Consistency evaluation of benchmarks used for causal discovery

Anonymous ACL submission

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 mis-aligned 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 11 popular real-world benchmarks, for which our pipeline in total proceeds 38,081 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 the causal links between variables, being a fundamental 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 problem's target 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 performances, evaluated against synthetic or real-world benchmarks. However, a crucial dilemma of evaluation benchmarks hides these methods' effectiveness in handling real world challenges: A high-quality real-world benchmark dataset should contain a convincing ground truth causal graph, but this is challenging for many domains. As a result, many emerging causal discovery works rely on old real-world benchmarks, e.g., (lau, 1988; 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 of the popular real-world benchmarks contain out-of-date domain knowledge, potentially producing evaluation results that are inconsistent with the truth. This problem especially undermines the recently exploding new streamline 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.

A motivation example The benchmark Asia (lau, 1988), shown in Figure 6, is a popular causal discovery benchmark which is used 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 performances against the Asia benchmark, usually fully/almost recover the causal graph. However, domain research suggests that many of the bench-

mark's causal/conditional association relationships need update. For instance, the deep learning technique can now reveal bronchitis via x-ray scan (Chen et al., 2020; Ntiamoah et al., 2021; Nishino et al., 2014), contradicting against the benchmark, e.g., X-ray and bronchitis are mutually independent conditioned on smoking. Therefore, a high evaluation performance against such an out of date benchmark contrarily reveals problems of the method in practice.

In this work, we intend to examine the quality of popular real-world causal discovery benchmarks. First, we review the time independent causal discovery works published in artificial intelligence venues in recent years, and summarise the real-world benchmark usage 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 the meta-data from the arXiv. The number of LLM-based works drastically increases since 2023, expressing the necessity of systematically evaluating the causal discovery benchmarks.

Second, we investigate the alignment between the popular causal discovery benchmarks and corresponding domain research. The local Markov property of causal graphs entails a series of conditional association relationships between variables (the d-separation concept), which provide a natural connection to the controlled experimental analysis in most of the domain research works. We design a LLM-equipped pipeline to verify the consistency between the causal graphs and domain research papers. Practically, we evaluate 11 real-world benchmarks which are the most frequently used among the surveyed 69 benchmarks, and we analyse 38,081 research papers to mine the scientific evidence that supports or contradicts against the benchmarks.

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

Related works Several papers, e.g., (Brouillard et al., 2025; Gentzel et al., 2019), review the benchmark usage in causal discovery works, and summarise the benchmark types (i.e., synthetic or practical). Further before, (Glymour et al., 2019) encourages using semi-synthetic benchmarks to better align with the practice.

To the best of our knowledge, we are the first to systematically evaluate common benchmarks from the angle of domain research alignment. Our benchmark survey updates benchmark usage in the causal discovery domain, especially, benchmarks' popularity and usage in the emerging LLM-based causal discovery works. We postpone the detailed related work section to Appendix A.

2 Causal discovery benchmark revisit

We focus on semi-synthetic and real-world causal discovery benchmarks. Such datasets are constructed based on either human annotations, e.g., (Tu et al., 2019), statistical based methods using human selected features, e.g., (Spiegelhalter, 1992), or a combination of the two types of methods, e.g., (Sachs et al., 2005). As human judgment plays an important role in constructing such datasets, we call them empirical benchmarks. We evaluate the alignment with domain research of the popular benchmarks in Section 3.

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

2.1 Benchmark usage in AI publications

Paper selection We select papers published in main AI conferences in the latest 5 years. These conferences include NeurIPS, ICML, ICLR, AAAI, IJCAI, ACL, and EMNLP. The selection process consists of two rounds. In the first round, we filter published papers using keyword "causal", and manually select the papers that propose methods for *time-independent causal discovery*, based on papers' full content. In the second round, we check the benchmarks used in each paper's experimental, remove papers that purely use synthetic datasets, and record the empirical benchmarks used in the remaining papers. Note that we do not include papers that study causal relationship in language itself or among multimodal data.

Summary The selection process results in 96 papers with at least one empirical benchmark. There are in total 69 unique benchmarks used in these papers. The benchmark-paper details are listed in Appendix B.

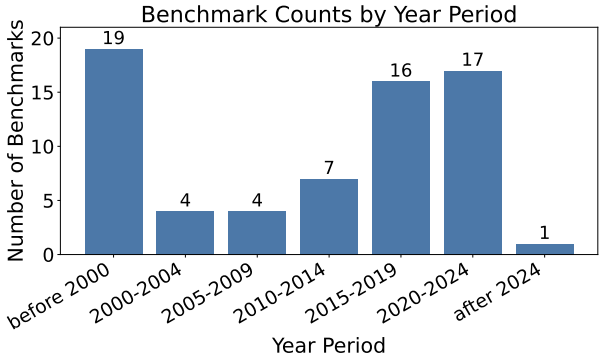

Figure 1: All benchmarks by year period.

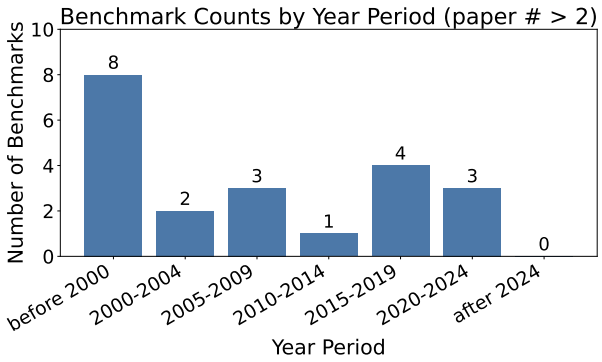

Figure 2: Benchmarks used in at least 3 papers by year period.

Figure 1 counts the number of benchmarks cited by these papers of a given year, and Figure 2 counts the number of frequently used benchmarks, i.e., those appear in at least three papers. 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. There are 34 new benchmarks emerged since 2015, indicating the community is aware of the insufficient benchmark problem in recent years.

21 out of 69 benchmarks are frequently used benchmarks (Figure 3). The distribution of the frequently used benchmarks is relatively even over years. Researchers still like to use many of the old benchmarks, observing 8 out of 21 are constructed before 2000, while 7 out of 21 are 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, expect the pairwise Tübingen dataset, are recorded in the bnlearn network repository[1], revealing its popularity in the domain. 15 among the 21 popular benchmarks are from the domains of medical science and biology.

[1]https://www.bnlearn.com/

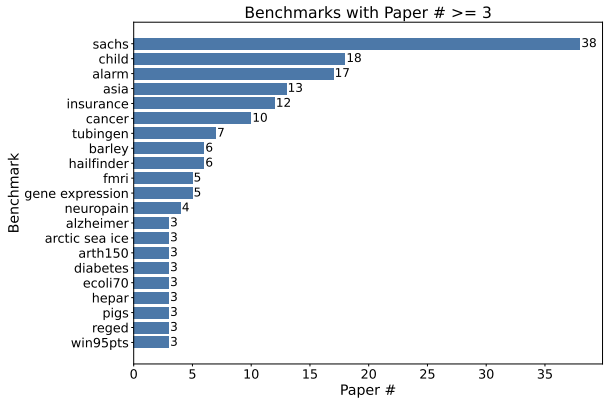

Figure 3: Number of papers that use frequently used benchmarks.

LLM-based methods LLMs have become a tool in causal discovery in recent years. Among the 96 papers, 10 papers propose LLM-based causal discovery methods.They were published very recently with 9 of them published since 2025. That is, LLM based methods make 31% of the surveyed causal discovery papers after 2024 (29 papers in total). There are 28 distinct benchmarks used in these 10 papers: these papers use diverse benchmarks (see details in Appendix B.2). However, the top 3 frequently used benchmarks in LLM-based papers are: Asia, Cencer, and Sachs, all of them from the bnlearn small network repository.

2.2 Emerging LLM-based causal discovery works

arXiv search process As new papers often appear in arXiv first, analyzing their benchmark use can reveal the trend 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 by two keyword sets: one for causal discovery and one for LLMs. Then, we submit the queries to the arXiv API to search papers of which the title or the abstract contain both keywords, expecting to retrieve LLM-based causal discovery papers.

We also extract benchmark usage information contained in the abstract through rule-based matching against an extended benchmark inventory constructed by LLMs (details shown in Appendix B.3).

Summary Figures 4 and 5 show the number of papers that propose LLM-based causal discovery

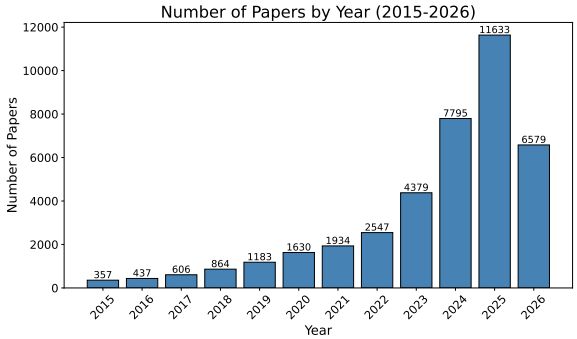

Figure 4: Number of papers/year by searching “causal discovery" and “LLM".

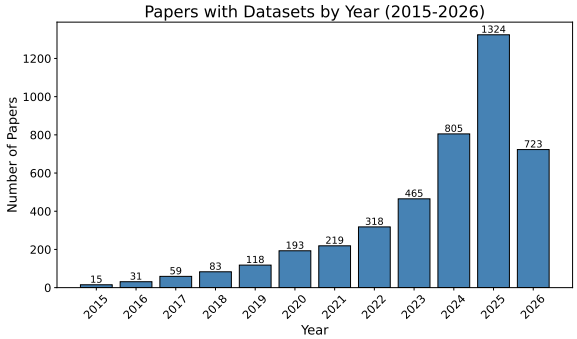

Figure 5: Number of papers/year by searching “causal discovery", “benchmark", and “LLM".

methods and LLM-based causal discovery with evaluation by empirical datasets, respectively. The two numbers have a similar trend: both of them stably increase from 2015 to 2025, and the trends have become drastic since 2022. This trend is strongly stimulated by the rapid advancement of LLMs in recent years. In 2025, there are up to 11633 LLM-based causal discovery papers, among which up to 1324 papers contain evaluation based on empirical benchmarks.

In Appendix B.3, the arXiv search results show the top 20 benchmarks and our survey benchmarks have overlapped 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, $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed edges. For each pair of variables $V_i, V_j \in \mathbf{V}$, (V_i, V_j) presenting in $\mathbf{E}$ denotes an directed edge from V_i to V_j exists, 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 $V_{ij}, V_{i(j+1)}$ are 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 $\mathbf{E}$. Typically, a DAG does not contain any cycle, i.e., 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., $pa_{\mathcal{G}}(V_i) = \{V \in \mathbf{V} \mid (V, V_i) \in E\}$, 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 $V_i \rightarrow V_j \leftarrow V_k$ and V_i and V_k are not adjacent in $\mathcal{G}$, we call V_j a collider.

Local Markov property An ubiquitous assumption for a causal graph 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 statistically independent.

d-separation The local Markov property entails a series of conditional association relationship between variables. The definition of *d-separation* characterises such conditional association relationship reflected from 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 of 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. Note that $\mathbf{Z}$ can be empty.

Example 3.1 *Figure 6 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*

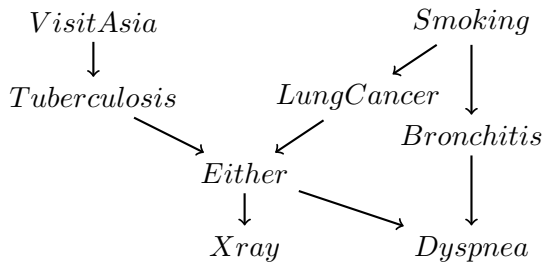

Figure 6: The causal graph of the Asia benchmark dataset: a semi-synthetic benchmark dataset.

$Smoking$, and therefore, we have that

$$Xray \perp Bronchitis \mid Smoking.$$

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 are always correlated, even 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 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 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 from the causal graph, and then, evaluate what the proportion of the retrieved papers contain inconsistent conclusions than the causal graph's conditional association relationships. This inconsistency rate measures the mis-alignment between the causal graph and the domain research.

3.2 Methodology

We design a LLM equipped pipeline to evaluate causal graphs' consistency 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 causal graph conditional association relationship 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 where all papers study the relationship between V_i and V_j, by searching 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 relevant papers in a database, download the accessible papers among the search result, and extract the text from the downloaded files to make papers LLM-reading ready.

A simple search query "V_i and V_j" is not sufficient for efficient search due to the following facts: (i) Inconsistent term using is normal in various research domains, e.g., the term "Smoking" in Example 3.1 may be expressed as "Tobacco use" or "Cigarette consumption". (ii) Many database uses keyword matching against the full paper in the search engine, and the search query may retrieve a large number of irrelevant papers. Hence, we employ the *query expansion* (Carpineto and Romano, 2012) technique to form a search query, and add an extra filter 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, it forms a query:

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

Upon the returned search results, the filter screens in papers of which the abstract contains terms from both sets $\{V_i^1, \cdots, V_i^{ki}\}$ and $\{V_j^1, \cdots, V_j^{kj}\}$. Then, in the final paper list, 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.

Conditional association extraction Algorithm 1 deterministically extracts the conditional association relationships from causal graph $\mathcal{G}$: (i) if V_i and V_j can be d-separated by an empty set, $V_i \perp V_j$; (ii) if V_i and V_j can be d-separated by set $\mathbf{Z}$, $V_i \perp V_j | \mathbf{Z}$;

and (iii) if V_i and V_j are adjacent in $\mathcal{G}$, they are always correlated. Observe that V_i and V_j can be d-separated by multiple sets, and we record all such conditional association relationships. Then, for each conditional association relationship, we formulate a query, e.g., "Is V_i and V_j are mutually independent conditioned on variables in $\mathbf{Z}$?"

Consistency verification With the paper contents and the conditional association queries prepared, we employ LLMs to verify if the each paper's conclusion is consistent with each conditional association query. Especially, we provide the mathematical intuition and example expressions of the each conditional association type in the prompt, and query the LLMs to verify the consistency (see prompt in Appendix C). Each LLM query may return "unknown" when the paper content cannot support the decision making, or "consistent"/"inconsistent" when the paper content is consistent/inconsistent with the query.

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 are used in at least 3 causal discovery papers. There are 21 popular benchmarks in total, and we exclude 10 benchmarks: (i) the pairwise benchmark (Tübingen) due to its simple conditional association type; (ii)) the large benchmarks (especially, Hailfinder, Gene expression, Neuropain, Arth150, Hepar2, Pigs, Reged, Barley, and WIN95PTS) due to our limited computation capacity.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 only return up to 500 the most relevant papers, and download the open access papers among the searched papers. In the 11 evaluated benchmarks, some of them have more than 200 combinations of variable pairs, and we randomly sample 200 variable pairs to evaluate. Therefore, the pipeline processes up to 100000 full papers for each benchmark[2].

Most of the benchmark causal graphs are from medical science and biology. For such causal graphs, we use PubMed[3] as the scientific database to search papers, and we download papers that are available in the open access database PMC. For causal graphs from the other domains, we use Openalex[4] to search papers, and download papers accessible through the institution's subscription databases. The downloaded pdf files are converted into json files by the Mistral ocr API.

