Proceedings of Machine Learning Research 323:1–21, 2026 5th Conference on Causal Learning and Reasoning

Information-theoretic signatures of causality in Bayesian networks and hypergraphs

Sung En Chiang SUNG.CHIANG24@IMPERIAL.AC.UK
Department of Mathematics, Imperial College London

Zhaolu Liu ZHAOLU.LIU16@IMPERIAL.AC.UK
Department of Mathematics, Imperial College London

Robert L. Peach PEACH_R@UKW.DE
Department of Neurology, University Hospital Würzburg
Department of Brain Sciences, Imperial College London

Mauricio Barahona M.BARAHONA@IMPERIAL.AC.UK
Department of Mathematics, Imperial College London

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Analyzing causality in multivariate systems involves establishing how information is generated, distributed and combined. Traditional causal discovery frameworks are capable of multivariate reasoning but their intrinsic pairwise graph topology restricts them to do so only indirectly by integrating multivariate information across pairwise edges. Higher-order information theory provides direct tools that can explicitly model higher-order interactions. In particular, Partial Information Decomposition (PID) allows the decomposition of the information that a set of sources provides about a target into redundant, unique, and synergistic components. Yet the mathematical connection between such higher-order information-theoretic measures and causal structure remains undeveloped. Here we establish the first theoretical correspondence between PID components and causal structure in both Bayesian networks and hypergraphs. We first show that in Bayesian networks unique information precisely characterizes direct causal neighbors, while synergy identifies collider relationships. This establishes a localist causal discovery paradigm in which the structure surrounding each variable can be recovered from its immediate informational footprint, eliminating the need for global search over graph space. Extending these results to more expressive causal representation, we prove that PID signatures in Bayesian hypergraphs differentiate parents, children, co-heads, and co-tails, revealing a novel collider effect unique to multi-tail hyperedges. Our results position PID as a rigorous, model-agnostic foundation for inferring both pairwise and higher-order causal structure, and introduce a fundamentally local information–theoretic viewpoint on causal discovery.

Keywords: higher-order interactions; causal discovery; bayesian networks; hypergraphs

1. Introduction

An important task in statistical and machine learning research is to understand how complex multivariate systems are structured: which variables influence which; how information is shared between them; in what ways their interactions extend beyond simple correlations. Classical structural models like Bayesian networks encode some of those relationships through directed edges, and are interpreted via conditional factorization and independence (Pearl, 2009; Spirtes et al., 2000). While these edge-based models in principle are able to capture complex higher-order interactions, inferring them requires aggregating information across the global structure since graphs intrinsically

© 2026 S.E. Chiang, Z. Liu, R.L. Peach & M. Barahona.

encode pairwise relations in local topology. Such higher-order interactions are crucial in many real-world domains, including neural, biological and social systems (Luppi et al., 2022; Schneidman et al., 2003; Iacopini et al., 2019) systems, where higher-order structures (e.g., hypergraphs or simplicial complexes) offer more accurate representations of the system and the data (Torres et al., 2021; Bick et al., 2023). Owing to this, in parallel, a rich body of work on higher-order information-theoretic measures from Shannon information theory Shannon (1948) has produced a range of tools for quantifying higher-order interactions (Liu et al., 2025; Rosas et al., 2019; Bell, 2003), yet their relationship to causality has remained conceptually underdeveloped.

Among information-theoretic higher-order measures, Partial Information Decomposition (PID) offers a principled framework for analyzing the multivariate information that a set of sources conveys about a target, separating it into redundant, unique, and synergistic components (Williams and Beer, 2010), and thereby capturing the higher-order information-theoretic behavior that traditional graphical models are unable to represent. However, despite extensive use across biomedicine (Varley et al., 2023; Luppi et al., 2024; Cang and Nie, 2020) and machine learning (Liang et al., 2023, 2024; Makkeh et al., 2025), PID has so far been primarily used as a descriptive metric, and lacks a structural interpretation that relates its components to graphical roles. Recent frameworks that introduce PID to causal analysis either rely on known temporal orderings to measure past-to-future information flow Mediano et al. (2025); Martínez-Sánchez et al. (2024) or are constrained by additive noise assumptions and single-tail restrictions (Enouen et al., 2025).

To address these limitations, here we provide the first systematic correspondence between PID components and structural roles in both pairwise and higher-order graph structures, yielding a localist interpretation of causality. Standard causal discovery frameworks (Kalisch and Bühlman, 2007; Pearl, 2009; Spirtes et al., 2000) often execute local conditional independence tests, yet their underlying mechanics are fundamentally global. Because they are designed to construct a consistent network-wide equivalence class, a single edge orientation inevitably triggers widespread constraint propagation across the graph to prevent cycles and false v-structures. Similarly, score-based methods may utilize mathematically decomposable metrics, but these scores ultimately serve merely as heuristics to guide an algorithmic search such as the hill-climbing algorithm Tsamardinos et al. (2003, 2006) under global structural constraints. In contrast, our PID-based localist perspective of causality focuses on one region of the system at a time: for any variable of interest, we study its immediate neighborhood by computing the relevant PID components to characterize the structural roles of the variables surrounding it in the (hyper)graph.

We start by developing this mathematical correspondence for Bayesian networks (Bishop, 2007). In particular, we find that unique information in PID corresponds to direct causal neighbors, and the edge directions are identified through the synergistic patterns, yielding a two-step procedure. We then extend these concepts to Bayesian hypergraphs (Javidian et al., 2020), a generalization of Bayesian networks that incorporates mixtures of directed and undirected relations and can naturally encode higher-order dependence. We establish an analogous correspondence between hypergraph structure and PID patterns: unique information characterizes membership in the tail or head of a hyperedge (or symmetric co-head relations) and synergy isolates co-tail structure, thus revealing the higher-order collider effect unique to hypergraphs. Together, these signatures specify the full informational footprint of a hyperedge. Furthermore, since a particular PID pattern may be compatible with multiple combinations of hyperedges, we developed a procedure for identifying hyperedges with local information patterns by selecting the largest (maximal) hyperedge consistent with the pattern.

Our contributions are: (i) We enrich the multivariate reasoning with fine-grained interaction decomposition by establishing a formal link between multivariate Shannon information theory and causal structures by mapping PID components to causal roles in graphical models. (ii) We show that this mapping yields a localist information-theoretic characterization of causal neighborhoods in Bayesian networks. (iii) We extend this framework to Bayesian hypergraphs, a higher-order representation that explicitly encode multivariate information. (iv) We provide procedures that can be implemented to provide systematic causal discovery in both Bayesian networks and hypergraphs using this localist perspective.

2. Preliminaries

2.1. Partial Information Decomposition

Let us consider a set of d source variables $S = \{S_1, \ldots, S_d\}$ and a target variable T. The Partial Information Decomposition framework provides a principled way to dissect the mutual information between the set S and T, $I(S;T)$, by splitting it into components that represent distinct modes of information sharing: the unique, redundant, and synergistic contributions from the sources to the target (Williams and Beer, 2010). At the core of the PID formalism lies a combinatorial object, the redundancy lattice (Williams and Beer, 2010). A lattice is a set equipped with a partial order such that every pair of elements in the set has a well-defined meet and join (Rota, 1964; Birkhoff, 1940). In the PID setting, the underlying set $\mathcal{A}(S)$ is the collection of antichains drawn from the powerset of the powerset of S. Formally, each element $\alpha \in \mathcal{A}(S)$ is a collection of non-empty subsets $A_i \subseteq S$ such that no $A_i \in \alpha$ is a subset of another (Williams and Beer, 2010):

$$\mathcal{A}(S) = \{\alpha \in \mathcal{P}_1(\mathcal{P}_1(S)) : \forall A_i, A_j \in \alpha, A_i \not\subset A_j\} \ ,$$

where $\mathcal{P}_1(S)$ denotes the powerset of S excluding the empty set. The partial order $\preceq$ on $\mathcal{A}(S)$ is defined by

$$\alpha \preceq \beta \quad \Longleftrightarrow \quad \forall B \in \beta, \exists A \in \alpha \text{ such that } A \subseteq B \, .$$

This ordering reflects containment: $\alpha \preceq \beta$ if every subset in β contains at least one subset in α so that the information represented by β subsumes that represented by α.

