entropy

MDPI

Article

Causally Informative Entropic Inequalities within Families of Distributions with Shared Marginals

Daniel Chicharro

Department of Computer Science, City St George's, University of London, Northampton Square, London EC1V 0HB, UK; chicharro31@yahoo.es

Abstract

The joint probability distribution of observable variables from a system is constrained by the underlying causal structure. In the presence of hidden variables, untestable independencies that involve hidden variables lead to testable causally-imposed inequality constraints for observable variables, whose violation can reject the compatibility of a causal structure with data. One type of causally informative inequalities is entropic inequalities, which appear in the space of entropic terms associated with the distribution of observable variables. We derive a new type of minimum information (minInf) entropic inequalities that substantially increases causal inference power. These new entropic inequalities appear when considering the constraints that the causal structure imposes on entropic terms determined by information minimization within families of distributions that preserve sets of marginals shared with the original distribution. We introduce a new family of minInf data processing inequalities and a procedure to recursively combine different types of data processing inequalities to create tighter testable entropic inequalities. We extensively illustrate the applicability of this procedure in the instrumental causal scenario, integrating the new inequalities with standard instrumental entropic inequalities constructed with multivariate instrumental sets. We also provide additional examples with other types of entropic inequalities, such as the Information Causality and Groups-Decomposition inequalities.

Keywords: causality; directed acyclic graphs; causal discovery; structure learning; marginal scenarios; hidden variables; mutual information; entropic inequalities; data processing inequalities; maximum entropy; minimum information; instrumental inequality; shannon entropy cone; information causality

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

Check for updates

Academic Editor: Xiaoguang Gao

Received: 12 January 2026
Revised: 7 April 2026
Accepted: 13 April 2026
Published: 20 April 2026

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Understanding which causal structures are compatible with a set of observational data is a common question in science. The underlying causal structure of a system creates constraints on the probability distribution of variables generated from it [1–3], which helps to reversely infer which causal structures are compatible with the data. Causal learning algorithms based on conditional independencies [1,2,4] reconstruct a partially oriented graph [5] 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, in most real-world scenarios, the components of a system are only partially observed, and the presence of hidden variables

creates dependencies among the observable variables that limit the degree to which Markov equivalence classes narrow down the set of causal structures compatible with the data.

Beyond statistical independencies in the joint distribution, the underlying causal structure can also be reflected in other equality constraints imposed to the observable variables. These constraints comprise functional equality constraints [6,7] and independencies that originate from further assumptions about the functional form of the generative mechanisms of the variables [3,8–13]. Additionally, nonverifiable conditional independencies that involve hidden variables can manifest themselves through inequality constraints that only involve observable variables [14–16]. Unlike equality constraints, inequality constraints provide necessary but not sufficient conditions for the compatibility of data with a certain causal structure. Data violations of inequalities enforced by a causal structure allow discarding that causal structure as the one generating the data. Accordingly, causal inference power is increased when deriving tighter inequalities. Causally informative inequalities comprise inequalities derived in the probability space, such as Bell-type inequalities [17,18], instrumental inequalities [19,20], and interventional inequalities [21], as well as entropic inequalities derived in the space of the entropic terms associated with the observable variables [14,22].

Causally informative entropic inequalities can be derived with two alternative approaches. One approach is to derive a specific entropic inequality departing from a concrete equality that involves hidden variables and then using the conditional independencies associated with the causal structure to derive for the two sides of the equality upper and lower bounds which do not contain hidden variables [23–26]. In order to derive a testable inequality that only contains observable variables, this approach relies on the data processing (DP) inequality [27] to replace hidden variables by less informative observable variables. In a second approach, all the testable causally informative inequalities imposed by a causal structure are derived reducing the set of equalities and inequalities that characterize the whole system –comprising observable and hidden variables– to the marginal scenario associated with the observable variables [14,24]. This approach combines the inequalities that define the Shannon entropic cone [28], i.e., associated with the nonnegativity, monotonicity, and submodularity properties of entropy, and all additional independence constraints related to all the variables in the causal structure. Subsequently, variable elimination is performed to extract the resulting constraints that only involve observable variables. While this marginalization problem is algorithmically solvable [14], its implementation for large systems is challenging and furthermore does not provide an explicit constructive recipe that allows tracing the resulting inequalities in terms of the existing conditional independencies.

These methods to construct causally informative entropic inequalities traditionally rely on entropic terms associated with the original joint distribution of the observable variables. However, specifically for the so-called Groups-Decomposition inequalities [25,26], it has been shown [26] that new as well as tighter more informative inequalities can be derived if the entropic terms of the original joint distribution are combined with maximum entropy entropic terms. Chicharro and Nguyen [26] introduced a DP inequality for the maximum entropy measure of *unique information* [29], a measure originally proposed to decompose mutual information into redundant, unique, and synergistic components [30]. The maximum entropy unique information is defined by an information minimization within a family of distributions constrained to share some marginals of the original distribution of the observable variables. In this work, we capitalize on an extended combination of the original entropic terms and additional entropic terms defined by information minimization within a broader set of different families. We introduce novel procedures to obtain new and tighter entropic inequalities incorporating these additional entropic terms.

To motivate our derivations, we first proceed with the first approach to derive causally informative entropic inequalities. That is, we focus on concrete causal structures and apply an explicit procedure that involves DP inequalities. Specifically, we focus on instrumental entropic inequalities [24], which appear in the causal scenario of instrumental variables [19]. We derive new instrumental entropic inequalities comprising maximum entropy unique information terms. This allows us to characterize a procedure to recursively combine different DP inequalities to create tighter instrumental entropic inequalities. We then introduce a much wider family of DP inequalities for information terms determined by constrained minimization. These minimum information (minInf) terms are defined within families of distributions that share sets of marginals of the original joint distribution. MinInf DP inequalities are then used to derive tighter instrumental inequalities. Subsequently, we indicate how new entropic inequalities can be derived not only from the sequential application of new minInf DP inequalities, but also as a marginalization problem.

Overall, the minInf DP inequalities that we develop, and the procedure to sequentially combine them, provide a general tool to derive new types of entropic inequalities and to extend existing ones thanks to the incorporation of additional information terms to obtain tighter lower bounds. To illustrate the generality of these tools, we finally also examine how other well-known types of causally informative entropic inequalities [23,25,26] can equally be extended into inequalities with an increased causal inference power. While entropic inequalities have also been formulated for quantum systems [23,31], in this work we restrict our derivation to classical Shannon entropy measures. The Discussion section comments on potential extensions.

This paper is organized as follows. In Section 2, we review existing results relevant for our work. In Section 3.1, we derive instrumental entropic inequalities with unique information terms. In Section 3.2, we compare the new inequalities to standard instrumental inequalities with multivariate instrumental sets, identifying conditions in which the new inequalities provide additional causal inference power. In Section 3.3, we show that causal inference power is increased not only using the DP inequality of unique information instead of the standard DP inequality, but also iteratively combining them. In this way, we identify a procedure to iteratively combine multiple DP inequalities. This procedure is further developed in Section 3.4, where we introduce a general type of DP inequalities for minInf information terms and combine them recursively to add observable minInf information terms as lower bounds of information terms with hidden variables. In Section 3.5, we apply this procedure specifically to build more causally informative instrumental entropic inequalities. Section 3.6 reframes the use of minInf information terms for causal learning with the optics of a marginalization problem. We indicate how to extend the Shannon entropy cone to minInf Shannon entropy cones that jointly characterize minInf families. Marginalization of the hidden variables within this joint space produces also the causally informative entropic inequalities that contain minInf entropic terms. Finally, to provide broader examples of applicability of our methods, in Section 3.7 we show how to extend two other types of causally informative entropic inequalities, namely Groups-Decompositions inequalities [25,26] and the Information Causality inequality [23,31].

2. Methods

In this section we review the relation between causal graphs and conditional independencies, the standard data processing inequality, the standard instrumental entropic inequality, as well as the formulation of minimum mutual information quantities, comprising a measure of maximum entropy unique information.

2.1. Causal Graphs and Conditional Independencies

We review Directed Acyclic Graphs (DAGs) and the relation between causal structures and dependencies. A DAG $G = (\bar{\mathbf{V}}; \mathcal{E})$ associated with a set of random variables $\bar{\mathbf{V}} = \{\bar{V}_1, \ldots, \bar{V}_n\}$ consists of nodes $\bar{\mathbf{V}}$ and edges $\mathcal{E}$ between the nodes, where $\bar{V}$ refers both to a variable and its corresponding node. Note that in general $\bar{\mathbf{V}}$ can comprise both observable and hidden variables; we will later use specifically letter U for hidden variables. The set of edges $\mathcal{E}$ contains $(\bar{V}_i; \bar{V}_j) \in \mathcal{E}$ for each arrow $\bar{V}_i \to \bar{V}_j$, which indicates a causal connection in the system generating the variables. The structure of edges in the graph removing arrowheads is called the skeleton of the graph. The graph is acyclic because we consider causal mechanisms not to be instantaneous and any causal cycle spans in time.

A path in G is a sequence of (at least two) distinct nodes $\bar{V}_1, \ldots, \bar{V}_m$, such that there is an edge between $\bar{V}_k$ and $\bar{V}_{k+1}$ for all $k = 1, \ldots, m-1$. If all edges are directed as $\bar{V}_k \to \bar{V}_{k+1}$ the path is a causal or directed path. A node $\bar{V}_i$ is a collider in a path if it has incoming arrows $\bar{V}_{i-1} \to \bar{V}_i \leftarrow \bar{V}_{i+1}$ and is a noncollider otherwise. If there is an arrow $\bar{V}_i \to \bar{V}_j$, then $\bar{V}_i$ is a parent of $\bar{V}_j$, and $\bar{V}_j$ is a child of $\bar{V}_i$. A node $\bar{V}_i$ is called an ancestor of $\bar{V}_j$ if there is a directed path from $\bar{V}_i$ to $\bar{V}_j$. Conversely, in this case $\bar{V}_j$ is a descendant of $\bar{V}_i$. We use bidirected arcs $\bar{V}_i \leftrightarrow \bar{V}_j$ to indicate the presence of a nondirected path between $\bar{V}_i$ and $\bar{V}_j$ consisting only of hidden noncolliders.

A causal graph accurately represents the generative mechanisms of a system when a variable $\bar{V}_i$ is a parent of another variable $\bar{V}_j$ if and only if it is an argument of an underlying functional equation that captures the mechanisms that generate $\bar{V}_j$. This creates a relation between the conditional independencies that hold between variables in the system and a graphical criterion of separability between the nodes, called *d-separation* [32]. The criterion of d-separation states that 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 $p(\bar{\mathbf{V}})$ 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 X and Y are d-separated given $\mathbf{S}$ –denoted by $X \perp_G Y|\mathbf{S}$.

The inference of the causal structure of a system from data generated from the system relies on this link between the causal structure and independencies. Causal learning algorithms that use conditional independencies to reconstruct a partially oriented graph [1,2,4] rely on the assumption of faithfulness in order to determine the skeleton of the graph and to apply rules of orientation of the edges. On the other hand, in the case that causally informative inequalities are used to rule out causal structures [14,24,26], it is only required to assume the substantially weaker assumption that d-separability implies conditional independence. Under this assumption, if a causal structure implies the set of independencies that lead to the fulfillment of the inequality, its violation allows discarding that causal structure. The assumption of faithfulness is not required because if unfaithful independencies are present in the data, which do not follow from the causal structure, this may decrease the power to reject causal structures, but does not lead to incorrect rejections.

Note that the assumption that graphical separability implies statistical conditional independence is substantially weaker than the converse assumption that statistical conditional independence implies graphical separability. A counterexample of the latter is the X-OR logical gate. On the other hand, if the causal graph reflects the underlying structure of mechanisms involved in generating the variables, all statistical dependencies need to originate from some paths of influence between the variables. If some variables are conditionally dependent while the graph indicates that they are d-separated, then the graph must be misrepresenting the paths that create the observed dependence.

2.2. The Data Processing Inequality

The data processing inequality (DP inequality) of mutual information indicates that information cannot be increased in a Markov chain [27].

Lemma 1 (Data processing inequality of conditional mutual information)**.** *Let* $\bar{\mathbf{Z}}$, $\mathbf{D}$, $\mathbf{D}'$, *and* $\mathbf{E}$ *be four nonoverlapping sets of variables. If* $\bar{\mathbf{Z}} \perp \mathbf{D}'|\mathbf{D}\mathbf{E}$, *then it follows that* $I(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D}'|\mathbf{E}) = I(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{E}) \geqslant I(\bar{\mathbf{Z}}; \mathbf{D}'|\mathbf{E})$.

While the DP inequality is often used only to refer to the inequality between the information carried by $\mathbf{D}$ and $\mathbf{D}'$, we will also apply the equality of the information carried by $\{\mathbf{D}, \mathbf{D}'\}$ and $\mathbf{D}$ alone.

2.3. The Instrumental Entropic Inequality

We here revise the instrumental entropic inequality [24]. We provide its full derivation because this helps to identify how new entropic inequalities can be derived. Consider the causal structures of Figure 1A, with all variables observable except U hidden. We use a notation of the variables consistent with the role of Z, X, Y, and U in the work that introduced the instrumental inequality [19] and its entropic formulation [24]. The diagram represents several causal structures, depending on how the dashed edges are instantiated (or removed) as additional causal connections. These additional connections are constrained by the acyclic nature of the causal graph, for example $Z \to Y$ or $Z \leftrightarrow Y$ are valid, while $Z \leftarrow Y$ is not, since it would lead to the existence of a cycle. For all the causal structures of Figure 1A, no conditional independencies between variables in $\{X, Y, Z\}$ exist that involve conditioning only on observable variables. Accordingly, the reconstruction of the causal structure based on conditional independencies, for example using the PC algorithm of Spirtes et al. [1], results in all cases in a reconstructed graph in which nodes $\{X, Y, Z\}$ are all connected. In particular, also for the causal structure in which the dashed edges are removed, a reconstructed edge $Z - Y$ is obtained, even if not present in the actual skeleton of that causal structure. This is due to the fact that blocking the path $Z \to X \to Y$ by conditioning on X activates the path $Z \to X \leftarrow U \to Y$.

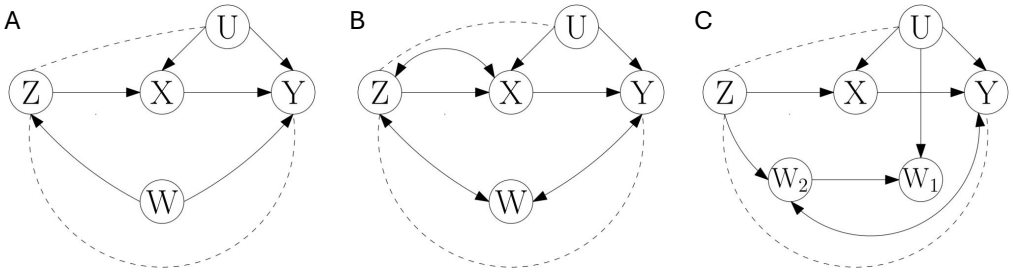


Figure 1. Examples of causal structures within the instrumental inequality scenario. All variables are observable except U hidden. Each graph represents several causal structures, corresponding to instantiations of the dashed edges under the constraints of acyclic graphs. Instantiations comprise arrows in one or the other direction, a bidirectional arc indicating the presence of a hidden parent, a combination of an arrow and bidirectional arc, or the removal of the dashed edge, corresponding to the lack of a direct connection. (**A**) Example of the standard instrumental entropic inequality (Section 2.3). (**B**) Example of the instrumental entropic inequality with a term of unique information (Section 3.1). (**C**) Example in which an instrumental entropic inequality with unique information provides causal inference power additional to the one obtainable from standard instrumental entropic inequalities (Section 3.2).

The instrumental entropic inequality provides a causally informative test to reject the compatibility of a data set with this causal structure in which the dashed edges are not

present. Even if no independencies between $\{Z, X, Y\}$ exist in the marginal scenario in which U is hidden, this causal structure contains untestable independencies that involve the hidden variable U, comprising $Z \perp U|W$ and $Z \perp Y|UXW$. These untestable independencies impose additional constraints that manifest themselves in an inequality between information terms, namely the instrumental entropic inequality. For later convenience, we formulate the standard instrumental entropic inequality allowing for a multivariate $\mathbf{Z}$:

Proposition 1 (Instrumental entropic inequality)**.** *Consider the variables* $\mathbf{Z}$*,* X*,* Y*,* $\mathbf{B}_0$*, and* U*, all observable except* U *a hidden variable. Consider that the causal structure is such that, for all* $Z_i \in \mathbf{Z}$*, no pair from* $\{Z_i, X, Y\}$ *is separable given that* U *is hidden. Consider that the causal structure imposes the existence of the nontestable independencies* $\mathbf{Z} \perp U|\mathbf{B}_0$ *and* $\mathbf{Z} \perp Y|UX\mathbf{B}_0$*. These independencies result in the testable inequality*

$$H(X|\mathbf{B}_0) \geqslant I(\mathbf{Z}; X|\mathbf{B}_0) + I(\mathbf{Z}; Y|\mathbf{B}_0, X). \tag{1}$$

Proof. The mutual information $I(\mathbf{Z}; U, X|\mathbf{B}_0)$ can be decomposed applying the chain rule in two alternative orders. If information with U is considered first

$$I(\mathbf{Z}; U, X|\mathbf{B}_0) \overset{(a)}{=} I(\mathbf{Z}; U|\mathbf{B}_0) + I(\mathbf{Z}; X|\mathbf{B}_0, U) \overset{(b)}{\leqslant} H(X|\mathbf{B}_0). \tag{2}$$

Equality (a) applies the chain rule of mutual information. Inequality (b) holds because $I(\mathbf{Z}; U|\mathbf{B}_0) = 0$, given the independence $\mathbf{Z} \perp U|\mathbf{B}_0$, and by definition $I(\mathbf{Z}; X|\mathbf{B}_0, U)$ is smaller than or equal to $H(X|\mathbf{B}_0, U)$, which by monotonicity of entropy under conditioning is smaller than or equal to $H(X|\mathbf{B}_0)$. Considering now the chain rule with X first,

$$I(\mathbf{Z}; U, X|\mathbf{B}_0) \overset{(a)}{=} I(\mathbf{Z}; X|\mathbf{B}_0) + I(\mathbf{Z}; U|\mathbf{B}_0, X) \overset{(b)}{\geqslant} I(\mathbf{Z}; X|\mathbf{B}_0) + I(\mathbf{Z}; Y|\mathbf{B}_0, X). \tag{3}$$

Equality (a) applies the chain rule of mutual information. Inequality (b) holds because $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ implies the DP inequality $I(\mathbf{Z}; U|\mathbf{B}_0, X) \geqslant I(\mathbf{Z}; Y|\mathbf{B}_0, X)$. Combining the upper bound $H(X|\mathbf{B}_0)$ and the lower bound $I(\mathbf{Z}; X|\mathbf{B}_0) + I(\mathbf{Z}; Y|\mathbf{B}_0, X)$ proves the testable inequality. □

The instrumental entropic inequality holds in Figure 1A with, $\mathbf{Z} = Z$ and $\mathbf{B}_0 = W$. It equally holds with changes in $Z - W - Y$, as long as W is a noncollider. Note that while the instrumental inequality gets its name from the possibility to use Z as a *causal intervention instrument* that can be manipulated (intervened) to estimate the causal effect that X has on Y [20,33,34] in fact the inequality is equally fulfilled with $Z \leftrightarrow X$, since this does not alter the independence $Z \perp U|W$. That is, the instrumental inequality also holds for causal structures where Z is not a causal intervention instrument. Since in this work we study causal structure learning and not the identification of causal effects, we will refer to a set of variables as a *causal discovery instrumental set* purely based on the fulfillment of the independence $\mathbf{Z} \perp U|\mathbf{B}_0$. The reason not to involve the independence $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ in this criterion will become clear in Section 3.2. Furthermore, since there is no possible confusion within this work, we will abbreviate *causal discovery instrumental set* simply as *instrumental set*.

Proposition 1 states a straightforward extended version of the basic instrumental entropic inequality in the sense that it comprises a multivariate $\mathbf{Z}$. We will refer to this inequality as the *standard instrumental entropic inequality*. This multivariate version will be needed for comparison with the new types of instrumental entropic inequalities we will introduce. Note that we purposely have excluded further straightforward generalizations, such as a multivariate $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{U}$. We will add more generalizations in the Results section, but this simple version is suited to identify the key components of the inequality and its relation to the causal structure, as will be examined in Section 3.2.

Importantly, a causal structure that fulfills the nontestable independencies $\mathbf{Z} \perp U|\mathbf{B}_0$ and $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ imposes the fulfillment of the inequality of Proposition 1 to any data set generated from that causal structure. In that case, we will say that the inequality is *causally* fulfilled. On the other hand, for a data set generated with another causal structure, the entropic inequality may equally be fulfilled, even if its fulfillment was not imposed by the causal structure. In that case, we will say that the inequality is *statistically* fulfilled. The causal inference power of an inequality emanates from the possibility to discard a causal structure that imposes the causal-fulfillment of the inequality when the violation of the inequality is verified from data. As mentioned in Section 2.1, in order to be able to reject causal structures based on the violation of causally informative inequalities, we will work under the assumption that causal separability (d-separation) implies statistical conditional independence. That is, we assume that for a causal structure that causally imposes an inequality, the inequality indeed is fulfilled because the causal structure creates the independencies that lead to the inequality.

2.4. Constrained Minimum Mutual Information and Maximum Entropy Unique Information

Problems of constrained optimization of information-theoretic quantities within families that share marginal distributions often appear in the study of communication channels [35]. Furthermore, minimum information methods have been proposed for machine learning and signal processing tasks [36,37] as a generalization of the maximum entropy principle [38]. As developed in the Results below, our extension of causally informative entropic inequalities relies on the use of minimum information (minInf) terms that are defined within families of distributions that share sets of marginals with the original joint distribution.

In general, minimization constraints can comprise both inequality and equality equations that are imposed to joint distributions of variables. A set of constraints determines a family ΔP of probability distributions within which the mutual information term of interest is to be minimized, namely among all distributions compatible with the fulfillment of the constraints. In this work, we focus on minInf terms defined within families of distributions that preserve sets of marginals of the joint distribution associated with a data set. In general, given a joint distribution $P(\bar{\mathbf{V}})$ for $\bar{\mathbf{V}}$ variables, a minInf term is defined as

$$\min_{Q \in \Delta P} I_Q(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3), \tag{4}$$

where $\bar{\mathbf{V}}_i$, $i = 1, \ldots, 3$ are subsets of $\bar{\mathbf{V}}$ and ΔP is the family of distributions determined by the set of constraints imposed to the distributions. Accordingly, the minimization within the family corresponds to a constrained minimization subject to the constraints that define the family. When the constraints impose the preservation of a set of marginal distributions of the original $P(\bar{\mathbf{V}})$, that is, when $Q \in \Delta P$ is subject to $Q(\bar{\mathbf{V}}_S) = P(\bar{\mathbf{V}}_S)$ for a certain number of subsets S of $\bar{\mathbf{V}}$, then the constraints constitute a set of affine linear equality constraints on the joint distribution $Q(\bar{\mathbf{V}})$.

In this work we prove some general properties of minInf information terms that render them useful for the derivation of more powerful causally informative testable entropic inequalities. These properties include a data processing inequality for minInf terms (Proposition 5) and a procedure to iteratively combine minInf data processing inequalities (Theorem 1). We then use these properties for the construction of some specific entropic inequalities, such as extended instrumental entropic inequalities. To the best of our knowledge, although the use of minimum information quantities appears in the study of communication channels and has been formulated in machine learning problems as

mentioned above, the properties we introduce have not been derived before, and the use of minInf terms to derive causally informative entropic inequalities is new.

As a precedent to this work, Chicharro and Nguyen [26] showed how to apply to causal structure learning a measure of maximum entropy unique information previously introduced in [29]. Maximum entropy measures correspond to a specific subcase of minInf terms of the form of Equation (4). Concretely, when $P(\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_3)$ is among the preserved marginals in ΔP, then the entropy $H_Q(\bar{\mathbf{V}}_1|\bar{\mathbf{V}}_3)$ is fixed, and the minimization of $I_Q(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2|\bar{\mathbf{V}}_3)$ is equivalent to the maximization of $H_Q(\bar{\mathbf{V}}_1|\bar{\mathbf{V}}_2, \bar{\mathbf{V}}_3)$.

We now revise the definition of the maximum entropy unique information [29] and the relevant properties derived in Chicharro and Nguyen [26]. This will serve as reference to generalize more general data processing inequalities for minInf terms. The concept of *unique information* was originally introduced [30] as part of a nonnegative decomposition of the joint mutual information that a set of *predictor* variables has about a *target* variable $\bar{\mathbf{Z}}$. In the simplest scenario with two (possibly multivariate) predictors $\{\mathbf{D}_1, \mathbf{D}_2\}$, the unique information of predictor $\mathbf{D}_i$ with respect to the *reference* predictor $\mathbf{D}_j$ quantifies the exclusive information about $\bar{\mathbf{Z}}$ obtained from $\mathbf{D}_i$ and not from $\mathbf{D}_j$. The other components of the decomposition quantify redundant and synergistic information terms. Alternative formulations of this decomposition of mutual information have been introduced, e.g., [29,39–41], using definitions alternative to the maximum entropy formulation of unique information of [29]. However, the application of the maximum entropy unique information measure to causal inference [26] does not rely on its embedding within the framework of [30], but only on certain properties that render it useful to derive testable causally informative information inequalities. We here revise its definition and these properties. In general, we will use *unique information* to refer concretely to the maximum entropy unique information.

While the measure was originally introduced in the bivariate unconditional case [29], we here revise the conditional unique information measure as presented in [26]. For sets of variables $\bar{\mathbf{Z}}$, $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{O}_1$, the unique information of *predictor* $\mathbf{D}_1$ with respect to the *reference* $\mathbf{D}_2$ about the *target* $\bar{\mathbf{Z}}$, conditioning on $\mathbf{O}_1$, is defined as

$$I(\bar{\mathbf{Z}}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1) \equiv \min_{Q \in \Delta P} I_Q(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{E}), \tag{5}$$

where $\mathbf{E} \equiv \{\mathbf{D}_2, \mathbf{O}_1\}$, and ΔP is the family of distributions on $\{\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{D}_2, \mathbf{O}_1\}$ that preserve the marginals $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{D}_2, \mathbf{O}_1)$ of the original $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{D}_2, \mathbf{O}_1)$. The notation I_Q is used to indicate that the mutual information is calculated on the probability distribution Q. We use $I(\bar{\mathbf{Z}}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1)$ to refer to the unique information of $\mathbf{D}_1$ with reference $\mathbf{D}_2$, conditioning on $\mathbf{O}_1$, compared to $I(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{D}_2, \mathbf{O}_1)$, which is the standard conditional information of $\mathbf{D}_1$ conditioning on $\{\mathbf{D}_2, \mathbf{O}_1\}$. Note that the constraints on ΔP are such that they divide the conditioning set $\mathbf{E}$ into the variables $\mathbf{O}_1$ included in $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{O}_1)$ and the variables $\mathbf{D}_2$, that appear only in the marginal $P(\bar{\mathbf{Z}}, \mathbf{D}_2, \mathbf{O}_1)$, where $\mathbf{D}_1$ is excluded. We enumerate $\mathbf{O}_1$ with subindex 1 because when dealing with more general minInf terms in Section 3.4 we will distinguish multiple subsets of $\mathbf{E}$ preserved in different marginal distributions. We use the notation $\mathbf{D}_1 \backslash\backslash \mathbf{D}_2$ instead of the notation $\mathbf{D}_1 \backslash \mathbf{D}_2$ originally introduced by [29] to differentiate it from the set notation $\mathbf{D}_1 \backslash \mathbf{D}_2$, which indicates the subset of variables in $\mathbf{D}_1$ that is not contained in $\mathbf{D}_2$. The measure defined in Equation (5) is equivalently a maximum entropy measure because the information minimization can equally be formulated as an entropy maximization, since all distributions within ΔP preserve the conditional entropy $H(\bar{\mathbf{Z}}|\mathbf{D}_2, \mathbf{O}_1)$.

Several properties are important for its use into causally informative inequalities. First, by construction [29], the conditional unique information is bounded as

$$\min\{I(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{O}_1), I(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{D}_2, \mathbf{O}_1)\} \geqslant I(\bar{\mathbf{Z}}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1) \geqslant 0. \tag{6}$$

Second, the unique information is monotonic in the predictor argument:

Lemma 2 (Monotonicity of unique information in the predictor argument)**.** *The maximum entropy conditional unique information is monotonic on its second argument, corresponding to the non-referent predictor:*

$$I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D_2}|\mathbf{O}_1) \geqslant I(\bar{\mathbf{Z}};\mathbf{D}\backslash\backslash\mathbf{D_2}|\mathbf{O}_1). \tag{7}$$

In relation to Equation (5), $\mathbf{D}_1 = \{\mathbf{D},\mathbf{D}'\}$. This property was derived in Lemma 3 of [42] for the unconditional case and extended to the conditional case in Lemma 2 of [26]. The proof is provided in Appendix A.

