Large Language Models as Tools to Improve Bayesian Causal Discovery

Bruna Bazaluk[1] **Benjie Wang**[2] **Denis Deratani Mauá**[1] **Flavio S Correa da Silva**[1]

[1]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil
[2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

Abstract

Causal discovery is the task of automatically inferring causal structures, typically from observational data. Recently, there has been much interest in utilizing domain knowledge from large language models (LLM) in causal discovery. However, existing LLM-based approaches only output a single directed acyclic graph (DAG) without uncertainty, which can be unreliable. In this work, we investigate using LLMs alongside Bayesian structure learning (BSL) methods for causal discovery, which output a distribution of possible graphs. In particular, we propose to harness the domain knowledge from the LLMs in the prior distribution over graphs, in place of uninformed priors or human expertise. Our experiments show that LLM-informed priors can improve the performance of Bayesian structure learning methods.

1 INTRODUCTION

Causal Discovery (CD) is the task of automatically discovering causal structure, typically represented as a directed acyclic graph (DAG), from data generated from the causal system. It has proven useful in various fields, including economics [Imbens, 2004], biology [Wen et al., 2023], chemistry [Bi et al., 2023], and several others. Hence, there are a wide range of algorithms to solve this problem. For instance, constraint-based methods use conditional independencies found in the data [Colombo and Maathuis, 2014, Zhang, 2008]. On the other hand, score-based methods define a score function over DAGs, and search this space for the best-scoring DAG [Zheng et al., 2018, Alonso-Barba et al., 2013].

Most causal discovery algorithms output a point estimate of the true causal graph. However, in real-life applications of CD, it is often important to assess the uncertainty about the selected model. This allows the analyst, for example, to estimate probabilities of certain properties (e.g., that some variables is a cause of some other), or consider different plausible causal mechanisms consistent with the observations. Following this philosophy, some works propose CD in a Bayesian manner [Friedman and Koller, 2000, Cundy et al., 2021, Lorch et al., 2021]. Bayesian Structure Learning algorithms define a posterior distribution of possible causal DAGs given a prior distribution and likelihood function, that is:

$$p(G, \Theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(G)p(\Theta|G)p(\mathcal{D}|G, \Theta), \qquad (1)$$

where $p(G)$ is the prior distribution over the possible graph structures, $p(\Theta|G)$ is the prior over the BN parameters, and $p(\mathcal{D}|G, \Theta)$ is a likelihood function for the dataset $\mathcal{D}$ given graph and parameters. Sampling from or tractably representing this distribution is challenging as the marginal data distribution, $p(\mathcal{D})$, is computationally hard to estimate (e.g., the number of DAG structures to average over is super-exponential). As such, most algorithms rely on approximate inference.

There has also been much recent interest in exploiting the domain knowledge contained within large language models (LLMs) to aid CD. Some works focus on using the LLM alone to infer the causal graph [Long et al., 2022, Jiralerspong et al., 2024], while others incorporate LLM as an oracle within existing causal discovery approaches [Ban et al., 2025, 2023, Liu et al., 2024, Li et al., 2025].

The goal of this work is to use LLMs as a knowledge base of expert domain knowledge to improve existing Bayesian CD algorithms. For this purpose, we prompt GPT-4o [OpenAI et al., 2024] and DeepSeek-V3 [DeepSeek-AI et al., 2025] about each pair of edges and use the outputs as a prior distribution for DiBS [Lorch et al., 2021], a differentiable Bayesian structure learning algorithm based on variational inference.

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

2 RELATED WORK

Bayesian Causal Discovery Due to the intractability of the posterior over causal graphs in high dimensions, Bayesian approaches typically approximate Equation (1) typically resort to Markov Chain Monte Carlo Madigan and York [1995], Heckerman et al. [2006] or variational inference [Cundy et al., 2021, Lorch et al., 2021]. The output of these algorithms consists of a sample of graphs or variational representation of the posterior, which can be used for Bayesian model averaging for inference of downstream causal effects [Toth et al., 2022]. In this paper, we focus on the DiBS Bayesian structure learner [Lorch et al., 2021], which is a differentiable framework for that operates in the continuous space of a latent probabilistic graph representation. DiBS produces a set of (approximate) samples from either the marginal posterior over DAG structures or the joint posterior over DAG structures and parameters (Equation (1)).

LLMs for Causal Discovery Some works in this area focus on extracting the whole graph directly from the LLM. Jiralerspong et al. [2024], for example, focuses on prompting the LLM edge by edge using breadth-first search (BFS), which uses a linear number of queries. Vashishtha et al. [2023], on the other hand, proposes some prompting strategies asking the LLMs about three variables at a time. There are also recent works focusing on using LLMs to select intervention targets for CD [Li et al., 2025]. Kiciman et al. [2024] empirically study the use of LLMs in causal discovery tasks and find that LLM-based methods, on average, outperform state-of-the-art data-driven CD algorithms in well-known datasets or common sense knowledge; though they can also exhibit unexpected failure modes. The prompts we use in the present work are inspired by examples from this paper.

3 METHODOLOGY

In Bayesian structure learning, a key step is the specification of the prior over causal graphs. Typically, this involves either using either simple, uninformed priors [Eggeling et al., 2019], or edge probabilities obtained from human experts; the former does not incorporate any domain knowledge, while the latter is not always available. In this work, we propose to instead elicit this prior information from large language models. In particular, we propose to extract prior probailities for each individual edge from the LLM.