The key link between the redundancy lattice and PID is that any element $\beta \in \mathcal{A}(S)$ is associated with a redundancy measure $I_\cap(\beta;T)$ that quantifies the amount of information about T that is redundantly available from any $B \in \beta$. Crucially, $I_\cap(\beta;T)$ can be expressed as a sum of partial information atoms, $\Pi_R(\alpha;T)$, which represent the irreducible units of shared information and can be viewed as the Zeta function in the Möbius inversion theorem (Rota, 1964):

$$I_\cap(\beta;T) = \sum_{\alpha \preceq \beta} \Pi_R(\alpha;T) \, . \tag{1}$$

Inverting Equation (1) via Möbius recovers the partial information atoms from the redundancy function (Williams and Beer, 2010).

In the case of two sources, $S = \{S_1, S_2\}$, the redundancy lattice has four elements: $\{\{S_1\},\{S_2\}\}$, $\{\{S_1\}\}$, $\{\{S_2\}\}$ and $\{\{S_1, S_2\}\}$, and the four corresponding partial information atoms can be interpreted as redundant, unique and synergistic components:

$$\begin{aligned} I(S_1, S_2; T) = I_\cap(\{\{S_1, S_2\}\}, T) = \mathrm{Red}(\{\{S_1\},\{S_2\}\};T) + \mathrm{Uniq}(\{\{S_1\}\};T) \\ + \mathrm{Uniq}(\{\{S_2\}\};T) + \mathrm{Syn}(\{\{S_1, S_2\}\};T) \, . \end{aligned}$$

Red denotes the redundant contribution of the two sources; Uniq, information uniquely attributable to one source; and Syn, synergistic information available only when both are taken together.

Remark: The self-redundancy axiom (Williams and Beer, 2010) implies that the mutual information between sources and target is equal to the redundancy measure of the maximum element in the redundancy lattice, i.e., $I(S,T) = I_\cap(\{\{S_1,\ldots,S_d\}\},T)$.

Remark: For $d > 2$, the situation is more intricate. Although inverting Eq. (1) via Möbius inversion allows in principle the recovery of the partial information atoms from the redundancy measure, the size of the redundancy lattice grows according to the Dedekind numbers (Jäkel, 2023), which makes computation infeasible beyond $d > 3$. The structures of the redundancy lattice for $d = 2$ and $d = 3$ are shown in Figure 3 in Appendix A.1.

2.2. Bayesian network

A natural choice to encode the interactions within a set of variables is through a graph $\mathcal{G} = (V, E)$, where the set of vertices V is associated with the variables and the set of edges $E \subseteq V \times V$ capture the interactions between variables. When $\mathcal{G}$ is directed, an edge is an ordered pair $(X_i, X_j) \in E$, which indicates that X_i is a parent of X_j, or equivalently, X_j is a child of X_i. The set of parents (children) of node X_i in the graph is denoted as $\mathrm{pa}(X_i)$ ($\mathrm{ch}(X_i)$). A directed acyclic graph (DAG) is a directed graph with no sequence of directed edges that forms a closed loop.

Causal structural graphs are formulated as Bayesian networks (Bishop, 2007), i.e., DAGs where the d vertices represent random variables $\{X_1,\ldots,X_d\}$ with a distribution function $\mathbb{P}(X_1,\ldots,X_d)$. Then $\mathbb{P}$ can be factorized as:

$$\mathbb{P}(X_1,\ldots,X_d) = \prod_{i=1}^{d} \mathbb{P}(X_i \mid \mathrm{pa}(X_i))\,.$$

See Figure 1a for an example. The d-separation criterion then provides a systematic method for translating the Bayesian network into a set of conditional independence properties defined by the DAG structure (Bishop, 2007).

While Bayesian networks are suitable for representing pairwise, directed relationships, their intrinsic edge-based topology is not designed to explicitly encode higher-order interactions. To naturally capture these complex, higher-order directed and undirected relationships, we must turn to a more expressive causal graphical representation, the Bayesian hypergraph.

2.3. Bayesian hypergraph

A hypergraph $\mathcal{H} = (V, \mathcal{E})$ is an extension of pairwise graphs where each hyperedge $\varepsilon \in \mathcal{E}$ is associated with a subset of the nodes in V with arbitrary cardinality. In a directed hypergraph, every hyperedge is an ordered pair of subsets $\varepsilon = (T(\varepsilon), H(\varepsilon))$ such that the tail $T(\varepsilon)$ and head $H(\varepsilon)$ do not overlap, and the direction is from tail to head. (Note that the head or tail could be the empty set.) Then we say that $t \in T(\varepsilon)$ is a parent of $h \in H(\varepsilon)$, or, equivalently, h is a child of t. Furthermore, two vertices $X_i, X_j \in V$ are said to be co-heads (co-tails) when there exists a directed hyperedge ε whose head (tail) contains both vertices. A partially directed cycle in $\mathcal{H}$ is a sequence of vertices $(X_1, X_2, \ldots, X_k)$ with $X_{k+1} = X_1$ such that each X_i is either a co-head or a parent of its successor in the sequence, X_{i+1}, for all $1 \le i \le k$. A directed acyclic hypergraph is a directed hypergraph with no partially directed cycles. A hyperedge ε_1 is contained in hyperedge ε_2

if $T(\varepsilon_1) \cup H(\varepsilon_1) \subset T(\varepsilon_2) \cup H(\varepsilon_2)$. When a hyperedge is not contained by any other hyperedges it is called irreducible. See Javidian et al. (2020) for definitions.

A Bayesian hypergraph is a probabilistic model defined via a directed acyclic hypergraph, where the nodes represent random variables and the hyperedges encapsulate dependencies between sets of variables (Javidian et al., 2020). The Bayesian hypergraph provides a principled integration of directed and undirected relationships through a two-level hierarchical factorization structure, as follows. First, the variables are coarse-grained into *chain components* τ; these are equivalence classes where X_i and X_j belong to the same τ if there exists a sequence of distinct variables $(X_i, \ldots, X_j)$ such that each consecutive pair in the sequence are co-heads. A canonical DAG is formed by taking the chain components of $\mathcal{H}$ as nodes and a directed edge from a chain component τ_1 to another chain component τ_2 is present if there exists $u \in \tau_1$, $v \in \tau_2$ and a hyperedge ε such that $u \in T(\varepsilon)$ and $v \in H(\varepsilon)$. The joint distribution then factorizes similarly to a Bayesian network over the set of chain components, $\mathcal{D}$:

$$\mathbb{P}(X_1, \ldots, X_d) = \prod_{\tau \in \mathcal{D}} \mathbb{P}(X_\tau \mid X_{\text{pa}(\tau)}) \,,$$

where X_τ denotes the set of variables in chain component τ, and $\text{pa}(\tau)$ represents the parents of chain component τ in the canonical DAG of the hypergraph. Second, within each τ, the conditional distribution functions can be factorized into non-negative potential functions, akin to those in Markov random fields (Bishop, 2007), each specified by an irreducible hyperedge. This leads to the following factorization over the set of irreducible hyperedges in τ, $\mathcal{M}_\tau$:

$$\mathbb{P}(X_\tau \mid X_{\text{pa}(\tau)}) \propto \prod_{\varepsilon \in \mathcal{M}_\tau} \psi_\varepsilon(X_\varepsilon) \,,$$

where $\psi_\varepsilon(X_\varepsilon)$ denotes the potential function associated with the irreducible hyperedge ε. An example of Bayesian hypergraph with its factorization is shown in Figure 1b.

