entropy

MDPI

Article

Causal Structure Learning with Conditional and Unique Information Groups-Decomposition Inequalities

Daniel Chicharro [1,*] and Julia K. Nguyen [2]

[1] Artificial Intelligence Research Centre, Department of Computer Science, City, University of London, London EC1V 0HB, UK
[2] Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
* Correspondence: chicharro31@yahoo.es

Abstract: The causal structure of a system imposes constraints on the joint probability distribution of variables that can be generated by the system. Archetypal constraints consist of conditional independencies between variables. However, particularly in the presence of hidden variables, many causal structures are compatible with the same set of independencies inferred from the marginal distributions of observed variables. Additional constraints allow further testing for the compatibility of data with specific causal structures. An existing family of causally informative inequalities compares the information about a set of target variables contained in a collection of variables, with a sum of the information contained in different groups defined as subsets of that collection. While procedures to identify the form of these groups-decomposition inequalities have been previously derived, we substantially enlarge the applicability of the framework. We derive groups-decomposition inequalities subject to weaker independence conditions, with weaker requirements in the configuration of the groups, and additionally allowing for conditioning sets. Furthermore, we show how constraints with higher inferential power may be derived with collections that include hidden variables, and then converted into testable constraints using data processing inequalities. For this purpose, we apply the standard data processing inequality of conditional mutual information and derive an analogous property for a measure of conditional unique information recently introduced to separate redundant, synergistic, and unique contributions to the information that a set of variables has about a target.

Keywords: causality; directed acyclic graphs; causal discovery; structure learning; causal structures; marginal scenarios; hidden variables; mutual information; unique information; entropic inequalities; data processing inequality

MSC: 62H22; 62D20; 94A15; 94A17

Citation: Chicharro, D.; Nguyen, J.K. Causal Structure Learning with Conditional and Unique Information Groups-Decomposition Inequalities. *Entropy* **2024**, *26*, 440. https://doi.org/10.3390/e26060440

Academic Editor: Richard D. Gill

Received: 13 March 2024
Revised: 12 May 2024
Accepted: 17 May 2024
Published: 23 May 2024

1. Introduction

The inference of the underlying causal structure of a system using observational data is a fundamental question in many scientific domains. The causal structure of a system imposes constraints on the joint probability distribution of variables generated from it [1–4], and these constraints can be exploited to learn the causal structure. Causal learning algorithms based on conditional independencies [1,2,5] allow the construction of a partially oriented graph [6] that represents the equivalence class of all causal structures compatible with the set of conditional independencies present in the distribution of the observable variables (the so-called Markov equivalence class). However, without restrictions on the potential existence and structure of an unknown number of hidden variables that could account for the observed dependencies, Markov equivalence classes may encompass many causal structures compatible with the data.

Conditional independencies impose equality constraints on a joint probability distribution; namely, an independence results in the equality between conditional and unconditional probability distributions, or equivalently, in a null mutual information between

independent variables. In addition to the information from independencies between the observed variables, causal information can also be obtained from other functional equality constraints [7], such as dormant independencies that would occur under active interventions [8]. Further causal inference power can be obtained incorporating assumptions on the potential form of the causal mechanisms in order to exploit additional independencies associated with hidden substructures within the generative model [9,10], or independencies related to exogenous noise terms [11–13]. Other approaches have studied the identifiability of specific parametric families of causal models [3,14]. However, these methods only provide additional inference power if the actual causal mechanisms conform to the required parametric form.

Beyond equality constraints, the causal structure may also impose inequality constraints on the distribution of the data [15,16], which reflect non-verifiable independencies involving hidden variables. Figure 1 illustrates this distinction between pairs of causal structures distinguishable based on independence constraints (Figure 1A,B) and causal structures that may be discriminated based on inequality constraints (Figure 1C,D). The structures of Figure 1A,B belong to different Markov equivalence classes because in Figure 1A variables V_1 and V_2 are independent conditioned on S, while in Figure 1B, to obtain an independence it is required to further the condition on V_3. On the other hand, the structures of Figure 1C,D belong to the same equivalence class because no independencies exist between the observable variables $V_i, i = 1, 2, 3$. Nonetheless, if the hidden variables were also observable, these structures would be distinguishable. In Figure 1D, all the dependencies between the observable variables are caused by a single hidden variable U, while in Figure 1C dependencies are created pairwise by different hidden variables. In this case, a testable inequality constraint involving the observable variables reflects the non-verifiable independencies that involve also hidden variables. Intuitively, in Figure 1C, the inequality constraint imposes an upper bound on the overall degree of dependence between the three variables, given that these dependencies arise only in a pairwise manner, while in Figure 1D no such bound exists.

Importantly, unlike equality constraints, inequality constraints provide necessary but not sufficient conditions for the compatibility of data with a certain causal structure. While a certain hypothesized causal structure—like in Figure 1C—may impose the fulfillment of a given inequality intrinsically from its structure, other causal structures—like in Figure 1D—can generate data that, given a particular instantiation of the causal mechanisms, also fulfill the inequality. Accordingly, the causal inference power of inequality constraints lies in the ability to reject hypothesized causal structures that would intrinsically require the fulfillment of an inequality when that inequality is not fulfilled by the data. This means that tighter inequalities have more inferential power, giving the capacity to discard more causal structures.

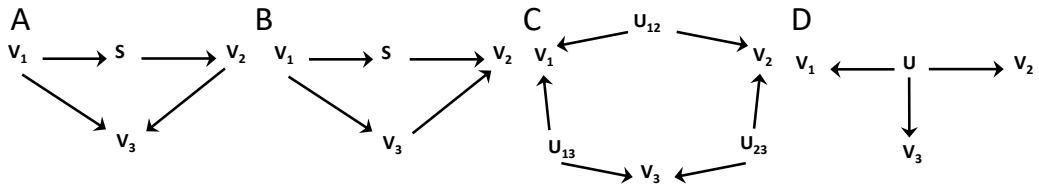

Figure 1. Examples of causal structures distinguishable from independencies (**A**,**B**) and structures that may only be discriminated based on inequality constraints (**C**,**D**). In this case, the structure in (**C**), and not the one in (**D**), intrinsically imposes a constraint due to dependencies between the observable variables $V_i, i = 1, 2, 3$ arising only from pairwise dependencies with hidden common causes.

Two main classes of inequality constraints have been derived. The first class corresponds to inequality constraints in the probability space, which comprise tests of compatibility such as Bell-type inequalities [17,18], instrumental inequalities [19,20], and inequalities that appear on identifiable interventional distributions [21]. The second class corresponds to inequalities involving information-theoretic quantities. The relation between

these probabilistic and entropic inequalities has been examined in [22]. One approach to construct entropic inequalities combines the inequalities defining the Shannon entropic cone, i.e., associated with the non-negativity, monotonicity, and submodularity properties of entropy, and additional independence constraints related to the causal structure [23,24]. Additional causally informative inequalities can be derived if considering the so-called Non-Shannon inequalities [25,26]. When the causal structure to be tested involves hidden variables, all non-trivial entropic inequalities in the marginal scenario associated with the set of observable variables can be derived with an algorithmic procedure [23,24] that projects the set of inequalities of all variables into inequalities that only involve the subset of observable variables.

As an alternative approach, information-theoretic inequality constraints can be derived by an explicit analytical formulation [24,27]. In particular, [27] introduced inequalities comparing the information about a target variable contained in a whole collection of variables with a weighted sum of the information contained in groups of variables corresponding to subsets of the collection. Two procedures were introduced to select the composition of these groups. In a first type of inequalities, the composition of the groups is arbitrarily determined, but an inequality only exists under some conditions of independence between the chosen variables, whose fulfillment reflects the underlying causal structure. In a second type, no conditions are required for the existence of an inequality, but the groups must be ancestral sets; that is, must contain all other variables that have a causal effect on any given element of the group. In both cases, [27] showed that the coefficients in the weighted sum of the information contained in groups of variables are determined by the number of intersections between the groups.

In this work, we build upon the results of [27] and generalize their framework of groups-decomposition inequalities in several ways. First, we generalize both types of inequalities to the conditional case, when the inequalities involve conditional mutual information measures instead of unconditional ones. While this extension is trivial for the first type of inequalities, we show that for the second type it requires a definition of augmented ancestral sets. Second, we formulate more flexible conditions of independence for which the first type of inequalities exists. Third, we add flexibility to the construction of the ancestral sets that appear in the second type of inequalities. We show that, given a causal graph and a conditioning set of variables used for the conditional mutual information measures, alternative inequalities exist when determining ancestors in subgraphs that eliminate causal connections from different subsets of the conditioning variables. Furthermore, we determine conditions in which an inequality also holds when removing subsets of ancestors from the whole set of variables, hence relaxing for the second type of inequalities the requirement that the groups correspond to ancestral sets.

Apart from these generalizations, we expand the power of the approach of [27] by considering inequalities whose existence is determined by the partition into groups of a collection of variables that also contains hidden variables. That is, hidden variables can appear not only as hidden common ancestors of the collection but also as part (or even all) of the variables in the collection for which the inequality is defined. To render operational the use of inequalities derived from collections containing hidden variables, we develop procedures that allow mapping those inequalities into testable inequalities that only involve observable variables. While this mapping can be carried out by simply applying the monotonicity of mutual information to remove hidden variables from the groups, this does not work when all variables in the collection are hidden. We show that data processing inequalities [28] can be applied to obtain testable inequalities also in this case, or applied to obtain tighter inequalities than those obtained by simply removing the hidden variables. We illustrate how testable inequalities whose coefficients in the weighted sum depend on intersections among subsets of hidden variables instead of among subsets of observable variables can result into tighter inequalities with higher inferential power.

In order to derive testable groups-decomposition inequalities, we do not only apply the standard data processing inequality of conditional mutual information [28], but we

derive an additional data processing inequality for the so-called *unique information* measure introduced in [29]. This measure was introduced in the framework of a decomposition of mutual information into redundant, unique, and synergistic information components [30]. Recently, alternative decompositions have been proposed to decompose the joint mutual information that a set of predictor variables has about a target variable into redundant, synergistic, and unique components [31–35] (among others). These alternative decompositions generally differ in the quantification of each component and differ in whether the measures fulfill certain properties or axioms. However, in our work, we do not apply the unique information measure of [29] as part of a decomposition of the joint mutual information. Instead, we show that it provides an alternative data processing inequality that holds for different causal configurations than the standard data processing inequality of conditional mutual information. In this way, the unique information data processing inequality increases the capability to eliminate hidden variables in order to obtain testable groups-decomposition inequalities. Accordingly, the groups-decomposition inequalities we derive can contain unique information terms apart from the standard mutual information and entropy measures that appear when considering the constraints of the Shannon entropic cone [23,24].

We envisage the application of the causally informative tests here proposed in the following way. Given a data set, a hypothesized causal structure is selected to test its compatibility with the data. First, the set of inequality constraints enforced by that causal structure is determined. Second, their fulfillment is evaluated from the data and the causal structure is discarded if some inequality does not hold. In the first step, the determination of the set of groups-decomposition inequalities enforced by a causal structure requires at different levels the verification of conditional independencies. This is the case, for example, with the conditional independencies that are necessary conditions for the existence of the first type of inequalities introduced by [27]. If all variables involved were observable, this verification could be conducted directly from the data. However, as mentioned above, we here consider groups-decomposition inequalities that may contain hidden variables as part of the collection of variables, which precludes this direct verification. For this reason, we will work under the assumption that statistical independencies can be assessed from the structure of the causal graph, namely with the graphical criterion of separability between nodes in the graph known as *d-separation* [36]. That is, we will rely on the assumption that graphical separability is a sufficient condition for statistical independence and hence characterize the set of groups-decomposition inequalities enforced by a causal structure without using the data. Data would only be used in the second step, in which the actual fulfillment of the inequalities is evaluated.

This paper is organized as follows. In Section 2, we review previous work relevant for our contributions. In Section 3.1, we formulate the data processing inequality for the unique information measure. In Section 3.2, we generalize the first type of inequalities of [27], formulating for the conditional case more general conditions of independence for which a groups-decomposition inequality exists. We also apply data processing inequalities to derive testable groups-decomposition inequalities when collections include hidden variables. In Section 3.3, we generalize the second type of inequalities of [27] as outlined above. In Section 4, we discuss the connection of this work with other approaches to causal structure learning and point to future continuations and potential applications. The Appendix contains proofs of the results (Appendices A and B) and a discussion of the relations between conditional independencies and d-separations required so that the inequalities here derived are applicable to test causal structures (Appendix C).

2. Previous Work on Information-Theoretic Measures and Causal Graphs Relevant for Our Derivations

In this section we review properties of information-theoretic measures and concepts of causal graphs relevant for our work. In Section 2.1, we review basic inequalities of the mutual information and in Section 2.2 the definition and relevant properties of the

unique information measure of [29]. We then review in Section 2.3 Directed Acyclic Graphs (DAGs) and their relation to conditional independence through the graphical criterion of *d-separation* [36,37]. Finally, we review the inequalities introduced by [27] to test causal structures from information decompositions involving sums of groups of variables (Section 2.4). We do not aim to more broadly review other types of information-theoretic inequalities [23,24] also used for causal inference. The relation with these other types will be considered in the Discussion.

2.1. Mutual Information Inequalities Associated with Independencies

We present in Lemma 1 two well-known inequalities that will be used in our derivations. This lemma corresponds to Lemma 1 in [27]. For completion, we provide the proof of the lemma.

Lemma 1. *The mutual information fulfills the following inequalities in the presence of the corresponding independencies:*

(i) (Conditional mutual information data processing inequality): Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{B}'$, *and* $\mathbf{D}$ *be four sets of variables. If* $I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{D}) = 0$, *then it follows that* $I(\mathbf{A};\mathbf{B}|\mathbf{D}) \geqslant I(\mathbf{A};\mathbf{B}'|\mathbf{D})$.

(ii) (Increase through conditioning on independent sets): Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, *and* $\mathbf{Y}$ *be four sets of variables. If* $I(\mathbf{A};\mathbf{C}|\mathbf{B}) = 0$, *then* $I(\mathbf{Y};\mathbf{A}|\mathbf{B}) \leqslant I(\mathbf{Y};\mathbf{A}|\mathbf{B},\mathbf{C})$.

Proof. (i) is proven applying, in two different orders, the chain rule of the mutual information to $I(\mathbf{A};\mathbf{B},\mathbf{B}'|\mathbf{D})$:

$$I(\mathbf{A};\mathbf{B},\mathbf{B}'|\mathbf{D}) = I(\mathbf{A};\mathbf{B}|\mathbf{D}) + I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{D}) = I(\mathbf{A};\mathbf{B}'|\mathbf{D}) + I(\mathbf{A};\mathbf{B}|\mathbf{B}',\mathbf{D}).$$

Since $I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{D}) = 0$ and the mutual information is non-negative, this implies the inequality. To prove (ii), the chain rule is applied in different orders to $I(\mathbf{Y},\mathbf{C};\mathbf{A}|\mathbf{B})$:

$$I(\mathbf{Y},\mathbf{C};\mathbf{A}|\mathbf{B}) = I(\mathbf{C};\mathbf{A}|\mathbf{B}) + I(\mathbf{Y};\mathbf{A}|\mathbf{B},\mathbf{C}) = I(\mathbf{Y};\mathbf{A}|\mathbf{B}) + I(\mathbf{C};\mathbf{A}|\mathbf{B},\mathbf{Y}).$$

Since $I(\mathbf{C};\mathbf{A}|\mathbf{B}) = 0$ and the mutual information is non-negative, this implies the inequality. □

2.2. Definition and Properties of the Unique Information

The concept of *unique information* as part of a decomposition of the joint mutual information $I(\mathbf{Y};\mathbf{X})$ that a set of predictor variables $\mathbf{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ has about a target (possibly multivariate) variable $\mathbf{Y}$ was introduced in [30]. In the simplest case of two predictors $\{\mathbf{X}_1,\mathbf{X}_2\}$, this framework decomposes the joint mutual information about $\mathbf{Y}$ into four terms, namely the redundancy of $\mathbf{X}_1$ and $\mathbf{X}_2$, the unique information of $\mathbf{X}_1$ with respect to $\mathbf{X}_2$, the unique information of $\mathbf{X}_2$ with respect to $\mathbf{X}_1$, and the synergy between $\mathbf{X}_1$ and $\mathbf{X}_2$. The predictors share the redundant component, the synergistic one is only obtained by combining the predictors, and unique components are exclusive to each predictor. Several information measures have been proposed to define this decomposition, aiming to comply with a set of desirable properties which were not all fulfilled by the original proposal [29,31–33]. However, in this work we will not study the whole decomposition but specifically apply the bivariate measure of unique information introduced in [29]. In Section 3.1, we derive a data processing inequality for this measure and in Section 3.2 we show how it can help to obtain testable groups-decomposition inequalities for causal structures for which the standard data processing inequality of the mutual information would not allow elimination of the hidden variables. In this Section, we review the definition of the unique information measure of [29], we provide a straightforward generalization to a conditional unique information measure, and state a monotonicity property that will be used to derive the data processing inequality of the unique information. The unique information of $\mathbf{X}_1$ with respect to $\mathbf{X}_2$ about $\mathbf{Y}$ was defined as

$$I(\mathbf{Y};\mathbf{X}_1 \backslash\backslash \mathbf{X}_2) \equiv \min_{Q\in\Delta_P} I_Q(\mathbf{Y};\mathbf{X}_1|\mathbf{X}_2), \tag{1}$$

where Δ_P is defined as the set of distributions on $\{\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2\}$ that preserve the marginals $P(\mathbf{Y}, \mathbf{X}_1)$ and $P(\mathbf{Y}, \mathbf{X}_2)$ of the original distribution $P(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2)$. The notation I_Q is used to indicate that the mutual information is calculated on the probability distribution Q. We use $I(\mathbf{Y}; \mathbf{X}_1 \backslash\backslash \mathbf{X}_2)$ to refer to the unique information of $\mathbf{X}_1$ with respect to $\mathbf{X}_2$, compared to $I(\mathbf{Y}; \mathbf{X}_1 | \mathbf{X}_2)$, which is the standard conditional information of $\mathbf{X}_1$ given $\mathbf{X}_2$. We use the notation $\mathbf{X}_1 \backslash\backslash \mathbf{X}_2$ instead of the notation $\mathbf{X}_1 \backslash \mathbf{X}_2$ introduced by [29] to differentiate it from the set notation $\mathbf{X}_1 \backslash \mathbf{X}_2$, which indicates the subset of variables in $\mathbf{X}_1$ that is not contained in $\mathbf{X}_2$, since we will also be using this set notation. This unique information measure is a maximum entropy measure, since all distributions within Δ_P preserve the conditional entropy $H(\mathbf{Y}|\mathbf{X}_2)$, and hence the minimization is equivalent to a maximization of the conditional entropy $H(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2)$. The rationale that supports this definition is that the unique information of $\mathbf{X}_1$ with respect to $\mathbf{X}_2$ about $\mathbf{Y}$ has to be determined by the marginal probabilities $P(\mathbf{Y}, \mathbf{X}_1)$ and $P(\mathbf{Y}, \mathbf{X}_2)$, and cannot depend on any additional structure in the joint distribution that contributes to the dependence between $\{\mathbf{X}_1, \mathbf{X}_2\}$ and $\mathbf{Y}$ [29]. This additional contribution is removed by minimizing within Δ_P.

In a straightforward generalization, we define the conditional unique information given another set of variables $\mathbf{Z}$ as

$$I(\mathbf{Y}; \mathbf{X}_1 \backslash\backslash \mathbf{X}_2 | \mathbf{Z}) \equiv \min_{Q \in \Delta_{P'}} I_Q(\mathbf{Y}; \mathbf{X}_1 | \mathbf{X}_2, \mathbf{Z}), \tag{2}$$

where $\Delta_{P'}$ is the set of distributions on $\{\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Z}\}$ that preserve the marginals $P(\mathbf{Y}, \mathbf{X}_1, \mathbf{Z})$ and $P(\mathbf{Y}, \mathbf{X}_2, \mathbf{Z})$ of the original $P(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Z})$. By construction [29], the conditional unique information is bounded as

$$\min\{I(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}), I(\mathbf{Y}; \mathbf{X}_1 | \mathbf{X}_2, \mathbf{Z})\} \geqslant I(\mathbf{Y}; \mathbf{X}_1 \backslash\backslash \mathbf{X}_2 | \mathbf{Z}) \geqslant 0. \tag{3}$$

This is consistent with the intuition of the decomposition that the unique information is a component exclusive of X_1. In Lemma 2, we present a type of monotonicity fulfilled by the conditional unique information. This result is a straightforward extension to the conditional case of the one stated in Lemma 3 of [38]. We include the full proof because it will be useful in the Results section to prove a related data processing inequality for the unique information. To better suit our subsequent use of notation, we consider the two predictors to be $\mathbf{Z}_1$ and $\{\mathbf{X}_1, \mathbf{X}_1'\}$, and the conditioning set to be $\mathbf{Z}_2$.

