Article

Topological Information Data Analysis: Poincare-Shannon Machine and statistical physic of finite heterogeneous systems

Pierre Baudot [1,†*], Monica Tapia [2,†] and Jean-Marc Goaillard [3,†]

[1] Inserm UNIS UMR1072 - Université Aix-Marseille; pierre.baudot@gmail.com
[2] Inserm UNIS UMR1072 - Université Aix-Marseille; monica.tapia-pacheco@univ-amu.fr
[3] Inserm UNIS UMR1072 - Université Aix-Marseille; Jean-marc.goaillard@univ-amu.fr
[*] Correspondence: pierre.baudot@gmail.com; Tel.: +33-4-91-69-89-54
[†] Current address: Inserm UMRS 1072, Faculté de Médecine - Secteur Nord Université Aix Marseille, 51, Boulevard Pierre Dramard, 13015 Marseille, France

Received: date; Accepted: date; Published: date

Abstract: This paper establishes methods that quantify the structure of statistical interactions within a given data set using the characterization of information theory in cohomology by finite methods, and provides their expression in terms of statistical physic and machine learning. Following [1–3], we show directly that k multivariate mutual-informations (I_k) are k-coboundaries. The k-cocycles are given by $I_k = 0$, which generalize statistical independence to arbitrary dimension k. The topological approach allows to investigate Shannon's information in the multivariate case without the assumptions of independent identically distributed variables. We develop the computationally tractable subcase of simplicial information cohomology represented by entropy H_k and information I_k landscapes. The I_1 component defines a self-internal energy functional U_k, and $(-1)^k I_{k,k\geq 2}$ components define the contribution to a free energy functional G_k of the k-body interactions. The set of information paths in simplicial structures is in bijection with the symmetric group and random processes, provides a topological expression of the 2nd law and points toward a discrete Noether theorem (1st law). The local minima of free-energy, related to conditional information negativity and the non-Shannonian cone of Yeung [4], characterize a minimum free energy complex. This complex formalizes the minimum free-energy principle in topology, provides a definition of a complex system, and characterizes a multiplicity of local minima that quantifies the diversity observed in biology. Finite data size effects and estimation bias severely constrain the effective computation of the information topology on data, and we provide simple statistical tests for the undersampling bias and for the k-dependences following [5]. We give an example of application of these methods to genetic expression and cell-type classification. The maximal positive I_k identifies the variables that co-vary the most in the population, whereas the minimal negative I_k identifies clusters and the variables that differentiate-segregate the most. The methods unravel biologically relevant I_{10} with a sample size of 41. It establishes generic methods to quantify the epigenetic information storage and a unified epigenetic unsupervised learning formalism.

Keywords: Information theory; Cohomology; Algebraic Topology; Topological Data Analysis; Genetic Expression; Epigenetics; Machine Learning; Statistical Physic; Multivariate Mutual-Information; Complex Systems; Biodiversity

> *"When you use the word information, you should rather use the word form"*
>
> René Thom

Contents

0. Introduction

This paper is based on the information cohomology framework developed in [1–3], and relies on a theorem establishing uniquely the usual entropy and multi-variate mutual-information (I_k) as the first class and co-boundaries of a Hochschild cohomology theory, respectively, with finite (non-asymptotic) methods (except the logarithm). In a first part, we establish the coboundary nature of mutual-informations more directly by computing the cohomology in higher degrees and use the Hodge decomposition of Gerstenhaber and Shack [6] to encompass odd and even I_k coboundaries in a single (hyper) cohomology. It allows to generalize statistical independence to arbitrary degrees k such that the "0" of k-mutual-information I_k coincides with k-cocycle condition. Altogether, this first part theoretically justifies the functions estimated in the data analysis, and consistently settles their statistical, mathematical nature and interpretations used in the article.

In a second part, we develop the computationally tractable subcases of simplicial information cohomology. To represent such a structure, we define entropy (H_k) and information (I_k) landscapes, which are the basis of the data analysis. They consist in an exhaustive computation and representation of all the information elements of the semi-lattice of subsets. We define the I_1 component as a (isotherm) self-internal energy functional U_k and $(-1)^k I_{k,k\geq 2}$ components as the contribution to a free energy functional G_k of the k-body interactions. In the special case of Gibbs distributions (exponential family), this allows to recover the fundamental equation of thermodynamic, $TH_k = U_k - G_k$. It gives a topological expression (free of any metric assumption) of the free energy functional in purely informational terms that holds for arbitrary empirical data and small systems (non asymptotic and with variables not necessarily identically distributed). We then settle the space of paths on the simplicial information structure (or landscape), relating the existence of local minima to information inequalities and conditional mutual-information negativity. In simplicial structures, the set of information paths is in bijection with the symmetric group and random processes, allowing to derive a topological formulation of the 2nd law of thermodynamic, generalizing Cover's theorem [7]. These minima allow to define the maximal (length) chains that characterize a minimum free energy complex. Hence, in simple terms, our approach implements and formalizes the minimum energy principle on a topological ground, allowing to characterize multiple local minima, such as those arising in complex-frustrated systems [8,9]. Altogether, this second part provides the machine-learning and statistical physic interpretation of the cohomology and establishes a topological version of Boltzmann and Helmholtz machines, which would therefore deserve the name of Poincaré-Shannon machine. Notably, such machine replaces the usual geodesic approach in gradient descent by the homotopical consideration of information paths.

In a third part, we present the application of the methods to genetic expression, which consists in two dual unsupervised learning tasks: cell type and gene module detection. The k-tuples with maximal positive I_k identify the variables that co-vary the most in the population, whereas the k-tuples with minimal negative I_k identify clusters and the variables that differentiate-segregate the most the population. The partial estimation of the minimum free energy complex identifies relevant k-tuples up to the dimension 10 for cell modules, and up to the dimension 6 for gene modules, despite a small sample size of $m = 41$ and $m = 111$, respectively. Notably, the algorithm retrieves the two cell types in the data with low errors. It establishes that statistical dependences and multivariate I_k beyond pairwise and even in high dimension are biologically relevant and functional. We discuss the finite, discrete and "naive" methods allowing to accurately estimate dependences in small sample size with respect to previous MaxEnt studies, notably the work of Margolin and colleagues [10]. We conclude that information topology provides i) a mathematical unified model of epigenetic learning allowing to quantify the information storage and structure in cell and gene assemblies, and ii) a formalization and quantification of complex systems beyond complex networks.

The materials and methods are dedicated to the effective computation of information topology on data, focused primarily on the investigation of finite data size effects and estimation bias. We establish a simple method to evaluate the undersampling regime by computing the dimension at which the estimation of informations becomes too biased. We also provide a test of k-dependence generalizing the exact test of Pethel and Hahs [5], allowing to assess the significance of the estimated I_k values. We provide a tractable algorithm approximating the free energy complex.

This presentation underlines the important limits of the methods: methods and results exposed here are partial, because i) for computational reasons we restricted the data analysis to tractable simplicial information (up to dimension 21) and information path subcases, ii) we employed a basic procedure of probability estimation of the data that can also mask some of the statistical relationships, iii) of the small sample size used. This article is almost self-content from the theory to the application (at the exception of finite probability foundations that can be found in [11], and the theorem 1 in [1–3]). An important part of the developments has no claim for novelty but aims to underline the convergence of different formalisms, by deriving known results notably in information theory from a topological point of view (with less assumptions). We believe that the highlighted precisions, references, open questions and limitations will help on the long term the researches on the topic. As the topics of the article spans several domains, we provide some of the bibliographical references and problematics within separated introductory paragraphs within each chapter.

1. Information Cohomology

1.1. Information functions (definitions)

The information functions used in [1] and the present study were originally defined by Shannon [12] and Kullback [13] and further generalized and developed by Hu Kuo Ting [14] and Yeung [4] (see also McGill [15]). These functions include entropy, noted $H_1 = H(X;P)$, joint entropy, noted $H_k = H(X_1, ..., X_k; P)$, mutual-information noted $I_2 = I(X_1; X_2; P)$, multivariate k-mutual-information, noted $I_k = I(X_1; ...; X_k; P)$ and the conditional entropy and mutual information, noted $Y.H_k = H(X_1, ..., X_k|Y; P)$ and $Y.I_k = I(X_1; ...; X_k|Y; P)$. The classical expression of these functions is the following (using $k = -1/\ln 2$, the usual bit unit):

- The Shannon-Gibbs entropy of a single variable X_j is defined by [12]:

$$H_1 = H(X_j; P_{X_j}) = k \sum_{x \in [N_j]} p(x) \ln p(x) = k \sum_{i=1}^{N_j} p_i \ln p_i \tag{1}$$

  where $[N_j] = \{1, ..., N_j\}$ denotes the alphabet of X_j.
- The relative entropy or Kullback-Liebler divergence, which was also called "discrimination information" by Kullback [13], is defined for two probability mass function $p(x)$ and $q(x)$ by:

$$\begin{aligned} D(p(x)||q(x)) = D(X; p(x)||q(x)) &= k \sum_{x \in \mathscr{X}} p(x) \ln \frac{q(x)}{p(x)} \\ &= H(X; p(x), q(x)) - H(X; p(x)) \end{aligned} \tag{2}$$

  where $H(X; p(x), q(x))$ is the cross-entropy and $H(X; p(x))$ the Shannon entropy. It hence generates as a special case minus entropy, taking the deterministic constant probability $q(x) = 1$. With the convention $k = -1/\ln 2$, $D(p(x)||q(x))$ is always positive or null.

- The joint entropy is defined for any joint-product of k random variables $(X_1,...,X_k)$ and for a probability joint-distribution $\mathbb{P}_{(X_1,...,X_k)}$ by [12]:

$$\begin{aligned} H_k &= H(X_1,...,X_k;P_{X_1,...,X_k}) \\ &= k\sum_{x_1,...,x_k\in[N_1\times...\times N_k]}^{N_1\times...\times N_k} p(x_1.....x_k)\ln p(x_1.....x_k) \\ &= k\sum_{i,j,...,k}^{N_1,...,N_k} p_{\underbrace{ij...k}_{\text{k indices}}} \ln p_{ij...k} \end{aligned} \tag{3}$$

where $[N_1\times...\times N_k] = \{1,...,N_j\times...\times N_k\}$ denotes the alphabet of $(X_1,...,X_k)$.
- The mutual information of two variables X_1, X_2 is defined as [12]:

$$I(X_1;X_2;P_{X_1,X_2}) = k\sum_{x_1,x_2\in[N_1\times N_2]}^{N_1\times N_2} p(x_1.x_2)\ln\frac{p(x_1)p(x_2)}{p(x_1.x_2)} \tag{4}$$

And it can be generalized to k-mutual-information (also called co-information) using the alternated sums given by equation 17, as originally defined by McGill [15] and Hu Kuo Ting [14], giving:

$$I_k = I(X_1;...;X_k;P) = k\sum_{x_1,...,x_k\in[N_1\times...\times N_k]}^{N_1\times...\times N_k} p(x_1.....x_k)\ln\frac{\prod_{I\subset[k];card(I)=i;i \text{ odd}} p_I}{\prod_{I\subset[k];card(I)=i;i \text{ even}} p_I} \tag{5}$$

For example, the 3-mutual information is the function:

$$I_3 = k\sum_{x_1,x_2,x_3\in[N_1\times N_2\times N_3]}^{N_1\times N_2\times N_3} p(x_1.x_2.x_3)\ln\frac{p(x_1)p(x_2)p(x_3)p(x_1.x_2.x_3)}{p(x_1.x_2)p(x_1.x_3)p(x_2.x_3)} \tag{6}$$

For $k\geq 3$, I_k can be negative [14].
- The total correlation introduced by Watanabe [16], called integration by Tononi and Edelman [17] or multi-information by Studený and Vejnarova [18] and Margolin and colleagues [10], which we note $C_k(X_1;...X_k;P)$, is defined by:

$$\begin{aligned} C_k = C_k(X_1;...X_k;P) = \sum_{i=1}^{k} H(X_i) - H(X_1;...X_k) = \sum_{i=2}^{k}(-1)^i \sum_{I\subset[n];card(I)=i} I_i(X_I;P) \\ = k\sum_{x_1,...,x_k\in[N_1\times...\times N_k]}^{N_1\times...\times N_k} p(x_1....x_k)\ln\frac{p(x_1...x_k)}{p(x_1)...p(x_k)} \end{aligned} \tag{7}$$

For two variables the total correlation is equal to the mutual-information ($C_2 = I_2$). The total correlation has the nice property of being a relative entropy 2 between marginal and joint-variable and hence to be always non-negative.
- The conditional entropy of X_1 knowing (or given) X_2 is defined as [12]:

$$\begin{aligned} X_2.H_1 = H(X_1|X_2;P) = k\sum_{x_1,x_2\in[N_1\times N_2]}^{N_1*N_2} p(x_1.x_2)\ln p_{x_2}(x_1) \\ = k\sum_{x_2\in\mathscr{X}_2}^{N_2} p(x_2).\left(\sum_{x_1\in\mathscr{X}_1}^{N_1} p_{x_2}x_1\ln p_{x_2}x_1\right) \end{aligned} \tag{8}$$

doi:10.20944/preprints201804.0157.v1

Conditional joint-entropy, $X_3.H(X_1, X_2)$ or $(X_1, X_2).H(X_3)$, is defined analogously by replacing the marginal probabilities by the joint probabilities.

- The conditional mutual information of two variables X_1, X_2 knowing a third X_3 is defined as [12]:

$$X_3.I_2 = I(X_1; X_2|X_3; P) = k \sum_{x_1,x_2,x_3 \in [N_1 \times N_2 \times N_3]}^{N_1 \times N_2 \times N_3} p(x_1.x_2.x_3) \ln \frac{p_{x_3}(x_1)p_{x_3}(x_2)}{p_{x_3}(x_1, x_2)} \tag{9}$$

Conditional mutual information generates all the preceding information functions as subcases, as shown by Yeung [4]. We have the theorem : if $X_3 = \Omega$ then it gives the mutual information, if $X_2 = X_1$ it gives conditional entropy, and if both conditions are satisfied, it gives entropy. Notably, we have $I_1 = H_1$.

We now give the few information equalities and inequalities that are of central use in the homological framework, in the information diagrams and for the estimation of the informations from the data.
We have the chain rules (see [19] for proofs):

$$H(X_1; X_2; P) = H(X_1; P) + X_1.H(X_2; P) = H(X_2; P) + X_2.H(X_1; P) \tag{10}$$

$$I(X_1; X_2; P) = H(X_1; P) - X_2.H(X_1; P) = H(X_2; P) - X_1.H(X_2; P) \tag{11}$$

That we can write more generally (where the hat denotes the omission of the variable):

$$H(X_1; ...; \widehat{X_i}; ...; X_{k+1}; P) = H(X_1; ...; X_{k+1}; P) - (X_1; ...; \widehat{X_i}; ...; X_{k+1}).H(X_i; P) \tag{12}$$

That we can write in short $H_{k+1} - H_k = (X_1, ...X_k).H(X_{k+1})$

$$I(X_1; ...; \widehat{X_i}; ...; X_{k+1}; P) = I(X_1; ...; X_{k+1}; P) + X_i.I(X_1; ...; \widehat{X_i}; ...; X_{k+1}; P) \tag{13}$$

That we can write in short $I_{k-1} - I_k = X_k.I_{k-1}$, generating the chain rule 10 as special case.

These two equations provide recurrence relationships that give an alternative formulation of the chain rules in terms of a chosen path on the lattice of information structures:

$$H_k = H(X_1, ..., X_k; P) = \sum_{i=1}^{k} (X_1, ..., X_{i-1}).H(X_i; P) \tag{14}$$

where we assume $H(X_1; P) = X_0.H(X_1; P)$ and hence that X_0 is the greatest element $X_0 = \Omega$.

$$I_k = I(X_1; ...; X_k; P) = I(X_1) - \sum_{i=2}^{k} X_i.I(X_1; ...; X_{i-1}) \tag{15}$$

We have the alternated sums or inclusion-exclusion rules [1,14,20]:

$$H_n(X_1, ..., X_n; P) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{I \subset [n]; card(I)=i} I_i(X_I; P) \tag{16}$$

$$I_n(X_1; ...; X_n; P) = \sum_{i=1}^{n} (-1)^{i-1} \sum_{I \subset [n]; card(I)=i} H_i(X_I; P) \tag{17}$$

For example: $H_3(X_1, X_2, X_3) = I_1(X_1) + I_1(X_2) + I_1(X_3) - I_2(X_1; X_2) - I_2(X_1; X_3) - I_2(X_2; X_3) + I_3(X_1; X_2; X_3)$

The chain rule of mutual-information goes together with the following inequalities discovered by Matsuda [20]. For all random variables $X_1; ..; X_k$ with associated joint probability distribution P we have:

- $X_k.I(X_1;..;X_{k-1};P) \geq 0$ if and only if $I(X_1;..;X_{k-1};P) \geq I(X_1;..;X_k;P)$ (in short: $I_{k-1} \geq I_k$)
- $X_k.I(X_1;..;X_{k-1};P) < 0$ if and only if $I(X_1;..;X_{k-1};P) < I(X_1;..;X_k;P)$ (in short: $I_{k-1} < I_k$)

that fully characterize the phenomenon of information negativity as an increasing or diverging sequence of mutual information.

1.2. A long march through information topology

From the mathematical point of view, a motivation of information topology is to capture the ambiguity theory of Galois, which is the essence of group theory or discrete symmetries (see André's reviews [21,22]), and Shannon's information uncertainty theory in a common framework, a path already paved by some results on information inequalities (see Yeung's results [23]) and in algebraic geometry. In the work of Cathelineau, [24], entropy first appeared in the computation of the degree one homology of the group $SL(2,\mathbb{C})$ with coefficients in the adjoint action by choosing a pertinent definition of the derivative of the Bloch-Wigner dilogarithm. It could be shown that the functional equation with 5-terms of the dilogarithm implies the functional equation of entropy with 4-terms. Kontsevitch [25] discovered that a finite truncated version of the logarithm appearing in cyclotomic studies also satisfied the functional equation of entropy, suggesting a higher degree generalization of information, analog to polylogarithm, and hence showing that the functional equation of entropy holds in p and 0 field characteristics. Elbaz-Vincent and Gangl used algebraic means to construct this information generalization which holds over finite fields [26], and where information functions appear as derivations [27]. After entropy appeared in tropical and idempotent semi-ring analysis in the study of the extension of Witt semiring to the characteristic 1 limit [28], Marcolli and Thorngren developed thermodynamic semiring, and entropy operad that could be constructed as deformation of the tropical semiring [29]. Introducing Rota-Baxter algebras, it allowed to derive a renormalization procedure [30]. Baez, Fritz and Leinster defining the category of finite probability and using Fadeev axiomatization, could show that the only family of functions that has the functorial property is Shannon information loss [31,32]. Boyom, basing his approach on information and Koszul geometry, developed a more geometrical view of statistical models that notably considers foliations in place of the random variables [33]. Introducing a deformation theoretic framework, and chain complex of random variables, Drumond-Cole, Park and Terilla [34–36] could construct a homotopy probability theory for which the cumulants coincide with the morphisms of the homotopy algebras. A probabilistic framework, used here, was introduced in [1], and generalized to Tsallis entropies by Vigneaux [3]. The diversity of the formalisms employed in these independent but convergent approaches is astonishing. So, to the question what is information topology, it is only possible to answer that it is under development at the moment. The results of Catelineau, Elbaz-Vincent and Gangl inscribed information into the theory of motives, which according to Beilison's program is a mixed Hodge-Tate cohomology [37]. All along the development of the application to data, following the cohomology developed by [1,3] on an explicit probabilistic basis, we aimed to preserve such a structure and unravel its expression in information theoretic terms. Moreover, following Aomoto's results [38,39], the actual conjecture [1] is that the higher classes of information cohomology should be some kind of polylogarithmic k-form (k-differential volume that are symmetric and additive, and that correspond to the cocycle conditions for the cohomology of Lie groups [38]). The following developments suggest that these higher information groups should be the families of functions satisfying the functional equations of k-independence $I_k = 0$, a rather vague but intuitive view that can be tested in special cases.

1.3. Information structures and coboundaries

This section justifies the choice of functions and algorithm, the topological nature of the data analysis and the approximations we had to concede for computation. We refer to [1,3] for a detailed description of the construction and generalizations.

The characterization of information functions such as entropy and k-mutual-information by homological means underlines their mathematical foundation and universality, and justifies that they constitute pertinent candidates to estimate the structure of data. In the general formulation of information cohomology, the random variables are partitions of the atomic probabilities of a finite probability space $(\Omega, \mathcal{B}, P)$ (e.g. all their equivalence classes). The **Joint-Variable** (X_1, X_2) is the less fine partition that is finer than X_1 and X_2; the whole lattice of partitions Π [40] corresponds to the lattice of joint random variables [1,41]. Then, a general **information structure** is defined to be the triple (Ω, Π, P). A more modern and general expression in category theory and topos is given in [1,3]. $(X_1, ..., X_k; P)$ designates the image law of the probability P by the measurable function of joint variables $(X_1, ..., X_k)$. Figure 1 gives a simple example of the lattice of partitions for 4 atomic probabilities, with the simplicial sublattice used for data analysis. Atomic probabilities are also illustrated in Figure 14.

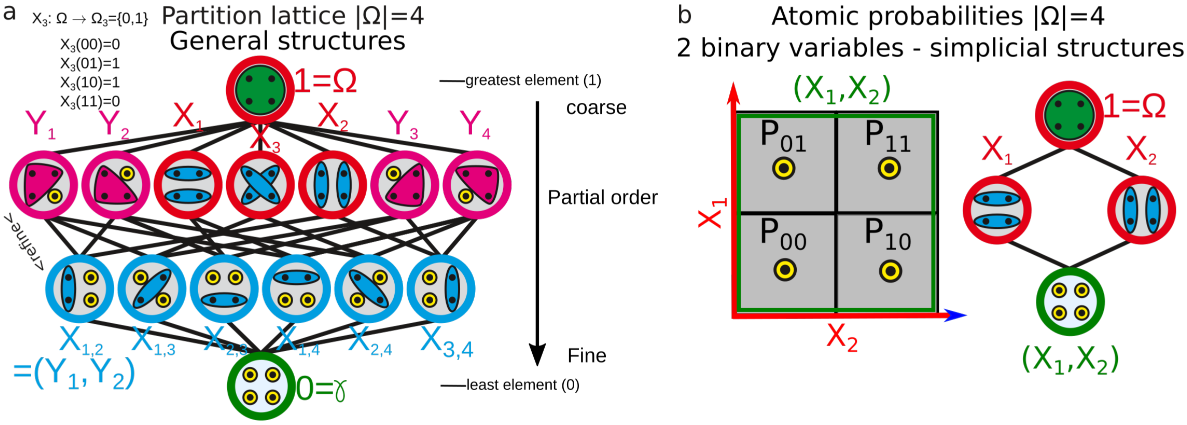

Figure 1. Example of general and simplicial information structures. a, Example of lattice of random variables (partitions): the lattice of partitions of atomic-elementary events for a sample space of 4 atomic elements $|\Omega| = 4$ (for example two coins and $\Omega = \{00, 01, 10, 11\}$), each element being denoted by a black dot in the circles representing the random variables. The joint operation of Random Variables noted (X, Y) or $X \otimes Y$ of two partitions is the less fine partition Z that is finer than X and Y (Z divides Y and X, or Z is the greatest common divisor of Y and X). It is represented by the coincidence of two edges of the lattices. The joint operation has an identity element noted $1 = \Omega$ (that we will note 0 thereafter), with $X, 1 = X, \Omega = X$ and is idempotent $(X, X) = X^2 = X$. The structure is a partially ordered set (poset) with a refinement relation. **b,** Illustration of the simplicial structure (sublattice) used for the data analysis ($|\Omega| = 4$ as previously).

On this general information structure, we consider the real module of all measurable functions $F(X_1, ..., X_k; P)$, and the conditioning-expectation by Y of measurable functions as the action of Y on the functional module, noted $Y.F(X_1, ..., X_k; P)$, such that it corresponds to the usual definition of conditional entropy (equ. 8). We define our complexes of measurable functions of random variables $X^k = F(X_1, ..., X_k; P)$, and the cochain complexes (X^k, ∂^k) as :

$$0 \rightarrow X^0 \xrightarrow{\partial^0} X^1 \xrightarrow{\partial^1} X^2 \xrightarrow{\partial^2} ...X^{k-1} \xrightarrow{\partial^{k-1}} X^k$$

where ∂^k is the left action co-boundary that Hochschild proposed for associative and ring structures [42]. A similar construction of random variable complex was given by Drumond-Cole, Park and Terilla [34,35]. We consider also the two other directly related cohomologies that are defined by considering a trivial left action [1] and a symmetric (left and right) action [6,43,44] of conditioning:

- The left action Hochschild-information coboundary and cohomology (with trivial right action, left panel of Figure 2):

$$\begin{aligned}(\partial^k)F(X_1;X_2;...;X_{k+1};P) = {} & X_1.F(X_2;...;X_{k+1};P) \\ & + \sum_{i=1}^{k}(-1)^i F(X_1;X_2;...;(X_i,X_{i+1});...;X_{k+1};P) \\ & + (-1)^{k+1}F(X_1;...;X_k;P)\end{aligned} \tag{18}$$

  This coboundary, with a trivial right action, is the usual coboundary of Galois cohomology ([45], p.2), and in general it is the coboundary of homological algebra obtained by Cartan and Eilenberg [46] and MacLane [47] (non homogenous bar complex).
- The "topological-trivial" Hochschild-information coboundary and cohomology: considering a trivial left action in the preceding setting , e.g. $X_1.F(X_2;...;X_{k+1}) = F(X_2;...;X_{k+1})$. It is the subset of the preceding case, which is invariant under the action of conditioning. We obtain the topological coboundary (∂_t^k) [1]:

$$\begin{aligned}(\partial_t^k)F(X_1;X_2;...;X_{k+1};P) = {} & F(X_2;...;X_{k+1};P) \\ & + \sum_{i=1}^{k}(-1)^i F(X_1;X_2;...;(X_i,X_{i+1});...;X_{k+1};P) \\ & + (-1)^{k+1}F(X_1;...;X_k;P)\end{aligned} \tag{19}$$

- The symmetric Hochschild-information coboundary and cohomology: as introduced by Gerstenhaber and Shack [6], Kassel [44] (p.13) and Weibel [43] (chap.9), we consider a symmetric (left and right) action of conditioning, that is $X_1.F(X_2;...;X_{k+1}) = F(X_2;...;X_{k+1}).X_1$. The left action module is essentially the same as considering a symmetric action bimodule [6,43,44]. We hence obtain the following symmetric coboundary (∂_*^k):

$$\begin{aligned}(\partial_*^k)F(X_1;X_2;...;X_{k+1};P) = {} & X_1.F(X_2;...;X_{k+1};P) \\ & + \sum_{i=1}^{k}(-1)^i F(X_1;X_2;...;(X_i,X_{i+1});...;X_{k+1};P) \\ & + (-1)^{k+1}X_{k+1}.F(X_1;...;X_k;P)\end{aligned} \tag{20}$$

Based on these definitions, Baudot and Bennequin [1] computed the first homology class in the left action Hochschild-information cohomology case and the coboundaries in higher degrees. We introduce here the symmetric case, and detail the higher degree cases by direct specialization of the co-boundaries formulas, such that it appears that information functions and chain rules are homological by nature. The main results are summarized in Figure 2.

For notation clarity, we omit the probability in the writing of the functions, and when specifically stated replace their notation F by their usual corresponding informational function notation H, I.

