Evaluating bivariate causal statements based on mutual compatibility

Erik Jahn [1 2] **Dominik Janzing** [2]

Abstract

In many real-world systems, causal ground truth is hard to obtain, making it challenging to evaluate expert statements about causal effects. In this paper, we propose new methods for evaluating lists of bivariate causal statements in the context of linear structural causal models and causal graphs. Our approach is based on a mathematical formalization of mutual compatibility, measuring the extent to which a collection of bivariate statements is consistent with a single plausible multivariate causal model. We provide theoretical and empirical evidence that our methods can successfully distinguish correct and incorrect causal statements, and we demonstrate their practical applicability by analyzing causal statements obtained from large language models. Our work aims to provide a first foundation for assessing the reliability of causal information derived from human experts or artificial intelligence in settings where alternative forms of validation are unavailable.

1. Introduction

Causal inference in critical applied fields such as medicine or economics still relies heavily on expert knowledge (Gururaghavendran & Murray, 2024). Such knowledge is often expressed in the form of pairwise cause–effect judgments. Indeed, there is empirical evidence that humans find it easier to assess bivariate causal relations than to infer entire multivariate causal graphs (Tatlidil et al., 2025). A similar pattern can be observed when querying large language models (LLMs) for causal relationships (Kiciman et al., 2024), an approach that has attracted substantial recent attention (Ma, 2025). While using LLMs for causal inference has shown impressive practical success, there are virtually no guarantees regarding the quality of their outputs. In contrast, traditional causal discovery algorithms typically come with well-understood consistency guarantees in the large-sample limit, provided their assumptions hold—nothing comparable exists for LLMs (or humans). This raises a fundamental question: how can we trust causal statements originating from humans or AI models with limited interpretability?

In many real-world settings, obtaining causal ground truth through experimentation is either infeasible or prohibitively expensive, leaving no straightforward answer. Unlike standard machine learning problems, performance often cannot simply be evaluated on a test set. A recent approach (Faller et al., 2024) instead proposes evaluating multivariate causal models in the absence of causal ground truth based on self-compatibility: the extent to which they generate causal models on subsets of variables that do not contradict one another. While this framework applies whenever one has access to a causal discovery method capable of producing model estimates for arbitrary subsets of variables, we focus on an even more restricted setting in which only bivariate causal statements are available. This scenario is directly motivated by causal knowledge elicited from humans or large language models, which often struggle to provide reliable estimates of full multivariate causal graphs. Inspired by Faller et al. (2024), we propose to evaluate a list of bivariate causal statements over a set of n variables based on their mutual compatibility.

We develop this idea at two levels of causal information. First, we consider bivariate causal statements that quantify total pairwise causal effects, assuming a linear causal ground truth. We show that, in the absence of faithfulness, any arbitrary list of such bivariate causal statements corresponds to a unique multivariate causal model whose marginalizations exactly coincide with the given statements. As a consequence, there are no hard compatibility constraints for such lists beyond, possibly, consistency with an acyclic causal ordering. This motivates measuring compatibility based on the plausibility of the uniquely induced multivariate causal model. We postulate that a plausible multivariate causal model should exhibit less confounding than its bivariate marginal models, as it can explain more of the observed correlations through causal pathways involving observed variables that are unobserved in the marginals. We formalize this intuition mathematically, leading to a measure of model plausibility. In synthetic experiments, we demonstrate that this measure successfully distinguishes correct

[1]California Institute of Technology, USA [2]Amazon Research Tübingen, Germany. Correspondence to: Erik Jahn <ejahn@caltech.edu>.

Preprint. January 21, 2026.

from incorrect bivariate causal statements. We further apply our approach to causal statements over real-world variables obtained from LLMs and show that the resulting scores vary substantially across different model architectures.

Second, we consider the case where only qualitative bivariate causal statements are available, indicating the presence or absence of a causal effect and of confounding. Under the faithfulness assumption, such statements are subject to hard graphical compatibility constraints, following Faller et al. (2024). In this setting, we develop a compatibility score that measures the number of violations of these constraints. Synthetic experiments again show that this score reliably approximates the true number of errors in a list of bivariate causal statements.

The remainder of the paper is structured as follows. In Section 2, we develop our approach for measuring the compatibility of quantitative bivariate causal statements. Section 3 addresses graphical causal statements. Additional details on proofs and experiments are provided in the appendix.

2. Soft compatibility for linear bivariate causal statements

2.1. Linear structural equation models

In this section, the causal models that we consider are *linear structural equation models (SEMs)*. A linear SEM for n observable variables $X_1, \dots X_n$ is specified by a probability distribution of a vector of n noise variables $\mathbf{N} = (N_1, \dots, N_n)$ and a matrix of causal coefficients $\Gamma \in \mathbb{R}^{n \times n}$. We make the common *acyclicity* assumption and require Γ to be strictly lower-triangular. Then, the vector of observable variables $\mathbf{X} = (X_1, \dots, X_n)$ is modeled by the equation

$$\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N}. \tag{1}$$

Since Γ is lower-triangular, $(I - \Gamma)$ is invertible, so $\mathbf{X} = (I - \Gamma)^{-1}\mathbf{N}$ is the unique solution to equation (1). We do not assume the entries of the noise vector to be independent, therefore allowing for unmeasured confounding.

The causal interpretation of the SEM is that it does not only determine the probability distribution of $\mathbf{X}$ but also its distribution under arbitrary interventions. An *(atomic) intervention* (or experiment) sets a variable X_j to a specific value $x_j \in \mathbb{R}$. The distribution of the other variables is then obtained from solving the equation

$$\mathbf{X}' = (I - \mathrm{diag}(\mathbf{e}_j))\Gamma \mathbf{X}' + (I - \mathrm{diag}(\mathbf{e}_j))\mathbf{N} + x_j \mathbf{e}_j, \tag{2}$$

Where I is the $(n \times n)$-dimensional identity matrix, and $\mathbf{e}_j$ is the j'th unit vector. Note that equation (2) indeed results in the value of X_j being x_j and leaves the equations (i.e. the causal mechanism) for all other variables X_i the same. We denote the probability distribution of the solution $\mathbf{X}'$ by $P(\mathbf{X} \mid \mathrm{do}\, X_j = x_j)$ and accordingly its expectation by $\mathbb{E}[\mathbf{X} \mid \mathrm{do}\, X_j = x_j]$. SEMs can be graphically represented by acyclic directed mixed graphs (ADMGs) as follows:

Definition 2.1. Consider the SEM $\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N}$. We define the ADMG G corresponding to $(\Gamma, \mathbf{N})$ as the graph on $X_1, \dots, X_n$ that has

- a directed edge from X_i to X_j iff $\Gamma_{ji} \neq 0$;
- a bidirected edge between X_i and X_j iff N_i and N_j are not independent.

We finish this section, by showing how SEMs are marginalized.

Lemma 2.2 (marginalized SEM, see for instance Lemma 7 in Hyttinen et al. (2012))**.** *Consider the SEM $\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N}$ and let $\mathbf{Y}, \mathbf{Z}$ be a partition of the variables of $\mathbf{X}$. Then, there exists an error vector $\tilde{\mathbf{N}}$ such that all the observational and interventional distributions of the solution to the SEM*

$$\mathbf{Y} = \left(\Gamma_{\mathbf{YY}} + \Gamma_{\mathbf{YZ}}(I - \Gamma_{\mathbf{ZZ}})^{-1}\Gamma_{\mathbf{ZY}}\right)\mathbf{Y} + \tilde{\mathbf{N}}.$$

align with the observational and interventional distributions of $\mathbf{X}$, marginalized to the variables $\mathbf{Y}$.

