IGDA: INTERACTIVE GRAPH DISCOVERY THROUGH LARGE LANGUAGE MODEL AGENTS

Alex Havrilla*
School of Mathematics
Georgia Institute of Technology
Atlanta, GA, 30308, USA
ahavrilla3@gatech.edu

David Alvarez-Melis
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA
dam@seas.harvard.edu

Nicolo Fusi
Microsoft Research
Cambridge, MA 02142
fusi@microsoft.com

ABSTRACT

Large language models (**LLMs**) have emerged as a powerful method for discovery. Instead of utilizing numerical data, LLMs utilize associated variable *semantic metadata* to predict variable relationships. Simultaneously, LLMs demonstrate impressive abilities to act as black-box optimizers when given an objective f and sequence of trials. We study LLMs at the intersection of these two capabilities by applying LLMs to the task of *interactive graph discovery*: given a ground truth graph G^* capturing variable relationships and a budget of I edge experiments over R rounds, minimize the distance between the predicted graph $\hat{G}_R$ and G^* at the end of the R-th round. To solve this task we propose **IGDA**, a LLM-based pipeline incorporating two key components: 1) an LLM uncertainty-driven method for edge experiment selection 2) a local graph update strategy utilizing binary feedback from experiments to improve predictions for unselected neighboring edges. Experiments on eight different real-world graphs show our approach often outperforms all baselines including a state-of-the-art numerical method for interactive graph discovery. Further, we conduct a rigorous series of ablations dissecting the impact of each pipeline component. Finally, to assess the impact of memorization, we apply our interactive graph discovery strategy to a complex, new (as of July 2024) causal graph on protein transcription factors, finding strong performance in a setting where memorization is impossible. Overall, our results show IGDA to be a powerful method for graph discovery complementary to existing numerically driven approaches.

1 INTRODUCTION

Given a set of variables $X_1, ..., X_n$, the *graph discovery* task involves finding a graph G^* on the nodes $X_1, ..., X_n$ whose edges capture causal relationships between the *parent* (source) and *child* (target). Often, observational data can be collected for the variables $X_1, ..., X_n$. This data can then be used to predict an initial graph G_0 using statistical causal discovery techniques (Spirtes & Zhang, 2016). Recently, large language models (**LLMs**) have emerged as a competitive alternative method for predicting causal graphs (Kıcıman et al., 2024; Abdulaal et al., 2024; Chen et al., 2024). Unlike pre-existing statistical methods, LLMs require no observational data (Kıcıman et al., 2024), instead relying purely on semantic metadata such as variable names and descriptions. Another related line a work (Yang et al., 2024) investigates the abilities of LLMs to act as in-context black-box optimizers. Given an objective function f and an evaluation budget B, the LLM is tasked with finding a maximizer x^* of f by sequentially proposing queries $\{x_i\}_{i=1}^B$ and observing their

*Work done during internship at Microsoft Research.

associated values $\{f(x_i)\}_{i=1}^{B}$. Taken together, these directions suggest a powerful new application of LLMs: *interactive graph discovery*.

Given an initial predicted graph $\hat{G}_0$ and a series of experiment rounds $1, ..., R$, the interactive graph discovery problem involves minimizing some distance $d(\hat{G}_k, G^*)$ between the predicted graph $\hat{G}_k$ at round k and the true graph G^* (unknown to the learner) through a sequence of targeted experiments on edges $e = (X, Y)$ testing the effect of the parent variable X on the child variable Y. The edge experiment operation is kept purposefully abstract, requiring only that binary feedback be given indicating the presence or absence of an edge. In practice this operation can be implemented via any number of experimental procedures (e.g. via hard interventions in the formal causal sense (Pearl, 2009) or empirical methods such as randomized controlled trials (Sibbald & Roland, 1998)). The IGD problem setup captures the process researchers go through everyday when designing and prioritizing experiments, guided by their prior experience, to study numerous potential relationships between any number of variables.

The interactive graph discovery problem requires the agent to solve two key sub-tasks:

1. **Experiment selection**: Selecting which edges (X_i, X_j) to target for experimentation in the next round.
2. **Graph updates**: Updating the predicted graph from $\hat{G}_{k-1}$ to $\hat{G}_k$ given binary feedback based on the outcome of the previous experiments.

We propose to solve this task with the Interactive Graph Discovery Agent (**IGDA**): a novel LLM agent uncertainty-driven approach as an alternative to existing statistical methods (Olko et al., 2024; Scherrer et al., 2022). While statistical models can work well in some settings, they crucially rely on the abundance of domain specific observational and interventional numerical data. For many problems, such data might be hard or impossible acquire. LLMs, however, potentially contain relevant latent knowledge derived from vast amounts of variable semantic metadata contained in their pre-training or internet corpora. Further, we find that, via a combination of broad background knowledge and reasoning abilities, advanced LLMs (Grattafiori et al., 2024) are capable of updating their predictions and confidences when presented with experimental feedback revealing unexpected relationships between a subset of edges. This makes LLM based approaches a powerful alternative to statistical methods when numerical data is not available.

In particular, IGDA predicts and maintains uncertainty estimates for each unknown edge $e \in \hat{G}_k$. Edges are then selected for experimentation by prioritizing those with the highest uncertainty. When feedback is received on the selected edges, pairwise-local updates on both edge predictions and uncertainty estimates are performed for each edge in $\hat{G}$ sharing a parent or child variable with an experimented edge. This process continues for R rounds with I edges selected for experimentation each round. We benchmark IGDA on eight real world graphs, finding uncertainty driven selection with local updates outperforms baselines. In summary, we make the following contributions:

- The interactive graph discovery problem as a novel setting for evaluating LLM capabilities.
- LLM-based uncertainty-guided edge experiment selection as a policy for prioritizing edge experimentation.
- A local update strategy for robustly updating the predicted graph G_k with binary experiment feedback.
- Ablations rigorously evaluating the contribution of each pipeline component and other discovery strategies.

2 Background and Related Work

Causal Discovery and LLMs. The causal discovery task involves learning causal relationships from observed empirical data (Peters et al., 2017; Spirtes & Zhang, 2016). Many proposed algorithms exist (Spirtes et al., 1993; Yu et al., 2019; Nauta et al., 2019; Zheng et al., 2018; Chickering, 2002) attempting to solve the causal discovery problem. However, these methods are known to struggle on real world graphs where observations are noisy or common structural assumptions are violated (Chevalley et al., 2023; Tu et al., 2019).

Recently, LLMs have emerged as an alternative approach to causal discovery (Kıcıman et al., 2024; Abdulaal et al., 2024; Vashishtha et al., 2023; Li et al., 2024; Lampinen et al., 2023). Kıcıman et al. (2024) first investigated the capability of LLMs to act as zero-shot causal discovery agents using only semantic information and pairwise prompting on each variable pair. Follow-up work (Abdulaal et al., 2024) further improves LLM predictions with observational data by selecting for predictions which maximize data likelihood. Vashishtha et al. (2023) utilize *triplet prompting* to prevent cycles when the causal graph is acyclic. They show only a topological ordering on variables is required for many common causal reasoning tasks (Chu et al., 2023). Other works (Zhou et al., 2024; Chen et al., 2024) benchmark LLMs across a range of causality related tasks including causal discovery and causal inference confirming that LLMs struggle with integrating numerical data.

Another line of work more related to our proposed interactive causal discovery problem studies how to incorporate background knowledge into causal discovery algorithms (Meek, 2013). Define a set of *background knowledge* as the tuple $\mathcal{K} = (F, R)$, where F specifies a set of "forbidden" graph edges and R specifies a set of "required" graph edges. Meek (2013) presents an algorithm for constructing a causal graph consistent with $\mathcal{K}$ by leveraging an assumed structural directed acyclic graph (DAG) property. Building on Meek (2013), Chickering (2002) proposes a greedy search algorithm that performs well in practice.

Most related are statistical methods from the causal discovery literature which aim to efficiently choose a sequence of interventions to discover causal structure (Scherrer et al., 2022; Olko et al., 2024). In particular, Gradient based Interventional Targeting (**GIT**) (Olko et al., 2024) utilizes existing neural causal discovery methods (Lippe et al., 2022) to learn a distribution over possible graph structures and variable assignments. For each round of intervention, GIT prioritizes variables whose simulated interventional distribution have large gradient with respect to the structural training loss.