1.3.1. first degree (k=1)

For the first degree $k = 1$, we have the following results:

- The left 1-co-boundary is $(\partial^1)F(X_1;X_2) = X_1.F(X_2) - F(X_1,X_2) + F(X_1)$. The 1-cocycle condition $(\partial^1)F(X_1;X_2) = 0$ gives $F(X_1,X_2) = F(X_1) + X_1.F(X_2)$, which is the chain rule of information shown in equation 10. Then, following Kendall [48] and Lee [49], it is possible to recover the functional equation of information and to characterize uniquely, up to the arbitrary multiplicative constant k, the entropy (equation 1) as the first class of cohomology [1,3]. This main theorem allows us to obtain the other information functions in what follows. Marcolli and Thorngren [29], Leinster, Fritz and Baez [31,32] obtained independently an analog result using

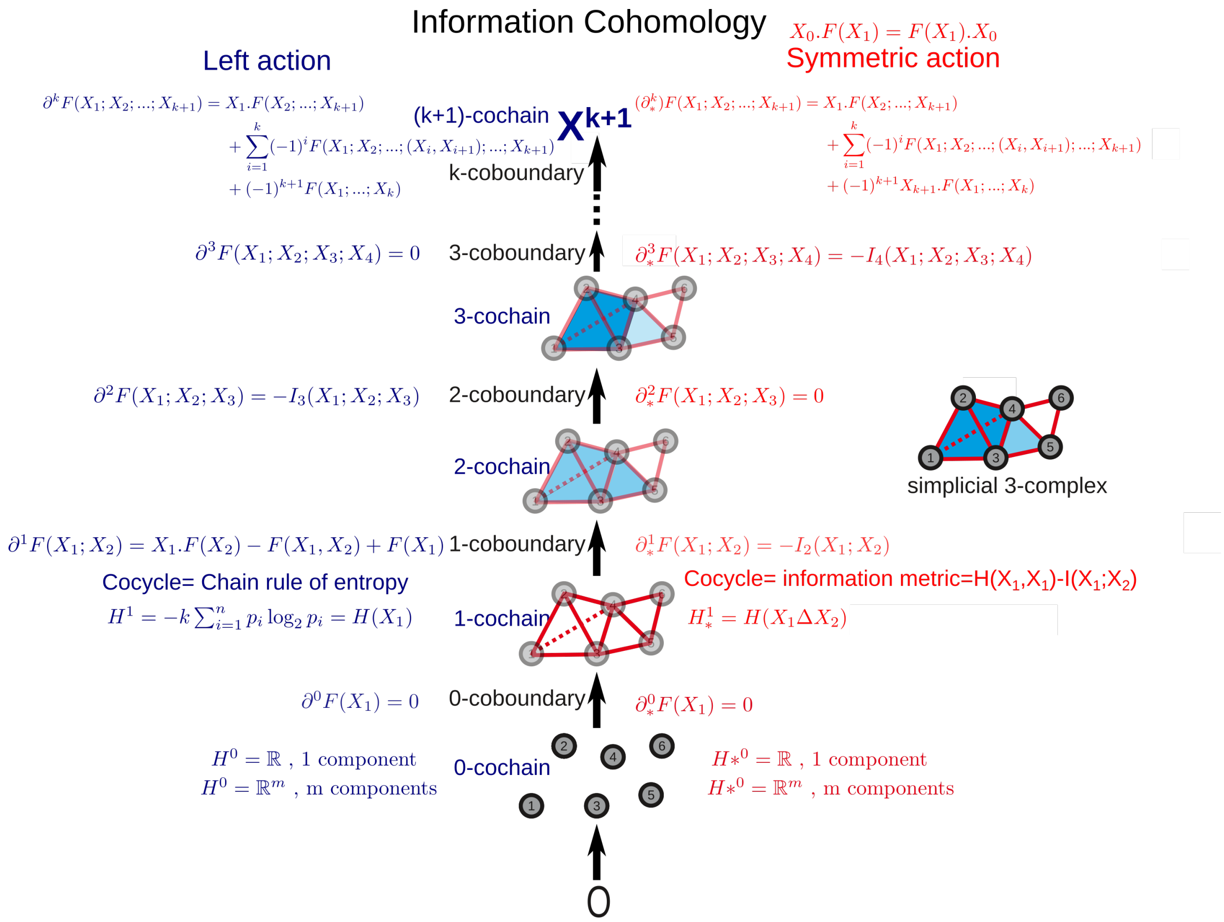
Information Cohomology
$X_0.F(X_1) = F(X_1).X_0$
Left action
Symmetric action
$\partial^k F(X_1;X_2;...;X_{k+1}) = X_1.F(X_2;...;X_{k+1}) + \sum_{i=1}^{k}(-1)^i F(X_1;X_2;...;(X_i,X_{i+1});...;X_{k+1}) + (-1)^{k+1}F(X_1;...;X_k)$
(k+1)-cochain
X^{k+1}
k-coboundary
$(\partial_*^k)F(X_1;X_2;...;X_{k+1}) = X_1.F(X_2;...;X_{k+1}) + \sum_{i=1}^{k}(-1)^i F(X_1;X_2;...;(X_i,X_{i+1});...;X_{k+1}) + (-1)^{k+1}X_{k+1}.F(X_1;...;X_k)$
$\partial^3 F(X_1;X_2;X_3;X_4) = 0$
3-coboundary
$\partial_*^3 F(X_1;X_2;X_3;X_4) = -I_4(X_1;X_2;X_3;X_4)$
3-cochain
$\partial^2 F(X_1;X_2;X_3) = -I_3(X_1;X_2;X_3)$
2-coboundary
$\partial_*^2 F(X_1;X_2;X_3) = 0$
2-cochain
simplicial 3-complex
$\partial^1 F(X_1;X_2) = X_1.F(X_2) - F(X_1,X_2) + F(X_1)$
1-coboundary
$\partial_*^1 F(X_1;X_2) = -I_2(X_1;X_2)$
Cocycle= Chain rule of entropy
$H^1 = -k\sum_{i=1}^n p_i \log_2 p_i = H(X_1)$
1-cochain
Cocycle= information metric=H(X1,X1)-I(X1;X2)
$H_*^1 = H(X_1 \Delta X_2)$
$\partial^0 F(X_1) = 0$
0-coboundary
$\partial_*^0 F(X_1) = 0$
$H^0 = \mathbb{R}$, 1 component
$H^0 = \mathbb{R}^m$, m components
0-cochain
$H*^0 = \mathbb{R}$, 1 component
$H*^0 = \mathbb{R}^m$, m components
1
2
3
4
5
6
0

Figure 2. Coboundaries and Cocycles of the left and symmetric action information cohomologies. illustration summarizing the important theorems of the left (left, blue) and symmetric action (right, red) information cohomologies. For illustration, in the center is depicted a sequence of simplicial information complex, a cochain complex of information introduced in next chapter, the corresponding simplicial complex is drawn on the right. Here, we will only consider the case $H^0 = \mathbb{R}$, which corresponds to 1 connected component.

measure-preserving function and characteristic one Witt construction, respectively. In these various theoretical settings, this result extends to relative entropy [1,29,32], and Tsallis entropies [3,29].

- The topological 1-coboundary (∂_t^1) is $(\partial_t^1)F(X_1;X_2) = F(X_2) - F(X_1,X_2) + F(X_1)$, which corresponds to the definition of mutual information $(\partial_t^1)F(X_1;X_2) = I(X_1;X_2) = H(X_1) + H(X_2) - H(X_1,X_2)$, and hence I_2 is a topological 1-coboundary.
- The symmetric 1-coboundary (∂_*^1) is $(\partial_*^1)F(X_1;X_2) = X_1.F(X_2) - F(X_1,X_2) + X_2.F(X_1)$, which corresponds to the negative of the pairwise mutual information $(\partial_*^1)F(X_1;X_2) = X_2.H(X_1) + X_1.H(X_2) - H(X_1,X_2) = -I(X_1;X_2)$, and hence $-I_2$ is a symmetric 1-coboundary. Moreover, the 1-cocycle condition $(\partial_*^1)F(X_1;X_2) = 0$ characterizes functions satisfying $F(X_1,X_2) = X_2.F(X_1) + X_1.F(X_2)$, which corresponds to the information pseudo-metric discovered by Shannon [50], Rajski [51], Zurek [52] and Bennett [53], and has further been applied for hierarchical clustering and finding categories in data by Kraskov and Grassberger [54]: $H(X_1 \triangle X_2) = X_2.H(X_1) + X_1.H(X_2) = H(X_1,X_2) - I(X_1;X_2)$. Therefore, up to an arbitrary scalar multiplicative constant k, the information pseudo-metric $H(X_1 \triangle X_2)$ is the first class of symmetric cohomology. This pseudo metric is represented in the Figure . It generalizes to

pseudo k-volumes that we define by $V_k = H_k - I_k$ (particularly interesting symmetric functions computed by the provided software).

1.3.2. Second degree (k=2)

For the second degree $k = 2$, we have the following results:

- The left 2-co-boundary is $\partial^2 F(X_1;X_2;X_3) = X_1.F(X_2;X_3) - F((X_1,X_2);X_3) + F(X_1;(X_2,X_3)) - F(X_1;X_2)$, which corresponds to minus the 3-mutual information $\partial^2 F(X_1;X_2;X_3) = X_1.I(X_2;X_3) - I((X_1,X_2);X_3) + I(X_1;(X_2,X_3)) - I(X_1;X_2) = -I(X_1;X_2;X_3)$, and hence $-I_3$ is left 2-coboundary.
- The topological 2-coboundary is $(\partial_t^2)F(X_1;X_2;X_3) = F(X_2;X_3) - F((X_1,X_2);X_3) + F(X_1;(X_2,X_3)) - F(X_1;X_2)$, which corresponds in information to $\partial_t^2 F(X_1;X_2;X_3) = I(X_2;X_3) - I((X_1,X_2);X_3) + I(X_1;(X_2,X_3)) - I(X_1;X_2) = 0$, and hence the topological 2-coboundary is always null-trivial.
- The symmetric 2-coboundary is $(\partial_*^2)F(X_1;X_2;X_3) = X_1.F(X_2;X_3) - F((X_1,X_2);X_3) + F(X_1;(X_2,X_3)) - X_3.F(X_1;X_2)$, which corresponds in information to $\partial_*^2 F(X_1;X_2;X_3) = X_1.I(X_2;X_3) - I((X_1,X_2);X_3) + I(X_1;(X_2,X_3)) - X_3.I(X_1;X_2) = 0$, and hence the symmetric 2-coboundary is always null-trivial.

1.3.3. Third degree (k=3)

For the third degree $k = 3$, we have the following results:

- The left 3-co-boundary is $\partial^3 F(X_1;X_2;X_3;X_4) = X_1.F(X_2;X_3;X_4) - F((X_1,X_2);X_3;X_4) + F(X_1;(X_2,X_3);X_4) - F(X_1;X_2;(X_3,X_4)) + F(X_1;X_2;X_3)$, which corresponds in information to $\partial^3 F(X_1;X_2;X_3;X_4) = X_1.I(X_2;X_3;X_4) - I((X_1,X_2);X_3;X_4) + I(X_1;(X_2,X_3);X_4) - I(X_1;X_2;(X_3,X_4)) + I(X_1;X_2;X_3) = 0$, and hence the left 3-coboundary is always null-trivial.
- The topological 3-coboundary is $\partial_t^3 F(X_1;X_2;X_3;X_4) = F(X_2;X_3;X_4) - F((X_1,X_2);X_3;X_4) + F(X_1;(X_2,X_3);X_4) - F(X_1;X_2;(X_3,X_4)) + F(X_1;X_2;X_3)$, which corresponds in information to $\partial_t^3 F(X_1;X_2;X_3;X_4) = I(X_2;X_3;X_4) - I((X_1,X_2);X_3;X_4) + I(X_1;(X_2,X_3);X_4) - I(X_1;X_2;(X_3,X_4)) + I(X_1;X_2;X_3) = I(X_1;X_2;X_3;X_4)$, and hence I_4 is a topological 3-coboundary.
- The symmetric 3-coboundary is $(\partial_*^3)F(X_1;X_2;X_3;X_4) = X_1.F(X_2;X_3;X_4) - F((X_1,X_2);X_3;X_4) + F(X_1;(X_2,X_3);X_4) - F(X_1;X_2;(X_3,X_4)) + X_4.F(X_1;X_2;X_3)$, which corresponds in information to $\partial_*^3 F(X_1;X_2;X_3;X_4) = X_1.I(X_2;X_3;X_4) - I((X_1,X_2);X_3;X_4) + I(X_1;(X_2,X_3);X_4) - I(X_1;X_2;(X_3,X_4)) + X_4.I(X_1;X_2;X_3) = -I(X_1;X_2;X_3;X_4)$, and hence $-I_4$ is a symmetric 3-coboundary.

1.3.4. Higher degrees

For $k = 4$, we obtain $\partial^4 F(X_1;X_2;X_3;X_4;X_5) = -I_5$, and $\partial_t^5 F(X_1;X_2;X_3;X_4;X_5) = 0$, and $\partial_*^5 F(X_1;X_2;X_3;X_4;X_5) = 0$. For arbitrary k, the symmetric coboundaries are just the opposite of the topological coboundaries $\partial_t^k = -\partial_*^k$. It is possible to generalize to arbitrary degrees [1] by remarking that we have:

- For even degrees $2k$: we have $I_{2k} = -\partial_t I_{2k-1}$ and then $I_{2k} = \partial_t \partial \partial_t ... \partial \partial_t H$ with $2k-1$ boundary terms. In conclusion, we have:

$$\partial^{2k} = -I_{2k+1} \text{ and } \partial_*^{2k} = -\partial_t^{2k} = 0 \tag{21}$$

- For odd degrees $2k+1$: $I_{2k+1} = -\partial I_{2k-1}$ and then $I_{2k+1} = -\partial \partial_t \partial ... \partial \partial_t H$ with $2k$ boundary terms. In conclusion, we have:

$$\partial^{2k-1} = 0 \text{ and } \partial_*^{2k-1} = -\partial_t^{2k} = -I_{2k} \tag{22}$$

Since $I_k(0) = 0$, such structures with alternated "zeros" enforce the fundamental relation $\partial^k \partial^{k-1} = 0$, and also $\partial_*^k \partial_*^{k-1} = 0$, that defines a cohomological theory. $I_k(0) = 0$, or $F(\emptyset) = 0$ is indeed a theorem ([55] p.3) that follows directly from disjoint additivity axiom $F(X_1 \cup X_2) = F(X_1) + F(X_2)$ and empty set definition [56]. We hence consider together with disjoint additivity axiom, the multiplicative expression of the empty set, e.g. the non-contradiction (consistency) principle stated by Boole called idempotence $X.(1-X) = 0$ [57], as the fundamental axioms of the mathematical theory of (classical) information.

1.3.5. Information double complex

So far, following homological construction, the mutual informations in odd and even degrees appeared in two separate but related cohomologies. Following Gerstenhaber and Shack who gave a Hodge decomposition of Hochschild cohomology [6] and Loday [58], it is possible to merge the previous cohomology theories in a single one by combining them in a double complex $X^{\bullet,\bullet}$, explicitly the triplet $(X^{\bullet,\bullet}, \partial, \partial_*) = (X^{k',k''}, \partial^{k',k''}, \partial_*^{k',k''})$, $(k',k'') \in \mathbb{N} \times \mathbb{N}$ represented by the following section of the diagram:

$$\begin{tikzcd}
X^{k,0} \arrow[r, "\partial^{k,0}"'] & X^{k,1} \arrow[r, "\partial^{k,1}"'] & X^{k,2} \arrow[r, "\partial^{k,2}"'] & X^{k,3} \arrow[r, "\partial^{k,3}"'] & \cdots \arrow[r, "\partial^{k,k-1}"'] & X^{k,k} \\
\vdots \arrow[u, "\partial_*^{k-1,0}"'] & \vdots \arrow[u, "\partial_*^{k-1,1}"'] & \vdots \arrow[u, "\partial_*^{k-1,2}"'] & \vdots \arrow[u, "\partial_*^{k-1,3}"'] & & \vdots \arrow[u, "\partial_*^{k-1,k}"'] \\
X^{2,0} \arrow[u, "\partial_*^{2,0}"'] \arrow[r, "\partial^{2,0}"'] & X^{2,1} \arrow[u, "\partial_*^{2,1}"'] \arrow[r, "\partial^{2,1}"'] & X^{2,2} \arrow[u, "\partial_*^{2,2}"'] \arrow[r, "\partial^{2,2}"'] & X^{2,3} \arrow[u, "\partial_*^{2,3}"'] \arrow[r, "\partial^{2,3}"'] & \cdots \arrow[r, "\partial^{2,k-1}"'] & X^{2,k} \arrow[u, "\partial_*^{2,k}"'] \\
X^{1,0} \arrow[u, "\partial_*^{1,0}"'] \arrow[r, "\partial^{1,0}"'] & X^{1,1} \arrow[u, "\partial_*^{1,1}"'] \arrow[r, "\partial^{1,1}"'] & X^{1,2} \arrow[u, "\partial_*^{1,2}"'] \arrow[r, "\partial^{1,2}"'] & X^{1,3} \arrow[u, "\partial_*^{1,3}"'] \arrow[r, "\partial^{1,3}"'] & \cdots \arrow[r, "\partial^{1,k-1}"'] & X^{1,k} \arrow[u, "\partial_*^{1,k}"'] \\
X^{0,0} \arrow[u, "\partial_*^{0,0}"'] \arrow[r, "\partial^{0,0}"'] & X^{0,1} \arrow[u, "\partial_*^{0,1}"'] \arrow[r, "\partial^{0,1}"'] & X^{0,2} \arrow[u, "\partial_*^{0,2}"'] \arrow[r, "\partial^{0,2}"'] & X^{0,3} \arrow[u, "\partial_*^{0,3}"'] \arrow[r, "\partial^{0,3}"'] & \cdots \arrow[r, "\partial^{0,k-1}"'] & X^{0,k} \arrow[u, "\partial_*^{0,k}"']
\end{tikzcd} \tag{23}$$

where ∂ and ∂_* denote the left and symmetric coboundaries, respectively. This step is natural in information terms since odd and even degrees of mutual information alternate in the two coboundaries. Hence, in the case where we have only one component in the cohomology (connected), the left and symmetric coboundaries anti-commute and we have:

$$\partial^k \partial_*^k + \partial_*^k \partial^k = 0 \tag{24}$$

Proof: if k is even, $k = 2a$, we have $\partial^k \partial_*^k = -I_{2a+1}(0) = 0$, and $\partial_*^k \partial^k = -F_0(-I_{2a+1}) = 0$ since we have only one constant F_0 (connected) and it sends to zero. If k is odd $k = 2a+1$, we have $\partial^k \partial_*^k = F_0(-I_{2a}) = 0$ since we have only one constant F_0 (connected) and it sends to zero, and $\partial_*^k \partial^k = -I_{2a}(0) = 0\,\square$.
The total complex is defined by:

$$X_{\text{Tot}}^k = \oplus_{k'+k''=k} X_{\text{Tot}}^{k',k''} \tag{25}$$

with coboundary $\partial_{tot}^k = \partial^k + (-1)^k \partial_*^k$ and hence the coboundary of the total complex of information is :

$$\partial_{tot}^k = (-1)^{k+1} I_{k+1} \tag{26}$$

Moreover, the mutual information chain rules $I_{k-1} - I_k = X_k.I_{k-1}$ and $I_k = I(X_1;...;X_k;P) = I(X_1) - \sum_{i=2}^k X_i.I(X_1;...;X_{i-1})$ (equation 15) show that conditional mutual information is the coface map of the total complex. In usual notations, $\partial^k = \sum_{i=0}^k (-1)^i d^i$, where $d^i = (-1)^{i+1} X_i.I(X_1;...;X_{i-1})$ is the coface map, and $d^0 = I(X_1)$. Figure 6 gives a graphical representation of this coface map on the information landscapes.

Hence we have established the following theorem (to simplify the notations with respect to usual information theory, we do not write the constant $X_0 = 0$ in the information structure, although it is present in all what follows):

Theorem 1.1. *Consider an information structure (Ω, Π, P) with a n-complex of cochain $F(X_1, ..., X_n; P) = X^n$, then odd mutual-informations are minus even left coboundaries $\partial^{2k} = -I_{2k+1}$, even mutual-informations are minus odd symmetric coboundaries $\partial_*^{2k-1} = -I_{2k}$ [1], and k-mutual-informations are alternate sign $k-1$-coboundaries of the total complex $(X^{\bullet,\bullet}, \partial, \partial_*)$, $\partial^k_{tot} = (-1)^{k+1} I_{k+1}$.*

In conclusion, the total complex allows us to handle the usual information functions introduced in the preceding section in a single cohomology that exhibits a Hodge structure, necessary in the motivic theory where entropy originally appeared, according to the work of Gangl and Elbaz-Vincent, following Kontsevitch and Cathelineau [24,26,27]. We underline that the methods employed to characterize information functions are finite (at the exception of the logarithm function, all objects are finite and methods non-asymptotic; for example, no Stirling approximation is made) and make no positivity or convexity assumptions (at the exception of the probability definition [11]).

1.3.6. k-independence cocycle

It is easy to show that two random-variables X_1, X_2 are statistically independent if and only if $I_2 = I(X_1, X_2) = 0$ (ex: [19], p.27). Moreover, $I(X_1, X_2)$ is more general than the correlation coefficient used in statistics and statistical physics since $I(X_1, X2) = 0$ implies a null correlation coefficient ($I(X_1, X_2) = 0 \Rightarrow \rho_{X_1,X_2} = 0$, $\rho_{X_1,X_2} = \frac{\mathrm{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}$), whereas the converse does not hold in general. This suggests the following definition of k-independence, weaker than the mutual-independence [11].
k-independence (definition): given an information structure (Ω, Π, P), the k random variables $X_1, ..., X_k$ in this structure are k-independent if $I_k = I(X_1, ..., X_k) = 0$.
Then, n variables are mutually independent if and only if $\forall k \leq n$, $I_k = 0$ (the proof is direct from the definition of I_k given in equation 5). This definition of k-independence is directly related to the one obtained using the total correlation C_k proposed by Studený and Vejnarova [16,18] and Margolin and colleagues [10], as expressed in equation 7. C_k quantifies the cumulative dependencies across the dimensions, I_k quantifies the dependencies arising at each dimension. Both definitions are justified, have their interest and are closely related as we will see in section 2.2.1. We focused in this section on I_k because of its straightforward cohomological meaning and characterization. Such a choice implies the appearance of information negativity that we review and characterize more in depth in section 3.1. As a probabilistic interpretation and conclusion, the information cohomology hence quantifies statistical dependences at all degrees and the obstruction to factorization. Moreover, k-independence coincides with cocycles. We therefore expect that the higher cocycles of information, conjectured to be polylogarithmic forms [1,26,27], are characterized by the functional equations $I_k = 0$, and quantify statistical k-independence.

2. Simplicial Information Cohomology

2.1. Simplicial substructures of information

The general information structure, relying on the information functions defined on the whole lattice of partitions, encompasses all possible statistical dependences and relations, since by definition it considers all possible equivalent classes on a probability space. One could hence expect this general structure to provide the promising theoretical framework for classification tasks on data, and this is probably true in theory. However, this general case hardly allows any interesting computational investigation as it implies an exhaustive (free) exploration of computational complexity following Bell's combinatoric in $\mathcal{O}(exp(exp(N^n)))$ for n N-ary variables. This fact was already remarked in the study of aggregation for Artificial Intelligence by Lamarche-Perrin and colleagues [59]. At each order k, the number of k-joint-entropy and k-mutual-information to evaluate is given by Stirling numbers of the second kind $S(n,k)$ that sum to Bell number B_n, $B_n = \sum_{k=0}^{n} S(n,k)$. For example, considering 16 variables that can take 8 values each, we have $8^{16} = 2^{48} \approx 3.10^{14}$ atomic probabilities and the partition lattice of variables exhibits around $e^{e^{2^{48}}-1} \geq 2^{200}$ elements to compute. Such computational reef can be decreased by considering the sample-size m, which is the number of trials, repetitions or points that is used to effectively estimate the empirical probability. It restricts the computation to $\mathcal{O}(exp(exp(m))$, which remains insurmountable in practice with our current classical Turing machines. To circumvent this computational barrier, data analysis is developed on the simplest and oldest subcase of Hochschild cohomology: the simplicial cohomology, which we hence call the simplicial information cohomology and structure, and which corresponds to a subcase of cohomology and structure introduced previously (see Figure 1b.). It corresponds to the example 1 and 4 in [1]. For simplicity, we note also the simplicial information structure (Ω, Δ^n, P), $\Delta^n = (X_1, ..., X_n; P)$, as we will not come back to the general setting. Joint (X_1, X_2) and meet $(X_1; X_2)$ operations on random variables are the usual joint-union and meet-intersection of Boolean algebra and define two opposite-dual monoids, generating freely the lattice of all subsets and its dual. The combinatorics of the simplicial information structure follow binomial coefficients and, for each degree k in an information structure of n variables, we have $\binom{n}{k} = \frac{n!}{k!(n-k!)}$ elements that are in one to one correspondence with the k-faces (the k-tuples) of the n-simplex of random variables (or its barycentric subdivision). It is a (simplicial) substructure of the general structure since any finite lattice is a sub-lattice of the partition lattice [60]. This lattice embedding and the fact that simplicial cohomology is a special case of Hochschild cohomology can be inferred directly from their coboundary expression and has been explicitly formalized in homology: notably, Gerstenhaber and Shack showed that a functor, noted $\Sigma \mapsto k_\Sigma!$, induces an isomorphism between simplicial and Hochschild cohomology $H^\bullet(\Sigma, k) \cong H^\bullet(k_\Sigma!, k_\Sigma!)$ [61]. A simplicial complex $X^k = F(X_1, ..., X_k; P)$ of measurable functions is any subcomplex of this simplex Δ^n with $k \leq n$, and any simplicial complex can be realized as a subcomplex of a simplex (see Steenrod [62] p.296). The information landscapes presented in Figure 4 illustrate an example of such a lattice/information structure. Moreover in this ordinary homological structure, the degree obviously coincides with the dimension of the data space (the data space is in general $\mathbb{R}^n$, the space of "co-ordinate" values of the variables). This homological (algebraic, geometric and combinatorial) restriction to the simplicial subcase can have some important statistical consequences. In practice, whereas the consideration of the partition lattice ensured that no reasonable (up to logical equivalence) statistical dependences could be missed (since all the possible equivalence classes on the atomic probabilities were considered), the monoidal simplicial structure unavoidably misses some possible statistical dependences as shown and exemplified by James and Crutchfield [63].

2.2. Topological Self and Free Energy of k-body interacting system - Poincaré-Shannon Machine

2.2.1. Topological Self and Free Energy of k-body interacting system

The basic idea behind the development of topological quantum field theories (see Schwarz, Atiyah and Witten [55,64,65]) was to define the action and energy functionals on a purely topological ground, independently of any metric assumptions, and to derive from this the correlation functions or partition functions. The very basic principle goes back to Noether, who associated on homological ground any continuous symmetry to a conserved quantity [66]. Here, in our elementary model for applied purposes, we define, in the special case of classical and discrete probability, the k-mutual-information I_k (that we prove to be discrete $(k-1)$-differential operators and statistically more general than correlation functions, cf. 1.3.6 [67]), as the contribution of the k-body interactions to the energy functional. Some further observations support such a definition: i) as stated in [1] (Th.D), the signed mutual-informations $(-1)^k I_k$ defining energy are sub-harmonic, a kind of weak convexity ii) in the next sections, we define the paths of information and show that they are equivalent to the discrete symmetry group iii) from the empirical point of view, Figure 7 shows that these energy functionals estimated on real data behave as expected for usual k-body homogeneous formalism such as Van-Der-Walls model, or more refined Density Functional Theory (DFT) [68,69]. These definitions, given in the context of simplicial structures, generalize to the case of partitions lattice, and altogether provide the usual thermodynamical and machine-learning expressions and interpretation of mutual-information quantities: some new methods free of metric assumptions. There are two qualitatively and formally different components in the I_k :

- **Self-internal information energy (definition):** for $k = 1$, I_1 and their sum in an information structure expressed in equation 16, namely $\sum_{T\subset[n];card(T)=1} I_1(X_T; P)$, are a self-interaction component, since it sums over marginal information-entropy $I_1(X_i) = H_1(X_i)$. We call the first dimension mutual information component $U(X_1, ..., X_n; P_N)$ the self information or internal information energy, in analogy to usual statistical physic and notably DFT:

  $$U(X_1, ..., X_n; P_N) = \sum_{i=1}^{n} I_1(X_i; P_N) \tag{27}$$

  Note that in the present context, which is discrete and where the interactions do not depend on a metric, the self-interaction does not diverge, which is a usual problem with metric continuous formalism and was the original motivation for regularization and renormalization infinite corrections, considered by Feynman and Dirac as the mathematical default of the formalism [70,71].
- **k-free-energy and total-free-energy (definition)**: for $k \geq 2$, $(-1)^k I_k$ and their sum in an information structure (equation 16) quantify the contribution of the k-body interactions. We call the k$^{\text{th}}$ dimension mutual information component $(-1)^k I_k$, given in equation 5, the k-free-information-energy. We call the (cumulative) sum over dimensions of these k-free-information-energies starting at pairwise interactions (dimension 2), the total n-free-information-energy, and note it $G(X_1, ..., X_n; P_N)$:

  $$G(X_1, ..., X_n; P_N) = \sum_{i=2}^{n} (-1)^{i-1} \sum_{I\subset[n];card(I)=i} I_i(X_I; P_N) = C_n(X_1; ...X_n; P_N) \tag{28}$$

  The total free-energy is the total correlation (equation 7) introduced by Watanabe in 1960 [16] that quantifies statistical dependence in the work of Studený and Vejnarova [18] and Margolin and colleagues [10], and among other examples consciousness in the work of Tononi and Edelman [17]. In agreement with the results of Baez and Pollard in their study of biological dynamics using out-of-equilibrium formalism [72], the total free-energy is a relative entropy. Moreover, whereas I_k energy component can be negative, the C_k total energy component is always non-negative. Each $(-1)^k I_k$ term in the free energy can be understood as a free energy correction accounting for the k-body interactions.

