Quantifying high-order interdependencies via multivariate extensions of the mutual information

Fernando E. Rosas,[1,2,*] Pedro A. M. Mediano,[3] Michael Gastpar,[4] and Henrik J. Jensen[1,5]

[1]*Centre of Complexity Science and Department of Mathematics, Imperial College London, London SW7 2AZ, England, United Kingdom*

[2]*Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, England, United Kingdom*

[3]*Department of Computing, Imperial College London, London SW7 2AZ, England, United Kingdom*

[4]*School of Computer and Communication Sciences, École polytechnique fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland*

[5]*Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan*

(Received 31 December 2018; published 13 September 2019)

This paper introduces a model-agnostic approach to study statistical synergy, a form of emergence in which patterns at large scales are not traceable from lower scales. Our framework leverages various multivariate extensions of Shannon's mutual information, and introduces the *O-information* as a metric that is capable of characterizing synergy- and redundancy-dominated systems. The O-information is a symmetric quantity, and can assess intrinsic properties of a system without dividing its parts into "predictors" and "targets." We develop key analytical properties of the O-information, and study how it relates to other metrics of high-order interactions from the statistical mechanics and neuroscience literature. Finally, as a proof of concept, we present an exploration on the relevance of statistical synergy in Baroque music scores.

DOI: 10.1103/PhysRevE.100.032305

I. INTRODUCTION

A unique opportunity in the era of "big data" is to make use of the abundant available data to deepen our understanding of the high-order interdependencies that are at the core of complex systems. Plentiful data are nowadays available about, e.g., the orchestrated activity of multiple brain areas, the relationship between various econometric indices, and the interactions between different genes. What allows these systems to be more than the sum of their parts is not in the nature of the parts, but in the structure of their interdependencies [1]. However, quantifying the "synergy" of different complex systems is challenging, especially in scenarios where the number of parts is large but far below the thermodynamic limit.

The relevance of synergistic relationships and other high-order interactions has been thoughtfully demonstrated in the literature of theoretical neuroscience. For example, studies on neural coding have shown that neurons can carry redundant, complementary, or synergistic information—the latter corresponding to neurons that are uninformative individually but informative when considered together [2,3]. Also, studies on retina cells suggest that high-order Hamiltonians are necessary for representing neurons firing in response to natural images, while pairwise interactions suffice for neurons responding to less structured stimuli [4]. Lastly, neuroimaging analyses have pointed out the compatibility of local differentiation and global integration of different brain areas, and suggested this to be a key capability for enabling high cognitive functions [5,6]. Various metrics have been proposed to capture these high-order features in data, including the *redundancy-synergy index* [7,8] (and corresponding extensions [9–11]), *connected information* [12], *neural complexity* [13], and *integrated information* [14,15]. While being capable of capturing features of biological relevance, most of these metrics have *ad hoc* definitions motivated by specific research agendas, and have few theoretical guarantees [16].

A promising approach for addressing high-order interdependencies is *partial information decomposition* (PID), which distinguishes different "types" of information that multiple predictors convey about a target variable [17–19]. In this framework, *statistical synergies* are structures (or relationships) that exist in the whole but cannot be seen in the parts, this being rooted in the elementary fact that variables can be pairwise independent while being globally correlated. Unfortunately, the adoption of PID has been hindered by the lack of agreement on how to compute the components of the decomposition, despite numerous recent efforts [20–23]. Furthermore, although practical applications of PID have been reported [24–26], the applicability of the framework is restricted by the rapid growth of the number of terms for large systems.

The crux of multivariate interdependencies is that information-theoretic descriptions of such phenomena are not straightforward, as extensions of Shannon's classical results to general multivariate settings have proven elusive [27]. The most well-established multivariate extensions of Shannon's mutual information are the *total correlation* [28] and the *dual total correlation* [29], which provide suitable metrics of overall correlation strength. Their values, however, differ in ways that are hard to understand [30], even gaining the adjective of "enigmatic" among scholars [31,32]. Other popular extension of the mutual information is the *interaction information* [33], which is a signed measure obtained by applying the inclusion-exclusion principle to the Shannon entropy [34,35]. Although this metric provides insightful results when applied to three variables, it is not easily interpretable when applied to larger groups [17].

This paper proposes to study multivariate interdependency via two dual perspectives: as *shared randomness* and as *collective constraints* [36]. This setup leads to the *O-information*

*f.rosas@imperial.ac.uk

(shorthand for "information about organizational structure"), which—following Occam's razor—points out which of these perspectives provides a more parsimonious description of the system. The O-information is found to coincide with the interaction information for the case of three variables, while providing a more meaningful extension for larger system sizes.

We show how the O-information captures the dominant characteristic of multivariate interdependency, distinguishing redundancy-dominated scenarios where three or more variables have copies of the same information and synergy-dominated systems characterized by high-order patterns that cannot be traced from low-order marginals. In contrast with existing quantities that require a division between predictors and target variables, the O-information is—to the best of our knowledge—the first symmetric quantity that can give account of intrinsic statistical synergy in systems of more than three parts. Moreover, as the computational complexity of the O-information scales gracefully with system size, our framework provides a scalable approach for applying PID principles to large systems, suitable for practical data analysis.

In the following, Sec. II introduces the notions of shared randomness and collective constraints, and Secs. III and IV present the O-information and its fundamental properties. Section V compares the O-information with other metrics of high-order effects, and Sec. VI presents a case study on music scores. Finally, Sec. VII summarizes our main conclusions.

II. FUNDAMENTALS

This section introduces two fundamental perspectives from which one can develop an information-theoretic description of a system, and explains how they enable novel perspectives to study interdependency.

A. Entropy and negentropy

For every outside there is an inside and for every inside there is an outside. And although they are different, they always go together.

Alan Watts, *Myth of Myself*

Following the Bayesian interpretation of information theory, we define the *information contained in a system* as the average amount of data that an observer would gain after determining its configuration—i.e., after measuring it [37]. If each possible configuration is to be represented by a distinct sequence of bits, source coding theory (see Chap. 5 of Ref. [38]) shows that an optimal (i.e., shortest) labeling depends on prior information available before the measurement. Information, hence, refers to how the state of knowledge of the observer changes after the system is measured, quantifying the amount of bits that are revealed through this process [39].

For concreteness, let us consider an observer measuring a system composed by n discrete variables, $\boldsymbol{X}^n = (X_1, \ldots, X_n)$. If the observer only knows that each variable X_j can take values over a finite alphabet $\mathcal{X}_j$ of cardinality $|\mathcal{X}_j|$, the amount of information needed to specify the state of X_j is $\log|\mathcal{X}_j|$ (logarithms are calculated using base 2 unless specified otherwise). In contrast, if the observer knows that the system's behavior follows a probability distribution $p_{\boldsymbol{X}^n}$, then the average amount of information in the system reduces to the *entropy* $H(\boldsymbol{X}^n) := -\sum_{\boldsymbol{x}^n} p_{\boldsymbol{X}^n}(\boldsymbol{x}^n)\log p_{\boldsymbol{X}^n}(\boldsymbol{x}^n)$ [37]. The difference

$$\mathcal{N}(\boldsymbol{X}^n) := \sum_{j=1}^{n}\log|\mathcal{X}_j| - H(\boldsymbol{X}^n) \tag{1}$$

is known as *negentropy* [40], and corresponds to the information about the system that is disclosed by the knowledge of the statistics, before any measurement takes place.

Probability distributions are, from this perspective, a compendium of soft and hard constraints that reduce the effective phase space that the system can explore (hard constraints completely forbid some configurations; soft constraints make them improbable). Consequently, a given distribution divides the phase space in an admissible region quantified by the entropy, and an inadmissible region quantified by the negentropy [41]. Each part describes the system's structure from a different point of view: the entropy refers to what the system can do, while the negentropy refers to what it cannot do.

B. The two faces of interdependency

1. Collective constraints

In the same way as $\mathcal{N}(\boldsymbol{X}^n)$ quantifies the strength of the overall constraints that rule the system, the constraints that affect individual variables are captured by the *marginal negentropies* $\mathcal{N}(X_j) := \log|\mathcal{X}_j| - H(X_j)$. Intuitively, the constraints that affect the whole system are richer than individual constraints, as the latter do not take into account collective effects. Their difference,

$$\begin{aligned}\mathrm{TC}(\boldsymbol{X}^n) :&= \mathcal{N}(\boldsymbol{X}^n) - \sum_{j=1}^{n}\mathcal{N}(X_j)\\ &= \sum_{j=1}^{n}H(X_j) - H(\boldsymbol{X}^n),\end{aligned} \tag{2}$$

quantifies the strength of the "collective constraints." This quantity is known as *total correlation* [28] (or *multi-information* [42]). By rewriting this relationship as $\mathcal{N}(\boldsymbol{X}^n) = \sum_j \mathcal{N}(X_j) + \mathrm{TC}(\boldsymbol{X}^n)$ one finds that the constraints prescribed by the distribution are of two types: constraints confined to individual variables, and collective constraints that restrict groups of two or more variables.

Example 1. Consider X_1 and X_2 to be binary random variables with $p_{X_1,X_2}(0,1) = p_{X_1,X_2}(1,0) = 1/2$. This distribution divides the total information (two bits) into $H(X_1, X_2) = 1$ and $\mathcal{N}(X_1, X_2) = 1$. Moreover, $\mathcal{N}(X_1) = \mathcal{N}(X_2) = 0$ and therefore $\mathrm{TC}(X_1, X_2) = \mathcal{N}(X_1, X_2) = 1$, confirming that the constraints act on both X_1 and X_2.

As a contrast, consider Y_1 and Y_2 binary random variables with distribution $p_{Y_1,Y_2}(0,0) = p_{Y_1,Y_2}(1,0) = 1/2$. In this case $\mathcal{N}(Y_1) = 0$ while $\mathcal{N}(Y_2) = \mathcal{N}(Y_1, Y_2) = 1$, showing that the only constraint in this system acts solely over Y_2. Accordingly, for this case $\mathrm{TC}(Y_1, Y_2) = 0$.

2. Shared randomness

As we did for $\mathcal{N}(\boldsymbol{X}^n)$, let us decompose $H(\boldsymbol{X}^n)$ in individual and collective components. To do this, we introduce the quantity $R_j = H(X_j|\boldsymbol{X}^n_{-j})$ as a metric of how independent X_j is from the rest of the system $\boldsymbol{X}^n_{-j} =$

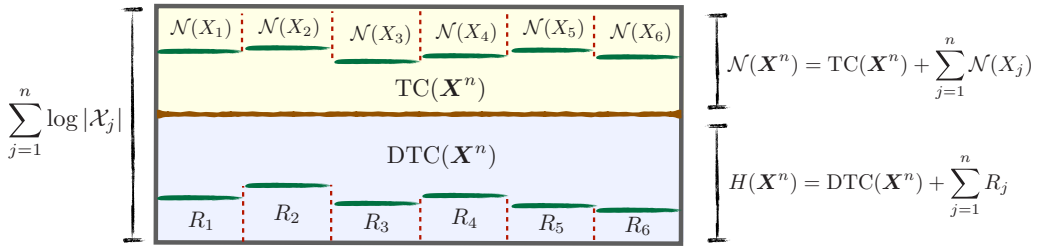

FIG. 1. The total information that can be stored in the system $\boldsymbol{X}^n$ ($\sum_{j=1}^{n} \log|\mathcal{X}_j|$) is decomposed by a given state of knowledge (i.e., a probability distribution) into two parts: what is determined by the constraints [the negentropy, $\mathcal{N}(\boldsymbol{X}^n)$] and what is not instantiated until an actual measurement takes place [the entropy, $H(\boldsymbol{X}^n)$]. Both terms can be further decomposed into their individual and collective components, yielding different perspectives on interdependency seen as either collective constraints [measured by the total correlation $\mathrm{TC}(\boldsymbol{X}^n)$] or shared randomness [corresponding to the dual total correlation $\mathrm{DTC}(\boldsymbol{X}^n)$].