Remark: The patterns of conditional independence in Bayesian hypergraphs differ from Bayesian networks. In Bayesian networks, conditioning on a child induces conditional dependence among all parents (Fig. 1a). This is not necessarily the case in Bayesian hypergraphs since conditioning on the child activates dependence only between co-tails within the same hyperedge leaving other parent pairs conditionally independent (Fig. 1b). Hence Bayesian hypergraphs are strictly more expressive than Bayesian networks in both their probabilistic factorization patterns and their graphical representations.

Remark: Another key advantage of directed hypergraphs is that the co-head relationship provides a natural framework for accommodating indeterminacy in the causal direction between variables. When conditional independence queries reveal dependencies without clear directional information, hypergraphs can represent these relationships through co-head structures that capture the symmetric aspects of the interaction while preserving the overall causal hierarchy.

3. PID Signatures of Causal Structures

In this section, we first introduce theoretical conditions to establish a bidirectional correspondence between causal roles and information-theoretic signatures for Bayesian networks. We then generalize these results to Bayesian hypergraphs, where the richer structural relationships enable more nuanced PID signatures for causal configurations. The proofs can be found in Appendix B. We start

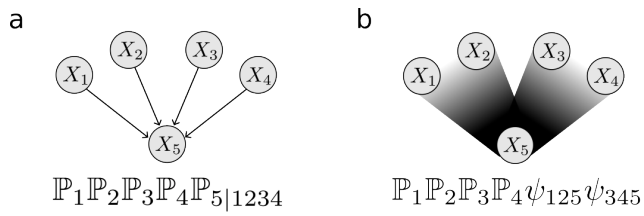

Figure 1: **Factorization of distributions according to Bayesian network and Bayesian hypergraphs.** (a) A Bayesian network where X_1, X_2, X_3 and X_4 are the parents of X_5. (b) A Bayesian hypergraph where X_1 and X_2 (similarly X_3 and X_4) are co-parents of X_5. Conditioning on X_5 only induces dependence between $\{X_1, X_2\}$ and $\{X_3, X_4\}$ whereas in (a) conditioning on X_5 induces dependence among all parent pairs. The factorizations (up to a normalizing constant) are given below each structure. Hyperedges are shaded directionally from (co-)tail (light) to (co-)head (dark).

by defining our problem, desiderata[1] and assumptions, and establishing our theoretical building blocks.

Desideratum 1 (Non-negative PID atoms) *It is desirable that the choice of redundancy measure leads to non-negative atoms for all elements of the redundancy lattice:* $\Pi_R(\alpha; T) \geq 0, \forall \alpha \in \mathcal{A}(S)$.

As there is indeterminacy in the definition of the redundancy measure in PID (Lizier et al., 2018; Rauh et al., 2014), Desideratum 1 ensures that each lattice element contributes a meaningful (non-negative) amount of information to the PID decomposition (Eq. (1)). For bivariate PID (i.e., $d = 2$ source variables), it is possible to decompose $I(S_1, S_2; T)$ into *non-negative* unique, synergistic, and redundant components following Bertschinger et al. (2014). For $d > 2$, however, Lyu et al. (2025) demonstrated that no redundancy measure can always guarantee non-negativity of all PID components while satisfying other intuitive desiderata, e.g., independent additivity guaranteeing that two independent variables have zero redundant information.

To circumvent this difficulty, in this paper we employ a formalism based on bivariate PID but apply it to systems involving d variables by building 'supervariables' that contain the complement set of any two variables of interest. This allows us to exploit the well-defined bivariate PID components (unique, synergistic, redundant) of a variable but now with respect to the rest of variables. For a set of variables V with target variable $X_k \in V$ and source variable $X_i \in V \setminus \{X_k\}$, we thus define the shorthand notation:

$$\Pi_R^{V'}(\{\{X_i\}\}; X_k) \equiv \Pi_R^{\{X_i, X'\}}(\{\{X_i\}\}; X_k)\,,$$

where $X' = V \setminus \{X_i, X_k\}$ is the supervariable containing all variables except X_i and the target X_k, and $V' = \{X_i, X'\}$ is the bivariate source set for the PID decomposition. By grouping the rest of variables into X', this notation ensures that the unique information represents only what X_i can provide that is truly unavailable from any combination of the remaining variables. This formulation enables PID causal formalization for d variables while maintaining the theoretical guarantees of bivariate PID. Note that increasing the source variables from S to $S' \supset S$ increases the lattice elements that precede any particular element in the lattice ordering; hence it can occur that previously

1. Desiderata are properties desired for the PID measures to fit in our causal discovery framework, whereas the assumptions are precise mathematical statements we accept as true to prove the theorems.

unique information is reclassified as redundant in the expanded source set. Therefore, our notation for the PID components includes explicitly the source set S, i.e., $\Pi_R^S(\alpha; T)$.

First, we establish a fundamental result on the relationship between conditional mutual information and unique information, namely, that the vanishing of conditional mutual information implies zero unique information.

Theorem 1 (Conditional independence and unique information) *Let $C \subseteq V$ be a conditioning set, and $X_i, X_j \in V \setminus C$. Given Desideratum 1, if there exists C such that $I(X_i; X_j \mid C) = 0$, then X_j has zero unique information when X_i is set as the target variable, i.e., $\Pi_R^{V'}(\{\{X_j\}\}; X_i) = 0$.*
Proof *See Section B.1 of the appendix.* ■

This theorem is a crucial building block for establishing the connection between direct causal neighbors and unique information patterns. For any variable that is not a direct causal neighbor of the target, we can always construct some conditioning set that renders their conditional mutual information zero, and through this theorem we can then establish that such variables contribute zero unique information to the target.

Another key condition that should be satisfied is the monotonicity of unique information.

Desideratum 2 (Monotonicity of unique information) *Let X_1, X_2 be the source variables, and for any additional variable K, let $X_2' = \{X_2\} \cup \{K\}$ be the supervariable that combines K and X_2. We then require: $\Pi_R^{\{X_1, X_2'\}}(\{\{X_1\}\}; T) \leq \Pi_R^{\{X_1, X_2\}}(\{\{X_1\}\}; T)$.*

In other words, Desideratum 2 formalizes the intuitive notion that, in bivariate PID, augmenting one source variable with additional information cannot increase the unique information contribution of the other source variable. The supervariable X_2' contains all the information that X_2 originally had plus additional information from K. Therefore, any information previously unique to X_1 (relative to X_2) may now be partially or fully captured by X_2', but no new unique information for X_1 should emerge from this augmentation. This monotonicity property ensures that the notion of uniqueness remains coherent as we expand the information available to competing sources. Note how the use of supervariable X_2' allows us to remain in a bivariate PID framework.

To establish the bidirectional correspondence between causal roles and their PID signatures, we also require the following constraint on the joint probability distribution over all variables.

