Human-Aided Discovery of Ancestral Graphs

Tiago da Silva[1] **Eliezer de Souza da Silva**[1] **António Góis**[2] **Dominik Heider**[3]
Samuel Kaski[4,5] **Diego Mesquita**[1*] **Adèle Helena Ribeiro**[3*]
[1]School of Applied Mathematics, Getulio Vargas Foundation;
[2]Mila Quebec AI Institute, Université de Montréal;
[3]Institute of Medical Informatics, University of Münster;
[4]Department of Computer Science, Aalto University;
[5]Department of Computer Science, University of Manchester.

Abstract

In data-scarce situations, causal discovery (CD) algorithms often produce unreliable causal relationships that may conflict with expert knowledge, especially in the presence of latent confounders. Additionally, most CD methods lack adequate uncertainty quantification, hindering users' ability to evaluate and refine results. To address these issues, we present a fully probabilistic CD method referred to as Ancestral GFlowNets (AGFNs). In a nutshell, AGFNs sample ancestral graphs (AGs) proportionally to a score-based belief distribution, allowing users to assess the uncertainty of the discovered causal relationships. On top of that, we design an elicitation framework that enables the incorporation of human knowledge into the inference process via importance sampling. Notably, our approach naturally accommodates CD on data sets with latent confounding and potentially heterogeneous data types, a setting that has received little attention from the literature. Finally, experimental results with observational data show that our method effectively samples from distributions over AGs and significantly enhances inference quality with human aid.

1 Introduction

Causal discovery (CD) algorithms are essential to uncover complex cause-and-effect relationships in observational studies. When latent confounders are present, causal discovery is facilitated by encoding causal models as **Ancestral Graphs (AGs)**. AGs effectively encode ancestral (causal) relationships without explicitly representing unobserved variables [Richardson and Spirtes, 2002]. CD algorithms typically rely on observational data to infer the models most likely to have generated it, known as the Markov Equivalence Class (MEC). However, their reliability significantly decreases when data is scarce, as the inferred statistical relationships may not correspond to the true causal model. This discrepancy constitutes a violation of the *faithfulness* assumption [Zhang and Spirtes, 2016], which is especially pronounced in the presence of latent confounding. For example, constraint-based CD algorithms may rely on falsely identify independencies arising from insufficient statistical power, resulting in erroneous edge orientations [Zhang and Spirtes , 2008, Zhalama et al., 2017, Ng et al., 2021]. Similarly, score-based algorithms may identify optimal structures for the observed data but still misrepresent the true ground-truth MEC [Ogarrio et al., 2016].

This *extended abstract* presents a pragmatic approach to enhancing robustness, trustworthiness, and transparency in CD by incorporating uncertainty quantification and mechanisms for iterative improvement through human feedback. Further details of the method are available in the full paper.

Method. We propose a fully probabilistic CD framework called *Ancestral GFlowNet* (AGFN) that generates a data-driven distribution over AGs and iteratively updates it by actively engaging with a human expert. AGFN initially samples AGs based on a score function that measures goodness-of-fit on observational data, thereby encapsulating the epistemic uncertainty around the inference process.

[*]Shared last authorship.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

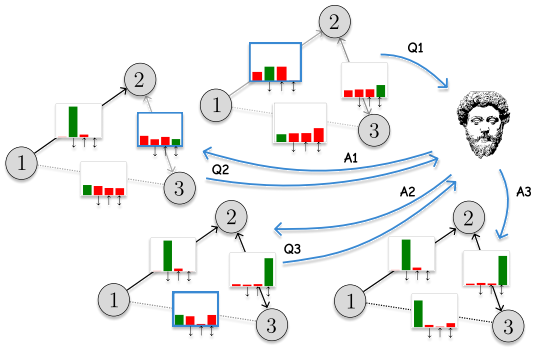

Figure 1: **Human-in-the-loop probabilistic CD.** We fit an AGFN using data-driven scores that quantify AGs fitness to the data. Then, refine it iteratively by i) querying (Q) experts on informative variable pairs and ii) updating the belief with potentially noisy feedback. Histograms on edges show marginals, with ground truth in green. The belief increasingly focuses on the true AG, $1 \rightarrow 2 \leftrightarrow 3$.