2.2. Bivariate causal statements

Definition 2.3 (Bivariate causal statement)**.** A *bivariate causal statement* for a pair of variables (X_i, X_j) specifies

1. a causal direction $i \to j$;
2. a linear causal coefficient $\alpha_{ij} \in \mathbb{R}$.

Our goal is to develop an approach to evaluate the quality of a complete list of $\binom{n}{2}$ bivariate causal statements for a fixed set of variables $X_1, \dots, X_n$ based on their mutual compatibility. We will assume that all causal directions are compatible with the ordering of $X_1, \dots, X_n$ given by their indices, i.e. each statement says $i \to j$ for $i < j$. See Section 3 for how to handle violations of this assumption.

Moreover, we assume that the joint probability distribution of $(X_1, \dots, X_n)$ is known (e.g. observed from samples). Note that then, a bivariate causal statement assigning the causal coefficient α_{ij} to the pair (X_i, X_j) is equivalently proposing the bivariate structural equation model

$$X_j = \alpha_{ij} X_i + \tilde{N}_{ij}, \tag{3}$$

where the distribution of $\tilde{N}_{ij}$ is simply given by the distribution of $X_j - \alpha_{ij} X_i$.

Our first result is that in this setting, where the causal orderings are compatible, there are no further *hard compatibility*

constraints: any complete list of bivariate causal statements corresponds to a unique multivariate SEM whose pairwise marginals agree with the proposed bivariate SEMs. To show this, let us encode a list of bivariate statements by the unit lower-triangular matrix $A \in \mathbb{R}^{n\times n}$ whose entries are given by

$$A_{ii} = 1, \qquad A_{ij} = \alpha_{ji} \ \ (i > j), \qquad A_{ij} = 0 \ \ (i < j).$$

Lemma 2.4 (Existence of a unique compatible SEM)**.** *Let A be a unit lower-triangular matrix encoding a complete list of bivariate causal statements consistent with the ordering of $(X_1, \dots, X_n)$ and define $\Gamma = I - A^{-1}$. Then, the multivariate structural equation model*

$$\mathbf{X} = \Gamma\mathbf{X} + \mathbf{N},$$

where the distribution of $\mathbf{N}$ is defined as the distribution of $(I - \Gamma)\mathbf{X}$, is the unique linear SEM whose pairwise marginal submodels are given by

$$X_j = \alpha_{ij}X_i + \tilde{N}_{ij}.$$

The proof of Lemma 2.4 can be found in Appendix A.1.

2.3. Confounding postulate

Given a system of n observed variables $(X_1, \dots, X_n)$ with an arbitrary distribution, Lemma 2.4 tells us that any list of bivariate causal claims for this system is indeed explained by a potential multivariate causal model. Hence, without further experimentation or knowledge of causal ground truth, strict falsification of these bivariate claims is not possible. Instead, we propose to measure the *degree of compatibility* of a list of bivariate causal claims based on how *plausible* the unique composite multivariate SEM is that they induce. Of course, plausibility of a causal model is not simple to define and any such attempt needs to be subject to debate. In this paper, we suggest a necessary criterion for plausibility based on the following postulate:

Assumption 2.5. *A generic multivariate causal model should have a smaller or equal amount of confounding compared to all its pairwise marginal models.*

This postulate is based on the following intuition: whenever one marginalizes a multivariate causal model, some variables change from observed to unobserved and these variables introduce additional confounding in the marginal model. Hence, the amount of confounding in the marginal model generally increases, unless there is strong cancellation between the new and the old confounding. This intuition is aligned with the basic machine learning principle that including additional features reduces bias. While logically independent from the faithfulness condition, our postulate is strongly related to it. Indeed, examples of causal models that violate Assumption 2.5 usually rely on (near-) faithfulness violations. The faithfulness condition is a standard assumption in many causal discovery settings, usually justified by the fact that the set of unfaithful causal models has measure zero for any continuous parametrization. However, the reader might also be familiar with the fact that near-faithfulness violations can be quite common (Uhler et al., 2013) and might wonder if this is evidence that Assumption 2.5 is less likely to hold. We argue that, while *some* near-faithfulness violations may likely occur in large causal models, *most* causal paths should not cancel with each other and this should still make the total amount of confounding in the marginal models larger compared to the multivariate model. We provide further empirical and theoretical evidence in favor of our postulate in sections 2.5 and 2.6.

2.4. Confounding measures

To translate our postulate into a computable score, we need to mathematically formalize the amount of confounding in a causal model first. There is no uniquely accepted notion for this (see Reddy & Balasubramanian (2024)). We need a measure that makes it possible to compare the confounding in a multivariate model to the confounding in its pairwise marginals. In a bivariate SEM $X_j = \alpha_{ij}X_i + \tilde{N}_{ij}$, we have

$$\mathrm{Cov}(X_i, X_j) = \alpha_{ij}\,\mathrm{Var}(X_i) + \mathrm{Cov}(X_i, \tilde{N}_{ij}).$$

That is, the covariance of X_i and X_j naturally decomposes into two terms, the first of which can be attributed to the causal effect from X_i to X_j and the second of which is due to confounding between X_i and X_j. This leads to the following natural definition:

Definition 2.6 (amount of confounding in bivariate models)**.** For a bivariate linear SEM given by $X_j = \alpha_{ij}X_i + \tilde{N}_{ij}$, we define the amount of confounding between X_i and X_j as

$$\begin{aligned}\mathcal{C}(X_i, X_j, \alpha_{ij}) &= \mathrm{Cov}(X_i, \tilde{N}_{ij})^2 \\ &= \left(\mathrm{Cov}(X_i, X_j) - \alpha_{ij}\,\mathrm{Var}(X_i)\right)^2.\end{aligned}$$

We simply square the part of the covariance between X_i and X_j that can be attributed to hidden variables to get a confounding score that is non-negative and reaches its minimum at zero if and only if the noise term $\tilde{N}_{ij}$ of X_j is uncorrelated with X_i. Up to normalization, this notion has been studied before in the bivariate setting (Janzing & Schölkopf, 2018; Janzing & Schölkopf, 2018). In order to extend it now to the multivariate setting, consider a SEM $\mathbf{X} = \Gamma\mathbf{X} + \mathbf{N}$. According to Wright's path tracing rules (Wright, 1934), the covariance between X_i and X_j decomposes as

$$\mathrm{Cov}(X_i, X_j) = \sum_{k \le i} \mathrm{Var}(X_k) \cdot \sum_{\substack{P_1: k \to i \\ P_2: k \to j \\ P_1 \cap P_2 = \emptyset}} \Gamma^{P_1} \Gamma^{P_2} + \sum_{\ell \neq k \le j} \mathrm{Cov}(N_\ell, N_k) \cdot \sum_{\substack{P_1: \ell \to i \\ P_2: k \to j \\ P_1 \cap P_2 = \emptyset}} \Gamma^{P_1} \Gamma^{P_2} \tag{4}$$

In the sums above, each $P : k \to i$ for $k \le i$ runs over all possible directed paths $(t_1, \ldots, t_m)$ in the ADMG that represents the SEM, where $k = t_1 < t_2 < \cdots < t_m = i$. Two paths $P_1 = (t_1, \ldots, t_m)$ and $P_2 = (q_1, \ldots, q_r)$ are disjoint, i.e. $P_1 \cap P_2 = \emptyset$ if they do not share a vertex, except for an endpoint. Moreover, for $P = (t_1, \ldots, t_m)$, we define

$$\Gamma^P = \prod_{a=1}^{m} \Gamma_{t_{a+1}, t_a}$$

to be the product of the causal coefficients along the path P. Here, the intuition is that the first term in equation (4) quantifies the contribution of all direct causal paths and all back-door paths through observed variables to the covariance of X_i and X_j (i.e. all paths that d-connect X_i and X_j in the ADMG, using only directed edges), whereas the second term quantifies the contribution of all back-door paths through at least one unobserved vertex to the covariance (i.e. all d-connecting paths between X_i and X_j that use at least one bidirected edge).