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[5] 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[6].

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

In the entire process, our pipeline processes 38,081 full papers (see details in Table 1). We measure the inconsistency rates of these papers on benchmarks.

4.1 Inconsistency rates and paper distribution

Figure 7 shows the benchmark with the lowest inconsistency rate (26.70%) is Sachs, which potentially best aligns with the domain research, and it is the most frequently used benchmark (38 surveyed papers use it). The second and third benchmarks

[2]The actual paper number is much fewer 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.

[3]https://pubmed.ncbi.nlm.nih.gov/

[4]https://openalex.org/

[5]https://www.nlm.nih.gov/research/umls/index.html

[6]To show the accuracy of the LLM-based consistency verifier, we randomly sample 100 retrieved papers, and manually check the correctness of the answer. The verifier achieves an accuracy of 0.9. See details in Appendix C.3

benchmarks	# nodes	# edges	# node pairs	# retrieved papers	benchmark 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
ecoli70	46	70	200	2741	2005

Table 1: Details of selected benchmark datasets.

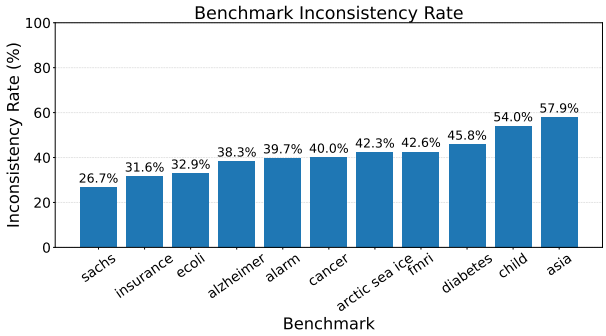

Figure 7: Benchmarks' inconsistency rates.

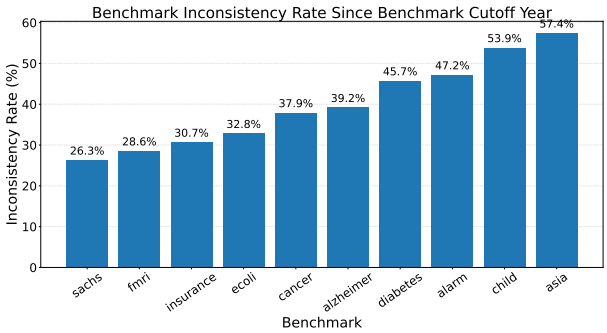

Figure 8: Benchmarks' inconsistency rates since the construction year.

are Insurance and Ecoli, with inconsistency rates 31.6% and 32.9%, respectively.

In contrast, Diabetes, Child, and Asia have the highest inconsistency rates (45.8%, 54.0%, and 57.9%, correspondingly). It is worth noting that all of them are constructed before 2000.

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

We also use the construction year as a cutoff year to measure the inconsistency rate. When only considering papers published after each benchmark's construction year, the inconsistency rates vary significantly as shown in Figure 8. Fmri's consistency rate reduces from 42.6% to 28.6%, with the ranking promotes to the second. In contrast, Alarm's inconsistency rate increases from 39.7% to 47.2%. All prior year 2000 benchmarks, except for Insurance, are ranked the last, i.e., Diabetes, Alarm, Child, and Asia.

Figures 9 and 10 show the number of retrieved relevant papers (i.e., papers do not return "Unknown") and inconsistent papers over years, respectively. Observe that for most of the benchmarks, both paper numbers start increasing since 2001, and achieve the highest from 2011 to 2025. Interestingly, many of the benchmarks have a relatively lower number of inconsistent papers around their publication year compared to the number in the preceding and subsequent years, especially, Alarm has the most obvious trend (see the detailed inconsistent paper numbers in Appendix C.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 grow along time as the problem understanding progresses. Figure 8 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 the dated benchmarks help the devel-

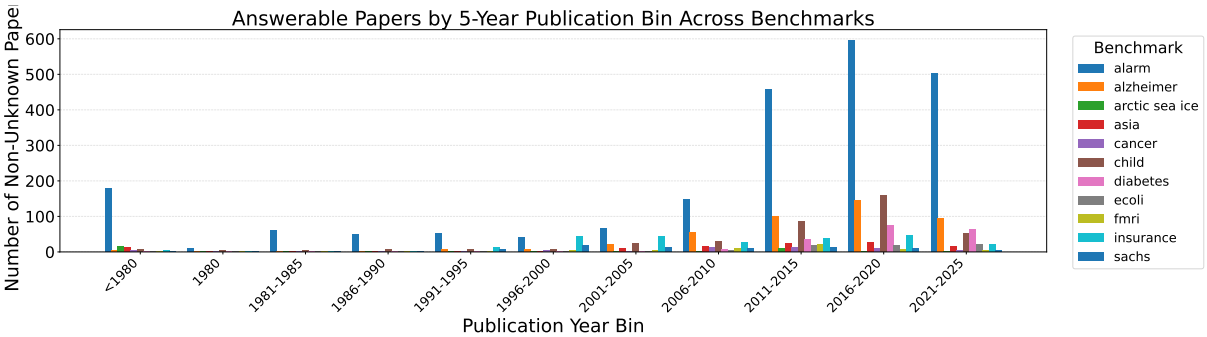

Figure 9: Benchmarks' relevant paper number per year period.

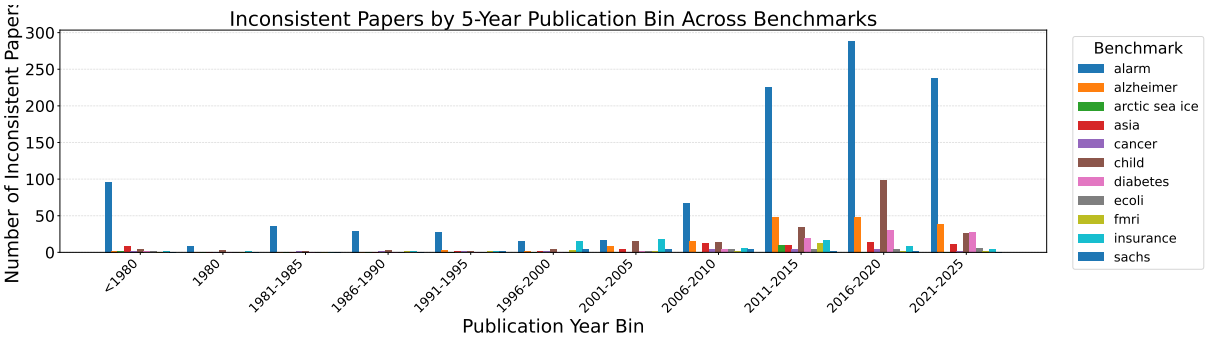

Figure 10: Benchmarks' inconsistent paper number per year period.

opment of better causal discovery methods.

Hypothesis on growing inconsistency For almost benchmarks, the inconsistent paper number increases after benchmarks' publication year, whilst the inconsistency rates do not have such a trend. We conjecture that the subsequent higher inconsistent paper number is due to richer study 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 are latent confounders previously. Such studies may not invalidate the previous benchmarks, but they consider additional factors missing in previous studies, which may lead to the inconsistency.

Benchmark construction process matters The Sachs benchmark achieves the lowest inconsistency rates in both Figures 7 and 8. 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 guarantees its alignment with the domain knowledge. We advocate 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 they align with changing domain knowledge. We designed a 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 alignment levels with the domain research. Especially, the mis-alignment issue is more severe on old benchmarks.

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

Limitations

The limitations of our work is mainly in two folds.

(1) **The accuracy of the LLM paper verifier** To verify the consistency between a causal graph and each research paper from the vast paper pool, the use of LLMs is a necessity. However, employing LLMs unavoidably introduces noise in the process. Empirically, we find that a more advanced LLM tool can significantly improve our evaluation result. However, this also requires a more generous budget for the computation or LLM API callings. Due to the vast size of our retrieved paper pool, we deploy a small-size open-source LLM (Qwen3-30B-A3B-Thinking-2507) in the experiment. We conjecture that the result would be further improved under the use of a more powerful LLM tool.

(2) **Paper retrieving efficiency** In the current version, we simply send a searching query, which is formulated by a query expansion technique, to scientific databases' search engines. We aim to obtain a comprehensive and unbiased paper list. However, the component's paper retrieval efficiency is still low, forcing us to analyse more papers. Scientific databases also have varying paper retrieving capability. PubMed performs well on biomedical paper retrieval, however, the other domains rely on different databases, and we find the database quality also bottlenecks our evaluation.

Moreover, the access to full papers is still significantly limited, presenting another main challenge for vast domain paper based experiments such as ours.

References

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

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

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

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

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

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

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

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

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

Ingo A Beinlich, Henri Jacques Suermondt, R Martin Chavez, and Gregory F Cooper. 1989. 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*, pages 247–256. Springer.

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

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

Philippe Brouillard, Chandler Squires, Jonas Wahl, Karen Sachs, Alexandre Drouin, Dhanya Sridhar, and 1 others. 2025. The landscape of causal discovery data: Grounding causal discovery in real-world applications. In *Causal Learning and Reasoning*, pages 834–873. PMLR.

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

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

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, and 1 others. 2020. Diagnosis of common pulmonary diseases in children by x-ray images and deep learning. *Scientific reports*, 10(1):17374.

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

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

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

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, and 1 others. 2011. The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654):1–28.

Cristina Conati, Abigail S Gertner, Kurt VanLehn, and Marek J Druzdzel. 1997. 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*, pages 231–242. Springer.

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

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

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

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

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

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

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

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

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

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

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

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

Tao Feng, Lizhen Qu, Niket Tandon, and Gholamreza Haffari. 2025a. 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)*, pages 9400–9428.

Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. 2025b. 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)*, pages 9565–9590.

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

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, and 1 others. 2021. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341.

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

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

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

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

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

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

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

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

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

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

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

Marianne J Heins, Hans Knoop, William J Burk, and Gijs Bleijenberg. 2013. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

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

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

Kristian Kristensen and Ilse A Rasmussen. 2002. 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.

Chanhui Lee, Juhyeon Kim, YongJun Jeong, Yoonseok Yeom, Juhyun Lyu, Jung-Hee Kim, Sangmin Lee, Sangjun Han, Hyeokjun Choe, Soyeon Park, and 1 others. 2024. On incorporating prior knowledge extracted from pre-trained language models into causal discovery. In *Causality and Large Models@ NeurIPS 2024*.

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

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

Mingjia Li, Hong Qian, Tian-Zuo Wang, Min Zhang, Aimin Zhou, and 1 others. 2025a. Strong and weak identifiability of optimization-based causal discovery in non-linear additive noise models. In *Forty-second International Conference on Machine Learning*.

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

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

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

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

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

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

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

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

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

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

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

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

Alessandro Magrini, Stefano Di Blasi, Federico Mattia Stefanini, and 1 others. 2017. 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.

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

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

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

Ehsan Mokhtarian, Mohmmadsadegh Khorasani, Jalal Etesami, and Negar Kiyavash. 2023. 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, pages 12260–12268.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102.

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

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

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

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

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

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

Rainer Opgen-Rhein and Korbinian Strimmer. 2007. 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.

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

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, and 1 others. 2010. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209.

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

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, and 1 others. 2015. Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1):8885.

Jie Qiao, Zhengming Chen, Jianhua Yu, Ruichu Cai, and Zhifeng Hao. 2024a. 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, pages 20516–20523.

Jie Qiao, Yu Xiang, Zhengming Chen, Ruichu Cai, and Zhifeng Hao. 2024b. 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, pages 20524–20531.

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

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

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

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

Amartya Roy, N Devharish, Shreya Ganguly, and Kripabandhu Ghosh. 2025. 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.

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

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

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

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

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

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

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

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

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

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

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

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search*. The MIT Press.

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

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

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. 2024. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*.

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

Quang-Duy Tran, Bao Duong, Phuoc Nguyen, and Thin Nguyen. 2025. Identifying causal direction via variational bayesian compression. *arXiv preprint arXiv:2505.07503*.

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

Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. 2022. 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.

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

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

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

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

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

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

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

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

JN Warwick, TL Sellers, SR Talbot, AJ Cawthorn, and WB Ford. 1994. 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).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Additional related works

Causal graph benchmarks 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 benchmark usage of the three types, and reviews the real-world and semi-synthetic benchmarks. Earlier than 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 the practice, compared to using purely synthetic data.

Our work focuses on semi-synthetic and real-world benchmarks. We summarise the detailed usage of such benchmarks, especially, with an update on the benchmark usage in LLM-based causal discovery methods, and evaluate their aligningness with real-world knowledge.

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.

B Additional information of literature review

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.

B.1 Full list of the benchmarks and causal discovery papers

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 and Kasiviswanathan, 2021; Cundy et al., 2021; Wang et al., 2025a; Li et al., 2025a; Elahi et al., 2024; Wang et al., 2024; Kaltenpoth and 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)

2. Child
   Original paper (Spiegelhalter, 1992)
   Papers using it: (Shen et al., 2025; Peyrard and 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 and 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 and 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 and Nicholson, 2010)
   Papers using it: (Feng et al., 2025a,b; Shahverdikondori et al., 2024; Peyrard and 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 and 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 and 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 and 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 and 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 and 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 and Kong, 1999)
    Papers using it: (Wang et al., 2022; Ling et al., 2025b)

24. Mildew
    Original paper (Jensen and 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 and 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. (no?)Chemistry image
    Original paper (Schäfer and Strimmer, 2005)
    Papers using it: tba

47. climatic analysis
    Original paper (Compo et al., 2011)
    Papers using it: (Liu et al., 2024a)

48. COVID-19
    Original paper Link
    Papers using it: (Wang et al., 2025b; Vashishtha et al., 2025)

49. Credit
    Original paper (Quinlan, 1987)
    Papers using it: (Li et al., 2022)

50. dream
    Original paper (Kalainathan and Goudet, 2019)
    Papers using it: (Roy et al., 2025)

51. football
    Original paper Link
    Papers using it: (Qiao et al., 2024b)

52. G7
    Original paper (Demirer et al., 2018)
    Papers using it: (Jalaldoust et al., 2022)

53. General social survey
    Original paper Link
    Papers using it: (Li et al., 2025b)

54. IHDP
    Original paper (Hill, 2011)
    Papers using it: (Ashman et al., 2023)

55. Lucas
    Original paper (Lucas et al., 2004)
    Papers using it: (Roy et al., 2025)

56. Magnetic
    Original paper (Hwang et al., 2024)
    Papers using it: (Zhao et al., 2025)

57. (tba)Micro24
    Original paper (Schäfer and Strimmer, 2005)
    Papers using it: tba

58. (tba)Micro25
    Original paper (Schäfer and Strimmer, 2005)
    Papers using it: tba

