# Realizing LLMs' Causal Potential Requires Science-Grounded, Novel Benchmarks


## Abstract

Recent claims of strong performance by Large Language Models (LLMs) on causal discovery tasks are undermined by a critical flaw: many evaluations rely on benchmarks likely included in LLMs' pretraining data, raising concerns that apparent success reflects memorization rather than genuine reasoning. This risks creating a misleading narrative that LLM-only methods, which ignore observational data, outperform classical statistical approaches. We challenge this view by asking whether LLMs truly reason about causal structure, how such reasoning can be measured reliably without leakage, and whether LLMs can be trusted for causal discovery in real scientific domains. We argue that realizing their potential for accelerating scientific discovery requires two shifts: developing robust evaluation protocols based on recent, unseen scientific studies to avoid dataset leakage, and designing hybrid methods that combine LLM-derived world knowledge with statistical approaches. To this end, we outline a practical recipe for constructing causal graphs from post-training scientific publications, ensuring evaluations remain leakage-free while encompassing both established and novel causal relationships.

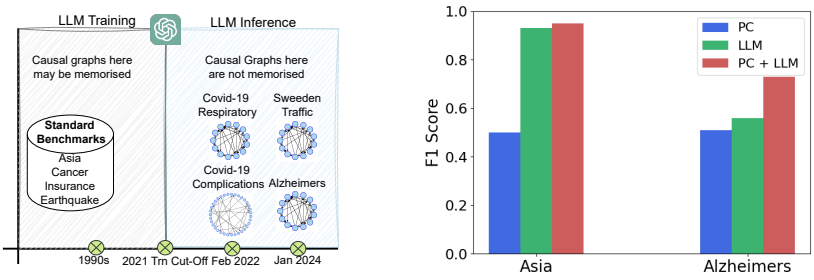

## 1 Introduction

Causal discovery, the task of learning an underlying causal graph, is a cornerstone of causal inference. It enables identification of adjustment variables for treatment effect estimation [44, 37], and reveals pathways of interventions in interventional and counterfactual analyses [36, 38]. Classical approaches rely on observational data, including constraint-based methods that use statistical tests to infer conditional independencies [48, 47, 15, 49], score-based methods that optimize a goodness-of-fit score over candidate graphs [16, 10, 35, 58], and functional-causal models exploiting assumptions such as additive noise or non-Gaussian residuals [17, 55, 45]. Yet, these methods face inherent limitations: observational data alone cannot disambiguate causal direction between dependent variables without strong assumptions or external supervision [13, 14, 19, 25].

Recent advances in Large Language Models (LLMs) have sparked interest in leveraging their encoded world knowledge for causal discovery [27, 3, 31, 52]. However, evaluations often rely on well-known benchmarks such as `BNLearn`, which likely appeared in LLM pretraining corpora, raising concerns that reported success reflects memorization rather than reasoning [53]. Indeed, Jin et al. [23, 22] show that when domain knowledge is stripped away, LLMs fail to infer causal relationships reliably. This casts doubt on the narrative that LLM-only approaches can outperform statistical methods by ignoring observational data.

Despite this, LLMs hold promise for supporting scientific discovery. Scientific studies often mix well-established and novel variables: even if LLMs only *recall* known relationships, they can accelerate graph construction, while any genuine reasoning ability would add further value. To realize this potential, we argue that two challenges must be addressed: **(1)** Developing principled evaluation benchmarks that eliminate dataset leakage by using recent scientific studies. **(2)** Designing hybrid methods that integrate LLM-derived knowledge with data-driven inference. Our work contributes to both directions: we introduce a recipe for obtaining leakage-free, science-grounded benchmarks

Preprint. Under review.




(a) Causal Graphs Timeline

(b) Comparing Asis and Alz. Datasets

Figure 1: **(a)** Our novel benchmarks were created by scientists post-2021 with expert consensus, unlike BNLearn graphs from the 1990s likely memorized by LLMs. Using pre-2021 LLM checkpoints ensures fair evaluation on unseen graphs. **(b)** Comparing PC, LLM-BFS, and hybrid PC+LLM on Asia vs. Alzheimers (post-2021, unseen) shows the performance gap between PC and LLM diminishes significantly on unseen data.

| **Dataset** | **M1** | | | | **M2** | | | | **M3 (Nodes)** | | | | **M3 (Edges)** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% |
| Asia | 0.75 | 0.91 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.25 | **1.00** | **1.00** | **1.00** | 0.00 | **1.00** | **1.00** | **1.00** |
| Cancer | **1.00** | **1.00** | **1.00** | 0.5 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.80 | **1.00** | **1.00** | 0.86 | **1.00** | 0.67 |
| Earthquake | 0.60 | 0.75 | 0.67 | 0.50 | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | 0.86 | **1.00** | 1.00 | **1.00** | **1.00** | **1.00** | **1.00** |
| Child | **1.00** | 0.11 | 0.06 | 0.53 | 0.44 | 0.55 | 0.44 | 0.36 | **1.00** | 0.85 | 0.89 | 0.00 | 0.17 | 0.00 | 0.30 | 0.16 |
| Insurance | 0.06 | 0.11 | 0.45 | 0.59 | 0.36 | 0.44 | 0.24 | 0.00 | 0.21 | 0.25 | 0.92 | 0.77 | 0.00 | 0.00 | 0.00 | 0.00 |
| Alarm | **1.00** | 0.72 | 0.79 | 0.89 | 0.49 | 0.38 | 0.12 | 0.00 | 0.97 | 0.72 | 0.92 | 0.00 | 0.43 | 0.10 | 0.00 | 0.00 |

Table 1: F1 scores for memorization tests (M1–M3) across datasets at varying context levels ($\alpha$).

leveraging recent publications that are released after the training cut-offs of the LLMs, and we demonstrate that hybrid methods combining LLM predictions with classical algorithms outperform both approaches individually (Figure 1).