Definition 2.7 (amount of confounding in multivariate models)**.** For a multivariate linear SEM given by $\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N}$, we define the amount of confounding between X_i and X_j as

$$\mathcal{C}(X_i, X_j, \Gamma) = \left(\sum_{\ell \neq k \le j} \mathrm{Cov}(N_\ell, N_k) \cdot \sum_{\substack{P_1: \ell \to i \\ P_2: k \to j \\ P_1 \cap P_2 = \emptyset}} \Gamma^{P_1} \Gamma^{P_2} \right)^2$$

Note that for two variables, this definition exactly aligns with Definition 2.6. Now, we can quantify the compatibility of a list of bivariate statements as the degree to which Assumption 2.5 holds.

Definition 2.8 (compatibility score)**.** Let A be a unit lower-triangular matrix encoding a complete list of bivariate causal statements consistent with the ordering of $(X_1, \ldots, X_n)$ and define $\Gamma = I - A^{-1}$. We define the *compatibility score*:

$$\mathrm{comp}(A) = \sum_{i<j} \mathcal{C}(X_i, X_j, A_{ij}) - \mathcal{C}(X_i, X_j, \Gamma).$$

Note that the exact size of the score depends on the normalization of the variables. When applying the score in practice, we will always normalize the variances of all variables to 1. Still, the magnitude of the score can be hard to interpret, but its sign very clearly relates to Assumption 2.5. The compatibility score is negative if and only if Assumption 2.5 is violated. Hence, we view a negative compatibility score for a list of bivariate causal claims as "soft" evidence against the validity of these statements.

2.5. Theoretical analysis

A negative compatibility score can only be used as evidence against a list of bivariate causal statements if most "plausible" causal models indeed satisfy Assumption 2.5. Our first argument supporting this claim is by showing that the expected compatibility score of the true bivariate causal statements is indeed positive for a wide range of plausible distributions on causal models. Here, we consider causal models given by linear SEMs, where the distribution of the noise variables is Gaussian and centered at zero. Hence, the model is fully described by a strictly lower triangular matrix Γ of causal coefficients and a symmetric positive definite covariance matrix Σ_N of the noise variables.

Assumption 2.9. We consider probability distributions over (Γ, Σ_N) satisfying two natural properties:

1. (Unbiasedness) For all i, j, we have $\mathbb{E}[\Gamma_{ij}] = 0$;
2. (Independence of causal mechanisms) The noise correlation matrix and all the entries of Γ are mutually independent.
3. (Non-degeneracy) We have $\Sigma_N \succ 0$ almost surely and $\mathrm{Var}(\Gamma_{ij}) > 0$ for all $i > j$.

Theorem 2.10. *For $n \ge 3$, let (Γ, Σ_N) specify a random n-dimensional linear Gaussian causal model drawn from a distribution that satisfies Assumption 2.9. Define $A = (I - \Gamma)^{-1}$ to be the matrix of total bivariate causal effects. Then,*

$$\mathbb{E}\left[\mathrm{comp}(A)\right] > 0.$$

2.6. Experiments

Our first results are from synthetic experiments. We sample causal ground truth in the form of linear Gaussian SEMs with n observed variables, m hidden variables and sparsity parameter p as follows:

1. Draw random causal coefficients for $n + m$ variables from a normal distribution and error variances for each variable from an exponential distribution;
2. Set each causal coefficient to 0 independently with probability $1 - p$;

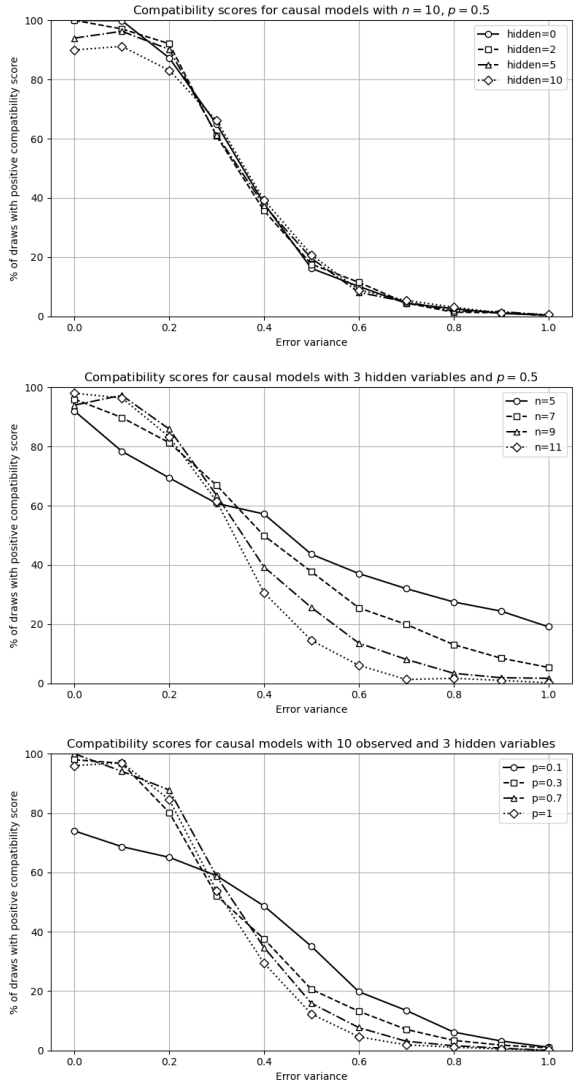

Figure 1. Percentage of positive compatibility scores for lists of bivariate causal statements on synthetic linear models with increasing amount of error.