In contrast to these works, our proposed algorithm utilizes LLMs to reason about the semantic/physical, as opposed to formal/structural, relationships between variables and edges in causal graphs. For this reason we are not required to make any structural assumptions on an underlying DAG, as is common in the causal discovery literature. This is desirable as in practice many real-world causal graphs are cyclic and poorly structured (Zhu et al., 2024; Huang et al., 2021). Additionally our method does not rely on observational or interventional data for real world graphs which may be expensive to acquire but crucial for good performance with statistical methods.

LLMs as Optimizers. Another growing line of work utilizes LLMs as black-box optimizers (Yang et al., 2024; Roohani et al., 2024). Yang et al. (2024) introduce the notion of an LLM as a generic optimizer and use it to optimize performance objectives stemming from a range of tasks including linear regression and mathematical word problems (Cobbe et al., 2021). Other works (Madaan et al., 2023; Havrilla et al., 2024) examine the self-refinement capabilities of LLMs where the LLM must reason and self-improve on earlier responses. A growing number of papers apply LLMs to optimal experiment design and discovery (Roohani et al., 2024; AI4Science & Quantum, 2023; Gao et al., 2024; Majumder et al., 2024; Jansen et al., 2024). Roohani et al. (2024) apply LLMs to gene discovery tasks which aim to find highly-influential parent genes affecting the regulation of a downstream target gene. Majumder et al. (2024); Jansen et al. (2024) both present benchmarks evaluating the ability of LLMs to perform real-world and synthesized discovery tasks.

3 METHOD

Setup. As input we are given a set of variables $X_1, ..., X_n$ with associated metadata including variable names and variable descriptions. We use the notation $Y \to X$ to indicate when variable Y has a direct effect on variable X and the set of parents of a variable X as $Pa(X) = \{X_i : X_i \to X\}$. We can then consider the directed ground truth graph $G^* = \{(X_i, X_j) : X_i \in Pa(X_j)\}$ with unlabeled and unweighted edges. The only assumed graph structure is simplicity i.e. no self-edges or multi-edges. No additional structure on the graph (such as acyclicity) is assumed. We can frame the prediction of G^* as an edge-wise binary classification problem over the complete graph K_n, where an edge (X_i, X_j) has the label $l_{ij} = 1$ if $X_i \to X_j$ and $l_{ij} = 0$ otherwise. G^* can then be written as a collection of ground truth labelings $G^* = \{(X_i, X_j, l_{ij}) : 1 \leq i \neq j \leq n\}$.

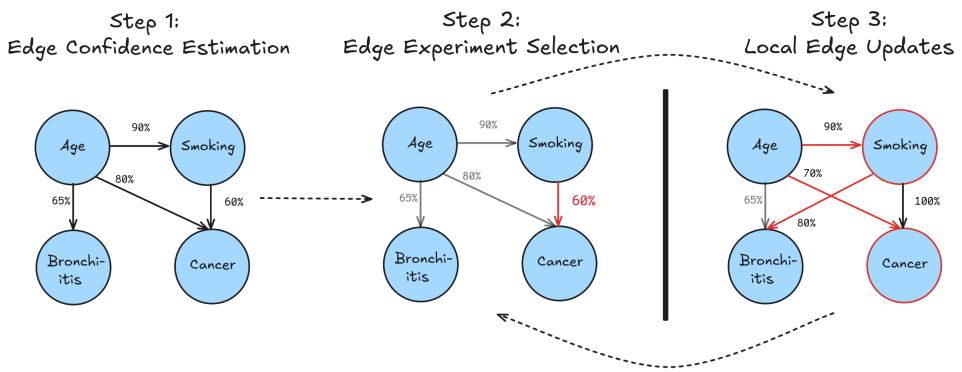

Figure 1: Diagram of the interactive graph discovery process through LLMs. The process begins by predicting edges and confidences for each edge. Interactive discovery then proceeds by selecting the most uncertain edges for experimentation. The LLM then updates its predictions and confidences for edges adjacent to the selected edge. Note: only edges predicted as present are shown.

The *interactive graph discovery* task then aims to learn G^* by interacting with the discovery environment via *experiments* on each edge (X_i, X_j). We define an *experiment* on an edge (X_i, X_j) as an operation revealing the ground truth label $l_{i,j}$. This experiment operation is purposefully kept abstract for generality and could correspond to any number of real-world experimental experimental strategies including formal do operations (Pearl, 2009) or empirical randomized control trials (Sibbald & Roland, 1998). Interactive graph discovery then proceeds in two phases:

> **Phase 1 (Zero-shot prediction):** Produce an initial graph prediction $\hat{G}_0$ using available variables $X_1, ..., X_n$ plus semantic metadata.
>
> **Phase 2 (Interactive Discovery):** Over a series of R rounds, propose I edge experiments on (X_i, X_j) each round and receive binary feedback on l_{ij}. Use this to produce an updated prediction $\hat{G}_{r-1} \to \hat{G}_r$

We evaluate the accuracy of a prediction $\hat{G}$ using the F1 objective, i.e.

$$F1(G^*, \hat{G}) = \frac{2 \cdot \texttt{Precision}_{\hat{G}} \cdot \texttt{Recall}_{\hat{G}}}{\texttt{Precision}_{\hat{G}} + \texttt{Recall}_{\hat{G}}}$$

where $\texttt{Precision}_{\hat{G}}$ and $\texttt{Recall}_{\hat{G}}$ are computed with the label predictions $(X_i, X_j, \hat{l}_{ij}) \in \hat{G}$ and l_{ij} as ground truth. The goal of the interactive discovery process is then to maximize $F1(G^*, \hat{G}_R)$.

Method. Our proposed method IGDA begins by generating a zero-shot graph prediction $\hat{G}_0$. A prediction for each variable pair (X_i, X_j), $1 \leq i \neq j \leq n$, is generated by prompting an LLM to reason about $X_i \to X_j$ in a manner similar to the pairwise-prompting strategy utilized in Kıcıman et al. (2024). In addition, we prompt the LLM to reason about its confidence in the prediction and output a confidence score from 1 - 100. Section B shows the exact prompt used. To obtain a reliable confidence estimate we sample the LLM $K = 16$ times. We denote the initial confidence for (X_i, X_j) as c^0_{ij} and set it to be the (signed) average over $K = 16$ output confidences. The initial edge label l^0_{ij} is then taken as the boolean $l^0_{ij} = \mathbf{1}_{c^0_{ij} \geq 0}$. This gives us the initial prediction $\hat{G}_0$.

Next, in each experimentation round $r \leq R$, we sort the confidence scores $\{c^r_{ij} : 1 \leq i, j \leq n\}$ by absolute value and experiment on the I edges with the lowest absolute confidence (and highest uncertainty). This reveals the ground truth labels l_{ij} for for each experimented edge (X_i, X_j). Using this feedback, we update the predicted edge labels for experimented edges to $l^{r+1}_{ij} = l_{ij}$ and the confidences to $c^{r+1}_{ij} = 100$. Additionally, we prompt the LLM, conditioned on the ground truth label l_{ij}, to update its prediction and confidence for each edge (X_i, X_k) or (X_l, X_j), $1 \leq k, l \leq n$ which shares a node with (X_i, X_j) and has absolute confidence less than 100. We call each update to an edge (X_l, X_k) a *local update*. It may be that an edge (X_l, X_k) is adjacent to multiple experimented