59. Munin
    Original paper (Andreassen et al., 1989)
    Papers using it: (Dong et al., 2025; Wang et al., 2022)

60. Pacific walker circulation
    Original paper (Runge et al., 2019)
    Papers using it: (Liu and Kuang, 2023)

61. Pathfinder
    Original paper (Heckerman et al., 1992)
    Papers using it: (Zhang et al., 2021)

62. Pharmacokinetics
    Original paper (Grzegorzewski et al., 2021)
    Papers using it: (Li et al., 2024a)

63. Physics
    Original paper (Lee et al., 2024)
    Papers using it: (Kang et al., 2026)

64. Protein
    Original paper Link
    Papers using it: (Peyrard and West, 2020)

65. Respiratory disease
    Original paper (Long et al., 2022)
    Papers using it: (Feng et al., 2025a)

66. Sangiovese
    Original paper (Magrini et al., 2017)
    Papers using it: (Abdulaal et al., 2024)

67. SCM
    Original paper (Ferro et al., 2015)
    Papers using it: (Aglietti et al., 2023; Sussex et al., 2022)

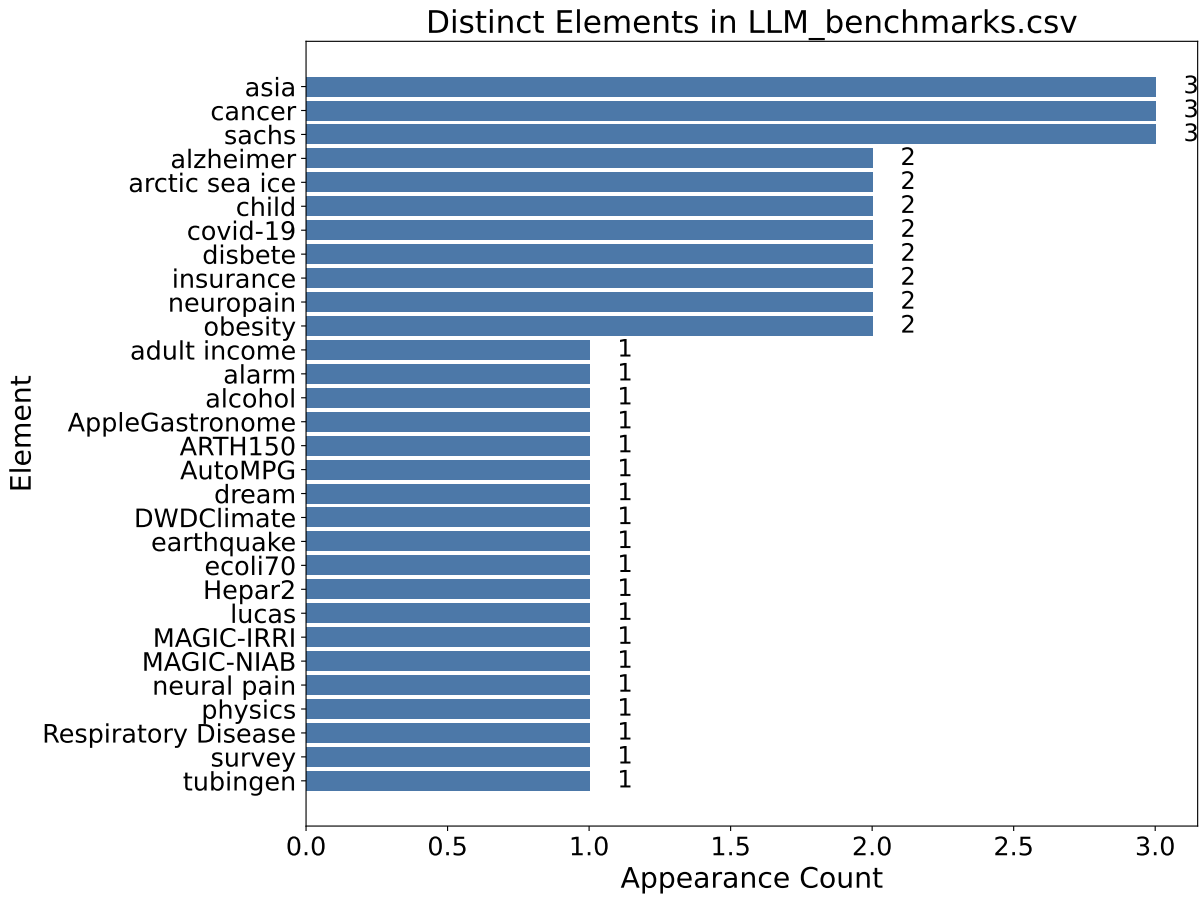

Figure 11: Benchmark numbers used in LLM-based papers.

68. Sepsis
    Original paper (Matot and Sprung, 2001)
    Papers using it: (Liu et al., 2024b)

69. 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)

70. Water
    Original paper (Jensen et al., 1989)
    Papers using it: (Mokhtarian et al., 2023)

71. World value survey
    Original paper Link
    Papers using it: (Dai et al., 2025)

72. Yahoo
    Original paper: (Janzing et al., 2010)
    Papers using it: (Zeng et al., 2021)

B.2 LLM-based causal discovery papers

Figure 11 shows the benchmarks used in our surveyed LLM-based causal discovery works, with the number of papers that use each benchmark.

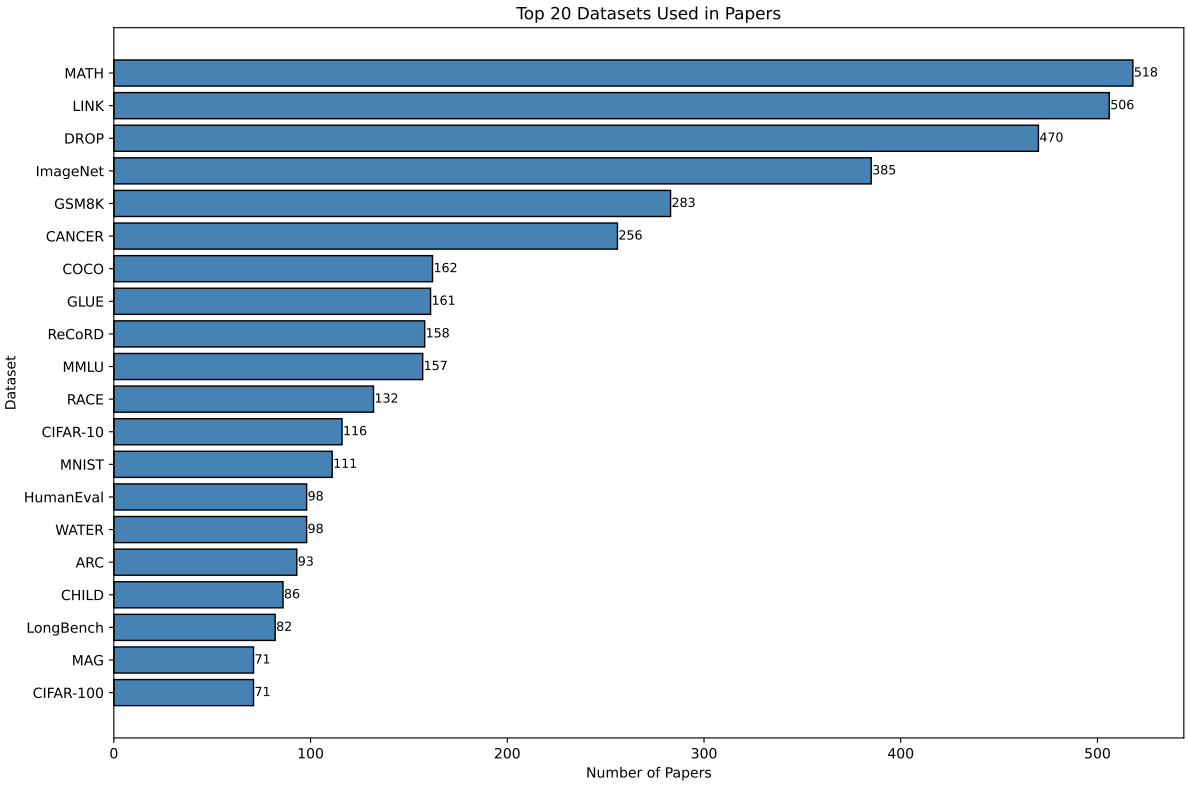

Figure 12: Top 20 benchmarks used in arXiv LLM-based causal discovery papers.

B.3 LLM-based causal discovery papers benchmarks on arXiv

Figure 12 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.

C Additional information of the benchmark evaluation process

C.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.

C.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 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 13 to 23 show the inconsistent paper number per year for all evaluated benchmarks.

C.3 LLM-based paper verifier accuracy

A pipeline component that potentially brings high noise is the LLM-based consistency verifier which employs the LLMs to analyse papers and decide the papers' consistency with the conditional associations. We randomly sample 100 retrieved papers, and manually verify the correctness of the LLM-based consistency verifier's answers. The accuracy achieves 90%, i.e., the LLM-based consistency verifier correctly analyses 90 out of the 100 randomly sampled papers. The follows are the details of manual check on the LLM-based verifier answers.

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
    $SepSet \leftarrow \emptyset$.
    $\mathbf{D} \leftarrow \{\mathbf{V}' \subseteq \mathbf{V} \setminus \{X, Y\} \mid \forall P \in \mathbf{P}, P$ 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

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.

- **Benchmark name:** alarm
  Pair name: Intubation Stroke volume
  Query: independent
  Paper title: Tracheal intubation in traumatic brain injury: a multicentre prospective observational study
  Answer: unknown
  LLM thought: The paper examines the association between tracheal intubation (prehospital and in-hospital) and functional outcome (Glasgow Outcome Scale Extended) in traumatic brain injury patients, adjusting for confounders such as age, GCS, pupil reactivity, etc. The adjusted odds ratios for prehospital intubation (OR=1.01, p=0.96) and in-hospital intubation (OR=0.86, p=0.28) are not statistically significant, indicating no significant marginal correlation between intubation and functional outcome. However, the query asks about correlation between Intubation and Stroke volume, but Stroke volume is not mentioned in the paper. Therefore, based on the available data, there is no information about Stroke volume, so the answer should be 'Unknown'. However, the instructions require answering for the factors Intubation and Stroke volume. Since Stroke volume is not present, we cannot establish any relationship. But given the output must be one of the three

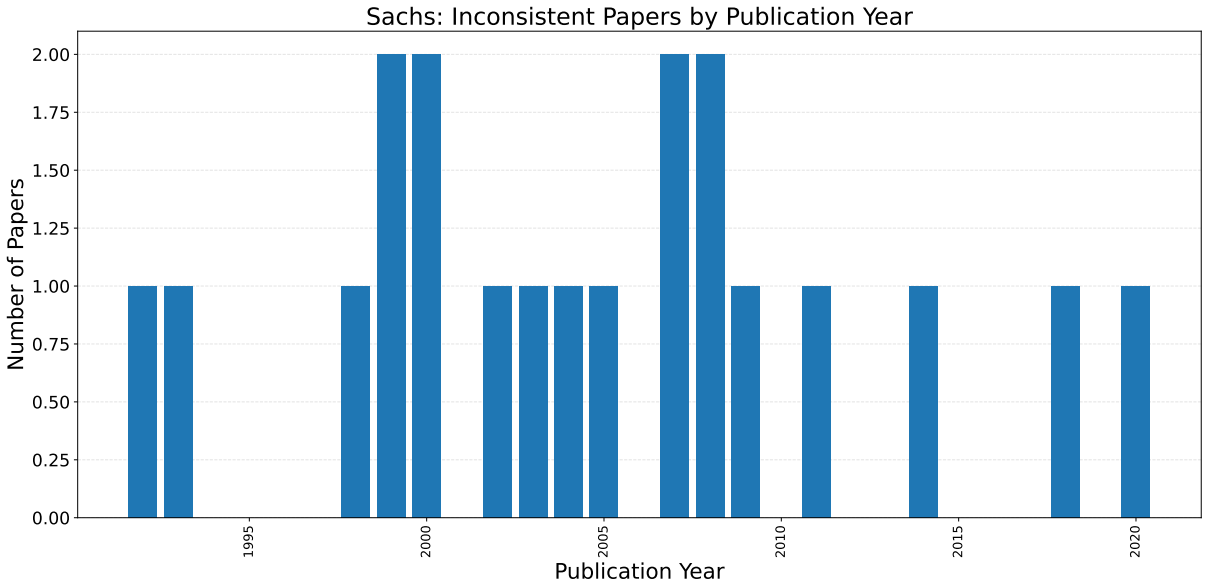

Figure 13: Inconsistent paper number per year of Sachs.

choices, and there is no data, the answer is 'Unknown'.
Label: True

- **Benchmark name:** alarm
  Pair name:
  Query: independent
  Paper title: Pneumonia in acute stroke patients fed by nasogastric tube
  Answer: correlated
  LLM thought: The paper reports a significant association between intubation (endotracheal intubation) and pneumonia, and pneumonia is associated with stroke severity, but does not directly analyze the marginal independence between intubation and stroke volume. However, Table 1 shows that patients with pneumonia had significantly higher NIHSS scores (reflecting stroke severity) and more frequently required intubation, implying a correlation. The query asks about intubation and stroke volume; stroke volume is not directly measured, but NIHSS is a measure of stroke severity. The paper shows that intubation is correlated with stroke severity (NIHSS), thus intubation and stroke volume (as stroke severity) are correlated.
  Label: False

- **Benchmark name:** alarm
  Pair name:
  Query: independent
  Paper title: Identifying dysphagia in the intensive care unit: Validation of the Swedish version of the Gugging swallowing screen—Intensive care unit
  Answer: unknown
  LLM thought: The paper studies dysphagia screening in ICU patients, with intubation as an inclusion criterion (all patients were intubated >48h). Stroke volume is not mentioned; instead, stroke or neurological conditions are admission diagnoses. The query asks about marginal independence between Intubation and Stroke volume. The paper does not analyze the relationship between these two factors; it treats intubation as a given condition and does not report any statistical test linking intubation duration or presence to stroke volume. Without data analysis on their correlation, we cannot conclude they are correlated.
  Label: True

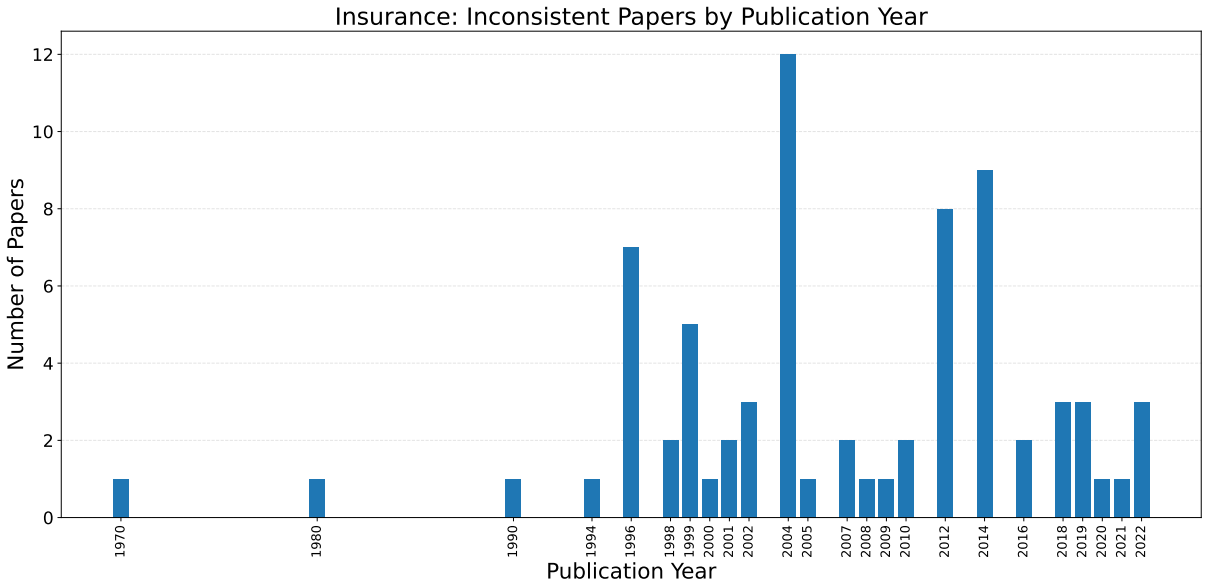

Figure 14: Inconsistent paper number per year of Insurance.