## 2 Limitations of Existing Benchmarks and the Case for Science-Grounding

We critically examine existing causal discovery benchmarks to assess their suitability for LLM-based causal discovery. Our analysis reveals that many benchmarks are compromised by memorization, necessitating the development of novel, science-grounded datasets. For example, popular datasets such as `BNLearn` are often memorized by LLMs, undermining fair evaluation. Detecting memorization is particularly challenging for closed-source models like GPT-4, where training data is unknown. Prior work [7] shows that overlap with training data does not necessarily imply memorization, as it depends on factors such as model size and data frequency in the corpus. Hence, explicit memorization tests are needed, and current techniques typically rely on carefully designed prompts that reveal partial data and test whether LLMs reproduce missing parts verbatim. Such methods have proven effective for tabular [6], image [29], and text data [34, 5, 18, 8, 28]. Building on this, our paper extends the idea to causal graphs by designing reconstruction-based tests, where LLMs are prompted with partial graph information and evaluated on their ability to infer missing structures across three natural categories:

**M1** Given the dataset name and a random $\alpha\%$ subset of nodes, predict the remaining nodes.
**M2** Given dataset name, the full list of nodes, and an $\alpha\%$ of edges, identify the remaining edges.
**M3** Given a dataset and an $\alpha$-subgraph, predict the remaining graph.

Table 1 shows F1 scores for M1–M3 across datasets and context levels ($\alpha$), with prompts detailed in Appendix E. Several trends emerge: (i) many datasets yield near-perfect F1 even at $\alpha = 0$, strongly indicating memorization; (ii) M2 achieves high accuracy even with only node lists, questioning the value of traversal-based strategies like LLM-BFS; (iii) performance drops as graph size grows, evident in `Child` and `Insurance`; and (iv) these patterns collectively cast doubt on existing benchmarks and emphasize the need for leakage-free alternatives.

In summary, our experiments show that LLMs can reproduce `BNLearn` [42] graphs with near-perfect accuracy, strongly indicating memorization and undermining their credibility as benchmarks.

| Causal Graph | Nodes | Edges | Colliders | In-Degree Min | In-Degree Median | In-Degree Max | Longest Path |
|---|---|---|---|---|---|---|---|
| Alzheimer's | 11 | 19 | 1 | 0 | 2 | 4 | 5 |
| COVID-19 Respiratory | 11 | 20 | 1 | 0 | 2 | 4 | 7 |
| Sweden Transport | 11 | 10 | 3 | 0 | 1 | 3 | 3 |
| COVID-19 Complications | 63 | 138 | 23 | 0 | 2 | 7 | 23 |

Table 2: Characteristics of novel science datasets introduced in our work.

| Dataset | M1 | | | | M2 | | | | M3 (Nodes) | | | | M3 (Edges) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% | 0% | 25% | 50% | 75% |
| Alz. | 0.00 | 0.11 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.62 | 0.67 | 0.80 | 0.00 | 0.34 | 0.23 | 0.00 |
| C19-small | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.91 | 1.00 | 0.00 | 0.00 | 0.12 | 0.00 |
| C19-large | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Sweden | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.67 | 0.67 | 0.00 | 0.00 | 0.00 | 0.06 | 0.07 |

Table 3: Results of memorization tests conducted on the novel science datasets.

### 2.1 Science-Grounded Datasets

To ensure fair evaluation, we construct new benchmarks grounded in recent scientific literature. Prior work [23, 9, 22] often relies on synthetic or toy scenarios with limited real-world relevance. Inspired by the original `BNLearn` construction, we design datasets that reflect the complexity of scientific studies. Our methodology involves: **(a)** identifying recent studies that provide or imply causal graphs, and **(b)** extracting source data when available, or generating synthetic data otherwise.

We curated four datasets from their corresponding papers: **(1)** *Alzheimer's* [1], modeling causal relations among clinical phenotypes and MRI-derived radiological features; **(2)** *COVID-19 Respiratory* [33], capturing disease progression within the respiratory system; **(3)** *COVID-19 Complications* [33], extending (2) to multi-organ variables including heart and kidneys; and **(4)** *Sweden Traffic* [57], describing bus delay propagation from Google Transit Feed Specification data. Dataset statistics are reported in Table 2, with detailed descriptions in App C. Since these graphs lack observational data, we generate both linear and non-linear *synthetic* observational datasets (App D).

**Memorization Tests.** Applying reconstruction tests to these datasets yields F1 scores near zero (Table 3), indicating science-grounded datasets are substantially less prone to leakage and better suited for fair benchmarking.

## 3 Call for Hybrid Methods

We evaluate LLM-only methods on the four science-grounded datasets, highlighting their limitations and motivating hybrid approaches that combine LLMs with statistical methods. Experiments span state-of-the-art techniques including GES [10], NOTEARS [58], PC [48], FCI [47], Direct LiNGAM [46], ICA LiNGAM [45], ANM, and LLM-based methods **LLM Pairwise** [27] and **LLM BFS** (details in Appendix B).

**LLM-Only Methods Fall Short on Novel Science Datasets.** Table 4 shows LLM-only methods achieve markedly lower accuracy on science-grounded datasets compared to standard `BNLearn` benchmarks [24]. F1 scores fall below 0.3 on Sweden Transport and Covid-19 Complications, and remain under 0.6 on Covid-19 Respiratory and Alzheimers. Among statistical baselines, LiNGAM variants perform best. For LLM-only methods, LLM BFS proves most effective while requiring fewer prompts than pairwise querying, though it struggles with coherence on the largest benchmark (Covid-19 Respiratory Complications), where statistical methods also degrade.

**Hybrid Methods Bridge the Gap.** We evaluate LLM+PC, which uses LLM BFS to generate a prior graph $\mathcal{G}_{\text{prior}}$ guiding PC during skeleton discovery and edge orientation. Prior edges $X \to Y$ or $X \leftarrow Y$ prevent PC from removing $X \leftrightarrow Y$, even when statistical tests detect conditional independence. Table 4 shows hybrid variants using Fisher's Z-test and KCI consistently achieve the highest F1 scores, outperforming LLM-only and statistical baselines while maintaining robustness across datasets. We now explore several LLM+PC variants.