Entropy is given by the alternated sums of information (equation 16), which then read as the usual isotherm thermodynamic relation:

$$H_n(X_1, ..., X_n; P_N) = U(X_1, ..., X_n; P_N) - G(X_1, ..., X_n; P_N) \tag{29}$$

This information theoretic formulation of thermodynamic relation follows Jaynes [73,74], Landauer [75], Wheeler [76], and Bennett's[77] original work, and is general in the sense that it is finite and discrete, and holds independently of the assumption of the system being in equilibrium or not, i.e. for whatever finite probability. In more probabilistic terms, it does not assume that the variables are identically distributed, a condition that is required for the application of classical central limit theorems

(CLT) to obtain the normal distributions in the asymptotic limit [78]. In the special case where one postulates that the probability follows the equilibrium Gibbs distribution, which is also the maximum entropy distribution [79,80], the expression of the joint-entropy ($k = -1/\ln 2$) allows to recover the equilibrium fundamental relation, as usually achieved in statistical physic (see Adami and Cerf [81] and Kapranov [82] for more details). Explicitly, let's consider the Gibb's distribution:

$$p(X_1 = x_1, ..., X_n = x_n) = p\underbrace{_{ij...n}}_{\text{n indices}} = \frac{1}{Z} e^{-\beta E_{ij...n}/k_B T} \tag{30}$$

where $E_{ij...n}$ is the energy of the elementary-atomic probability $p_{ij...n}$, k_B is Boltzmann constant, T the temperature and $Z = \sum_{i,j,..,n}^{N_i.N_j...N_n} e^{-E_{ij...n}/k_B T}$ is the partition function, such that $\sum_{i,j,..,n}^{N_i.N_j...N_n} p_{ij...n} = 1$. Since $H(X_1, ..., X_n) = k \sum_{i,j,..,n}^{N_i.N_j...N_n} p_{ij...n} \ln p_{ij...n}$, equals the thermodynamic entropy function S up to the arbitrary Landauer constant factor $k_B \ln 2$, $S = k_B \ln 2 H(X_1, ..., X_n)$, the entropy for Gibbs distribution gives:

$$H(X_1, ..., X_n)/k = \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n}/k_B T + \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} \ln Z = (\langle E \rangle - G)/k_B T \tag{31}$$

, which gives the expected thermodynamical relation:

$$k_B T \ln 2 . H(X_1, ..., X_n) = \langle E \rangle - G = U - G \tag{32}$$

, where G is the free-energy $G = -k_B T \ln Z$.
In the general case of arbitrary random variables (not necessarily iid) and discrete probability space, the identification of marginal informations with internal energy:

$$\sum_{k=1}^{n} H(X_k) = \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n} \tag{33}$$

implies by direct algebraic calculus that:

$$\sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n} = - \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} \ln \left(\prod_{k=i}^{n} p_{\bullet\bullet...k...\bullet} \right) \tag{34}$$

, where the marginal probability $p_{\bullet\bullet...k...\bullet}$ is the sum over all probabilities for which $X_k = x_k$. It is hence tempting to identify the elementary-atomic energies $E_{ij...n}$ with the elementary marginal informations $\ln p_{\bullet\bullet...k...\bullet}$. This is achieved uniquely by considering that such an elementary energy function must satisfy the additivity axiom (extensivity): $(E(X_i = x_i, Xj = x_j) = E_{i,j} = E_{ij} = E_i + E_j)$, which is the functional equation of the logarithm. The original proof goes back at least to Kepler, an elementary version was given by Erdos [83], and in Information theory terms can be found in the proofs of uniqueness of "single event information function" by Aczel and Darokzy ([84], p.3). It establishes the following proposition:

Theorem 2.1. *Given a simplicial information structure, the elementary energies satisfying the extensivity axiom are the functions:*

$$E_{ij...n} = k \sum_{k=i}^{n} \ln p_{\bullet\bullet...k...\bullet} \tag{35}$$

, where k is an arbitrary constant settled to $k = -1/\ln 2$ *for units in bit.*

The geometric meaning of these elementary energies as log of marginal elementary probability volumes (locally Euclidean) is illustrated in Figure 3 and further underlines that $I_{k,k\geq 2}$ are volume corrections accounting for the statistical dependences among marginal variables.

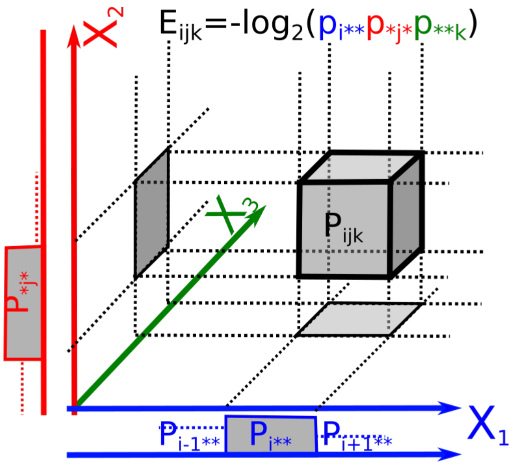

Figure 3. Elementary energy as logarithm of locally Euclidean probability volumes. Example of an elementary energy E_{ijk} associated to a probability p_{ijk} using the same data space representation as in Supplementary Figure 5 and Figure 14 ($n = 3$ variables). The histograms of the marginal distributions of each variable are plotted beside the axes.

Examples: i) in the example of 3 binary random variables ($n = 3, N_1 = N_2 = N_3 = 2$, three variables of Bernoulli) illustrated in Figure 9, we have $E_{000} = -\ln(p_{0\bullet\bullet}p_{\bullet 0\bullet}p_{\bullet\bullet 0})$, $E_{000} = -\ln(p_{000} + p_{010} + p_{001} + p_{011}) - \ln(p_{000} + p_{100} + p_{001} + p_{101}) - \ln(p_{000} + p_{100} + p_{010} + p_{110})$ and in the configuration of negative-entangled-Borromean information of the Figure 9a-right, we obtain $E_{000} = 3$ in bit units, and similarly $E_{001} = E_{010} = E_{011} = E_{101} = E_{110} = E_{111} = 3$, and we hence recover $U = \sum_{i,j,k}^{8} p_{ijk}E_{ijk} = \sum_{i=1}^{3} H(X_i) = 3$ bits. Note that the limit $0 \ln 0 \sim 0$ avoids singularity of elementary energies.
ii) in the special case of identically distributed variables, $p_{\bullet\bullet\ldots k\ldots\bullet} = p_{\bullet\ldots j\ldots\bullet\bullet}$, we have $E_{ij\ldots n} = nk \ln p_{\bullet\bullet\ldots k\ldots\bullet}$ and hence the marginal Gibbs distribution: $p_{\bullet\bullet\ldots k\ldots\bullet} = e^{\frac{E_{ij\ldots n}}{nk}}$.
iii) for independent identically distributed variable (non-interacting), we have $G_n = 0$, and hence:

$$H_n(X_1, ..., X_n; P_N) = U(X_1, ..., X_n; P_N) = nH(X_i) \tag{36}$$

iv) considering the variables to be the $6n$ variables of the phase space, with one variable of position and one variable of momentum per body (noted $(X_k^1, X_k^2, X_k^3, P_k^1, P_k^2, P_k^3)$ for the kth body), it is possible to re-express the semi-classical formalism according to which the entropy formulation is (Landau and Lifshitz [85],p.22):

$$H_{6n}(X_1^1, X_1^2, X_1^3, P_1^1, P_1^2, P_1^3, ..., P_n^3; P_N) = \log\left(\frac{\Delta X \Delta P}{(2\pi\hbar)^{6n}}\right) \tag{37}$$

It is achieved by identifying the internal and free energy as following:

$$\langle E \rangle = -6n \log(2\pi\hbar) \tag{38}$$

$$G = -\log(\Delta X \Delta P) \tag{39}$$

This identifies the elementary volumes/probabilities with the Planck constant, the quantum of action (the consistency in the units is realized in section 2.4.3 by the introduction of time). The quantum of action can be illustrated by considering in the Figure 3 that it is the surface of the square/rectangle for two conjugate variables (considered as position and momentum). In this setting, $\Delta X \Delta P$ quantifies

doi:10.20944/preprints201804.0157.v1

the non-extensivity of the volume in the phase-space due to interactions, or in other words, the mutual-informations account for the consideration of the dependence of the subsystems considered as opened and exchanging energy. As noted by Baez and Pollard, the relative entropy provides a quantitative measure of how far from equilibrium the whole system is [72]. The (real) random variables are the classical analogs of the observables in quantum physic, and can be considered as synonymous of classical observables. The basic principle of such expression of information theory in physic is known at least since Jaynes's work [74,86]. However, a more complete (semi) classical expression of the information structures presented here could be done.

As a conclusion, information topology applies, without imposing metric or symplectic or contact structures, to physical formalism of n-body interacting systems relying on empirical measures. Considering the $3n$ or $6n$ dimensions (degrees of freedom) of a configuration or a phase space as random variables, it is possible to recover the (semi) classical statistical physic formalism. We emphasize in section 5.5 that an important part of the results in statistical physic and thermodynamic relies on taking the infinite dimension case ($n \rightarrow \infty$), the asymptotic limit, or the "continuous space" case ($N \rightarrow \infty$) and identically distributed variables, so that the finite case investigated here can be understood as a micro-heterogeneous thermodynamic. So the main interest of this section relative to thermodynamic is the proposition that I_k functions quantify the contribution of the k-body interactions to the free-energy functional that still holds in small heterogeneous systems. It is also interesting to discuss the status of the analog of the temperature variable in the present formalism which is played by the graining, which is the size N_i of the alphabet of a variable X_i (cf. section 5.5.3). In usual thermodynamic we have $H(X^n; P_N) = T.S(X^n)$, and to stay consistent, temperature shall be a functional inverse of the graining N, lowest temperature being the finest grain (large N), highest temperature being the coarsest graining (small N).

2.2.2. Machine learning - Poincare-Shannon Machine

The energetic expression of I_k functions allows a direct interpretation of the information topology in terms of machine learning. The original work based on spin networks by Hopfield [87] formalized fully recurrent networks as n binary random variables ($N = 2$). Ackley, Hinton and Sejnowski [88] followed up by imposing the Markov Field condition, allowing the introduction of conditional independence to handle network structures with hidden layers and hidden nodes. The result, the Boltzmann or Helmholtz machine [89], relies on the maximum entropy or free-energy minimization principle, and originally on minimizing the relative entropy between the network and environmental states [88]. On the side of statistics, Efron established the statistical curvature, which vanishes for exponential families and is positive for nonexponential families [90]. Cencov characterized the Fisher information metric on statistical models as the only invariant Riemannian metric under diffeomorphism [91], recently generalized to the infinite dimensional case by Ay, Jost, Le and Schwachhöfer [92]. Following these results, Amari introduced a more general geometrical and statistical formalization of the machine learning methods, notably introducing Fisher metric and giving the natural gradient descent algorithm [93,94]. Although they point toward the existence of a topological formalism, these methods are geometric as they require a predefined metric or distance to reach the minimum of the energy functional. The present method is topological and avoids the introduction a priori of such a metric: rather, a family of Shannon's pseudometric emerges from the formalism as a first cohomological class (cf. section 1.3.1): considering a symmetric action of conditioning, we obtain Shannon's metric parametrized by a scalar multiplicative constant, much like for the Cayley-Klein-Hilbert metric [95] or the metric of Cartan for 3D conformal anallagmatic space [96]. Notably, the notion of geodesic used in machine learning is replaced by the homotopical notion of path, that we will introduce in the next section. Moreover, The generalization of binary neuronal-ising models to arbitrary multivalued variables with N_i values corresponds to the generalization of spin models to Potts models [97]. Such generalization is pertinent for biological systems in general, since

coding and meaningful variables in biology are usually non-digital and only in quite exceptional cases binary. We propose to call the algorithm and data analysis presented here, the Poincaré-Shannon machine, since it implements simplicial homology and information theory in a single framework, applied effectively to empirical data.
The topological methods can also be understood in the perspective of classical statistical tests and data analysis for multivariate data. Benzecri formalized the factorial correspondence analysis [98], a special case of principal component analysis with Euclidian χ^2-metric, the χ^2 test relying on the assumption of independent normally or identically distributed data. The topological information approach gives a preliminary view of factorial analysis when one does not assume a Euclidian χ^2-metric and an identical distribution of the variables. The section 5.5.2, which investigates and discusses Pethel and Hahs test of dependence [5], provides a preliminary presentation in practice of what could be called topological factorial analysis, however with important open questions toward such an achievement.
As a summary and more generally, all the works we just cited point out that the geometrical formalization of probability that occurred within the last decades led to a situation that is analog to the discovery of the existence and relevance of non-euclidian deterministic geometries one century before (see Poincare's review of 1891 [99]), and that information theory provides the functions to quantify such geometries. This analogy may be deepened by proposing that the Axiom of Choice (or the resulting Excluded Third) plays in the present probability context the role of Euclide's 5th axiom in the genesis of modern geometry. Interestingly, the motivation in physic of the associated theories, going from the deterministic to the probabilistic view, also followed a partial shift from the celestial motivation of relativistic mechanic, to the more down-to-earth and constructive aims of biological, neuroscience and ecological systems modeling. The "frontiers" of physic may be much closer to ourselves than previously thought and could be bounded by what humans (or machines) can reasonably understand and infer from observation.

On the side of algebraic topology, the identification of the topological structures of dataset has motivated important researches following the development of persistent homology [100–102]. Combining in a single framework statistical and topological structures remains an active challenge of data analysis that already gave some interesting results [103,104]. Some recent works have proposed information theoretical approaches grounded on homology, defining persistent entropy [105,106], graph's topological entropy [107], spectral entropy [108], or multilevel Integration entropies [109]. In regard to current topological data analysis methods, the main peculiarity of the methods presented here is to be based on atomic probabilities rather than on data points, and on an intrinsically probabilistic framework: the differential operators are fundamental maps in probability-information theory. On the data analysis side, it provides new algorithm and tools for Topological Data Analysis allowing to rank, detect clusters, functional modules and to make dimensionality reduction; all these classical tasks in data analysis have indeed a deep homological meaning.

2.3. *k-Entropy and k-Information landscapes*

Information landscapes (definition): The information landscapes are a representation of the (semi)-lattice of information structures where each element is represented as a function of its corresponding value of entropy or mutual information. In abscissa are the dimensions k and in ordinate the values of information functions of a given subset of k variables.
In data science terms, these landscapes are a visualization of high-dimensional data. In information theory terms, it provides a representation of Shannon's work on lattice [50] further developed by Han [110]. H_k and I_k, as real continuous functions, provide a ranking of the lattices at each dimension k. It is the ranking, i.e. the relative values of information, which matters and comes out of the homological approach, rather than the absolute values. The principle of H_k and I_k landscapes is illustrated in Figure 4 for $n = 4$. H_k and I_k analyse quantify the variability-randomness and statistical dependences at all dimensions k, respectively, from 1 to n, n being the total number of variables under study. The

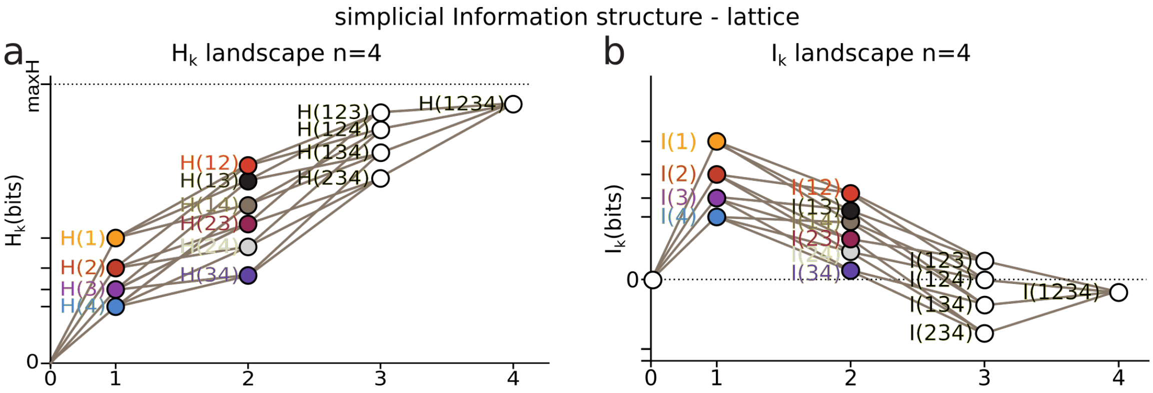

Figure 4. Entropy and information landscapes. a, illustration of the principle of entropy H_k landscape and **b,** of a mutual-information I_k landscape for $n = 4$ random variables. The lattice of the simplicial information structure is depicted with grey lines.

H_k landscape represents the values of joint entropy for all k-tuples of variables as a function of the dimensions k, the number of variables in the k-tuple, together with the associated edges-paths of the lattice (in grey). The I_k landscape represents the values of mutual-information for all k-tuples of variables as a function of the dimension k, which is the number of variables in the k-tuple.

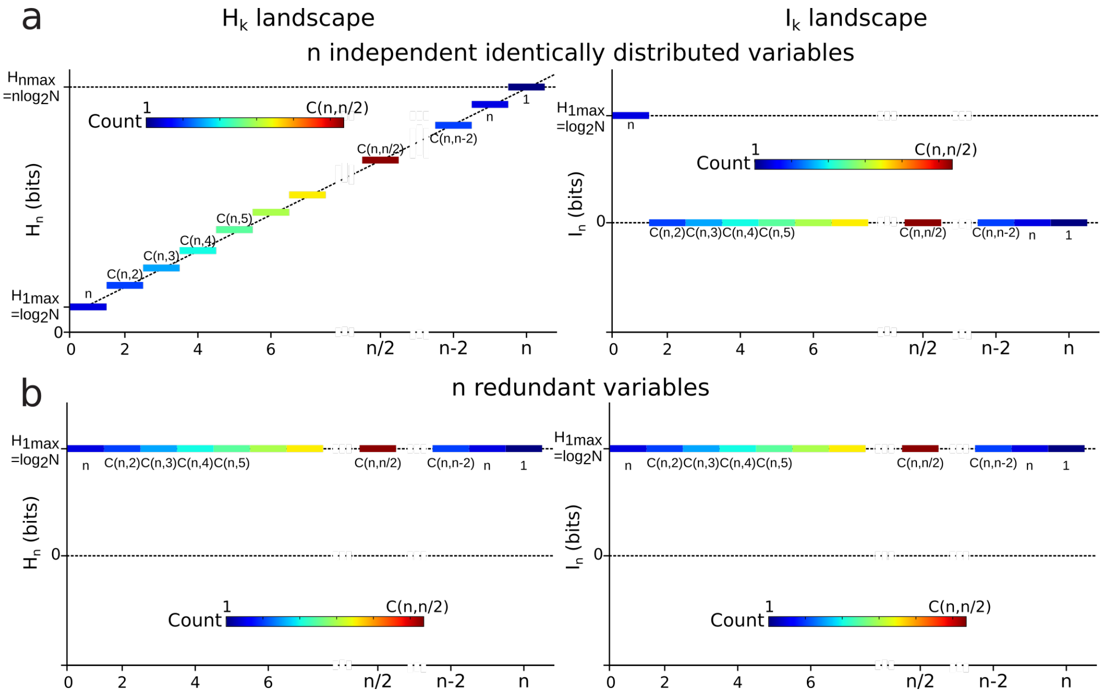

Figure 5. Theoretical examples of entropy and information landscapes. a, H_k and I_k landscapes for n independent identically distributed variables. The degeneracy of H_k and I_k values is represented by a color code: the number of k-tuples having the same information value. **b,** H_k and I_k landscapes for n fully redundant variables. Such variables are equivalent from the information point of view, they are identically distributed and fully dependent.

Figure 5 gives two theoretical extremal examples of such landscapes, one for independent identically distributed variables (totally disordered) and one for fully dependent identically distributed variables (totally ordered). The degeneracy of H_k and I_k values is given by the binomial coefficient (color code in Figure 5), hence allowing to derive the normal exact expression of the information landscapes in the assymptotic infinite dimensional limit ($n \to \infty$) by application of Laplace-Lemoivre theorem. These are theoretical extremal examples: H_k and I_k landscapes effectively computed and estimated on biological data with finite sample are shown in Figure 11, and in practice the finite sample size (m) may impose some bounds on the landscapes (see section 5.5).

2.4. Information Paths and Minimum Free Energy Complex

In this section we establish that information landscapes and paths encode directly all the basic equalities, inequalities and functions of information theory and allow us to obtain the minimum free energy complex that we estimate on data.

2.4.1. Information Paths

Information path (definition): On the discrete simplicial information lattice Δ_k, we define a path of degree k as a sequence of edges of the lattice that begins at the least element of the lattice (the identity-constant "0"), travels along edges from vertex to vertex of increasing dimension and ends at the greatest element of the lattice of dimension k. Information paths could also be called information chains, as they are generated by the chain rules of information (equation 14).
From the algebraic point of view, information paths are defined on both joint-entropy and meet-mutual information semi-lattices, and the usual joint-entropy and mutual-information functions are defined on each element of such paths. Entropy path and information path of degree k are noted HP_k and IP_k, respectively, and the set of all information paths is noted $\mathcal{HP}_k = \{HP_i\}_{i\in 1,\dots,k!}$ for the entropy paths, and $\mathcal{IP}_k = \{IP_i\}_{i\in 1,\dots,k!}$ for the mutual-information paths. We have the theorem:

Theorem 2.2. *The two sets of all information paths $\mathcal{HP}_k$ and $\mathcal{IP}_k$ in the simplicial information structure Δ_k are both in bijection with the symmetric group S_k. Notably, there are $k!$ information paths in Δ_k.*

Proof: by simple enumeration, an edge of dimension m connects $k-m$ edges of dimension $m+1$, the number of paths is hence $(k-0).(k-1)....(k-k+2).(k-k+1) = k!$, hence the conclusion □.
A given path can be identified with **a permutation** or **a total order** by extracting the missing variable in a previous node when increasing the dimension, for example the mutual-information path in Δ_4: $IP_i = 0 \to (0, X_2) \to (0, X_1, X_2) \to (X_1, X_2, X_4) \to (0, X_1, X_2, X_3, X_4)$ can be noted as the permutation σ:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 2 & 1 & 4 & 3 \end{pmatrix} \text{ or } (01234) \xrightarrow{\sigma} (02143) \tag{40}$$

We note an information path with arrows, giving for the previous example $IP_i = (0 \to X_2 \to X_1 \to X_4 \to X_3)$. These paths shall be seen as the automorphisms of $\{1,2.....k\} = [k]$ and the space of entropy and mutual information paths can be endowed with the structure of two opposite symmetric groups S_k and S_k^{opp} . The equivalence of the set of paths and symmetric group only holds for the subcase of simplicial structures, and the information paths in the lattice of partition are obviously much richer. More precisely, the subset of simplicial information paths in the lattice of partitions corresponds to the automorphisms of the lattice. It is known that the finite symmetric group is the automorphism group of the finite partition lattice [111]; such remark being a preliminary basis of the construction of cyclic homology upon Eulerian partitions by Loday [58], so the only novelty of this theorem is its appearance in the context of information, thermodynamic and probability theory. The geometrical realization of information paths $\mathcal{IP}_k$ and $\mathcal{HP}_k$ consists in two dual permutohedron (see Postnikov [112]), and gives the informational version of the work of Matúš on conditional probability and permutohedron [113].

2.4.2. Derivatives, inequalities and conditional mutual information negativity

Derivatives of information paths:

In the information landscapes, the paths HP_i and IP_i are piecewise linear functions $IP_i(k)$ with $IP_i(k) = I_k$ where I_k is the mutual-information of the k-tuple of variables pertaining to the path IP_i. We define the first derivatives of the paths for both entropy and mutual information structures as piecewise linear functions:

First derivative of entropy path: the first derivative of an entropy path $HP_i(k)$ is the conditional information $(X_1, ..., X_{k-1}).H(X_k; \mathbb{P})$:

$$\frac{dHP_i(k)}{dk} = H(X_1, ..., X_k; \mathbb{P}) - H(X_1, ..., X_{k-1}; \mathbb{P}) = (X_1, ..., X_{k-1}).H(X_k; \mathbb{P}) \tag{41}$$

This derivative is illustrated in the graph of Figure 6a. It represents an arbitrary element and path-edge of the landscape for a dimension k to $k+1$. It implements the chain rule of entropy $H_{k+1} - H_k = (X_1; ...; \widehat{X_i}; ...; X_{k+1}).H(X_i)$ (equation 12), and in homology provides a diagram where conditional entropy is a simplicial coface map $(X_1; ...; \widehat{X_i}; ...; X_{k+1}).H(X_i) = d^i : X^k \rightarrow X^{k+1}$, as a special case of section 1.3.5.

First derivative of mutual information path: the first derivative of an information path $IP_i(k)$ is minus the conditional information $(X_k).I(X_1, ..., X_{k-1}; \mathbb{P})$:

$$\frac{dIP_i(k)}{dk} = I(X_1, ..., X_k; \mathbb{P}) - I(X_1, ..., X_{k-1}; \mathbb{P}) = -X_k.I(X_1, ..., X_{k-1}; \mathbb{P}) \tag{42}$$

This derivative is illustrated in the graph of Figure 6b. It represents an arbitrary element and path-edge of the landscape for a dimension k to $k+1$. It implements the chain rule of mutual-information $I_{k-1} - I_k = X_k.I_{k-1}$ (equation 13), and in homology provides a diagram where minus the conditional mutual-information is a simplicial coface map $X_i.I(X_1; ...; \widehat{X_i}; ...; X_{k+1}) = d^i : X^k \rightarrow X^{k+1}$, introduced in section 1.3.5.

Bounds of the derivatives and information inequalities

The slope of entropy paths is bounded by the usual conditional entropy bounds ([4] p.27-28). Its minimum is 0 and is achieved in the case where X_{k+1} is a deterministic function of $(X_1, ..., X_k)$ (lower dashed red line in Figure 6a). Its global upperbound is max $H_{k+1} = k.\ln(N_1...N_{k+1})$ and its sharp bound given by $(X_1; ...; \widehat{X_i}; ...; X_{k+1}).H(X_i) \leq H(X_i)$ is achieved in the case where X_{k+1} is independent of $X_1, ..., X_k$ (we have $H_{k+1} = H_k + H(X_{k+1})$ (higher dashed red line in Figure 6a). Hence, any entropy path lies in the (convex) entropy cone defined by the 3 points labeled H_k, min H_{k+1} and max H_{k+1}: the 3 vertices of the cone depicted as a red surface in Figure 6a and called the Shannonian Cone following Yeung's seminal work [114].

The behavior of a mutual-information path and the bounds of its slope are richer and more complex than the preceding conditional entropy:

- For $k = 2$, the conditional information is the conditional entropy $X_i.I(X_j) = X_i.H(X_j)$ and has the same usual bounds $0 \leq X_i.I(X_j) \leq I(X_j)$.
- For $k = 3$ the conditional mutual-information $X_i.I(X_j; X_h)$ is always positive or null $X_i.I(X_j; X_h) \geq 0$ and hence $I_2 \geq I_3$ ([4] p.26, the opposite of th.2.40 p.30), whereas the higher limit is given by $X_i.I(X_j; X_h) \geq \min\left(X_i.H(X_j), X_i.H(X_h)\right)$ ([20] th.2.17), with equality iff X_j and X_h are conditionally independent given X_i, and implying that the slope from $k = 2$ to $k = 3$ increases in the I_k landscape.
- For $k > 3$, $X_k.I(X_1; ..; X_{k-1})$ can be negative as a consequence of the preceding inequalities. In terms of information landscape this negativity means that the slope is positive, hence that the information path has crossed a critical point, a minimum. As expressed by theorem 1.1,

doi:10.20944/preprints201804.0157.v1

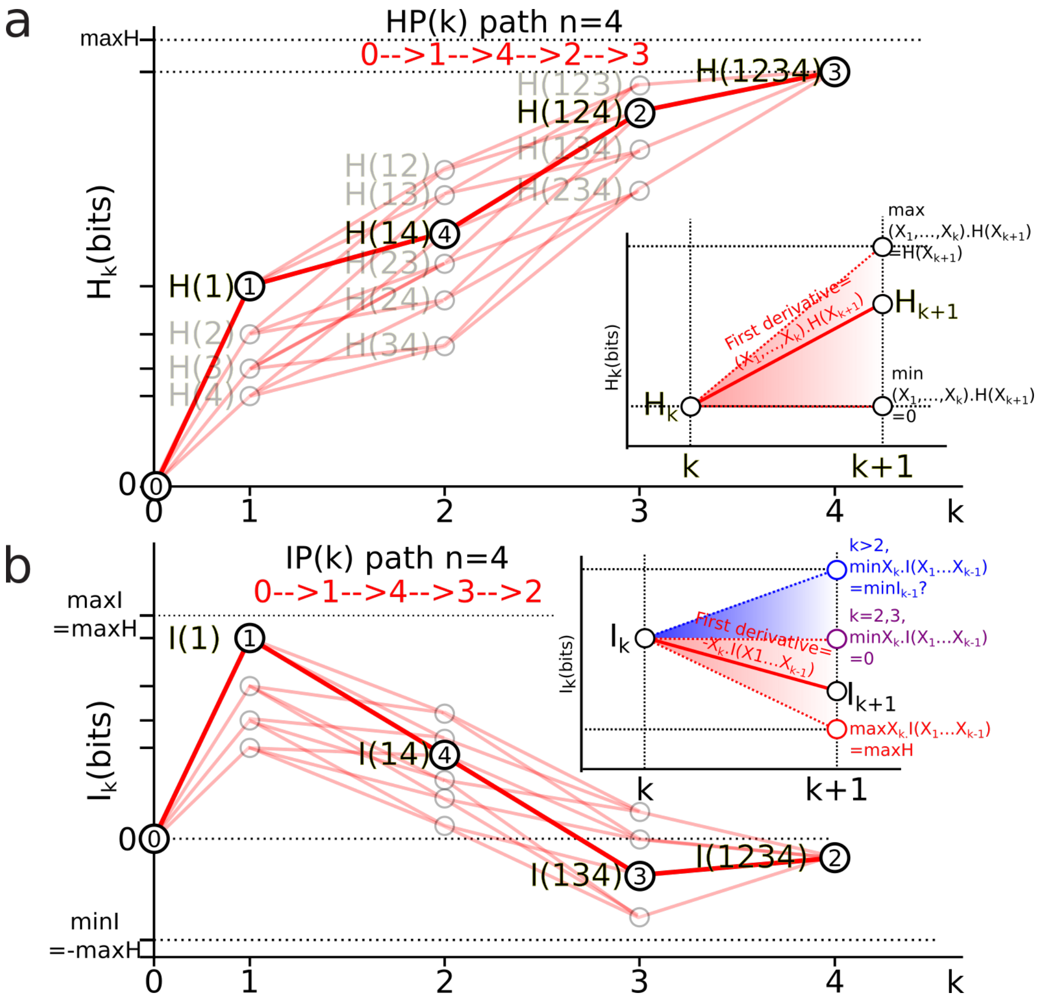

Figure 6. Entropy and information paths. Illustration of an entropy path $HP_i = 0 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 3$ **(a)** and of a mutual information path $IP_i = 0 \rightarrow 1 \rightarrow 4 \rightarrow 3 \rightarrow 2$ **(b)** for $n = 4$ random variables (see text).