3. given the remaining causal coefficients, calculate the covariance matrix of all $n + m$ variables, assuming independence between the error terms, resulting in a linear SEM without confounding;

4. Select m variables to be the hidden variables at random and marginalize the SEM to the remaining n variables.

Given the randomly generated causal ground truth, we can now compute the compatibility scores of lists of bivariate causal statements with decreasing quality by starting with the true bivariate causal statements and adding random error with increasing variance σ. The results can be seen in Figure 3. For each data point, we draw 50 different causal models and for each causal model, we draw the random error in the bivariate statements 20 times. Then, we report the percentage of draws with a positive compatibility score. We get three different figures for varying each of the main model parameters - the number m of hidden variables, the number n of observed variables, and the sparsity p - separately, while keeping the other two fixed. First of all, across basically all tested combinations of model parameters, one can see that our compatibility score successfully distinguishes correct from incorrect statements, with the percentage of statements with positive compatibility scores strictly decreasing as the error increases. In particular, the data points at error zero provide further empirical evidence for the validity of Assumption 2.5 as most true sampled causal models have positive compatibility scores. The only slight exception is when the models become very sparse ($n = 10$, $p = 0, 1$ in the last figure). This can be explained by the fact that very sparse multivariate models do not have many back-door paths, therefore making the amount of confounding in the multivariate model very similar to the amount of confounding in the bivariate models. In general, the compatibility score is quite unaffected by the number of hidden variables in the true causal model - but it distinguishes more successfully between correct and incorrect causal statements as the number of observed variables and the density of the true causal model increases. This is aligned with our intuition, as in both cases the number of observed back-door paths in the multivariate model increases, which are unobserved in the bivariate models and therefore add to the difference of the confounding between the multivariate and the bivariate models.

For our second set of experiments, we evaluated the compatibility scores for lists of bivariate causal statements that were obtained from large language models (LLMs) for real-world data. We chose the following five variables from the gapminder dataset (Gapminder Foundation, 2023):

1. GDP per capita,
2. average number of years that women spent in school,
3. average fertility rate,
4. average child mortality rate,
5. average life expectancy.

The dataset consists of datapoints for the years from 1970-2009 and for 179 different countries. For each pair of variables above, we asked multiple large language models to estimate the total causal effect between the variables based on their description and the empirical correlation matrix (see full prompt details in Appendix C). The results can be seen in Figure 2. While some LLMs do not even beat a baseline obtained from choosing causal coefficients between 1 and -1 at random, others do remarkably well on

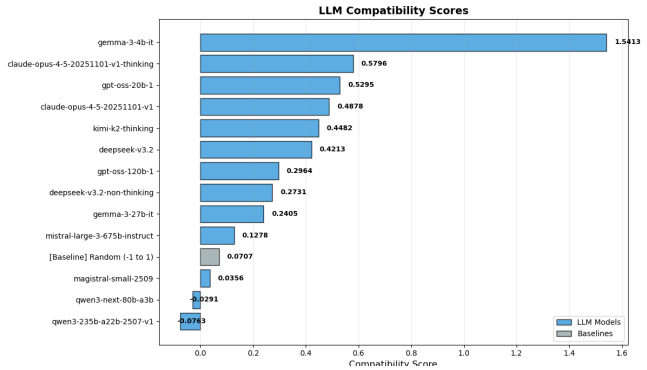

Figure 2. Compatibility scores for lists of bivariate statements obtained from different LLMs.

the dataset, with Google's Gemma 3B model obtaining the highest score.

3. Hard compatibility for graphical bivariate causal statements

3.1. Graphical causal statements

In situations where exact quantitative causal statements are infeasible to obtain, it can still be desirable to make qualitative causal statements about the existence and direction causal effects. In this section, the causal models we consider are no longer linear SEMs, but instead acyclic directed mixed graphs (ADMGs) (that could correspond to an SEM, see Definition 2.1). In general, an ADMG is a graph on a vertex set V that can have both directed edges and bidirected edges and the set of directed edges is acyclic. ADMGs have been widely studied as graphical causal models for causal structures with confounding (Pearl & Verma, 1995; Spirtes et al., 2001). In general, directed edges in the ADMG correspond to directed causal effects and bidirected edges correspond to correlation induced by confounding. Moreover, there is a well-defined notion of marginalization for ADMGs.

Definition 3.1. Let v, w be vertices in an ADMG G. A path between v, w is a *confounding path* if both v and w are adjacent to an arrowhead of the path and no intermediate vertex is adjacent to two arrowheads of the path (e.g. $v \leftrightarrow \rightarrow \rightarrow w$ or $v \leftarrow \leftrightarrow \rightarrow w$).

Definition 3.2 (marginal ADMG, see Richardson et al. (2023))**.** Let G be an ADMG on the vertex set $V \cup L$, where V and L are disjoint. The marginal ADMG H on V is the ADMG with vertex set V that contains all edges of G within V and additionally

1. a directed edge $v \to w$, whenever there is a directed path from v to w in G with intermediate vertices in L;
2. a bidirected edge $v \leftrightarrow w$, whenever there is a confounding path from v to w in G with intermediate vertices in L.

Under the faithfulness condition, the ADMG that corresponds to a marginalized SEM is exactly obtained by marginalizing the ADMG of the full SEM.

Definition 3.3 (bivariate graphical causal statements)**.** A *bivariate graphical causal statement* for two variables X_i, X_j specifies an ADMG on these two variables, i.e. the existence and direction of a direct causal effect and the existence of confounding. Given a complete list of $\binom{n}{2}$ bivariate graphical causal statements for n variables $X_1, \ldots, X_n$, we define the *statement graph* $\mathcal{G}$ to be the mixed graph that is simply the union of all bivariate ADMGs.

Already Faller et al. (2024) introduced the concept of graphical compatibility, stating that a list of small causal ADMGs on a common vertex set is *graphically compatible* if they coincide with the marginalizations of a single large ADMG on the vertex set. In our specific situation, we have the following characterization of graphical compatibility

Lemma 3.4. *A complete list of graphical bivariate causal statements with statement graph $\mathcal{G}$ is graphically compatible if and only if*

1. *the directed part of $\mathcal{G}$ is acyclic;*
2. *the directed part of $\mathcal{G}$ is transitively closed, i.e. if there is a directed path from X_i to X_j, then there must be an edge from X_i to X_j;*
3. *if there is a confounding path between X_i and X_j, then there must be a bidirected edge between X_i and X_j.*

3.2. Compatibility score

Let $d(\mathcal{G}, \mathcal{G}^*)$ denote the Hamming distance between two mixed graphs, that is the number of edge deletions and additions needed to transform $\mathcal{G}$ into $\mathcal{G}^*$.

Definition 3.5. For a statement graph $\mathcal{G}$ summarizing a complete list of bivariate graphical causal statements, we define the compatibility score $\mathrm{comp}(\mathcal{G})$ to be the minimum Hamming distance $\min_{\mathcal{G}^*} d(\mathcal{G}, \mathcal{G}^*)$ over all mixed graphs $\mathcal{G}^*$ that satisfy the three properties stated in Lemma 3.4.

This is similar to the graphical incompatibility score defined by Faller et al. (2024) but not quite the same. In our case, we do not have an inferred multivariate ADMG to which we can compare the bivariate ADMGs. Instead, we compare with the ADMG that is consistent with as many bivariate graphical statements as possible. This definition also extends naturally to the case where we do not have a complete list of bivariate graphical causal statements. Given an incomplete statement graph $\hat{\mathcal{G}}$, the compatibility score

$\mathrm{comp}(\hat{\mathcal{G}})$ can be defined as the minimum Hamming distance $\min_{\mathcal{G},\mathcal{G}^*} d(\mathcal{G},\mathcal{G}^*)$, where $\mathcal{G}$ runs over all mixed graphs that extend $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ runs over all mixed graphs with the properties in Lemma 3.4.