**Ablation 1: Dropping Edges.** We evaluate post-processing LLM+PC by pruning edges using statistical tests. After ensuring acyclicity, we identify witness sets for remaining edges, perform conditional independence tests, and remove the top $\alpha\%$ edges with lowest $p$-values ($\alpha = 0\%$ is unaltered LLM+PC; higher $\alpha$ yields sparser graphs). Table 5 shows edge pruning consistently

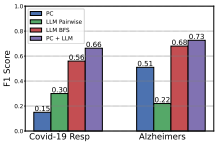
| Methods | Covid-19 Resp. | | | Alzheimers | | | Sweden Transport | | | Covid-19 Compl. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 |
| GES | 0.25 | 0.10 | 0.14 | 0.08 | 0.05 | 0.06 | 0.27 | 0.27 | 0.27 | - | - | |
| PC(Fisherz) | 0.14 | 0.05 | 0.07 | 0.50 | 0.52 | 0.51 | 0.54 | 0.60 | 0.57 | 0.05 | 0.03 | 0.04 |
| PC(KCI) | 0.33 | 0.10 | 0.15 | 0.36 | 0.21 | 0.27 | 0.28 | 0.4 | 0.33 | 0.05 | 0.01 | 0.02 |
| ICA LiNGAM | 0.44 | 0.2 | 0.28 | 0.58 | 0.52 | 0.55 | **0.71** | 0.50 | 0.59 | **0.07** | 0.01 | 0.01 |
| Direct LiNGAM | 0.33 | 0.10 | 0.15 | 0.50 | 0.10 | 0.17 | 0.62 | 0.50 | 0.55 | 0.00 | 0.00 | |
| ANM | 0.44 | 0.20 | 0.28 | 0.30 | 0.15 | 0.20 | 0.22 | 0.2 | 0.21 | 0.04 | 0.04 | 0.04 |
| FCI | 0.30 | 0.15 | 0.20 | 0.42 | 0.26 | 0.32 | 0.50 | 0.3 | 0.38 | 0.02 | 0.03 | 0.03 |
| LLM pairwise | 0.26 | 0.35 | 0.30 | 0.17 | 0.31 | 0.22 | 0.20 | 0.50 | 0.29 | - | - | |
| LLM BFS | **0.90** | 0.45 | 0.60 | **0.69** | 0.47 | 0.56 | 0.25 | 0.4 | 0.31 | 0.06 | 0.04 | 0.05 |
| PC(Fisherz) + LLM | 0.64 | **0.80** | **0.71** | 0.58 | 0.78 | 0.66 | 0.64 | **0.70** | **0.67** | 0.06 | **0.07** | **0.07** |
| PC(KCI) + LLM | **0.90** | 0.45 | 0.60 | 0.64 | **0.84** | **0.73** | 0.50 | 0.50 | 0.50 | **0.07** | 0.05 | 0.06 |

Table 4: Results on Non-Linear Observational Dataset. GES and LLM-pairwise are compute-intensive methods and were not feasible to run for the larger Covid-19 Complications dataset.

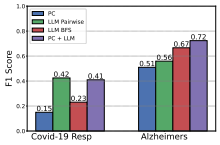
degrades F1 scores across datasets, indicating the original LLM+PC output should be retained without aggressive post-hoc pruning.

| $\alpha$% Edges | COVID-19 | | | Alzheimer's | | | Sweden | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| 0 | **0.64** | **0.80** | **0.71** | 0.58 | **0.78** | **0.66** | 0.63 | **0.70** | **0.67** |
| 5 | **0.64** | 0.70 | 0.67 | 0.58 | 0.74 | 0.65 | 0.60 | 0.60 | 0.60 |
| 10 | 0.62 | 0.65 | 0.63 | 0.57 | 0.68 | 0.62 | **0.66** | 0.60 | 0.63 |
| 25 | 0.55 | 0.50 | 0.52 | 0.53 | 0.53 | 0.53 | 0.62 | 0.50 | 0.55 |
| 50 | 0.42 | 0.25 | 0.31 | **0.61** | 0.42 | 0.50 | 0.4 | 0.2 | 0.27 |

Table 5: Removing edges based on p-Value




(a) GPT-4 Turbo




(b) Llama 3.1

Table 6: Evaluation of GPT-4-Turbo and Llama 3.1

| | P | R | F1 |
|---|---|---|---|
| PC | 0.54 | 0.60 | 0.57 |
| PC+LLM | 0.64 | **0.70** | 0.67 |
| PC+LLM (-ve prior) | **0.70** | **0.70** | **0.70** |

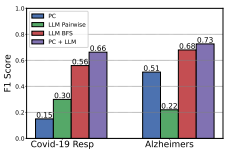
Table 7: Sweden dataset with expert-provided ground-truth negative prior

| $\alpha$% | COVID-19 Resp. | | | Alz. | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| 100 (LLM) | **0.90** | 0.45 | 0.60 | **0.69** | 0.47 | 0.56 |
| 0 (PC+LLM) | 0.57 | **0.70** | **0.63** | 0.56 | **0.70** | **0.62** |
| 25 | 0.60 | 0.62 | 0.61 | 0.56 | 0.61 | 0.58 |
| 50 | 0.60 | 0.54 | 0.50 | 0.55 | 0.55 | 0.55 |
| 75 | 0.67 | 0.48 | 0.56 | 0.61 | 0.49 | 0.54 |

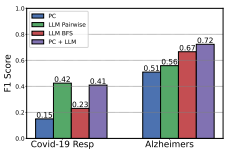
Table 8: PC+LLM performance under varying levels of randomly sampled LLM-derived negative priors

**Ablation 2: Incorporating Priors on Missing Edges.** We evaluate whether PC should use both positive and negative edge priors. Negative edges come from expert knowledge or LLM inferences (e.g., edges not in top-$k$ predictions). We modify LLM+PC to forcibly remove negative prior edges during skeleton discovery. Ground-truth negative priors improve performance with precision gains without recall loss (Sweeden Traffic, Tab. 6). However, LLM-derived negative priors show inconsistent improvements due to noise. Varying $\alpha$ (percentage of non-LLM edges used as negative priors) shows standard PC+LLM ($\alpha = 0$) achieves best F1 compared to LLM-only ($\alpha = 100$) in Tab. 7. Results demonstrate negative priors benefit performance only when high-quality.

