An efficient search-and-score algorithm for ancestral graphs using multivariate information scores

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We propose a greedy search-and-score algorithm for ancestral graphs, which include directed as well as bidirected edges, originating from unobserved latent variables. The normalized likelihood score of ancestral graphs is estimated in terms of multivariate information over relevant subsets of vertices, $\boldsymbol{C}$, that are connected through collider paths confined to the ancestor set of $\boldsymbol{C}$. For computational efficiency, the proposed two-step algorithm relies on local information scores limited to the close surrounding vertices of each node (step 1) and edge (step 2). This computational strategy is shown to outperform state-of-the-art causal discovery methods on challenging benchmark datasets.

1 Introduction

The likelihood function plays a central role in the selection of a graphical model $\mathcal{G}$ based on observational data $\mathcal{D}$. Given N independent samples from $\mathcal{D}$, the likelihood $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ that they might have been generated by the graphical model $\mathcal{G}$ is given by [1],

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{1}{Z_{\mathcal{D},\mathcal{G}}} \exp\Big(-NH(p,q)\Big) \tag{1}$$

where $H(p,q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$ is the cross-entropy between the empirical probability distribution $p(\boldsymbol{x})$ of the observed data $\mathcal{D}$ and the theoretical probability distribution $q(\boldsymbol{x})$ of the model $\mathcal{G}$ and $Z_{\mathcal{D},\mathcal{G}}$ a data- and model-dependent factor ensuring proper normalization condition for finite dataset. In short, Eq.1 results from the asymptotic probability that the N independent samples, $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}$, are drawn from the model distribution, $q(\boldsymbol{x})$, *i.e.* $\mathcal{L}_{\mathcal{D}|\mathcal{G}} \equiv q(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}) = \prod_i q(\boldsymbol{x}^{(i)})$, rather than the empirical distribution, $p(\boldsymbol{x})$. This leads to, $\log \mathcal{L}_{\mathcal{D}|\mathcal{G}} = \sum_i \log q(\boldsymbol{x}^{(i)})$, which converges towards $N \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x}) = -N\, H(p,q)$ in the large sample size limit, $N \to \infty$, with $\log Z_{\mathcal{D},\mathcal{G}} = \mathcal{O}(\log N)$.

The structural constraints of the model $\mathcal{G}$ translate into the factorization form of the theoretical probability distribution, $q(\boldsymbol{x})$ [2–6]. In particular, the probability distribution of Bayesian networks (BN) factorizes in terms of conditional probabilities of each variable given its parents, as $q_{\text{BN}}(\boldsymbol{x}) = \prod_i q(x_i|\mathbf{pa}_{X_i})$, where $\mathbf{pa}_{X_i}$ denote the values of the parents of node X_i in $\mathcal{G}$, $\mathbf{Pa}_{X_i}$. For Bayesian networks, the factors of the model distribution, $q(x_i|\mathbf{pa}_{X_i})$, can be directly estimated with the empirical conditional probabilities of each node given its parents as, $q(x_i|\mathbf{pa}_{X_i}) \equiv p(x_i|\mathbf{pa}_{X_i})$, leading to the well known estimation of the likelihood function in terms of conditional entropies $H(X_i|\mathbf{Pa}_{X_i}) = -\sum_{\boldsymbol{x}} p(x_i, \mathbf{pa}_{X_i}) \log p(x_i|\mathbf{pa}_{X_i})$,

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\text{BN}}} = \frac{1}{Z_{\mathcal{D},\mathcal{G}_{\text{BN}}}} \exp\Big(-N \sum_{X_i \in \boldsymbol{V}}^{\text{vertices}} H(X_i|\mathbf{Pa}_{X_i})\Big) \tag{2}$$

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

This paper concerns the experimental setting for which some variables of the underlying Bayesian model are not observed. This frequently occurs in practice for many applications. We derive an explicit likelihood function for the class of ancestral graphs, which include directed as well as bidirected edges, arising from the presence of unobserved latent variables. Tian and Pearl 2002 [7] showed that the probability distribution of such graphs factorizes into c-components including subsets of variables connected through bidirected paths (*i.e.* containing only bidirected edges). Richardson 2009 [6] later proposed a refined factorization of the model distribution of the broader class of acyclic directed mixed graphs in terms of conditional probabilities over "head" and "tail" subsets of variables within each ancestrally closed subsets of vertices. However, unlike with Bayesian networks, the contributions of c-components or head-and-tail factors to the likelihood function cannot simply be estimated in terms of empirical distribution $p(\boldsymbol{x})$, as shown below. This leaves the likelihood function of ancestral graphs difficult to estimate from empirical data, in general, although iterative methods have been developped when the data is normally distributed [8–13].

The present paper provides an explicit decomposition of the likelihood function of ancestral graphs in terms of multivariate cross-information over relevant 'ac-connected' subsets of variables, Figs. 1., which do not rely on the head-and-tail factorization but coincide with the parametrizing sets [14] derived from the head-and-tail factorization. It suggests a natural estimation of these revelant contributions to the likelihood function in terms of empirical distribution $p(\boldsymbol{x})$. This result extends the likelihood expression of Bayesian Networks (Eq. 2) to include the effect of unobserved latent variables and enables the implementation of a greedy search-and-score algorithm for ancestral graphs. For computational efficiency, the proposed two-step algorithm relies on local information scores limited to the close surrounding vertices of each node (step 1) and edge (step 2). This computational strategy is shown to outperform state-of-the-art causal discovery methods on challenging benchmark datasets.

2 Theoretical results

2.1 Multivariate cross-entropy and cross-information

The theoretical result of the paper (Theorem 1) is expressed in terms of multivariate cross-information derived from multivariate cross-entropies through the Inclusion-Exclusion Principle. The same expressions can be written between multivariate information and multivariate entropies by simply substituting $q(\{x_i\})$ with $p(\{x_i\})$ in the equations below and will be used to estimate the likelihood function of ancestral graphs (Proposition 3).

As recalled above, the cross-entropy between m variables, $\boldsymbol{V} = \{X_1, \cdots, X_m\}$, is defined as,

$$H(\boldsymbol{V}) \;=\; -\sum_{\{x_i\}} p(x_1, \cdots, x_m) \log q(x_1, \cdots, x_m) \tag{3}$$

where $p(\{x_i\})$ is the empirical joint probability distribution of the variables $\{X_i\}$ and $q(\{x_i\})$ the joint probability distribution of the model. Bayes formula, $q(\{x_i\}, \{y_j\}) = q(\{x_i\}|\{y_j\})\, q(\{y_j\})$, directly translates into the definition of conditional cross-entropy through the decomposition,

$$H(\{X_i\}, \{Y_j\}) = H(\{X_i\}|\{Y_j\}) + H(\{Y_j\}) \tag{4}$$

Multivariate (cross) information, $I(\boldsymbol{V}) \equiv I(X_1; \cdots; X_m)$, are defined from multivariate (cross) entropies through Inclusion-Exclusion formulas over all subsets of variables [15–18] as,

$$\begin{aligned}
I(X) &= H(X)\\
I(X;Y) &= H(X) + H(Y) - H(X,Y)\\
I(X;Y;Z) &= H(X) + H(Y) + H(Z) - H(X,Y) - H(X,Z) - H(Y,Z) + H(X,Y,Z)\\
I(\boldsymbol{V}) &= -\sum_{\boldsymbol{S} \subseteq \boldsymbol{V}} (-1)^{|\boldsymbol{S}|} H(\boldsymbol{S})
\end{aligned} \tag{5}$$

where the semicolon separators are needed to distinguish multipoint (cross) information from joint variables as $\{X, Z\}$ in $I(\{X,Z\};Y) = I(X;Y) + I(Z;Y) - I(X;Y;Z)$. Below, implicit separators between non-conditioning variables in multivariate (cross) information will always correspond to semicolons, *e.g.* as in $I(\boldsymbol{V})$ in Eq. 5. Unlike multivariate (cross) entropies, which are always positive,

$H(X_1, \cdots, X_k) \geqslant 0$, multivariate (cross) information, $I(X_1; \cdots; X_k)$, can be positive or negative for $k \geqslant 3$, while they remain always positive for $k < 3$, *i.e.* $I(X;Y) \geqslant 0$ and $I(X) \geqslant 0$.

In turn, multivariate (cross) entropies can be expressed through the Principle of Inclusion-Exclusion into the same expression form but in terms of multivariate (cross) information,

$$H(\boldsymbol{V}) \;=\; -\sum_{\boldsymbol{S} \subseteq \boldsymbol{V}} (-1)^{|\boldsymbol{S}|} I(\boldsymbol{S}), \tag{6}$$

Conditional multivariate (cross) information $I(\boldsymbol{V}|Z)$ are defined similarly as multivariate (cross) information $I(\boldsymbol{V})$ but in terms of conditional (cross) entropies as,

$$I(\boldsymbol{V}|Z) \;=\; -\sum_{\boldsymbol{S} \subseteq \boldsymbol{V}} (-1)^{|\boldsymbol{S}|} H(\boldsymbol{S}|Z) \tag{7}$$

Eqs. 5 & 7 lead to a decomposition rule relative to a variable Z, Eq. 8, which can be conditioned on a set of joint variables, $\boldsymbol{A} = \{A_1, \cdots, A_m\}$, with implicit comma separators for conditioning variables in Eq. 9,

$$I(\boldsymbol{V}) \;=\; I(\boldsymbol{V}|Z) + I(\boldsymbol{V};Z) \tag{8}$$

$$I(\boldsymbol{V}|\boldsymbol{A}) \;=\; I(\boldsymbol{V}|Z,\boldsymbol{A}) + I(\boldsymbol{V};Z|\boldsymbol{A}) \tag{9}$$

Alternatively, conditional (cross) information, such as $I(X;Y|\boldsymbol{A})$, can be expressed in terms of non-conditional (cross) entropies using Eq. 4,

$$\begin{aligned} I(X;Y|\boldsymbol{A}) \;&=\; H(X|\boldsymbol{A}) + H(Y|\boldsymbol{A}) - H(X,Y|\boldsymbol{A}) \\ &=\; H(X,\boldsymbol{A}) + H(Y,\boldsymbol{A}) - H(X,Y,\boldsymbol{A}) - H(\boldsymbol{A}) \end{aligned} \tag{10}$$

which can in turn be expressed in terms of non-conditional (cross) information as,

$$\begin{aligned} I(X;Y|\boldsymbol{A}) &= I(X;Y) - \cdots (-1)^k \sum_{i_1 < \cdots < i_k} I(X;Y;A_{i_1};\cdots;A_{i_k}) + \cdots (-1)^m I(X;Y;A_1;\cdots;A_m) \\ &= \sum_{\boldsymbol{S'} \subseteq \boldsymbol{S}}^{X,Y \in \boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}), \end{aligned} \tag{11}$$

where $\boldsymbol{S} = \{X,Y\} \cup \boldsymbol{A}$. This corresponds, up to an opposite sign, to *all (cross) information terms including both X and Y* in the expression of the multivariate (cross) entropy, $H(X,Y,\boldsymbol{A})$, Eq. 6.

2.2 Graphs and connection criteria

2.2.1 Directed mixed graphs and ancestral graphs

Two vertices are said to be **adjacent** if there is an edge (of any type) between them, $X *\!\!-\!\!* Y$, where $*$ stands for any (head or tail) end mark. X and Y are said to be **neighbors** if $X - Y$, **parent** and **child** if $X \rightarrow Y$ and **spouses** if $X \longleftrightarrow Y$ in $\mathcal{G}$.