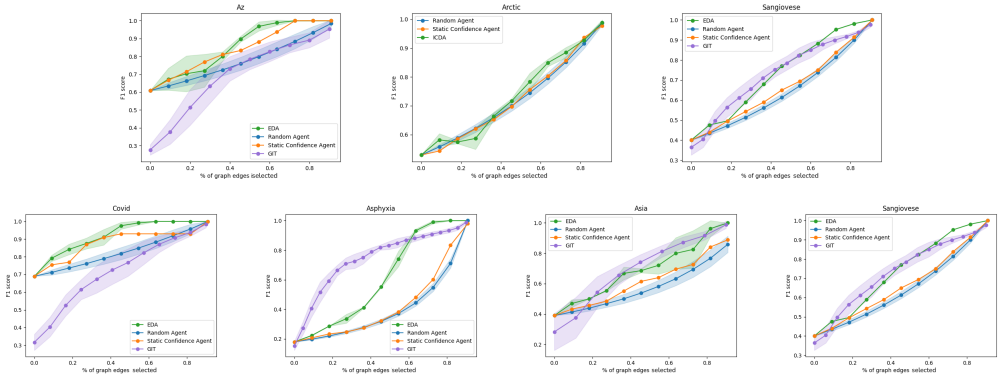

Figure 2: Results on real world graphs showing F1 score of the predicted graph against percentage of edges in the graph selected. IGDA almost always outperforms both the random baseline and static selection via uncertainty. Note: static confidence selection without local updates is deterministic and thus has no confidence intervals. Additionally, GIT is not reported on the Arctic graph because the grpah is cyclic.

edges $(X_{i_1}, X_{j_1}), (X_{i_2}, X_{j_2})$ in a single round and thus receives multiple local updates. To manage these cases we set the next confidence c_{lk}^{r+1} to the (signed) average of all individual local updates to c_{lk}^{r}. Then we set $l_{lk}^{r+1} = \mathbf{1}_{c_{lk} \geq 0}$ as before. This continues until the final round R is reached.

We call the complete discovery pipeline the *Interactive Graph Discovery Agent* (IGDA). A diagram of the full pipeline is shown in Figure 1. We report all prompts in sec:prompts.

4 RESULTS

We evaluate our approach on seven real-world graphs. The graphs range in size from 8 to 30 nodes (variables) and vary widely in structure (some are acyclic while others are cyclic). Details for each graph can be found in Appendix C. To produce initial zero-shot graph predictions $\hat{G}_0$ for all graphs we utilize pairwise causal prompting as in Kıcıman et al. (2024) with `Meta-Llama-3-70B-Instruct` (Grattafiori et al., 2024) as the base LLM. We chose `Meta-Llama-3-70B-Instruct` as at the time of our experiments it was the best open-source model with advanced reasoning and instruction following capabilities For the interactive discovery phase we then initialize all methods using $\hat{G}_0$. We compare our method against several baselines:

Random selection: Starting from $\hat{G}_0$ we randomly select edges for experimentation. After receiving binary feedback we update incorrect predictions on experiment edges for the next round. We do not allow edges to be selected for experimentation twice.

Static confidence selection: We select edges for experimentation based on the initial confidence scores c_{ij}. No updates are performed beyond fixing incorrect predictions in the experimentation set.

Gradient-based Intervention Targeting (GIT): We adapt the statistical GIT method (Olko et al., 2024) by selecting the node at each round which has a) not already been selected and b) has the largest loss gradient under a neural causal model (Lippe et al., 2022) trained with all available observational and interventional training data. We initially train the model with 5000 observational datapoints sampled from the ground-truth graph. 100 additional interventional datapoints on the experiment node are sampled from the ground-truth graph and added to the training set after each round of experimentation.

`Meta-Llama-3-70B-Instruct` is used as the base LLM when applicable. To assess performance, we plot the mean F1 score, averaged over five independent runs, against the percentage of edges selected in each graph. Results are shown in Figure 2.

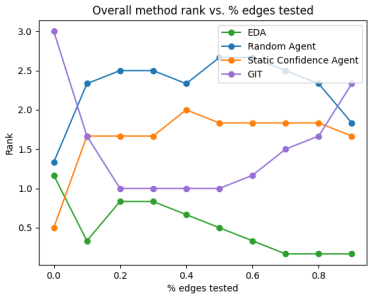

Figure 3: Average rank of each method when numbered from 0 to 2 across each timestep on each graph. The full LLM driven update agent consistently achieves rank 0 across all timesteps. Note: **lower is better**.

Uncertainty driven experiment selection with local updates performs best. Uncertainty driven experiment selection with the LLM utilizing experimental feedback for local updates performs best on nearly all graphs. Further, it outperforms the random selection baselines at nearly every round on every graph, at times by up to 0.5 absolute F1 score. The only exception to this is the Arctic sea ice graph where local updates initially perform poorly. We attribute this to the highly cyclic and thus harder-to-predict graph structure. Additionally, the method significantly outperforms the statistical GIT baseline on both Az and Covid graphs and remains competitive on the rest. Figure 3 plots the average rank of all methods over all timesteps, confirming IGDA's strong performance. Notably, even on graphs where the LLM proposes a poor zero-shot initial prediction, the LLM is able to recover quickly, converging to the correct structure with local updates. This suggests the LLM is able to effectively utilize experiment feedback even when lacking detailed domain knowledge.

Local updates can outperform random selection even with few experiments. Allowing the LLM to make local edge updates using experiment feedback quickly improves the predicted graph even when relatively few edges are selected. This behavior is particularly desirable, as in practice it may be expensive to experiment on even a small fraction of all edges. On some graphs, where the initial LLM confidence estimates are good, the static confidence selection baseline without local updates is also able to quickly outperform random selection. Yet, even when the initial confidence estimates are subpar, local updates compensate and allow for the prediction to quickly improve with just a few edge experiments. This again demonstrates the broad effectiveness of local updates even when initial predictions are poor.

Static uncertainty driven selection performs better than random selection. Despite not fully utilizing experimental feedback, static uncertainty driven selection still outperforms the random selection baseline on five out of seven graphs. This method performs particularly well on AZ and Covid graphs where the initial LLM predictions are already reasonably good. On these graphs static uncertainty selection quickly outperforms randomly selection and is competitive even with local updates. This shows that, on a subset of the graphs, the LLM's confidence in its predictions are well-calibrated, allowing our selection policy to prevent wasting experiments on edges which are most likely already correct. However, we also see the LLM's confidence estimates can be poorly calibrated on graphs for which the initial predictions are inaccurate. See for example the Asphyxia and Neuropathic pain graphs, which start with initial F1 score less than 0.2. On these graphs the static confidence selection component struggles to outperform the random baseline.

GIT performance heavily depends on availability of both observational and interventional data With ample data (5000 observational samples and 100 interventional samples per node) the statistical GIT methods performs well on most graphs where it is applicable (i.e. the graph is acyclic). However, we find this good performance heavily depends on the availability of such data, with decreases in both observational and interventional sample sizes significantly impacting results. In Figure 18 we plot the performance of GIT on the Asphyxia graph with varying amounts of data demonstrating this effect. Results on more graphs are presented in the appendix. In contrast, IGDA does not depend at all on the availability of numerical observational or interventional data. Instead,

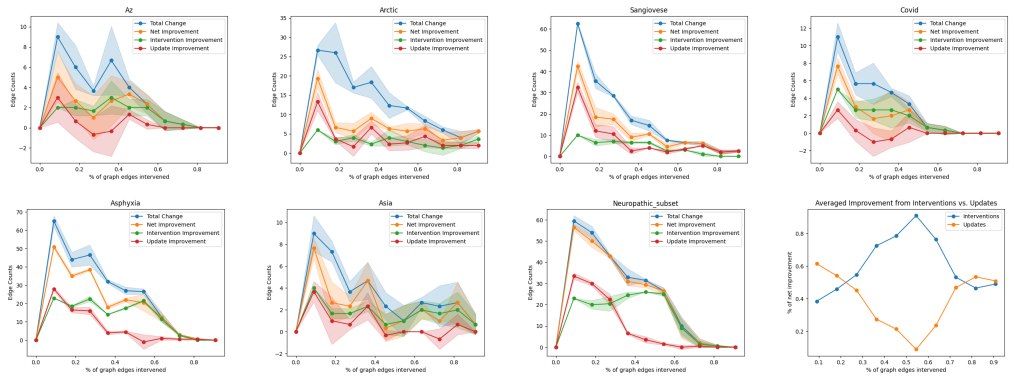

Figure 4: % Improvement from experiments vs. LLM prediction updates across timesteps. Improvement directly from LLM updates peaks early but then falls off. Improvement from experiments stays constant or improves with more experiments as confidence scores become better calibrated.

IGDA relies on the complementary availability of semantic metadata of graph variables within either its pretraining dataset or on the internet.

In an effort to better understand the factors behind IGDA's success we conduct a number of ablations in the following section.