$(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$. According to distributed source coding theory (see Chap. 10.5 of Ref. [27]), R_j corresponds to the data contained in X_j that cannot be extracted from measurements of other variables [43]. The quantity $\sum_{j=1}^{n} R_j$ is known as the *residual entropy* [44] (originally introduced under the name of *erasure entropy* [45,46]), and quantifies the total information that can only be accessed by measuring a specific variable, i.e., the amount of "nonshared randomness." Accordingly, the difference

$$\mathrm{DTC}(\boldsymbol{X}^n) := H(\boldsymbol{X}^n) - \sum_{j=1}^{n} R_j \tag{3}$$

is known as *dual total correlation* [29] (being also known as *binding information* [32,44] and *excess entropy* [47]), and refers to the part of the joint entropy that is shared by two or more variables—equivalently, information that can be obtained by measuring more than one specific variable. As the entropy corresponds to the randomness within the system, the dual total correlation quantifies the "shared randomness" that exists among the variables.

Example 2. Let us consider X_1, X_2 and Y_1, Y_2 from Example 1. For the former system one finds that $R_1 = R_2 = 0$ and hence $\mathrm{DTC}(X_1, X_2) = H(X_1, X_2) = 1$, which means that the randomness within the system can be retrieved from measuring either X_1 or X_2. In contrast, when considering Y_1, Y_2 one finds that $R_2 = 0$ and $R_1 = H(Y_1, Y_2) = 1$, and hence $\mathrm{DTC}(Y_1, Y_2) = 0$. This implies that the randomness of the system can be retrieved by measuring only Y_1.

Wrapping up, one can rewrite Eq. (1) using Eqs. (2) and (3) and express the total information encoded in the system described by $\boldsymbol{X}^n$ in terms of constraints and randomness:

$$\begin{aligned}\sum_{j=1}^{n} \log|\mathcal{X}_j| &= \mathcal{N}(\boldsymbol{X}^n) + H(\boldsymbol{X}^n) \\ &= \underbrace{\left[\mathrm{TC}(\boldsymbol{X}^n) + \sum_{j=1}^{n} \mathcal{N}(X_j)\right]}_{\text{Collective and individual constraints}} + \underbrace{\left[\mathrm{DTC}(\boldsymbol{X}^n) + \sum_{j=1}^{n} R_j\right]}_{\text{Shared and private randomness}}.\end{aligned}$$

This decomposition is illustrated in Fig. 1.

III. INTRODUCING THE O-INFORMATION

A. Definition and basic properties

The TC and DTC provide complementary metrics of interdependence strength. Following Occam's razor, one might ask which of these perspectives allows for a shorter (i.e., more parsimonious) description. This is answered by the following definition:

Definition 1. The O-information of the system described by the random vector $\boldsymbol{X}^n$ is defined as

$$\begin{aligned}\Omega(\boldsymbol{X}^n) &:= \mathrm{TC}(\boldsymbol{X}^n) - \mathrm{DTC}(\boldsymbol{X}^n) \\ &= (n-2)H(\boldsymbol{X}^n) + \sum_{j=1}^{n} [H(X_j) - H(\boldsymbol{X}^n_{-j})].\end{aligned} \tag{4}$$

Intuitively, $\Omega(\boldsymbol{X}^n) > 0$ states that the interdependencies can be more efficiently explained as shared randomness, while $\Omega(\boldsymbol{X}^n) < 0$ implies that viewing them as collective constraints can be more convenient. Note that $\Omega(\boldsymbol{X}^n)$ was first introduced as "enigmatic information" in Ref. [31], although now that its properties have been revealed we choose to give it a more appropriate name.

To develop some insight about the O-information, let us compare it with the *interaction information* [48], which is a signed metric defined according to the inclusion-exclusion principle by

$$I(X_1; X_2; \ldots; X_n) := \sum_{\boldsymbol{\gamma} \subseteq \{1,\ldots,n\}} (-1)^{|\boldsymbol{\gamma}|+1} H(\boldsymbol{X}^{\boldsymbol{\gamma}}), \tag{5}$$

where the sum is over all the subsets of indices $\boldsymbol{\gamma} \subseteq \{1, \ldots, n\}$, with $|\boldsymbol{\gamma}|$ being the cardinality of $\boldsymbol{\gamma}$ and $\boldsymbol{X}^{\boldsymbol{\gamma}}$ the vector of all variables with indices in $\boldsymbol{\gamma}$. For $n = 2$, Eq. (5) reduces to the well-known *mutual information*

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2).$$

For $n = 3$, Eq. (5) gives

$$\begin{aligned}I(X_1; X_2; X_3) &= I(X_i; X_j) - I(X_i; X_j | X_k) \\ &= I(X_i; X_j) + I(X_i; X_k) - I(X_i; X_j, X_k)\end{aligned} \tag{6}$$

for $\{i, j, k\} = \{1, 2, 3\}$, which is known to measure the difference between synergy and redundancy [17], and has found applications in a range of scenarios including genetic networks [49], neural signals [7], and engineered communication systems [50]. Specifically, redundancy dominates when $I(X_1; X_2; X_3) \geqslant 0$; e.g., if X_1 is a Bernoulli random variable

with $p = 1/2$ and $X_1 = X_2 = X_3$, then $I(X_1;X_2;X_3) = 1$. In contrast, synergy dominates when $I(X_1;X_2;X_3) \leqslant 0$, corresponding to statistical structures that are present in the full distribution but not in the pairwise marginals. For example, if Y_1 and Y_2 are independent Bernoulli variables with $p = 1/2$ and $Y_3 = Y_1 + Y_2 \pmod 2$ (i.e., an `xor` logic gate) then $I(Y_1;Y_2;Y_3) = -1$, since these variables are pairwise independent while globally correlated [51]. Unfortunately, for $n \geqslant 4$ the coinformation no longer reflects the balance between redundancy and synergy (see Sec. V of Ref. [17]).

To contrast with the interaction information, the next lemma presents some basic properties of Ω (the proofs are left for the reader).

Lemma 1. The O-information satisfies the following properties:

(i) Ω does not depend on the order of $X_1, \ldots, X_n$.
(ii) $\Omega(X_1, X_2) = 0$ for any $p_{X_1X_2}$.
(iii) $\Omega(X_1, X_2, X_3) = I(X_1;X_2;X_3)$ for any p_{X^3}.

Property (i) shows that Ω reflects an intrinsic property of the system, without the need of dividing the variables in groups with differentiated roles (e.g., targets vs predictors, or input vs output). Property (ii) confirms that Ω captures only interactions that go beyond pairwise relationships. Finally, property (iii) shows that when $n = 3$ the O-information is equal to $I(X_1;X_2;X_3)$. Interestingly, a direct calculation shows that if $n > 3$ then in general $\Omega(\boldsymbol{X}^n) \neq I(X_1;X_2;\ldots;X_n)$.

At this stage, one might wonder if the O-information could provide a metric for quantifying the balance of redundancy and synergy, as the interaction information does for $n = 3$. Intuitively, one could expect redundant systems to have small $\mathrm{DTC}(\boldsymbol{X}^n)$ due to the multiple copies of the same information that exist in the system, while having large values of $\mathrm{TC}(\boldsymbol{X}^n)$ because of the constraints that are needed to ensure that the variables remain correlated. On the other hand, synergistic systems are expected to have small values of $\mathrm{TC}(\boldsymbol{X}^n)$ due to the few high-order constraints that rule the system, while having larger values of $\mathrm{DTC}(\boldsymbol{X}^n)$ due to the weak low-order structure. These insights are captured in the following definition, which is supported by multiple findings presented in the following sections.

Definition 2. If $\Omega(\boldsymbol{X}^n) > 0$ we say that the system is *redundancy dominated*, while if $\Omega(\boldsymbol{X}^n) < 0$ we say it is *synergy dominated*.

In previous work we used another metric to assess synergy- and redundancy-dominated systems [52]. Appendix A provides an analytical and numerical account of the consistency between these two metrics.

B. Information decompositions

This section presents information decompositions that deepen our understanding of the O-information. In the following, we first introduce the partition lattice, which is then used to build decompositions of the TC, DTC, and Ω. Information lattices have also been explored in Ref. [53].

1. The lattice of partitions

Let us characterize the possible ways in which one can sequentially decompose the system described by $\boldsymbol{X}^n$. For this, let us consider partitions $\pi = (\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2|\ldots|\boldsymbol{\alpha}_m)$ of the set of indices $\{1, \ldots, n\}$, which are collections of *cells* $\boldsymbol{\alpha}_j = \{\alpha_j^1, \ldots, \alpha_j^{l(j)}\} \subset \{1, \ldots, n\}$ that are disjoint and satisfy $\bigcup_{j=1}^m \boldsymbol{\alpha}_j = \{1, \ldots, n\}$. The collection of all possible partitions of $\{1, \ldots, n\}$, denoted by $\mathcal{P}_n$, has a lattice structure [54] enabled by the partial ordering introduced by the refinement relationship, in which $\pi_2 \succeq \pi_1$ if π_2 is *finer* [55] than π_1 (or, equivalently, if π_1 is *coarser* than π_2). A partition π_2 is said to *cover* π_1 if $\pi_2 \succeq \pi_1$ and it is not possible to find another partition π_3 such that $\pi_2 \succeq \pi_3 \succeq \pi_1$ [56]. For this partial order relationship, $\pi_{\text{source}} = (12\ldots n)$ is the unique infimum of $\mathcal{P}_n$, and $\pi_{\text{sink}} = (1|2|\ldots|n)$ is the unique supremum of $\mathcal{P}_n$.

A directed acyclic graph (DAG) $\mathcal{G}_n$ can be built, where the nodes are the partitions in $\mathcal{P}_n$, and a directed edge exists from π_1 to π_2 if and only if π_2 covers π_1 [57]. A path $\mathtt{p}$ in $\mathcal{G}_n$ joining two partitions π_a and π_b is a sequence of nodes $\mathtt{p} = (\pi_1, \ldots, \pi_L)$, where $\pi_1 = \pi_\mathrm{a}$, $\pi_L = \pi_\mathrm{b}$, and π_{i+1} covers π_i for all $i \in \{1, \ldots, L-1\}$. The collection of all paths from π_a to π_b is denoted by $\mathtt{P}(\pi_\mathrm{a}, \pi_\mathrm{b})$ [58]. If the edge joining π_1 and π_2 has a weight $\upsilon(\pi_1, \pi_2)$ associated, then the corresponding *path weight* of $\mathtt{p} = (\pi_1, \ldots, \pi_L)$ is merely the summation of all edge weights along $\mathtt{p}$:

$$W(\mathtt{p};\upsilon) := \sum_{k=1}^{L-1} \upsilon(\pi_k, \pi_{k+1}). \tag{7}$$

2. Lattice decompositions of $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$

Let us build some useful weight functions over $\mathcal{G}_n$. We first assign to each node $\pi = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_L) \in \mathcal{P}_n$ the value

$$H(\pi) := H\left(\prod_{j=1}^{L} p_{\boldsymbol{X}^{\boldsymbol{\alpha}_j}}\right) = \sum_{j=1}^{L} H(\boldsymbol{X}^{\boldsymbol{\alpha}_j})$$

with $\boldsymbol{X}^{\boldsymbol{\alpha}_j} = (X_{\alpha_j^1}, \ldots, X_{\alpha_j^{l(j)}})$, which corresponds to the entropy of the probability distribution $\prod_{j=1}^{L} p_{\boldsymbol{X}^{\boldsymbol{\alpha}_j}}$ that includes interdependencies within cells, but not across cells. To each edge of $\mathcal{G}_n$ we assign a weight

$$\upsilon_\mathrm{h}(\pi_1, \pi_2) := H(\pi_2) - H(\pi_1). \tag{8}$$

Since $H(\pi_\mathrm{a}) \geqslant H(\pi_\mathrm{b})$ if $\pi_\mathrm{a} \succeq \pi_\mathrm{b}$, one can represent $\mathcal{G}_n$ under υ_h by placing nodes with more cells in higher layers (see the upper half of Fig. 2).

Alternatively, let us now consider the residual entropy of $\pi = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_m) \in \mathcal{P}_n$, which is given by $R(\pi) := \sum_{k=1}^{m} R_{\boldsymbol{\alpha}_k}$, with

$$R_{\boldsymbol{\alpha}_k} := H(\boldsymbol{X}^{\boldsymbol{\alpha}_k}|\boldsymbol{X}^{\boldsymbol{\alpha}_1}, \ldots, \boldsymbol{X}^{\boldsymbol{\alpha}_{k-1}}, \boldsymbol{X}^{\boldsymbol{\alpha}_{k+1}}, \ldots, \boldsymbol{X}^{\boldsymbol{\alpha}_m}).$$

The above generalizes the notion of residual entropy per individual variable given in Sec. II B 2 [59]. With this, we introduce weights to each edge of $\mathcal{G}_n$ based on residuals, given by

$$\upsilon_\mathrm{r}(\pi_1, \pi_2) := R(\pi_1) - R(\pi_2). \tag{9}$$

As residual entropy decreases when the partition is refined (see Appendix B), in this case one can illustrate the corresponding DAG by placing nodes with more cells in lower positions (see lower half of Fig. 2).