A **path** in $\mathcal{G}$ is a sequence of distinct vertices $V_1, \ldots, V_n$ consecutively adjacent in $\mathcal{G}$, as, $V_1 *\!\!-\!\!* V_2 *\!\!-\!\!* \cdots *\!\!-\!\!* V_{n-1} *\!\!-\!\!* V_n$. In particular, a **collider path** between V_1 and V_n has the form $V_1 *\!\!\rightarrow V_2 \longleftrightarrow \cdots \longleftrightarrow V_{n-1} \leftarrow\!\!* V_n$ and a **directed path** corresponds to $V_1 \rightarrow V_2 \rightarrow \cdots \rightarrow V_n$.

X is called an **ancestor** of Y and Y a **descendant** of X if $X = Y$ or there is a **directed path** from X to Y, $X \rightarrow \cdots \rightarrow Y$. $\mathbf{An}_{\mathcal{G}}(Y)$ denotes the **set of ancestors** of Y in $\mathcal{G}$. By extension, for any subset of vertices, $\boldsymbol{C} \subseteq \boldsymbol{V}$, $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C})$ denotes the set of ancestors for all $Y \in \boldsymbol{C}$ in $\mathcal{G}$.

A **directed mixed graph** is a vertex-edge graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ that can contain two types of edges: directed ($\rightarrow$) and bidirected ($\longleftrightarrow$) edges.

A **directed cycle** occurs in $\mathcal{G}$ when $X \in \mathbf{An}_{\mathcal{G}}(Y)$ and $X \leftarrow Y$. An **almost directed cycle** occurs when $X \in \mathbf{An}_{\mathcal{G}}(Y)$ and $X \longleftrightarrow Y$.

Definition 1. An **ancestral graph** is a directed mixed graph:
- *i)* without directed cycles;
- *ii)* without almost directed cycles.

An **ancestral graph** is said to be **maximal** if every missing edge corresponds to a structural independence. If an ancestral graph $\mathcal{G}$ is not maximal, there exists a unique maximal ancestral graph $\bar{\mathcal{G}}$ by adding bidirected edges to $\mathcal{G}$ [8].

2.2.2 *ac*-connecting paths and *ac*-connected subsets

Let us now define **ancestor collider connecting paths** or ***ac*-connecting paths**, which entail simpler path connecting criterion than the traditional **m-connecting criterion**, discussed in the Appendix A. Yet, ***ac*-connecting paths** and ***ac*-connected subsets** will turn out to be directly relevant to characterize the likelihood decomposition and Markov equivalent classes of ancestral graphs.

Definition 2. [ac-connecting path] An ac-connecting path between X and Y given a subset of variables $\boldsymbol{C}$ (possibly including X and Y) is a collider path, $X \ast\!\!\to Z_1 \longleftrightarrow \cdots \longleftrightarrow Z_K \leftarrow\!\!\ast\, Y$, with all $Z_i \in \mathbf{An}_{\mathcal{G}}(\{X,Y\} \cup \boldsymbol{C})$, that is, with Z_i in $\boldsymbol{C}$ or connected to $\{X,Y\} \cup \boldsymbol{C}$ by an ancestor path, *i.e.* $Z_i \to \cdots \to T$ with $T \in \{X,Y\} \cup \boldsymbol{C}$.

Definition 3. [ac-connected subset] A subset $\boldsymbol{C}$ is said to be ac-connected if $\forall X, Y \in \boldsymbol{C}$, X and Y are connected (through any type of edge) or there is an ac-connecting path between X and Y given $\boldsymbol{C}$.

2.3 Likelihood decomposition of ancestral graphs

Theorem 1. [**likelihood of ancestral graphs**] *The cross-entropy $H(p,q)$ and likelihood $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$ of an ancestral graph $\mathcal{G}$ is decomposable in terms of multivariate cross-information, $I(\boldsymbol{C})$, summed over all ac-connected subsets of variables, $\boldsymbol{C}$ (Definition 3),*

$$\begin{aligned} H(p,q) &= -\sum_{\boldsymbol{C}\subseteq \boldsymbol{V}}^{\mathrm{ac-connected}} (-1)^{|\boldsymbol{C}|} I(\boldsymbol{C}) \\ \mathcal{L}_{\mathcal{D}|\mathcal{G}} &= \frac{1}{Z_{\mathcal{D},\mathcal{G}}} \exp\Big(N \sum_{\boldsymbol{C}\subseteq \boldsymbol{V}}^{\mathrm{ac-connected}} (-1)^{|\boldsymbol{C}|} I(\boldsymbol{C}) \Big) \end{aligned} \tag{12}$$

where N is the number of iid samples in the dataset $\mathcal{D}$ and $Z_{\mathcal{D},\mathcal{G}}$ a data- and model-dependent normalization constant.

The proof of Theorem 1 is left to the Appendix B. It is based on a partition of the cross-entropy (Eq. 6) into cross-information contributions from ac-connected and non-ac-connected subsets of variables, which do not rely on head-and-tail factorizations. Hu and Evans [14] proposed an equivalent result (Proposition 3.3 in [14]) with a proof using head-and-tail decomposition to define parametrizing sets, which happen to coincide with the ac-connected sets defined here (Definition 3). Theorem 1 characterizes in particular the Markov equivalence class of ancestral graphs [8, 19–24] as,

Corollary 2. *Two ancestral graphs are Markov equivalent if and only if they have the same ac-connected subsets of vertices.*

Note, in particular, that Eq. 12 holds for *maximal ancestral graphs* (MAG), for which all pairs of ac-connected variables are connected by an edge, and their Markov equivalent representatives, the *partial ancestral graphs* (PAG) [8, 25–27].

Proposition 3. The likelihood decomposition of ancestral graphs (Eq. 12, Theorem 1) can be estimated by replacing the model distribution q by the empirical distribution p in the retained multivariate cross-information terms $I(\boldsymbol{C})$ corresponding to all *ac*-connected subsets of variables, $\boldsymbol{C}$.

Hence, Proposition 3 amounts to estimating all relevant cross-information terms in the likelihood function with the corresponding multivariate information terms computed from the available data, while assuming by construction that the model distribution obeys all local and global conditional independences entailed by the ancestral graph. The corresponding factorization of the model distribution can be expressed in terms of empirical distribution, assuming positive distributions, see Appendix C.

Fig. 1 illustrates the cross-entropy decomposition for a few graphical models in terms of cross-information contributions from their ac-connected subsets of vertices. In particular, an unshielded non-collider (*e.g.* $X \to Z \to W$, Fig. 1A), is less likely (*i.e.* higher cross-entropy) than an unshielded collider or 'v-structure' (*e.g.* $X \to Z \leftarrow W$, Fig. 1B), if the corresponding three-point information term is negative, $I(X;Z;W) < 0$, in agreement with earlier results [28, 29]. However, this early approach, exploiting the sign and magnitude of three-point information to orient v-structures, does not include higher order terms involving multiple v-structures, which can lead to orientation conflicts between unshielded triples, in practice. Resolving such orientation conflicts requires to include

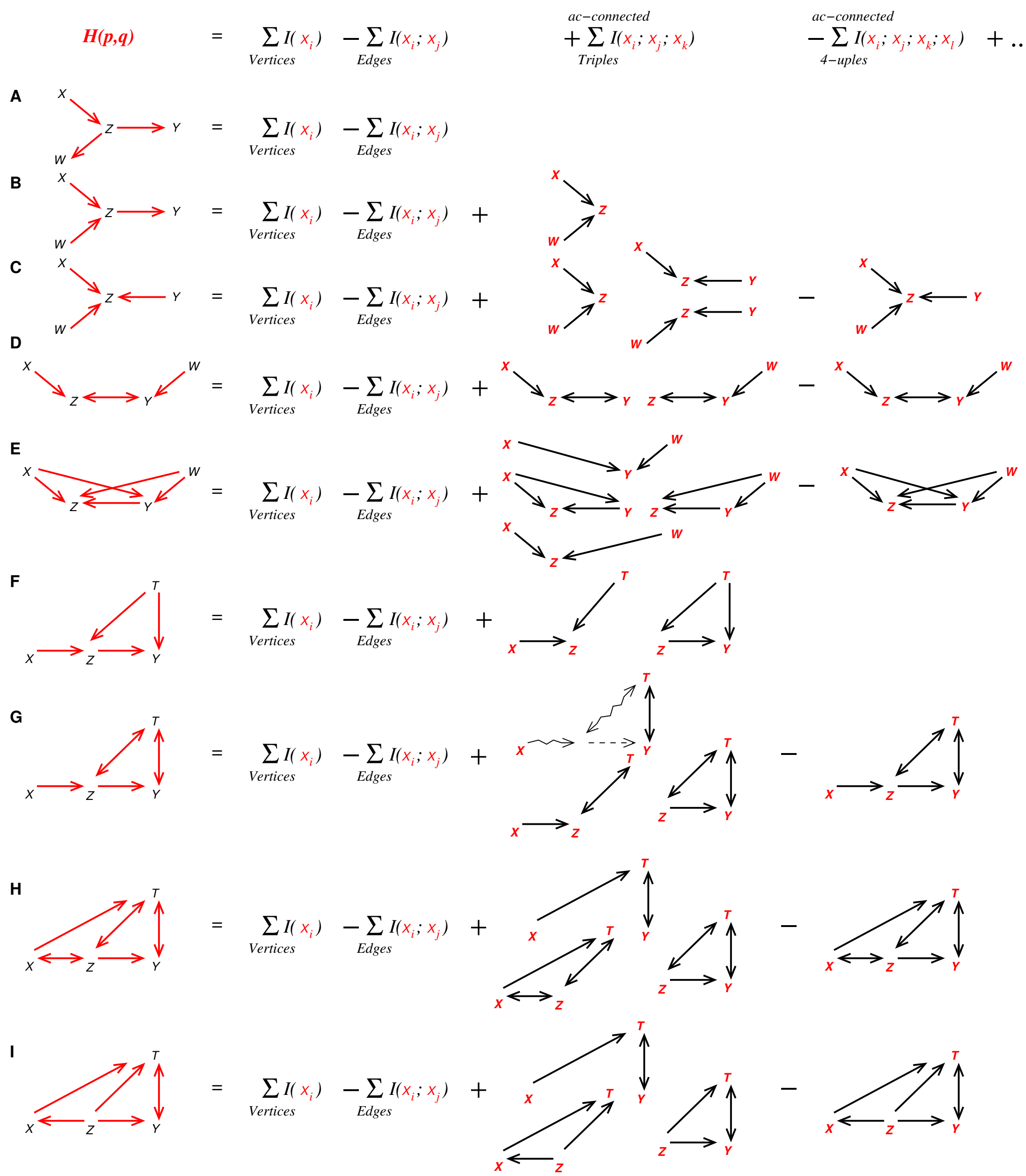

Figure 1: **Cross-entropy decomposition of ancestral graphs.** Examples of cross-entropy decomposition of ancestral graphs (red edges, lhs) in terms of relevant multivariate cross-information contributions $I(\boldsymbol{C})$ with $\boldsymbol{C} \subseteq \boldsymbol{V}$ (red nodes, rhs). Simple graphs: (**A**) without unshielded colliders, (**B**) with a single or non-overlapping unshielded colliders, (**C**) with overlapping unshielded colliders through three or more (conditionally) independent parents or (**D**) through a two-(or more)-collider path. (**E**) Bayesian graph corresponding to the head-and-tail factorization of the two-collider path in (D) estimated using the empirical distribution $p(.)$, see Appendix C. (**F**) Simple Bayesian graph not Markov equivalent to an ancestral graph (**G**) sharing the same edges and unshielded collider [24]. Solid black edges correspond to direct connections or collider paths confined to the corresponding ac-connected subset $\boldsymbol{C}$, while wiggly edges indicate collider paths extending beyond $\boldsymbol{C}$ yet indirectly connected to $\boldsymbol{C}$ by an ancestor path, marked with dashed edges, see Definition 2. By contrast, graphs **H** and **I** illustrate the fact that collider paths may not be unique nor conserved between two Markov equivalent graphs (*i.e.* sharing the same cross-information terms) [24].

