entropy

MDPI

Article

The Poincaré-Shannon Machine: Statistical Physics and Machine Learning Aspects of Information Cohomology

Pierre Baudot [1,2,†]

[1] Inserm UNIS UMR1072—Université Aix-Marseille, 13015 Marseille, France; pierre.baudot@mediantechnologies.com; Tel.: +33-6-19-39-10-39
[2] Median Technologies, 06560 Valbonne, France
† Current address: Median Technologies, Les Deux Arcs, 1800 Route des Crêtes, 06560 Valbonne, France.

Received: 5 July 2019; Accepted: 3 September 2019; Published: 10 September 2019



Abstract: Previous works established that entropy is characterized uniquely as the first cohomology class in a topos and described some of its applications to the unsupervised classification of gene expression modules or cell types. These studies raised important questions regarding the statistical meaning of the resulting cohomology of information and its interpretation or consequences with respect to usual data analysis and statistical physics. This paper aims to present the computational methods of information cohomology and to propose its interpretations in terms of statistical physics and machine learning. In order to further underline the cohomological nature of information functions and chain rules, the computation of the cohomology in low degrees is detailed to show more directly that the k multivariate mutual information (I_k) are $(k-1)$-coboundaries. The $(k-1)$-cocycles condition corresponds to $I_k = 0$, which generalizes statistical independence to arbitrary degree k. Hence, the cohomology can be interpreted as quantifying the statistical dependences and the obstruction to factorization. I develop the computationally tractable subcase of simplicial information cohomology represented by entropy H_k and information I_k landscapes and their respective paths, allowing investigation of Shannon's information in the multivariate case without the assumptions of independence or of identically distributed variables. I give an interpretation of this cohomology in terms of phase transitions in a model of k-body interactions, holding both for statistical physics without mean field approximations and for data points. The I_1 components define 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 (the total correlation) of the k-body interactions. A basic mean field model is developed and computed on genetic data reproducing usual free energy landscapes with phase transition, sustaining the analogy of clustering with condensation. The set of information paths in simplicial structures is in bijection with the symmetric group and random processes, providing a trivial topological expression of the second law of thermodynamics. The local minima of free energy, related to conditional information negativity and conditional independence, 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. I give an interpretation of this complex in terms of unsupervised deep learning where the neural network architecture is given by the chain complex and conclude by discussing future supervised applications.

Keywords: algebraic topology; machine learning; information theory; statistical physics; deep neural networks; unsupervised learning; multivariate mutual information; statistical dependences; k-body interactions; synergy; clustering

Contents

1 Introduction **3**
1.1 Observable Physics of Information . . . 3
1.2 Statistical Interpretation: Hierarchical Independences and Dependences Structures . . . 3
1.3 Statistical Physics Interpretation: K-Body Interacting Systems . . . 4
1.4 Machine Learning Interpretation: Topological Deep Learning . . . 5

2 Information Cohomology **6**
2.1 A Long March through Information Topology . . . 6
2.2 Information Functions (Definitions) . . . 7
2.3 Information Structures and Coboundaries . . . 10
2.3.1 First Degree ($k = 1$) . . . 12
2.3.2 Second Degree ($k = 2$) . . . 12
2.3.3 Third Degree ($k = 3$) . . . 12
2.3.4 Higher Degrees . . . 13

3 Simplicial Information Cohomology **13**
3.1 Simplicial Substructures of Information . . . 13
3.2 Topological Self and Free Energy of K-Body Interacting System-Poincaré-Shannon Machine 14
3.3 k-Entropy and k-Information Landscapes . . . 18
3.4 Information Paths and Minimum Free Energy Complex . . . 19
3.4.1 Information Paths (Definition) . . . 19
3.4.2 Derivatives, Inequalities and Conditional Mutual-Information Negativity . . . 20
3.4.3 Information Paths Are Random Processes: Topological Second Law of Thermodynamics and Entropy Rate . . . 21
3.4.4 Local Minima and Critical Dimension . . . 23
3.4.5 Sum over Paths and Mean Information Path . . . 24
3.4.6 Minimum Free Energy Complex . . . 25

4 Discussion **27**
4.1 Statistical Physics . . . 27
4.1.1 Statistical Physics without Statistical Limit? Complexity through Finite Dimensional Non-Extensivity . . . 27
4.1.2 Naive Estimations Let the Data Speak . . . 28
4.1.3 Discrete Informational Analog of Renormalization Methods: No Mean-Field Assumptions Let the Objects Differentiate . . . 29
4.1.4 Combinatorial, Infinite, Continuous and Quantum Generalizations . . . 29
4.2 Data Science . . . 29
4.2.1 Topological Data Analysis . . . 29
4.2.2 Unsupervised and Supervised Deep Homological Learning . . . 30
4.2.3 Epigenetic Topological Learning—Biological Diversity . . . 31

5 Conclusions **32**

References **33**

> *"Now what is science? ...it is before all a classification, a manner of bringing together facts which appearances separate, though they are bound together by some natural and hidden kinship. Science, in other words, is a system of relations. ...it is in relations alone that objectivity must be sought. ...it is relations alone which can be regarded as objective. External objects... are really objects and not fleeting and fugitive appearances, because they are not only groups of sensations, but groups cemented by a constant bond. It is this bond, and this bond alone, which is the object in itself, and this bond is a relation."*
>
> H. Poincaré

1. Introduction

The present paper aims to provide a comprehensive introduction and interpretation in terms of statistics, statistical physics and machine learning of the information cohomology theory developed in References [1,2]. It presents the computational aspects of the application of information cohomology to data presented in Reference [3] and in the associated paper [2], which consists of an unsupervised classification of cell types or gene modules and provides a generic model for epigenetic co-regulation and differentiation.

1.1. Observable Physics of Information

In its application to empirically measured data, information cohomology is at the cross-roads of both data analysis and statistical physics and this article aims to give some keys to its interpretation within those two fields, which could be quoted as "passing the information between disciplines" in reference to Mezard's review [4]. Just as topos have been used as a communication bridge allowing the translation of theorems between different domains and therefore to unify mathematical theories [5], information cohomology can help in further unraveling some equivalences between different disciplines (e.g., statistical physics and machine learning) and shall play a foundational role in both as already proposed by Doering and Isham considering only probability structures [6,7]. In doing so, information theory goes one step toward a general mathematical and physical theory of communication (or of measure and observation), where information conservation is an isomorphism. In other terms, following the paths of topology, this paper pursues the work that started with the studies of Brillouin [8], Jaynes [9,10], Landauer [11], Penrose [12], Wheeler [13], and Bennett [14], of an informational theory of physics that would be furthermore restricted to a theory of data: an austere empirical theory [15] with a minimum set of priors or axioms, a physics that "let the data speak". It is the axiom of observability ("*concepts which correspond to no conceivable observation should be eliminated from physics*" [16], p. 264) which imposes that statistical physics and data sciences shall not be dissociated but unified in their foundations. Thermodynamics established that free energy is the energy that can be effectively used, whereas entropy was also called "lost heat". The same holds generically on whatever data with mutual information and total correlations, some kind of relative or shared information: mutual information is the information that can be effectively used for pattern detection and classification. It hence appears that knowledge is a form of energy [17] and the results suggest that there are important resources of such information-energy in the k-body dependences even beyond pairwise interactions.

1.2. Statistical Interpretation: Hierarchical Independences and Dependences Structures

First, in order to provide a mathematical interpretation of information cohomology, I give a brief bibliographical overview of its multiple and diverse origins coming notably from the theory of motive and recall the open conjecture on the higher classes. In References [1,2], entropy was characterized uniquely as the first cohomology class in a topos theory and this result could be extended to quantum information, Kullback-Leibler divergence and cross-entropy (Proposition 4 [1]) and have been further developed and extended by Vigneaux [18,19] to Tsallis entropies and differential entropy following some preliminary results of Marcolli and Thorngren [20]. Here, we show directly by computing

the cohomology in the low degrees that multivariate k-mutual information, denoted I_k, are $(k-1)$-coboundaries. Recalling a result of the associated paper (Theorem 2 [2]), establishing that n random variables are independent if and only if all the subsets of k-mutual information vanish ($k \geq 2$), allows one to conclude that the $(k-1)$-cocycles condition corresponds to $I_k = 0$, which generalizes statistical independence to an arbitrary degree k. The interpretation and meaning of information cohomology is hence the quantification of the statistical dependences and of the obstruction to factorization—a result confirmed by the work of Mainiero, with a different approach [21]. Notably, the introduction of a symmetric action of conditioning, following Gerstenhaber and Schack [22], gives a 1-cocycle condition that characterizes the famous information pseudo-metric discovered by Shannon, Rajski, Zurek, Crutchfield and Bennett [23–28] and that can be generalized to the k-multivariate case by new symmetric non-negative information functions, the pseudo k-volumes $V_k = H_k - I_k$.

1.3. Statistical Physics Interpretation: K-Body Interacting Systems

Second, to provide an interpretation of information cohomology in terms of statistical physics, this paper settles information structures in the context of a generic k-body interacting system. One can remark that the setting of information cohomology is equivalent to the Potts models that generalizes the spin models to arbitrary multivalued variables (see Reference [29] for review) and considers all possible k-ary statistical interactions in a similar way as the multispin interaction models do (k-spin interaction models that generalize pairwise and nearest-neighbor models, see Reference [30] and reference therein). I describe the computational and combinatorial restrictions from the lattice of partitions to the simplicial sub-lattice that allows one to compute in practice the information cohomology on data as in References [2,3] and that defines entropy and information landscapes and their respective paths that provide the discrete informational analog of path integrals. The combinatorics of possible interactions that are computed by the information landscapes is equivalent to the computational *"exponential wall"* encountered in many particle studies, notably density functional theory (DFT), as exposed by Kohn [31]. 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 total free-energy functional G_k of the k-body interactions (i.e., the total correlation). The total free energy is a Kullback-Leibler divergence and a special case (symmetric in the interacting body) of the free energy introduced by Baez and Pollard [32] (see also the appendix of the associated paper [2]). These definitions allow the recovery of usual equilibrium or semi-classical expressions in special cases. The set of all first critical points of information paths—a conditional-independence condition—gives a construction of the minimum free energy complex. Mutual information negativity, also called synergy [33]—a phenomenon known to provide the signature of frustrated states in glasses since the work of Matsuda [34]—is related here in the context of the more general conditional mutual information to a kind of (discrete) first-order transition, analogous to smooth phase transition in small systems [35], yet seen topologically as the critical points of a simplicial complex.

