. We conjecture that its construction process mainly contributes to its stable quality. Sachs et al. (2005) runs a Bayesian network computational method to reconstruct a basic causal graph. Then, the work incorporates domain expertise to revise the basic causal graph and obtain the benchmark causal graph. Such a hybrid construction process helps guarantee its alignment with domain knowledge. We advocate that the domain treat this hybrid process as a standard in benchmark construction.

5 Conclusion

We revisited popular benchmarks used in causal discovery works in recent years and evaluated them based on how well they align with evolving domain knowledge. We designed an LLM-powered pipeline to retrieve a large number of research papers, against which the pipeline evaluates the benchmarks. Our results show that the popular benchmarks have varying levels of alignment with the domain research. In particular, the misalignment issue is more severe for old benchmarks.

Future works include improving paper retrieval performance in terms of both efficiency and accuracy. Ultimately, causal discovery research should develop a systematic method to track changes in the knowledge underlying existing benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for []. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results], and [Generate synthetic data sets, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [edit a research paper to improve readability, summarize or analyse existing literature, and identify relevant literature]. We have reviewed all AI-assisted work. For instance, we have carefully checked LLM refined parts, and we manually evaluated LLM's literature analysis results which is a key component of our experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

We specify the implementation details in Section 3.3. Code is accessible in the anonymous repository: `https://anonymous.4open.science/r/causal_benchmark_verification-B8F8/README.md`.

REFERENCES

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, 1988.

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel Castro, and Daniel Alexander. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 57559–57610, 2024.

Bruce Abramson, John Brown, Ward Edwards, Allan Murphy, and Robert L Winkler. Hailfinder: A bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1): 57–71, 1996.

Raghavendra Addanki and Shiva Kasiviswanathan. Collaborative causal discovery with atomic interventions. *Advances in Neural Information Processing Systems*, 34:12761–12773, 2021.

Virginia Aglietti, Alan Malek, Ira Ktena, and Silvia Chiappa. Constrained causal bayesian optimization. In *International Conference on Machine Learning*, pp. 304–321. PMLR, 2023.

Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

Steen Andreassen, Finn V Jensen, Stig Kjær Andersen, Björn Falck, Uffe Kjærulff, Marianne Woldbye, Aage R Sørensen, Annelise Rosenfalck, and Finn Jensen. Munin: an expert emg assistant. In *Computer-aided electromyography and expert systems*, pp. 255–277. Pergamon Press, 1989.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36:1738–1763, 2023.

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural admg learning. *arXiv preprint arXiv:2303.12703*, 2023.

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*, pp. 247–256. Springer, 1989.

Alexis Bellot, Junzhe Zhang, and Elias Bareinboim. Scores for learning discrete causal graphs with unobserved confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11043–11051, 2024.

John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, 1997.

Philippe Brouillard, Chandler Squires, Jonas Wahl, Karen Sachs, Alexandre Drouin, Dhanya Sridhar, et al. The landscape of causal discovery data: Grounding causal discovery in real-world applications. In *Causal Learning and Reasoning*, pp. 834–873. PMLR, 2025.

Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, and Kun Zhang. Causal discovery with latent confounders based on higher-order cumulants. In *International conference on machine learning*, pp. 3380–3407. PMLR, 2023.

Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *Acm Computing Surveys (CSUR)*, 44(1):1–50, 2012.

Kai-Chi Chen, Hong-Ren Yu, Wei-Shiang Chen, Wei-Che Lin, Yi-Chen Lee, Hung-Hsun Chen, Jyun-Hong Jiang, Ting-Yi Su, Chang-Ku Tsai, Ti-An Tsai, et al. Diagnosis of common pulmonary diseases in children by x-ray images and deep learning. *Scientific reports*, 10(1):17374, 2020.

Wei Chen, Xiaokai Huang, Zijian Li, Ruichu Cai, Zhiyi Huang, and Zhifeng Hao. Individual causal structure learning from population data. In *IJCAI*, pp. 7109–7117, 2024.

Xinshi Chen, Haoran Sun, Caleb Ellington, Eric Xing, and Le Song. Multi-task learning of order-consistent causal graphs. *Advances in Neural Information Processing Systems*, 34:11083–11095, 2021.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Gilbert P Compo, Jeffrey S Whitaker, Prashant D Sardeshmukh, Nobuki Matsui, Robert J Allan, Xungang Yin, Byron E Gleason, Russell S Vose, Glenn Rutledge, Pierre Bessemoulin, et al. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137 (654):1–28, 2011.

Cristina Conati, Abigail S Gertner, Kurt VanLehn, and Marek J Druzdzel. On-line student modeling for coached problem solving using bayesian networks. In *User Modeling: Proceedings of the Sixth International Conference UM97 Chia Laguna, Sardinia, Italy June 2–5 1997*, pp. 231–242. Springer, 1997.

Zijun Cui, Naiyu Yin, Yuru Wang, and Qiang Ji. Empirical bayesian approaches for robust constraint-based causal discovery under insufficient data. *arXiv preprint arXiv:2206.08448*, 2022.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Haoyue Dai, Peter Spirtes, and Kun Zhang. Independence testing-based approach to causal discovery under measurement error and linear non-gaussian models. *Advances in Neural Information Processing Systems*, 35:27524–27536, 2022.

Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent variable causal discovery under selection bias. *arXiv preprint arXiv:2512.11219*, 2025.

Mert Demirer, Francis X Diebold, Laura Liu, and Kamil Yilmaz. Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1):1–15, 2018.

Anish Dhir, Samuel Power, and Mark Van Der Wilk. Bivariate causal discovery using bayesian model selection. *arXiv preprint arXiv:2306.02931*, 2023.

Anish Dhir, Ruby Sedgwick, Avinash Kori, Ben Glocker, and Mark Van Der Wilk. Continuous bayesian model selection for multivariate causal discovery. *arXiv preprint arXiv:2411.10154*, 2024.

Shuyu Dong, Michele Sebag, Kento Uemura, Akito Fujii, Shuang Chang, Yusuke Koyanagi, and Koji Maruhashi. Dcilp: A distributed approach for large-scale causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16345–16353, 2025.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representations*, volume 2024, pp. 43084–43118, 2024.

Bao Duong, Sunil Gupta, and Thin Nguyen. Causal discovery via bayesian optimization. *arXiv preprint arXiv:2501.14997*, 2025.

Muhammad Qasim Elahi, Lai Wei, Murat Kocaoglu, and Mahsa Ghasemi. Adaptive online experimental design for causal discovery. In *International Conference on Machine Learning*, pp. 12385–12408. PMLR, 2024.

Elias Eulig, Atalanti A Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26778–26786, 2025.

Tao Feng, Lizhen Qu, Niket Tandon, and Gholamreza Haffari. Iris: An iterative and integrated framework for verifiable causal discovery in the absence of tabular data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9400–9428, 2025a.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. On the reliability of large language models for causal discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9565–9590, 2025b.

Ana Ferro, Francisco Pina, Milton Severo, Pedro Dias, Francisco Botelho, and Nuno Lunet. Use of statins and serum levels of prostate specific antigen. *Acta Urológica Portuguesa*, 32(2):71–77, 2015.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

Jensen FV. *An introduction to Bayesian networks*. London, UK: UCL Press, 1996.

Erdun Gao, Ignavier Ng, Mingming Gong, Li Shen, Wei Huang, Tongliang Liu, Kun Zhang, and Howard Bondell. Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35:5024–5038, 2022.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Jan Grzegorzewski, Janosch Brandhorst, Kathleen Green, Dimitra Eleftheriadou, Yannick Duport, Florian Barthorscht, Adrian Köller, Danny Yu Jia Ke, Sara De Angelis, and Matthias König. Pk-db: pharmacokinetics database for individualized and stratified computational modeling. *Nucleic acids research*, 49(D1):D1358–D1364, 2021.

Xianjie Guo, Kui Yu, Lin Liu, and Jiuyong Li. Fedcsl: A scalable and accurate approach to federated causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12235–12243, 2024a.

Xianjie Guo, Kui Yu, Hao Wang, Lizhen Cui, Han Yu, and Xiaoxiao Li. Sample quality heterogeneity-aware federated causal discovery through adaptive variable space selection. In *IJCAI*, pp. 4071–4079, 2024b.

Muralikrishnna Guruswamy Sethuraman and Faramarz Fekri. Differentiable cyclic causal discovery under unmeasured confounders. *Advances in Neural Information Processing Systems*, 38: 174111–174147, 2026.

Isabelle Guyon, Olivier Goudet, and Diviyan Kalainathan. Evaluation methods of cause-effect pairs. In *Cause Effect Pairs in Machine Learning*, pp. 27–99. Springer, 2019.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

David Earl Heckerman, Eric J Horvitz, and Bharat N Nathwani. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine*, 31(02):90–105, 1992.

Marianne J Heins, Hans Knoop, William J Burk, and Gijs Bleijenberg. The process of cognitive behaviour therapy for chronic fatigue syndrome: which changes in perpetuating cognitions and behaviour are related to a reduction in fatigue? *Journal of psychosomatic research*, 75(3):235–241, 2013.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery from multiple data sets with non-identical variable sets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10153–10161, 2020.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in big Data*, 4:642182, 2021.

Inwoo Hwang, Yunhyeok Kwak, Suhyung Choi, Byoung-Tak Zhang, and Sanghack Lee. Fine-grained causal dynamics learning with quantization for improving robustness in reinforcement learning. *arXiv preprint arXiv:2406.03234*, 2024.

Amirkasra Jalaldoust, Kateřina Hlaváčková-Schindler, and Claudia Plant. Causal discovery in hawkes processes by minimum description length. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6978–6987, 2022.

Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 479–486, 2010.

Allan Leck Jensen and Finn Verner Jensen. Midas-an influence diagram for management of mildew in winter wheat. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pp. 349–356, 1996.