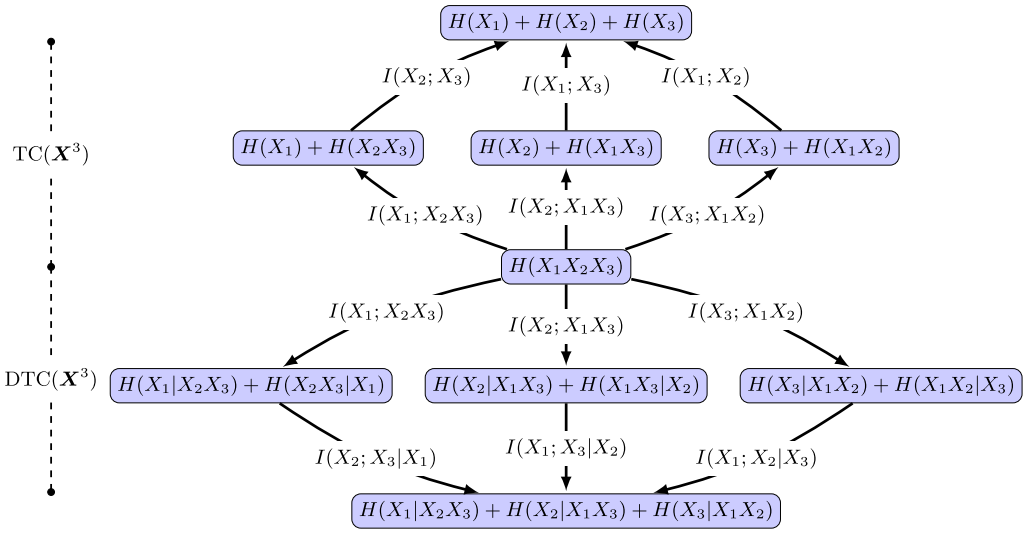

FIG. 2. Double diamond diagram with the possible sequences of binary partitions of three variables. Every path from the source node ($H(\boldsymbol{X}^3)$) to the two sink nodes $[H(X_1)+H(X_2)+H(X_3)$ and $H(X_1|X_2X_3)+H(X_2|X_1X_3)+H(X_3|X_1X_2)]$ corresponds to a decomposition of either $\mathrm{TC}(\boldsymbol{X}^3)$ or $\mathrm{DTC}(\boldsymbol{X}^3)$.

Conveniently, for every edge v_h and v_r correspond to a mutual information or a conditional mutual information term, respectively. This is illustrated in the edges of Fig. 2 and formalized in Appendices C and D.

The next result shows that the weights v_h and v_r provide decompositions for $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$, respectively.

Lemma 2. Every path $\mathtt{p} \in \mathtt{P}(\pi_\mathrm{source}, \pi_\mathrm{sink})$ provides the following decompositions:

$$\mathrm{TC}(\boldsymbol{X}^n) = W(\mathtt{p}; v_\mathrm{h}), \quad \mathrm{DTC}(\boldsymbol{X}^n) = W(\mathtt{p}; v_\mathrm{r}).$$

Proof. See Appendix C. ■

Example 3. For the case of $n = 3$, there are three paths joining source and sink:

$$\mathtt{p}_1 = \{(123), (1|23), (1|2|3)\},$$
$$\mathtt{p}_2 = \{(123), (2|13), (1|2|3)\},$$
$$\mathtt{p}_3 = \{(123), (3|12), (1|2|3)\}.$$

Lemma 2 shows that $\mathrm{TC}(\boldsymbol{X}^3) = W(\mathtt{p}_i; v_\mathrm{h})$ and $\mathrm{DTC}(\boldsymbol{X}^3) = W(\mathtt{p}_i; v_\mathrm{r})$ for $i \in \{1, 2, 3\}$, which provides the following decompositions:

$$\mathrm{TC}(\boldsymbol{X}^3) = I(X_i; X_j, X_k) + I(X_j; X_k),$$
$$\mathrm{DTC}(\boldsymbol{X}^3) = I(X_i; X_j, X_k) + I(X_j; X_k|X_i).$$

3. *Lattice decomposition of* $\boldsymbol{\Omega(X^n)}$

Let us now leverage the results presented in the previous subsection to develop decompositions for the O-information. For this, let us first introduce a new assignment of weights for the edges of $\mathcal{G}_n$, given by

$$v_\mathrm{s}(\pi_1, \pi_2) := v_\mathrm{h}(\pi_1, \pi_2) - v_\mathrm{r}(\pi_1, \pi_2). \tag{10}$$

In contrast with Eqs. (8) and (9), these weights can attain negative values. The following key result shows that the weights v_s provide a decomposition of $\Omega(\boldsymbol{X}^n)$.

Proposition 1. Every path $\mathtt{p} \in \mathtt{P}(\pi_\mathrm{source}, \pi_\mathrm{sink})$ provides the following decomposition:

$$\Omega(\boldsymbol{X}^n) = W(\mathtt{p}; v_\mathrm{s}). \tag{11}$$

Moreover, Eq. (11) is a sum of interaction information terms of the form in Eq. (6).

Proof. See Appendix D. ■

This finding extends property (iii) of Lemma 1 by showing that the O-information can always be expressed as a sum of interaction information terms of three sets of variables (see Corollary 1 below for an explicit example of this). As a consequence, the O-information inherits the capabilities of the triple interaction information for reflecting the balance between synergies and redundancies, and is applicable to systems of any size. This decomposition of the O-information is analogous to the one introduced in Ref. [10] for the redundancy-synergy index.

An inconvenient feature of partition lattices is that they grow superexponentially with system size [60], and hence heuristic methods for exploring them are necessary. A particularly interesting subfamily of $\mathtt{P}(\pi_\mathrm{source}, \pi_\mathrm{sink})$ is composed of the "assembly paths," which have the form (up to relabeling)

$$\mathtt{p}_\mathrm{a} = \{(12\ldots n), (12\ldots(n-1)|n), \ldots, (1|2|\ldots|n)\}. \tag{12}$$

These paths can be thought of as the process of first separating X_n from the rest of the system, then X_{n-1}, and so on. Conversely, by considering them backwards, one can think of these paths as first connecting X_1 and X_2, then connecting X_3 to $\boldsymbol{X}^2$, and so on—i.e., as assembling the system by sequentially placing its pieces together. The following corollary of Proposition 1 presents useful decompositions of $\mathrm{TC}(\boldsymbol{X}^n)$, $\mathrm{DTC}(\boldsymbol{X}^n)$, and $\Omega(\boldsymbol{X}^n)$ in terms of assembly paths.

Corollary 1. For an assembly path as given in Eq. (12), the corresponding decompositions of the TC, DTC, and O-information are

$$\mathrm{TC}(\boldsymbol{X}^n) = \sum_{i=2}^{n} I(X_i; \boldsymbol{X}^{i-1}), \tag{13}$$

$$\mathrm{DTC}(\boldsymbol{X}^n) = I(X_n; \boldsymbol{X}^{n-1}) + \sum_{j=2}^{n-1} I\left(X_j; \boldsymbol{X}^{j-1} | \boldsymbol{X}_{j+1}^n\right), \quad (14)$$

$$\Omega(\boldsymbol{X}^n) = \sum_{k=2}^{n-1} I\left(X_k; \boldsymbol{X}^{k-1}; \boldsymbol{X}_{k+1}^n\right), \quad (15)$$

with $\boldsymbol{X}_k^n = (X_k, X_{k+1}, \ldots, X_n)$ and $\boldsymbol{X}^k = (X_1, \ldots, X_k)$.

As a concluding remark, let us note that the decompositions presented by Corollary 1 are valid for any relabeling of the indices (i.e., any ordering of the system's variables). This property is a direct consequence of the lattice construction developed in this subsection, which plays an important role in the following sections.

IV. UNDERSTANDING THE O-INFORMATION

By definition, $\Omega > 0$ implies that the interdependencies are better described as shared randomness, while $\Omega < 0$ implies that they are better explained as collective constraints. In this section we explore this further, examining what the magnitude of Ω tells us about the system.

Through this section we use the shorthand notation $|\mathcal{X}| := \max_{j=1,\ldots,n} |\mathcal{X}_j|$ for the cardinality of the largest alphabet in $\boldsymbol{X}^n$.

A. Characterizing extreme values of Ω

Let us explore the range of values that the O-information can attain. As a first step, Lemma 3 provides bounds for $\mathrm{TC}(\boldsymbol{X}^n)$, $\mathrm{DTC}(\boldsymbol{X}^n)$, and $\Omega(\boldsymbol{X}^n)$.

Lemma 3. The following bounds hold.

(1) $(n-1)\log|\mathcal{X}| \geqslant \mathrm{TC}(\boldsymbol{X}^n) \geqslant 0$.

(2) $(n-1)\log|\mathcal{X}| \geqslant \mathrm{DTC}(\boldsymbol{X}^n) \geqslant 0$.

(3) $n\log|\mathcal{X}| \geqslant \mathrm{TC}(\boldsymbol{X}^n) + \mathrm{DTC}(\boldsymbol{X}^n) \geqslant 0$.

(4) $(n-2)\log|\mathcal{X}| \geqslant \Omega(\boldsymbol{X}^n) \geqslant (2-n)\log|\mathcal{X}|$.

Moreover, these bounds are tight.

Proof. See Appendix G. ■

Let us introduce some nomenclature. A random binary vector $\boldsymbol{X}^n$ is said to be a "n-bit copy" if X_1 is a Bernoulli random variable with parameter $p = 1/2$ (i.e., a *fair coin*) and $X_1 = X_2 = \cdots = X_n$. Also, a random binary vector $\boldsymbol{X}^n$ is said to be an "n-bit `xor`" if $\boldsymbol{X}^{n-1}$ are i.i.d. fair coins and $X_n = \sum_{j=1}^{n-1} X_j \pmod 2$. Our next result shows that these two distributions attain the upper and lower bounds of the O-information.

Proposition 2. Let $\boldsymbol{X}^n$ be a binary vector with $n \geqslant 3$. Then, the following holds:

(1) $\Omega(\boldsymbol{X}^n) = n-2$, if and only if $\boldsymbol{X}^n$ is a n-bit copy.

(2) $\Omega(\boldsymbol{X}^n) = 2-n$, if and only if $\boldsymbol{X}^n$ is a n-bit `xor`.

Proof. See Appendix F. ■

Corollary 2. The same proof can be used to confirm that for variables with $|\mathcal{X}_1| = \cdots = |\mathcal{X}_n| = m$ the maximum $\Omega(\boldsymbol{X}^n) = (n-2)\log m$ is attained by variables which are a copy of each other, while the minimum $\Omega(\boldsymbol{X}^n) = (2-n)\log m$ corresponds to when $\boldsymbol{X}^{n-1}$ are independent and uniformly distributed and $X_n = \sum_{j=1}^{n-1} X_j \pmod m$.

Proposition 2 points out an important difference between the O-information and the interaction information: if $\boldsymbol{X}^n$ is an n-bit `xor` then $\Omega(\boldsymbol{X}^n) = 2-n$ is consistently negative and decreasing with n, while $I(X_1; \ldots; X_n) = (-1)^{n+1}$ oddly oscillates between -1 and $+1$. This result also points out the convenience of merging $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$ into $\Omega(\boldsymbol{X}^n)$, as only the latter has the n-bit copy and the n-bit `xor` as unique extremes.

Finally, note that Ω is continuous over small changes in $p_{\boldsymbol{X}^n}$, as it can be expressed as a linear combination of Shannon entropies (see Definition 1). Therefore, Proposition 2 guarantees that distributions that are similar to an n-bit copy have a positive O-information, while distributions close to an n-bit `xor` have negative O-information.

B. Statistical structures across scales

In this section we study how the O-information is related to statistical structures of subsets of $\boldsymbol{X}^n$—i.e., structures at different scales of the system. For simplicity, we assume in this subsection that $|\mathcal{X}|$ is finite.