This approach ensures that the highest-scoring AGs are sampled more often while also allowing for the sampling of AGs that are less compliant with the (potentially unfaithful) data. Then, we leverage the AG samples to devise an active elicitation framework and a procedure for updating the AGFN distribution based on feedback from external experts.

While most probabilistic CD relies on Markov Chain Monte Carlo (MCMC) methods [e.g., Silva and Ghahramanir, 2009, Silva, 2013], AGFNs samples AGs using the formalism of *Generative Flow Networks* [GFlowNets; Bengio et al., 2021a,b], which are generative models that learn to sample from a distribution in proportion to a specified reward – in this case, defined by the score function. Notably, the score function is a hyperparameter of AGFN, enabling users to choose different functions without changing the overall method. Our *human-in-the-loop elicitation framework* allows AGFN to probe the user regarding the existence and nature (confounding/ancestral) of a maximally informative causal relation — subsequently updating our beliefs to incorporate the (potentially noisy) feedback in the process. Furthermore, we use importance sampling to update our initial beliefs with the human feedback, avoiding retraining GFlowNets repeatedly.

We conduct experiments using the BIC score for linear Gaussian causal models to validate our approach. Specifically, we evaluate: i) our ability to accurately sample from score-based beliefs over AGs; ii) how our samples compare to those from bootstrapped versions of state-of-the-art (SOTA) CD methods; and iii) the effectiveness of our active knowledge elicitation framework with simulated human input. Our results show that AGFN: i) accurately samples from our beliefs over AGs; ii) consistently includes AGs with low structural error among its top-scored samples; and iii) significantly enhances performance metrics (i.e., SHD and BIC) when incorporating human feedback.

In summary, our **contributions** are:

1. We introduce AGFN, the first score-based CD algorithm that integrates observational data with potentially noisy human feedback while offering uncertainty quantification and handling latent confounding.
2. We develop an active elicitation framework that enables AGFN to optimally interact with experts by sequentially selecting the most informative questions and effectively incorporating human feedback.
3. We devise an importance sampling scheme to update AGFN samples following expert feedback, eliminating the need to retrain AGFN for sampling from the updated belief;
4. We evaluate our method on various CD tasks, achieving competitive results with state-of-the-art (SOTA) methods while effectively refining quality through iterative human feedback.

2 Ancestral GFlowNets

AGFNs sample AGs using the formalism of *Generative Flow Networks* (GFlowNets). AGFN generates a distribution over AGs proportionally to a reward function, defined by a score-based measure such as the Bayesian Information Criterion (BIC) [Foygel and Drton, 2010].

Generative Flow Networks. GFlowNets sample structured objects by defining a trajectory of states $s \in \mathcal{S}$ guided by transition probabilities $\pi_F(s'|s)$. The forward transition probability, assigned to

each terminating state (an AG in our case), is proportional to the reward $R(s')$. The goal is to sample AGs proportionally to their reward. The flow-matching condition ensures that the flow entering any state equals the flow leaving it, allowing AGFNs to generate valid AGs efficiently. For AGs, the reward is based on a score function such as BIC:

$$R(\mathcal{G}) = \exp\left(\frac{\mu - U(\mathcal{G})}{\sigma}\right),$$

where μ and σ are constants ensuring numerical stability, and $U(\mathcal{G})$ is the BIC score of graph $\mathcal{G}$.

Score-Based Belief. AGFN samples graphs proportionally to the reward based on the chosen score function, allowing flexibility in handling different data types. We use the extended BIC score for linear Gaussian models:

$$U(\mathcal{G}) = -2\log L(\mathcal{G}) + |\mathbf{E}|\log N,$$

where $L(\mathcal{G})$ is the likelihood and $|\mathbf{E}|$ is the number of edges in the graph.

