LARGE LANGUAGE MODELS FOR ZERO-SHOT INFERENCE OF CAUSAL STRUCTURES IN BIOLOGY

Izzy Newsham[1]* **Luka Kovačević**[1]* **Richard Moulange**[1]*
Nan Rosemary Ke[2, 3] **Sach Mukherjee**[4,1]
[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[2]DeepMind, London, UK
[3]Mila, Montreal, Canada
[4]DZNE & University of Bonn, Bonn, Germany
{izzy.newsham, sach.mukherjee}@mrc-bsu.cam.ac.uk

ABSTRACT

Genes, proteins and other biological entities influence one another via causal molecular networks. Causal relationships in such networks are mediated by complex and diverse mechanisms, through latent variables, and are often specific to cellular context. It remains challenging to characterise such networks in practice. Here, we present a novel framework to evaluate the ability of large language models (LLMs) for zero-shot inference of causal relationships in biology. In particular, we systematically evaluate causal claims obtained from an LLM using real-world interventional data. This is done over one hundred variables and thousands of causal hypotheses. Furthermore, we consider several prompting and retrieval-augmentation strategies, including large, and potentially conflicting, collections of articles. Our results show that with tailored augmentation and prompting, even relatively small LLMs can capture meaningful aspects of causal structure in biological systems. This supports the notion that LLMs could act as orchestration tools in biological discovery, by helping to distil current knowledge in ways amenable to downstream analysis. Our approach to assessing LLMs with respect to experimental data is relevant for a broad range of problems at the intersection of causal learning, LLMs and scientific discovery.

1 INTRODUCTION

Discovery in many scientific disciplines is complex and often expensive and time-consuming. In biology, the large number of interacting components creates an enormous space of potential experiments to perform. Experimental plans, as well as joint experimental-computational workflows, are usually informed by the existing literature (e.g. on a particular disease or biomolecular pathway), but this step is itself challenging. If large language models (LLMs) are capable of reducing even a small fraction of the work needed to elucidate biological mechanisms, or compress literature-derived knowledge into priors for *in silico* models, they have the potential to significantly accelerate biological discovery and the understanding of disease mechanisms.

To date, often without fine-tuning on biological data, LLMs have been applied in biochemistry as experimental orchestration tools (M. Bran et al., 2024), AI bioinformaticians (Ding et al., 2024), collaborative multi-agent teams of AI scientists (Swanson et al., 2024) or as a source for pre-trained embeddings for downstream prediction tasks (Chen & Zou, 2024). However, to accelerate science, LLMs must be capable of elucidating causal relationships and this is highly relevant in the context of studying molecular networks underpinning disease biology.

In experimental settings conducive to LLM-driven orchestration, researchers often face large experimental search spaces. Tasks may include identifying molecular pathways associated with a transcriptional signature, inferring regulatory pathways linked to a gene target, or designing molecules that target a specific gene. LLMs that are capable of elucidating causal gene-gene relationships would be

*Contributed equally.

useful in collapsing the experimental space by identifying potential (i) upstream regulators, (ii) direct effects and (iii) key regulatory pathways. However, for any such downstream application it would be essential to first understand how to evaluate whether LLMs are able to infer *causal* relationships between molecular entities.

Perturbation screens are fundamental tools in biology that allow researchers to observe inherently causal relationships by intervening experimentally on specific molecular entities. Contemporary gene perturbation experiments— notably those using CRISPR-based gene editing tools—offer the possibility of carrying out interventions in a systematic fashion (spanning potentially large numbers of intervention targets) and to record changes brought about by such interventions in a global fashion. These experimental approaches have developed rapidly and include a wide range of specific protocols (Jinek et al., 2012; Qi et al., 2013; Piatek et al., 2015; Dixit et al., 2016). Such experiments can be carried out in specific biological contexts (e.g. particular cell types and epigenetic backgrounds), offering a scalable way to explore causal relationships that may themselves be context-dependent. Data from such experiments offer an opportunity to understand, in a systematic manner, whether LLMs are capable of elucidating causal relationships between genes.

Our contributions. We leverage data from interventional experiments to generate a causal ground-truth to assess whether LLMs can infer gene regulatory networks in a zero-shot fashion, without prior exposure to the experimental data. Specifically, we focus on evaluating the ability of LLMs to identify directed gene–gene causal relationships, a key step towards their wider application in applied causal inference tasks in biology. Importantly, our experimental design tests LLM output with respect to empirical, interventional experiments, mimicking a real-world scientific workflow. This paves the way for applying LLMs to generate context-specific priors for causal models of biology and helps to evaluate their potential in orchestrating experiments aimed at studying gene regulation. We offer the following contributions:

1. We develop a benchmarking approach for assessing capabilities of LLMs in eludicating (directed) causal relationships in biology;
2. We explore a range of retrieval-augmented strategies, including ways to specify biological or experimental context as well as providing guidance at the node or gene level;
3. Finally, the LLM-based output is compared to standard knowledge-driven prior specification methods.

By combining insights from causal structure learning, Perturb-seq datasets, and advanced text-mining tools like PubTator 3.0 (Wei et al., 2024) and the STRING database (Szklarczyk et al., 2023), this work aims to quantify the ability of current LLMs to infer causal gene–gene relationships and build a framework to help evaluate the role of LLMs in causal discovery within complex biological systems.

2 BACKGROUND

We provide a brief introduction to single-cell perturbation screens in Section 2.1. We then introduce the basic terminology and notation from causal structure learning that will be used throughout this paper in Section 2.2. Finally, Section 2.3 describes related work.

2.1 SINGLE-CELL PERTURBATION SCREENS

Advances in RNA sequencing technologies allow the measurement of thousands of molecular readouts in single cells, referred to as single-cell RNA sequencing (scRNA-seq). These advances underpin protocols such as Perturb-seq (Dixit et al., 2016), which combine large-scale interventions (e.g. gene knockouts) with measurement of gene expression at the single-cell level. Perturb-seq and related protocols allow researchers to explore causal effects at scale, providing rich datasets for studying gene networks and causal regulatory interactions.

2.2 CAUSAL STRUCTURE LEARNING

Inferring the existence of causal relationships is a fundamental question in science. Causal structure learning (CSL) methods (Heinze-Deml et al., 2018) aim to identify these causal relationships from

observational and interventional data. These relationships are usually represented by a causal graph $\mathcal{G} = (V, E)$, which is composed of vertices $V = [d]$ and directed edges $E = \{i \to j : X_i \in X_{\text{pa}(j)}\}$. An edge $i \to j$ implies a causal relationship between X_i and X_j.

In this context, if there exists a directed path $i \to \ldots \to j$, then node i is an ancestor of node j, and conversely, node j is a descendant of node i. Here, we use the term ancestral or indirect causal graph to refer to a graph where each edge $x_i \to x_j$ implies that x_i is an ancestor of x_j (rather than a parent). We can also compute the transitive closure of $\mathcal{G}$, denoted by $\mathcal{G}^+ = (V^+, E^+)$, such that if there is a path between $i, j \in V$ then there is a direct edge $i \to j \in E^+$. Indirect edges are a relevant notion in analyzing experimental perturbations of the kind seen in genetic screens, since these capture total causal effects of perturbation on a given node (Hill et al., 2019).

2.3 Related work

Previous research has demonstrated that LLMs are capable of extracting causal relationships between variables from short, well-structured sentences (Kıcıman et al., 2023; Nie et al., 2023; Romanou et al., 2023; Jin et al., 2024). However, these tasks typically lack the complexity, noise, and contradictions typically found in scientific literature.