In the next proposition we present some fundamental restrictions between the total correlation of subsystems and the value of $\Omega(\boldsymbol{X}^n)$.

Proposition 3. If $\Omega(\boldsymbol{X}^n) \geqslant 0$, then for all $m \in [n-1]$

$$\min_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}}) \geqslant \Omega(\boldsymbol{X}^n) - (n-m-1)\log|\mathcal{X}|. \quad (16)$$

If $\Omega(\boldsymbol{X}^n) \leqslant 0$, then for all $m \in [n-1]$

$$\max_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}}) \leqslant \Omega(\boldsymbol{X}^n) + (n-2)\log|\mathcal{X}|. \quad (17)$$

Both bounds are tight if $|\Omega| \geqslant (n-m+1)\log|\mathcal{X}|$.

Proof. See Appendix G. ■

Corollary 3. The following bounds hold for all $\boldsymbol{\gamma} \subseteq \{1, \ldots, n\}$ with $|\boldsymbol{\gamma}| = m$:

$$\min\left\{m-1, \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} + (n-2)\right\} \geqslant \frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|} \geqslant \max\left\{0, \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} - (n-m-1)\right\}.$$

Corollary 3 shows that positive values of Ω constrain subgroups to be correlated: if $\Omega(\boldsymbol{X}^n) \geqslant (n-m-1)\log|\mathcal{X}|$ then all groups of m or more variables must have some statistical dependency. Negative values of Ω, on the other hand, impose limits on the allowed correlation strength: if $\Omega(\boldsymbol{X}^n) \leqslant -(n-m-1)\log|\mathcal{X}|$ then the correlation of all groups of m or more variables is upper bounded. As an example, for $|\mathcal{X}| = 2$ and $m = 2$ the bounds given in Corollary 3 are

$$\max\{1, \Omega(\boldsymbol{X}^n) + n - 2\} \geqslant I(X_i; X_j) \geqslant \min\{0, \Omega(\boldsymbol{X}^n) - (n-3)\},$$

for all $i, j \in \{1, \ldots, n\}$, which shows that the bounds related to Ω are only active when $n-3 \leqslant |\Omega| \leqslant n-2$.

In conclusion, the sign of Ω determines whether the constraint is a lower or upper bound, and $|\Omega|$ determines which scales of the system are affected, with smaller groups being harder to constrain—i.e., requiring higher absolute values of Ω. The relationship between the system's scales and the values of Ω is illustrated in Fig. 3.

The next result corresponds to the converse of Corollary 3, and shows how interactions at different scales limit the achievable values of Ω.

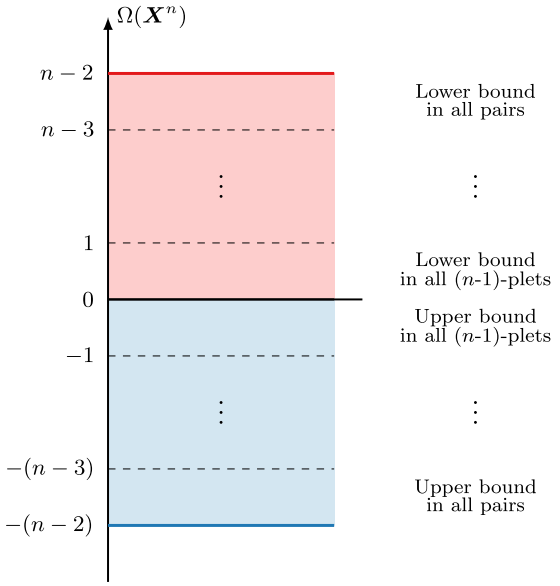

FIG. 3. Diagram of how values of the O-information impose limits on the strength of interactions—as measured by $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$—at different scales. Positive (negative) values of Ω put lower (upper) bounds on subsets of $\boldsymbol{X}^n$, and higher absolute values of Ω put bounds on subsystems of smaller sizes.

Corollary 4. For a given $\boldsymbol{\gamma} \subset \{1, \ldots, n\}$ with $|\boldsymbol{\gamma}| = m$, the following bounds on Ω hold:

$$n - m - 1 + \frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|} \geqslant \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} \geqslant -(n-2) + \frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|}.$$

By comparing it with Lemma 3, this result shows that a large $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ does not allow Ω to reach its lower bound. On the other hand, small values of $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ decrease the upper bound, forbidding high values of Ω. Additionally, note that fixing the value of only one subset of m variables reduces the range of values of Ω from $2(n-2)$ to $2(n-2)-(m-1)$. The following example illustrates these findings.

Example 4. Let us consider a system $\boldsymbol{X}^n$ of binary variables, two of which are related by the marginal distribution

$$p_{X_1X_2}(x_1, x_2) = \frac{(1-\eta)^{1-|x_1-x_2|}\eta^{|x_1-x_2|}}{2}.$$

That is, X_1 and X_2 are fair coins linked by a binary symmetric channel with crossover probability η (see Sec. 7 of Ref. [38]). Hence, $\mathrm{TC}(\boldsymbol{X}^2) = I(X_1; X_2) = 1 - H(\eta)$, with $H(\eta) = -\eta\log\eta - (1-\eta)\log(1-\eta)$ being the binary entropy function. By considering $m = 2$, Corollary 4 states that

$$n - 2 - H(\eta) \geqslant \Omega(\boldsymbol{X}^n) \geqslant -[n - 3 + H(\eta)],$$

which is illustrated in Fig. 4. Moreover, using Eq. (15) one can verify that the upper bound (solid red line) is attained when $X_2 = X_3 = \cdots = X_n$, while the lower bound (solid blue line) is attained when $X_3, \ldots, X_{n-1}$ are independent fair coins and $X_n = \sum_{j=1}^{n-1} X_j \pmod 2$ [61].

C. Ω as a superposition of tendencies

This subsection explores sufficient conditions that make a system have a small O-information. As a preliminary step, the next result shows that Ω is additive for systems with independent subsystems.

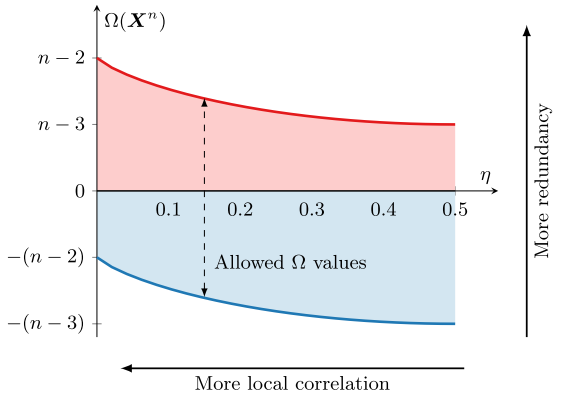

FIG. 4. Bounds of the O-information when two variables are connected via a binary symmetric channel with crossover probability η (see Example 4).

Lemma 4. If $p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) = \prod_{k=1}^m p_{\boldsymbol{X}^{\boldsymbol{\alpha}_k}}(\boldsymbol{x}^{\boldsymbol{\alpha}_k})$ for some partition $\pi = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_m)$, then

$$\Omega(\boldsymbol{X}^n) = \sum_{k=1}^m \Omega(\boldsymbol{X}^{\boldsymbol{\alpha}_k}).$$

Proof. Let us consider the case $\pi = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, as the general case is then guaranteed by induction. Using Eqs. (13) and (14) it is direct to check that, due to the independence, $\mathrm{TC}(\boldsymbol{X}^n) = \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\alpha}_1}) + \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\alpha}_2})$ and $\mathrm{DTC}(\boldsymbol{X}^n) = \mathrm{DTC}(\boldsymbol{X}^{\boldsymbol{\alpha}_1}) + \mathrm{DTC}(\boldsymbol{X}^{\boldsymbol{\alpha}_2})$. Then, the desired result follows from induction on the number of cells and the definition of Ω. ∎

Corollary 5. $\Omega(\boldsymbol{X}^n) = 0$ for all systems the joint distribution of which can be factorized as

$$p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) = \prod_{k=1}^{n/2} p_{X_{2k-1}X_{2k}}(x_{2k-1}, x_{2k}). \tag{18}$$

Proof. Using Eq. (18) and Lemma 4 we find that

$$\Omega(\boldsymbol{X}^n) = \sum_{k=1}^{n/2} \Omega(X_{2k-1}, X_{2k}) = 0,$$

where the last equality is a consequence of the O-information being zero for sets of two variables, as shown in Proposition 1. ∎

Corollary 5 states that having disjoint pairwise interactions is a sufficient condition for $\Omega = 0$ to hold. However, this condition is not necessary: from Lemma 4 we can see that a system composed by redundant ($\Omega > 0$) and synergistic ($\Omega < 0$) subsystems can attain zero net O-information due to "destructive interference."

As a consequence, the O-information can be understood as the result of a superposition of behaviors of subsystems. Therefore, $\Omega = 0$ can take place in two qualitatively different scenarios: systems in which redundancies and synergies are balanced, or systems with only disjoint pairwise effects. Some of these cases can be resolved by considering the information diagram of $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$ (see Fig. 2), or by studying the O-information of parts of the system. However, it is

important to remark that redundancy and synergy can coexist either in disjoint subsystems or within the same variables. An insightful example of the latter case can be found in Sec. 2 of Ref. [62].

As a final remark, note that systems where pairwise interdependencies are overlapping (e.g., pairwise maximum entropy models [63]) cannot be factorized as required by Corollary 5, and hence can have either positive or negative O-information [64].

V. RELATIONSHIP WITH OTHER NOTIONS OF HIGH-ORDER EFFECTS

A. High-order interactions in statistical mechanics

A popular approach to address high-order interactions in the statistical physics literature is via Hamiltonians that include interaction terms with three or more variables [12]. For example, systems of n spins (i.e., $\mathcal{X}_i = \{-1, 1\}$ for $i = 1, \ldots, n$) that exhibit kth-order interactions are usually represented by probability distributions of the form

$$p_{X^n}(\boldsymbol{x}^n) = \frac{e^{-\beta \mathcal{H}_k(\boldsymbol{x}^n)}}{Z}, \tag{19}$$

where β is the inverse temperature, Z is a normalization constant, and $\mathcal{H}(\boldsymbol{x}^n)$ is a Hamiltonian given by

$$\mathcal{H}_k(\boldsymbol{x}^n) = -\sum_{i=1}^{n} J_i x_i - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} J_{i,j} x_i x_j \cdots - \sum_{|\boldsymbol{\gamma}|=k} J_{\boldsymbol{\gamma}} \prod_{i\in\boldsymbol{\gamma}} x_i,$$

with the last sum running over all subsets $\boldsymbol{\gamma} \subseteq \{1, \ldots, n\}$ of size $|\boldsymbol{\gamma}| = k$. According to Eq. (19), configurations with lower $\mathcal{H}_k(\boldsymbol{x}^n)$ are more likely to be visited. Note that J_i quantify external influences acting over individual spins, while $J_{\boldsymbol{\gamma}}$ for $|\boldsymbol{\gamma}| \geqslant 2$ represent the strength of the interactions; in particular, if $J_{i,k} > 0$ then the pair X_i, X_k tend to be aligned, while if $J_{i,k} < 0$ they tend to be antialigned. As a matter of fact, $\boldsymbol{X}^n$ are independent if and only if $J_{\boldsymbol{\gamma}} = 0$ for all $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \geqslant 2$. Models with kth-order interactions have been studied via the maximum entropy principle [12], information geometry [65], and PID [66].

Considering the results presented in previous sections, one could expect that systems with high-order interactions (i.e., large k) should attain lower values of Ω than systems with low-order interactions (i.e., small k). To confirm this hypothesis, we studied ensembles of systems with kth-order interactions, and analyzed how the value of Ω is influenced by k. For this, we considered random Hamiltonians with $J_{\boldsymbol{\gamma}}$ drawn i.i.d. from a standard normal distribution and $\beta = 0.1$.

In agreement with intuition, results show that Ω is usually very close to zero for $k = 2$, and becomes negative as k grows (Fig. 5). These results suggest that the notion of synergy measured by Ω is consistent with the traditional ideas of high-order interactions from statistical physics.