**Ablation 3: Extensions using Open-Source LLMs.** We explore open-source LLMs instead of proprietary models, investigating: **1)** training domain-specific models for causal inference, and **2)** end-to-end integration with discovery methods. Both GPT-4-Turbo and Llama 3.1 maintain competitive performance on our benchmarks, with Llama 3.1 showing interesting behavior where pairwise comparison strategies outperform BFS methods. These results demonstrate our benchmarks remain relevant for evaluating newer language models, supporting evaluation framework robustness across different model architectures and training approaches.




(a) GPT-4 Turbo




(b) Llama 3.1

## 4 Conclusion

We showed that many benchmarks in LLM-based causal discovery are compromised by leakage and fail to test genuine reasoning. To address this, we introduced a lightweight strategy for building robust, science-grounded benchmarks. Our results challenge the view that LLM-only methods suffice, instead demonstrating the promise of hybrid approaches that combine LLMs with observational data, pointing to a feasible way of using LLMs in scientific causal discovery.

## References


[1] Ahmed Abdulaal, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C Castro, Daniel C Alexander, et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[3] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023.

[4] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*, 2023.

[5] Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[6] Sebastian Bordt, Harsha Nori, and Rich Caruana. Elephants never forget: Testing language models for memorization of tabular data, 2024.

[7] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

[8] Bowen Chen, Namgi Han, and Yusuke Miyao. A multi-perspective analysis of memorization in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11190–11209, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu. Clear: Can language models really understand causal graphs? *arXiv preprint arXiv:2406.16605*, 2024.

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

[11] Oscar Clivio, Divyat Mahajan, Perouz Taslakian, Sara Magliacane, Ioannis Mitliagkas, Valentina Zantedeschi, and Alexandre Drouin. Learning to defer for causal discovery with imperfect experts, 2025.

[12] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[13] A.P. Dawid. Beware of the DAG! *NeurIPS Workshop on Causality*, 2008.

[14] D. Freedman and P. Humphreys. Are there algorithms that discovery causal structure? *Synthese*, 121, 1999.

[15] K. Fukumizu and A. Gretton. Kernel measures of conditional dependence. *Electronic Proceedings of Neural Information Processing Systems*, 2008.

[16] D. Geiger and D. Heckerman. Learning Gaussian networks. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, 1994.

[17] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[18] Jing Huang, Diyi Yang, and Christopher Potts. Demystifying verbatim memorization in large language models. *arXiv preprint arXiv:2407.17817*, 2024.

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

[20] Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Jarvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. *arXiv preprint arXiv:1309.6836*, 2013.

[21] D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10), 2010.

[22] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024.

[23] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024.

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

[25] Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3):1587–1595, 2022.

[26] Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

[27] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

[28] Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. A comprehensive analysis of memorization in large language models. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[29] Nicky Kriplani, Minh Pham, Gowthami Somepalli, Chinmay Hegde, and Niv Cohen. Solidmark: Evaluating image memorization in generative models. *arXiv preprint arXiv:2503.00592*, 2025.

[30] Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

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

[32] Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023.

[33] Steven Mascaro, Yue Wu, Owen Woodberry, Erik P Nyberg, Ross Pearson, Jessica A Ramsay, Ariel O Mace, David A Foley, Thomas L Snelling, Ann E Nicholson, et al. Modeling covid-19 disease processes by remote elicitation of causal bayesian networks from medical experts. *BMC Medical Research Methodology*, 23(1):76, 2023.

[34] Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. Analyzing memorization in large language models through the lens of model attribution, 2025.

[35] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pages 368–379. PMLR, 2016.

[36] J. Pearl and J. Mackenzie. *The book of why*. Basic Books, USA, 2018.

[37] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

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

[40] Joseph D Ramsey. Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

[41] Thomas Richardson. *Feedback models: Interpretation and discovery*. PhD thesis, Ph. D. thesis, Carnegie Mellon, 1996.

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

[43] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), 2020.

[44] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

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

[46] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Y. Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth A. Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

[47] Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

[48] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[49] James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245, 1980.

[50] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

[51] T. Verma and J. Pearl. Equivalence and synthesis of causal models. *Computer Science Department, UCLA*, 1991.

[52] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Probing for correlations of causal facts: Large language models and causality. 2022.

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

[54] C. Zhang, B. Chen, and J. Pearl. A simultaneous discover-identify approach to causal inference in linear models. *Proceedings of the 34th International Conference on Artificial Intelligence*, 2020.

[55] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

[56] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press.

[57] Qi Zhang, Zhenliang Ma, Yancheng Ling, Zhenlin Qin, Pengfei Zhang, and Zhan Zhao. Causal graph discovery for urban bus operation delays: A case study in stockholm. *Transportation Research Record*, page 03611981241306754, 2025.

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

[59] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR, 26–28 Aug 2020.

# A Code

We release the anonymous code at the url : https://anonymous.4open.science/r/novographs-5005/

# B Background: Types of causal discovery algorithms

We categorize related work into: (a) *data-driven* methods, which rely solely on observational datasets to infer the causal graph, (b) *LLM-based* methods, which rely solely on prompt responses, and (c) *hybrid* methods, which use both LLMs and observational datasets.

**Data-Driven Methods.** Constraint-based methods for causal discovery, such as the PC algorithm [48] and the FCI algorithm [47], identify causal relationships by testing conditional independencies. Variants of these methods [12, 41, 20, 51, 54, 43] aim to improve scalability and accommodate different assumptions. While some of these methods offer asymptotic consistency guarantees, their performance in practice often depends on the power of the statistical hypothesis tests applied to determine conditional independencies from observational data, a factor we examine in our experiments. Standard tests include Fisher's z-test [49] for linear dependencies and kernel-based tests [56] for non-linear dependencies. Other methods include score-based methods [10, 35, 21, 40] that optimize a score function over graphs, including recent versions based on continuous optimization [58]; and parametric methods that assume parametric assumptions about the functional relationships among nodes in a causal graph, e.g., assuming non-gaussian noise [45, 30].