Assumption 1 (Persistent relevance) *If $I(X_i; T \mid C) > 0$ for all conditioning sets $C \subseteq S \setminus \{X_i\}$, then X_i must possess positive unique information about target T, i.e., $\Pi_R(\{\{X_i\}\}; T) > 0$.*

This property addresses pathological cases in PID where a variable might always maintain some dependency with the target, yet contribute zero unique information. For example, consider two independent coin flips X_1 and X_2, and let $X_3 = X_1 \oplus X_2$ such that X_1 and X_2 will only have synergistic information about X_3 due to the XOR gate. Given another independent coin flip X_4, construct new variables $X_1' = \{X_1, X_4\}, X_2' = \{X_2, X_4\}, X_3' = \{X_3, X_4\}$. Setting X_3' as the target, X_1' and X_2' will only have redundant information and synergistic information, as X_4 is perfectly replicated across three variables, which means any information about X_4 provided by X_1' can also be provided by X_2', where as information about X_3 can only be attained synergistically. Since X_3 and X_4 together determine X_3', there is no room for unique information from either variable.

3.1. PID signatures of causality in Bayesian networks

3.1.1. DIRECT CAUSAL NEIGHBORS AND UNIQUE INFORMATION

We now present one of our central results establishing the correspondence between unique information and causal neighbors. In this section, we assume the standard faithfulness condition (Spirtes et al., 2000), i.e., all statistical independencies in the joint distribution correspond precisely to the d-separations in the underlying Bayesian network.

Theorem 2 (Causal relationships and unique information) *Under Assumption 1, for any X_i and X_j in a Bayesian network, the unique information of X_j about X_i is positive iff X_j is a parent or child of X_i.*
Proof *See Section B.2 of the appendix.* ■

This result formalizes our intuition that a parent exerts an influence on the child variable that cannot be derived from any other source, while the child is influenced by the parent variable in ways that reflect information unavailable elsewhere in the system. Importantly, Theorem 2 supports a localist interpretation of causal roles. In contrast to standard causal learning approaches, which proceed by narrowing down causal models through systematic conditional independence queries across the entire variable set, this allows us to focus locally on a particular variable of interest and identify those variables that contribute unique information to it, which are either its parents or children.

Remark: The estimation of redundant information in PID, which depends on multiple variables simultaneously, can be reused across different targets. This shared structure allows the unique information for several variables to be inferred from one redundancy computation, thus reducing computational cost. Such reuse has no analogue in standard causal methods based on conditional independence tests, and suggests new possibilities for more efficient structural discovery algorithms. An example of redundancy reuse is shown in Appendix A.2.

3.1.2. CO-PARENT AND SYNERGY

After identifying the PID signature of parents and children, we now seek to distinguish between these two roles. We do this by identifying the characteristic PID signature of causal co-parents and introducing appropriate theoretical conditions to ensure that this identification is bidirectional, i.e., co-parent relationships both imply and are implied by specific PID patterns.

Before proving our result, we state the necessary assumptions. First, we introduce the collider amplification assumption which formalizes the intuition that the collider effect should always increase mutual information between co-parents, regardless of whether they have nonzero unconditional mutual information beforehand.

Assumption 2 (Collider amplification) *For every collider $X_i \to X_j \leftarrow X_k$ in a Bayesian network, we assume that the following holds: $I(X_i; X_k \mid X_j) > I(X_i; X_k)$.*

Remark: Clearly, the pure V-structure satisfies $I(X_i; X_k) = 0$ by d-separation and Assumption 2 holds directly. Yet additional dependencies between X_i and X_k induced by other variables in the graph might result in $I(X_i; X_k) > 0$. When this happens, d-separation alone cannot specify by itself the relative magnitude of $I(X_i; X_k)$ versus $I(X_i; X_k \mid X_j)$.

Second, a key property that should be satisfied by any reasonable redundancy measure is that adding information to some variable will not reduce the value of any PID atoms that contains it:

Desideratum 3 (Information monotonicity) *Given $X \in S$ and another disjoint set of random variables $K \cap S = \emptyset$, let: (i) X' be the supervariable that combines X and K; (ii) S' be the source set obtained by replacing X with X'; (iii) α' be the lattice element in the bivariate redundancy lattice obtained from α by replacing X with X'. We require $\Pi_R^{S'}(\alpha') \geq \Pi_R^{S}(\alpha)$ for all α that contains a subset that contains X.*

Since X' contains all the information that X has (and potentially more from K), replacing X with X' in any lattice element should preserve or increase the corresponding PID component. This property ensures that when we replace X with the augmented variable X', any synergistic, redundant, or unique information will maintain access to the original information in X while potentially gaining additional information from K.

We now present our result that identifies co-parents using synergy. This result complements Theorem 2, which linked positive unique information to parents and children.

Theorem 3 (Co-parent identification via synergy) *Under Assumptions 1 and 2, for $X_i, X_k \in V$, X_k is a co-parent of X_i if and only if:*

(i) X_k has zero unique information about X_i: $\Pi_R^{V'}(\{\{X_k\}\}; X_i) = 0$;

(ii) There exists $X_j \in V$ with $\Pi_R^{V'}(\{\{X_j\}; X_i\}) > 0$ such that X_k and X_j have positive synergy about X_i: $\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) > 0$, where X_j is the joint child of X_k and X_i.

Note that the zero unique information is relative to the set of all variables V, while positive synergy is only relative to $\{X_j, X_k\}$.

Proof *See Section B.3 of the appendix.* ■

Identifying the co-parent PID signature enables us to confirm whether a variable X_j is a child of the target variable X_i. Conversely, to verify that X_j is a parent of X_i, we reverse the roles by setting X_j as the target and attempting to identify a co-parent of X_j with X_i serving as their joint child.

Remark: While the identification of co-parents requires finding pairs of variables corresponding to the collider and the co-parent, we can narrow down the set of possible colliders through Theorem 2, which drastically improves efficiency if the graph is sparse. In particular, we only have to check for co-parent structures between two variables X_i and X_j when their direct neighborhoods $N(X_i)$ and $N(X_j)$ intersect since the collider has unique information to both X_i and X_j.

By combining Theorems 2 and 3, which allow direct causal neighbor and co-parent identification, we provide a systematic causal discovery procedure for Bayesian networks based on PID.

Procedure 1 (PID-based causal discovery for Bayesian networks) *Under Assumptions 1 and 2 and Desiderata 1–3 the causal structure of a Bayesian network can be identified through the following three-step procedure:*

1. ***Direct Causal Neighbor Identification.*** *For each $X_i \in V$, set X_i as the target for PID analysis and identify all variables with positive unique information about X_i (Theorem 2).*

2. ***Directional Confirmation via Co-parent Detection.*** *For each $X_i \in V$ and each of its direct causal neighbors $X_j \in N(X_i)$, attempt to identify a co-parent variable such that unique information about X_i is zero and synergistic information with X_j is positive (Theorem 3). The identification of such a co-parent X_k confirms that $X_j \in ch(X_i)$ (equivalently, $X_i \in pa(X_j)$), thereby establishing the causal direction $X_i \to X_j$.*

3. ***Remaining Direction Resolution.*** *After co-parent relationships have been used to direct edges, apply remaining heuristics (e.g., acyclicity constraints) to complete any indeterminate causal relationships that could not be resolved through co-parent identification.*

3.2. PID signatures in Bayesian hypergraphs

In this section, we extend our results linking causality and PID to Bayesian hypergraphs. Crucially, we identify the PID signatures of variables within individual hyperedges that capture both the directed relationships between tails and heads, as well as the symmetric relationships among co-heads and co-tails. Here we only consider interactions that can be faithfully represented by a Bayesian hypergraph, such that all statistical independencies correspond exactly to the conditional independence properties dictated by the hypergraph factorization.