B. Complexity and integration

In their seminal 1994 article, Tononi, Edelman, and Sporns devised a measure of complexity (henceforth called *TSE complexity*) to describe the interplay between local segregation and global integration [5,13]. The TSE complexity is

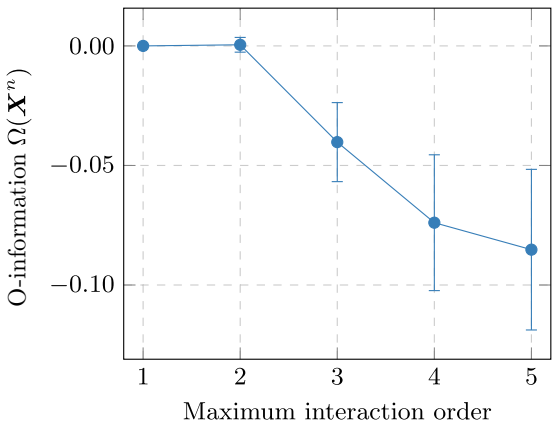

FIG. 5. Mean value and confidence intervals of ensembles of systems of $n = 5$ spins with randomly generated Hamiltonians. By including high-order interaction terms, net synergy increases and Ω decreases.

defined as

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) := \sum_{k=1}^{n} \left[\frac{k}{n}\mathrm{TC}(\boldsymbol{X}^n) - C_n(k)\right], \tag{20}$$

where $C_n(k) = \binom{n}{k}^{-1} \sum_{|\boldsymbol{\gamma}|=k} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ is the average total correlation of the subsets $\boldsymbol{\gamma} \subseteq \{1, \ldots, n\}$ of size $|\boldsymbol{\gamma}| = k$. By measuring the convexity of $C_n(k)$ as a function of k, the TSE complexity attempts to distinguish scenarios that exhibit "relative statistical independence of small subsets of the system [...] and significant deviations from independence of large subsets" ([13], Abstract), in the same spirit as our motivation behind Ω above.

To study the relationship between the TSE complexity and the O-information, it is useful to consider an alternative expression of the former:

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) = \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^{-1} \sum_{|\boldsymbol{\gamma}|=k} I\left(\boldsymbol{X}^{\boldsymbol{\gamma}}; \boldsymbol{X}^n_{-\boldsymbol{\gamma}}\right),$$

where $X^n_{-\boldsymbol{\gamma}}$ represents all the variables that are not in $\boldsymbol{\gamma}$, and $\lfloor \cdot \rfloor$ is the floor function. Motivated by this expression, let us introduce the quantity [67]

$$\Sigma(\boldsymbol{X}^n) := \mathrm{TC}(\boldsymbol{X}^n) + \mathrm{DTC}(\boldsymbol{X}^n) = \sum_{i=1} I\left(X_i; \boldsymbol{X}^n_{-i}\right). \tag{21}$$

By noting the similarities between Eqs. (21) and (21), together with the fact that $C_{\mathrm{TSE}}(\boldsymbol{X}^3) = \frac{1}{3}[C(\boldsymbol{X}^3) + \mathrm{DTC}(\boldsymbol{X}^3)]$, we can hypothesize that, qualitatively,

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) \propto \Sigma(\boldsymbol{X}^n). \tag{22}$$

Monte Carlo simulations show that this approximation is justified: when evaluated on distributions p_{X^n} sampled uniformly at random from the probability simplex, the correlation between Σ and C_{TSE} is consistently above 0.97 (Fig. 6). Moreover, Σ outperforms other proposed approximations of the TSE complexity [68].

Figure 6 and Eq. (22) suggest that the TSE complexity is large when either the shared randomness or the collective constraints are large. As a more direct example, we evaluate C_{TSE}

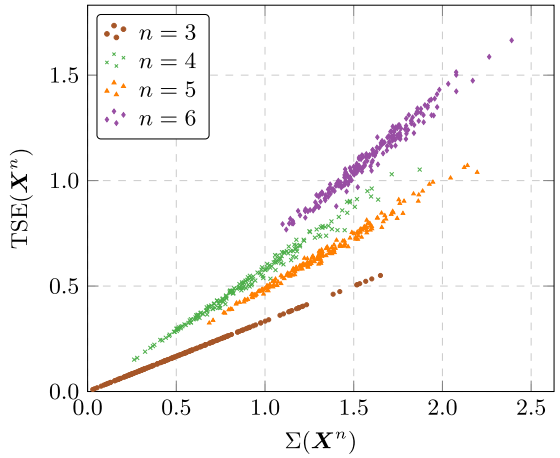

FIG. 6. The sum of the TC and DTC (denoted by Σ) is an accurate approximation of the TSE complexity. Each dot corresponds to probability distribution over n binary variables, which are sampled uniformly at random from the corresponding probability simplex.

in a distribution given by a linear mixture of the distributions of a three-bit copy and a three-bit `xor`, showing that C_{TSE} has exactly the same value in both extremes, and hence that it conflates redundancy with synergy (Fig. 7).

Taken together, our results show that the TSE complexity is a good metric of overall integration between parts of the system, but it generally fails to discriminate high- from low-order phenomena. Overall, the fact that

$$\Omega = \mathrm{TC} - \mathrm{DTC}, \quad C_{\mathrm{TSE}} \propto \mathrm{TC} + \mathrm{DTC} \tag{23}$$

suggests that the TSE complexity and the O-information are complementary, corresponding to an insightful "change of basis" from an elementary constraints vs randomness representation. Effectively, while both TC and DTC provide two measures of roughly the same phenomenon (interdependency

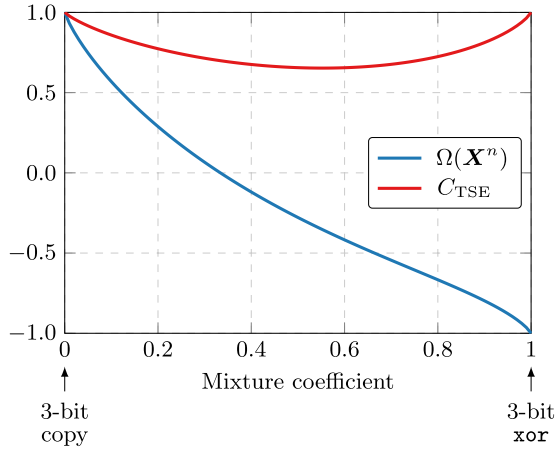

FIG. 7. C_{TSE} (upper line) and Ω (lower line) evaluated on a distribution resulting from a linear mixture between a copy (left) and an `xor` (right), showing that the TSE complexity conflates synergy and redundancy. The figure shows the case $n = 3$, but results are qualitatively similar for larger systems.

strength), Ω and C_{TSE} refer to different aspects: C_{TSE} gives an overarching account of the strength of the interdependencies within $\boldsymbol{X}^n$, and Ω indicates whether these correlations are predominantly redundant or synergistic.

VI. CASE STUDY: BAROQUE MUSIC SCORES

To illustrate the proposed framework in a data-driven application, this section presents a study of the multivariate statistics of musical scores from the Baroque period. In the following, Sec. VI A describes the procedure to obtain and analyze the data, and Sec. VI B discusses numerical results.

A. Method description

1. Data

Our analysis focuses on two sets of repertoire: the well-known chorales for four voices by Johann Sebastian Bach (1685–1750), and Opuses 1 and 3–6 by Arcangelo Corelli (1653–1713). All of these works correspond to the Baroque period (approx. 1600–1750), which is characterized by elaborate counterpoint between melodic lines. Baroque music usually exhibits a balance in the interest and richness of the parts of all the involved instruments, contrasting with the subsequent Classic (1730–1820) and Romantic (1780–1910) periods where higher voices tend to take the lead.

Our analysis is based on the electronic scores publicly available at [69]. We focused on scores with four melodic lines: four voices (soprano, alto, tenor, and bass) in the case of Bach's chorales, and four string instruments (first violin, second violin, viola, and cello) in the case of Corelli's pieces. The scores were preprocessed in PYTHON using the MUSIC21 package [70], which allowed us to select only the pieces written in major mode and to transpose them to C major. The melodic lines were transformed into a time series of 13 possible values (one for each note plus one for the silence), using the smallest rhythmic duration as the time unit. This generated $\approx 4 \times 10^4$ four-note chords for the chorales, and $\approx 8 \times 10^4$ for Corelli's pieces. With these data, the joint distribution of the values for the four-note chords was estimated using their empirical frequency [71].

2. Research questions and tools

We focus on the multivariate statistics of the harmonic structures of these pieces. In particular, we ask to what extent the notes played simultaneously by different instruments are redundant or synergistic. Our analysis focuses exclusively on harmony and chords, leaving melodic properties to future studies.

Let us denote by $\boldsymbol{X}^4$ the random vector of notes, where $|\mathcal{X}| = 13$. We first compute the marginal entropy of each voice, $H(X_k)$, which is an indicator of harmonic richness. We also compute the O-information of the ensemble $\Omega(\boldsymbol{X}^4)$, which determines the dominant behavior. Interestingly, for $n = 4$ the decomposition in Eq. (15) yields

$$\Omega(\boldsymbol{X}^4) = I(X_i; X_j; X_k, X_l) + I(X_k; X_l; X_i, X_j)$$

for $\{i, j, k, l\} = \{1, 2, 3, 4\}$. One can gain a fine-grained view of Ω by considering these interaction information terms, which can be seen as local contributions to Ω. More formally,

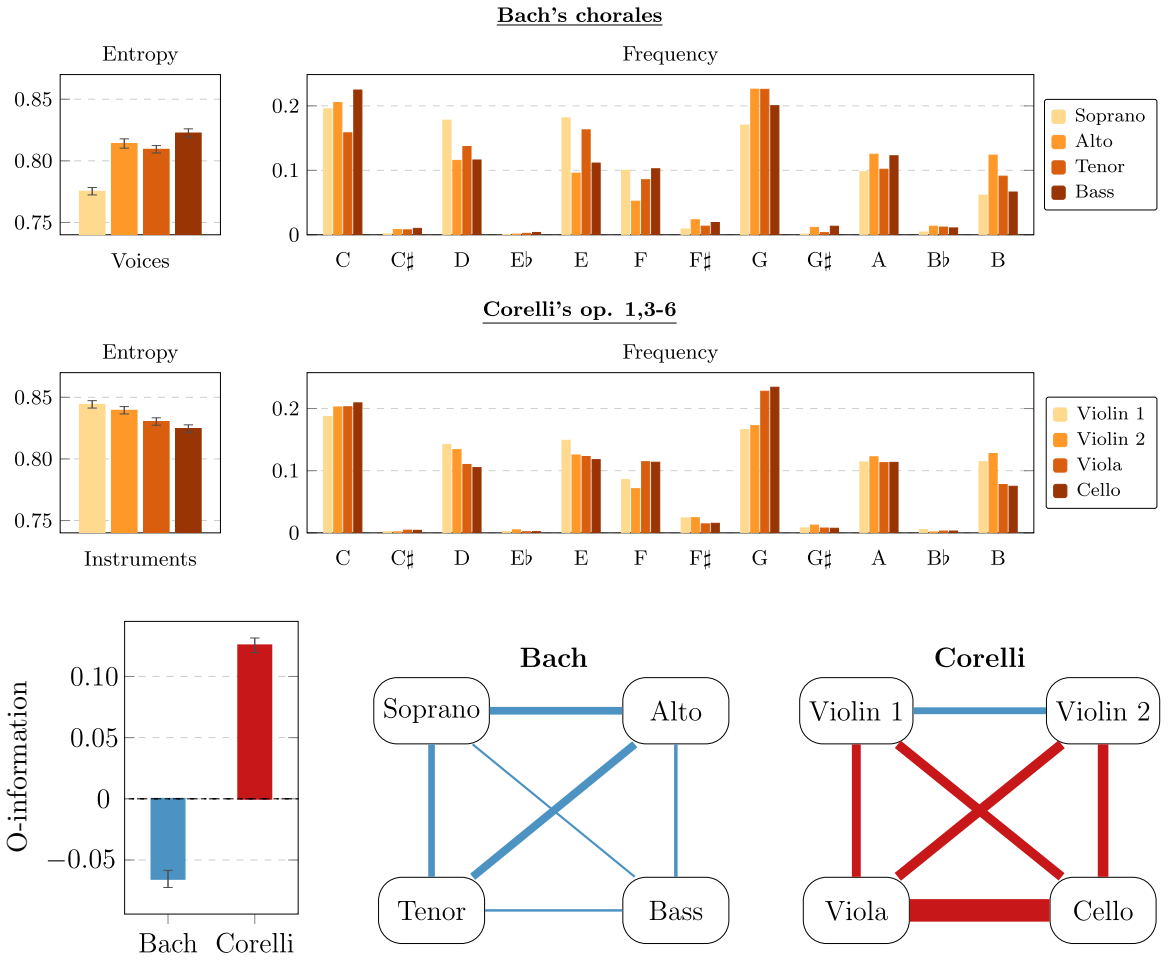

FIG. 8. Top: Entropy of the frequencies of appearance of each note in the studied pieces of Bach and Corelli, measured in muts (logarithm to base 13); standard errors were estimated via circular block-bootstrap. While the higher voices in Corelli have higher entropy, Bach's soprano has a lower entropy than all other voices. Bottom: Global O-information (left) and networks of local O-information (middle, right) with red reflecting redundancy (Ω, $w_{ij} > 0$) and blue reflecting synergy (Ω, $w_{ij} < 0$). Links characterize the triple interaction between each part of the corresponding dyad and the rest of the system. While Bach's chorales are synergy dominated, the pieces of Corelli are strongly redundant (mainly due to the viola and cello).

we define the *local O-information* between X_i and X_j as

$$\omega_{ij}(\boldsymbol{X}^n) := I\big(X_i; X_j; \boldsymbol{X}^n_{-ij}\big). \tag{24}$$

Please note that $\omega_{ij}(\boldsymbol{X}^n)$ refers to a relationship between a triplet: two individual variables (X_i and X_j) and the rest of the system ($\boldsymbol{X}^n_{-ij}$). Interestingly, these local terms could be of the opposite sign to the global $\Omega(\boldsymbol{X}^n)$, indicating local synergy (or redundancy) between some components within a predominantly redundant (or synergistic) system.