Märtens et al. (2024) demonstrated that LLM-derived embeddings can improve the performance of generative models for perturbation prediction. While this highlights the utility of LLMs in representation learning for biology, it does not directly address their ability to perform causal inference tasks. By focusing specifically on causal retrieval and reasoning, our study fills this gap.

Other research in evaluation for CSL from Perturb-seq experiments has focused on evaluating CSL from Perturb-seq data directly (Chevalley et al., 2022; Kovačević et al., 2024), or considered the perturbation prediction problem from a non-causal perspective (Wu et al., 2024; Szałata et al., 2024). In Chevalley et al. (2022) the authors apply a similar hypothesis testing procedure to generate ground truth causal graphs from Perturb-seq data, however, they focus on data-driven CSL methods.

To evaluate LLM performance, in addition to testing against experimental data, we benchmark also against traditional knowledge-driven methods. This includes the STRING database (Szklarczyk et al., 2023), a database that integrates known and predicted protein-protein interactions from multiple sources, including experimental data, computational predictions, and text mining. Unlike standard LLMs, STRING relies on curated biological networks and statistical association scores, making it highly structured and reliable for established interactions but less adaptable to novel or ambiguous queries. Our work contrasts this classical knowledge-driven approach with the broader reasoning capabilities of LLMs.

3 Methodology

To evaluate the ability of LLMs to retrieve causal information about gene regulatory relationships, we compare the causal relationships identified by the LLM to those that can be inferred from Perturb-seq experiments. In Section 3.1, we describe how the causal ground truth is constructed from perturbation data. Section 3.2 briefly describes our prompting scheme, then Section 3.3 defines the approach for evaluating the LLM-derived causal graph.

3.1 Constructing a causal ground truth

Problem setting. For each Perturb-seq experiment with N cells we observe the pairs $\{\mathbf{x}^i, v_i\}_{i=1}^N$, where $\mathbf{x}^i \in \mathbb{R}^d$ is the gene expression for d genes and $v_i \in V \cup \{I_0\}$ denotes the perturbation target, with I_0 being the non-targeting perturbation (i.e. no gene is knocked out). Note that we assume each perturbation targets a single gene. We denote the univariate interventional distribution under intervention v_i for the readout gene j as $\mathbf{X}_j^{v_i}$.

Hypothesis testing. To identify the genes downstream of a given gene k, we test the null hypothesis $H_0 : \mathbf{X}_j^{v_k} \sim \mathbf{X}_j^{I_0}$ (the intervened and unintervened univariate distributions for gene j, respectively $\mathbf{X}_j^{v_k}$ and $\mathbf{X}_j^{I_0}$, are identical) against the alternative hypothesis $H_1 : \mathbf{X}_j^{v_k} \not\sim \mathbf{X}_j^{I_0}$. The Mann–Whitney U test (Mann & Whitney, 1947) is used for these comparisons, with the Benjamini–Hochberg

correction (Benjamini & Hochberg, 1995) to control for the false discovery rate across multiple hypothesis tests. A fixed significance level of $\alpha = 0.05$ is used to determine significant differences.

This procedure identifies a set of differentially expressed genes[1] for gene k denoted by Δ_k, where each $j \in \Delta_k$ represents a gene that significantly changes following an intervention on gene k. Formally, we define:

$$\Delta_k = \{j \mid p_j^k < \alpha\}, \tag{1}$$

where p_j^k is the corrected p-value from the hypothesis test comparing the distributions of $X_j^{v_k}$ and $X_j^{I_0}$.

As shown in Figure 1, for each $j \in \Delta_k$ we draw an ancestral edge $k \to j$, signifying that gene k causally influences gene j. Repeating this procedure across all $k \in \{1, \ldots, d\}$ yields our baseline ancestral graph $\mathcal{G}$. Each directed edge in $\mathcal{G}$ is interpreted as causal but ancestral and possibly indirect, as defined in Section 2.2.

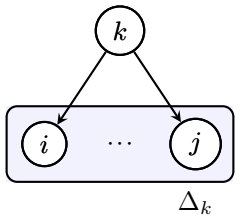

Figure 1: Directed edges are drawn between the perturbed gene k and the set of genes $\Delta_k = \{i, \ldots, j\}$ that change significantly under experimental intervention on k.

3.2 LLM PROMPTING FOR ANCESTRAL CAUSAL GRAPHS

The LLM is prompted to predict the probability of a causal relationship between each pair of genes $i, j \in V$, which we call the *query* genes. For example[2]:

> `User`: Please quantify the extent to which gene i has a causal effect on gene j. Return your answer as a two decimal place probability between 0 and 1.

The LLM output is parsed automatically to retrieve the predicted probability. If no probability is present, it is set to 0, however, this affected only $< 0.33\%$ across all prompts in our experiments.

We repeat this process for all possible pairs of d genes to obtain a matrix of probabilities $P \in \mathbb{R}^{d \times d}$ where each entry $P[i, j] = p_{i \to j}$ represents the probability of an edge $i \to j$. In general, this matrix is not symmetric since the probability of an edge $i \to j$ may differ from $j \to i$. For performance metrics that require binary calls on edges, including those used in this work, the matrix P is thresholded at γ to yield a predicted graph $\hat{\mathcal{G}}_\gamma$ or its transitive closure $\hat{\mathcal{G}}_\gamma^+$. Crucially, after each prompt, the LLM is reinstatiated to prevent biasing the output.

3.3 EVALUATION

Given a probability matrix P obtained by repeatedly prompting an LLM and a ground truth $\mathcal{G}$, we evaluate performance using Area Under the Receiver Operating Characteristic curve (AUROC). We calculate the AUROC over all non-diagonal items in P, where the binary labels are given by the corresponding items in the adjacency matrix of $\mathcal{G}$. In Appendix C, we also present results from considering transitive closures of the ground truth and predicted graphs.

[1] Differentially expressed genes are those that are significantly different between two conditions. In this case, the two conditions are before and after a targeted (CRISPRi) intervention.

[2] Details of our prompting strategy are given in Appendix A.

4 EXPERIMENTS

4.1 DATASET

The Perturb-seq dataset generated by Replogle et al. (2022) contains data from more than 2.5 million human cells on cell-lines[3] K562 and RPE1. This dataset has been used in several causal modelling papers to study the performance of causal models in perturbation prediction (Chevalley et al., 2022).

Since a causal graph with no self-loops and d nodes requires $d(d-1)$ LLM queries to construct the causal graph we consider only the 100 most commonly referenced cancer relevant genes. The process of selecting these genes is explained in Appendix B.

The data is filtered to exclude lowly expressed genes and cells, and z-normalised with respect to the unperturbed (control) samples,

$$\tilde{\mathbf{x}}^i = \frac{\mathbf{x}^i - \mu_0}{\sigma_0}, \tag{2}$$

where μ_0 is the mean and σ_0 is the standard deviation of $\mathbf{X}^{I_0}$. The filtered and z-normalised dataset $\{\tilde{\mathbf{x}}^i, p^i\}_{i=1}^N$ is then used to generate $\mathcal{G}$ as detailed in Section 3.1.