Claus Skaanning Jensen and Augustine Kong. Blocking gibbs sampling for linkage analysis in large pedigrees with many loops. *The American Journal of Human Genetics*, 65(3):885–901, 1999.

FV Jensen, U Kjærulff, KG Olesen, and J Pedersen. An expert system for control of waste water treatment—a pilot project. Technical report, Technical report, Judex Datasystemer A/S, Aalborg, 1989. In Danish, 1989.

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

Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In *International Conference on Machine Learning*, pp. 15639–15654. PMLR, 2023.

Jiyeon Kang, Songseong Kim, Chanhui Lee, Doyeong Hwang, Joanie Chung, Yunkyung Ko, Sumin Lee, Sungwoong Kim, and Sungbin Lim. Score-informed neural operator for enhancing ordering-based causal discovery. *Advances in Neural Information Processing Systems*, 38:113109–113151, 2026.

Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Anirudh Goyal, Jorg Bornschein, Melanie Rey, Theophane Weber, Matthew Botvinic, Michael Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. *arXiv preprint arXiv:2204.04875*, 2022.

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*, 2023.

Murat Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Advances in Neural Information Processing Systems*, 36:74140–74179, 2023.

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

Kristian Kristensen and Ilse A Rasmussen. The use of a bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *computers and electronics in agriculture*, 33(3):197–217, 2002.

Chanhui Lee, Juhyeon Kim, YongJun Jeong, Yoonseok Yeom, Juhyun Lyu, Jung-Hee Kim, Sangmin Lee, Sangjun Han, Hyeokjun Choe, Soyeon Park, et al. On incorporating prior knowledge extracted from pre-trained language models into causal discovery. In *Causality and Large Models@ NeurIPS 2024*, 2024.

Loka Li, Haoyue Dai, Hanin Al Ghothani, Biwei Huang, Jiji Zhang, Shahar Harel, Isaac Bentwich, Guangyi Chen, and Kun Zhang. On causal discovery in the presence of deterministic relations. *Advances in Neural Information Processing Systems*, 37:130920–130952, 2024a.

Loka Li, Ignavier Ng, Gongxu Luo, Biwei Huang, Guangyi Chen, Tongliang Liu, Bin Gu, and Kun Zhang. Federated causal discovery from heterogeneous data. *arXiv preprint arXiv:2402.13241*, 2024b.

Mingjia Li, Hong Qian, Tian-Zuo Wang, Min Zhang, Aimin Zhou, et al. Strong and weak identifiability of optimization-based causal discovery in non-linear additive noise models. In *Forty-second International Conference on Machine Learning*, 2025a.

Yan Li, Rui Xia, Chunchen Liu, and Liang Sun. A hybrid causal structure learning algorithm for mixed-type data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7435–7443, 2022.

Zheng Li, Zeyu Liu, Feng Xie, Hao Zhang, Chunchen Liu, and Zhi Geng. Local identifying causal relations in the presence of latent variables. In *Forty-second International Conference on Machine Learning*, 2025b.

Jiaxuan Liang, Jun Wang, Guoxian Yu, Shuyin Xia, and Guoyin Wang. Multi-granularity causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13727–13735, 2024.

Zhaolong Ling, Mengxiang Guo, Xingyu Wu, Debo Cheng, Peng Zhou, Tianci Li, and Zhangling Duan. Gradient-based causal feature selection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, Montreal, QC, Canada*, pp. 16–22, 2025a.

Zhaolong Ling, Honghui Peng, Yiwen Zhang, Debo Cheng, Xingyu Wu, Peng Zhou, and Kui Yu. Hybrid local causal discovery. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5725–5733, 2025b.

Zhaolong Ling, Jiale Yu, Yiwen Zhang, Debo Cheng, Peng Zhou, Xingyu Wu, Bingbing Jiang, and Kui Yu. Local causal discovery without causal sufficiency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18737–18745, 2025c.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

Chenxi Liu and Kun Kuang. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*, pp. 21756–21777. PMLR, 2023.

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

Mingzhou Liu, Xinwei Sun, Yu Qiao, and Yizhou Wang. Causal discovery via conditional independence testing with proxy variables. In *International Conference on Machine Learning*, pp. 31866–31889. PMLR, 2024b.

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.

Peter JF Lucas, Linda C Van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care, 2004.

Alessandro Magrini, Stefano Di Blasi, Federico Mattia Stefanini, et al. A conditional linear gaussian network to assess the impact of several agronomic settings on the quality of tuscan sangiovese grapes. *Biometrical Letters*, 54(1):25–42, 2017.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. *Advances in Neural Information Processing Systems*, 36:75595–75622, 2023.

Idit Matot and Charles L Sprung. Definition of sepsis. *Intensive care medicine*, 27(Suppl 1):S3–S9, 2001.

Dominik Meier, Sujai Hiremath, Promit Ghosal, and Kyra Gan. When additive noise meets unobserved mediators: Bivariate denoising diffusion for causal discovery. *Advances in Neural Information Processing Systems*, 38:30917–30959, 2026.

Osman A Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8975–8982, 2021.

Ehsan Mokhtarian, Mohmmadsadegh Khorasani, Jalal Etesami, and Negar Kiyavash. Novel ordering-based approaches for causal structure learning in the presence of unobserved variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12260–12268, 2023.

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(32):1–102, 2016.

Yang Ni. Bivariate causal discovery for categorical data via classification with optimal label permutation. *Advances in neural information processing systems*, 35:10837–10848, 2022.

Mizuki Nishino, Harumi Itoh, and Hiroto Hatabu. A practical approach to high-resolution ct of diffuse lung disease. *European journal of radiology*, 83(1):6–19, 2014.

Xueyan Niu, Xiaoyun Li, and Ping Li. Learning cluster causal diagrams: An information-theoretic approach. In *IJCAI*, pp. 4871–4877, 2022.

Prince Ntiamoah, Sanjay Mukhopadhyay, Subha Ghosh, and Atul C Mehta. Recycling plastic: diagnosis and management of plastic bronchitis among adults. *European Respiratory Review*, 30 (161), 2021.

Mateusz Olko, Michał Zajac, Aleksandra Nowak, Nino Scherrer, Yashas Annadani, Stefan Bauer, Łukasz Kuciński, and Piotr Miłoś. Trust your ∇: Gradient-based intervention targeting for causal discovery. *Advances in Neural Information Processing Systems*, 36:50617–50647, 2023.

Agnieszka Onisko. Probabilistic causal models in medicine: Application to diagnosis of liver disorders. In *Ph. D. dissertation, Inst. Biocybern. Biomed. Eng., Polish Academy Sci., Warsaw, Poland*, 2003.

Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):37, 2007.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

Maxime Peyrard and Robert West. A ladder of causal distances. *arXiv preprint arXiv:2005.02480*, 2020.

Russell A Poldrack, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, et al. Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1): 8885, 2015.

Jie Qiao, Zhengming Chen, Jianhua Yu, Ruichu Cai, and Zhifeng Hao. Identification of causal structure in the presence of missing data with additive noise model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20516–20523, 2024a.

Jie Qiao, Yu Xiang, Zhengming Chen, Ruichu Cai, and Zhifeng Hao. Causal discovery from poisson branching structural causal model using high-order cumulant with path analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20524–20531, 2024b.

J. R. Quinlan. Credit Approval. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5FS30.

J Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pp. 236–243, 1993.

Yoram Reich and Steven Joseph Fenves. *Incremental learning for capturing design expertise*. [Carnegie Mellon University], Engineering Design Research Center, 1989.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

Amartya Roy, N Devharish, Shreya Ganguly, and Kripabandhu Ghosh. Causal-llm: A unified one-shot framework for prompt-and data-driven causal graph discovery. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:8259–8279, 2025.

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

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

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

M Scutari. Bayesian networks, magic populations and multiple trait prediction. In *Invited Talk at the 5th International Conference on Quantitative Genetics (ICQG 2016)*, 2016.

Marco Scutari, Phil Howell, David J Balding, and Ian Mackay. Multiple quantitative trait analysis using bayesian networks. *Genetics*, 198(1):129–137, 2014.

Marco Scutari, Maintainer Marco Scutari, and Hiton-PC MMPC. Package 'bnlearn'. *Bayesian network structure learning, parameter learning and inference, R package version*, 4(1), 2019.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6371–6387. PMLR, 2023.

Mohammad Shahverdikondori, Ehsan Mokhtarian, and Negar Kiyavash. Qwo: Speeding up permutation-based causal discovery in ligams. *Advances in Neural Information Processing Systems*, 37:55915–55942, 2024.

ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. Exploring multi-modal data with tool-augmented llm agents for precise causal discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 636–660, 2025.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 10 (1):2975, 2020.

David J Spiegelhalter. Learning in probabilistic expert systems. *Bayesian statistics*, 4:447–465, 1992.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158. doi: 10.7551/mitpress/1754.001.0001. URL https://doi.org/10.7551/mitpress/1754.001.0001.

Alexander Statnikov, Sisi Ma, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R Peskin, and Constantin F Aliferis. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *The Journal of Machine Learning Research*, 16(1): 3219–3267, 2015.

Scott Sussex, Caroline Uhler, and Andreas Krause. Near-optimal multi-perturbation experimental design for causal structure learning. *Advances in Neural Information Processing Systems*, 34: 777–788, 2021.

Scott Sussex, Anastasiia Makarova, and Andreas Krause. Model-based causal bayesian optimization. *arXiv preprint arXiv:2211.10257*, 2022.

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. *arXiv preprint arXiv:2402.01454*, 2024.

Jean-François Ton, Dino Sejdinovic, and Kenji Fukumizu. Meta learning for causal direction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9897–9905, 2021.

Quang-Duy Tran, Bao Duong, Phuoc Nguyen, and Thin Nguyen. Identifying causal direction via variational bayesian compression. *arXiv preprint arXiv:2505.07503*, 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.

Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. In *ICLR 2022-The Tenth International Conference on Learning Representations (Virtual), Apr 25th-29th, 2022*. International Conference on Learning Representations, ICLR, 2022.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):43, 2006.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *International Conference on Learning Representations*, volume 2025, pp. 95330–95371, 2025.

Wenjie Wang, Biwei Huang, Feng Liu, Xinge You, Tongliang Liu, Kun Zhang, and Mingming Gong. Optimal kernel choice for score function-based causal discovery. In *International Conference on Machine Learning*, pp. 50650–50668. PMLR, 2024.

X Wang, Y Du, S Zhu, L Ke, Z Chen, J Hao, and J Wang. Ordering-based causal discovery with reinforcement learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp. 3566–3573. IJCAI International Joint Conferences on Artificial Intelligence Organization, 2021.

Yunxia Wang, Fuyuan Cao, Kui Yu, and Jiye Liang. Efficient causal structure learning from multiple interventional datasets with unknown targets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8584–8593, 2022.

Yunxia Wang, CAO Fuyuan, Kui Yu, and Jiye Liang. Federated causal structure learning with non-identical variable sets. In *Forty-second International Conference on Machine Learning*, 2025a.

Zidong Wang, Fei Liu, Qi Feng, Qingfu Zhang, and Xiaoguang Gao. Llm-enhanced score function evolution for causal structure learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9086–9094, 2025b.

Zidong Wang, Xi Lin, Chuchao He, and Xiaoguang Gao. Robust causal discovery under imperfect structural constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36757–36765, 2026.

JN Warwick, TL Sellers, SR Talbot, AJ Cawthorn, and WB Ford. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. sea fisheries division. *Sea Fisheries Division, Technical*, (48), 1994.

Feng Xie, Zheng Li, Peng Wu, Yan Zeng, Chunchen Liu, and Zhi Geng. Local causal structure learning in the presence of latent variables. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54511–54530, 2024.

Dezhi Yang, Guoxian Yu, Jun Wang, Zhengtian Wu, and Maozu Guo. Reinforcement causal structure learning on order graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10737–10744, 2023.

Dezhi Yang, Xintong He, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Jinglin Zhang. Federated causality learning with explainable adaptive optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16308–16315, 2024.

Huiyang Yi, Yanyan He, Duxin Chen, Mingyu Kang, He Wang, and Wenwu Yu. The robustness of differentiable causal discovery in misspecified scenarios. *arXiv preprint arXiv:2510.12503*, 2025.

Naiyu Yin, Tian Gao, Yue Yu, and Qiang Ji. Effective causal discovery under identifiable heteroscedastic noise model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16486–16494, 2024.

Shenbao Yu, Yifeng Zeng, Fan Yang, and Yinghui Pan. Causal-driven skill prerequisite structure discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20604–20612, 2024.

Yan Zeng, Shohei Shimizu, Ruichu Cai, Feng Xie, Michio Yamamoto, and Zhifeng Hao. Causal discovery with multi-domain lingam for latent factors. In *Causal Analysis Workshop Series*, pp. 1–4. PMLR, 2021.

An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. Boosting differentiable causal discovery via adaptive sample reweighting. *arXiv preprint arXiv:2303.03187*, 2023a.

Hao Zhang, Kun Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. Testing independence between linear combinations for causal discovery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 6538–6546, 2021.

Hao Zhang, Shuigeng Zhou, Kun Zhang, and Jihong Guan. Residual similarity based conditional independence test and its application in causal discovery. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 5942–5949, 2022.

Hao Zhang, Yewei Xia, Yixin Ren, Jihong Guan, and Shuigeng Zhou. Differentially private nonlinear causal discovery from numerical data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12321–12328, 2023b.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*, 2024.

Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, and Mengyue Yang. Curious causality-seeking agents learn meta causal world. *arXiv preprint arXiv:2506.23068*, 2025.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Appendix

B Additional related works

LLM-based causal discovery Compared to statistical-based causal discovery methods, LLM-based methods are particularly vulnerable to benchmarks' non-aligningness with real-world knowledge. We list several representatives of recent works due to the vast number of new works.

A LLM-based method is either purely LLM-based or hybrid (of LLM-based and statistical-based). Pure LLM-based methods typically send LLMs pairwise causal relationship queries, e.g., Long et al. (2022). Jiralerspong et al. (2024) reduces this strategy's complexity by a breadth-first search algorithm. Zhang et al. (2024) adds retrieval augmented generation (RAG) to utilise domain research in pure LLM-based method. Vashishtha et al. (2025) additionally queries LLMs the causal orders to retain stability.

A hybrid causal discovery method, e.g., Roy et al. (2025), combines LLM query and statistical-based method. Feng et al. (2025a) and Shen et al. (2025) use LLMs as a data extractor when tabular data is insufficient.

Moreover, Kiciman et al. (2023) and Feng et al. (2025b) examine LLMs' performance in causal discovery, e.g., the reasons of LLMs' success and failure, and LLMs' reliability.

C Additional information of literature review

C.1 The PRISMA paper selection process

We conduct a structured literature review following the PRISMA guidelines for study identification and screening. We consider papers published in major artificial intelligence conferences during 2021-2025, including NeurIPS, ICML, ICLR, AAAI, IJCAI, ACL, and EMNLP.

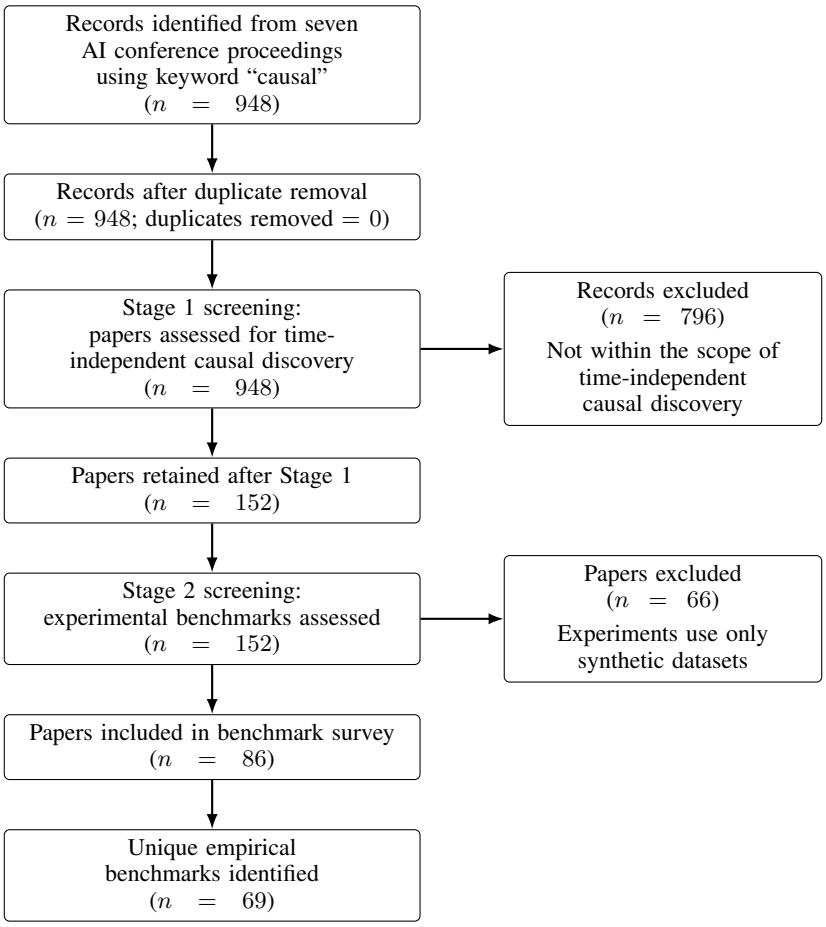


Figure 8: PRISMA flow diagram of the paper-selection process for the causal discovery benchmark survey.

In the identification stage, we retrieve published papers containing the keyword "causal" in the title. This search returns 948 records. As the searches are conducted over non-overlapping conference proceedings, no duplicate records are identified or removed. Therefore, all 948 records proceed to screening.

The screening process consists of two stages. In the first stage, we manually inspect the paper introductions and retain those that propose methods for *time-independent causal discovery*. Papers outside this scope, including works that do not study causal discovery and works focusing on causal relationships in language, multimodal, or time series data, are excluded. After this stage, 152 papers remain, while 796 papers are excluded.

In the second stage, we inspect the experimental evaluation of each remaining paper and identify the benchmarks used. We exclude papers whose experiments use only synthetic datasets, since our analysis focuses on empirical benchmarks, i.e., semi-synthetic or real-world benchmarks whose construction involves empirical data or domain knowledge. This stage excludes 66 papers and results in 86 papers containing at least one empirical benchmark.

The complete paper-selection procedure is summarized by the PRISMA flow diagram in Figure 8. From the final set of 86 papers, we identify 69 unique empirical benchmarks. Detailed paper–benchmark associations are provided in Appendix C.

C.2 FULL LIST OF THE BENCHMARKS AND CAUSAL DISCOVERY PAPERS

The follows are all benchmark datasets we reviewed in the causal discovery publications. Each item is a dataset, where the "original paper" is the resource paper of the dataset, and the "papers using it" element lists all causal discovery publications that use it in the evaluation process. Note that the original paper might not be the only actual resource paper, for instance, for the Gene expression dataset, Sethuraman et al. (2023) reports a causal graph based on interventional genetic data in Frangieh et al. (2021), thus forming a complete benchmark for causal discovery. However, the original papers of some datasets correspond to the dataset resource instead of the causal graph proposal paper.

Some datasets have different versions. Instances include the Ecoli dataset versions, e.g., Ecoli70 and Ecoli100. The Alzheimer has several versions, e.g., Poldrack et al. (2015), Shen et al. (2020), and several others. We list different versions into one item.