information contributions from higher-order ac-connected subgraphs, such as star-like ac-connected subsets including three or more parents, Fig. 1C. Similarly, the cross-entropies of collider paths involving several colliders also include higher-order terms, as with the simple example of a two-collider path, Fig. 1D. By contrast, the cross-entropy based on the head-and-tail factorization of the same two-collider path, *i.e.* $q(x,z,y,w) = q(z,y|x,w)q(x)q(w)$ [6], is found to be equivalent to the cross-entropy of a Bayesian graph without bidirected edge, Fig. 1E, when estimated with the empirical distribution $p(.)$, see Appendix C. This observation illustrates the difficulty to estimate the likelihood functions of ancestral graphs using head-and-tail factorization.

Further examples of graphical models, Figs. 1F-I, show the relative simplicity of the decomposition with only few (non-trivial) ac-connected contributing subsets $\boldsymbol{C}$ with $|\boldsymbol{C}| \geqslant 3$, as compared to the much larger number of non-ac-connected non-contributing subsets, that cancel each other by construction due to conditional independence constraints of the underlying model. Note, in particular, that most contributing multivariate information $I(\boldsymbol{C})$ only concern direct connections or collider paths within a single component subgraph induced by $\boldsymbol{C}$ (solid line edges in Fig. 1). However, occasionally, collider paths extending beyond $\boldsymbol{C}$ into $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C}) \setminus \boldsymbol{C}$ (marked with wiggly edges) with corresponding ancestor path(s) (marked with dashed edges) do occur, as shown in Fig. 1G.

In addition, the present information-theoretic decomposition of the likelihood of ancestral graphs can readily distinguish their Markov equivalence classes according to Corollary 2. For instance, the ancestral graphs of Fig. 1F and Fig. 1G, despite sharing the same edges and the same unshielded collider ($X \rightarrow Z \leftarrow T$), turn out not to be Markov equivalent, as discussed in [24]. Indeed, their cross-entropy decompositions differ by two ac-connected contributing terms: a three-point cross information $I(X;Y;T)$ with a collider path not confined in $\boldsymbol{C}$ (*i.e.* $X \rightsquigarrow Z \leftrightsquigarrow T \longleftrightarrow Y$ and corresponding ancestor path $Z \dashrightarrow Y$) and a four-point information term $I(X;Y;Z;T)$ due to the two-collider path ($X \rightarrow Z \longleftrightarrow T \longleftrightarrow Y$). More quantitatively, it shows that the graph of Fig. 1G with a two-collider path is more likely than the graph of Fig. 1F whenever $I(X;Y;T) - I(X;Y;Z;T) = I(X;Y;T|Z) = I(X;Y|Z) - I(X;Y|Z,T) < 0$. Finally, the Markov equivalent graphs of Fig. 1H and Fig. 1I, also due to [24], illustrate the fact that the actual ancestor collider path between unconnected pairs does not need to be unique nor conserved between Markov equivalent graphs (as long as their cross-entropies share the same multivariate cross-information decomposition).

3 Efficient search-and-score causal discovery using local information scores

The likelihood estimation of ancestral graphs (Theorem 1 and Proposition 3) enables the implementation of a search-and-score algorithm for this broad class of graphs, which has attracted a number of contributions recently [11–13, 30–32]. Our specific objective is not to develop an exact method limited to simple graphical models with a few nodes and small datasets but to implement an efficient and reliable heuristic method applicable to more challenging graphical models and large datasets.

Indeed, search-and-score structure learning methods need to rely on heuristic rather than exhaustive search, in general, given that the number of ancestral graphs grows super-exponentially as the number of vertices increases. This can be implemented for instance with a Monte Carlo algorithmic scheme with random restarts, which efficiently probes relevant graphical models. Here, we opt, instead, to use the prediction of an efficient hybrid causal discovery method, MIIC [29, 33, 34], as starting point for a subsequent search-and-score approach based on the proposed likelihood estimation of ancestral graphs (Eq. 12 and Proposition 3).

Moreover, while the likelihood decomposition of ancestral graphs may involve extended ac-connected subsets of variables, as illustrated in Fig. 1, we aim to implement a computationally efficient search-and-score causal discovery method based on approximate local scores limited to the close surrounding vertices of each node and edge. Yet, while MIIC only relies on unshielded triple scores, the novel search-and-score extension, MIIC_search&score, uses also higher-order local information scores to compare alternative subgraphs, as detailed below.

The proposed method is shown to outperform MIIC and other state-of-the-art causal discovery methods on challenging datasets including latent variables.

3.1 MIIC, an hybrid causal discovery method based on unshielded triple scores

MIIC is an hybrid causal discovery method combining constraint-based and information-theoretic frameworks [29, 35]. Unlike traditional constraint-based methods [4, 5], MIIC does not directly attempt to uncover conditional independences but, instead, iteratively substracts the most significant three-point (conditional) information contributions of successive contributors, A_1, A_2, ..., A_n, from the mutual information between each pair of variables, $I(X;Y)$, as,

$$I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \cdots - I(X;Y;A_n|\{A_i\}_{n-1}) = I(X;Y|\{A_i\}_n) \quad (13)$$

where $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$ is the *positive* information contribution from A_k to $I(X;Y)$ [28, 36]. Conditional independence is eventually established when the residual conditional mutual information on the right hand side of Eq. 13, $I(X;Y|\{A_i\}_n)$, becomes smaller than a complexity term, *i.e.* $k_{X;Y|\{A_i\}}(N) \geqslant I(X;Y|\{A_i\}_n) \geqslant 0$, which dependents on the considered variables and sample size N.

This leads to an undirected skeleton, which MIIC then (partially) orients based on the sign and amplitude of the regularized conditional 3-point information terms [28, 29]. In particular, negative conditional 3-point information terms, $I(X;Y;Z|\{A_i\}) < 0$, correspond to the signature of causality in observational data [28] and lead to the prediction of a v-structure, $X \rightarrow Z \leftarrow Y$, if X and Y are not connected in the skeleton. By contrast, a positive conditional 3-point information term, $I(X;Y;Z|\{A_i\}) > 0$, implies the absence of a v-structure and suggests to propagate the orientation of a previously directed edge $X \rightarrow Z - Y$ as $X \rightarrow Z \rightarrow Y$.

In practice, MIIC's strategy to circumvent spurious conditional independences significantly improves recall, that is, the fraction of correctly recovered edges, compared to traditional constraint-based methods [28, 29]. Yet, MIIC only relies on unshielded triple scores to reliably uncover significant contributors and orient v-structures, as outlined above. MIIC has been recently improved to ensure the consistency of the separating set in terms of indirect paths in the final skeleton or (partially) oriented graphs [37, 34] and to improve the reliably of predicted orientations [33, 34].

The predictions of this recent version of MIIC, which include three type of edges (directed, bidirected and undirected), have been used as starting point for the subsequent local search-and-score method implemented in the present paper.

3.2 New search-and-score method based on higher-order local information scores

Starting from the structure predicted by MIIC, as detailed above, MIIC_search&score method proceeds in two steps.

3.2.1 Step 1: Node scores for edge orientation priming and edge removal

The first step consists in minimizing a node score corresponding to the local normalized log likelihood of each node w.r.t. its possible parents or spouses amongst the connected nodes predicted by MIIC. To this end, the node score assesses the conditional entropy of each node w.r.t. a selection of parents, spouses or neighbors, $\mathbf{Pa'}_{X_i} \subseteq \mathbf{Pa}_{X_i} \cup \mathbf{Sp}_{X_i} \cup \mathbf{Ne}_{X_i}$, and a factorized Normalized Maximum Likelihood (fNML) regularization [28], see Appendix D for details,

$$\mathrm{Score_n}(X_i) = H(X_i|\mathbf{Pa'}_{X_i}) + \frac{1}{N}\sum_{j}^{q_{x_i}} \log \mathcal{C}_{n_j}^{r_{x_i}} \quad (14)$$

where q_{x_i} corresponds to the combination of levels of $\mathbf{Pa'}_{X_i}$, while r_{x_i} is the number of levels of X_i, and n_j the number of samples corresponding to a particular combination of levels j in each summand, with $\sum_j n_j = N$, the total number of samples. $\log \mathcal{C}_{n_j}^{r_{x_i}}$ is the fNML regulatization cost summed over all combinations of levels, q_{x_i}, [38, 39], see Appendix D.

This first algorithm is looped over each node, priming the orientations of their surrounding edges (as directed, bidirected or undirected), until convergence. Edges without orientation priming at either extremity are removed at the end of Step 1.

3.2.2 Step 2: Edge orientation scores

The second step consists in minimizing an edge orientation score corresponding to the local normalized log likelihood of each edge w.r.t. its nodes' parents and spouses inferred in Step 1. To this end, the edge score assesses the conditional information and a fNML complexity cost with respect to the type of orientation, given three sets of parents and spouses of X and Y, *i.e.* $\mathbf{Pa'}_{X\setminus Y} = \mathbf{Pa}_X \cup \mathbf{Sp}_X \setminus Y$, $\mathbf{Pa'}_{Y\setminus X} = \mathbf{Pa}_Y \cup \mathbf{Sp}_Y \setminus X$ and $\mathbf{Pa'}_{XY} = \mathbf{Pa'}_{X\setminus Y} \cup \mathbf{Pa'}_{Y\setminus X}$ with their corresponding combinations of levels, $q_{y\backslash x}$, $q_{x\backslash y}$ and q_{xy}. These orientation scores, listed in Table 1, include symmetrized fNML complexity terms to enforce Markov equivalence, if X and Y share the same parents or spouses (excluding X and Y), see Appendix D. Indeed, all three scores become equals if $\mathbf{Pa}'_{Y\backslash X} = \mathbf{Pa}'_{X\backslash Y} = \mathbf{Pa}'_{XY}$ implying also the same combinations of parent and spouse levels, $q_{y\backslash x} = q_{x\backslash y} = q_{xy}$.

This second algorithm is looped over each edge to compute an orientation score decrement, given the orientations of its surrounding edges. The orientation change corresponding to the largest orientation score decrement is then chosen at each iteration until convergence or until a limit cycle is reached and stopped at the lowest sum of local orientation scores.

Table 1: Local scores for the orientation of a single directed or bidirected edge.