Lemma 2. *The maximum entropy conditional unique information is monotonic on its second argument, corresponding to the non-conditioning predictor, as follows:*

$$I(\mathbf{Y}; \mathbf{X}_1, \mathbf{X}_1' \backslash\backslash \mathbf{Z_1} | \mathbf{Z_2}) \geqslant I(\mathbf{Y}; \mathbf{X}_1 \backslash\backslash \mathbf{Z_1} | \mathbf{Z_2}).$$

Proof. Consider the distribution $P_{1,1'} \equiv P(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_1', \mathbf{Z}_1, \mathbf{Z}_2)$ and its marginal $P_1 \equiv P(\mathbf{Y}, \mathbf{X}_1, \mathbf{Z}_1, \mathbf{Z}_2)$. Consider any distribution $Q_{1,1'} \in \Delta_{P_{1,1'}}$ and its marginal Q_1 on $(\mathbf{Y}, \mathbf{X}_1, \mathbf{Z}_1, \mathbf{Z}_2)$. Then $Q_1 \in \Delta_{P_1}$. By monotonicity of the mutual information, $I_{Q_{1,1'}}(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}_1, \mathbf{Z}_2)$ is lower than or equal to $I_{Q_{1,1'}}(\mathbf{Y}; \mathbf{X}_1, \mathbf{X}_1' | \mathbf{Z}_1, \mathbf{Z}_2)$. Since $I_{Q_{1,1'}}(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}_1, \mathbf{Z}_2)$ does not have $\mathbf{X}_1'$ as an argument, it is equal to the information calculated on its marginal $I_{Q_1}(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}_1, \mathbf{Z}_2)$. Since this holds for any distribution in $\Delta_{P_{1,1'}}$, it holds in particular for the distribution $Q^*_{1,1'}$ that minimizes $I(\mathbf{Y}; \mathbf{X}_1, \mathbf{X}_1' | \mathbf{Z}_1, \mathbf{Z}_2)$ in $\Delta_{P_{1,1'}}$. Since Q_1^* belongs to Δ_{P_1}, the minimum of $I(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}_1, \mathbf{Z}_2)$ in Δ_{P_1} is equal to or smaller than $I_{Q_1^*}(\mathbf{Y}; \mathbf{X}_1 | \mathbf{Z}_1, \mathbf{Z}_2)$ and hence equal to or smaller than $I_{Q^*_{1,1'}}(\mathbf{Y}; \mathbf{X}_1, \mathbf{X}_1' | \mathbf{Z}_1, \mathbf{Z}_2)$. □

2.3. Causal Graphs and Conditional Independencies

We here review basic notions of Directed Acyclic Graphs (DAGs) and the relation between causal structures and dependencies. Consider a set of random variables $\mathbf{V} = \{\mathrm{V}_1, \ldots, \mathrm{V}_n\}$. A DAG $G = (\mathbf{V}; \mathcal{E})$ consists of nodes $\mathbf{V}$ and edges $\mathcal{E}$ between the nodes. The graph contains $V_i \to \mathrm{V}_j$ for each $(\mathrm{V}_i; \mathrm{V}_j) \in \mathcal{E}$. We refer to V as both a variable and its corresponding node.

Causal influences can be represented in acyclic graphs given that causal mechanisms are not instantaneous and causal loops can be spanned using separate time-indexed variables. A path in G is a sequence of (at least two) distinct nodes $V_1, \ldots, V_m$, such that there is an edge between V_k and V_{k+1} for all $k = 1, \ldots, m-1$. If all edges are directed as $V_k \rightarrow V_{k+1}$ the path is a causal or directed path. A node V_i is a collider in a path if it has incoming arrows $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ and is a noncollider otherwise. A node V_i is called a parent of V_j if there is an arrow $V_i \rightarrow V_j$. The set of parents is denoted $\mathbf{Pa}_{V_j}$. A node V_i is called an ancestor of V_j if there is a directed path from V_i to V_j. Conversely, in this case V_j is a descendant of V_i. For convenience, we define the set of ancestors $an_G(V_i)$ as including V_i itself, and the set of descendants $D_G(V_i)$ as also containing V_i itself.

The link between generative mechanisms and causal graphs relies on the fact that in the graph a variable V_i is a parent of another variable V_j if and only if it is an argument of an underlying functional equation that captures the mechanisms that generate V_j; that is, an argument of $V_j := f_{V_j}(\mathbf{Pa}_{V_j}, \varepsilon_{V_j})$, where ε_{V_j} captures additional sources of stochasticity exogenous to the system. If a DAG constitutes an accurate representation of the causal mechanisms, an isomorphic relation exists between the conditional independencies that hold between variables in the system and a graphical criterion of separability between the nodes, called *d-separation* [36]. Two nodes X and Y are *d-separated* given a set of nodes $\mathbf{S}$ if and only if no $\mathbf{S}$-active paths exist between X and Y. A path is active given the conditioning set $\mathbf{S}$ ($\mathbf{S}$-active) if no noncollider in the path belongs to $\mathbf{S}$ and every collider in the path either is in $\mathbf{S}$ or has a descendant in $\mathbf{S}$. A causal structure G and a generated probability distribution $p(\mathbf{V})$ are *faithful* [1,2] to one another when a conditional independence between X and Y given $\mathbf{S}$—denoted by $X \perp_P Y|\mathbf{S}$—holds if and only if there is no $\mathbf{S}$-active path between them; that is, if X and Y are d-separated given $\mathbf{S}$—denoted by $X \perp_G Y|\mathbf{S}$. Accordingly, faithfulness is assumed in the algorithms of causal inference [1,2] that examine conditional independencies to characterize the Markov equivalence class of causal structures that share a common set of independencies. A well-known example of a system that is unfaithful to its causal structure is the exclusive-OR (X-OR) logic gate, whose output is independent of the two inputs separately but dependent on them jointly.

In contrast to the algorithms that infer Markov equivalence classes, we will show that the applicability of the groups-decomposition inequalities here studied relies on the assumption that d-separability is a sufficient condition for conditional independence. That is, instead of an *if and only if* relation between d-separability and conditional independence, as required in the faithfulness assumption, it is enough to assume that d-separability implies conditional independence. As we further discuss in Appendix C, this is a substantially weaker assumption, since usually faithfulness is violated due to the presence of independencies that are incompatible with the causal structure. This is the case, for example, of the X-OR logic gate, for which faithfulness is violated because the inputs are separately independent of the output despite each having an arrow towards the output in the corresponding causal graph. Conversely, the X-OR gate complies with d-separability being a sufficient condition for independence, since in the graph only the input nodes are d-separated and the corresponding input variables of the X-OR gate are independent. Despite only requiring that d-separability implies independence, to simplify the presentation of our results in the main text we will assume faithfulness and indistinctively use $X \perp Y|\mathbf{S}$ to indicate statistical independence and graphical separability, instead of distinguishing between $X \perp_P Y|\mathbf{S}$ and $X \perp_G Y|\mathbf{S}$. In Appendix C, we will more closely examine how in the proofs of our results the sufficient condition of d-separability for conditional independencies is enough. An important implication of independencies following from d-separability is that, if variables $\{X_1, X_2\}$ are separately independent from Y—namely $Y \perp X_1$ and $Y \perp X_2$—because of the lack of any connection between node Y and both nodes X_1 and X_2, then $\{X_1, X_2\}$ cannot be jointly dependent on Y, namely $Y \not\perp \{X_1, X_2\}$ cannot occur. This is because d-separability between node Y and the set of nodes $\{X_1, X_2\}$ is determined by separately considering the lack of active paths between Y and each node X_1 and X_2. Since the set of paths between Y and $\{X_1, X_2\}$ is the union of the paths between

Y and both X_1 and X_2, considering $\{X_1, X_2\}$ jointly does not add new paths that could create a dependence of Y with $\{X_1, X_2\}$. A dependence can only be created by conditioning on some other variable, which could activate additional paths by activating a collider.

2.4. Inequalities for Sums of Information Terms from Groups of Variables

We now review two results in [27] that are at the foundation of our results. The first corresponds to their Proposition 1. We provide a slightly more general formulation that is useful for subsequent extensions.

Proposition 1. *(Decomposition of information from groups with conditionally independent non-shared components): Consider a collection of groups $\mathbf{A}_{[n]} \equiv \{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$, where each group $\mathbf{A}_i$ consists of a subset of observable variables $\mathbf{A}_i \subset \mathcal{O}$, being $\mathcal{O}$ the set of all observable variables. For every $\mathbf{A}_i \in \mathbf{A}_{[n]}$, define d_i as the maximal value such that $\mathbf{A}_i$ has a non-empty intersection where it intersects jointly with $d_i - 1$ other distinct groups out of $\mathbf{A}_{[n]}$. Consider a conditioning set $\mathbf{Z}$ and target variables $\mathbf{Y}$. If each group is conditionally independent given $\mathbf{Z}$ from the non-shared variables in each other group (i.e., $\mathbf{A}_i \perp \mathbf{A}_j \backslash \mathbf{A}_i | \mathbf{Z}, \forall i, j$), then the conditional information that $\mathbf{A}_{[n]}$ has about the target variables $\mathbf{Y}$ given $\mathbf{Z}$ is bounded from below by*

$$I(\mathbf{Y}; \mathbf{A}_{[n]} | \mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{A}_i | \mathbf{Z}).$$

Proof. The proof is presented in Appendix A. It is a generalization to the conditional case of the proof of Proposition 1 in [27] and a slight generalization that allows for dependencies to exist between variables shared by two groups as long as dependencies with non-shared variables do not exist. □

An illustration of Proposition 1 for the unconditional case is presented in Figure 3 of [27], together with further discussion. In Section 3.2 we will provide further illustrations for the extensions of Proposition 1 that we introduce. We will use $\mathbf{d} \equiv \{d_1, \ldots, d_n\}$ to indicate the maximal values for all groups. We will add a subindex $\mathbf{d}_{\mathbf{A}_{[n]}}$ to specify the collection if different collections are compared. A trivial refinement of Proposition 1 would consider $I(\mathbf{Y}; \mathbf{A}_{[n]} \backslash \mathbf{Z} | \mathbf{Z})$ and for each group $I(\mathbf{Y}; \mathbf{A}_i \backslash \mathbf{Z} | \mathbf{Z})$. This may lead to a tighter lower bound by decreasing some values in $\mathbf{d}$ if some intersections between groups occur in $\mathbf{Z}$. We do not present this refinement in order to simplify the presentation.

The second result from [27] that we will be relying on is their Theorem 1. We present a version that is slightly reduced and modified, which is more convenient in order to relate to our own results.

Theorem 1. *(Decomposition of information in ancestral groups.) Let G be a DAG model that includes nodes corresponding to the variables in a collection of groups $\mathbf{A}_{[n]} \equiv \{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$, which is a subset all observable variables $\mathcal{O}$. Let $an_G(\mathbf{A}_{[n]}) \equiv \{an_G(\mathbf{A}_1), \ldots, an_G(\mathbf{A}_n)\}$ be the collection of ancestors of the groups, as determined by G. For every ancestral set of a group, $an_G(\mathbf{A}_i)$, let $d_i(G)$ be maximal, such that there is a non-empty joint intersection of $an_G(\mathbf{A}_i)$ and other $d_i(G) - 1$ distinct ancestral sets out of $an_G(\mathbf{A}_{[n]})$. Let $\mathbf{Y}$ be a set of target variables. Then the information of $an_G(\mathbf{A}_{[n]})$ about $\mathbf{Y}$ is bounded as*

$$H(\mathbf{Y}) \geqslant I(\mathbf{Y}; an_G(\mathbf{A}_{[n]})) \geqslant \sum_{i=1}^{n} \frac{1}{d_i(G)} I(\mathbf{Y}; an_G(\mathbf{A}_i)).$$

Proof. The original proof can be found in [27]. □

In contrast to Proposition 1, a generalization to the conditional mutual information is not trivial and will be developed in Section 3.3. We will also propose additional generalizations regarding which graph to use to construct the ancestral sets and conditions

to exclude some ancestors from the groups. In their work, [27] conceptualized $\mathbf{Y}$ as corresponding to leaf nodes in the graph, for example providing some noisy measurement of $\mathbf{A}_{[n]}$, with $\mathbf{Y} = \mathbf{A}_{[n]}$ being the case of perfect measurement. While this conceptualization guided their presentation, their results were general, and here we will not assume any concrete causal relation between $\mathbf{Y}$ and $\mathbf{A}_{[n]}$. We have slightly modified the presentation of Theorem 1 from [27] to add the upper bound and to remove some additional subcases with extra assumptions presented in their work. The upper bound is the standard upper bound of mutual information by entropy [28]. In the Results, we will also be interested in cases in which $an_G(\mathbf{A}_{[n]})$ contains hidden variables, so that $I(\mathbf{Y}; an_G(\mathbf{A}_{[n]}))$ cannot be estimated. Given the monotonicity of mutual information, the terms from each ancestral group can be lower bounded by the information in the observable variables within each group and $H(\mathbf{Y})$ is used as a testable upper bound.

There are two main differences between Proposition 1 and Theorem 1. First, Theorem 1 does not impose conditions of independence for the inequality to hold. Second, while the value d_i of each group $\mathbf{A}_i$ is determined in Proposition 1 by the overlap between groups, with no influence of the causal structure relating the variables, on the other hand in Theorem 1 the value $d_i(G)$ depends on the causal structure, since it is determined from the intersections between ancestral sets. Despite these differences, given the relation between causal structure and independencies reviewed in Section 2.3, both types of inequalities can have causal inference power to test the compatibility of certain causal structures with data.

3. Results

In Section 3.1, we introduce a data processing inequality for the conditional unique information measure of [29]. In Section 3.2, we develop new information inequalities involving groups of variables and examine how data processing inequalities can help to derive testable inequalities in the presence of hidden variables. In Section 3.3, we develop new information inequalities involving ancestral sets. The application of these inequalities for causal structure learning is discussed. As justified in the proofs of our results (Appendices A and B) and further discussed in Appendix C, our derivations of groups-decomposition inequalities only rely on the assumption that d-separability implies conditional independence. No further assumptions are used in our work, in particular, our application of the unique information measures of [29] does not require any assumption regarding the precise distribution of the joint mutual information among redundancy, unique, and synergistic components.

3.1. Data Processing Inequality for Conditional Unique Information

Proposition 2. *(Conditional unique information data processing inequality): Let* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{B}'$, $\mathbf{D}$, *and* $\mathbf{E}$ *be five sets of variables. If* $I(\mathbf{A}; \mathbf{B}'|\mathbf{B}, \mathbf{E}) = 0$, *then* $I(\mathbf{A}; \mathbf{B}, \mathbf{B}' \backslash\backslash \mathbf{D}|\mathbf{E}) = I(\mathbf{A}; \mathbf{B} \backslash\backslash \mathbf{D}|\mathbf{E}) \geqslant I(\mathbf{A}; \mathbf{B}' \backslash\backslash \mathbf{D}|\mathbf{E})$.

Proof. Let $P_{BB'} \equiv P(\mathbf{A}, \mathbf{B}, \mathbf{B}', \mathbf{D}, \mathbf{E})$ be the original distribution of the variables and define $\Delta_{P_{BB'}}$ as the set of distributions on $\{\mathbf{A}, \mathbf{B}, \mathbf{B}', \mathbf{D}, \mathbf{E}\}$ that preserve the two marginals $P(\mathbf{A}, \mathbf{B}, \mathbf{B}', \mathbf{E})$ and $P(\mathbf{A}, \mathbf{D}, \mathbf{E})$. Let $P_B \equiv P(\mathbf{A}, \mathbf{B}, \mathbf{D}, \mathbf{E})$ be the marginal of $P_{BB'}$ and Δ_{P_B} be the set of distributions that preserve the marginals $P(\mathbf{A}, \mathbf{B}, \mathbf{E})$ and $P(\mathbf{A}, \mathbf{D}, \mathbf{E})$. By the definition of unique information (Equation (2))

$$\begin{aligned} I(\mathbf{A}; \mathbf{B}, \mathbf{B}' \backslash\backslash \mathbf{D}|\mathbf{E}) &\equiv \min_{Q_{BB'} \in \Delta_{P_{BB'}}} I_{Q_{BB'}}(\mathbf{A}; \mathbf{B}, \mathbf{B}'|\mathbf{D}, \mathbf{E}) \overset{(a)}{=} \\ &\min_{Q_{BB'} \in \Delta_{P_{BB'}}} \Big[I_{Q_{BB'}}(\mathbf{A}; \mathbf{B}|\mathbf{D}, \mathbf{E}) + I_{Q_{BB'}}(\mathbf{A}; \mathbf{B}'|\mathbf{B}, \mathbf{D}, \mathbf{E}) \Big] \overset{(b)}{=} \\ &\min_{Q_{BB'} \in \Delta_{P_{BB'}}} \Big[I_{Q_B}(\mathbf{A}; \mathbf{B}|\mathbf{D}, \mathbf{E}) + I_{Q_{BB'}}(\mathbf{A}; \mathbf{B}'|\mathbf{B}, \mathbf{D}, \mathbf{E}) \Big]. \end{aligned} \tag{4}$$

Equality (a) follows from the chain rule of mutual information. Equality (b) holds because $I_{Q_{BB'}}(\mathbf{A};\mathbf{B}|\mathbf{D},\mathbf{E})$ does not depend on $\mathbf{B}'$ and can be calculated with the marginal Q_B, marginalizing $Q_{BB'}$ on $\mathbf{B}'$. Note that $Q_B \in \Delta_{P_B}$. Since $I_{P_{BB'}}(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{E})$ is null, $P(\mathbf{A},\mathbf{B},\mathbf{B}',\mathbf{E})$ factorizes as $P(\mathbf{B}'|\mathbf{B},\mathbf{E})P(\mathbf{A},\mathbf{B},\mathbf{E})$. For any distribution $\tilde{Q}_B \in \Delta_{P_B}$, which preserves $P(\mathbf{A},\mathbf{D},\mathbf{E})$ and $P(\mathbf{A},\mathbf{B},\mathbf{E})$, a distribution can be constructed as $\tilde{Q}_{BB'} \equiv P(\mathbf{B}'|\mathbf{B},\mathbf{E})\tilde{Q}_B$, such that $\tilde{Q}_{BB'} \in \Delta_{P_{BB'}}$, since $\tilde{Q}_{BB'}$ continues to preserve $P(\mathbf{A},\mathbf{D},\mathbf{E})$ and $P(\mathbf{A},\mathbf{B},\mathbf{B}',\mathbf{E})$ is preserved by construction. Also by construction, $I_{\tilde{Q}_{BB'}}(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{D},\mathbf{E}) = 0$ for any $\tilde{Q}_{BB'}$ created from any $\tilde{Q}_B \in \Delta_{P_B}$. In particular, this holds for the distribution $\tilde{Q}^*_{BB'}$, constructed from $\tilde{Q}^*_B$ that minimizes $I_{\tilde{Q}_B}(\mathbf{A};\mathbf{B}|\mathbf{D},\mathbf{E})$, which determines $I(\mathbf{A};\mathbf{B}\backslash\backslash\mathbf{D}|\mathbf{E})$. The distribution $\tilde{Q}^*_{BB'}$ minimizes the first term in the r.h.s of Equation (4) and, given the non-negativity of mutual information, it also minimizes the second term, hence providing the minimum in $\Delta_{P_{BB'}}$. Accordingly, $I(\mathbf{A};\mathbf{B},\mathbf{B}'\backslash\backslash\mathbf{D}|\mathbf{E}) = I(\mathbf{A};\mathbf{B}\backslash\backslash\mathbf{D}|\mathbf{E})$. The monotonicity of unique information on the non-conditioning predictor (Lemma 2) leads to $I(\mathbf{A};\mathbf{B},\mathbf{B}'\backslash\backslash\mathbf{D}|\mathbf{E}) \geqslant I(\mathbf{A};\mathbf{B}'\backslash\backslash\mathbf{D}|\mathbf{E})$. □

A related data processing inequality has already been previously derived for the unconditional unique information in the case of $I(\mathbf{A},\mathbf{D};\mathbf{B}'|\mathbf{B}) = 0$, with $\mathbf{E} = \varnothing$ [39]. Differently, Proposition 2 formulates a data processing inequality for the case $I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{E}) = 0$. When $\mathbf{E} = \varnothing$, Proposition 2 states a weaker requirement for the existence of an inequality, given the *decomposition axiom* of the mutual information [27]. As we will now see in Section 3.2, Proposition 2 will allow us to apply the unique information data processing inequality in cases in which $I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{E}) = 0$. In particular, $I(\mathbf{A};\mathbf{B},\mathbf{B}'\backslash\backslash\mathbf{D}|\mathbf{E}) \geqslant I(\mathbf{A};\mathbf{B}'\backslash\backslash\mathbf{D}|\mathbf{E})$ allows us to obtain a lower bound when $\mathbf{B}$ contains hidden variables that we want to eliminate in order to have a testable groups-decomposition inequality. In contrast, the application of the standard data processing inequality of the mutual information $I(\mathbf{A};\mathbf{B},\mathbf{B}'|\mathbf{D},\mathbf{E}) \geqslant I(\mathbf{A};\mathbf{B}'|\mathbf{D},\mathbf{E})$ requires $I(\mathbf{A};\mathbf{B}'|\mathbf{B},\mathbf{D},\mathbf{E}) = 0$, and hence the two types of data processing inequalities may be applicable in different cases to eliminate $\mathbf{B}$. This will be fully appreciated in Propositions 5 and 6. Note that this application of the unique information measure of Equation (2) to eliminate hidden variables is not restrained by the role of the measure in the mutual information decomposition and by considerations about which alternative decompositions optimally quantify the different components [30,35].