$X_k.I(X_1;..;X_{k-1}) < 0$ iff $I_k < I_{k+1}$, meaning that the mutual information estimation diverges. The minima correspond to zeros of conditional information (conditional independence) and hence detect cocycles in the data. The appendix A.1 presents the upper and lower (conjectured) bounds of conditional mutual information. The results on information inequalities define as "Shannonian" [115–117] the set of inequalities that are obtained from conditional information positivity ($X_i.I(X_j;X_h) \geq 0$) by linear combination, which forms a convex "positive" cone after closure. "Non-Shannonian" inequalities could also be exhibited [115][116], hence defining a new convex cone that includes and is strictly larger than the Shannonian set. Following Yeung's nomenclature and to underline the relation with his work, we call the positive conditional mutual-information cone (surface colored in red in Figure 6b) the "Shannonian" cone and the negative conditional mutual-information cone (surface colored in blue in figure 6b) the "non-Shannonian" cone.

In appendix A.2 we give the representation in information landscape of the generalization of conditioning to joint conditional entropy and meet conditional information. In appendix A.3, we give the characterization of Hu Kuo Ting of Markov chains in terms of mutual informations and their representation in information landscape. In appendix A.4, we give an interpretation of conditional mutual information positivity/negativity, in terms of stability/instability, and provide a gluing of H_k and I_k landscape in a single landscape consisting in a twist (Tate-like).

2.4.3. Information paths are random processes: topological 2nd law of thermodynamic and entropy rate

So far, at the exception of the action of conditioning, the information structures were presented quite statically. Here we present the dynamical aspects of information structures. Information paths provide directly the standard definition of a stochastic process and it imposes how the time arrow appears in the homological framework, how time series can be analyzed, how entropy rates can be defined (etc.).

Random (stochastic) process (definition [118]): A random process $\{X_t, t \in T\}$ is a collection of random variables on the same probability space $(\Omega, \mathcal{B}, P)$ and the index set T is a totally ordered set. A stochastic process is a collection of random variables indexed by time, the probabilistic version of a time series. Considering each symbol of a time series as a random variable, the definition of a random-stochastic process corresponds to the unique information paths HP_i and IP_i which total order is the time order of the series. We have the following lemma:

Lemma 1. *(Stochastic process and information paths): Let (Ω, Δ^k, P) be a simplicial information structure, then the set of entropy paths $\mathcal{HP}_k$ and of mutual-information paths $\mathcal{IP}_k$ are in one to one correspondence with the set of stochastic processes $\{X_t, t \in T, |T| = k\}$.*

Proof: direct from the definitions □.

As we previously stated, these paths are also automorphisms of $\{1, 2.....k\} = [k]$. We obtain immediately the topological version of the second law of thermodynamic, which improves the result of Cover [19]:

Theorem 2.3. *(Stochastic process and information paths): Let (Ω, Δ^k, P) be a simplicial information structure, then the entropy of a stochastic process can only increase with time.*

Proof: given the correspondence we just established, the statement is equivalent to $H(X_1, ..., X_k) \geq H(X_1, ..., X_{k-1})$, which is a direct consequence of conditional entropy positivity and the chain rule of information with $k = -1/\ln 2$. The generalization with respect to the stationary Markov condition used by Cover comes from the remark that in any case the indexing set of the variable is a total order. Note that the homological formalism imposes an "initial" minimally low entropy state $H(0) = I(0) = 0$ (a usual assumption in physic), the constant and zero degree homology, which has to have at least 1 component to talk about the cohomology □.

Remark: the meaning of this theorem in common terms was summarized by Gabor and Brillouin: *"you can't have something for nothing, not even an observation"* [119]. This increase in entropy is illustrated in Figure 11. The usual stochastic approach of time series assumes a Markov chain structure, imposing peculiar statistical dependences that restrict memory effects (cf. section A.3). The consideration of stochastic processes without restriction allows any kind of dependences and arbitrary long historical and "non trivial" memory. From the biological point of view it formalizes the phenomenon of arbitrary long-lasting memory. From the physical point of view, without proof, such a framework appears as a classical analog of the consistent or decoherent histories developed notably by Griffiths [120], Omnes [121], and Gell-Mann and Hartle [122]. The information structures impose a stronger constraint of a totally ordered set (or more generally a weak ordering) than the preorder imposed by Lieb and Yngvason [123] to derive the second law. It is also interesting to note that even in this classical probability framework, the entropy cone (the topological cone depicted in Figure 6a) imposed by information inequalities, when considered with this time ordering, is a time-like cone (much-like the special relativity cone), but with the arguably remarkable fact that we did not introduce any metric.

A natural question, regarding directly the formalization of information inequalities into category theory, is: can this theorem be generalized by the consideration of data processing inequalities?

The stochastic process definition allows to define the finite and asymptotic information rate: the finite information rate r of an information path HP_i is $r = \frac{H_k}{k}$. The asymptotic information rate r of an information path HP_i is $r = \lim_{k \to \infty} \frac{H_k}{k}$. It requires the generalization of the present formalism to the infinite dimensional setting or infinite information structures, which is not trivial and will be investigated in further work. We underline that no stationarity, ergodicity or Markovian (and all the more iid memoryless processes) assumptions have been made in the formalism. Altogether, the formalization of information paths provides an elementary setting to tackle the problematics raised by Gromov on symmetry, probability, entropy [124]. Question: recently Baez and Fong published a Noether Theorem for Markov Processes [125]; can we derive a Noether theorem for random discrete processes in general, that is for all the symmetric group S_n using the present construction? Such a theorem would provide the topological expression of the first law of thermodynamic. Such a question was asked by Neuenschwander [126], and related to this aim, Mansfield gave a Noether theorem for finite elements [127].

2.4.4. Local minima and critical dimension

The derivative of information paths allows to establish the lemma on which is based the information path analysis. A critical point is said to be non-trivial if at this point the sign of the derivative of the path, i.e. the conditional information, changes.

Lemma 2. ***(local minima of information paths):*** *if $X_k.I(X_1;..;X_{k-1}) < 0$ then all paths from 0 to I_k passing by I_{k-1} have at least one local minimum. In order for an information path to have a non-trivial critical point, it is necessary that $k > 3$, the smallest possible dimension of a critical point being $k = 3$.*

Proof: it is a direct consequence of the definitions of paths and of conditional 2-mutual-information $X_k.I_2$ positivity ($X_k.I_2 \geq 0$, cf. theorem 2.4.2.2)□.

Note that by definition a local minimum can be a global minimum. We will call, if it exists, the dimension k of the first local minimum of an information path the **first informational critical dimension** of the information path IP_i, and note it k_{i_1}. This allows us to define maximal information paths:
Positive information path (definition): A positive information path is an information path from 0 to a given I_k corresponding to a given k-tuple of variables such that $I_k < I_{k-1} < ... < I_1$.
Maximal Positive information path (definition): A maximal positive information path is a positive information path of maximal length. More formally, a maximal positive information path is a positive information path that is not a proper subset of positive information paths.

The definitions make coincide positive information paths and maximal positive information path with chains (faces) and maximal chains (facets), respectively. The maximal positive information path stops at the first local minimum of an information path, if it exists. The first informational critical dimension k_{i_1} of a time series IP_i, whenever it exists, gives a quantification of the duration of the memory of the system.

2.4.5. Sum over paths and mean information path

As previously, for $k = 1$, $IP_i(1)$ can be identified with the self-internal energy and for $k \geq 2$, $IP_i(k)$ corresponds to the k-free-energy of a single path IP_i. The chain rule of mutual information (equation

15) and the derivative of an IP_i path (equation 41) implies that the k-free-energy can be obtained from a single path:

$$I_k = I(X_1; ...; X_k; P) = I(X_1) - \sum_{i=2}^{k} X_i.I(X_1; ...; X_{i-1}) = IP_i(1) + \sum_{j=2}^{k} \frac{dIP_i(j)}{dj} \tag{43}$$

Hence, the global thermodynamical relation 29 can be understood as the sum over all paths, the sum over informational histories: the classical, discrete and informational version of the path integrals in statistical physic [128]. Indeed considering an inverse relation between time and dimension $t = \frac{1}{n}$ in the probability expression 2.2.1 for iid processes gives the usual expression of a unitary evolution operator $p_{\bullet\bullet ...k...\bullet} = e^{\frac{t.E_{ij...n}}{k}}$. Free-information-energy integrates over the simplicial structure of the whole lattice of partitions over degrees $k \geq 2$, which further justifies its free energy name.

In order to obtain a single state function instead of a group of $k!$ paths-functions, we can compute the mean behavior of the information structure, which is achieved by defining the mean H_k and I_k, noted $\langle H_k \rangle$ and $\langle I_k \rangle$:

$$\langle H_k \rangle = \frac{\sum_{T \subset [n]; card(T)=k} H_k(X_T; P)}{\binom{n}{k}} \tag{44}$$

and

$$\langle I_k \rangle = \frac{\sum_{T \subset [n]; card(T)=k} I_k(X_T; P)}{\binom{n}{k}} \tag{45}$$

For example, considering $n = 3$, then $\langle I_2 \rangle = \frac{I(X_1;X_2)+I(X_1;X_3)+I(X_2;X_3)}{3}$. This defines the mean mutual information path and a mean entropy path noted $\langle HP \rangle(k)$ and $\langle IP \rangle(k)$ in the information landscape. As previously, $\langle IP \rangle(1)$ can be identified with the mean self-internal energy $U(X^n_{hom}; P_N)$ and for $k > 1$ $\langle IP \rangle(k)$ to the mean k-free-information-energy $G(X^n_{hom}; P_N)$, giving the usual isotherm relation:

$$H(X^n_{hom}; P_N) = U(X^n_{hom}; P_N) - G(X^n_{hom}; P_N) \tag{46}$$

The computation of the mean paths corresponds to an idealized information structure X^n_{hom} for which all the variables would be identically distributed, would have the same entropy, and would share the same mutual information I_k at each dimension k: a homogeneous information structure, with homogeneous high-dimension k-body interactions. Like usually achieved in physic notably in mean-field theory (for example Weiss [129] or Hartree), it aims to provide a single function summarizing the average behavior of the system (we will see that in practice it misses the important biological structures, pointing out the constitutive heterogeneity of biological systems see 3.2.0.1). Prefiguring the application to data exposed in the next chapter, the $\langle IP \rangle(k)$ paths estimated on genetic expression data set are shown for population A and population B neurons in Figure 7. We quantified the gene expression levels for 41 genes in two populations of cells (A or B) as presented in section 5.1. We estimated H_k and I_k landscapes for these two populations and for two sets of genes ("genes of interest" and "non relevant") according to the computational and estimation methods presented in section 3. The available computational power restricts the analysis to a maximum of $n = 21$ variables (or 21 dimensions), and imposed us to divide the genes between the two classes "genes of interest" and "non relevant". The 21 genes of interest were selected within the 41 quantified genes according to their known specific involvement in the function of population A cells.
Figure 7 exhibits the critical phenomenon usually encountered in condensed matter physic, like the example of Van-der-Walls interactions [130]. Like any I_k path, $\langle IP \rangle(k)$ can have a first minimum with a critical dimension k_{i_1} that could be called the homogeneous critical dimension. For the 21 genes of interest (whose expression levels, given the literature, are expected to be linked in these cell types) the $\langle I_k \rangle$ path exhibits a clear minimum at the critical dimension $k_{i_1} = 4$ for population A neurons and $k_{i_1} = 5$ population B neurons, reproducing the usual free-energy potential in the condensed phase for

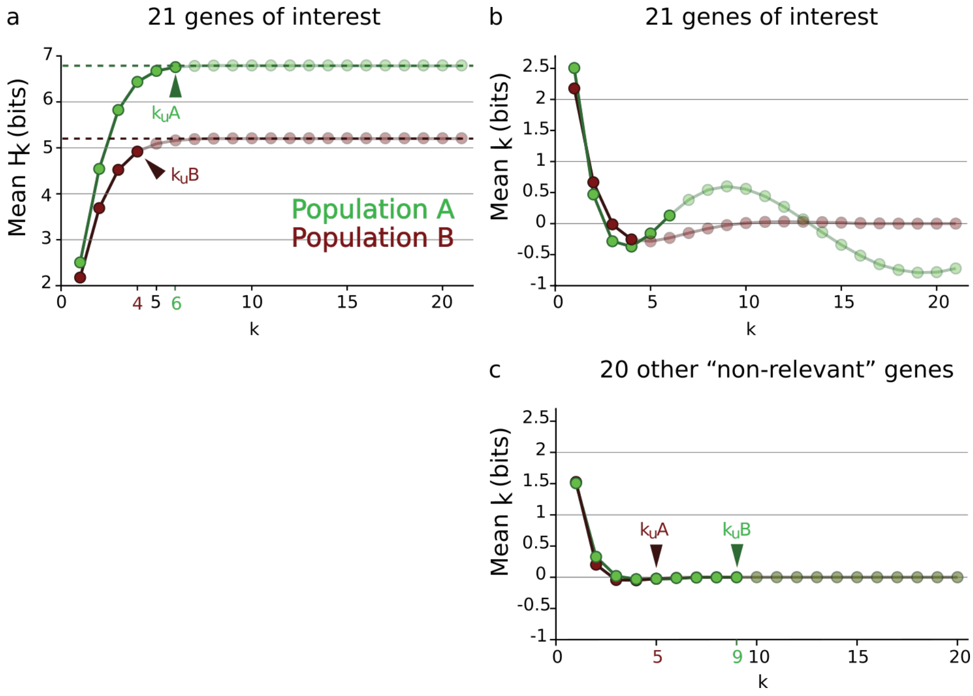

Figure 7. Example of mean entropy and information paths of gene expression. a, Mean entropy path $\langle H_k \rangle$ for the 21 genes of interest for population A (green line) and population B neurons (red line). **b,** Mean information path $\langle I_k \rangle$ for the same pool of genes. **c,** Mean information path $\langle I_k \rangle$ for the rest of 20 genes ("non relevant"). The undersampling dimension introduced in section 5.5.1 is depicted with arrows.

which n-body interactions are non-negligible. For the 20 other genes, less expected to be related in these cell types, the $\langle I_k \rangle$ path exhibits a monotonic decrease without a non-trivial minimum, which corresponds to the usual free-energy potential in the uncondensed-disordered phase for which the n-body interactions are negligible. Indeed, as shown in the work of Xie and colleagues [131], the tensor network renormalization approach of n-body interacting quantum systems gives rise to an expression of the free-energy as a function of the dimension of the interactions, in the same way than achieved here.

2.4.6. Minimum free energy complex

The analysis of information paths that we now propose aims to determine all the first critical points of information paths, in other words to determine all the information paths for which conditional information stays positive, and all first local minima of the information landscape. Such an exhaustive characterization would give a good description of the landscape and of the complexity of the measured system. The qualitative reason for considering only the first extrema for the data analysis is that, beyond that point, mutual information diverges (as section 2.4.4 explains) and the maximal positive information paths correspond to stable functional modules in the application to data (gene expression). A more mathematical justification is that they define the facets of a complex in our simplicial structure, which we will call the minimum energy complex of our information structure, underlining that this complex is the formalization of the minimum free energy principle in a degenerate case.

We now obtain the main theorem that our information path analysis aims to characterize empirically:

Theorem 2.4. ***(Minimum free energy complex):*** *the set of all positive information paths forms a simplicial complex that we call the minimum free energy complex. Moreover, the dimension-degree of the minimum free* *energy complex is the maximum of all the first informational critical dimensions ($d = \max k_{i_1}$), if it exists, or the dimension of the whole simplicial structure n. The minimum free energy complex is noted X^{+d}. A necessary condition for this complex not to be a simplex is that its dimension is greater or equal to four ($d \geq 4$).*

Proof: It is known that there is a one to one correspondence between simplicial complexes and their set of maximal chains (facets) (see [132] p.95 for example). The last part follows from Lemma 1. □.

In simple words, the maximal faces, e.g. the maximal positive information paths, encode all the structures of the minimum free energy complex. Figure 8 illustrates one of the simplest examples of a minimum free energy complex that is not a simplex, of dimension four in a five-dimensional simplicial structure of information Δ_5.

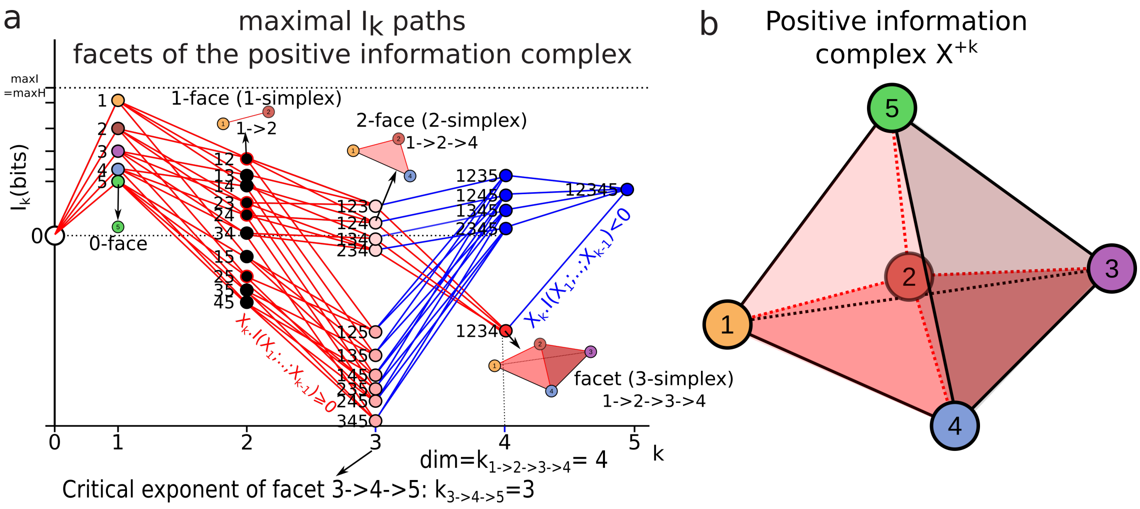

Figure 8. Example of maximal I_k paths in an I_k landscape for $n = 5$ together with its corresponding minimum free information energy complex. a, maximal I_k paths in an I_k landscape for $n = 5$. The maximum positive information paths are depicted in red, for example the paths $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ but also $4 \rightarrow 3 \rightarrow 2 \rightarrow 1$, $3 \rightarrow 4 \rightarrow 5$, and $1 \rightarrow 2 \rightarrow 5$ are maximum positive information paths, that is facets/maximal chains. The facet $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ is a 3-simplex while $3 \rightarrow 4 \rightarrow 5$ is a 2-simplex with critical dimension $k_{3 \rightarrow 4 \rightarrow 5} = 3$. The usual dimension of the simplex is used here, but we could have augmented it by one, since we added the constant element "0" to the algebra (pointed space), such that the usual simplicial dimension and the critical dimension correspond. The maximal critical dimension of the positive information paths is the dimension of the complex and hence $d(X^{+k}) = d(1 \rightarrow 2 \rightarrow 3 \rightarrow 4) = 4$. **b,** The minimum free energy complex corresponding to the preceding maximal i_k paths. It is a subcomplex of the 4-simplex also called the 5-cell with only one 4 dimensional cell among the five depicted as the bottom tetrahedron $\{1234\}$ with darker red volume. It has 5 vertices, 10 edges, 10 2-faces and one 3-face (cell), hence its Euler characteristic is $\chi(X^{+k}) = 5 - 10 + 10 - 1 = 4$ and its minimum free energy characteristic characteristic is: $H^{+k}(X^{+k}) = \sum_{X_i \in X^{+k}}^{5} I(X_i) - \sum_{(X_i;X_j) \in X^{+k}}^{10} I(X_i; X_j) + \sum_{(X_i;X_j;X_h) \in X^{+k}}^{10} I(X_i; X_j; X_h) - I(X_1; X_2; X_3; X4)$

We define the minimum free energy characteristic as:

$$H^{+k}(X^{+k}; P) = \sum_{i=1}^{k} (-1)^{i-1} \sum_{I \subset X^{+}; card(I) = i} I_i(X_I; P) \tag{47}$$

, where the component with dimension higher than one is a free energy. In the example of Figure 8 it gives:

$$\begin{aligned} H^{+k}(X^{+k}) = & \sum_{X_i \in X^{+k}}^{5} I(X_i) - \sum_{(X_i;X_j)\in X^{+k}}^{10} I(X_i;X_j) \\ & + \sum_{(X_i;X_j;X_h)\in X^{+k}}^{10} I(X_i;X_j;X_h) - I(X_1;X_2;X_3;X4) \end{aligned} \tag{48}$$

We propose that this complex defines a complex system:
Complex system (definition): A complex system is a minimum free energy complex.
It has the merit to provide a formal definition of complex systems as simple as the definition of an abstract simplicial complex can be, and to be quite consensual with respect to some of the approaches in this domain, as reviewed by Newman [133]. Notably, it provides a formal basis to define some of the important concepts in complex systems: emergence being the coboundary map, imergence the boundary map, synergy being information negativity, organization scales being the ranks of random-variable lattices, a collective interaction being a local minimum of free-energy, diversity being the multiplicity of these minima quantified by the number of facets, a network being a 1-complex, a network of network being a 1-complex in hyper-cohomology.
The interpretation in terms of sum over paths in the complex is direct as it sums over paths until they diverge, the divergence being here the negativity of conditional mutual information. We called it the minimum free energy complex but could have called it instead the positive or instantaneous complex because its facets appear as the boundaries of the "present" structure, but it obviously contains all the past-history and the memory of the structure (notably encoded in the negative I_k that are necessarily non-Markovian). The topological formalization of the minimum energy allows the coexistence of numerous local minima, a situation usually encountered in complex systems (slow aging) such as frustrated glasses and K-sat problems [8,134] which settings correspond here to the case of n binary random variables, $N_1 = ... = N_2 = 2$. The existence of the frustration effect, due to the multiplicity of these local minima in the free energy landscape [135], has also been one of the main difficulties of the condensed matter theory. Matsuda could show that I_k negativity is a signature of frustration [20]. The first axioms of DFT consider that probability densities of n' elementary bodies are each in a 3-dimensional space [68,69], defining a whole simplicial structure of dimension $n = 3n'$, commonly called the configuration space. When considered with the physical axiom of a configuration space, the theorem 2.4 implies that, while the minimum free information energy complex of an elementary body can only be a simplex, the configuration space of n' elementary bodies can be a complex with (quite) arbitrary topology. In simple terms, this settles the elementary components of the configuration space as 3-simplices, which composition can give arbitrarily complicated k-complexes. This idea is in resonance with the triangulations of space-time that arose notably from the work of Wheeler [76] and Penrose [136], like spin foams [137] and causal sets [138], while we only considered here classical probabilities.

3. Application to Gene expression data - detection of cell types and gene modules

The developments and tests of the estimation of simplicial information topology on data is made on a genetic expression dataset of two cell types obtained as described in the section material and methods 5. The result of this quantification of gene expression is represented in "Heat maps" and allows two kinds of analysis:

- The analysis with genes as variables: in this case the "Heat maps" correspond to (m, n) matrices D (presented in the section 5.2) together with the labels (population A or population B) of the cells. The data analysis consists in the detection of gene modules.

- The analysis with cells (neurons) as variables: in this case the "Heat maps" correspond to the transposed matrices D^T (presented in the Figure 11) together with the labels (population A or population B) of the cells. The data analysis consist in the detection of cell types.

The theoretical and estimation framework presented here has been developed for a finite dimension n without asymptotic corrections. This section justifies empirically this finite dimensional naive estimation and presents an example of information topology analysis allowing a comparison with the previous MaxEnt approaches that originally inspired this work. As investigated in detail in the materials and methods section 5.5, the effective computation of information topology on data is confronted with two severe but classical problems, i) the computational complexity of the algorithm and ii) the finite sample size in relatively high dimensional space, the curse of dimension. The materials and methods section presents the important computational restrictions we had to concede and the statistical tests we developed to circumvent the undersampling problem. In consequence, we underline that the presented results are not exact but partial approximations of the dependence structure (a simplicial subcase) and of the minimum energy complex.

3.1. Mutual-information negativity, clusters and links

We give first a short introduction to information decompositions, to the appearance of negative information and related phenomena. The interpretation of information negativity, like any missing quantity in physic, is not obvious. Here, we established that entropy and the information decomposition originally proposed by Hu Kuo Ting in 1962 [14] (see also [1,15,16,18,139]) are characterized uniquely when one considers a discrete probability space equipped with a structure of multivariate random variables. As demonstrated by Hu Kuo Ting [14], for $k \geq 3$, I_k can be negative. The possible negativity of I_k functions arising from this "classical" decomposition has posed serious problems of interpretations, and it was the main argument for the theoretical studies to reject such a function for measuring information dependences and statistical interactions (although the total C_k are positive). Notably, it motivated the proposition of non-negative decomposition by Williams and Beer [140] and of "unique information" by Bertschinger and colleagues [141,142], or Griffith and Koch [143]. These partial decompositions of information are the subject of many recent investigations notably with applications to the development of neural network [144], and neuromodulation [145]. The idea of a negative information quantifying certainty (or order) instead of uncertainty (or disorder) was developed by Brillouin [146], Wiener who defined information with a negative sign [147], and Schrödinger in his book "what is life?" where he argued that "living systems feed upon negentropy" [148] or "free energy" (in side note). According to our identification of I_k with free-energy, negative I_k is a negative energy component. Originally in physic, negative energies appeared as the solution of Dirac equation and were identified as positron, motivating the Dirac's Hole theory [149]. In a classical context, negative pressure had appeared at the minimum of free-energy in the critical phase of gaz with n-body interactions, the Van-der-Walls model [130], which further allowed to derive the Casimir force [150]. Feynman dedicated an article to classical negative probabilities [151] with the aim of solving the problem of infinities raised by renormalization in quantum field theories. Adami and Cerf then showed that quantum conditional entropies and I_2 can be negative, and that it happens precisely for entangled systems according to Bell inequalities [152,153]. The framework of Adami and Cerf provides a natural way to generalize the present classical structures to the quantum case (see also [1]) and allows the interpretation that classical I_k negativity detects classical entanglement-like relations. According to their results, Bell inequalities impose that the "non-Shannonian" (quantum) cone in information landscapes should happen below the third dimension: in contrast to what we presented for the classical minimum energy complex 2.4.6, quantum entanglement should allow quantum minimum free energy complexes that are not simplex to happen in dimension below 3 (a promising topological insight into the possible non-locality of the configuration space). In spin glasses study, Matsuda showed that I_3 negativity provides a signature of frustrated interactions [20,154], a result recently extended by Sootla and colleagues [155]. Interestingly, Matsuda also showed that the maximum of I_2 detects the phase transition of ising systems, in agreement with renormalization predictions [154](see also [155]). In biological studies, Brenner and colleagues have observed and quantified an equivalent definition of I_3 negativity in the spiking activity of neurons and called it synergy [156], and Anastassiou and colleagues unraveled I_3 negativity within gene expression, suggesting cooperativity in gene regulation and new targets for cancer therapy [157,158]. It can be easily shown that modules with negative I_k are necessarily non-Markovian (cf. appendix A.3) and that negativity hence can characterize some form of distributed memory. From the logical and computational point of view, I_3 negativity has been associated to XOR boolean operator relation between 3 binary variables [159,160], as illustrated by the configurations detected by the two minima of I_3 in Figure 9, namely $\{001, 010, 100, 111\}$ and $\{000, 011, 101, 110\}$. Such interpretation is partial and only holds for binary variables.