**Leveraging LLMs for learning Causal Graphs.** There is a growing interest in augmenting observational data with meta-knowledge, aiming for improved causal predictions [1]. Large Language Models (LLMs) offer a promising source of such augmentation, requiring minimal manual effort. For instance, the pairwise approach [27, 52, 32] finds the causal graph using prompts like "Does A cause B?" for each pair of nodes, then coalesces the graph based on the responses. While effective, this method requires $O(n^2)$ prompts for $n$ nodes, making it costly. Alternative approaches [24] reduce prompt complexity by building the graph with a breadth-first search. Another recent approach considers querying LLMs over triplet of variables [50].

**Hybrid Approaches.** ALCM [26] is a recent approach that begins with the PC algorithm and subsequently queries the LLM to validate each edge predicted by the PC. Other methods in this category initiate with a prior LLM-based graph and adjust it using observational data [4, 2] or use LLM as a post-processing critic for data-based output [31, 50]. [11] introduces a method that adaptively defers to either expert (LLM) recommendations or data-driven causal discovery based on their reliability. In their work, [24] presented a variant that incorporates the $p$-values from statistical tests into the prompts while constructing the causal graph. However, the authors found that the inclusion of $p$-values does not yield any improvement over their standalone LLM variant. This indicates that merely adding superficial data statistics to the prompts is less effective, highlighting the necessity for explicit mechanisms to integrate LLM and data-driven graph predictions, and for testing such mechanisms on non-memorized benchmarks.

However, almost all of the above studies use popular, existing graph datasets such as bnlearn for evaluation of LLM-based methods. In the next section, we show why such evaluation is not reliable.

# C Detailed Description of the Datasets

In this section we discuss the four causal graphs, each developed in a recent publication through careful expert elicitation and consensus. Key statistics for these graphs are summarized in Table 2. As new LLMs are introduced, the recipe can be repeated to generate more novel datasets.

**Alzheimer's Graph** The first dataset is the Alzheimer's graph from [1], developed with input from five domain experts. It includes two broad categories of variables: clinical phenotypes (e.g., age, sex, education) and radiological features extracted from MRI scans (e.g., brain and ventricular volumes), as illustrated in Fig. 3. The consensus graph was built by retaining only those edges that were agreed upon by at least two of the five experts. As highlighted in Figure 21 of [1], there is substantial disagreement among the individual expert graphs, underscoring the difficulty for automated methods such as LLMs to infer a consensus graph. Although the graph's structure was developed independently, its variables align with a subset of those used in the Alzheimer's Disease Neuroimaging Initiative [39].

**COVID-19 Respiratory Graph** The second graph models the progression of COVID-19 within the respiratory system, as introduced in [33]. It tracks the disease's path from initial viral entry to pulmonary dysfunction and symptomatic manifestations. The graph was developed through iterative elicitation sessions involving 7–12 domain experts and released on medRxiv in February 2022. Figure 2 presents the graph with color-coded nodes corresponding to different stages of infection: viral entry (pink), lung mechanics (yellow), infection-induced complications (orange), and observable symptoms (cyan). Each variable captures a phase in the progression from infection to respiratory distress. The graph was refined through group workshops and follow-ups, followed by independent expert validation to ensure consensus and accuracy.

**COVID-19 Complications Graph** The third dataset extends the respiratory model to include systemic complications resulting from COVID-19, again from [33]. This graph captures how the virus can affect organs beyond the lungs, such as the heart, liver, kidneys, and vascular system. It includes variables like vascular tone, blood clotting, cardiac inflammation, and ischemia, while retaining key pulmonary indicators such as hypoxemia and hypercapnia (see Fig. 2). Constructed using a similar expert elicitation process, this graph focuses on mapping primary pathways that lead to severe complications, including immune overreactions and multi-organ failure. It distinguishes between observable variables used in clinical monitoring and latent variables that reflect complex physiological states. With 63 nodes and 138 edges, this is the most complex of the four graphs and presents a challenging testbed for causal discovery algorithms.

**The Sweden Traffic Dataset** The Sweden traffic dataset was introduced in a recent study [57] aimed at modeling bus delay propagation through a causal graph. Each node corresponds to a variable that influences delays, such as `arrival_delays`, `dwell_time`, and `scheduled_travel_time`. Unlike the previous three studies, a notable feature of this work is that the true graph is not known since it deals with real-world bus traffic data. Instead, the authors provide expert annotations specifying a subset of edge that should definitely exist, and a subset that are forbidden. Thus, the ground-truth contains not only *positive* edges that should be present in the causal graph but also *negative* edges that must be absent. The dataset is inspired by the General Transit Feed Specification (GTFS), a standardized format for public transit schedules and geographic data. As such, benchmarking causal discovery methods on this dataset holds promise for informing real-world applications in transportation systems analysis.

## D Description of the Synthetic Observational Data

For datasets where source data is unavailable, we generate synthetic observational data based on expert-designed causal graphs, following the approach used in the `BNLearn` benchmark. We consider two settings: **(a) Linear** and **(b) Non-Linear**, differing in the form of structural equations used for each node. Data is generated in topological order over the graph. Root nodes are sampled as $\mathbf{x}_i \sim \mathcal{N}(0, 1)$. For non-root nodes, we use: $\mathbf{x}_i \sim f_i(\text{Pa}_i) + \epsilon_i$ where $\text{Pa}_i$ denotes the values of the parents of node $i$, and $\epsilon_i \sim \mathcal{N}(0, 1)$ is an exogenous noise term. In the **Linear** setting, $f_i(\text{Pa}_i) = \mathbf{w}^\top \text{Pa}_i$ with weights $\mathbf{w}$ drawn from $\mathcal{U}(0, 2)$ to ensure consistent scaling across graph depths. In the **Non-Linear** setting, $f_i$ is parameterized by a randomly initialized 3-layer MLP with ReLU activations and four neurons per hidden layer: $f_i(\text{Pa}_i) = \text{MLP}(\text{Pa}_i)$ This setup enables flexible modeling of complex non-linear relationships, as in prior work [59].