- **Benchmark name:** alarm
  Pair name:
  Query: independent
  Paper title: International Survey of Mechanical Thrombectomy Stroke Systems of Care During COVID-19 Pandemic
  Answer: independent
  LLM thought: The paper compares infection rates between centers with preferred intubation (PI) and preferred non-intubation (PNI) for MT, and finds no significant difference (p=0.60), indicating marginal independence between Intubation policy and Stroke volume (as a proxy for infection rate).
  Label: False

- **Benchmark name:** alzheimer
  Pair name: AV45 Education level
  Query: independent
  Paper title: Both hyper- and hypo-activation to cognitive challenge are associated with increased beta-amyloid deposition in healthy aging: A nonlinear effect
  Answer: independent
  LLM thought: The paper reports a t-test comparing education level between low and high amyloid groups (dichotomized by SUVR cutoff) and finds no significant difference (p >.146), indicating no association between AV45 (amyloid burden) and Education level, so they are independent marginally.
  Label: True

- **Benchmark name:** alzheimer
  Pair name:
  Query:
  Paper title: Use of the PET ligand florbetapir for in vivo imaging of pancreatic islet amyloid deposits in human islet amyloid polypeptide transgenic mice
  Answer: unknown
  LLM thought: The paper contains data analysis (as it reports statistical tests and results), but it does not involve the variable 'Education level' at all. The study is on mice and does not collect human education level data, so the mutual independence of AV45 and Education level cannot be determined from this paper.
  Label: True

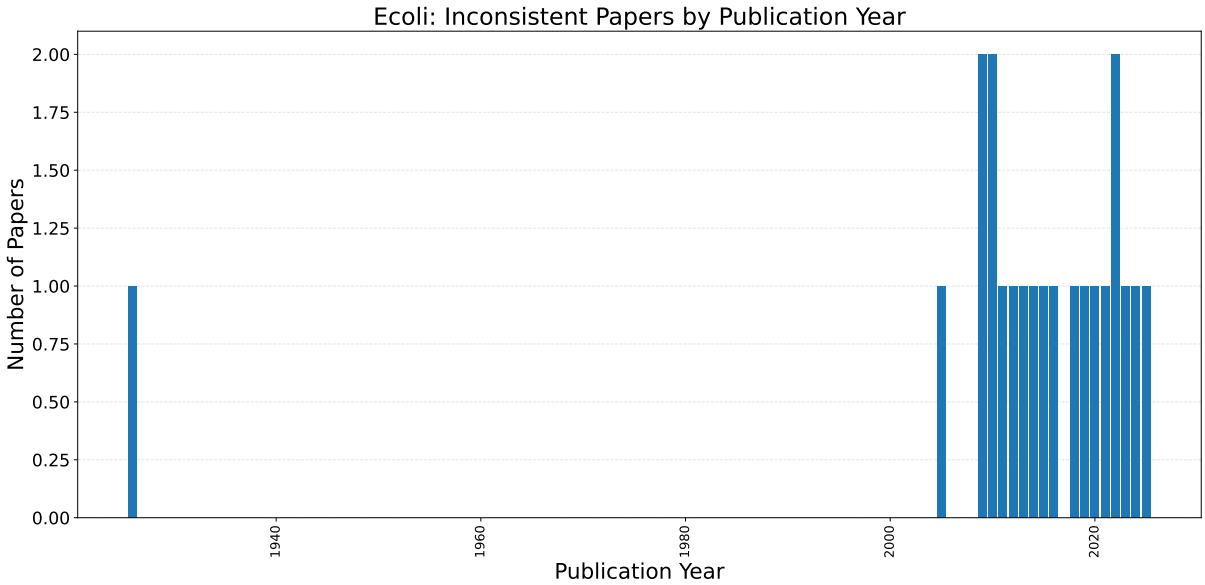

Figure 15: Inconsistent paper number per year of Ecoli.

- **Benchmark name:** alzheimer
  Pair name: AV45 Slice number
  Query: independent
  Paper title: Longitudinal Trajectories of the Cognitive Function Index in the A4 Study
  Answer: unknown
  LLM thought: The paper contains data analysis but does not include the variables 'AV45' and 'Slice number' as defined in the query. The amyloid measure is referred to as 'FBP Centiloid' or 'FBP SUVr', not 'AV45', and 'Slice number' is not mentioned at all.
  Label: True

- **Benchmark name:** alzheimer
  Pair name: Brain MRI APOE4
  Query: AV45
  Paper title: APOE 4 Is Associated with Disproportionate Progressive Hippocampal Atrophy in AD
  Answer: correlated
  LLM thought: The paper investigates the relationship between APOE4 and hippocampal atrophy rates while adjusting for whole-brain atrophy rates. The statistical analysis shows that after adjusting for whole-brain atrophy (AV45 is not mentioned, but whole-brain atrophy is used as a covariate), the difference in hippocampal atrophy rates between APOE4 carriers and non-carriers remains statistically significant in AD and MCI-P. This indicates that Brain MRI (hippocampal atrophy rates) and APOE4 are not independent when conditioned on whole-brain atrophy (which serves as the covariate). The paper does not mention AV45, but the analysis conditions on whole-brain atrophy, which is the relevant covariate for the query.
  Label: False

- **Benchmark name:** alzheimer
  Pair name:
  Query:
  Paper title: APOE 4 associated with preserved executive function performance and maintenance of temporal and cingulate brain volumes in younger adults
  Answer: unknown
  LLM thought: The paper does not contain AV45 data, so it cannot condition on AV45. Therefore, the query cannot be answered.

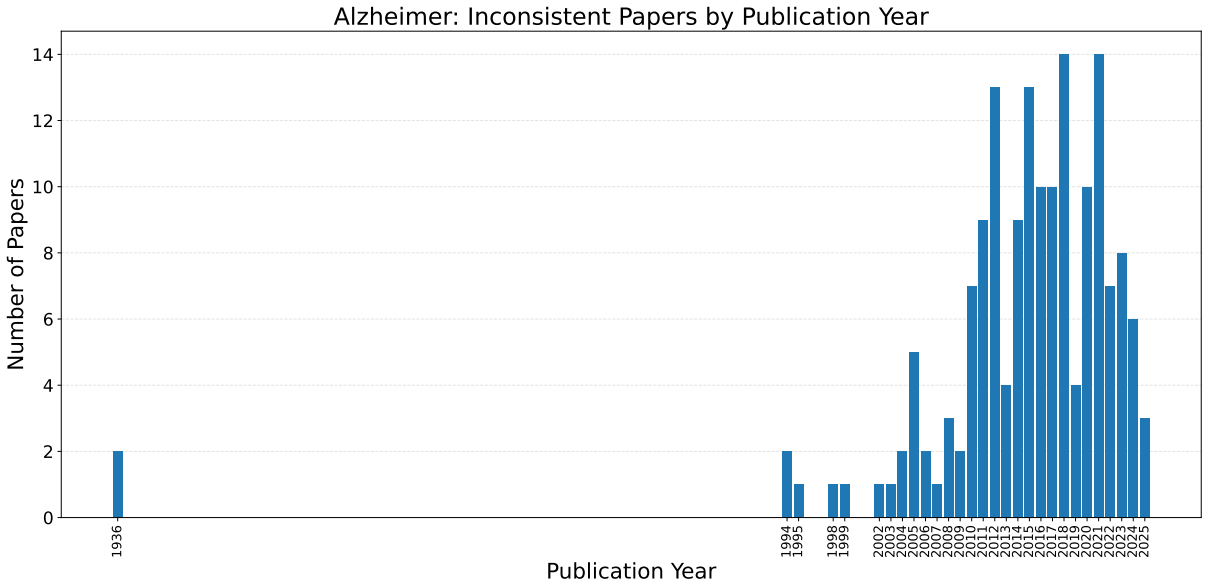

Figure 16: Inconsistent paper number per year of Alzheimer.

Label: True

- **Benchmark name:** alzheimer
  Pair name: Brain volume AV45
  Query: direct
  Paper title: Hippocampal subfield volumes and preclinical Alzheimer's disease in 408 cognitively normal adults born in 1946
  Answer: independent
  LLM thought: The paper studies hippocampal volumes (total and subfields) and amyloid (AV45). The query's 'Brain volume' is interpreted as total hippocampal volume. The paper reports a non-significant correlation (p=0.075) for total hippocampal volume after conditioning on covariates (age, WMHV, sex, TIV), so the correlation is not present in the conditional sense.
  Label: True

- **Benchmark name:** alzheimer
  Pair name: Brain volume Education level
  Query: independent
  Paper title: Cerebrospinal Fluid Biomarkers, Education, Brain Volume and Future Cognition
  Answer: unknown
  LLM thought: The paper analyzes the interaction of brain volume and education with CSF biomarkers in predicting cognitive impairment, but does not test the marginal association between brain volume and education. Therefore, the data analysis does not provide information to answer the query.
  Label: True

- **Benchmark name:** alzheimer
  Pair name: Education level MOCA score
  Query: direct
  Paper title: Functional deficits and other psychiatric associations with abnormal scores on the Montreal Cognitive Assessment (MoCA) in older HIV-infected patients
  Answer: independent
  LLM thought: The paper's adjusted analysis (controlling for covariates such as non-white race, gait speed, and social support) shows no significant association between education level and MOCA score, indicating conditional independence. The unadjusted analysis showed a significant association, but

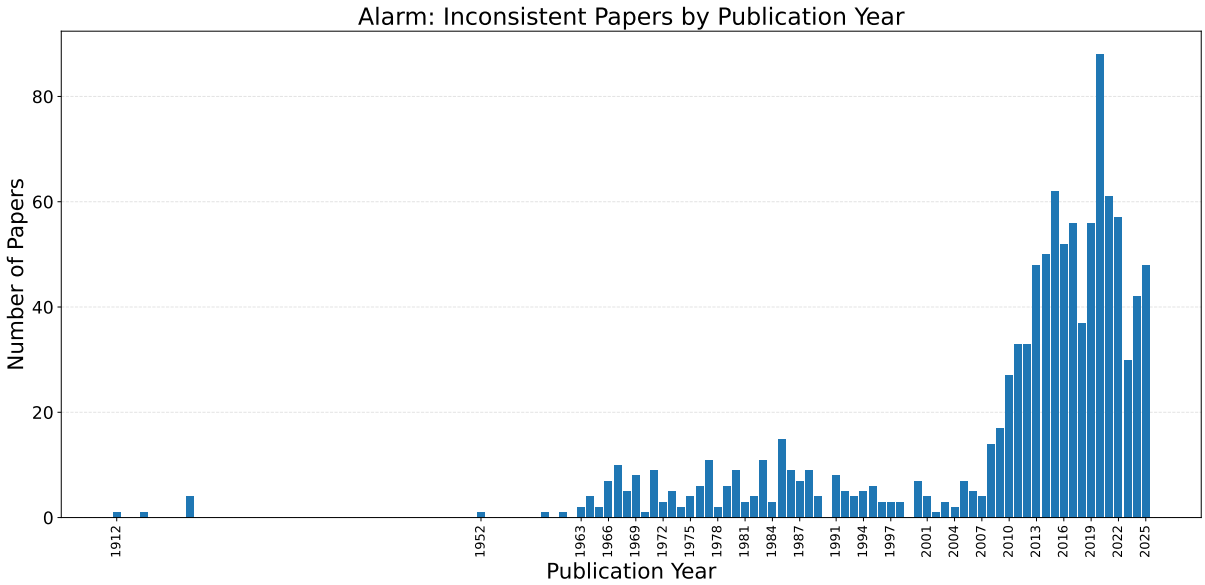

Figure 17: Inconsistent paper number per year of Alarm.

the adjusted model removed education level, meaning the correlation is not present after conditioning on other factors.
Label: True

- **Benchmark name:** alzheimer
  Pair name: P tau Education level
  Query: independent
  Paper title:
  Answer: independent
  LLM thought: The paper reports a single-factor linear regression analysis (marginal association) between education level and p-tau217 in Table 2, with a p-value of 0.068, which is not significant at the 0.05 level. Therefore, the factors are mutually independent marginally.
  Label: True

- **Benchmark name:** alzheimer
  Pair name: Sex AV45
  Query: independent
  Paper title: Longitudinal Evolution of Financial Capacity and Cerebral Tau and Amyloid Burden in Older Adults with Normal Cognition or Mild Cognitive Impairment
  Answer: unknown
  LLM thought: The paper uses gender as a covariate in the main analysis but does not test the association between gender and amyloid (AV45) in the marginal distribution. There is no statistical test reported for the independence of Sex and AV45.
  Label: True

- **Benchmark name:** alzheimer
  Pair name:
  Query: independent
  Paper title: The relationship between leisure time physical activity patterns, Alzheimer's disease markers and cognition
  Answer: independent
  LLM thought: The paper reports a marginal association test between Sex and A status (AV45) in Table 1 with a p-value of 0.4 (not significant), indicating no evidence of association, so they are

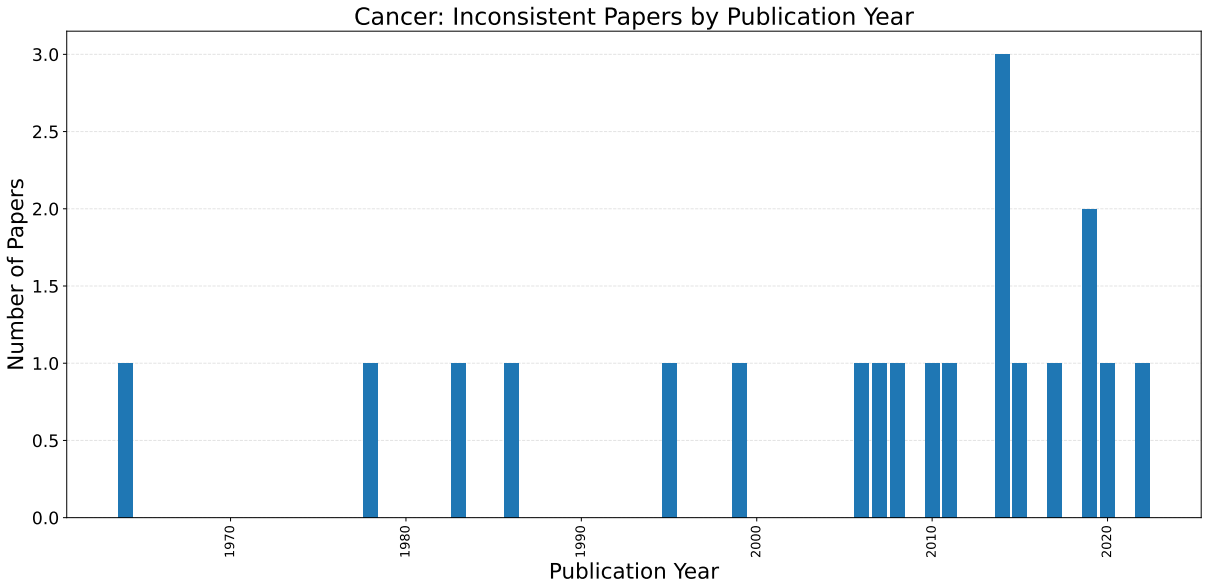

Figure 18: Inconsistent paper number per year of Cancer.