There is an ongoing discussion in the literature about how to calculate the uncertainty over an LLM's response to a question [Tanneru et al., 2024]. Directly prompting the LLM to respond with its uncertainty was shown not to be very accurate [Xiong et al., 2024]. Some other works use the model's logits to estimate its uncertainty over its answer [Ma et al., 2025]. However, logits are not available for the best performing closed-source LLMs such as GPT-4o.

Thus, we choose to work with a Monte Carlo estimate, in which we use 10 sampled LLM generations for each edge. To avoid degenerate probabilities, we replace 1 with 0.99 and 0 with 1e-10. As a result, we obtain a probabilistic adjacency matrix $P \in (0,1)^{d\times d}$ where d is the number of variables. The original prior distribution used in DiBS assumes that each edge exists independently with a given probability. We substitute this prior with the edge probabilities given by the LLM's output:

$$p(G) \propto \prod_{(i,j)\in E} P_{ij} \prod_{(i,j)\notin E} (1 - P_{ij}). \tag{2}$$

4 EXPERIMENTS AND RESULTS

We use the Sachs protein dataset [Sachs et al., 2005], which was used in the DiBS paper, the MAGIC-NIAB [Scutari et al., 2014] dataset, and a custom dataset based on common sense knowledge that we call Summer. Full details about all of the datasets and causal graphs can also be found in Appendix A.

```
You are an expert on the human immune system
    cell, who responds only in numbers.
```

```
You are an expert on the human immune system
cell. You are investigating the cause-and-effect
    relationships between a set of observed
      variables representing proteins and
 phospholipids: "Raf", "Mek", "Plcg", "PIP2",
  "PIP3", "Erk", "Akt", "PKA", "PKC", "P38",
  "Jnk". Your task is to determine the causal
  relationship between the variables "Raf" and
  "Mek". Answer "0" if "Raf" and "Mek" are not
 statistically associated. Answer "1" if "Raf"
   directly causes "Mek". Answer "2" if "Mek"
            directly causes "Raf".
```

Figure 1: Sachs prompt example. The first box contains the instruction given to the model, and the second the content.

For the LLM prior extraction, we prompt 2 different LLMs: GPT-4o [OpenAI et al., 2024] and DeepSeek-V3 [DeepSeek-AI et al., 2025]. We show a prompt example for the Sachs dataset in Figure 1. According to the LLM's outputs, we build probabilistic adjacency matrices that are used as priors in DiBS. Figures 2, 3 and 4 show the matrices, where each entry represents the probability of the respective edge. Green represents 100% and dark blue 0%.

For all experiments, we run DiBS with 30 starting latent particles and 3000 iterations. We repeat the experiment 10

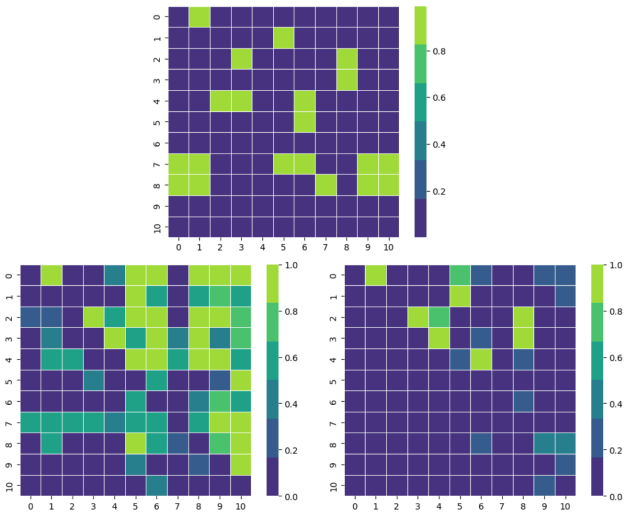

Figure 2: Sachs' true adjacency matrix, followed by DeepSeek's prior on the left and GPT's prior on the right.

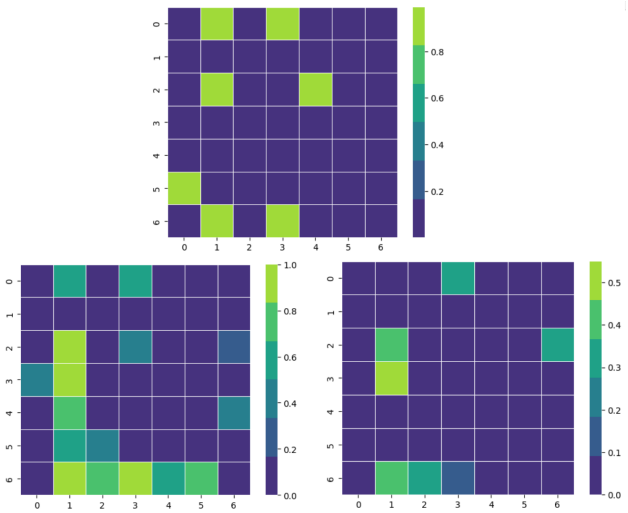

Figure 3: MAGIC-NIAB's true adjacency matrix, followed by DeepSeek's prior on the left and GPT's prior on the right.

times and calculate the expected structural Hamming distance ($\mathbb{E}$-SHD) for each of them, defined by:

$$\mathbb{E}\text{-SHD}(p, G^*) := \sum_G p(G|\mathcal{D})\ \text{SHD}(G, G^*), \qquad (3)$$

where G^* is the ground truth graph, and $p(G|\mathcal{D})$ is the approximate posterior over graphs given by DiBS (in this case, an empirical sample of 30 graphs). We also compute the AUROC, which is a standard metric that treats Bayesian structure learning as a probabilistic binary classification problem for each edge.