3 Human-in-the-Loop Causal Discovery

AGFNs integrate human expertise through a *human-in-the-loop (HITL)* framework that refines the belief distribution over AGs by querying experts about specific relationships between pairs of variables. As probing these experts might be a potentially costly operation that may require human intervention, we sensibly select a relationship r that maximally reduces our uncertainty over the structure of the true causal diagram at each iteration of the HITL pipeline. More specifically, let $p_\theta(\mathcal{G})$ be AGFN's sampling distribution and $\mathbf{f}_K = (f_{r_k})_{k=1}^K$ be the received feedbacks regarding the relations $\{r_k\}_{k=1}^K$. In this context, we select a relation r that maximizes the cross-entropy-based acquisition function

$$a_{K+1}(r) = -\mathbb{E}_{f_r \sim q(\cdot|\mathbf{f}_K)}\left[\mathbf{H}(q(\mathcal{G}|\mathbf{f}_K \cup \{f_r\}), q(\mathcal{G}|\mathbf{f}_K))\right]$$

in which $q(\mathcal{G}|\mathbf{f}_K \cup \{f_r\})$ (see below) is the posterior belief and $\mathbf{H}$ is the cross-entropy. Intuitively, maximizing a_{K+1} corresponds to minimizing an upper bound of the entropy (uncertainty) of $q(\mathcal{G}|\mathbf{f}_K \cup f_r)$.

Knowledge elicitation. To allow for the incorporation of potentially noisy feedback into AGFN, we define a probabilistic model over the expert's responses to our queries. For this, we model a noisy knowledge on a relation $r = \{V_i, V_j\}$ between the variables V_i and V_j with $i < j$ as a categorical random variable ω_r indexing the tuple $(\emptyset, \rightarrow, \leftarrow, \nleftrightarrow)$, e.g., $\omega_r = 2$ indicates that $V_i \rightarrow V_j$ is the true relationship. Then, the expert's feedback f_r on r is a noisy realization of ω_r under the expert's model,

$$\omega_r \sim \text{Cat}(\boldsymbol{\rho}_r), \tag{1}$$

$$f_r|\omega_r \sim \text{Cat}\left(\delta_{\omega_r} \cdot \pi + (\mathbf{1} - \delta_{\omega_r}) \cdot \left(\frac{1-\pi}{3}\right)\right), \tag{2}$$

$$p(\omega_{r_k}|f_{r_k}) = \text{Cat}\left(\frac{\boldsymbol{\rho}_r}{\eta_r} \odot \left(\pi \cdot \delta_{f_r} + \left(\frac{1-\pi}{3}\right) \cdot (\mathbf{1} - \delta_{f_r})\right)\right), \tag{3}$$

in which $\boldsymbol{\rho}_r = (\rho_{r,1}, \rho_{r,2}, \rho_{r,3}, \rho_{r,4})$ represents our prior beliefs about the relation, $\pi \in [0,1]$ reflects the feedback's reliability, and δ_k is the k-the line of the identity matrix in $\mathbb{R}^4$. Heuristically, f_r matches ω_r with probability π and is otherwise uniformly distributed among the incorrect alternatives.

Human-driven inference refinement. Given a set $\mathbf{f}_K$ of expert-provided feedbacks, we define a posterior belief $q(\cdot|\mathbf{f}_K)$ over the AGs as the product between AGFN's sampling distribution p_θ and the expert's model p in Equation (3) [Hinton, 2002]. To approximate the expectation of a test function h on the space of AGs (e.g., an indicator of a causal relationship) under q, we utilize an importance sampling estimator having p_θ as proposal. More specifically, we sample i.i.d. $\{\mathcal{G}_t\}_{t=1}^T \sim p_\theta(\cdot)$ and let

$$q(\mathcal{G}|\mathbf{f}_K) \propto p_\theta(\mathcal{G}) \prod\nolimits_{1 \le k \le K} p(\omega_{r_k}|f_{r_k}) \quad \text{and} \quad \mathbb{E}_q[h(\mathcal{G})] \approx \sum_{t=1}^{T} \frac{q(\mathcal{G}^{(t)}|\mathbf{f}_K)}{p_\theta(\mathcal{G}^{(t)})} h(\mathcal{G}^{(t)}), \tag{4}$$