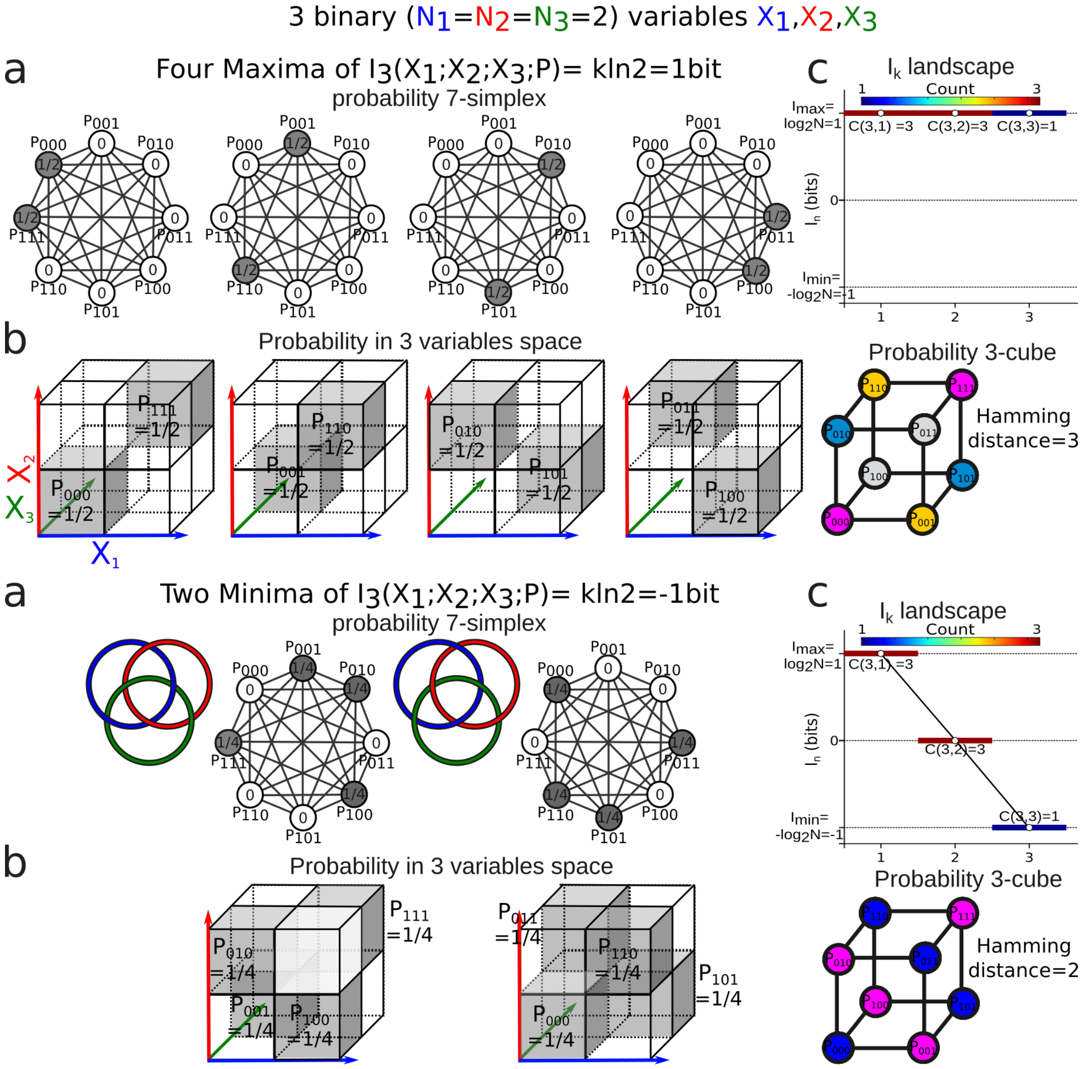

Figure 9. **Example of the 4 maxima (left panel) and of the 2 minima of I_3 for 3 binary variables** **a,** informal representation of the 7-simplex of probability associated with 3 binary variables. The values of the atomic probabilities that achieve the extremal configurations are noted in each vertex. **b,** Representation of the associated probabilities in the data space of the 3 variables for these extremal configurations. **c,** Information I_k landscapes of these configurations (top). Representation of these extremal configurations on the probability cube. The colors represents the non-nul atomic probability of each extremal configuration (bottom).

Figure 9 illustrates that such minimal negative information configurations provide a clear example of purely emergent and collective interactions analog to Borromean links in topology, since it cannot be detected from any pairwise investigation or 2-dimensional observation. The variables are pairwise independent but dependent at 3. In general I_k negativity detects such effects of projection on lower dimensions, and illustrates the main difficulty when going from dimension 2 to 3 in information theory. The example given in Figure 9 provides a simple example of this dimensional effect in the data space: the alternated clustering of the data corresponding to I_3 negativity cannot be detected by the projections onto whichever subspace of pair of variables, since the variables are pairwise independent. For N-ary variables the negativity becomes much more complicated, with more degeneracy of the minima and maxima of I_k.

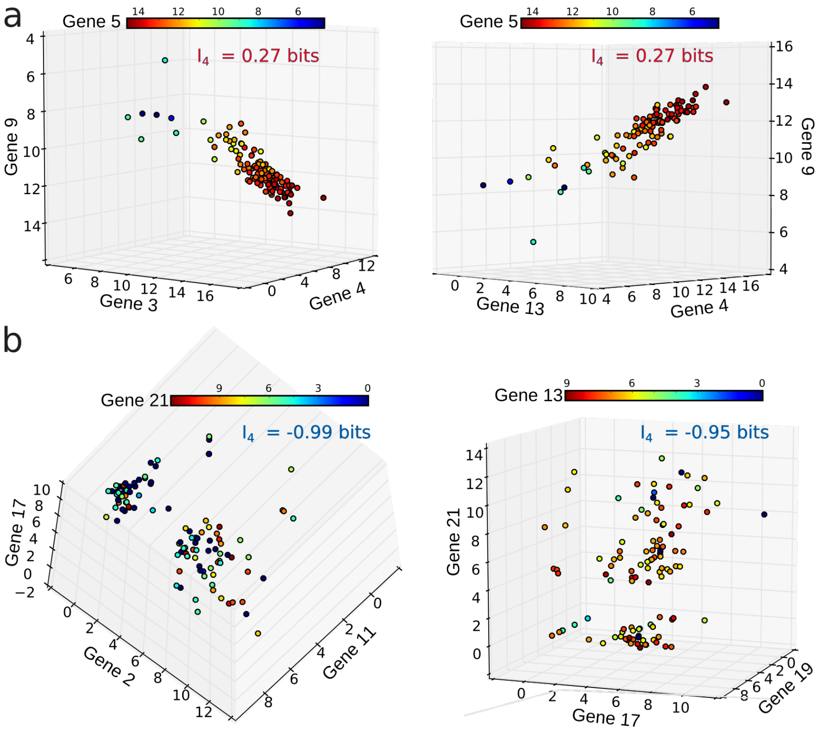

Figure 10. Examples of some of 4-modules (quaduplets) with the highest (positive) and lowest (negative) I_4 of gene expression represented in the data space. a, Two 4-modules of genes sharing among the highest positive I_4 of the gene expression data set (cf. 5.1). The data are represented in the data space of the measured expression of the 4 variables-genes. The fourth dimension-variable is color coded. **b,** Two 4-modules of genes sharing among the lowest negative I_4. All the modules were found to be significant according to the dependence test introduced in section 5.5.2, except the module $\{17, 19, 21, 13\}$. The identified extremal modules (different) give similar patterns of dependences [161].

In order to illustrate the theoretical examples of Figure 9 on real data, considering the data set of gene expression (matrix D), we plotted some quadruplets of genes sharing some of the highest (positive) and lowest (negative) I_4 values in the data space of the variables (Figure 10). Figure 10 shows that in the data space, I_k negativity identifies the clustering of the data points, or in other words, the modules (k-tuples) for which the data points are segregated into condensate clusters. Such a data clustering is indeed treated here as an arbitrary data point generalization of the condensation phenomenon studied in statistical physic. As expected theoretically, I_k positivity identifies co-variations of the variables, even in cases of non-linear relations, as shown by Reshef and colleagues [67] in the pairwise case. In classical physic, considering the variables of the configuration space, the phenomenon of condensation into droplets provides an obvious example of negativity and clustering. As a result, the interpretation of the negativity of I_k is that it provides a signature and quantification of the variables that segregate or differentiate the measured population. Question: recently Kauffman expressed the relation of link homology and Khovanov homology with simplicial homotopy theory [162]. Can we express link homology and Khovanov homology in information cohomology and generalize the configurations of minimal I_3 to higher k in such a framework? A positive answer should recover the knot and link expressions of the statistical physic models, namely Potts model, obtained by Jones [163].

3.2. Cell type detection - comparison with previous MaxEnt studies

Example of cell type recognition with a low sample size $m = 41$, dimension $n = 20$, and graining $N = 9$.

The information landscape and path analysis corresponding to the analysis with cells as variables is illustrated in Figure 11. It comes to consider the cells as a realization of gene expression rather than the converse, the "selfish gene" point of view [164]. In this case, the data analysis task is to recover blindly the pre-established labels of cell types (population A and population B) from the topological data analysis. The heat-map transpose matrix of $n = 20$ cells with $m = 41$ genes is represented in Figure 11a. We took $n = 20$ neurons among the 148 within which 10 were pre-identified as population A neurons (in green) and 10 were pre-identified as population B neurons (in dark red), and ran the analysis on the 41 gene expression with a graining of $N = 9$ values (cf. section 5.1). The dimension above which the estimations of information become too biased due to the finite sample size is given by the undersampling dimension $k_u = 11$ (p value 0.05, cf. section 5.5.1). The landscapes turn out to be very different from the extremal (totally disordered and totally ordered) homogeneous (identically distributed) theoretical cases presented in Figure 4. This discrepancy illustrates a general principle of the organization of living systems "in between crystal and smoke", originally proposed by Atlan [165]. The I_k landscape shown in Figure 11c exhibits two clearly separated components. The scaffold below represents the tuple corresponding to the maximum of I_{10}: it corresponds exactly to the 10 neurons pre-identified as being population A neurons. As introduced in previous section 3.1 and further shown in Figure 11, the k-tuples presenting the highest and lowest information (I_k) values are the most relevant modules biologically. Hence, biologically relevant modules are the furthest from the mean values (the $\langle IP \rangle (k)$ path, cf. section 2.4.5), and that a mean or homogeneous formalism will unavoidably miss such extremal structures. This further underlines the constitutive heterogeneity of biological systems.

As detailed in materials and methods 5.4, the computation of the minimum free energy complex being NP hard, we applied a heuristic that only considers information paths with maximal or minimal slope at each vertex of a path, in order to detect the maximal positive information path (the facets of the complex, cf. 2.4.4), and the result can only give a partial estimation of the complex and of its facets. The maximum (in red) and minimum (in blue) I_k information paths identified by the algorithm are represented in Figure 11d. The scaffold below represents the two tuples corresponding to the two longest maximum paths in each component: the longest (noted Max IP_{11} in green) IP_{11} contains the 10 neurons pre-identified as population A and 1 "error" neuron pre-identified as population B. We restricted the longest maximum path to the undersampling dimension $k_u = 11$, but this path reached $k = 14$ with erroneous classifications. The second longest maximum path (noted Max IP_{11} in red) IP_{11} contains the 10 neurons pre-identified as population B and 1 neuron pre-identified as population A that is hence erroneously classified by the algorithm. Altogether the information landscape shows that population A neurons constitutes a quite homogenous population, whereas the population B neurons corresponds to a more heterogeneous population of cells, a fact that was already known and reported in the biological studies of these populations. The histograms of the distributions of I_k for $k = 1, .., 12$, shown in Figure 11e are clearly bimodal and the insets provide a magnification on the population A component. As detailed in the section materials and methods 5.5.2, we developed a test based on the random shuffles of the data points that leaves the marginal distributions unchanged, as proposed by [5]. It estimates if a given I_k significantly differs from a randomly generated I_k, a test of the specificity of the k-dependence. The shuffled distributions and the significance value for $p = 0.1$ are depicted by the black lines and the doted lines, as in Figure 16. As underlined in

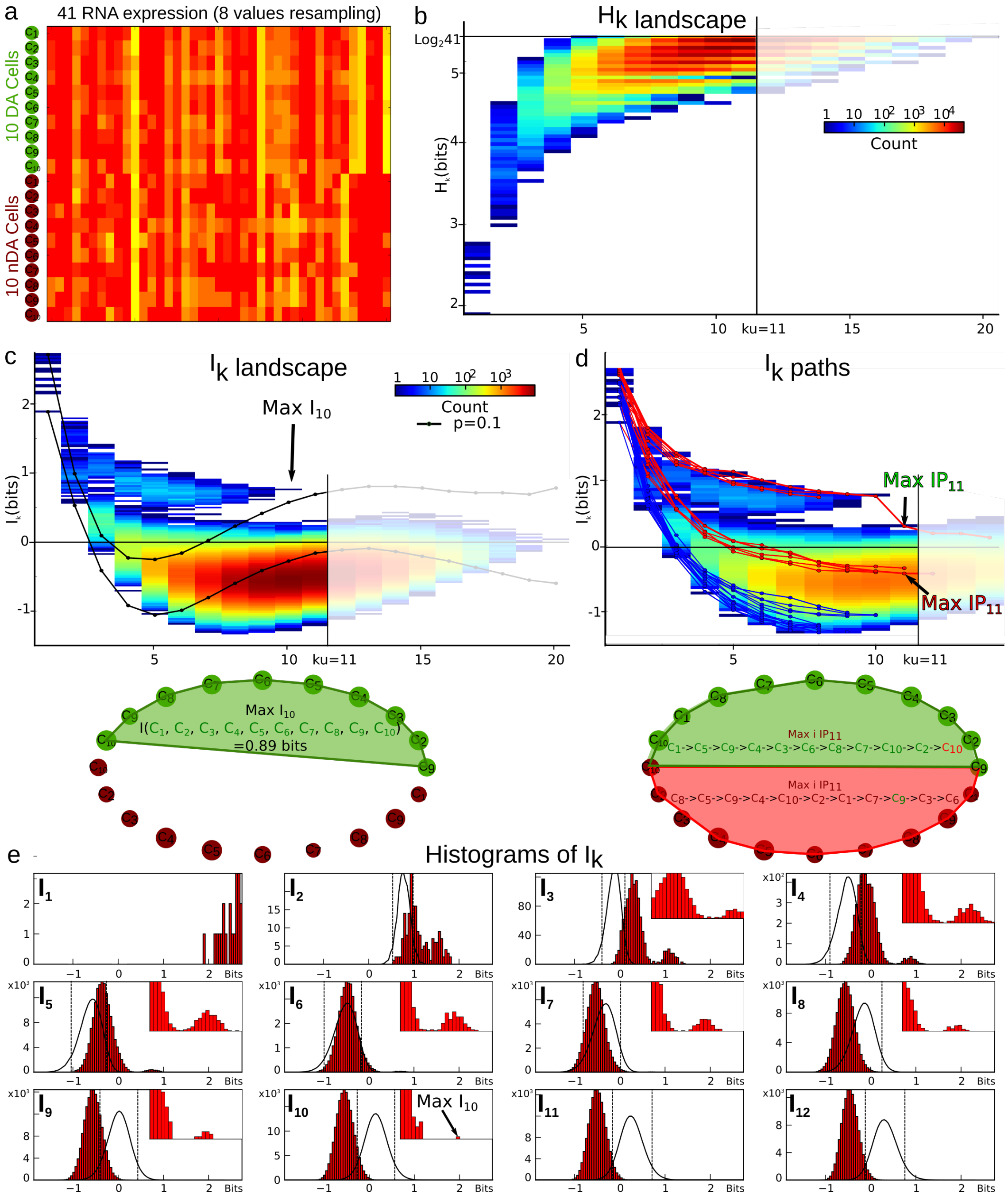

Figure 11. Example of a I_k landscape and path analysis. a, heatmap (transpose of matrix D) of $n = 20$ neurons with $m = 41$ genes.**b,** the corresponding H_k landscape. **c,** the corresponding I_k landscape **d,** maximum (in red) and minimum (in blue) I_k information paths. **e,** histograms of the distributions of I_k for $k = 1, .., 12$. See text for details.

conclusion 4.3 and illustrated in the histograms of Figure 11e and in [161], if these results show that higher dependencies can be important, they do not mean that pairwise or marginal informations are not: the consideration of higher dependencies can only improve the efficiency of the detection obtained from pairwise or marginal considerations. Thermodynamic established that free energy is the energy that can be effectively used, whereas entropy was also called lost heat. The same applies to data with mutual-informations: mutual-informations are the informations that can be effectively used for pattern detection and segmentation. These results suggest that there are important resources of such information-energy in the k-body dependences.

Comparison with previous MaxEnt studies

The analysis we just showed was achieved with 41 points in 20 dimensions and N=9. The usual estimation of information on data [166] stated that a very large amount of data is required to estimate probability and information [167,168]. The justification is based on a maximum entropy assumption, namely that the distribution respects a maximum entropy principle under constraints such as first or second moment. Then, under the assumption of an iid process, it is possible to estimate the size of the typical set in 2^H, as it bounds the probability. In this condition, the estimation is considered as biased if the sample size m is below, or of the order of the size of the typical set. Applying Bayesian reasoning, Nemenman and colleagues [167] proposed methods for finite bias corrections based on the inferred probability distribution obtained by maximizing the entropy under the constraints of observed marginals and pairwise (or higher) statistics. Applying the calculation of a reasonable sampling size according to the priors and method of [166] in the example of Figure 11 gives $m > 2^{49.27} = 6.8.10^{14}$ trials, since the entropy computed on cell's marginal distributions considered as independent (the sum of the marginal entropies) is 49.27 bits in this example. However, despite the strikingly low number of points ($41 << 6.8.10^{14}$) in a relatively high-dimensional space, it provides impressive results regarding the preclassification, demonstrating that multivariate information analysis can be achieved in what previous methods would have considered to be the undersampling regime. Such conclusion agrees with the report of Margolin and colleagues [10], as discussed later.

4. Discussion

4.1. Complexity through finite dimension non-extensivity

Statistical physic without statistical limit?

The measure of entropy and information rate on data (the evolution of entropy H_k when the number of variables k increases) has a long history. Originally, in the work of Strong and colleagues [166], and as usual in information theory and statistical physic, it was considered that the "true" entropy was given in the asymptotic limit $\lim_{n\to\infty} H_n$ under stationarity or stronger assumptions. As explained in section 1.3 (see also the note of Kontsevitch [25], and the work of Baez, Fritz and Leinster [31]), entropy does not need asymptotic or infinite assumptions such as Stirling approximation to be derived, an observation that we tried to understand, explore, and exploit here. Rather than being interested in the asymptotic limit (the infinite dimensional case) and absolute values of information, the present analysis focuses on the finite version of the "slow approach of the entropy to its extensive asymptotic limit" that Grassberger [169], and then Bialek, Nemenman and Tishby proposed to be "a sign of complexity" [170], "complexity through non-extensivity" (see also Tsallis [171]). In short, we consider the non-extensivity of information before considering its asymptotic limit. Considering a statistical physic without statistical limit could be pertinent for the study of "small" systems, which concerns biological systems. Their small size allows them to harness thermal fluctuations, and impose their investigation with out-of equilibrium methods, as exposed in the work of Ritort and colleagues, reviewed in [172]. The H_k and I_k landscapes presented here give a detailed expression of the "signs of complexity" and non-extensivity, for such small size systems (finite dimension k), and give a finite dimensional geometric view of the "slow approach of the entropy to its extensive asymptotic limit". In a sense, what replaces here the large number limits, Avogadro number consideration (etc.), is the combinatorial explosions of the different possible interactions: in the same way as in Van Der Walls paradigm, a combinatorial number of weak interactions can lead to a strong global interaction. In practice, like for any empirical investigation, the finite dimensional case imposes restrictions on the

conclusions, and basically reminds us that we have not measured everything. Some relevant random variables for the observed system may be absent from the analysis, and adding such variables could reveal new interactions that are effectively constitutive and relevant to the system. Among all possible structures of data, one is universal: data and empirical measures are discrete and finite, as emphasized by Born [173], and fully justify the cohomological approach used here, which was originally constructed to handle in a common framework the Lie and Galois theory, continuous and discrete symmetries. Such finite elementary formalism provides a basis, such that a handling of asymptotic limits may be safely achieved by algebraic extensions, exactly following Hilbert's recommendations on infinite formalism derivations [174]: "adjoin to finite statements ideal statements", algebraically from p to 0 characteristic.
Questions : can we extend the formalism to infinite (or profinite) dimensional structures of information, such that it allows to recover central limit theorems and asymptotic equipartition property in the associated special case? Note that such an extension may erase an important part of what biologists usually consider as relevant structures, as further discussed. Does the asymptotic limit of the graining $N \to \infty$ allow to recover the Lie group cohomology?

Naive estimations let the data speaks.

One of the striking results of this data analysis concerns the relatively low sample size ($m = 41$ and $m = 111$ for the analysis with cells as variable and with genes as variables respectively) required to obtain satisfying results in relatively high dimensions ($k = 10$ and $k = 6$ respectively). Satisfying results means here that they predict already known results reported in the biological literature, or in agreement with experts labels. In [10], Nemenman and colleagues, who developed the problematic of the sampling problem, state in the introduction "entropy may be estimated reliably even when inferences about details of the underlying probability distribution are impossible. Thus the direct estimation of dependencies has a chance even for undersampled problems" and conclude that "a major advantage of our definition of statistical dependencies in terms of the MaxEnt approximations is that it can be applied even when the underlying distributions are undersampled". The present analysis agrees and confirms their conclusion. The method applied here is quite elementary. It does not make assumptions of an expected or true distribution, of maximum entropy distribution or pairwise interaction Hamiltonian, coupling constant or metric, of stationarity or ergodicity or iid process, Markov chain, or underlying network structure (...) or whatever prior that would speak in place of the data. It just considers numerical empirical probabilities as expressed by Kolmogorov axioms ([11] chap.1), which he called the "generalized fields of probability" because it does not assume the 6th axiom of continuity. Rather than fixing a model with priors, the present formalism allows the raw data to impose freely their specific structure to the model, what is usually called the naive approach or naive estimation. If one accepts that a frequentist theory and interpretation of probability is mathematically valid ([11], chap.1), one may then conclude that a frequentist theory of entropy and information may also hold, and moreover directly fulfills the usual requirement of observability in theoretical physic [173]. This frequentist elementary consideration is not trivial mathematically notably when considered from the number theoretic point of view. For example, the combinatoric of integer partitions of m could be investigated in the general information structure (partition) context, which up to our knowledge has not been achieved in the context of probability and information.

A dual interplay between cell type and gene module detection?

The two kinds of analysis, cell type and gene module detection, are dual in the sense of linear algebra, since they apply respectively on a matrix D and its transpose D^T. In a sense, gene interactions tell the identity and diversity of cell populations, and cell interactions tell the identity and diversity of functional gene expression modules. We believe this dual interplay to be quite generic in complex systems. However this duality is beyond the present work, as we do not have yet the formalism allowing to combine the informational analysis of a matrix D and its transpose D^T in a single framework. In the present case where $n \neq m$, they are not isomorphic in general, and the question of the relation between their respective analyse is left as an open question concerning linear algebra. Such a study of the swapping between m and n will obviously also involve the undersampling dimension estimation.

4.2. Epigenetic topological learning - biological diversity

In place of the MaxEnt principle, we proposed an almost synonymous least energy principle equivalent here to a homological complex (finite and without metric assumptions). Mathematically, we took profit of the fact that whether the maximum of entropy functional is always unique and in a sense normative, the minima of I_k functionals exhibit a rich structure of degeneracy, generated by the "non-Shannonian set" [115–117] and conjectured to be at least as rich as topological links can be. We proposed that this multiplicity of minima accounts for biological diversity, or more precisely that the number of facets of this complex quantifies the diversity in the system. The application to cell type identification 3.2.0.1 gives a preliminary validation of such quantification. Moreover, the definition of a complex system as the minimum free-energy complex given in section 2.4.6, underlining that diversity is just the multiplicity of the minima, is in agreement with Waddington's original work [175] (see Figure 12b). In the allegory of Waddington's epigenetic landscapes, whatever the ball, it will always fall down, a statement that can be assimilated to the second law of thermodynamic. But doing so, it will be able to take different paths: diversity comes from the multiple minima. The explanation by Waddington of such landscape is a "complex system of interactions" that can be formalized by the minimum free energy complex with interactions corresponding to the I_k. Moreover, formalisms assuming that the variables are identically distributed, as for the homogeneous systems described in the section on mean paths 2.4.5, will display a single first minima (one facet, a simplex), and hence no diversity. Sharing the same aims, Teschendorff and Enver, and then Jin and colleagues, proposed an alternative interpretation of Waddington's landscape in terms of signaling entropy [176] and of probability transitions [177], respectively.

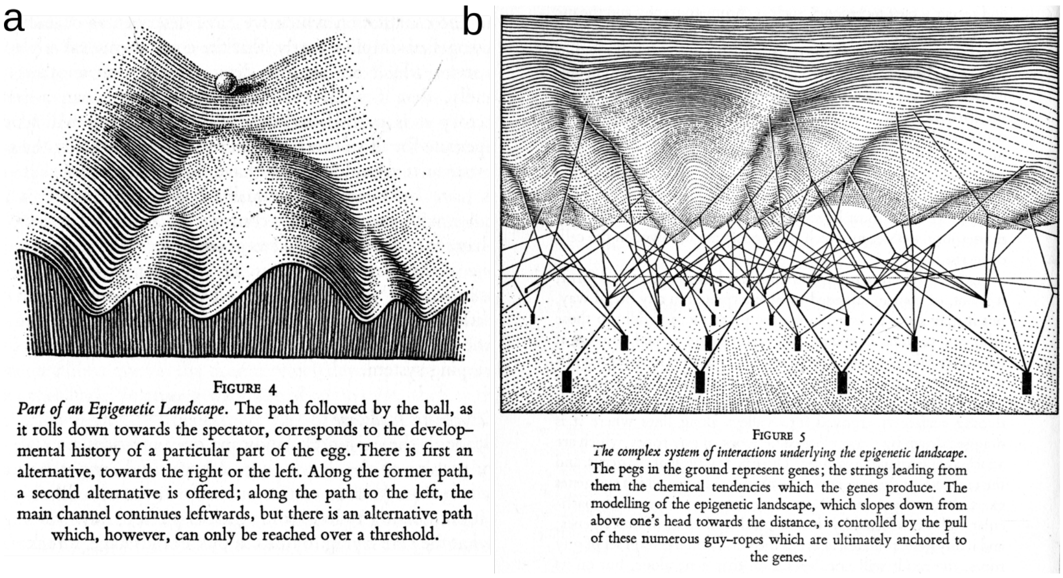

FIGURE 4

Part of an Epigenetic Landscape. The path followed by the ball, as it rolls down towards the spectator, corresponds to the developmental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

FIGURE 5

The complex system of interactions underlying the epigenetic landscape. The pegs in the ground represent genes; the strings leading from them the chemical tendencies which the genes produce. The modelling of the epigenetic landscape, which slopes down from above one's head towards the distance, is controlled by the pull of these numerous guy–ropes which are ultimately anchored to the genes.

Figure 12. The epigenetic landscape of Waddington. a, The epigenetic landscape of Waddington, a path of the ball in this landscape illustrates a cell developmental fate. **b,** "The complex system of interactions underlying the epigenetic landscape" with Waddington's original legends [175].

Following Thom's topological morphogenetic view of Waddington's work [178], we propose that I_k landscape, paths and minimum free energy complex provide a possible informational formalization of Waddington's epigenetic complex landscape and cell fates (cf. Figure 12). This formalization of

doi:10.20944/preprints201804.0157.v1

Waddington's epigenetic view is consistent with the machine learning formalization of Hebbian epigenetic plasticity. From the pure formal view, the models of Hebbian neural learning like Hopfield's network, Boltzmann machines, the Infomax models proposed by Linsker, Nadal and Parga, Bell and Sejnowski [179–181]) can be viewed as binary variables subcases of a generic N-ary variable epigenetic developmental process. For example, Potts model were implemented for the simulation of cell-based morphogenesis by Glazier and colleagues [182]. Hence the topological approach can allow the treatment of neural learning and development on the ground of a common epigenetic formalism, in agreement with biological results pointing out the continuum and "entanglement" of the biological processes underlying development and learning [183]. In terms of the current problematics of neuroscience, such generalization allows on a formal level to consider an analog coding in place of a digital coding, and the methods developed here can be applied to studies investigating (discrete) analog coding.

Moreover, following all the work of these last decades on the application of statistical physic to biological systems (some of them cited in this article), we propose that the epigenetic process implements the first two laws of thermodynamics, which weak topological versions are proposed to hold in the raw data space (without phase space or symplectic structure, cf. section 2.4.3). As previously underlined, the condition for such an inscription of living organism dynamic into classical statistical physic to be legitimate is that the considered variables correspond to phase space variables.

4.3. Information beyond pairwise interaction and homological complex beyond complex networks.

During the last decades, there have been important efforts in trying to evaluate the pairwise and higher order interactions in neuronal and biological measurements, notably to extract the undergoing collective dynamic. Following some application of the MaxEnt principle on ising spin models to neural data [184,185], most of the studies concluded that pairwise statistics are sufficient to capture the global collective dynamics, leading to the "pairwise sufficiency" paradigm (see Merchan and Nemenman for presentation [168]). Closely related to the present study, Margolin, Wang, Califano and Nemenman have started to investigate multivariate dependences of higher order [10] with MaxEnt methods, using the total-correlation or multi-information C_k (the total free-energy, cf. equation 7 and section on k-independence 1.3.6). These functions are indeed the same as the one defined by Tononi and Edelman [17] to measure consciousness and complexity in brain dynamics. They are also implemented in the Infotopo program. Since the I_k and C_k functions are closely related (cf. 7 and section 1.3.6), in practice, the methods exposed here provide some formal developments and views on their methods, combinatorial, homological and free energy aspects, and reveal information negativity and its clustering correlate. The formalism developed here underlines that biological structures can be seen as discrete-finite random processes, that can be equivalently seen as symmetric permutation groups or finite symmetries, with an action of expectation-conditioning. As illustrated in Figure 11, the present analysis shows that in the expression of 41 genes of interest of population A neurons, the higher order statistical interactions are non-negligible and have a simple functional meaning of collective module, a cell type. We believe such conclusion to be generic in biology. More precisely, we believe that biological structures are higher-order statistical interactions, and that these interactions provide the signature of their memory engramming. As we previously stated, the quantification of the information storage applied here to genes can be considered as a generic epigenetic memory characterization, resulting of a developmental-learning process. The consideration of higher dimensional statistical dependences increases combinatorially the number of possible information modules engrammed by the system. It hence provides an appreciable capacity reservoir for information storage and for differentiation, for diversity. For finite systems and data, the relative contributions of k-interactions shall depend on the system. These finite dimensional interactions are hopefully not universal in the usual sense of statistical physic: the peculiar biological mechanisms and functions of population A neurons are not the same as the ones of cells from other tissue.