Some datasets are updated regularly, and we list the latest version, e.g., the APM (Application performance monitoring program of AWS), and the General social survey.

Some papers claim that two datasets, namely, the Earthquake and the Survey, are practical. However, we cannot find any realistic resource of these two datasets, and therefore, we neglect these two datasets in our list.

1. Sachs
   Original paper Sachs et al. (2005)
   Papers using it: Eulig et al. (2025); Shen et al. (2025); Roy et al. (2025); Shahverdikondori et al. (2024); Kang et al. (2026); Aglietti et al. (2023); Olko et al. (2023); Annadani et al. (2023); Perry et al. (2022); Dai et al. (2022); Addanki & Kasiviswanathan (2021); Cundy et al. (2021); Wang et al. (2025a); Li et al. (2025a); Elahi et al. (2024); Wang et al. (2024); Kaltenpoth & Vreeken (2023); Rolland et al. (2022); Yi et al. (2025); Ke et al. (2022); Zhang et al. (2023a); Lippe et al. (2021); Wang et al. (2026); Yin et al. (2024); Liang et al. (2024); Yang et al. (2024); Guo et al. (2024a); Bellot et al. (2024); Yang et al. (2023); Ling et al. (2025b;a); Chen et al. (2024); Niu et al. (2022); Wang et al. (2021); Mameche et al. (2023)
2. Child
   Original paper Spiegelhalter (1992)
   Papers using it: Shen et al. (2025); Peyrard & West (2020); Olko et al. (2023); Wang et al. (2025a; 2024); Vashishtha et al. (2025); Duong et al. (2025); Ke et al. (2022); Lippe et al. (2021); Guo et al. (2024a); Zhang et al. (2023b; 2022); Wang et al. (2022); Ling et al. (2025b); Guo et al. (2024b); Cui et al. (2022)
3. Alarm
   Original paper Beinlich et al. (1989)
   Papers using it: Akbari et al. (2021); Roy et al. (2025); Xie et al. (2024); Li et al. (2022); Olko et al. (2023); Wang et al. (2025a); Duong et al. (2025); Lippe et al. (2021); Guo et al. (2024a); Zhang et al. (2023b; 2022); Wang et al. (2022); Zhang et al. (2021); Ling et al. (2025b); Guo et al. (2024b); Cui et al. (2022)
4. Asia
   Original paper lau (1988)
   Papers using it: Shen et al. (2025); Roy et al. (2025); Shahverdikondori et al. (2024); Olko et al. (2023); Kocaoglu (2023); Addanki & Kasiviswanathan (2021); Vashishtha et al. (2025); Duong et al. (2025); Ke et al. (2022); Lippe et al. (2021); Bellot et al. (2024); Zhang et al. (2023b; 2022)
5. Insurance
   Original paper Binder et al. (1997)
   Papers using it: Mokhtarian et al. (2023); Akbari et al. (2021); Feng et al. (2025a;b); Peyrard & West (2020); Wang et al. (2025a); Guo et al. (2024a); Zhang et al. (2022); Wang et al. (2022); Zhang et al. (2021); Guo et al. (2024b); Cui et al. (2022)
6. Cancer
   Original paper Korb & Nicholson (2010)
   Papers using it: Feng et al. (2025a;b); Shahverdikondori et al. (2024); Peyrard & West

(2020); Olko et al. (2023); Vashishtha et al. (2025); Duong et al. (2025); Lippe et al. (2021); Zhang et al. (2023b; 2022)

7. Barley
   Original paper Kristensen & Rasmussen (2002)
   Papers using it: Akbari et al. (2021); Wang et al. (2025a); Ling et al. (2025c); Zhang et al. (2022; 2021); Ling et al. (2025b)
8. Hailfinder
   Original paper Abramson et al. (1996)
   Papers using it: Mokhtarian et al. (2023); Akbari et al. (2021); Li et al. (2022); Zhang et al. (2021); Ling et al. (2025b); Cui et al. (2022)
9. Fmri hippocampus
   Original paper Poldrack et al. (2015)
   Papers using it: Li et al. (2024b); Chen et al. (2024); Zeng et al. (2021)
10. Alzheimer
    Original paper Petersen et al. (2010); Shen et al. (2020)
    Papers using it: Abdulaal et al. (2024); Feng et al. (2025a); Vashishtha et al. (2025)
11. Arctic sea ice
    Original paper Huang et al. (2021)
    Papers using it: Abdulaal et al. (2024); Feng et al. (2025b); Kiciman et al. (2023)
12. Diabetes
    Original paper Long et al. (2022)
    Papers using it: Feng et al. (2025a;b); Lippe et al. (2021)
13. Ecoli70(100)
    Original paper Schäfer & Strimmer (2005)
    Papers using it: Mokhtarian et al. (2023); Akbari et al. (2021); Kang et al. (2026); Chen et al. (2021)
14. Gene expression
    Original paper Sethuraman et al. (2023)
    Papers using it: Guruswamy Sethuraman & Fekri (2026); Xie et al. (2024); Li et al. (2025b); Guo et al. (2024a); Ling et al. (2025b)
15. Hepar2
    Original paper Onisko (2003)
    Papers using it: Mokhtarian et al. (2023); Roy et al. (2025); Li et al. (2022)
16. Pigs
    Original paper FV (1996)
    Papers using it: Lippe et al. (2021); Guo et al. (2024a); Ling et al. (2025b)
17. Reged
    Original paper Statnikov et al. (2015)
    Papers using it: Mian et al. (2021); Kaltenpoth & Vreeken (2023); Guo et al. (2024b)
18. Andes
    Original paper Conati et al. (1997)
    Papers using it: Xie et al. (2024); Li et al. (2022)
19. Arth150
    Original paper Opgen-Rhein & Strimmer (2007)
    Papers using it: Mokhtarian et al. (2023); Akbari et al. (2021); Kang et al. (2026)
20. Auto mpg
    Original paper Quinlan (1993)
    Papers using it: Eulig et al. (2025); Shen et al. (2025)
21. Carpo
    Original paper Link
    Papers using it: Mokhtarian et al. (2023); Akbari et al. (2021)
22. HK stock
    Original paper Huang et al. (2020)
    Papers using it: Li et al. (2024b); Cai et al. (2023)

23. Link
    Original paper Jensen & Kong (1999)
    Papers using it: Wang et al. (2022); Ling et al. (2025b)
24. Mildew
    Original paper Jensen & Jensen (1996)
    Papers using it: Xie et al. (2024); Li et al. (2022)
25. Neuropain
    Original paper Tu et al. (2019)
    Papers using it: Liu et al. (2024a); Feng et al. (2025a); Kiciman et al. (2023); Vashishtha et al. (2025)
26. Obesity
    Original paper Long et al. (2022)
    Papers using it: Feng et al. (2025a;b)
27. Syntren
    Original paper Van den Bulcke et al. (2006)
    Papers using it: Dhir et al. (2024); Ling et al. (2025b)
28. WIN95PTS
    Original paper Heckerman et al. (1995)
    Papers using it: Xie et al. (2024); Wang et al. (2022); Zhang et al. (2021)
29. CE-Tueb
    Original paper Mooij et al. (2016)
    Papers using it: Dhir et al. (2023)
30. CE-cha
    Original paper Guyon et al. (2019)
    Papers using it: Dhir et al. (2023)
31. Cognition and Aging in the Chronic Fatigue Syndrome
    Original paper Heins et al. (2013)
    Papers using it: Qiao et al. (2024a)
32. DWDClimate
    Original paper Mooij et al. (2016)
    Papers using it: Shen et al. (2025)
33. MAGIC-IRRI
    Original paper Scutari (2016)
    Papers using it: Kang et al. (2026)
34. MAGIC-NIAB
    Original paper Scutari et al. (2014)
    Papers using it: Kang et al. (2026)
35. New York Times
    Original paper Link
    Papers using it: Liu et al. (2024a)
36. Pittsburgh Bridges
    Original paper Reich & Fenves (1989)
    Papers using it: Ni (2022)
37. Categorical Cause-Effect Pairs
    Original paper Ni (2022)
    Papers using it: Ni (2022)
38. Abalone
    Original paper Warwick et al. (1994)
    Papers using it: Ni (2022)
39. Alcohol
    Original paper Long et al. (2022)
    Papers using it: Feng et al. (2025b)

40. Algibra I
    Original paper Link
    Papers using it: Yu et al. (2024)
41. APM
    Original paper Link
    Papers using it: Eulig et al. (2025)
42. Apple gastronome
    Original paper Liu et al. (2024a)
    Papers using it: Feng et al. (2025a)
43. Big five
    Original paper Link
    Papers using it: Dai et al. (2025); Dong et al. (2024)
44. Brain tumor
    Original paper Link
    Papers using it: Liu et al. (2024a)
45. Chemical
    Original paper Ke et al. (2021)
    Papers using it: Zhao et al. (2025)
46. climatic analysis
    Original paper Compo et al. (2011)
    Papers using it: Liu et al. (2024a)
47. COVID-19
    Original paper Link
    Papers using it: Wang et al. (2025b); Vashishtha et al. (2025)
48. Credit
    Original paper Quinlan (1987)
    Papers using it: Li et al. (2022)
49. Dream3
    Original paper Kalainathan & Goudet (2019)
    Papers using it: Roy et al. (2025); Gao et al. (2022); Sussex et al. (2021)
50. football
    Original paper Link
    Papers using it: Qiao et al. (2024b)
51. G7
    Original paper Demirer et al. (2018)
    Papers using it: Jalaldoust et al. (2022)
52. General social survey
    Original paper Link
    Papers using it: Li et al. (2025b)
53. IHDP
    Original paper Hill (2011)
    Papers using it: Ashman et al. (2023)
54. Lucas
    Original paper Lucas et al. (2004)
    Papers using it: Roy et al. (2025)
55. Magnetic
    Original paper Hwang et al. (2024)
    Papers using it: Zhao et al. (2025)
56. Munin
    Original paper Andreassen et al. (1989)
    Papers using it: Dong et al. (2025); Wang et al. (2022)
57. Pacific walker circulation
    Original paper Runge et al. (2019)
    Papers using it: Liu & Kuang (2023)