4 Experiments

Our experiments evaluate AGFN's ability to model distributions over AGs, compare its performance against SOTA CD algorithms, and demonstrate the effectiveness of incorporating human feedback.
AGFN performs comparably to CD baselines. We generate 10 datasets with 500 independent samples from randomly parametrized linear Gaussian SCMs for the causal diagrams, including:

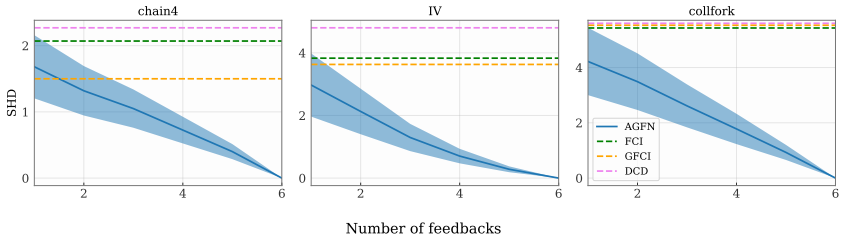

Figure 2: **Human-aided AGFN outperforms CD baselines after a single feedback** across the considered datasets. The HITL pipeline results (blue) show the SHD averaged over 30 HITL simulations.

(i) `chain4` ($W \to X \to Y \to Z$), (ii) `IV` ($W \to X \to Y; X \leftarrow Z \to Y$), and (iii) `collfork` ($X \to Z \leftarrow W \to Y; X \dashleftarrow\dashrightarrow W; Z \dashleftarrow\dashrightarrow Y$), each representing increasingly complex structures with latent confounding. AGFN is compared against the **baselines** FCI [Spirtes et al., 2001], GFCI [Ogarrio et al., 2016], ACI [Magliacane et al., 2016], DCD [Bhattacharya et al., 2021], and N-ADMG [Ashman et al., 2023], comprising constraint-based, score-based, and hybrid approaches to CD. Table 1 shows the top-scoring samples sampled by AGFN achieve a SHD lower than or comparable to the baselines'.

	chain4	IV	collfork
FCI	2.07 ± 2.00	3.83 ± 2.90	5.43 ± 1.87
GFCI	$\mathbf{1.50} \pm 1.63$	3.63 ± 3.16	5.53 ± 2.11
ACI	5.77 ± 2.66	8.58 ± 2.16	8.02 ± 2.18
DCD	2.27 ± 1.46	4.80 ± 2.17	5.60 ± 2.13
N-ADMG	4.38 ± 0.81	6.08 ± 1.77	6.87 ± 0.93
AGFN	2.00 ± 1.55	$\mathbf{3.50} \pm 3.29$	$\mathbf{4.90} \pm 2.70$

Table 1: Structural Hamming Distance (SHD, $\downarrow$) comparison across methods, with lower SHD indicating better performance. Notably, AGFN performs comparably or better than the baselines.

Inference quality drastically improves with human knowledge. To assess the improvements enacted by the incorporation of an expert's feedback into the inference process, Figure 2 exhibits the expected SHD of AGFN's top-scoring samples during each step of the iterative refinement procedure. Strikingly, AGFN outperforms all baselines after issuing a single feedback from the expert. This highlights the effectiveness of our framework for HITL CD. Also, note that AGFN is the only method that can be seamlessly integrated into a HITL pipeline among the considered baselines.

Comparison with DAG-GFlowNet. To illustrate the importance of accounting for latent confounding in CD problems, we compare AGFN against DAG-GFlowNets [Deleu et al., 2022], which assumes causal sufficiency. As expected, Table 2 shows AGFN is more accurate than DAG-GFlowNet when applied to latently confounded datasets. Obviously, however, we can easily constraint the search space of AGFN to DAGs (instead of AGs) by making out the addition of bidirected edges from its generative process if we are confident that latent confounding is not an issue.