To further settle the thermodynamical interpretation of information cohomology, it is relevant to wonder what the cohomological expression of the first and second principles of thermodynamics could be. The set of information paths in simplicial structures is in bijection with the symmetric group and random processes, providing a trivial topological expression of the second law of thermodynamic as a consequence of entropy convexity, improving the theorem of Cover that needed to assume Markov conditions [36]. Thanks to the theorem of Noether [37], the first principle is expressed in terms of continuous symmetries and her theorem has been restated in more modern homological terms for finite elements by Mansfield [38] and for Markov chains by Baez and Fong [39]. Hence, the expression of the first principle in information cohomology, let as a conjecture here, should take the form of a Noether theorem for random discrete processes—that is, for the entire symmetric group S_n, a question already asked by Neuenschwander [40].

As presented by Kadanov [41] or in References [42,43], close to phase transition (notably in three dimensions), the failure of the mean field theories introduced by Van der Waals, Maxwell and Landau led to the development of renormalization group methods. Renormalization methods

intrinsically rely on asymptotic unrealistic assumptions such as an infinite number of particles [41,43] and neglect infinite quantities, which has raised fundamental criticism [44,45]. In order to compute an analog of a mean field model like van der Waals interactions [46–48], we define and compute the mean information paths that correspond to the homogeneous case of identically distributed random variables. On a dataset of expected interacting genes, the mean information path reproduces the usual behavior of the free energy in the condensed phase (i.e., with a critical point), while for genes that are less expected to interact, the 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. However, compared with the non-averaged original information paths as presented in References [2,3], it is clear that this analog mean-field approach erases the multiplicity of critical points and the diversity or richness of the complexes.

Hence, with respect to statistical physics, the main novelty brought by the information cohomology approach is the introduction of purely finite and discrete methods that can account for transition phenomena in heterogeneous systems without mean field assumptions and of new measures of correlations (i.e., multivariate mutual information that generalizes the usual correlation coefficients to non-linear relations) [2,3]. Among the work left for further studies in this preliminary line, I conjecture that the infinite dimensional continuous extension of the information cohomology formalism should be equivalent to renormalization methods, while I underline that, in practice, there is no need of these physically unrealistic assumptions.

1.4. Machine Learning Interpretation: Topological Deep Learning

The interpretation of the information cohomology in terms of machine learning and data analysis first relies on the definition of random variables as partition and of information structures and complexes on the lattice of partition [1]. Partitions are equivalent to equivalence classes (see Ellerman [49,50] for review) and hence the complex of random variables spans all possible equivalence classes of the data point. Therefore, the information cochain complex deserves the function of a universal classifier. Moreover, these information structures defined on the whole lattice of partitions encompass all possible statistical dependences and relations, since by definition it considers all possible equivalent classes on a probability space and hence fully answers to the problem raised by James and Crutchfield [51], who remarked on some partitions that are not distinguished by the I_k simplicial structure. The combinatorics of these structures forbid any computation in practice, until quantum computers become available and the severe restriction to the simplicial case of the cohomology with complexity $\mathcal{O}(2^n))$ implies that not all statistical dependencies can be estimated, as shown by James and Crutchfield [51]. The interpretation makes two phenomena coincide (i.e., condensation in statistical physics and clustering of data points in data science), which are signed by information and conditional information negativity as studied in the companion paper [2]. This generalizes the idea and results obtained on networks by the team of Grassberger [52].

On the side of applied algebraic topology, the identification of the topological structures of a dataset has motivated important research following the development of persistent homology [53–55]. Combining statistical and topological structures in a single framework remains an active challenge of data analysis that has already yielded some interesting results [56,57]. Some recent works have proposed information theoretical approaches grounded on homology, defining persistent entropy [58,59], graph topological entropy [60], spectral entropy [61] or multilevel integration entropies [62]. The present work is formally different and arose independent of persistence and provides a cohomology intrinsically based on probability for which the invariants are arguably the most important features of statistics and free of any metric assumption. The interpretation in terms of deep neural networks is also straightforward. It is preliminarily developed and applied here and in Reference [2] for the unsupervised case, while the supervised subcase is only briefly discussed here and left for further work. See Reference [63] for a presentation and preliminary applications to the digit images database of the Mixed National Institute of Standards and Technology (MNIST). The

informational approach of topological data analysis provides a direct probabilistic and statistical analysis of the structure of a dataset which allows the gap with neural network analysis to be bridged and may be a step toward their formalization in mathematics and the characterization of the network architecture necessary for a given dataset. The original work based on spin networks by Hopfield [64] formalized fully recurrent networks as n binary random variables ($N = 2$). Ackley, Hinton and Sejnowski [65] 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 [66]—relies on the maximum entropy or free-energy minimization principle and originally on minimizing the relative entropy between the network and environmental states [65]. Indeed, Reference [67] and references therein (notably see the whole opus of Marcolli with application to linguistics [68]) provides a review of the relevance of homology to artificial or natural cognition. Considering neurons as binary random variables or more generally N-ary variables (corresponding to a rate coding hypothesis) in the present context provides a homologically constrained approach of those neural networks, where the first input layer is represented by the marginal (single variable, degree 1 component) while hidden layers are associated to higher degrees. In a very naive sense, higher cohomological degrees distinguish higher-order patterns (or higher-dimensional patterns in the simplicial case), just as receptive fields of convolutional neural networks recognize higher-order features when going to higher depth-rank of neural layers as described in David Marr's original sketch [69] and now implemented efficiently in deep network structures. Notably, the notion of geodesic used in machine learning is replaced by the homotopical notion of path. On the data analysis side, it provides a new algorithm and tools for topological data analysis allowing one to rank and detect clusters and functional modules, and to make dimensionality reduction; indeed, all these classical tasks in data analysis have a direct homological meaning. I propose to call the data analysis method presented here the Poincaré-Shannon machine, since it implements simplicial homology (see Poincaré's Analysis Situs [70]) and information theory in a single framework (see Shannon's theory of communication [71]), applied effectively to empirical data.

2. Information Cohomology

This section provides a short bibliographical note on the inscription of information and probability theory within homological theories Section 2.1. We also recall the definition of information functions Section 2.2 and provide a short description of information cohomology computed in the low degrees Section 2.3, such that the interpretation of entropy and mutual information within Hochschild cohomology appears straightforward and clear. There are no new results in this section but I hope to provide a more simple and helpful presentation for some researchers outside the field of topology of what can be found in References [1,18,19] that should be considered for more precise and detailed exposition.

2.1. 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 [72,73]) and Shannon's information uncertainty theory in a common framework—a path already paved by some results on information inequalities (see Yeung's results [74]) and in algebraic geometry. In the work of Cathelineau, [75], entropy first appeared in the computation of the degree-one homology of the discrete 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 five terms of the dilogarithm implies the functional equation of entropy with four terms. Kontsevitch [76] 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 [77], where information functions appear as derivations [78]. After entropy appeared in tropical and idempotent semi-ring analysis in the study of the extension of Witt semiring to the characteristic 1 limit [79], Marcolli and Thorngren developed the thermodynamic semiring, an entropy operad that could be constructed as a deformation of the tropical semiring [20]. Introducing Rota–Baxter algebras, it allowed the derivation of a renormalization procedure [80]. In defining the category of finite probability and using Fadeev axiomatization, Baez, Fritz and Leinster could show that the only family of functions that has the functorial property is Shannon information loss [81,82]. Basing his approach on information and Koszul geometry, Boyom developed a more geometrical view of statistical models that notably considers foliations in place of the random variables [83]. Introducing a deformation theoretic framework and chain complex of random variables, Drumond-Cole, Park and Terilla [84–86] constructed a homotopy probability theory for which the cumulants coincide with the morphisms of the homotopy algebras. The probabilistic framework used here was introduced in Reference [1] and generalized to Tsallis entropies by Vigneaux [18,19]. The diversity of the formalisms employed in these independent but convergent approaches is astonishing. So, as 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 [87]. All along the development of the application to data, following the cohomology developed by References [1,18] 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 [88,89], the actual conjecture [1] is that the higher classes of information cohomology should be some kind of polylogarithmic k-form (k-differential volumes that are symmetric and additive and that correspond to the cocycle conditions for the cohomology of Lie groups [88]). 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.

2.2. Information Functions (Definitions)

The information functions used in Reference [1] and the present study were originally defined by Shannon [71] and Kullback [90] and further generalized and developed by Hu Kuo Ting [91] and Yeung [92] (see also McGill [93]). These functions include entropy, denoted $H_1 = H(X; P)$; joint entropy, denoted $H_k = H(X_1, ..., X_k; P)$; mutual information, denoted $I_2 = I(X_1; X_2; P)$; multivariate k-mutual information, denoted $I_k = I(X_1; ...; X_k; P)$; and the conditional entropy and mutual information, denoted $Y.H_k = H(X_1, ..., X_k|Y; P)$ 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 [71]:

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

  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 [90], is defined for two probability mass functions $p(x)$ and $q(x)$ by:

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

where $H(X; p(x), q(x))$ is the cross-entropy and $H(X; p(x))$ the Shannon entropy. It hence generates minus entropy as a special case, 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 [71]:

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

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 [71]:

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

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 [93] and Hu Kuo Ting [91], 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 [91].