The problem of exactly computing $\mathrm{comp}(\mathcal{G})$ is NP-complete. This is because it is strictly harder than the following two NP-complete problems (in fact, these problems are even hard to approximate):

1. MINIMUM FEEDBACK ARC: Given a directed graph G, find the minimum number of edge deletions to make it acyclic. See (Karp, 1972) for a hardness proof.
2. TRANSITIVITY EDITING: Given a directed graph G, find the minimum number of edge deletions and additions to make it transitively closed. See Weller et al. (2012) for a hardness proof.

Does this mean that computing $\mathrm{comp}(\mathcal{G})$ is hopeless? Intuitively, the computational hardness of the optimization problems above comes from regimes where the size of the optimal solution set is large. However, when $\mathrm{comp}(\mathcal{G})$ is large, we do not need to accurately compute or even approximate it: any sufficiently large lower bound already certifies that the graph $\mathcal{G}$ is highly incompatible and therefore provides falsification for the bivariate causal statements.

In the regime, where $\mathrm{comp}(\mathcal{G})$ is small, exact computation becomes in principle tractable. Indeed, a brute-force algorithm can decide in polynomial time whether $\mathrm{comp}(\mathcal{G})$ is smaller than a constant bound k or not. However, since k does appear in the exponent of the run-time of this algorithm, it quickly becomes impractical. Motivated by this insight, we design a heuristic algorithm for computing $\mathrm{comp}(\mathcal{G})$ that prioritizes correctness in the low-score regime while remaining computationally efficient.

3.3. Heuristic computation of the compatibility score

The true optimal solution for the minimization problem in Definition 3.5 depends on the complex interaction between all three properties of Lemma 3.4. As a first heuristic simplification, we propose a decoupling into three optimization problems that can be solved separately. This means, given a mixed graph $\mathcal{G}$ with directed part $D = D(\mathcal{G})$, we go through the following steps:

1. Identify a small set of edges $E_{\mathrm{cycles}} \subseteq E(D)$ whose deletion makes D acyclic (i.e. approximate the minimum feedback arc problem with instance D).
2. Find a small set of edge deletions E_{del} and edge additions E_{add} such that $(D \cup E_{\mathrm{add}}) \setminus (E_{\mathrm{cycles}} \cup E_{\mathrm{del}})$ is acyclic and transitively closed (i.e. approximate the transitivity editing problem with instance $D \setminus E_{\mathrm{cycles}}$)
3. Find a small set of bidirected edge deletions B_{del} and bidirected edge additions B_{add} such that the graph $(\mathcal{G} \cup E_{\mathrm{add}} \cup B_{\mathrm{add}}) \setminus (E_{\mathrm{cycles}} \cup E_{\mathrm{del}} \cup B_{\mathrm{del}})$ also satisfies property 3 of Lemma 3.4.

Algorithm 1 Greedy Feedback Arc Set (GreedyFAS)

Input: Directed graph $D = (V, E)$
Output: Feedback arc set $E_{\mathrm{cycles}} \subseteq E$
Initialize empty lists $L \leftarrow [\,]$, $R \leftarrow [\,]$
Initialize $H \leftarrow D$
while $V(H) \neq \emptyset$ **do**
  if H has a source v (i.e. $d_H^{\mathrm{in}}(v) = 0$) **then**
    Remove v from H
    Append v to the end of L
  else if H has a sink v (i.e. $d_H^{\mathrm{out}}(v) = 0$) **then**
    Remove v from H
    Prepend v to the beginning of R
  else
    Choose $v \in V(H)$ maximizing $d_H^{\mathrm{out}}(v) - d_H^{\mathrm{in}}(v)$
    Remove v from H
    Append v to the end of L
  end if
end while
Let π be the linear ordering given by concatenation $L \circ R$
$E_{\mathrm{cycles}} \leftarrow \{(u \to v) \in E : \pi(u) > \pi(v)\}$
return E_{cycles}

After these three steps, we report $|E_{\mathrm{cycles}}|+|E_{\mathrm{add}}|+|E_{\mathrm{del}}|+|B_{\mathrm{add}}|+|B_{\mathrm{del}}|$ as the heuristic approximation for $\mathrm{comp}(\mathcal{G})$. Note that apart from splitting the problem of computing $\mathrm{comp}(\mathcal{G})$ into three different subproblems, this approach introduces another simplification: it considers the set of directed edges fixed after the first two steps and only modifies bidirected edges in the third step to obtain property 3 of Lemma 3.4. While this might introduce some additional error in the final solution, this approach allows for fast heuristic solutions for each step.

STEP 1

To solve the MINIMUM FEEDBACK ARC problem we use the GreedyFAS algorithm (see Algorithm 1), which runs in linear time, is easy to implement and has good empirical performance (Eades et al., 1993; Simpson et al., 2016). In the description of the algorithm, $d_H^{\mathrm{in}}(v)$ denotes the in-degree of a vertex v with respect to the graph H and similarly, $d_H^{\mathrm{out}}(v)$ denotes the out-degree.

STEP 2

Let $D' = D \setminus E_{\mathrm{cycles}}$ be the acyclic directed graph obtained from step 1. We design our own greedy algorithm to solve the transitivity editing problem, see Algorithm 2. Here $tc(D')$ denotes the *transitive closure* of D', which is the

Algorithm 2 Greedy Transitivity Editing (GreedyTE)

Input: Acyclic directed graph $D = (V, E)$
Output: Edge sets E_{del} and E_{add} such that $D' = (V, (E \cup E_{\text{add}}) \setminus E_{\text{del}})$ is transitively closed
Initialize $E_{\text{del}} \leftarrow \emptyset$
Initialize $H \leftarrow D$
repeat
  Initialize $bestGain \leftarrow 0$
  Initialize $e^\star \leftarrow \emptyset$
  for each edge $e \in E(H)$ **do**
    Compute $gain \leftarrow |tc(H)| - |tc(H \setminus e)| - 1$
    if $gain > bestGain$ **then**
      $bestGain \leftarrow gain$
      $e^\star \leftarrow e$
    end if
  end for
  if $bestGain > 0$ **then**
    $E_{\text{del}} \leftarrow E_{\text{del}} \cup \{e^\star\}$
    $H \leftarrow H \setminus e^\star$
  end if
until $bestGain = 0$
Define $D' \leftarrow D \setminus E_{\text{del}}$
Set $E_{\text{add}} \leftarrow E(tc(D')) \setminus E(D')$
return $(E_{\text{del}}, E_{\text{add}})$

smallest digraph that contains D and is transitively closed. The idea behind Algorithm 2 is to greedily delete edges that reduce the distance from the current digraph to its transitive closure as much as possible. Once that distance cannot be reduced anymore, one simply adds all the remaining edges needed to get to the transitive closure.

STEP 3

After executing the first two steps, we now have transformed the directed part D of a statement graph $\mathcal{G}$ into a DAG $D' = (D \setminus (E_{\text{cycles}} \cup E_{\text{del}})) \cup E_{\text{add}}$. that satisfy the first two properties of Lemma 3.4. To complete our heuristic computation of $\text{comp}(\mathcal{G})$, we now keep the directed part D' fixed and only operate on the bidirected edges. Let us define the *confounding path closure* $cpc(G)$ of a mixed graph G to be the graph obtained from G by adding bidirected edges between each pair of vertices that is connected by a confounding path. Now, we can essentially reuse Algorithm 2 after replacing the transitive closure with the confounding path closure to compute an approximation to the closest graph to G that satisfies the third property of Lemma 3.4. We state the precise algorithm as Algorithm 3 in Appendix B.