58. Pathfinder
    Original paper Heckerman et al. (1992)
    Papers using it: Zhang et al. (2021)
59. Pharmacokinetics
    Original paper Grzegorzewski et al. (2021)
    Papers using it: Li et al. (2024a)
60. Physics
    Original paper Lee et al. (2024)
    Papers using it: Kang et al. (2026)
61. Protein
    Original paper Link
    Papers using it: Peyrard & West (2020)
62. Respiratory disease
    Original paper Long et al. (2022)
    Papers using it: Feng et al. (2025a)
63. Sangiovese
    Original paper Magrini et al. (2017)
    Papers using it: Abdulaal et al. (2024)
64. SCM
    Original paper Ferro et al. (2015)
    Papers using it: Aglietti et al. (2023); Sussex et al. (2022)
65. Sepsis
    Original paper Matot & Sprung (2001)
    Papers using it: Liu et al. (2024b)
66. Tübingen
    Original paper Mooij et al. (2016)
    Papers using it: Ni (2022); Meier et al. (2026); Tran et al. (2025); Kiciman et al. (2023); Tu et al. (2022); Yin et al. (2024); Ton et al. (2021)
67. Water
    Original paper Jensen et al. (1989)
    Papers using it: Mokhtarian et al. (2023)
68. World value survey
    Original paper Link
    Papers using it: Dai et al. (2025)
69. Yahoo
    Original paper: Janzing et al. (2010)
    Papers using it: Zeng et al. (2021)

C.3 LLM-BASED CAUSAL DISCOVERY PAPERS

Figure 9 shows the benchmarks used in our surveyed LLM-based causal discovery works, with the number of papers that use each benchmark.

C.4 LLM-BASED CAUSAL DISCOVERY PAPERS BENCHMARKS ON ARXIV

Figure 10 shows the top 20 benchmarks that are mostly used in recent LLM-based causal discovery works on arXiv. The numbers are the papers that use each benchmark.

D ADDITIONAL INFORMATION OF THE BENCHMARK EVALUATION PROCESS

D.1 CONDITIONAL ASSOCIATION RELATIONSHIP EXTRACTION ALGORITHM

Algorithm 1 is used to extract the conditional association queries from a causal graph for a given variable pair.

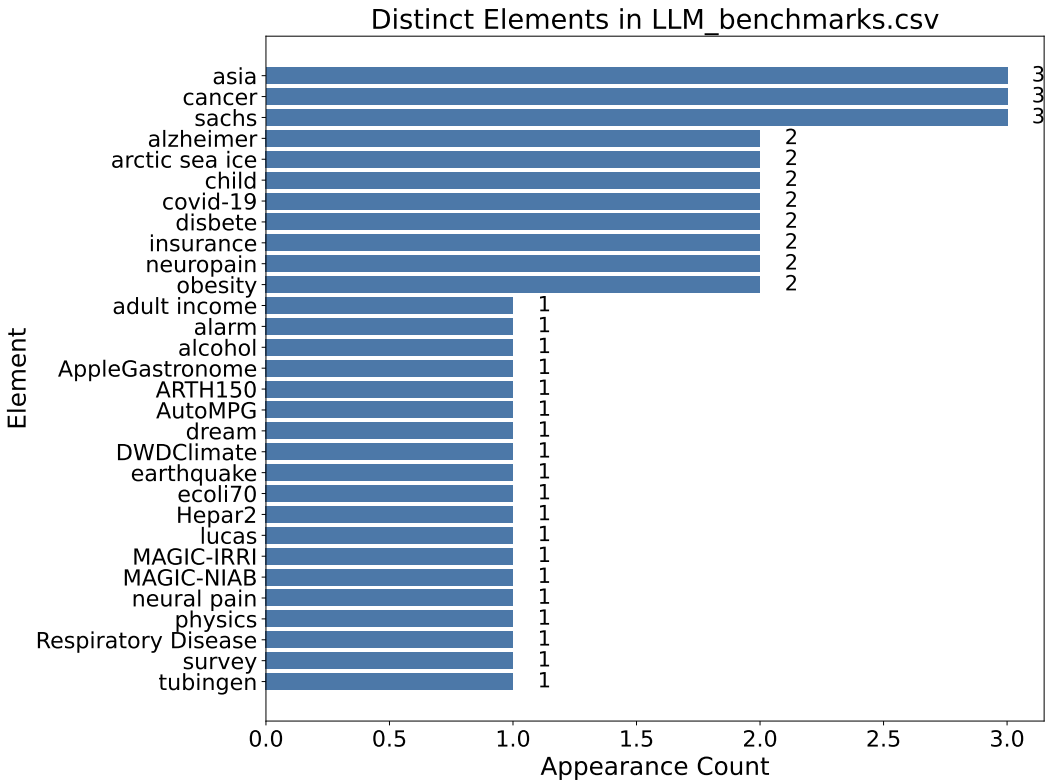


Figure 9: Benchmark numbers used in LLM-based papers.

Algorithm 1 Conditional association relationship extraction

Require: Causal graph $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, and target variable pair $X, Y \in \mathbf{V}$

if X and Y are adjacent in $\mathcal{G}$ **then**
  Return: X and Y are always correlated.
else
  $\mathbf{P} \leftarrow$ all collider-free paths between X and Y.
  if $\mathbf{P} = \emptyset$ **then**
    Return: X and Y are mutually independent.
  else
    $SepSet \leftarrow \emptyset$.
    $\mathbf{D} \leftarrow \{\mathbf{V}' \subseteq \mathbf{V} \setminus \{X, Y\} \mid \forall P \in \mathbf{P}, P \text{ contains at least one variable in } \mathbf{V}'\}$.
    for $\mathbf{Z} \in \mathbf{D}$ **do**
      if $\nexists \mathbf{Z}' \subset \mathbf{Z}$ such that $\mathbf{Z}' \in \mathbf{D}$ **then**
        $SepSet.append(\mathbf{Z})$
      end if
    end for
    Return: [X and Y are mutually independent conditioned on $\mathbf{Z}$ for $\mathbf{Z}$ in $SepSet$]
  end if
end if

D.2 Analysed paper distribution

As shown in Table 2, a large part of the retrieved papers do not contain sufficient information to answer the conditional association query, i.e., return answer "Unknown". This high rate of "Unknown" papers is because of the paper search mechanism of the scientific databases: the keyword matching rule is inefficient to retrieve target papers. Note that the Unknown papers may be higher than the actual number of the irrelevant papers. The reason is that a given variable pair may have

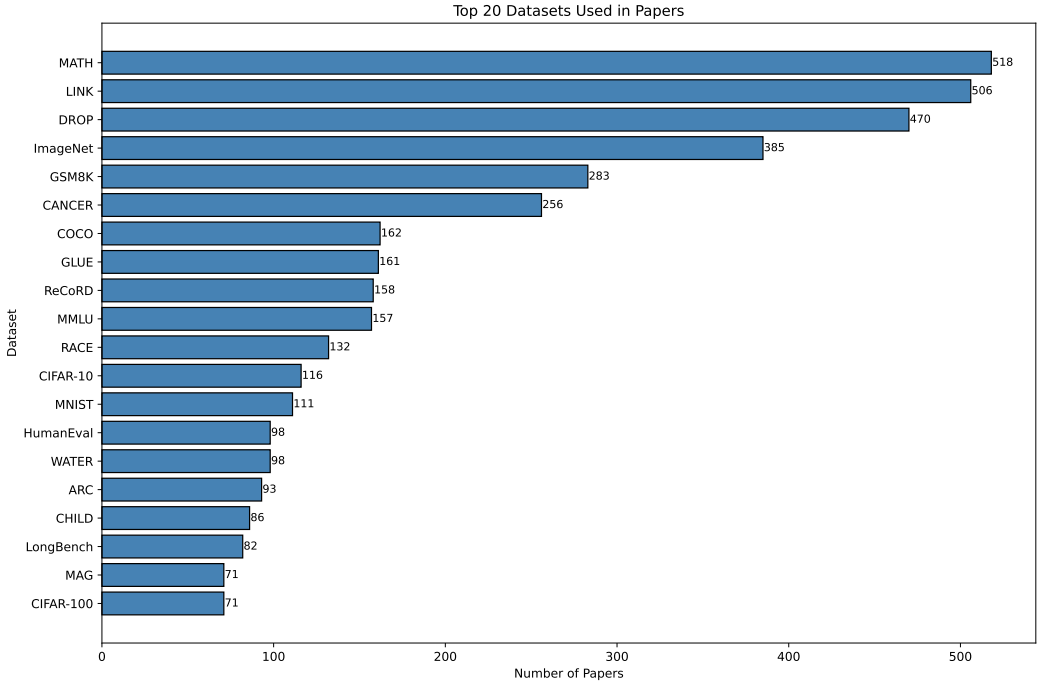


Figure 10: Top 20 benchmarks used in arXiv LLM-based causal discovery papers.

multiple conditional association queries, since the pair can be d-separated by multiple variable sets. A paper may contain information to verify the consistency of one query, but for the other queries, the answer is "Unknown".

Figures 11 to 21 show the inconsistent paper number per year for all evaluated benchmarks.