- The total correlation introduced by Watanabe [94] called integration by Tononi and Edelman [95] or multi-information by Studený and Vejnarova [96] and Margolin and colleagues [97], which we denote $C_k(X_1; ...X_k; P)$, 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 favorable property of being a relative entropy 2 between marginal and joint-variable and hence of being always non-negative.

- The conditional entropy of X_1 knowing (or given) X_2 is defined as [71]:

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

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 [71]:

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

Conditional mutual information generates all the preceding information functions as subcases, as shown by Yeung [92]. 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 information from the data.

We have the chain rules (see Reference [36] for proofs):

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

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

which we can write more generally as (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 $H_{k+1} - H_k = (X_1, ...X_k).H(X_{k+1})$

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

which we can write in short $I_{k-1} - I_k = X_k.I_{k-1}$, generating the chain rule (10) as a 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 [1,34,91]:

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

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

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 [34]. For all random variables $X_1; ...; X_k$ with associated joint probability distribution P, we have the theorem due to Matsuda [34]

- $X_k.I(X_1;...;X_{k-1};P) \geq 0$ 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: $I_{k-1} < I_k$),

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

2.3. 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 the computation. 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 [98] corresponds to the lattice of joint random variables [1,99]. 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 References [1,18]. $(X_1, ..., X_k; P)$ 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 four atomic probabilities, with the simplicial sublattice used for data analysis. Atomic probabilities are also illustrated in a figure in the associated paper [2].

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, denoted $Y.F(X_1, ..., X_k; P)$, such that it corresponds to the usual definition of conditional entropy (Equation (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:

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

where ∂^k is the left action co-boundary that Hochschild proposed for associative and ring structures [100]. A similar construction of a random variable complex was given by Drumond-Cole, Park and Terilla [84,85]. We also consider the two other directly related cohomologies defined by considering a trivial left action [1] and a symmetric (left and right) action [22,101,102] of conditioning:

- The left action Hochschild-information coboundary and cohomology (with trivial right action):

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

This coboundary, with a trivial right action, is the usual coboundary of Galois cohomology ([103], p. 2) and in general it is the coboundary of homological algebra obtained by Cartan and Eilenberg [104] and MacLane [105] (non-homogenous bar complex).

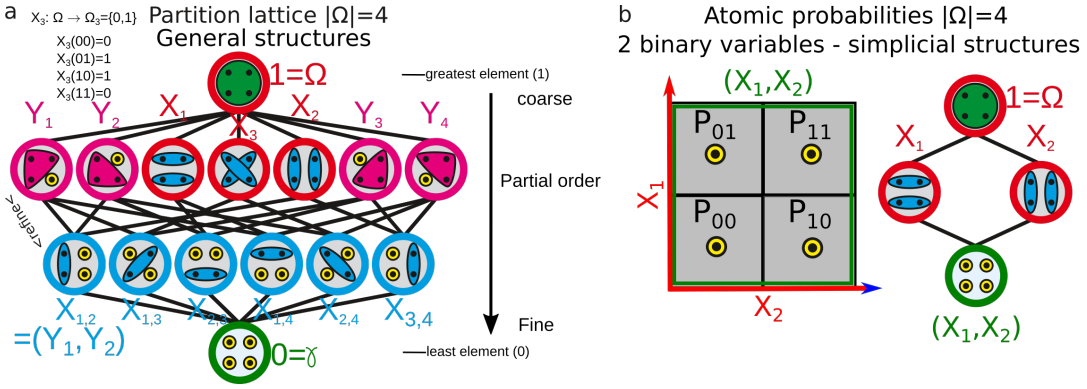


Figure 1. Example of general and simplicial information structures. (**a**) Example of lattice of random variables (partitions): the lattice of partitions of atomic-elementary events for a sample space of four atomic elements $|\Omega| = 4$ (e.g., two coins and $\Omega = \{00, 01, 10, 11\}$), each element being denoted by a black dot in the circles representing the random variables. The joint operation of random variables denoted (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 denoted $1 = \Omega$ (that we will denote 0 hereafter), 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).

- The "topological-trivial" Hochschild-information coboundary and cohomology: consider 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 [22], Kassel [102] (p. 13) and Weibel [101] (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 [22,101,102]. 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-boundary formulas, such that it appears that information functions and chain rules are homological by nature. 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.

2.3.1. 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 [106] and Lee [107], 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 [1,18]. This main theorem allows us to obtain the other information functions in what follows. Marcolli and Thorngren [20] and the group of Leinster, Fritz and Baez [81,82] independently obtained an analog result using a measure-preserving function and a characteristic one Witt construction, respectively. In these various theoretical settings, this result extends to relative entropy [1,20,82] and Tsallis entropies [18,20].
- The topological 1-coboundary (∂^1_t) is $(\partial^1_t)F(X_1; X_2) = F(X_2) - F(X_1, X_2) + F(X_1)$, which corresponds to the definition of mutual information $(\partial^1_t)F(X_1; X_2) = I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$ 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 [23], Rajski [24], Zurek [25] and Bennett [26] and has further been applied for hierarchical clustering and finding categories in data by Kraskov and Grassberger [27]: $H(X_1 \triangle X_2) = X_2.H(X_1) + X_1.H(X_2) = H(X_1, X_2) - I(X_1; X_2)$. 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 Figure 3. It generalizes to pseudo k-volumes that we define by $V_k = H_k - I_k$ (particularly interesting symmetric nonnegative functions computed by the provided software).

2.3.2. 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 the left 2-coboundary.
- The topological 2-coboundary is $(\partial^2_t)F(X_1; X_2; X_3) = F(X_2; X_3) - F((X_1, X_2); X_3) + F(X_1; (X_2, X_3)) - F(X_1; X_2)$, which corresponds in information to $\partial^2_t F(X_1; X_2; X_3) = I(X_2; X_3) - I((X_1, X_2); X_3) + I(X_1; (X_2, X_3)) - I(X_1; X_2) = 0$ 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.

2.3.3. 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.

2.3.4. 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:

- 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:

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

- 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:

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

In References [2,108] (Theorem 2), we show that the mutual independence of n variables is equivalent to the vanishing of all I_k functions for all $2 \leq k \leq n$. 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 [1,77,78], are characterized by the functional equations $I_k = 0$ and quantify statistical k-independence.

3. Simplicial Information Cohomology

3.1. 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 a 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 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 [109]. 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. This computational reef can be decreased by considering the sample size m, which is the number of trials, repetitions or points 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 Examples 1 and 4 in Reference [1], and to the python scripts shared on Github (cf. Supplementary Materials). 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, freely generating 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 subdivisions). It is a (simplicial) substructure of the general structure, since any finite lattice is a sub-lattice of the partition lattice [110]. This lattice embedding and the fact that simplicial cohomology is a special case of Hochschild cohomology can also be inferred directly from their coboundary expression and has been explicitly formalized in homology: notably, Gerstenhaber and Shack showed that a functor, denoted $\Sigma \mapsto k_\Sigma!$, induces an isomorphism between simplicial and Hochschild cohomology $H^\bullet(\Sigma, k) \cong H^\bullet(k_\Sigma!, k_\Sigma!)$ (see [111] for precisions). 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 p. 296 in Reference [112]). The information landscapes presented in Section 3.3 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 [51].

3.2. Topological Self and Free Energy of K-Body Interacting System-Poincaré-Shannon Machine

Topological Self and Free Energy of *K*-Body Interacting Systems

The basic idea behind the development of topological quantum field theories [113–115] 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. Here, in an elementary model for applied purposes, we define, in the special case of classical and discrete probability, the k-mutual information I_k (that generalize the correlation functions to nonlinear relation [116]), as the contribution of the k-body interactions to the energy functional. Some further observations support such a definition: (i) as stated in Reference [1], the signed mutual information $(-1)^k I_k$ 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, the results in Section 3.4.5 shows that these energy functionals estimated on real data behave as expected for usual k-body homogeneous formalism such as the van der Waals model or more refined density functional theory (DFT) [117,118]. Given in the context of simplicial structures, these definitions generalize to the case of a 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, that give the two following definitions.

Definition 1. *Self-internal 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 they sum 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 energy, in analogy to usual statistical physics and notably DFT:*

$$U(X_1,...,X_n;P_N) = \sum_{i=1}^{n} I_1(X_i;P_N). \tag{23}$$

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 [44,45].

Definition 2. *k-free-energy and total-free-energy: 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 kth 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 denote it $G(X_1,...,X_n;P_N)$:*

$$G(X_1,...,X_n;P_N) = \sum_{i=2}^{n}(-1)^{i-1} \sum_{I\subset[n];card(I)=i} I_i(X_I;P_N) = C_n(X_1;...X_n;P_N). \tag{24}$$

The total free energy is the total correlation (Equation (7)) introduced by Watanabe in 1960 [94] that quantifies statistical dependence in the work of Studený and Vejnarova [96] and Margolin and colleagues [97] and among other examples consciousness in the work of Tononi and Edelman [95]. In agreement with the results of Baez and Pollard in their study of biological dynamics using out-of-equilibrium formalism [32] and the appendix of the companion paper on Bayes free energy [2], the total free energy is a relative entropy. The consideration that free energy is the peculiar case of total correlation within the set of relative entropies accounts for the fact that the free energy shall be a symmetric function of the variables associated to the various bodies (e.g., $f(X;Y) = f(Y;X)$ in the pairwise interaction case). Moreover, whereas the I_k energy component can be negative, the G_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:

$$H_n(X_1,...,X_n;P_N) = U(X_1,...,X_n;P_N) - G(X_1,...,X_n;P_N). \tag{25}$$

This information-theoretic formulation of thermodynamic relation follows Jaynes [9,10], Landauer [11], Wheeler [119] and Bennett's [14] 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 any finite probability). In more probabilistic terms, it does not assume that the variables are identically distributed—a required condition for the application of classical central limit theorems (CLTs) to obtain the normal distributions in the asymptotic limit [120]. In the special case where one postulates that the probability follows the equilibrium Gibbs distribution, which is also the maximum entropy distribution [121,122], the expression of the joint entropy ($k = -1/\ln 2$) allows recovery of the equilibrium fundamental relation, as usually achieved in statistical physics (see Adami and Cerf [123] and Kapranov [124] for more details). Explicitly, let us consider Gibbs' distribution:

$$p(X_1 = x_1,...,X_n = x_n) = p_{\underbrace{ij...n}_{\text{n indices}}} = \frac{1}{Z} e^{-\beta E_{ij...n}/k_B T}, \tag{26}$$

where $E_{ij...n}$ is the energy of the elementary-atomic probability $p_{ij...n}$, k_B is Boltzmann's constant, T is the temperature and $Z = \sum_{i,j,..,n}^{N_i.N_j...N_n} e^{-E_{ij...n}/k_BT}$ 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 the Gibbs distribution gives:

$$H(X_1, ..., X_n)/k = \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n}/k_BT + \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} \ln Z = (\langle E \rangle - G)/k_BT, \tag{27}$$

which gives the expected thermodynamical relation:

$$k_BT \ln 2.H(X_1, ..., X_n) = \langle E \rangle - G = U - G, \tag{28}$$

where G is the free energy $G = -k_BT \ln Z$.