3.4. Experiments

For our experiments, we sample causal ground truth with n observed variables, m hidden variables and sparsity parameter p as follows:

1. Draw a uniformly random permutation on $\{1, \dots, n+m\}$ to get a causal ordering;
2. Set each directed edge that is consistent with the ordering independently with probability p to get a causal DAG;
3. Choose the set of m hidden variables at random and marginalize the graph to the remaining n observed variables according to Definition 3.2;
4. Obtain the correct bivariate causal statements by further marginalizing the causal graph to each possible pair of variables.

For each list of correct bivariate causal statements, we now introduce a fixed amount of random errors, by randomly selecting a pair of variables and flipping one of the three possible edges between them (each pair of variables has one possible bidirected edge and two possible directed edges - by flipping, we either delete the edge if it was present, or we add the edge if it was absent). The resulting (erroneous) graph is then the statement graph $\mathcal{G}$, to which we apply our heuristic method for computing $\text{comp}(\mathcal{G})$. The compatibility scores for different numbers of errors and variations of the model parameters are displayed in Figure **??**. Ideally, each of the plots would show only straight lines with slope 1 - meaning that the compatibility score directly reflects the true number of errors in the list of statements. This is indeed approximately the case in many regimes, showing that the compatibility score is a good proxy for the true quality of the causal statements. Only when the graph becomes too large or too dense, the compatibility score tends to overestimate the number of errors, indicating that our heuristic algorithms can only find suboptimal solutions. On the other hand, for small or very sparse models, the compatibility score slightly underestimates the number of errors, which can happen in the case where another causal model explains the erroneous statements better than the actual ground truth. While sparsity and model size directly affects the quality of the results, the compatibility scores seem to be quite robust with respect to changes in the number of hidden variables.

4. Discussion

We developed two complementary approaches to assess bivariate causal statements based on their mutual compatibility. For graphical statements, we approximately compute how many of the given statements can be explained by a single consistent multivariate causal graph under the assumptions of acyclicity and faithfulness. The closer this number is to the total number of statements, the higher the resulting

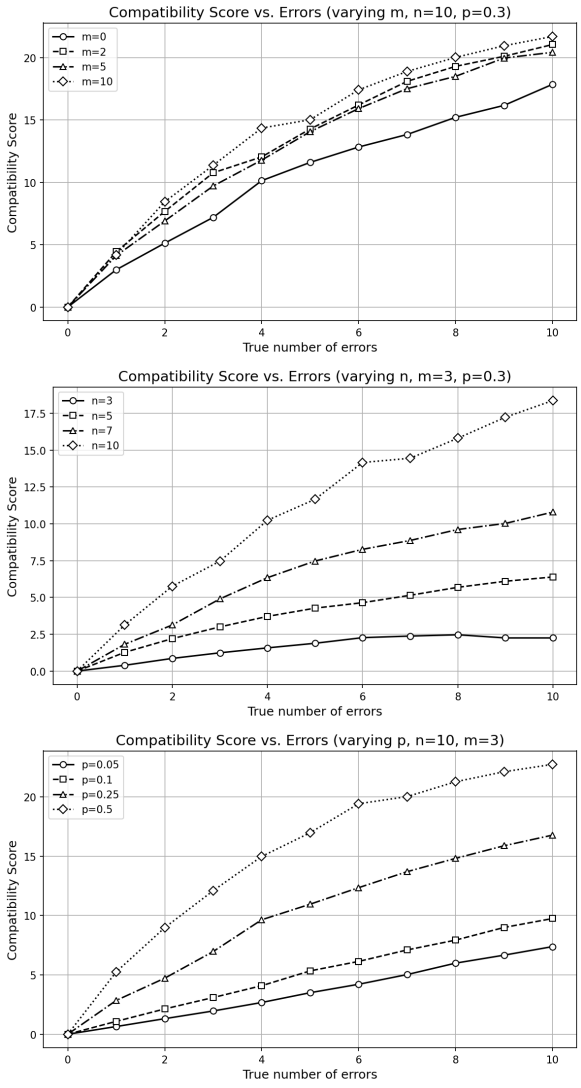

Figure 3. Average compatibility scores for lists of graphical bivariate causal statements with increasing number of errors.

compatibility score. In contrast, in the absence of faithfulness and when quantitative causal statements are available, we measure compatibility by evaluating how well the unique composite multivariate model satisfies our plausibility criterion, namely that it should exhibit less confounding than its bivariate marginals. Importantly, a high compatibility score does not imply causal validity; however, a low score provides concrete evidence of internal inconsistency. To our knowledge, these are the first approaches to evaluating bivariate causal statements without access to either causal ground truth or a causal estimate over the full variable set.

A promising direction for future work is to extend our framework to settings in which causal statements are available not only for pairs of variables but also for small subsets, and where the collection of statements may be incomplete. While graphical compatibility is likely to generalize relatively directly to such settings, extending soft compatibility for quantitative statements may require new ideas, as the existence of a unique composite multivariate model can no longer be guaranteed.

Finally, our experiments with LLM-generated causal statements demonstrate that compatibility scores already provide a meaningful additional dimension for comparing lists of causal claims that would otherwise be indistinguishable in the absence of causal ground truth. We hope that this work can serve as a foundation for improving the reliability and assessment of causal statements in such settings.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Eades, P., Lin, X., and Smyth, W. A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters*, 47(6):319–323, 1993. ISSN 0020-0190. doi: https://doi.org/10.1016/0020-0190(93)90079-O. URL https://www.sciencedirect.com/science/article/pii/002001909390079O.

Faller, P. M., Vankadara, L. C., Mastakouri, A. A., Locatello, F., and Janzing, D. Self-compatibility: evaluating causal discovery without ground truth. In *Proc. 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 4132–4140. PMLR, 2024.

Gapminder Foundation. Gapminder data. https://www.gapminder.org/data/, 2023. Accessed: 2026-01-20.

Gururaghavendran, R. and Murray, E. J. Can algorithms replace expert knowledge for causal inference? a case study on novice use of causal discovery. *American Journal of Epidemiology*, 194(5):1399–1409, 08 2024. ISSN 0002-9262. doi: 10.1093/aje/kwae338. URL https://doi.org/10.1093/aje/kwae338.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(109):3387–3439, 2012. URL http://jmlr.org/papers/v13/hyttinen12a.html.

Janzing, D. and Schölkopf, B. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):1–27, March 2018. doi: 10.

1515/jci-2017-0013. URL https://ideas.repec.org/a/bpj/causin/v6y2018i1p27n2.html.

Janzing, D. and Schölkopf, B. Detecting non-causal artifacts in multivariate linear regression models. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2245–2253. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/janzing18a.html.

Karp, R. M. *Reducibility among Combinatorial Problems*, pp. 85–103. Springer US, Boston, MA, 1972. ISBN 978-1-4684-2001-2. doi: 10.1007/978-1-4684-2001-2_9. URL https://doi.org/10.1007/978-1-4684-2001-2_9.

Kiciman, E., Ness, R. O., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (TMLR)*, August 2024. Selected for presentation at ICLR 2025.