Since $X_1, \ldots, X_4$ take values among alphabets of cardinality $|\mathcal{X}| = 13$, we perform all computations employing logarithms to base 13, so that $H(X_k) \leqslant 1$ for all $k \in \{1, \ldots, 4\}$. We call this unit a *mut*, for *musical bit*.

B. Results

By studying the entropies of each voice, our results confirm that the four voices in these Baroque scores tend to have similar harmonical richness (Fig. 8, top left). In fact, their values are similar (although slightly lower) than $\log_{13} 7 \approx 0.845$ muts, which corresponds to a uniform distribution over the seven notes of a major scale (notes without sharp or flat). Also, our results show that the entropies in the music of Corelli are higher for instruments with higher register (i.e., the violins). In contrast, in Bach's music the soprano has significantly less entropy than the other voices. This could be related with the fact that these pieces were made to be used in public religious services [72], with the soprano conveying a melodic line that was intended to be sung by the attendees—and hence its structure is simpler to make it easy to sing.

Most strikingly, our analyses of the multivariate structure of the pieces show that Bach's chorales have negative O-information, suggesting that the harmonic structure of these pieces is dominated by synergistic effects (Fig. 8, bottom left). This result is further confirmed by the fact that all the local O-information terms are negative, which means that the pairwise dependence between any pair of voices is comparatively smaller than the global dependencies that exists within the group (see Table I).

In contrast, Corelli's pieces have positive O-information, suggesting that they are dominated by a redundant component. Interestingly, the local O-information has a positive value

TABLE I. Multivariate statistics of Baroque repertoire. For each pair of voices or instruments, we report the mutual information (MI), conditional mutual information (CMI), and local O-information (ω_{ij}). Quantities are measured in musical bits, or *muts* (logarithm to base 13). Standard errors were estimated via circular block-bootstrap, and in all cases are below the least significant figure shown in the table.

Bach's chorales				
		MI	CMI	ω_{ij}
Soprano	Alto	0.14	0.19	−0.05
Soprano	Tenor	0.12	0.16	−0.04
Soprano	Bass	0.15	0.16	−0.02
Alto	Tenor	0.17	0.22	−0.05
Alto	Bass	0.15	0.17	−0.02
Tenor	Bass	0.15	0.17	−0.02
Corelli's Opuses 1 and 3–6				
		MI	CMI	ω_{ij}
Violin 1	Violin 2	0.071	0.115	−0.04
Violin 1	Viola	0.086	0.028	0.06
Violin 1	Cello	0.095	0.034	0.06
Violin 2	Viola	0.118	0.054	0.07
Violin 2	Cello	0.107	0.039	0.07
Viola	Cello	0.630	0.460	0.17

for all pairs except for violins 1 and 2. The strongest O-information is the one between viola and cello, indicating that the parts of these two instruments are highly redundant.

The redundancy in the pieces of Corelli might be related to compositional practices for instrumental music in the Baroque period. In fact, the original score of many of the studied pieces was written for only three parts: two soloists and a bass line called "basso continuo." This bass line was supposed to be interpreted in different ways by the bass instruments, which in this case correspond to viola and cello. Therefore, it is fair to say that these instruments are redundant, as both of them are carrying the same bass line. Despite this redundancy, the relationship between the violins is still synergistic, which is appropriately captured by the negative value of their local O-information.

The dominance of synergy in the case of Bach could be the consequence of an artistic purpose. In effect, in the Baroque period the aim was that each voice should introduce unique elements into the piece. This goal could be easily achieved by superposing unrelated melodies; however, the overall result is arguably of limited interest due to the lack of global coordination. In contrast, a synergistic structure serves the Baroque ideal better, as it provides global constraints that ensure collective coherence while imposing weak pairwise constraints.

VII. CONCLUSION

We introduced $\Omega(\boldsymbol{X}^n)$ as the difference between the strength of the collective constraints and the shared randomness in a multivariate system $\boldsymbol{X}^n$. We argued that Ω captures the net balance between statistical synergy and redundancy, since (i) it is a sum of triple interaction information, (ii) it is maximized (minimized) by an n-bit copy (xor), and (iii) it imposes bounds over the interdependency allowed at different scales. According to this framework, synergistic systems are characterized by a large amount of shared randomness regulated by weak collective constraints, this being consistent with recent approaches to study emergence based on constructive logic [73]. Moreover, in deriving Ω, we also provided a joint source of explanation for three long-standing extensions of Shannon's mutual information (TC, DTC, and interaction information) in terms of shared randomness and collective constraints. The proposed framework is straightforward to generalize to continuous variables and apply to neural data, which will be done in a separate publication.

From the PID perspective, the O-information can be understood as a difference between redundancies and synergies. While the presented framework does not refine our current understanding of PID, it allows us to apply PID principles to large systems and circumvent some of the prohibitive scaling properties of PID. Additionally, the local O-information can be employed to identify subsystems with interesting high-order properties, which can guide the application of PID analyses while avoiding the need of computing the PID of the whole system.

The O-information was compared to other notions of high-order effects, most notably the TSE complexity [13]. We found that TSE does not measure statistical synergy as such, but total correlation strength. Moreover, our analysis suggests that Ω and TSE are complementary metrics: TSE gives an overarching account of the strength of the interdependencies within $\boldsymbol{X}^n$, and the O-information reveals whether these correlations are predominantly redundant or synergistic. We take this as a step towards a multidimensional framework that allows for a finer and more subtle taxonomy of complex systems.

The proposed framework was applied to Baroque music scores and found that Bach's chorales, unlike pieces by some of his contemporaries, are strongly synergistic as measured by Ω. Informally, we can speculate about the artistic role of synergy: synergistic music (like Bach's) allows each voice to contribute unique material while ensuring an overall harmonious integration of the ensemble. This delicate balance has an intriguing similarity with the coexistence of integration and differentiation in brain activity [5,6], suggesting unexplored relationships between music structure and neural organization.

ACKNOWLEDGMENTS

Part of this work was carried out while F.E.R. was at the School of Computer and Communication Sciences of École polytechnique fédérale de Lausanne (EPFL) as an academic visitor. The authors thank Shamil Chandaria, Alberto Pascual, and Nicolas Rivera for insightful discussions. F.E.R. was supported by the European Union's H2020 research and innovation program, under Marie Skłodowska-Curie Grant No. 702981.

APPENDIX A: COMPATIBILITY BETWEEN Ω AND PRIOR WORK

In prior work [52], we introduced $\psi(k)$ as

$$\psi(k) := \max_{j\in\{1,\ldots,n\}} \max_{\substack{\gamma\subseteq\{1,\ldots,n\}\\ |\gamma|=k,\, j\notin\gamma}} I(X_j; \boldsymbol{X}^{\gamma}).$$

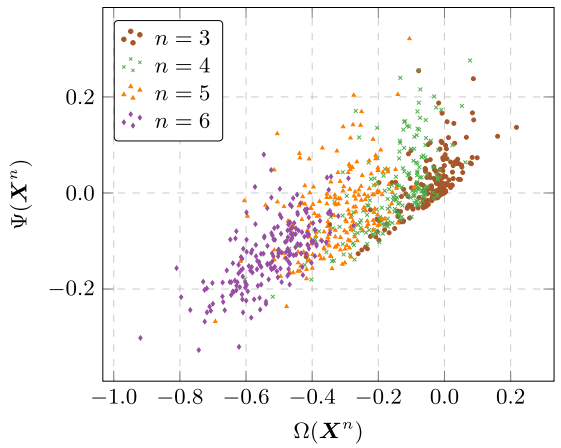

FIG. 9. The O-information and Ψ—introduced in our previous work [52]—have good agreement.

The growth profile of this nondecreasing function was taken as an indicator of the leading quality of the interdependency structure of $\boldsymbol{X}^n$, convexity being associated with statistical synergy, and concavity being associated with redundancy (see Definition 2 of Ref. [52]).

The relationship between these ideas and the ones developed in this paper can be established by noting that convexity in $\psi(k)$ implies that small scales of the system are relatively independent while large scales show correlation, which—due to the results of Sec. IV B—is the key characteristic of synergy-dominated systems. Conversely, concavity in $\psi(k)$ implies that some small groups of variables are highly correlated, which implies a relatively high value of $\mathrm{TC}(\boldsymbol{X}^n)$ and $\Omega(\boldsymbol{X}^n)$.

To enable a quantitative comparison between $\psi(k)$ and Ω, one can quantify the convexity or concavity of the former by measuring the distance from $\psi(k)$ to a straight line joining $\psi(1)$ and $\psi(n)$ as

$$\Psi(\boldsymbol{X}^n) := \sum_{k=1}^{n} \left[\psi(k) - \left(\frac{k}{n}[\psi(n) - \psi(1)] + \psi(1) \right) \right].$$

We computed Ω and Ψ of binary systems of different sizes generated randomly from a uniform distribution over the corresponding probability simplex. Our results show a good agreement between these two metrics (see Fig. 9), which confirms the analytic reasoning presented above.

In summary, Ω can be regarded as a formalization of the intuitive notions introduced in Ref. [52]. Moreover, Ω possesses more theoretical properties than Ψ and requires the calculation of a smaller number of terms.

APPENDIX B: $R(\pi)$ DECREASES FOR FINER PARTITIONS

Lemma 5. Let us consider two partitions $\pi_\mathrm{a} = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_K)$ and $\pi_\mathrm{b} = (\boldsymbol{\beta}_1|\ldots|\boldsymbol{\beta}_J)$ such that $\pi_\mathrm{b} \succeq \pi_\mathrm{a}$. Then, $R(\pi_\mathrm{b}) \leqslant R(\pi_\mathrm{a})$.

Proof. Let us assume that $\pi_\mathrm{a} = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_K\}$, $\pi_\mathrm{b} = (\boldsymbol{\beta}_1|\ldots|\boldsymbol{\beta}_J)$ such that $\pi_\mathrm{b} \succeq \pi_\mathrm{a}$, and consider a path $\mathrm{p} = (\pi_1, \ldots, \pi_L)$ in $\mathrm{P}(\pi_\mathrm{a}, \pi_\mathrm{b})$ so that $\pi_1 = \pi_\mathrm{a}$ and $\pi_L = \pi_\mathrm{b}$. To prove the lemma suffices to show that $R(\pi_{j+1}) \leqslant R(\pi_j)$ for $j = 1, \ldots, L-1$. As $\pi_1, \ldots, \pi_n$ are related by covering relationships, one just needs to prove the inequality for two partitions such that one covers the other.