In the general case of arbitrary random variables (not necessarily i.i.d.) and discrete probability space, the identification of marginal information with internal energy

$$\sum_{k=1}^{n} H(X_k) = \sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n} \tag{29}$$

implies by direct algebraic calculus that:

$$\sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} E_{ij...n} = -\sum_{i,j,..,n}^{N_1.N_2...N_n} p_{ij...n} \ln \left(\prod_{k=i}^{n} p_{\bullet\bullet...k...\bullet} \right), \tag{30}$$

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 information $\ln p_{\bullet\bullet...k...\bullet}$. 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 [125] and in information theoretic terms can be found in the proofs of uniqueness of "single event information function" by Aczel and Darokzy ([126], p. 3). It establishes the following proposition:

Theorem 1. *Given a simplicial information structure, the elementary energies satisfying the extensivity axiom are the functions:*

$$E_{ij...n} = k \sum_{k=i}^{n} \ln p_{\bullet\bullet...k...\bullet}, \tag{31}$$

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

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

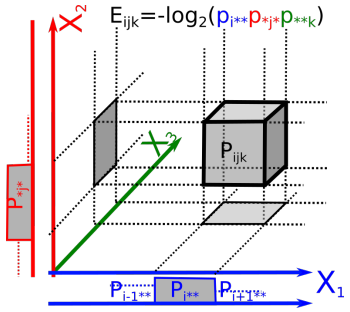


Figure 2. Elementary energy as logarithm of locally Euclidean probability volumes. Example of an elementary energy E_{ijk} associated to a probability p_{ijk} ($n = 3$ variables). The histograms of the marginal distributions of each variable are plotted beside the axes.

Examples: (i) In the example of three binary random variables ($n = 3, N_1 = N_2 = N_3 = 2$, three variables of Bernoulli) illustrated in the figure of the associated paper [2], we have $E_{000} = -\ln(p_{0\bullet\bullet}p_{\bullet0\bullet}p_{\bullet\bullet0})$, $E_{000} = -\ln(p_{000} + p_{010} + p_{001} + p_{011}) - \ln(p_{000} + p_{100} + p_{001} + p_{101}) - \ln(p_{000} + p_{100} + p_{010} + p_{110})$ and in the configuration of negative-entangled-Borromean information of the figure of the associated paper [2], 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:

$$p_{\bullet\bullet\ldots k\ldots\bullet} = e^{\frac{E_{ij\ldots n}}{nk}}. \tag{32}$$

(iii) For independent identically distributed variables (non-interacting), we have $G_n = 0$ and hence:

$$H_n(X_1, ..., X_n; P_N) = U(X_1, ..., X_n; P_N) = nH(X_i). \tag{33}$$

(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 (denoted $(X_k^1, X_k^2, X_k^3, P_k^1, P_k^2, P_k^3)$ for the kth body), it is possible to re-express the semi-classical formalism, according to which the entropy formulation is (p. 22, [127]):

$$H_{6n}(X_1^1, X_1^2, X_1^3, P_1^1, P_1^2, P_1^3, ..., P_n^3; P_N) = \log\left(\frac{\Delta X \Delta P}{(2\pi\hbar)^{6n}}\right). \tag{34}$$

This is achieved by identifying the internal and free energy as follows:

$$\langle E \rangle = -6n \log(2\pi\hbar), \tag{35}$$

$$G = -\log(\Delta X \Delta P). \tag{36}$$

This identifies the elementary volumes/probabilities with the Planck constant, the quantum of action (the consistency in the units is realized in Section 3.4.3 by the introduction of time). The quantum of action can be illustrated by considering in Figure 2 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 the non-extensivity of the volume in the phase-space due to interactions or in other words, the mutual information accounts 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 [32]. The basic principle of this expression of information theory in physics has been known at least since Jaynes's work [9,10].

As a conclusion, information topology applies—without imposing metric, symplectic or contact structures—to the 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 physics formalism. 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. In usual thermodynamics 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, the lowest temperature being the finest grain (large N) and the highest temperature being the coarsest grain (small N).

3.3. k-Entropy and k-Information Landscapes

Definition 3. *Information Landscapes: 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 the information functions of a given subset of k variables.*

In data science terms, these landscapes provide a visualization of the potentially high-dimensional structure of the data points. In information theoretic terms, it provides a representation of Shannon's work on lattices [23], further developed by Han [128]. 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 3 for $n = 4$. H_k and I_k analyze and 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 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 3 gives two theoretical extremal examples of such landscapes: one for independent and 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 3), hence allowing one to derive the normal exact expression of the information landscapes in the asymptotic 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 a finite sample are shown in References [2,3,108] and in practice the finite sample size (m) may impose some bounds on the landscapes.

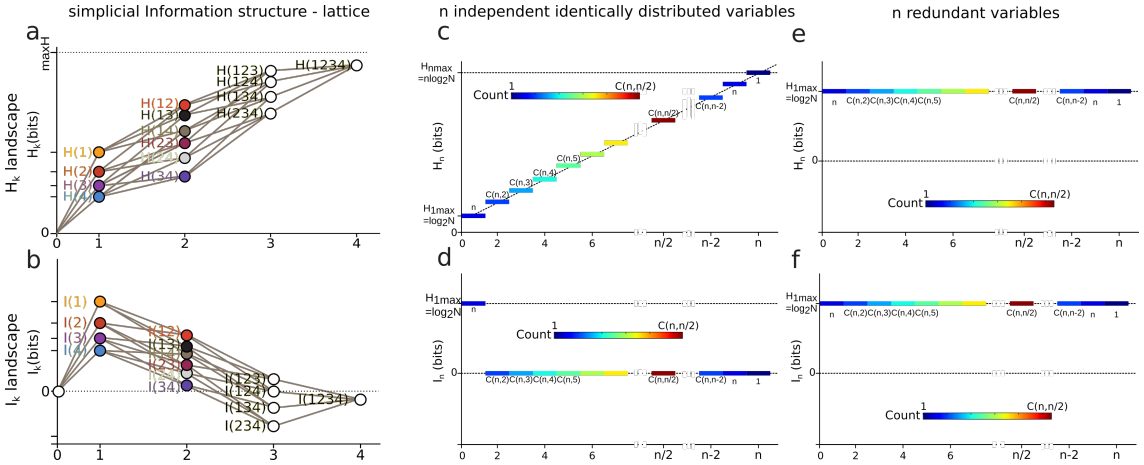


Figure 3. Entropy and information landscapes. (**a**) Illustration of the principle of an entropy H_k landscape and (**b**) of a mutual information I_k landscape for $n = 4$ random variables. The lattice of the simplicial information structure is depicted with grey lines. Theoretical examples of entropy and information landscapes. (**c**,**d**) H_k and I_k landscapes for n independent and 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. (**e**,**f**) 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.

3.4. Information Paths and Minimum Free Energy Complex

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

3.4.1. Information Paths (Definition)

Definition 4. *Information Paths: 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 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. The entropy path and information path of degree k are denoted HP_k and IP_k, respectively and the set of all information paths is denoted $\mathcal{HP}_k = \{HP_i\}_{i\in 1,\ldots,k!}$ for the entropy paths and $\mathcal{IP}_k = \{IP_i\}_{i\in 1,\ldots,k!}$ for the mutual-information paths.*

We have the theorem:

Theorem 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:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 2 & 1 & 4 & 3 \end{pmatrix} \text{ or } (01234) \xrightarrow{\sigma} (02143). \tag{37}$$

We note an information path with arrows, giving for the previous example $IP_i = (0 \rightarrow X_2 \rightarrow X_1 \rightarrow X_4 \rightarrow X_3)$. 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 [129]. The geometrical realization of information paths $\mathcal{IP}_k$ and $\mathcal{HP}_k$ consists of two dual permutohedra (see Postnikov [130]) and gives the informational version of the work of Matúš on conditional probability and permutohedra [131].

3.4.2. 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 $(X_1, ..., X_{k-1}).H(X_k; \mathbb{P})$:

$$\frac{dHP_i(k)}{dk} = H(X_1, ..., X_k; \mathbb{P}) - H(X_1, ..., X_{k-1}; \mathbb{P}) = (X_1, ..., X_{k-1}).H(X_k; \mathbb{P}). \tag{38}$$

This derivative is illustrated in the graph of Figure 4a. 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 simplicial special case of Hochschild coboundaries (Section 2.3).

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

$$\frac{dIP_i(k)}{dk} = I(X_1, ..., X_k; \mathbb{P}) - I(X_1, ..., X_{k-1}; \mathbb{P}) = -X_k.I(X_1, ..., X_{k-1}; \mathbb{P}). \tag{39}$$

This derivative is illustrated in the graph of Figure 4b. 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 2.3.