Some important questions concern the behavior of the informational structure when the dimension n tends to infinity (question 4.1.0.1). It is possible that in the infinite dimensional limit, all these different biological structures simplify and falls in a common universal class accounted by a simple pairwise interaction Hamiltonian poised at criticality, as proposed by Mora and Bialek [186]. This would explain the discrepancy of their proposition, that "models based on pairwise interactions be powerful enough to capture the behavior of biological systems"[187], with our observations. Based on infinite model axioms and priors, it may be possible to argue that the observations made here are finite size effects, and even to consider them as artefacts. Based on finite model and axioms following Galois, and centrally using the beautiful mathematical construction of derivatives that holds in the finite domains (the coboundaries of Galois's cohomology, appearing here as mutual-informations), we argue that biological structures may also be understood as finite size effects, a natural fact rather than an artefact. In our opinion, it shall not be understood that what we presented in this article is different in essence from the MaxEnt approach, we just unraveled its topological nature, avoiding unnecessary assumptions and further underlining that the consideration of higher interactions may improve the precision and performance of studies restricted to pairwise interactions. The precise contribution of higher order is indeed directly quantified by the I_k values in the landscapes and paths. Figure 13 further illustrates the gain and the importance of considering higher statistical interactions, using the previous example of cells pre-identified as 10 population A and 10 population B cells ($n = 20$, $m = 47$, $N = 9$). The plots are the finite and discrete analogs of Gibbs's original representation of entropy vs. energy [188]. Whereas pairwise interactions ($k = 2$) cannot (or very hardly) distinguish the population A and population B cell types, the maximum of I_{10} unambiguously identifies the population A.

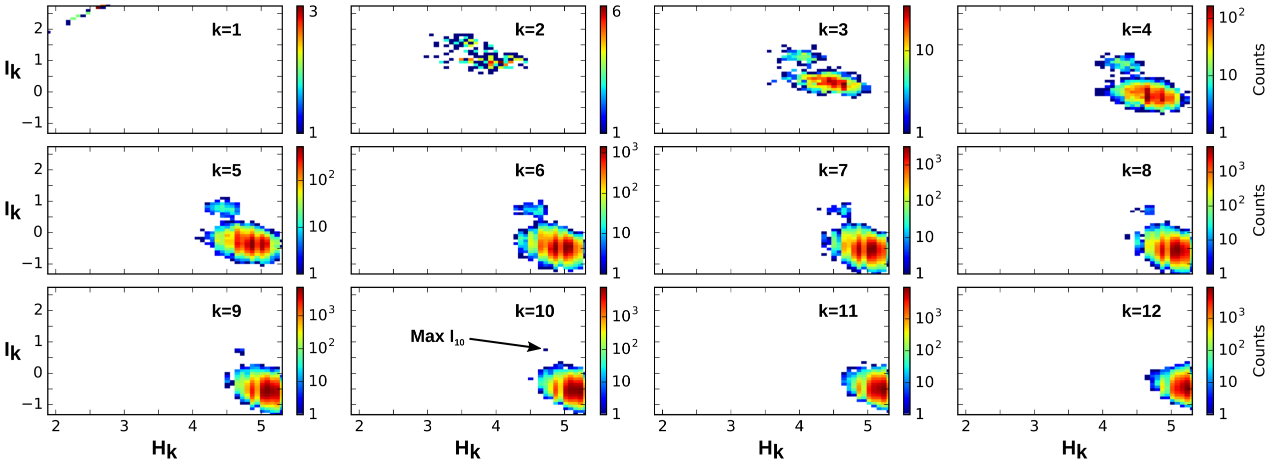

Figure 13. $H_k - I_k$ **landscape: Gibbs-Maxwell's entropy vs. energy representation.** H_k and I_k are plotted in abscissa and ordinate respectively for dimension $k = 1, ..., 12$ for the same data and setting as in Figure 11 ($n = 20$ cells, $m = 47$ genes, $N = 9$, $k_u = 11$). Compare the difficulty in identifying the 2 cells types from the pairwise $k = 2$ landscape to the $k = 10$ landscape.

A major part of the research in quantitative and theoretical biology during the last decades has been dedicated to biological interaction networks (protein, genetic expression, neural networks...). From the physical point of view, none of these systems are networks (1 dimensional graphs). It should be reminded that network structure is a simplifying assumption, a kind of first order approximation. Cohomological complexes of random variables and multivariate mutual information generalize the complex network descriptions by considering and quantifying higher statistical interactions, which appear here (e.g. Figure 13) to be biologically relevant and functional.

5. Materials and Methods - Computation of Simplicial Information Cohomology

5.1. The dataset: quantified genetic expression in two cell types

The quantification of genetic expression was performed using microfluidic qPCR technique on single dopaminergic (DA) and non-dopaminergic (nDA) neurons isolated from two midbrain structures, the Substantia Nigra pars compacta (SNc) and the neighboring Ventral Tegmental Area (VTA), extracted from adult TH-GFP mice (transgenic mice expressing the Green Fluorescent Protein under the control of the Tyrosine Hydroxylase promoter). The precise protocols of extraction, quantification, and identification are detailed in [161]. This technique allowed us to quantify in a single cell the levels of expression of 41 genes chosen for their implication in neuronal activity and identity of dopaminergic (DA) neurons. The SNc neurons (DA) were identified based on GFP fluorescence (TH expression). This identification was further confirmed based on the expression levels of *Th* and *Slc6a3* genes, which are established markers of DA metabolism. The quantification of the expression of the 41 genes ($n = 41$) was achieved in 111 neurons ($m = 111$) identified as dopaminergic (DA) and in 37 neurons ($m = 37$) identified as non-dopaminergic (nDA). In this article, for readability purpose, we replaced the name of the genes by gene numbers and the cell type DA by population A, and the cell type nDA by population B. The dataset is available in supplementary material 5.5.3.

5.2. Probability estimation

The presentation of probability estimation procedure is achieved on matrices D (genes as variables), and it is the same in the case of the analysis of the matrices D^T (cells as variables). It is illustrated in Figure 14 for the simple case of 2 random variables taken from the dataset of gene expression presented in section 5.1, namely the expression of two genes Gene5 and Gene21 in $m = 111$ population A cells. Our probability estimation corresponds to a step of the integral estimation procedure of Riemann.

We write the heatmap as a (m, n) matrix D and its real coefficients $x_{ij} \in \mathbb{R}$, $i \in \{1..m\}$, $j \in \{1...n\}$: the columns of D span the m repetitions-trials (here the m neurons) and the rows of D spans the n variables (here the n genes). We also note, for each variable X_j, the minimum and maximum values measured as $\min x_j = \min_{1 \leq i \leq m} x_{ij}$ and $\max x_j = \max_{1 \leq i \leq m} x_{ij}$.

We consider the space in the intervals $[\min x_j, \max x_j]$ for each variable X_j and divide it into $N_1.N_2...N_n$ boxes, on which it is possible to estimate the atomic probabilities by elementary counting. We note each n-dimensional box by an n-tuple of integers $\{a_1, ..., a_n\}$ where $\forall i \in \{1, ..., n\}$, $a_i \in \{1, ..., N_i\}$, and writing the min and the max of a box on each variable X_j (the jth co-ordinate of the vertex of the box) as $\text{bmin}_j = \min x_j + \frac{(a_j - 1)(\max x_j - \min x_j)}{N_j}$ and $\text{bmax}_j = \min x_j + \frac{(a_j)(\max x_j - \min x_j)}{N_j}$, then the atomic probabilities can be defined using Dirac function δ as:

$$\begin{aligned} &P\left(\text{bmin}_1 \leq X_1 \leq \text{bmax}_1, \text{bmin}_2 \leq X_2 \leq \text{bmax}_2, ..., \text{bmin}_n \leq X_n \leq \text{bmax}_n\right) \\ &\quad = \sum_{i=1}^{m} \frac{\delta_i}{m}, \; \delta_i = \begin{cases} 0, \text{ if } \text{bmin}_1 > x_{i1} \text{ or } x_{i1} > \text{bmax}_1 \text{ ...or } \text{bmin}_n > x_{in} \text{ or } x_{in} > \text{bmax}_n \\ 1, \text{ if } \text{bmin}_1 \leq x_{i1} \leq \text{bmax}_1 \text{ and...and } \text{bmin}_n \leq x_{in} \leq \text{bmax}_n \end{cases} \end{aligned} \tag{49}$$

For two variables, using the definition of conditioning $P_X(Y) = \frac{P(X.Y)}{P(X)}$ and in the general case using the theorem of total probability [11] ($P(X) = \sum_{i=0}^{N} P(A_i.X) = \sum_{i=0}^{N} P(A_i).P_{A_i}(X)$), we can marginalize, or geometrically project on lower dimensions, to obtain all the probabilities corresponding to subsets of variables, as illustrated in Figure 14. For example, with short notation, the probability

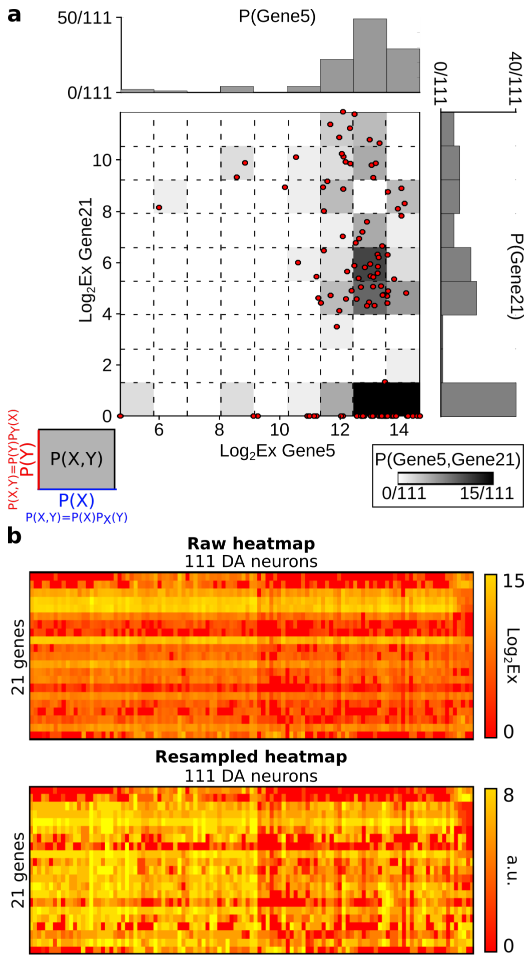

Figure 14. Principles of probability estimation for 2 random variables. a, illustration of the basic procedure used in practice to estimate the probability densitiy for the two genes ($n = 2$) Gene5 and Gene21 in 111 population A neurons ($m = 111$) using a graining of 9 ($N_1 = N_2 = 9$). The data points corresponding to the 111 observations are represented as red dots, and the graining is depicted by the 81-box grid ($N_1.N_2$). The borders of the graining interval are obtained by considering the maximum and minimum measured values for each variable, and data are then sampled regularly within this interval with N_i values. Projections of the data points on lower dimensional variable subspaces (X_1 and X_2 axes here) are obtained by marginalization, giving the marginal probability laws for the 2 variables X_1 and X_2 ($P_{X_i,N_i,m}$) ; represented as histograms above the X_1-axis for Gene21 and on the right of the X_2-axis for Gene21). b, heatmaps representing the levels of expression of the 21 genes of interest on a $\log_2 Ex$ scale (top, raw heatmap) and after resampling with a graining of 9 (bottom, $N_1 = N_2 = ... = N_{21} = 9$).

associated to the marginal variable X_i being in the interval $[\text{bmin}_i, \text{bmax}_i]$ is obtained by direct summation:

$$P\left(\text{bmin}_i \leq X_i \leq \text{bmax}_i\right) = \sum_{i=1}^{N_1...\widehat{N_i}...N_n} P\left(\text{bmin}_1 \leq X_1 \leq \text{bmax}_1, \text{bmin}_2 \leq X_2 \leq \text{bmax}_2, ..., \text{bmin}_n \leq X_n \leq \text{bmax}_n\right) \quad (50)$$

In the example of Figure 14, the probability of the level of *Th* being in the 4th box is:

$$P\left(8 \leq \mathrm{Th} \leq 9.8\right) = \sum_{i=0}^{8} P\left(8 \leq \mathrm{Th} \leq 9.8, \mathrm{bmin}_2 \leq \mathrm{Calb1} \leq \mathrm{bmax}_2\right) = 2/111 + 2/111 \quad (51)$$

In terms of geometry, the probability defines an $N_1.N_2...N_n - 1$ dimensional simplex $\Delta_{N_1.N_2...N_n-1}$ (the -1 accounts for the normalization equation $\sum P_i = 1$, which moreover imposes that the geometry is affine). In the example of Figure 14, we have an 80-dimensional probability simplex Δ_{80}, each k-face of the simplex representing the boolean algebra of the joint-probabilities, which is equivalent in the finite case to the sigma-algebra. In our analysis, we have chosen $N_1 = N_2 = ... = N_n = 9$ and this choice is justified in section 5.5 using Reshef and colleagues criterion [67] and undersampling constraints.

In summary, our probability estimation and data analysis depend on n (the number of random variables), on m (the number of observations), and on $N_1, ..., N_i$ (the graining). The merit of this method is its simplicity (few assumptions, no priors on the distributions) and low computational cost. There exist different methods that can significantly improve this basic probability estimation, but we leave this for future investigation. The graining given by the numbers $N_1.N_2...N_n$ and the sample size m are important parameters of the analysis explored in section 5.5.

5.3. Computation of k-Entropy and k-Information landscapes

The computational exploration of the simplicial sublattice has a complexity in $\mathcal{O}(2^n)$ ($2^n = \sum_{k=1}^{n} \binom{n}{k}$). In this simplicial setting we can exhaustively estimate information functions on the simplicial information structure, that is joint-entropy H_k and mutual-informations I_k at all dimensions $k \leq n$ and for every k-tuple, with a standard commercial personal computer (a laptop with processor Intel Core i7-4910MQ CPU @ 2.90GHz × 8, even though the program currently uses only one CPU) up to $k = n = 21$ in a reasonable time ($\approx$ 3 hours). Using the expression of joint-entropy (equation 3) and the probability obtained using equation 49 and marginalization, it is possible to compute the joint-entropy and marginal entropy of all the variables. The alternated expression of n-mutual information given by equation 17 then allows a direct evaluation of all these quantities. The definitions, formulas and theorems are sufficient to obtain the algorithm. We moreover provide the Information Topology program INFOTOPO-V1.2 under opensource licence on github depository at https://github.com/pierrebaudot/INFOTOPO. Information Topology is a program written in Python (compatible with Python 3.4.x), with a graphic interface built using TKinter [189], plots drawn using Matplotlib [190], calculations made using NumPy [191], and scaffold representations drawn using NetworkX [192]. It computes all the results on information presented in the current study, including the information paths and approximation of the minimum information energy complex, statistical tests of I_k values described in the next sections, but also k-pseudovolumes A3 and the finite entropy rate $\frac{H_k}{k}$. The input is an excel table containing the data values, e.g. the matrix D with the first row and column containing the labels. Here, we limited our analysis to $n = 21$ genes of specific biological interest.

5.4. Computation of the minimum free energy complex

Computationally the exploration of the set of information paths is hard, since finding a global functional extremum or all the first critical dimensions k_{i_1} falls into the NP-hard class and the computational complexity of information minimizing algorithms is $\mathcal{O}(n!)$, although it can be reduced using dynamic programing to $\mathcal{O}(n^2 2^n)$. Evaluating a function on this simplicial information structure, in order for example to find its extrema, comes to explore the $n!$ paths, or if we restrict to paths of dimension k, the $n!/(n-k)!$ paths starting at the least element and ending at one of the edges of

dimension k. Even in this partial case, the computation becomes impossible in practice with usual personal computational resources. We hence have to find a heuristic to reduce significantly the computational exploration.
From the biological point of view, not all the local minima appear equally interesting, some of them potentially corresponding to information paths with low values of mutual information I_k, while biologically relevant functional modules appear to correspond to the highest and lowest values of I_k. Each k-face of the simplicial structure connects to $n-k$ different $k+1$-faces by the conditional information $X_{k+1}.I(X1;..;X_k)$, and we chose to explore only the two connections (morphisms) with lowest and highest positive values of $X_{k+1}.I(X1;..;X_k)$. This local heuristic finds the maximal positive information paths that display the highest and lowest I_k values at each element of a path. It starts at one of the I_1, takes the path of maximum or of minimum slope and iterates until it stops at the minimum (whenever the conditional mutual information starts to be negative) and then ranks the paths as a function of their length. The paths with highest I_k values are called maximum I_k paths while the paths with lowest I_k values are called minimum I_k paths. It is very fast with a computational complexity in $\mathcal{O}(2n)$ once the landscape is computed. Considering the chain rule 15, it finds the paths that keep on maximizing or minimizing the free-energy component of a single path $I_k - I(X_1) = -\sum_{i=2}^{k} X_i.I(X_1;...;X_{i-1})$ with the constraint that the conditional mutual information stays positive. It hence finds some biologically relevant maximal positive information paths, but without any guarantee to find them all. We underline that this algorithm is a biologically driven heuristic and we have no theorem justifying its necessity, uniqueness. It gives a partial estimation of the minimum free energy complex that may be richer and of greater dimensionality than the results presented here. Considering the case of cells as variables, Figure 11d presents the longest paths of maximum and minimum I_k paths. In the case where the maximum paths identified by the algorithm are the same up to permutation, as it happens for some paths shown in Figure 11d, it indicates that attributing a particular ordering to the module has no pertinence.

5.5. Finite methods: sample size m, graining N and dimension n

The computational hardnesses of the homological estimation are far from being the last problems in the estimation. The data in practice are always finite, and probability estimation in high dimension can be severely biased by the limited size of the sample (the number of empirical points, m, or trials). In order to circumvent these biases, we propose two statistical tests that bound the estimated values to significant values in the dimension and in the information values (the abscissa and ordinates of the landscapes). These methods have the merit to be very simple for biological use.

5.5.1. Undersampling dimension

The information data analysis presented here depends on the two variables N and m. The finite size of the sample m is known to impose an important bias in the estimation of information quantities: in high-dimensional data analysis, it is quoted as the Hugues phenomenon [193] and in entropy estimation it has been called the sampling problem since the seminal work of Strong and colleagues [166–168]. Paninski developed a complete framework for information estimation [194], which unfortunately does not apply here, since an important part of his results assumes binary variables and independent identically distributed (iid) processes such that central limit theorems can apply and asymptotic convergence is obtained. The asymptotic convergence and divergence would be very interesting to investigate from the cohomological point of view, but it requires the development of the infinite dimensional case, a whole open subject. However, as argued previously, our intuition is that the conditional mutual information and its sign will be the central function in such a study. Following the original presentation of the sampling problem by Strong and colleagues [166], the extreme cases of sampling are given by:

- When $N_1 = N_2 = ... = N_n = 1$, there is a single box Ω and $P(\Omega) = m/m = 1$ and we have $H_k = I_k = 0, \forall k \in 0, ..., n$. The case where $m = 1$ is identical. This fixes the lower bound of our analysis in order not to be trivial; we need $m \geq 2$ and $N_1 = N_2 = ... = N_n \geq 2$.
- When $N_1.N_2...N_n$ are such that only one data point falls into a box, m of the values of atomic probabilities are $1/m$ and $N_1.N_2...N_n - m$ are null as a consequence of equation 50, and hence we have $H_n = \log_2 m$.

The latter case is the extreme signature of undersampling. Whenever this happens for a given k-tuple, all the HP_k paths passing by this k-tuple will stay on the same information values since conditional entropy is non-negative: we have $H_k = H_{k+1}$ or equivalently $(X_1, ..., X_k)H(X_{k+1}) = 0$, and all $k + l$-tuples are deterministic (a function of) with respect to the k-tuple. This is typically the case illustrated in Figure 11: adding a new variable to an undersampled k-tuple is equivalent to adding the deterministic variable "0" since the probability remains unchanged $(1/m)$.

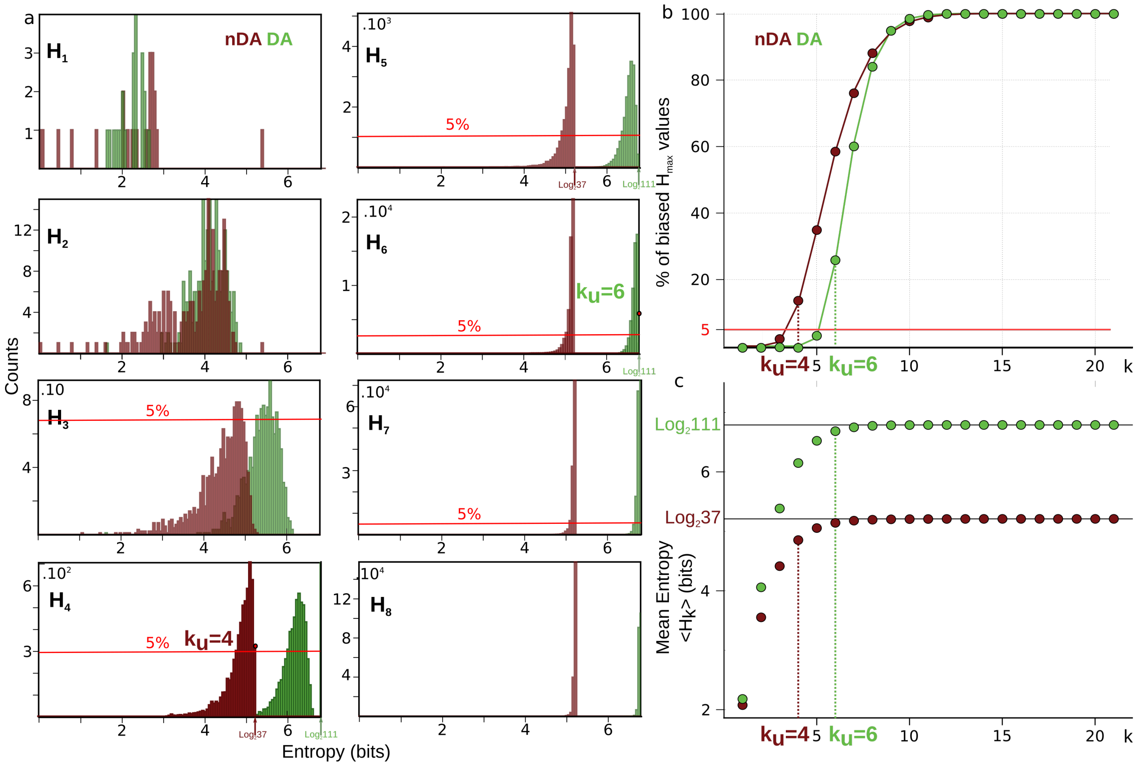

Figure 15. Determination of undersampling dimension k_u. **a,** distributions of H_k for $m = 111$ population A neurons (green) and $m = 37$ population B neurons (dark red) for $k = 1, .., 6$. The horizontal red line represents the threshold we have fixed to 5 percent of the total number of k-tuples. **c,** Plot of the percent of maximum entropy $H_k = \ln m$ biased values as a function of the dimension k. The horizontal red line represents the threshold fixed to 5 percent, giving $k_u = 6$ for population A and $k_u = 4$ for population B neurons. **c,** The mean $\langle HP \rangle(k)$ paths for these two populations of neurons, the maximum entropy $H_k = \ln m$ is represented by plain horizontal lines.

Considering the analysis of cells as variables (matrix D^T), the signature of this undersampling is the saturation at $H_k = \log_2 41$ observed in the H_k landscape in Figure 15b, starting at $k = 5$ for some 5-tuples of neurons. Considering the analysis of genes as variables (matrix D), the mean entropy computed also shows this saturation at $H_k = \log_2 111$ for population A neurons and $H_k = \log_2 37$ for population B neurons. This undersampling also affects mutual-information and conditional mutual-information estimation by adding a combinatorial number of 0 values (independence) to the I_k landscape, which is observable in the theoretical Figure A1a,b in appendix. We propose to define a dimension k_u as the dimension for which the probability p_u of having the H_k at the biased value of

$H_k = \log_2 m$ is above 5 percent ($p_u = 0.05$). As shown for the analysis of cells as variables in Figure 15, this basic estimation gives here $k_u = 6$ for population A neurons and $k_u = 4$ for population B neurons. The information structures identified by our methods beyond these values can be considered as unlikely to have a biological or physical meaning and shall not be interpreted. Since undersampling affects mainly the distribution of I_k values close to 0 value, the maxima and minima of I_k and the maximal and minimal information paths below k_u are the least affected by the sampling problem and the low sample size. This will be illustrated in the next sections.

5.5.2. k-dependence test

Pethel and Hahs [5] have constructed an exact test of 2-dependence for any pair of variables, not necessarily binary or iid. Iid condition is usually assumed for the χ^2 test, but does not seem relevant for biological observations (we give an example here with genetic expression, see [161] for the distributions of all gene variables exemplifying this fact). It allows to test the significance of the estimated I_2 values given a finite sample size m, the null hypothesis being that $I_2 = 0$ or 2-independence according to Pethel and Hahs (we provide below another interpretation). We follow here their presentation of the problem, and provide an extension of their test to arbitrary k (higher dimensions), with the null hypothesis being the k-independence $I_k = 0$, according to their interpretation. Even in the lowest dimensions, and below the undersampling bound, the values of I_k estimated from a finite sample size m are considered as biased [5]. As depicted in Figure 4a, if one considers an infinite sample ($m \to \infty$) of n independent variables, we then have for all $k \geq 2$ $I_k = 0$. However, if we randomly shuffle the values such that the marginal distributions for each variables X_i are preserved, the estimated I_k can be very different from 0, with distributions of I_k values not centered on 0. Figure 16 illustrates an example of such bias with $m = 111$ for the analysis with genes as variables. Reproducing the method of Pethel and Hahs [5], we designed a shuffling procedure of the n variables which consists in randomly permuting the measured values (co-ordinates) of each variable one by one in the matrix D or D^T (geometrically, a "random" permutation of the co-ordinates of each data point, point by point). Such a shuffle leaves marginal probabilities invariant while Pethel and Hahs [5] postulate that statistical dependences between the variables should be destroyed (see the alternative interpretation bellow). Figure 16 gives an example of the joint and marginal distributions before and after shuffle for two genes. Extending the 2-test of [5] to $k \geq 2$, the I_k values obtained after shuffling provide the distribution of the null hypothesis, k-independence ($I_k = 0$) according to [5]. The task is hence to compute many shuffles, 10.000 in [5], in order to obtain these "null" distributions. The exact procedure of Pethel and Hahs [5] would require to obtain such "null" distribution for all the 2^n tuples, which would require a number of shuffled trials impossible to obtain computationally. We hence propose a global test that consists in computing 17 different shuffles of the 21 genes, giving "null" distribution of shuffled I_k values composed of $21 \times \binom{n}{k}$. For example, the test of 2-dependence and 3-dependence will be against a null distribution with $21 * 210 = 3750$ I_2 values and $21 * 1330 = 22610$ I_3 values respectively. We fix a p value above which we reject the null hypothesis (a significance level, fixed at $p = 0.05$ in [5]), allowing to determine the statistical significance thresholds as information values for which the integral of the null distribution reaches the significance level $p = 0.05$. This holds for $k = 2$, as described in [5], but since for $k \geq 2$ I_k can be negative, the test becomes symmetric on the distribution, and hence for $k \geq 2$ we choose a significance level of $p = 0.1$ in order to stay consistent with the 2-dependence test. The "null" distributions and the threshold given by the significance p-value of rejection are illustrated in Figure 16d. If the observed values of I_k are above or below these threshold values, we reject the null hypothesis.

In practice, a random generator is used to generate the random permutations (here the NumPy generator [191]), and the present method is not exempt from the possibility that it generates statistical dependences in the higher degrees.