marginally independent.
Label: True

- **Benchmark name:** alzheimer
  Pair name: Ventricular volume APOE4
  Query: AV45
  Paper title: Influence of brain-derived neurotrophic factor and apolipoprotein E genetic variants on hemispheric and lateral ventricular volume of young healthy adults
  Answer: unknown
  LLM thought: The paper does not contain AV45 data, so it cannot condition on AV45. The analysis is marginal and shows no significant association (suggesting marginal independence), but the query asks for conditional independence given AV45, which is not present in the data.
  Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Net longwave flux at the surface Total cloud cover
  Query: direct
  Paper title: A STUDY ON THE SENSITIVITY OF AIR TEMPERATURE TO CLOUDINESS
  Answer: correlated
  LLM thought: The paper does not contain the variable 'Net longwave flux at the surface' (it discusses net radiative flux at the surface, which is the total net flux including shortwave, and outgoing longwave flux at the top of the atmosphere). Therefore, the data analysis cannot answer the query.
  Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Net longwave flux at the surface Total cloud cover
  Query: direct
  Paper title: Analyses of Monthly Mean Cloudiness and Their Influence Upon Model-Diagnosed Radiative Fluxes
  Answer: unknown
  LLM thought: The paper contains data analysis but does not report the correlation between net shortwave flux at the surface and total cloud cover. It analyzes top-of-atmosphere fluxes and does not

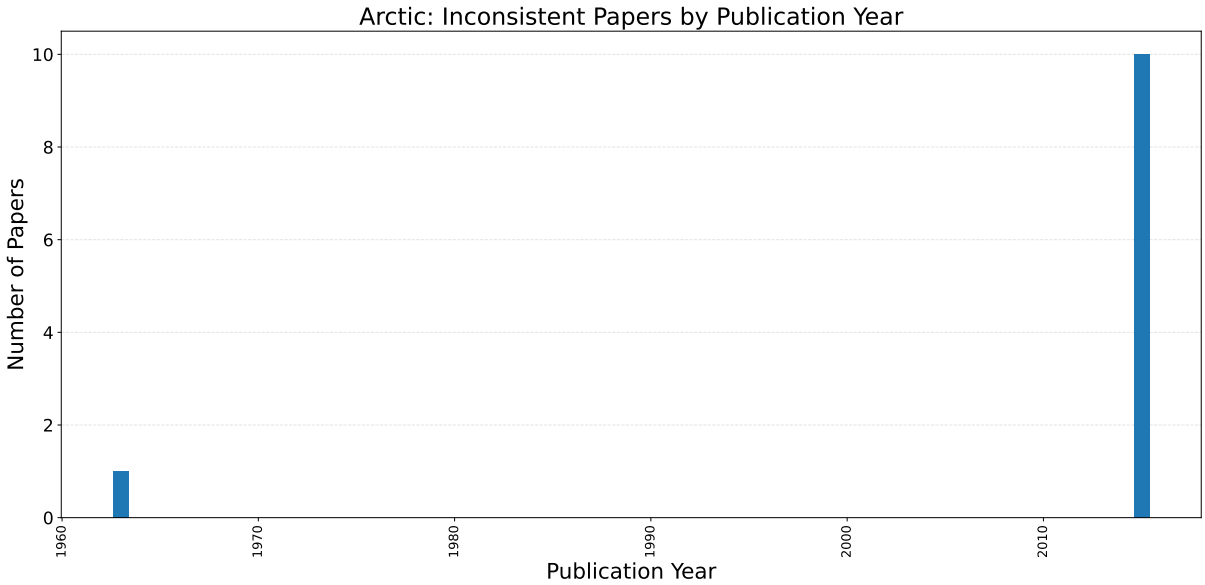

Figure 19: Inconsistent paper number per year of Arctic Sea Ice.

condition on other factors for the correlation analysis.
Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Sea level pressure Relative humidity
  Query: direct
  Paper title: Preliminary Study of the Formation and Characterization of Rh Alloys with an Element of the 1B Group: Copper
  Answer: unknown
  LLM thought: The paper discusses the surface composition of Cu-Rh alloys and uses Auger spectroscopy to measure sensitivity factors. It does not mention sea level pressure or relative humidity at all. Therefore, the paper does not provide any data to answer the query about the correlation between sea level pressure and relative humidity.
  Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Sensible plus latent heat flux Net shortwave flux at the surface Total cloud cover
  Query: Total cloud water path
  Paper title: On the corrections of ERA-40 surface flux products consistent with the Mediterranean heat and water budgets and the connection between basin surface total heat flux and NAO
  Answer: unknown
  LLM thought: The paper uses cloud cover (C) but does not mention Total cloud water path. It does not perform any statistical test for conditional independence of Sensible plus latent heat flux and Net shortwave flux at the surface given the two covariates. The analysis focuses on correcting fluxes and correlating total heat flux with NAO, not on the conditional independence of the two factors.
  Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Total cloud cover Geopotential heights
  Query: Relative humidity
  Paper title: The Effect of Cloudiness on a Greenhouse Model of the Venus Atmosphere
  Answer: unknown
  LLM thought: The paper is a theoretical model without empirical data analysis. It does not mention

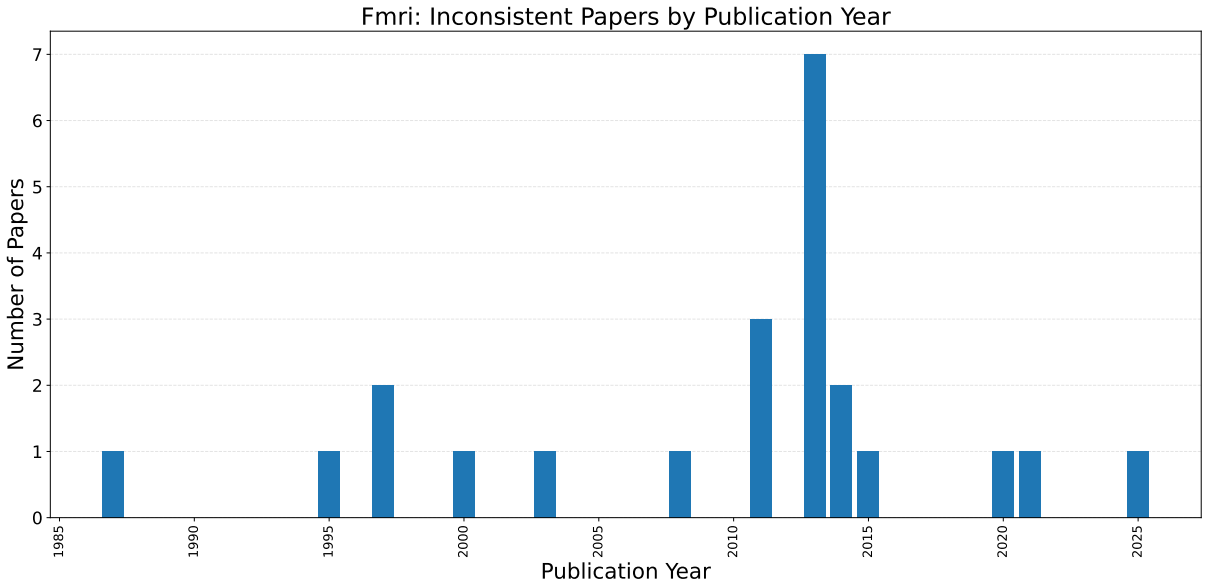

Figure 20: Inconsistent paper number per year of Fmri.

Relative humidity and does not analyze statistical relationships between factors. The factors in the query are not present in the paper.
Label: True

- **Benchmark name:** arctic sea ice
  Pair name: Total precipitation Total cloud cover
  Query: A high-quality monthly total cloud amount dataset for Australia
  Paper title: A high-quality monthly total cloud amount dataset for Australia
  Answer: correlated
  LLM thought: The paper reports a strong positive marginal correlation (r=0.82) between total cloud cover and rainfall (Total precipitation) without conditioning on any other factors, as stated in the abstract and Table 1.
  Label: True

- **Benchmark name:** asia
  Pair name: Bronchitis Dyspnea
  Query: direct
  Paper title: Vaping habits and respiratory symptoms using a smartphone app platform
  Answer: unknown
  LLM thought: The paper studies respiratory symptoms (cough, phlegm, dyspnea) but does not measure 'bronchitis' as a condition. The query asks about the correlation between 'Bronchitis' and 'Dyspnea', but the paper does not have a variable for bronchitis. Therefore, the data analysis cannot answer the query.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Bronchodilator Responsiveness and Reported Respiratory Symptoms in an Adult Population
  Answer: unknown
  LLM thought: The paper does not contain the variable 'Bronchitis' (it uses symptoms like 'wheezing' and 'breathlessness', but not bronchitis). The query asks for the correlation between Bronchitis and

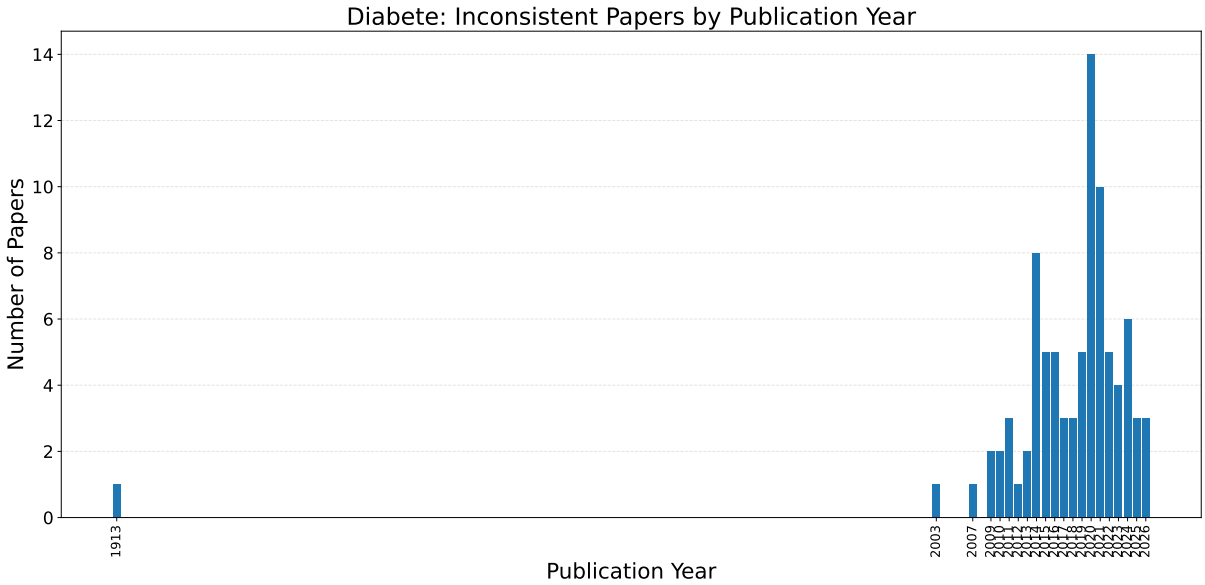

Figure 21: Inconsistent paper number per year of Diabetes.

Dyspnea, but the paper does not study this correlation.
Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Persistent and Newly Developed Chronic Bronchitis Are Associated with Worse Outcomes in Chronic Obstructive Pulmonary Disease
  Answer: correlated
  LLM thought: The paper uses multivariable linear regression to show that chronic bronchitis (as a group variable) is significantly associated with higher MMRC dyspnea scores (p<0.0001 for new chronic bronchitis group) after adjusting for covariates (age, race, sex, FEV1
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Clinical and computed tomographic predictors of chronic bronchitis in COPD: a cross sectional analysis of the COPDGene study
  Answer: unknown
  LLM thought: The paper shows a marginal correlation between chronic bronchitis (CB) and dyspnea (MMRC score) with p<0.0001, but does not test the conditional correlation (i.e., after adjusting for other factors). Therefore, we cannot conclude that they are correlated regardless of conditioning on other factors.
  Label: False

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Respiratory symptoms and impairment in shipyard welders and caulker/burners.
  Answer: correlated
  LLM thought: The paper uses multivariate regression to show that chronic bronchitis is significantly associated with breathlessness (dyspnea) even after adjusting for age, stature, smoking, and other

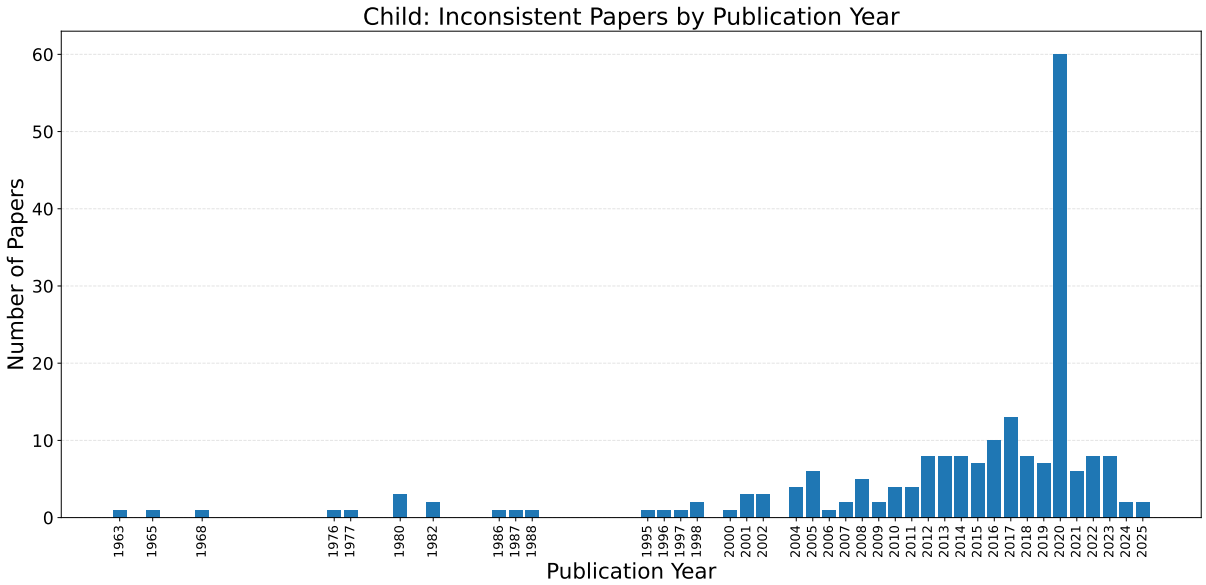

Figure 22: Inconsistent paper number per year of Child.