4.1 Ablations

Impact of experiment improvements versus update improvements As a starting point we define the *net graph improvement* in a round r as the difference between the number of edges correctly classified in in $\hat{G}_r$ versus in $\hat{G}_{r-1}$. If an edge (X_i, X_j) is correctly classified in $\hat{G}_r$ but not in $\hat{G}_{r-1}$ we say it has been *improved*. Recall there are two potential mechanisms of improvement for (X_i, X_j): 1) (X_i, X_j) was selected for experimentation in the previous round $r-1$ and feedback on the experiment was received at the start of round r 2) The prediction for (X_i, X_j) was updated by the LLM after receiving experiment feedback for an adjacent edge (X_k, X_l). We call the former improvements *experiment improvements* and the latter *update improvements*. In a given round r we are interested in how much of the net improvement for a graph is due to experiment improvements versus update improvements. To examine this, we plot both quantities in Figure 4 for the discovery processes discussed in the previous section. In addition, we plot the net graph improvement and total number of edges changed from each round.

In all seven graphs we see both the total number of changed edges and the net improved edges peak at the first round and then decay towards zero. Notably, on some graphs there is a significant gap between net improvement and total change, indicating many edges changed during dynamic updates are misclassified after previously being correctly classified. This decline in total and net change is reflected in the number of update improvements which peak early and sharply decline to zero. This observation supports our intuition above that allowing the LLM to dynamically update edge predictions without direct experimental feedback on the edge can dramatically improve performance at small percentages of experiments. In contrast, experiment improvement accounts for a smaller percentage (less than 40%) of edge improvements early on. However, in most graphs the number of experiment improvements stays nearly constant until at least 50% of edges are already selected. As a result, improvement from experiments grows to account for 90% of all edge improvements for rounds performed during this period. This demonstrates improvements from experiment and updates complement each other, with **update improvement driving net improvement early and experiment improvement driving net improvement later on**.

Our analysis here also confirms the effectiveness of allowing the LLM agent to update both the prediction **and** confidence for an edge. Even when only considering improvements from experiments when doing local updates, we see a major improvement over the static confidence baseline. This suggests the **updates made to edge confidence scores are equally important in achieving good performance**, allowing for sustained experiment improvement throughout the discovery process.

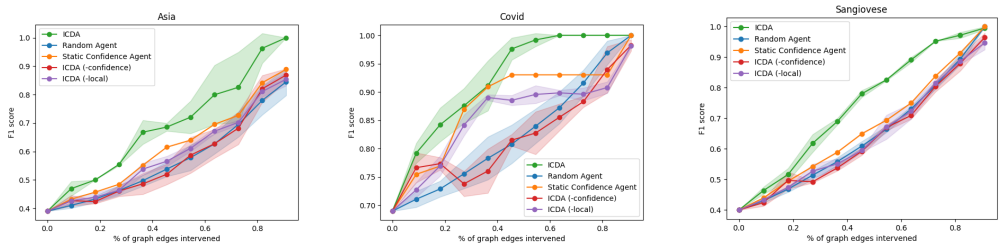

Figure 5: Ablating confidence based edge selection and local update prompting.

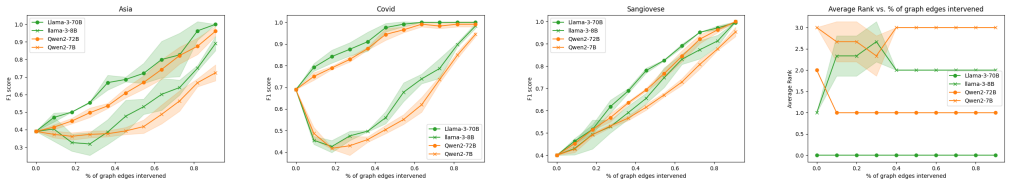

Figure 6: Performance of LLM driven interactive discovery on different sized models. Small LLMs (8B params) underperform the random baseline.

Impact of Confidence Based Selection and Local Prompting We now ablate the impact of two key components of our discovery strategy: 1) confidence based edge selection and 2) local update prompting. To ablate 1) we directly prompt the LLM to generate a list of edges to experiment on instead of selecting via confidence. This requires us to put the entire current predicted graph $\hat{G}_r$ in-context. When dynamically updating $\hat{G}_r$ after receiving experimental feedback we remove all confidence estimates but retain the local prompting strategy. To ablate 2) we retain the same confidence edge selection proposed but replace local update prompts after with a single global update prompt containing the current prediction $\hat{G}_r$ and all recently received experiment feedback. We report the results of running the interactive discovery process with these methods in Figure 5.

We find both ablations struggle to perform better than the random baseline. Local updates without confidence selection perform well early on but fall off quickly. F1 score on the Covid graph even regresses after the initial improvements, likely due to incorrect local updates and a poor experiment selection policy. This suggests in addition to providing a strong experiment selection procedure, maintaining running confidence estimates for each edge reduces the variance of local updates from experiment feedback. Turning to the ablation for local prompting, we again find performance not much better than the random baseline. Surprisingly, even on Covid where the static confidence selection performs well, confidence based selection + global updates still struggles. This indicates the base LLM is not able to correctly update the predicted graph when giving everything in context at once. This further motivate the practical importance of the local prompting procedure, which greatly simplifies the context the LLM must consider in each model call. Additionally, we note that for large enough graphs, putting everything in context is simply not feasible. By contrast, local prompting is easily scalable to larger graphs, albeit at a quadratic cost.

Impact of the LLM Model Size The above experiments exclusively use a single base LLM (`Meta-Llama-3-70B-Instruct`) to perform both the initial round of zero-shot edge predictions and dynamically update edge predictions/confidences using experiment feedback. Now, we examine the impact of changing both the base model size and type. In Figure 6 we initialize the discovery process with zero-shot predictions made by `Meta-Llama-3-70B-Instruct` and run local updates using the smaller `Meta-Llama-3-8B-Instruct` as well as two models from the Qwen2 series.

We find the original `Meta-Llama-3-70B-Instruct` consistently performs best on all graphs at every time step. The other 70B model, `Qwen2-72B-Instruct`, performs similarly but consistently worse. In contrast, on the Asia and Covid graphs, both 8B models perform worse than even the random baseline. Surprisingly `Meta-Llama-3-8B-Instruct` performs reasonably well

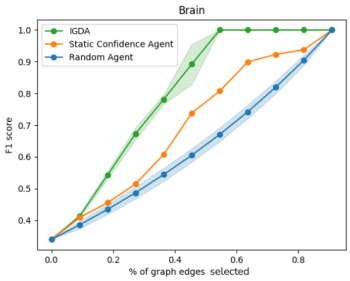

Figure 7: Performance curves of uncertainty driven selection + local prompting vs. baselines on the Brain graph (Zhu et al., 2024) recently published in July 2024. Both LLM-driven methods perform well despite the graph not possibly being in the LLM's training data. Note: GIT is not reported because the graph is cyclic.

on the Sangiovese graph, performing similarly even to the 9x larger Qwen2 70B model. Overall however these results indicate performance on the interactive graph discovery task can be substantially improved with model scale.

We next investigate the performance of different models on the initial zero-shot edge prediction task. Using the pairwise confidence estimation prompt in Section B we prompt each of four models to produce a zero-shot prediction $\hat{G}_0$ with edge confidence values. Using the predicted confidence estimates we run greedy static confidence selection procedure as in 4. Ranks for each selection procedure averaged over all graphs are plotted in Figure 9. F1 scores in each graph are reported in Figure 8 in the Appendix.

Impact of Memorization The success of LLMs in discovery stems from their immense background knowledge acquired during pre-training. This background knowledge informs the model during edge prediction and confidence calibration, allowing for strong performance even zero-shot. However, if benchmark graphs are contained verbatim in pre-training data, memorization becomes a significant confounding factor. To investigate to what extent memorization impacts performance we find a recently published graph (published in July 2024) from Zhu et al. (2024) modeling the gene regulatory network underlying 29 protein transcription factors. Because `Meta-Llama-3-70B-Instruct` finished training in 2023 this graph is guaranteed to be memorization free. Figure 7 plots the performance of uncertainty driven edge selection + local updates compared to the static selection and random baseline.