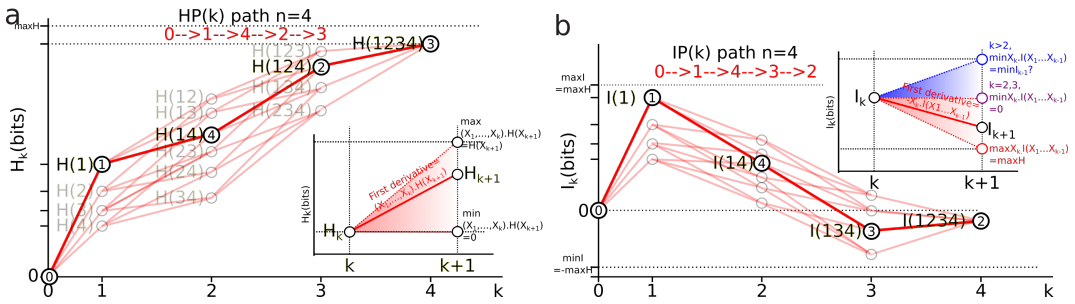


Figure 4. Entropy and information paths. Illustration of an entropy path $HP_i = 0 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 3$ (**a**) and of a mutual-information path $IP_i = 0 \rightarrow 1 \rightarrow 4 \rightarrow 3 \rightarrow 2$ (**b**) for $n = 4$ random variables (see text).

Bounds of the derivatives and information inequalities

The slope of entropy paths is bounded by the usual conditional entropy bounds ([92] pp. 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 4a). Its global upper bound 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 4a). Hence, any entropy path lies in the (convex) entropy cone defined by the three points labeled H_k, $\min H_{k+1}$ and $\max H_{k+1}$: the three vertices of the cone depicted as a red surface in Figure 4a and called the Shannonian cone following Yeung's seminal work [132]. 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 $I_2 \geq I_3$ ([92], p. 26, the opposite of Theorem 2.40, p. 30), whereas the higher limit is given by $X_i.I(X_j; X_h) \geq \min\left(X_i.H(X_j), X_i.H(X_h)\right)$ ([34] 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 due to Matsuda [34], $X_k.I(X_1; ...; X_{k-1}) < 0$ iff $I_k < I_{k+1}$. The minima correspond to zeros of conditional information (conditional independence) and hence detect cocycles in the data. The results on information inequalities define as "Shannonian" [133–135] 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 [133,134], 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 (the surface colored in red in Figure 4b) the "Shannonian" cone and the negative conditional mutual-information cone (the surface colored in blue in Figure 4b) the "non-Shannonian" cone.

3.4.3. Information Paths Are Random Processes: Topological Second Law of Thermodynamics and Entropy Rate

Here we present the dynamical aspects of information structures. Information paths directly provide 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 and so forth.

Definition 5. *Random (stochastic) process ([136]): 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. 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. Considering each symbol of a time series as a random variable, the definition of a stochastic process corresponds to the unique information paths HP_i and IP_i, whose total order is the time order

of the series. More formally, we can prove that the number of different total order on the finite set $T = 1, 2, ..., k$ with k elements is $k!$, such that we can establish a one-to-one correspondence of total orders on T with permutations on T and by Theorem 2 with entropy and information paths. Let T be a finite set with k elements and $\leq$ a total order relation on T. Consider that $k = 1$, then the set T contains only 1 element and any relation is trivially a total order on T. Now, consider that $k > 1$ and that $T = \{t_1, t_2, ..., t_n\}$. Suppose that $x \in T$. By the definition of a total order (Definition 9, p. 146 Bourbaki [137]), since $\leq$ is a total order on T then for y in T and where $x \neq y$ we have that $x < y$ or $y < x$. If $x < y, \forall y \in T$ then we define x to be the minimum element in T. If $x \neq y, \forall y \in T$ then there exists a $y \in T$ such that $y < x$ and such a process can be achieved recurrently until we obtain this minimal element. Without loss of generality, consider that t_1 is this minimal element. We then take the set $T \setminus \{t_1\}$ and repeat the preceding reasoning to find a minimal element t_2 of $T \setminus \{t_1\}$ and by recurrence we obtain $t_1 < t_2 < ... < t_k$ and hence that there are $k!$ total orders in T. □

In other words, these paths are the automorphisms of $\{1, 2.....k\} = [k]$. We immediately obtain a topological version of the second law of thermodynamics, which follows from an elementary convexity property and improves the result of Cover [36]:

Theorem 3. *(Stochastic process and information paths): Let (Ω, Δ^k, P) be a simplicial information structure and let $\mathscr{X}_T = \{X_t, t \in T\}$ be a stochastic process defined by a collection of random variables on the same probability space $(\Omega, \mathcal{B}, P)$ with cardinality $|\mathscr{X}_T| = t$, where the index set T is a totally ordered set, then the entropy $H(X_T; P)$, where X_T is the joint-random variable of the variables in $\mathscr{X}_T$ and can only increase or stay constant with t.*

Proof. given the correspondence we just established in Lemma 1, the statement is equivalent to $H(X_1, ..., X_k) \geq H(X_1, ..., X_{k-1})$ or $H(X_1, ..., X_k) - H(X_1, ..., X_{k-1}) \geq 0$ which by the chain rule of information (12) with $k = -1/\ln 2$ gives $(X_1, ..., X_{k-1}).H(X_k) \geq 0$. It is hence sufficient to prove the non-negativity of conditional entropy which can be found in Yeung ([92], p. 27). Consider the definition of conditional entropy given in 8, that we denote for simplicity:

$$H(X_k|(X_1, ..., X_{k-1}); P) = k \sum_{(x_1,...,x_{k-1})}^{N_1 \times ... \times N_{k-1}} p(x_1, ..., x_{k-1}).H(X_k|((X_1, ..., X_{k-1}) = (x_1, ..., x_{k-1}))), \quad (40)$$

with $(x_1, ..., x_{k-1}) \in (\mathscr{X}_1 \times ... \times \mathscr{X}_{k-1})$. It is hence sufficient to show that $H(X_k|((X_1, ..., X_{k-1}) = (x_1, ..., x_{k-1}))) \geq 0)$, which follows from the fact that for $0 \leq p_{(x_1,...,x_{k-1})}(x_k) \leq 1$, we have $k \ln(1/p_{(x_1,...,x_{k-1})}(x_k)) \geq 0$ with $k = -1/\ln 2$, which follows from the concavity of the logarithm. The generalization with respect to the stationary Markov condition on $\mathscr{X}_T$ used by Cover comes from the remark that in any case the indexing set of the variable is a total order. □

Remark 1. *The equality $H(X_1, ..., X_k) = H(X_1, ..., X_{k-1})$ corresponds to a statistical independence condition $I_2((X_1, ..., X_{k-1}); X_k) = 0$ and to an equilibrium condition. Note that the homological formalism imposes an "initial" minimally low entropy state $H(0) = I(0) = 0$, which is a usual assumption in physics and which corresponds to the constant and zero degree homology, which has to have at least one component to talk about the cohomology. The meaning of this theorem in common terms was summarized by Gabor and Brillouin: "you cannot have something for nothing, not even an observation" [138]. This increase in entropy is illustrated in Figure 5a. The usual stochastic approach of time series assumes a Markov chain structure, imposing peculiar statistical dependences that restrict memory effects (cf. associated paper [2] proposition 8). 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 [139], Omnes [140] and Gell-Mann and*

Hartle [141]. 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 [142] 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 4a) 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.

The stochastic process definition allows definition of the finite and asymptotic information rate:

Definition 6. *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 also let the question of the expression in information cohomology of the first principle as an open problem. Question: recently Baez and Fong published a Noether Theorem for Markov processes [39]; can we derive a Noether theorem for random discrete processes in general, that is, for all the symmetric groups S_n using the present construction? Such a theorem would provide the topological expression of the first law of thermodynamics. This question was asked by Neuenschwander [40] and related to this aim, Mansfield gave a Noether theorem for finite elements [38].

3.4.4. Local Minima and Critical Dimension

The derivative of information paths allows establishment of the lemma on which information path analysis is based. 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 3.4.2.2 [92]). □

Note that, by definition, a local minimum can be a global minimum. If it exists, we will call the dimension k of the first local minimum of an information path the first informational critical dimension of the information path IP_i and denote it k_{i_1}. This allows us to define maximal information paths:

Definition 7. *Positive information path: 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 $I_k < I_{k-1} < ... < I_1$.*

Definition 8. *Maximal positive information path: 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 positive information paths and maximal positive information paths coincide 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.