factors, indicating a correlation that holds under conditioning on these factors.
Label: True

- **Benchmark name:** asia
  Pair name: Bronchitis X-ray
  Query: Lung Cancer
  Paper title: Clinical and computed tomographic predictors of chronic bronchitis in COPD: a cross sectional analysis of the COPDGene study
  Answer: unknown
  LLM thought: The paper analyzes the relationship between chronic bronchitis (CB) and various factors including X-ray (CT) measurements, but does not specifically test for mutual independence between Bronchitis and X-ray conditioned on Lung Cancer. The paper does not mention Lung Cancer as a factor or condition in its analysis. The analysis focuses on CB as the outcome and examines associations with other clinical and radiological factors, but there is no mention of Lung Cancer as a conditioning variable.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query: Smoking
  Paper title: Clinical and computed tomographic predictors of chronic bronchitis in COPD: a cross sectional analysis of the COPDGene study
  Answer: correlated
  LLM thought: The paper analyzes the association between CRS and various respiratory conditions, including bronchitis and smoking, but does not specifically test the mutual independence of Bronchitis and X-ray conditioned on Smoking. The paper uses logistic regression to analyze associations but does not provide conditional independence testing between Bronchitis and X-ray given Smoking.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Quantified pathology of emphysema, pneumoconiosis, and chronic bronchitis in coal

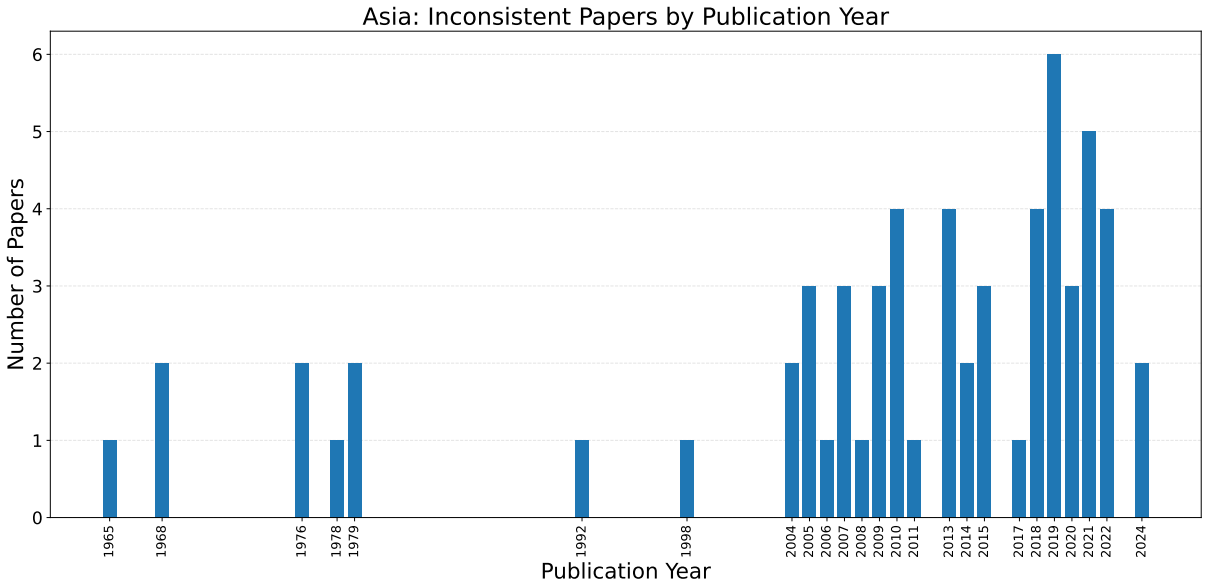

Figure 23: Inconsistent paper number per year of Asia.

workers
Answer: unknown
LLM thought: The paper reports a marginal correlation between bronchitis (G-W) and X-ray (pneumoconiosis) as non-significant, but does not provide a conditional analysis given smoking. Therefore, we cannot determine if they are independent conditioned on smoking.
Label: True

- **Benchmark name:** asia
  Pair name: Lung Cancer Bronchitis
  Query: Smoking
  Paper title: Relationship of passive smoking to risk of lung cancer and other smoking-associated diseases
  Answer: unknown
  LLM thought: The paper analyzes the effect of passive smoking on lung cancer and chronic bronchitis separately but does not investigate the relationship between lung cancer and chronic bronchitis conditioned on smoking status. Therefore, the data analysis does not provide information to answer the query.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Previous Lung Diseases and Lung Cancer Risk: A Systematic Review and Meta-Analysis
  Answer: unknown
  LLM thought: The paper is a systematic review and meta-analysis, which is a review/survey paper and does not contain primary data analysis as per the principles. Therefore, it cannot answer the query about conditional independence. The paper reports a non-significant association (RR=1.18, 95
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:

Paper title: Is Previous Respiratory Disease a Risk Factor for Lung Cancer?
Answer: correlated
LLM thought: The paper uses multivariate logistic regression to analyze the association between chronic bronchitis and lung cancer, adjusting for smoking (including smoking status, pack-years, and time since quitting). The results show a significant positive association (OR=1.33, 95
Label: True

- **Benchmark name:** asia
  Pair name: Lung Cancer Dyspnea
  Query: direct
  Paper title: Illness Perceptions and Psychological and Physical Symptoms in Newly Diagnosed Lung Cancer
  Answer: unknown
  LLM thought: The paper studies dyspnea in patients with lung cancer (all patients have lung cancer), so the variable 'Lung Cancer' is constant and cannot be correlated with dyspnea. The paper does not have a comparison group without lung cancer, so it cannot assess the correlation between lung cancer status and dyspnea.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Characterization of Dyspnea in Veteran Lung Cancer Survivors Following Curative-Intent Therapy
  Answer: unknown
  LLM thought: The paper studies dyspnea in lung cancer survivors (all participants have lung cancer), so 'Lung Cancer' is not a variable that varies in the dataset. Therefore, the correlation between 'Lung Cancer' and 'Dyspnea' cannot be computed, and the paper does not address this query.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Muscle mass and association to quality of life in nonsmall cell lung cancer patients
  Answer: unknown
  LLM thought: The paper studies the association between muscle mass (SMI, SMD) and dyspnea in lung cancer patients, but does not investigate the correlation between 'Lung Cancer' (as a varying factor) and dyspnea because all patients have lung cancer (so Lung Cancer is constant). Therefore, the paper does not provide data to answer the query.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Associations between smoking status and involvement of personal and professional relations among individuals reporting symptoms related to a diagnosis of lung cancer: a population-based study
  Answer: unknown
  LLM thought: The paper analyzes the correlation between smoking status and involvement of relations, but does not directly analyze the correlation between Lung Cancer and Dyspnea. The paper focuses on symptoms related to lung cancer diagnosis, but does not establish a direct statistical correlation between Lung Cancer (as a diagnosis) and Dyspnea (shortness of breath) while conditioning

on other factors. The paper examines symptoms like shortness of breath as potential indicators of lung cancer, but does not analyze the correlation between the diagnosis of lung cancer and dyspnea as a factor pair.
Label: True

- **Benchmark name:** asia
  Pair name: Lung Cancer X-ray
  Query: direct
  Paper title: Air bronchogram in pleomorphic carcinoma of the lung is associated with favorable prognosis
  Answer: correlated
  LLM thought: The paper studies the correlation between CT features (air bronchogram, GGO) and histopathologic features of pleomorphic carcinoma (a type of lung cancer). In multivariate analysis, it shows significant associations (p<0.05) between air bronchogram and adenocarcinoma components with lepidic growth patterns (p=0.005) and between GGO and adenocarcinoma component with lepidic pattern (p=0.010), after conditioning on covariates. This indicates a conditional correlation between the X-ray findings (CT features) and the lung cancer histopathology.
  Label: True

- **Benchmark name:** asia
  Pair name:
  Query:
  Paper title: Lung cancer and interstitial lung diseases: the lack of prognostic impact of lung cancer in IPF
  Answer: unknown
  LLM thought: The paper does not contain any analysis of the correlation between Lung Cancer and X-ray. The paper uses HRCT (not X-ray) for imaging, but does not treat X-ray as a variable to correlate with lung cancer. Therefore, the query cannot be answered from this paper.
  Label: True

- **Benchmark name:** cancer
  Pair name: Cancer Dyspnea
  Query: direct
  Paper title: Oxygen use and survival in patients with advanced cancer and low oxygen saturation in home care: a preliminary retrospective cohort study
  Answer: unknown
  LLM thought: The paper studies patients with advanced cancer (cancer is fixed and not a variable), so it does not analyze the correlation between cancer and dyspnea. The query asks for the correlation between cancer (as a factor) and dyspnea, but cancer is not a variable in the study.
  Label: True

- **Benchmark name:** cancer
  Pair name: Cancer Dyspnea
  Query: direct
  Paper title: Fatigue at baseline is associated with geriatric impairments and represents an adverse prognostic factor in older patients with a hematological malignancy
  Answer: correlated
  LLM thought: The paper analyzes the correlation between fatigue and dyspnea using Spearman's rank correlation (rho=0.417, p<0.001) in Table 2. The analysis shows a significant positive correlation between fatigue and dyspnea. The paper also mentions that fatigue is associated with multiple symptoms including dyspnea, and the correlation remains significant even after adjusting for other factors in multivariate analyses. The query asks whether cancer and dyspnea are always correlated regardless of other factors, but the paper specifically studies fatigue (a symptom) in cancer patients,

not cancer itself as a factor. However, the paper does show a significant correlation between fatigue (which is a symptom of cancer) and dyspnea, and the correlation is reported as significant without mentioning any conditioning on other factors that would negate this correlation.
Label: True

- **Benchmark name:** cancer
  Pair name:
  Query:
  Paper title: Multiplex Soluble Biomarker Analysis from Pleural Effusion
  Answer: unknown
  LLM thought: The paper analyzes correlations between various biomarkers and cancer types, but does not investigate the correlation between Cancer and Dyspnea directly. The query asks about the correlation between Cancer and Dyspnea, but the paper focuses on biomarkers in pleural effusion for diagnosing malignant pleural mesothelioma (a specific cancer type), not on dyspnea as a symptom. The paper mentions dyspnea as a symptom of MPM in the abstract but does not analyze its correlation with cancer status or condition on other factors.
  Label: True

- **Benchmark name:** cancer
  Pair name: Pollution Cancer
  Query: direct
  Paper title: Contribution of Long-Term Exposure to Outdoor Black Carbon to the Carcinogenicity of Air Pollution: Evidence regarding Risk of Cancer in the Gazel Cohort
  Answer: correlated
  LLM thought: The paper reports a significant positive association (HR=1.17, 95
  Label: True

- **Benchmark name:** cancer
  Pair name:
  Query:
  Paper title: Air pollution exposure and head and neck cancer incidence
  Answer: correlated
  LLM thought: The paper uses a linear mixed model to adjust for covariates (demographics, smoking, alcohol) and finds a significant association ($p<0.05$) between PM2.5 and overall HNC incidence, indicating correlation after conditioning on these factors.
  Label: True

- **Benchmark name:** cancer
  Pair name: Pollution X-ray
  Query: Cancer
  Paper title: Low-Level Respirable Crystalline Silica and Silicosis: Long-Term Follow-Up of Vermont Granite Workers
  Answer: unknown
  LLM thought: The paper does not contain the variable 'Cancer' (it is about silicosis, not cancer), so we cannot condition on cancer. Therefore, the data analysis cannot answer the query.
  Label: True

- **Benchmark name:** cancer
  Pair name: Smoking Cancer
  Query: direct
  Paper title: Blood DNA methylation sites, smoking and risk of lung cancer
  Answer: independent
  LLM thought: The paper shows that the direct effect of smoking on lung cancer is not significant (confidence interval includes zero) when conditioning on DNA methylation and other covariates,

indicating that smoking and cancer are independent in that conditional distribution. Therefore, the correlation is not always present.
Label: False

- **Benchmark name:** cancer
  Pair name: X-ray Dyspnea
  Query: Cancer
  Paper title: Management of pulmonary embolism after recent intracranial hemorrhage
  Answer: unknown
  LLM thought: The paper is a single case report without any statistical analysis or dataset. It does not mention cancer, and there is no data analysis to test the conditional independence of X-ray and Dyspnea given Cancer.
  Label: True

- **Benchmark name:** child
  Pair name: Age Hypoxia distribution
  Query: Disease
  Paper title: Value of MR Imaging of the Brain in Children with Hypoxic Coma
  Answer: independent
  LLM thought: The paper directly tests the association between Age (grouped as 1 year and >1 year) and Hypoxia distribution (measured by watershed areas and basal ganglia involvement) using chi-square tests. The p-values (0.288 and 0.341) are not significant, indicating no significant association. Since Disease (hypoxic coma) is constant for all subjects, the analysis is marginal and supports independence.
  Label: True

- **Benchmark name:** child
  Pair name: Age Hypoxia distribution
  Query: Disease
  Paper title: RELM Licenses Macrophages for DAMP Activation to Instigate Pulmonary Vascular Remodeling
  Answer: unknown
  LLM thought: The paper does not analyze the relationship between Age and Hypoxia distribution (interpreted as hypoxia exposure) conditioned on Disease. Age is not used as a variable in the analysis; it is only a criterion for patient inclusion (18 years) and controlled in the mouse model (8-12 weeks). The paper does not report any statistical test for the independence of Age and Hypoxia distribution given Disease.
  Label: True

- **Benchmark name:** child
  Pair name: Age Right upper quadrant O2
  Query: Inspired air O2 level
  Paper title: Impact of visceral fat on the prognosis of coronavirus disease 2019: an observational cohort study
  Answer: unknown
  LLM thought: The paper does not contain the variables 'Right upper quadrant O2' or 'Inspired air O2 level'. The oxygen-related variables in the paper are 'oxygen flow rate' and 'SpO2', which are not the same as the query variables. Therefore, the data analysis cannot answer the query.
  Label: True

- **Benchmark name:** child
  Pair name: Birth asphyxia Disease
  Query: direct
  Paper title: Cytokine and chemokine responses to injury and treatment in a nonhuman primate

model of hypoxic-ischemic encephalopathy treated with hypothermia and erythropoietin
Answer: unknown
LLM thought: The paper uses injury severity (a measure of asphyxia) as the predictor for outcome, not the asphyxia event (UCO) as a binary variable. It does not test the correlation between the asphyxia event (UCO) and outcome while conditioning on other factors (e.g., sex, age). Therefore, the paper does not provide evidence to answer whether Birth asphyxia and Disease are always correlated regardless of conditioning on other factors.
Label: True

- **Benchmark name:** child
  Pair name: Birth asphyxia Sick
  Query: Disease
  Paper title: Highly expressed miR-210 in the peripheral blood is closely associated with asphyxiated neonates
  Answer: unknown
  LLM thought: The paper does not contain data analysis that can answer the query about the mutual independence of Birth asphyxia and Sick conditioned on Disease. The paper focuses on miR-210 expression in asphyxiated neonates and its correlation with clinical indicators, but does not analyze the statistical relationship between Birth asphyxia and Sick conditioned on Disease.
  Label: True