Benchmark	Correlated	Independent	Inconsistent	Unknown
alarm	1262	904	1054	23667
alzheimer	150	286	167	5105
arctic sea ice	14	12	11	422
asia	74	33	62	1333
cancer	34	16	20	648
child	212	173	208	4490
diabetes	189	3	87	252
ecoli	19	51	23	6981
fmri	19	35	23	2103
insurance	70	177	78	7709
sachs	14	61	20	3797

Table 2: Paper number distribution for each evaluated benchmarks.

D.3 LLM-BASED PAPER VERIFIER ACCURACY

A pipeline component that potentially brings noise is the LLM-based consistency verifier which employs the LLMs to analyse papers and decide the papers' consistency with the conditional association relationships. To deliver convincing results, we evaluate the performance of the LLM-based consistency verifier by randomly sampling 100 retrieved papers, and manually verifying the correctness of the LLM-based consistency verifier's answers.

Specifically, we read each of the sampled papers, and identify whether the statistical/machine learning analysis reveals the association relationship between the query variables V_i and V_j. Then,

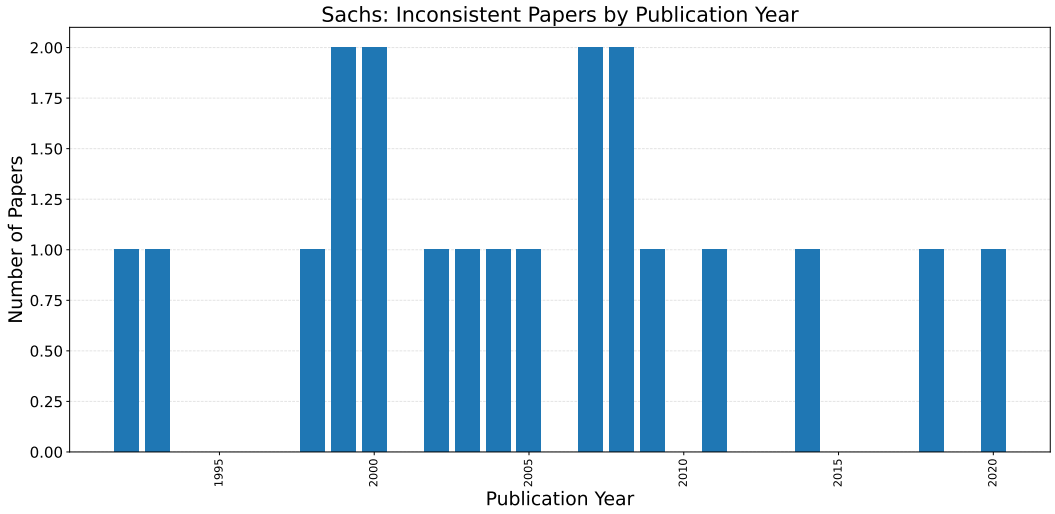


Figure 11: Inconsistent paper number per year of Sachs.

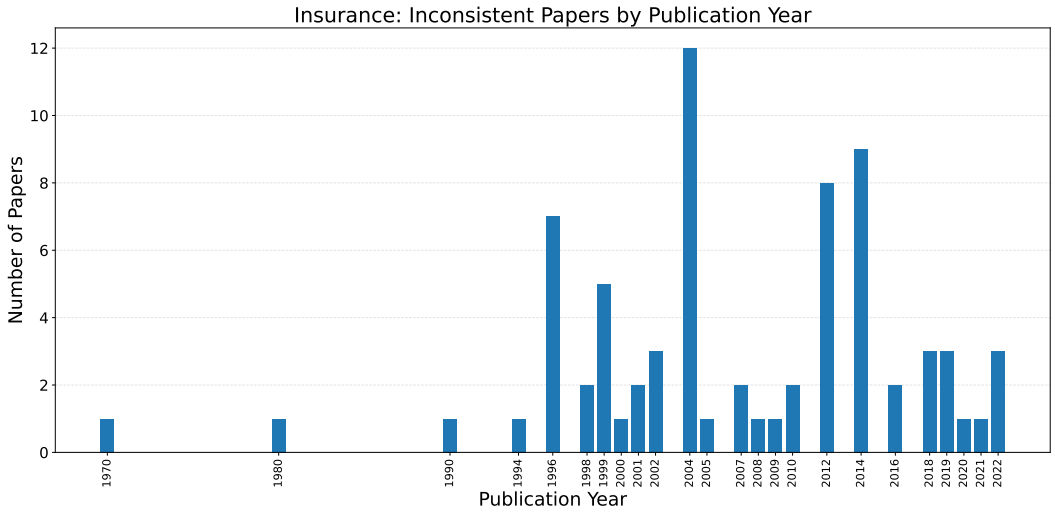


Figure 12: Inconsistent paper number per year of Insurance.

based on the experimental settings, we identify the controlled covariate set $\mathbf{Z}$, and verify whether $\mathbf{Z} \in \mathbf{V} \setminus \{V_i, V_j\}$ holds. Lastly, we compare the conditional association relationship extracted from the paper and $V_i \overset{\mathcal{G}}{-} V_j | \mathbf{Z}$, and verify the correctness of the answer.

The accuracy achieves 90%, i.e., the LLM-based consistency verifier correctly analyses 90 out of the 100 randomly sampled papers. We provide the detailed manual evaluation results in `https://anonymous.4open.science/r/causal_benchmark_verification-B8F8/LLM_consistency_verifier_manual_evaluation_labels.pdf`.

D.4 LLM PROMPT

You are a scientific research assistant analyzing a paper on {pair_info}. Your task is to determine if this paper contains data analysis that can answer the {query}. You must return ONLY a JSON object with the structure specified in the user's instructions.

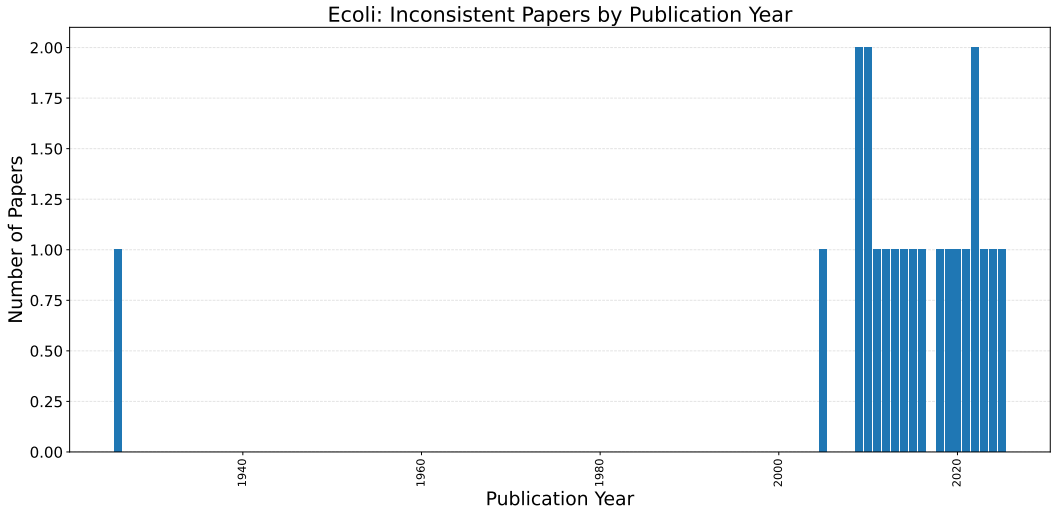


Figure 13: Inconsistent paper number per year of Ecoli.

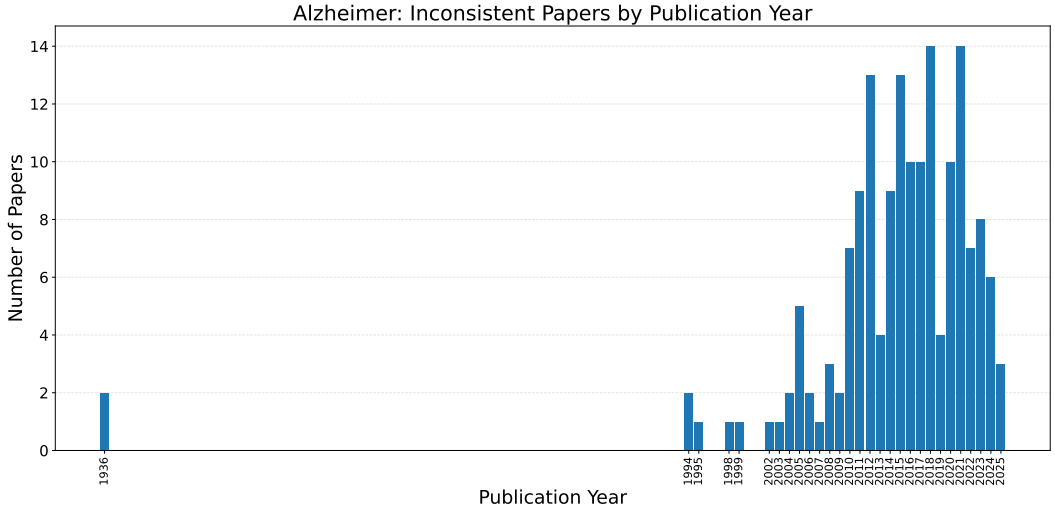


Figure 14: Inconsistent paper number per year of Alzheimer.

First, you need to thoroughly read the following "Paper content", and then to determine whether the paper contains data analysis information based on the following principles. You are an experienced researcher and good at analysing research papers. Think step by step to answer the {query}.

"has_data_analysis determination principles":
A paper contains data analysis information if:
1. It has a concrete description of the data used in the study or data collection methods.
2. It has a detailed illustration on the statistical methods or machine learning based methods, used to analyse the data.
3. It contains detailed reports, for example, the analysis results and significance (for example, p values), on the above analysis methods and data.
4. *IMPORTANT* Do check whether the paper introduces any numerical dataset and implements corresponding analysis on the dataset. A review/survey paper does not contain data analysis.