Edge	Information	Symmetrized fNML complexity (Markov equivalent)
$X \rightarrow Y$	$-I(X;Y\|\mathbf{Pa}'_{Y\backslash X})$	$\frac{1}{2N}\Big(\sum_j^{q_{x\backslash y}r_y} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{x\backslash y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\backslash x}r_x} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{y\backslash x}} \log \mathcal{C}_{n_j}^{r_y}\Big)$
$X \leftarrow Y$	$-I(X;Y\|\mathbf{Pa}'_{X\backslash Y})$	$\frac{1}{2N}\Big(\sum_j^{q_{x\backslash y}r_y} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{x\backslash y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\backslash x}r_x} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{y\backslash x}} \log \mathcal{C}_{n_j}^{r_y}\Big)$
$X \leftrightarrow Y$	$-I(X;Y\|\mathbf{Pa}'_{XY})$	$\frac{1}{2N}\Big(\sum_j^{q_{xy}r_y} \log \mathcal{C}_{n_j}^{r_x} - \sum_j^{q_{xy}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{yx}r_x} \log \mathcal{C}_{n_j}^{r_y} - \sum_j^{q_{yx}} \log \mathcal{C}_{n_j}^{r_y}\Big)$

4 Experimental results

We first tested whether MIIC_search&score orientation scores (Table 1) effectively predicts bidirected orientations on three simple ancestral models, Fig. 3, when the end nodes do not share the same parents (Fig. 3, Model 1), share some parents (Fig. 3, Model 2) or when the bidirected edge is part of a longer than two-collider paths (Fig. 3, Model 3). The prediction of the edge orientation scores are summarized in Table 3, Appendix E, and show good predictions for large enough datasets.

Beyond these simple examples, focussing on the discovery of bidirected edges in small toy models of ancestral graphs, we also analyzed more challenging benchmarks from the bnlearn repository [40], Fig. 2. They concern ancestral graphs obtained by hiding up to 20% of variables in Bayesian Networks of increasing complexity (number of nodes and parameters), such as Alarm (37 nodes, 46 links, 509 parameters), Insurance (27 nodes, 52 links, 984 parameters), and Barley (48 nodes, 84 links, 114,005 parameters). We then assessed causal discovery performance in terms of *Precision*, $TP/(TP + FP)$, and *Recall*, $TP/(TP + FN)$, relative to the theoretical PAGs, while counting as false positive (FP), all correctly predicted edges but without or with a different orientation as the directed or bidirected edges of the PAG.

Fig. 2 compares MIIC_search&score performance to MIIC results used as starting point for MIIC_search&score and to FCI [41]. MIIC and MIIC_search&score settings were set as described in section 3 above. The open-source MIIC R package (v1.5.2, GPL-3.0 license) was obtained at `https://github.com/miicTeam/miic_R_package`. FCI from the python causal-learn package (v0.1.3.8, MIT license) [41] was obtained at `https://github.com/py-why/causal-learn` and run with G^2-conditional independence test and default parameter $\alpha = 0.05$.

Overall, MIIC_search&score is found to outperform MIIC in terms of edge precision with little to no decrease in edge recall, Fig. 2, demonstrating the benefit of MIIC_search&score's rationale to improve MIIC predictions by extending MIIC information scores from unshielded triples to higher-order information contributions. These originate from ac-connected subsets including nodes with more than two parents or spouses, or ac-connected subsets including two-collider paths. MIIC_search&score is also found to outperform FCI on complex ancestral benchmark networks with many parameters, such as Barley (114,005 parameters), Fig. 2. However, FCI is found to reach similar or better precision scores on easier benchmarks with fewer parameters (*i.e.* Alarm and Insurance), although its recall remains usually lower than MIIC_search&score, especially at small sample size, as expected for a purely constraint-based causal discovery approach.

Importantly, the benchmark PAGs used to score the causal discovery results with increasing proportions of latent variables, Fig. 2, include not only bidirected edges originating from hidden common causes but also additional directed or undirected edges arising, in particular, from indirect effects of

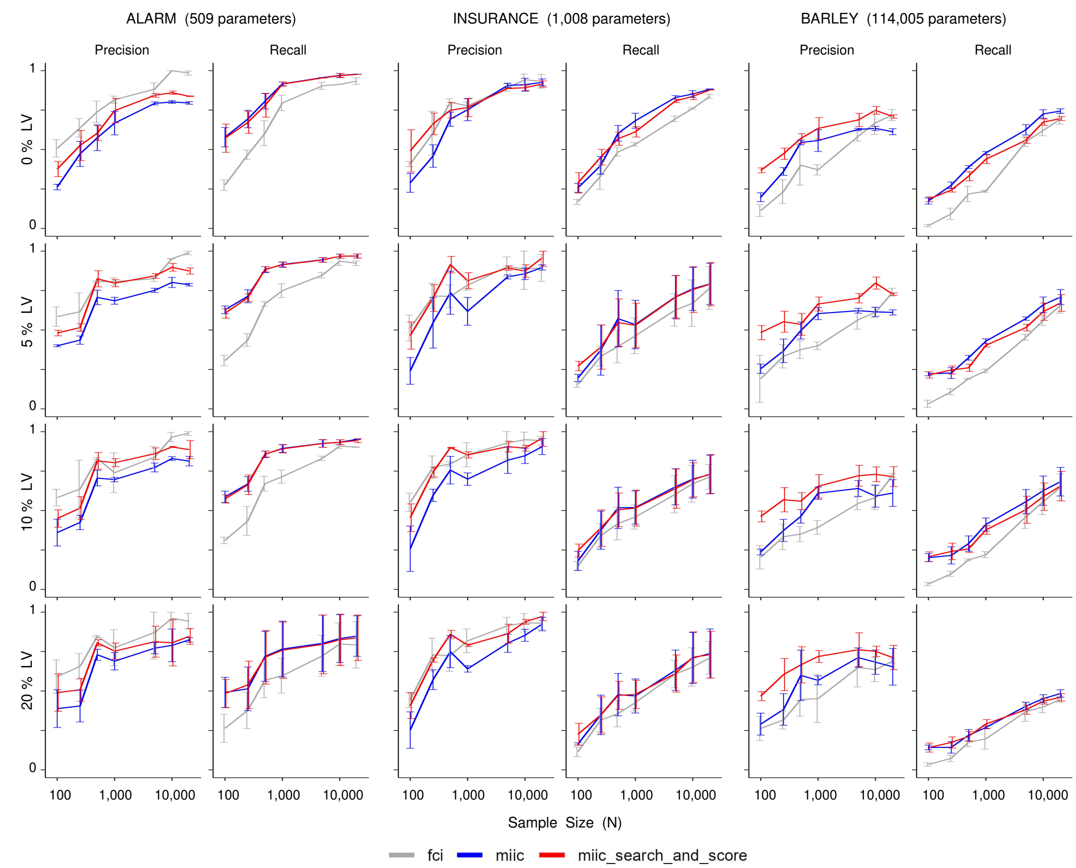

Figure 2: **Benchmark results on ancestral graphs of increasing complexity.** Benchmark results on ancestral graphs obtained by hiding 0%, 5%, 10% or 20% of variables in Bayesian Networks of increasing complexity (see main text): Alarm (lhs), Insurance (middle), and Barley (rhs). MIIC_search&score results are compared to MIIC results used as starting point for MIIC_search&score and FCI [41]. Causal discovery performance is assessed in terms of *Precision* and *Recall* relative to the theoretical PAGs, while counting as false positive all correctly predicted edges but without or with a different orientation as the directed or bidirected edges of the PAG. Error bars ($\pm\sigma$): standard deviations.

hidden variables with observed parents. Irrespective of their orientations, all these additional edges originating from indirect effects of hidden variables generally correspond to weaker effects (*i.e.* lower mutual information of indirect effects due to the Data Processing Inequality) and are more difficult to uncover than the edges of the original graphical model without hidden variables.

5 Limitations

The main limitation of the paper concerns the local scores used in the search-and-score algorithm, which are limited to ac-connected subsets of vertices with a maximum of two-collider paths.

While this approach could be extended to higher-order information contributions including three- or more collider paths, it allows for a simple two-step search-and-score scheme at the level of individual nodes (step 1) and edges (step 2), as detailed in section 3. This already shows a significant improvement in causal discovery performance (*i.e.* combing good precision and good recall on challenging benchmarks) as compared to existing state-of-the-art methods.

References

[1] D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).

[2] J. Pearl, A. Paz, Graphoids: A graph-based logic for reasoning about relevance relations, or when would x tell you more about y if you already know z, *Tech. rep.*, UCLA Computer Science Department (1985).

[3] J. Pearl, *Probabilistic reasoning in intelligent systems* (Morgan Kaufmann, San Mateo, CA, 1988).

[4] J. Pearl, *Causality: models, reasoning and inference* (Cambridge University Press, 2009), second edn.

[5] P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction, and Search* (MIT press, , 2000), second edn.

[6] T. S. Richardson, A factorization criterion for acyclic directed mixed graphs, *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09 (AUAI Press, Arlington, VA, USA, 2009), p. 462–470.

[7] J. Tian, J. Pearl, A general identification condition for causal effects, *Proceedings of the National Conference on Artificial Intelligence* (Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002), pp. 567–573.

[8] T. Richardson, P. Spirtes, Ancestral graph markov models. *Ann. Statist.* **30**, 962–1030 (2002).

[9] M. Drton, M. Eichler, T. S. Richardson, Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research* **10**, 2329–2348 (2009).

[10] R. J. Evans, T. S. Richardson, Maximum likelihood fitting of acyclic directed mixed graphs to binary data, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, UAI'10 (AUAI Press, Corvallis, OR, USA, 2010).

[11] S. Triantafillou, I. Tsamardinos, Score-based vs constraint-based causal learning in the presence of confounders, *CFA@UAI* (2016).

[12] K. Rantanen, A. Hyttinen, M. Järvisalo, Maximal ancestral graph structure learning via exact search, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, C. de Campos, M. H. Maathuis, eds. (PMLR, 2021), vol. 161 of *Proceedings of Machine Learning Research*, pp. 1237–1247.

[13] T. Claassen, I. G. Bucur, Greedy equivalence search in the presence of latent confounders, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, J. Cussens, K. Zhang, eds. (PMLR, 2022), vol. 180 of *Proceedings of Machine Learning Research*, pp. 443–452.

[14] Z. Hu, R. Evans, Faster algorithms for markov equivalence, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, J. Peters, D. Sontag, eds. (PMLR, 2020), vol. 124 of *Proceedings of Machine Learning Research*, pp. 739–748.

[15] W. J. McGill, Multivariate information transmission. *Trans. of the IRE Professional Group on Information Theory (TIT)* **4**, 93-111 (1954).

[16] H. K. Ting, On the amount of information. *Theory Probab. Appl.* **7**, 439-447 (1962).

[17] T. S. Han, Multiple mutual informations and multiple interactions in frequency data. *Information and Control* **46**, 26-45 (1980).

[18] R. W. Yeung, A new outlook on shannon's information measures. *IEEE transactions on information theory* **37**, 466–474 (1991).

[19] P. Spirtes, T. Richardson, A polynomial time algorithm for determinint dag equivalence in the presence of latent variables and selection bias, *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics* (1996).

[20] T. Richardson, Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* **30**, 145-157 (2003).