	DAG-GFlowNet	AGFN
Causal diagram 1 ($A \to B \dashleftarrow\dashrightarrow C \leftarrow D$)	3.80	**1.67**
Causal diagram 2 ($X \to A \to Y$ and $A \dashleftarrow\dashrightarrow B \dashleftarrow\dashrightarrow Y$)	6.40	**5.27**
Causal diagram 3 ($A \dashleftarrow\dashrightarrow B \dashleftarrow\dashrightarrow C \dashleftarrow\dashrightarrow D$ and $A \dashleftarrow\dashrightarrow D$)	8.22	**7.87**

Table 2: SHD ($\downarrow$) comparison: AGFN vs DAG-GFlowNet under latent confounding.

5 Conclusion

We introduced AGFN, a novel and robust framework for probabilistic CD in the presence of latent confounding that effectively incorporates potentially noisy human feedback through an optimal elicitation strategy. In short, AGFN samples AGs based on a score function defining a belief distribution that encapsulates our epistemic uncertainty around the inference process. Albeit our experiments were constrained to Foygel and Drton [2010]'s extended BIC score, this is merely a design choice that does not restrain AGFN's applicability. For discrete data, for instance, Drton and Richardson [2008]'s score could be considered. Importantly, we empirically demonstrated that AGFN outperforms baseline methods in terms of SHD and BIC after receiving a single feedback from an expert. Overall, this work underlines the potential of human-driven probabilistic methods for CD problems.

Acknowledgements

Tiago da Silva, Eliezer Silva and Diego Mesquita acknowledge the support of Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro FAPERJ (SEI-260003/000709/2023), São Paulo Research Foundation (FAPESP, grant 2023/00815-6), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grant 404336/2023-0). Adèle Ribeiro and Dominik Heider were supported by the LOEWE program of the State of Hesse (Germany) in the Diffusible Signals research cluster and by the German Federal Ministry of Education and Research (BMBF) [031L0267A] (Deep Insight). António Góis acknowledges the support by Samsung Electronics Co., Ltd. Samuel Kaski was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI), EU Horizon 2020 (European Network of AI Excellence Centres ELISE, grant agreement 951847), UKRI Turing AI World-Leading Researcher Fellowship (EP/W002973/1).

We also acknowledge the computational resources provided i) by the Aalto Science-IT Project from Computer Science IT, and ii) FGV TIC.

References

M. Ashman, C. Ma, A. Hilmkil, J. Jennings, and C. Zhang. Causal reasoning in the presence of latent confounders via neural ADMG learning. In *International Conference on Learning Representations (ICLR)*, 2023.

E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Y. Bengio, T. Deleu, E. J. Hu, S. Lahlou, M. Tiwari, and E. Bengio. GFlowNet Foundations. *arXiv preprint*, 2021b.

R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitser. Differentiable causal discovery under unmeasured confounding. In *Artificial Intelligence and Statistics (AISTATS)*, 2021.

T. Deleu, A. Góis, C. C. Emezue, M. Rankawat, S. Lacoste-Julien, S. Bauer, and Y. Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2022.

M. Drton and T. S. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2008.

R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing (NeurIPS)*, 2010.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002.

S. Magliacane, T. Claassen, and J. M. Mooij. Ancestral causal inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

I. Ng, Y. Zheng, J. Zhang, and K. Zhang. Reliable causal discovery with improved exact search and weaker assumptions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Probabilistic Graphical Models (PGM)*, 2016.

T. Richardson and P. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 2002.

R. Silva. A MCMC approach for learning the structure of gaussian acyclic directed mixed graphs. In *Statistical Models for Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 343–351. Springer, 2013.

R. Silva and Z. Ghahramanir. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research (JMLR)*, 2009.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2001.

Zhalama, J. Zhang, and W. Mayer. Weakening faithfulness: some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics*, 3(2):93–104, 2017. ISSN 2364-4168. doi: 10.1007/s41060-016-0033-y.

J. Zhang and P. Spirtes. The three faces of faithfulness. *Synthese*, 193(4):1011–1027, 2016. ISSN 1573-0964. doi: 10.1007/s11229-015-0673-9.

J. Zhang and P. Spirtes Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, Jun 2008.