Ma, J. Causal inference with large language model: A survey. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5886–5898, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.327. URL https://aclanthology.org/2025.findings-naacl.327/.

Pearl, J. and Verma, T. S. A theory of inferred causation. In Prawitz, D., Skyrms, B., and Westerståhl, D. (eds.), *Logic, Methodology and Philosophy of Science IX*, volume 134 of *Studies in Logic and the Foundations of Mathematics*, pp. 789–811. Elsevier, 1995. doi: https://doi.org/10.1016/S0049-237X(06)80074-1. URL https://www.sciencedirect.com/science/article/pii/S0049237X06800741.

Reddy, A. G. and Balasubramanian, V. N. Detecting and measuring confounding using causal mechanism shifts. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1): 334 – 361, 2023. doi: 10.1214/22-AOS2253. URL https://doi.org/10.1214/22-AOS2253.

Simpson, M., Srinivasan, V., and Thomo, A. Efficient computation of feedback arc set at web-scale. *Proc. VLDB Endow.*, 10(3):133–144, November 2016. ISSN 2150-8097. doi: 10.14778/3021924.3021930. URL https://doi.org/10.14778/3021924.3021930.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. The MIT Press, 01 2001. ISBN 9780262284158. doi: 10.7551/mitpress/1754.001.0001. URL https://doi.org/10.7551/mitpress/1754.001.0001.

Tatlidil, S., Sloman, S. A., Basu, S., Tran, T., Saxena, S., Kim, M. H., and Bahar, I. A comparison of methods to elicit causal structure. *Frontiers in Cognition*, 4:1544387, 2025.

Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436 – 463, 2013. doi: 10.1214/12-AOS1080. URL https://doi.org/10.1214/12-AOS1080.

Weller, M., Komusiewicz, C., Niedermeier, R., and Uhlmann, J. On making directed graphs transitive. *Journal of Computer and System Sciences*, 78(2):559–574, 2012. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.07.001. URL https://www.sciencedirect.com/science/article/pii/S002200001100078X. Games in Verification.

Wright, S. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2957502.

A. Appendix

A.1. Proofs

Lemma 2.4 (Existence of a unique compatible SEM)**.** *Let A be a unit lower-triangular matrix encoding a complete list of bivariate causal statements consistent with the ordering of $(X_1, \dots, X_n)$ and define $\Gamma = I - A^{-1}$. Then, the multivariate structural equation model*

$$\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N},$$

where the distribution of $\mathbf{N}$ is defined as the distribution of $(I - \Gamma)\mathbf{X}$, is the unique linear SEM whose pairwise marginal submodels are given by

$$X_j = \alpha_{ij} X_i + \tilde{N}_{ij}.$$

Proof. Consider an arbitrary SEM $\mathbf{X} = \Gamma \mathbf{X} + \mathbf{N}$. By Lemma 2.2, marginalizing to the pair (X_i, X_j) with $i < j$ results in the SEM

$$\begin{pmatrix} X_i \\ X_j \end{pmatrix} = (\Gamma_{YY} + \Gamma_{YZ}(I - \Gamma_{ZZ})^{-1}\Gamma_{ZY}) \begin{pmatrix} X_i \\ X_j \end{pmatrix} + \tilde{\mathbf{N}},$$

where $Y = i, j$ and $Z = \{1, \dots, n\} \setminus \{i, j\}$. Since Γ is strictly lower triangular, we have that

$$\Gamma_{YY} + \Gamma_{YZ}(I - \Gamma_{ZZ})^{-1}\Gamma_{ZY} = \begin{pmatrix} 0 & 0 \\ r & 0 \end{pmatrix},$$

for some entry $r \in \mathbb{R}$. To calculate what this entry r is, note that we have according to the block matrix inversion formula using the Schur complement that

$$(I - \Gamma)^{-1}_{YY} = (I - \Gamma_{YY} - \Gamma_{YZ}(I - \Gamma_{ZZ})^{-1}\Gamma_{ZY})^{-1} = \begin{pmatrix} 1 & 0 \\ -r & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix}.$$

This implies $r = (I - \Gamma)^{-1}_{ji}$, so the pairwise marginal submodels of an SEM with causal coefficients Γ are given by

$$X_j = (I - \Gamma)^{-1}_{ji} X_i + \tilde{N}_{ij}.$$

Requiring that all these submodels align with $X_j = \alpha_{ij} X_i + \tilde{N}_{ij}$ is therefore equivalent to $(I - \Gamma)^{-1} = A$ or $\Gamma = I - A^{-1}$, which completes the proof. □

Theorem 2.10. *For $n \geq 3$, let (Γ, Σ_N) specify a random n-dimensional linear Gaussian causal model drawn from a distribution that satisfies Assumption 2.9. Define $A = (I - \Gamma)^{-1}$ to be the matrix of total bivariate causal effects. Then,*

$$\mathbb{E}\left[\mathrm{comp}(A)\right] > 0.$$

Proof. Let (Γ, Σ_N) be an n-dimensional linear Gaussian causal model drawn from a distribution that satisfies Assumption 2.9. and define $A = (I - \Gamma)^{-1}$ to be the matrix of total bivariate causal effects. For $1 \leq i < j \leq n$, let

$$B_{ij} := \sum_{k<i} \mathrm{Var}(X_k) \cdot \sum_{\substack{P_1 : k \to i \\ P_2 : k \to j \\ P_1 \cap P_2 = \emptyset}} \Gamma^{P_1} \Gamma^{P_2}$$

to be the part of the covariance between X_i, X_j coming from back-door paths through observed variables and

$$\varepsilon_{ij} := \sum_{\ell \neq k \leq j} \mathrm{Cov}(N_\ell, N_k) \cdot \sum_{\substack{P_1 : \ell \to i \\ P_2 : k \to j \\ P_1 \cap P_2 = \emptyset}} \Gamma^{P_1} \Gamma^{P_2}$$

to be the part of the covariance between X_i, X_j coming from back-door paths that are not fully observed. First, we show that B_{ij} and ε_{ij} are uncorrelated. For$i = 1$, we have $B_{ij} = 0$, so this is trivially true. For $i > 1$, we get

$$\mathbb{E}[B_{ij}\varepsilon_{ij}] = \sum_{k<i} \sum_{\ell \neq m \leq j} \sum_{\substack{P_1: k \to i \\ P_2: k \to j \\ P_1 \cap P_2 = \emptyset}} \sum_{\substack{Q_1: \ell \to i \\ Q_2: m \to j \\ Q_1 \cap Q_2 = \emptyset}} \mathbb{E}\left[\mathrm{Var}(X_k) \cdot \mathrm{Cov}(N_l, N_m) \cdot \Gamma^{P_1}\Gamma^{P_2}\Gamma^{Q_1}\Gamma^{Q_2}\right] \tag{5}$$