If the paper contains data analysis information, then you need to determine whether the data analysis can answer the {query}. Typically, we are interested in the statistical correlation between the main

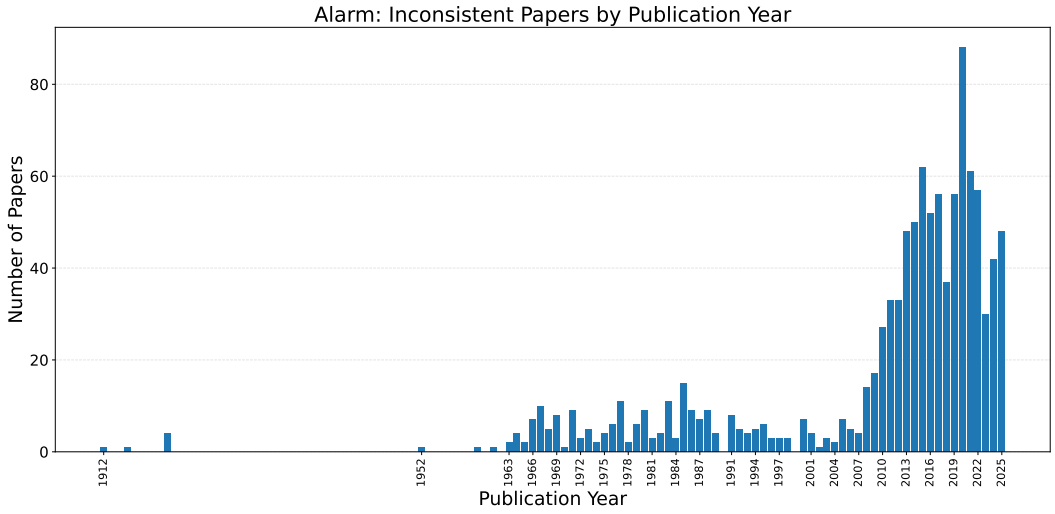


Figure 15: Inconsistent paper number per year of Alarm.

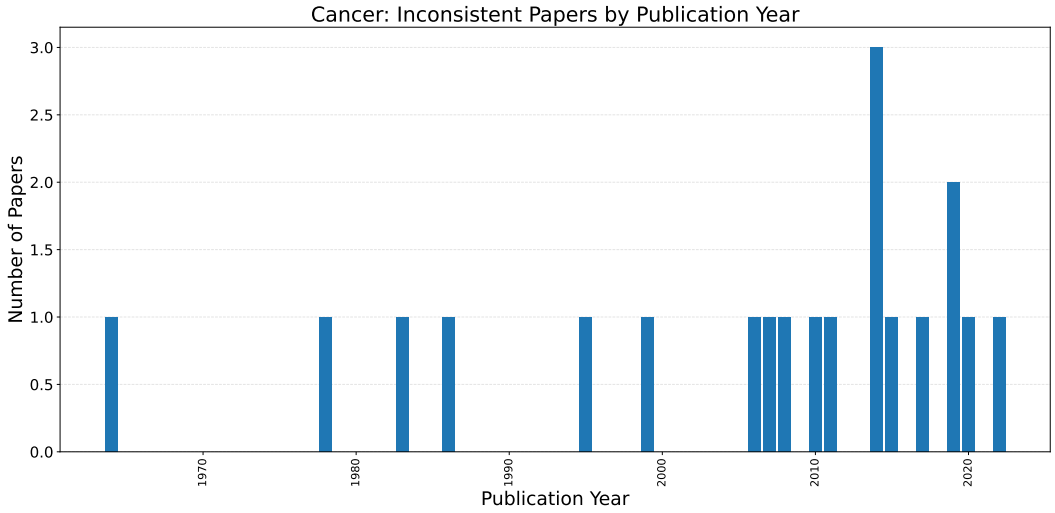


Figure 16: Inconsistent paper number per year of Cancer.

factors of {pair_info}, possibly conditioned on a set of other factors (called covariates). Then, you can answer the query {query} by the following principles:

1. The query may ask whether two factors are correlated or independent conditioned on some other factors. This query is to ask the statistical correlational relationship between the factor pair conditioned on other factors.
2. The statistical correlation between the factor pair can be revealed by statistical tests or machine learning results based on data described in the paper. 3. We can conclude that the pair of factors are correlated if the paper shows a significant statistical test result for a correlation test (for example Pearson test or regression), otherwise we can conclude that the factor pair are mutually independent. 4. When the query is asking the statistical correlation between the factor pair conditioned on a set of factors, we consider the statistical correlation of the conditional density. 5. Sometimes, the conditioned factors are called covariates. A research usually consider adjusting statistical analysis by covariates, and under such processes, we can confirm the work investigates the correlation (correlated or independent) of the factor pair conditioned on the covariates. 6. The correlation of the conditional density means: if we fix the conditioned factors constant, are the pair of factors mutually independent or correlated? 7. In the research paper, the conditional correlation may be considered

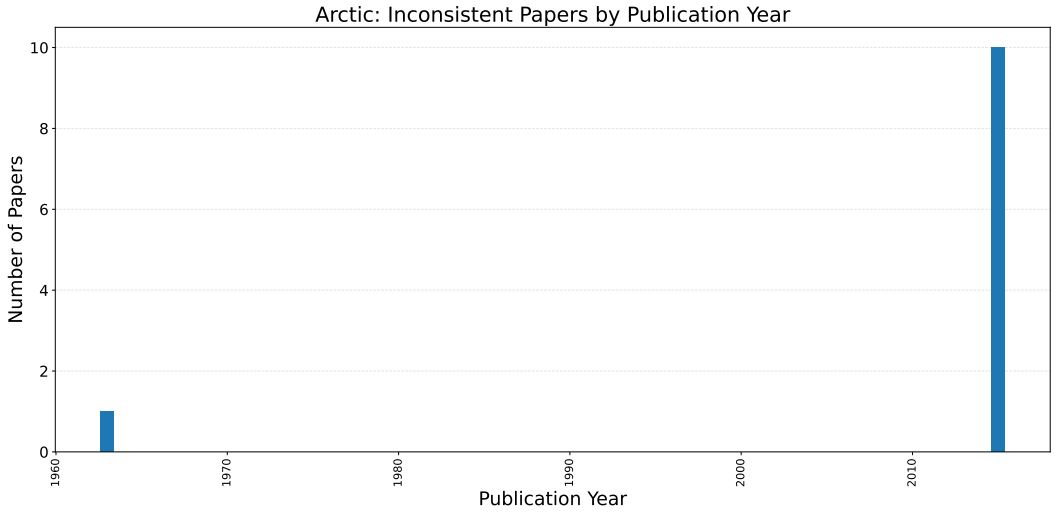


Figure 17: Inconsistent paper number per year of Arctic Sea Ice.

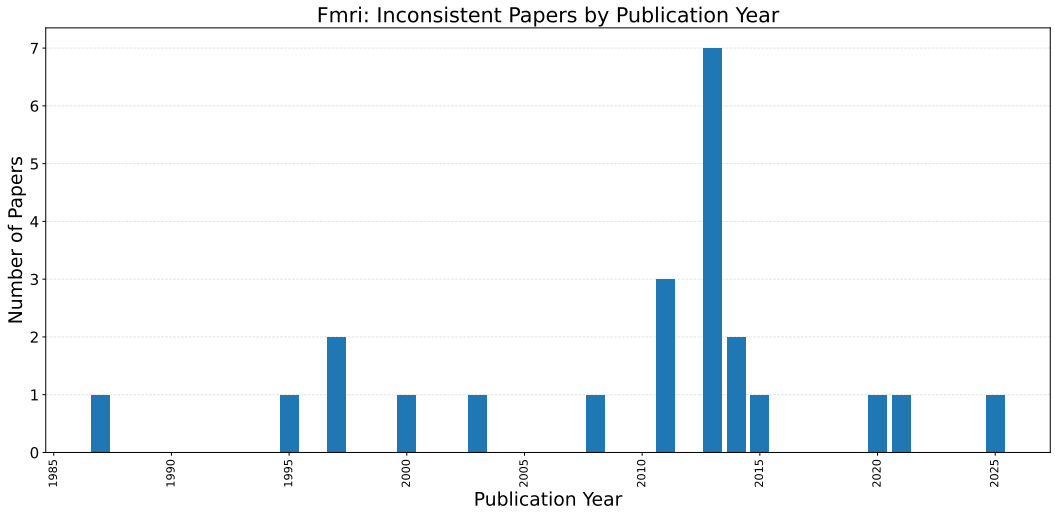


Figure 18: Inconsistent paper number per year of Fmri.

as: in the data collection experiments (e.g., random trial experiments), researchers may control the value of some factors (the covariates), and collect corresponding data for the main factor pair. In such cases, we can obtain reliable conclusion about the conditional correlation. 8. Reason indication of covariates. For example, factor Smoking may be strongly correlated with age. If a data sample has an age under 18, it may indicate that the sample does not smoke. 9. When it is possible to reason indication of factor, do not implement the reasoning for {pair_info}. 10. Two factors are mutually independent marginally if the marginal density of the two factors are statistically independent, but the independence may not hold when conditioned on other factors. 11. If the query asks whether two factors are correlated directly, you do not need to consider the conditional factors, that is, the covariates.

IMPORTANTLY, you need to get clear about the following points: 1. Check and remember the covariates mentioned in the paper's method description. 2. Be clear about the main factors. 3. The covariates may be correlated with or independent from the main factors. This should not influence our investigation on the correlation between the main factors. 4. Be careful whether the factors studied in the paper are really those in {pair_info} and query.