As a baseline to assess the LLM's elicited prior, we compute the $\mathbb{E}$-SHD over both GPT's and DeepSeek's estimated distributions (that are used to define the BSL prior). In

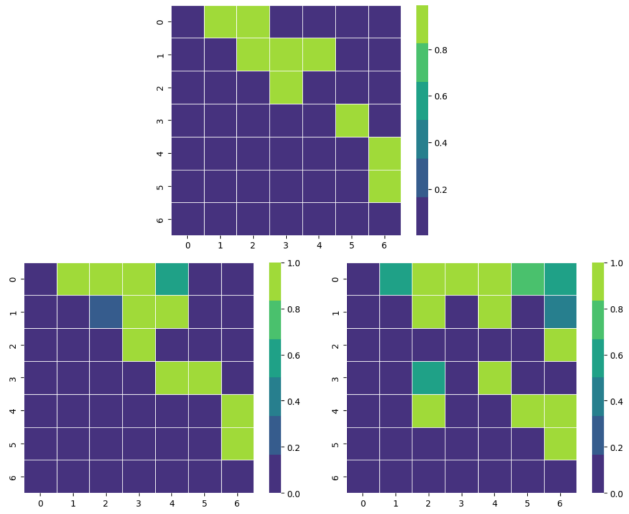

Figure 4: Summer's true adjacency matrix, followed by DeepSeek's prior on the left and GPT's prior on the right.

particular, we sample 30 random graphs according to the estimated edge probabilities, rejecting those that are not acyclic. Then, we compute the (average) SHD between the remaining DAG samples and the ground truth graph. Table 1 shows the results.

Analyzing the results for the **Sachs** dataset, we observe that, in both cases (GPT and DeepSeek), the prior distribution is as close (if not closer) to the ground truth graph as the approximate posterior (inferred by DiBS) in expectation, as measured by E-SHD. However, the highest AUROC is achieved by DiBS using DeepSeek's prior, while GPT's prior alone results in the lowest AUROC, even though it has the best SHD. This can be explained by the results seen in Figure 2, where GPT predicts very few edges (leading to a low E-SHD since the true graph is fairly sparse), but does not accurately capture uncertainty over the remaining edges.

For the **MAGIC-NIAB** dataset, we construct a case where causal sufficiency is violated and so data-driven approaches assuming suffiency are expected to perform poorly. In particular, we select a subset of the 44 variables and consider the subgraph induced by those variables (see Appendix A.2 for details). In this case, there are many unobserved confounders, and so, as expected, DiBS discovers significantly more edges than are actually present in the causal graph. On the other hand, the LLM priors achieve better performance at distinguishing the true causal edges due to their domain knowledge. Interestingly, we also observe that DiBS with either of the LLM priors performs similarly to DiBS using the original edge prior. This suggests that, at least for DiBS, prior information is insufficient to correct for the data-generating assumptions being violated.

Given that Sachs and MAGIC-NIAB are existing datasets/-causal graphs that may have appeared in the pretraining

Table 1: Experimental result for causal discovery on Sachs, MAGIC-NIAB and Summer datasets; we show the mean and standard deviation over 10 runs.

Sachs		
MODEL	E-SHD	AUROC
DiBS	23.4 ± 0.5	0.58 ± 0.04
DiBS + GPT	21.7 ± 0.5	0.64 ± 0.03
DiBS + DeepSeek	26.7 ± 0.6	$\mathbf{0.67 \pm 0.06}$
GPT (prior)	$\mathbf{18.07 \pm 0.25}$	0.57 ± 0.005
DeepSeek (prior)	24.97 ± 0.55	0.63 ± 0.009

Magic-Niab		
MODEL	E-SHD	AUROC
DiBS	13.76 ± 1.1	0.43 ± 0.05
DiBS + GPT	16.07 ± 2.68	0.45 ± 0.04
DiBS + DeepSeek	16.24 ± 2.75	0.46 ± 0.04
GPT (prior)	7.03 ± 0.18	0.5 ± 0
DeepSeek (prior)	$\mathbf{5.07 \pm 0.25}$	$\mathbf{0.70 \pm 0.01}$

Summer		
MODEL	E-SHD	AUROC
DiBS	12.17 ± 0.98	0.80 ± 0.05
DiBS + GPT	12.42 ± 0.38	0.80 ± 0.03
DiBS + DeepSeek	9.68 ± 1.25	0.89 ± 0.04
GPT (prior)	8.97 ± 0.18	0.71 ± 0.01
DeepSeek (prior)	$\mathbf{3.07 \pm 0.25}$	$\mathbf{0.91 \pm 0.02}$

data of LLMs, we chose to construct a novel commonsense causal graph, **Summer**. For this graph, incorporating DeepSeek's prior substantially improves the original model, both in terms of SHD and AUROC. In contrast, GPT's prior yields results comparable to those of the original DiBS prior. Nevertheless, in the case of GPT, the posterior distribution is closer to the ground truth than the prior when considering AUROC alone.

Overall, for the three datasets, the best SHD results were from the prior distributions alone; for both the MAGIC-NIAB and Summer datasets, using DeepSeek alone yields the best overall performance in terms of both E-SHD and AUROC. This indicates that modern LLMs are highly effective at solving simpler CD tasks, particularly discovering small graphs in well-known domains.

One limitation of using (E-)SHD as a metric is that a DAG with some correctly placed edges and some incorrect ones may have the same SHD as a graph in which all edges are incorrectly placed. Additionally, it does not reward accurate estimation of uncertainty. Considering AUROC alone, we observe that for the Summer and Sachs datasets, incorporating GPT prior through BSL significantly improved the results. For the Sachs dataset, the best AUROC among all tested methods is given by using DiBS with DeepSeek's prior distribution.