Consider $\pi_1, \pi_2 \in \mathcal{P}_n$ such that π_2 covers π_1. As both partitions differ only in one elementary refinement, let us without loss of generality assume that the refinement is done on the last cell of π_1; i.e., $\pi_1 = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_m)$ and $\pi_2 = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_{m-1}|\tilde{\boldsymbol{\alpha}}_m|\tilde{\boldsymbol{\alpha}}_{m+1})$ so that $\tilde{\boldsymbol{\alpha}}_m \cup \tilde{\boldsymbol{\alpha}}_{m+1} = \boldsymbol{\alpha}_m$ and $\tilde{\boldsymbol{\alpha}}_m \cap \tilde{\boldsymbol{\alpha}}_{m+1} = \varnothing$. Then

$$\begin{aligned} R(\pi_1) - R(\pi_2) &= R_{\boldsymbol{\alpha}_m} - \left(R_{\tilde{\boldsymbol{\alpha}}_m} + R_{\tilde{\boldsymbol{\alpha}}_{m+1}}\right) \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m}; \boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}} | \boldsymbol{X}^{\boldsymbol{\alpha}_1} \ldots \boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}) \\ &\geqslant 0, \end{aligned}$$

proving the desired result.

APPENDIX C: PROOF OF LEMMA III B 2

Proof. Consider a path $\mathrm{p} \in \mathrm{P}(\pi_\mathrm{source}, \pi_\mathrm{sink})$, so that $\mathrm{p} = (\pi_1, \ldots, \pi_L)$ with $\pi_1 = \pi_\mathrm{source}$ and $\pi_L = \pi_\mathrm{sink}$. Then, by using Eqs. (7) and (8), a direct calculation shows that

$$\begin{aligned} W(\mathrm{p}; \upsilon_h) &= \sum_{j=1}^{L-1} [H(\pi_{j+1}) - H(\pi_j)] \\ &= H(\pi_\mathrm{sink}) - H(\pi_\mathrm{source}) \\ &= \sum_{i=1}^{n} H(X_i) - H(\boldsymbol{X}^n). \end{aligned}$$

Similarly, using Eqs. (7) and (9) gives

$$\begin{aligned} W(\mathrm{p}; \upsilon_\mathrm{r}) &= \sum_{j=1}^{L-1} [R(\pi_j) - R(\pi_{j+1})] \\ &= R(\pi_\mathrm{source}) - R(\pi_\mathrm{sink}) \\ &= H(\boldsymbol{X}^n) - \sum_{i=1}^{n} H(X_i | \boldsymbol{X}^n_{-i}). \end{aligned}$$

Both results make use of the fact that $W(\mathrm{p}; \upsilon_\mathrm{h})$ and $W(\mathrm{p}; \upsilon_\mathrm{r})$ are telescopic sums and all but the first and last terms cancel out. ■

APPENDIX D: PROOF OF PROPOSITION 1

Proof. Let us consider a path $\mathrm{p} \in \mathrm{P}(\pi_\mathrm{source}, \pi_\mathrm{sink})$. Then,

$$\begin{aligned} W(\mathrm{p}; \upsilon_\mathrm{s}) &= \sum_{j=1}^{L} \upsilon_\mathrm{s}(\pi_j, \pi_{j+1}) \\ &= \sum_{j=1}^{L} \upsilon_\mathrm{h}(\pi_j, \pi_{j+1}) - \sum_{k=1}^{L} \upsilon_\mathrm{r}(\pi_k, \pi_{k+1}) \\ &= \mathrm{TC}(\boldsymbol{X}^n) - \mathrm{DTC}(\boldsymbol{X}^n) = \Omega(\boldsymbol{X}^n), \end{aligned} \tag{D1}$$

which proves the first part of the theorem.

Thanks to Eq. (D1), one can prove the second part of the theorem by showing that if $\pi_\mathrm{a}, \pi_\mathrm{b} \in \mathcal{P}_n$ such that $\pi_\mathrm{b} \succeq \pi_\mathrm{a}$ then $\upsilon_\mathrm{s}(\pi_1, \pi_2)$ is equal to an interaction information. To show this, first note that if $\pi_\mathrm{b} \succeq \pi_\mathrm{a}$ then both

partitions differ only in one elementary refinement. Without loss of generality, we assume that the refinement is done on the last cell, such that $\pi_a = (\boldsymbol{\alpha}_1|\dots|\boldsymbol{\alpha}_m)$ and $\pi_b = (\boldsymbol{\alpha}_1|\dots|\boldsymbol{\alpha}_{m-1}|\tilde{\boldsymbol{\alpha}}_m|\tilde{\boldsymbol{\alpha}}_{m+1})$ such that $\tilde{\boldsymbol{\alpha}}_m \cap \tilde{\boldsymbol{\alpha}}_{m+1} = \varnothing$ and $\tilde{\boldsymbol{\alpha}}_m \cup \tilde{\boldsymbol{\alpha}}_{m+1} = \boldsymbol{\alpha}_m$. Then,

$$\begin{aligned} v_s(\pi_a, \pi_b) &= v_h(\pi_a, \pi_b) - v_r(\pi_a, \pi_b) \\ &= [H(\pi_b) - H(\pi_a)] - [R(\pi_a) - R(\pi_b)] \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m}; \boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}}) - I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m}; \boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}} | \boldsymbol{X}^{\boldsymbol{\alpha}_1} \dots \boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}) \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m}; \boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}}; \boldsymbol{X}^{\boldsymbol{\alpha}_1} \dots \boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}), \end{aligned}$$

which proves the desired result. ■

APPENDIX E: PROOF OF LEMMA IV A

Proof. Let us first note that

$$\log|\mathcal{X}| \geqslant I(X_i; X_j|X_k) \geqslant 0, \tag{E1}$$

$$\log|\mathcal{X}| \geqslant I(X_i; X_j; X_k) \geqslant -\log|\mathcal{X}|, \tag{E2}$$

for all $i, j, k \in \{1, \dots, n\}$. Above, Eq. (E2) follows from noting that $I(X_i; X_j; X_k) = I(X_i; X_j) - I(X_i; X_j|X_k)$, and applying the bounds in Eq. (E1). The proposition is proved by applying these inequalities on Eqs. (13), (14), (15), and (21). Finally, the tightness of the bounds is a direct consequence of the tightness of Eqs. (E1) and (E2). ■

APPENDIX F: PROOF OF PROPOSITION 2

Proof. Let us first prove the first statement. By considering $\boldsymbol{X}^n$ to be an n-bit copy, a direct calculation using Eqs. (13) and (14) shows that $\mathrm{TC}(\boldsymbol{X}^n) = n - 1$ and $\mathrm{DTC}(\boldsymbol{X}^n) = 1$, and therefore the upper bound is attained. To prove the converse, let us start by assuming that $\Omega(\boldsymbol{X}^n) = n - 2$. By applying (E2) to each term in (15), it is clear that $I(X_j; \boldsymbol{X}^{j-1}; \boldsymbol{X}^n_{j+1}) = 1$ holds for all $j \in \{1, \dots, n\}$. In particular $I(X_2; X_1; \boldsymbol{X}^n_3) = 1$ holds, which due to Eq. (15) implies that $I(X_2; X_1|\boldsymbol{X}^n_3) = 0$ and hence $I(X_2; X_1) = 1$, which in turns implies that X_1 and X_2 are Bernoulli distributed with parameter $p = 1/2$, and also that $X_1 = X_2$. By relabeling the variables and following the same rationale one can prove that all pairs of variables are equal to each other, which proves that $\boldsymbol{X}^n$ is an n-bit copy.

Let us prove the second statement. By considering now $\boldsymbol{X}^n$ to be a n-bit `xor`, using Eqs. (13) and (14) it is direct to check that $\mathrm{TC}(\boldsymbol{X}^n) = 1$ and $\mathrm{DTC}(\boldsymbol{X}^n) = n - 1$, and hence the lower bound is attained. To prove the converse, let us assume that $\boldsymbol{X}^n$ is such that $\Omega(\boldsymbol{X}^n) = 2 - n$. By considering the bounds given by Eq. (E2) in Eq. (15), this implies that $I(X_j; \boldsymbol{X}^{j-1}; \boldsymbol{X}^n_{j+1}) = -1$ for all $j \in \{2, \dots, n-1\}$, and in particular $I(\boldsymbol{X}^{n-2}; X_{n-1}; X_n) = -1$. Due to Eq. (E2), this implies in turn that $I(\boldsymbol{X}^{n-2}; X_{n-1}) = 0$, and via relabeling one can prove that $\boldsymbol{X}^{n-1}$ are jointly independent. Moreover, $I(\boldsymbol{X}^{n-2}; X_{n-1}; X_n) = -1$ also implies that $I(X_{n-1}; X_n|\boldsymbol{X}^{n-2}) = 1$, which implies that

$$I(\boldsymbol{X}^{n-1}; X_n) = I(X_{n-1}; X_n|\boldsymbol{X}^{n-2}) + I(\boldsymbol{X}^{n-2}; X_n) = 1.$$

This equality implies that X_n is Bernoulli distributed with $p = 1/2$, and that X_n is a deterministic function of $\boldsymbol{X}^{n-1}$. Moreover, the fact that $I(X_1; X_n|\boldsymbol{X}^{n-1}_2) = 1$ implies that, for given $\boldsymbol{X}^{n-1}_2$, X_n is a function of X_1, while via relabelling one finds that $I(X_1; X_n) = 0$. Since the only functions with these properties are functions isomorphic to an n-variate `xor`, this proves the desired result. ■

APPENDIX G: PROOF OF PROPOSITION 3

The following proof uses Lemma 6, which is stated and proved afterwards in this Appendix.

Proof. To prove Eq. (16), first note that

$$\Omega(\boldsymbol{X}^n) = \mathrm{TC}(\boldsymbol{X}^{n-1}) - \mathrm{DTC}(\boldsymbol{X}^{n-1}|X_n) \leqslant \mathrm{TC}(\boldsymbol{X}^{n-1}).$$

Then, the inequality follows from a direct application of Lemma 6. As $\mathrm{TC}(\boldsymbol{X}^m) \geqslant 0$, the equality becomes nontrivial when

$$\Omega(\boldsymbol{X}^n) - (n - m - 1)\log|\mathcal{X}| \geqslant 0.$$

To prove Eq. (17), note that by using Eqs. (13), (14), and (15) one can find that

$$\begin{aligned} \Omega(\boldsymbol{X}^n) &= \mathrm{TC}(\boldsymbol{X}^m) - \mathrm{DTC}\big(\boldsymbol{X}^m|\boldsymbol{X}^n_{m+1}\big) \\ &\quad + \sum_{j=m+1}^{n-1} I\big(X_j; \boldsymbol{X}^{j-1}; \boldsymbol{X}^n_{j+1}\big) \\ &\geqslant \mathrm{TC}(\boldsymbol{X}^m) - (n-2)\log|\mathcal{X}|. \end{aligned}$$

Above, the inequality is due to $I(X_j; \boldsymbol{X}^{j-1}; \boldsymbol{X}^n_{j+1}) \leqslant \log|\mathcal{X}|$ and $\mathrm{DTC}(\boldsymbol{X}^m|\boldsymbol{X}^n_{m+1}) \leqslant (m-1)\log|\mathcal{X}|$. As the above relationship does not depend on the labeling of the X's, this proves Eq. (17). As $\mathrm{TC}(\boldsymbol{X}^m) \leqslant (m-1)\log|\mathcal{X}|$, the equality becomes nontrivial when

$$\Omega(\boldsymbol{X}^n) + (n-2)\log|\mathcal{X}| \leqslant (m-1)\log|\mathcal{X}|.$$

■

Lemma 6. If $|\mathcal{X}| = \min_{i=1,\dots,n}|\mathcal{X}_i|$, then

$$\min_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^\gamma) \geqslant \mathrm{TC}(\boldsymbol{X}^n) - (n-m)\log|\mathcal{X}|.$$

Proof. A direct calculation using Eq. (13) shows that

$$\begin{aligned} \mathrm{TC}(\boldsymbol{X}^n) &= \mathrm{TC}(\boldsymbol{X}^m) + \sum_{j=m+1}^{n} I(X_j; \boldsymbol{X}^{j-1}) \\ &\leqslant \mathrm{TC}(\boldsymbol{X}^m) + (n-m)\log|\mathcal{X}|. \end{aligned}$$

As the labeling of the indices can be modified without changing this result, this suffices to prove the desired result. ■

[1] J. P. Crutchfield, The calculi of emergence, Physica D **75**, 11 (1994).