Fix a summand in the expression above, determined by specific choices of $k, \ell, m, P_1, P_2, Q_1, Q_2$. Since Q_1, Q_2 are disjoint paths with all different endpoints, they cannot fully contain the paths P_1, P_2 as those intersect in the common endpoint k. Hence, there must exist an entry Γ_{rs} with $r > k$ that only appears once in the product $\Gamma^{P_1}\Gamma^{P_2}\Gamma^{Q_1}\Gamma^{Q_2}$. Note that $\mathrm{Var}(X_k)$ only depends on Σ_N and entries Γ_{pq} with $p \leq k$. By part (2) of Assumption 2.9, the entry Γ_{rs} is therefore independent from all other factors in the product $\mathrm{Var}(X_k) \cdot \mathrm{Cov}(N_l, N_m) \cdot \Gamma^{P_1}\Gamma^{P_2}\Gamma^{Q_1}\Gamma^{Q_2}$. This implies

$$\mathbb{E}\left[\mathrm{Var}(X_k) \cdot \mathrm{Cov}(N_l, N_m) \cdot \Gamma^{P_1}\Gamma^{P_2}\Gamma^{Q_1}\Gamma^{Q_2}\right] = \mathbb{E}\left[\Gamma_{rs}\right] \cdot \mathbb{E}\left[\mathrm{Var}(X_k) \cdot \mathrm{Cov}(N_l, N_m) \cdot \Gamma^{P_1}\Gamma^{P_2}\Gamma^{Q_1}\Gamma^{Q_2} \cdot \frac{1}{\Gamma_{rs}}\right] = 0,$$

by part (1) of Assumption 2.9. Since this argument holds for all summands of equation (5), we get

$$\mathbb{E}[B_{ij}\varepsilon_{ij}] = 0.$$

Now, recall equation (4) stating that

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Var}(X_i) \cdot A_{ji} + B_{ij} + \varepsilon_{ij}.$$

Therefore, we get for the bivariate confounding between X_i and X_j

$$\mathcal{C}(X_i, X_j, A_{ji}) = (B_{ij} + \varepsilon_{ij})^2$$

and the multivariate confounding is given by

$$\mathcal{C}(X_i, X_j, \Gamma) = \varepsilon_{ij}^2.$$

Hence, we can write

$$\mathbb{E}[\mathrm{comp}(A)] = \mathbb{E}\left[\sum_{i<j}(B_{ij} + \varepsilon_{ij})^2 - \varepsilon_{ij}^2\right] = \sum_{i<j}\left(\mathbb{E}\left[B_{ij}^2\right] + 2\,\mathbb{E}\left[B_{ij}\varepsilon_{ij}\right]\right) = \sum_{i<j}\mathbb{E}[B_{ij}^2] \geq 0.$$

For $i = 2, j = 3$, we get from part (2) of Assumption 2.9

$$\mathbb{E}[B_{23}^2] = \mathbb{E}[\mathrm{Var}(X_1)^2 \cdot \Gamma_{21}^2 \cdot \Gamma_{31}^2] = \mathbb{E}[\Sigma_{N,11}^2 \cdot \Gamma_{21}^2 \cdot \Gamma_{31}^2] = \mathbb{E}[\Sigma_{N,11}^2] \cdot \mathbb{E}[\Gamma_{21}^2] \cdot \mathbb{E}[\Gamma_{31}^2].$$

By part (3) of Assumption 2.9, we have that $\Sigma_{N,11}^2 > 0$ almost surely, and $\mathbb{E}[\Gamma_{21}^2] = \mathrm{Var}(\Gamma_{21}) > 0, \mathbb{E}[\Gamma_{31}^2] = \mathrm{Var}(\Gamma_{31}) > 0$. This implies $\mathbb{E}[B_{23}^2] > 0$ and therefore also

$$\mathbb{E}[\mathrm{comp}(A)] > 0.$$

□

B. Algorithms

C. Experiment Details

To obtain lists of bivariate causal statements from LLMs (see Figure 2), we used the following prompts:

System prompt:

Algorithm 3 Greedy Confounding Path Closure (GreedyCPC)

Input: Mixed graph G with bidirected edge set $B = B(G)$
Output: Bidirected edge sets $B_{\mathrm{del}} \subseteq B$ and B_{add} such that deleting B_{del} from G and adding B_{add} makes it equal to its confounding path closure
Initialize $B_{\mathrm{del}} \leftarrow \emptyset$
Initialize $H \leftarrow G$
repeat
  Initialize $bestGain \leftarrow 0$
  Initialize $e^{\star} \leftarrow \emptyset$
  for each edge $e \in B(H)$ **do**
    Compute $gain \leftarrow |cpc(H)| - |cpc(H \setminus e)| - 1$
    if $gain > bestGain$ **then**
      $bestGain \leftarrow gain$
      $e^{\star} \leftarrow e$
    end if
  end for
  if $bestGain > 0$ **then**
    $B_{\mathrm{del}} \leftarrow B_{\mathrm{del}} \cup \{e^{\star}\}$
    $H \leftarrow H \setminus e^{\star}$
  end if
until $bestGain = 0$
Define $G' \leftarrow G \setminus E_{\mathrm{del}}$
Set $B_{\mathrm{add}} \leftarrow B(cpc(G')) \setminus B(G')$
return $(B_{\mathrm{del}}, B_{\mathrm{add}})$

```
You are a causality expert, tasked to estimate standardized TOTAL causal effects between
country development indicators. Return your answer in HTML format:

<answer>CAUSAL_COEFFICIENT: <number></answer>

For example: <answer>CAUSAL_COEFFICIENT: 0.35</answer> or
<answer>CAUSAL_COEFFICIENT: -0.62</answer>

The causal coefficient quantifies the expected change in the effect variable in standard
deviations, given an intervention that changes the cause variable by 1 standard deviation.
It includes the effect of all direct causal pathways from the cause to the effect variable.
Do not assume away confounding; use realistic domain knowledge. No other text.
```

User prompt:

```
    I have observational data on 5 country-level development indicators measured 1970-2009.
    Correlation matrix:

                  gdp_per_capita  women_education  fertility  child_mortality  life_expectancy
gdp_per_capita          1.000000         0.548420  -0.513966        -0.517020         0.589402
women_education         0.548420         1.000000  -0.855142        -0.823446         0.787732
fertility              -0.513966        -0.855142   1.000000         0.823476        -0.802140
child_mortality        -0.517020        -0.823446   0.823476         1.000000        -0.896291
life_expectancy         0.589402         0.787732  -0.802140        -0.896291         1.000000

Variable descriptions:
'gdp_per_capita': 'GDP per capita (economic output per person in USD)',
'women_education': 'Mean years of schooling for women of reproductive age (15-44)',
```

```
'fertility': 'Total fertility rate (children per woman)',
'child_mortality': 'Child mortality rate (deaths per 1000 live births, ages 0-5)',
'life_expectancy': 'Life expectancy at birth (years)'

YOUR TASK:
Estimate the total linear causal coefficient for the causal effect of
{cause_var} on {effect_var}.
```

For the experiment, we send the user prompt above separately for each ordered pair of variables, according to the assumed causal ordering GDP -¿ women education -¿ fertility -¿ child mortality -¿ life expectancy. We send each request 5 times and take the average of the coefficients from each valid answer. In case that the output of the LLM does not contain a recognizable coefficient, we send the following message as a follow-up (keeping the chat history from the first prompt):

```
Please provide your final answer in HTML format like this:
<answer>CAUSAL_COEFFICIENT: <number></answer>
```

After sending the follow-up message, almost all LLM answers contained a valid coefficient with the notable exception of the gpt-oss-20b model, where often only 1 or 2 out of 5 prompts received a valid answer. For all prompts, we used the following parameters:

- Maximum number of output tokens: 2048
- temperature: 0.6
- top P: 0.7.