5 CONCLUSION AND FUTURE WORK

In this work, we use two state-of-the-art LLMs to build distributions of graphs to be used as priors for a Bayesian Structure Learning framework that infers posterior distributions of graphs given some dataset. We show that, if the constructed LLM prior is close enough to the ground truth graph, it is possible to combine the knowledge from an LLM with a Bayesian Structure Learning framework to yield better results than using the framework alone.

For the Sachs [Sachs et al., 2005] dataset, GPT-4o's prior was closer to the ground truth than the posterior distribution. An important observation is that Sachs [Sachs et al., 2005] is one of the most commonly used datasets to test CD algorithms. Therefore, it may have had a strong presence on the LLM's training set, which leads to the LLM being more accurate in its responses. However, we also see good results when using our custom dataset, which is build over common-sense knowledge. Thus, future work could examine experimenting with different BSL algorithms and larger causal graphs. One could also further investigate the distributional uncertainty provided by the LLM posterior beyond SHD and AUROC.

References

Juan I. Alonso-Barba, Luis Delaossa, Jose A. GáMez, and Jose M. Puerta. Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *Int. J. Approx. Reasoning*, 54(4):429–451, June 2013. ISSN 0888-613X. doi: 10.1016/j.ijar.2012.09.004. URL https://doi.org/10.1016/j.ijar.2012.09.004.

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, 2023. URL https://arxiv.org/abs/2306.16902.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen. Llm-driven causal discovery via harmonized prior. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1943–1960, 2025. doi: 10.1109/TKDE.2025.3528461.

Mitul N. Bavaria, Shi Jin, Ramesh M. Ray, and Leonard R. Johnson. The mechanism by which mek/erk regulates jnk and p38 activity in polyamine depleted iec-6 cells during apoptosis. *Apoptosis*, 19:467 – 479, 2013. URL https://api.semanticscholar.org/CorpusID:254251061.

Xiaotian Bi, Deyang Wu, Daoxiong Xie, Huawei Ye, and Jinsong Zhao. Large-scale chemical process

causal discovery from big data with transformer-based deep learning. *Process Safety and Environmental Protection*, 173:163–177, 2023. ISSN 0957-5820. doi: https://doi.org/10.1016/j.psep.2023.03.017. URL https://www.sciencedirect.com/science/article/pii/S0957582023002100.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, January 2014. ISSN 1532-4435.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery, 12 2021.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Ralf Eggeling, Jussi Viinikka, Aleksis Vuoksenmaa, and Mikko Koivisto. On structure priors for learning bayesian networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1687–1695. PMLR, 2019.

Nir Friedman and Daphne Koller. Being bayesian about network structure. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, page 201–210, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607099.

David Heckerman, Christopher Meek, and Gregory Cooper. *A Bayesian Approach to Causal Discovery*, pages 1–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-33486-6. doi: 10.1007/3-540-33486-6_1. URL https://doi.org/10.1007/3-540-33486-6_1.

Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651. URL https://doi.org/10.1162/003465304323023651.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024. URL https://openreview.net/forum?id=5RBUTx75yr.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (TMLR)*, August 2024. Selected for presentation at ICLR 2025.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. Can large language models help experimental design for causal discovery?, 2025. URL https://arxiv.org/abs/2503.01139.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022. URL https://openreview.net/forum?id=LQQoJGw8JD1.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: differentiable bayesian structure learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence, 2025. URL https://arxiv.org/abs/2502.00290.

David Madigan and Jeremy York. Bayesian graphical models for discrete data. *International Statistical Review*, 63: 215–232, 1995.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet,

Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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

Marco Scutari, Phil Howell, David J Balding, and Ian Mackay. Multiple quantitative trait analysis using bayesian networks. *Genetics*, 198(1):129–137, 09 2014. ISSN 1943-2631. doi: 10.1534/genetics.114.165704. URL https://doi.org/10.1534/genetics.114.165704.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR, 2024.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35: 16261–16275, 2022.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *CoRR*, October 2023. URL https://www.microsoft.com/en-us/research/publication/causal-inference-using-llm-guided-discovery/.

Yujian Wen, Jielong Huang, Shuhui Guo, Yehezqel Elyahu, Alon Monsonego, Hai Zhang, Yanqing Ding, and Hao Zhu. Applying causal discovery to single-cell analyses using causalcell. *eLife*, 12:e81464, may 2023. ISSN 2050-084X. doi: 10.7554/eLife.81464. URL https://doi.org/10.7554/eLife.81464.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16–17):1873–1896, November 2008. ISSN 0004-3702. doi: 10.1016/j.artint.2008.08.001. URL https://doi.org/10.1016/j.artint.2008.08.001.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9492–9503, Red Hook, NY, USA, 2018. Curran Associates Inc.

Large Language Models as Tools to Improve Bayesian Causal Discovery (Supplementary Material)

Bruna Bazaluk[1] **Benjie Wang**[2] **Denis Deratani Mauá**[1] **Flavio S Correa da Silva**[1]

[1]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil
[2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA

A DATASETS

A.1 SACHS

Sachs [Sachs et al., 2005] is a famous real-world dataset about protein signaling, and one of the most used to test CD algorithms. It has 7467 samples of 11 gaussian variables.

The ground-truth graph, suggested by the original paper, can be seen in Figure 5. However, there is still a discussion among experts regarding some edges [Bavaria et al., 2013].