Third, a DP inequality was derived in Chicharro and Nguyen [26] for the maximum entropy unique information:

Lemma 3 (Conditional unique information data processing inequality)**.** *Let* $\bar{\mathbf{Z}}$, $\mathbf{D}$, $\mathbf{D}'$, $\mathbf{D_2}$, *and* $\mathbf{O}_1$ *be five nonoverlapping sets of variables. If* $I(\bar{\mathbf{Z}};\mathbf{D}'|\mathbf{D},\mathbf{O}_1) = 0$, *then* $I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D_2}|\mathbf{O}_1) = I(\bar{\mathbf{Z}};\mathbf{D}\backslash\backslash\mathbf{D_2}|\mathbf{O}_1) \geqslant I(\bar{\mathbf{Z}};\mathbf{D}'\backslash\backslash\mathbf{D_2}|\mathbf{O}_1)$.

Again, the proof is provided in Appendix A to serve as guidance for posterior extensions. This DP inequality is analogous to the standard DP inequality of the mutual information (Lemma 1), except that it requires the independence $\bar{\mathbf{Z}} \perp \mathbf{D}'|\mathbf{D}\mathbf{O}_1$ instead of $\bar{\mathbf{Z}} \perp \mathbf{D}'|\mathbf{D}\mathbf{E}$, with $\mathbf{E} = \{\mathbf{D_2},\mathbf{O}_1\}$. Note that all the variables involved in the independence $\bar{\mathbf{Z}} \perp \mathbf{D}'|\mathbf{D}\mathbf{O}_1$ are included in the marginal $P(\bar{\mathbf{Z}},\mathbf{D}',\mathbf{D},\mathbf{O}_1)$ preserved in the family ΔP that defines $I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D_2}|\mathbf{O}_1)$.

3. Results

To develop how minInf terms can be used in causally informative entropic inequalities, we start from the scenario of the standard instrumental entropic inequality (Figure 1A) and consider changes in the causal structure. We first derive (Section 3.1) an instrumental entropic inequality that applies the DP inequality of unique information. We then address the embedding of this new type of inequalities together with standard instrumental inequalities derived with multivariate instrumental sets and we illustrate that they can provide additional causal inference power (Section 3.2). In Section 3.3, we examine instrumental entropic inequalities in which the DP inequality of conditional mutual information and unique information are combined. This analysis reveals how different types of DP inequalities can recursively be applied. In Section 3.4, we introduce a type of DP inequalities for minInf terms, which encompasses as subcases the DP inequalities of conditional mutual information and unique information. We show how to recursively apply these DP inequalities to obtain sums of observable information terms as lower bounds of unobservable information terms. In Section 3.5, we apply this procedure to construct more powerful instrumental entropic inequalities. In Section 3.6, we reexamine more broadly the derived minInf inequalities from a geometrical perspective, in connection with Shannon entropy cones. Finally, in Section 3.7, we apply the procedures developed in Section 3.4 to other types of entropic inequalities beyond the instrumental inequality scenario.

3.1. Instrumental Entropic Inequalities with Maximum Entropy Unique Information Terms: The Case with One Data Processing Inequality Applied

We start considering how to construct instrumental entropic inequalities with the causal structures of Figure 1B. Again, the graph displays several causal structures depending on the instantiation of the dashed edges. Similar to the case of Figure 1A, a requirement for any instrumental entropic inequalities to be causally fulfilled is that the dashed edges between Z and Y as well as between Z and U are removed. Therefore, we focus on this case

with no edges between Z and Y and between Z and U. A difference between Figure 1A,B is that W is a noncollider in Figure 1A, leading to $Z \not\perp Y|UX$ and $Z \perp Y|UXW$ while W is a collider in Figure 1B, such that $Z \perp Y|UX$ and $Z \not\perp Y|UXW$. For Figure 1B, the instrumental inequality of Proposition 1 can be applied with $\mathbf{Z} = Z$ and $\mathbf{B}_0 = \varnothing$. The required independencies $\mathbf{Z} \perp U|\mathbf{B}_0$ and $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ are fulfilled, namely they correspond to $Z \perp U$ and $Z \perp Y|UX$. This leads to

$$H(X) \geqslant I(Z;X) + I(Z;Y|X). \tag{8}$$

On the contrary, the fact that W is a collider leads to a dependence $Z \not\perp Y|UXW$, and hence in Figure 1B Proposition 1 cannot be applied selecting $\mathbf{B}_0 = W$. Note that being able to condition on $\mathbf{B}_0 = W$ would be advantageous because, following the derivation of Proposition 1, it would lead to a tighter upper bound $H(X|W)$ instead of $H(X)$. Since what prevents deriving an instrumental entropic inequality with $\mathbf{B}_0 = W$ is that $Z \not\perp Y|UXW$, as opposed to $Z \perp Y|UX$, we can consider if using the unique information DP inequality is useful in this case. This is because the unique information has a DP inequality (Lemma 3) that differs from the one of conditional mutual information (Lemma 1) in that it is associated with a conditional independence that excludes the reference variables from the conditioning set. This type of exclusion is precisely what is needed to use $Z \perp Y|UX$ instead of $Z \not\perp Y|UXW$. We first state a general formulation of an instrumental entropic inequality that uses the unique information and we will then go back to the example of Figure 1B.

Proposition 2 (Instrumental entropic inequality with maximum entropy unique information)**.** *Consider the variables* $\mathbf{Z}$, X, Y, $\mathbf{B}_0$, *and* U, *all observable except* U *a hidden variable. Consider that the causal structure is such that, for all* $Z_i \in \mathbf{Z}$, *no pair from* $\{Z_i, X, Y\}$ *is separable given that* U *is hidden. Consider an exclusive partition* $\mathbf{B}_0 = \{\mathbf{B}_1, \mathbf{B}_2\}$. *Consider that the causal structure imposes the nontestable independencies* $\mathbf{Z} \perp U|\mathbf{B}_0$ *and* $\mathbf{Z} \perp Y|UX\mathbf{B}_1$. *These independencies result in the testable inequality*

$$H(X|\mathbf{B}_0) \geqslant I(\mathbf{Z};X|\mathbf{B}_0) + I(\mathbf{Z};Y\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X). \tag{9}$$

Proof. The proof is analogous to the one of Proposition 1. Again, the departing quantity is $I(\mathbf{Z};U,X|\mathbf{B}_0)$. Using the chain rule to decompose $I(\mathbf{Z};U,X|\mathbf{B}_0)$ as the sum of $I(\mathbf{Z};U|\mathbf{B}_0)$ and $I(\mathbf{Z};X|\mathbf{B}_0,U)$, the independence $\mathbf{Z} \perp U|\mathbf{B}_0$ allows deriving $H(X|\mathbf{B}_0)$ as upper bound, as in Equation (2). For the lower bound, instead of Equation (3) that applies the DP inequality of conditional information, the DP inequality of unique information is applied:

$$\begin{aligned} I(\mathbf{Z};U,X|\mathbf{B}_0) &\overset{(a)}{=} I(\mathbf{Z};X|\mathbf{B}_0) + I(\mathbf{Z};U|\mathbf{B}_0,X) \overset{(b)}{\geqslant} I(\mathbf{Z};X|\mathbf{B}_0) + I(\mathbf{Z};U\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X) \overset{(c)}{\geqslant} \\ &I(\mathbf{Z};X|\mathbf{B}_0) + I(\mathbf{Z};Y\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X). \end{aligned} \tag{10}$$

Equality (a) applies the chain rule of mutual information. Inequality (b) applies the definition of the unique information as a contribution smaller than or equal to the conditional mutual information (Equation (6)). Inequality (c) holds because $\mathbf{Z} \perp Y|UX\mathbf{B}_1$ by Lemma 3 implies the DP inequality of the unique information $I(\mathbf{Z};U\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X) \geqslant I(\mathbf{Z};Y\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X)$. Combining the upper bound $H(X|\mathbf{B}_0)$ and the lower bound $I(\mathbf{Z};X|\mathbf{B}_0) + I(\mathbf{Z};Y\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X)$ proves the testable inequality. □

In the example of Figure 1B, Proposition 2 applies with $\mathbf{Z} = Z$, $\mathbf{B}_1 = \varnothing$, and $\mathbf{B}_2 = W$ and results in

$$H(X|W) \geqslant I(Z;X|W) + I(Z;Y\backslash\backslash W|X). \tag{11}$$

The inequality of Equation (11) is causally imposed when the causal structure creates the independencies $Z \perp U|W$ and $Z \perp Y|UX$. These independencies would also exist if in Figure 1B variables X and W were connected by an arc $X \leftrightarrow W$. On the other hand, $W \rightarrow X$ would produce $Z \not\perp Y|UX$, and $X \rightarrow W$ would produce $Z \not\perp U|W$.

Comparing Equations (8) and (11), the first is derived with $\mathbf{B}_0 = \varnothing$ and the second with $\mathbf{B}_0 = W$, $\mathbf{B}_1 = \varnothing$, and $\mathbf{B}_2 = W$. To better appreciate the factors that determine their power, we can rewrite them passing the first term at the r.h.s. to the l.h.s.:

$$H(X|Z) \geqslant I(Z;Y|X) \tag{12a}$$

$$H(X|Z,W) \geqslant I(Z;Y\backslash\backslash W|X). \tag{12b}$$

In general, these inequalities are complementary. For example, consider that it is to be tested the compatibility of a data set with the causal structure in Figure 1B with no edge $Z - U$ and no edge $Z - Y$. This causal structure creates independencies $Z \perp U$, $Z \perp U|W$, and $Z \perp Y|UX$, and hence causally imposes both inequalities of Equations (8) and (11). Therefore, the violation of any of the two inequalities suffices to discard the causal structure. Comparing their form, Equation (12b) has a smaller or equal upper bound than Equation (12a), given the monotonicity of entropy under conditioning. However, it also has a smaller or equal lower bound, since the unique information is upper-bounded by the unconditional mutual information (Equation (6)). This means that, for a concrete data set that has been generated from another causal structure that does not impose the fulfillment of the inequalities, any of the two inequalities can be violated while the other is not, so that their use is complementary for causal inference. Using $Z \perp U|W$ instead of $Z \perp U$ allows decreasing the upper bound, but when W is a collider between Z and Y such that $Z \not\perp Y|UXW$, a smaller observable lower bound is derived with the unique information and $Z \perp Y|UX$.

3.2. Instrumental Entropic Inequalities with Multivariate Instrumental Sets

So far, the example of Figure 1B presented an application of Proposition 2 with a univariate instrument $\mathbf{Z} = Z$. However, to establish that the new type of inequalities of Proposition 2 contributes additional causal inference power to the standard instrumental entropic inequalities, we also need to examine the standard instrumental inequalities with multivariate instrumental sets that exist for the same causal structure. To do so, we first highlight three key elements of the structure of instrumental entropic inequalities.

The derivation of both Propositions 1 and 2 departs from the quantity $I(\mathbf{Z};U,X|\mathbf{B}_0)$. The first key element is that the two required types of independencies play separate roles in the derivation of instrumental inequalities: $\mathbf{Z} \perp U|\mathbf{B}_0$ is used to derive the observable upper bound, while the other independence is used to derive the observable lower bound. In more detail, $\mathbf{Z} \perp U|\mathbf{B}_0$ is used after $I(\mathbf{Z};U,X|\mathbf{B}_0)$ is separated into $I(\mathbf{Z};U|\mathbf{B}_0)$ and $I(\mathbf{Z};X|\mathbf{B}_0,U)$. The other required independence is used to derive the observable lower bound thanks to a DP inequality, which is applied to $I(\mathbf{Z};U|\mathbf{B}_0,X)$. The different DP inequality applied is what differentiates Propositions 1 and 2.

The second key element is better appreciated rewriting the inequalities of Propositions 1 and 2 passing the the first term of the r.h.s. to the l.h.s.:

$$H(X|\mathbf{Z},\mathbf{B}_0) \geqslant I(\mathbf{Z};Y|\mathbf{B}_0,X) \tag{13a}$$

$$H(X|\mathbf{Z},\mathbf{B}_0) \geqslant I(\mathbf{Z};Y\backslash\backslash\mathbf{B}_2|\mathbf{B}_1,X). \tag{13b}$$

Written like this, the two inequalities have upper bound $H(X|\mathbf{Z},\mathbf{B}_0)$, with a conditioning set $\{\mathbf{Z},\mathbf{B}_0\}$ that does not differentiate between $\mathbf{Z}$ and $\mathbf{B}_0$, which appear in different arguments of $\mathbf{Z} \perp U|\mathbf{B}_0$. Therefore, regarding the upper bound, there is an invariance under the exchange of variables between the *causal discovery instrumental set* $\mathbf{Z}$ and the conditioning

set $\mathbf{B}_0$. Alternative instrumental entropic inequalities that would require independencies $\{\mathbf{Z}, \mathbf{B}_4\} \perp U|\mathbf{B}_3$, with $\mathbf{B}_0 = \{\mathbf{B}_3, \mathbf{B}_4\}$, would all lead to the same upper bound $H(X|\mathbf{Z}, \mathbf{B}_0)$. Accordingly, instrumental inequalities with multivariate instrumental sets obtained under this invariance need to be considered in order to assess the additional causal inference power provided by the new type of inequalities introduced in Proposition 2.

We can see an example of multivariate instrumental set in Figure 1B, again focusing on the causal structure with no connections $Z - U$ and $Z - Y$. Variable W, which in the derivation of Equation (12b) is assigned to $\mathbf{B}_0 = W$, can be assigned to the instrumental set, leading to the bivariate instrumental set $\mathbf{Z} = \{Z, W\}$ and to $\mathbf{B}_0 = \varnothing$. The set $\{Z, W\}$ fulfills $\mathbf{Z} \perp U|\mathbf{B}_0$, namely $\{Z, W\} \perp U$. On the other hand, $\{Z, W\}$ does not fulfill the other independence condition $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ required in Proposition 1, since $\{Z, W\} \not\perp Y|UX$ due to the direct connection between W and Y.

This leads us to the third key element of the construction of instrumental entropic inequalities. In Propositions 1 and 2, there is a single set $\mathbf{Z}$ that appears in both independencies, namely $\mathbf{Z}' = \mathbf{Z}$ in either $\mathbf{Z} \perp U|\mathbf{B}_0$ and $\mathbf{Z}' \perp Y|UX\mathbf{B}_0$, or in $\mathbf{Z} \perp U|\mathbf{B}_0$ and $\mathbf{Z}' \perp Y|UX\mathbf{B}_1$. However, this constraint is not necessary and can be relaxed. In the derivation of an observable lower bound, a DP inequality could be applied to any subset $\mathbf{Z}' \subseteq \mathbf{Z}$. This is captured in the following Proposition. For later convenience, we now also consider multivariate variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{U}$:

Proposition 3 (Chainlike instrumental entropic inequalities with multivariate instrumental sets)**.** *Consider variables* $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{B}_0$, *and* $\mathbf{U}$, *all observable except* $\mathbf{U}$ *hidden variables. Consider that the causal structure is such that for at least a* $Z_i \in \mathbf{Z}$ *there is a nonempty subset* $\mathbf{X}_i \in \mathbf{X}$ *and* $\mathbf{Y}_i \in \mathbf{Y}$ *such that no pair in* $\{Z_i, \mathbf{X}_i, \mathbf{Y}_i\}$ *is separable with* $\mathbf{U}$ *hidden. Consider an exclusive partition in r parts of the multivariate instrumental set* $\mathbf{Z}$ *given by* $\mathbf{Z}^{[r]} = \{\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_r\}$, *with* $\mathbf{Z}_0 = \varnothing$. *Consider that the causal structure imposes the nontestable independence* $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$. *This independence creates a nontestable instrumental entropic inequality*

$$H(\mathbf{X}|\mathbf{B}_0) \geqslant I(\mathbf{Z}; \mathbf{X}|\mathbf{B}_0) + I(\mathbf{Z}; \mathbf{U}|\mathbf{B}_0, \mathbf{X}) = I(\mathbf{Z}; \mathbf{X}|\mathbf{B}_0) + \sum_{j=1}^{r} I(\mathbf{Z}_j; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[j-1]}), \tag{14}$$

with $\mathbf{Z}^{[j-1]} = \{\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_{j-1}\}$, *where nontestability is due to the nonestimable components of the lower bound. A nontrivial testable instrumental entropic inequality exists if for at least one term* $I(\mathbf{Z}_j; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[j-1]})$ *at least one conditional independence exists that enables a data processing inequality to substitute that term by an estimable lower bound that contains some variables in* $\mathbf{Y}$ *and does not contain* $\mathbf{U}$.

Proof. The derivation of the upper bound is the same as in Propositions 1 and 2. Starting from $I(\mathbf{Z}; \mathbf{U}, \mathbf{X}|\mathbf{B}_0)$ the chain rule is applied and the upper bound derived with $I(\mathbf{Z}; \mathbf{U}|\mathbf{B}_0) = 0$ thanks to $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$. The nonestimable lower bound follows from a direct application of the chain rule equality of conditional mutual information to separate $\{\mathbf{U}, \mathbf{X}\}$ into the terms $I(\mathbf{Z}; \mathbf{X}|\mathbf{B}_0)$ and $I(\mathbf{Z}; \mathbf{U}|\mathbf{B}_0, \mathbf{X})$, followed by the chain rule to separate $\mathbf{Z}$ with the partition $\mathbf{Z}^{[r]}$. Finally, the proposition states that it suffices that at least one term of the sum in the lower bound can be replaced by at least one observable information term so that a nontrivial testable inequality is obtained, dropping all remaining terms in the lower bound that contain hidden variables. This replacement is possible when applying at least one DP inequality to at least a term $I(\mathbf{Z}_j; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[j-1]})$. Of course, a testable instrumental inequality is also obtained if more than one term can be replaced by observable lower bounds. □

Note that Proposition 3 does not specify the form of the conditional independencies and associated DP inequalities applied to obtain observable lower bounds. New procedures

to do so are to be specified in Section 3.4. The advantage of this formulation is that it distinguishes between the variables $\mathbf{Z}$ that appear in the condition $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$ and the subsets $\mathbf{Z}_j$ that are involved in the independencies associated with DP inequalities to derive observable terms in the lower bound. This explains why in Section 2.3 we defined *causal discovery instrumental sets* based only on $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$.

We can now reconsider the multivariate instrumental set of Figure 1B that assigns $\mathbf{Z} = \{Z, W\}$ and $\mathbf{B}_0 = \varnothing$. As mentioned, $\{Z, W\} \perp U$ allows deriving an upper bound $H(X|Z, W)$. The reason why Proposition 1 could not be applied is that $\{Z, W\} \not\perp Y|UX$ due to the direct connection between W and Y. However, Proposition 3 can be applied to the example of Figure 1B selecting a partition with $r = 2$, $\mathbf{Z}^{[2]} = \{\varnothing, \{Z\}, \{W\}\}$. Then $I(Z, W; U|X)$ is decomposed into $I(Z; U|X) + I(W; U|X, Z)$. It now suffices to apply the DP inequality based on $Z \perp Y|UX$ to obtain a lower bound $I(Z; U|X) \geqslant I(Z; Y|X)$, and hence to obtain the instrumental entropic inequality

$$H(X) \geqslant I(Z, W; X) + I(Z; Y|X). \tag{15}$$

The test with this inequality subsumes both tests of Equations (8) and (11), derived from Propositions 1 and 2, respectively. This is because it can be rewritten as $H(X|Z, W) \geqslant I(Z; Y|X)$, and hence combines the upper bound of Equation (12b) and the lower bound of Equation (12a).

The consideration of instrumental inequalities with multivariate instrumental sets discards that the new type instrumental entropic inequality of Proposition 2 provides additional causal inference power in the case of Figure 1B. More generally, we show in Appendix J that for a multivariate instrumental set $\mathbf{Z}$ that fulfills an independence $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$, no additional power can be gained from tests that use only a subset of $\mathbf{Z}$ as instrumental set, removing the rest of variables or transferring some variables of $\mathbf{Z}$ into the conditioning set $\mathbf{B}_0$. However, a causal structure may be such that for $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2\}$ it holds $\mathbf{Z}_2 \perp \mathbf{U}|\mathbf{B}_0\mathbf{Z}_1$ and yet $\mathbf{Z} \not\perp \mathbf{U}|\mathbf{B}_0$. Or with the opposite perspective, using the notation of Proposition 2 with $\mathbf{B}_0 = \{\mathbf{B}_1, \mathbf{B}_2\}$, the causal structure may be such that $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$, but $\{\mathbf{Z}, \mathbf{B}_2\} \not\perp \mathbf{U}|\mathbf{B}_1$. In this case, Proposition 2 can add causal inference power.

An example of this is illustrated in Figure 1C. With no direct connection between Z and U, the independence $Z \perp U|W_1W_2$ holds. On the other hand, $\{Z, W_1\} \not\perp U|W_2$, $\{Z, W_2\} \not\perp U|W_1$, and $\{Z, W_1, W_2\} \not\perp U$. If any of these independencies existed, the upper bound $H(X|Z, W_1, W_2)$ would be equally obtained from the corresponding multivariate instrumental set, because of the invariance of the upper bound to exchanges between $\mathbf{Z}$ and $\mathbf{B}_0$. In those cases, the variables of $\{W_1, W_2\}$ included in the instrumental set could be marginalized instead of requiring the use of a unique information with both variables in the reference argument. However, since these other independencies do not exist, the instrumental inequality constructed from Proposition 2 using $Z \perp U|W_1W_2$ and $Z \perp Y|UX$ adds additional causal inference power with the test

$$H(X|W_1, W_2) \geqslant I(Z; X|W_1, W_2) + I(Z; Y\backslash\backslash\{W_1, W_2\}|X). \tag{16}$$

In this section, we have addressed the issue of whether the new type of entropic inequalities with unique information terms can add causal inference power when tested together with standard instrumental inequalities that use related multivariate instrumental sets. Our objective was to ensure that the new type of inequalities is not trivially subsumed. To do so, we have contemplated two factors that determine when a new inequality test adds causal inference power. First, that there is some hypothesized causal structure of interest that causally imposes the new type of inequality, possibly together with other inequalities. Second, that the form of the new inequality is such that a probability distribution can exist for which the new test is rejected while no other causally-imposed inequality test

is simultaneously rejected. On the other hand, if the new inequality is such that when violated also another inequality is always violated, then it does not add additional power. In Appendix B, we provide a full formal statement of the if and only if conditions under which a new entropic inequality test adds additional power to a set of other tests.

By examining multivariate instrumental sets, we have provided an example in Figure 1C for which Proposition 2 provides additional causal inference power. More broadly, we have seen in Proposition 3 that a nontestable entropic inequality is associated with each instrumental set. The identification of a set of variables as an instrumental set implies the existence of an observable upper bound, and the lower bound can be decomposed as a sum of nonobservable information terms. Nontrivial testable entropic inequalities are obtained when finding observable lower bounds of some of these terms. Proposition 3 accommodates the inequalities derived in Propositions 1 and 2. Compared to those, it relaxes the condition of Propositions 1 and 2 that requires that no pair from $\{Z_i, X, Y\}$ is separable when U is hidden, for all $Z_i \in \mathbf{Z}$. This is because Proposition 3 does not cover only the specific application of a type of DP inequality associated with a specific conditional independence, contrarily to Propositions 1 and 2 that are linked to $\mathbf{Z} \perp Y|UX\mathbf{B}_0$ and $\mathbf{Z} \perp Y|UX\mathbf{B}_1$, respectively. Sections 3.3–3.5 will generalize procedures to convert nontestable instrumental entropic inequalities into testable ones.

3.3. Instrumental Entropic Inequalities with Mutual Information and Maximum Entropy Unique Information Terms: The Case with Two Data Processing Inequalities Applied

We have seen above that, when a single conditional independence is used to derive a lower bound, the unique information DP inequality only increases causal inference power when the variables in the reference argument of the unique information are not part of a valid instrumental set (i.e., no transfer from $\mathbf{B}_2$ as specified in Proposition 2 to $\mathbf{Z}$ creates a valid instrumental set). We now show that this limitation does not occur when two DP inequalities are used to add terms in the lower bound. We show that unique information terms can be added in the lower bound not only instead of conditional information terms, but in addition to them, resulting in an increase of causal inference power. We use this scenario to illustrate the procedure that will then be generalized to combine an arbitrary number of DP inequalities, comprising DP inequalities of conditional mutual information, unique information, and of minInf information terms, as will be introduced in Section 3.4.

Proposition 4 (Instrumental entropic inequalities with conditional mutual information and unique information terms)**.** *Consider variables* $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2\}$, $\mathbf{B}_0 = \{\mathbf{B}_1, \mathbf{B}_2\}$, *and* $\mathbf{U}$, *all observable except* $\mathbf{U}$ *hidden variables. Consider that the causal structure is such that for at least a* $Z_i \in \mathbf{Z}$ *there is a nonempty subset* $\mathbf{X}_i \in \mathbf{X}$ *and* $\mathbf{Y}_i \in \mathbf{Y}$ *such that no pair in* $\{Z_i, \mathbf{X}_i, \mathbf{Y}_i\}$ *is separable when* $\mathbf{U}$ *is hidden. Consider that the causal structure imposes the nontestable independencies* $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$, $\mathbf{Z} \perp \mathbf{Y}_1|\mathbf{U}\mathbf{X}\mathbf{B}_0$, *and* $\mathbf{Z} \perp \mathbf{Y}_2|\mathbf{U}\mathbf{X}\mathbf{B}_1\mathbf{Y}_1\backslash\bar{\mathbf{Y}}$, *with* $\bar{\mathbf{Y}} \subseteq \mathbf{Y}_1$. *These independencies result in the testable inequality*

$$H(\mathbf{X}|\mathbf{B}_0) \geqslant I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0) + I(\mathbf{Z};\mathbf{Y}_1|\mathbf{B}_0,\mathbf{X}) + I(\mathbf{Z};\mathbf{Y}_2\backslash\backslash\{\mathbf{B}_2,\bar{\mathbf{Y}}\}|\mathbf{B}_1,\mathbf{X},\mathbf{Y}_1\backslash\bar{\mathbf{Y}}). \tag{17}$$

Proof. The upper bound is derived as in Proposition 1, given the independence $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$. To derive the lower bound, start with $I(\mathbf{Z};\mathbf{U}|\mathbf{B}_0,\mathbf{X})$ after extracting $I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0)$ with the chain rule of mutual information:

$$\begin{aligned} &I(\mathbf{Z};\mathbf{U}|\mathbf{B}_0,\mathbf{X}) \overset{(a)}{=} I(\mathbf{Z};\mathbf{U},\mathbf{Y}_1|\mathbf{B}_0,\mathbf{X}) \overset{(b)}{=} I(\mathbf{Z};\mathbf{Y}_1|\mathbf{B}_0,\mathbf{X}) + I(\mathbf{Z};\mathbf{U}|\mathbf{B}_0,\mathbf{X},\mathbf{Y}_1) \overset{(c)}{\geqslant} \\ &I(\mathbf{Z};\mathbf{Y}_1|\mathbf{B}_0,\mathbf{X}) + I(\mathbf{Z};\mathbf{U}\backslash\backslash\{\mathbf{B}_2,\bar{\mathbf{Y}}\}|\mathbf{B}_1,\mathbf{X},\mathbf{Y}_1\backslash\bar{\mathbf{Y}}) \overset{(d)}{\geqslant} I(\mathbf{Z};\mathbf{Y}_1|\mathbf{B}_0,\mathbf{X}) + \\ &I(\mathbf{Z};\mathbf{Y}_2\backslash\backslash\{\mathbf{B}_2,\bar{\mathbf{Y}}\}|\mathbf{B}_1,\mathbf{X},\mathbf{Y}_1\backslash\bar{\mathbf{Y}}). \end{aligned} \tag{18}$$

Equality (a) applies the DP inequality of conditional mutual information (Lemma 1) thanks to $\mathbf{Z} \perp \mathbf{Y}_1|\mathbf{UXB}_0$. Equality (b) applies the chain rule equality of mutual information. Inequality (c) holds from the definition of unique information, which has conditional mutual information as an upper bound (Equation (6)). Inequality (d) applies the DP inequality of unique information with $\mathbf{Z} \perp \mathbf{Y}_2|\mathbf{UXB}_1\mathbf{Y}_1\backslash\tilde{\mathbf{Y}}$. □

The examples of Figure 2A,B illustrate how the two types of DP inequalities are combined. For simplicity of the explanations, in contrast to Figure 1, in Figure 2 we only represent individual causal structures (no dashed connections). Our objective here is not to derive all existing inequalities in these graphs, but to illustrate the procedure to construct inequalities combining the two types of DP inequalities and to examine the additional causal inference power they can provide. In these causal structures, select the instrumental set $\mathbf{Z} = Z$ and conditioning set $\mathbf{B}_0 = \{W_1, W_2\}$. With this selection, both causal structures impose inequalities of the type of Proposition 4.