[21] R. A. Ali, T. S. Richardson, Markov equivalence classes for maximal ancestral graphs, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002), pp. 1–9.

[22] R. A. Ali, T. S. Richardson, P. Spirtes, J. Zhang, Towards characterizing markov equivalence classes for directed acyclic graphs with latent variables, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'05 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005).

[23] J. Tian, Generating markov equivalent maximal ancestral graphs by single edge replacement, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'05 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005).

[24] R. A. Ali, T. S. Richardson, P. Spirtes, Markov equivalence for ancestral graphs. *Ann. Statist.* **37**, 2808–2837 (2009).

[25] T. Richardson, P. Spirtes, Scoring ancestral graph models, *Tech. rep.* (1999). Available as Technical Report CMU-PHIL 98.

[26] J. Zhang, A characterization of markov equivalence classes for directed acyclic graphs with latent variables, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'07 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007).

[27] J. Zhang, On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **172**, 1873-1896 (2008).

[28] S. Affeldt, H. Isambert, Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information, *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands* (2015), pp. 42–51.

[29] L. Verny, N. Sella, S. Affeldt, P. P. Singh, H. Isambert, Learning causal networks with latent variables from multivariate information in genomic data. *PLoS Comput. Biol.* **13**, e1005662 (2017).

[30] B. Andrews, G. F. Cooper, T. S. Richardson, P. Spirtes, The m-connecting imset and factorization for admg models, Preprint (2022). Arxiv 2207.08963.

[31] Z. Hu, R. J. Evans, Towards standard imsets for maximal ancestral graphs. *Bernoulli* **30** (2024).

[32] Z. Hu, R. Evans, A fast score-based search algorithm for maximal ancestral graphs using entropy, Preprint (2024). Arxiv 2402.04777.

[33] V. Cabeli, H. Li, M. da Câmara Ribeiro-Dantas, F. Simon, H. Isambert, Reliable causal discovery based on mutual information supremum principle for finite datasets, *WHY21, 35rd Conference on Neural Information Processing Systems* (NeurIPS, 2021).

[34] M. d. C. Ribeiro-Dantas, H. Li, V. Cabeli, L. Dupuis, F. Simon, L. Hettal, A.-S. Hamy, H. Isambert, Learning interpretable causal networks from very large datasets, application to 400, 000 medical records of breast cancer patients. *iScience* **27**, 109736 (2024).

[35] V. Cabeli, L. Verny, N. Sella, G. Uguzzoni, M. Verny, H. Isambert, Learning clinical networks from medical records based on information estimates in mixed-type data. *PLoS Comput. Biol.* **16**, e1007866 (2020).

[36] S. Affeldt, L. Verny, H. Isambert, 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. *BMC Bioinformatics* **17** (2016).

[37] H. Li, V. Cabeli, N. Sella, H. Isambert, Constraint-based causal structure learning with consistent separating sets. *Advances in Neural Information Processing Systems (NeurIPS)* **32** (2019).

[38] P. Kontkanen, P. Myllymäki, A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* **103**, 227–233 (2007).

[39] T. Roos, T. Silander, P. Kontkanen, P. Myllymäki, Bayesian network structure learning using factorized nml universal models, *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)* (IEEE Press, 2008).

[40] M. Scutari, Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.* **35**, 1–22 (2010).

[41] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, K. Zhang, Causal-learn: Causal discovery in python. *Journal of Machine Learning Research* **25**, 1–8 (2024).

[42] Y. M. Shtarkov, Universal sequential coding of single messages. *Problems of Information Transmission* **23**, 3–17 (1987).

[43] J. Rissanen, I. Tabus, *Adv. Min. Descrip. Length Theory Appl.* (MIT Press, 2005), pp. 245–264.

[44] W. Szpankowski, *Average case analysis of algorithms on sequences* (John Wiley & Sons, , 2001).

[45] P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, H. Tirri, Efficient computation of stochastic complexity. *in: C. Bishop, B. Frey (Eds.) Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics* **103**, 233–238 (2003).

[46] P. Kontkanen, Computationally efficient methods for mdl-optimal density estimation and data clustering, Ph.D. thesis (2009).

[47] D. M. Chickering, A Transformational Characterization of Equivalent Bayesian Network Structures, *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, 1995), pp. 87–98.

[48] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, P. Bühlmann, Causal inference using graphical models with the r package pcalg. *J. Stat. Softw.* **47**, 1–26 (2012).

Appendix / supplemental material

A Preliminaries: connection and separation criteria

A.1 *m*-connection *vs* *m'*-connection criteria

An ancestral graph can be interpreted as encoding a set of conditional indepencence relations by a graphical criterion, called m-separation, based on the concept of m-connecting paths, which generalizes the separation criteria of Markov and Bayesian networks to ancestral graphs.

Definition 4. [m-connecting path] A path $\boldsymbol{\pi}$ between X and Y is m-connecting given a (possibly empty) subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with $X, Y \notin \boldsymbol{C}$) if:

i) its non-collider(s) are not in $\boldsymbol{C}$, and

ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{C})$.

Definition 5. [m-separation criterion] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be m-separated by $\boldsymbol{C}$, noted $\boldsymbol{A} \perp_m \boldsymbol{B} | \boldsymbol{C}$, if there is no m-connecting path between any vertex in $\boldsymbol{A}$ and any vertex in $\boldsymbol{B}$ given $\boldsymbol{C}$.

The probabilistic interpretation of ancestral graph is given by its global and pairwise Markov properties (which are equivalent [8]): if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated by $\boldsymbol{C}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$ and $\forall X \in \boldsymbol{A}$ and $\forall Y \in \boldsymbol{B}$, there is a probability distribution P faithful to $\mathcal{G}$ such that their conditional mutual information vanishes, *i.e.* $I_P(X;Y|\boldsymbol{C}) = 0$, also noted $X \perp\!\!\!\perp_P Y | \boldsymbol{C}$.

However, as discussed above, the proof of Theorem 1 will require to introduce a weaker m'-connection criterion defined below.

Definition 6. [m'-connecting path] A path $\boldsymbol{\pi}$ between X and Y is m'-connecting given a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with X, Y possibly in $\boldsymbol{C}$) if:

i) its non-collider(s) are not in $\boldsymbol{C}$, and

ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$.

Note, in particular, that an m-connecting path is necessary an m'-connecting path but that the converse is not always true. For example, the path $X \to Z \longleftrightarrow T \longleftrightarrow Y$ in Fig. 1G (with $Z \to Y$) is an m'-connecting path given T (as $Z \in \mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup T)$) but not an m-connecting path given T (as $Z \notin \mathbf{An}_{\mathcal{G}}(T)$).

However, Richardson and Spirtes 2002 [8] have shown the following lemma,

Lemma 4. [Corollary 3.15 in [8]] *In an ancestral graph $\mathcal{G}$, there is a m'-connecting path $\boldsymbol{\mu}$ between X and Y given $\boldsymbol{C}$ if and only if there is a (possibly different) m-connecting path $\boldsymbol{\pi}$ between X and Y given $\boldsymbol{C}$.*

Hence, Lemma 4 implies that m'-separation and m-separation criteria are in fact equivalent, as an absence of m'-connecting paths implies an absence of m-connecting paths and vice versa. This enables to reformulate the m-separation criterion above as,

Definition 7. [m'-separation (and m-separation) criteria] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be m'-separated (or m-separated) by $\boldsymbol{C}$, if all paths from any $X \in \boldsymbol{A}$ to any $Y \in \boldsymbol{B}$ have either

i) a non-collider in $\boldsymbol{C}$, or

ii) a collider *not* in $\mathbf{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{C})$.

The probabilistic interpretation of an ancestral graph is given by its (global) Markov property: if $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated (or m'-separated) by $\boldsymbol{C}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$, noted as, $\boldsymbol{A} \perp_m \boldsymbol{B} | \boldsymbol{C}$.

A.2 *ac*-connecting paths and *ac*-connected subsets

Let us now recall the definition of **ancestor collider connecting paths** or ***ac*-connecting paths**, which is directly relevant to characterize the likelihood decomposition and Markov equivalent classes of ancestral graphs (Theorem 1). We give here a different yet equivalent definition of *ac*-connecting paths as defined in the main text (Definition 2) in order to underline the similarities and differencies with the notion of m'-connecting path (Definition 6).

Definition 8. [ac-connecting path] A path $\boldsymbol{\pi}$ between X and Y is an ac-connecting path given a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ (with X and Y possibly in $\boldsymbol{C}$) if:

i) $\boldsymbol{\pi}$ does not have any noncollider, and

ii) its collider(s) are in $\mathbf{An}_{\mathcal{G}}(\{X,Y\} \cup \boldsymbol{C})$.

Hence, more simply (following Definition 2 in the main text), an ac-connecting path given $\boldsymbol{C}$ is a collider path, $X \ast\!\!\rightarrow Z_1 \leftrightarrow \cdots \leftrightarrow Z_K \leftarrow\!\!\ast\, Y$, with all $Z_i \in \mathbf{An}_{\mathcal{G}}(\{X,Y\} \cup \boldsymbol{C})$, *i.e.* with Z_i in $\boldsymbol{C}$ or connected to $\{X,Y\} \cup \boldsymbol{C}$ by an ancestor path, $Z_i \rightarrow \cdots \rightarrow T$ with $T \in \{X,Y\} \cup \boldsymbol{C}$.

Definition 9. [ac-separation criterion] The subsets $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be ac-separated by $\boldsymbol{C}$ if there is no ac-connecting path between any vertex in $\boldsymbol{A}$ and any vertex in $\boldsymbol{B}$ given $\boldsymbol{C}$.

Previous definitions and Lemma 4 readily lead to the following corollary between the different connection and separation criteria:

Corollary 5.

i) m*-connecting path* $\pi \implies m'$*-connecting path* π

ii) ac*-connecting path* $\pi \implies m'$*-connecting path* π

iii) m*-separation* $\iff m'$*-separation*

iv) m/m'*-separation* $\implies ac$*-separation*

Finally, we recall the notion of $\boldsymbol{ac}$**-connected subset** (Definition 3 in the main text), which is central for the decomposition of the likelihood of ancestral graphs (Theorem 1): A subset $\boldsymbol{C}$ is said to be ac-connected if $\forall X, Y \in \boldsymbol{C}$, there is an ac-connecting path between X and Y w.r.t. $\boldsymbol{C}$.

B Proof of Theorem 1.

In order to prove that the likelihood function of an ancestral graph, Eq. 12, contains all and only the ac-connected subsets of vertices in $\mathcal{G}$ (Definition 3), we will first show (***i***) that all non-ac-connected subsets $\boldsymbol{S'}$ are included in a cancelling combination of multivariate information terms $I(X;Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \boldsymbol{S} = \{X,Y\} \cup \boldsymbol{A}$. Conversely, we will then show (***ii***) that cancelling combinations of multivariate information terms associated to pairwise conditional independence, $I(X;Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'} \subseteq \boldsymbol{S}}^{X,Y \in \boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}) = 0$ do not contain any ac-connected subset $\boldsymbol{S'}$. Finally, we will prove (***iii***) that the information terms which appear in multiple cancelling combinations from different pairwise independence constraints do not modify the multivariate information decomposition of the likelihood function of ancestral graphs, Eq. 12, as these shared/overlapping terms in fact all cancel through more global Markov independence relationships involving higher order (three or more points) vanishing multivariate information terms, such as $I(X;Y;Z|\boldsymbol{A}) = 0$.

i) Let's first prove that all non-ac-connected subsets $\boldsymbol{S'}$ are included in at least one cancelling combination of multivariate information terms, $I(X;Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \{X,Y\} \cup \boldsymbol{A}$.