3.2. Inequalities Involving Sums of Information Terms from Groups

In this section, we extend Proposition 1 in several ways. Propositions 3–6 present subsequent generalizations, all subsumed by Proposition 6. We present these generalizations progressively to better appreciate the new elements. For these Propositions, examples are displayed in Figures 2 and 3 and explained in text after the enunciation of each Proposition. Which Proposition is illustrated by each example is indicated in the figure caption and in the main text. The objective of these generalizations is twofold: First, to derive new testable inequalities for causal structures not producing a testable inequality from Proposition 1. Second, to find inequalities with higher inferential power, even when some already exist. These objectives are achieved introducing inequalities with less constringent requirements of conditional independence and using data processing inequalities to substitute certain variables from $\mathbf{A}_{[n]}$, so that the conditions of independence are fulfilled or the number of intersections is reduced and lower values in $\mathbf{d}$ are obtained. The first extension relaxes the conditions $\mathbf{A}_i \perp \mathbf{A}_j\backslash\mathbf{A}_i|\mathbf{Z}\ \forall i,j$ required in Proposition 1:

Proposition 3. *(Weaker conditions of independence through group augmentation for a decomposition of information from groups with conditionally independent non-shared components): Consider a collection of groups* $\mathbf{A}_{[n]}$*, a conditioning set* $\mathbf{Z}$*, and target variables* $\mathbf{Y}$ *as in Proposition 1. Consider that for each group* $\mathbf{A}_i$ *a group* $\mathbf{B}_i$ *exists, such that* $\mathbf{A}_i \subseteq \mathbf{B}_i$ *and* $\mathbf{B}_i$ *can be partitioned in two disjoint subsets* $\mathbf{B}_i = \{\mathbf{B}_i^{(1)}, \mathbf{B}_i^{(2)}\}$ *such that* $\mathbf{B}_i^{(1)}$ *fulfills the conditions of independence* $\mathbf{B}_i^{(1)} \perp \mathbf{B}_j^{(1)}\backslash\mathbf{B}_i^{(1)}|\mathbf{Z}$ *and* $\mathbf{B}_i^{(2)}$ *the conditions* $\mathbf{B}_i^{(2)} \perp \mathbf{B}_j\backslash\mathbf{B}_i^{(2)}|\mathbf{B}_i^{(1)}\mathbf{Z}\ \forall i,j$*, and such that* $\mathbf{B}_{[n]}^{(1)} \equiv \{\mathbf{B}_1^{(1)},\ldots,\mathbf{B}_n^{(1)}\}$ *and* $\mathbf{B}_{[n]}^{(2)} \equiv \{\mathbf{B}_1^{(2)},\ldots,\mathbf{B}_n^{(2)}\}$ *are disjoint. Define the maximal values* $d_{\mathbf{B}_i}$ *like in Proposition 1 but for*

the augmented groups $\mathbf{B}_{[n]} \equiv \{\mathbf{B}_1, \ldots, \mathbf{B}_n\}$. *Then, the conditional information that* $\mathbf{B}_{[n]}$ *has about the target variables* $\mathbf{Y}$ *given* $\mathbf{Z}$ *is bounded from below by:*

$$I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z}).$$

Proof. The proof is provided in Appendix A. □

The contribution of Proposition 3 is to relax the conditional independence requirements $\mathbf{A}_i \perp \mathbf{A}_j \backslash \mathbf{A}_i | \mathbf{Z}$. Analogous conditions remain for $\mathbf{B}_i^{(1)}$, but $\mathbf{B}_i^{(2)}$ needs to fulfill the conditions $\mathbf{B}_i^{(2)} \perp \mathbf{B}_j \backslash \mathbf{B}_i^{(2)} | \mathbf{B}_i^{(1)} \mathbf{Z}$ $\forall i, j$. This means that the variables in $\mathbf{B}_i^{(1)}$ are used to separate the variables in $\mathbf{B}_i^{(2)}$ from other groups. If $\mathbf{B}_i^{(2)}$ is empty for all i, Proposition 3 reduces to Proposition 1.

Another difference between Propositions 1 and 3 regards the role of hidden variables. Assume that each $\mathbf{A}_i$ is formed by $\{\mathbf{V}_i, \mathbf{U}_i\}$, where $\mathbf{U}_i$ are hidden variables and $\mathbf{V}_i$ observable variables. In Proposition 1, the requirement that the variables are observable is not fundamental and could be removed. However, to obtain a testable inequality, monotonicity of mutual information would need to be applied to reduce each term $I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z})$ to its estimable lower bound $I(\mathbf{Y}; \mathbf{V}_i|\mathbf{Z})$ that does not contain the hidden variables $\mathbf{U}_i$. On the other hand, the fulfillment of $\mathbf{A}_i \perp \mathbf{A}_j \backslash \mathbf{A}_i | \mathbf{Z}$ implies $\mathbf{V}_i \perp \mathbf{V}_j \backslash \mathbf{V}_i | \mathbf{Z}$, and reducing $\mathbf{A}_i$ to $\mathbf{V}_i$ can only decrease the number of intersections, and hence $\mathbf{d}_{\mathbf{V}_{[n]}}$ values are equal or smaller than $\mathbf{d}_{\mathbf{A}_{[n]}}$. Therefore, with Proposition 1, there is no advantage in including hidden variables. When testing Proposition 1 for a hypothesis of the underlying causal structure (and related independencies), it is equally or more powerful to use $\mathbf{V}_{[n]}$ than $\mathbf{A}_{[n]}$.

This changes in Proposition 3, since $\mathbf{B}_i^{(1)}$ appears in the conditioning side of the independencies that constrain $\mathbf{B}_i^{(2)}$. If hidden variables within $\mathbf{B}_i^{(1)}$ are necessary to create the independencies for $\mathbf{B}_i^{(2)}$, it is not possible to reduce each group to its subset of observable variables. Note that, for a hypothesized causal structure, whether the independence conditions required by Proposition 3 are fulfilled can be verified without observing the hidden variables by using the d-separation criterion on the causal graph, assuming d-separation implies independence. The actual estimation of mutual information values is only needed when testing an inequality from the data.

If $\mathbf{B}_{[n]}$ includes hidden variables, in general $I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z})$ cannot be estimated and $H(\mathbf{Y}|\mathbf{Z})$ is used as an upper bound. For the r.h.s. of the inequality, a lower bound is obtained by monotonicity of the mutual information, removing the hidden variables. In general, a testable inequality has the form

$$H(\mathbf{Y}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{V}_i|\mathbf{Z}), \tag{5}$$

with $\mathbf{V}_i \subseteq \mathbf{B}_i$ being the observable variables within each group. In the case that $I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z}) = I(\mathbf{Y}; \mathbf{V}_{[n]}|\mathbf{Z})$, that is, if the hidden variables do not add information, then a testable tighter upper bound is available using $I(\mathbf{Y}; \mathbf{V}_{[n]}|\mathbf{Z})$. Importantly, the values $\mathbf{d}_{\mathbf{B}_{[n]}}$ are determined using the groups in $\mathbf{B}_{[n]}$. Since $\mathbf{A}_i \subseteq \mathbf{B}_i$, group augmentation comes at the price that $\mathbf{d}_{\mathbf{B}_{[n]}}$ are equal or higher than $\mathbf{d}_{\mathbf{A}_{[n]}}$, but the conditional independence requirements may not be fulfilled without it. Note also that the partition $\mathbf{B}_i = \{\mathbf{B}_i^{(1)}, \mathbf{B}_i^{(2)}\}$ is not known a priori, but determined in the process of finding suitable augmented groups that fulfill the conditions.

We examine some examples before further generalizations. Throughout all figures, we will read independencies from the causal structures using d-separation, assuming faithfulness. In Figure 2A, consider groups $\mathbf{A}_1 = \{V_1, V_2\}$ and $\mathbf{A}_2 = \{V_3, V_4\}$, and $\mathbf{Z} = \varnothing$. Proposition 1 is not applicable due to $V_2 \not\perp V_3$. Augmenting the groups to $\mathbf{B}_1^{(1)} = \mathbf{B}_2^{(1)} = \{U\}$, $\mathbf{B}_1^{(2)} = \{V_1, V_2\}$, and $\mathbf{B}_2^{(2)} = \{V_3, V_4\}$ the conditions of Proposition 3 are fulfilled,

as can be verified by d-separation. Coefficients are determined by $\mathbf{d} = \{2, 2\}$ due to the intersection of the groups in U. Note that hidden variables are not restricted to be hidden common ancestors, and here U is a mediator between V_2 and V_3. In Figure 2B, consider groups $\mathbf{A}_1 = \{V_1\}$, $\mathbf{A}_2 = \{V_3\}$, $\mathbf{A}_3 = \{V_5\}$, which do not fulfill the conditions of Proposition 1. Augmenting the groups to $\mathbf{B}_1^{(1)} = \{U_2, U_4\}$, $\mathbf{B}_1^{(2)} = \{V_1\}$, $\mathbf{B}_2^{(1)} = \{U_2\}$, $\mathbf{B}_2^{(2)} = \{V_3\}$, $\mathbf{B}_3^{(1)} = \{U_4\}$, and $\mathbf{B}_3^{(2)} = \{V_5\}$ the conditions are fulfilled. Maximal intersection values are $\mathbf{d} = \{2, 2, 2\}$. In both examples the upper bound is $H(Y)$ since $I(Y; \mathbf{B}_{[n]})$ cannot be estimated due to hidden variables.

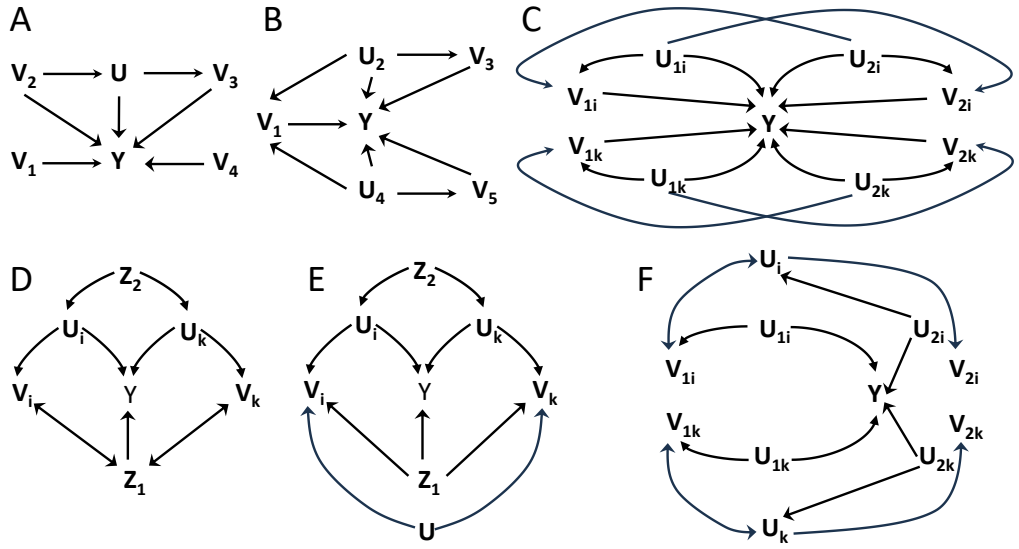

Figure 2. Examples of applications of Proposition 3 (**A**–**C**) and Proposition 4 (**D**–**F**) to obtain testable inequalities. The causal graphs allow verifying if the required conditional independence conditions are fulfilled by using d-separation. Variable Y is the target variable, observable variables are denoted by V, hidden variables by U, and conditioning variables by Z. For all examples, the composition of groups is described in the main text. For graphs using subindexes i, k to display two concrete groups, those are representative of the same causal structure for all groups that compose the system. In those graphs, variables with no subindex have the same connectivity with all groups. Bidirectional arrows indicate common hidden parents not included in any group.

We also consider scenarios with more groups. Figure 2C represents $2N$ groups organized in pairs, with subindexes i, k indicating two particular pairs. The $2N$ groups are defined in pairs, with $\mathbf{A}_{1j} = \{V_{1j}\}$ and $\mathbf{A}_{2j} = \{V_{2j}\}$, $j = 1, \ldots, N$. The causal structure is the same across pairs, but the mechanisms generating the variables beyond the causal structure can possibly differ. Proposition 1 is not fulfilled since $V_{1j} \not\perp V_{2j}$. Groups can be augmented to $\mathbf{B}_{j'j}^{(1)} = \{U_{1j}, U_{2j}\}$, $\mathbf{B}_{j'j}^{(2)} = \{V_{j'j}\}$, for $j' = 1, 2$. Proposition 3 then holds with $d = 2$ for all $2N$ groups. The pairs of groups contribute to the sum as $1/2[I(Y; V_{1j}, U_{1j}, U_{2j}) + I(Y; V_{2j}, U_{1j}, U_{2j})]$, which in the testable inequality of the form of Equation (5) reduces to $1/2[I(Y; V_{1j}) + I(Y; V_{2j})]$. The upper bound to the sum of $2N$ terms is $H(Y)$. This inequality provides causal inference power because $V_{1j} \perp V_{2j} | U_{1j}, U_{2j}$ for all j is not directly testable. As previously indicated, the inference power of an inequality emanates from the possibility to discard causal structures that do not fulfill it. Note that for this system an alternative is to define N groups instead of $2N$ groups, each as $\mathbf{A}'_j = \{V_{1j}, V_{2j}\}$. In this case Proposition 1 is already applicable with the coefficients being all 1, since $V_{1i}, V_{2i} \perp V_{1j}, V_{2j}$ for all $i \neq j$. For this inequality, each of the N groups contributes with $I(Y; V_{1j}) + I(Y; V_{2j} | V_{1j})$, and since there are no hidden variables the l.h.s. is $I(Y; \mathbf{A}'_{[n]})$. However, this latter inequality holds for any causal structure that fulfills $V_{1i}, V_{2i} \perp V_{1j}, V_{2j}$ for all $i \neq j$. Given that these independencies do not involve hidden variables, they are di-

rectly testable from data, so that the latter inequality does not provide additional inference power, in contrast to the former one.

We now continue with further generalizations. Group augmentation in Proposition 3 cannot decrease the values of the maximal number of intersections. We now describe how the data processing inequalities in Lemma 1(i) and Proposition 2 can be used to substitute variables within the groups, potentially reducing the number of intersections. We start with the data processing inequality for the conditional mutual information.

Proposition 4. *(Decomposition of information from groups modified with the conditional mutual information data processing inequality): Consider a collection of groups* $\mathbf{A}_{[n]}$*, a conditioning set* $\mathbf{Z}$*, and target variables* $\mathbf{Y}$ *as in Proposition 1. Consider that for some group* $\mathbf{A}_i$ *a group* $\mathbf{B}_i$ *exists such that* $\mathbf{Y} \perp \mathbf{A}_i \backslash \mathbf{B}_i | \mathbf{B}_i \mathbf{Z}$*, with* $\mathbf{A}_i \backslash \mathbf{B}_i \neq \varnothing$*. Define* $\mathbf{B}_{[n]}$ *as the collection of groups that replaces* $\mathbf{A}_i$ *by* $\mathbf{B}_i$ *for those following the previous independence condition. If* $\mathbf{B}_{[n]}$ *fulfills the conditions of Proposition 3, the inequality derived for* $\mathbf{B}_{[n]}$ *also provides an upper bound for the sum of the information provided by the groups in* $\mathbf{A}_{[n]}$*:*

$$H(\mathbf{Y}|\mathbf{Z}) \geqslant I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z}).$$

Proof. The proof applies Proposition 3 to $\mathbf{B}_{[n]}$ followed by the data processing inequality of Lemma 1(i) to each term within the sum in which $\mathbf{A}_i$ and $\mathbf{B}_i$ are different. Given that $\mathbf{Y} \perp \mathbf{A}_i \backslash \mathbf{B}_i | \mathbf{B}_i \mathbf{Z}$ implies $I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) \geqslant I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z})$, their sum is also smaller or equal. □

Proposition 3 envisaged cases in which the conditions of independence of Proposition 1 were not fulfilled for a collection $\mathbf{A}_{[n]}$ and augmentation allowed fulfilling weaker conditions, even if with higher $\mathbf{d}_{\mathbf{B}_{[n]}}$ values compared to $\mathbf{d}_{\mathbf{A}_{[n]}}$. Proposition 4 is useful not only when the conditions of independence are not fulfilled for $\mathbf{A}_{[n]}$, but more generally if some values in $\mathbf{d}_{\mathbf{B}_{[n]}}$ are lower than in $\mathbf{d}_{\mathbf{A}_{[n]}}$, hence providing a tighter inequality. Including hidden variables in $\mathbf{B}_{[n]}$ is beneficiary when replacing observed by hidden variables leads to fewer intersections. The procedures of Proposition 3 and 4 can be combined, that is, starting with $\mathbf{A}_{[n]}$ that contains only observable variables, a new collection can be constructed adding new variables and removing others from $\mathbf{A}_{[n]}$, ending with $\mathbf{B}_{[n]}$ that contains both observable and hidden variables. The collection $\mathbf{B}_{[n]}$ fulfilling the conditions of Proposition 3 may even contain only hidden variables, and a testable inequality is obtained as long as the data processing inequality allows calculating observable lower bounds for all terms in the sum.

Figure 2D–F are examples of Proposition 4. Again we consider cases with N groups with equal causal structure and use indexes i, k to represent two concrete groups. In Figure 2D, with $\mathbf{A}_j = \{V_j\}$, Proposition 3 does not apply for $\mathbf{A}_{[n]}$ conditioning on $\{Z_1, Z_2\}$ because $V_i \not\perp V_j | Z_1, Z_2$, for all i, j. However, given that $Y \perp V_j | U_j, Z_1, Z_2$, each V_j can be replaced to build $\mathbf{B}_j = \{U_j\}$, and since $U_i \perp U_j | Z_1, Z_2$, for all i, j Proposition 3 applies after using Proposition 4 to create $\mathbf{B}_{[n]}$. A testable inequality is derived with upper bound $H(Y|Z_1, Z_2)$ and a sum of terms $I(Y; V_j|Z_1, Z_2)$, each being a lower bound of $I(Y; U_j|Z_1, Z_2)$ given the data processing inequality that follows from $Y \perp V_j | U_j, Z_1, Z_2$. The coefficients are $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$. Therefore, in this case Proposition 4 results in an inequality when no inequality held for $\mathbf{A}_{[n]}$. In Figure 2E, the same procedure relies on $Y \perp V_j | U_j, Z_1, Z_2$ and $U_i \perp U_j | Z_1, Z_2$ to use $\mathbf{B}_j = \{U_j\}$ to create a testable inequality with l.h.s. $H(Y|Z_1, Z_2)$ and the sum of terms $I(Y; V_j|Z_1, Z_2)$ in the r.h.s. with $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$. Note that by U, which has no subindex, we represent in Figure 2E a hidden common driver of all N groups, not only the displayed i, k. In this example Proposition 3 could have been directly applied without using Proposition 4 if augmenting $\mathbf{A}_j = \{V_j\}$ to $\mathbf{B}'_j = \{V_j, U\}$, with $\mathbf{B}'^{(1)}_j = \{U\}$ and $\mathbf{B}'^{(2)}_j = \{V_j\}$, since $V_i \perp V_j | U, Z_1, Z_2$. However, $\mathbf{d}_{\mathbf{B}'_{[n]}} = \mathbf{N}$, since all groups $\mathbf{B}'_j$ intersect in U. Therefore, in this case an inequality already exists without applying Proposition 4, but its use allows

replacing $\mathbf{d}_{\mathbf{B}'_{[n]}} = \mathbf{N}$ by $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$, hence creating a tighter inequality with higher causal inference power.

In Figure 2F, again we consider $2N$ groups, consisting of N pairs with the same causal structure across pairs and indices i, k representing two of these pairs. For groups $\mathbf{A}_{j'j} = \{V_{j'j}\}$, with $j' = 1, 2$ and $j = 1, ..., N$, Proposition 3 is directly applicable for $\mathbf{B}^{(1)}_{j'j} = \{U_j\}$ and $\mathbf{B}^{(2)}_{j'j} = \{V_{j'j}\}$, with $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{2}$. The data processing inequalities associated with $Y \perp V_{j'j}|U_{j'j}$ allow applying Proposition 4 to obtain an inequality for the groups $\mathbf{B}'_{j'j} = \mathbf{B}'^{(1)}_{j'j} = \{U_{j'j}\}$, which $\mathbf{d}_{\mathbf{B}'_{[n]}} = \mathbf{1}$.

Proposition 4 relies on the data processing inequality of the conditional mutual information. The data processing inequality of unique information can also be used for the same purpose, and both data processing inequalities can be combined applying them to different groups.

Proposition 5. *(Decomposition of information from groups modified using across different groups the conditional or unique information data processing inequality): Consider a collection of groups* $\mathbf{A}_{[n]}$*, a conditioning set* $\mathbf{Z}$*, and target variables* $\mathbf{Y}$ *as in Proposition 1. Consider a subset of groups such that for* $\mathbf{A}_i$ *a group* $\mathbf{B}_i$ *exists such that, for some* $\mathbf{Z}^{(1)}_i \subseteq \mathbf{Z}$, $\mathbf{Y} \perp \mathbf{A}_i \backslash \mathbf{B}_i | \mathbf{B}_i \mathbf{Z}^{(1)}_i$*, with* $\mathbf{A}_i \backslash \mathbf{B}_i \neq \varnothing$*. Define* $\mathbf{B}_{[n]}$ *as the collection of groups that replaces* $\mathbf{A}_i$ *by* $\mathbf{B}_i$ *for those following the previous independence conditions. Define* $\mathbf{Z}^{(1)}_i \equiv \mathbf{Z}$ *for the unaltered groups and* $\mathbf{Z}^{(2)}_i \equiv \mathbf{Z} \backslash \mathbf{Z}^{(1)}_i$ *for all groups. If* $\mathbf{B}_{[n]}$ *fulfills the conditions of Proposition 3, the inequality derived for* $\mathbf{B}_{[n]}$ *also provides an upper bound for a sum combining conditional and unique information terms for different groups in* $\mathbf{A}_{[n]}$*:*

$$H(\mathbf{Y}|\mathbf{Z}) \geqslant I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) \geqslant$$
$$\sum_{\{i:|\mathbf{Z}^{(2)}_i|=0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z}) + \sum_{\{i:|\mathbf{Z}^{(2)}_i|>0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i \backslash\backslash \mathbf{Z}^{(2)}_i | \mathbf{Z}^{(1)}_i).$$

Proof. The proof applies Proposition 3 to $\mathbf{B}_{[n]}$ and then both types of data processing inequalities depending on which one holds for different groups:

$$\begin{aligned} \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) &\overset{(a)}{\geqslant} \sum_{\{i:|\mathbf{Z}^{(2)}_i|=0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) + \sum_{\{i:|\mathbf{Z}^{(2)}_i|>0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i \backslash\backslash \mathbf{Z}^{(2)}_i | \mathbf{Z}^{(1)}_i) \overset{(b)}{\geqslant} \\ \sum_{\{i:|\mathbf{Z}^{(2)}_i|=0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z}) &+ \sum_{\{i:|\mathbf{Z}^{(2)}_i|>0\}} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i \backslash\backslash \mathbf{Z}^{(2)}_i | \mathbf{Z}^{(1)}_i). \end{aligned} \tag{6}$$

Inequality (a) follows from the unique information always being equal to or smaller than the conditional mutual information (Equation (3)). Inequality (b) applies the conditional mutual information data processing inequality to those groups with $\mathbf{A}_i$ different than $\mathbf{B}_i$ but $|\mathbf{Z}^{(2)}_i| = 0$, and the unique information data processing inequality to those groups with $|\mathbf{Z}^{(2)}_i| > 0$. □

Proposition 5 is useful when the conditions of independence required to apply Proposition 3 do not hold for $\mathbf{A}_{[n]}$. It can also be useful to obtain inequalities with higher causal inferential power if $\mathbf{d}_{\mathbf{B}_{[n]}}$ are smaller than $\mathbf{d}_{\mathbf{A}_{[n]}}$, even if Proposition 3 is directly applicable. By definition, the terms $I(\mathbf{Y}; \mathbf{A}_i \backslash\backslash \mathbf{Z}^{(2)}_i | \mathbf{Z}^{(1)}_i)$ are equal to or smaller than $I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z})$, which can only decrease the lower bound, but the data processing inequality may hold only for the unique information and not the conditional information term. Note that the partition

$\{\mathbf{Z}_i^{(1)}, \mathbf{Z}_i^{(2)}\}$ can be group-specific and selected such that data processing inequalities can be applied.