In the example of Figure 2A, given $\mathbf{B}_0 = \{W_1, W_2\}$, the upper bound $H(X|W_1, W_2)$ is derived with $Z \perp U|W_1W_2$. Following Proposition 4, the term with Y_1 in the r.h.s., $I(Z; Y_1|W_1, W_2, X)$, is obtained with $Z \perp Y_1|UXW_1W_2$, from the mutual information DP inequality. The term with Y_2, $I(Z; Y_2\backslash\backslash\{W_2, Y_1\}|W_1, X)$, is obtained with $Z \perp Y_2|UXW_1$, from the unique information DP inequality. The derivation corresponds to the assignment in Equation (17) of $\mathbf{B}_1 = W_1$, $\mathbf{B}_2 = W_2$, and $\tilde{\mathbf{Y}} = Y_1$. Note that it is necessary to exclude $\{W_2, Y_1\}$ from the conditioning set because $Z \not\perp Y_2|UXW_1\mathbf{S}$, for any nonempty $\mathbf{S} \subseteq \{W_2, Y_1\}$. In Figure 2B, the upper bound and the term with Y_1 are derived in the same way. On the other hand, the term $I(Z; Y_2\backslash\backslash W_2|W_1, X, Y_1)$ results from $Z \perp Y_2|UXW_1Y_1$, with $\tilde{\mathbf{Y}} = \varnothing$. The unique information is chosen with reference variable W_2 as opposed to the reference variable $\{W_2, Y_1\}$ for Figure 2A. This difference is due to $W_2 \to Y_1$ in Figure 2A, which renders Y_1 a descendant of the collider W_2 in $Z - W_2 - Y_2$, which needs to be removed from conditioning to create an independence. Note that while for simplicity Figure 2A,B show specific causal structures, Proposition 4 also applies under certain changes of these examples. The same inequalities apply in any graph in which arrows are assigned such that W_1 continues to be a noncollider in $Z - W_1 \to U$ and $Z - W_1 \to Y_1$, or if either $W_1 - U$ or $W_1 - Y_1$ is not present. The graphs could also comprise $W_1 \to W_2$ or $Z - W_1 \to Y_2$, with W_1 a noncollider.

We will now illustrate that instrumental inequalities of the type of Proposition 4 can add extra causal inference power. To do so, we show that, for example in Figure 2A, this type of instrumental entropic inequalities is not subsumed by any standard instrumental entropic inequality that uses a multivariate instrumental set. We here justify the additional causal inference power verifying that no standard instrumental inequality can jointly use the DP inequalities that introduce variables Y_1 and Y_2. This guarantees that it does not happen that the new type of instrumental inequality is only violated in cases in which already a standard instrumental inequality is violated. To complement this reasoning, in Appendix D we examine concrete numerical examples in which no rejection occurs for tests based on standard instrumental inequalities, while rejections are obtained when incorporating unique information terms as in Proposition 4.

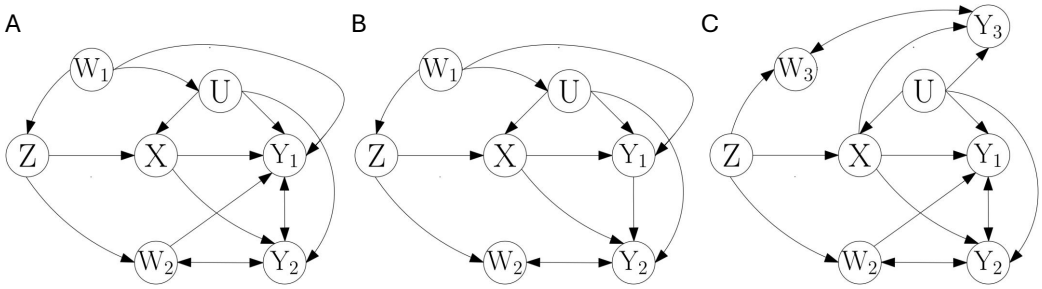


Figure 2. Examples of causal structures that causally impose instrumental entropic inequalities derived by the combination of several types of DP inequalities. All variables are observable except U hidden. (**A**,**B**) Examples of causal structures that impose inequalities of the type of Proposition 4, containing conditional mutual information and unique information terms in the lower bound. (**C**) Example in which an instrumental entropic inequality is causally imposed following Proposition 6, containing a sum of minInf terms in its lower bound.

Consider the multivariate instrumental sets that can be used as an alternative given the selection of the instrumental set $\mathbf{Z} = Z$ and conditioning set $\mathbf{B}_0 = \{W_1, W_2\}$ in Figure 2A,B. We know from Section 3.2 that the same upper bound would hold for any valid instrumental set obtained from a transfer between these sets, namely with $\mathbf{Z} = \{Z, \mathbf{S}\}$, and $\mathbf{B}_0 = \{W_1, W_2\} \backslash \mathbf{S}$, for $\mathbf{S} \subseteq \{W_1, W_2\}$. Given the direct connection between W_1 and U, no set containing W_1 fulfills the criterion $\mathbf{Z} \perp \mathbf{U} | \mathbf{B}_0$ to be an instrumental set. On the other hand, $\{Z, W_2\} \perp U | W_1$, so that $\mathbf{Z} = \{Z, W_2\}$ is a valid instrumental set with $\mathbf{B}_0 = W_1$.

We now focus specifically on Figure 2A. Starting from $I(Z, W_2; U | W_1, X)$, we examine different partitions of $\mathbf{Z} = \{Z, W_2\}$ that can be selected to apply Proposition 3. One partition is $\mathbf{Z}^{[2]} = \{\varnothing, \{Z\}, \{W_2\}\}$, which results in terms $I(Z; U | W_1, X)$ and $I(W_2; U | W_1, X, Z)$. Given that W_2 has direct connections to Y_1 and Y_2, and hence it is nonseparable from them, no DP inequality can be applied to $I(W_2; U | W_1, X, Z)$. For $I(Z; U | W_1, X)$, the DP inequality associated with $Z \perp Y_2 | UXW_1$ can be applied. However, it is not possible to introduce Y_1, which requires conditioning on W_2 in $Z \perp Y_1 | UXW_1W_2$. The opposite partition $\mathbf{Z}^{[2]} = \{\varnothing, \{W_2\}, \{Z\}\}$ results in terms $I(W_2; U | W_1, X)$ and $I(Z; U | W_1, X, W_2)$. Again, no DP inequality is applicable to $I(W_2; U | W_1, X)$. We see that the remaining term $I(Z; U | W_1, X, W_2)$ corresponds to the term used as starting point when applying Proposition 4. This shows that in this example starting from the multivariate instrumental set comes back to $\mathbf{Z} = Z$, and $\mathbf{B}_0 = \{W_1, W_2\}$.

The key element that leads the combination of the mutual information and unique information DP inequalities to add causal inference power is the intertwined requirements in the independencies $Z \perp Y_1 | UXW_1W_2$ and $Z \perp Y_2 | UXW_1$. In Figure 2A, conditioning on W_2 is necessary to separate Z from Y_1, since W_2 is a noncollider in $Z - W_2 - Y_1$. At the same time, W_2 cannot appear in the conditioning set to separate Z and Y_2, since it is a collider in $Z - W_2 - Y_2$. This means that W_2 cannot simply be marginalized to exploit jointly the two independencies $Z \perp Y_1 | UXW_1W_2$ and $Z \perp Y_2 | UXW_1$. It needs to first appear in the conditioning set (when applying the DP inequality of conditional mutual information) and then be excluded from the conditioning set (leading to the application of the unique information DP inequality).

This analysis of Figure 2A highlights the difference with the scenario addressed in Section 3.1, in which a single DP inequality was applied. With a single DP inequality, the unique information DP inequality can only contribute to increase causal inference power when the variables that appear in the reference argument of the unique information cannot be part of a valid instrumental set. On the other hand, when combining DP inequalities, it is the intertwined structure of the independencies associated with different types of DP inequalities what requires their combination. To further highlight this point, in

Appendix C we compare in more detail Figure 2A,B. For Figure 2B, the inequality of the type of Proposition 4 derived with $\mathbf{Z} = Z$ and $\mathbf{B}_0 = \{W_1, W_2\}$ does not add causal inference power to the instrumental inequality derived with $\mathbf{Z} = \{Z, W_2\}$ and $\mathbf{B}_0 = W_1$ that relies only on the DP inequality of conditional mutual information. The key difference is that in Figure 2A conditioning on W_2 is necessary to create the independence between Z and Y_1, while in Figure 2B the independence $Z \perp Y_1|UXW_1$ also holds. This does not create the intertwined structure of $Z \perp Y_1|UXW_1W_2$ and $Z \perp Y_2|UXW_1$, which require respectively the conditioning on W_2 and non conditioning on W_2 (see Appendix C for details).

In this section, we have examined how the DP inequality of conditional mutual information and of unique information can be used sequentially to introduce new observable information terms in the lower bound of an instrumental entropic inequality. Note that the chainlike instrumental entropic inequality of Proposition 3 accommodates the use of Proposition 4. Proposition 3 indicates potential partitions of $\mathbf{Z}$ into $\mathbf{Z}^{[r]} = \{\varnothing, \mathbf{Z}_1, \ldots, \mathbf{Z}_r\}$, while Proposition 4 describes a procedure to derive observable information terms that can be introduced in parallel starting separately from different summands $I(\mathbf{Z}_j; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[j-1]})$ of the r.h.s. of Equation (14).

In the next section we will see that the sequential addition of observable information terms can be extended with a more general type of minInf DP inequalities. With Proposition 4, we have seen that the combination of the DP inequality of conditional mutual information and unique information allows sequentially including and then removing from the conditioning set variables that are required to create an independence between $\mathbf{Z}$ and $\mathbf{Y}_1$, but that preclude from creating an independence between $\mathbf{Z}$ and $\mathbf{Y}_2$. This is achieved because, while the DP inequality of conditional mutual information operates in the original joint distribution $P(\mathbf{Z}, \mathbf{U}, \mathbf{B}_0, \mathbf{X}, \mathbf{Y})$, the DP inequality of unique information operates within the family of distributions that only preserve $P(\mathbf{Z}, \mathbf{U}, \mathbf{B}_1, \mathbf{X}, \mathbf{Y}_1\backslash\tilde{\mathbf{Y}}, \mathbf{Y}_2)$ and $P(\mathbf{Z}, \mathbf{B}_0, \mathbf{X}, \mathbf{Y}_1)$. It is the exclusion of $\mathbf{B}_2 = \mathbf{B}_0\backslash\mathbf{B}_1$ and of $\tilde{\mathbf{Y}}$ from $P(\mathbf{Z}, \mathbf{U}, \mathbf{B}_1, \mathbf{X}, \mathbf{Y}_1\backslash\tilde{\mathbf{Y}}, \mathbf{Y}_2)$ what allows exploiting an independence with only $\mathbf{B}_1$ in the conditioning set instead of $\mathbf{B}_0$. With the same logic, further relaxations of which marginals are preserved will allow us to sequentially combine more DP inequalities.

As a last remark, so far the introduction of new types of instrumental entropic inequalities (Propositions 2 and 4) has been accompanied by the comparison to related standard intrumental entropic inequalities with multivariate instrumental sets. This was necessary to verify that the new entropic inequalities do provide additional causal inference power. The examples of Figures 1C and 2A show that indeed additional causal inference power can be gained either because of the lack of validity of corresponding multivariate instrumental sets (Figure 1C), or because of the intertwinement between the conditioning sets that appear in different independencies, which requires the application of the unique information DP inequality (Figure 2A). In the next sections, we will not proceed in the same way, and instead we will exclusively focus in developing instrumental entropic inequalities that add more minInf information terms in the lower bound. The verification that this addition can further increase causal inference power follows from the same logic of these previous examples. Numerical examples will be provided in Appendix H to illustrate that the addition of more minInf terms together with unique information terms increases causal inference power. It is out of the scope of this work to provide a full taxonomy of when instrumental inequalities that exploit certain types of DP inequalities are subsumed by instrumental inequalities that only exploit a subset of those types of DP inequalities. Only in Appendix J, we derive a hierarchy between specific types of instrumental entropic inequalities with related instrumental sets.

3.4. Recursive Use of Data Processing Inequalities to Add Observable minInf Information Terms as Lower Bounds of Information Terms with Hidden Variables

We have identified the DP inequality of unique information as the key property that allows increasing causal inference power using unique information terms. This raises the question of whether analogous DP inequalities exist for other minInf information terms defined with other sets of constraints on the preserved marginals and if so, how to recursively use these DP inequalities to insert additional observable information terms into entropic inequalities. We now show that indeed there is such a DP inequality for a more general form of minInf information terms.

This section contains our core results of how to exploit minInf DP inequalities. We have used the instrumental entropic inequality to vertebrate our presentation, but in Section 3.6 we will describe a wider framework for the finding of new entropic inequalities and in Section 3.7 we will provide further examples of the applicability of the tools here developed. To help differentiate between general results and results specific of the instrumental inequality scenario, we continue to use a different notation of variables specific for the instrumental scenario, separate from the notation used for general results. We present a general DP inequality for minInf terms using the same notation of the DP inequalities of mutual information (Lemma 1) and unique information (Lemma 3). We then show how to iteratively combine minInf DP inequalities to add new observable terms into entropic inequalities. In Section 3.5, we will show how to apply these tools concretely to the instrumental scenario.

Proposition 5 (Data processing inequality in predictor variables of minInf information terms preserving sets of marginals)**.** *Let* $\bar{\mathbf{Z}}$, $\mathbf{D}$, $\mathbf{D}'$, $\mathbf{E}$, *and* $\mathbf{E}_2$ *be five nonoverlapping sets of variables. Consider a probability distribution* $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}', \mathbf{E}, \mathbf{E}_2)$ *and the family of distributions* $\Delta P_{\mathbf{D}\mathbf{D}'}$ *that share the set of marginals* $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}', \mathbf{O}_1)$ *and* $P(\bar{\mathbf{Z}}, \mathbf{O}_i)$ *for* $i = 2, \ldots, m$, *where* $\mathbf{O}^{[m]} = \{\mathbf{O}_0, \mathbf{O}_1, \ldots, \mathbf{O}_m\}$ *is a collection of subsets* $\mathbf{O}_i \subseteq \{\mathbf{E}, \mathbf{E}_2\}$ *and* $\mathbf{O}_0 = \varnothing$. *If the distribution* $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}', \mathbf{E}, \mathbf{E}_2)$ *is such that* $\bar{\mathbf{Z}} \perp_P \mathbf{D}'|\mathbf{D}\mathbf{O}_1$, *then*

$$\min_{Q \in \Delta P_{\mathbf{D}\mathbf{D}'}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D}', \mathbf{E}_2|\mathbf{E}) = \min_{Q \in \Delta P_{\mathbf{D}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{E}_2|\mathbf{E}) \geqslant \min_{Q \in \Delta P_{\mathbf{D}'}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}', \mathbf{E}_2|\mathbf{E}), \tag{19}$$

where $\Delta P_{\mathbf{D}}$ *is the family of distributions that preserve* $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{O}_1)$ *and* $P(\bar{\mathbf{Z}}, \mathbf{O}_i)$ *for* $i = 2, \ldots, m$, *and* $\Delta P_{\mathbf{D}'}$ *is the family of distributions that preserve* $P(\bar{\mathbf{Z}}, \mathbf{D}', \mathbf{O}_1)$ *and* $P(\bar{\mathbf{Z}}, \mathbf{O}_i)$ *for* $i = 2, \ldots, m$.

Proof. Given the chain rule of mutual information

$$\begin{aligned} \min_{Q \in \Delta P_{\mathbf{D}\mathbf{D}'}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D}', \mathbf{E}_2|\mathbf{E}) = \min_{Q \in \Delta P_{\mathbf{D}\mathbf{D}'}} \big[& I_Q(\bar{\mathbf{Z}}; \mathbf{E}_2|\mathbf{E}) + \\ & I_Q(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{E}, \mathbf{E}_2) + I_Q(\bar{\mathbf{Z}}; \mathbf{D}'|\mathbf{E}, \mathbf{E}_2, \mathbf{D}) \big], \end{aligned} \tag{20}$$

and

$$\min_{Q \in \Delta P_{\mathbf{D}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{E}_2|\mathbf{E}) = \min_{Q \in \Delta P_{\mathbf{D}}} \big[I_Q(\bar{\mathbf{Z}}; \mathbf{E}_2|\mathbf{E}) + I_Q(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{E}, \mathbf{E}_2) \big]. \tag{21}$$

Now consider a distribution that minimizes Equation (21), namely

$$Q^*(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{E}, \mathbf{E}_2) \equiv \arg \min_{Q \in \Delta P_{\mathbf{D}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{E}_2|\mathbf{E}). \tag{22}$$

Construct $\bar{Q}(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}', \mathbf{E}, \mathbf{E}_2) \equiv P(\mathbf{D}'|\mathbf{D}, \mathbf{O}_1) Q^*(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{E}, \mathbf{E}_2)$. Given that $Q^* \in \Delta P_{\mathbf{D}}$, it preserves the marginals $P(\bar{\mathbf{Z}}, \mathbf{O}_i)$ for $i = 2, \ldots, m$ and $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{O}_1)$. Furthermore, $\bar{Q}$ by construction preserves $\bar{\mathbf{Z}} \perp_P \mathbf{D}'|\mathbf{D}\mathbf{O}_1$, which means that it preserves $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}', \mathbf{O}_1)$, and hence $\bar{Q} \in \Delta P_{\mathbf{D}\mathbf{D}'}$, since all other constraints to preserve marginals are the same in $\Delta P_{\mathbf{D}\mathbf{D}'}$ and $\Delta P_{\mathbf{D}}$. Since the first two terms in the sum of Equation (20) do not depend on $\mathbf{D}'$ their minimization

is the same in $\Delta P_{\mathbf{D}}$ or $\Delta P_{\mathbf{DD'}}$ and $Q^*(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{E}, \mathbf{E}_2)$ minimizes their sum, which is equal to the one in Equation (21). By construction of $\bar{Q}$, the independence $\{\bar{\mathbf{Z}}, \mathbf{E}, \mathbf{E}_2\} \backslash \mathbf{O}_1 \perp_{\bar{Q}} \mathbf{D'} | \mathbf{D}\mathbf{O}_1$ holds and hence, using the weak union axiom of semi-graphoids for mutual information [25,43], also the independence $\bar{\mathbf{Z}} \perp_{\bar{Q}} \mathbf{D'} | \mathbf{D}\mathbf{E}\mathbf{E}_2$ holds. This means that the last term in the sum of Equation (20) is zero for $\bar{Q}$. Therefore, $\bar{Q}$ minimizes the r.h.s. of Equation (20), which is equal to the r.h.s. of Equation (21), so that $\min_{Q \in \Delta P_{\mathbf{DD'}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'}, \mathbf{E}_2 | \mathbf{E})$ is equal to $\min_{Q \in \Delta P_{\mathbf{D}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{E}_2 | \mathbf{E})$, with the minima reached by $\bar{Q}(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{E}, \mathbf{E}_2)$ and $Q^*(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{E}, \mathbf{E}_2)$, respectively. Furthermore, monotonicity of mutual information guarantees that information can only decrease when removing variable $\mathbf{D}$ from $\{\mathbf{D}, \mathbf{D'}\}$, namely $I_{\bar{Q}}(\bar{\mathbf{Z}}; \mathbf{D'}, \mathbf{E}_2 | \mathbf{E})$ is smaller than or equal to $I_{\bar{Q}}(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'}, \mathbf{E}_2 | \mathbf{E})$. Finally, $\min_{Q \in \Delta P_{\mathbf{D'}}} I_Q(\bar{\mathbf{Z}}; \mathbf{D'}, \mathbf{E}_2 | \mathbf{E})$ by definition is smaller than or equal to $I_{\bar{Q}}(\bar{\mathbf{Z}}; \mathbf{D'}, \mathbf{E}_2 | \mathbf{E})$. □

Proposition 5 encompasses Lemmas 1 and 3 as subcases. The DP inequality of conditional mutual information is subsumed with $m = 1$, $\mathbf{E}_2 = \varnothing$, and $\mathbf{O}_1 = \mathbf{E}$, such that $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{O}_1)$ corresponds to the joint original distribution. The DP inequality of unique information is subsumed with $m = 2$, $\mathbf{E}_2 = \varnothing$, $\mathbf{O}_2 = \mathbf{E}$, $\mathbf{O}_1 \subset \mathbf{O}_2$. This results in a unique information $I(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'} \backslash\backslash \mathbf{D}_2 | \mathbf{O}_1)$, with $\mathbf{D}_2 = \{\mathbf{O}_2 \backslash \mathbf{O}_1\}$, as in Lemma 3.

Given this DP inequality for minInf information terms, we now describe how it can be used to iteratively add new observable information terms in a lower bound of a minInf information term containing as predictor hidden variables. We start with an example to gain some intuition of the procedure. For this purpose, we recap how the DP inequalities of mutual information and unique information are sequentially combined in Proposition 4, concretely in the example of Figure 2A. We then point out how a similar procedure can be used to sequentially combine more minInf DP inequalities, using Figure 2C as an example.

The first row of Table 1 summarizes how the relaxation of preserved marginals allows combining the DP inequalities of mutual information and unique information in Figure 2A to sequentially insert Y_1 and Y_2. The key aspect of this relaxation is that only the variables involved in the independence $Z \perp Y_2 | UXW_1$ are preserved in the marginal $P(Z, U, W_1, X)$ that includes the hidden variable. This allows applying the unique information DP inequality to insert Y_2, while $Z \not\perp Y_2 | UXW_1W_2$ does not allow applying the mutual information DP inequality. The second row of Table 1 summarizes the analogous procedure applied to Figure 2C to combine the DP inequalities of mutual information and unique information to sequentially insert Y_1 and Y_2. A more detailed examination of the corresponding instrumental entropic inequality that holds for Figure 2C will be examined in Section 3.5. Here, our interest is to motivate that the same procedure of relaxation of the preserved marginals allows applying a third minInf DP inequality to insert Y_3.

Table 1. Sequential relaxation of preserved marginals when combining minInf DP inequalities. (i) Combination of the DP inequality of mutual information and of unique information in Figure 2A. (ii) Combination of the DP inequality of mutual information and of unique information in Figure 2C. (iii) Combination with an additional minInf DP inequality in Figure 2C.

	Prior Use of a DP Inequality	**Relaxation of Preserved Marginals**	**Subsequent Use of a DP Inequality**
(i)	$P(Z, U, Y_1, W_1, W_2, X)$ $Z \perp Y_1 \| UXW_1W_2$	$\{P(Z, U, W_1, X), P(Z, Y_1, W_1, W_2, X)\}$	$\{P(Z, U, Y_2, W_1, X), P(Z, Y_1, W_1, W_2, X)\}$ $Z \perp Y_2 \| UXW_1$
(ii)	$P(Z, U, Y_1, W_2, W_3, X)$ $Z \perp Y_1 \| UXW_2W_3$	$\{P(Z, U, W_3, X), P(Z, Y_1, W_2, W_3, X)\}$	$\{P(Z, U, Y_2, W_3, X), P(Z, Y_1, W_2, W_3, X)\}$ $Z \perp Y_2 \| UXW_3$
(iii)	$\{P(Z, U, Y_2, W_3, X),$ $P(Z, Y_1, W_2, W_3, X)\}$ $Z \perp Y_2 \| UXW_3$	$\{P(Z, U, Y_2, X), P(Z, Y_2, W_3, X),$ $P(Z, Y_1, W_2, W_3, X)\}$	$\{P(Z, U, Y_3, Y_2, X), P(Z, Y_2, W_3, X),$ $P(Z, Y_1, W_2, W_3, X)\}$ $Z \perp Y_3 \| UXY_2$

This is shown in the third row of Table 1. The preserved marginals $\{P(Z, U, Y_2, W_3, X), P(Z, Y_1, W_2, W_3, X)\}$ in the third column of (ii), which allow applying the DP inequality of unique information, are the departing set in the first column of (iii). Then a new relaxation of the marginals divides $P(Z, U, Y_2, W_3, X)$ into $\{P(Z, U, Y_2, X), P(Z, Y_2, W_3, X)\}$. This allows preserving in $P(Z, U, Y_2, X)$ only the variables involved in the independence $Z \perp Y_3|UXY_2$, while $Z \not\perp Y_3|UXW_3Y_2$. The other marginal $P(Z, Y_1, W_2, W_3, X)$ is left unchanged during the relaxation. We then recognize in the structure of the marginals preserved after the iterative application of the relaxations the pattern of constraints of the families of distributions considered in Proposition 5, namely the fact that the hidden variables and conditioning variables involved in the subsequent conditional independence to be exploited are the only ones included together with $\bar{\mathbf{Z}}$ in the marginal distribution that plays the role of $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{O}_1)$. Accordingly, the minInf DP inequality of Proposition 5 is used to insert Y_3.

We now formalize how DP inequalities of minInf terms that are determined by sequential relaxations of the preserved marginals can be combined:

Theorem 1 (Iterative addition of observable minInf information terms to lower bounds of unobservable minInf information terms)**.** *Consider nonoverlapping sets of variables* $\bar{\mathbf{Z}}$, $\mathbf{E}$, *and* $\bar{\mathbf{U}}$ *with all observable except* $\bar{\mathbf{U}}$ *hidden variables. For* $k \geqslant 1$, *consider a nonempty collection of observable nonoverlapping sets of variables* $\mathbf{A}^{[k]} = \{\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_k\}$, *with* $\mathbf{A}_0 = \varnothing$. *Consider a collection* $\bar{\mathbf{Z}}^{[k]} = \{\bar{\mathbf{Z}}_0, \bar{\mathbf{Z}}_1, \ldots, \bar{\mathbf{Z}}_k\}$ *and a collection* $\check{\mathbf{Z}}^{[k]} = \{\check{\mathbf{Z}}_0, \check{\mathbf{Z}}_1, \ldots, \check{\mathbf{Z}}_k\}$ *such that* $\bar{\mathbf{Z}}_0 = \bar{\mathbf{Z}}$, $\check{\mathbf{Z}}_0 = \varnothing$, *and* $\check{\mathbf{Z}}_j \subset \bar{\mathbf{Z}}_{j-1}$, $\bar{\mathbf{Z}}_j \subseteq \bar{\mathbf{Z}}_{j-1} \backslash \check{\mathbf{Z}}_j$, *for* $j = 1, \ldots, k$. *Consider a collection* $\bar{\mathbf{U}}^{[k]} = \{\bar{\mathbf{U}}_0, \bar{\mathbf{U}}_1, \ldots, \bar{\mathbf{U}}_k\}$ *such that* $\bar{\mathbf{U}}_0 = \bar{\mathbf{U}}_1 = \bar{\mathbf{U}}$ *and* $\bar{\mathbf{U}}_j \subseteq \bar{\mathbf{U}}_{j-1}$, *for* $j = 1, \ldots, k$. *Consider the collections of sets of variables* $\bar{\mathbf{B}}^{[k]} = \{\bar{\mathbf{B}}_0, \bar{\mathbf{B}}_1, \ldots, \bar{\mathbf{B}}_k\}$ *and* $\mathbf{C}^{[k]} = \{\mathbf{C}_0, \mathbf{C}_1, \ldots, \mathbf{C}_k\}$, *with* $\bar{\mathbf{B}}_0 = \mathbf{E}$, $\mathbf{C}_0 = \varnothing$, *and iteratively constructed such that* $\bar{\mathbf{B}}'_1 \subseteq \{\mathbf{A}_0, \bar{\mathbf{B}}_0\} = \mathbf{E}$, $\bar{\mathbf{B}}_1 = \{\bar{\mathbf{B}}'_1, \check{\mathbf{Z}}_1\}$, $\mathbf{C}_1 = \{\mathbf{A}_0, \bar{\mathbf{B}}_0, \check{\mathbf{Z}}_1\} = \{\mathbf{E}, \check{\mathbf{Z}}_1\}$, *and for* $j > 1$, $\bar{\mathbf{B}}'_j \subseteq \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$, $\bar{\mathbf{B}}_j = \{\bar{\mathbf{B}}'_j, \check{\mathbf{Z}}_j\}$, *and* $\mathbf{C}_j = \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}, \check{\mathbf{Z}}_j\}$, *so that* $\bar{\mathbf{B}}_j \subseteq \mathbf{C}_j$, *for* $j = 1, \ldots, k$. *Consider a joint distribution* $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}^{[k]}, \mathbf{E})$. *Consider the family of distributions* ΔP_{k-1} *preserving* $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$ *and* $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, *for* $j = 1, \ldots, k-1$. *Consider the family of distributions* ΔP_k *preserving* $P(\bar{\mathbf{Z}}_k, \bar{\mathbf{U}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *and* $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, *for* $j = 1, \ldots, k$. *If* $\bar{\mathbf{Z}}_k \perp_P \mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$, *then*

$$\min_{Q \in \Delta P_{k-1}} I_Q(\bar{\mathbf{Z}}_{k-1}; \bar{\mathbf{U}}_{k-1}|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k-1]}) \geqslant \min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}) + \min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \bar{\mathbf{U}}_k|\mathbf{E}, \mathbf{A}^{[k]}, \check{\mathbf{Z}}^{[k]}). \tag{23}$$

Proof. The proof is provided in Appendix E. □

Theorem 1 provides a way to iteratively add additional observable terms at the lower bound of nonobservable information terms with hidden variables. A full understanding of how it proceeds can be gained with its proof. In the rest of this section, we highlight its main properties, describing the transition from the exploitation of independence $\bar{\mathbf{Z}}_{k-1} \perp_P \mathbf{A}_{k-1}|\bar{\mathbf{U}}_{k-1}\bar{\mathbf{B}}_{k-1}$ in iteration $k-1$ to the exploitation of $\bar{\mathbf{Z}}_k \perp_P \mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$ in iteration k. We start with the simplest scenario, in which $\bar{\mathbf{Z}}_j = \bar{\mathbf{Z}}$, $\bar{\mathbf{U}}_j = \bar{\mathbf{U}}$, and $\check{\mathbf{Z}}_j = \varnothing$, for $j = 1, \ldots, k$. This case already allows appreciating the core of the recursiveness. It corresponds to the scenario in which the DP inequalities being used all have the same target variable $\bar{\mathbf{Z}}$ and rely on the same set of hidden variables $\bar{\mathbf{U}}$. Given that $\check{\mathbf{Z}}_j = \varnothing$ for $j = 1, \ldots, k$, the iterative construction of $\bar{\mathbf{B}}^{[k]}$ and $\mathbf{C}^{[k]}$ can be simplified to $\bar{\mathbf{B}}_1 \subseteq \{\mathbf{A}_0, \bar{\mathbf{B}}_0\} = \mathbf{E}$, $\mathbf{C}_1 = \{\mathbf{A}_0, \bar{\mathbf{B}}_0\} = \mathbf{E}$, and for $j > 1$, $\bar{\mathbf{B}}_j \subset \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$ and $\mathbf{C}_j = \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$. The auxiliary variables $\bar{\mathbf{B}}'_j$ are not needed when $\check{\mathbf{Z}}_j = \varnothing$ for $j = 1, \ldots, k$ because their role is only to add $\check{\mathbf{Z}}_j$ in $\bar{\mathbf{B}}_j = \{\bar{\mathbf{B}}'_j, \check{\mathbf{Z}}_j\}$. Furthermore, for $j > 1$, we have $\bar{\mathbf{B}}_j \subset \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$ instead of $\bar{\mathbf{B}}_j \subseteq \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$ because an equality $\bar{\mathbf{B}}_j = \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$ leads to applying a DP inequality in step $j-1$ with the independence $\bar{\mathbf{Z}} \perp_P \mathbf{A}_{j-1}|\bar{\mathbf{U}}\bar{\mathbf{B}}_{j-1}$ and in step j with $\bar{\mathbf{Z}} \perp_P \mathbf{A}_j|\bar{\mathbf{U}}\mathbf{A}_{j-1}\bar{\mathbf{B}}_{j-1}$. These two steps can then be

merged in a new step $j-1$ that jointly adds $\{\mathbf{A}_{j-1}, \mathbf{A}_j\}$ given $\bar{\mathbf{Z}} \perp_P \mathbf{A}_{j-1}\mathbf{A}_j|\bar{\mathbf{U}}\bar{\mathbf{B}}_{j-1}$, based on the *contraction* axiom of semi-graphoids [25,43].