4.2 LLM PROMPTING SETTINGS

For all our experiments, we use `Gemma2-9B-it` (Team et al., 2024), from the state-of-the-art small-model family at time of publication. This is because it is open-source, so we could freely test different experimental settings in the early project exploration stage, and relatively small, so that it required only one GPU to run. Given that some of our experiments—those that introduce long additional literature contexts—took up to 32h to run, we were unable to straightforwardly use larger models and these restrictions are relevant also to prospective scientific use-cases.

In addition to evaluating the performance of LLM-based retrieval without prior or contextual information, we consider additions to our original prompt that allow us to condition the LLM's output on the cellular context, measurement modality, gene function and experimental protocol. We reason that providing additional information about the experimental and biological context could help the LLM reach an embedding space more closely related to the experimental context.

The full list of LLM prompts used for both the experiments for inferring causal direction and full causal graphs can be found in Table 1.

4.3 INFERRING CAUSAL DIRECTION

As an initial illustrative example, we investigate the LLM's ability to infer the causal direction between the example gene pair ATR and CD47. For this pair, $\mathcal{G}$ includes the causal edge ATR $\rightarrow$ CD47, but not CD47 $\rightarrow$ ATR (i.e. the experiments support the former but not the latter).

Naive inference. With no contextual information, the LLM returns $p_{\text{ATR}\rightarrow\text{CD47}} = 0.25$ and $p_{\text{CD47}\rightarrow\text{ATR}} = 0.15$, as visualised in 2**a**. This simple two gene problem provides an example where the LLM is capable of distinguishing causal direction (here, given that there is a causal relationship) with only the names of the two genes (systematic results follow below).

Cancer contextual information. Next, we explore whether conditioning the LLM on the relevant cancer type could reinforce its causal assessment. We add *'in human CML (Chronic myelogenous leukemia)'* to the prompt, which specifies the disease type of the K562 cells on which the experiments were performed. This improves the LLM's predicted scores yielding $p_{\text{ATR}\rightarrow\text{CD47}} = 0.35$ and $p_{\text{CD47}\rightarrow\text{ATR}} = 0.05$, as visualised in 2**b**.

Gene-wise contextual information. We consider several approaches to conditioning on gene-relevant information. First, we consider general independent descriptions of ATR and CD47, provided by RefSeq (O'Leary et al., 2016), which leads to significantly lower predicted probability scores for both causal directions ($p_{\text{ATR}\rightarrow\text{CD47}} = 0$, $p_{\text{CD47}\rightarrow\text{ATR}} = 0.02$), as shown in Figure 2**c**.

[3]Cell-lines are a class of immortalised human cells commonly used in biology as tractable laboratory models.

Setting	Description	Context Type	Inference task
`naive`	Default prompt.	None	Causal direction (4.3) & graph (4.4)
`cancer`	Default prompt with context about the type of cancer that the cells come from.	Experimental	Causal direction (4.3) & graph (4.4)
`gene-desc`	Default prompt with independent information about each of the query genes.	Query genes	Causal direction (4.3)
`literature`	Default prompt with extracts from the literature describing the causal relationship between the query genes.	Query genes	Causal direction (4.3) & graph (4.4)
`false`	Default prompt with a statement denying the true causal relationship between the query genes.	Query genes	Causal direction (4.3)
`contradict`	Default prompt with multiple contradictory statements about the causal relationship between the query genes.	Query genes	Causal direction (4.3)
`mRNA`	Default prompt with context about the type of measurement being taken, which in this case is mRNA abundance or gene expression.	Experimental	Causal graph (4.4)
`evidence`	Default prompt with encouragement for the LLM to consider various types of evidence present in the literature.	Experimental	Causal graph (4.4)
`cancer` + `mRNA`	Combination of the above prompt settings.	Experimental	Causal graph (4.4)
`cancer` + `mRNA` + `evidence`	Combination of the above prompt settings.	Experimental	Causal graph (4.4)
`cancer` + `mRNA` + `experiment`	Combination of above prompt settings with additional context about the type of CRISPRi experiment carried out.	Experimental	Causal graph (4.4)

Table 1: Prompt settings with a description of what they should contain and the task where they are used. The table also includes the type of information that each setting targets whether regarding the query genes or the experimental setting. Examples of each prompt setting can be found in Appendix A.

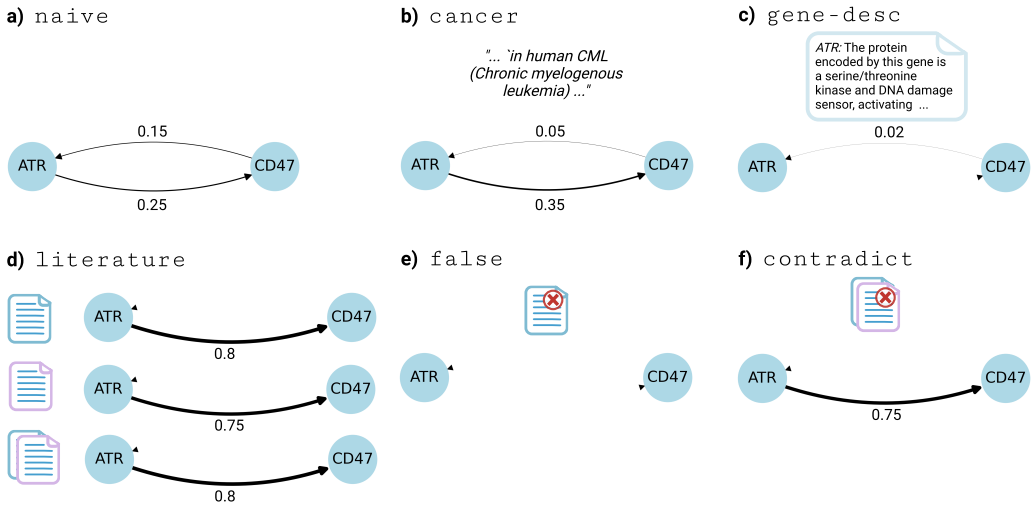

Figure 2: Outputs for inferring causal direction with different prompt contexts, for the example gene pair ATR and CD47.

Now we consider relevant context from the literature comprised of passages[4] where the causal link between ATR and CD47 is directly referenced. The probabilities with this additional context are $p_{\text{ATR}\to\text{CD47}} \geq 0.75$ and $p_{\text{CD47}\to\text{ATR}} = 0$. This is visualised in 2**d** and shows that provided sufficient information directly, LLMs can correctly predict causal direction. It has been shown previously by Jin et al. (2024) that LLMs are capable of this kind of causal reasoning. This underlines again that LLM performance on causal tasks is dependent on the provided evidence.

Contradictory contextual information. To explore this further, we examine the effect of incorrect sentences and contradictory sentences. As expected, the incorrect sentence describing no causal link from ATR to CD47 results in a predicted score of 0 in both directions (Figure 2**e**). When provided with contradictory sentences, the LLM is still able to predict the correct causal edge with $p_{\text{ATR}\to\text{CD47}} = 0.75$ and $p_{\text{CD47}\to\text{ATR}} = 0$ (Figure 2**f**).

4.4 Inferring a causal graph

Having shown that LLMs are capable of distinguishing causal direction given that there is a causal relationship, we move to the more challenging problem of inferring a full causal graph for the genes specified in Section 4.1.

Figure 3 shows the results for `Gemma2` with varying degrees of gene-wise contextual information on the y-axis and experimental contextual information on the x-axis, as described in Table 1. Each cell shows the mean AUROC score for this level of context with the standard error in parentheses. The distributions of the AUROC scores are visualised in Figure A.5.