Figure 7 shows our confidence driven selection + local update approach performs very well even on graphs with minimal memorization contamination. As previously observed, local prediction updates allow for fast improvement over the random baseline even with a small number of experiments. Surprisingly, the static confidence selection approach also works well here. This indicates zero-shot edge confidence scores can be well calibrated on graphs with no contamination from memorization. We additionally note this graph has a complex structure with many cycles of varying lengths. This shows our method performs well even on graphs which strongly violate often assumed DAG conditions.

5 Conclusions and Future Work

In this work we proposed IGDA as a novel application of LLMs to interactive graph discovery. Our experiments confirm the proposed IGDA method significantly outperforms baselines. Our ablations confirm both uncertainty driven edge selection and local updates using experiment feedback as importantly contributing to the method's good performance. Further, this method is complementary to existing statisical methods for experiment design or causal discovery (e.g. GIT (Olko et al., 2024)). Statistical methods utilize available observational/interventional numerical data to make predictions and confidence estimates whereas IGDA utilizes available variable semantic metadata to make pre-

dictions and confidence estimates. Designing a method leveraging both numerical and semantic variable data is promising future work.

Ethics Statement As with any work studying generative models, we note generative modeling can suffer from pre-existing biases in the training data. This behavior may help propagate existing societal biases present today.

Reproducibility Statement This work utilizes only open-source models and datasets making it 100% reproducible. All benchmark graphs are documented in Appendix Section C. We additionally plan to release code in the case of an acceptance.

REFERENCES

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pAoqRlTBtY.

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4, 2023. URL https://arxiv.org/abs/2311.07361.

Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu. Clear: Can language models really understand causal graphs?, 2024. URL https://arxiv.org/abs/2406.16605.

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data, 2023. URL https://arxiv.org/abs/2210.17283.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002. URL https://api.semanticscholar.org/CorpusID:1191614.

Zhixuan Chu, Jianmin Huang, Ruopeng Li, Wei Chu, and Sheng Li. Causal effect estimation: Recent advances, challenges, and opportunities, 2023. URL https://arxiv.org/abs/2302.00848.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,

Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan

Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. Glore: When, where, and how to improve llm reasoning via global and local refinements, 2024. URL https://arxiv.org/abs/2402.10963.

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, 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.642182. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.642182.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents, 2024. URL https://arxiv.org/abs/2406.06769.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Andrew Kyle Lampinen, Stephanie C Y Chan, Ishita Dasgupta, Andrew J Nam, and Jane X Wang. Passive learning of active causal strategies in agents and language models, 2023. URL https://arxiv.org/abs/2305.16183.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenweu Zhu. Realtcd: Temporal causal discovery from interventional data with large language model, 2024. URL https://arxiv.org/abs/2404.14786.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints, 2022. URL https://arxiv.org/abs/2107.10483.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark.

Discoverybench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

Christopher Meek. Causal inference and causal explanation with background knowledge, 2013. URL https://arxiv.org/abs/1302.4972.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Mach. Learn. Knowl. Extr.*, 1:312–340, 2019. URL https://api.semanticscholar.org/CorpusID:68070067.

Mateusz Olko, Michał Zajac, Aleksandra Nowak, Nino Scherrer, Yashas Annadani, Stefan Bauer, Łukasz Kuciński, and Piotr Miłoś. Trust your ∇: Gradient-based intervention targeting for causal discovery, 2024. URL https://arxiv.org/abs/2211.13715.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Yusuf Roohani, Jian Vora, Qian Huang, Zachary Steinhart, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments, 2024. URL https://arxiv.org/abs/2405.17631.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C. Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. Learning neural causal models with active interventions, 2022. URL https://arxiv.org/abs/2109.02429.

B Sibbald and M Roland. Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, 316(7126):201, jan 1998. doi: 10.1136/bmj.316.7126.201.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x. URL https://doi.org/10.1186/s40535-016-0018-x.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. 1993. URL https://api.semanticscholar.org/CorpusID:117765107.

Ruibo Tu, Kun Zhang, Bo Christer Bertilson, Hedvig Kjellström, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation, 2019. URL https://arxiv.org/abs/1906.01732.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery, 2023. URL https://arxiv.org/abs/2310.15117.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers, 2024. URL https://arxiv.org/abs/2309.03409.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks, 2019. URL https://arxiv.org/abs/1904.10098.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018. URL https://arxiv.org/abs/1803.01422.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. Causalbench: A comprehensive benchmark for causal learning capability of large language models, 2024. URL https://arxiv.org/abs/2404.06349.

Yuehua Zhu, Panayiotis V Benos, and Maria Chikina. A hybrid constrained continuous optimization approach for optimal causal discovery from biological data. *Bioinformatics*, 40(Supplement$_2$) : $ii87 - -ii97, 092024. ISSN1367 - 4811. doi : .$ URL https://doi.org/10.1093/bioinformatics/btae411.

A STATIC CONFIDENCE SELECTION OVER MULTIPLE MODELS

Figure 8 reports the results of applying static confidence experiment selection using various models.

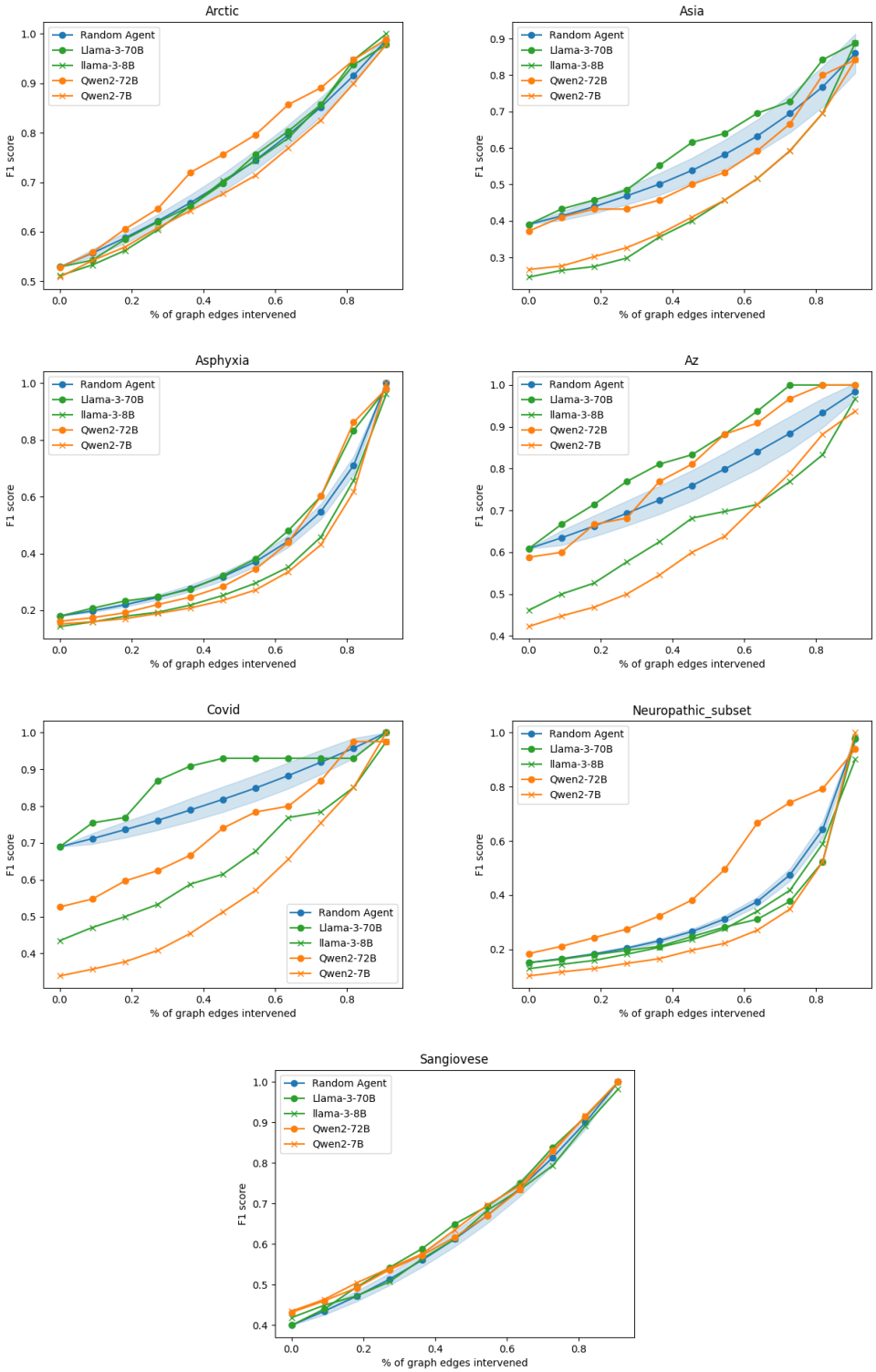

Figure 8: Static confidence selection over multiple models.