Figure 3A shows an example of the application of the data processing inequality of unique information. For $\mathbf{A}_j = \{V_j\}$, Proposition 3 does not apply to $I(Y; \mathbf{A}_{[n]}|Z)$ because $V_i \not\perp V_k|Z$. The data processing inequality of conditional mutual information does not hold with $Y \not\perp V_i|U_iZ$. This data processing inequality could be used adding to U_i the latent common parent in $Y \leftrightarrow Z$, but this variable would be shared by all augmented groups $\mathbf{B}_i$, leading to an intersection of all N groups. Alternatively, the data processing inequality holds for the unique information with $I(Y; U_j \backslash\backslash Z) \geqslant I(Y; V_j \backslash\backslash Z)$, and $U_i \perp U_j|Z$ for all $i \neq j$. Proposition 5 is applied with $\mathbf{Z}_j^{(1)} = \varnothing$, $\mathbf{Z}_j^{(2)} = \{Z\}$, and $\mathbf{B}_j = \mathbf{B}_j^{(1)} = \{U_j\}$, $\forall j$. This leads to an inequality with $H(Y|Z)$ as upper bound and the sum of terms $I(Y; V_j \backslash\backslash Z)$ at the r.h.s. with coefficients determined by $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$. In Figure 3B, taking $\mathbf{A}_j = \{V_{j1}, V_{j2}\}$ $\forall j$ and defining the conditioning set $\mathbf{Z} = \{Z, Z_1, ..., Z_N\}$, we have $V_{i2} \not\perp V_{k2}|\mathbf{Z}$ and $V_{j1}, V_{j2} \not\perp Y|U_j\mathbf{Z}$. On the other hand, $V_{j1}, V_{j2} \perp Y|U_j\mathbf{Z}\backslash Z_j$, so that the data processing can be applied with the unique information and Proposition 5 is applied with $\mathbf{Z}_j^{(1)} = \mathbf{Z}\backslash Z_j$, $\mathbf{Z}_j^{(2)} = \{Z_j\}$ and $\mathbf{B}_j = \mathbf{B}_j^{(1)} = \{U_j\}$. An inequality exists given that $U_i \perp U_k|\mathbf{Z}$, and the testable inequality has an upper bound $H(Y|\mathbf{Z})$ and at the r.h.s. the sum of terms $I(Y; V_{j1}V_{j2} \backslash\backslash Z_j|\mathbf{Z}\backslash Z_j)$, with $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$.

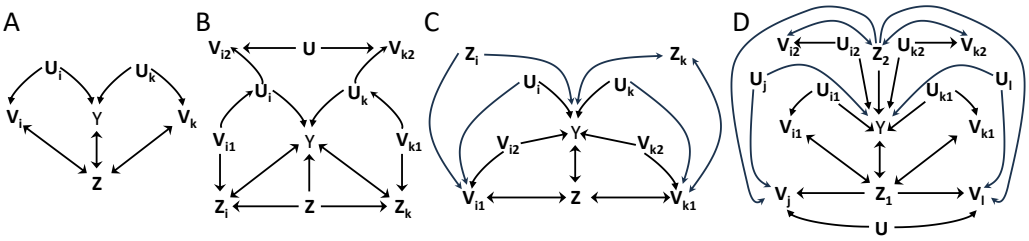

Figure 3. Examples of the application of Proposition 5 (**A**–**C**) and Proposition 6 (**D**) to obtain testable inequalities. Notation is analogous to Figure 2. The composition of groups is described in the main text.

In Figure 3C, we examine an example in which groups differ in the causal structure of the conditioning variable Z_j: For the groups of the type of group i, Z_i is a common parent of Y and V_{i1}. For the groups of the type of k, Z_k is a collider in a path between Y and V_{k1}. Consider M groups of the former type and $N - M$ of the latter. We examine the existence of an inequality for groups defined as $\mathbf{A}_j = \{V_{j1}, V_{j2}\}$ $\forall j$, with $\mathbf{Z} = \{Z, Z_1, \ldots, Z_N\}$. Proposition 3 cannot be applied to $I(Y; \mathbf{A}_{[n]}|\mathbf{Z})$ because $V_{i1} \not\perp V_{j1}|\mathbf{Z}$ for all $i \neq j$. The mutual information data processing inequality is not applicable to substitute V_{j1} because $Y \not\perp V_{j1}|U_jV_{j2}\mathbf{Z}$. However, for the M groups like i, the independence $Y \perp V_{j1}|U_jV_{j2}\mathbf{Z}\backslash Z$ leads to the data processing inequality $I(Y; U_jV_{j2} \backslash\backslash Z|\mathbf{Z}\backslash Z) \geqslant I(Y; V_{j1}V_{j2} \backslash\backslash Z|\mathbf{Z}\backslash Z)$. For these groups, $\mathbf{Z}_j^{(1)} = \mathbf{Z}\backslash Z$ and $\mathbf{Z}_j^{(2)} = \{Z\}$. For the $N - M$ groups like k, the independence $Y \perp V_{j1}|U_jV_{j2}\mathbf{Z}\backslash\{Z, Z_j\}$ leads to $I(Y; U_jV_{j2} \backslash\backslash Z, Z_j|\mathbf{Z}\backslash\{Z, Z_j\}) \geqslant I(Y; V_{j1}V_{j2} \backslash\backslash Z, Z_j|\mathbf{Z}\backslash\{Z, Z_j\})$. For these groups $\mathbf{Z}_j^{(1)} = \mathbf{Z}\backslash\{Z, Z_j\}$ and $\mathbf{Z}_j^{(2)} = \{Z, Z_j\}$. In all cases the modified groups are $\mathbf{B}_j = \mathbf{B}_j^{(1)} = \{U_j, V_{j2}\}$, which fulfill the requirement $U_j, V_{j2} \perp U_i, V_{i2}|\mathbf{Z}$ for all $i \neq j$ needed to apply Proposition 3. The testable inequality that follows from Proposition 5 has upper bound $H(Y|\mathbf{Z})$ and in the sum at the r.h.s. has M terms of the form $I(Y; V_{j1}V_{j2} \backslash\backslash Z|\mathbf{Z}\backslash Z)$ and $N - M$ terms of the form $I(Y; V_{j1}V_{j2} \backslash\backslash Z, Z_j|\mathbf{Z}\backslash\{Z, Z_j\})$. The coefficients are determined by $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$.

Proposition 5 combines both types of data processing inequalities, but only across different groups. Our last extension of Proposition 1 combines both types across and within groups. For each group, we introduce a disjoint partition into m_i subgroups $\mathbf{A}_i = \{\mathbf{A}_i^{(1)}, \ldots, \mathbf{A}_i^{(m_i)}\}$ and define $\mathbf{A}_i^{(0)} \equiv \varnothing$. Subgroups are analogously defined for $\mathbf{Z}_i$, also with $\mathbf{Z}_i^{(0)} \equiv \varnothing$. In general, for any ordered set of vectors we use $\mathbf{V}_i^{[k]} \equiv \{\mathbf{V}_i^{(0)}, \mathbf{V}_i^{(1)}, \ldots, \mathbf{V}_i^{(k)}\}$ to refer to all elements up to k, where in general $\mathbf{V}_i^{(0)}$ can be nonempty.

Proposition 6. *(Decomposition of information from groups modified with the conditional or unique information data processing inequality across and within groups): Consider a collection of groups* $\mathbf{A}_{[n]}$*, a conditioning set* $\mathbf{Z}$*, and a target variable* $\mathbf{Y}$ *as in Proposition 1. Consider that for each group* $\mathbf{A}_i$ *there are disjoint partitions* $\mathbf{A}_i = \{\mathbf{A}_i^{(1)}, \ldots, \mathbf{A}_i^{(m_i)}\}$ *and* $\mathbf{Z} = \{\mathbf{Z}_i^{(1)}, \ldots, \mathbf{Z}_i^{(m_i)}\}$*, and a collection of sets of additional variables* $\mathbf{C}_i = \{\mathbf{C}_i^{(0)}, \mathbf{C}_i^{(1)}, ..., \mathbf{C}_i^{(m_i-1)}\}$*, such that* $\mathbf{Y} \perp \mathbf{A}_i^{(k)} | \mathbf{C}_i^{[k]} \mathbf{Z}_i^{[k]} \mathbf{A}_i \backslash \mathbf{A}_i^{[k]}$ *for* $k = 1, \ldots, m_i - 1$*. Define the collection* $\mathbf{B}_{[n]}$ *with the modified groups* $\mathbf{B}_i = \{\mathbf{C}_i, \mathbf{A}_i^{(m_i)}\}$*. If* $\mathbf{B}_{[n]}$ *fulfills the conditions of Proposition 3, the inequality derived for* $\mathbf{B}_{[n]}$ *also provides an upper bound for sums combining conditional and unique information terms for different groups in* $\mathbf{A}_{[n]}$*:*

$$H(\mathbf{Y}|\mathbf{Z}) \geqslant I(\mathbf{Y}; \mathbf{B}_{[n]}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) \geqslant$$

$$\sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{C}_i^{[k_i]} \mathbf{A}_i \backslash \mathbf{A}_i^{[k_i]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k_i]} | \mathbf{Z}_i^{[k_i]}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{A}_i \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{(1)} | \mathbf{Z}_i^{(1)}),$$

for $k_i \in \{1, \ldots, m_i - 1\}$*.*

Proof. The proof is provided in Appendix A. □

If $m_i = 1$ for all i, then $\mathbf{A}_i^{(1)} = \mathbf{A}_i$, $\mathbf{Z}_i^{(1)} = \mathbf{Z}$, $\mathbf{B}_i = \{\mathbf{C}_i^{(0)}, \mathbf{A}_i\}$, and Proposition 6 reduces to Proposition 3. If $m_i = 2$ and $\mathbf{Z}_i^{(1)} = \mathbf{Z}$ for all i, we recover Proposition 4, with $\mathbf{B}_i = \{\mathbf{C}_i, \mathbf{A}_i^{(2)}\}$. If $m_i = 2$ for all i and $\mathbf{Z}_i^{(1)} \subset \mathbf{Z}$ for some i, we recover Proposition 5, with $\mathbf{B}_i = \{\mathbf{C}_i, \mathbf{A}_i^{(2)}\}$ and $\mathbf{Z}_i^{(2)} = \mathbf{Z} \backslash \mathbf{Z}_i^{(1)}$. Like for previous propositions, some groups may be unmodified such that $\mathbf{B}_i = \mathbf{A}_i$.

The tightest inequality results from maximizing across $k_i \in \{1, \ldots, m_i - 1\}$ each term in the sum. In the proof of Proposition 6 in Appendix A we show that, when increasing $k_i \in \{1, \ldots, m_i - 1\}$, the terms $I(\mathbf{Y}; \mathbf{C}_i^{[k]} \mathbf{A}_i \backslash \mathbf{A}_i^{[k]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k]} | \mathbf{Z}_i^{[k]})$ are monotonically increasing. However, in general $\mathbf{C}_i$ can contain hidden variables, which means that, to obtain a testable inequality, for each $k_i \in \{1, \ldots, m_i - 1\}$ each term needs to be substituted by its lower bound that quantifies the information in the subset of observable variables. For each group, the optimal k_i leading to the tightest inequality will depend on the subset of observable variables $\mathbf{V}_i^{(k_i)} \subseteq \{\mathbf{C}_i^{[k_i]}, \mathbf{A}_i \backslash \mathbf{A}_i^{[k_i]}\}$ and the corresponding values of $I(\mathbf{Y}; \mathbf{V}_i^{(k_i)} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k_i]} | \mathbf{Z}_i^{[k_i]})$.

Figure 3D shows an example of application of Proposition 6. Like in Figure 3C, there are two types of groups with different causal structure. M groups have the structure of the variables with indexes i, k, and $\mathbf{A}_{j'} = \{V_{j'1}, V_{j'2}\}$. The other $N - M$ groups have the structure of the variables with indexes l, j, and $\mathbf{A}_{j'} = \{V_{j'}\}$. The conditioning set selected is $\mathbf{Z} = \{Z_1, Z_2\}$. Proposition 3 cannot be applied directly because $V_{i1} \not\perp V_{k1} | \mathbf{Z}$ for all $i \neq k$ within the M groups, and $V_j \not\perp V_l | \mathbf{Z}$ for all $j \neq l$ within the $N - M$ groups. Proposition 6 applies as follows. For the $N - M$ groups, $m_{j'} = 2$ with $\mathbf{A}_{j'}^{(1)} = \{V_{j'}\}$, $\mathbf{A}_{j'}^{(2)} = \varnothing$, $\mathbf{Z}_{j'}^{(1)} = \mathbf{Z}$, and $\mathbf{B}_{j'} = \mathbf{C}_{j'}^{(1)} = \{U_{j'}\}$. The independencies $\mathbf{Y} \perp \mathbf{A}_i^{(k)} | \mathbf{C}_i^{[k]} \mathbf{Z}_i^{[k]} \mathbf{A}_i \backslash \mathbf{A}_i^{[k]}$ for $k = 1, \ldots, m_i - 1$ correspond in this case to $Y \perp V_{j'} | \mathbf{Z} U_{j'}$, for $k = 1$. For the other M groups, $m_{j'} = 3$ with $\mathbf{A}_{j'}^{(1)} = \{V_{j'1}\}$, $\mathbf{A}_{j'}^{(2)} = \{V_{j'2}\}$, $\mathbf{A}_{j'}^{(3)} = \varnothing$, $\mathbf{Z}_{j'}^{(1)} = \{Z_2\}$, $\mathbf{Z}_{j'}^{(2)} = \{Z_1\}$, $\mathbf{C}_{j'}^{(1)} = \{U_{j'1}\}$, $\mathbf{C}_{j'}^{(2)} = \{U_{j'2}\}$, and $\mathbf{B}_{j'} = \{U_{j'1}, U_{j'2}\}$. The independencies involved are $Y \perp V_{j'1} | Z_2, U_{j'1}, V_{j'2}$, for $k = 1$, and $Y \perp V_{j'2} | \mathbf{Z}, U_{j'1}, U_{j'2}$, for $k = 2$.

Proposition 6 applies because with $\mathbf{B}_{[n]}$ defined as $\mathbf{B}_{j'} = \{U_{j'}\}$ for the $N - M$ groups and $\mathbf{B}_{j'} = \{U_{j'1}, U_{j'2}\}$ for the M groups, the requirements of independence of Proposition 3 are fulfilled, in particular $\mathbf{B}_i \perp \mathbf{B}_j | \mathbf{Z}$ for all $i \neq j$. The terms $I(Y; \mathbf{B}_{j'} | \mathbf{Z})$ for the $N - M$ groups are $I(Y; U_{j'} | Z_1, Z_2)$ and are substituted by lower bounds $I(Y; V_{j'} | Z_1, Z_2)$ in the testable inequality. For the M groups, we have the subsequent sequence of inequalities: $I(Y; U_{j'1}, U_{j'2} | Z_1, Z_2) \geqslant I(Y; U_{j'1}, V_{j'2} | Z_1, Z_2) \geqslant I(Y; U_{j'1}, V_{j'2} \backslash\backslash Z_1 | Z_2) \geqslant I(Y; V_{j'1}, V_{j'2} \backslash\backslash Z_1 | Z_2)$. The first inequality follows from the independence for $k = 2$, the second from the unique

information being equal or smaller than the conditional information, and the third from the independence for $k = 1$. Considering that a testable inequality can only contain observable variables, for the M groups the terms in the sum can be $I(Y; V_{j'1}, V_{j'2} \backslash\backslash Z_1 | Z_2)$ or $I(Y; V_{j'2} | Z_1, Z_2)$, depending on which one is higher. The coefficients are determined by $\mathbf{d}_{\mathbf{B}_{[n]}} = \mathbf{1}$ and the resulting testable inequality has upper bound $H(Y|Z_1, Z_2)$.

Overall, Propositions 4–6 further extend the cases in which groups-decomposition inequalities of the type of Proposition 1 can be derived. Our Proposition 1 extends Proposition 1 of [27] to allow conditioning sets, Proposition 3 further weakens the conditions of independence required in Proposition 1, and Propositions 4–6 use data processing inequalities to obtain testable inequalities from groups-decompositions derived comprising hidden variables, which can be more powerful than inequalities directly derived without comprising hidden variables. In Figures 2 and 3, we have provided examples of causal structures for which these new groups-decompositions inequalities exist. In all these cases, the use of our groups-decomposition inequalities increases the set of available inequality tests that can be used to reject hypothesized causal structures underlying data.

3.3. Inequalities Involving Sums of Information Terms from Ancestral Sets

We now examine inequalities involving ancestral sets as in Theorem 1 of Steudel and Ay [27], which we reviewed in our Theorem 1 (Section 2.4). We extend this theorem allowing for a conditioning set $\mathbf{Z}$ and adding flexibility on how ancestral sets are constructed, as well as allowing the selection of reduced ancestral sets that exclude some variables. Like for Theorem 1, we will use $an_G(\mathbf{A}_{[n]}) \equiv \{an_G(\mathbf{A}_1), \ldots, an_G(\mathbf{A}_n)\}$ to indicate the collection of all ancestral sets in graph G from the collection of groups $\mathbf{A}_{[n]} \equiv \{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$.

The extension of Theorem 1 to allow for a conditioning set $\mathbf{Z}$ requires an extension of the notion of ancestral set that will be used to determine the coefficients in the inequalities. The intuition for this extension is that conditioning on $\mathbf{Z}$ can introduce new dependencies between groups, in particular when a variable $Z_j \in \mathbf{Z}$ is a common descendant of several ancestral groups, and hence conditioning on it activates paths in which it is a collider. The coefficients need to take into account that common information contributions across ancestral groups can originate from these new dependencies. At the same time, conditioning can also block paths that created dependencies between the ancestral groups. To also account for this, we will not only consider ancestral sets in the original graph G, but in any graph $G' = G_{\underline{\mathbf{Z}'}}$, with $\mathbf{Z}' \subseteq \mathbf{Z}$. The graph $G_{\underline{\mathbf{Z}'}}$ is constructed by removing from G all the outgoing arrows from nodes in $\mathbf{Z}'$. This has an effect equivalent to conditioning on $\mathbf{Z}'$ with regard to eliminating dependencies enabled by paths through $\mathbf{Z}'$ in which the variables in $\mathbf{Z}'$ are noncolliders, since removing those arrows deactivates the paths. To account for these effects of conditioning on $\mathbf{Z}$, for each $Z_j \in \mathbf{Z}$ we define an augmented ancestral set of the groups $\mathbf{A}_i \in \mathbf{A}_{[n]}$ as follows:

$$an_{G'}(\mathbf{A}_i; Z_j) \equiv \begin{cases} an_{G'}(\mathbf{A}_i) & \text{if } an_{G'}(\mathbf{A}_i) \perp an_{G'}(Z_j) \cap an_{G'}(\mathbf{A}_{[n]}) | \mathbf{Z} \\ an_{G'}(\mathbf{A}_i) \cup (an_{G'}(Z_j) \cap an_{G'}(\mathbf{A}_{[n]})) & \text{otherwise.} \end{cases} \tag{7}$$

We then define the set $\mathbf{S}(G'; Z_j) \equiv \{\mathbf{A}_i \in \mathbf{A}_{[n]} : an_{G'}(\mathbf{A}_i) \not\perp an_{G'}(Z_j) \cap an_{G'}(\mathbf{A}_{[n]}) | \mathbf{Z}\}$, that is, the set of groups that have some ancestor not independent from some ancestor of Z_j that is also ancestor of $\mathbf{A}_{[n]}$, given $\mathbf{Z}$.

For each $\mathbf{A}_i$, let $d_i(G'; Z_j)$ be the maximal number such that a non-empty intersection exists between $an_{G'}(\mathbf{A}_i; Z_j)$ and $d_i(G'; Z_j) - 1$ other distinct augmented ancestral sets of $\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_{d_i(G'; Z_j)-1}}$. Furthermore, we define $d_i(G'; \mathbf{Z})$ as the maximum for all $Z_j \in \mathbf{Z}$:

$$d_i(G'; \mathbf{Z}) \equiv \max_{Z_j \in \mathbf{Z}} d_i(G'; Z_j). \tag{8}$$

We will use $\mathbf{d}(G';\mathbf{Z})$ to refer to the whole set of maximal values for all groups. If required, we will use $\mathbf{d}_{\mathbf{A}_{[n]}}(G';\mathbf{Z})$ to specify that the collection is $\mathbf{A}_{[n]}$.