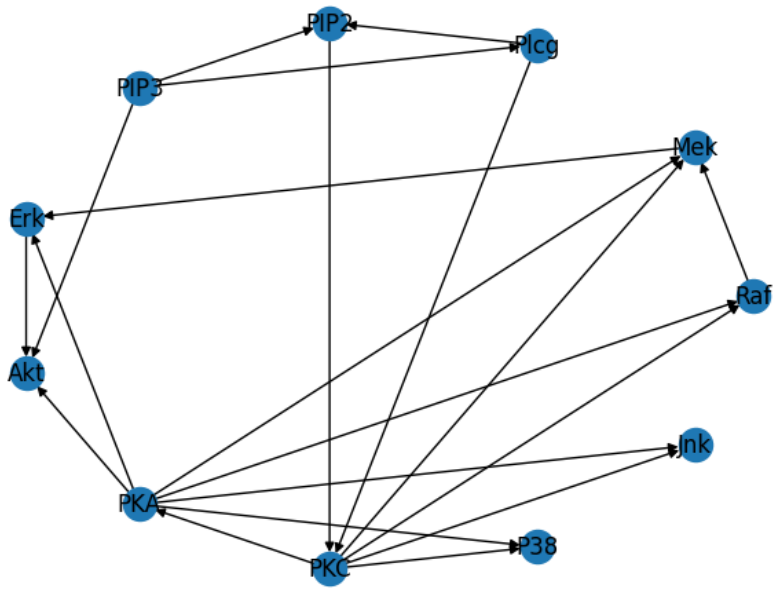

Figure 5: Sachs ground-truth graph.

A.2 MAGIC-NIAB

MAGIC-NIAB [Scutari et al., 2014] is a much smaller real-world dataset, with only 601 samples and with 44 variables. Among them, 7 are wheat genetic traits and the others are genes. For this work we considered only the 7 traits as nodes.

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

The ground-truth graph, suggested by the original paper, can be seen in Figure 6.

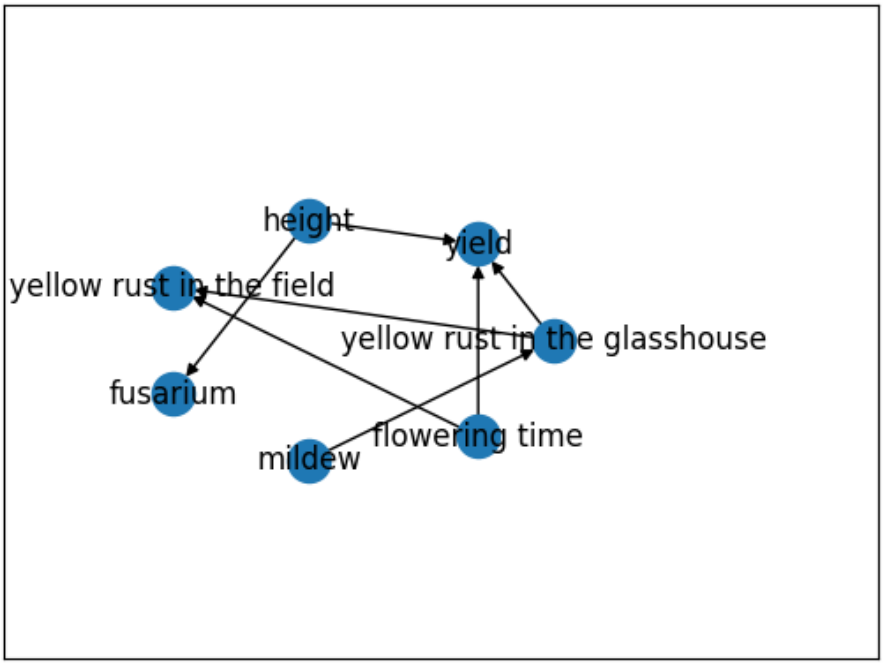

Figure 6: MAGIC-NIAB ground-truth graph.

A.3 SUMMER

Summer is a custom dataset built specifically for this work. It has 2001 samples of 7 variables that represent a common-sense knowledge about summertime. Firstly, we defined the 7 variables. Then, we drew the causal graph representing the causal relationships. This graph is shown in Figure 7. The graph was validated by a group of 15 people.

With the ground-truth graph in hand, we generate the Gaussian data using the following R code:

```
library(data.table)
library(ggplot2)
library(simDAG)

dag <- empty_dag()

dag <- dag +
  node("summer", type="rnorm", mean=0, sd=1) +
  node("sun", type="gaussian", parents=c("summer"), betas=c(runif(n=1, min
      =50, max=100)), intercept=runif(n=1, min=1, max=10), error=runif(n=1,
      min=0, max=5)) +
  node("vacation", type="gaussian", parents=c("summer", "sun"), betas=runif(n
      =2, min=50, max=100), intercept=runif(n=1, min=1, max=10), error=runif(
      n=1, min=0, max=5)) +
  node("beach", type="gaussian", parents=c("sun", "vacation"), betas=runif(n
      =2, min=50, max=100), intercept=runif(n=1, min=1, max=10), error=runif(
      n=1, min=0, max=5)) +
  node("sunburn", type="gaussian", parents=c("sun"), betas=c(runif(n=1, min
      =50, max=100)), intercept=runif(n=1, min=1, max=10), error=runif(n=1,
      min=0, max=5)) +
  node("drowning", type="gaussian", parents=c("beach"), betas=c(runif(n=1,
```

```
        min=50, max=100)), intercept=runif(n=1, min=1, max=10), error=runif(n
        =1, min=0, max=5)) +
    node("hospital", type="gaussian", parents=c("sunburn", "drowning"), betas=
        runif(n=2, min=50, max=100), intercept=runif(n=1, min=1, max=10), error
        =runif(n=1, min=0, max=5))

summary(dag)

set.seed(42)
sim_dat <- sim_from_dag(dag=dag, n_sim=2000)
```

The parameters were generated pseudo-randomly by R functions.