## E Prompts for Memorization Tasks

**Prompt Template for M1 Task**

You are provided with the name of the `bnlearn` dataset: `{dataset_name}` and the following nodes: `{given_nodes}`. Give me the remaining nodes. Strictly output the nodes in the format: `['node1', 'node2', 'node3']`.

**Note:** Add `bnlearn` if it is a `bnlearn` dataset.

Prompt Template for M2 Task

You are provided with the name of the bnlearn dataset: `{dataset_name}`, all nodes: `{all_nodes}`, and the following edges: `{given_edges}`. Give me the remaining edges of the graph. Strictly output the edges in the format: `[['node1', 'node2'], ['node1', 'node3'], ['node2', 'node3']]`.

**Note:** Add `bnlearn` if it is a `bnlearn` dataset.

Prompt Template for M3 Task

You are provided with the name of the bnlearn dataset: `{dataset_name}`, the following nodes: `{given_nodes}`, and the following edges: `{given_edges}`. Give me the remaining nodes and edges. Strictly output the nodes and edges in the format and do not add any text before or after the list:
`{'remaining_nodes': ['node1', 'node2', 'node3'], 'remaining_edges': [['node1', 'node2'], ['node1', 'node3'], ['node2', 'node3']]}`

**Note:** Add `bnlearn` if it is a `bnlearn` dataset.

# F Visualizations of Novel Sciences benchmark

The Covid-19 respiratory dataset represents the full pathway of Covid-19's impact on the body, organized into six distinct subsystems: vascular, pulmonary, cardiac, system-wide, background, and other organs. This dataset provides a comprehensive view of Covid-19's effects as observed across various aspects of human anatomy.

The complexity of this dataset stems from the high level of interconnections between the subsystems, resulting in a dense causal graph structure with 63 nodes and 138 edges. This density, along with numerous collider structures, makes it exceptionally challenging to analyze, even with advanced statistical algorithms and causal discovery methods.

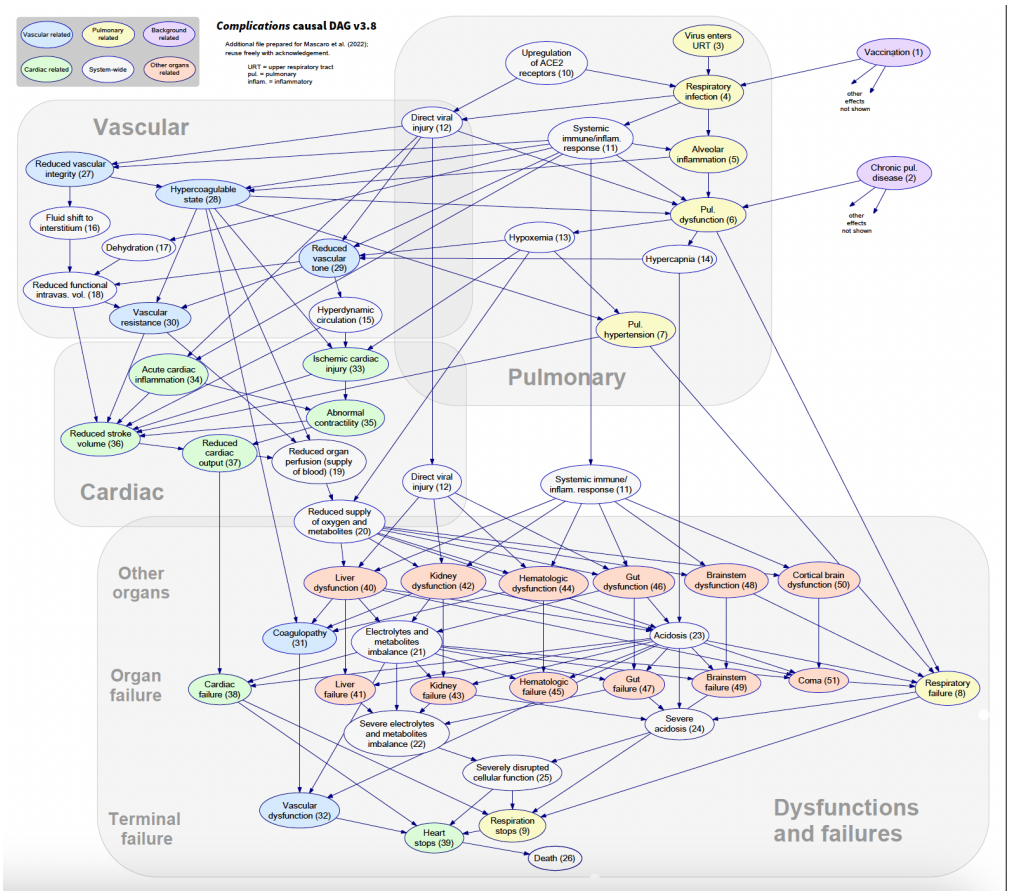
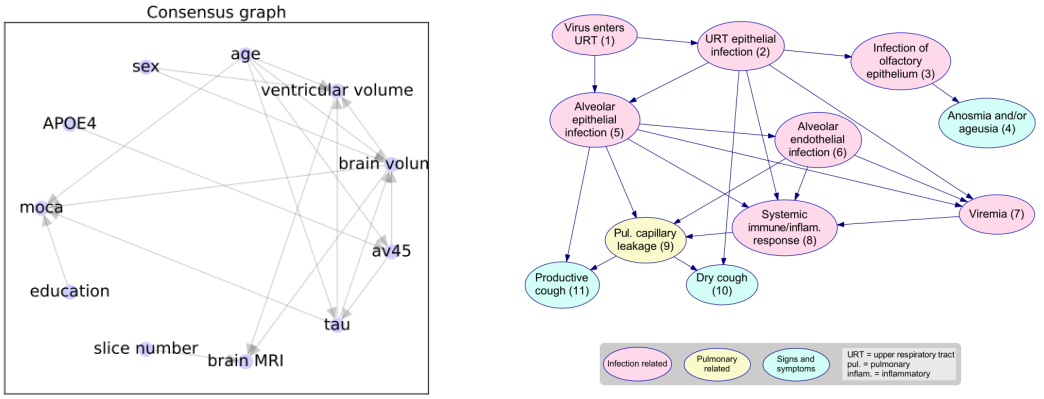



Figure 2: Covid-19 Complications Graph, reproduced from [33].