In Figure 4A–D, we consider examples to understand the rationale of how $\mathbf{d}_{\mathbf{A}_{[n]}}(G';\mathbf{Z})$ is determined in inequalities with a conditioning $\mathbf{Z}$. In Figure 4A, for groups $\mathbf{A}_1 = \{V_1\}$ and $\mathbf{A}_2 = \{V_2\}$, the augmented ancestral sets on graph G are $an_G(\mathbf{A}_1;Z) = \{V_1, Z\}$ and $an_G(\mathbf{A}_2;Z) = \{V_2, Z\}$, which intersect on Z and $d_i(G;\mathbf{Z}) = 2$ for $i = 1,2$. However, Z is a noncollider in the path creating a dependence between V_1 and V_2, and conditioning on Z renders them independent, so that $d_i(G;\mathbf{Z}) = 2$ overestimates the amount of information the groups may share after conditioning. Alternatively, selecting $G_{\underline{Z}}$ the ancestral sets are $an_{G_{\underline{Z}}}(\mathbf{A}_1;Z) = \{V_1\}$ and $an_{G_{\underline{Z}}}(\mathbf{A}_2;Z) = \{V_2\}$, which do not intersect and $d_i(G_{\underline{Z}};\mathbf{Z}) = 1$ for $i = 1,2$ when calculated following Equation (7). A priori, we do not know which graph $G' = G_{\underline{\mathbf{Z}'}}$, $\mathbf{Z}' \subseteq \mathbf{Z}$, results in a tighter inequality. Here we see that $G_{\underline{Z}}$ leads to an inequality with more causal inference power than G for Figure 4A. In Figure 4B, Z is a collider between V_1 and V_2, so that conditioning on Z creates a dependence between the groups. If the values d_i were determined from the standard ancestral sets, in this case $an_G(\mathbf{A}_i) = an_{G_{\underline{Z}}}(\mathbf{A}_i) = \{V_i\}$, for $i = 1,2$, which do not intersect, leading to unit coefficients. However, the augmented ancestral sets following Equation (7) are $an_G(\mathbf{A}_i;Z) = an_{G_{\underline{Z}}}(\mathbf{A}_i;Z) = \{V_1, V_2\}$ for $i = 1,2$, so that $d_i(G;\mathbf{Z}) = d_i(G_{\underline{Z}};\mathbf{Z}) = 2$. This illustrates that the augmented ancestral sets are necessary to properly determine the coefficients in inequalities with conditioning sets $\mathbf{Z}$, in this case reflecting that $I(Y;V_1|Z)$ and $I(Y;V_2|Z)$ can have redundant information.

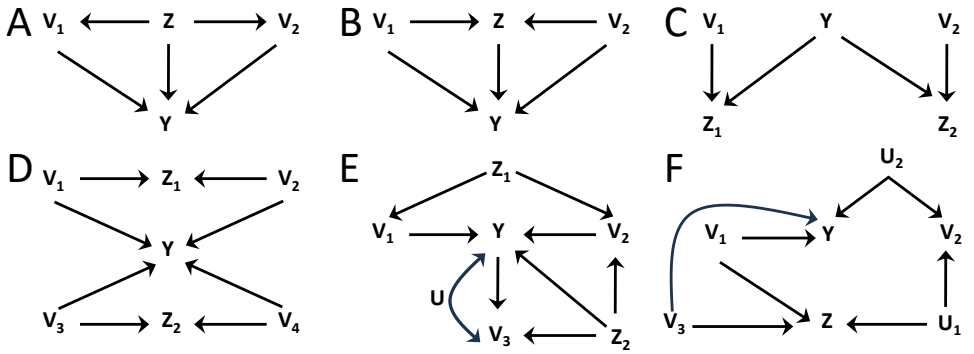

Figure 4. Inequalities involving sums of information terms from ancestral sets. (**A**–**D**) Examples to illustrate the definition of augmented ancestral sets (Equations (7) and (8)). (**E**,**F**) Examples of the application of Theorem 2 to obtain testable inequalities.

Figure 4C shows a scenario in which conditioning creates dependencies of Y with V_1 and V_2, which were previously independent. The standard ancestral sets $an_{G'}(\mathbf{A}_1) = \{V_1\}$ and $an_{G'}(\mathbf{A}_2) = \{V_2\}$ would not intersect in any $G' = G_{\underline{\mathbf{Z}'}}$, with $\mathbf{Z}' \subseteq \{Z_1, Z_2\}$ and would lead to unit values for d_i. On the other hand, the augmented ancestral sets are $an_{G'}(\mathbf{A}_i;Z_j) = \{V_i\}$ for $i = j$ and $an_{G'}(\mathbf{A}_i;Z_j) = \{V_1, V_2\}$ for $i \neq j$, for all $G' = G_{\underline{\mathbf{Z}'}}$, with $\mathbf{Z}' \subseteq \{Z_1, Z_2\}$. This results in $d_i(G';\mathbf{Z}) = 2$ in all cases, which appropriately captures that the two groups can have common information about Y when conditioning on $\{Z_1, Z_2\}$. The example of Figure 4D illustrates why each value $d_i(G';Z_j)$ is determined separately (Equation (7)) first, and only after is the maximum calculated (Equation (8)). Four groups are defined as $\mathbf{A}_i = V_i$ for $i = 1,\ldots,4$. If $d_i(G';\mathbf{Z})$ were to be determined directly from Equation (7) but using $\mathbf{Z} = \{Z_1, Z_2\}$, instead of using separately Z_1 and Z_2, then for all the ancestral sets the augmented ancestral set would include all variables, since $an_{G'}(\mathbf{Z}) \cap an_{G'}(\mathbf{A}_{[n]})$ is equal to $an_{G'}(\mathbf{A}_{[n]})$. This would lead to $d_i = 4, \forall i$. However, that determination would overestimate how many groups become dependent when conditioning on $\mathbf{Z}$, since Z_1 creates a dependence between V_1 and V_2 and Z_2 between V_3 and V_4, but no dependencies across these pairs are created. The determination of $\mathbf{d}(G';\mathbf{Z}) = \mathbf{2}$ from Equations (7) and (8) properly leads to a tighter inequality than the one obtained if considering jointly both conditioning variables.

Equipped with this extended definition of $\mathbf{d}_{\mathbf{A}_{[n]}}(G';\mathbf{Z})$, we now present our generalization of Theorem 1:

Theorem 2. *Let G be a DAG model containing nodes corresponding to a set of (possibly hidden) variables $\mathcal{X}$. Let $\mathbf{Y} \in \mathcal{X}$ be a set of observable target variables, and $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$ a conditioning set of observable variables, with $\mathbf{Z} \subset \mathcal{X}$. Let $\mathbf{A}_{[n]} = \{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ be a collection of (possibly overlapping) groups of (possibly hidden) variables $\mathbf{A}_i \subset \mathcal{X}$. Consider a DAG G' selected as $G' = G_{\underline{\mathbf{Z}'}}$ with $\mathbf{Z}' \subseteq \mathbf{Z}$, constructed by removing from graph G all the outgoing arrows from nodes in $\mathbf{Z}'$. Following Equation (7), define an augmented ancestral set in G' for each group $\mathbf{A}_i \in \mathbf{A}_{[n]}$ for each variable in the conditioning set, $Z_j \in \mathbf{Z}$. Following Equation (8), determine $d_i(G';\mathbf{Z})$ for each group, given the intersections of the augmented ancestral sets $an_{G'}(\mathbf{A}_i; Z_j)$. Select a variable $W_0 \in an_{G'}(\mathbf{A}_{[n]})$ and a group of variables $\mathbf{W} \subseteq D_{G'}(W_0) \cap an_{G'}(\mathbf{A}_{[n]})$, possibly $\mathbf{W} = \varnothing$. Define the reduced ancestral sets $\tilde{an}_{G'}(\mathbf{A}_i) \equiv an_{G'}(\mathbf{A}_i) \backslash \mathbf{W}$ for each $\mathbf{A}_i \in \mathbf{A}_{[n]}$, and the reduced collection $\tilde{an}_{G'}(\mathbf{A}_{[n]}) \equiv an_{G'}(\mathbf{A}_{[n]}) \backslash \mathbf{W}$. The information about $\mathbf{Y}$ in this reduced collection when conditioning on $\mathbf{Z}$ is bounded from below by*

$$I(\mathbf{Y}; \tilde{an}_{G'}(\mathbf{A}_{[n]}) | \mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i(G';\mathbf{Z})} I(\mathbf{Y}; \tilde{an}_{G'}(\mathbf{A}_i) | \mathbf{Z}). \tag{9}$$

Proof. The proof is provided in Appendix B. □

Theorem 2 provides several extensions of Theorem 1. First, it allows for a conditioning set $\mathbf{Z}$. Second, given a hypothesis of the generative causal graph G underlying the data, Theorem 2 can be applied to any $G' = G_{\underline{\mathbf{Z}'}}$ with $\mathbf{Z}' \subseteq \mathbf{Z}$, and hence offers a set of inequalities potentially adding causal inference power. As we have discussed in relation to Figure 4A–D, the selection of G' that leads to the tightest inequality in some cases will be determined by the causal structure, but in general it also depends on the exact probability distribution of the variables. Third, Theorem 2 allows excluding some variables $\mathbf{W}$ from the ancestral sets, although imposing constraints in the causal structure of $\mathbf{W}$. The role of these constraints is clear in the proof at Appendix B. The case of Theorem 1 corresponds to $\mathbf{Z} = \varnothing$, $\mathbf{W} = \varnothing$, and $G' = G$.

Excluding some variables $\mathbf{W}$ can be advantageous. For example, if $\mathbf{Y}$ is univariate and it overlaps with some ancestral sets, as it is the case when some groups include descendants of Y, then the upper bound $I(Y; an_{G'}(\mathbf{A}_{[n]})|\mathbf{Z})$ is equal to $H(Y|\mathbf{Z})$ and also $I(Y; an_{G'}(\mathbf{A}_i)|\mathbf{Z})$ is equal to $H(Y|\mathbf{Z})$ for all ancestral sets that include Y. Excluding $\mathbf{W} = Y$ provides a tighter upper bound $I(Y; an_{G'}(\mathbf{A}_{[n]}) \backslash Y|\mathbf{Z})$ and may provide more causal inferential power. Another scenario in which a reduced collection can be useful is when excluding $\mathbf{W}$ removes all hidden variables from $an_{G'}(\mathbf{A}_{[n]})$, such that $\tilde{an}_{G'}(\mathbf{A}_{[n]})$ is observable, giving $I(\mathbf{Y}; \tilde{an}_{G'}(\mathbf{A}_{[n]})|\mathbf{Z})$ as a testable upper bound instead of $H(\mathbf{Y}|\mathbf{Z})$. When comparing inequalities with different sets $\mathbf{W}$, in some cases the form of the causal structure and the specification of which variables are hidden or observable will a priori determine an order of causal inference power among the inequalities. However, like for the comparison across $G' = G_{\underline{\mathbf{Z}'}}$ with $\mathbf{Z}' \subseteq \mathbf{Z}$, in general the power of the different inequalities depends on the details of the generated probability distributions. Formulating general criteria to rank inequalities with different $\mathbf{Z}$, G', and $\mathbf{W}$ in terms of their inferential power is beyond the scope of this work.

Note that we have formulated Theorem 2 explicitly allowing for hidden variables. Also, in Theorem 1 (as a subcase of Theorem 2) the restriction of $\mathbf{A}_{[n]}$ being observable variables can be removed. In any case, the inclusion of hidden variables can only increase the causal inference power if combined with data processing inequalities to obtain a testable inequality. Propositions 4–6 indicate how to possibly tighten an inequality derived from Proposition 1 by substituting $\mathbf{A}_{[n]}$ by a new collection $\mathbf{B}_{[n]}$ that, including hidden variables, leads to $\mathbf{d}_{\mathbf{B}_{[n]}}$ smaller than $\mathbf{d}_{\mathbf{A}_{[n]}}$. The same application of data processing inequalities of the unique and conditional mutual information can be used for Theorem 2 to determine a $\mathbf{B}_{[n]}$ with $\mathbf{d}_{\mathbf{B}_{[n]}}(G';\mathbf{Z})$ smaller than $\mathbf{d}_{\mathbf{A}_{[n]}}(G';\mathbf{Z})$. The use of data processing inequalities is necessary because they allow substituting some of the observable variables by hidden

variables, instead of only adding hidden variables. When only adding variables, the number of intersections between ancestral groups can only increase, hence not decreasing $\mathbf{d}(G';\mathbf{Z})$. On top of this, a testable inequality replaces information terms of ancestral groups by their lower bounds given by observable subsets of variables. This means that, adding hidden variables, the testable inequality will contain the same information terms of the observable variables, but possibly smaller coefficients, hence resulting in a looser inequality. This is not the case any more when hidden variables are not added but instead substitute some of observable variables, thanks to data processing inequalities. This substitution may decrease the number of intersections between ancestral groups, and the coefficients in the sum may be higher. We will not describe this procedure in detail, since the use of data processing inequalities is analogous to their use in Propositions 4–6.

We now illustrate the application of Theorem 2. In Figure 4E, with $\mathbf{Z} = \{Z_1, Z_2\}$, the conditions of independence required by Proposition 6 do not hold for any set of groups, either $\mathbf{A}_i = \{V_i\}$, $i = 1,2,3$, or, with $i \neq j \neq k$, $\mathbf{A}_1 = \{V_i, V_j\}$, $\mathbf{A}_2 = \{V_i, V_k\}$ or $\mathbf{A}_1 = \{V_i, V_j\}$, $\mathbf{A}_2 = \{V_k\}$. No data processing inequalities can be applied to replace some variables to fulfill the conditions. On the other hand, Theorem 2 can always be applied, since it does not require the fulfillment of some conditions of independence. For example, for $\mathbf{A}_i = \{V_i\}$, $i = 1,2,3$ and for $G' = G_{\underline{Z_1 Z_2}}$, we have $an_{G'}(V_1) = \{V_1\}$, $an_{G'}(V_2) = \{V_2\}$, $an_{G'}(V_3) = \{V_1, V_2, V_3, U, Y\}$, and following Equation (7) $an_{G'}(V_1; Z_j) = \{V_1\}$, $an_{G'}(V_2; Z_j) = \{V_2\}$, and $an_{G'}(V_3; Z_j) = \{V_1, V_2, V_3, U, Y\}$, for $j = 1,2$. This leads to $\mathbf{d}(G';\mathbf{Z}) = \{2,2,3\}$. For illustration purpose, we focus on $\mathbf{W}$ equal to $\{Y, U\}$ or any of its subsets. In all cases $\tilde{an}_{G'}(V_i) = an_{G'}(V_i)$, for $i = 1,2$, contributing terms $1/2 I(Y; V_1|Z_1, Z_2)$ and $1/2 I(Y; V_2|Z_1, Z_2)$. For $\mathbf{W} = \{Y, U\}$ or $\mathbf{W} = \{Y\}$, the contribution of the observable lower bound of the third group is $1/3 I(Y; V_1, V_2, V_3|Z_1, Z_2)$. For $\mathbf{W} = \{U\}$ or $\mathbf{W} = \varnothing$, the third group contributes $1/3 H(Y|Z_1, Z_2)$. For $\mathbf{W} = \{Y, U\}$, $\tilde{an}_{G'}(\mathbf{A}_{[n]}) = \{V_1, V_2, V_3\}$, which is observable and the upper bound is $I(Y; V_1, V_2, V_3|Z_1, Z_2)$. For any other subset of $\{Y, U\}$ the upper bound in the testable inequality is $H(Y|Z_1, Z_2)$. Because the terms in the sum for groups 1 and 2 are equal for all the $\mathbf{W}$ compared, in this case it can be checked that selecting $\mathbf{W} = \{Y, U\}$ leads to the tightest inequality. This example illustrates the utility of being able to construct inequalities for reduced ancestral sets.

While in the previous example only Theorem 2 and not Proposition 6 was applicable, more generally, a causal structure will involve the fulfillment of a set of inequalities, some obtained using Proposition 6 and some using Theorem 2. Which inequalities have higher inferential power will depend on the causal structure and the exact probability distribution of the data, and neither Theorem 2 nor Proposition 6 are more powerful a priori. In Figure 4F, Proposition 6 cannot be applied using $\mathbf{A}_i = \{V_i\}$, $i = 1,2,3$ and conditioning on Z, because $V_i \not\perp\!\!\!\perp V_j|Z, \forall i, j$ and no data processing inequalities help to substitute these variables. On the other hand, Theorem 2 can be applied with $\mathbf{A}_i = \{V_i\}$, leading to $an_{G'}(V_1) = \{V_1\}$, $an_{G'}(V_3) = \{V_3\}$, and $an_{G'}(V_2) = \{V_2, U_1, U_2\}$, for all $G' = G_{\underline{\mathbf{Z}'}}$. The augmented ancestral sets are $an_{G'}(V_1; Z) = \{V_1, V_3, U_1\} = an_{G'}(V_3; Z)$, and $an_{G'}(V_2; Z) = \{V_1, V_2, V_3, U_1, U_2\}$, also for all G', resulting in $\mathbf{d}(G'; Z) = \mathbf{3}$. Focusing on the case of $\mathbf{W} = \{Y, U_2\}$, or any subset of it, in all cases the associated testable inequality has $H(Y|Z)$ as upper bound and in the r.h.s. the sum of terms $1/3 I(Y; V_i|Z)$, $i = 1,2,3$. Alternatively, defining $\mathbf{A}_1 = \{V_1, V_3, U_1\}$ and $\mathbf{A}_2 = \{V_2, U_1\}$, Proposition 3 is applicable with the two groups intersecting in U_1 and $V_1, V_3 \perp\!\!\!\perp V_2|Z, U_1$. The associated testable inequality has the same upper bound $H(Y|Z)$ and in the r.h.s. the sum of terms $1/2 I(Y; V_1, V_3|Z)$ and $1/2 I(Y; V_2|Z)$. In this case, which inequality has more causal inferential power will depend on the exact distribution of the data.

Overall, Theorem 2 extends Theorem 1 of [27], allowing conditioning sets and providing more flexible conditions to form the groups. In the examples of Figure 4, we have illustrated how Theorem 2 substantially increases the number of groups-decomposition inequalities that can be tested to reject hypothesized causal structures to be compatible with a certain data set.

4. Discussion

We have presented several generalizations of the type of groups-decomposition inequalities introduced by [27], which compare the information about target variables contained in a collection of variables with a weighted sum of the information contained in subsets of the collection. These generalizations include an extension to allow for conditioning sets and methods to identify existing inequalities that involve collections and subsets selected with less restrictive criteria. This comprises less restrictive conditions of independence, the use of ancestral sets from subgraphs of the causal structure of interest, and the removal of some variables from the ancestral sets. We have also shown how to exploit inequalities identified for collections containing hidden variables—which are not directly testable—by converting them into testable inequalities using data processing inequalities.

Our use of data processing inequalities to derive testable groups-decomposition inequalities when collections contain hidden variables is not entirely new. We found inspiration for this approach in the proof of Theorem 1 in [24]. This theorem derives a causally informative inequality from a particular type of causal structure, namely common ancestor graphs in which all dependencies between observable variables are caused by hidden common ancestors. The inequality presented in the theorem corresponds to the setting of a univariate target variable and groups composed by different single observable variables. In their simplest case, each hidden ancestor only has two children, which are observable variables. Their proof uses the mutual information data processing inequality to convert a sum of information terms involving the observable variables into a sum of terms involving the hidden ancestors. The final inequality can equally be proven applying our Proposition 4 by deriving an inequality for the collection of hidden variables and then converting it into a testable inequality using data processing inequalities. The same final inequality can also be derived as an application of our Theorem 2 followed by the use of the data processing inequality.

We have expanded the applicability of data processing inequalities by showing that this type of inequality also holds for conditional unique information measures [29]. For a given causal structure, a testable causally informative inequality may be obtained substituting hidden variables by observable variables thanks to the data processing inequality of the unique information, in cases in which the data processing inequality of mutual information is not applicable. As shown in Proposition 6, the unique information data processing inequalities are particularly powerful for deriving groups-decomposition inequalities with a conditioning set, since they can iteratively be applied to replace different subsets of hidden variables by observable variables choosing which variables are kept as conditioning variables and which ones are taken as reference variables for different unique information measures. This use of unique information indicates how other types of information-theoretic measures could be similarly incorporated to derive causally informative inequalities. Recent developments in the decomposition of mutual information into redundant, synergistic, and unique contributions [30] provide candidate measures whose utility for this purpose needs to be further explored [31–35,40,41] (among others). Furthermore, while this type of decomposition has been extensively debated recently [35,42,43], aspects of its characterization are still unsolved and an understanding of how the terms are related to the causal structure can provide new insights.

One particular domain in which our generalizations can be useful is to study causal interactions among dynamical processes [23,44], for which causal interactions are characterized from time series both in the temporal [45] and spectral domain [46–48]. When studying high-dimensional multivariate dynamical processes, such as brain dynamics (e.g., [49–51]) or econometric data [52,53], an important question is to determine whether correlations between time series are related to causal influences or to hidden common influences. For highly interconnected systems with many hidden variables, the number of independencies may be small, hence providing limited information about the causal structure. In this case, inequality constraints can help to substantially narrow down the set of causal structures compatible with the data. Accordingly, our generalization to formulate

conditional inequalities may play an important role in combination with measures to quantify partial dependencies between time series [54,55]. We expect this approach to be easily adaptable to non-stationary time-series, as it is often the case in the presence of unit roots and co-integrated time series [56–58]. This can be carried out by selecting collections and groups consistent with the temporal partitioning in non-stationary information-theoretic measures of causality in time-series [59,60]. Another area to extend the applicability of our proposal is to study non-classical quantum systems [16,61–63]. In this case, an extended d-separation criterion [64] and adapted faithfulness considerations [65] have been proposed to take into account the particularities of quantum systems. Further exploration will be required to determine if and how our derivations that rely on d-separation leading to statistical independence (Appendix C) are also applicable when considering generalized causal structures for quantum systems.