For this simplest scenario, we now highlight the core of the recursiveness. The independencies of step $k-1$ and k are $\bar{\mathbf{Z}} \perp_P \mathbf{A}_{k-1}|\bar{\mathbf{U}}\bar{\mathbf{B}}_{k-1}$ and $\bar{\mathbf{Z}} \perp_P \mathbf{A}_k|\bar{\mathbf{U}}\bar{\mathbf{B}}_k$, respectively. The family ΔP_{k-1} preserves $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$ and $P(\bar{\mathbf{Z}}, \mathbf{C}_j)$ for $j = 1, \ldots, k-1$, and given that $\mathbf{C}_k = \{\mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1}\}$, it hence preserves $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1}) = P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{C}_k)$. In iteration $k-1$, variables $\mathbf{A}_{k-1}$ are introduced using Proposition 5 with $\bar{\mathbf{Z}} \perp_P \mathbf{A}_{k-1}|\bar{\mathbf{U}}\bar{\mathbf{B}}_{k-1}$. Here the variables $\{\mathbf{Z}, \mathbf{D}, \mathbf{D}', \mathbf{O}_1, \mathbf{E}, \mathbf{E}_2\}$ of Proposition 5 are assigned as $\{\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1}, \{\mathbf{E}, \mathbf{A}^{[k-2]}\}, \varnothing\}$. The family ΔP_{k-1} plays the role of $\Delta P_{\mathbf{D}\mathbf{D}'}$ in Equation (19). When moving from ΔP_{k-1} to ΔP_k, the preservation of $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1}) = P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{C}_k)$ is loosen to the preservation of two of its marginals, namely $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}, \mathbf{C}_k)$. The first is a marginal because $\bar{\mathbf{B}}_k \subseteq \mathbf{C}_k$. The second is a marginal because $\bar{\mathbf{U}}$ is removed. Now $\bar{\mathbf{U}}$ only appears in $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \bar{\mathbf{B}}_k)$. The variables $\mathbf{A}_k$ are introduced analogously to $\mathbf{A}_{k-1}$, using Proposition 5 now with $\bar{\mathbf{Z}} \perp_P \mathbf{A}_k|\bar{\mathbf{U}}\bar{\mathbf{B}}_k$. Here $\{\mathbf{Z}, \mathbf{D}, \mathbf{D}', \mathbf{O}_1, \mathbf{E}, \mathbf{E}_2\}$ are assigned as $\{\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{A}_k, \bar{\mathbf{B}}_k, \{\mathbf{E}, \mathbf{A}^{[k-1]}\}, \varnothing\}$. The family ΔP_k plays the role of $\Delta P_{\mathbf{D}\mathbf{D}'}$ in Equation (19). The second term at the r.h.s. of Equation (23) has the same form as the one at the l.h.s., replacing $k-1$ by k. Comparing the two independencies used in steps $k-1$ and k, $\bar{\mathbf{B}}_k$ used in $\bar{\mathbf{Z}} \perp_P \mathbf{A}_k|\bar{\mathbf{U}}\bar{\mathbf{B}}_k$ is a subset of $\bar{\mathbf{B}}_{k-1}$ used in $\bar{\mathbf{Z}} \perp_P \mathbf{A}_{k-1}|\bar{\mathbf{U}}\bar{\mathbf{B}}_{k-1}$, except for the possible addition of variables from $\mathbf{A}_{k-1}$. This follows the same pattern already seen in Figure 2A with $Z \perp Y_1|UXW_1W_2$ and $Z \perp Y_2|UXW_1$, where $\mathbf{A}_1 = \{Y_1\}$, $\mathbf{A}_2 = \{Y_2\}$, $\bar{\mathbf{B}}_1 = \{X, W_1, W_2\}$, and $\bar{\mathbf{B}}_2 = \{X, W_1\}$, or in Figure 2B with $Z \perp Y_1|UXW_1W_2$ and $Z \perp Y_2|UXW_1Y_1$, where $\mathbf{A}_1$, $\mathbf{A}_2$, and $\bar{\mathbf{B}}_1$ are the same, and $\bar{\mathbf{B}}_2 = \{X, W_1, Y_1\}$.

We now provide an overview of the rest of scenarios. Concrete examples are described in Section 3.5 and in Appendix F. These other scenarios comprise cases in which $\bar{\mathbf{Z}}_j$ or $\bar{\mathbf{U}}_j$ are not constant for $j = 1, \ldots, k$. The cases with non constant $\bar{\mathbf{U}}_j$, given that $\bar{\mathbf{U}}_j \subseteq \bar{\mathbf{U}}_{j-1}$, correspond to cases in which some of the hidden variables $\bar{\mathbf{U}}$ are marginalized before applying subsequent DP inequalities. This happens when in step j the variables $\bar{\mathbf{U}}_{j-1} \backslash \bar{\mathbf{U}}_j$ are colliders or descendants of colliders in paths that lead to $\bar{\mathbf{Z}}_j \not\perp_P \mathbf{A}_j|\bar{\mathbf{U}}_{j-1}\bar{\mathbf{B}}_j$ as opposed to $\bar{\mathbf{Z}}_j \perp_P \mathbf{A}_j|\bar{\mathbf{U}}_j\bar{\mathbf{B}}_j$. An example will be shown in Figure A2B.

Similarly, if $\check{\mathbf{Z}}_j = \varnothing$ for $j = 1, \ldots, k$, the cases with non constant $\bar{\mathbf{Z}}_j$ for some j correspond to cases in which some target variables are marginalized to apply subsequent DP inequalities. This is because the relations $\check{\mathbf{Z}}_j \subset \bar{\mathbf{Z}}_{j-1}$, $\bar{\mathbf{Z}}_j \subseteq \bar{\mathbf{Z}}_{j-1} \backslash \check{\mathbf{Z}}_j$, for $j = 1, \ldots, k$ simplify to $\bar{\mathbf{Z}}_j \subseteq \bar{\mathbf{Z}}_{j-1}$ when $\check{\mathbf{Z}}_j = \varnothing$ for $j = 1, \ldots, k$. This happens in step j when $\bar{\mathbf{Z}}_{j-1} \not\perp_P \mathbf{A}_j|\bar{\mathbf{U}}\bar{\mathbf{B}}_j$, as opposed to $\bar{\mathbf{Z}}_j \perp_P \mathbf{A}_j|\bar{\mathbf{U}}\bar{\mathbf{B}}_j$, because the variables $\bar{\mathbf{Z}}_{j-1} \backslash \bar{\mathbf{Z}}_j$ have active paths reaching $\mathbf{A}_j$. An example will be shown in Figure A2C.

Finally, the case in which $\check{\mathbf{Z}}_j \neq \varnothing$ for some j covers cases in which conditioning on $\check{\mathbf{Z}}_j \subset \bar{\mathbf{Z}}_{j-1}$ is necessary to create the independence $\bar{\mathbf{Z}}_j \perp_P \mathbf{A}_j|\bar{\mathbf{U}}\bar{\mathbf{B}}_j$, with $\check{\mathbf{Z}}_j \subseteq \bar{\mathbf{B}}_j$. Accordingly, in step j the variables $\check{\mathbf{Z}}_j$ are moved from target variables to conditioning variables using a chain rule of the information terms. An example will be shown in Figure A2D. Furthermore, the marginalization of some variables in $\bar{\mathbf{U}}$, the marginalization of some variables in $\bar{\mathbf{Z}}$, and the conditioning on some $\check{\mathbf{Z}}_j$ can co-occur in the same step j, leading to the final general formulation of Theorem 1. The commonality to all scenarios is that in Equation (23), while the term at the l.h.s. is not observable, the first term at r.h.s. does not depend on any hidden variable and hence leads to an observable term by relaxing the preservation of $P(\bar{\mathbf{Z}}_k, \bar{\mathbf{U}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ to $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$. Furthermore, the second term of the r.h.s. has the same form as the term at the l.h.s., which means that Theorem 1 can be applied recursively.

3.5. Instrumental Entropic Inequalities with Sums of minInf Information Terms

We now show how to use Theorem 1 to create testable entropic inequalities from the nontestable inequalities of Proposition 3. Reexamining the terms $I(\mathbf{Z}_j; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[j-1]})$,

$j = 1, \ldots, r$, that appear in the r.h.s. of Equation (14), we see that each of these terms can be the starting point to iterate the addition of observable information terms using Theorem 1. The application of Theorem 1 to a nonempty subset of a partition $\mathbf{Z}^{[r]} = \{\varnothing, \mathbf{Z}_1, \ldots, \mathbf{Z}_r\}$ converts a nontestable instrumental entropic inequality from Proposition 3 into testable.

Proposition 6 (Testable instrumental entropic inequalities from the iterative application of data processing inequalities to minInf information terms)**.** *Consider nonoverlapping sets of variables* $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{B_0}$, *and* $\mathbf{U}$, *all observable except* $\mathbf{U}$ *hidden variables. Consider that the joint distribution of these variables is generated from a causal structure that creates the independence* $\mathbf{Z} \perp \mathbf{U}|\mathbf{B_0}$, *which leads to a nontestable entropic inequality of the form* $H(\mathbf{X}|\mathbf{B_0}) \geqslant I(\mathbf{Z};\mathbf{X}|\mathbf{B_0}) + I(\mathbf{Z};\mathbf{U}|\mathbf{B_0},\mathbf{X})$. *Consider an exclusive partition in r parts of the instrumental set* $\mathbf{Z}$ *given by* $\mathbf{Z}^{[r]} = \{\varnothing, \mathbf{Z}_1, \ldots, \mathbf{Z}_r\}$, *such that* $I(\mathbf{Z};\mathbf{U}|\mathbf{B_0},\mathbf{X})$ *is separated in the sum of r nonestimable information terms* $I(\mathbf{Z}_k;\mathbf{U}|\mathbf{B}_0,\mathbf{X},\mathbf{Z}^{[k-1]})$, $k = 1, \ldots, r$. *Select* $\mathbf{E}_k = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[k-1]}\}$. *Consider a nonoverlapping subset* $\tilde{\mathbf{Z}}^{[q]} = \{\varnothing, \tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_q\} \subseteq \mathbf{Z}^{[r]}$, $0 < q \leqslant r$, *such that each* $\tilde{\mathbf{Z}}_l \in \tilde{\mathbf{Z}}^{[q]}$ *corresponds to a different* $\mathbf{Z}_i \in \mathbf{Z}^{[r]}$. *Consider that for each* $\mathbf{Z}_i \in \tilde{\mathbf{Z}}^{[q]}$ *it is possible to iteratively apply Theorem 1 with an initial assignment of its inputs* $\{\bar{\mathbf{Z}}, \mathbf{E}, \bar{\mathbf{U}}\}$ *as* $\{\mathbf{Z}_i, \mathbf{E}_i, \mathbf{U}_i\}$, *with* $\mathbf{U}_i \subseteq \mathbf{U}$. *Accordingly, for each* $\mathbf{Z}_i \in \tilde{\mathbf{Z}}^{[q]}$, *it is possible to construct collections* $\mathbf{A}_i^{[n_i]}$, $\bar{\mathbf{Z}}_i^{[n_i]}$, $\check{\mathbf{Z}}_i^{[n_i]}$, $\bar{\mathbf{U}}_i^{[n_i]}$, $\bar{\mathbf{B}}_i^{[n_i]}$, *and* $\mathbf{C}_i^{[n_i]}$, *with* $n_i > 0$, *which are associated with sets of independencies* $\bar{\mathbf{Z}}_{ij} \perp_P \mathbf{A}_{ij}|\bar{\mathbf{U}}_{ij}\bar{\mathbf{B}}_{ij}$, *for* $j = 1, \ldots, n_i$, *which are imposed by the causal structure. This leads to the testable instrumental entropic inequality*

$$H(\mathbf{X}|\mathbf{B_0}) \geqslant I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0) + \sum_{\mathbf{Z}_i \in \tilde{\mathbf{Z}}^{[q]}} \sum_{j=1}^{n_i} \min_{Q \in \Delta P_{ij}} I_Q(\bar{\mathbf{Z}}_{ij};\mathbf{A}_{ij}|\mathbf{E}_i, \mathbf{A}_i^{[j-1]}, \check{\mathbf{Z}}_i^{[j]}), \tag{24}$$

where each family of distributions ΔP_{ij} *preserves* $P(\bar{\mathbf{Z}}_{ij}, \mathbf{A}_{ij}, \bar{\mathbf{B}}_{ij})$ *and* $P(\bar{\mathbf{Z}}_{ik}, \mathbf{C}_{ik})$, *for* $k = 1, \ldots, j$.

Proof. Proposition 6 follows directly from the iterative application of Theorem 1 to a subset of the nonobservable information terms in the sum of Equation (14). In each case, the theorem is applied starting from a different set of variables $\{\bar{\mathbf{Z}}, \mathbf{E}, \bar{\mathbf{U}}\}$, namely $\{\mathbf{Z}_i, \mathbf{E}_i, \mathbf{U}_i\}$, where $\mathbf{Z}_i \in \mathbf{Z}^{[r]}$ corresponds to some $\tilde{\mathbf{Z}}_l \in \tilde{\mathbf{Z}}^{[q]}$. The variables in $\mathbf{U}$ not included in $\mathbf{U}_i$ are marginalized. Theorem 1 describes the properties that need to fulfill the collections $\mathbf{A}_i^{[n_i]}$, $\bar{\mathbf{Z}}_i^{[n_i]}$, $\check{\mathbf{Z}}_i^{[n_i]}$, $\bar{\mathbf{U}}_i^{[n_i]}$, $\bar{\mathbf{B}}_i^{[n_i]}$, and $\mathbf{C}_i^{[n_i]}$. The requirements $q > 0$ and $n_i > 0$ ensure that at least one observable information term is added in the lower bound, such that a nontrivial entropic inequality is testable. The form of the resulting testable inequality is determined by which independencies are imposed by the causal structure of interest, that is, which sets of independencies $\bar{\mathbf{Z}}_{ij} \perp_P \mathbf{A}_{ij}|\bar{\mathbf{U}}_{ij}\bar{\mathbf{B}}_{ij}$, for $i = 1, \ldots, q$, $j = 1, \ldots, n_i$ are combined to apply DP inequalities that add estimable information terms at the lower bound. □

The inequality of Proposition 6 encompasses the ones of Proposition 1, 2, and 4. It may be asked why the term $I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0)$ is always separated before starting to apply DP inequalities. In fact, in the case of Proposition 1 where the standard DP inequality is applied, this is not a differentiating factor, since the r.h.s. of Equation (1) is equal to $I(\mathbf{Z};\mathbf{X},\mathbf{Y}|\mathbf{B}_0)$. However, when a DP inequality is applied in combination with relaxations of the constraints on the marginals to be preserved, this changes. As elaborated after the proof of Theorem 1 in Appendix E, in order to obtain a lower bound as tight as possible the constraints on the marginals should always be as strong as possible, while loose enough to allow the application of the subsequent DP inequalities. Given that the term $I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0)$ is observable, a minimization after a relaxation of the constraints that would not preserve $P(\mathbf{Z},\mathbf{X},\mathbf{B}_0)$ results in an equal or smaller lower bound. This means that, to obtain the highest lower bound, DP inequalities should be applied starting with $I(\mathbf{Z};\mathbf{U}|\mathbf{B}_0,\mathbf{X})$, after the separation

of $I(\mathbf{Z};\mathbf{X}|\mathbf{B}_0)$. Accordingly, we will further illustrate in Figure A2A that $\mathbf{X}$ and $\mathbf{B}_0$ play the same role in the derivation of an estimable lower bound.

We now examine in detail the application of Proposition 6 to the example of Figure 2C. The causal structure of Figure 2C is analogous to the one of Figure 2A, with some differences: It contains an additional predictor Y_3 and a new conditioning variable W_3. The conditioning variable W_1 has been removed for simplicity of the figure, but could be left as in Figure 2A with no qualitative effect in our reasoning. For simplicity of the explanation, we now focus on the construction of an instrumental entropic inequality with instrumental set $\mathbf{Z} = \{Z\}$ and conditioning set $\mathbf{B}_0 = \{W_2, W_3\}$. We do so because this suffices to illustrate the iterative application of minInf DP inequalities with Proposition 6. See Appendix I for a more detailed analysis of this example. With $\mathbf{Z} = \{Z\}$ and $\mathbf{B}_0 = \{W_2, W_3\}$, we apply Proposition 6 with $\mathbf{A}^{[n]} = \{\varnothing, \{Y_1\}, \{Y_2\}, \{Y_3\}\}$, using the independencies $Z \perp Y_1|UXW_2W_3$, $Z \perp Y_2|UXW_3$, and $Z \perp Y_3|UXY_2$. The derived entropic inequality is

$$\begin{aligned} H(X|W_2, W_3) \geqslant & I(Z; X, Y_1|W_2, W_3) + \min_{Q\in\Delta P_2} I_Q(Z; Y_2|W_2, W_3, X, Y_1) + \\ & \min_{Q\in\Delta P_3} I_Q(Z; Y_3|W_2, W_3, X, Y_1, Y_2), \end{aligned} \tag{25}$$

where ΔP_2 preserves the marginals $\{P(Z, W_3, X, Y_2), P(Z, W_2, W_3, X, Y_1)\}$ and ΔP_3 preserves the marginals $\{P(Z, X, Y_2, Y_3), P(Z, W_3, X, Y_2), P(Z, W_2, W_3, X, Y_1)\}$. The second term in the r.h.s. is the minInf term that corresponds to $I(Z; Y_2 \backslash\backslash \{W_2, Y_1\}|W_3, X)$. Variable Y_1 is inserted thanks to $Z \perp Y_1|UXW_2W_3$, using the standard DP inequality of Lemma 1. Variable Y_2 is inserted thanks to $Z \perp Y_2|UXW_3$, using the DP inequality of unique information of Lemma 3. Finally, variable Y_3 is inserted thanks to $Z \perp Y_3|UXY_2$, using a minInf DP inequality of the form of Proposition 5.

Our objective with this example was to illustrate the iterative insertion of estimable information terms in the lower bound. As mentioned above, an extended analysis of instrumental entropic inequalities for the causal structure of Figure 2C is presented in Appendix I. This extended presentation will cover alternative entropic inequalities that are derived with the multivariate instrumental set $\mathbf{Z} = \{Z, W_2, W_3\}$, when using different partitions to create chainlike instrumental entropic inequalities of the form of Equation (14). Concretely, the inequality of Equation (25) is subsumed by the one of Equation (A17e). Appendix I provides further evidence of the increase in causal inference power thanks to the addition of minInf terms, since they allow obtaining tighter lower bounds thanks to the combination of conditional independencies that cannot be jointly used in a standard instrumental inequality.

This completes our extension of instrumental entropic inequalities, from the standard form reviewed in Proposition 1, to the minInf instrumental entropic inequalities of Proposition 6. We have used the specific scenario of instrumental inequalities to vertebrate the presentation of our core contributions, namely the theoretical derivation of DP inequalities for minInf information terms (Proposition 5) and the iterative procedure to combine them (Theorem 1). In the rest of our Results, we will more broadly show how to apply these tools to derive other entropic inequalities apart from instrumental inequalities.

While our main contribution focuses on the theoretical derivation of the properties of minInf terms that render them useful for causal structure learning, in Appendix G we also discuss their estimation. As we explain, this estimation constitutes a non-convex minimization problem [44] and a general implementation is out of the scope of this work. Nonetheless, in Lemma A2 we recast minInf terms separating a convex and non-convex component of the minimization problem. We use this approach to extend the numerical examples presented in Appendix D, which show the gain in causal inference power obtained with Proposition 4. In this way, in Appendix H we also provide numerical examples in

which it is only thanks to the addition of a second minInf term, like in Equation (25), that a rejection is obtained when testing the entropic inequality.

3.6. The Region of minInf Shannon Entropy Cones

In previous sections we have derived entropic inequalities with increased causal inference power by introducing minInf DP inequalities. We here more broadly reformulate this derivation from a geometrical perspective. In a geometrical perspective of entropy [22], entropy values associated with a set of variables $\bar{\mathbf{V}} = \{\bar{V}_1, \bar{V}_2, \ldots, \bar{V}_n\}$ are represented as a point in a $\mathbb{R}^{2^n}$ space. In more detail, given a joint distribution $P(\bar{\mathbf{V}})$, for the set of indices $[n] = \{1, 2, \ldots, n\}$ associated with the variables, an entropy value is obtained for each subset of indexes $S \subseteq [n]$. This entropy value corresponds to the joint entropy $H(\bar{\mathbf{V}}_S)$ of the marginal probability distribution $P(\bar{\mathbf{V}}_S)$. Given that the *power set* of subsets of $[n]$ contains 2^n subsets, a vector constructed with all entropy values $H(\bar{\mathbf{V}}_S)$ lies in a $\mathbb{R}^{2^n}$ space. The region in this space containing all points obtainable from probability distributions, the *entropy cone*, forms a convex cone (Theorem 15.5 in [22]), but has an unknown explicit characterization. However, an approximation of this region is given by the *Shannon entropy cone*, which includes all points that comply with the following linear inequality constraints:

$$H(\varnothing) = 0, \tag{26a}$$

$$H(\bar{\mathbf{V}}_T) \geqslant H(\bar{\mathbf{V}}_S) \text{ if } S \subseteq T, \tag{26b}$$

$$I(\bar{\mathbf{V}}_S; \bar{\mathbf{V}}_T | \bar{\mathbf{V}}_{S \cap T}) \geqslant 0, \tag{26c}$$

where S and T are two subsets of $[n]$. These inequalities are known as the *basic inequalities* [22,28] and can be expressed as linear inequalities only involving entropy terms, hence introducing constraints among different components of the entropy vectors. The basic inequalities impose requirements for any well-defined probability distribution, namely the nonnegativity of entropy (Equation (26a,b)), the monotonicity of entropy (Equation (26b)), and the nonnegativity of conditional mutual information (Equation (26c)), associated with the submodularity of entropy.

These basic inequalities are constraints that apply to any entropic vector created from a well-defined probability distribution. Furthermore, if a joint probability distribution of interest is generated under additional constraints, such as the compatibility with a certain causal structure, then the set of independencies induced by the causal structure adds extra equality constraints to the basic inequalities, namely in the form of conditional mutual information terms with zero values. In the presence of hidden variables, the cancelation of conditional mutual information terms involving the hidden variables is not verifiable. However, given the set combining the basic inequalities and the causally-induced equalities, it has been shown [14,24] that causally informative entropic inequalities, such as the standard instrumental entropic inequality, are derived by marginalization of the hidden variables to obtain inequalities that only involve observable variables. This marginalization has been algorithmically implemented using Fourier-Motzkin elimination, a standard linear programming algorithm for the elimination of variables from systems of inequalities [14].

The derivation of entropic inequalities with minInf terms can be formulated as an analogous marginalization problem, but starting from the region of a *minInf Shannon entropy cone*, which generalizes the region of the Shannon entropy cone from individual distributions to families of distributions sharing sets of constraints. In more detail, consider a minInf term

$$\min_{Q \in \Delta P} I_Q(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3), \tag{27}$$

where $\bar{\mathbf{V}}_i$, $i = 1, \ldots, 3$ are sets of variables, without specification of which variables are observable or hidden. The family of distributions ΔP is defined as preserving a set of marginals from a joint original distribution $P(\bar{\mathbf{V}})$, with $\{\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, \bar{\mathbf{V}}_3\} \subseteq \bar{\mathbf{V}}$. The minimum within ΔP determines a set of distributions, at least one, that reach the minimum. Select a distribution $Q^*(\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, \bar{\mathbf{V}}_3)$ among the ones reaching the minimum and consider the region that its entropic vector can occupy. This region is restricted by basic inequalities (Equation (26)) and also by the constraints intrinsic to the definition of the minInf term in Equation (27). First, there is an additional constraint $I_{Q^*}(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3) \leqslant I_P(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3)$, associated with the minimization, since $P \in \Delta P$ and Q^* is a minimum. Second, there is a constraint $H_P(\bar{\mathbf{V}}_S) = H_{Q^*}(\bar{\mathbf{V}}_S)$ for any $\bar{\mathbf{V}}_S$ that appears in one of the marginal distributions preserved in ΔP.

In the presence of additional constraints induced by a causal structure, the same constraints are imposed to P and Q^* for those independencies associated with variables whose joint marginals are preserved. For example, for a causally-induced independence $\bar{\mathbf{V}}_1 \perp \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3$, the constraints $I_P(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3) = 0$ and $I_{Q^*}(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3) = 0$ are imposed if $P(\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, \bar{\mathbf{V}}_3)$ is preserved in ΔP. Overall, the entropic vector associated with the original distribution P and the vector associated with Q^* are coupled. Furthermore, the set of constraints that characterizes the accessible region for entropic vectors also includes the minInf DP inequalities. This is because the derivation of these DP inequalities results from the definitions in terms of the minimization operator (see proofs of Lemma 3 and Proposition 5), that is, the DP inequality holds specifically for distributions at the minimum. Without imposing the DP inequalities as constraints to bound the region accessible to entropic vectors, the entropic vector of $Q^*(\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, \bar{\mathbf{V}}_3)$ could correspond to any distribution within ΔP fulfilling $I_{Q^*}(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3) \leqslant I_P(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3)$, without corresponding to the minimum.