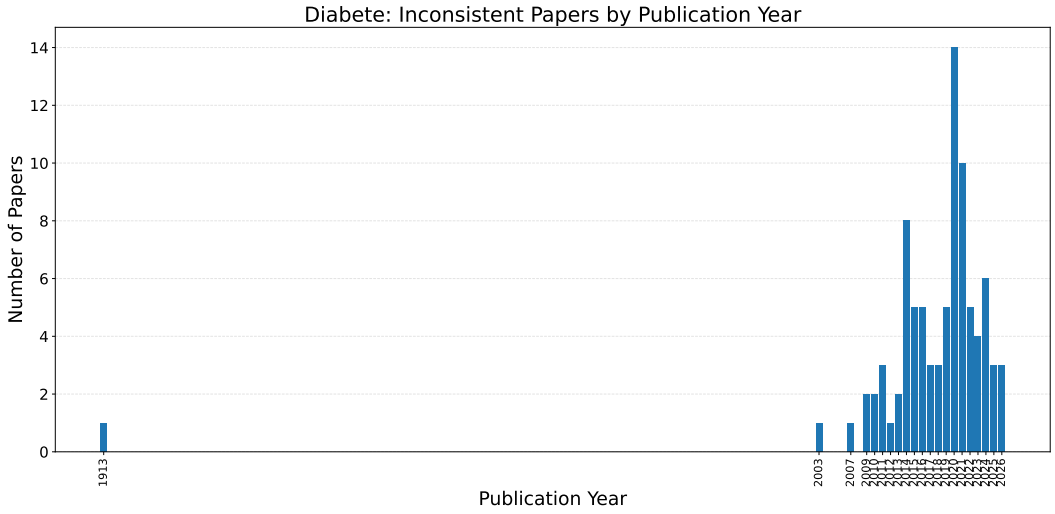


Figure 19: Inconsistent paper number per year of Diabetes.

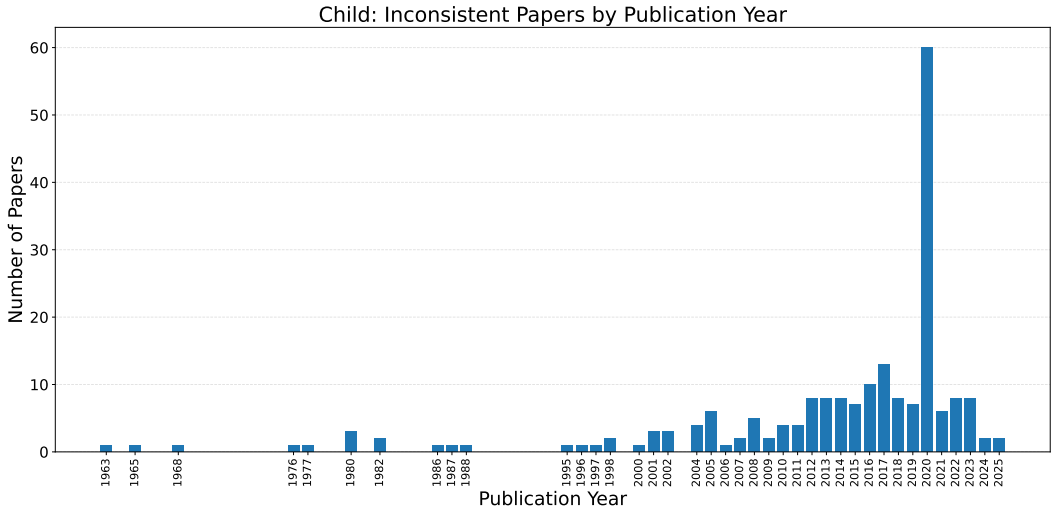


Figure 20: Inconsistent paper number per year of Child.

”Required information” 1. Answer the query with a single choice among: ”Correlated”, ”Independent”, and ”Unknown”. 2. A brief thinking process of how you obtain the conclusion. 3. Reference sentences from the original paper for the conclusion.

Paper content (possibly truncated): “‘ {paper_text} “‘

[IMPORTANT and MUST] Return ONLY a JSON object with this structure: {{ ”results”:{{ ”has_data_analysis”: true or false, ”answer of the query”: Independent or Correlated or Unknown, ”brief thinking process”: the brief reasoning process of how you obtain the answer, ”reference”: the reference contents from the original paper corresponding to your answer. }} }} ”””

D.5 Additional results

Figures 22 and 23 show the numbers of retrieved papers and relevant papers per year, respectively, for all evaluated benchmarks.

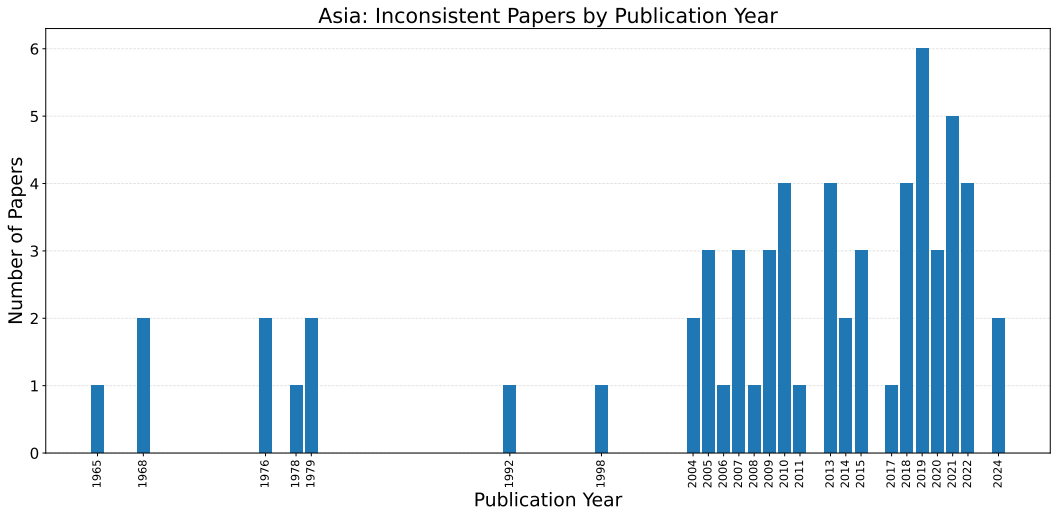


Figure 21: Inconsistent paper number per year of Asia.

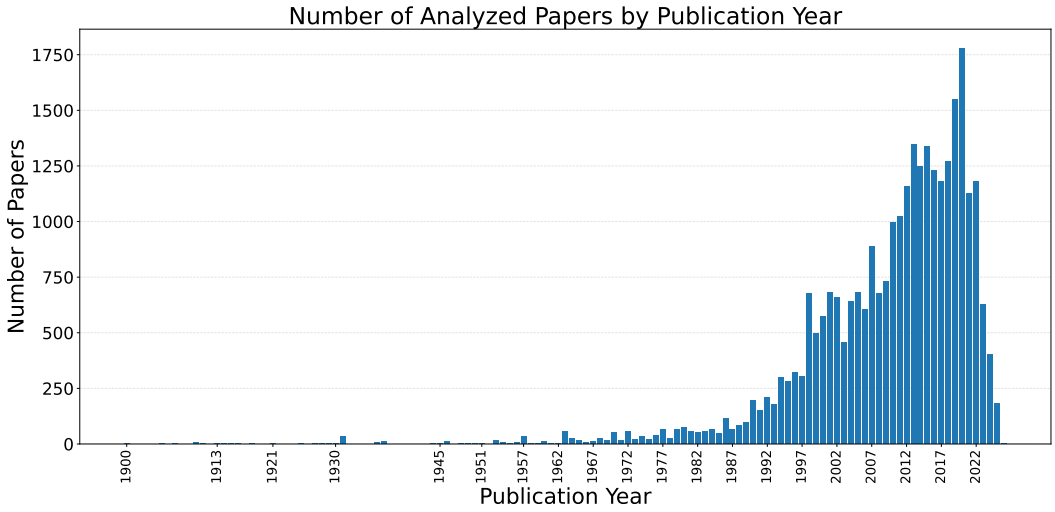


Figure 22: Retrieved papers by year.

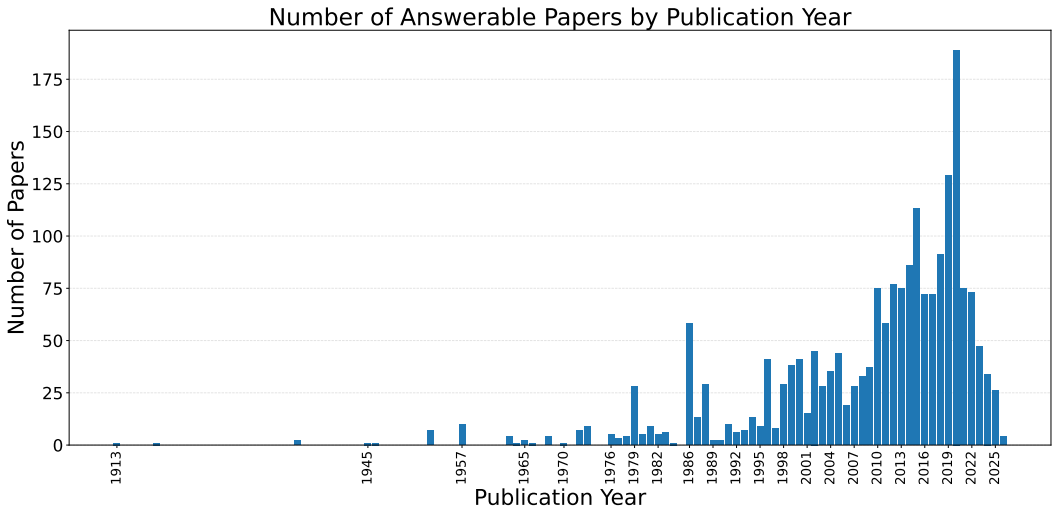


Figure 23: Retrieved papers that contain relevant information by year.