[2] E. Schneidman, W. Bialek, and M. J. Berry, Synergy, redundancy, and independence in population codes, J. Neurosci. **23**, 11539 (2003).

[3] P. E. Latham and S. Nirenberg, Synergy, redundancy, and independence in population codes, revisited, J. Neurosci. **25**, 5195 (2005).

[4] E. Ganmor, R. Segev, and E. Schneidman, Sparse low-order interaction network underlies a highly correlated and learnable

neural population code, Proc. Natl. Acad. Sci. USA **108**, 9679 (2011).

[5] G. Tononi, G. M. Edelman, and O. Sporns, Complexity and coherency: Integrating information in the brain, Trends Cognit. Sci. **2**, 474 (1998).

[6] D. Balduzzi and G. Tononi, Integrated information in discrete dynamical systems: Motivation and theoretical framework, PLoS Comput. Biol. **4**, e1000091 (2008).

[7] I. Gat and N. Tishby, Synergy and redundancy among brain cells of behaving monkeys, in *Advances in Neural Information Processing Systems* (MIT press, Denver, 1999), pp. 111–117.

[8] G. Chechik, A. Globerson, M. J. Anderson, E. D. Young, I. Nelken, and N. Tishby, Group redundancy measures reveal redundancy reduction in the auditory pathway, in *Advances in Neural Information Processing Systems* (MIT press, Vancouver, 2002), pp. 173–180.

[9] V. Varadan, D. M. Miller III, and D. Anastassiou, Computational inference of the molecular logic for synaptic connectivity in C. elegans, Bioinformatics **22**, e497 (2006).

[10] L. M. A. Bettencourt, V. Gintautas, and M. I. Ham, Identification of Functional Information Subgraphs in Complex Networks, Phys. Rev. Lett. **100**, 238701 (2008).

[11] S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo, Expanding the transfer entropy to identify information circuits in complex systems, Phys. Rev. E **86**, 066211 (2012).

[12] E. Schneidman, S. Still, M. J. Berry, and W. Bialek, Network Information and Connected Correlations, Phys. Rev. Lett. **91**, 238701 (2003).

[13] G. Tononi, O. Sporns, and G. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system, Proc. Natl. Acad. Sci. USA **91**, 5033 (1994).

[14] A. B. Barrett and A. K. Seth, Practical measures of integrated information for time-series data, PLoS Comput. Biol. **7**, e1001052 (2011).

[15] P. A. M. Mediano, A. K. Seth, and A. B. Barrett, Measuring integrated information: Comparison of candidate measures in theory and simulation, Entropy **21**, 17 (2018).

[16] An exception is the connected information, which can be elegantly derived from principles of information geometry [65] however, there are no known methods to compute this metric from data.

[17] P. L. Williams and R. D. Beer, Nonnegative decomposition of multivariate information, arXiv:1004.2515.

[18] V. Griffith and C. Koch, Quantifying synergistic mutual information, *Guided Self-Organization: Inception* (Springer, New York, 2014), pp. 159–190.

[19] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips, Partial information decomposition as a unified approach to the specification of neural goal functions, Brain and Cognition **112**, 25 (2017).

[20] A. B. Barrett, Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems, Phys. Rev. E **91**, 052802, (2015).

[21] R. A. Ince, Measuring multivariate redundant information with pointwise common change in surprisal, Entropy **19**, 318 (2017).

[22] R. James, J. Emenheiser, and J. Crutchfield, Unique information via dependency constraints, J. Phys. A, **52**, 014002 (2019).

[23] C. Finn and J. T. Lizier, Pointwise partial information decomposition using the specificity and ambiguity lattices, Entropy **20**, 297 (2018).

[24] T. M. Tax, P. A. M. Mediano, and M. Shanahan, The partial information decomposition of generative neural network models, Entropy **19**, 9 (2017).

[25] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, Quantifying information modification in developing neural networks via partial information decomposition, Entropy **19**, 9 (2017).

[26] L. Faes, D. Marinazzo, and S. Stramaglia, Multiscale information decomposition: exact computation for multivariate Gaussian processes, Entropy **19**, 408 (2017).

[27] A. El Gamal and Y.-H. Kim, *Network Information Theory* (Cambridge University, Cambridge, England, 2011).

[28] S. Watanabe, Information theoretical analysis of multivariate correlation, IBM J. Res. Dev. **4**, 66 (1960).

[29] T. S. Han, Linear dependence structure of the entropy space, Inform. and Control **29**, 337 (1975).

[30] Results about structural properties of systems with low values of total correlation or dual total correlation have been recently reported in Ref. [74].

[31] R. G. James, C. J. Ellison, and J. P. Crutchfield, Anatomy of a bit: Information in a time series observation, Chaos **21**, 037109 (2011).

[32] V. S. Vijayaraghavan, R. G. James, and J. P. Crutchfield, Anatomy of a spin: The information-theoretic structure of classical spin systems, Entropy **19**, 214 (2017).

[33] W. J. McGill, Multivariate information transmission, Psychometrika **19**, 97 (1954).

[34] H. K. Ting, On the amount of information, *Theory of Probability and its Applications* (SIAM, Pennsylvania, US, 1962), pp. 439–447.

[35] R. W. Yeung, A new outlook on Shannon's information measures, IEEE Trans. Inf. Theory **37**, 466 (1991).

[36] This distinction might not have been stressed in the past because most studies focus on bivariate interactions between two sets of variables, for which these two effects are equivalent and equal to the mutual information. However, for interactions involving three or more variables these perspectives differ.

[37] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University, Cambridge, England, 2003).

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 2012).

[39] For a quantum-mechanical treatment of this notion, see Chap. 2 of Ref. [75].

[40] L. Brillouin, The negentropy principle of information, J. Appl. Phys. **24**, 1152 (1953).

[41] This observation can be made rigorous via the Shannon-McMillan-Breiman theorem (see Sec. 3 of Ref. [38]).

[42] M. Studený and J. Vejnarová, The multiinformation function as a tool for measuring stochastic dependence, *Learning in Graphical Models* (Springer, New York, 1998), pp. 261–297.

[43] In fact, a direct calculation shows that the variables $\boldsymbol{X}^n$ are independent if and only if $\sum_j R_j = H(\boldsymbol{X}^n)$.

[44] S. A. Abdallah and M. D. Plumbley, A measure of statistical complexity based on predictive information with application to finite spin systems, Phys. Lett. A **376**, 275 (2012).

[45] S. Verdú and T. Weissman, Erasure entropy, *IEEE International Symposium on Information Theory* (IEEE, New York, 2006), pp. 98–102.

[46] S. Verdú and T. Weissman, The information lost in erasures, IEEE Trans. Inf. Theory **54**, 5030 (2008).

[47] E. Olbrich, N. Bertschinger, N. Ay, and J. Jost, How should complexity scale with system size? Eur. Phys. J. B **63**, 407 (2008).

[48] The interaction information is closely related to the I measures [35], the coinformation [76], and the multiscale complexity [77].

[49] A. A. Margolin, K. Wang, A. Califano, and I. Nemenman, Multivariate dependence and genetic networks inference, IET systems biology **4**, 428 (2010).

[50] I. Kontoyiannis and B. Lucena, Mutual information, synergy and some curious phenomena for simple channels, in *Proceedings of the International Symposium on Information Theory, 2005* (IEEE, New York, 2005), pp. 1651–1655.

[51] F. Rosas, V. Ntranos, C. J. Ellison, S. Pollin, and M. Verhelst, Understanding interdependency through complex information sharing, Entropy **18**, 38 (2016).

[52] F. Rosas, P. A. Mediano, M. Ugarte, and H. J. Jensen, An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems, Entropy **20**, 10 (2018).

[53] P. Baudot, M. Tapia, and J.-M. Goaillard, Topological information data analysis: Poincare-Shannon machine and statistical physics of finite heterogeneous systems, Preprints **2018**, 2018040157, doi: 10.20944/preprints201804.0157.v1.

[54] A lattice is a partially ordered set with a unique infimum and supremum. For more details on this construction, see Ref. [78].

[55] If $\pi_1, \pi_2 \in \mathcal{P}_n$ with $\pi_1 = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_r)$ and $\pi_2 = (\boldsymbol{\beta}_1|\ldots|\boldsymbol{\beta}_s)$, π_1 is *finer* than π_2 if for each $\boldsymbol{\alpha}_i$ exists $\boldsymbol{\beta}_k$ such that $\boldsymbol{\alpha}_i \subset \boldsymbol{\beta}_k$.

[56] It is direct to see that π_2 covers π_1 if and only if it is an elementary refinement, i.e., π_2 can be obtained from π_1 by dividing one cell of π_1 in two. Hence, if π_2 covers π_1 then $|\pi_2| = |\pi_1| + 1$, where $|\pi|$ is the number of (nonempty) cells of π.

[57] Put simply, there is an edge from π_1 to π_2 if π_2 results from taking π_1 and splitting one of its cells in two.

[58] It is direct to check that $\pi_b \succ \pi_a$ if and only if $\mathtt{P}(\pi_a, \pi_b) \neq \varnothing$. Moreover, all $\mathtt{p} \in \mathtt{P}(\pi_a, \pi_b)$ have the same length, given by $|\mathtt{p}| = ||\pi_b| - |\pi_a||$, where $|\mathtt{p}|$ is the number of edges in the path.

[59] In effect, $R_{\boldsymbol{\alpha}_k}$ represents the portion of the entropy of the kth cell that is not shared with other cells.

[60] The number of the nodes of $\mathcal{G}_n$ grows with the *Bell numbers*, known for their superexponential growth rate [79]. To find the number of paths in $\mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})$, note that if one starts from the sink and moves towards the source every step corresponds to merging two cells into one. Therefore, as selecting two out of m cells gives $\binom{m}{2}$ choices, the total number of paths is given by

$$|\mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})| = \sum_{m=2}^{n} \binom{m}{2} = \frac{n!(n-1)!}{2^{n-1}},$$

which grows faster than the Bell numbers.

[61] Interestingly, despite the correlation between X_1 and X_2, an n-bit `xor` still enables the most synergistic configuration attainable.

[62] R. G. James and J. P. Crutchfield, Multivariate dependence beyond Shannon information, Entropy **19**, 531 (2017).

[63] R. Cofré, C. Maldonado, and F. Rosas, Large deviations properties of maximum entropy Markov chains from spike trains, Entropy **20**, 573 (2018).

[64] For a detailed discussion of this issue for the case of three variables see Sec. 5 of Ref. [51].

[65] S.-I. Amari, Information geometry on hierarchy of probability distributions, IEEE Trans. Inf. Theory **47**, 1701 (2001).

[66] E. Olbrich, N. Bertschinger, and J. Rauh, Information decomposition and synergy, Entropy **17**, 3501 (2015).

[67] The quantity TC + DTC has been introduced in the context of time series analysis as *local exogenous information*, being characterized as a "very mutual information" [31].

[68] In Fig. 2 of Ref. [5] the DTC (under the name "interaction complexity") is proposed as a metric "related but not identical to neural complexity." Numerical evaluations show that the sum of TC and DTC, as proposed in (22), is a more accurate approximation for the TSE complexity (results not shown).

[69] http://kern.ccarh.org.

[70] http://web.mit.edu/music21.

[71] Regularization methods (such as Laplace smoothing) were found to have strong effects on the results. We decided not to use such methods, as some chords (e.g., C-C♯-D-D♯) are just not going to take place in the Baroque repertoire.

[72] R. Taruskin, *Music in the Early Twentieth Century: The Oxford History of Western Music* (Oxford University, New York, 2006).

[73] A. Pascual-Garcia, A constructive approach to the epistemological problem of emergence in complex systems, PLoS ONE **13**, e0206489 (2018).

[74] T. Austin, Measures of correlation and mixtures of product measures, arXiv:1809.10272.

[75] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University, New York, 2002).

[76] A. J. Bell, The co-information lattice, in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation* (Springer, Berlin, Heidelberg, 2003).

[77] Y. Bar-Yam, Multiscale complexity/entropy, Adv. Comp. Systems **7**, 47 (2004).

[78] R. P. Stanley, *Enumerative Combinatorics*, Cambridge Studies in Advanced Mathematics Vol. 1 (Cambridge University, Cambridge, England, 2012).

[79] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions* (Springer, New York, 2012).