This type of coupling among entropic vectors does not uniquely result from constraints involving the original distribution P. Consider two families ΔP and $\Delta \tilde{P}$, with $\Delta \tilde{P} \subseteq \Delta P$, and a term $I_Q(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3)$ to be minimized. Consider distributions $\tilde{Q}^*$ and Q^* that reach the minimum within $\Delta \tilde{P}$ and ΔP, respectively. Since $\Delta \tilde{P} \subseteq \Delta P$, this means that there is a constraint $I_{\tilde{Q}^*}(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3) \geqslant I_{Q^*}(\bar{\mathbf{V}}_1; \bar{\mathbf{V}}_2 | \bar{\mathbf{V}}_3)$. Furthermore, there are also constraints $H_{\tilde{Q}^*}(\bar{\mathbf{V}}_S) = H_{Q^*}(\bar{\mathbf{V}}_S)$ for any $\bar{\mathbf{V}}_S$ that appears in the shared preserved marginals of ΔP and $\Delta \tilde{P}$. The same happens with the additional constraints induced by a causal structure. If both ΔP and $\Delta \tilde{P}$ preserve the marginal distribution containing the variables associated with a causally-imposed conditional independence, this results in information terms with zero values, creating a further coupling between entropic vectors corresponding to distributions in the two families.

Overall, there is a coupling between entropic vectors, comprising the one of the original probability distribution and those of the probability distributions defined in terms of minimizations within families that preserve certain sets of marginals. The resulting set of constraints includes constraints of different types. First, the basic inequalities of Equation (26), which apply to all distributions. Second, inequalities related to the definition of the minInf terms within a family of distribution preserving a set of marginals. This includes equalities between entropies when two families of distributions share marginals that are preserved. It also includes inequalities between information terms when they result from the minimization within families that are one a subset of the other. Third, it includes causally-induced constraints. This includes equalities (information terms with zero value) that apply to the original distribution and to any minInf distribution that preserves the joint marginal of variables involved in a conditional independence. This implies the DP inequalities of minInf terms.

The region accessible to the vectors compatible with a certain causal structure can thus be characterized in two dual ways. Given the selection of M minInf distributions

resulting from M minimizations, one possibility is to describe the region as a set of $M+1$ interdependent entropic vectors within the $\mathbb{R}^{2^n}$ space of entropies. Another possibility is to define a $\mathbb{R}^{(2^n)(M+1)}$ space in which the entropic vectors of the original joint distribution P and of the M additional joint distributions associated with the minInf terms are appended. The latter representation constructs the minInf Shannon entropy cone.

A formal characterization of the region of minInf Shannon entropy cones is beyond the aim of this paper. However, the considerations above suggest how an algorithmic entropic characterization of causal structures [14,24] can be extended to exploit the constraints that exist in minInf Shannon entropy cones. Since the constraints involving minInf terms also constitute a linear system of equalities and inequalities, the procedure used to derive testable inequalities by the marginalization of hidden variables [14] can be extended to include minInf terms. Note that the minimization operations involved in the identification of the minInf distributions are not to be solved as part of the linear system. They are reflected in the system by the inclusion of the constraints associated with the definition of the minInf terms. This guarantees that, after the marginalization, the reduced system will contain entropic inequalities that express relations between estimable minInf terms, which can then be tested. The minInf instrumental entropic inequalities introduced in previous sections are one example of inequalities that would be obtained with this algorithmic approach. The implementation of this extended procedure is left for future work.

3.7. Other Types of Entropic Inequalities with minInf Information Terms

The implementation of a procedure to algorithmically derive causally informative inequalities with minInf terms is out of the scope of this work. However, we here point to two other well-known types of causally informative entropic inequalities that can be extended thanks to the minInf DP inequalities introduced in Section 3.4. We do not aim to provide a full presentation of these entropic inequalities, but to reframe them in a form that allows appreciating their extensions.

The first type of inequalities that we extend is the Groups-Decomposition (GD) inequalities [25,26]. We keep the notation of [26] to facilitate the comparison. This type of inequalities relates the information that a collection of variables has about a set of target variables $\mathbf{Y}$ with a weighted sum of the information contained in different groups defined as subsets of that collection. Two subtypes of GD inequalities were introduced in [25]. The existence of an inequality of the first subtype (GD1) imposes certain conditions of independence between the groups and determines the weights based on the overlap between them. The second subtype (GD2) has no requirements of independence, but only applies to collections and groups that constitute ancestral sets, that is, sets including all ancestors of their members. Several extensions were introduced in [26], comprising a relaxation of the required conditions of independence for GD1 and more flexibility in the configuration of groups for GD2. These extensions also included a generalization to allow for conditioning sets and the use of the DP inequalities of Lemmas 1 and 3 to derive testable GD inequalities from collections containing hidden variables.

We here present GD inequalities in a form that highlights how to apply the tools developed in Section 3.4, hence using minInf DP inequalities to increase their causal inference power. For simplicity, we restrict ourselves to GD1 inequalities, because the presentation of the second subtype is more mathematically heavy. We extend GD1 inequalities following Proposition 3 of [26], while an extension of GD2 analogously follows from their Theorem 2. For the purpose of this extension, we explicitly differentiate observable variables $\mathbf{V}$ and hidden variables $\mathbf{U}$. The inequality considers a target set of variables $\mathbf{Y}$ and a conditioning set $\mathbf{Z}$. It also considers a collection of variables $\mathbf{B}_{[n]} = [\mathbf{B}_1, \ldots, \mathbf{B}_n]$ formed by n groups,

possibly overlapping. Each group can contain observable and hidden variables, such that $\mathbf{B}_i = \{\mathbf{V}_i, \mathbf{U}_i\}$. The GD1 inequality states that

$$H(\mathbf{Y}|\mathbf{Z}) \geqslant \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} I(\mathbf{Y}; \mathbf{B}_i|\mathbf{Z}) = \sum_{i=1}^{n} \frac{1}{d_{\mathbf{B}_i}} [I(\mathbf{Y}; \mathbf{V}_i|\mathbf{Z}) + I(\mathbf{Y}; \mathbf{U}_i|\mathbf{Z}, \mathbf{V}_i)], \tag{28}$$

where $d_{\mathbf{B}_i} - 1$ is the number of groups that intersect with $\mathbf{B}_i$. The inequality holds if the groups fulfill the following conditions of independence: Given disjoint partitions $\mathbf{B}_i = \{\mathbf{B}_i^{(1)}, \mathbf{B}_i^{(2)}\}$, $\mathbf{B}_i^{(1)} \perp \mathbf{B}_j^{(1)} \backslash \mathbf{B}_i^{(1)} | \mathbf{Z}$ and $\mathbf{B}_i^{(2)} \perp \mathbf{B}_j \backslash \mathbf{B}_i^{(2)} | \mathbf{B}_i^{(1)} \mathbf{Z}$, for all $i \neq j$. For each group, the term $I(\mathbf{Y}; \mathbf{V}_i|\mathbf{Z})$ can be estimated, while the term $I(\mathbf{Y}; \mathbf{U}_i|\mathbf{Z}, \mathbf{V}_i)$ is analogous to the terms with hidden variables that appear in Proposition 3. Ref. [26] used the DP inequalities of mutual information and unique information to derive tighter estimable lower bounds. An example of a causal structure for which this inequality is causally fulfilled is shown in Figure 3A. In this case, the structure of a Common Ancestors (CM) graph [24] is obtained after conditioning on Z, with all dependencies between observable variables mediated by hidden variables. A GD inequality exists selecting $\mathbf{B}_i = \mathbf{B}_i^{(1)} = \{U_i\}$, such that $\mathbf{B}_i^{(1)} \perp \mathbf{B}_j^{(1)} | Z$ holds for all $i \neq j$. The standard DP inequality is applied to each group, thanks to $Y \perp V_i | Z U_i$, which leads to a testable GD inequality. Consider now the role of the terms $I(\mathbf{Y}; \mathbf{U}_i|\mathbf{Z}, \mathbf{V}_i)$ if the structure of Figure 3A is embedded as part of a larger causal structure for which other minInf DP inequalities can be applied. In that case, analogously to Figures 1 and 2, the terms $I(\mathbf{Y}; \mathbf{U}_i|\mathbf{Z}, \mathbf{V}_i)$ can be the starting point to iteratively add estimable information terms at the lower bound of the inequality, with a procedure analogous to the one enabled by Propositions 3 and 6.

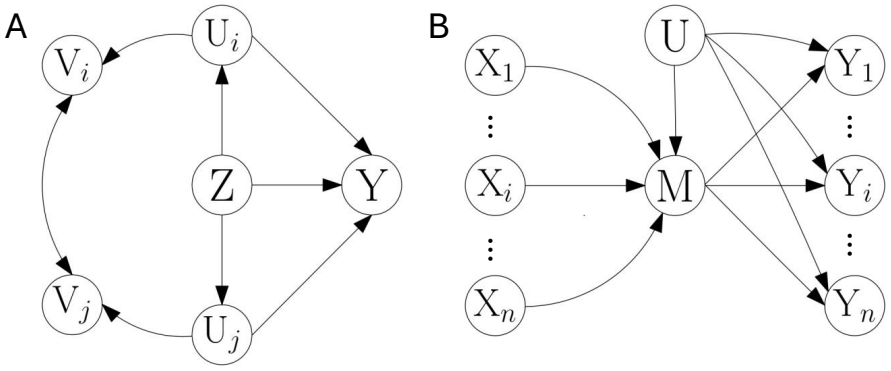


Figure 3. Other types of entropic inequalities with a form that allows applying minInf DP inequalities to add estimable information terms in the lower bound with the procedure developed in Section 3.4. (**A**) An example of causal structure associated with Groups-Decomposition (GD) inequalities [25,26]. The graph shows two representative groups $\mathbf{B}_i = \{U_i\}$, $\mathbf{B}_j = \{U_j\}$ out of a collection $\mathbf{B}_{[n]} = [\mathbf{B}_1, \ldots, \mathbf{B}_n]$, with all groups having the same structure of connectivity. An entropic inequality of the form of Equation (28) holds. (**B**) Causal structure associated with the Information Causality (IC) inequality [23,31] in the case of classical systems. An entropic inequality of the form of Equation (29) holds.

The second type of inequality to be generalized is the Information Causality (IC) inequality [23,31]. This inequality differs from the ones considered so far in that it contemplates a marginal scenario defined not only in terms of the presence of hidden variables, but also in terms of restrictions regarding which observable variables are jointly observable. The motivation of this marginal scenario is that the IC inequality was conceived to comprise also quantum systems. However, we here focus on its derivation for classical systems, since the consideration of quantum systems would require examining how to possibly adapt our results to the case in which Shannon entropy is substituted by von Neumann entropy [31]. In particular, we follow the generalization introduced by [31]. For classical

systems, the derivation of this IC inequality can be understood in relation to the causal structure of Figure 3B. We have kept the notation used in [31] to facilitate a comparison with their derivations. The only exception is that we use U for the hidden variable, consistently with our previous derivations, while in [31] it corresponds to variable B. All variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, and M are observable, with U hidden. However, the marginal scenario is defined as imposing further constraints of observability, such that the variables in $\mathbf{Y}$ have mutually exclusive observability. Only marginal distributions of the form $p(\mathbf{X}, M, Y_i)$ are observable, for each $Y_i \in \mathbf{Y}$. Accordingly, for an inequality to be testable, it can only contain information terms estimable from these marginals. We now present the IC inequality in a form that highlights how it can be extended based on the tools developed in Section 3.4:

$$\begin{aligned} H(M) - H(\mathbf{X}) + \sum_{i=1}^{n} H(X_i) &\overset{(a)}{\geqslant} \left[I(X_1; M) + \sum_{i=2}^{n} I(X_i; X_1, M) \right] + \\ \left[I(X_1; U|M) + \sum_{i=2}^{n} I(X_i; U|X_1, M) \right] &\overset{(b)}{\geqslant} \left[I(X_1; M) + \sum_{i=2}^{n} I(X_i; X_1, M) \right] + \\ \left[I(X_1; Y_1|M) + \sum_{i=2}^{n} I(X_i; Y_i|X_1, M) \right] &+ \left[I(X_1; U|M, Y_1) + \sum_{i=2}^{n} I(X_i; U|X_1, M, Y_i) \right]. \end{aligned} \tag{29}$$

Note that the selection of X_1 is arbitrary and without loss of generality it can be replaced by any other variable in $\mathbf{X}$. The detailed derivation of inequality (a) can be found in Equations (16)–(21) of [31] and follows from Lemma 2 in [24]. Apart from basic properties of entropy, the derivation relies on the causally-imposed independence $\mathbf{X} \perp U$. On the other hand, inequality (b) is derived thanks to the DP inequalities of mutual information associated with $X_1 \perp Y_1|UM$ and $X_i \perp Y_i|UMX_1$, which provide observable lower bounds of the terms $I(X_1; U|M)$ and $I(X_i; U|X_1, M)$. More generally, these terms can be the starting point to iteratively add estimable information terms at the lower bound of the inequality, with a procedure analogous to the one enabled by Propositions 3 and 6. To highlight this, we have kept in the lower bound the nonestimable information terms that contain U and that are dropped in the final testable inequality of [31]. If the structure of Figure 3B is embedded as part of a larger causal structure such that other types of minInf DP inequalities can be applied, an iterative addition of observable information terms in the lower bound can proceed as in Proposition 6.

4. Discussion

In this work we have explored how causally informative entropic inequalities can be extended to contain minimum information (minInf) entropic terms. We have first examined how to combine the standard data processing (DP) inequality and a DP inequality for the maximum entropy unique information [26,29] to create new and tighter instrumental entropic inequalities [24]. In this way, we identified a procedure to recursively combine different types of DP inequalities to introduce additional observable terms in the lower bound of information terms that contain hidden variables. We then introduced a DP inequality for a general type of minInf information terms defined by information minimization within families of distributions that preserve sets of marginals shared with the original distribution. We have shown how to recursively apply these DP inequalities to exploit sets of independencies with different conditioning sets. We then have used this procedure to build minInf instrumental entropic inequalities that provide additional causal inference power. While our development of causally informative entropic inequalities uses as vector for its derivation the instrumental causal scenario, the procedure presented in Theorem 1

to recursively combine DP inequalities is general. We have exemplified this for two other types of entropic inequalities, namely the Groups-Decomposition inequality [25,26], and the Information Causality inequality [23,31].

More generally, we have also reframed the use of minInf entropic terms to derive causally informative entropic inequalities from a geometrical perspective [22]. Entropic inequalities can be systematically derived as a marginalization problem [14]. To derive causally informative inequalities that only involve entropic terms of the original distribution, this marginalization operates on the set of constraints that defines the Shannon entropy cone [22], in combination with the additional conditional independence constraints imposed by the causal structure. We have indicated that, in order to incorporate minInf entropic terms in this geometrical approach, Shannon entropy cones can be extended to not only contain entropic terms from the original distribution, but to jointly combine the minInf entropic terms associated with a set of families of distributions that share sets of marginals with it. The derivation of causally informative inequalities with minInf terms can then proceed analogously as a marginalization problem, now departing from a set of constraints that comprises those that the conditional independencies in the causal structure impose to the minInf entropic terms, such as the minInf DP inequalities we have derived. While we have conceptualized this procedure, its implementation in a linear programming algorithm [14,22] remains out of the scope of this work.

For the instrumental causal scenario, we have provided a range of examples of causal structures for which causal inference power is increased with the new tests, through the examples of the Figures in the main text and further examples in Appendices F and I. With an increasing number of causally informative inequality tests, an important question is how to determine minimal sets of tests that preserve the overall causal inference power. For this purpose, we have stated a criterion to determine when a new inequality test can add causal inference power to a set of tests (Appendix B). We have used this criterion to derive a hierarchy for instrumental inequalities with multivariate instrumental sets that encompasses the comparison of certain types of tests. On the other hand, we have also provided examples for which various minInf entropic inequalities are complementary among them, potentially providing each additional power.

Our contribution has focused on the derivation of causally informative minInf entropic inequalities for which the involved families of distributions are determined by constraints on shared marginals. Alternatively, the minInf terms could be determined comprising other types of constraints. Specifically for the unique information term developed within the decomposition of mutual information into redundant, unique, and synergistic components, alternative approaches [39,41,45] define unique information within families that combine other constraints. Future work should determine if alternative entropic inequalities are derived with these measures and the degree to which they contribute to increase causal inference power.

Our approach can also benefit the study of causal interactions in dynamical systems [14,46]. High-dimensional multivariate dynamical processes appear in many domains of interest, such as brain dynamics, e.g., [47–49] or econometrics [50,51]. For time-series, methods of causal inference both in the temporal [52] and spectral domain [53,54] predominantly rely on the detection of conditional independencies between observable signals, and hence are affected by hidden variables due to the partial observability of complex systems. Our treatment of the instrumental causal scenario suggests that our methods can help in these cases, given that concatenated instrumental-like causal motifs frequently appear in temporal dynamics, as represented for example by autoregressive moving-average models [52]. The unique information measure has already been applied to neural data, e.g., [55,56]. However, applicability of entropic inequalities with minInf terms to dynamical

systems will require further study of the scalability of our methods to high-dimensional data. The applicability of entropic inequalities to time-series will also need to incorporate further assumptions such as stationarity [57] to make the approach operational.

Beyond applications to classical systems, it remains to be seen how to adapt the formulation of minInf entropic inequalities for causal inference in the quantum domain [16,31,58,59]. In our work, we have considered how to extend Information Causality inequalities [23] using minInf DP inequalities, but we focused on their formulation for classical systems [31]. Quantum mutual information defined in terms of von Neumann entropy [60] fulfills the standard DP inequality as well as the chain rule property, which is involved in the derivation of Theorem 1. Nonetheless, future work is required to pursue an adaptation to quantum systems.

In this work we have focused on causally informative entropic inequalities. We have mostly focused on the instrumental causal scenario to introduce new minInf entropic inequalities and examine them jointly with standard inequalities constructed with multivariate instrumental sets. More broadly, an important question regards the embedding of this type of causally informative entropic inequalities with other approaches to test the compatibility of causal structures with data [61–63]. The fact that the new inequalities appear as constraints only in the extended minInf Shannon entropy cones indicates that they cannot be reduced to constraints in the standard Shannon entropy cone of the original joint distribution. However, a separate question is how to construct minimal sets of inequality constraints with equivalent inference power, and how these sets would combine or prioritize the use of minInf tests and tests that operate in an inflated standard Shannon entropy cone [64]. The integration of different families of tests under the criterion that determines when the addition of a new test increases the causal inference power of a set of tests stands as a goal to determine the boundaries of distinguishably between causal structures from data.

Funding: This research received no external funding.

Data Availability Statement: No datasets used. Code used can be provided upon request.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proofs of Monotonicity and Data Processing Inequality of the Unique Information

Proof of Lemma 2. Consider the distribution $P_{\mathbf{DD'}} \equiv P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{D}_2, \mathbf{O}_1)$ and its marginal $P_{\mathbf{D}} \equiv P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}_2, \mathbf{O}_1)$. Consider the family $\Delta P_{\mathbf{DD'}}$ of distributions that preserve $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{D}_2, \mathbf{O}_1)$, and the family $\Delta P_{\mathbf{D}}$ that preserves $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{D}_2, \mathbf{O}_1)$. Consider any distribution $Q_{\mathbf{DD'}} \in \Delta P_{\mathbf{DD'}}$ and its marginal $Q_{\mathbf{D}}$ on $\{\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D}_2, \mathbf{O}_1\}$. Then $Q_{\mathbf{D}} \in \Delta P_{\mathbf{D}}$. By monotonicity of the mutual information, the information $I_{Q_{\mathbf{DD'}}}(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{D}_2, \mathbf{O}_1)$ is lower than or equal to $I_{Q_{\mathbf{DD'}}}(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'}|\mathbf{D}_2, \mathbf{O}_1)$. Since $I_{Q_{\mathbf{DD'}}}(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{D}_2, \mathbf{O}_1)$ does not have $\mathbf{D'}$ as an argument, it is equal to $I_{Q_{\mathbf{D}}}(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{D}_2, \mathbf{O}_1)$. Since this holds for any distribution in $\Delta P_{\mathbf{DD'}}$, it holds in particular for the distribution $Q^*_{\mathbf{DD'}}$ that minimizes $I(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'}|\mathbf{D}_2, \mathbf{O}_1)$ in $\Delta P_{\mathbf{DD'}}$. Since $Q^*_{\mathbf{D}}$ belongs to $\Delta P_{\mathbf{D}}$, the minimum of $I(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{D}_2, \mathbf{O}_1)$ in $\Delta P_{\mathbf{D}}$ is equal to or smaller than $I_{Q^*_{\mathbf{D}}}(\bar{\mathbf{Z}}; \mathbf{D}|\mathbf{D}_2, \mathbf{O}_1)$ and hence equal to or smaller than $I_{Q^*_{\mathbf{DD'}}}(\bar{\mathbf{Z}}; \mathbf{D}, \mathbf{D'}|\mathbf{D}_2, \mathbf{O}_1)$. □

Proof of Lemma 3. Let $P_{\mathbf{DD'}} \equiv P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{E})$, with $\mathbf{E} = \{\mathbf{D}_2, \mathbf{O}_1\}$, be the original distribution of the variables and define $\Delta P_{\mathbf{DD'}}$ as the family of distributions on $\{\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{E}\}$ that preserve the two marginals $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{D'}, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{E})$. Let $P_{\mathbf{D}} \equiv P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{E})$ be the marginal of $P_{\mathbf{DD'}}$ and $\Delta P_{\mathbf{D}}$ be the family of distributions that preserve the marginals $P(\bar{\mathbf{Z}}, \mathbf{D}, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{E})$. By definition of unique information (Equation (5))

$$I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D}_2|\mathbf{O}_1) \equiv \min_{Q_{\mathbf{DD}'}\in\Delta P_{\mathbf{DD}'}} I_{Q_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'|\mathbf{E}) \overset{(a)}{=} \min_{Q_{\mathbf{DD}'}\in\Delta P_{\mathbf{DD}'}} \Big[I_{Q_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}|\mathbf{E}) + I_{Q_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}'|\mathbf{D},\mathbf{E}) \Big] \overset{(b)}{=} \min_{Q_{\mathbf{DD}'}\in\Delta P_{\mathbf{DD}'}} \Big[I_{Q_{\mathbf{D}}}(\bar{\mathbf{Z}};\mathbf{D}|\mathbf{E}) + I_{Q_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}'|\mathbf{D},\mathbf{E}) \Big]. \tag{A1}$$

Equality (a) follows from the chain rule of mutual information. Equality (b) holds because $I_{Q_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}|\mathbf{E})$ does not depend on $\mathbf{D}'$ and can be calculated with $Q_{\mathbf{D}}$, marginalizing $Q_{\mathbf{DD}'}$ on $\mathbf{D}'$. Note that $Q_{\mathbf{D}} \in \Delta P_{\mathbf{D}}$. Since $I_{P_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}'|\mathbf{D},\mathbf{O}_1) = 0$, $P(\bar{\mathbf{Z}},\mathbf{D},\mathbf{D}',\mathbf{O}_1)$ factorizes as $P(\mathbf{D}'|\mathbf{D},\mathbf{O}_1)P(\bar{\mathbf{Z}},\mathbf{D},\mathbf{O}_1)$. For any distribution $\tilde{Q}_{\mathbf{D}} \in \Delta P_{\mathbf{D}}$, which preserves $P(\bar{\mathbf{Z}},\mathbf{E})$ and $P(\bar{\mathbf{Z}},\mathbf{D},\mathbf{O}_1)$, a distribution can be constructed as $\tilde{Q}_{\mathbf{DD}'} \equiv P(\mathbf{D}'|\mathbf{D},\mathbf{O}_1)\tilde{Q}_{\mathbf{D}}$, such that $\tilde{Q}_{\mathbf{DD}'} \in \Delta P_{\mathbf{DD}'}$, since $\tilde{Q}_{\mathbf{DD}'}$ continues to preserve $P(\bar{\mathbf{Z}},\mathbf{E})$, and $P(\bar{\mathbf{Z}},\mathbf{D},\mathbf{D}',\mathbf{O}_1)$ is preserved by construction. Also by construction, $I_{\tilde{Q}_{\mathbf{DD}'}}(\bar{\mathbf{Z}},\mathbf{D}_2;\mathbf{D}'|\mathbf{D},\mathbf{O}_1) = 0$ and hence $I_{\tilde{Q}_{\mathbf{DD}'}}(\bar{\mathbf{Z}};\mathbf{D}'|\mathbf{D},\mathbf{E}) = 0$ for any $\tilde{Q}_{\mathbf{DD}'}$ created from any $\tilde{Q}_{\mathbf{D}} \in \Delta P_{\mathbf{D}}$. In particular, this holds for the distribution $\tilde{Q}^*_{\mathbf{DD}'}$ constructed from $\tilde{Q}^*_{\mathbf{D}}$ that minimizes $I_{\tilde{Q}_{\mathbf{D}}}(\bar{\mathbf{Z}};\mathbf{D}|\mathbf{E})$, which determines $I(\bar{\mathbf{Z}};\mathbf{D}\backslash\backslash\mathbf{D}_2|\mathbf{O}_1)$. The distribution $\tilde{Q}^*_{\mathbf{DD}'}$ minimizes the first term in the r.h.s. of Equation (A1) and, given the nonnegativity of mutual information, it also minimizes the second term, hence providing the minimum in $\Delta P_{\mathbf{DD}'}$. Accordingly, $I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D}_2|\mathbf{O}_1) = I(\bar{\mathbf{Z}};\mathbf{D}\backslash\backslash\mathbf{D}_2|\mathbf{O}_1)$. The monotonicity of the unique information in its second argument (Lemma 2) leads to $I(\bar{\mathbf{Z}};\mathbf{D},\mathbf{D}'\backslash\backslash\mathbf{D}_2|\mathbf{O}_1) \geqslant I(\bar{\mathbf{Z}};\mathbf{D}'\backslash\backslash\mathbf{D}_2|\mathbf{O}_1)$. □

Appendix B. Sets of Entropic Inequality Tests with Complementary Causal Inference Power

A set of causal inequality tests contains complementary inequalities if each inequality can potentially contribute to increase the causal inference power, that is, each test can potentially discard a causal structure not rejected by the other tests. We here formalize this idea:

Remark A1 (Lack of additional causal inference power of an inequality for causal structures compatible with a set of independencies)**.** *Consider a set of observable variables* $\mathbf{V}$ *that fulfill a set of testable conditional independencies* $\mathbf{I}_{\mathbf{V}}$*. Consider a set of causal inequality tests* $\mathbf{T}_{\mathbf{V}}$ *that provide sufficient conditions to potentially reject some causal structure from the set* $\mathbb{G}(\mathbf{I}_{\mathbf{V}})$ of causal structures compatible with $\mathbf{I}_{\mathbf{V}}$*. An inequality test* $t_{\mathbf{V}} \in \mathbf{T}_{\mathbf{V}}$ *does not provide additional causal inference power if and only if: (i) All the causal structures in* $\mathbb{G}(\mathbf{I}_{\mathbf{V}})$ *causally impose the fulfillment of the inequality in* $t_{\mathbf{V}}$ *or (ii) The inequality is not causally imposed by all causal structures in* $\mathbb{G}(\mathbf{I}_{\mathbf{V}})$ *but for each causal structure in* $\mathbb{G}(\mathbf{I}_{\mathbf{V}})$ *that causally imposes the fulfillment of the inequality in* $t_{\mathbf{V}}$ *there is another subset of inequality tests* $\mathbf{T}'_{\mathbf{V}} \subset \mathbf{T}_{\mathbf{V}}$ *(not necessarily the same for all causal structures) such that the inequalities in* $\mathbf{T}'_{\mathbf{V}}$ *are also causally imposed and the statistical rejection of* $t_{\mathbf{V}}$ *from data is sufficient for the rejection of at least a* $t'_{\mathbf{V}} \in \mathbf{T}'_{\mathbf{V}}$ *from data.*

Remark A1 indicates that a test $t_{\mathbf{V}}$ can only increase causal inference power if it can potentially discard a causal structure that could not otherwise be discarded by other tests. This can occur in two ways. First, there is some causal structure for which no other inequality is causally imposed by that structure. In this case, $t_{\mathbf{V}}$ is the only test that if rejected allows discarding the causal structure. Second, for all causal structures that impose the fulfillment of $t_{\mathbf{V}}$ there is also another subset $\mathbf{T}'_{\mathbf{V}}$ whose fulfillment is imposed by the causal structure. In this case, $t_{\mathbf{V}}$ only adds power if it exists a causal structure, other than the ones for which $t_{\mathbf{V}}$ is causally imposed, such that potentially no other inequality from $\mathbf{T}'_{\mathbf{V}}$ is violated when tested in data generated with that causal structure, while $t_{\mathbf{V}}$ is violated. On the other hand, if the violation of $t_{\mathbf{V}}$ always leads to the violation of some other $t'_{\mathbf{V}} \in \mathbf{T}'_{\mathbf{V}}$, there is no context in which $t_{\mathbf{V}}$ allows discarding a causal structure that is not already discarded by another test.