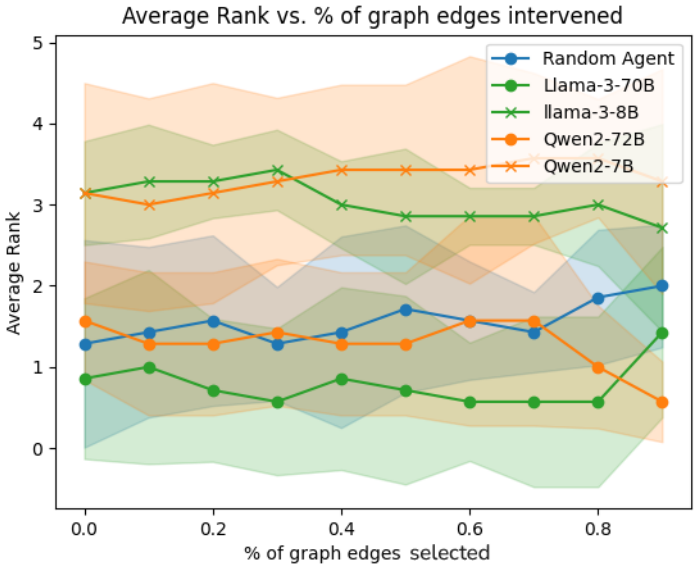

Figure 9: Static confidence based selection ranks for different models averaged across graphs. `Meta-Llama-3-70B-Instruct` is the only model to consistently outperform random guessing. Note: **lower is better**.

B PROMPTS

Zero-shot Confidence Estimation Prompt

[ht] {task_description} Your goal is to understand the direct causal parents of {target}. Another variable is a direct causal parent of {target} if an experiment on the variable affects {target} and there are no other causal parents between the variable and {target}. Now, you must determine whether {parent} is a causal parent of {target}. Here is a list of all other variables to consider:
{variables_info}
Do some brainstorming, comparing relevant characteristics of both variables and then print your judgment at the end of your response enclosed in the tags decision YES/NO/decision . Print YES if {parent} is causal. Otherwise print NO. You should also print your confidence from a scale from 1 - 100 (with 100 being most confident) in the tags confidence ... /confidence .
Information about {target}: {target_info}
Information about {parent}: {parent_info}

Parent Update Prompt

You are a causal discovery expert. You have been given the following list of variables and tasked with predicting the true causal graph through a sequence of experiments on edges.
{variables_info}
Note: each edge has an associated confidence value from 1 - 100. The presence of an edge is represented as (A− >B,CONFIDENCE) where A is the parent and B is the child. The absence of an edge is represented as (NOT A− >B, CONFIDENCE)
From one experiment you have discovered {experiment_feedback} Previously you predicted {experiment_prediction}
Now you should update your belief about the other edges of {parent} based on the results of the experiment. Consider the predicted edge
{other_edge_prediction}
Now you should reason about how to update your belief about the above edge based on the experiment. This means you can either keep your confidence the same, update your confidence, or change your prediction entirely. At the end of your response give your updated prediction at the end of your response in the format decision PARENT/NOT CAUSAL/decision¿ confidence CONFIDENCE/confidence . Print 'PARENT' if the edge should be present and 'NOT CAUSAL' if the edge should be absent.
You should do this in three steps.
Step 1: Brainstorm what physical causal connection there may be, if any.
Step 2: Reason about what the experiment feedback tells you. Think carefully about how similar the new child is to the experimental child.
Step 3: Give your final decision.

Child Update Prompt

You are a causal discovery expert. You have been given the following list of variables and tasked with predicting the true causal graph through a sequence of experiments on edges.
{variables_info}
Note: each edge has an associated confidence value from 1 - 100. The presence of an edge is represented as (A− > B,CONFIDENCE) where A is the parent and B is the child. The absence of an edge is represented as (NOT A− >B, CONFIDENCE)
From one experiment you have discovered {experiment_feedback} Previously you predicted {experiment_prediction}
Now you should update your belief about the other edges of {child} based on the results of the experiment. Consider the predicted edge
{other_edge_prediction}
Now you should reason about how to update your belief about the above edge based on the experiment. This means you can either keep your confidence the same, update your confidence, or change your prediction entirely. At the end of your response give your updated prediction at the end of your response in the format decision PARENT/NOT CAUSAL/decision confidence CONFIDENCE/confidence . Print 'PARENT' if the edge should be present and 'NOT CAUSAL' if the edge should be absent.
You should do this in three steps.
Step 1: Brainstorm what physical causal connection there may be, if any.
Step 2: Reason about what the experiment feedback tells you. Think carefully about how similar the new parent is to the experiment parent.
Step 3: Give your final decision.

C CAUSAL GRAPHS

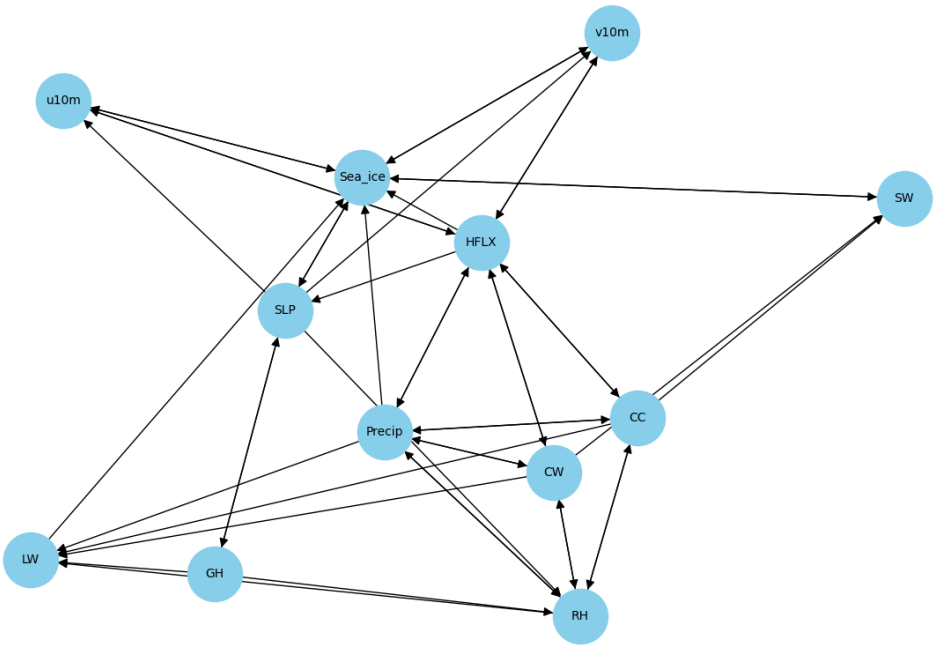

Figure 10: Arctic sea ice causal graph.