If $\boldsymbol{S'}$ is a non-ac-connected subset, there is at least one disconnected pair X and Y for which each path π_j between X and Y contains either some collider(s) not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ or, if all colliders along π_j are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, there must be some non-collider(s) at node(s) $\boldsymbol{Z}_j$ but not necessarily in $\boldsymbol{S'}$. Let's define $\boldsymbol{S} = \boldsymbol{S'} \cup_j \boldsymbol{Z}_j$. X and Y can be shown to be m-separated given $\boldsymbol{S} \setminus \{X,Y\}$, as for each path π_j between X and Y, its non-collider(s) are in $\boldsymbol{S}$ at node(s) $\boldsymbol{Z}_j$ (when all collider(s) along π_j are in $\boldsymbol{S'}$) or there is some collider(s) not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which are not in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ either. The latter statement is proven by contradiction assuming that there is a collider at $Z \notin \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ such that $Z \in \mathbf{An}_{\mathcal{G}}(\boldsymbol{S})$. There is therefore a directed path $Z \rightarrow \cdots \rightarrow W$ with $W \in \boldsymbol{S}$. Hence, $W \in \boldsymbol{S'}$ or there is a noncollider at $W \in \boldsymbol{Z}_j$ which is on a path π_j between X and Y along which all colliders are in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ by construction of $\boldsymbol{S}$. This leads by induction to $Z \rightarrow \cdots \rightarrow W \rightarrow \cdots \rightarrow T$ where $T \in \boldsymbol{S'}$ and thus $Z \in \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which is a contradiction. Hence, all non-ac-connected subsets $\boldsymbol{S'}$ are included in a cancelling combination of multivariate information terms $I(X;Y|\boldsymbol{A}) = 0$, with $X, Y \in \boldsymbol{S'}$ and $\boldsymbol{S'} \subseteq \boldsymbol{S} = \{X,Y\} \cup \boldsymbol{A}$.

ii) Conversely, we will now show that cancelling combinations of multivariate information terms associated to pairwise conditional independence, $I(X;Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'} \subseteq \boldsymbol{S}}^{X,Y \in \boldsymbol{S'}} (-1)^{|\boldsymbol{S'}|} I(\boldsymbol{S'}) = 0$, do not contain any ac-connected subset $\boldsymbol{S'}$.

We will prove it by contradiction assuming that there exists a subset $\boldsymbol{W} \subseteq \boldsymbol{A}$, such that $\boldsymbol{S'} = \{X,Y\} \cup \boldsymbol{W}$ is ac-connected. In particular, there should be an ac-connecting path between X and Y

confined to $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$ and thus to $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S}) \supseteq \mathbf{An}_{\mathcal{G}}(\boldsymbol{S'})$, which is an m'-connecting path between X and Y given $\boldsymbol{A}$, contradicting the above hypothesis of m'-separation given $\boldsymbol{A}$, *i.e.* $I(X;Y|\boldsymbol{A}) = 0$. The use of m'-separation, *i.e.* the absence of m'-connecting paths with colliders in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{S})$ rather than m-connecting paths with colliders in $\mathbf{An}_{\mathcal{G}}(\boldsymbol{A})$, is necessary here, see Definitions 4 and 6. Hence, no ac-connected subset $\boldsymbol{S'}$ is included in cancelling combinations of multivariate information terms associated to pairwise conditional independence, $I(X;Y|\boldsymbol{A}) = \sum_{\boldsymbol{S'}\subseteq\boldsymbol{S}}^{X,Y\in\boldsymbol{S'}}(-1)^{|\boldsymbol{S'}|}I(\boldsymbol{S'}) = 0$.

iii) Finally, we will show that the information terms which appear in multiple cancelling combinations from different pairwise independence constraints do not modify the multivariate information decomposition of the likelihood function of ancestral graphs, Eq. 12, as these shared/overlapping terms in fact all cancel through more global Markov independence relationships involving higher order (three or more points) vanishing multivariate information terms, such as $I(X;Y;Z|\boldsymbol{A}) = 0$.

This result requires to use an ordering of the nodes, $X_k \succ X_j \succ X_i$, that is compatible with the directed edges of the ancestral graph assumed to have no undirected edges, *i.e.* $X_j \notin \mathbf{An}(X_i)$ if $X_j \succ X_i$. Under this ordering, higher order nodes $X_k \succ X_i \succ X_j$ can be a priori excluded from all separating sets $\boldsymbol{A}_{ij}$ of pairs of lower order nodes, *i.e.* if $I(X_i;X_j|\boldsymbol{A}_{ij}) = 0$ then $X_k \notin \boldsymbol{A}_{ij}$.

In particular, the two pairwise conditional independence relations $I(X_k;X_\ell|\boldsymbol{A}_{k\ell}) = 0$, with $X_\ell \succ X_k$, and $I(X_i;X_j|\boldsymbol{A}_{ij}) = 0$, with $X_j \succ X_i$, do not share any multivariate information terms, if $X_\ell \neq X_j$. Indeed, as $I(X_k;X_\ell|\boldsymbol{A}_{k\ell})$ contains all information terms including both X_k and X_ℓ as well as every subset (possibly empty) of $\boldsymbol{A}_{k\ell}$, none of them includes X_j if $X_\ell \succ X_j$. Therefore $I(X_k;X_\ell|\boldsymbol{A}_{k\ell})$ does not contain any information term of $I(X_i;X_j|\boldsymbol{A}_{ij})$ which contains both X_i and X_j as well as every subset (possibly empty) of $\boldsymbol{A}_{ij}$. This property eliminates all multiple counting of multivariate informations terms shared if $X_\ell \neq X_j$. Note that this result does not hold in general for ancestral graphs including undirected edges.

Hence, the issue of redundant multivariate information terms in the likelihood decomposition, Eq. 12, is related to the conditional independences of two or more pairs, $\{X_i, X_r\}$, $\{X_j, X_r\}$, ..., $\{X_\ell, X_r\}$, sharing the same higher order node, X_r. However, this situation also entails a more global Markov independence constraint between X_r and $\{X_i, X_j, \cdots, X_\ell\}$, given a separating set $\boldsymbol{A}$, which can be decomposed into more local independence constraints using the chain rule and the decomposition rules of multivariate information (Eq. 9),

$$
\begin{aligned}
0 &= I(\{X_i, X_j, \cdots, X_\ell\}; X_r|\boldsymbol{A})\\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A},X_i)\big) + \big[I(X_k;X_r|\boldsymbol{A},X_i,X_j)\big] + \cdots + I(X_\ell;X_r|\boldsymbol{A},\cdots)\\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A}) - I(X_i;X_j;X_r|\boldsymbol{A})\big)\\
&\quad + \big[I(X_k;X_r|\boldsymbol{A},X_i) - I(X_j;X_k;X_r|\boldsymbol{A},X_i)\big] + \cdots + I(X_\ell;X_r|\boldsymbol{A},\cdots)\\
&= \big(I(X_i;X_r|\boldsymbol{A}) + I(X_j;X_r|\boldsymbol{A}) - I(X_i;X_j;X_r|\boldsymbol{A})\big)\\
&\quad + \big[I(X_k;X_r|\boldsymbol{A}) - I(X_j;X_k;X_r|\boldsymbol{A}) - I(X_i;X_k;X_r|\boldsymbol{A}) + I(X_i;X_j;X_k;X_r|\boldsymbol{A})\big] + \cdots
\end{aligned}
$$

where all the conditional multivariate information terms vanish by induction due to the non-negativity of (conditional) mutual information. In particular, the conditional multivariate information terms in the last expression, *i.e.* between X_r and each subset of $\{X_i, X_j, \cdots, X_\ell\}$ given the separating set $\boldsymbol{A}$, all vanish. This result can be readily extended to any subsets $\{X_r, X_s, \cdots, X_z\}$ (conditionally) independent of $\{X_i, X_j, \cdots, X_\ell\}$ given a separating set $\boldsymbol{A}$, *i.e.* $I(\{X_i, X_j, \cdots, X_\ell\}; \{X_r, X_s, \cdots, X_z\}|\boldsymbol{A}) = 0$. Hence, as the final conditional multivariate cross information terms of the decomposition all vanish while not sharing any subsets of variables, it proves the absence of redundancy and a global cancellation of non-ac-connected subsets (from pairwise and higher order conditional independence relations) in the likelihood function of ancestral graphs without undirected edges, Eq. 12.

Hence, only ac-connected subsets effectively contribute to the cross-entropy of an ancestral graph with only directed and bidirected edges, Eq. 12. $\square$

C Factorization of the probability distribution of ancestral graphs

C.1 Factorization resulting from Theorem 1 and Proposition 3

Before presenting the factorization of the model distribution of ancestral graphs resulting from Theorem 1 and Proposition 3, it is instructive to obtain an equivalent factorization for Bayesian graphs, assuming a positive empirical distributions, $p(x_1, \cdots, x_m) = \prod_{i=1}^{m} p(x_i|x_{i-1}, \cdots, x_1) > 0$,

$$\begin{aligned} q(x_1, \cdots, x_m) &= \prod_{i=1}^{m} q(x_i|\mathbf{pa}_{x_i}) = \prod_{i=1}^{m} p(x_i|\mathbf{pa}_{x_i}) \\ &= p(x_1, \cdots, x_m) \prod_{i=1}^{m} \frac{p(x_i|\mathbf{pa}_{x_i})}{p(x_i|x_{i-1}, \cdots, x_1)} \\ &= p(x_1, \cdots, x_m) \prod_{i=1}^{m} \frac{p(x_i|\mathbf{pa}_{x_i}) p(\boldsymbol{x}_{i-1} \backslash \mathbf{pa}_{x_i}|\mathbf{pa}_{x_i})}{p(x_i, \boldsymbol{x}_{i-1} \backslash \mathbf{pa}_{x_i}|\mathbf{pa}_{x_i})} \end{aligned} \tag{15}$$

This leads to the following alternative expressions for the cross-entropy $H(p, q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$ in terms of multivariate entropy and information, which only depend on the empirical joint distribution $p(\boldsymbol{x})$,

$$\begin{aligned} H(p, q) &= \sum_{i=1}^{m} H(x_i|\mathbf{Pa}_{X_i}) \\ &= H(X_1, \cdots, X_m) + \sum_{i=1}^{m} I(X_i; \boldsymbol{X}_{i-1} \backslash \mathbf{Pa}_{X_i}|\mathbf{Pa}_{X_i}) \end{aligned} \tag{16}$$

where $\sum_{i=1}^{m} I(X_i; \boldsymbol{X}_{i-1} \backslash \mathbf{Pa}_{X_i}|\mathbf{Pa}_{X_i})$ can be decomposed, using the chain rule and Eq. 11, into unconditional multivariate information terms, which exactly cancel all the multivariate information of the non-ac-connected subsets of variables in the multivariate entropy decomposition, Eq. 6.

Note, however, that this result obtained for Bayesian networks requires an explicit factorization of the global model distribution, $q(\boldsymbol{x})$, in terms of the empirical distribution, $p(\boldsymbol{x})$, which is not known and presumably does not exist, in general, for ancestral graphs.