Besides the extension to particular domains, an important question yet to be addressed regards the relation between the causal inferential power of different inequalities. Our proposal considerably enlarges the number of groups-decomposition inequalities of the type of [27] available to test the compatibility of a causal structure with a given data set. We have seen in our analysis some examples of how, under certain conditions, the causal structure imposes an ordering to the power of alternative inequalities. Future work should aim to derive broader criteria to rank the inferential power of inequalities, for example in terms of the relation between the conditioning sets or the constituency of the groups that appear in each inequality. Formulating criteria to rank the inferential power of different inequalities would help to simplify the set of inequalities that needs to be tested when the compatibility of a certain causal structure with the data is to be examined.

Apart from a characterization of how groups-decomposition inequalities are related among themselves, future work should also examine the relation and embedding of this type of inequalities with those derived with other approaches. In our understanding, the algorithmic projection procedure of [23,24] could equally retrieve some of the inequalities here described, but without the advantage of having a constructive procedure to derive the form of an inequality directly reading a causal graph, and instead requiring costly computations that may limit the derivation of inequalities for large systems. The incorporation of constraints for other types of information-theoretic measures, such as constraints involving unique information measures, would require an extension of the algorithmic approach. Among other approaches, the so-called *Inflation technique* [66] stands out as capable of providing asymptotically sufficient tests of causal compatibility [67]. The inflation method creates a new causal structure with multiple copies of the original structure and symmetry constraints on the ancestral properties of the different copies, in such a way that testable constraints on the inflated graph can be mapped back to the compatibility of the original causal structure. However, despite the ongoing advances in its theoretical developments and implementation [68], to our knowledge it is not straightforward to identify the order of inflation and the specific inflation structure adequate to discriminate between certain causal structures. The availability of inequalities easily derived by reading the original causal structure can also be helpful in combination with the inflation method, in order to discard as many candidate causal structures as possible before the design of additional inflated graphs. The connection with other approaches [69–74] also deserves further investigation, ultimately to determine minimal sets of inequality constraints with equivalent inferential power.

Beyond the derivation of existing testable causally informative inequalities, a crucial issue for their application is the implementation of the corresponding tests. This implementation depends on the estimation of information-theoretic measures from data. A ubiquitous challenge for the application of mutual information measures is that they are positively biased and their estimation is data-demanding [75,76]. These biases scale with the dimensionality of the variables, and hence can hinder the applicability of information-theoretic inequalities for large collections of variables, or for variables with high cardinality. However, recent advances in the estimation of mutual information for high-dimensional

data can help to attenuate these biases [77]. Furthermore, the implementation of the tests can take advantage of the existence of both upper-bound and lower-bound estimators of mutual information [78], using opposite bounds at the two sides of the inequalities. These technical aspects of the implementation of the tests are important to apply all types of information-theoretic inequalities [23–27,71]. Despite these common challenges, our extension of groups-decomposition inequalities does not come at the price of having to test inequalities that intrinsically are more difficult to estimate. Our contribution can substantially increase the number of inequalities available to be tested, and we have provided examples in Figures 2–4 of new inequalities in which—in particular thanks to the use of data processing inequalities—the dimensionality of the collections is not increased. Future work is required to determine how to efficiently combine all available tests. In the goal to determine minimal sets of inequality tests that are maximally informative, the statistical power of the tests will need to be considered together with their discrimination power among causal structures.

Author Contributions: All authors contributed to the design of the research. The research was carried out by D.C. The manuscript was written by D.C. with the contribution of J.K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proofs of Propositions 1, 3, and 6

Proof of Proposition 1. Given a collection $\mathbf{A}_{[n]} = \{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$, define $\mathbf{X}_{[r]}$ as the set of r variables that are part of at least a group in $\mathbf{A}_{[n]}$. We have that

$$\begin{aligned} I(\mathbf{Y}; \mathbf{A}_{[n]}|\mathbf{Z}) &\overset{(a)}{=} I(\mathbf{Y}; \mathbf{X}_{[r]}|\mathbf{Z}) \overset{(b)}{=} \sum_{k=1}^{r} I(\mathbf{Y}; X_k|\mathbf{X}_{[k-1]}, \mathbf{Z}) \overset{(c)}{\geqslant} \\ &\sum_{k=1}^{r} I(\mathbf{Y}; X_k|\mathbf{X}_{[k-1]}, \mathbf{Z}) \left(\sum_{\mathbf{A}_i : X_k \in \mathbf{A}_i} \frac{1}{d_i} \right) \overset{(d)}{=} \sum_{i=1}^{n} \frac{1}{d_i} \sum_{X_k \in \mathbf{A}_i} I(\mathbf{Y}; X_k|\mathbf{X}_{[k-1]}, \mathbf{Z}) \overset{(e)}{\geqslant} \\ &\sum_{i=1}^{n} \frac{1}{d_i} \sum_{X_k \in \mathbf{A}_i} I(\mathbf{Y}; X_k|(\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}) \overset{(f)}{=} \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{A}_i|\mathbf{Z}) \end{aligned} \tag{A1}$$

Equality (a) follows from $\mathbf{X}_{[r]}$ containing the same variables as $\mathbf{A}_{[n]}$. Equality (b) follows from the iterative application of the chain rule for mutual information, where $X_{[0]} \equiv \varnothing$ and $\mathbf{X}_{[k-1]} = \{X_0, \ldots, X_{k-1}\}$. Inequality (c) follows from the definition of d_i as maximal, such that the number of groups that contain X_k is equal or smaller than d_i for all $\mathbf{A}_i$ containing X_k, and hence $\sum_{\mathbf{A}_i : X_k \in \mathbf{A}_i} 1/d_i \leqslant 1$. Equality (d) groups together into the inner sum variables within the same group. Inequality (e) follows from Lemma 1(ii). In more detail, $\mathbf{A}_i \perp \mathbf{A}_j \backslash \mathbf{A}_i | \mathbf{Z}\ \forall i, j$, combined with the weak union property of independencies [27], ensures that for each $X_k \in \mathbf{A}_i$, $X_k \perp (\mathbf{X}_{[k-1]} \cap \mathbf{A}_j) \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}, \forall j \neq i$. Assuming faithfulness, this implies $X_k \perp \mathbf{X}_{[k-1]} \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$. Lemma 1 (ii) applies with $A = X_k$, $\mathbf{B} = \{(\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}\}$, and $\mathbf{C} = \mathbf{X}_{[k-1]} \backslash \mathbf{A}_i$. Equality (f) follows applying the chain rule within each group $\mathbf{A}_i$. □

Proposition 1 of [27] is included in the case $\mathbf{Z} = \varnothing$. The faithfulness assumption allows relaxing their assumption $X_k \perp X_{[r]} \backslash X_k \forall k$ to $\mathbf{A}_i \perp \mathbf{A}_j \backslash \mathbf{A}_i | \mathbf{Z}\ \forall i, j$. A tighter bound can be obtained in some cases if some variables are trimmed. In particular, for a variable X', $\mathbf{A}_j$ can be trimmed to $\mathbf{A}_j \backslash X'$ for all groups such that $I(\mathbf{Y}; \mathbf{A}_j|\mathbf{Z}) = I(\mathbf{Y}; \mathbf{A}_j \backslash X'|\mathbf{Z})$ and possibly lower d_j values can be obtained after trimming. We do not explicitly include this trimming process in the definition of d_j to simplify the formulation.

Proof of Proposition 3. Consider a collection of groups $\mathbf{B}_{[n]} = \{\mathbf{B}_1, \ldots, \mathbf{B}_n\}$, each with a partition in disjoint subsets $\mathbf{B}_i = \{\mathbf{B}_i^{(1)}, \mathbf{B}_i^{(2)}\}$ that fulfill the conditions $\mathbf{B}_i^{(1)} \perp \mathbf{B}_j^{(1)} \backslash \mathbf{B}_i^{(1)} | \mathbf{Z}$ and $\mathbf{B}_i^{(2)} \perp \mathbf{B}_j \backslash \mathbf{B}_i^{(2)} | \mathbf{B}_i^{(1)} \mathbf{Z}$ $\forall i, j$, and such that $\mathbf{B}_{[n]}^{(1)} = \{\mathbf{B}_1^{(1)}, \ldots, \mathbf{B}_n^{(1)}\}$ and $\mathbf{B}_{[n]}^{(2)} = \{\mathbf{B}_1^{(2)}, \ldots, \mathbf{B}_n^{(2)}\}$ are disjoint. Define $\mathbf{X}_{[r_k]}^{(k)}$ as the set of r_k variables part of at least a group in $\mathbf{B}_{[n]}^{(k)}$, for $k = 1, 2$. We have that

$$
\begin{aligned}
& I(\mathbf{Y}; \mathbf{B}_{[n]} | \mathbf{Z}) \overset{(a)}{=} I(\mathbf{Y}; \mathbf{X}_{[r_1]}^{(1)}, \mathbf{X}_{[r_2]}^{(2)} | \mathbf{Z}) \overset{(b)}{=} \\
& \sum_{X_k \in \mathbf{X}_{[r_1]}^{(1)}} I(\mathbf{Y}; X_k | \mathbf{X}_{[k-1]}^{(1)}, \mathbf{Z}) + \sum_{X_k \in \mathbf{X}_{[r_2]}^{(2)}} I(\mathbf{Y}; X_k | \mathbf{X}_{[k-1]}^{(2)}, \mathbf{X}_{[r_1]}^{(1)}, \mathbf{Z}) \overset{(c)}{\geqslant} \\
& \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{B}_i^{(1)} | \mathbf{Z}) + \sum_{i=1}^{n} \frac{1}{d_i} \sum_{X_k \in \mathbf{B}_i^{(2)}} I(\mathbf{Y}; X_k | \mathbf{X}_{[k-1]}^{(2)}, \mathbf{X}_{[r_1]}^{(1)}, \mathbf{Z}) \overset{(d)}{\geqslant} \\
& \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{B}_i^{(1)} | \mathbf{Z}) + \sum_{i=1}^{n} \frac{1}{d_i} \sum_{X_k \in \mathbf{B}_i^{(2)}} I(\mathbf{Y}; X_k | (\mathbf{X}_{[k-1]}^{(2)} \cap \mathbf{B}_i^{(2)}), \mathbf{B}_i^{(1)}, \mathbf{Z}) \overset{(e)}{=} \\
& \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{B}_i^{(1)} | \mathbf{Z}) + \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{B}_i^{(2)} | \mathbf{B}_i^{(1)} \mathbf{Z}) \overset{(f)}{=} \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \mathbf{B}_i | \mathbf{Z}).
\end{aligned}
\tag{A2}
$$

Equality (a) holds because $\{\mathbf{X}_{[r_1]}^{(1)}, \mathbf{X}_{[r_2]}^{(2)}\}$ contains the same variables as $\mathbf{B}_{[n]}$. Equality (b) is an iterative application of the chain rule. Inequality (c) is as follows: For the sum in $\mathbf{X}_{[r_1]}^{(1)}$, steps (c) to (f) of Equation (A1) are all combined, substituting sets $\mathbf{A}_i$ by $\mathbf{B}_i^{(1)}$ and given that these variables fulfill conditions of independence equivalent to Proposition 1. For the sum in $\mathbf{X}_{[r_2]}^{(2)}$, only steps (c) and (d) of Equation (A1) are applied, substituting sets $\mathbf{A}_i$ by $\mathbf{B}_i^{(2)}$. Inequality (d) holds applying Lemma 1 (ii). In more detail, $\mathbf{B}_i^{(2)} \perp \mathbf{B}_j \backslash \mathbf{B}_i^{(2)} | \mathbf{B}_i^{(1)} \mathbf{Z}$ $\forall i, j$ combined with the weak union property of independencies [27] mean that for each $X_k \in \mathbf{B}_i^{(2)}$, $X_k \perp \{\mathbf{B}_j^{(1)}, (X_{[k-1]}^{(2)} \cap \mathbf{B}_j^{(2)}) \backslash \mathbf{B}_i^{(2)}\} | (X_{[k-1]}^{(2)} \cap \mathbf{B}_i^{(2)}), \mathbf{B}_i^{(1)}, \mathbf{Z}$ $\forall j \neq i$. Assuming faithfulness, this implies $X_k \perp \{(\mathbf{X}_{[r_1]}^{(1)} \backslash \mathbf{B}_i^{(1)}), (X_{[k-1]}^{(2)} \backslash \mathbf{B}_i^{(2)})\} | (X_{[k-1]}^{(2)} \cap \mathbf{B}_i^{(2)}), \mathbf{B}_i^{(1)}, \mathbf{Z}$. Accordingly, Lemma 1 (ii) applies with $A = X_k$, $\mathbf{B} = \{(X_{[k-1]}^{(2)} \cap \mathbf{B}_i^{(2)}), \mathbf{B}_i^{(1)}, \mathbf{Z}\}$, and $\mathbf{C} = \{(\mathbf{X}_{[r_1]}^{(1)} \backslash \mathbf{B}_i^{(1)}), (X_{[k-1]}^{(2)} \backslash \mathbf{B}_i^{(2)})\}$. Equalities (e) and (f) follow from the chain rule of mutual information. □

Before continuing with the proof of Proposition 6, we formulate in Lemma A1 a property of the unique information that will be used in the proof.

Lemma A1. *(Conditioning on reference variables increases conditional unique information): The conditional unique information $I(\mathbf{Y}; \mathbf{X} \backslash\backslash \mathbf{Z}_1 \mathbf{Z}_2 | \mathbf{Z}_3)$ is smaller than or equal to $I(\mathbf{Y}; \mathbf{X} \backslash\backslash \mathbf{Z}_1 | \mathbf{Z}_2 \mathbf{Z}_3)$, where $\mathbf{Z}_2$ moves from the set of reference predictors of the unique information to the conditioning set.*

Proof of Lemma A1. The unique information $I(\mathbf{Y}; \mathbf{X} \backslash\backslash \mathbf{Z}_1 \mathbf{Z}_2 | \mathbf{Z}_3)$ is by definition (Equation (2)) the minimum information $I(\mathbf{Y}; \mathbf{X} | \mathbf{Z}_1 \mathbf{Z}_2 \mathbf{Z}_3)$ among the distributions that preserve $P(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_3)$ and $P(\mathbf{Y}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$, and $I(\mathbf{Y}; \mathbf{X} \backslash\backslash \mathbf{Z}_1 | \mathbf{Z}_2 \mathbf{Z}_3)$ is the minimum information $I(\mathbf{Y}; \mathbf{X} | \mathbf{Z}_1 \mathbf{Z}_2 \mathbf{Z}_3)$ among the distributions that preserve $P(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$ and $P(\mathbf{Y}, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$. Since the latter constraints subsume the former ones, the minimum can only be equal or higher. □

Proof of Proposition 6. For iterations $k = 1, ..., m_i - 1$, consider the following:

$$I(\mathbf{Y}; \mathbf{C}_i^{[k-1]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k-1]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k-1]} | \mathbf{Z}_i^{[k-1]}) \overset{(a)}{\leqslant} I(\mathbf{Y}; \mathbf{C}_i^{[k]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k-1]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k-1]} | \mathbf{Z}_i^{[k-1]}) \overset{(b)}{\leqslant} I(\mathbf{Y}; \mathbf{C}_i^{[k]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k]} | \mathbf{Z}_i^{[k]}) \overset{(c)}{=} I(\mathbf{Y}; \mathbf{C}_i^{[k]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k]} | \mathbf{Z}_i^{[k]}). \tag{A3}$$

Inequality (a) holds from monotonicity, information cannot decrease if adding $\mathbf{C}_i^{(k)}$ to $\mathbf{C}_i^{[k-1]}$. Inequality (b) holds from Lemma A1, moving $\mathbf{Z}_i^{(k)}$ from the set of reference predictors of the unique information to the conditioning set. Equality (c) follows from $\mathbf{A}_i \backslash \mathbf{A}_i^{[k-1]} = \{\mathbf{A}_i^{(k)}, \mathbf{A}_i \backslash \mathbf{A}_i^{[k]}\}$ and the assumption in Proposition 6 that $\mathbf{Y} \perp \mathbf{A}_i^{(k)} | \mathbf{C}_i^{[k]} \mathbf{Z}_i^{[k]} \mathbf{A}_i \backslash \mathbf{A}_i^{[k]}$ holds. Accordingly, the unique information is preserved removing $\mathbf{A}_i^{(k)}$ (Proposition 2). This leads to the inequality $I(\mathbf{Y}; \mathbf{C}_i^{[k-1]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k-1]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k-1]} | \mathbf{Z}_i^{[k-1]}) \leqslant I(\mathbf{Y}; \mathbf{C}_i^{[k]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k]} | \mathbf{Z}_i^{[k]})$. Equation (A3) iterated for $k = 1, \ldots, m_i - 1$ leads to $I(\mathbf{Y}; \mathbf{B}_i \backslash\backslash \mathbf{Z}_i^{(m_i)} | \mathbf{Z}_i^{[m_i-1]})$, with $\mathbf{B}_i = \{\mathbf{C}_i^{[m_i-1]}, \mathbf{A}_i^{(m_i)}\}$. Finally, this unique information by construction is smaller than $I(\mathbf{Y}; \mathbf{B}_i | \mathbf{Z})$. The terms $I(\mathbf{Y}; \mathbf{A}_i \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{(1)} | \mathbf{Z}_i^{(1)})$ are obtained removing $\mathbf{C}_i^{[1]}$ from $I(\mathbf{Y}; \mathbf{C}_i^{[k]}\mathbf{A}_i \backslash \mathbf{A}_i^{[k-1]} \backslash\backslash \mathbf{Z} \backslash \mathbf{Z}_i^{[k]} | \mathbf{Z}_i^{[k]})$ by monotonicity, from step $k = 1$. □

Appendix B. Proof of Theorem 2

Proof of Theorem 2. The proof proceeds by induction like the proof of Theorem 1 in [27]. To render the notation less heavy, we simplify $an_{G'}(\mathbf{A}_i)$ to $an(\mathbf{A}_i)$ and $d_i(G'; \mathbf{Z})$ to d_i, with both G' and $\mathbf{Z}$ fixed. Define $\mathbf{V}_Z = \{\mathbf{V}_Z^{(1)}, \ldots, \mathbf{V}_Z^{(m)}\}$, with $\mathbf{V}_Z^{(j)} \equiv (an(Z_j) \cap an(\mathbf{A}_{[n]})) \backslash \mathbf{W}$. Without loss of generality, for $j = 1, \ldots, m$ we sequentially apply the chain rule to separate the information that each subset $\mathbf{V}_Z^{(j)}$ provides about $\mathbf{Y}$ after the chain rule has already been applied to $\mathbf{V}_Z^{[j-1]} \equiv \{\mathbf{V}_Z^{(1)}, ..., \mathbf{V}_Z^{(j-1)}\}$. At the j-th iteration, we obtain

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) = I(\mathbf{Y}; \mathbf{V}_Z^{(j)} | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) + I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}_Z^{[j]}). \tag{A4}$$

The iterative induction step proceeds as follows. Assume that the inequality of Theorem 2 holds for

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}_Z^{[j]}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}_Z^{[j]}). \tag{A5}$$

We show that then the inequality also holds for $I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}_Z^{[j-1]})$. First, if $\mathbf{V}_Z^{(j)} \subseteq \{\mathbf{Z}, \mathbf{V}_Z^{[j-1]}\}$ then $\{\mathbf{Z}, \mathbf{V}_Z^{[j-1]}\} = \{\mathbf{Z}, \mathbf{V}_Z^{[j]}\}$ and Equation (A5) already provides the desired inequality. We continue with $\mathbf{V}_Z^{(j)} \nsubseteq \{\mathbf{Z}, \mathbf{V}_Z^{[j-1]}\}$. Split the sum in Equation (A5) into two sums, one containing groups in the set $\mathbf{S}(G'; Z_j)$ (see its definition below Equation (7)), and the other groups not in $\mathbf{S}(G'; Z_j)$. For the sake of simplifying notation, we use $\mathbf{S}_{Z_j}$ for $\mathbf{S}(G'; Z_j)$, given that G' is fixed. We first consider the sum of groups in $\mathbf{S}_{Z_j}$:

$$\begin{aligned}
&\sum_{\mathbf{A}_i \in \mathbf{S}_{Z_j}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}_Z^{[j]}) \overset{(a)}{=} \\
&\sum_{\mathbf{A}_i \in \mathbf{S}_{Z_j}} \frac{1}{d_i} \Big[I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i), \mathbf{V}_Z^{(j)} | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) - I(\mathbf{Y}; \mathbf{V}_Z^{(j)} | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) \Big] \overset{(b)}{\geqslant} \\
&\left[\sum_{\mathbf{A}_i \in \mathbf{S}_{Z_j}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i), \mathbf{V}_Z^{(j)} | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) \right] - I(\mathbf{Y}; \mathbf{V}_Z^{(j)}) | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) \overset{(c)}{\geqslant} \\
&\left[\sum_{\mathbf{A}_i \in \mathbf{S}_{Z_j}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}) \right] - I(\mathbf{Y}; \mathbf{V}_Z^{(j)}) | \mathbf{Z}, \mathbf{V}_Z^{[j-1]}).
\end{aligned} \tag{A6}$$