Note that Remark A1 does not consider the estimation properties of the information measures, and it is out of the focus of our work to address estimation considerations. For example, consider that $t_{\mathbf{V}}$ has in its upper bound the sum of upper bounds from a subset $\mathbf{T}'_{\mathbf{V}}$ of causally-imposed inequalities, and in its lower bound the sum of the corresponding lower bounds. Then the violation of $t_{\mathbf{V}}$ implies the violation of at least a test $t'_{\mathbf{V}} \in \mathbf{T}'_{\mathbf{V}}$, and hence this type of test consisting on additions of tests never adds causal inference power for causal graphs that causally impose all inequalities in $\mathbf{T}'_{\mathbf{V}}$. This does not preclude the fact that, despite this lack of additional power, a test consisting on an addition of tests may in practice still be useful when considering the estimation properties of the measures, the size of the data set, and practical considerations about the false rejection rate and false acceptance rate that are desirable.

Appendix C. A Further Comparison of Instrumental Inequalities in Figure 2B

We here analyze in further detail Figure 2B in order to provide an example in which, in contrast to Figure 2A, no additional causal inference power is gained combining the DP inequality of conditional mutual information and the one of unique information. As mentioned in Section 3.3, we compare the inequality resulting from $\mathbf{Z} = Z$, $\mathbf{B}_0 = \{W_1, W_2\}$ and the one resulting from $\mathbf{Z} = \{Z, W_2\}$, $\mathbf{B}_0 = \{W_1\}$. The obtained inequalities are:

$$H(X|W_1, W_2) \geqslant I(Z; X|W_1, W_2) + I(Z; Y_1|W_1, W_2, X) + I(Z; Y_2 \backslash\backslash W_2|W_1, X, Y_1), \tag{A2a}$$

$$H(X|W_1) \geqslant I(Z, W_2; X|W_1) + I(Z, W_2; Y_1|W_1, X) + I(Z; Y_2|W_1, X, Y_1). \tag{A2b}$$

Inequality A2a corresponds to Equation (17), with $\mathbf{Z} = Z$, $\mathbf{B}_0 = \{W_1, W_2\}$, $\mathbf{B}_1 = W_1$, $\mathbf{B}_2 = W_2$, and $\bar{\mathbf{Y}} = \varnothing$. Inequality A2b is derived with $\mathbf{Z} = \{Z, W_2\}$, $\mathbf{B}_0 = \{W_1\}$, starting from $I(Z, W_2; U|W_1, X)$ instead of from $I(Z; U|W_1, W_2, X)$. In Equation (A2a), the term with Y_1 is introduced thanks to $Z \perp Y_1|UXW_1W_2$, while in Equation (A2b) the term with Y_1 is introduced thanks to $\{Z, W_2\} \perp Y_1|UXW_1$. Both terms are obtained with the DP inequality of conditional mutual information. The difference appears in the derivation of the term with Y_2. We have seen in the proof of Proposition 4 that in Equation (A2a) this term is obtained thanks to the DP inequality of unique information, removing W_2 from the conditioning set to exploit $Z \perp Y_2|UXW_1Y_1$, oppositely to $Z \not\perp Y_2|UXW_1W_2Y_1$. In contrast, in Equation (A2b), the term with Y_2 is obtained directly from the marginalization of W_2 in $I(Z, W_2; U|W_1, X, Y_1)$, which allows applying again the DP inequality of conditional mutual information to exploit $Z \perp Y_2|UXW_1Y_1$. As expected, given the invariance of the upper bound to an exchange of variables between $\mathbf{Z}$ and $\mathbf{B}_0$, the same upper bound $H(X|W_1, W_2, Z)$ is obtained in Equation (A2a,b) after moving the first term of the r.h.s. to the l.h.s. On the other hand, $I(Z; Y_1|W_1, W_2, X)$ is smaller or equal than $I(Z, W_2; Y_1|W_1, X)$ by the chain rule of mutual information, and $I(Z; Y_2 \backslash\backslash W_2|W_1, X, Y_1)$ is smaller or equal than $I(Z; Y_2|W_1, X, Y_1)$ by construction of the unique information (Equation (6)). Therefore, the inequality of Equation (A2b) is always violated when the one of Equation (A2a) is violated.

This analysis highlights that in Figure 2A it is the fact that $Z \perp Y_1|UXW_1W_2$ holds but $Z \not\perp Y_1|UXW_1$ what introduces the necessity to use the DP inequality of the unique information combined with the one of the conditional mutual information. Here, since $\{Z, W_2\} \perp Y_1|UXW_1$ holds, there is no need to condition on W_2 to introduce the term with Y_1, and hence no need to remove W_2 from the conditioning set using the unique information DP inequality to exploit $Z \perp Y_2|UXW_1$. Overall, $\{Z, W_2\} \perp Y_1|UXW_1$ and $Z \perp Y_2|UXW_1$ are combined through the marginalization of W_2, while the conditioning sets in $Z \perp Y_1|UXW_1W_2$ and $Z \perp Y_2|UXW_1$ require the combination of the different types of DP inequalities.

Appendix D. Estimation and Numerical Examples of Entropic Inequalities with Unique Information Terms

In Section 3.3 we have discussed the examples of causal structures of Figure 2A,B, which causally impose instrumental entropic inequalities of the form of Proposition 4. Here we provide some numerical examples of violations of the inequalities, that is, examples in which any causal structure that causally imposes their fulfillment can be discarded as the generative structure that underlies the observed variables.

Apart from the standard estimation of entropy and mutual information terms, a test of the form of Equation (17) also requires the estimation of the maximum entropy unique information. In general, the estimation of minInf information terms corresponds to a constrained minimization problem, where the constraints on the preservation of the marginals constitute a set of affine equality constraints. Therefore, whether the optimization problem is convex or non-convex depends on whether the mutual information term to be minimized is a convex function of the probability distributions within the family ΔP [65].

Specifically for the case of the unique information (Equation (5)), the mutual information minimization problem on $I(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{O}_1, \mathbf{D}_2)$ can be recast as an entropy maximization problem since the constraints of ΔP impose the preservation of $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{E})$, with $\mathbf{E} = \{\mathbf{D}_2, \mathbf{O}_1\}$. Concretely, preserving $P(\bar{\mathbf{Z}}, \mathbf{E})$ makes the entropy $H(\bar{\mathbf{Z}}|\mathbf{E})$ constant within ΔP, and hence minimizing $I(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{O}_1, \mathbf{D}_2)$ corresponds to maximizing $H(\bar{\mathbf{Z}}|\mathbf{E}, \mathbf{D}_1)$. This conditional entropy is concave on ΔP [29], and hence the constrained minimization of the mutual information is a convex optimization problem, which guarantees convergence towards a global minimum. Several software packages are available to numerically estimate the maximum entropy unique information [66] (https://github.com/dit/dit (accessed on 12 April 2026)), ref. [67] (https://github.com/Abzinger/MAXENT3D_PID (accessed on 12 April 2026)), or [55] (https://github.com/epiasini/SubPID/ (accessed on 12 April 2026)).

Apart from applications of convex optimization tools to estimate unique information terms, for some specific types of multivariate probability distributions, such as Gaussian distributions, expressions of the unique information have been analytically derived [68,69] in terms of standard mutual information terms. We will here focus on these cases to facilitate the examination of numerical examples. As we will further discuss in Appendix G, the estimation of other minInf terms that are not the unique information constitutes a non-convex optimization problem. By concentrating on Gaussian distributions we will provide some numerical examples also including other minInf terms, hence examining the effect of iteratively adding more terms to an entropic inequality, as it is done in Equation (25). We note that the focus of our work is the theoretical derivation of data processing inequalities and entropic inequalities with minInf information terms, and that the full implementation of numerical optimization tools to estimate these terms is beyond our scope.

For a multivariate Gaussian system with a univariate target variable $\bar{Z}$ it has been shown [68] that the maximum entropy unique information $I(\bar{Z}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1)$ is determined in terms of the mutual information terms $I(\bar{Z}; \mathbf{D}_1|\mathbf{O}_1)$ and $I(\bar{Z}; \mathbf{D}_2|\mathbf{O}_1)$. In more detail, for ΔP that preserves the marginals $P(\bar{Z}, \mathbf{D}_1, \mathbf{O}_1)$ and $P(\bar{Z}, \mathbf{E})$, with $\mathbf{E} = \{\mathbf{D}_2, \mathbf{O}_1\}$, in this case the unique information can be expressed as

$$I(\bar{Z}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1) = \min_{Q \in \Delta P} I_Q(\bar{Z}; \mathbf{D}_1|\mathbf{E}) = \max\{I(\bar{Z}; \mathbf{D}_1|\mathbf{O}_1) - I(\bar{Z}; \mathbf{D}_2|\mathbf{O}_1), 0\}, \tag{A3}$$

that is, $I(\bar{Z}; \mathbf{D}_1 \backslash\backslash \mathbf{D}_2|\mathbf{O}_1)$ is the additional information that $\mathbf{D}_1$ has, or zero if $\mathbf{D}_2$ has more information than $\mathbf{D}_1$. Furthermore, for Gaussian variables entropy terms are determined by the second-order moments of the distributions, e.g., [70]. Concretely, the terms that appear in an instrumental entropic inequality of the form of Equation (17) are such that the conditional entropy only depends on the determinant of the corresponding conditional

covariance matrix and the information terms are determined by corresponding partial correlation coefficients [27].

To illustrate the additional causal inference power gained with entropic inequalities of the form of Proposition 4, we analyze a causal structure as in Figure 2A, with the addition of a hidden confounder $X \leftrightarrow W_2$. For simplicity, $W_1 = \varnothing$ is chosen, to avoid an additional conditioning variable. For this structure, we compare two entropic inequalities that are causally imposed:

$$H(X|Z) \geqslant I(Z;Y_2|X) \tag{A4a}$$

$$H(X|Z,W_2) \geqslant I(Z;Y_1|X,W_2) + I(Z;Y_2\backslash\backslash\{W_2,Y_1\}|X) = \tag{A4b}$$
$$I(Z;Y_1|X,W_2) + \max\{I(Z;Y_2|X) - I(Z;W_2,Y_1|X), 0\}.$$

The inequality of Equation (A4a) is a standard entropic inequality that follows from $Z \perp U$ and $Z \perp Y_2|UX$. The inequality of Equation (A4b) is an entropic inequality of the form of Equation (17) that follows from $Z \perp U|W_2$, $Z \perp Y_2|UX$, and $Z \perp Y_1|UXW_2$. We will refer to the tests associated with these inequalities as Test 1 and 2, respectively. The last expression at the r.h.s. of Equation (A4b) is specific for Gaussian variables, following Equation (A3). When $I(Z;Y_2|X)$ is bigger than $I(Z;W_2,Y_1|X)$ the inequality of Equation (A4b) has the form

$$H(X|Z,W_2) \geqslant I(Z;Y_2|X) - I(Z;W_2|X). \tag{A5}$$

Comparing Equations (A4a) and (A5), we see that for Gaussian variables with a univariate $\mathbf{Z}$ the r.h.s. is always smaller in Test 2. It is for this reason that we here examine a causal structure with the additional connection $X \leftrightarrow W_2$, since in Figure 2A $W_2 \perp X|Z$ results in $H(X|Z,W_2) = H(X|Z)$ and Test 2 would always have less power than Test 1. Note that this relation between the tests is specific of the form of Test 2 in this Gaussian case.

In Figure A1 we compare the results of Tests 1 and 2 when additional connections $Z \to Y_1$ and $Z \to Y_2$ are added. The addition of the connection $Z \to Y_2$ leads to $Z \not\perp Y_2|UX$, and hence Test 1 from Equation (A4a) is not causally fulfilled and can be violated. Similarly, either the addition of $Z \to Y_2$ or $Z \to Y_1$ can lead to the violation of Test 2 caused by $Z \not\perp Y_2|UX$ or $Z \not\perp Y_1|UXW_2$.

We generate the variables as a system of linear equations with Gaussian noise. Given the large number of parameters in the system, we keep fixed the values of most parameters and examine concrete settings in which the strength of the connection $Z \to Y_2$, determined by the coefficient a_{y_2z}, is changed together with a single other parameter. In more detail, we set a default configuration in which all coefficients in the system are set to 1. The standard deviation of all hidden confounders is also set to 1, and their mean set to zero. Apart from Z, that has no parents, all other nodes are generated having also some exogenous independent noises, all of which are generated with mean zero and by default standard deviation 0.25. In each row of Figure A1 we explore different modifications of this default configuration. In all rows the coefficient a_{y_2z} is changed from 1 to 50 with 0.5 increments. In the first row, we also modify the strength of the confounder U' in $X \leftrightarrow W_2$, choosing $b \equiv a_{xu'} = a_{w_2u'}$. The explored values appear in the abscissa of Figure A1C. In the second row, we set $b = 2$ and modify the strength of connections $Z \to X$, $Z \to W_2$, and $U \to X$, choosing $c \equiv a_{xz} = a_{w_2z} = a_{xu}$. The explored values appear in the abscissa of Figure A1F. In the third row, the default configuration is modified only changing the standard deviation of the exogenous noises (v), which was fixed to 0.25. The explored values appear in the abscissa of Figure A1I.

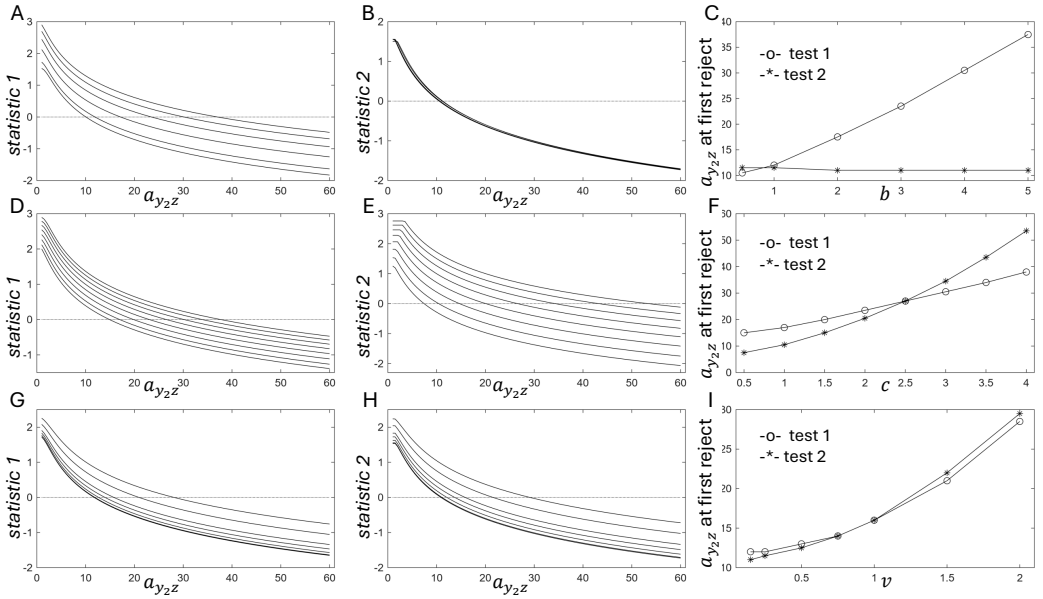


Figure A1. Examples of applications of entropic inequalities to reject the compatibility of causal structures with data. The fulfillment of inequalities in Equation (A4) is tested for a causal structure like in Figure 2A, with the addition of a connection $X \leftrightarrow W_2$ and connections $Z \rightarrow Y_1$, $Z \rightarrow Y_2$. Each row illustrates changes in the results of the entropic inequality tests for a concrete set of modifications of the default configuration, as it is described in the main text. The first column shows the values of the statistic of Test 1, corresponding to the difference between the l.h.s. and r.h.s. of Equation (A4a). The second column shows the values of the statistic of Test 2, the difference between the l.h.s. and r.h.s. of Equation (A4b). The third column shows the minimal strength of a_{y_2z} that results in a negative statistic and hence in the violation of a test and the rejection of causal structures that causally impose its fulfillment. Note that the concrete configuration to which each line in the first and second column belongs is retrieved from the value of the parameters in the abscissas of the third column.

Figure A1 provides numerical examples of how Test 2, the entropic inequality that includes the unique information term, does increase causal inference power. We have chosen configurations in which the strength of $a_{y_1z} = 1$ is not enough by itself for a rejection in Test 2. This is reflected in the positive values of the tests when a_{y_2z} is small, and indicates that for none of all the configuration studied Test 2 is rejected if removing the term $I(Z; Y_2 \backslash\backslash \{W_2, Y_1\}|X)$ from the r.h.s. of the inequality. Accordingly, all obtained negative values of Test 2, and hence the additional causal power, come from the addition of the unique information term. We can compare the rejections of Test 1 and 2 when increasing the strength of a_{y_2z}. We see in Figure A1C that for any value of b higher than 1 Test 2 is more powerful than Test 1. Also in Figure A1F,I Test 2 is more powerful than Test 1 for some range of the explored configurations. In all these cases, causal information is gained for configurations for which Test 1 would not allow rejecting the causal structures that result in $Z \perp Y_2|UX$, because the strength of a_{y_2z} is not sufficiently high.

In Appendix H we will resume this analysis to illustrate how additional causal inference power can be gained incorporating additional minInf terms. For this purpose, in Appendix G we address the estimation of other minInf terms.

Appendix E. Proof of Theorem 1

We here provide the proof of Theorem 1. As a preliminary, we introduce a chain rule inequality for minInf mutual information terms.

Lemma A1 (Chain rule inequality for minInf mutual information terms)**.** *Given sets of variables* $\mathbf{V}_1$, $\mathbf{V}_2$, $\mathbf{V}_3$, *and* $\mathbf{E}$, *and a family of distributions* ΔP *preserving some marginals of the distribution* $P(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \mathbf{E})$,

$$\min_{Q\in\Delta P} I_Q(\mathbf{V_1}; \mathbf{V}_2, \mathbf{V}_3|\mathbf{E}) \geqslant \min_{Q\in\Delta P} I_Q(\mathbf{V}_1; \mathbf{V}_2|\mathbf{E}) + \min_{Q\in\Delta P} I_Q(\mathbf{V_1}; \mathbf{V}_3|\mathbf{E}, \mathbf{V}_2). \tag{A6}$$

Proof. The chain rule inequality follows from the chain rule equality that applies to all distributions $Q \in \Delta P$ and the fact that the sum of the minima is smaller than or equal to the minimum of the sum. □

Note that Lemma A1 holds independently of which is the set of marginals to be preserved in ΔP. The mutual information is symmetrical in its first and second argument, and the chain rule equality of mutual information is equally applied to the first or second argument. This symmetry implies that the chain rule inequality of Lemma A1 also holds when used to separate variables in the first or second argument. This is the case even if the definition of which marginals are preserved breaks the symmetry between the arguments, e.g., as it happens in the definition of the unique information, which identifies the first argument as the set of target variables.

We now prove Theorem 1.

Proof of Theorem 1. We first consider the case of $k = 1$ when $\bar{\mathbf{B}}'_1 = \{\mathbf{A}_0, \bar{\mathbf{B}}_0\} = \mathbf{E}$, which leads to $\bar{\mathbf{B}}_1 = \{\mathbf{E}, \check{\mathbf{Z}}_1\} = \mathbf{C}_1$. The family ΔP_0 preserves $P(\bar{\mathbf{Z}}, \bar{\mathbf{U}}, \mathbf{E})$, given $\bar{\mathbf{Z}}_0 = \bar{\mathbf{Z}}$, $\bar{\mathbf{U}}_0 = \bar{\mathbf{U}}$. On the other hand, ΔP_1 preserves $P(\bar{\mathbf{Z}}_1, \bar{\mathbf{U}}, \mathbf{A}_1, \mathbf{E}, \check{\mathbf{Z}}_1)$, given $\bar{\mathbf{U}}_1 = \bar{\mathbf{U}}$. Furthermore, $\mathbf{A}^{[0]} = \varnothing$ and $\check{\mathbf{Z}}^{[0]} = \varnothing$, and hence at the l.h.s. of Equation (23) for k = 1, $\min_{Q\in\Delta P_0} I_Q(\bar{\mathbf{Z}}_0; \bar{\mathbf{U}}_0|\mathbf{E}, \mathbf{A}^{[0]}, \check{\mathbf{Z}}^{[0]}) = I(\bar{\mathbf{Z}}; \bar{\mathbf{U}}|\mathbf{E})$. Since $\{\bar{\mathbf{Z}}_1, \check{\mathbf{Z}}_1\} \subseteq \bar{\mathbf{Z}}$, the chain rule of mutual information guarantees that $I(\bar{\mathbf{Z}}; \bar{\mathbf{U}}|\mathbf{E}) \geqslant I(\bar{\mathbf{Z}}_1; \bar{\mathbf{U}}|\mathbf{E}, \check{\mathbf{Z}}_1)$. Subsequently, the standard DP inequality of Lemma 1 allows adding $\mathbf{A}_1$ to $\bar{\mathbf{U}}$ given the independence $\bar{\mathbf{Z}}_1 \perp_P \mathbf{A}_1|\bar{\mathbf{U}}\bar{\mathbf{B}}_1$, with $\bar{\mathbf{B}}_1 = \{\mathbf{E}, \check{\mathbf{Z}}_1\}$. Therefore, $I(\bar{\mathbf{Z}}_1; \bar{\mathbf{U}}|\mathbf{E}, \check{\mathbf{Z}}_1) = I(\bar{\mathbf{Z}}_1; \bar{\mathbf{U}}, \mathbf{A}_1|\mathbf{E}, \check{\mathbf{Z}}_1)$. The chain rule of mutual information allows separating in a sum the terms $I(\bar{\mathbf{Z}}_1; \mathbf{A}_1|\mathbf{E}, \check{\mathbf{Z}}_1)$ and $I(\bar{\mathbf{Z}}_1; \bar{\mathbf{U}}|\mathbf{E}, \mathbf{A}_1, \check{\mathbf{Z}}_1)$. Given that ΔP_1 preserves $P(\bar{\mathbf{Z}}_1, \bar{\mathbf{U}}, \mathbf{A}_1, \mathbf{E}, \check{\mathbf{Z}}_1)$, these two terms correspond to the terms in the r.h.s. of Equation (23) for k = 1.

We now consider the case of $k = 1$ with $\bar{\mathbf{B}}'_1 \subset \{\mathbf{A}_0, \bar{\mathbf{B}}_0\} = \mathbf{E}$ and the case of any $k > 1$. We follow three steps. Step k.1 involves intermediate families that allow moving from the preservation of $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$, which enables the application of the DP inequality associated with $\bar{\mathbf{Z}}_{k-1} \perp_P \mathbf{A}_{k-1}|\bar{\mathbf{U}}_{k-1}\bar{\mathbf{B}}_{k-1}$, to the preservation of $P(\bar{\mathbf{Z}}_k, \bar{\mathbf{U}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$, which allows applying the DP inequality associated with $\bar{\mathbf{Z}}_k \perp_P \mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$. Step k.1 proceeds as follows: we compare the family ΔP_{k-1} that preserves $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$ and $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k-1$, with a family $\Delta P'_{k-1}$ defined as preserving $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \bar{\mathbf{B}}_k)$, $P(\bar{\mathbf{Z}}_{k-1}, \mathbf{C}_k)$, and $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$ for $j = 1, \ldots, k-1$. By construction, $\mathbf{C}_k = \{\mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1}, \check{\mathbf{Z}}_k\}$ and $\check{\mathbf{Z}}_k \subset \bar{\mathbf{Z}}_{k-1}$, so $P(\bar{\mathbf{Z}}_{k-1}, \mathbf{C}_k)$ is a marginal of $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$. Similarly, by construction $\bar{\mathbf{B}}_k \subseteq \mathbf{C}_k$ so $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \bar{\mathbf{B}}_k)$ is also a marginal of $P(\bar{\mathbf{Z}}_{k-1}, \bar{\mathbf{U}}_{k-1}, \mathbf{A}_{k-1}, \bar{\mathbf{B}}_{k-1})$. This means that $\Delta P_{k-1} \subseteq \Delta P'_{k-1}$, since the constraints of $\Delta P'_{k-1}$ are a subset of the constraints of ΔP_{k-1}. Accordingly, the term at the l.h.s. of Equation (23) is such that

$$\min_{Q\in\Delta P_{k-1}} I_Q(\bar{\mathbf{Z}}_{k-1}; \bar{\mathbf{U}}_{k-1}|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k-1]}) \geqslant \min_{Q\in\Delta P'_{k-1}} I_Q(\bar{\mathbf{Z}}_{k-1}; \bar{\mathbf{U}}_{k-1}|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k-1]}). \tag{A7}$$

We then further modify the constraints according to

$$\begin{aligned}
&\min_{Q\in\Delta P'_{k-1}} I_Q(\bar{\mathbf{Z}}_{k-1};\bar{\mathbf{U}}_{k-1}|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k-1]}) \overset{(a)}{\geqslant}\\
&\min_{Q\in\Delta P'_{k-1}} I_Q(\bar{\mathbf{Z}}_k,\check{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k-1]}) \overset{(b)}{\geqslant}\\
&\min_{Q\in\Delta P'_{k-1}} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}) \overset{(c)}{\geqslant} \min_{Q\in\Delta P''_{k-1}} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}).
\end{aligned} \tag{A8}$$

Step (a) holds because $\{\bar{\mathbf{Z}}_k,\check{\mathbf{Z}}_k\}\subseteq\bar{\mathbf{Z}}_{k-1}$ and $\bar{\mathbf{U}}_k\subseteq\bar{\mathbf{U}}_{k-1}$, so that the information between the subsets can only be smaller than or equal to the information between the sets. In step (b), the chain rule inequality for minInf mutual information terms (Lemma A1) is used to move $\check{\mathbf{Z}}_k$ to the conditioning set. In step (c), the constraints of $\Delta P'_{k-1}$ on $P(\bar{\mathbf{Z}}_{k-1},\bar{\mathbf{U}}_{k-1},\bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}_{k-1},\mathbf{C}_k)$ are loosed in the marginalized variables of $\bar{\mathbf{Z}}_{k-1}\backslash\{\bar{\mathbf{Z}}_k,\check{\mathbf{Z}}_k\}$ and $\bar{\mathbf{U}}_{k-1}\backslash\bar{\mathbf{U}}_k$, which now do not appear in the information term. This loosening results in a family $\Delta P''_{k-1}$ such that $\Delta P'_{k-1}\subseteq\Delta P''_{k-1}$ with constraints on $P(\bar{\mathbf{Z}}_k,\bar{\mathbf{U}}_k,\bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}_k,\mathbf{C}_k)$. Note that $\check{\mathbf{Z}}_k$ is not explicitly written as variables contained in these distributions because $\check{\mathbf{Z}}_k\subseteq\bar{\mathbf{B}}_k\subseteq\mathbf{C}_k$. The latter constraint on $P(\bar{\mathbf{Z}}_k,\mathbf{C}_k)$ can be grouped with $P(\bar{\mathbf{Z}}_j,\mathbf{C}_j)$, for $j=1,\ldots,k-1$, resulting in $P(\bar{\mathbf{Z}}_j,\mathbf{C}_j)$, for $j=1,\ldots,k$.

We then compare the constraints of $\Delta P''_{k-1}$ to the ones of the family ΔP_k. The constraints on $P(\bar{\mathbf{Z}}_j,\mathbf{C}_j)$, for $j=1,\ldots,k$ are common. ΔP_k also has a constraint on $P(\bar{\mathbf{Z}}_k,\bar{\mathbf{U}}_k,\mathbf{A}_k,\bar{\mathbf{B}}_k)$, which contains $\mathbf{A}_k$, as opposed to the constraint $P(\bar{\mathbf{Z}}_k,\bar{\mathbf{U}}_k,\bar{\mathbf{B}}_k)$ of $\Delta P''_{k-1}$. That is, overall step $k.1$ leads to a lower bound with an information term in which $\mathbf{A}_k$ can be introduced thanks to the independence $\bar{\mathbf{Z}}_k\perp_P\mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$.

This insertion of $\mathbf{A}_k$ is done in step $k.2$:

$$\min_{Q\in\Delta P''_{k-1}} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}) = \min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k,\mathbf{A}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}). \tag{A9}$$

The minInf DP inequality of Proposition 5 is applied with variables $\{\mathbf{Z},\mathbf{D},\mathbf{D}',\mathbf{O}_1,\mathbf{E},\mathbf{E}_2\}$ assigned as $\{\bar{\mathbf{Z}}_k,\bar{\mathbf{U}}_k,\mathbf{A}_k,\bar{\mathbf{B}}_k,\{\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}\},\varnothing\}$. ΔP_k plays the role of $\Delta P_{\mathbf{DD}'}$ in Equation (19) and $\Delta P''_{k-1}$ the role of $\Delta P_{\mathbf{D}}$.