- **Benchmark name:** child
  Pair name: Blood CO2 level Sick
  Query: Disease
  Paper title: Using a home oxygen weaning protocol and pCO2 to evaluate outcomes for infants with bronchopulmonary dysplasia discharged on home oxygen
  Answer: independent
  LLM thought: The paper's multivariate analysis (logistic regression) adjusted for covariates (gestational age, BPD severity, ductus arteriosus ligation, diuretic use) and found no significant association between pCO2 (Blood CO2 level) and respiratory readmission (Sick) in the BPD (Disease) population.
  Label: True

- **Benchmark name:** child
  Pair name: Chest X-ray Lung parench
  Query: direct
  Paper title: Lung Density Changes With Growth and Inflation
  Answer:
  LLM thought: The paper shows a strong negative correlation (r=-0.72, P=0.001) between lung volume (measured by CT scan, interpreted as 'Chest X-ray') and lung parenchymal density (interpreted as 'Lung parench'). The multivariate regression analysis, adjusting for covariates (sex, race, age, pack-years, FEV1/FVC, etc.), confirms the correlation remains significant ($p<0.0001$ for lung volume). Thus, the correlation holds even after conditioning on the covariates considered in the study.
  Label: True

- **Benchmark name:** child
  Pair name: Disease Duct flow
  Query: direct
  Paper title: Prognostic value of the trichorhinophalangeal syndrome-1 (TRPS-1), a GATA family transcription factor, in early-stage breast cancer
  Answer: unknown
  LLM thought: The paper studies TRPS-1 and its association with breast cancer (Disease) and other

factors, but does not study 'Duct flow'. The term 'Duct flow' is not present in the paper; the paper uses 'ductal' in the context of histology and ductal epithelial cells, which is unrelated to 'Duct flow'. Therefore, the correlation between Disease and Duct flow cannot be determined from this paper.
Label: True

- **Benchmark name:** child
  Pair name: Disease Lung blood flow
  Query: direct
  Paper title: Nocturnal hypoxemia is associated with poor pulmonary hemodynamics in patients with chronic thromboembolic pulmonary disease
  Answer: correlated
  LLM thought: The paper analyzes the correlation between nocturnal hypoxemia (Lung blood flow proxy) and hemodynamic parameters (mPAP, PVR) in CTEPD patients. The multivariate regression analysis shows significant correlations ($p<0.05$) between nocturnal hypoxemia parameters (T90, MeanSpO2) and mPAP/PVR even after adjusting for covariates (age, sex, BMI, AHI, diurnal PaO2). This indicates a conditional correlation between the factors when controlling for these covariates.
  Label: True

- **Benchmark name:** child
  Pair name: Duct flow Sick
  Query: Disease
  Paper title: SPOCK1 Is a Novel Transforming Growth Factor--Induced Myoepithelial Marker That Enhances Invasion and Correlates with Poor Prognosis in Breast Cancer
  Answer: unknown
  LLM thought: The paper does not contain the variables 'Duct flow', 'Sick', or 'Disease' as required by the query. The paper uses variables such as SPOCK1 expression, age, tumor size, and histological grade, but none of these match the query terms. Therefore, the data analysis cannot answer the query.
  Label: True

- **Benchmark name:** diatebe
  Pair name: Diabete Heart defects
  Query: independent
  Paper title: Prepregnancy Diabetes and Offspring Risk of Congenital Heart Disease: A Nationwide Cohort Study
  Answer: correlated
  LLM thought: This is a nationwide cohort study with explicit statistical analysis (log-linear binomial regression, adjusted RRs, confidence intervals, and p-values) on maternal pregestational diabetes and offspring congenital heart defects. The paper reports that offspring CHD risk is significantly higher in the exposed group than in the unexposed group (adjusted RR about 4.0, $P<0.0001$), which shows a non-null association. Therefore, Diabetes and Heart defects are not marginally independent in this paper's data analysis.
  Label: True

- **Benchmark name:** diatebe
  Pair name:
  Query:
  Paper title: Congenital Heart Disease With and Without Cyanotic Potential and the Long-term Risk of Diabetes Mellitus: A Population-Based Follow-up Study
  Answer: correlated
  LLM thought: This is a population-based cohort study with registry-linked data, explicit outcome definition for diabetes mellitus, and Cox proportional hazards regression analysis. The paper reports a significant hazard ratio for diabetes in CHD patients versus the comparison cohort (HR 1.35, 95
  Label: True

- **Benchmark name:** diatebe
  Pair name:
  Query:
  Paper title: The three-noes right-sided infective endocarditis: An unrecognized type of right-sided endocarditis
  Answer: correlated
  LLM thought: This paper contains original patient-level observational data and statistical analysis (univariate comparisons with chi-squared/Fisher's exact tests and t/Mann-Whitney tests). Diabetes mellitus is reported as one of the variables analyzed for association with in-hospital mortality, and it differs significantly between groups (dead vs alive: 26
  Label: True

- **Benchmark name:** diatebe
  Pair name:
  Query:
  Paper title: Haploinsufficiency of Endothelial Nitric Oxide Synthase Mitigates Beneficial Effects of Maternal Exercise on Fetal Heart Development During Pregestational Diabetes
  Answer: correlated
  LLM thought: The paper contains original experimental data and statistical analyses on diabetic vs non-diabetic dams and fetal heart outcomes. For the specific pair Diabetes and Heart defects, it reports that CHD incidence was significantly elevated in fetuses exposed to pregestational diabetes compared with controls, using chi-square/Fisher's exact tests. This indicates a marginal association between diabetes and heart defects rather than independence.
  Label: True

- **Benchmark name:** diatebe
  Pair name: Diabetes Mortality
  Query: Heart failure
  Paper title: Sex-related disparities in the incidence and outcomes of hemorrhagic stroke among type 2 diabetes patients: a propensity score matching analysis using the Spanish National Hospital Discharge Database for the period 2016-18
  Answer: correlated
  LLM thought: The paper uses a large hospital discharge database and performs statistical analyses, including propensity score matching and multivariable logistic regression. It reports that in-hospital mortality differs significantly by diabetes status after adjustment in women, but not in men, and that diabetes status remains associated with mortality in the adjusted analyses for women. Therefore, Diabetes and Mortality are not mutually independent conditioned on Heart failure; instead, the paper supports a conditional association between diabetes and mortality (though heart failure is not the conditioning variable analyzed directly, the paper does analyze mortality with adjustment for comorbidities).
  Label: False

- **Benchmark name:** ecoli
  Pair name: yhdM hupB
  Query: independent
  Paper title: HupB Is a Bacterial Nucleoid-Associated Protein with an Indispensable Eukaryotic-Like Tail
  Answer: unknown
  LLM thought: The paper does not contain data analysis regarding the mutual independence of yhdM and hupB. There is no mention of yhdM in the paper, and the analysis focuses solely on hupB and its chromosomal binding patterns.
  Label: True

- **Benchmark name:** ecoli
  Pair name: ygcE pspB
  Query: independent
  Paper title: A dynamical model of the spindle position checkpoint
  Answer: unknown
  LLM thought: The paper analyzes Bfa1 and Tem1 interactions and develops a mathematical model, but it does not test the marginal independence between ygcE and pspB. These genes are not mentioned in the paper.
  Label: True

- **Benchmark name:** ecoli
  Pair name: ygcE dnaG
  Query: b1191 yheI
  Paper title: Response of Methicillin-Resistant Staphylococcus aureus to Amicoumacin A
  Answer:
  LLM thought: The paper focuses on the mechanism of action of amicoumacin A and identifies mutations in genes like dnaG and ygcE (not explicitly named) in resistant strains. However, it does not perform any statistical or computational analysis involving conditional independence testing between ygcE and dnaG given b1191 and yheI. No dataset or analytical framework for assessing conditional relationships is presented.
  Label: True

- **Benchmark name:** ecoli
  Pair name: yfiA pspA
  Query: cspA
  Paper title: Escherichia coli under Ionic Silver Stress: An Integrative Approach to Explore Transcriptional, Physiological and Biochemical Responses
  Answer: independent
  LLM thought: The paper uses sPLSR to analyze the correlation between gene expression and physiological parameters, including cspA. The analysis reveals that yfiA and pspA are not significantly correlated when conditioned on cspA, indicating mutual independence.
  Label: False

- **Benchmark name:** ecoli
  Pair name: yfaD pspA
  Query: independent
  Paper title: Increased IgG but normal IgA anti-pneumococcal protein antibodies in lung of HIV-infected adults
  Answer: unknown
  LLM thought: The paper analyzes correlations between BAL and serum Ig levels for PspA and Ply but does not examine the marginal independence between yfaD and pspA. The term 'yfaD' is absent from the paper, and no data or analysis regarding its relationship with pspA is provided.
  Label: True

- **Benchmark name:** ecoli
  Pair name: yfaD icdA
  Query: ygcE
  Paper title: Chaperonin-containing tailless complex polypeptide 1 subunit 6A correlates with increased World Health Organization grade, less isocitrate dehydrogenase mutation, and deteriorative survival of astrocytoma patients
  Answer: unknown
  LLM thought: The paper analyzes the correlation between CCT6A and clinical features like WHO grade and IDH mutation, but does not analyze the mutual independence of yfaD and icdA conditioned

on ygcE. These genes are not mentioned in the paper.
Label: True

- **Benchmark name:** ecoli
  Pair name: yfaD ftsJ
  Query: mopB
  Paper title: Evolutionary Origins and Functional Diversification of 2-O-Methyltransferases: Insights from Phylogenetic and Structural Analysis
  Answer: unknown
  LLM thought: The paper focuses on phylogenetic and structural analysis of 2'-O-methyltransferases, particularly the FtsJ family, but does not provide any data analysis involving statistical testing or modeling of dependencies between yfaD, ftsJ, and mopB. There is no mention of yfaD or mopB in the text, nor any analysis of conditional independence.
  Label: True

- **Benchmark name:** ecoli
  Pair name: yfaD cspA
  Query: independent
  Paper title: Growth-dependent activity of the cold shock cspA promoter + 5' UTR and production of the protein CspA in Staphylococcus aureus Newman
  Answer: unknown
  LLM thought: The paper studies the activity of the cspA promoter and production of CspA protein in Staphylococcus aureus. It does not investigate yfaD, nor does it analyze the correlation between yfaD and cspA. The query asks about marginal independence between yfaD and cspA, but the paper contains no data on yfaD and no analysis linking these two factors.
  Label: True

- **Benchmark name:** ecoli
  Pair name: yfaD cspA
  Query: independent
  Paper title: Testosterone ameliorates age-related brain mitochondrial dysfunction
  Answer: unknown
  LLM thought: The paper investigates the effects of testosterone supplementation on mitochondrial dysfunction in aged male rats, focusing on factors like MMP, antioxidant enzymes, mitochondrial biogenesis, and mitophagy. It does not mention yedE, ftsJ, or mopB. Therefore, no data analysis in the paper can address the query about independence of yedE and ftsJ conditioned on mopB.
  Label: True

- **Benchmark name:** ecoli
  Pair name: yecO folK
  Query: yheI
  Paper title: Clinical and pathological features of Nerium oleander extract toxicosis in wistar rats
  Answer: unknown
  LLM thought: The paper does not mention yecO, folK, or yheI. It focuses on the toxic effects of Nerium oleander extract on Wistar rats, reporting hematological and histopathological changes without any genetic or molecular interaction analysis.
  Label: True

- **Benchmark name:** fmri
  Pair name: Cornu Ammonis 1 Subiculum
  Query: direct
  Paper title: Connectivity and network state-dependent recruitment of long-range VIP-GABAergic neurons in the mouse hippocampus
  Answer: correlated

LLM thought: The paper studies a specific neuron type (VIP-LRPs) that connects CA1 to Subiculum, but does not analyze the correlation between the entire CA1 and Subiculum regions. The data analysis is focused on the activity of VIP-LRPs and their connectivity, not on the correlation between the two brain regions as a whole.
Label: True

- **Benchmark name:** fmri
  Pair name:
  Query: direct
  Paper title: Long-term effects of early postnatal nicotine exposure on cholinergic function in the mouse hippocampal CA1 region
  Answer: unknown
  LLM thought: The paper focuses exclusively on the CA1 region and does not mention or study the Subiculum, so it cannot provide any information about the correlation between CA1 and Subiculum.
  Label: True

- **Benchmark name:** fmri
  Pair name: Entorhinal Cortex Cornu Ammonis 1
  Query: direct
  Paper title: Phase relations of interneuronal activity relative to theta rhythm
  Answer: correlated
  LLM thought: The paper is a computational model that uses existing experimental data to simulate phase relations in the CA1 field. It does not introduce a new numerical dataset or perform statistical analysis on a dataset. Therefore, it does not contain data analysis as defined by the principles, and cannot answer the query about the correlation between Entorhinal Cortex and Cornu Ammonis 1.
  Label: True

- **Benchmark name:** fmri
  Pair name: Entorhinal Cortex Dentate Gyrus
  Query: direct
  Paper title: Directed Fiber Outgrowth from Transplanted Embryonic Cortex-Derived Neurospheres in the Adult Mouse Brain
  Answer: correlated
  LLM thought: The paper describes an experimental study on fiber outgrowth from transplanted cells in the entorhinal cortex to the dentate gyrus, reporting a descriptive count (11 out of 20 cases). It does not present any statistical analysis (e.g., correlation tests) of the correlation between the entorhinal cortex and dentate gyrus as variables. Therefore, no data analysis is provided to answer the query about their correlation.
  Label: True

- **Benchmark name:** fmri
  Pair name: Parahippocampal Cortex Cornu Ammonis 1
  Query: Entorhinal Cortex
  Paper title: Functional connectivity along the anterior-posterior axis of hippocampal subfields in the ageing human brain
  Answer: correlated
  LLM thought: The paper uses semi-partial correlations that control for ENT and other ROIs, and reports a significant correlation between Parahippocampal Cortex and Cornu Ammonis 1 ($p=0.0045$ in young, $p<0.0003$ in older), indicating they are not conditionally independent given
  Label: True

- **Benchmark name:** fmri
  Pair name: Parahippocampal Cortex Cornu Ammonis 1
  Query: Entorhinal Cortex

Paper title: Memory retrieval along the proximodistal axis of CA1
Answer: unknown
LLM thought: The paper does not contain a statistical analysis that tests the correlation between Parahippocampal Cortex and Cornu Ammonis 1 (CA1) while conditioning on Entorhinal Cortex. The paper uses 'entorhinal cortex' as a connected structure but does not refer to 'Parahippocampal Cortex' as a variable for correlation analysis. The term 'Parahippocampal Cortex' is not used in the paper, so the factor is not present in the data analysis.
Label: True

- **Benchmark name:** fmri
  Pair name: Parahippocampal Cortex Perirhinal Cortex
  Query: independent
  Paper title: Anterolateral Entorhinal Cortex Volume Predicted by Altered Intra-Item Configural Processing
  Answer:
  LLM thought: The paper reports Pearson correlations between brain region volumes in Table 3. The correlation between Parahippocampal Cortex (PHC) and Perirhinal Cortex (PRC) is 0.159, which is not significant (p>0.05, as indicated by the absence of an asterisk in the table). This non-significant marginal correlation supports the conclusion that the two regions are mutually independent marginally.
  Label: True