The best performance is an AUROC $= 0.625$, via prompting with the experimental context of `cancer` + `mRNA` with no gene-specific information. Providing `literature` evidence near-uniformly reduces the LLM performance, especially when there are additional caveats that the literature may not generalise to the specific context of the experiments in question. We hypothesise that this is because this further prompting introduces additional uncertainty and leads the LLM to become less likely to commit to high-probability outputs. When not including gene-specific information, `cancer` +`mRNA` yield predictions with AUROC > 0.6. Including additional details on experimental protocol generally improves performance slightly.

Interestingly, repeating these experiments with chain-of-thought (CoT) reasoning (Wei et al., 2022) does not improve AUROC scores. Figure A.1 shows this for a simple variant of CoT and Figure A.2 for a more detailed, guided variant of CoT, which shows that CoT decreases the performance in many cases. Again, we suspect this is because chain-of-thought reasoning encourages the LLM to be less confident for or against a causal relationship.

Gene-wise contextual information. At least in the present setting of a small LLM, in the context of specific ground-truth experiments, gene descriptions and literature information appear to not be useful for predicting causal regulatory relationships observed in the experiments. We believe that this may be due to the lack of literature available for mechanisms impacted by CRISPRi interventions used to generate the Replogle et al. (2022) dataset. The prompts used for the final row of results in Figure 3 emphasise that the gene-specific contextual information should only be used as supplementary information and yet this worsens performance across prompt settings.

To investigate why gene-wise contextual information worsens performance, we quantify the correlation between $\mathcal{G}$ and the literature evidence: there is no correlation between $\mathcal{G}$ and the literature (Boschloo's one-sided exact test: $p = 0.2075$; Boschloo (1970)). That is, the presence of literature evidence for a gene pair is not significantly associated with whether that gene pair is causal, as described in Appendix B. This result, over a large number of causal relationships, underlines the limitations of relying on traditional literature mining approaches in the context of elucidating causal relationships at large-scale in specific biological contexts.

We also find that the gene descriptions, even when provided as supplementary information, strongly influence the LLM's predictions. When examining the outputs of the LLM with guided CoT, we find that the model heavily focuses on the distinct functions of the genes and how their pathways are not directly interconnected. For example, one of the outputs for the causal gene pair ATR and CD47 states: "There is no readily apparent direct relationship between ATR's role in DNA damage response

[4]Full gene context passages can be found in Table A.2.

and repair, and CD47's function in cell adhesion, calcium signaling, and thrombospondin binding" and "A lack of known direct interactions or regulatory pathways linking ATR to CD47 suggests a low probability of a causal effect".

Comparison with a knowledge-driven baseline. Next, we investigate how the predictions from `Gemma2` compare to a method that does not utilise LLMs but also draws from existing literature-based knowledge on gene interactions and associations. We obtain the scores given by the STRING database (as described in Appendix B), which integrates multiple gene interaction and association databases and literature-derived associations, and evaluate them against an undirected version of $\mathcal{G}$. This results in an AUROC of 0.460, demonstrating that the predictions by `Gemma2` outperform the predictions obtained from STRING for this task regardless of the prompt setting used. It is important to note however, that the STRING database provides symmetric association scores, which are not intended to provide causal information.

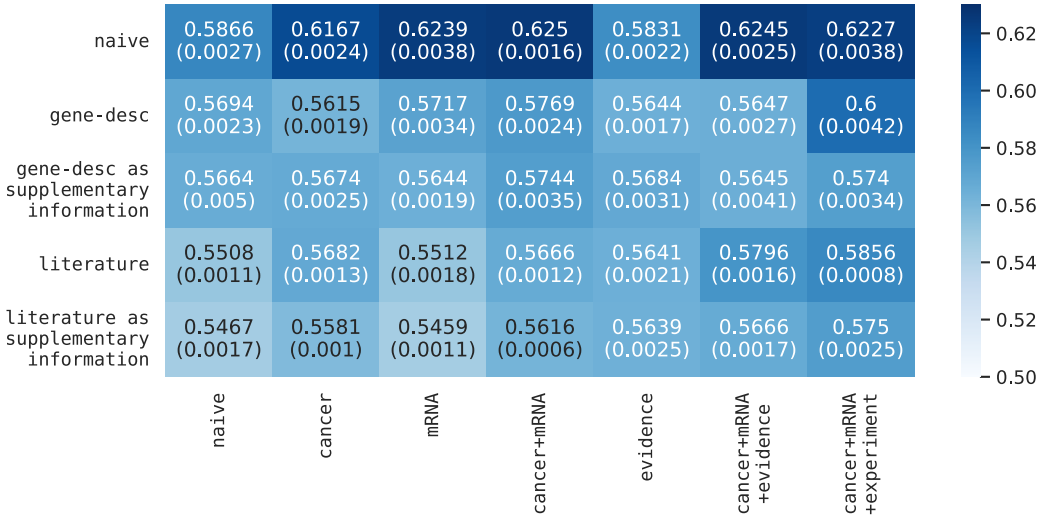

Figure 3: Results on Gemma2 for all combinations of prompt variants (different contexts along each column, different gene-specific information along each row). The results are shown as the mean AUROC over 10 repetitions, with the standard error given in brackets.

5 Discussion

In conclusion, we find that even a small LLM is capable of inferring causal gene–gene relationships better than random chance and in an entirely zero-shot manner, with no input experimental data. `Gemma2` performs considerably better than a database-based zero-shot baseline based on the STRING database. While the highest AUROC of 0.625 is not large in absolute terms, this is an entirely automated, zero-shot method, which requires no biological knowledge nor any empirical data for the task at hand. The LLM output can be straightforwardly integrated as a prior into any downstream causal structure learning task. Due to the hyperexponential space of possible causal graphs, even a weak prior that could steer data driven methods has the potential to signficantly narrow this space.

Our results highlight the importance of context-specific information. Potentially, more sophisticated retrieval-augmented generation strategies—with careful constraining of the associated context—may improve performance. We found that providing the experimental context improved performance yet gene-specific information did not, suggesting that the latter lacked specificity to enable accurate inference with respect to the target ground truth of interest. Indeed, this underlines the need for novel approaches, that go beyond classical literature mining, in the context of elucidating causal relationships in biology. Future work could search explicitly for experimentally-relevant literature—in this case, relating to CML cells and mRNA measurements—rather than any biological or experimental context with associated gene measurement.

Although we spent significant effort on prompt engineering, we nevertheless expect that it might be possible to further improve results by providing longer context and optimising the prompts further.

This is particularly the case for the chain-of-thought reasoning. Moreover, larger 'vanilla' LLMs or RL-imbued reasoning-style LLMs (such as o1; Jaech et al. (2024)) may perform better. As with all recent LLM advances, this work represents a lower-bound on the state-of-the-art and we are excited to see how far future LLMs can provide more accurate zero-shot CSL priors and better quantify causal gene–gene relationship at scale.

ACKNOWLEDGMENTS

This work was partly supported by the UK Medical Research Council (MC_UU_00040/5 & MC_UU_00002/17), the Helmholtz Association AI Project "UNITY" and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust).

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

RD Boschloo. Raised conditional level of significance for the $2\times$ 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9, 1970.

Yiqun Chen and James Zou. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv*, pp. 2023–10, 2024.

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2022.

Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, et al. Automating exploratory proteomics research via language models. *arXiv preprint arXiv:2411.03743*, 2024.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866, 2016.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5(1):371–391, 2018.

Steven M Hill, Chris J Oates, Duncan A Blythe, and Sach Mukherjee. Causal learning via manifold regularization. *Journal of Machine Learning Research*, 20(127):1–32, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096):816–821, 2012.

Jonathan Kans. Entrez Direct: E-utilities on the Unix Command Line. Entrez Programming Utilities Help [Internet], Apr 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK179288/. [Updated 2024 Apr 4].

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

Luka Kovačević, Izzy Newsham, Sach Mukherjee, and John Whittaker. Simulation-based Benchmarking for Causal Structure Learning in Gene Perturbation Experiments. *arXiv preprint arXiv:2407.06015*, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturbation models with LLM-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36:78360–78393, 2023.

Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016.

Agnieszka Piatek, Zahir Ali, Hatoon Baazim, Lixin Li, Aala Abulfaraj, Sahar Al-Shareef, Mustapha Aouida, and Magdy M Mahfouz. RNA-guided transcriptional regulation in planta via synthetic dC as9-based transcription factors. *Plant Biotechnology Journal*, 13(4):578–589, 2015.

Liviu Pirvan and Shamith A Samarajiwa. Pangaea: A modular and extensible collection of tools for mining context dependent gene relationships from the biomedical literature. *bioRxiv*, pp. 2020–04, 2020.

Lei S Qi, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. Crab: Assessing the strength of causal relationships between real-world events. *arXiv preprint arXiv:2311.04284*, 2023.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. *bioRxiv*, pp. 2024–11, 2024.

Artur Szałata, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Jason Fong, Sunil Kuppasani, Richard Lieberman, Tianyu Liu, Javier A Mas-Rosario, Rico Meinl, et al. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. *Advances in Neural Information Processing Systems*, 37, 2024.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, pp. gkae235, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking Machine Learning Models for Cellular Perturbation Analysis. *arXiv preprint arXiv:2408.10609*, 2024.

APPENDIX

A PROMPTING DETAILS

For each pair of genes—in the 9,900 non-identical pairs constructed from the 100 genes pre-selected for their relevance to cancer (see Appendix B)—we used the following default prompt. Here, we have provided the example with genes ATR and CD47:

> `User`: You are a professional biologist who is an expert in assessing whether one gene has a causal effect or not on another. You reply concisely and always follow instructions exactly. Below, you will be asked to quantify the level at which one gene has a causal effect on another by returning a 2 decimal place number between 0 and 1, where 0 indicates the first gene has no causal effect on the second and 1 indicates the first gene has a very strong causal effect on the second. Here is an example:
> [Begin example]
> Q: Please quantify the extent to which [Gene A] has a causal effect on [Gene B]. Return your answer as a 2 decimal place probability between 0 and 1.
> LLM output: Probability = the 2dp probability
> [End example]
> IMPORTANT: DO NOT REPLY IN ANY OTHER WAY. YOUR REPLY MUST END WITH "Probability = " AND THEN A 2 DECIMAL PLACE PROBABILITY. Remember, you are a highly accurate expert biologist who answers concisely, follows instructions and returns only 2dp probabilities as answers.
> Q: Please quantify the extent to which ATR has a causal effect on CD47. Return your answer as a 2 decimal place probability between 0 and 1.
> Probability =

We constructed 105 different variants of this prompt by considering three different types of augmentation: the *experimental context* that underpins the ground truth data we used to evaluate the LLM inferences, *specific information on the gene–gene pair* in question and requests for *chain-of-thought reasoning*. After drafting the prompts ourself, we used GPT-4o (Hurst et al., 2024) to improve the prompts.

We investigated providing six additional experimental contexts to better guide the LLM to the correct biologically-relevant part of its embedding space. These included clarifying that the prediction should be based on a specific type of cancer, that the causal relationship was identified using mRNA gene expression data, and sometimes additional experimental or biological details about causal gene–gene experiments. See Table A.1 for the specific modifications we made to the prompts. We also provided four different types of gene-specific information (see Table A.2). The first two straightforwardly drew on gene descriptions from RefSeq, with or without an additional note clarifying that this was only supplementary information. The third and fourth types provided literature evidence using the PubTator 3.0 system, again with or without a clarifying note. Finally, we experimented with two

chain-of-thought reasoning extensions, which are described in Table A.3. The maximum number of new tokens was set differently for different chain of thought variants: 500 for guided, 200 for simple and 10 for no chain-of-thought.

For around 40% of the gene pairs, no PubTator associations were found. For such cases—for example HMGB1 and RUNX1—the PubTator part of the prompt was replaced with

> A search of PubTator for research articles relating to HMGB1 and RUNX1 did not yield relevant results. However, the absence of findings in this search does not rule out the possibility of an association, as data may exist in other resources or contexts not captured by PubTator.

to clarify. If more than 100 PubTator associations were found, only the first 100 were given due to memory limitations.

B GENE DATA

Cancer genes. We used a text-mining procedure to obtain 100 well-known cancer genes. Specifically, we used the Entrez Direct tool (Kans, 2013) and Pangaea (Pirvan & Samarajiwa, 2020) to parse abstracts from PubMed articles using the search term "cancer". From these abstracts, we identified the 100 most commonly referenced genes that were both perturbed and measured in the Replogle dataset.

Gene descriptions. We used the Entrez Direct tool (Kans, 2013) to retrieve the Entrez gene summary for each gene.

Literature evidence For each gene pair and each of the following relation types: "associate", "interact", "positive_correlate", "negative_correlate", we queried the PubTator 3.0 search API (Wei et al., 2024) and extracted the resulting text to produce the literature evidence data. We ran a one-sided Boschloo's exact test on the literature evidence data to investigate whether the presence of literature evidence was significantly associated with whether a particular gene–gene pair had a causal relationship in either direction. Specifically, we split the gene pairs into two groups: gene pairs with at least one sentence in the literature evidence, and gene pairs with no such sentences. We constructed a contingency table (see Table A.4): A one-sided Boschloo's exact test gives a nonsignificant p-value of 0.2075, which suggests the presence of literature evidence is not significantly associated with whether the gene–gene pair is causal.

STRING scores. We downloaded the STRING database (Szklarczyk et al., 2023) for *Homo sapien* and extracted the combined scores for each gene–gene pair. We constructed a $d \times d$ matrix from the scores (which is symmetric since STRING scores are symmetric), with 0-entry where no STRING score was found.

C EVALUATION UNDER TRANSITIVE CLOSURES

In the main text, we have assumed that the LLM infers *ancestral* causal relationships, since our prompt does not specify that direct causal edges are needed but rather whether gene A has a causal effect on gene B. This causal effect could occur due to a direct causal relationship (i.e. $A \to B$) or through an indirect path (i.e. $A \to C \to B$).

Thus, the LLM's predictions might represent direct edges or even a mix of ancestral and direct edges, hence to assess against an indirect ground truth it may be helpful to consider the transitive closure of the LLM-derived graph. We therefore considered the transitive closure of $\hat{\mathcal{G}}_\gamma$, denoted by $\hat{\mathcal{G}}_\gamma^+$, and compared this with the ground truth $\mathcal{G}$. We also considered comparing $\hat{\mathcal{G}}_\gamma^+$ to the transitive closure of the ground truth, $\mathcal{G}^+$.