Remark: As Bayesian hypergraphs generalizes Bayesian networks, being faithful to a Bayesian network implies faithfulness to a Bayesian hypergraph, but the reverse does not hold. Under standard faithfulness assumptions, neither Bayesian networks nor Bayesian hypergraphs are fully designed to represent strict logical XORs (Marx et al., 2021).

Remark: To illustrate interactions that standard Bayesian networks fail to faithfully represent but Bayesian hypergraphs can, consider the following example of a partial collider. Suppose X_1, X_2, X_3, and X_4 are independent coin flips, and $X_5 = (Y_1, Y_2)$ where $Y_1 = X_1$ OR X_2 and $Y_2 = X_3$ OR X_4. Then, conditioning on X_5 activates both colliders, rendering X_1 dependent on X_2, and X_3 dependent on X_4. Crucially, the pair $\{X_1, X_2\}$ remains conditionally independent of $\{X_3, X_4\}$ given X_5. A standard Bayesian network cannot faithfully capture this structure; conditioning on the common child X_5 would incorrectly induce dependence among all parent pairs, resulting in the structure shown in Fig. 1a. A Bayesian hypergraph, however, correctly encodes these as two disjoint causal mechanisms meeting at X_5, preserving the conditional independence between the distinct groups through its structured factorization, as illustrated in Fig. 1b.

Similar to Bayesian networks, we also require the assumption of collider amplification for hypergraphs, with the notion of co-parents replaced by co-tails.

Assumption 3 (Hypergraph collider amplification) *Given a directed hyperedge ε in Bayesian hypergraph $\mathcal{H}$ such that $X_i, X_k \in T(\varepsilon)$ and $X_j \in H(\varepsilon)$, the following holds:*

$$I(X_i; X_k \mid X_j) > I(X_i; X_k)\,.$$

The hypergraph collider amplification property states that the collider effect only activates for variables that are co-tails within the same directed hyperedge. This restriction is necessary due to the factorization structure of directed hypergraphs. For instance, X_1 and X_3 are not co-tails of any hyperedge in the Bayesian hypergraph in Figure 1b; hence conditioning on X_5 does not create conditional dependence between X_1 and X_3, as they belong to separate potential functions ψ_{125} and ψ_{345}. This is in contrast to the Bayesian network in Figure 1a, where conditioning on X_5 creates conditional dependence between any pair of parent variables. Therefore Assumption 2 requires variables to be joint causes (co-tails) of the same effect (head) within a single directed hyperedge.

Next, we introduce our theorem on the relationship between unique information and causal neighbors in hypergraphs. This is the Bayesian hypergraph parallel of Theorem 2.

Theorem 4 (Hypergraph unique information characterization) *Under Assumption 1, for any variable X_i and X_j in a Bayesian hypergraph, the unique information of X_j about X_i is positive iff X_j is a parent, child, or co-head of X_i.*

Proof *See Section B.4 of the appendix.* ■

In Bayesian hypergraphs, the variables with positive unique information correspond not only to parents and children, but also to co-heads because the factorization of Bayesian hypergraphs contains both asymmetric causal relationships (parent/child) and symmetric relationships (co-heads).

We now present the Bayesian hypergraph analogue of Theorem 3. For hypergraphs, the collider effect only activates among co-tails, hence we replace co-parents in Theorem 3 with co-tails.

Theorem 5 (Hypergraph co-tail identification via synergy) *Under Assumptions 1 and 3, X_k is a co-tail of variable X_i if and only if:*

(i) X_k has zero unique information about X_i: $\Pi_R^{V'}(\{\{X_k\}\}; X_i) = 0$;

(ii) There exists $X_j \in V$ with $\Pi_R^{V'}(\{\{X_j\}\}; X_i) > 0$ such that X_k and X_j have positive synergy about X_i: $\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) > 0$.

Proof *See Section B.5 of the appendix.* ■

By combining our results on Bayesian hypergraph unique information and co-tail identification (Theorems 4 and 5), we can pin down the PID signature of a hyperedge as follows:

Theorem 6 (PID signatures of hyperedges) *Under Assumptions 1 and 3, for any directed hyperedge ε with head $H(\varepsilon)$ and tail $T(\varepsilon)$, the following PID signatures hold:*

(i) For every $X_i \in T(\varepsilon)$ and $X_j \in H(\varepsilon)$, any co-tail $X_k \in T(\varepsilon) \setminus \{X_i\}$ has positive synergy with X_j about X_i: $\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) > 0$.

(ii) For every $X_i \in T(\varepsilon)$ and $X_j \in H(\varepsilon)$, the head X_j has positive unique information about X_i: $\Pi_R^{V'}(\{\{X_j\}\}; X_i) > 0$.

(iii) Any pair of co-heads $X_j, X_k \in H(\varepsilon)$ have positive unique information about each other: $\Pi_R^{V'}(\{\{X_j\}\}; X_k) > 0$ and $\Pi_R^{V'}(\{\{X_k\}\}; X_j) > 0$.

Proof *See Section B.6 of the appendix.* ■

While each hyperedge implies particular PID signatures, a particular PID signature may be compatible with multiple combinations of hyperedges. To resolve this ambiguity, we introduce the notion of maximal hyperedge.

Definition 7 (Maximal hyperedge) *A directed hyperedge $\varepsilon^* \in \mathcal{E}$ is called maximal if extending either its head $H(\varepsilon^*)$ or tail $T(\varepsilon^*)$ would cause it to violate the PID signature properties in Theorem 6. Formally, ε^* is maximal if there does not exist a hyperedge ε' with either: $T(\varepsilon^*) \subset T(\varepsilon')$ and $H(\varepsilon^*) = H(\varepsilon')$, or $T(\varepsilon^*) = T(\varepsilon')$ and $H(\varepsilon^*) \subset H(\varepsilon')$ such that ε' also satisfies the PID signature properties in Theorem 6.*

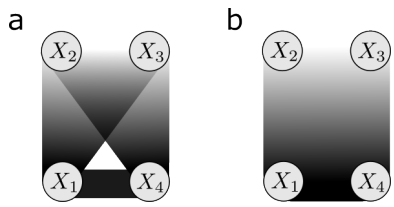

Figure 2: **Example of maximal hyperedge construction.** The Bayesian hypergraph with three hyperedges in (a) has the same conditional independence properties as the one in (b). Yet (b) provides a more parsimonious representation using the maximal hyperedge construction.

The maximal hyperedge construction provides a canonical representation that consolidates redundant hyperedge specifications. For example, consider the Bayesian hypergraph in Figure 2 with three hyperedges: ε_1 with $T(\varepsilon_1) = \{X_1, X_2\}$ and $H(\varepsilon_1) = \{X_3\}$; ε_2 with $T(\varepsilon_2) = \{X_1, X_2\}$ and $H(\varepsilon_2) = \{X_4\}$; ε_3 with $T(\varepsilon_3) = \emptyset$ and $H(\varepsilon_3) = \{X_3, X_4\}$. These can be absorbed into a single maximal hyperedge: ε^* with $T(\varepsilon^*) = \{X_1, X_2\}$ and $H(\varepsilon^*) = \{X_3, X_4\}$, which results in the same conditional independence properties while providing a more parsimonious representation of the hypergraph structure.