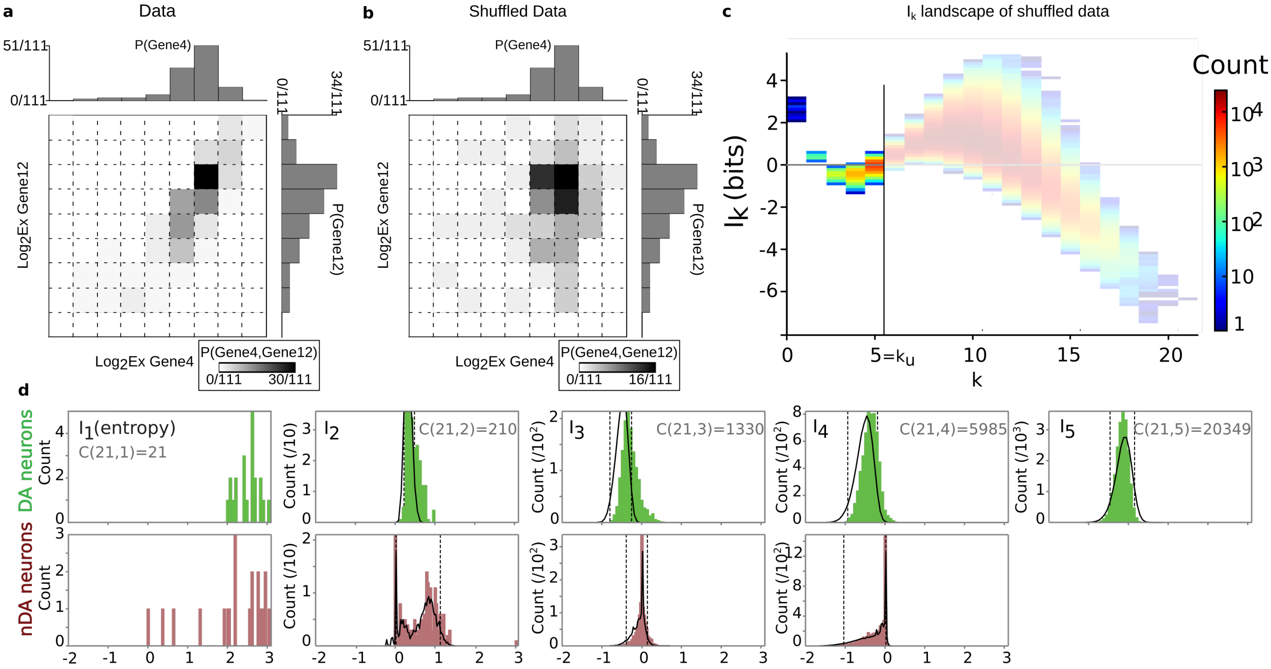

Figure 16. Probability and Information landscape of shuffled data. The figure corresponds to the case of analysis with genes as variables. **a,** joint and marginal distributions of two genes (genes 4 and 12) for $m = 111$ population A neurons. **b,** joint and marginal distributions after a shuffling of the values of expression of each gene. **c,** the estimated I_k landscape for the expression of 21 genes after shuffling. **d,** histograms representing the distribution of I_k values for all the degrees until $k = 5$ for population B. The total number of combinations C(n,k) for each degree (number of pairs for I_2; number of triplets for I_3, etc) is given in gray. The averaged shuffled values of information obtained with 17 shuffles are represented on each histogram as a black line, and the statistical significance threshold values for $p = 0.1$ are represented as vertical dotted lines.

Interpretation of the dependence test. The original interpretation of the test by Pethel and Hahs was that the null hypothesis corresponded to independent distributions. This view was motivated by the statement that "permutation destroys any dependence that may have existed between the datasets but preserves symbol frequencies". However, while derangements (permutations without fixed point) may destroy statistical dependencies, the other permutations, like the identity, include fixed points and therefore may preserve some of the statistical dependences. The definition of independence holds for finite and discrete variables as settled by Kolmogorov in his axiomatization ([11]), without requiring asymptotic m considerations: with full certainty it is possible to say if a finite m outcomes of a pair of dice are independent or not. Let's consider the simplest example of two fully redundant binary variables (two dices) like any pair of variable of Figure 9c, with atomic probability $\{P_{00} = 1/2, P_{01} = 0, P_{10} = 0, P_{11} = 1/2\}$, such configuration can be obtained with 2 trials ($m = 2$) and the two variables are fully dependent. The random shuffling will give any of the 2 possible configurations leaving invariant the marginals, namely the previous one (identity permutation) and $\{P_{00} = 0, P_{01} = 1/2, P_{10} = 1/2, P_{11} = 0\}$. In these two cases the variables are fully dependent, and hence random shuffles express the two possible fully dependent configurations. Considering more complex examples, we see immediately that in the finite context, random shuffles generate all possible statistical dependences that preserve the marginal. In fact, the handling of the concept of randomness in the context of finite and statistical objects is delicate. In the finite setting, the common deterministic definition of the randomness K of a string is non-computable, and since Zvonkin and Levin [195], it is well known that Shannon entropy of binary iid variables equals the averaged randomness-complexity K in the limit of infinitely long strings (see Th. 5.18 for precise statement [195]). We propose that for a

given finite m, permutations express all the possible statistical dependences that preserve symbol frequencies (much-like some finite random monkeys typewriter that write any books, in the article of E.Borel on statistical mechanic and irreversibility [196]). This statement basically corresponds to what we observe in Figure 16. Hence we propose that in finite context the null-hypothesis corresponds to a random k-dependence. The meaning of the presented test is hence a selectivity or specificity test: a test of an I_k of given k-tuple against a null hypothesis of "randomly" selected k-statistical dependences that preserve the marginals and m. Moreover, consistent with a frequentist view of probability, the direct values of I_k can be interpreted as a statistical significance test, against the null hypothesis of k-independence $I_k = 0$. We propose that in the finite context, topological information data analysis is itself a statistical significance analysis of the dependences, and that the rank established by homological tools indicates the most significant k-dependences. In the finite context, the question of the stability of the estimations of the dependences, that is their dependence on m and hence, how much one can trust the estimated dependencies, can be achieved using variational methods introduced in next section.
The formalism developed here heavily relies on the assimilation of Galois ambiguity and permutation groups with probabilistic-informational uncertainties, and proposes an alternative interpretation of the shuffling methods. The context of the general information structure is simpler with respect to the effects of the permutations of the probability atoms than the simplicial case investigated here: the work of Fresse established that the lattice of partitions is directly equipped with an action of the symmetric group [41].

5.5.3. Sampling size and graining landscapes - stability of minimum energy complex estimation

Figure 17 gives a first simple study of how robust the paths of maximum length are with respect to the variations of m and N, in the case of the analysis of genes as variables. The limit $N \rightarrow \infty$ recovers Riemann integration theory and gives the differential entropy with the correcting additive factor N (theorem 8.3.1 [19]).

The information paths of maximal length identified by our algorithm are relatively stable in the range of $N = 5, 7, 9, 11$ and $m = 34, 56, 89, 111$ where the m cells where taken arbitrarily among the 111 population A neurons. If we consider that the paths that only differ by the ordering of the variables are equivalent, then the stability of the two first paths is further and largely improved. The undersampling dimension obtained in these conditions is $k_u(m = 34) = 5$, $k_u(m = 56) = 6$, $k_u(m = 89) = 6$, $k_u(m = 111) = 6$ and $k_u(N = 5) = 8$, $k_u(N = 7) = 7$, $k_u(N = 9) = 6$, $k_u(N = 11) = 5$. In general, information landscapes can be investigated with the additional dimensions of N and m together with n. It allows to define our landscapes as iso-graining landscapes and to study the appearance of critical points in this larger space in the same way as isotherms in usual thermodynamic. In practice, to study more precisely the variations of information depending on N and m and to obtain a 2-dimensional representation, we plot the mean information as a function of N and m together with n, as presented in Figure 18a. We call the obtained landscapes the iso-graining I_k landscapes. The choice of a specific graining N can be done using this representation: a "pertinent" graining should be at a critical point of the landscape, consistent with the proposition of the work of Reshef and colleagues [67], who used maximal information coefficient (MI_2C) depending on the graining (with a more elaborated graining procedure) to detect pairwise associations. We have chosen to illustrate the landscapes with $N = 9$ according to this criterion and the undersampling criterion, because the I_2 values are close to their maximal values and the sampling size is not too limiting, with a $k_u = 6$ (see Figure 18a). Moreover, this choice of graining size $N = 9$ is sufficiently far from the critical point to ensure that we are in the condensed phase where interactions are expected. It is well below the analog of the critical temperature (the critical graining size), which according to the Figure 18a happens at $N_c = 3$ (the N for which the critical points cease to be trivial). In general, there is no reason why there should be only one "pertinent" graining (for example in glasses, the slow aging can be modeled

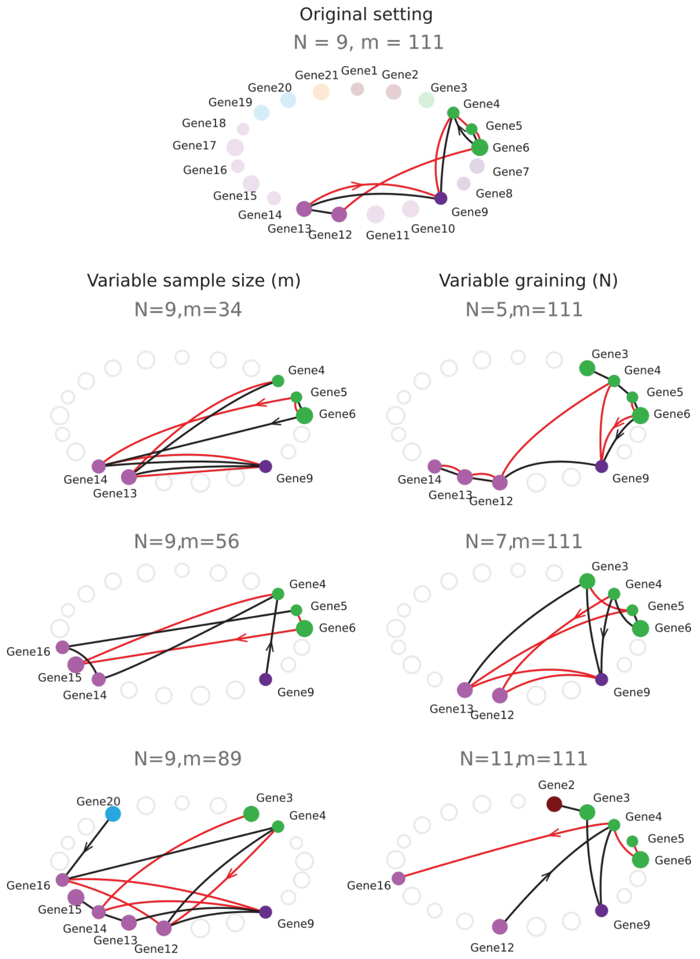

Figure 17. Effect of changing sample size and graining on the identification of gene modules. The figure corresponds to the case of analysis with genes as variables for the population A neurons. The positive I_k paths of maximum length were computed for a variable number of cells (m,left column) and a variable graining (N, right column). For clarity, only the two positive paths of maximum length are represented (first in red, second in black) for each parameter setting and the direction of each path is indicated by arrowheads. The two positive paths of maximum length for the original setting ($N = 9$, $m = 111$) are represented on the scaffold at the top of the figure for comparison. Smaller samples of cells (one random pick of 34, 56 and 89 cells) and larger ($N = 11$) or smaller ($N = 5, N = 7$) graining than the original ($N = 9$) were tested. Although slight differences in paths can be seen (especially for $N = 11$), most of the parameter combinations identify gene modules that strongly overlap with the module identified using the original setting.

with several effective temperatures [197]). The graining algorithm could be improved by applying direct methods of probability density estimation [198]. Finer methods of estimation (graining) have been developed by Reshef and colleagues [67] in order to estimate pairwise mutual-information, with interesting results. Their algorithm presents a lower computational complexity than the estimation on the lattice of partitions, but a higher complexity than the simple one applied here.
What we call the iso-sampling size I_k landscapes is presented in Figure 18b for mean I_k. Such investigation is also important since it monitors what is usually considered as the convergence (or divergence) in probability of the informations. For the estimations below the k_u represented here, the information estimations are quite constant as a function of m, indicating the stability of the estimation with respect to the sample size.

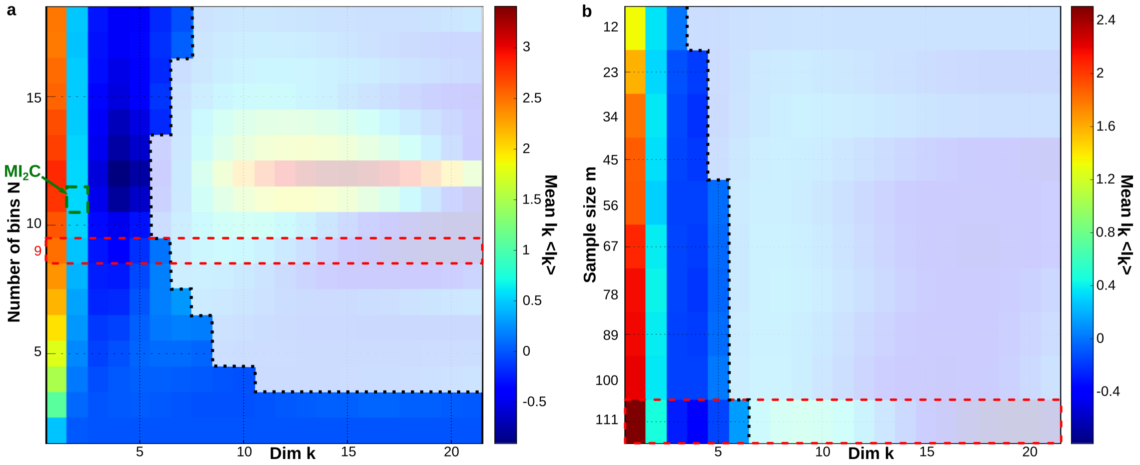

Figure 18. Iso-sample-size (m) and iso-graining mean $\langle IP\rangle(k)$ landscapes. The figure corresponds to the case of analysis with genes as variables for the population A neurons. **a,** The mean $\langle IP\rangle(k)$ paths are presented for $N = 2, ..., 18$ and $n = 21$ genes for the $m = 111$ population A neurons. The "undersampling" region beyond the k_u is shaded in white and delimited by black dotted line (the k_u was undetermined for $N = 2, 3$). For $N = 2$ the mean $\langle IP\rangle(k)$ path has no non-trivial minimum (monotonically decreasing). This $N = 2$ iso-graining is analog to the non condensed disordered phase of non interacting bodies, $\forall k > 1$, $\langle IP\rangle(k) \approx 0$. All the other mean $\langle IP\rangle(k)$ paths have non-trivial critical dimensions. The condition $N = 9$, $m = 111$ used for the analysis is surrounded by dotted red lines. It was chosen to be in the condensed phase above the critical graining, here $N_c = 3$, close to the criterion of maximal mutual information coefficient MI_2C proposed by Reshef and colleagues (bin surrounded by green dotted line) and with a not too low undersampling dimension. **b,** The mean $\langle IP\rangle(k)$ paths are presented for $m = 111, 100, ..., 12$ population A neurons and $n = 21$ genes with a number of bins $N = 9$.

Supplementary Materials: The following are available at www.mdpi.com/link: Table S1: Dataset.
The software Infotopo is available at https://github.com/pierrebaudot/INFOTOPO

Acknowledgments: This work was funded by the European Research Council (ERC consolidator grant 616827 *CanaloHmics* to J.M.G.; supporting M.T. and P.B.), developed at UNIS Inserm 1072 - Université Aix-Marseille , and thanks previously to supports and hostings since 2007 of Max Planck Institute for Mathematic in the Sciences (MPI-MIS) and Complex System Instititute Paris-Ile-de-France (ISC-PIF) and Institut de Mathématiques de Jussieu - Paris Rive Gauche (IMJ-PRG). This work is dedicated to D.Bennequin and addresses a deep and warm acknowledgement to the researchers who encouraged, supported and helped its realization: H.Atlan, F.Barbaresco, H.Bénali, P.Bourgine, F.Chavane, A.Chenciner, M.Gromov, J.Jost, G.Marrelec, A.Mohammad-Djafari, JP.Nadal, J.Petitot, A.Sarti, J.Touboul. A partial version of this work has been deposited in the method section of Bioarxiv 168740 in July 2017.

Author Contributions: P.B. wrote the paper and analysed the data; M.T. performed the experiments; M.T. and J.M.G. conceived and designed the experiments; P.B., M.T. and J.M.G participated in the conception of the analysis.

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

Abbreviations

The following abbreviations are used in this manuscript:

iid	independent identicaly distributed
DA	Dopaminergic neuron
nDA	non Dopaminergic neuron
H_k	Multivariate k-joint Entropy
I_k	Multivariate k-Mutual-Information
C_k	Multivariate k-total-correlation
MI_2C	Maximal 2-mutual-Information Coefficient

Appendix A Conditional mutual-information: bounds, Markov chains and HI landscapes

Appendix A.1 Bounds of conditional mutual-information

The upper bound of conditional mutual information is given by:

$$X_k.I(X_1;..;X_{k-1}) \leq \min_{i\in[k-1]} (X_k.H(X_i)) \tag{A1}$$

with equality given by the configurations for which the variables $X_1, .., X_{k-1}$ are equivalent to $X_1 \sim ... \sim X_{k-1}$ when X_k is given. The lower bound appears less obvious, and we conjecture that:

$$\min I(X_1;..;X_{k-1}) \leq X_k.I(X_1;..;X_{k-1}) \tag{A2}$$

We obtain the equality by considering that the k-th variable is the constant-deterministic variable $X_k = 0$ together with the minimal configuration of $I(X_1;..;X_{k-1})$. Figure A1 illustrates examples of these conjectured extreme cases and of conditional mutual information negativity for $k = 4, 5$.

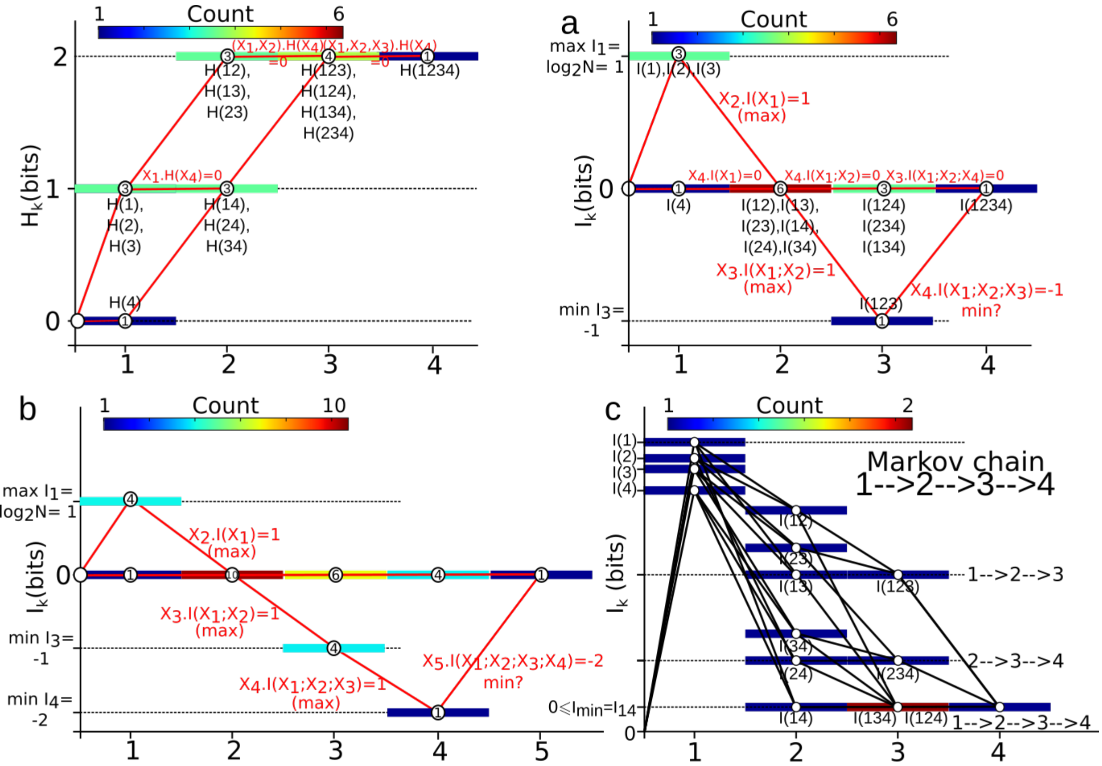

Figure A1. Theoretical examples of information landscapes with maximal and conjectured minimal conditional information, and of Markov chains. a, For $n = 4$, we consider 3 binary variables which are pairwise independent and shares a minimal I_3 (the configuration of minimal negative I_3 is studied more in depth in section 3.1 and Figure 9) together with a 4th variable chosen as the "0" deterministic variable. The left panel represents the H_k landscape and the right panel represents the I_k landscape. **b,** For $n = 5$, the same setting and configuration together with a 4th variable such that all triplets are the same as the example of minimum of I_3 and a 5th variable chosen as the trivial "0" deterministic variable. **c,** Example of a Markov chain $X_1 \to X_2 \to X_3 \to X_4$. In the example, the data processing inequality gives: $I(X_1;X_3) \leq I(X_1;X_2)$, $I(X_1;X_3) \leq I(X_2;X_3)$, $I(X_2;X_4) \leq I(X_3;X_4)$, $I(X_2;X_4) \leq I(X_2;X_3)$, $I(X_1;X_4) \leq I(X_2;X_4)$, $I(X_1;X_4) \leq I(X_1;X_3)$, $I(X_1;X_2;X_4) \leq I(X_2;X_3;X_4)$, $I(X_1;X_2;X_4) \leq I(X_1;X_2;X_3)$

doi:10.20944/preprints201804.0157.v1

Appendix A.2 Joint conditional entropy and meet conditional information

As illustrated in Figure A2, using the chain rule of entropy (equation 12) and of mutual information (equation 13), it is possible to generalize conditional entropy and conditional mutual information to joint conditional entropy and meet conditional information. They are given by the following equations, for any $j < k$:

$$H(X_1,...,X_k) - H(X_1,...,X_j) = (X_1,...,X_j).H(X_{j+1},...,X_k) \tag{A3}$$

and

$$I(X_1;...;X_k) - I(X_1;...;X_j) = -(X_{j+1};...;X_k).I(X_1;...;X_j) \tag{A4}$$

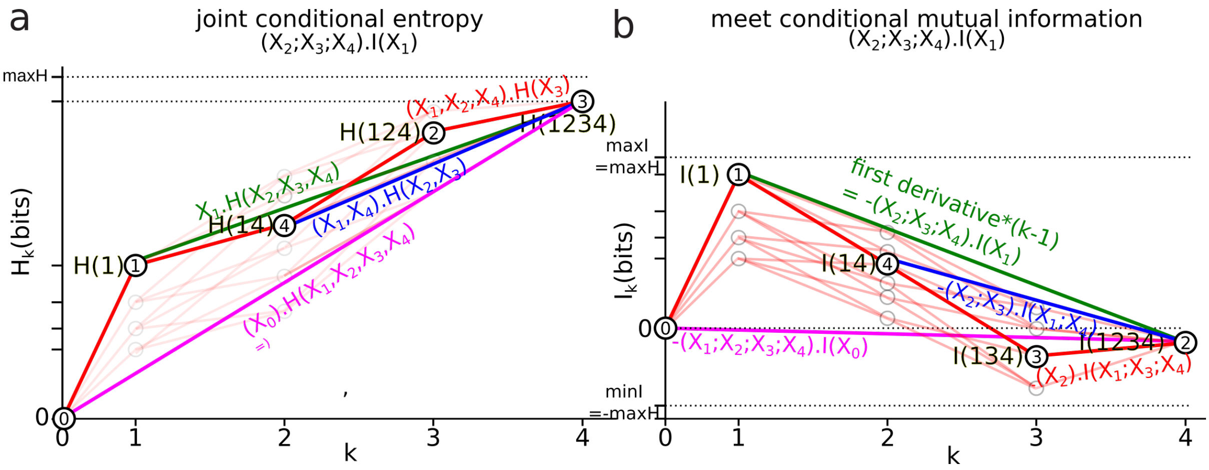

Figure A2. Joint conditional entropy and meet conditional information. a, illustration of the joint conditional entropy in H_k landscape for $n = 4$, as a vectorial addition (see text). **b,** illustration of the meet conditional information in I_k landscape for $n = 4$, as a vectorial addition (see text).

Appendix A.3 Markov chains information structures

Markov chains and conditional independence provide the foundation for important classes of models in statistical physic and machine learning, like Brownian motion, or ϵ-machine Markovian computational mechanics [199], or the field of algebraic statistics following the work of Drton, Sturmfels and Sullivant [200] based on conditional independence (see also the introduction to machine learning). It is hence important to restate the theorem of Hu Kuo Ting [14], which characterizes Markov Chains in terms of pairwise-mutual-information, further underlining that Markov chains are just a partial subset of all the possible information relations.

Theorem A.1. ***Markov chain and extremal Mutual information equivalence (Theorem 3, Hu kuo Ting [14]):*** *the sequence $X_0, ..., X_{n+1}$ forms a Markov chain if and only if $\forall I \subset [n]; card(I) = i, i \geq 1$ we have $I_{i+2}(X_0; X_I; X_{n+1}; \mathbb{P}) = I_2(X_0; X_{n+1}; \mathbb{P})$*

See Hu kuo Ting [14] for proof. In simple words, all higher mutual informations are equal to the mutual information between X_0 and X_{n+1}) and as a consequence, all Markov chains will always display non-negative mutual-informations. More precisely, we have the theorem: if $X_1 \rightarrow ... \rightarrow X_n$ forms a Markov Chain then $0 \leq I_2(X_1; X_n) \leq I_k$ for all I_k (proof follows from Hu's theorem 3 and data processing inequalities). Figure A1c shows how Markov chains are characterized by I_k landscape, giving an illustration of Hu's theorem 3 [14].

Appendix A.4 Conditional mutual information positivity-negativity: stability-instability

We now justify that I_k and H_k landscapes can be conceived as a single landscape, which allows an intuitive interpretation of the negativity and positivity of conditional mutual-information as information loss and gain, respectively. Conditional information negativity quantifies what would be the opposite of our usual concept of dependence, a dependence loss, a divergence of mutual-information quantities with respect to the dimension. Using the previous conventions of information, this "information loss" corresponds in equation to $I_{k+1} \geq I_k$, which direct interpretation is exactly the opposite. To circumvent this counterintuitive behavior, it is possible to glue both I_k and H_k landscapes by applying the theorem $I_1 = H_1$ such that the resulting HI'_k landscape still respects the partial order given by the inclusion, as depicted in Figure A3.

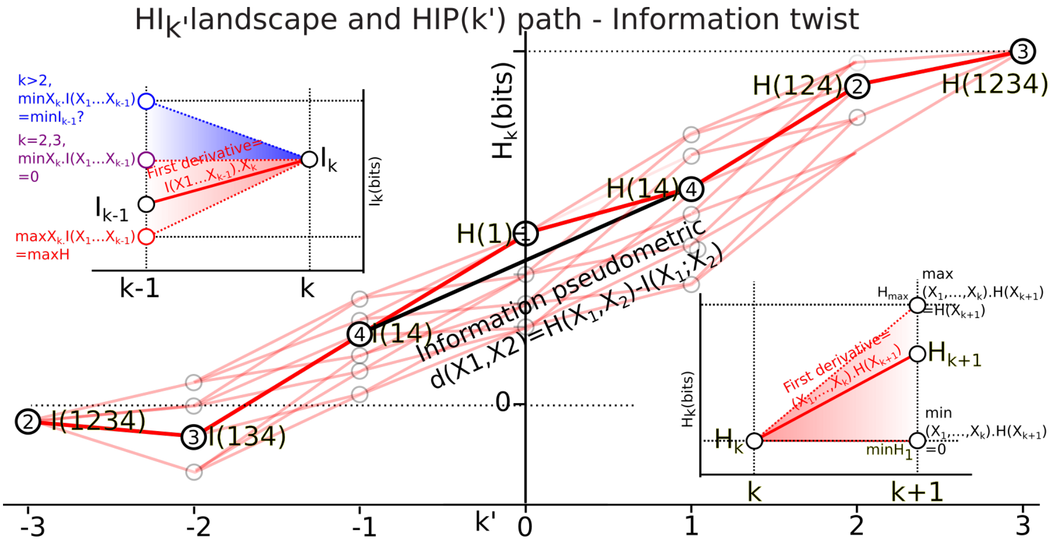

Figure A3. HI_k **landscape.** Illustration of the HI_k landscape, constructed using a twist information in I_k landscape for $n = 4$ which glues the H_k and I_k at the dimension 1, while reversing the I_k dimensions (see text). It moreover illustrates what is the symmetric left and right action of conditioning that we introduced in the first section. The information pseudometric is the slope from I_2 to H_2. It generalizes to the k-pseudovolumes defined by $V_k = H_k - I_k$

The dimensions are shifted by -1 and the degrees of the I_k landscape are multiplied by -1 (more precisely $k'_{HI} = (-1)^{2k_I+1}(k_I - 1)$ and $k'_{HI} = (-1)^{2k_H}(k_H - 1)$). This twist is a classical operation in homology (like the Tate twist) and allows to encode the homology in the negative dimensions and the cohomology in the positive dimensions, making a single theory out of two dual theories and can be related to the construction of the double complex exposed in the first section. The mathematical motivation is to recover the topological structure where entropy appeared in the early works of Catelineau [24], Gangl and Elbaz-Vincent [26], motiv theory. According to Beilinson and colleagues [37] such cohomology should exhibit a Hodge-Tate structure and hence a Tate twist. We underline that such a structure can have a concrete informational implementation and interpretation. Algebraically, the construction of a single landscape is achieved by the fact that the following square commutes, and its iteration generates long exact sequences of cohomology:

$$\begin{tikzcd}
 & H(X_1,X_2) \arrow[dl] \arrow[dr] & \\
H(X_1) \arrow[dr] & & H(X_2) \arrow[dl] \\
 & I(X_1;X_2) &
\end{tikzcd} \tag{A5}$$

Proof: this is an expression of the associativity of conditioning $X_1.(X_2.F(Y)) = X_2.(X_1.F(Y))$, cf. proof of lemma 1 in [1].□ With such a unified landscape (cf. Figure 6c) we have the reversed mutual information chain rule $I_k - I_{k-1} = X_k.I(X_1;...X_{k-1})$ and the negativity of conditional mutual information is now appearing as a decrease of mutual information $I_k < I_{k-1}$, as intuitively expected.