In the final step $k.3$, the chain rule inequality for minInf information terms from Lemma A1 is used to separate the observable information term that does not contain $\bar{\mathbf{U}}_k$:

$$\begin{aligned}
\min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k,\mathbf{A}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]}) \geqslant &\min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\mathbf{A}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]})+\\
&\min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k]},\check{\mathbf{Z}}^{[k]}).
\end{aligned} \tag{A10}$$

The first term $\min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\mathbf{A}_k|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k]})$ only involves observable variables and an estimable information term is obtained relaxing the preservation of $P(\bar{\mathbf{Z}}_k,\bar{\mathbf{U}}_k,\mathbf{A}_k,\bar{\mathbf{B}}_k)$ to the preservation of $P(\bar{\mathbf{Z}}_k,\mathbf{A}_k,\bar{\mathbf{B}}_k)$. The second term $\min_{Q\in\Delta P_k} I_Q(\bar{\mathbf{Z}}_k;\bar{\mathbf{U}}_k|\mathbf{E},\mathbf{A}^{[k]},\check{\mathbf{Z}}^{[k]})$ has the same form of the term at the l.h.s. of Equation (23), which is $\min_{Q\in\Delta P_{k-1}} I_Q(\bar{\mathbf{Z}}_{k-1};\bar{\mathbf{U}}_{k-1}|\mathbf{E},\mathbf{A}^{[k-1]},\check{\mathbf{Z}}^{[k-1]})$, but with k instead of $k-1$. This completes iteration k. □

Overall, we can summarize the three steps of each iteration k in the following way. Step $k.1$ loosens the constraints on the marginals preserved in the family of distributions. This loosening is the minimum amount required so that the next conditional independence can be used. In fact, the same procedure could be applied with $\mathbf{C}_j\subseteq\{\mathbf{A}_{j-1},\mathbf{B}_{j-1},\check{\mathbf{Z}}_j\}$, but that would lead to weaker constraints $P(\bar{\mathbf{Z}}_j,\mathbf{C}_j)$ for $j=1,\ldots,k$ in the minimization,

and hence to an equal or smaller lower bound. Step k.2 corresponds to the application of the DP inequality of Proposition 5. It can be verified that in the scenarios included in Theorem 1, Proposition 5 is always applied with $\mathbf{E}_2 = \varnothing$. Further generalizations with a nonempty $\mathbf{E}_2$ are left for future work since they do not qualitatively add to the procedures here developed. Finally, step k.3 applies the chain rule inequality of minInf information terms (Lemma A1) to separate the observable information term.

Appendix F. Examples of Applications of Theorem 1

In this section, we further examine scenarios comprised in the iterative application of Theorem 1 when used in Proposition 6. We start with the example of Figure A2A, which illustrates that $\mathbf{X}$ and the conditioning set $\mathbf{B}_0$ play the same role in terms of the application of DP inequalities to derive an estimable lower bound. In Figure 2 we have seen examples in which variables belonging to $\mathbf{B}_0$ are excluded at a certain iteration k from the set $\bar{\mathbf{B}}_k$ that in Theorem 1 appears in $\bar{\mathbf{Z}}_k \perp_P \mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$. For example, in Figure 2A, with $\mathbf{B}_0 = \{W_1, W_2\}$, we have derived the term $I(Z; Y_2\backslash\backslash\{W_2, Y_1\}|W_1, X)$ relaxing the preservation of $P(Z, U, Y_1, W_1, W_2, X)$ to $P(Z, U, W_1, X)$, given that $Z \perp Y_2|UXW_1$, while $Z \not\perp Y_2|UXW_1W_2Y_1$. The same relaxation is applicable to exclude variables from $\mathbf{X}$, like variables from $\mathbf{B}_0$. This is because in Proposition 6 the derivation of the estimable lower bound starts from $I(\mathbf{Z}; \mathbf{U}|\mathbf{B_0}, \mathbf{X})$, and hence $\{\mathbf{B_0}, \mathbf{X}\}$ jointly appear as part of all $\mathbf{E}_k = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}^{[k-1]}\}$. Figure A2A shows an example in which an inequality is derived relaxing the preservation of $P(Z, U, Y_1, X_1, X_2)$ to $P(Z, U, X_1)$, hence excluding $X_2 \in \mathbf{X}$. In more detail, the inequality

$$H(X_1, X_2|Z) \geqslant I(Z; Y_1|X_1, X_2) + I(Z; Y_2\backslash\backslash\{X_2, Y_1\}|X_1) \tag{A11}$$

is derived selecting the instrumental set $\mathbf{Z} = \{Z\}$, the conditioning set $\mathbf{B_0} = \varnothing$, and $\mathbf{X} = \{X_1, X_2\}$. The upper bound is obtained with $Z \perp U$, and the terms in the lower bound are derived with $Z \perp Y_1|UX_1X_2$ and $Z \perp Y_2|UX_1$, respectively. To introduce Y_2, the minInf constraints are relaxed to exclude $\{X_2, Y_1\}$ because they are a collider and a descendant of a collider in a path between Z and Y_2.

Figure A2B shows an example in which the iterative application of DP inequalities requires the marginalization of some hidden variable. The inequality

$$H(X_1, X_2|W, Z) \geqslant I(Z; Y_1|X_1, X_2, W) + I(Z; Y_2\backslash\backslash\{X_1, Y_1, W\}|X_2) \tag{A12}$$

is derived selecting $\mathbf{Z} = \{Z\}$, $\mathbf{B_0} = \{W\}$, $\mathbf{U} = \{U_1, U_2\}$, and $\mathbf{X} = \{X_1, X_2\}$. The upper bound is derived with $Z \perp \{U_1, U_2\}|W$. The estimable terms in the lower bound are derived thanks to $Z \perp Y_1|U_1U_2X_1X_2W$ and $Z \perp Y_2|U_2X_2$, respectively. In particular, after the application of the first DP inequality the term $I(Z; U_1, U_2|X_1, X_2, W, Y_1)$ is obtained. A second DP inequality cannot be applied directly from this term without first marginalizing U_1. This is because $\{W, X_1, Y_1, U_1\}$ cannot be in the conditioning set since W is a collider in a path between Z and Y_2, and $\{X_1, Y_1, U_1\}$ are its descendants. The hidden variable U_1 is marginalized given that by monotonicity $I(Z; U_2|X_1, X_2, W, Y_1)$ is equal to or smaller than $I(Z; U_1, U_2|X_1, X_2, W, Y_1)$. The other variables $\{W, X_1, Y_1\}$ are excluded from the conditioning set relaxing the minInf constraints from $P(Z, U_2, Y_1, W, X_1, X_2)$ to $P(Z, U_2, X_2)$.

https://doi.org/10.3390/e28040472

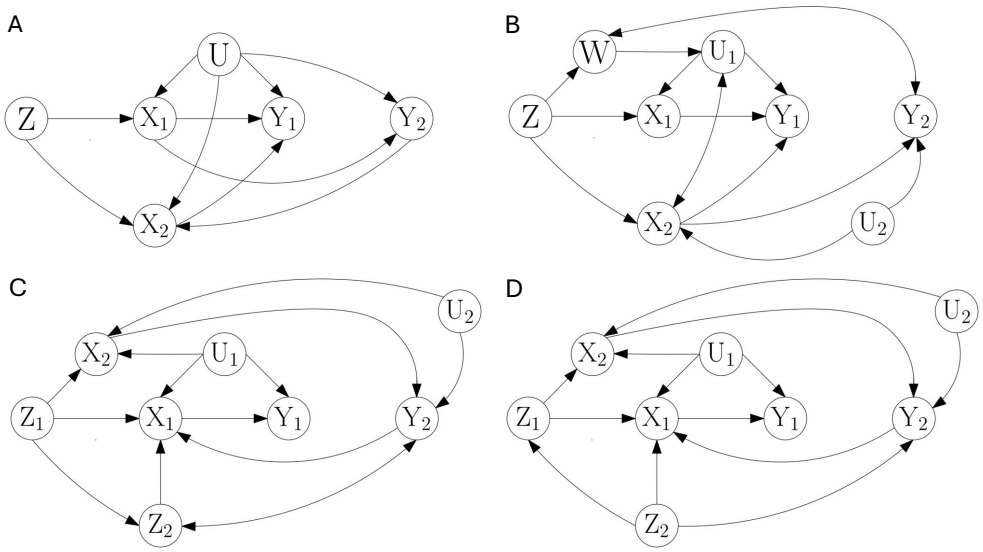


Figure A2. Examples for the application of Theorem 1 in the derivation of instrumental entropic inequalities of Proposition 6. All variables are observable except **U** hidden. These examples are described in Appendix F. (**A**) Causal structure related to Equation (A11). (**B**) Causal structure related to Equation (A12). (**C**) Causal structure related to Equation (A13). (**D**) Causal structure related to Equation (A14).

Figure A2C shows an example in which the iterative application of DP inequalities requires the marginalization of some variable from the instrumental set **Z**. The inequality

$$H(X_1, X_2|Z_1, Z_2) \geqslant I(Z_1, Z_2; Y_1|X_1, X_2) + I(Z_1; Y_2 \backslash\backslash \{X_1, Y_1\}|X_2) \tag{A13}$$

is derived selecting $\mathbf{Z} = \{Z_1, Z_2\}$, $\mathbf{B_0} = \varnothing$, $\mathbf{U} = \{U_1, U_2\}$, and $\mathbf{X} = \{X_1, X_2\}$. The upper bound follows from $Z_1Z_2 \perp U_1U_2$. The terms in the lower bound are derived with DP inequalities associated with $Z_1Z_2 \perp Y_1|U_1U_2X_1X_2$ and $Z_1 \perp Y_2|U_1U_2X_2$. After the application of the first DP inequality a term $I(Z_1, Z_2; U_1, U_2|X_1, X_2, Y_1)$ is obtained. Since Z_2 is not separable from Y_2 by conditioning on any subset of $\{U_1, U_2, X_1, X_2, Y_1\}$, the term is marginalized to $I(Z_1; U_1, U_2|X_1, X_2, Y_1)$. Subsequently, $\{X_1, Y_1\}$ are excluded from conditioning by relaxing the minInf constraints, given that they are a collider or descendant of a collider in paths between Z_1 and Y_2.

Finally, Figure A2D shows an example in which the iterative application of DP inequalities requires conditioning on some variable from the instrumental set **Z**, as opposed to the example of Figure A2C in which variable $Z_2 \in \mathbf{Z}$ was marginalized. The causal structure in Figure A2D is the same as in Figure A2C except that Z_2 is a noncollider instead of a collider in $Z_1 - Z_2 - Y_2$. The inequality

$$H(X_1, X_2|Z_1, Z_2) \geqslant I(Z_1, Z_2; Y_1|X_1, X_2) + I(Z_1; Y_2 \backslash\backslash \{X_1, Y_1\}|X_2, Z_2) \tag{A14}$$

is again derived selecting $\mathbf{Z} = \{Z_1, Z_2\}$, $\mathbf{B_0} = \varnothing$, $\mathbf{U} = \{U_1, U_2\}$, and $\mathbf{X} = \{X_1, X_2\}$. The upper bound follows from $Z_1Z_2 \perp U_1U_2$ and the first DP inequality is applied with $Z_1Z_2 \perp Y_1|U_1U_2X_1X_2$. However, now the second DP inequality is associated with $Z_1 \perp Y_2|U_1U_2X_2Z_2$. After the application of the first DP inequality the term $I(Z_1, Z_2; U_1, U_2|X_1, X_2, Y_1)$ is obtained. As in Figure A2C, no other DP inequality can be applied to $\{Z_1, Z_2\}$, since Z_2 is adjacent to Y_2. Contrarily to Figure A2C, now Z_2 cannot be marginalized but needs to be moved to conditioning in order to separate Z_1 and Y_2. Accordingly, the chain rule is applied to $I(Z_1, Z_2; U_1, U_2|X_1, X_2, Y_1)$ to separate $I(Z_2; U_1, U_2|X_1, X_2, Y_1)$ and $I(Z_1; U_1, U_2|X_1, X_2, Y_1, Z_2)$. The first term is dropped, and the

second term is used to apply the DP inequality associated with $Z_1 \perp Y_2|U_1U_2X_2Z_2$ after relaxing the preservation of the marginals to exclude $\{X_1, Y_1\}$ from the conditioning set.

Overall, in Figure A2 we have provided additional examples that illustrate how instrumental entropic inequalities can be derived from Proposition 6 by relaxing the preservation of marginals that include a subset of $\mathbf{X}$ (Figure A2A), by marginalizing on a subset of the hidden variables $\mathbf{U}$ (Figure A2B), by marginalizing on a subset of the instrumental set $\mathbf{Z}$ (Figure A2C), or by moving part of $\mathbf{Z}$ to the conditioning set (Figure A2D). For all these examples, it can be verified that the instrumental entropic inequalities of Equations (A11)–(A14) provide additional causal inference power according to the criterion of Remark A1. For the sake of space, a detailed verification of this criterion is not presented, in particular because our objective here was to further illustrate the versatility within Theorem 1 to build sequences $\bar{\mathbf{Z}}^{[n]}$, $\bar{\mathbf{B}}^{[n]}$, and $\bar{\mathbf{U}}^{[n]}$ associated with independencies $\bar{\mathbf{Z}}_k \perp_P \mathbf{A}_k|\bar{\mathbf{U}}_k\bar{\mathbf{B}}_k$ for $k = 1, \ldots, n$.

Appendix G. Estimation of minInf Information Terms

As discussed in Appendix D, the maximum entropy unique information is a special case of a minInf information term in that its estimation constitutes a convex optimization problem. In general, the estimation of minInf information terms requires non-convex optimization techniques. To see this, consider an observable term $\min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ as they appear in Theorem 1, where ΔP_k preserves the marginals $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k$. We can compare this with the specific case of the unique information as discussed in Appendix D, for which the term to be minimized is $I_Q(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{E})$, with ΔP that preserves $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{O}_1)$ and $P(\bar{\mathbf{Z}}, \mathbf{E})$. The key property that renders the estimation of the unique information a convex optimization problem is that preserving $P(\bar{\mathbf{Z}}, \mathbf{E})$ fixes $H(\bar{\mathbf{Z}}|\mathbf{E})$ constant within ΔP, so that minimizing $I_Q(\bar{\mathbf{Z}}; \mathbf{D}_1|\mathbf{E})$ corresponds to maximizing $H_Q(\bar{\mathbf{Z}}|\mathbf{E}, \mathbf{D}_1)$, which is concave in ΔP. However, in general, when minimizing $I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, the constraints of ΔP_k are not such that $H_Q(\bar{\mathbf{Z}}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ is constant in ΔP_k, since $P(\bar{\mathbf{Z}}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ is not one of the marginals comprised in $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k$. Therefore, while the constraints on the preservation of the marginals always constitute a set of affine equality constraints, the mutual information term to be minimized is not a convex function of the probability distributions within the family of distributions ΔP_k [65].

Accordingly, in general the estimation of minInf information terms requires non-convex optimization methods [44,71,72]. The form of the mutual information objective function as a difference of entropies suggests that the implementation may benefit from methods developed for successive convex approximation, specifically for differences of convex functions [73]. An alternative approach relies on the use of copula methods to construct minimum information joint distributions [74,75], and further work would need to explore how to expand their use for the set of constraints of the minInf terms of the form of Proposition 5. More generally, the determination of information terms defined in non-convex optimization problems is common in network information theory [35]. Non-convex optimization problems appear in multi-terminal communication channels such as Broadcast channels [76], Gray-Wyner networks [77], or Interference channels [78], as well as in problems of confidential and secure communication (e.g., [79]). We expect that the data processing inequalities we have derived can find applications in other domains of information theory, and benefit from estimation methods developed in those domains. However, a full implementation to estimate minInf information terms is beyond the scope of this work, which focuses on the theoretical derivation of data processing inequalities and entropic inequalities with minInf terms. Nonetheless, to advance in their estimation,

we here reexpress the definition of minInf terms of the sort that appear in Theorem 1 with a formulation that separates a convex and a non-convex component of the minimization problem. Based on this formulation, in Appendix H we resume the numerical analysis of examples with Gaussian systems, showing that the addition of other minInf terms together with the unique information can further increase causal inferential power.

Lemma A2 (MinInf information terms as minimal unique information terms)**.** *Consider a mutual information term* $I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k | \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ *defined as in Theorem 1, and to be minimized within the family of distributions* ΔP_k *that preserves the marginals* $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *and* $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$*, for* $j = 1, \ldots, k$*, with the collections* $\bar{\mathbf{Z}}^{[k]}$ *and* $\mathbf{C}^{[k]}$ *constructed as in Theorem 1. The minimization problem can be reexpressed as the minimization of a unique information term:*

$$\min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k | \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}) = \min_{\bar{Q} \in \Delta P_k} I_{\bar{Q}}(\bar{\mathbf{Z}}_k; \mathbf{A}_k \backslash\backslash \{\{\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}\} \backslash \bar{\mathbf{B}}_k\} | \bar{\mathbf{B}}_k), \tag{A15}$$

where $\bar{Q}(\mathbf{A}_k, \bar{\mathbf{Z}}^{[k]}, \mathbf{C}^{[k]})$ *has marginals* $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *equal to* $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *and* $\bar{Q}(\bar{\mathbf{Z}}^{[k]}, \mathbf{C}^{[k]})$ *that preserves all the marginals* $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$*, for* $j = 1, \ldots, k$*.*

Proof. By construction, $\mathbf{C}_1 = \{\mathbf{E}, \check{\mathbf{Z}}_1\}$. Also by construction, $\mathbf{C}_j = \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}, \check{\mathbf{Z}}_j\}$. This means that all the variables in the conditioning set $\{\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}\}$ appear at least in one of the preserved marginals $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k$. On the other hand, $\mathbf{A}_k$ only appears in the marginal $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$. Therefore, the fulfillment of the constraints for a distribution $\bar{Q}(\mathbf{A}_k, \bar{\mathbf{Z}}^{[k]}, \mathbf{C}^{[k]})$ can be separated into $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ being equal to $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and the other constraints imposed to the marginal $\bar{Q}(\bar{\mathbf{Z}}^{[k]}, \mathbf{C}^{[k]})$. We can hence separate the minimization within ΔP_k in two steps. First, a minimization that involves the selection of a concrete marginal $\bar{Q}(\bar{\mathbf{Z}}^{[k]}, \mathbf{C}^{[k]})$ compatible with $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k$. Subsequently, the minimization operates among the distributions $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ compatible with the marginals $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k) = P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and the predetermined $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$. Note that also by construction, given $\bar{\mathbf{B}}'_j \subseteq \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}\}$ and $\bar{\mathbf{B}}_j = \{\bar{\mathbf{B}}'_j, \check{\mathbf{Z}}_j\}$, then $\bar{\mathbf{B}}_k \subseteq \{\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}\}$. Therefore, the preservation of the two marginals $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ has the form of the constraints that define a unique information. Concretely, following Equation (5), the form of a unique information term is recovered with the assignments of $\bar{\mathbf{Z}}$ as $\bar{\mathbf{Z}}_k$, $\mathbf{D}_1$ as $\mathbf{A}_k$, $\mathbf{D}_2$ as $\{\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}\} \backslash \bar{\mathbf{B}}_k$, and $\mathbf{O}_1$ as $\bar{\mathbf{B}}_k$, such that $P(\bar{\mathbf{Z}}, \mathbf{D}_1, \mathbf{O}_1)$ of Equation (5) corresponds to $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}, \mathbf{D}_2, \mathbf{O}_1)$ corresponds to $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$. □

Appendix H. Numerical Examples with Additional minInf Information Terms

We now resume the analysis of Appendix D to numerically study the entropic inequality of Equation (25), associated with Figure 2C. We use Lemma A2 to rewrite Equation (25) and we also express the unique information terms specifically for Gaussian variables:

$$\begin{aligned}
&H(X|Z, W_2, W_3) \geqslant I(Z; Y_1 | W_2, W_3, X) + \min_{Q \in \Delta P_2} I_Q(Z; Y_2 | W_2, W_3, X, Y_1) + \\
&\min_{Q \in \Delta P_3} I_Q(Z; Y_3 | W_2, W_3, X, Y_1, Y_2) \overset{(a)}{=} I(Z; Y_1 | W_2, W_3, X) + I(Z; Y_2 \backslash\backslash \{W_2, Y_1\} | W_3, X) + \\
&\min_{\bar{Q} \in \Delta P_3} I_{\bar{Q}}(Z; Y_3 \backslash\backslash \{W_2, W_3, Y_1\} | X, Y_2) \overset{(b)}{=} I(Z; Y_1 | W_2, W_3, X) + \max\{I(Z; Y_2 | W_3, X) - \\
&I(Z; W_2, Y_1 | W_3, X), 0\} + \min_{\bar{Q} \in \Delta P_3} \max\Big\{ I_{\bar{Q}}(Z; Y_3 | X, Y_2) - I_{\bar{Q}}(Z; W_2, W_3, Y_1 | X, Y_2), 0 \Big\},
\end{aligned} \tag{A16}$$

where ΔP_2 preserves the marginals $\{P(Z, W_3, X, Y_2), P(Z, W_2, W_3, X, Y_1)\}$ and ΔP_3 preserves the marginals $\{P(Z, X, Y_2, Y_3), P(Z, W_3, X, Y_2), P(Z, W_2, W_3, X, Y_1)\}$. Equality (a) uses Lemma A2 to reexpress the second minInf term as the minimization of a unique information estimated on distributions $\bar{Q} \in \Delta P_3$ that factorize as $\bar{Q} = P(Z, X, Y_2, Y_3)Q(W_2, W_3, Y_1|Z, X, Y_2, Y_3)$ and preserve the marginals $P(Z, W_3, X, Y_2)$ and $P(Z, W_2, W_3, X, Y_1)$. Equality (b) expresses the unique information terms specifically for Gaussian variables, given their form in Equation (A3).

We proceed analogously to Appendix D, simulating variables generated with a system that conforms to the causal structure of Figure 2C, but with additional direct connections $Z \to Y_1$, $Z \to Y_2$, and $Z \to Y_3$. We also keep the connection $X \leftrightarrow W_2$ so that other than for the inclusion of the new observable variables W_3 and Y_3, the generative process of the observable variables is the same as in Appendix D. If not for the connections $Z \to Y_1$, $Z \to Y_2$, and $Z \to Y_3$, the system would causally fulfill the inequality of Equation (25), but these connections can lead to violations. We again generate the variables as a system of linear equations with Gaussian noise. Following the same strategy of Appendix D, we keep fixed the values of most parameters and examine concrete settings in which the strength of the connection $Z \to Y_3$ changes, as determined by the coefficient a_{y_3z}. In more detail, we again select a default configuration in which the standard deviation of all hidden confounders is set to 1, and their mean set to zero. Again all nodes apart from Z are generated having also some exogenous independent noises, all of which are generated with mean zero and standard deviation v. All coefficients associated with the connections are again by default set to 1, if not indicated otherwise. For the connections exclusive of Figure 2C in comparison to Figure 2A, we set $a_{w_3z} = a_{w_3u''} = a_{y_3u''} = 2$, where U'' is the hidden confounder in $W_3 \leftrightarrow Y_3$.

Since in this Appendix we are interested in examining the additional causal inference power gained with the extra minInf term associated with Y_3 in Equation (A16), we focus on configurations for which the strength of both $Z \to Y_1$ and $Z \to Y_2$ is not enough to violate the inequality. We fixed $a_{y_1z} = a_{y_2z} = 1$ and verified that a violation does not occur due to these connections. Accordingly, we here examine configurations in which in Equation (A16) the sum of $I(Z; Y_1|W_2, W_3, X)$ and $I(Z; Y_2 \backslash\backslash \{W_2, Y_1\}|W_3, X)$ at the r.h.s. is smaller than the upper bound $H(X|Z, W_2, W_3)$, so that it is the additional minInf term associated with Y_3 the one that determines whether a violation occurs. Apart from exploring different strengths of a_{y_3z}, we extend the analysis of Figure A1G–I, and examine configurations with different values of the standard deviation v of the exogenous noises.

To find the minimum of $\bar{Q} \in \Delta P_3$ we proceed by exploring the space of joint distributions compatible with the preservation of the marginals $\{P(Z, X, Y_2, Y_3), P(Z, W_3, X, Y_2), P(Z, W_2, W_3, X, Y_1)\}$. This preservation is reflected in fixed entries of the covariance matrix, which for Gaussian variables determines all information terms. Given the symmetry of the covariance matrix, the preservation of these marginals results in 5 remaining degrees of freedom to explore joint distributions within ΔP_3. We sampled joint distributions with valid covariance matrices covering the range of the non-fixed entries of the matrix with a grid of 50 samples along each of the degrees of freedom, hence probing in the order of 3×10^8 joint distributions. Figure A3 shows the results of testing the inequality of Equation (25) across configurations with varying a_{y_3z} and v. As in Figure A1, we display the test statistic corresponding to the upper bound minus the lower bound of the inequality, such that a violation occurs for negative values.

Figure A3 shows that violations of the inequality occur with an increasing strength of $Z \to Y_3$. The terms $I(Z; Y_1|W_2, W_3, X)$ and $I(Z; Y_2 \backslash\backslash \{W_2, Y_1\}|W_3, X)$, as well as the upper bound $H(X|Z, W_2, W_3)$, are constant to changes in a_{y_3z}, and hence the decrease of the statistic and the occurrence of negative values is due to $\min_{Q \in \Delta P_3} I_Q(Z; Y_3|W_2, W_3, X, Y_1, Y_2)$.

This illustrates that the addition of new minInf terms in the instrumental entropic inequality provides additional causal inference power.

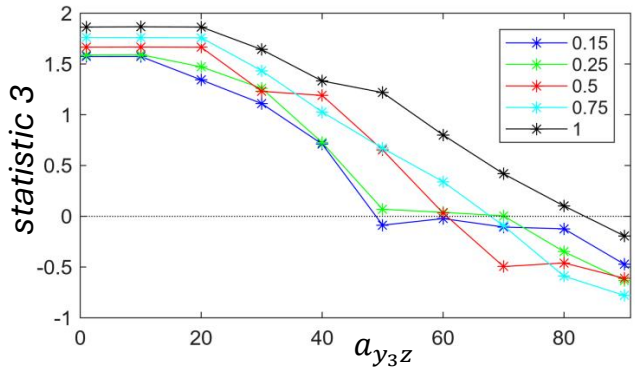


Figure A3. Examples of violations of the instrumental entropic inequality of Equation (25). Multivariate Gaussian systems are generated as described in Appendix H. The statistic of the test associated with Equation (A16) is displayed as a function of the strength a_{y_3z} of the connection $Z \to Y_3$ for different standard deviations v of the exogenous noises of the variables.

Nonetheless, the lesser smoothness of the curves in Figure A3 compared to Figure A1 reflects the difficulty to estimate minInf terms in general. In Lemma A2, we have shown how to separate convex and non-convex parts of this estimation, and here we have restricted the simulations to multivariate Gaussian variables in order to benefit from the known form of unique information in these systems [68]. In our simulations the family of distributions to be explored is characterized by five degrees of freedom of the covariance matrix, making an exhaustive exploration of the family manageable, although computationally expensive. More broadly, non-convex optimization methods need to be adopted for the estimation of minInf terms, as described in Appendix G. The main contribution of this work has been the theoretical development of how minInf terms can be used to increase the causal inference power of entropic inequalities. The derived minInf data processing inequalities open a line of research to extend further types of entropic inequalities beyond the ones here considered. In order to make these new inequalities applicable in general, future work will also need to develop optimization methods that allow a reliable estimation of minInf terms.

Apart from the specific challenge of estimating minInf information terms due to the non-convexity of the minimization problem, the estimation from finite data sets is expected to involve additional difficulties ubiquitous for information-theoretic measures. While bias-correction methods have been widely studied for the standard measures [80,81], only recently bias-correction methods have been studied for the maximum entropy unique information [56]. To apply entropic inequalities with minInf terms to systems with many variables or variables with large cardinality, the analysis of bias corrections will need to be extended. Nonetheless, even for complex graphs that include many nodes, causally informative entropic inequalities may imply only subsets of nodes and therefore still be implementable.

Appendix I. Complementary Instrumental Entropic Inequalities in Figure 2C

We here extend the characterization of causally-fulfilled instrumental entropic inequalities in the causal structure of Figure 2C. In Section 3.5, we examined the inequality of Equation (25) as an example of application of Proposition 6 for which three DP inequalities are iteratively applied, starting from $I(Z; U|X, W_2, W_3)$, with $\mathbf{Z} = \{Z\}$ and $\mathbf{B}_0 = \{W_2, W_3\}$. We here instead examine instrumental entropic inequalities derived with $\mathbf{Z} = \{Z, W_2, W_3\}$ and $\mathbf{B}_0 = \varnothing$.