Based on Theorem 6 and Definition 7, we develop a systematic procedure for Bayesian hypergraph causal discovery that leverages information-theoretic patterns to identify canonical causal representations.

Procedure 2 (PID-based causal discovery for Bayesian hypergraph) *Under Assumptions 1 and 2 and Desiderata 1– 3, the following three-step procedure identifies the Bayesian hypergraph structure:*

1. ***Candidate Search****: Search for subsets $T \subseteq V$ (potential tails) and $H \subseteq V$ (potential heads) with $T \cap H = \emptyset$ that satisfy the hyperedge PID signature properties of Theorem 6.*

2. ***Maximal Extension****: For each candidate hyperedge (T, H) that satisfies the PID signature, expand both T and H by adding variables while maintaining the hyperedge signature properties, continuing until no further expansion is possible without violating the signature.*

3. ***Canonical Selection and Iterative Construction****: Select the maximal hyperedge (T^*, H^*) according to Definition 7. Continue this search process iteratively until no additional maximal hyperedges can be identified. The collection of maximal hyperedges specifies the causal structure of the underlying Bayesian hypergraph.*

4. Related Works

An important branch of causal discovery stems from algorithmic information theory (AIT) (Mian et al., 2021; Mameche et al., 2022; Janzing and Schölkopf, 2010; Kaltenpoth and Vreeken, 2019), which utilizes the Kolmogorov complexity, an inherently noncomputable metric, to identify edge directions via the algorithmic Markov conditions. In contrast, our framework is grounded in Shannon information theory, leveraging the fine-grained components of mutual information to identify the causal roles of nodes, a formulation that can be practically evaluated using existing estimators (Venkatesh et al., 2023; Pakman et al., 2021; Barrett, 2015). While a primary focus of AIT-based

methods is to pinpoint the true causal graph from among the DAGs within a Markov equivalence class, our PID-based method does not attempt to break Markov equivalence. Furthermore, PID's higher-order nature enables causal discovery in more expressive Bayesian hypergraphs, extending beyond the pairwise limitations of current AIT methods.

5. Discussion

Our theoretical results establish a direct correspondence between PID signatures and structural causal configurations. This provides a localist perspective that links higher-order information-theoretic measure to structural and causal roles, offering insights that extend beyond standard causal discovery procedures. In particular, we show that these correspondences extend naturally to Bayesian hypergraphs, yielding an information-theoretic characterization of higher-order causal structure.

Several challenges and limitations remain. Because PID ultimately encodes conditional mutual information, it cannot distinguish causal structures that are observationally Markov-equivalent. A natural direction for future work is to develop an AIT-formulated redundancy measure satisfying our desiderata so that our framework can be readily extended to uncover the true causal graph. Furthermore, although our bivariate reduction avoids the combinatorial growth of full multivariate PID, the practical use of our framework still depends on efficient estimation of unique and synergistic information. Existing PID measures often involve high-dimensional optimization or rely on restrictive parametric assumptions, limiting their scalability and applicability. Developing estimation strategies that enable reliable PID-based inference from finite samples is therefore an important direction for future work. Another natural extension is to broaden our structural correspondence beyond Bayesian networks and hypergraphs to more expressive causal frameworks, thereby expanding the scope of the localist perspective developed here.

Acknowledgments

RP was supported by the Deutsche Forschungsgemeinschaft (DFG, Project-ID 424778381, TRR 295).

References

Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Phys. Rev. E*, 91:052802, 2015.

Anthony J Bell. The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, volume 2003, 2003.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. ISSN 1099-4300. doi: 10.3390/e16042161.

Christian Bick, Elizabeth Gross, Heather A Harrington, and Michael T Schaub. What are higher-order networks? *SIAM Review*, 65(3):686–731, 2023.

Garrett Birkhoff. *Lattice theory*, volume 25. American Mathematical Soc., 1940.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738.

Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):2084, 2020.

James Enouen, Yujia Zheng, Ignavier Ng, Yan Liu, and Kun Zhang. Higher-order causal structure learning with additive models. *arXiv preprint arXiv:2511.03831*, 2025.

Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nature communications*, 10(1):1–9, 2019.

Christian Jäkel. A computation of the ninth dedekind number. *Journal of Computational Algebra*, 6:100006, 2023.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Mohammad Ali Javidian, Zhiyu Wang, Linyuan Lu, and Marco Valtorta. On a hypergraph probabilistic graphical model. *Annals of Mathematics and Artificial Intelligence*, 88(9):1003–1033, 2020.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

David Kaltenpoth and Jilles Vreeken. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2019.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Zhaolu Liu, Mauricio Barahona, and Robert L Peach. Information-theoretic measures on lattices for high-order interactions. *The International Conference on Artificial Intelligence and Statistics*, 28, 2025.

Joseph T Lizier, Nils Bertschinger, Jürgen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy*, 20(4):307, 2018.

Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Negin Holland, Tim D Fryer, John T O'Brien, James B Rowe, David K Menon, Daniel Bor, and Emmanuel A Stamatakis. A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, 25(6):771–782, 2022.

Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Judith Allanson, John Pickard, Robin L Carhart-Harris, Guy B Williams, Michael M Craig, Paola Finoia, Adrian M Owen, et al. A synergistic workspace for human consciousness revealed by integrated information decomposition. *Elife*, 12:RP88173, 2024.

Aobo Lyu, Andrew Clark, and Netanel Raviv. Multivariate partial information decomposition: Constructions, inconsistencies, and alternative measures. *arXiv preprint arXiv:2508.05530*, 2025.

Abdullah Makkeh, Marcel Graetz, Andreas C Schneider, David A Ehrlich, Viola Priesemann, and Michael Wibral. A general framework for interpretable neural learning based on local information-theoretic goal functions. *Proceedings of the National Academy of Sciences*, 122 (10):e2408125122, 2025.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Discovering invariant and changing mechanisms from data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1242–1252, 2022.

Álvaro Martínez-Sánchez, Gonzalo Arranz, and Adrián Lozano-Durán. Decomposing causality into its synergistic, unique, and redundant components. *Nature Communications*, 15(1):9296, 2024.

Alexander Marx, Arthur Gretton, and Joris M Mooij. A weaker faithfulness assumption based on triple interactions. In *Uncertainty in artificial intelligence*, pages 451–460. PMLR, 2021.

Pedro AM Mediano, Fernando E Rosas, Andrea I Luppi, Robin L Carhart-Harris, Daniel Bor, Anil K Seth, and Adam B Barrett. Toward a unified taxonomy of information dynamics via integrated information decomposition. *Proceedings of the National Academy of Sciences*, 122 (39):e2423297122, 2025.

Osman A Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8975–8982, 2021.

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. *Advances in neural information processing systems*, 34:20295–20307, 2021.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jürgen Jost. Reconsidering unique information: Towards a multivariate information decomposition. In *2014 IEEE International Symposium on Information Theory*, pages 2232–2236. IEEE, 2014.

Fernando E Rosas, Pedro AM Mediano, Michael Gastpar, and Henrik J Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3):032305, 2019.

Gian-Carlo Rota. On the foundations of combinatorial theory: I. theory of möbius functions. In *Classic Papers in Combinatorics*, pages 332–360. Springer, 1964.

Elad Schneidman, William Bialek, and Michael J Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Leo Torres, Ann S Blevins, Danielle Bassett, and Tina Eliassi-Rad. The why, how, and when of representations for complex systems. *SIAM Review*, 63(3):435–485, 2021.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Thomas F Varley, Maria Pope, Maria Grazia Puxeddu, Joshua Faskowitz, and Olaf Sporns. Partial entropy decomposition reveals higher-order structures in human brain activity. *arXiv preprint arXiv:2301.05307*, 2023.

Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina Ramirez, Greggory Heller, Severine Durand, Shawn Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. *Advances in Neural Information Processing Systems*, 36:74602–74635, 2023.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Appendix A. Additional Information

A.1. Redundancy Lattices

.

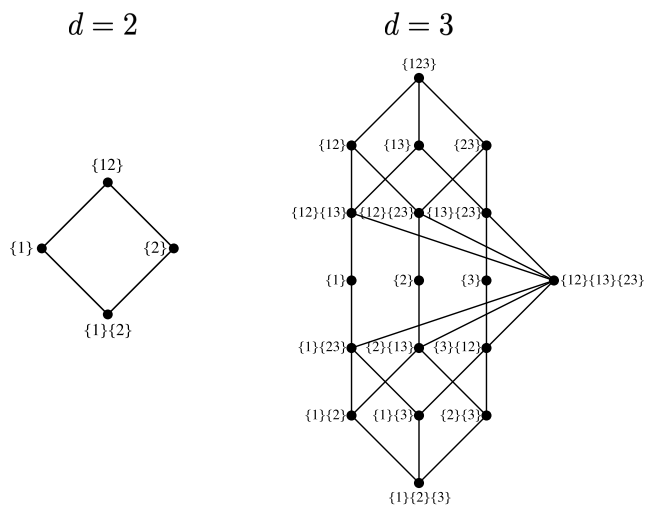

Figure 3: **Hasse diagrams of the redundancy lattice for $d = 2$ (left) and $d = 3$ (right) source variables.** Each node corresponds to an element of the redundancy lattice, and edges indicate the partial order defined by information containment. For $d = 2$, the lattice contains 4 elements, yielding the familiar bivariate PID structure. For $d = 3$, the lattice contains 18 elements, illustrating the rapid growth in lattice size as the number of sources increases. In general, the number of lattice elements grows according to the Dedekind numbers, leading to a combinatorial explosion for $d > 3$, which motivates the use of bivariate reductions via supervariables in this work.

A.2. Redundancy reuse

Suppose that we have three variables X_1, X_2, X_3. To identify the causal skeleton of their relationship we try to estimate the unique information, we can do this by utilizing the relationships

$$I(X_1; X_3) = \mathrm{Uniq}(X_1; X_3) + \mathrm{Red}(X_1, X_2; X_3)$$
$$I(X_2; X_3) = \mathrm{Uniq}(X_2; X_3) + \mathrm{Red}(X_1, X_2; X_3)$$

which gives

$$\mathrm{Uniq}(X_1; X_3) = I(X_1; X_3) - \mathrm{Red}(X_1, X_2; X_3)$$
$$\mathrm{Uniq}(X_2; X_3) = I(X_2; X_3) - \mathrm{Red}(X_1, X_2; X_3)$$

In other words, we only have to calculate the redundant information once and we can use that to estimate multiple unique information terms at once by subtracting it from mutual information.

Appendix B. Proofs

B.1. Proof of Theorem 1

Proof By using chain rule of conditional mutual information and decomposing each mutual information term into its PID atoms, we have:

$$I(X_j; X_i \mid C) = \sum_{\alpha \in \mathcal{A}} \Pi_{\mathbf{R}}(\alpha; X_i) \tag{2}$$

where $\mathcal{A} = \{\alpha \in \mathcal{L} : \alpha \preceq \{\{X_j, C\}\} \text{ and } \alpha \not\preceq \{\{C\}\}\}$.

By assumption, $I(X_i; X_j \mid C) = 0$, so:

$$\sum_{\alpha \in \mathcal{A}} \Pi_{\mathbf{R}}(\alpha; X_i) = 0 \tag{3}$$

The set $\mathcal{A}$ contains lattice elements that involve X_j but are not entirely determined by C. Crucially, this set includes $\{\{X_j\}\}$ since:

- $\{\{X_j\}\} \preceq \{\{X_j, C\}\}$ (as $\{X_j\} \subseteq \{X_j, C\}$)
- $\{\{X_j\}\} \not\preceq \{\{C\}\}$ (since $X_j \notin C$, we have $\{X_j\} \not\subseteq C$)

Therefore, $\{\{X_j\}\} \in \mathcal{A}$, and the sum includes $\Pi_{\mathbf{R}}(\{\{X_j\}\}; X_i)$.

Since all PID components are non-negative by our constraint on redundancy measures:

$$\Pi_{\mathbf{R}}(\alpha; X_i) \geq 0 \quad \forall \alpha \in \mathcal{L} \tag{4}$$

and the sum of non-negative terms equals zero, each individual term must be zero:

$$\Pi_{\mathbf{R}}(\alpha; X_i) = 0 \quad \forall \alpha \in \mathcal{A} \tag{5}$$

Therefore, X_j has zero unique information about target X_i. ■

B.2. Proof of Theorem 2

Proof **Direction 1** ($\Rightarrow$): If X_j is not a parent or child of X_i, then $\Pi_R^{V'}(\{\{X_j\}\}; X_i) = 0$.

We consider two cases:

Case 1: $X_j \notin MB(X_i)$ (not in the Markov blanket of X_i).

By the Markov property, variables outside the Markov blanket are conditionally independent of X_i given the Markov blanket:

$$X_j \perp\!\!\!\perp X_i \mid MB(X_i) \tag{6}$$

This gives $I(X_j; X_i \mid MB(X_i)) = 0$. In the bivariate PID decomposition with source variables $\{X_j, MB(X_i)\}$ (treating the Markov blanket as a single variable), this implies:

$$\Pi_R^{\{X_j, MB(X_i)\}}(\{\{X_j\}\}; X_i) = 0 \tag{7}$$

Since $MB(X_i) \subseteq V' = V \setminus \{X_i, X_j\}$, by the unique information monotonicity property (Definition 2):

$$Pi_R^{V'}(\{\{X_j\}\}; X_i) \leq \Pi_R^{\{X_j, MB(X_i)\}}(\{\{X_j\}\}; X_i) = 0 \tag{8}$$

Therefore, $\Pi_R^{V'}(\{\{X_j\}\}; X_i) = 0$.

Case 2: $X_j \in MB(X_i)$ but $X_j \notin \text{pa}(X_i) \cup \text{ch}(X_i)$.

Then X_j must be a co-parent of some child of X_i. Let $X_k \in \text{ch}(X_i)$ be a common child of X_i and X_j, so we have the V-structure $X_i \to X_k \leftarrow X_j$.

We can construct a conditioning set C as follows:

- Do not include X_k in C (to avoid activating the collider)
- For any other path between X_i and X_j, include appropriate variables in C to block those paths

By d-separation, with this choice of C:

$$X_j \perp\!\!\!\perp X_i \mid C \tag{9}$$

Since $C \subseteq V \setminus \{X_i, X_j\} = V'$, by the unique information monotonicity property:

$$\Pi_R^{V'}(\{\{X_j\}\}; X_i) \leq \Pi_R^{\{X_j, C\}}(\{\{X_j\}\}; X_i) = 0 \tag{10}$$

Therefore, $\Pi_R^{V'}(\{\{X_j\}\}; X_i) = 0$.

Direction 2 ($\Leftarrow$): If X_j is a parent or child of X_i, then $\Pi_{\mathbf{R}}(\{\{X_j\}\}; X_i) > 0$.