References

1. Baudot, P.; Bennequin, D. The Homological Nature of Entropy. *Entropy* **2015**, *17(5)*, 3253–3318.
2. Baudot, P.; Bennequin, D. Topological forms of information. *AIP Conf. Proc.* **2015**, *1641*, 213–221.
3. Vigneaux, J. The structure of information: from probability to homology. *arXiv:1709.07807* **2017**.
4. Yeung, R. *Information Theory and Network Coding.*; Springer, 2007.
5. Pethel, S.; Hahs, D. Exact Test of Independence Using Mutual Information. *Entropy* **2014**, *16*, 2839–2849.
6. Gerstenhaber, M.; Schack, S. A hodge-type decomposition for commutative algebra cohomology. *Journal of Pure and Applied Algebra* **1987**, *48*, 229–247.
7. Cover, T. Which Processes Satisfy the Second Law? *in: Physical Origins of Time Asymmetry, , eds. J. J. Halliwell, J. Perez-Mercader and W. H. Zurek* **1994**, pp. 98–107.
8. Mezard, M. Passing Messages Between Disciplines. *science* **2003**, *301*, 1686.
9. Mezard, M.; Montanari, A. *Information, Physics, and Computat*; Oxford University Press, 2009.
10. Margolin, A.; Wang, K.; Califano, A.; Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst Biol.* **2010**, *4*, 428–40.
11. Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitsrechnung.(English translation (1950): Foundations of the theory of probability.)*; Springer, Berlin (Chelsea, New York)., 1933.
12. Shannon, C.E. A Mathematical Theory of Communication. *The Bell System Technical Journal* **1948**, *27*, 379–423.
13. Kullback, S.; Leibler, R. On information and sufficiency. *Annals of Mathematical Statistics* **1951**, *22*, 79–86.
14. Hu, K.T. On the Amount of Information. *Theory Probab. Appl.* **1962**, *7(4)*, 439–447.
15. McGill, W. Multivariate information transmission. *Psychometrika* **1954**, *19*, p. 97–116.
16. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development* **1960**, *4*, 66–81.
17. Tononi, G.; Edelman, G. Consciousness and Complexity. *Science* **1998**, *282*, 1846–1851.
18. Studeny, M.; Vejnarova, J. The multiinformation function as a tool for measuring stochastic dependence. *in M I Jordan, ed., Learning in Graphical Models, MIT Press, Cambridge* **1999**, pp. 261–296.
19. Cover, T.; Thomas, J. *Elements of Information Theory.*; Wiley Series in Telecommunication, 1991.
20. Matsuda, H. Information theoretic characterization of frustrated systems. *Physica A: Statistical Mechanics and its Applications.* **2001**, *294 (1-2)*, 180–190.
21. Andre, Y. *Symétries I. Idées galoisiennes.*; Ircam online courses, 2007.
22. Andre, Y. Ambiguity theory, old and new. *arXiv:0805.2568* **2008**.
23. Yeung, R. On Entropy, Information Inequalities, and Groups. *Communications, Information and Network Security* **2003**, *Volume 712 of the series The Springer International Series in Engineering and Computer Science*, 333–359.
24. Cathelineau, J. Sur l'homologie de sl2 a coefficients dans l'action adjointe. *Math. Scand.* **1988**, *63*, 51–86.

25. Kontsevitch, M. The 11/2 logarithm. *Unpublished note. Reproduced in Elbaz-Vincent & Gangl, 2002 On poly(ana)logs I. Compositio Mathematica* **1995**.
26. Elbaz-Vincent, P.; Gangl, H. On poly(ana)logs I. *Compositio Mathematica* **2002**, *130(2)*, 161–214.
27. Elbaz-Vincent, P.; Gangl, H. Finite polylogarithms, their multiple analogues and the Shannon entropy. *To be published in* **2015**.
28. Connes, A.; Consani, C. Characteristic 1, entropy and the absolute point. *preprint arXiv:0911.3537v1.* **2009**.
29. Marcolli, M.; Thorngren, R. Thermodynamic Semirings. *arXiv 10.4171/JNCG/159* **2011**, *Vol. abs/1108.2874*.
30. Marcolli, M.; Tedeschi, R. Entropy algebras and Birkhoff factorization. *arXiv, Vol. abs/1108.2874* **2014**.
31. Baez, J.; Fritz, T.; Leinster, T. A Characterization of Entropy in Terms of Information Loss. *Entropy* **2011**, *13*, 1945–1957.
32. Baez, J.C.; Fritz, T. A Bayesian characterization of relative entropy. *Theory and Applications of Categories,* **2014**, *Vol. 29, No. 16*, p. 422–456.
33. Boyom, M. Foliations-Webs-Hessian Geometry-Information Geometry-Entropy and Cohomology. *Entropy* **2016**, *18*, 433.
34. Drummond-Cole, G.; Park, J.S.; Terilla, J. Homotopy probability theory I. *J. Homotopy Relat. Struct.* **2015**, *10*, 425–435.
35. Drummond-Cole, G.; Park, J.S.; Terilla, J. Homotopy probabilty Theory II. *J. Homotopy Relat. Struct.* **2015**, *10*, 623–635.
36. Park, J.S. Homotopy Theory of Probability Spaces I: Classical independence and homotopy Lie algebras. *arXiv:1510.08289* **2015**.
37. Beilinson, A.; Goncharov, A.; Schechtman, V.; Varchenko, A. Aomoto dilogarithms, mixed Hodge structures and motivic cohomology of pairs of triangles on the plane. *The Grothendieck Festschrift, vol. 1, in: Progr. Math., Birkhauser,* **1990**, *vol. 86*, 135–172.
38. Aomoto, K. Addition theorem of Abel type for Hyper-logarithms. *Nagoya Math. J.* **1982**, *Vol. 88*, 55–71.
39. Goncharov, A. Regulators. *Handbook of K-theory. pringer-Verlag Berlin Heidelberg. http://k-theory.org/handbook/* **2005**, pp. 297–324.
40. Andrews, G. *The Theory of Partitions*; Cambridge University Press, Cambridge, 1998.
41. Fresse, B. Koszul duality of operads and homology of partitionn posets. *Contemp. Math. Amer. Math. Soc.* **2004**, *346*, pp. 115–215.
42. Hochschild, G. On the cohomology groups of an associative algebra. *Annals of Mathematics. Second Series,* **1945**, *46*, 58–67.
43. Weibel, C. *An introduction to homological algebra*; Cambridge University Press, 1995.
44. Kassel, C. Homology and cohomology of associative algebras- A concise introduction to cyclic homology. *Advanced Course on non-commutative geometry* **2004**.
45. Tate, J. Galois Cohomology. *online course* **1991**.
46. Cartan, H.; Eilenberg, S. *Homological Algebra*; The Princeton University Press, Princeton, 1956.
47. Mac Lane, S. *Homology.*; Classic in Mathematics, Springer, Reprint of the 1975 edition, 1975.
48. Kendall, D. Functional Equations in Information Theory. *Z. Wahrscheinlichkeitstheorie* **1964**, *2*, p. 225–229.
49. Lee, P. On the Axioms of Information Theory. *The Annals of Mathematical Statistics* **1964**, *Vol. 35, No. 1*, pp. 415–418.
50. Shannon, C. A lattice theory of information. *Trans. IRE Prof. Group Inform. Theory* **1953**, *1*, 105–107.
51. Rajski, C. A metric space of discrete probability distributions. *Information and Control* **1961**, *4*, 371–377.
52. Zurek, W. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* **1989**, *341*, 119–125.
53. Bennett, C.; Gacs, P.; Ming Li, P.; Vitanyi, M.; Zurek, W. Information distance. *IEEE Transactions on Information Theory* **1998**, *44*, 1407–1423.
54. Kraskov, A. ; Grassberger, P. MIC: Mutual Information Based Hierarchical Clustering. *Information Theory and Statistical Learning. Springer ed. http://arxiv.org/abs/q-bio/0311039* **2009**, pp. 101–123.
55. Atiyah, M. Topological quantum field theory. *Publications mathématiques de l'I.H.E.S* **1988**, *68*, 175–186.
56. Bourbaki, N. *Theory of Sets - Elements of Mathematic.*; Addison Wesley publishing company. Hermann, 1968.
57. Boole, G. *An Investigation Of The Laws Of Thought On Which Are Founded The Mathematical Theories Of Logic And Probabilities.*; McMillan and Co., 1854.
58. Loday. Operations sur l'homologie cyclique des algebres commutatives. *Invent. math.* **1989**, *96*, 205–230.

59. Lamarche-Perrin, R.; Demazeau, Y.; Vincent, J. The Best-partitions Problem: How to Build Meaningful Aggregations? *Research Report RR-LIG-044 <hal-00947934>* **2013**, p. 18.
60. Pudlák, P. & Tůma, J. Every finite lattice can be embedded in a finite partition lattice. *Algebra Univ.* **1980**, *10*, 74–95.
61. Gerstenhaber, M.; Schack, S. Simplicial cohomology is Hochschild Cohomology. *Journal of Pure and Applied Algebra* **1983**, *30*, 143–156.
62. Steenrod, N. Products of Cocycles and Extensions of Mapping. *Annals of mathematics 2nd ser.* **1947**, *48*, 290–320.
63. James, R.; Crutchfield, J. Multivariate Dependence beyond Shannon Information. *Entropy* **2017**, *19*, 531.
64. Witten, E. Topological Quantum Field Theory. *Commun. Math. Phys.* **1988**, *117*, 353–386.
65. Schwarz, A. Topological Quantum Field Theory. *arXiv:hep-th/0011260v1* **2000**.
66. Noether, E. Invariant Variation Problems. M. A. Taveles English translation of "Invariante Variationsprobleme". *Nachr. d. Kenig. Gesellsch. d. Wiss. zu Gettingen, Math-phys. Klasse. Traduction in Transport Theory and Statistical Physics, 1 (3), 183-207 (1971).* **1918**, pp. 235–257.
67. Reshef, D.; Reshef, Y.; Finucane, H.; Grossman, S.; McVean, G.; Turnbaugh, P.; Lander, E.; Mitzenmacher, M.; Sabeti, P. Detecting Novel Associations in Large Data Sets. *Science* **2011**, *334*, 1518.
68. Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, 864–871.
69. Kohn, W.; Sham, L.J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, 1133–1138.
70. Feynman, R. *QED: The Strange Theory of Light and Matter*; Princeton University Press, 1985.
71. Dirac, P. *Directions in Physics*; John Wiley & Sons Inc, 1978.
72. Baez, J.; Pollard, S. Relative Entropy in Biological Systems. *Entropy* **2016**, *18*, 46.
73. Jaynes, E.T. Information Theory and Statistical Mechanics. *Physical Review. Series II* **1957**, *106 (4)*, pp. 620–630.
74. Jaynes, E. Information Theory and Statistical Mechanics II. *Physical Review. Series II* **1957**, *108 (2)*, pp. 171–190.
75. Landauer, R. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* **1961**, *5 (3)*, 183–191.
76. Wheeler, J. Information, Physics, quantum: the search for the links. *Proc. 3rd Int. Symp. Foundations of Quantum Mechanics, Tokyo* **1983**, pp. 354–368.
77. Bennett, C. Notes on Landauer's principle, Reversible Computation and Maxwell's Demon. *Studies in History and Philosophy of Modern Physics* **2003**, *34(3)*, 501–510.
78. Von Bahr, B. On the central limit theorem in Rk. *Ark. Mat.* **1967**, *7*, 61–69.
79. Ebrahimi, N.; Soofi, E.; Soyer, R. Multivariate maximum entropy identification, transformation, and dependence. *Journal of Multivariate Analysis* **2008**, *99*, 1217–1231.
80. Conrad, K. Probability distributions and maximum entropy. *Unpublished note. http://www.math.uconn.edu/ kconrad/blurbs/analysis/entropypost.pdf* **2005**.
81. Adami, C.; Cerf, N. Prolegomena to a non-equilibrium quantum statistical mechanics. *Chaos, Solitons & Fractals* **1999**, *10*, 1637–1650.
82. Kapranov, M. Thermodynamics and the moment map. *arXiv:1108.3472* **2011**.
83. Erdos, P. On the distribution function of additive functions. *Ann. of Math.* **1946**, *Vol. 47*, pp. 1–20.
84. Aczel, J.; Daroczy, Z. *On measures of information and their characterizations.*; Academic Press. Mathematics in science and engineering, 1975.
85. Landau, L.; Lifshitz, E. *Statistical Physics (Course of Theoretical Physics, Volume 5)*; Butterworth-Heinemann; 3 edition (1980), 1969.
86. Jaynes, E.T. *Information Theory and Statistical Mechanics*; Statistical Physics. New York: Benjamin. Ford, K., 1963.
87. Hopfield, J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *PNAS* **1982**, *79*, 2554–2558.
88. Ackley, D.; Hinton, G.; Sejnowski, T.J. A Learning Algorithm for Boltzmann Machines. *Cognitive Science* **1985**, *9*, 147–169.
89. Dayan, P.; Hinton, G.; Neal, R.; Zemel, R. The Helmholtz Machine. *Neural Computation* **1995**, *7*, 889–904.

90. Efron, B. Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency). *Annals of Statistics* **1975**, *3*, 1189–1242.
91. Cencov, N. *Statistical Decision Rules and Optimal Inference.*; Translations of Mathematical Monographs. Amer Mathematical Society, 1982.
92. Ay, N.; Jost, J.; Le, H.; Schwachhofer, L. Information geometry and sufficient statistics. *Probability Theory and Related Fields* **2015**, *162*, 327–364.
93. Amari, S. *Differential-geometrical methods in statistics.*; Lecture Notes in Statistics New York: Springer-Verlag., 1985.
94. Amari, S. Natural gradient works efficiently in learning. *Neural computation* **1998**, *10(2)*, 251–276.
95. Deza, M.M.; Deza, E. *Encyclopedia of Distances.*; Springer Science & Business Media, 2009.
96. Cartan, E. *Lecons sur la geometrie des espaces de Riemann, 2nd ed.*; Editions Jacques Gabay, 1946.
97. Wu, F. The Potts model. *Reviews of Modern Physics.* **1982**, *54*, 235–268.
98. Benzecri, J. *L'analyse des donnees 2. L'analyse de correspondence*; Dunod, 1973.
99. Poincare, H. Les geometries non-euclidiennes. *Revue generale* **1891**, *23*.
100. Lum, P.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **2013**, *3*.
101. Epstein, C.; Carlsson, G.; Edelsbrunner, H. Topological data analysis. *Inverse Probl.* **2011**, *27*, 120201.
102. Carlsson, G. Topology and data. *Bull. Amer. Math. Soc.* **2009**, *46*, p.255–308.
103. Niyogi, P.; Smale, S.; Weinberger, S. A Topological View of Unsupervised Learning from Noisy Data. *SIAM J. of Computing* **2011**, *20*, 646–663.
104. Buchet, M.; Chazal, F.; Oudot, S.; Sheehy, D. Efficient and Robust Persistent Homology for Measures. *arXiv:1306.0039* **2014**.
105. Chintakunta, H.; Gentimis, T.; Gonzalez-Diaz, R.; Jimenez, M.J.; Krim, H. An entropy-based persistence barcode. *Pattern Recognit.* **2015**, *48*, 391–401.
106. Merelli, E.; Rucco, M.; Sloot, P.; Tesei, L. Topological Characterization of Complex Systems: Using Persistent Entropy. *Entropy* **2015**, *17*, 6872–6892.
107. Tadic, B.; Andjelkovic, M.; Suvakov, M. The influence of architecture of nanoparticle networks on collective charge transport revealed by the fractal time series and topology of phase space manifolds. *J. Coupled Syst. Multiscale Dyn.* **2016**, *4*, 30–42.
108. Maletic, S.; Rajkovic, M. Combinatorial Laplacian and entropy of simplicial complexes associated with complex networks. *Eur. Phys. J.* **2012**, *212*, 77–97.
109. Maletic, S.; Zhao, Y. Multilevel Integration Entropies: The Case of Reconstruction of Structural Quasi-Stability in Building Complex Datasets. *Entropy* **2017**, *19*, 172.
110. Han, T.S. Linear dependence structure of the entropy space. *Information and Control.* **1975**, *vol. 29*, p. 337–368.
111. Bjorner, A. Continuous partition lattice. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6327–6329.
112. Postnikov, A. Permutohedra, Associahedra, and Beyond. *Int Math Res Notices. arXiv: math.CO/0507163.* **2009**, *2009(6)*, 1026–1106.
113. Matus. Conditional probabilities and permutahedron. *Annales de l'I.H.P. Probabilites et statistiques* **2003**, *39*, 687–701.
114. Yeung, R. Facets of entropy. *Communications in Information and Systems* **2015**, *15*, 87–117.
115. Yeung, R. A framework for linear information inequalities. *IEEE Transactions on Information Theory (New York)* **1997**, *43*, 1924–1934.
116. Zang, Z.; Yeung. On Characterization of Entropy Function via Information Inequalities. *IEEE transactions on information theory* **1997**, *44*, 1440–1452.
117. Matus, F. Infinitely Many Information Inequalities. *IEEE International Symposium on Information Theory.* **2007**.
118. Takacs, D. *Stochastic Processes problems and solutions*; John Wiley and Sons Inc, 1960.
119. Brillouin, L. *Scientific Uncertainty, and Information*; Academic Press, 2014.
120. Griffiths, R. Consistent Histories and the Interpretation of Quantum Mechanics. *J. Stat. Phys.* **1984**, *35*, 219.
121. Omnes, R. Logical reformulation of quantum mechanics I. Foundations. *Journal of Statistical Physics* **1988**, *53*, 893–932.

122. Gell-Mann, M.; Hartle, J. Quantum mechanics in the light of quantum cosmology. *W. H. Zurek (ed.), Complexity, entropy and the physics of information. Redwood City, Calif.: Addison-Wesley* **1990**, pp. 425–458.

123. Lieb, E. H.; Yngvason, J. A Guide to Entropy and the Second Law of Thermodynamics. *Notices of the American Mathematical Society. http://www.ams.org/notices/199805/lieb.pdf* **1998**, *Vol 45, N 5*, 571–581.

124. Gromov, M. Symmetry, Probabiliy, Entropy: Synopsis of the Lecture at MAXENT 2014. *Entropy* **2015**, *17*, 1273–1277.

125. Baez, J.; Fong, B. A Noether theorem for Markov processes. *Journal of Mathematical Physics* **2013**, *54*, 013301.

126. Neuenschwander, D. Noether's theorem and discrete symmetries. *Am. J. Phys.* **1995**, *63*, 489.

127. Mansfield, E. Noether's Theorem for Smooth, Difference and Finite Element Systems. *Foundations of Computational Mathematics. (Santander, 2005), L.M. Pardo, A. Pinkus, E. Suli, and M.J. Todd, eds., London Mathematical Society Lecture Note Series.* **2006**, *331*, 230–254.

128. Feynman, R. Space-Time Approach to Non-Relativistic Quantum Mechanics. *Reviews of Modern Physics* **1948**, *20*, 367–387.

129. Weiss, P. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.* **1907**, *6*, 661–690.

130. Parsegian, V. *Van der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists.*; Cambridge University Press., 2006.

131. Xie, Z.; Chen, J.; Yu, J.; Kong, X.; Normand, B.; Xiang, T. Tensor Renormalization of Quantum Many-Body Systems Using Projected Entangled Simplex States. *Phys. Rev. X* **2014**, *4*, 011025–1.

132. Tai Ha, H.; Van Tuyl, A. Resolutions of square-free monomial ideals via facet ideals: a survey. *Contemporary mathematics - American Mathematical Society - Algebra, Geometry and Their Interactions: International Conference Midwest* **2007**, *448*, 91–105.

133. Newman, M.E.J. Complex Systems: A Survey. *http://arxiv.org/abs/1112.1440v1* **2011**.

134. Mezard, M.; Montanari, A. *Information, Physics, and Computation*; Oxford University Press, 2009.

135. Vannimenus, J.; Toulouse, G. Theory of the frustration effect. II. Ising spins on a square lattice. *Journal of Physics C: Solid State Physics* **1977**, *10*.

136. Penrose, R. *Angular momentum : an approach to combinatorial space-time. In Quantum Theory and Beyondum*; Cambridge University Press. Ted Bastin p.151-180, 1971.

137. Rovelli, C. Notes for a brief history of quantum gravity. *arXiv:gr-qc/0006061v3* **2008**.

138. Sorkin, R. Finitary Substitute for Continuous Topology. *International Journal of Theoretical Physics* **1991**, *30*, 923–947.

139. Baudot, P. Natural computation: much ado about nothing? An intracellular study of visual coding in natural condition. Master's thesis, Paris 6 university, 2006.

140. Williams, P.; Beer, R. Nonnegative Decomposition of Multivariate Information. *arXiv:1004.2515v1* **2010**.

141. Olbrich, E.; Bertschinger, N.; Rauh, J. Information Decomposition and Synergy. *entropy* **2015**, *17*, 3501–3517.

142. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183.

143. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. *In Guided Self-Organization: Inception; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany* **2014**, pp. 159–190.

144. Wibral, M.; Finn, C.; Wollstadt, P.; Lizier, J.; Priesemann, V. Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition. *Entropy* **2017**, *19*, 494.

145. Kay, J.; Ince, R.; Dering, B.; Phillips, W. Partial and Entropic Information Decompositions of a Neuronal Modulatory Interaction. *Entropy* **2017**, *19*, 560.

146. Brillouin, L. *Science and Information theory*; Academic Press Inc New york, 1956.

147. Wiener, N. *Cybernetics: the Control and Communication in the Animal and the Machine [second edition]*; MIT, 1965.

148. Schrodinger, E. *What is Life?*; Based on lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943, 1944.

149. Dirac, P. Discussion of the infinite distribution of electrons in the theory of the positron. *Proc. Camb. Phil. Soc.* **1929**, *25*, 62.

150. Casimir, H. The Influence of Retardation on the London-van der Waals Forces. *Phys. Rev.* **1948**, *73*, 360.

151. Feynman, R. Negative probabilities. *In Quantum Implications: Essays in Honor of David Bohm , ed. F. D. Peat and B. Hiley, Routledge & Kegan Paul* **1987**, pp. 235–248.

152. Cerf, N.; Adami, C. Negative entropy and information in quantum mechanic. *Phys. Rev. Lett.* **1997**, *79*, 5194.
153. Cerf, N.; Adami, C. Entropic Bell Inequalities. *Phys. Rev. A* **1997**, *55-5*.
154. Matsuda, H.; Kudo, K.; Nakamura, R.; Yamakawa, O.; Murata, T. Mutual Information of Ising Systems. *International Journal of Theoretical Physics*, **1996**, *Vol. 35, No. 4*, 839.
155. Sootla, S.; Theis, D.; Vicente, R. Analyzing Information Distribution in Complex Systems. *Entropy* **2017**, *19*, 636.
156. Brenner, N.; Strong, S.; Koberle, R.; Bialek, W. Synergy in a Neural Code. *Neural Computation.* **2000**, *12*, 1531–1552.
157. Watkinson, J.; Liang, K.; Wang, X.; Zheng, T.; Anastassiou, D. Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information. *The Challenges of Systems Biology: Ann. N.Y. Acad. Sci.* **2009**, *1158*, 302–313.
158. Kim, H.; Watkinson, J.; Varadan, V.; Anastassiou, D. Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Medical Genomics* **2010**, *3:51*.
159. Schneidman, E.; Still, S.; Berry 2nd, M.; Bialek, W. Network information and connected correlations. *Phys Rev Lett.* **2003**, *1*, 238701.
160. James, R.; Crutchfield, J. Multivariate Dependence Beyond Shannon Information. *Santa Fe Institute Working Paper 16-09-XXX arXiv:1609.01233* **2016**.
161. Tapia, M.; Baudot, P.; Dufour, M.; Formizano-Treziny, C.; Temporal, S.; Lasserre, M.; Kobayashi, K.; J.M., G. Information topology of gene expression profile in dopaminergic neurons. *BioArXiv168740* **2017**.
162. Kauffman, L. Simplicial Homotopy Theory, Link Homology and Khovanov Homology. *arXiv:1701.04886v3* **2017**.
163. Jones, V. On Knot invariant related to some statistical mechanical models. *Pacific Journal of mathematics* **1989**, *137*, 311–334.
164. Dawkins, R. *The Selfish Gene.*; Oxford University Press. 1st ed., 1976.
165. Atlan, H. *Entre le cristal et la fumee. Essai sur l'organisation du vivant.*; Seuil, 1979.
166. Strong, S.; de Ruyter van Steveninck, R.; Bialek, W.; Koberle, R. On the application of information theory to neural spike trains. *Pac Symp Biocomput* **1998**, pp. 621–32.
167. Nemenman, I.; Bialek, W.; de Ruyter van Steveninck, R. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical review E* **2004**, *69*, 056111–.
168. Merchan, L.; Nemenman, I. On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks. *J Stat Phys* **2016**, *162*, 1294–1308.
169. Grassberger, P. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics* **1986**, *25*, 907–938.
170. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. *Physica A* **2001**, *302*, 89–99.
171. Tsallis, C. Entropic Nonextensivity : a possible measure of Complexity. *Chaos solitons and fractals* **2002**, *13*, 371–391.
172. Ritort, F. Nonequilibrium fluctuations in small systems: from physics to biology,. *Advances in Chemical Physics Ed. Stuart. A. Rice, Wiley publications* **2008**, *vol. 137*, 31–123.
173. Born, M. The statistical interpretation of quantum mechanics. *Nobel Lecture.* **1954**.
174. Hilbert, D. Sur l'infini. Hilbert's Lectures on the Infinite. *Traduit par Andre Weil Paris (1926). edited in David Hilbert's Lectures on the Foundations of Arithmetic and Logic 1917-1933. Springer* **1924**.
175. Waddington, C.H. *The Strategy of the Genes*; Routledge Library editions, 1957.
176. Teschendorff, A.; Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature communication* **2017**.
177. Jin, S.; MacLean, A.; Peng, T.; Nie, Q. scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics* **2018**, pp. 1–10.
178. Thom, R. *Stabilite struturelle et morphogenese*; deuxieme edition, InterEdition, Paris, 1977.
179. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105–117.
180. Nadal, J.-P. ; Parga, N. Sensory coding: information maximization and redundancy reduction. *Neural information processing, G. Burdet, P. Combe and O. Parodi Eds. World Scientific Series in Mathematical Biology and Medecine* **1999**, *Vol. 7*, p. 164–171.

181. Bell, A.; Sejnowski, T. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation* **1995**, *7*, *6*, 1129–1159.
182. Chen, N.; Glazier, J.; Izaguirre, J.; Alber, M. A parallel implementation of the Cellular Potts Model for simulation of cell-based morphogenesis. *Computer Physics Communications* **2007**, *176*, 670–681.
183. Galvan, A. Neural plasticity of development and learning. *Hum Brain Mapp* **2010**, *31*, 879–890.
184. Schneidman, E.; Berry 2nd, M.; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012.
185. Tkacik, G.; Marre, O.; Amodei, D.; Schneidman, E.; Bialek, W.; Berry, M.n. Searching for collective behavior in a large network of sensory neurons. *PLoS Comput Biol. 20* **2014**, *10*.
186. Mora, T.; Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **2011**, *144*, 268–302.
187. Bialek, W.; Ranganathan, R. Rediscovering the power of pairwise interactions. *Arxiv 0712.4397.* **2007**.
188. Gibbs, J. A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces. *Trans. of the connecticut Acad.* **1873**, *2*, 382–404.
189. Shipman, J. Tkinter Reference: a GUI for Python. . *New Mexico Tech Computer Center, Socorro, New Mexico* **2010**.
190. Hunter, J. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 22–30.
191. Van Der Walt, S.; Colbert, C.; Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.
192. Hagberg, A.; Schult, D.; Swart, P. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008). Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA)* **2008**, pp. 11–15.
193. Hughes, G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* **1968**, *14*, 55–63.
194. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation* **2003**, *15*, 1191–1253.
195. Zvontin, A.; Levin, L. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ. Math. Surv.* **1970**, *256*, 83–124.
196. Borel, E. La mechanique statistique et l'irreversibilite. *J. Phys. Theor. Appl.* **1913**, *3*, 189–196.
197. Kurchan, J. In and out of equilibrium. *Nature* **2005**, *433*, 222–225.
198. Scott, D. *Multivariate Density Estimation. Theory, Practice and Visualization.*; New York: Wiley, 1992.
199. Shalizi, C.; Crutchfield, J. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics* **2001**, *Vol. 104, Nos. 3/4*, 817.
200. Drton, M.; Sturmfels, B.; Sullivant, S. *Lectures on Algebraic Statistic*; Birkhauser Applied Probability and Statistics. https://math.berkeley.edu/ bernd/owl.pdf, 2009.