(a) Alzheimers

(b) Covid-19 Respiratory

Figure 3: Consensus causal graphs for Alzheimers benchmark reproduced from [1], and Covid-19 Respiratory dataset reproduced from [33].

### F.1 Sweden Urban Bus Operation Delays (Sweden Transport) Dataset Description

The Sweden Transport dataset [57] contains temporal and operational information from a public bus network. The variables in the dataset are defined as follows:

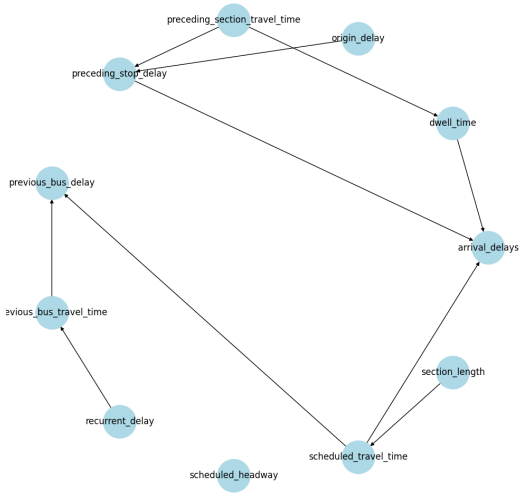



Figure 4: Causal graph obtained from the Sweden Urban Bus Operation Delays dataset.

| True assertion | False assertion |
|---|---|
| Preceding stop delay → arrival delay | Dwell time → preceding stop delay |
| Dwell time → arrival delay | Dwell time → preceding section travel time |
| Scheduled travel time → arrival delay | Scheduled headway → dwell time |
| Scheduled travel time → previous bus delay | Section length → preceding section travel time |
| Preceding section travel time → preceding stop delay | Preceding stop delay → preceding section travel time |
| Previous bus travel time → previous bus delay | Previous bus delay → previous bus travel time |
| Recurrent delay → previous bus travel time | Preceding stop delay → previous bus delay |
| Origin delay → preceding stop delay | Scheduled travel time → preceding section travel time |
| Preceding section travel time → dwell time | Section length → origin delay |
| Section length → scheduled travel time | Origin delay → previous bus delay |

Figure 5: Edges obtained from the Sweden Transport dataset. Both positive and negative causal edges are shown. These tables are quoted from the original paper [57] for ease of reference.

- **Arrival Delays**: Arrival delay of bus $j$ at stop $i$; the difference between the actual arrival time and the scheduled arrival time.
- **Dwell Time**: Actual dwell time at the preceding stop $(i-1)$; the difference between actual departure and arrival time at stop $i-1$ for bus $j$.
- **Preceding Section Travel Time**: Actual running time between stops $i-2$ and $i-1$; the difference between arrival at $i-1$ and departure from $i-2$.
- **Scheduled Travel Time**: Scheduled running time between stops $i-1$ and $i$; the difference between scheduled arrival at $i$ and scheduled departure from $i-1$.
- **Preceding Stop Delay**: Arrival delay of bus $j$ at stop $i-1$; the difference between actual and scheduled arrival time at stop $i-1$.
- **Previous Bus Delay**: Arrival delay (knock-on effect) of preceding bus $j-1$ at stop $i$; the difference between its actual and scheduled arrival time.
- **Previous Bus Travel Time**: Actual running time of bus $j-1$ between stops $i-1$ and $i$; used to indicate current traffic conditions.
- **Recurrent Delay**: Historical mean travel time of bus $j$ at stop $i$ during the same hour on weekdays; reflects recurrent congestion patterns.
- **Origin Delay**: Departure delay of bus $j$ at the first stop; the difference between actual and scheduled departure time.
- **Scheduled Headway**: Planned time interval between arrival times of buses $j-1$ and $j$ at stop $i$.
- **Section Length**: Distance between stop $i-1$ and $i$ (in metres).

Table 9: Results on Linear Observational Dataset.

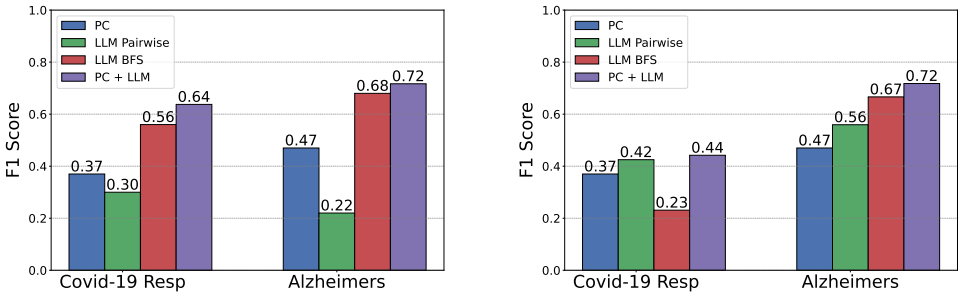
| methods | Covid-19 Resp. | | | Alzheimers | | | Sweden Transport | | | Covid-19 Compl. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** |
| GES | 0.16 | 0.20 | 0.18 | 0.26 | 0.26 | 0.26 | 0.21 | 0.30 | 0.25 | - | - | |
| PC(Fisherz) | 0.31 | 0.45 | 0.37 | 0.47 | 0.47 | 0.47 | 0.44 | **0.80** | 0.57 | 0.04 | 0.02 | 0.03 |
| PC(KCI) | 0.27 | 0.25 | 0.26 | 0.57 | 0.42 | 0.48 | **0.66** | **0.80** | **0.72** | 0.03 | 0.015 | 0.02 |
| NOTEARS | 0.13 | 0.10 | 0.11 | 0.16 | 0.26 | 0.20 | 0.16 | 0.20 | 0.18 | - | - | |
| ICA LiNGAM | 0.25 | 0.20 | 0.22 | 0.11 | 0.26 | 0.15 | 0.21 | 0.30 | 0.25 | 0.05 | 0.17 | 0.07 |
| Direct LiNGAM | 0.18 | 0.35 | 0.24 | 0.20 | 0.30 | 0.24 | 0.16 | 0.30 | 0.21 | 0.03 | 0.17 | 0.05 |
| ANM | 0.25 | 0.20 | 0.22 | 0.19 | 0.2 | 0.19 | 0 | 0 | - | 0.04 | **0.58** | **0.07** |
| FCI | 0.12 | 0.15 | 0.13 | 0.60 | 0.16 | 0.25 | 0.50 | 0.40 | 0.44 | 0.04 | 0.01 | 0.01 |
| LLM Pairwise | 0.26 | 0.35 | 0.30 | 0.17 | 0.31 | 0.22 | 0.20 | 0.50 | 0.29 | - | - | |
| LLM BFS | **0.90** | 0.45 | 0.60 | **0.69** | 0.47 | 0.56 | 0.25 | 0.40 | 0.31 | **0.06** | 0.04 | 0.05 |
| PC(Fisherz) + LLM | 0.46 | **0.70** | 0.56 | 0.54 | 0.68 | 0.60 | 0.53 | **0.80** | 0.64 | **0.06** | 0.06 | 0.06 |
| PC(KCI) + LLM | 0.63 | 0.60 | **0.61** | 0.60 | **0.78** | **0.68** | **0.66** | **0.80** | **0.72** | **0.06** | 0.05 | 0.05 |