Equality (a) follows from the chain rule. Inequality (b) follows from the definition of $d_i(G';\mathbf{Z})$ (in short d_i) in Equation (8). By construction $d_i(G';\mathbf{Z})$ is equal to or higher than all $d_i(G';Z_j)$ and $d_i(G';Z_j)$ is the maximal number of groups intersecting together with $an(\mathbf{A}_i;Z_j)$ (Equation (7)). For any group i within $\mathbf{S}(G';Z_j)$, $an(\mathbf{A}_i;Z_j)$ includes $an(Z_j)\cap an(\mathbf{A}_{[n]})$ and hence $d_i(G';Z_j)\geqslant|\mathbf{S}(G';Z_j)|$, so that $\sum_{\mathbf{A}_i\in\mathbf{S}_{Z_j}}1/d_i\leqslant 1$. Inequality (c) follows from the monotonicity property of the mutual information. For the other sum

$$\sum_{\mathbf{A}_i\notin\mathbf{S}_{Z_j}}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)|\mathbf{Z},\mathbf{V}_Z^{[j]})\overset{(a)}{\geqslant}\sum_{\mathbf{A}_i\notin\mathbf{S}_{Z_j}}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)\backslash\mathbf{V}_Z^{(j)}|\mathbf{Z},\mathbf{V}_Z^{[j-1]})\overset{(b)}{=}\sum_{\mathbf{A}_i\notin\mathbf{S}_{Z_j}}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)|\mathbf{Z},\mathbf{V}_Z^{[j-1]}). \tag{A7}$$

Inequality (a) follows from applying Lemma 1 (ii), with $\mathbf{A}=\tilde{an}(\mathbf{A}_i)\backslash\{\mathbf{Z},\mathbf{V}_Z^{[j]}\}$, $\mathbf{B}=\{\mathbf{Z},\mathbf{V}_Z^{[j-1]}\}$, and $\mathbf{C}=\mathbf{V}_Z^{(j)}\backslash\{\mathbf{Z},\mathbf{V}_Z^{[j-1]}\}$. Independence $\mathbf{A}\perp\mathbf{C}|\mathbf{B}$ holds because $\mathbf{A}_i\notin\mathbf{S}(G';Z_j)$ means $an(\mathbf{A}_i)\perp an(Z_j)\cap an(\mathbf{A}_{[n]})|\mathbf{Z}$ (Equation (7)), which implies $\tilde{an}(\mathbf{A}_i)\perp\mathbf{V}_Z^{(j)}|\mathbf{Z}$, given that $\mathbf{V}_Z^{(j)}\equiv(an(Z_j)\cap an(\mathbf{A}_{[n]}))\backslash\mathbf{W}$. Assuming faithfulness, since all the variables in $\mathbf{V}_Z^{[j-1]}$ are ancestors of $\mathbf{Z}$, conditioning on $\{\mathbf{Z},\mathbf{V}_Z^{[j-1]}\}$ does not create any new dependence (activating colliders) that did not exist conditioning on $\mathbf{Z}$. Equality (b) holds because given Equation (7) an overlap between $\tilde{an}(\mathbf{A}_i)\backslash\mathbf{Z}$ and $\mathbf{V}_Z^{(j)}\backslash\mathbf{Z}$ is in contradiction with $\mathbf{A}_i\notin\mathbf{S}(G';Z_j)$. Combining Equations (A6) and (A7) in the r.h.s of Equation (A5), we obtain that

$$I(\mathbf{Y};\tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z},\mathbf{V}_Z^{[j]})\geqslant\left[\sum_{i=1}^{n}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)|\mathbf{Z},\mathbf{V}_Z^{[j-1]})\right]-I(\mathbf{Y};\mathbf{V}_Z^{(j)}|\mathbf{Z},\mathbf{V}_Z^{[j-1]}). \tag{A8}$$

We then insert this inequality in Equation (A4) to obtain the final desired inequality:

$$I(\mathbf{Y};\tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z},\mathbf{V}_Z^{[j-1]})\geqslant\sum_{i=1}^{n}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)|\mathbf{Z},\mathbf{V}_Z^{[j-1]}). \tag{A9}$$

After subtracting $\mathbf{V}_Z=\{\mathbf{V}_Z^{(1)},\ldots,\mathbf{V}_Z^{(m)}\}$, the validity of the inequality of Theorem 2 depends on the validity of

$$I(\mathbf{Y};\tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z},an(\mathbf{Z})\cap\tilde{an}(\mathbf{A}_{[n]}))\geqslant\sum_{i=1}^{n}\frac{1}{d_i}I(\mathbf{Y};\tilde{an}(\mathbf{A}_i)|\mathbf{Z},an(\mathbf{Z})\cap\tilde{an}(\mathbf{A}_{[n]})). \tag{A10}$$

At each iterations, if $\tilde{an}(\mathbf{A}_{[n]})\backslash\{\mathbf{Z},\mathbf{V}_Z^{[j]}\}$ is empty, the corresponding assumption in Equation (A5) is trivially fulfilled and the proof ends. Otherwise, the proof by induction continues further subtracting variables from $\tilde{an}(\mathbf{A}_{[n]})\backslash an(\mathbf{Z})$. We define the set of groups whose ancestral set in G' overlaps with $\mathbf{W}$:

$$\mathbf{S}_\mathbf{W}\equiv\{\mathbf{A}_i\in\mathbf{A}_{[n]}:an_{G'}(\mathbf{A}_i)\cap\mathbf{W}\neq\varnothing\}. \tag{A11}$$

We select subsets of variables to be subtracted using the same criterion used in the proof of Theorem 1 of [27], but restricting the groups used as reference in each iteration to be in the complementary set $\overline{\mathbf{S}}_\mathbf{W}$, i.e., with $an(\mathbf{A}_i)=\tilde{an}(\mathbf{A}_i)$. In more detail, consider without loss of generality that in the first iteration the j-th group $\mathbf{A}_j$ is taken as reference. Define $\mathbf{V}^{(0)}\equiv\mathbf{V}_Z$, where $\mathbf{V}_Z=\{\mathbf{V}_Z^{(1)},...,\mathbf{V}_Z^{(m)}\}$ has already been subtracted from $\tilde{an}(\mathbf{A}_{[n]})$. With $\mathbf{A}_j$ as reference, find the joint intersection of $\tilde{an}(\mathbf{A}_j)\backslash\mathbf{V}^{(0)}$ with a maximal number of other groups $\tilde{an}(\mathbf{A}_{j'})\backslash\mathbf{V}^{(0)}$, $j'\neq j$. Define $\mathbf{S}_j^{(1)}$ as the set of groups in this intersection. The superindex indicates that this set is associated with the first iteration of this part of

the induction procedure, while the subindex indicates that the j-th group is the reference. The subindex will be omitted when the group used as reference is not relevant. Define $\mathbf{V}_j^{(1)} \equiv \bigcap_{\mathbf{A}_i \in \mathbf{S}_j^{(1)}} \tilde{an}(\mathbf{A}_i) \backslash \mathbf{V}^{(0)}$ as the set of variables contained in this intersection. This subset is subtracted in the first iteration. Analogously, consider that in the k-th iteration $\mathbf{V}^{[k-1]} \equiv \{\mathbf{V}^{(0)}, ..., \mathbf{V}^{(k-1)}\}$ has already been subtracted and the j-th group is taken as reference. Then $\mathbf{S}_j^{(k)}$ is determined by the joint intersection of $\tilde{an}(\mathbf{A}_j) \backslash \mathbf{V}^{[k-1]}$ with a maximal number of other groups $\tilde{an}(\mathbf{A}_{j'}) \backslash \mathbf{V}^{[k-1]}$, $j' \neq j$. The subset of variables subtracted in the k-th iteration is $\mathbf{V}_j^{(k)} \equiv \bigcap_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} \tilde{an}(\mathbf{A}_i) \backslash \mathbf{V}^{[k-1]}$. By construction $\mathbf{V}_j^{(k)} \subseteq \tilde{an}_{G'}(\mathbf{A}_{[n]}) \backslash \mathbf{V}^{[k-1]}$. Furthermore, $|\mathbf{S}_j^{(k)}| \leqslant d_j(G'; \mathbf{Z})$, since $d_j(G'; \mathbf{Z})$ is maximal among $d_j(G'; Z_i)$ for $i = 1, \ldots, m$ (Equation (8)) for all intersections of the augmented ancestral sets defined in Equation (7), while $\mathbf{S}_j^{(k)}$ is determined by only intersections with no support in $\mathbf{V}^{[k-1]}$ and only among the reduced ancestral sets. So far, we have described the selection of subsets to be subtracted. We now look at the iterative induction step when removing a subset $\mathbf{V}_j^{(k)}$ after the previous $k-1$ iterations have already been performed. Consider

$$\begin{aligned} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}^{[k-1]}) &\overset{(a)}{=} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) \\ &\overset{(b)}{=} I(\mathbf{Y}; \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) + I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}^{[k]}). \end{aligned} \tag{A12}$$

Equality (a) follows from $\mathbf{V}_j^{(k)} \subseteq \tilde{an}(\mathbf{A}_{[n]}) \backslash \mathbf{V}^{[k-1]}$. Equality (b) is an application of the chain rule. We now show that under the assumption that

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}^{[k]}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}^{[k]}), \tag{A13}$$

the analogous inequality holds for $I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]}) | \mathbf{Z}, \mathbf{V}^{[k-1]})$. We again break down the sum of the groups into two sums, one containing groups in $\mathbf{S}_j^{(k)}$ and the other the rest. We first consider the sum of groups in $\mathbf{S}_j^{(k)}$:

$$\begin{aligned} &\sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}^{[k]}) \overset{(a)}{=} \\ &\sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} \frac{1}{d_i} \Big[I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) - I(\mathbf{Y}; \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) \Big] \overset{(b)}{\geqslant} \\ &\left[\sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) \right] - I(\mathbf{Y}; \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}) \overset{(c)}{\geqslant} \\ &\left[\sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}^{[k-1]}) \right] - I(\mathbf{Y}; \mathbf{V}_j^{(k)} | \mathbf{Z}, \mathbf{V}^{[k-1]}). \end{aligned} \tag{A14}$$

Equality (a) follows from the chain rule. Inequality (b) holds because $|\mathbf{S}_j^{(k)}| \leqslant d_i(G'; \mathbf{Z})$ for all $\mathbf{A}_i \in \mathbf{S}_j^{(k)}$. This is because the intersection that determines $\mathbf{S}_j^{(k)}$ contains variables from $\mathbf{A}_j$ and from all other groups $\mathbf{A}_i \in \mathbf{S}_j^{(k)}$, and hence for all these groups it also determines $d_i(G'; \mathbf{Z})$ unless an intersection with more groups exists for $\mathbf{A}_i$. Given $|\mathbf{S}_j^{(k)}| \leqslant d_i(G'; \mathbf{Z})$ for all $\mathbf{A}_i \in \mathbf{S}_j^{(k)}$, it follows that $\sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} 1/d_i(G'; \mathbf{Z}) \leqslant \sum_{\mathbf{A}_i \in \mathbf{S}_j^{(k)}} 1/|\mathbf{S}_j^{(k)}| = 1$. Inequality (c)

follows from monotonicity of mutual information. We now consider the sum involving groups that do not belong to $\mathbf{S}_j^{(k)}$:

$$\sum_{\mathbf{A}_i \notin \mathbf{S}_j^{(k)}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}^{[k]}) \geqslant \sum_{\mathbf{A}_i \notin \mathbf{S}_j^{(k)}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i) | \mathbf{Z}, \mathbf{V}^{[k-1]}). \tag{A15}$$

The inequality holds applying Lemma 1(ii) with $\mathbf{A} = \tilde{an}(\mathbf{A}_i) \backslash \{\mathbf{Z}, \mathbf{V}^{[k]}\}$, $\mathbf{B} = \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$, and $\mathbf{C} = \mathbf{V}_j^{(k)} \backslash \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$. By construction, $\mathbf{V}_j^{(k)} \cap \{\mathbf{Z}, \mathbf{V}^{[k-1]}\} = \varnothing$ and hence $\mathbf{C} = \mathbf{V}_j^{(k)}$. Furthermore, $\tilde{an}(\mathbf{A}_i) \backslash \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ is equal to $\tilde{an}(\mathbf{A}_i) \backslash \{\mathbf{Z}, \mathbf{V}^{[k]}\}$ given that $\mathbf{A}_i \notin \mathbf{S}_j^{(k)}$. An intersection of $\tilde{an}(\mathbf{A}_i) \backslash \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ and $\mathbf{V}_j^{(k)}$ is contradictory with the definition of $\mathbf{V}_j^{(k)}$, since $|\mathbf{S}_j^{(k)}|$ is determined to be maximal, but would increase to $|\mathbf{S}_j^{(k)}| + 1$ if defined by that intersection, and that would lead to $\mathbf{A}_i \in \mathbf{S}_j^{(k)}$ instead. Lemma 1(ii) applies given the independence $\mathbf{A} \perp \mathbf{C} | \mathbf{B}$. We now prove that this independence holds. We proceed discarding the presence of all types of paths in G that would create a dependence $\mathbf{A} \not\perp \mathbf{C} | \mathbf{B}$. Under the faithfulness assumption, we examine the four different types of paths in G that could create a dependence. First, there is a variable $X_r \in \mathbf{C}$ and a variable $X_l \in \mathbf{A}$ with an active directed path in G from X_r to X_l, not blocked by $\mathbf{B}$. If this path is active in G conditioning on $\mathbf{B} = \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$, it also exists in any $G' = G_{\underline{\mathbf{Z}}'}$, with $\mathbf{Z}' \subseteq \mathbf{Z}$, since the removal of outgoing arrows has the same effect as conditioning for the paths in which the conditioning variables are noncolliders (i.e., do not have two incoming arrows). This active directed path means that X_r would be an ancestor of X_l in G'. Therefore, given $X_l \in \mathbf{A}$ and $X_r \in \mathbf{C}$, X_r itself would be part of $\tilde{an}_{G'}(\mathbf{A}_i) \backslash \mathbf{V}^{[k-1]}$. However, as argued above, an intersection of $\tilde{an}_{G'}(\mathbf{A}_i) \backslash \mathbf{V}^{[k-1]}$ and $\mathbf{V}_j^{(k)}$ is contradictory with $\mathbf{A}_i \notin \mathbf{S}_j^{(k)}$. Second, there is a variable $X_r \in \mathbf{C}$ and a variable $X_l \in \mathbf{A}$ with an active directed path in G from X_l to X_r, not blocked by $\mathbf{B}$. Again, this path being active in G when conditioning on $\mathbf{B} = \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$, means that it also exists in any $G' = G_{\underline{\mathbf{Z}}'}$, with $\mathbf{Z}' \subseteq \mathbf{Z}$. Therefore, X_l would be an ancestor of X_r in G'. This is again a contradiction with the definition of $\mathbf{V}_j^{(k)}$ because it could be redefined to include $|\mathbf{S}_j^{(k)}| + 1$ groups, since X_l would be an ancestor of all groups intersecting in $\mathbf{V}_j^{(k)}$. Third, there is a variable $X_r \in \mathbf{C}$, a variable $X_l \in \mathbf{A}$, and another variable X_h that is not part of $\mathbf{A}$ nor $\mathbf{C}$ with an active directed path in G from X_h to X_r and an active directed path from X_h to X_l, both not blocked by $\mathbf{B}$. This would also imply that these directed paths exist in $G' = G_{\underline{\mathbf{Z}}'}$, with $\mathbf{Z}' \subseteq \mathbf{Z}$, and hence X_h is an ancestor of $\mathbf{A}$ and $\mathbf{C}$ in G'. Since X_h is an ancestor of $\mathbf{A} = \tilde{an}(\mathbf{A}_i) \backslash \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ but by construction $X_h \notin \mathbf{A}$, this means that X_h has to be part of $\{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ or of $\mathbf{W}$, since any ancestor of $an(\mathbf{A}_i)$ is part of $an(\mathbf{A}_i)$. If $X_h \in \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$, conditioning on $\mathbf{B} = \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ would prevent from having active directed paths from X_h to X_r and from X_h to X_l, leading to a contradiction. We now consider the case $X_h \in \mathbf{W}$. Since X_h is an ancestor of $\mathbf{C} = V_j^{(k)}$, by construction of $V_j^{(k)}$, X_h is an ancestor of $\tilde{an}(\mathbf{A}_j) \backslash \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$. This means that $an(\mathbf{A}_j)$ includes $X_h \in \mathbf{W}$ which, given Equation (A11), is in contradiction with the criterion for selection of reference groups such that $\mathbf{A}_j \in \overline{\mathbf{S}}_{\mathbf{W}}$. In these three types of cases, an active path would exist despite conditioning on $\mathbf{B}$. In the last type, a path would be activated by conditioning on $\mathbf{B}$. At least one variable $X_h \in \mathbf{B} = \{\mathbf{Z}, \mathbf{V}^{[k-1]}\}$ has to be a collider or a descendant of a collider along the path that conditioning activates. Consider first that a single collider X_h is involved. For the collider to activate the path, it must exist an active directed subpath to X_h from a variable X_r that is part of $\mathbf{C}$ or part of its ancestor set in G'. Since this directed subpath is active in G when conditioning on $\mathbf{B}$, it is also active in G'. This means that X_r would be an ancestor of X_h in G'. If X_h is part of $\mathbf{Z}$ or part of $\mathbf{V}^{(0)} \equiv (an_{G'}(\mathbf{Z}) \cap an_{G'}(\mathbf{A}_{[n]})) \backslash \mathbf{W}$, then X_r being an ancestor of X_h means that it is part of $an_{G'}(\mathbf{Z}) \cap an_{G'}(\mathbf{A}_{[n]})$. Accordingly, by definition of $\mathbf{V}^{(0)}$, X_r would be part of $\mathbf{V}^{(0)}$ or of $\mathbf{W}$. The former option leads to a contradiction because

$\mathbf{V}^{(0)}$ has already been removed from $\tilde{an}(\mathbf{A}_{[n]})$ and is part of the conditioning variables, so that the subpath from X_r to X_h could not be part of the path activated by conditioning on the collider. The latter option, X_r being part of $\mathbf{W}$, is in contradiction with it being an ancestor of $\mathbf{C} = \mathbf{V}_j^{(k)}$, since this means being an ancestor of the group $\mathbf{A}_j$ taken as reference to build $\mathbf{V}_j^{(k)}$, which by construction is chosen from $\bar{S}_{\mathbf{W}}$. We continue considering that X_h is part of $\mathbf{V}^{(k')} \in \mathbf{V}^{[k-1]}$, for $0 < k' \leqslant k-1$. In this case, X_r being an ancestor of X_h would mean that either X_r is in $\mathbf{W}$ or it would have been possible to define $\mathbf{V}^{(k')}$ to include X_r. In the former case, this leads to a contradiction because for $0 < k' \leqslant k-1$ all $\mathbf{V}^{(k')}$ have been constructed taking as reference a group belonging to $\bar{S}_{\mathbf{W}}$. In the latter case, this leads to a contradiction because $\mathbf{V}^{(k')}$ is constructed to include all variables in the intersection with the maximum number of groups. The same reasoning holds if the activated path contains more than one collider from $\mathbf{B}$, by selecting the collider X_h closest to a variable in $\mathbf{C}$ along the path. Since for all four types of paths that could lead to $\mathbf{A} \not\perp \mathbf{C}|\mathbf{B}$ we reach a contradiction, $\mathbf{A} \perp \mathbf{C}|\mathbf{B}$ holds and Lemma 1(ii) can be applied to obtain the inequality in Equation (A15). Combining Equations (A14) and (A15) with the r.h.s. of Equation (A13), we obtain that

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \mathbf{V}^{[k]}) \geqslant \left[\sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \mathbf{V}^{[k-1]})\right] - I(\mathbf{Y}; \mathbf{V}_j^{(k)}|\mathbf{Z}, \mathbf{V}^{[k-1]}). \tag{A16}$$

We then insert this inequality in Equation (A12) to obtain the desired inequality:

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \mathbf{V}^{[k-1]}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \mathbf{V}^{[k-1]}). \tag{A17}$$

After the completion of these iterations, all variables in $(an(\mathbf{Z}) \cap an(\mathbf{A}_{[n]}))\backslash\mathbf{W}$ and in groups from $\overline{\mathbf{S}}_{\mathbf{W}}$ have been subtracted from $\tilde{an}(\mathbf{A}_{[n]})$. The proof ends if after some iteration $\tilde{an}(\mathbf{A}_{[n]})\backslash\mathbf{V}^{[k]}$ is empty. In particular, the proof ends if $\mathbf{W}$ is empty and hence all groups are already subtracted. Otherwise, assume that m' iterations have been carried out when finishing this step. The proof by induction continues with a single additional step for the remaining groups $\mathbf{S}_{\mathbf{W}}$. Select a single variable X_0 out of $\tilde{an}(\mathbf{A}_{[n]})\backslash\mathbf{V}^{[m']}$ that is only contained in groups in $\mathbf{S}_{\mathbf{W}}$ and apply the chain rule

$$\begin{aligned} &I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \mathbf{V}^{[m']}) = I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) + I(\mathbf{Y}; X_0|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) = \\ &I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) + I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0). \end{aligned} \tag{A18}$$

The iterative induction step should prove that if the inequality of the theorem holds for $I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0)$ it is also true for $I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \mathbf{V}^{[m']})$. We will prove this below. Before we show that the inequality

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \tag{A19}$$

always holds, and hence it provides the base case for the induction proof. The base case is true because

$$\begin{aligned} &\sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \overset{(a)}{=} \sum_{i: X_0 \in \tilde{an}(\mathbf{A}_i)} \frac{1}{d_i} I(\mathbf{Y}; X_0|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \\ &\overset{(b)}{=} I(\mathbf{Y}; X_0|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \left[\sum_{i: X_0 \in \tilde{an}(\mathbf{A}_i)} \frac{1}{d_i}\right] \overset{(c)}{\leqslant} I(\mathbf{Y}; X_0|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0). \end{aligned} \tag{A20}$$

Equality (a) holds because $I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0)$ is zero for the terms that do not contain X_0. Equality (b) holds because the information term is the same across the sum and can be factorized. Inequality (c) is justified as follows. Let N_0 be the number of groups that contain X_0, and hence that intersect in X_0. For these groups, $d_i(G'; \mathbf{Z})$ is higher than or equal to N_0. This means that $\sum_{i:X_0 \in \tilde{an}(\mathbf{A}_i)} 1/d_i(G'; \mathbf{Z}) \leqslant 1$. We now complete the proof of the last iterative induction step:

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \overset{(a)}{=} \sum_{\mathbf{A}_i \in \mathbf{S_W}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \tilde{an}(\mathbf{A}_{[n]})\backslash X_0) \overset{(b)}{=} \\
&\sum_{\mathbf{A}_i \in \mathbf{S_W}} \frac{1}{d_i} \Big[I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i), \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) - I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) \Big] \overset{(c)}{\geqslant} \\
&\left[\sum_{\mathbf{A}_i \in \mathbf{S_W}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i), \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) \right] - I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) \overset{(d)}{\geqslant} \\
&\left[\sum_{\mathbf{A}_i \in \mathbf{S_W}} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \mathbf{V}^{[m']}) \right] - I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}) \overset{(e)}{=} \\
&\left[\sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \mathbf{V}^{[m']}) \right] - I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})\backslash X_0|\mathbf{Z}, \mathbf{V}^{[m']}).
\end{aligned} \tag{A21}$$

Equality (a) holds because X_0 is selected to be contained only in groups in $\mathbf{S_W}$. Equality (b) follows from the chain rule and from $\mathbf{V}^{[m']} \subseteq \tilde{an}(\mathbf{A}_{[n]})\backslash X_0$. Inequality (c) holds because, for all $\mathbf{A}_i \in \mathbf{S_W}$, $d_i(G'; \mathbf{Z})$ is higher than or equal to $|\mathbf{S_W}|$, since their ancestral sets intersect at W_0, which is an ancestor of all variables in $\mathbf{W}$. This means that $\sum_{\mathbf{A}_i \in \mathbf{S_W}} 1/d_i(G'; \mathbf{Z}) \leqslant 1$. Inequality (d) follows from the monotonicity of mutual information, and equality (e) holds because $\tilde{an}(\mathbf{A}_i) \subseteq \{\mathbf{Z}, \mathbf{V}^{[m']}\}$ for all $\mathbf{A}_i \notin \mathbf{S_W}$. We use the last expression in Equation (A21) at the r.h.s of Equation (A19), and combine it with Equation (A18) to obtain

$$I(\mathbf{Y}; \tilde{an}(\mathbf{A}_{[n]})|\mathbf{Z}, \mathbf{V}^{[m']}) \geqslant \sum_{i=1}^{n} \frac{1}{d_i} I(\mathbf{Y}; \tilde{an}(\mathbf{A}_i)|\mathbf{Z}, \mathbf{V}^{[m']}). \tag{A22}$$

This completes the iterative induction step of the proof. Since the validity of the base case has also been proven, this completes the proof. □

Appendix C. On Required Assumptions Relating Independencies and d-Separation

In this Section, we discuss more closely the requirements on the relation between graphical d-separation and statistical independencies needed for the applicability of the derived inequality constraints. As indicated in Section 2.3, so far we have invoked the faithfulness assumption [1,2] in order to simplify the presentation, that is, we have not distinguished between $X \perp_P Y|\mathbf{S}$ and $X \perp_G Y|\mathbf{S}$. We will now make this distinction and reconsider all cases of the proofs of Appendices A and B where faithfulness has been invoked, showing that in fact it is only required to assume that d-separation is a sufficient condition for statistical independence.