3.4.5. 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 (38)) imply that the k-free-energy can be obtained from a single path:

$$I_k = I(X_1; ...; X_k; P) = I(X_1) - \sum_{i=2}^{k} X_i . I(X_1; ...; X_{i-1}) = IP_i(1) + \sum_{j=2}^{k} \frac{dIP_i(j)}{dj}. \tag{41}$$

Hence, the global thermodynamical relation (25) 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 physics [143]. Indeed, considering an inverse relation between time and dimension $t = \frac{1}{n}$ in the probability expression (32) 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 $k!$ path functions, we can compute the mean behavior of the information structure, which is achieved by defining the mean H_k and I_k, denoted $\langle H_k \rangle$ and $\langle I_k \rangle$:

$$\langle H_k \rangle = \frac{\sum_{T \subset [n]; card(T)=k} H_k(X_T; P)}{\binom{n}{k}}, \tag{42}$$

and

$$\langle I_k \rangle = \frac{\sum_{T \subset [n]; card(T)=k} I_k(X_T; P)}{\binom{n}{k}}. \tag{43}$$

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 denoted $\langle HP \rangle(k)$ and $\langle IP \rangle(k)$ in the information landscape. The case $k = 2$ of those functions introduced in Reference [108] is studied in Merkh and Montúfar [144] with a characterization of the degeneracy of their maxima and are called factorized mutual information. 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:

$$H(X^n_{hom}; P_N) = U(X^n_{hom}; P_N) - G(X^n_{hom}; P_N). \tag{44}$$

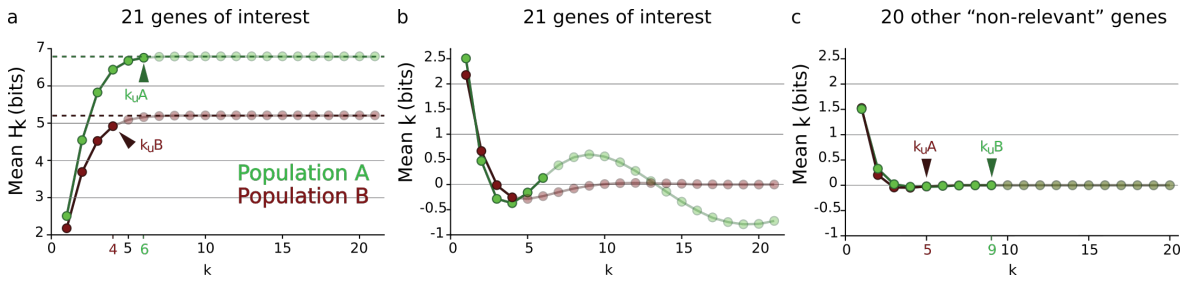


Figure 5. Example of mean entropy and information paths of gene expression. (**a**) Mean entropy path $\langle H_k \rangle$ for the 21 genes of interest for population A (green line) and population B neurons (red line). (**b**) Mean information path $\langle I_k \rangle$ for the same pool of genes. (**c**) Mean information path $\langle I_k \rangle$ for the remaining 20 genes ("non-relevant"). The undersampling dimension introduced in the associated paper [2] is depicted with arrows.

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. As is usually achieved in physics notably in mean-field theory (e.g., Weiss [145] 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 Section 4.1.3). Using the same dataset and results presented in References [2,3,108], the $\langle IP \rangle (k)$ paths estimated on a genetic expression data set are shown for two populations of neurons (A and B) in Figure 5. We quantified the gene expression levels for 41 genes in two populations of cells (A or B) as presented in References [2,3,108]. 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 References [2,3,108]. The available computational power restricted 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 5 exhibits the critical phenomenon usually encountered in condensed matter physics, like the example of van der Waals interactions [146]. 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 exhibited 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 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 exhibited 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 [147], 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 achieved here.

3.4.6. 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 that can also be interpreted as a conditional independence criterion. 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 explained in Section 3.4.4) 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 theorem that our information path analysis aims to characterize empirically:

Theorem 4. ***(Minimum free energy complex):*** *the set of all maximal 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 denoted X^{+d}. A necessary condition for this complex not to be a simplex is that its dimension is greater than 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 Reference [148], p. 95, for an example). The last part follows from Lemma 2. □

In simple words, the maximal faces (e.g., the maximal positive information paths) encode all the structures of the minimum free energy complex. Figure 6 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.

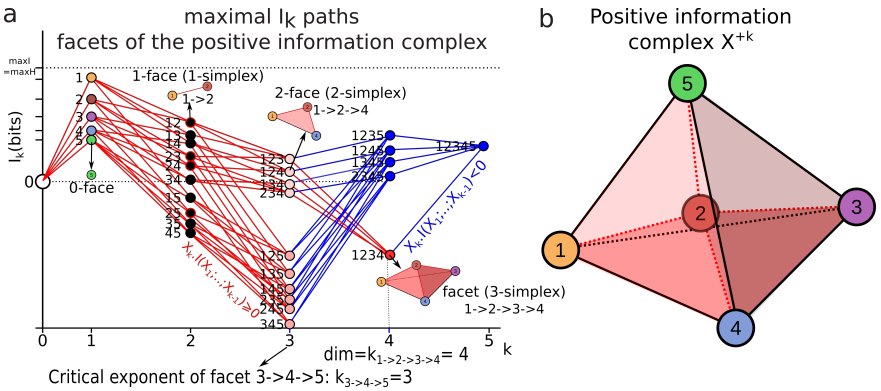


Figure 6. Example of maximal I_k paths in an I_k landscape for $n = 5$ together with its corresponding minimum free energy complex. (**a**) Maximal I_k paths in an I_k landscape for $n = 5$. The maximum positive information paths are depicted in red, for example, the paths $1 \to 2 \to 3 \to 4$ but also $4 \to 3 \to 2 \to 1$, $3 \to 4 \to 5$ and $1 \to 2 \to 5$ are maximum positive information paths (i.e., facets/maximal chains). The facet $1 \to 2 \to 3 \to 4$ is a 3-simplex while $3 \to 4 \to 5$ is a 2-simplex with critical dimension $k_{3 \to 4 \to 5} = 3$. 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 $d(X^{+k}) = d(1 \to 2 \to 3 \to 4) = 4$. (**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 four-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 1 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:

$$H^{+k}(X^{+k}; P) = \sum_{i=1}^{k} (-1)^{i-1} \sum_{I \subset X^{+}; card(I)=i} I_i(X_I; P), \tag{45}$$

where the component with dimension higher than one is a free energy. In the example of Figure 6, it gives:

$$\begin{aligned} H^{+k}(X^{+k}) = & \sum_{X_i \in X^{+k}}^{5} I(X_i) - \sum_{(X_i;X_j) \in X^{+k}}^{10} I(X_i; X_j) \\ & + \sum_{(X_i;X_j;X_h) \in X^{+k}}^{10} I(X_i; X_j; X_h) - I(X_1; X_2; X_3; X4). \end{aligned} \tag{46}$$

We propose that this complex defines a complex system:

Definition 9. *Complex system: 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 [149]. 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 networks 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 conditional independence. We called it the minimum free energy complex but could instead have called it 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 [4,150] 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 [151], has also been one of the main difficulties of condensed matter theory. Matsuda could show that I_k negativity is a signature of frustration [34]. The first axioms of DFT consider that probability densities of n' elementary bodies are each in a 3-dimensional space [117,118], defining a whole simplicial structure of dimension $n = 3n'$, commonly called the configuration space. When considered with the physical axiom of a configuration space, Theorem 4 implies that, while the minimum free 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, whose 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 [119] and Penrose [12], like spin foams [152] and causal sets [153], while here we only considered classical probabilities.

4. Discussion

4.1. Statistical Physics

4.1.1. Statistical Physics without Statistical Limit? Complexity through Finite Dimensional Non-Extensivity

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 [154] and as usual in information theory and statistical physics, 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 2.1 (see also the note of Kontsevitch [76], in the work of Baez, Fritz and Leinster [81]) and extensively in the statistical physics works of Niven [155–157], entropy does not need asymptotic or infinite assumptions such as the Stirling approximation to be derived. Rather, entropy appears without approximations in a discrete formalism equivalent to Galois cohomology [103] as a first cohomology class which has to be completed by invariants in the higher degrees. Here and in the associated paper [2], we have tried to understand, explore and exploit this observation. 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 [158] as well as Bialek, Nemenman

and Tishby proposed to be "a sign of complexity" [159], "complexity through non-extensivity" (see also Tsallis [160]). In short, we consider the non-extensivity of information before considering its asymptotic limit. Considering a statistical physics 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 Reference [161]. Finiteness and discreteness could be an essential property of the physics of complex living systems. 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 and so forth, is the combinatorial explosions of the different possible interactions: in the same way as in van der Waals paradigm, a combinatorial number of weak interactions can lead to a strong global interaction. Among all possible data structures, one is universal: data and empirical measures are discrete and finite, as emphasized by Born [16] and fully justify the cohomological approach used here originating in topos (designed by Grothendieck to hold the discrete and continuous in a single hand [162]), which was originally constructed to handle the Lie and Galois theory, as well as continuous and discrete symmetries, in a common framework. Notably, the principle originally enunciated by Willard Gibbs which considers that a phase transition is a singularity in thermodynamic behavior and occurs only in infinite systems is physically false: none of the observed systems are infinite and many of them—even small—present phase transition [35]. Cohomological methods allow the Gibbs transition principle to be corrected with respect to experience while leaving the thermodynamic theory untouched.

4.1.2. Naive Estimations Let the Data Speak

One of the striking results of the data analysis as presented here and in the associated paper [2] concerns the relatively low sample size ($m = 41$ and $m = 111$ for the analysis with cells as variables 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 Reference [97], Nemenman and colleagues, who developed the problematic of the sampling problem, state in the introduction that "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 methods applied here are quite elementary. They do 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 i.i.d. 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 ([163], chap. 1), which he called the "generalized fields of probability" because they do not assume the sixth axiom of continuity. Rather than fixing a model with priors, the present formalism allows the raw data to freely impose their specific structure to the model, which is usually called the naive approach or naive estimation. If one accepts that a frequentist theory and interpretation of probability is mathematically valid ([163], 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 physics recalled by Born in his Nobel lecture [16]. This frequentist elementary consideration is not mathematically trivial, 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 to our knowledge has not been achieved in the context of probability and information.

4.1.3. Discrete Informational Analog of Renormalization Methods: No Mean-Field Assumptions Let the Objects Differentiate

To our best knowledge, all previous studies that tried to quantify statistical dependences using information methods with more than three variables used total correlation [97,164] and crucially assumed that the interaction between the variables are homogeneous, which corresponds to the usual mean field assumption and to the identically distributed case of mean information (Section 3.4.5) presented here. The proposed combinatorial decomposition allows heterogeneous classes within the set of variables to be identified [2,3], which would not have been possible using homogeneous assumptions, just as renormalization methods allowed the failures of mean field models to be overcome [41].