We first compare $\hat{\mathcal{G}}_\gamma^+$ to the ground truth $\mathcal{G}$. The results are shown in Figure A.3**a**. This shows the AUROCs are generally lower than before, but here `gene-desc` and `literature` improve performance in some cases, contrasting with our previous results. The best performing prompt variant is `literature` and `cancer+mRNA+evidence` with an AUROC of 0.6074. This could imply that literature evidence helps `Gemma2` to identify direct causal edges, but is less helpful for

identifying the ancestral causal edges. We also compare $\hat{\mathcal{G}}_\gamma^+$ to the transitive closure of the ground truth, $\mathcal{G}^+$, and obtain similar results, as shown in Figure A.4**a**.

Figures A.3**b** and **c** (and A.4**b** and **c**) show how the AUCs change with simple and guided CoT, respectively. As before, the AUCs do not improve with CoT and in some cases the guided CoT decreases performance by up to 0.15.

D TRAINING DETAILS

Each inference experiment was completed on a single NVIDIA A100-SXM-80GB GPU, using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council. In particular, each GPU node contained four such GPUs, with two AMD EPYC 7763 64-Core 1.8GHz Processors and dual-rail Mellanox HDR200 InfiniBand interconnect.

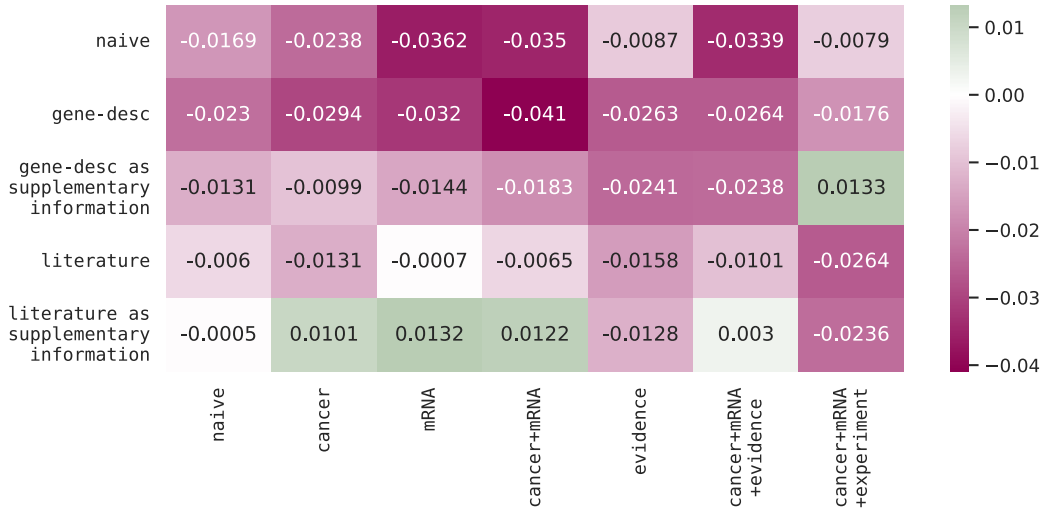

Figure A.1: Results on Gemma2 using simple chain of thought, compared to the results using no chain of thought (shown in Figure 3). Green indicates the simple CoT reached a higher AUROC than no CoT and pink indicates it reached a lower AUROC than no CoT.

Table A.1: Prompts with additional context

Context	Prompt addition
+cancer	... [Gene A] has a causal effect on [Gene B] **in human CML (Chronic myelogenous leukemia)** ... ... ATR has a causal effect on CD47 **in human CML (Chronic myelogenous leukemia)** ...
+mRNA	... [Gene A] has a causal effect on [Gene B] **in the context of gene expression (mRNA measurements)** ... ... ATR has a causal effect on CD47 **in the context of gene expression (mRNA measurements)** ...
+cancer+mRNA	... [Gene A] has a causal effect on [Gene B] **in human CML (Chronic myelogenous leukemia) in the context of gene expression (mRNA measurements)** ... ... ATR has a causal effect on CD47 **in human CML (Chronic myelogenous leukemia) in the context of gene expression (mRNA measurements)** ...
+extra detail	... a very strong causal effect on the second. **Use known scientific evidence, including experimental findings, gene pathway involvement, and established literature, to determine the causal effect. Prioritise experimental findings, such as knockout studies and mechanistic insights from direct gene interactions. Assess the strength of the evidence and consider the possibility of unknown factors. If causal relationships between two genes are unclear or lack sufficient evidence, return a probability reflective of this uncertainty.** Here is an example: ...
+cancer+mRNA+extra detail	... a very strong causal effect on the second. **Use known scientific evidence, including experimental findings, gene pathway involvement, and established literature, to determine the causal effect. Prioritise experimental findings, such as knockout studies and mechanistic insights from direct gene interactions. Assess the strength of the evidence and consider the possibility of unknown factors. If causal relationships between two genes are unclear or lack sufficient evidence, return a probability reflective of this uncertainty.** Here is an example: ... [Gene A] has a causal effect on [Gene B] **in human CML (Chronic myelogenous leukemia) in the context of gene expression (mRNA measurements)** ... ... ATR has a causal effect on CD47 **in human CML (Chronic myelogenous leukemia) in the context of gene expression (mRNA measurements)** ...
+Perturb-seq details	..a very strong causal effect on the second. **The evaluation is based on single-cell data from a Perturb-seq experiment conducted in K562 cells, a chronic myeloid leukemia (CML) cell line. In this experiment, gene perturbations are systematically introduced to assess their effects on the expression levels of other genes at single-cell resolution. Your task is to estimate the extent to which Gene B is affected by the knockdown of Gene A in this specific experimental context (without seeing the raw experimental data). The output probability should estimate the causal relationship inferred from the Perturb-seq data.** Here is an example: ... [Gene A] has a causal effect on [Gene B] **in the context of this Perturb-seq experiment** ... ... ATR has a causal effect on CD47 **in the context of this Perturb-seq experiment** ...

Table A.2: Prompts with gene-specific information, for the gene pair ATR and CD47