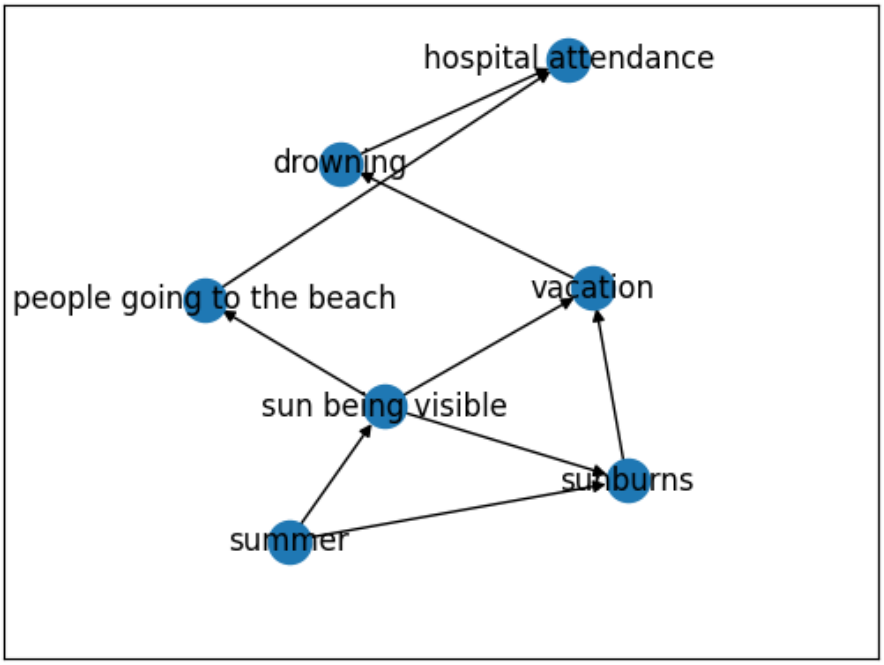

Figure 7: Summer ground-truth graph.

B ADJACENCY MATRICES HEATMAPS

Figures 8, 9 and 10 show Sachs, MAGIC-NIAB and Summer's true adjacency matrix, followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

Figures 11, 12 and 13 show Sachs, MAGIC-NIAB and Summer's priors: DeepSeek's on the left and GPT's on the right; followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

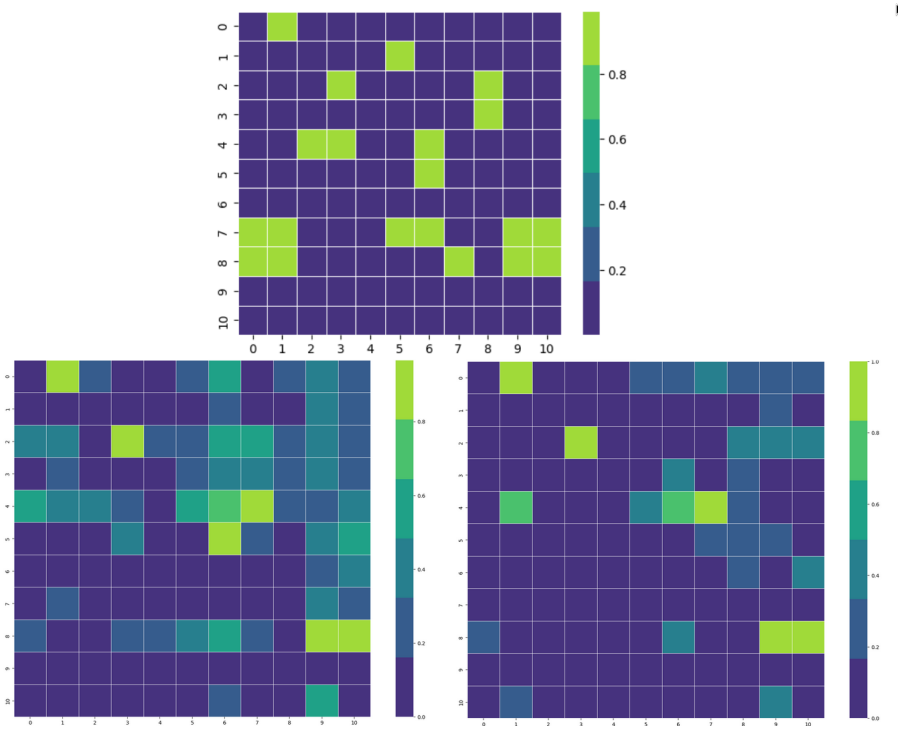

Figure 8: Sachs' true adjacency matrix, followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