4.1.4. Combinatorial, Infinite, Continuous and Quantum Generalizations

In place of the usual assumptions of statistical physics and in order to compute effectively information structures on the data, we had to concede a severe restriction to the simplest simplicial combinatoric, leaving the whole multinomial combinatoric as potential and relevant extensions of the present data analysis. It is likely (and left as an open question) that the generalization of the binomial combinatoric presented here to the multinomial case will allow extension of the theorem obtained in the associated paper that shows that informations provide co-ordinates on the probability simplex for binary variables (Theorem 3 in Reference [2]). Even larger extensions have already been achieved theoretically, notably by Vigneaux, who developed q-multinomial deformations of this combinatorics associated with Tsallis entropies [19,165,166]. Moreover, as shown in the PhD of Vigneaux, this finite and discrete elementary setting extends nicely to the infinite and continuous case [19] within cohomology theory. If those structures currently appear gigantic and out of computational reach, one should expect that the future generations of computing devices will allow us to investigate them. On the side of the generalization of the information cohomology to quantum information [1,19] (see also Maniero [21]), Adami and Cerf showed that quantum conditional entropies and I_2 can be negative and that it happens precisely for entangled systems according to Bell inequalities [167,168]. The framework of Adami and Cerf provides a natural way to generalize the present classical I_k structures to the quantum case 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 in Section 3.4.6, quantum entanglement should allow non-simplex quantum minimum free energy complexes to happen in dimension below 3—a promising topological insight into the possible non-locality of the configuration space.

4.2. Data Science

4.2.1. Topological Data Analysis

Thanks to the recent success of topological data analysis, a recurrent question has concerned the relation of information cohomology with persistence homology. The answer is not easy, since both appeared independently and are quite different in formalism, algorithms and results on the dataset. It is possible to provide an intuitive non-rigorous interpretation of the present work in persistence terms. Most of the persistent methods consist of approximating the birth and death of the Betti's numbers of the Čech complex obtained by considering balls around each data point while the radius of the ball grows. The Čech complex is given by the intersection of the balls and for combinatorial computational reasons, most of the algorithms restrict to pairwise intersections, giving the Vietoris–Ripps complex as

an approximation of the Čech complex. Our method focuses on the intersection of random variables rather than balls around data points: a theorem of Hu Kuo Ting [91] (Theorem 1 in Reference [2]) shows the equivalence of mutual-information functions with set theoretic finite measurable functions endowed with the intersection operator, formalizing the usual naive Venn diagram interpretation of mutual information. Hence, leaving the mathematical rigor to allow an interpretation of the current algorithm in the common language of persistence, I compute here an information-theoretic analog of the Čech complex and it is not excluded that this analogy can be made formal and notably to establish some nerve theorem for information structures (see Oudot for a review [169]). It turns out that "zero information intersections" is exactly equivalent to statistical independence (Theorem 2 in Reference [2]). Then, the balls can be viewed as a local estimation of the probability density under a metric space assumption, while computing the persistence is roughly analog to varying the graining of the probability estimation as achieved in Section 5.6 of Reference [2] and the maximal mutual-information coefficient method proposed by Reshef et al. [116]. Hence, in regard to current topological data analysis methods, the methods presented here provide an intrinsically probabilistic cohomological framework: the differential operators are fundamental maps in probability-information theory. As a consequence, no metric assumption is required a priori: in practice, it is possible to compute the homology, for example, on position variables and/or on qualitative variables such as "nice" and "not nice" or "Alice" and "Bob". The present method is topological and avoids the a priori introduction of such a metric; rather, a family of Shannon's pseudometric emerges from the formalism as a first cohomological class (Section 2.3.1). Considering a symmetric action of conditioning, we obtain Shannon's metric parametrized by a scalar multiplicative constant.

4.2.2. Unsupervised and Supervised Deep Homological Learning

As underlined in the Introduction, complexes of random variables and minimum free energy complex can be understood as providing a geometrically constrained architecture to deep neural networks, where the depth of the neural network corresponds to the dimension of the cochain complex and marginal variables and information I_1 correspond to the input layer. Notably, the multiplicity of facets forming the complexes allows consideration of neural networks with parallel layer architecture to analyze conditionally independent features of the input, as effectively achieved in real nervous systems, such as the macroscopic "where and what" (dorso and ventral, respectively) visual processing streams in the vertebrate cortex [170,171]. Moreover, the energy functional interpretation of mutual information I_k and of total correlation G_k functions directly follow and generalize the definitions of the hierarchical deep architectures of Boltzmann machines and of Helmholtz machines [65,66] that explicitly expressed free energy in terms of Kullback-Leibler (KL) divergences between layers. Notably, the introduction of the multiplicity decomposition of "energy functions" formalizes unsupervised learning in neural networks in terms of a combinatorial family of analytically independent functions I_k with independent gradients (Theorem 4 in Reference [2]): instead of a single energy and associated gradient descent, mutual information provides a multiplicity of gradients. The application scope is restricted to the general problem of unsupervised learning here but the supervised subcase can be proposed from this and is detailed with some examples of application to the digits MNIST images dataset presented here [63] and in a future publication. Notably, the mathematical depth of the back-propagation algorithm [172–174] comes from the fact that it implements the chain rule for derivation with respect to the parameters, allowing learning in the class of differentiable functions. In information cohomology, supervised learning appears as a subcase of unsupervised learning and defines the sublattice, information landscapes and complexes for which all chains contain the label variable X_i to be learned and the 2^{n-1} H_k , I_k which (sub)gradients are independent in open dense subsets of the probability subsimplex Δ_{X/X_i} (subsimplex obtained by conditioning on the parameters E_{X_i}, the conditional probability laws with respect to X_i). Notably, the marginal degree 1 layer is only composed of the label variable X_i, that correponds to the usual definition of an input, and the simplicial structure is translated by one degree. Then, the maximal depth of a deep neural network achieving

the classification given the data is the dimension of the information simplicial complex—a result that can be linked to the work of Montúfar on the dimension of restricted Boltzmann machine [175]. In this cohomological context, the back-propagation is implemented by the information chain rule and is forward as imposed the cohomology contra-variant nature. The relation to information geometry and Amari's natural gradient [176] follows from the relation of the Fisher information with the Hessian of the KL divergence [177] (or entropy [178]) and from the implementation of natural gradients for deep network training [179] and is left for further courageous investigations.

4.2.3. Epigenetic Topological Learning—Biological Diversity

In place of the MaxEnt principle, we proposed a least energy principle equivalent here to a homological complex (finite and without metric assumptions). Mathematically, profited from 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, corresponding to the "non-Shannonian set" [133–135] and conjectured to be at least as rich as topological links can be [2]. 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 presented in the associated paper [2] gives a preliminary validation of this quantification. Moreover, the definition of a complex system such as the minimum free energy complex given in Section 3.4.6, underlining that diversity is just the multiplicity of the minima, is in agreement with Waddington's original work [180] (see Figure 7b). 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 thermodynamics. Doing so, however, it will be able to take different paths: diversity comes from the multiple minima. Waddington's explanation of this landscape is as 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 (Section 3.4.5), 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 [181] and of probability transitions [182], respectively.

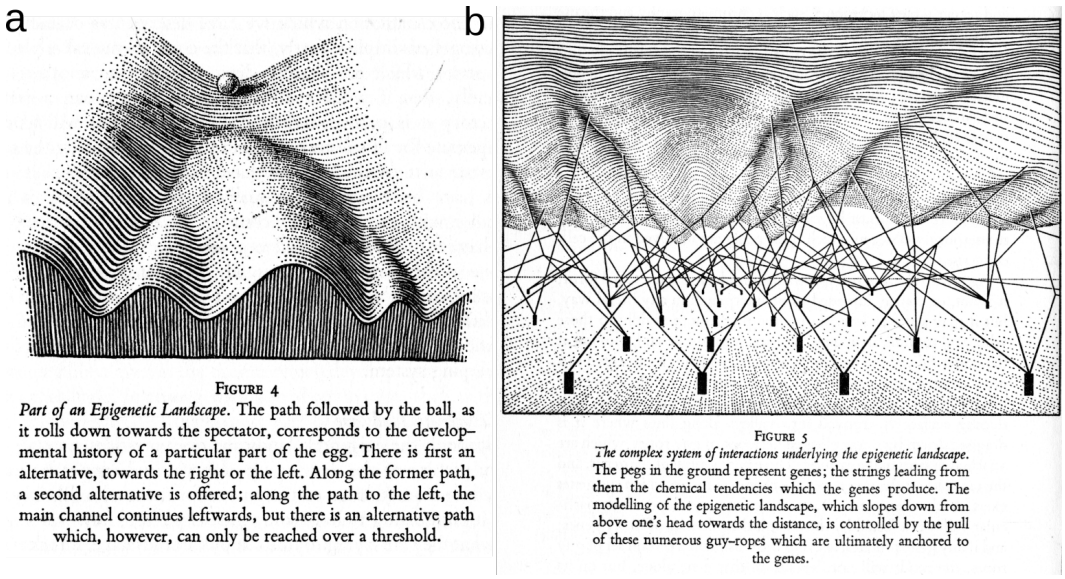


Figure 7. The epigenetic landscape of Waddington. (**a**) The epigenetic landscape of Waddington. A path of the ball in this landscape illustrates a cell's developmental fate. (**b**) "The complex system of interactions underlying the epigenetic landscape", with Waddington's original legends [180].

Following Thom's topological morphogenetic view of Waddington's work [183], 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 7). This formalization of 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, as well as Bell and Sejnowski [184–186]) can be viewed as binary variable subcases of a generic N-ary variable epigenetic developmental process. For example, Potts models were implemented for the simulation of cell-based morphogenesis by Glazier and colleagues [187]. 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 [188]. In terms of the current problematics of neuroscience, such generalization allows, on a formal level, consideration of 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 past decades on the application of statistical physics to biological systems (some of them cited in this article), we propose that the epigenetic process implements the first two laws of thermodynamics, weak topological versions of which are proposed to hold in the raw data space (without phase space or symplectic structure, cf. Section 3.4.3). As previously underlined, the condition for such an inscription of living organism dynamics into classical statistical physics to be legitimate is that the considered variables correspond to phase space variables.