Alternatively, assuming that the empirical and model distributions are positive ($\forall \boldsymbol{x}, p(\boldsymbol{x}) > 0$, $q(\boldsymbol{x}) > 0$), it is always possible to factorize them into factors associated to each (cross) information term in the (cross) entropy decomposition, Eq. 6, as,

$$q(\boldsymbol{x}) = \prod_{i=1}^{m} q(x_i) \times \prod_{i<j}^{m} \frac{q(x_i, x_j)}{q(x_i)q(x_j)} \times \prod_{i<j<k}^{m} \frac{q(x_i, x_j, x_k)q(x_i)q(x_j)q(x_k)}{q(x_i, x_j)q(x_i, x_k)q(x_j, x_k)} \times \cdots \tag{17}$$

where all the marginal distributions over a subset of variables, *e.g.* $q(x_i, x_j, x_k) = \sum_{\ell \neq i,j,k} q(\boldsymbol{x})$ or $p(x_i, x_j, x_k) = \sum_{\ell \neq i,j,k} p(\boldsymbol{x})$, cancel two-by-two by construction.

This can be illustrated on a simple example of a two-collider path including one bidirected edge, $X \rightarrow Z \longleftrightarrow Y \leftarrow W$ (Fig. 1D), valid for $q(.)$ and $p(.)$ alike,

$$\begin{aligned} q(x, z, y, w) = \; & q(x)\, q(z)\, q(y)\, q(w) \\ & \times \frac{q(x,z)}{q(x)\, q(z)} \frac{q(z,y)}{q(z)\, q(y)} \frac{q(y,w)}{q(y)\, q(w)} \frac{q(x,y)}{q(x)\, q(y)} \frac{q(x,w)}{q(x)\, q(w)} \frac{q(z,w)}{q(z)\, q(w)} \\ & \times \frac{q(x)\, q(z)\, q(y)\, q(x,z,y)}{q(x,z)\, q(x,y)\, q(z,y)} \frac{q(z)\, q(y)\, q(w)\, q(z,y,w)}{q(z,y)\, q(z,w)\, q(y,w)} \\ & \times \frac{q(x)\, q(z)\, q(w)\, q(x,z,w)}{q(x,z)\, q(x,w)\, q(z,w)} \frac{q(x)\, q(y)\, q(w)\, q(x,y,w)}{q(x,y)\, q(x,w)\, q(y,w)} \\ & \times \frac{q(x,z)\, q(z,y)\, q(y,w)\, q(x,y)\, q(x,w)\, q(z,w)\, q(x,z,y,w)}{q(x,z,y)\, q(x,z,w)\, q(x,y,w)\, q(z,y,w)\, q(x)\, q(y)\, q(z)\, q(w)} \end{aligned} \tag{18}$$

where all individual distribution marginals on subsets of variables, *e.g.* $q(x)$, $q(x, z)$, $q(x, z, y)$ (or $p(x)$, $p(x, z)$, $p(x, z, y)$), cancel two-by-two by construction, except $q(x, z, y, w)$ (or $p(x, z, y, w)$).

In addition and *only for the model distribution* $q(.)$, all ratios in gray in Eq. 18 also cancel due to Markov independence relations across non-ac-connected subsets (see proof of Theorem 1). This leaves a truncated factorization retaining all and only the ac-connected subsets of variables in the graph, which we propose to estimate on empirical data by substituting the remaining $q(.)$ terms by their empirical counterparts $p(.)$, see Proposition 3.

This leads to the following global factorization for $q(.)$ in terms of $p(.)$,

$$\begin{aligned}
q(x,z,y,w) \;&\equiv\; p(x)\,p(z)\,p(y)\,p(w)\,\frac{p(x,z)}{p(x)\,p(z)}\,\frac{p(z,y)}{p(z)\,p(y)}\,\frac{p(y,w)}{p(y)\,p(w)}\\
&\times\frac{p(x)\,p(z)\,p(y)\,p(x,z,y)}{p(x,z)\,p(x,y)\,p(z,y)}\,\frac{p(z)\,p(y)\,p(w)\,p(z,y,w)}{p(z,y)\,p(z,w)\,p(y,w)}\\
&\times\frac{p(x,z)\,p(z,y)\,p(y,w)\,p(x,y)\,p(x,w)\,p(z,w)\,p(x,z,y,w)}{p(x,z,y)\,p(x,z,w)\,p(x,y,w)\,p(z,y,w)\,p(x)\,p(y)\,p(z)\,p(w)}\\
&=\; p(x,z,y,w)\,\frac{p(x)\,p(y)}{p(x,y)}\,\frac{p(x)\,p(w)}{p(x,w)}\,\frac{p(z)\,p(w)}{p(z,w)}\\
&\times\frac{p(x,z)\,p(x,w)\,p(z,w)}{p(x)\,p(z)\,p(w)\,p(x,z,w)}\,\frac{p(x,y)\,p(x,w)\,p(y,w)}{p(x)\,p(y)\,p(w)\,p(x,y,w)}
\end{aligned}\tag{19}$$

where the terms in gray have been passed to the lhs of Eq. 18 applied to $p(.)$. This ultimately leads to the analog of the Bayesian Network factorization in Eq. 15 but for the two-collider path, $X \rightarrow Z \longleftrightarrow Y \leftarrow W$ (Fig. 1D),

$$q(x,z,y,w) \;\equiv\; p(x,z,y,w)\,\frac{p(x)\,p(w)}{p(x,w)}\,\frac{p(z|x)\,p(w|x)}{p(z,w|x)}\,\frac{p(x|w)\,p(y|w)}{p(x,y|w)}\tag{20}$$

where the last three factors "correct" the expression of $p(x,z,y,w)$ for the three (conditional) independences entailed by the underlying graph, that is, $X \perp W$, $Z \perp W|X$, and $X \perp Y|W$.

C.2 Relation to the head-and-tail factorizations

The head-and-tail factorizations of the model distribution of an acyclic directed mixed graph, introduced by Richardson 2009 [6], enable the parametrization of the joint probability distribution with independent parameters for ancestrally closed subsets of vertices.

For instance, the head-and-tail factorizations of the simple two-collider path including one bidirected edge, $X \rightarrow Z \longleftrightarrow Y \leftarrow W$, introduced above, Fig. 1D, are [6],

$$\begin{aligned}
q(x,w) &= q(x)\,q(w)\\
q(x,z) &= q(z|x)\;q(x)\\
q(y,w) &= q(y|w)\;q(w)\\
q(x,z,w) &= q(z|x)\;q(x)\,q(w)\\
q(x,y,w) &= q(y|w)\;q(w)\,q(x)\\
q(x,z,y,w) &= q(z,y|x,w)\;q(x)\,q(w)
\end{aligned}\tag{21}$$

Importantly, these head-and-tail factorizations imply additional relations such as $q(y|w) = q(y|x,w)$ (*i.e.* $X \perp Y|W$) obtained by comparing the last two relations in Eq. 21 after marginalizing $q(x,z,y,w)$ over z. However, such implicit conditional independence relations are *not verified by the empirical distribution* $p(.)$ *in general* and prevent the estimation of the head-and-tail factorizations by substituting the rhs $q(.)$ terms in Eq. 21 with their empirical counterparts $p(.)$, as in the case of Bayesian networks, Eq. 15.

Indeed, while the head-and-tail factorization relations, Eq. 21, obey the local and global Markov independence relations entailed by the graphical model, Fig. 1D, leading to the cancellation of all factors associated to non-ac-connected subsets in gray in Eq. 18, the remaining head-and-tail factors cannot be readily estimated with the empirical distribution $p(.)$.

In particular, the cross-entropy of the two-collider path of interest, Fig. 1D, obtained with the head-and-tail factorizations corresponds to[1] $H(p,q) = -\sum p(x,z,y,w)\log q(z,y|x,w)\,q(x)\,q(w)$. Then, estimating the $q(.)$ terms with their $p(.)$ counterparts leads to the cross-entropy of a Bayesian graph, Fig. 1E, with a different Markov equivalent class than the ancestral graph of interest, Fig. 1D. A similar discrepancy is obtained with a c-component factorization which leads to the cross-entropy of the Bayesian graph of Fig. 1E without edge $X \to Y$, corresponding to a different Markov equivalence class than the previous two graphs, Figs. 1D & E.

These examples illustrate the difficulty to exploit the c-component or head-and-tail factorizations to estimate the likelihood of ancestral graphs including bidirected edge(s).

D Node and edge scores based on Normalized Maximum Likelihood criteria

Search-and-score methods based on likelihood estimates need to properly account for finite sample size, as cross-entropy minimization leads to ever more complex models, resulting in model overfitting for finite datasets. While BIC regularization is valid in the asymptotic limit of very large datasets, it tends to overestimate finite size corrections, leading to lower recall, in general. In order to better take into account finite sample size, we used instead the (universal) Normalized Maximum Likelihood (NML) criteria [42, 43, 38, 39], which amounts to normalizing the likelihood function over all possible datasets with the same number N of samples.

Node score. We first used the factorized Normalized Maximum Likelihood (fNML) complexity [38, 39] to define a local score for each node X_i, which extends the decomposable likelihood of Bayesian graphs given each node's parents, Eq. 2, to all non-descendant neighbors, $\mathbf{Pa'}_{X_i}$,

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\mathrm{X_i}}} = e^{-N.\,\mathrm{Score_n}(X_i)} &= \frac{e^{-NH(X_i|\mathbf{Pa'}_{X_i})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(X_i|\mathbf{Pa'}_{X_i})}} && (22)\\
&= e^{-NH(X_i|\mathbf{Pa'}_{X_i}) - \sum_j^{q_i}\log \mathcal{C}^{r_i}_{n_j}} && (23)\\
&= e^{N\sum_j^{q_i}\sum_k^{r_i}\frac{n_{jk}}{N}\log\left(\frac{n_{jk}}{n_j}\right) - \sum_j^{q_i}\log \mathcal{C}^{r_i}_{n_j}} && (24)\\
&= \prod_j^{q_i} \frac{\prod_k^{r_i}\left(\frac{n_{jk}}{n_j}\right)^{n_{jk}}}{\mathcal{C}^{r_i}_{n_j}} && (25)
\end{aligned}$$

where n_{jk} corresponds to the number of data points for which X_i is in its kth state and its non-descendant neighbors in their jth state, with $n_j = \sum_k^{r_i} n_{jk}$. The universal normalization constant $\mathcal{C}^r_n$ is then computed by summing the numerator over all possible partitions of the n data points into a maximum of r subsets, $\ell_1 + \ell_2 + \cdots + \ell_r = n$ with $\ell_k \geqslant 0$,

$$\mathcal{C}^r_n = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n} \frac{n!}{\ell_1!\ell_2!\cdots\ell_r!}\prod_{k=1}^{r}\left(\frac{\ell_k}{n}\right)^{\ell_k} \quad (26)$$

which can in fact be computed in linear-time using the following recursion [38],

$$\mathcal{C}^r_n = \mathcal{C}^{r-1}_n + \frac{n}{r-2}\mathcal{C}^{r-2}_n \quad (27)$$

with $\mathcal{C}^1_n = 1$ for all n and applying Eq. 30 below for $r = 2$. However, for large n and r, $\mathcal{C}^r_n$ computation tends to be numerically unstable, which can be circumvented by implementing the recursion on parametric complexity ratios $\mathcal{D}^r_n = \mathcal{C}^r_n/\mathcal{C}^{r-1}_n$ rather than parametric complexities themselves [35] as,

$$\begin{aligned}
\mathcal{D}^r_n &= 1 + \frac{n}{(r-2)\mathcal{D}^{r-1}_n} && (28)\\
\log \mathcal{C}^r_n &= \sum_{k=2}^{r}\log \mathcal{D}^k_n && (29)
\end{aligned}$$