Gene-specific information	Prompt addition
Gene descriptions	… follows instructions and returns only 2dp probabilities as answers. **You may use the following contextual information:** **Description of ATR: The protein encoded by this gene is a serine/threonine kinase and DNA damage sensor, activating cell cycle checkpoint signaling upon DNA stress. The encoded protein can phosphorylate and activate several proteins involved in the inhibition of DNA replication and mitosis, and can promote DNA repair, recombination, and apoptosis. This protein is also important for fragile site stability and centrosome duplication. Defects in this gene are a cause of Seckel syndrome 1. [provided by RefSeq, Aug 2017]** **Description of CD47: This gene encodes a membrane protein, which is involved in the increase in intracellular calcium concentration that occurs upon cell adhesion to extracellular matrix. The encoded protein is also a receptor for the C-terminal cell binding domain of thrombospondin, and it may play a role in membrane transport and signal transduction. This gene has broad tissue distribution, and is reduced in expression on Rh erythrocytes. Alternatively spliced transcript variants have been found for this gene. [provided by RefSeq, Jul 2010]** Q: Please quantify the extent …
Gene descriptions as supplementary information	… follows instructions and returns only 2dp probabilities as answers. **You may use the following contextual information:** **Gene Descriptions: These provide independent functional information about each gene. Treat them as supplementary context, not definitive evidence of causality. While contrasting or compatible functions may inform your reasoning, rely primarily on broader biological knowledge, including regulatory pathways, experimental evidence, and known gene interactions.** **Description of ATR: …** **Description of CD47: …** **Note: The descriptions above are intended to provide auxiliary context about each gene's independent functions. They may suggest plausible functional relationships or incompatibilities but are not expected to directly indicate causality. Use them as supplementary information alongside your broader biological understanding to assess the causal relationship.** Q: Please quantify the extent …
Literature evidence	… follows instructions and returns only 2dp probabilities as answers. **You may use the following contextual information, which are extracted from research articles:** **We showed that treatment of tumor cells with a DNA-damage response (DDR) inhibitor targeting the ATR kinase limits the induction of CD47 and PD-L1 signals thus promoting increased antitumor abscopal activity in vivo.** **ATR inhibition may cause downregulation of programmed cell death 1 ligand 1 (PD-L1) and leukocyte surface antigen 47 (CD47), thereby giving a partial suppression of the PD-1/PD-L1 and SIRPα/CD47 immune checkpoints.** Q: Please quantify the extent …
Literature evidence as supplementary information	… follows instructions and returns only 2dp probabilities as answers. **You may use the following contextual information, extracted from research articles by PubTator. These provide associations or findings reported in the literature and are useful as supplementary evidence. However, they are not comprehensive or definitive sources for causal relationships. Consider the possibility of differences in biological contexts, experimental setups, or interpretation. Use them alongside your broader understanding of gene interactions, regulatory pathways, and experimental evidence.** **Literature evidence as supplementary information:** **We showed that treatment, … immune checkpoints.** **Note: The above associations from PubTator provide context about findings in specific studies. They may reflect particular experimental conditions or biological contexts that are not universally applicable. While they may suggest plausible evidence in favour or against a causal relationship, they should not be treated as the sole or definitive source of information. Always integrate this context with your broader biological knowledge and reasoning to assess the causal relationship accurately.** Q: Please quantify the extent …

Table A.3: Prompts with chain-of-thought

Chain of thought variant	Prompt addition
Simple	… Return your answer as a 2 decimal place probability between 0 and 1. **Please think step by step. Remember to first give one or two sentences of reasoning, and then YOU MUST include a 2dp probability at the end of your answer!** LLM output: **Reason = {Here you will include reasoning}** … Return your answer as a 2 decimal place probability between 0 and 1. **Please think step by step. Remember to first give one or two sentences of reasoning, and then YOU MUST include a 2dp probability at the end of your answer!** **Reason =**
Guided	… Return your answer as a 2 decimal place probability between 0 and 1. **Please think step by step. To do so first, generate evidence and reasoning in favour of a causal effect. Second, generate evidence and reasoning against a causal effect. Third, come to an overall conclusion about the evidence. Fourth, include a 2dp probability that reflects, overall, the extent to which [Gene A] has a causal effect on [Gene B]. YOU MUST include a 2dp probability at the end of your answer!** LLM output: **Evidence/reasoning in favour of causal effect = {Relevant evidence/reasoning in favour}** **Evidence/reasoning against a causal effect = {Relevant evidence/reasoning against}** **Overall conclusion = {A simple summary of the evidence/reasoning}** Probability = {the 2dp probability} …Return your answer as a 2 decimal place probability between 0 and 1. **Please think step by step. To do so first, generate evidence and reasoning in favour of a causal effect. Second, generate evidence and reasoning against a causal effect. Third, come to an overall conclusion about the evidence. Fourth, include a 2dp probability that reflects, overall, the extent to which ATR has a causal effect on CD47. YOU MUST include a 2dp probability at the end of your answer!** **Evidence**

Table A.4: Contingency table showing the relationship between STRINGdb-derived literature gene–gene associations and whether the corresponding edge is causal.

	No literature evidence	**Literature evidence present**
No causal edge	4060	5639
Causal edge	78	123

Figure A.2: Results on Gemma2 using guided chain of thought, compared to the results using no chain of thought (shown in Figure 3). Green indicates the guided CoT reached a higher AUROC than no CoT and pink indicates it reached a lower AUROC than no CoT.

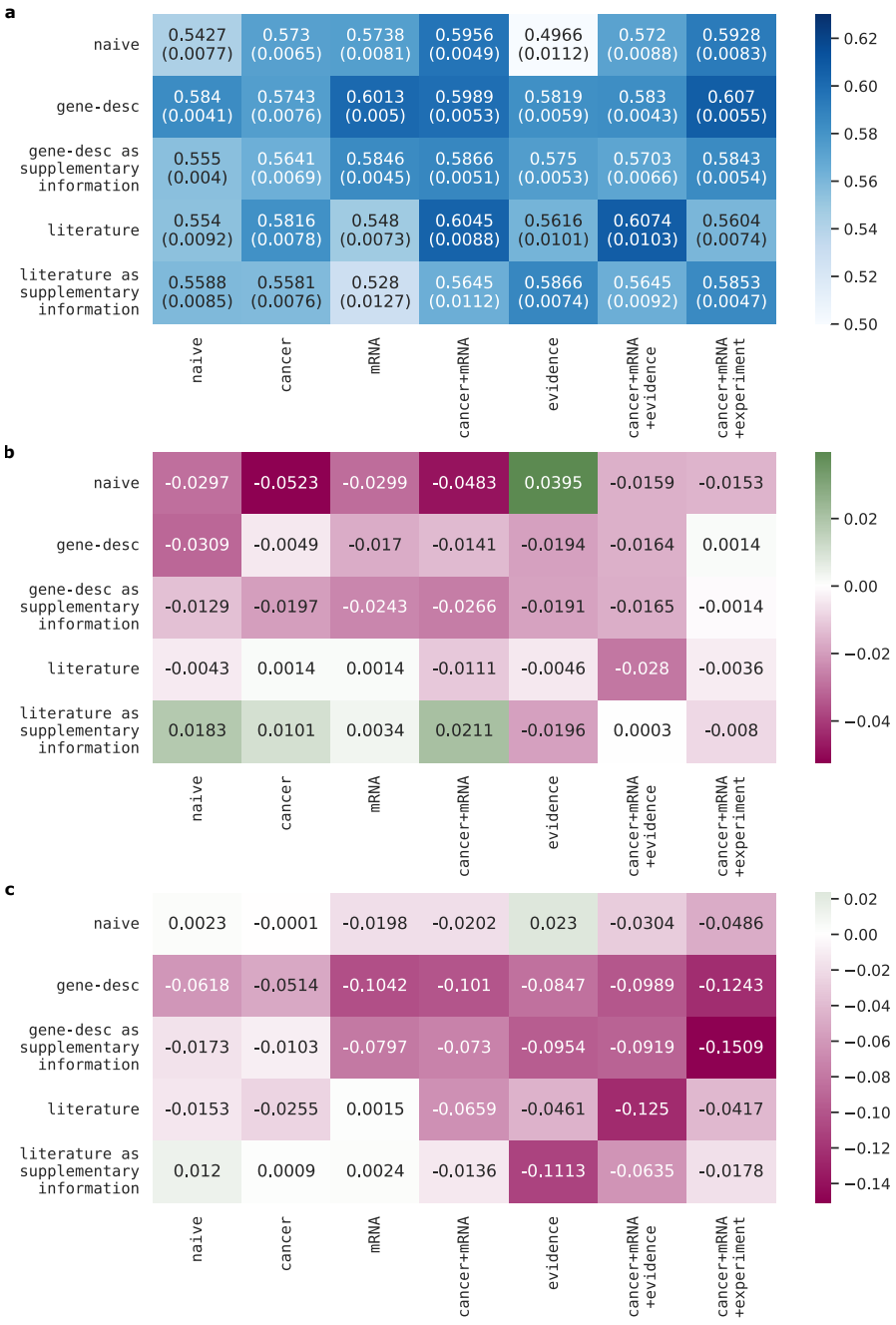

Figure A.3: Results on Gemma2 for all combinations of prompt variants when computing the transitive closure over the predictions. **a**) The results for no CoT, showing the mean AUROC over 10 repetitions, with the standard error given in brackets. **b**) The results for simple CoT, compared to the results using no CoT. **c**) The results for guided CoT, compared to the results using no CoT.