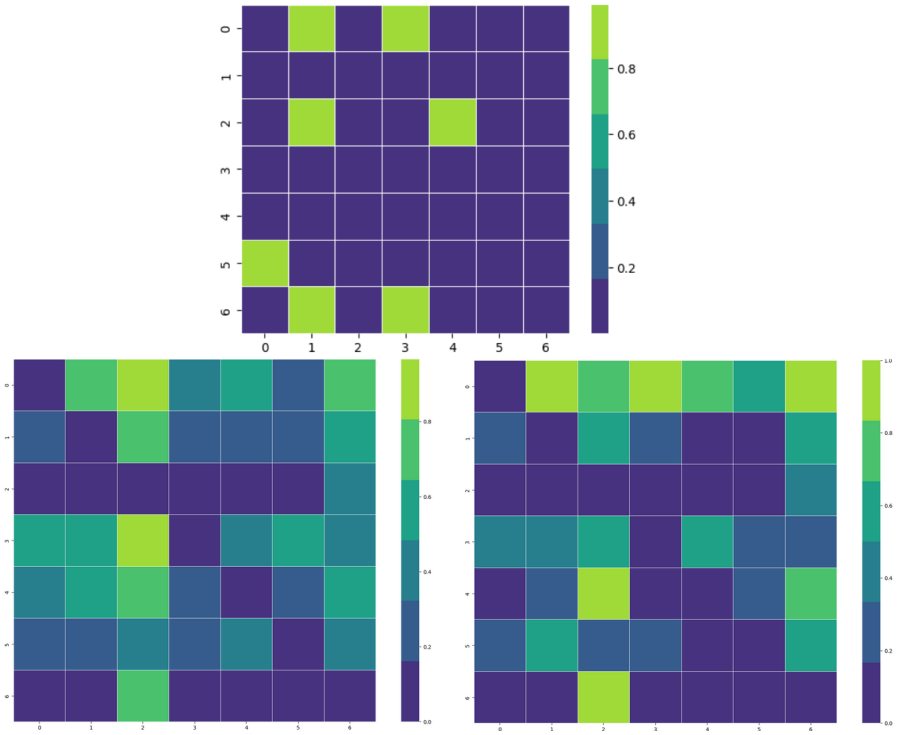

Figure 9: MAGIC-NIAB's true adjacency matrix, followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

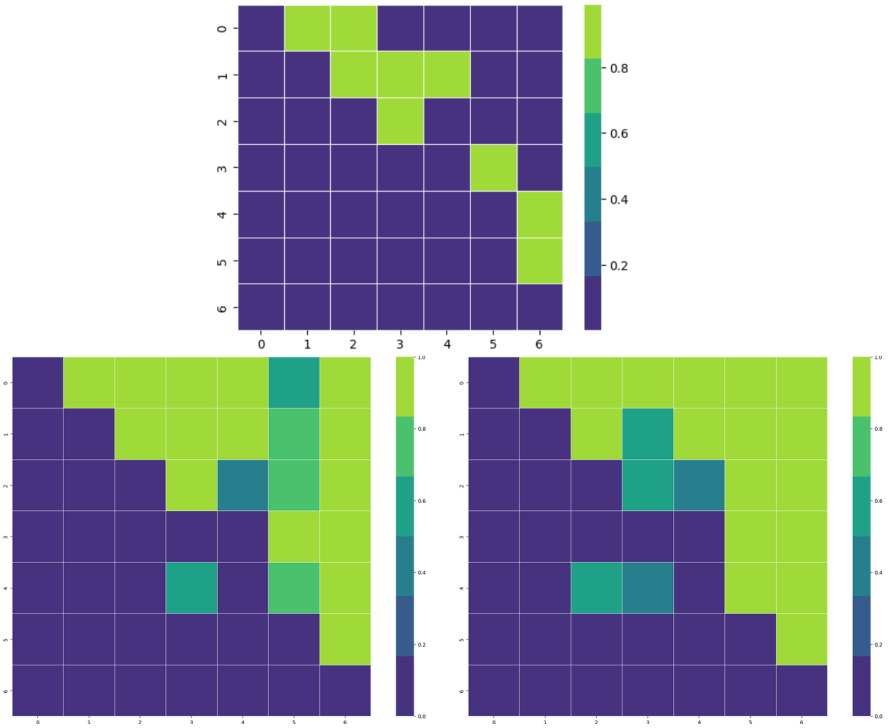

Figure 10: Summer's true adjacency matrix, followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

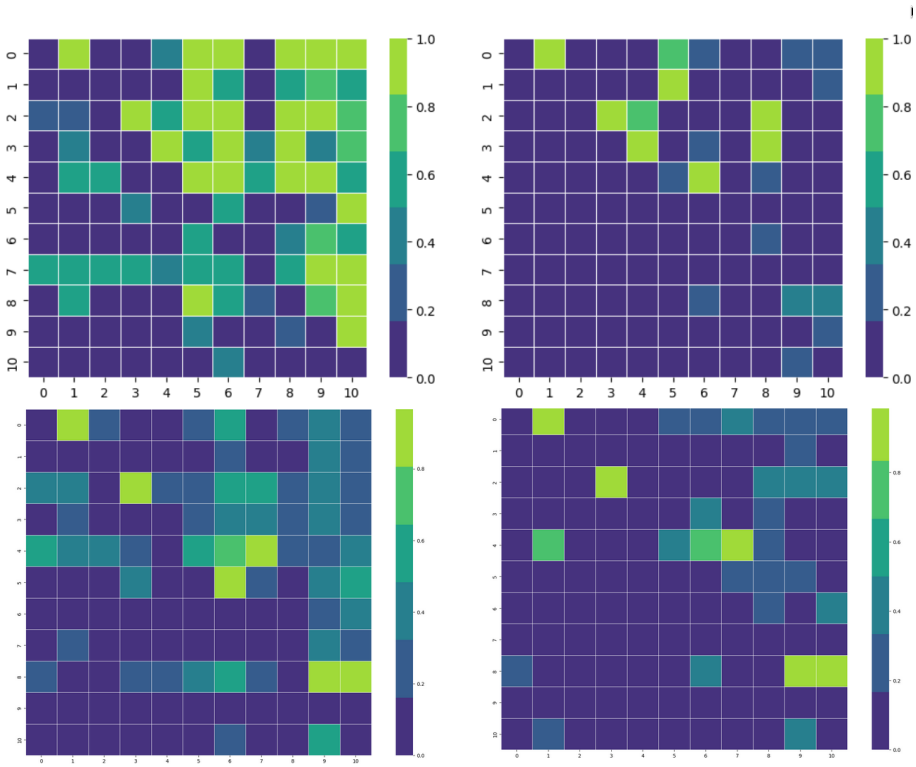

Figure 11: Sachs' priors: DeepSeek's on the left and GPT's on the right; followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