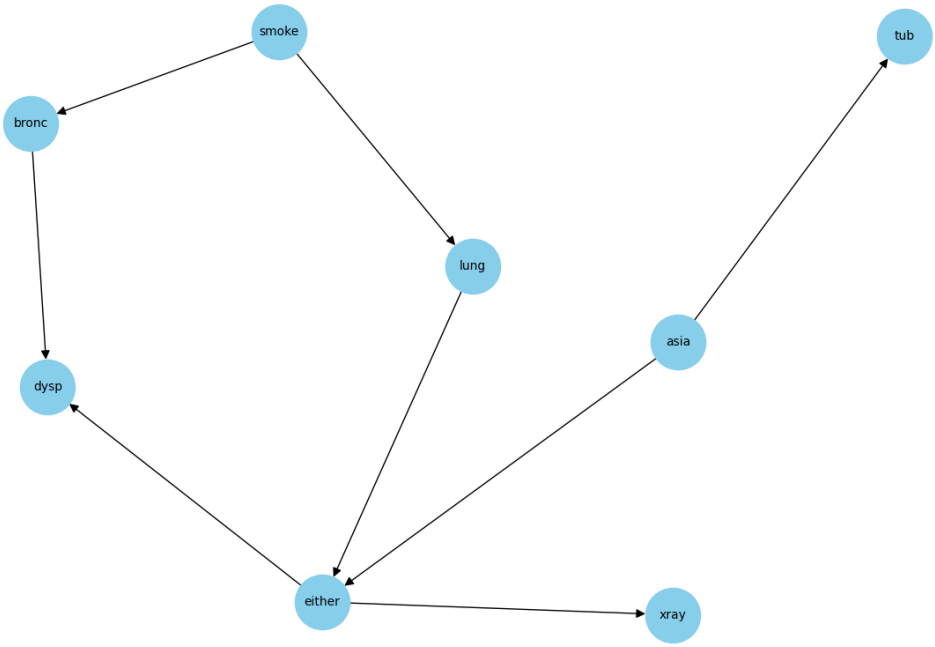

Figure 11: Asia causal graph.

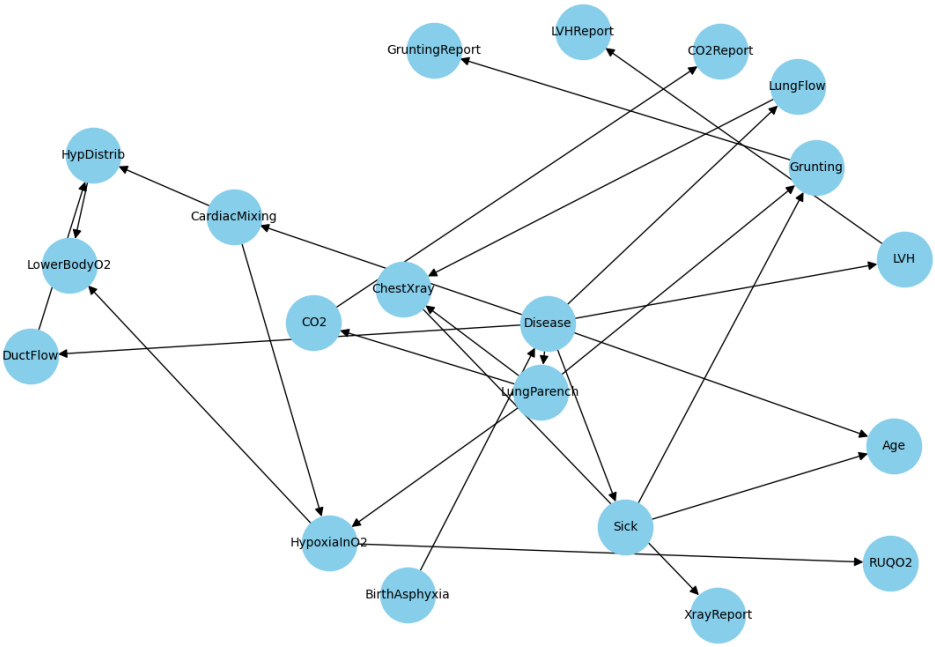

Figure 12: Asphyxia causal graph.

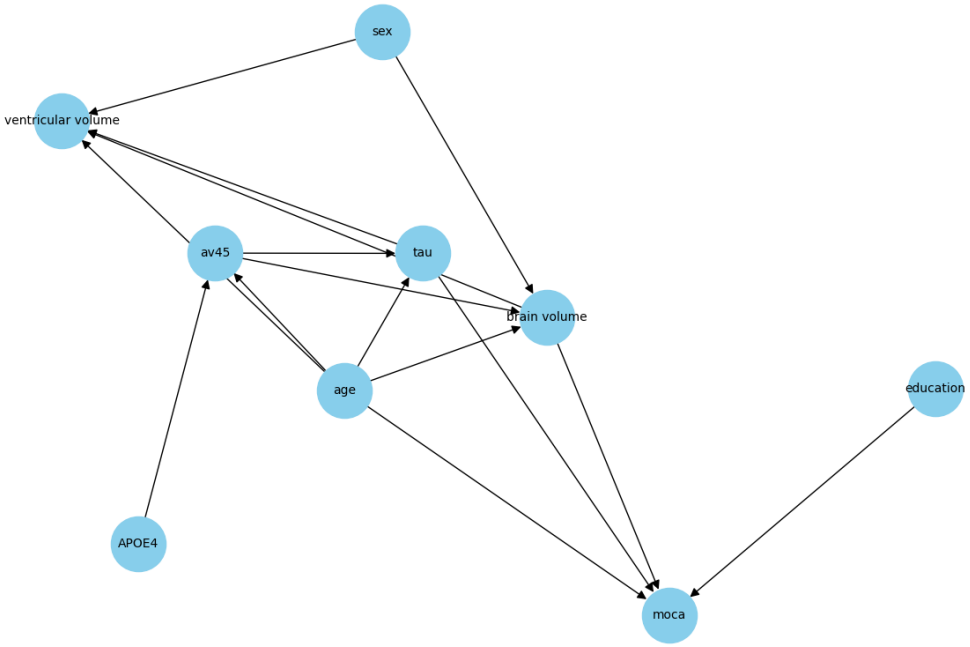

Figure 13: Alzheimers causal graph.

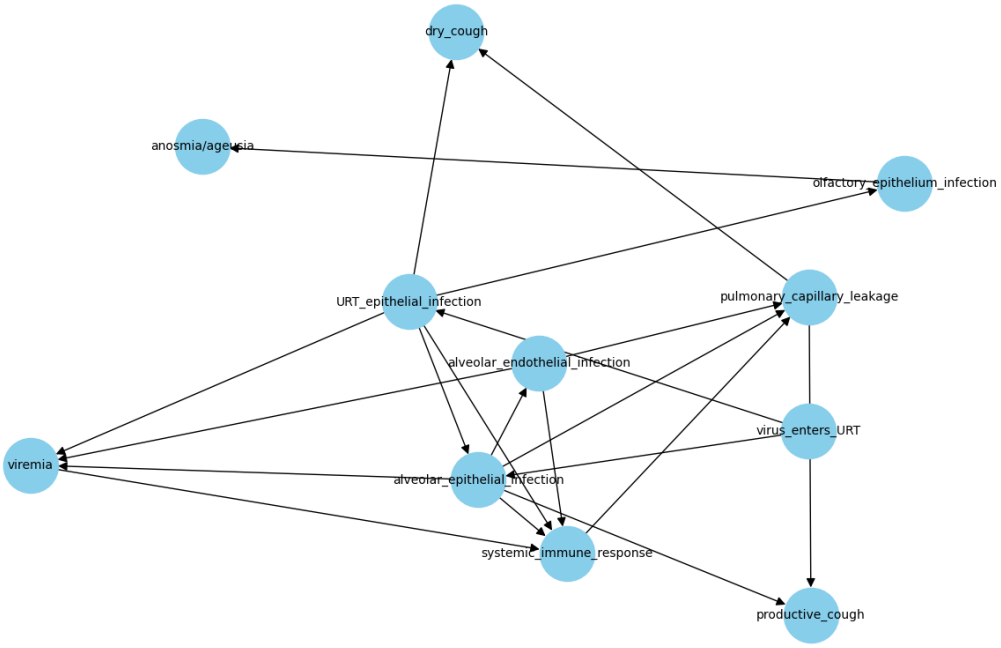

Figure 14: Covid causal graph.