## G Results on Linear Observational Dataset

Statistical methods were applied to linearly generated data, and results were obtained using GPT-4 with a 2021 cutoff, facilitating a comparison of performance between traditional algorithms, the LLM-based approach and our hybrid method.

## H Ablations using Linear dataset

We conduct ablation studies using GPT-4 Turbo and LLaMA 3.1 on linearly generated data and observed that our hybrid PC+LLM method outperforms both individual baselines. This demonstrates the advantage of combining PC's statistical rigor with LLM's contextual reasoning for causal discovery.




(a) GPT-4 Turbo (b) Llama 3.1

Figure 6: Evaluation of GPT-4-Turbo and Llama 3.1 models on Novel Sciences benchmarks. Notably, these models were trained after the release of these datasets, so there is a possibility that they may have encountered our datasets during training.

## I Experiments for Research Question: RQ4

We conduct a series of ablation studies to assess the robustness and generalization ability of our hybrid PC+LLM approach under various modifications to the data generation process.

**Ablation 1: MLP Depth.** We evaluate the impact of increasing the depth of the nonlinear generators by replacing 3-layer MLPs in our default setting with 5-layer MLPs. The results in Table 10 (left) indicate that performance remains consistent, suggesting insensitivity to architectural depth.

**Ablation 2: Noise Distribution.** To assess robustness under different exogenous noise assumptions, we replace the default $\mathcal{N}(0,1)$ noise with $\mathcal{N}(0,0.1)$ and $\mathcal{U}(0,1)$. As shown in Table 10 (right), PC+LLM consistently outperforms the PC baseline across all settings.

| Method | COVID-19 Resp. | | | Alzheimers | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| LLM | **0.90** | 0.45 | 0.60 | **0.69** | 0.47 | 0.56 |
| PC | 0.35 | 0.25 | 0.30 | 0.44 | 0.42 | 0.43 |
| PC + LLM | 0.73 | **0.55** | **0.63** | 0.60 | **0.78** | **0.68** |

| Noise | Method | COVID-19 Resp. | | | Alzheimers | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| $\mathcal{N}(0,0.1)$ | PC | 0.34 | 0.40 | 0.37 | 0.38 | 0.37 | 0.38 |
| | PC+LLM | **0.58** | **0.70** | **0.63** | **0.56** | **0.68** | **0.62** |
| $\mathcal{U}(0,1)$ | PC | 0.60 | 0.30 | 0.40 | 0.58 | 0.36 | 0.45 |
| | PC+LLM | **0.85** | **0.60** | **0.70** | **0.60** | **0.63** | **0.62** |

Table 10: Left: Effect of deeper MLPs on performance. Right: Performance under noisy LLM-derived priors.

**Ablation 3: MLP Initialization.** We compare three initialization strategies for MLP weights: uniform $\mathcal{U}(0,1)$, standard normal, and Xavier normal. As seen in Table 11 (left), the hybrid method retains its advantage across all configurations.

**Ablation 4: Linear Coefficient Sampling.** We vary the distribution used for sampling linear SEM coefficients, testing $\mathcal{U}(0,2)$, $\mathcal{N}(0,2)$, and $\mathcal{U}(-1,1)$. Table 11 (right) shows that PC+LLM consistently achieves superior recall and F1 scores.

| Init. | Method | COVID-19 Resp. | | | Alzheimers | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| Std Normal | PC | 0.44 | 0.20 | 0.28 | 0.35 | 0.37 | 0.36 |
| | PC+LLM | **0.73** | **0.55** | **0.63** | **0.50** | **0.57** | **0.54** |
| Xavier Normal | PC | 0.38 | 0.40 | 0.39 | 0.40 | 0.47 | 0.43 |
| | PC+LLM | **0.59** | **0.80** | **0.68** | **0.54** | **0.68** | **0.60** |

| Coeff. Dist. | Method | COVID-19 Resp. | | | Alzheimers | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| $\mathcal{N}(0,2)$ | PC | 0.14 | 0.20 | 0.16 | 0.44 | 0.42 | 0.43 |
| | PC+LLM | **0.66** | **0.70** | **0.68** | **0.59** | **0.68** | **0.63** |
| $\mathcal{U}(-1,1)$ | PC | 0.26 | 0.55 | 0.36 | 0.48 | 0.63 | 0.54 |
| | PC+LLM | **0.59** | **0.65** | **0.62** | **0.53** | **0.79** | **0.64** |

Table 11: Left: Performance across different MLP initializations. Right: Effect of different coefficient sampling distributions.

**In Summary,** these results collectively demonstrate the robustness and effectiveness of our method across a wide range of data-generating assumptions. Across all ablations, our PC+LLM hybrid approach consistently outperforms the standalone PC method. These experiments effectively illustrate the robustness of hybrid approaches.