If X_j is a parent or child of X_i, then there is a direct causal edge between them. By faithfulness, this direct causal connection implies $I(X_j; X_i \mid C) > 0$ for all possible conditioning sets $C \subseteq S \setminus \{X_j\}$. By the persistent relevance property, this implies $\Pi_{\mathbf{R}}(\{\{X_j\}\}; X_i) > 0$. ■

B.3. Proof of Theorem 3

Proof Direction 1 ($\Leftarrow$): If X_k is a co-parent of X_i, then the conditions hold.

Zero unique information: Since X_k is not a direct parent or child of X_i (but rather connected through the collider X_j), by Theorem 2, we have $\Pi_R^{V'}(\{\{X_k\}\}; X_i) = 0$.

Positive synergistic information: Consider the V-structure $X_k \to X_j \leftarrow X_i$, since X_j is a child of X_i, we have $\Pi_R(\{\{X_j\}\}) > 0$. By the collider amplification property:

$$I(X_k; X_i \mid X_j) > I(X_k; X_i) \tag{11}$$

From the two-source PID decomposition:

$$I(X_k; X_i \mid X_j) = \Pi_R^{\{X_j, X_k\}}(\{\{X_k\}\}; X_i) + \Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) \tag{12}$$

$$I(X_k; X_i) = \Pi_R^{\{X_j, X_k\}}(\{\{X_k\}, \{X_j\}\}; X_i) + \Pi_R^{\{X_j, X_k\}}(\{\{X_k\}\}; X_i) \tag{13}$$

Substituting into the collider amplification inequality:

$$\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) - \Pi_R^{\{X_j, X_k\}}(\{\{X_k\}, \{X_j\}\}; X_i) > 0 \tag{14}$$

Since redundant information is non-negative, $\Pi_R^{\{X_j, X_k\}}(\{\{X_k\}, \{X_j\}\}; X_i) \geq 0$, we conclude:

$$\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) > 0 \tag{15}$$

Direction 2 ($\Rightarrow$): If the conditions hold, then X_k is a co-parent of X_i.

We prove this by contradiction. Assume that variable X_k has zero unique information about X_i but is not a co-parent of X_i. We will show that there cannot exist any variable X_j with positive unique information such that X_k and X_j have positive synergistic information about X_i.

Since X_k has zero unique information about X_i, by Theorem 2, X_k is not a parent or child of X_i. Combined with our assumption that X_k is not a co-parent, this implies that X_k is outside the Markov blanket of X_i: $X_k \notin MB(X_i)$.

On the other hand, any variable X_j with positive unique information about X_i must be a parent or child of X_i by Theorem 2, which means $X_j \in MB(X_i)$.

By the Markov property, since $X_k \notin MB(X_i)$:

$$I(X_k; X_i \mid MB(X_i)) = 0 \tag{16}$$

This implies that in the bivariate PID decomposition with source variables $\{X_k, MB(X_i)\}$ (treating the Markov blanket as a single variable):

$$\Pi_R^{\{X_k, MB(X_i)\}}(\{\{X_k, MB(X_i)\}\}; X_i) = 0 \tag{17}$$

Since $X_j \in MB(X_i)$, by the information monotonicity property (Desideratum 3), we have:

$$\Pi_R^{\{X_k, MB(X_i)\}}(\{\{X_k, MB(X_i)\}\}; X_i) \geq \Pi_R^{\{X_k, X_j\}}(\{\{X_k, X_j\}\}; X_i) \tag{18}$$

Combining Eq. (17) and (18) above we have:

$$\Pi_R^{\{X_k, X_j\}}(\{\{X_k, X_j\}\}; X_i) = 0 \tag{19}$$

Therefore, X_k and X_j have zero bivariate synergistic information, contradicting the assumption that there exists positive synergistic information. Hence, X_k must be a co-parent of X_i. ■

B.4. Proof of Theorem 4

Proof Direction 1 ($\Leftarrow$): If X_j is a parent, child, or co-head of X_i, then $\Pi_R^{V'}(\{\{X_j\}\}; X_i) > 0$.

Parents and children: By the same argument as in 2, direct causal connections ensure $I(X_j; X_i \mid C) > 0$ for all conditioning sets $C \subseteq V \setminus \{X_j\}$. By persistent relevance, this implies positive unique information.

Co-heads: If $X_j \sim_H X_i$, they appear together in the head $H(h)$ of some directed hyperedge h. By the hypergraph factorization structure, this creates a direct statistical dependency through the shared potential function ψ_h that cannot be eliminated by conditioning on other variables, ensuring $I(X_j; X_i \mid C) > 0$ for all $C \subseteq V \setminus \{X_j\}$.

Direction 2 ($\Rightarrow$): If X_j is not a parent, child, or co-head of X_i, then $\Pi_R^{V'}(\{\{X_j\}\}; X_i) = 0$.

Outside Markov blanket: If $X_j \notin MB(X_i)$, then by the Markov property: $X_j \perp\!\!\!\perp X_i \mid MB(X_i)$, giving $I(X_j; X_i \mid MB(X_i)) = 0$. By the argument in Theorem 2, this implies zero unique information.

Co-tails: If X_j is a co-tail of X_i (i.e., $X_j \sim_T X_i$ but not co-head), then there exists a conditioning set C that excludes their common head variables but includes appropriate blocking variables for other paths. By the hypergraph Markov properties, this renders $X_j \perp\!\!\!\perp X_i \mid C$, yielding $I(X_j; X_i \mid C) = 0$ and thus zero unique information. ■

B.5. Proof of Theorem 5

Proof **Direction 1** ($\Leftarrow$): If X_k is a co-tail of X_i, then the conditions hold.

Zero multivariate unique information: Since X_k is not a parent, child, or co-head of X_i (but rather a co-tail), by the previous theorem, we have $\Pi_R^{V'}(\{\{X_k\}\}; X_i) = 0$.

Positive synergistic information: Consider the directed hyperedge h with $X_k, X_i \in T(h)$ and $X_j \in H(h)$. By the hypergraph collider amplification property:

$$I(X_k; X_i \mid X_j) > I(X_k; X_i) \tag{20}$$

Using the same PID decomposition argument as in Theorem 3, this inequality implies:

$$\Pi_R^{\{X_j, X_k\}}(\{\{X_j, X_k\}\}; X_i) > 0 \tag{21}$$

Since X_j is in the head of the hyperedge containing X_i in its tail, X_j is a child of X_i, ensuring $\Pi_R^{V'}(\{\{X_j\}\}; X_i) > 0$ by the previous theorem.

Direction 2 ($\Rightarrow$): If the conditions hold, then X_k is a co-tail of X_i.

We proceed by contradiction similar to Theorem 3. The key distinction is that the Markov blanket structure in hypergraphs differs from ordinary graphs. If X_k is not a co-tail of X_i, then even if they share common head variables, they belong to separate hyperedges in the factorization. As a result, conditioning on the head part of the hyperedge induces conditional independence between X_i and X_k due to the global markov property.This conditional independence, combined with the information monotonicity property, leads to zero bivariate synergistic information, contradicting the assumed positive synergy. Therefore, X_k must be a co-tail of X_i. ■

B.6. Proof of Theorem 6

Proof The theorem follows directly from our previous results:

Property 1: By Theorem 5, co-tail variables X_k and X_i exhibit positive bivariate synergistic information with their common head X_j.

Property 2: By Theorem 4, head variables X_j are children of tail variables X_i, ensuring positive unique information.

Property 3: By Theorem 4, co-head variables are directly connected through their shared hyperedge membership, guaranteeing mutual unique information. ■