In general, to characterize all existing instrumental entropic inequalities, we would need to apply Proposition 6 with all possible partitions $\mathbf{Z}^{[r]}$ for Proposition 3. However, our objective here is not to fully characterize the concrete case of Figure 2C, but to exemplify in more detail how to derive complementary instrumental inequalities with Proposition 6. For this reason, we focus on partitions $\mathbf{Z}^{[r]}$ that use the chain rule to decompose $\mathbf{Z} = \{Z, W_2, W_3\}$ separating individual variables sequentially. This results in six possible partitions, as shown in the second column of Table A1. Columns three to five show the three nonestimable information terms resulting from each partition. For each of them, we indicate associated independencies that allow applying DP inequalities to derive estimable lower bounds. The resulting instrumental entropic inequalities are given below, with the indexes of the inequalities mapping the indexes of rows in Table A1:

Table A1. Properties associated with the instrumental entropic inequalities presented in Appendix I, which are causally fulfilled by the causal structure of Figure 2C. The label of each row maps to subequations in Equation (A17). The second column indicates the order in which the chain rule is applied in Proposition 3, starting from $\mathbf{Z} = \{Z, W_2, W_3\}$ and $\mathbf{B}_0 = \varnothing$. Columns three to five provide the information terms in the chain decomposition and associated independencies that allow adding estimable information terms by applying Proposition 6.

(a)	$\{\{Z\},\{W_2\},\{W_3\}\}$	$I(Z;U\|X)$ $Z \perp \{Y_2, Y_3\}\|UX$	$I(W_2;U\|X,Z)$ $\varnothing$	$I(W_3;U\|X,Z,W_2)$ $\varnothing$
(b)	$\{\{Z\},\{W_3\},\{W_2\}\}$	$I(Z;U\|X)$ $Z \perp \{Y_2, Y_3\}\|UX$	$I(W_3;U\|X,Z)$ $\varnothing$	$I(W_2;U\|X,Z,W_3)$ $\varnothing$
(c)	$\{\{W_2\},\{Z\},\{W_3\}\}$	$I(W_2;U\|X)$ $W_2 \perp Y_3\|UX$	$I(Z;U\|X,W_2)$ $Z \perp \{Y_1Y_3\}\|UXW_2$ $Z \perp Y_2\|UXY_3$	$I(W_3;U\|X,Z,W_2)$ $\varnothing$
(d)	$\{\{W_3\},\{Z\},\{W_2\}\}$	$I(W_3;U\|X)$ $W_3 \perp Y_2\|UX$	$I(Z;U\|X,W_3)$ $Z \perp Y_2\|UXW_3$ $Z \perp Y_3\|UXY_2$	$I(W_2;U\|X,Z,W_3)$ $\varnothing$
(e)	$\{\{W_2\},\{W_3\},\{Z\}\}$	$I(W_2;U\|X)$ $W_2 \perp Y_3\|UX$	$I(W_3;U\|X,W_2)$ $W_3 \perp Y_1\|UXW_2$ $W_3 \perp Y_2\|UX$	$I(Z;U\|X,W_2,W_3)$ $Z \perp Y_1\|UXW_2W_3$ $Z \perp Y_2\|UXW_3$ $Z \perp Y_3\|UXY_2$
(f)	$\{\{W_3\},\{W_2\},\{Z\}\}$	$I(W_3;U\|X)$ $W_3 \perp Y_2\|UX$	$I(W_2;U\|X,W_3)$ $W_2 \perp Y_3\|UX$	$I(Z;U\|X,W_2,W_3)$ $Z \perp Y_1\|UXW_2W_3$ $Z \perp Y_3\|UXW_2Y_1$ $Z \perp Y_2\|UXY_3$

$$
\begin{aligned}
&H(X|Z,W_2,W_3) \geqslant \\
&\quad \geqslant I(Z;Y_2,Y_3|X) && \text{(A17a)}\\
&\quad \geqslant I(Z;Y_2,Y_3|X) && \text{(A17b)}\\
&\quad \geqslant I(W_2;Y_3|X) + I(Z;Y_1Y_3|X,W_2) + I(Z;Y_2\backslash\backslash\{W_2,Y_1\}|X,Y_3) && \text{(A17c)}\\
&\quad \geqslant I(W_3;Y_2|X) + I(Z;Y_2|X,W_3) + I(Z;Y_3\backslash\backslash W_3|X,Y_2) && \text{(A17d)}\\
&\quad \geqslant I(W_2;Y_3|X) + I(W_3;Y_1|X,W_2) + I(W_3;Y_2\backslash\backslash\{W_2,Y_1\}|X) + && \text{(A17e)}\\
&\qquad I(Z;Y_1|X,W_2,W_3) + I(Z;Y_2\backslash\backslash\{W_2,Y_1\}|X,W_3) + \\
&\qquad \min_{Q\in\Delta P_3} I_Q(Z;Y_3|X,W_2,W_3,Y_1,Y_2) \\
&\quad \geqslant I(W_3;Y_2|X) + I(W_2;Y_3\backslash\backslash W_3|X) + I(Z;Y_1|X,W_2,W_3) + && \text{(A17f)}\\
&\qquad I(Z;Y_3\backslash\backslash W_3|X,W_2,Y_1) + \min_{Q\in\Delta P_3'} I_Q(Z;Y_2|X,W_2,W_3,Y_1,Y_3),
\end{aligned}
$$

where ΔP_3 preserves $P(Z, X, W_2, W_3, Y_1)$, $P(Z, X, W_3, Y_2)$, and $P(Z, X, Y_2, Y_3)$, while $\Delta P_3'$ preserves $P(Z, X, W_2, W_3, Y_1)$, $P(Z, X, W_2, Y_1, Y_3)$, and $P(Z, X, Y_2, Y_3)$. Note that the upper bound is common to all inequalities. Each subequation should be read as comparing the l.h.s. with each individual r.h.s. with no order between the r.h.s. of the different subequations. Partitions (a) and (b) result in the same instrumental entropic inequality. It can be verified that the resulting five different inequalities of Equation (A17) provide complementary causal inference power. The inequality of Equation (A17e) subsumes the inequality of Equation (25). This is seen straightforwardly moving $I(Z, X|W_2, W_3)$ from the r.h.s. to the l.h.s. of Equation (25). Altogether, this further analysis illustrates that Proposition 6 allows deriving sets of instrumental entropic inequalities that exploit different combinations of independencies present in the causal structure, hence providing additional causal inference power to the standard instrumental entropic inequalities.

Appendix J. A Hierarchy Between Instrumental Entropic Inequalities Using Multivariate Instrumental Sets

We here consider the relation between instrumental entropic inequalities constructed with a multivariate instrumental set $\mathbf{Z}$ and instrumental entropic inequalities constructed by using as instruments only subsets of $\mathbf{Z}$. We focus on a more restricted scenario than the one of Remark A1. Instead of considering all causal structures $\mathbb{G}(\mathbf{I_V})$ compatible with an available set of testable conditional independencies $\mathbf{I_V}$, we consider a scenario in which, using a concrete set of hidden independencies, two specific types of inequalities are to be compared in their causal inference power to discard a single causal structure of interest:

Proposition A1 (A hierarchy of instrumental entropic inequalities)**.** *Consider nonoverlapping sets of variables* $\mathbf{Z}$, $\mathbf{B}_0$, $\mathbf{X}$, *and* $\mathbf{U}$, *all observable except* $\mathbf{U}$ *hidden variables. Consider that the causal structure of interest whose compatibility with data is to be tested is such that it creates an independence* $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$, *so that* $\mathbf{Z}$ *is a multivariate instrumental set. Consider a nonoverlapping partition* $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4\}$, *with* $\mathbf{Z}_1$ *nonempty, and* $\mathbf{Z}_2$, $\mathbf{Z}_3$, *and* $\mathbf{Z}_4$ *possibly empty. Consider that the causal structure also creates a set of m independencies that allow the use of* DP *inequalities by recursively applying Theorem 1 with initial inputs* $\bar{\mathbf{Z}} = \mathbf{Z}_1$ *and* $\mathbf{E} = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2\}$, *resulting in the introduction of estimable information terms with observable variables* $\mathbf{A}^{[m]}$. *In this case, an instrumental entropic inequality derived applying Theorem 1 with instruments* $\mathbf{Z}_1$, *and departing from* $I(\mathbf{Z}_1; \mathbf{U}, \mathbf{X}|\mathbf{B}_0, \mathbf{Z}_2, \mathbf{Z}_3)$, *does not add causal inference power to the instrumental inequality derived applying Theorem 1 with the whole instrumental set* $\mathbf{Z}$ *and departing from* $I(\mathbf{Z}; \mathbf{U}, \mathbf{X}|\mathbf{B}_0)$.

Proof. Given that the causal structure of interest fulfills $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$, the *weak union* axiom of semi-graphoids [25,43] guarantees that it also fulfills $\mathbf{Z}_1 \perp \mathbf{U}|\mathbf{B}_0\mathbf{Z}_2\mathbf{Z}_3$. The instrumental inequality developed using $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$ has as upper bound the entropy $H(\mathbf{X}|\mathbf{Z}, \mathbf{B}_0)$, while the inequality developed using $\mathbf{Z}_1 \perp \mathbf{U}|\mathbf{B}_0\mathbf{Z}_2\mathbf{Z}_3$, with $\mathbf{B}_0' = \{\mathbf{B}_0, \mathbf{Z}_2, \mathbf{Z}_3\}$, has as upper bound the entropy $H(\mathbf{X}|\mathbf{Z}\backslash\mathbf{Z}_4, \mathbf{B}_0)$. The observable lower bounds are obtained applying the DP inequalities departing from $I(\mathbf{Z}; \mathbf{U}|\mathbf{B}_0, \mathbf{X})$ and from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$, respectively.

Since the DP inequalities rely on independencies that are applied following Theorem 1 with $\bar{\mathbf{Z}} = \mathbf{Z}_1$ and $\mathbf{E} = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2\}$, this means that no variable from $\{\mathbf{Z}_3, \mathbf{Z}_4\}$ appears in the conditioning set of those independencies. Therefore, before the iterative insertion of observable terms, $I(\mathbf{Z}; \mathbf{U}|\mathbf{B}_0, \mathbf{X})$ is marginalized to $I(\mathbf{Z}_1, \mathbf{Z}_2; \mathbf{U}|\mathbf{B}_0, \mathbf{X})$. Furthermore, since $\bar{\mathbf{Z}} = \mathbf{Z}_1$, this means that $\mathbf{Z}_2$ only appears in the conditioning sets of the independencies used in the DP inequalities, and hence the departing term is further reduced to $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2)$.

We now compare this departing term $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2)$ used when $\mathbf{Z}$ is the instrumental set, and the departing term $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$, used when $\mathbf{Z}_1$ is the instrumental set. If $\mathbf{Z}_3$ is empty, then the lower bound obtained from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$ is the same as the one obtained from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2)$. In this case, since the upper bound $H(\mathbf{X}|\mathbf{Z}\backslash\mathbf{Z}_4, \mathbf{B}_0)$ is equal

to or higher than $H(\mathbf{X}|\mathbf{Z}, \mathbf{B}_0)$, the entropic inequality derived with instrumental set $\mathbf{Z}_1$ does not add causal inference power. If $\mathbf{Z}_3 \neq \varnothing$, then in $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$ the variables $\mathbf{Z}_3$ are part of the conditioning set but do not appear in any of the conditioning sets of the conditional independencies associated with the DP inequalities. This means that to apply the first DP inequality in the first iteration of Theorem 1 the marginals preserved are relaxed, separating $\mathbf{Z}_3$ in order to obtain the marginal $P(\bar{\mathbf{Z}}_1, \bar{\mathbf{U}}_1, \mathbf{A}_1, \bar{\mathbf{B}}_1)$, with $\bar{\mathbf{Z}}_1 \subseteq \mathbf{Z}_1$. That is, the marginal $P(\bar{\mathbf{Z}}_1, \mathbf{C}_1)$ preserved jointly with $P(\bar{\mathbf{Z}}_1, \bar{\mathbf{U}}_1, \mathbf{A}_1, \bar{\mathbf{B}}_1)$ if starting from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2)$, is replaced by $P(\bar{\mathbf{Z}}_1, \mathbf{C}'_1)$ with $\mathbf{C}'_1 = \{\mathbf{C}_1, \mathbf{Z}_3\}$ when starting from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$. After the first iteration the series of preserved marginals $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$ is the same as $P(\bar{\mathbf{Z}}_j, \mathbf{C}'_j)$, that is, $\mathbf{C}'_j = \mathbf{C}_j$ for $j > 1$, since they are determined from $P(\bar{\mathbf{Z}}_1, \bar{\mathbf{U}}_1, \mathbf{A}_1, \bar{\mathbf{B}}_1)$ recursively constructing them as $\mathbf{C}_j = \{\mathbf{A}_{j-1}, \bar{\mathbf{B}}_{j-1}, \check{\mathbf{Z}}_j\}$. Lemma A3 (see below) guarantees that each resulting observable term added to the lower bound with the application of each DP inequality will be equal or smaller starting from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3)$ than starting from $I(\mathbf{Z}_1; \mathbf{U}|\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2)$.

Assembling the upper and lower bounds, the upper bound obtained using the whole instrumental set $\mathbf{Z}$ is smaller than or equal to the one obtained using the subset $\mathbf{Z}_1$. On the other hand, each term in the lower bound is higher or equal with $\mathbf{Z}$, which means that the lower bound is higher or equal. Altogether, the instrumental inequality constructed with $\mathbf{Z}$ will always be violated when the one constructed with $\mathbf{Z}_1$ is violated. Therefore, the test with $\mathbf{Z}_1$ does not provide additional causal inference power. □

Note that Proposition A1 compares the power of two specific types of tests derived with an instrumental set $\mathbf{Z}$ or its subsets. Concretely, the comparison regards tests derived using the same set of hidden independencies to apply DP inequalities. On the other hand, this hierarchy does not preclude from other tests with an instrumental set $\mathbf{Z}_1 \subset \mathbf{Z}$ to add causal inference power. If with instrumental set $\mathbf{Z}$ in iteration j a certain independence $\bar{\mathbf{Z}}_j \perp \mathbf{A}_j|\bar{\mathbf{U}}_j\bar{\mathbf{B}}_j$, with $\bar{\mathbf{Z}}_j \subseteq \mathbf{Z}_1$ is exploited, alternatively, using as instrumental set $\mathbf{Z}_1$ another independence $\bar{\mathbf{Z}}'_j \perp \mathbf{A}'_j|\bar{\mathbf{U}}'_j\bar{\mathbf{B}}'_j\mathbf{Z}_{3j}$ could be exploited, with $\bar{\mathbf{Z}}'_j \subseteq \mathbf{Z}_1$ and $\mathbf{Z}_{3j} \subseteq \mathbf{Z}_3$. Contrarily to the case in which the same independencies are applied, now the inserted estimable information terms added in the lower bound are not comparable with Lemma A3.

We now present Lemma A3, which is used in the proof above. For the sake of space, instead of stating the property in Lemma A3 in general terms and then showing its application in Proposition A1, we directly present it as applied in the proof.

Lemma A3 (Decrease of information through conditioning for minInf mutual information terms preserving only marginals)**.** *Consider nonoverlapping sets of variables* $\mathbf{Z}$, $\mathbf{B}_0$, $\mathbf{X}$, *and* $\mathbf{U}$, *all observable except* $\mathbf{U}$ *hidden variables. Consider that* $\mathbf{Z} \perp \mathbf{U}|\mathbf{B}_0$. *Consider a nonoverlapping partition* $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4\}$, *with* $\mathbf{Z}_1$ *nonempty, and* $\mathbf{Z}_2$, $\mathbf{Z}_3$, *and* $\mathbf{Z}_4$ *possibly empty. Consider a set of independencies that allows constructing instrumental entropic inequalities by recursively applying Theorem 1 with initial inputs* $\bar{\mathbf{Z}} = \mathbf{Z}_1$ *and* $\mathbf{E} = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2\}$. *Consider a derived estimable term* $\min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, *where* ΔP_k *preserves* $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *and* $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, *for* $j = 1, \ldots, k$. *Consider another estimable term derived with Theorem 1 using the same set of independencies but starting with* $\bar{\mathbf{Z}} = \mathbf{Z}_1$ *and* $\mathbf{E}' = \{\mathbf{B}_0, \mathbf{X}, \mathbf{Z}_2, \mathbf{Z}_3\}$, *with the form* $\min_{Q \in \Delta P'_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}', \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, *where* $\Delta P'_k$ *preserves* $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ *and* $P(\bar{\mathbf{Z}}_j, \mathbf{C}'_j)$, *with* $\mathbf{C}'_j = \mathbf{C}_j$ *for* $j = 2, \ldots, k$ *and* $\mathbf{C}'_1 = \{\mathbf{C}_1, \mathbf{Z}_3\}$. *The minInf terms are such that*

$$\min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}) \geqslant \min_{Q \in \Delta P'_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}', \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}). \tag{A18}$$

Proof. Consider a distribution that minimizes the l.h.s. of Equation (A18)

$$Q^*(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}) \equiv \arg \min_{Q \in \Delta P_k} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}). \tag{A19}$$

For this distribution, the information $I_{Q^*}(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ does not depend on $\mathbf{E}'\backslash\mathbf{E} = \mathbf{Z}_3$. A joint distribution with $\mathbf{Z}_3$ can be created as $\bar{Q} \equiv P(\mathbf{Z}_3|\bar{\mathbf{Z}}_1, \mathbf{C}_1)Q^*(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, with $\bar{\mathbf{Z}}_1$ and $\mathbf{C}_1$ identified by the constraint of ΔP_k on the marginal $P(\bar{\mathbf{Z}}_1, \mathbf{C}_1)$. The decomposition of information into unique information and redundancy [29] can then be applied to $\bar{Q}$, obtaining the conditional unique information (Equation (5))

$$I_{\bar{Q}}(\bar{\mathbf{Z}}_k; \mathbf{A}_k \backslash\backslash \mathbf{Z}_3|\mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}) = \min_{Q \in \Delta\bar{Q}} I_Q(\bar{\mathbf{Z}}_k; \mathbf{A}_k|\mathbf{E}', \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]}), \tag{A20}$$

where by construction $\mathbf{E}' = \{\mathbf{E}, \mathbf{Z}_3\}$, and $\Delta\bar{Q}$ is the family of distributions that preserve $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$ and $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{E}', \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$. This unique information is by construction (Equation (6)) smaller than or equal to the l.h.s. of Equation (A18). We now compare it with the term in the r.h.s. of Equation (A18). The conditional information minimized is the same. $\Delta\bar{Q}$ preserves $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, which by construction is $Q^*(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \mathbf{E}, \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$. Since Q^* belongs to the family ΔP_k, this constraint implies preserving $P(\bar{\mathbf{Z}}_k, \mathbf{A}_k, \bar{\mathbf{B}}_k)$ and $P(\bar{\mathbf{Z}}_j, \mathbf{C}_j)$, for $j = 1, \ldots, k$. Furthermore, $\Delta\bar{Q}$ preserves $\bar{Q}(\bar{\mathbf{Z}}_k, \mathbf{E}', \mathbf{A}^{[k-1]}, \check{\mathbf{Z}}^{[k]})$, which given the construction of $\bar{Q}$ as $\bar{Q} = P(\mathbf{Z}_3|\bar{\mathbf{Z}}_1, \mathbf{C}_1)Q^*$ implies the preservation of the marginal $P(\bar{\mathbf{Z}}_1, \mathbf{C}_1, \mathbf{Z}_3) = P(\bar{\mathbf{Z}}_1, \mathbf{C}'_1)$. Accordingly, the constraints of $\Delta P'_k$ are superseeded by the constraints of $\Delta\bar{Q}$ and hence the information term can be further minimized within $\Delta P'_k$ in the r.h.s. of Equation (A18). □

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. Verma, T. *Graphical Aspects of Causal Models*; Technical Report R-191; Computer Science Department, UCLA: Los Angeles, CA, USA, 1993.
5. 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]
6. 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.
7. Tian, J.; Pearl, J. On the Testable Implications of Causal Models with Hidden Variables. In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, Edmonton, AB, Canada, 1–4 August 2002.
8. 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, 8–10 December 2008; pp. 689–696.
9. 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.
10. 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.
11. Chicharro, D.; Panzeri, S.; Shpitser, I. Conditionally-additive-noise models for structure learning. *arXiv* **2019**, arXiv:1905.08360.
12. Chicharro, D.; Besserve, M.; Panzeri, S. Causal learning with sufficient statistics: An information bottleneck approach. *arXiv* **2020**, arXiv:2010.05375. [CrossRef]
13. Parbhoo, S.; Wieser, M.; Wieczorek, A.; Roth, V. Information Bottleneck for Estimating Treatment Effects with Systematically Missing Covariates. *Entropy* **2020**, *22*, 389. [CrossRef] [PubMed]
14. Fritz, T.; Chaves, R. Entropic inequalities and marginal problems. *IEEE Trans. Inf. Theory* **2013**, *59*, 803–817. [CrossRef]
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*, 880–884. [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), Seattle, WA, 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 18th Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 13–16 July 2006.
22. Yeung, R.W. *Information Theory and Network Coding*; Springer: Berlin/Heidelberg, Germany, 2008.
23. Pawłowski, M.; Paterek, T.; Kaszlikowski, D.; Scarani, V.; Winter, A.; Zukowski, M. Information causality as a physical principle. *Nature* **2009**, *461*, 1101–1104. [CrossRef] [PubMed]
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, QC, Canada, 23–27 July 2014; pp. 112–121.
25. Steudel, B.; Ay, N. Information-Theoretic Inference of Common Ancestors. *Entropy* **2015**, *17*, 2304–2327. [CrossRef]
26. Chicharro, D.; Nguyen, J.K. Causal Structure Learning with Conditional and Unique Information Groups-Decomposition Inequalities. *Entropy* **2024**, *26*, 440. [CrossRef]
27. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2006.
28. Yeung, R.W. A framework for linear information inequalities. *IEEE Trans. Inform. Theory* **1997**, *43*, 1924–1934. [CrossRef]
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. [CrossRef]
31. Chaves, R.; Majenz, C.; Gross, D. Information-theoretic implications of quantum causal structures. *Nat. Commun.* **2015**, *6*, 5766. [CrossRef]
32. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]
33. Wright, P.G. *The Tariff on Animal and Vegetable Oils*; Macmillan: New York, NY, USA, 1928.
34. Balke, A.; Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* **1997**, *92*, 1171–1176. [CrossRef]
35. El Gamal, A.; Kim, Y.H. *Network Information Theory*; Cambridge University Press: Cambridge, UK, 2011.
36. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley and Sons: Hoboken, NJ, USA, 2001.
37. Globerson, A.; Tishby, N. The minimum information principle for discriminative learning. In Proceedings of the 20th conference on Uncertainty in Artificial Intelligence (UAI), Banff, AB, Canada, 7–11 July 2004; pp. 193–200.
38. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
39. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130. [CrossRef]
40. Ince, R.A.A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318–344. [CrossRef]
41. James, R.G.; Emenheiser, J.; Crutchfield, J.P. Unique Information via Dependency Constraints. *J. Phys. A Math. Theor.* **2019**, *52*, 014002. [CrossRef]
42. Rauh, J.; Bertschinger, N.; Olbrich, E.; Jost, J. Reconsidering unique information: Towards a multivariate information decomposition. In Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT 2014), Honolulu, HI, USA, 29 June–4 July 2014; pp. 2232–2236.
43. Dawid, A.P. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B* **1979**, *41*, 1–31. [CrossRef]
44. Horst, R.; Pardalos, P.M.; Thoai, N.V. *Introduction to Global Optimization: Nonconvex Optimization and Its Applications*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2000.
45. Chicharro, D. Quantifying multivariate redundancy with maximum entropy decompositions of mutual information. *arXiv* **2017**, arXiv:1708.03845v1.
46. Chicharro, D.; Ledberg, A. Framework to study dynamic dependencies in networks of interacting processes. *Phys. Rev. E* **2012**, *86*, 041901. [CrossRef] [PubMed]
47. 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]
48. 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]
49. Celotto, M.; Bím, J.; Tlaie, A.; De Feo, V.; Toso, A.; Lemke, S.M.; Chicharro, D.; Nili, H.; Bieler, M.; Donner, T.H.; 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.
50. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
51. Hiemstra, C.; Jones, J.D. Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. *J. Financ.* **1994**, *49*, 1639–1664.

52. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin/Heidelberg, Germany, 2006.
53. Geweke, J.F. Measurement of Linear Dependence and Feedback Between Multiple Time Series. *J. Am. Stat. Assoc.* **1982**, *77*, 304–313. [CrossRef]
54. 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. 195–219.
55. Pica, G.; Piasini, E.; Safaai, H.; Runyan, C.A.; Diamond, M.E.; Fellin, T.; Kayser, C.; Harvey, C.D.; Panzeri, S. Quantifying how much sensory information in a neural code is relevant for behavior. In Proceedings of the 31st Conference on Neural Information Processing System (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 3687–3697.
56. Koçillari, L.; Lorenz, G.M.; Engel, N.M.; Celotto, M.; Curreli, S.; Malerba, S.B.; Engel, A.K.; Fellin, T.; Panzeri, S. Finite-sampling bias correction for discrete Partial Information Decomposition. *bioRxiv* **2024**, bioRxiv:2024.06.04.597303.
57. Chicharro, D. On the spectral formulation of Granger causality. *Biol. Cybern* **2011**, *105*, 331–347. [CrossRef] [PubMed]
58. 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]
59. Tavakoli, A.; Pozas-Kerstjens, A.; Luo, M.; Renou, M.O. Bell nonlocality in networks. *Rep. Prog. Phys.* **2022**, *85*, 056001. [CrossRef] [PubMed]
60. Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information*; Cambridge University Press: Cambridge, UK, 2000.
61. 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.
62. Wolfe, E.; Spekkens, R.W.; Fritz, T. The Inflation Technique for Causal Inference with Latent Variables. *J. Caus. Inf.* **2019**, *7*, 20170020. [CrossRef]
63. Evans, R.J. Latent-free equivalent mDAGs. *arXiv* **2022**, arXiv:2209.06534. [CrossRef]
64. Navascués, M.; Elie Wolfe, E. The Inflation Technique Completely Solves the Causal Compatibility Problem. *J. Causal Infer.* **2020**, *8*, 70–91. [CrossRef]
65. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
66. James, R.G.; Ellison, C.J.; Crutchfield, J.P. dit: A Python package for discrete information theory. *J. Open Source Softw.* **2018**, *3*, 738. [CrossRef]
67. Makkeh, A.; Chicharro, D.; Theis, D.O.; Vicente, R. MAXENT3D_PID: An estimator for the maximum-entropy trivariate partial information decomposition. *Entropy* **2019**, *21*, 862. [CrossRef]
68. Barrett, A.B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **2015**, *91*, 052802. [CrossRef]
69. Goswami, C.; Merkley, A. Analytically deriving Partial Information Decomposition for affine systems of stable and convolution-closed distributions. In Proceedings of the 37th Conference on Neural Information Processing System (NIPS), Vancouver, BC, Canada, 10–15 December 2024; pp. 86749–86835.
70. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. [CrossRef]
71. Sun, Y.; Babu, P.; Palomar, D.P. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.* **2017**, *65*, 794–816. [CrossRef]
72. Danilova, M.; Dvurechensky, P.; Gasnikov, A.; Gorbunov, E.; Guminov, S.; Kamzolov, D.; Innokentiy, S. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*; Optimization and Its Applications; Nikeghbali, A., Pardalos, P.M., Raigorodskii, A.M., Rassias, M.T., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 191, pp. 79–163.
73. Le Thi, H.A.; Pham Dinh, T. DC Programming and DCA: Thirty years of developments. *Math. Program.* **2018**, *169*, 5–68. [CrossRef]
74. Bedford, T.; Wilson, K.J. On the construction of minimum information bivariate copula families. *Ann. Inst. Statist. Math.* **2014**, *66*, 703–723.
75. Sei, T.; Yano, K. Minimum information dependence modeling. *Bernoulli* **2024**, *30*, 2623–2643. [CrossRef]
76. Marton, K. A Coding theorem for the discrete memoryless broadcast channel. *IEEE Trans. Inf. Theory* **1979**, *25*, 306–311. [CrossRef]
77. Gray, R.M.; Wyner, A.D. Source coding for a simple network. *Bell Syst. Technol. J.* **1974**, *53*, 1681–1721. [CrossRef]
78. Han, T.S.; Kobayashi, K. A new achievable rate region for the interference channel. *IEEE Trans. Inform. Theory* **1981**, *27*, 49–60. [CrossRef]
79. Csiszár, I.; Körner, J. Broadcast Channels with Confidential Messages. *IEEE Trans. Inform. Theory* **1978**, *24*, 339–348. [CrossRef]

80. Treves, A.; Panzeri, S. The upward bias in measures of information derived from limited data samples. *Neural Comput.* **1995**, *7*, 399–407. [CrossRef]
81. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *17*, 1191–1253. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.