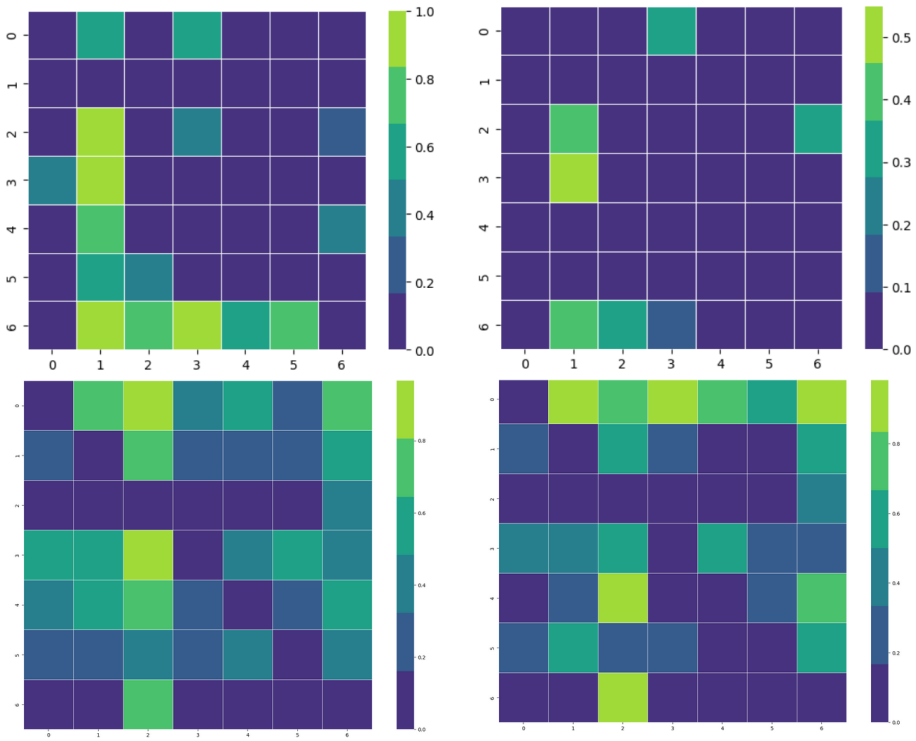

Figure 12: MAGIC-NIAB's priors: DeepSeek's on the left and GPT's on the right; followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.

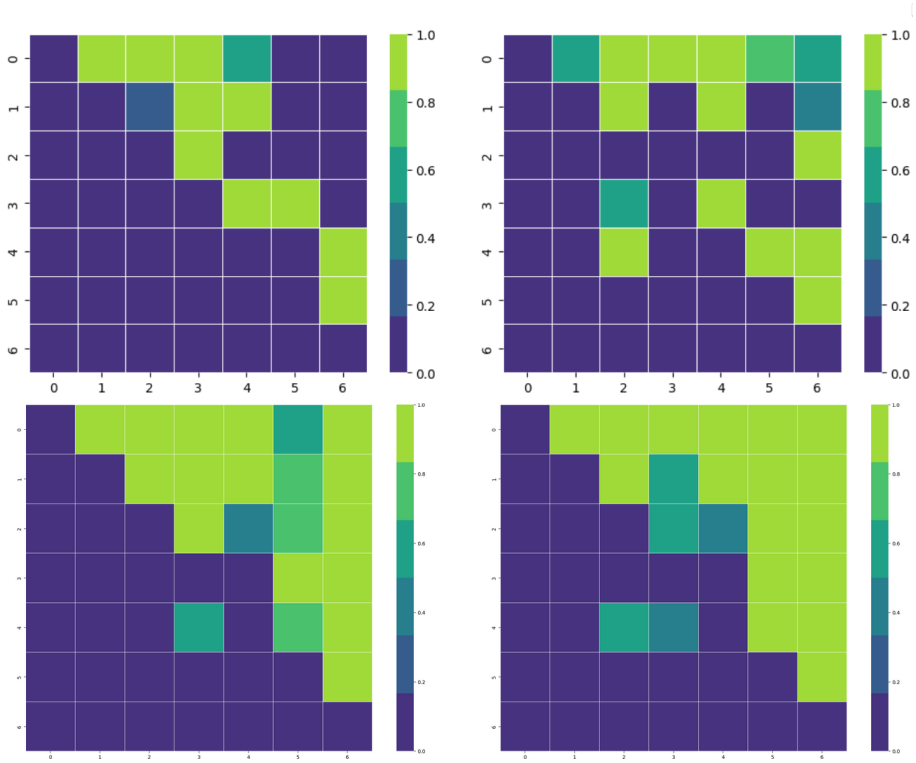

Figure 13: Summer's priors: DeepSeek's on the left and GPT's on the right; followed by DiBS+DeepSeek's posterior on the left and DiBS+GPT's posterior on the right.