We start with the role that the assumption of d-separation implying independence has in the proof of Propositions 1 and 3. As discussed in Section 1, we envisage the implementation of the tests such that conditional independence requirements of Proposition 1 or 3 are verified in terms of graphical separability for the hypothesized causal structure. In particular, a test from Proposition 1 is to be applied when verifying that for the selected collection and groups it holds that $\mathbf{A}_i \perp_G \mathbf{A}_j\backslash\mathbf{A}_i|\mathbf{Z}\ \forall i, j$. It is then assumed that this implies $\mathbf{A}_i \perp_P \mathbf{A}_j\backslash\mathbf{A}_i|\mathbf{Z}\ \forall i, j$. In the proof of Proposition 1, in step (e) of Equation (A1), Lemma 1(ii) has been applied invoking faithfulness to guarantee that for $X_k \in \mathbf{A}_i$, independencies $X_k \perp_P (\mathbf{X}_{[k-1]} \cap \mathbf{A}_j)\backslash\mathbf{A}_i|(\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}, \forall j \neq i$ imply the independence $X_k \perp_P \mathbf{X}_{[k-1]}\backslash\mathbf{A}_i|(\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$. However, while this implication needs to be assumed at the level of independencies, at the level of graphical separability,

$X_k \perp_G (\mathbf{X}_{[k-1]} \cap \mathbf{A}_j) \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$, $\forall j \neq i$ straightforwardly implies the joint separability $X_k \perp_G \mathbf{X}_{[k-1]} \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$. This is because the separability of $\mathbf{X}_{[k-1]} \backslash \mathbf{A}_i$ follows from the lack of active paths for each of the nodes it contains, and hence is equivalent to the separability of $(\mathbf{X}_{[k-1]} \cap \mathbf{A}_j) \backslash \mathbf{A}_i$ for all j, which jointly comprise the same nodes. The assumption that d-separation implies independence guarantees the independence $X_k \perp_P \mathbf{X}_{[k-1]} \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$ from $X_k \perp_G \mathbf{X}_{[k-1]} \backslash \mathbf{A}_i | (\mathbf{X}_{[k-1]} \cap \mathbf{A}_i), \mathbf{Z}$, without the need to more broadly require faithfulness. The proof of Proposition 3 relies on an analogous way on the assumption that d-separation implies independence, using it to guarantee the conditional independencies involving the subsets in $\mathbf{B}^{(1)}_{[n]}$ and $\mathbf{B}^{(2)}_{[n]}$. In step (d) of Equation (A2), the fact that separability for a joint set of nodes is straightforwardly guaranteed by the separability of each of its nodes is again applied and then mapped to the existence of an independence using this assumption. The fact that conditions $\mathbf{B}^{(1)}_i \perp_G \mathbf{B}^{(1)}_j \backslash \mathbf{B}^{(1)}_i | \mathbf{Z}$ and $\mathbf{B}^{(2)}_i \perp_G \mathbf{B}_j \backslash \mathbf{B}^{(2)}_i | \mathbf{B}^{(1)}_i \mathbf{Z}$ $\forall i, j$ can be verified using d-separation instead of estimating independencies from data is crucial in the case that the groups include hidden variables, which precludes the direct evaluation of these independencies.

The next result whose derivation relies on the assumption that d-separation implies independence is Theorem 2. In step (a) of Equation (A7), faithfulness was invoked to guarantee that conditioning on some ancestors of $\mathbf{Z}$ cannot create new dependencies that were not already created by conditioning on $\mathbf{Z}$ itself. In more detail, it was assumed that if the independence $\tilde{an}(\mathbf{A}_i) \perp_P \mathbf{V}^{(j)}_Z | \mathbf{Z}$ holds then also $\tilde{an}(\mathbf{A}_i) \perp_P \mathbf{V}^{(j)}_Z | \{\mathbf{Z}, \mathbf{V}^{[j-1]}_Z\}$ holds, where $\mathbf{V}^{[j-1]}_Z$ are by construction ancestors of $\mathbf{Z}$. Again, at the level of graphical separability this implication is straightforward and does not require any assumption. This is because by definition of d-separation a path is activated both when conditioning on a collider or on any descendant of the collider, and $\mathbf{V}^{[j-1]}_Z$ being ancestors of $\mathbf{Z}$ means that $\mathbf{Z}$ contains a descendant for each node in $\mathbf{V}^{[j-1]}_Z$. Accordingly, no assumption is needed to ensure $\tilde{an}(\mathbf{A}_i) \perp_G \mathbf{V}^{(j)}_Z | \{\mathbf{Z}, \mathbf{V}^{[j-1]}_Z\}$ from $\tilde{an}(\mathbf{A}_i) \perp_G \mathbf{V}^{(j)}_Z | \mathbf{Z}$. The assumption that d-separation implies independence is then used to ensure $\tilde{an}(\mathbf{A}_i) \perp_P \mathbf{V}^{(j)}_Z | \{\mathbf{Z}, \mathbf{V}^{[j-1]}_Z\}$ from $\tilde{an}(\mathbf{A}_i) \perp_G \mathbf{V}^{(j)}_Z | \{\mathbf{Z}, \mathbf{V}^{[j-1]}_Z\}$. Faithfulness is also invoked in the proof of Theorem 2 to justify the application of Lemma 1(ii) in Equation (A15). In this case, the existence of an independence $\mathbf{A} \perp_P \mathbf{C} | \mathbf{B}$ is directly justified in terms of the nonexistence of active paths in the graph, hence guaranteeing $\mathbf{A} \perp_G \mathbf{C} | \mathbf{B}$ and subsequently using the assumption that d-separation implies independence to derive $\mathbf{A} \perp_P \mathbf{C} | \mathbf{B}$.

The considerations above show that the assumption that d-separation implies statistical independence is enough to derive the existence of groups-decomposition inequalities under the conditions of Propositions 1 and 3, and of Theorem 2. Furthermore, if unfaithful independencies are present in the data that do not follow from the causal structure, this may decrease the power to reject causal structures testing the inequalities, but will not lead to incorrect rejections. This differs from the impact of unfaithful independencies on the inference of the Markov equivalence class from data [1,2]. In that case, unfaithful independencies can lead to an incorrect reconstruction of the skeleton of the graph or result in contradictory rules for edge orientation. The assumption that d-separation implies statistical independence is substantially weaker than the reverse assumption also included in the faithfulness assumption, namely that statistical independence implies d-separation. The X-OR logic gate is an example that the latter assumption can be violated. Conversely, if the causal graph is meant to reflect the underlying structure of actual physical mechanisms involved in generating the variables, all statistical dependencies need to originate from some paths of influence between the variables. Accordingly, a d-separation that does not lead to an independence can be taken as an indicator that some structure is missing in the causal graph, namely associated with the paths that create the observed dependence. In this regard, it is appropriate to reject a causal structure if it does not fulfill an inequality constraint because graphical separability is not reflected in the corresponding independencies found in the data.

References

1. Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, Prediction, and Search*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2000.
2. Pearl, J. *Causality: Models, Reasoning, Inference*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2009.
3. Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
4. Malinsky, D.; Danks, D. Causal discovery algorithms: A practical guide. *Philos. Compass* **2018**, *13*, e12470. [CrossRef]
5. Verma, T. *Graphical Aspects of Causal Models*; Technical Report R-191; Computer Science Department, UCLA: Los Angeles, CA, USA, 1993.
6. Zhang, J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* **2008**, *172*, 1873–1896. [CrossRef]
7. Tian, J.; Pearl, J. On the testable implications of causal models with hidden variables. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, 1–4 August 2002.
8. Verma, T.; Pearl, J. Equivalence and synthesis of causal models. In Proceedings of the Sixth Conference on Uncertainty in Artifial Intelligence, Cambridge, MA, USA, 27–29 July 1990; pp. 220–227.
9. Chicharro, D.; Besserve, M.; Panzeri, S. Causal learning with sufficient statistics: An information bottleneck approach. *arXiv* **2020**, arXiv:2010.05375.
10. Parbhoo, S.; Wieser, M.; Wieczorek, A.; Roth, V. Information bottleneck for estimating treatment effects with systematically missing covariates. *Entropy* **2020**, *22*, 389. [CrossRef]
11. Hoyer, P.O.; Janzing, D.; Mooij, J.M.; Peters, J.; Schölkopf, B. Nonlinear causal discovery with additive noise models. In Proceedings of the 21st Conference on Advances in Neural Information Processing Systems (NIPS 2008), Vancouver, BC, Canada, 8–11 December 2008; pp. 689–696.
12. Zhang, K.; Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, QC, Canada, 18–21 June 2009; pp. 647–655.
13. Chicharro, D.; Panzeri, S.; Shpitser, I. Conditionally-additive-noise models for structure learning. *arXiv* **2019**, arXiv:1905.08360
14. Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P.O.; Bollen, K. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.* **2011**, *12*, 1225–1248.
15. Evans, R.J. Graphs for margins of Bayesian networks. *Scand. J. Stat.* **2015**, *43*, 625. [CrossRef]
16. Weilenmann, M.; Colbeck, R. Analysing causal structures with entropy. *Proc. Roy. Soc. A* **2017**, *473*, 20170483. [CrossRef]
17. Bell, J.S. On the Einstein-Podolsky-Rosen paradox. *Physics* **1964**, *1*, 195–200. [CrossRef]
18. Clauser, J.F.; Horne, M.A.; Shimony, A.; Holt, R.A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **1969**, *23*, 880. [CrossRef]
19. Pearl, J. On the testability of causal models with latent and instrumental variables. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–20 August 1995; pp. 435–443.
20. Bonet, B. Instrumentality tests revisited. In Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI), San Francisco, CA, USA, 2–5 August 2001; pp. 48–55.
21. Kang, C.; Tian, J. Inequality constraints in causal models with hidden variables. In Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 13–16 July 2006; pp. 233–240.
22. Chaves, R.; Luft, L.; Gross, D. Causal structures from entropic information: Geometry and novel scenarios. *New J. Phys.* **2014**, *16*, 043001. [CrossRef]
23. Fritz, T.; Chaves, R. Entropic inequalities and marginal problems. *IEEE Trans. Inf. Theory* **2013**, *59*, 803–817. [CrossRef]
24. Chaves, R.; Luft, L.; Maciel, T.O.; Gross, D.; Janzing, D.; Schölkopf, B. Inferring latent structures via information inequalities. In Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence, Quebec City, QC, Canada, 23–27 July 2014; pp. 112–121.
25. Dougherty, R.; Freiling, C.; Zeger, K. Six new non-Shannon information inequalities. In Proceedings of the IEEE International Symposium on Information Theory, Seattle, WA, USA, 9–14 July 2006; pp. 233–236.
26. Weilenmann, M.; Colbeck, R. Non-Shannon inequalities in the entropy vector approach to causal structures. *Quantum* **2018**, *2*, 57. [CrossRef]
27. Steudel, B.; Ay, N. Information-theoretic inference of common ancestors. *Entropy* **2015**, *17*, 2304–2327. [CrossRef]
28. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2006.
29. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
30. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
31. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130. [CrossRef] [PubMed]
32. Ince, R.A.A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318. [CrossRef]
33. James, R.G.; Emenheiser, J.; Crutchfield, J.P. Unique Information via dependency constraints. *J. Phys. A Math. Theor.* **2019**, *52*, 014002. [CrossRef]
34. Ay, N.; Polani, D.; Virgo, N. Information decomposition based on cooperative game theory. *Kybernetika* **2020**, *56*, 979–1014. [CrossRef]
35. Kolchinsky, A. A novel approach to the partial information decomposition. *Entropy* **2022**, *24*, 403. [CrossRef]

36. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]
37. Geiger, D.; Verma, T.; Pearl, J. d-Separation: From theorems to algorithms. In Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, Amsterdam, The Netherlands, 18–20 August 1989; pp. 118–125.
38. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of the IEEE International Symposium on Information Theory (ISIT 2014), Honolulu, HI, USA, 29 June–4 July 2014; pp. 2232–2236.
39. Banerjee, P.K.; Olbrich, E.; Jost, J.; Rauh, J. Unique Informations and Deficiencies. *arXiv* **2019**, arXiv:1807.05103v3.
40. Chicharro, D.; Panzeri, S. Synergy and redundancy in dual decompositions of mutual information gain and information loss. *Entropy* **2017**, *19*, 71. [CrossRef]
41. Chicharro, D. Quantifying multivariate redundancy with maximum entropy decompositions of mutual information. *arXiv* **2017**, arXiv:1708.03845.
42. Pica, G.; Piasini, E.; Chicharro, D.; Panzeri, S. Invariant components of synergy, redundancy, and unique information among three variables. *Entropy* **2017**, *19*, 451. [CrossRef]
43. Chicharro, D.; Pica, G.; Panzeri, S. The identity of information: How deterministic dependencies constrain information synergy and redundancy. *Entropy* **2018**, *20*, 169. [CrossRef]
44. Chicharro, D.; Ledberg, A. Framework to study dynamic dependencies in networks of interacting processes. *Phys. Rev. E* **2012**, *86*, 041901. [CrossRef]
45. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin/Heidelberg, Germany, 2006.
46. Geweke, J.F. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.* **1982**, *77*, 304–313. [CrossRef]
47. Chicharro, D. On the spectral formulation of Granger causality. *Biol. Cybern.* **2011**, *105*, 331–347. [CrossRef]
48. Chicharro, D. Parametric and non-parametric criteria for causal inference from time-series. In *Directed Information Measures in Neuroscience*; Wibral, M., Vicente, R. Lizier, J.T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 195–223.
49. Brovelli, A.; Ding, M.; Ledberg, A.; Chen, Y.; Nakamura, R.; Bressler, S.L. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9849–9854. [CrossRef] [PubMed]
50. Brovelli, A.; Chicharro, D.; Badier, J.M.; Wang, H.; Jirsa, V. Characterization of cortical networks and corticocortical functional connectivity mediating arbitrary visuomotor mapping. *J. Neurosci.* **2015**, *35*, 12643–12658. [CrossRef] [PubMed]
51. Celotto, M.; Bím, J.; Tlaie, A.; De Feo, V.; Toso, A.; Lemke, S.M.; Chicharro, D.; Nili, H.; Bieler, M.; Hanganu-Opatz, I.L.; et al. An information-theoretic quantification of the content of communication between brain regions. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
52. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
53. Hiemstra, C.; Jones, J.D. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *J. Financ.* **1994**, *49*, 1639–1664.
54. Hlaváčková-Schindler, K.; Paluš, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time-series analysis. *Phys. Rep.* **2007**, *441*, 1–46. [CrossRef]
55. Geweke, J.F. Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.* **1984**, *79*, 907–915. [CrossRef]
56. Caporale, M.C.; Hassapis, C.; Pittis, N. Unit roots and long-run causality: Investigating the relationship between output, money and interest rates. *Econ. Model.* **1998**, *15*, 91–112. [CrossRef]
57. Caporale, M.C.; Pittis, N. Efficient estimation of cointegrating vectors and testing for causality in vector auto-regressions. *J. Econ. Surv.* **1999**, *13*, 3–35.
58. Hacker, R.S.; Hatemi, J.A. Tests for causality between integrated variables using asymptotic and bootstrap distributions: Theory and application. *Appl. Econ.* **2006**, *38*, 1489–1500. [CrossRef]
59. Massey, J.L. Causality, feedback and directed information. In Proceedings of the 1990 IEEE International Symposium Information Theory and Its Applications, Honolulu, HI, USA, 10–15 June 1990; Volume 27, pp. 303–305.
60. Amblard, P.O.; Michel, O. On directed information theory and Granger causality graphs. *J. Comput. Neurosci.* **2011**, *30*, 7–16. [CrossRef]
61. Chaves, R.; Majenz, C.; Gross, D. Information-theoretic implications of quantum causal structures. *Nat. Commun.* **2015**, *6*, 5766. [CrossRef]
62. Wolfe, E.; Schmid, D.; Sainz, A.B.; Kunjwal, R.; Spekkens, R.W. Quantifying Bell: The resource theory of nonclassicality of common-cause boxes. *Quantum* **2020**, *4*, 280. [CrossRef]
63. Tavakoli, A.; Pozas-Kerstjens, A.; Luo, M.; Renou, M.O. Bell nonlocality in networks. *Rep. Prog. Phys.* **2022**, *85*, 056001. [CrossRef] [PubMed]
64. Henson, J.; Lal, R.; Pusey, M.F. Theory-independent limits on correlations from generalized Bayesian networks. *New J. Phys.* **2014**, *16*, 113043. [CrossRef]
65. Wood, C.J.; Spekkens, R.W. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New J. Phys.* **2015**, *17*, 033002. [CrossRef]

66. Wolfe, E.; Spekkens, R.W.; Fritz, T. The Inflation Technique for causal inference with latent variables. *J. Caus. Inf.* **2019**, *7*, 20170020. [CrossRef]
67. Navascués, M.; Wolfe, E. The Inflation Technique completely solves the causal compatibility problem. *J. Causal Infer.* **2020**, *8*, 70–91. [CrossRef]
68. Boghiu, E.C.; Wolfe, E.; Pozas-Kerstjens, A. Inflation: A Python library for classical and quantum causal compatibility. *Quantum* **2023**, *7*, 996. [CrossRef]
69. Evans, R.J. Graphical methods for inequality constraints in marginalized DAGs. In Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Santander, Spain, 23–26 September 2012; pp. 1–6.
70. Fraser, T.C. A combinatorial solution to causal compatibility. *J. Causal Inference* **2020**, *8*, 22. [CrossRef]
71. Finkelstein, N.; Zjawin, B.; Wolfe, E.; Shpitser, I.; Spekkens, R.W. Entropic inequality constraints from e-separation relations in directed acyclic graphs with hidden variables. In Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, Online, 27–29 July 2021; pp. 1045–1055.
72. Evans, R.J. Latent-free equivalent mDAGs. *Algebr. Stat.* **2023**, *14*, 3–16. [CrossRef]
73. Khanna, S.; Ansanelli, M.M.; Pusey, M.F.; Wolfe, E. Classifying causal structures: Ascertaining when classical correlations are constrained by inequalities. *Phys. Rev. Res.* **2024**, *6*, 023038. [CrossRef]
74. Rodari, G.; Poderini, D.; Polino, E.; Suprano, A.; Sciarrino, F.; Chaves, R. Characterizing hybrid causal structures with the exclusivity graph approach. *arXiv* **2023**, arXiv:2401.00063.
75. Treves, A.; Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Comput.* **1995**, *7*, 399–407. [CrossRef]
76. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *17*, 1191–1253. [CrossRef]
77. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, R.D. Mutual Information Neural Estimation. In Proceedings of the Thirty-Fifth International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
78. Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A.A.; Tucker, G. On Variational Bounds of Mutual Information. In Proceedings of the Thirty-Sixth International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 5171–5180.