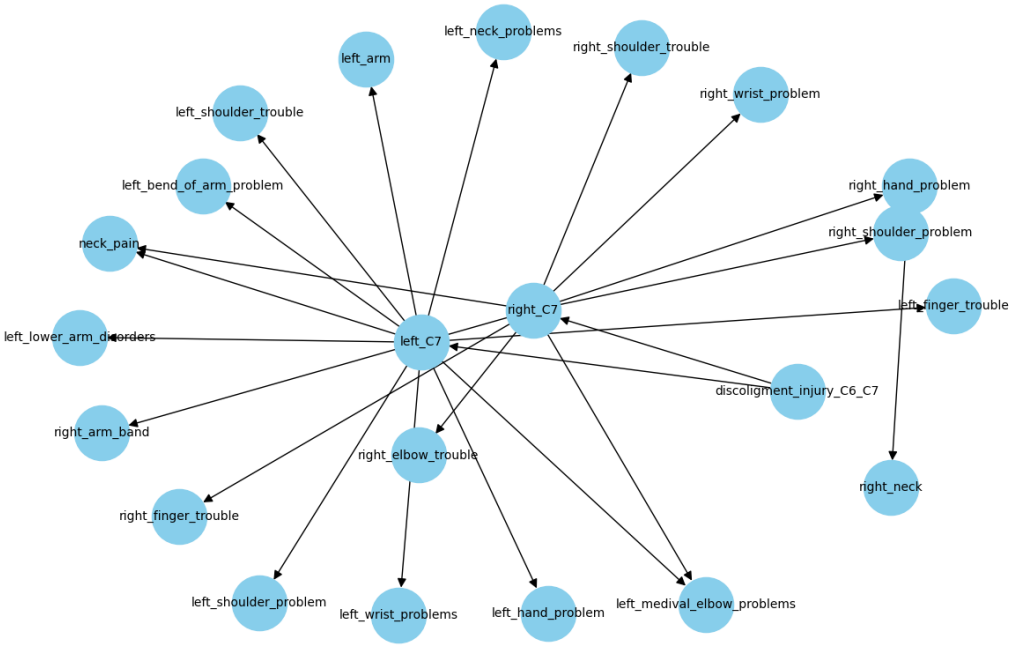

Figure 15: Neuropathic pain causal graph.

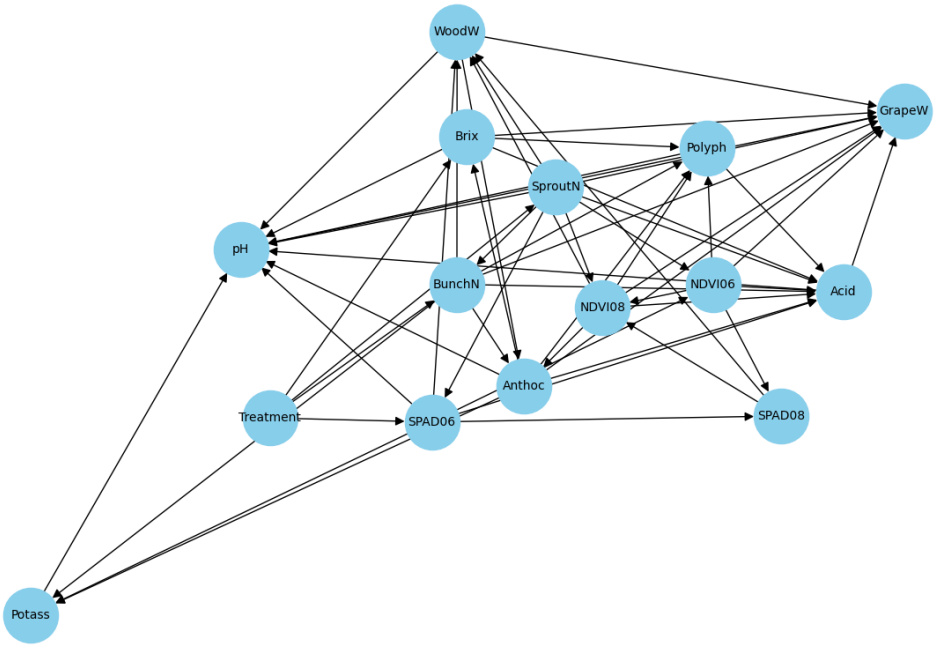

Figure 16: Sangiovese causal graph.

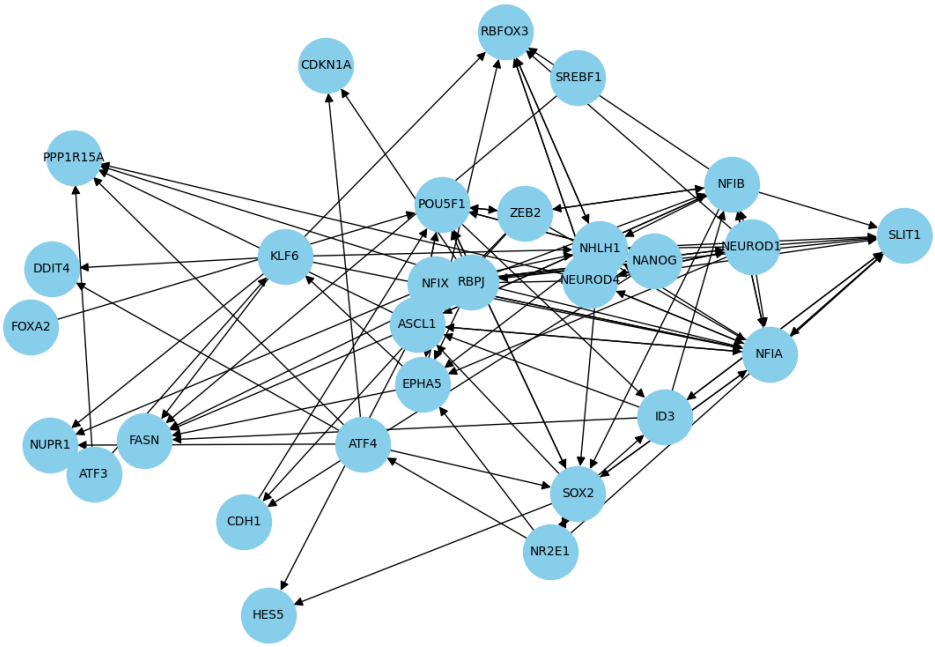

Figure 17: Brain causal graph.

D GIT Ablations

Figure 19 plot GIT performance (Olko et al., 2024) over six causal graphs with varying amounts of observational and interventional data.

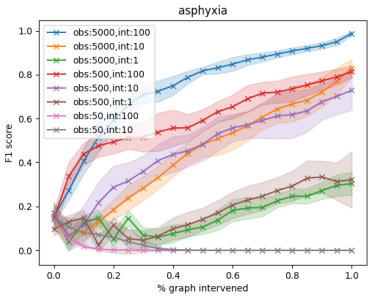

Figure 18: GIT with varying amounts of observational and interventional data. Decreasing either observational or interventional sample sizes can decrease performance by over 0.2 F1 score.

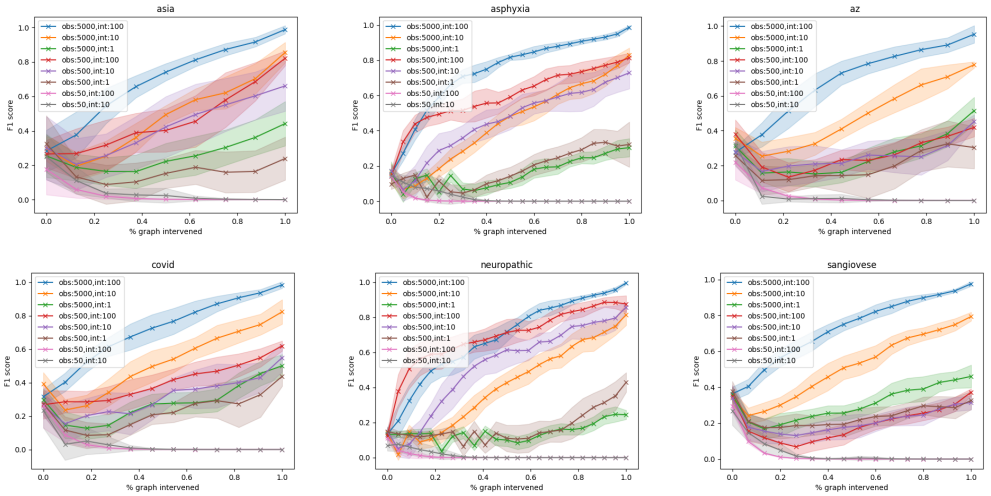

Figure 19: GIT ablations with varying amounts of observational and interventional data.