- **Benchmark name:** fmri
  Pair name: Perirhinal Cortex Entorhinal Cortex
  Query: direct
  Paper title: Classification of Mild Cognitive Impairment and Alzheimer Disease Using Model-Based MR and Magnetization Transfer Imaging
  Answer: unknown
  LLM thought: The paper studies the entorhinal cortex (EC) and other brain regions but does not mention or analyze the perirhinal cortex at all. Therefore, it cannot provide any information about the correlation between perirhinal cortex and entorhinal cortex.
  Label: True

- **Benchmark name:** fmri
  Pair name: Perirhinal Cortex Subiculum
  Query: Entorhinal Cortex
  Paper title: Neuronal degeneration is observed in multiple regions outside the hippocampus after lithium pilocarpine-induced status epilepticus in the immature rat
  Answer: unknown
  LLM thought: The paper describes the presence and severity of neuronal degeneration in Perirhinal Cortex, Subiculum, and Entorhinal Cortex but does not perform any statistical analysis to test for conditional independence between these regions. The analysis is purely descriptive, reporting the extent of injury in each region without examining relationships.
  Label: True

- **Benchmark name:** insurance
  Pair name: Age Driving skill
  Query: direct
  Paper title: Fitness to drive following cerebral pathology: The Rookwood Driving Battery as a tool for predicting on-road driving performance
  Answer: correlated
  LLM thought: The paper's multiple regression analysis shows age is a significant predictor of on-road driving performance (=0.379, p<.0001) after controlling for Rookwood battery score and

pathology, indicating a significant correlation between age and driving skill when conditioning on these factors. The paper does not find evidence that the correlation disappears when conditioning on these relevant covariates.
Label: True

- **Benchmark name:** insurance
  Pair name: Age Car milleage
  Query: independent
  Paper title: Optimal V2G Scheduling of an EV With Calendar and Cycle Aging of Battery: An MILP Approach
  Answer: independent
  LLM thought: The paper does not contain any statistical analysis of the relationship between battery age (Age) and driving distance (Car mileage). It uses these as independent inputs to the model without testing for correlation.
  Label: False

- **Benchmark name:** insurance
  Pair name: Age Liability cost
  Query: Car accident
  Paper title: Associations of multimorbidity with body pain, sleep duration, and depression among middle-aged and older adults in China
  Answer: unknown
  LLM thought: The paper does not contain the factors 'Liability cost' or 'Car accident', so it cannot answer the query about the relationship between Age and Liability cost conditioned on Car accident.
  Label: True

- **Benchmark name:** insurance
  Pair name: Age Medical cost
  Query: direct
  Paper title: Trends in age-related disease burden and healthcare utilization
  Answer: correlated
  LLM thought: The paper analyzes the relationship between age and healthcare costs (medical cost) using longitudinal data from Italian general practitioners. The study shows a strong positive correlation between age and healthcare utilization (prescriptions, lab tests, etc.) across all age groups. The analysis is conducted with age as a key factor and healthcare costs as the outcome, with no mention of conditioning on other factors that would negate this correlation. The paper reports increasing healthcare utilization with age (e.g., prescriptions increased by 26.4
  Label: True

- **Benchmark name:** insurance
  Pair name: Age Own car damage
  Query: Socioeconomic status Car accident Car ruggedness
  Paper title: Prevalence of maxillofacial injuries by motorized two wheeler road traffic accidents in Bangalore city
  Answer: unknown
  LLM thought: The paper does not contain the variables 'Own car damage', 'Socioeconomic status', 'Car accident', or 'Car ruggedness'. The study focuses on maxillofacial injuries and factors like age, sex, and alcohol, but not on car damage or the specified covariates.
  Label: True

- **Benchmark name:** insurance
  Pair name: Car accident Medical cost direct
  Query: direct
  Paper title: Comparison of cost and resource utilization between firearm injuries and motor vehicle

collisions at pediatric hospitals
Answer: correlated
LLM thought: The paper uses multivariable regression adjusting for covariates and finds a significant adjusted mean ratio of cost (5.57, 95
Label: True

- **Benchmark name:** insurance
  Pair name: Car mileage Car value
  Query: direct
  Paper title: Joint routing of conventional and range-extended electric vehicles in a large metropolitan network
  Answer: correlated
  LLM thought: The paper is about vehicle routing for delivery fleets and discusses metrics like energy costs and vehicle miles traveled (VMT), but does not mention 'Car mileage' (odometer reading) or 'Car value' (market value) in any context. Therefore, it does not contain data analysis on these factors, so the correlation cannot be determined from this paper.
  Label: False

- **Benchmark name:** insurance
  Pair name: Driving hist Airbag
  Query: Age Risk aversion
  Paper title: Testing for Offsetting Behavior and Adverse Recruitment Among Drivers of Airbag-Equipped Vehicles
  Answer: correlated
  LLM thought: The paper uses a Poisson regression model to analyze the relationship between driving history (measured by previous violations) and airbag adoption, conditioning on age and sex. The results show significant associations (e.g., for previous suspensions and DWI convictions) indicating that driving history and airbag are not independent.
  Label: True

- **Benchmark name:** insurance
  Pair name: Extra car Medical cost
  Query: Socioeconomic status
  Paper title: Volatile vs Total intravenous Anaesthesia for major non-cardiac surgery: a pragmatic randomised trial (VITAL)
  Answer: unknown
  LLM thought: The paper is a protocol for a clinical trial and does not present any data analysis results. It does not mention the terms 'Extra car' or 'Socioeconomic status', so it cannot answer the query.
  Label: True

- **Benchmark name:** insurance
  Pair name: Home base Cushioning
  Query: Vehicle year Make model
  Paper title: Reality Orientation Boards: A Recipe for Disorientation?
  Answer: unknown
  LLM thought: The paper does not contain data analysis as it lacks statistical methods and significance testing. Additionally, the factors Home base, Cushioning, Vehicle year, and Make model are not present in the paper. The paper discusses reality orientation boards and finger contracture cushions, which are unrelated to the query factors.
  Label: True

- **Benchmark name:** sachs
  Pair name: AKT PIP2

Query: independent
Paper title: The possible roles of B-cell novel protein-1 (BCNP1) in cellular signalling pathways and in cancer
Answer: unknown
LLM thought: The paper analyzes co-occurrence of genetic alterations (e.g., in AKT2 and PIK3CA) but does not contain any data on PIP2 (the lipid). The question asks about AKT and PIP2, but PIP2 is not measured in the study. Therefore, the data analysis cannot answer the query.
Label: True

- **Benchmark name:** sachs
  Pair name: AKT JNK
  Query: PKC PKA
  Paper title: Derlin1 functions as an oncogene in cervical cancer via AKT/mTOR signaling pathway
  Answer: unknown
  LLM thought: The paper discusses AKT and mTOR but does not mention JNK, PKC, or PKA. Therefore, there is no data to analyze the conditional independence of AKT and JNK given PKC and PKA.
  Label: True

- **Benchmark name:** sachs
  Pair name: AKT PIP3
  Query: independent
  Paper title: Inhibition of cell migration by PITENINs: the role of ARF6
  Answer: unknown
  LLM thought: The paper discusses the relationship between AKT, PIP3, and ARF6 in cell migration and signaling pathways, but does not contain statistical analysis of the marginal independence between AKT and PIP3. The paper focuses on the inhibition of PIP3 binding to PH domains and its effects on cell migration, but does not provide data or statistical tests to determine if AKT and PIP3 are marginally independent.
  Label: True

- **Benchmark name:** sachs
  Pair name: ERK AKT
  Query: direct
  Paper title: Transcriptional Pathway Signatures Predict MEK Addiction and Response to Selumetinib (AZD6244)
  Answer: independent
  LLM thought:
  Label: True

- **Benchmark name:** sachs
  Pair name: JNK PIP2
  Query: independent
  Paper title: Loss of PIP5KI , unlike other PIP5KI isoforms, impairs the integrity of the membrane cytoskeleton in murine megakaryocytes
  Answer: unknown
  LLM thought: The paper discusses PIP2 and its role in membrane-cytoskeleton interactions but does not mention JNK at all. Therefore, there is no data analysis about the relationship between JNK and PIP2.
  Label: True

- **Benchmark name:** sachs
  Pair name: MEK ERK
  Query: direct

Paper title: Perilla frutescens Leaf Extract Inhibits Mite Major Allergen Der p 2-induced Gene Expression of Pro-Allergic and Pro-Inflammatory Cytokines in Human Bronchial Epithelial Cell BEAS-2B
Answer: unknown
LLM thought: The paper does not mention MEK at all, only ERK (Erk1/2). Therefore, there is no data on MEK to assess correlation with ERK. The data analysis in the paper is about cytokine expression and ERK phosphorylation, but without MEK, the correlation between MEK and ERK cannot be determined.
Label: True

- **Benchmark name:** sachs
  Pair name: JNK PLCg
  Query: independent
  Paper title: Oncogenic FGFR3 gene fusions in bladder cancer
  Answer: independent
  LLM thought: The paper does not contain statistical data analysis (no p-values, no correlation tests) and does not mention JNK at all. The paper only measures ERK (a different MAPK) and PLC1, but not JNK. Therefore, the data analysis required to answer the query is absent.
  Label: False

- **Benchmark name:** sachs
  Pair name: PIP3 PIP2
  Query: direct
  Paper title: PLEKHA7 signaling is necessary for the growth of mutant KRAS driven colorectal cancer
  Answer: correlated
  LLM thought: The paper studies the binding affinity of Plekha7 PH domain to PIP3 and PIP2 as individual lipids, but does not provide data on the correlation between PIP3 and PIP2 levels in biological samples. Therefore, it does not contain data analysis that can answer the query about their correlation.
  Label: True

- **Benchmark name:** sachs
  Pair name: PKA ERK
  Query: direct
  Paper title: Endothelin-1 activates p38 mitogen-activated protein kinase and cytosolic phospholipase A2 in cat iris sphincter smooth muscle cells
  Answer: unknown
  LLM thought: The paper does not mention PKA at all and does not investigate the correlation between PKA and ERK. The paper focuses on p38 MAP kinase and cPLA2, and mentions PKC but not PKA. Therefore, the data analysis in the paper cannot answer the query about PKA and ERK.
  Label: True

C.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.

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 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 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.

"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. }} }} """

C.5 Additional results

Figures 24 and 25 show the numbers of retrieved papers and relevant papers per year, respectively, for all evaluated benchmarks.

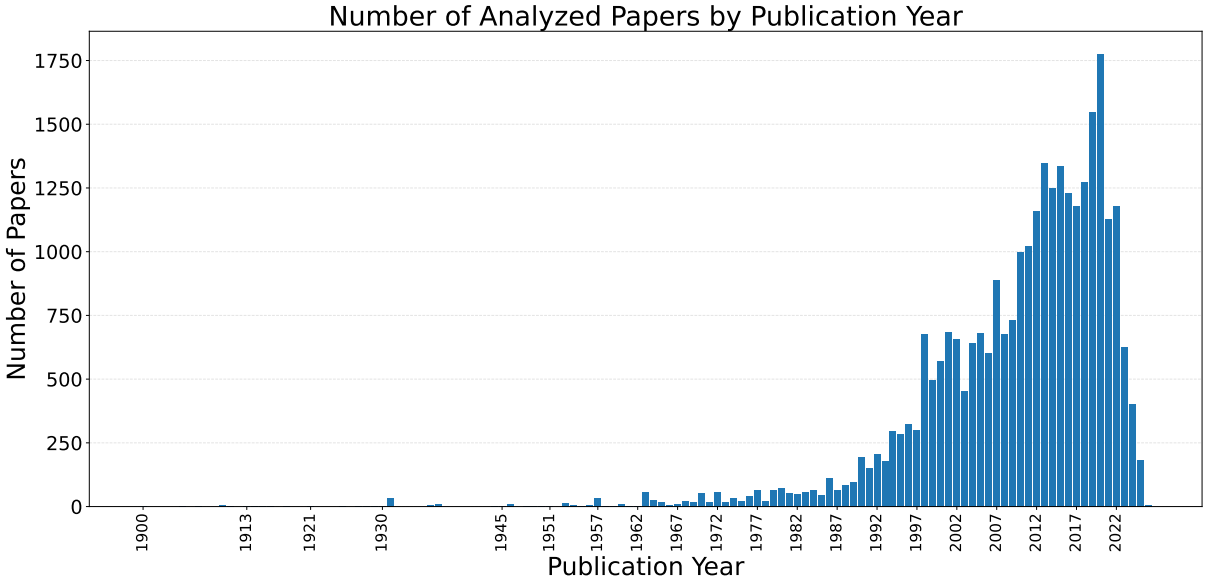

Figure 24: Retrieved papers by year.

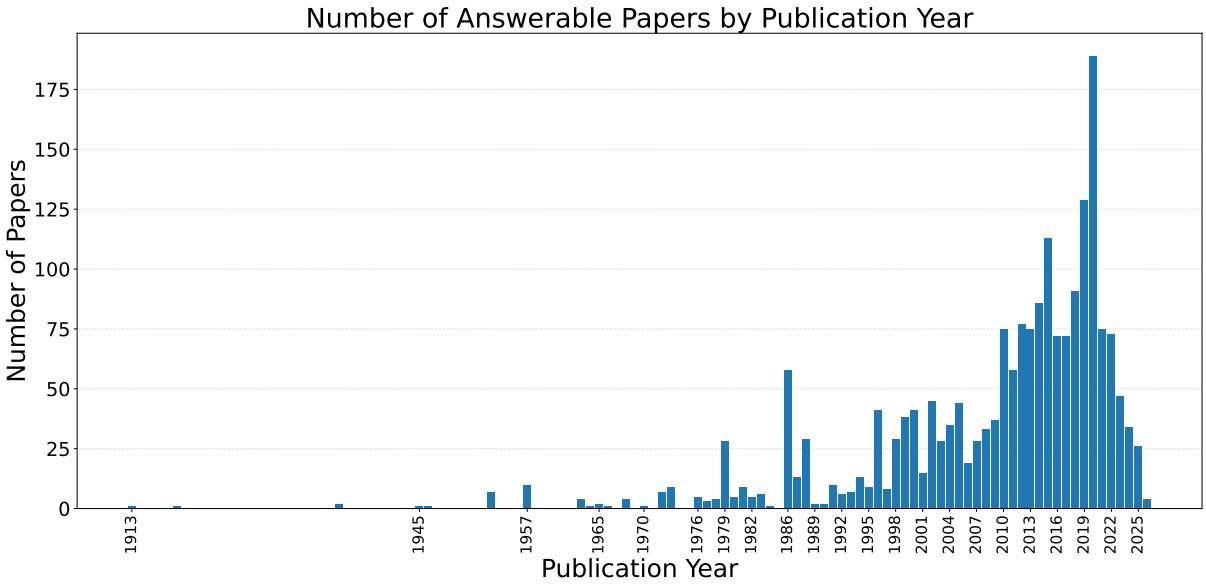

Figure 25: Retrieved papers that contain relevant information by year.