[1] Indeed, all terms in Eq. 18 actually cancel two-by-two by construction, *whatever their factorization expression*, except for the remaining joint-distribution over all variables, $q(x,z,y,w) = q(z,y|x,w)\,q(x)\,q(w)$.

for $r \geqslant 3$, with $\mathcal{C}_n^1 = 1$ and $\mathcal{C}_n^2 = \mathcal{D}_n^2$, which can be computed directly with the general formula, Eq. 26, for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{30}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) [44–46],

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{31}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{32}$$

This leads to the following local score for each node X_i, which is minimized over alternative combinations of non-descendant neighbors, $\mathbf{Pa'}_{X_i} \subseteq \mathbf{Pa}_{X_i} \cup \mathbf{Sp}_{X_i} \cup \mathbf{Ne}_{X_i}$, in the first step of the local search-and-score algorithm (step 1) detailed in the main text,

$$\mathrm{Score_n}(X_i) = H(X_i|\mathbf{Pa'}_{X_i}) + \frac{1}{N}\sum_{j}^{q_{x_i}} \log \mathcal{C}_{n_j}^{r_{x_i}} \tag{33}$$

Edge scores. We then defined several edge scores to optimize the orientation of each edge, $X - Y$, given its close surrounding vertices.

To this end, we first introduced a local score for node pairs which simply sums the node scores, Eq. 33, for each node. The resulting pair scores are listed in Table 2 for unconnected node pairs and for pairs of nodes connected by a directed edge, where $\mathbf{Pa'}_{X\backslash Y} = \mathbf{Pa}_X \cup \mathbf{Sp}_X \backslash Y$ and $\mathbf{Pa'}_{Y\backslash X} = \mathbf{Pa}_Y \cup \mathbf{Sp}_Y \backslash X$ with their corresponding combinations of levels, $q_{y\backslash x}$ and $q_{x\backslash y}$.

Table 2: Local scores for node pairs

Pair score	Information	fNML Complexity
$X \not\to Y$	$H(X\|\mathbf{Pa'}_{X\backslash Y}) + H(Y\|\mathbf{Pa'}_{Y\backslash X})$	$\frac{1}{N}\Big(\sum_j^{q_{x\backslash y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\backslash x}} \log \mathcal{C}_{n_j}^{r_y} \Big)$
$X \to Y$	$H(X\|\mathbf{Pa'}_{X\backslash Y}) + H(Y\|\mathbf{Pa'}_{Y\backslash X}, X)$	$\frac{1}{N}\Big(\sum_j^{q_{x\backslash y}} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\backslash x}\, r_x} \log \mathcal{C}_{n_j}^{r_y} \Big)$
$X \leftarrow Y$	$H(X\|\mathbf{Pa'}_{X\backslash Y}, Y) + H(Y\|\mathbf{Pa'}_{Y\backslash X})$	$\frac{1}{N}\Big(\sum_j^{q_{x\backslash y}\, r_y} \log \mathcal{C}_{n_j}^{r_x} + \sum_j^{q_{y\backslash x}} \log \mathcal{C}_{n_j}^{r_y} \Big)$

Then, edge scores for directed edges, $X \to Y$ and $Y \to X$, are defined w.r.t. to the edge removal score, $X \not\to Y$, by substracting the pair scores of unconnected pairs to the pair scores of directed edges, leading to the following edge orientation scores,

$$\mathrm{Score}(X \to Y) \quad = \quad -I(X;Y|\mathbf{Pa'}_{Y\backslash X}) + \frac{1}{N}\Big(\sum_{j}^{q_{y\backslash x}\, r_x} \log \mathcal{C}_{n_j}^{r_y} - \sum_{j}^{q_{y\backslash x}} \log \mathcal{C}_{n_j}^{r_y} \Big) \tag{34}$$

$$\mathrm{Score}(Y \to X) \quad = \quad -I(X;Y|\mathbf{Pa'}_{X\backslash Y}) + \frac{1}{N}\Big(\sum_{j}^{q_{x\backslash y}\, r_y} \log \mathcal{C}_{n_j}^{r_x} - \sum_{j}^{q_{x\backslash y}} \log \mathcal{C}_{n_j}^{r_x} \Big) \tag{35}$$

However, if $r_x \neq r_y$, the fNML complexities of these orientation scores are not identical for Markov equivalent edge orientations between nodes sharing the same parents (or spouses) [47], $\mathbf{Pa'}_{Y\backslash X} = \mathbf{Pa'}_{X\backslash Y} = \mathbf{Pa'}$ and $q_{y\backslash x} = q_{x\backslash y}$, despite sharing the same conditional mutual information,

$$I(X;Y|\mathbf{Pa'}) \quad = \quad \frac{1}{2}\Big(H(X|\mathbf{Pa'}) + H(Y|\mathbf{Pa'}, X)\Big) + \frac{1}{2}\Big(H(X|\mathbf{Pa'}, Y) + H(Y|\mathbf{Pa'})\Big) \tag{36}$$

This suggests to symmetrize the fNML complexities for edge orientation scores by averaging them over each directed orientation, as for the conditional information in Eq. 36, leading to the proposed fNML complexity for directed edges given in Table 1 in the main text.

For bidirected edges, the proposed local orientation score accounts for all ac-connected subsets in close vicinity of the bidirected edge, which concerns all subsets including either X and any combination (possibly void) of parents or spouses different from Y (*i.e.* corresponding to the information contributions $H(X|\mathbf{Pa}'_{X\backslash Y})$) or Y and any combination of parents or spouses different from X (*i.e.* corresponding to the information contributions $H(Y|\mathbf{Pa}'_{Y\backslash X})$) or, else, including both nodes X and Y plus any combination of their parents or spouses, corresponding to the following information contribution, $-I(X;Y|\mathbf{Pa}'_{XY})$, where $\mathbf{Pa}'_{XY} = \mathbf{Pa}'_{X\backslash Y} \cup \mathbf{Pa}'_{Y\backslash X}$. This last term, $-I(X;Y|\mathbf{Pa}'_{XY})$, contains all the remaining information contributions once the bidirected orientation score is given relative to the edge removal score (Table 2) as for the two directed orientation scores, above. Finally, the symmetrized fNML complexity associated with a bidirected edge should be computed with the whole set of conditioning parents or spouses, $\mathbf{Pa}'_{XY}$, as indicated in Table 1. Note that this bidirected orientation score becomes also Markov equivalent to the two directed orientation scores, as required, when the nodes share the same parents and spouses, *i.e.* $\mathbf{Pa}'_{XY} = \mathbf{Pa}'_{Y\backslash X} = \mathbf{Pa}'_{X\backslash Y}$ and $q_{xy} = q_{y\backslash x} = q_{x\backslash y}$ in Table 1.

E Toy models

Fig. 3 shows three simple ancestral models used to test MIIC_search&score orientation scores (Table 1) to effectively predict bidirected orientations when the end nodes do not share the same parents (Model 1), share some parents (Model 2) or when the bidirected edge is part of a longer than two-collider paths (Model 3).

The data is generated from the theoretical DAG using the rmvDAG function in the pcalg package [48]. Each node follows a normal distribution, and the data is discretized using bnlearn's discretize function using Hartemink's pairwise mutual information method [40]. For these toy models, the edge orientation scores are computed assuming the correct parents of each node.

The prediction of the edge orientation scores are summarized in Table 3 in % of replicates displaying directed edges (wrong) or bidirected edge (correct) as a function of increasing dataset size N.

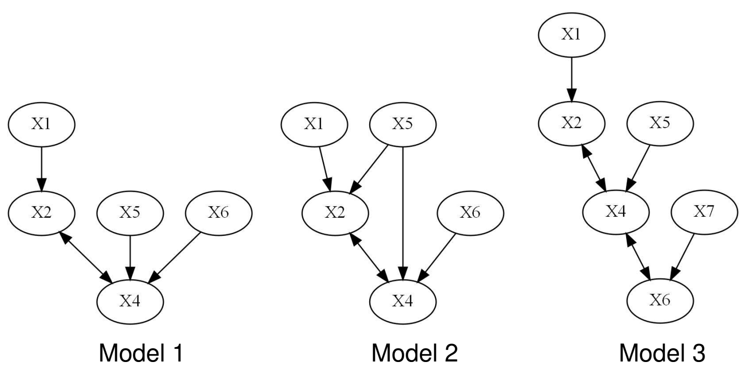

Figure 3: **Simple ancestral graphs.**

Table 3:	Model 1, $X_2 - X_4$			Model 2, $X_2 - X_4$			Model 3, $X_2 - X_4$			Model 3, $X_4 - X_6$		
N	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$	$\leftarrow$	$\rightarrow$	$\leftrightarrow$
1000	0	100	0	50	42	8	8	88	4	91.7	6.2	2.1
5000	0	68	32	18	2	80	2	80	18	76	24	0
10000	0	10	90	0	0	100	0	6	94	62	22	16
20000	0	0	100	0	0	100	0	0	100	2	0	98
35000	0	0	100	0	0	100	0	0	100	0	0	100
50000	0	0	100	0	0	100	0	0	100	0	0	100

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims of the paper are supported by the theoretical and experimental results shown in Figs. 1 & 2, respectively.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have added a Discussion & Limitation section at the end of the paper. The main limitation of the experimental results is the fact that we did not have sufficient time to perform many dataset replicates of the benchmark ancestral graphs. While the obtained statistics already support our main experimental results, we intend to perform more dataset replicates for the final version of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For the theoretical results (notably Theorem 1) we provide the full set of assumptions (section 2 and Appendix A)and a complete proof (Appendix B).

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided the full description of the experiments run in the paper (sections 2 & 3 and Appendix D). The open-source code reproducing the experimental results presented in the paper will be provided with the camera-ready version of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not include a new code with the initial submission, as it is not yet properly packaged at submission time, but we definitely intend to release this open-source code including proper annotation and userguide with the final camera-ready version of the paper. MIIC and FCI open-source packages used for benchmark comparison are already published and available on public servers.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provided the full description of the experiments run in the paper (sections 2 3 and Appendix D).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The 1-sigma error bars are plotted in Fig. 2. While these statistics already support our experimental results, we intend to perform more dataset replicates for the final version of the paper, which we did not have sufficient time to perform by the time of submission. This should reduce some error bars, in particular, those for the results displaying large error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resource used for all experiments is a simple laptop with intel i7 processors, 12 cores and 16 threads.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not use or produce sensitive data nor concern potentially harmful applications.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper does not use or produce sensitive data nor concern potentially harmful applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not use or produce sensitive data nor concern potentially harmful applications.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have credited all previously published resources (including license details) used in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not include a new code with the initial submission, as it is not yet properly packaged at submission time, but we definitely intend to release this open-source code including proper annotation and userguide with the final camera-ready version of the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.