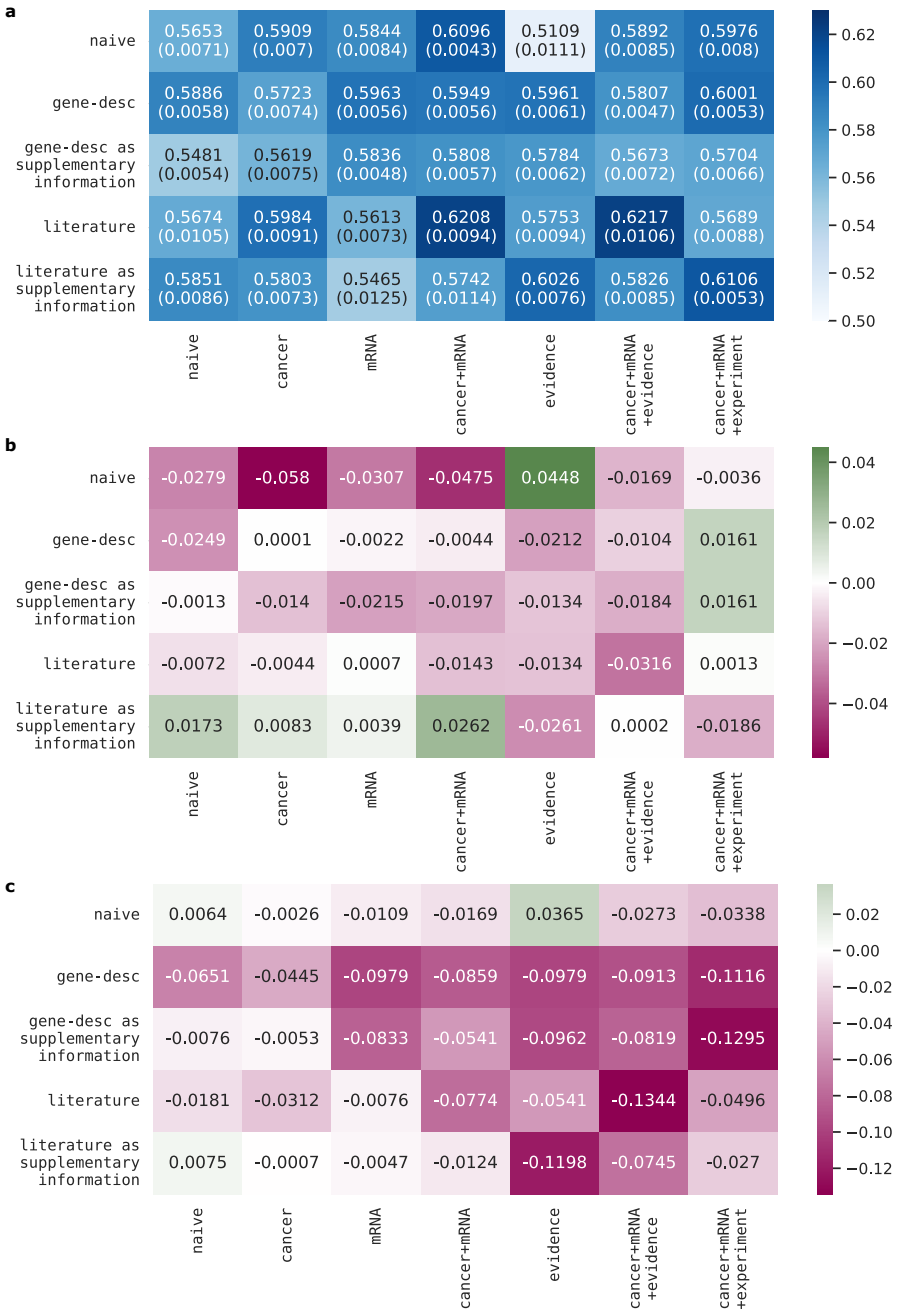

Figure A.4: Results on Gemma2 for all combinations of prompt variants when computing the transitive closure over both the predictions and the ground truth. **a**) The results for no CoT, showing the mean AUROC over 10 repetitions, with the standard error given in brackets. **b**) The results for simple CoT, compared to the results using no CoT. **c**) The results for guided CoT, compared to the results using no CoT.

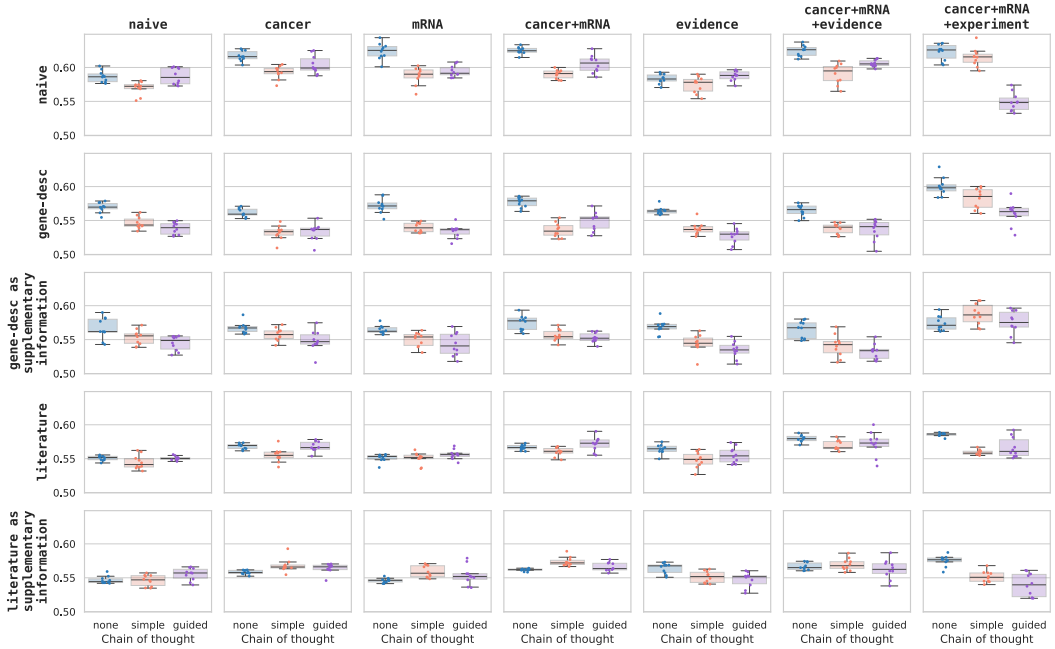

Figure A.5: Results on Gemma2 for all combinations of prompt variants (different contexts along each column, different gene-specific information along each row), shown as boxplots. The AUROCs for each of the 10 repetitions are plotted on the y axes.