5. Conclusions

In this paper we have proposed a unified view and account for classical statistical physics and machine learning methods like topological data analysis and deep neural networks under a common information cohomology framework. The particularity and main novelty of the methods are that they are finite and discrete and that such axioms, rather than being a default, can have fundamental mathematical (Galois theory [103]), probabilistic (Kolmogorov foundations of finite probability field [163]), physical (frequentist hypothesis and observability axiom [16]) and computational (directly computable combinatorial expressions) and machine learning (numerical pattern classification) meaning. From the statistical physics point of view, we developed within the information cohomology framework an expression of internal and free energy, of the second law and conjecturally of the first law of thermodynamics, while developing a generic k-body interaction model that can be viewed as a dicrete-finite and informational analog to renormalization methods (and hence more physically sound [44,45]), avoiding mean-field approximations. The application of these informational methods to genetic expression data reproduced the usual signatures of phase transition of mean-field k-body models [46–48]. From the machine learning point of view, this paper provides a combinatorial and computational expression of the information cohomology and notably of the simplicial subcase that can be effectively computed and underline its universal classifier function. The methods applied in References [2,3] offer new algorithms for topological data analysis which are intrinsically probabilistic, as well as acohomological insight into deep neural network architecture and training/learning algorithms, both supervised and unsupervised, while providing a generic biological model of epigenetic plasticity and development.

Supplementary Materials: The software Infotopo is available at https://github.com/pierrebaudot/INFOTOPO.

Funding: This work was funded by the European Research Council (ERC consolidator grant 616827 *CanaloHmics* to J.M. Goaillard) and Median Technologies, developed at UNIS Inserm 1072—Université Aix-Marseille.

Acknowledgments: Thanks previously to support and hosting since 2007 of the Max Planck Institute for Mathematics in the Sciences (MPI-MIS), the Complex Systems Institute Paris-Ile-de-France (ISC-PIF), and the Institut de Mathématiques de Jussieu - Paris Rive Gauche (IMJ-PRG). This work addresses a deep and warm

acknowledgement to the researchers who helped its realization: D. Bennequin, J.M. Goaillard, Hong Van le, G. Marrelec, M. Tapia, and J.P. Vigneaux; or supported and encouraged it: H. Atlan, F. Barbaresco, H. Bénali, P. Bourgine, F. Chavane, J. Jost, A. Mohammad-Djafari, J.P. 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 and preprints 2018 [108].

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

Abbreviations

The following abbreviations are used in this manuscript:

iid	Independent identically distributed
H_k	Multivariate k-joint Entropy
I_k	Multivariate k-mutual information
G_k	Multivariate k-total-correlation or k-multi-information

References

1. Baudot, P.; Bennequin, D. The Homological Nature of Entropy. *Entropy* **2015**, *17*, 3253–3318. [CrossRef]
2. Baudot, P.; Tapia, M.; Bennequin, D.; Goaillard, J.C. Topological Information Data Analysis. *Entropy* **2019**, *in press*. [CrossRef]
3. Tapia, M.; Baudot, P.; Formizano-Tréziny, C.; Dufour, M.; Temporal, S.; Lasserre, M.; Marquèze-Pouey, B.; Gabert, J.; Kobayashi, K.; Goaillard, J.M. Neurotransmitter identity and electrophysiological phenotype are genetically coupled in midbrain dopaminergic neurons. *Sci. Rep.* **2018**, *8*, 13637. [CrossRef] [PubMed]
4. Mézard, M. Passing Messages Between Disciplines. *Science* **2003**, *301*, 1686. [CrossRef] [PubMed]
5. Caramello, O. The unification of Mathematics via Topos Theory. *arXiv* **2010**, arXiv:1006.3930. Available online: https://arxiv.org/abs/1006.3930 (accessed 4 September 2019).
6. Doering, A.; Isham, C. Classical and quantum probabilities as truth values. *J. Math. Phys.* **2012**, *53*, 032101. [CrossRef]
7. Doering, A.; Isham, C. A Topos Foundation for Theories of Physics: I. Formal Languages for Physics. *J. Math. Phys.* **2008**, *49*, 053515. [CrossRef]
8. Brillouin, L. *Science and Information Theory*; Academic Press: Cambridge, MA, USA, 1956.
9. Jaynes, E. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *108*, 171–190. [CrossRef]
10. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
11. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
12. Penrose, R. Angular Momentum: An Approach to Combinatorial Space-Time. In *Quantum Theory and Beyondum*; Cambridge University Press: Cambridge, UK, 1971; pp. 151–180.
13. Wheeler, J. Information, Physics, Quantum: The Search for Links. In *Complexity, Entropy, and the Physics of Information*; Zureked, W.H., Ed.; CRC Press: New York, NY, USA, 1990.
14. Bennett, C. Notes on Landauer's principle, Reversible Computation and Maxwell's Demon. *Stud. Hist. Philos. Mod. Phys.* **2003**, *34*, 501–510. [CrossRef]
15. Wheeler, J.A. *Physics and Austerity, Law Without Law*; Numéro 122 de Pamhlets on Physics; Center for Theoretical Physics, University of Texas: Austin, TX, USA, 1982.
16. Born, M. The Statistical Interpretation of Quantum Mechanics. Nobel Lecture. 1954. Available online: https://www.nobelprize.org/prizes/physics/1954/born/lecture/ (accessed on 4 September 2019).
17. Baudot, P. Natural Computation: Much Ado about Nothing? An Intracellular Study of Visual Coding in Natural Condition. Ph.D. Thesis, Universit'e Pierre et Marie Curie-Paris VI, Paris, France, September 2006.
18. Vigneaux, J. The Structure of Information: From Probability to Homology. *arXiv* **2017**, arXiv:1709.07807. Available online: https://arxiv.org/abs/1709.07807 (accessed on 4 September 2019).
19. Vigneaux, J. Topology of Statistical Systems. A Cohomological Approach to Information Theory. Ph.D. Thesis, Paris 7 Diderot University, Paris, France, June 2019.
20. Marcolli, M.; Thorngren, R. Thermodynamic Semirings. *arXiv* **2011**, arXiv 10.4171/JNCG/159. Available online: https://arxiv.org/abs/1108.2874 (accessed on 4 September 2019).
21. Maniero, T. Homological Tools for the Quantum Mechanic. *arXiv* **2019**, arXiv:1901.02011. Available online: https://arxiv.org/abs/1901.02011 (accessed on 4 September 2019).

22. Gerstenhaber, M.; Schack, S. A hodge-type decomposition for commutative algebra cohomology. *J. Pure Appl. Algebr.* **1987**, *48*, 229–247. [CrossRef]
23. Shannon, C. The lattice theory of information. In *Transactions of the IRE Professional Group on Information Theory*; IEEE: Piscataway, NJ, USA, 1953; Volume 1, pp. 105–107.
24. Rajski, C. A metric space of discrete probability distributions. *Inform. Control* **1961**, *4*, 371–377. [CrossRef]
25. Zurek, W. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* **1989**, *341*, 119–125. [CrossRef]
26. Bennett, C.H.; Gács, P.; Li, M.; Vitányi, P.M.; Zurek, W.H. Information distance. *IEEE Trans. Inf. Theory* **1998**, *44*, 1407–1423. [CrossRef]
27. Kraskov, A.; Grassberger, P. MIC: Mutual Information Based Hierarchical Clustering. In *Information Theory and Statistical Learning*; Springer: Boston, MA, USA, 2009; pp. 101–123.
28. Crutchfield, J.P. Information and its metric. In *Nonlinear Structures in Physical Systems*; Springer: New York, NY, USA, 1990; pp. 119–130.
29. Wu, F. The Potts model. *Rev. Mod. Phys.* **1982**, *54*, 235–268. [CrossRef]
30. Turban, L. One-dimensional Ising model with multispin interactions. *J. Phys. A-Math. Theor.* **2016**, *49*, 355002. [CrossRef]
31. Kohn, W. Nobel Lecture: Electronic structure of matter—Wave functions and density functionals. *Rev. Mod. Phys.* **1999**, *71*, 1253. [CrossRef]
32. Baez, J.; Pollard, S. Relative Entropy in Biological Systems. *Entropy* **2016**, *18*, 46. [CrossRef]
33. Brenner, N.; Strong, S.P.; Koberle, R.; Bialek, W.; de Ruyter van Steveninck, R.R. Synergy in a neural code. *Neural Comput.* **2000**, *12*, 1531–52. [CrossRef] [PubMed]
34. Matsuda, H. Information theoretic characterization of frustrated systems. *Physica A* **2001**, *294*, 180–190. [CrossRef]
35. Dunkel, J.; Hilbert, S. Phase transitions in small systems: Microcanonical vs. canonical ensembles. *Physica A* **2006**, *370*, 390–406. [CrossRef]
36. Cover, T.; Thomas, J. *Elements of Information Theory*; John Wiley & Sons: Chichester, UK, 1991.
37. Noether, E. Invariant Variation Problems. *Transport Theor. Stat.* **1971**, *1*, 186–207. [CrossRef]
38. Mansfield, E.L. Noether's Theorem for Smooth, Difference and Finite Element Systems. In *Foundations of Computational Mathematics, Santander*; Cambridge University Press: Cambridge, UK, 2005; pp. 230–257.
39. Baez, J.; Fong, B. A Noether theorem for Markov processes. *J. Math. Phys.* **2013**, *54*, 013301. [CrossRef]
40. Neuenschwander, D. Noether's theorem and discrete symmetries. *Am. J. Phys.* **1995**, *63*, 489. [CrossRef]
41. Kadanoff, L.P. Phase Transitions: Scaling, Universality and Renormalization. Phase Transitions Dirac V2.4. 2010. Available online: https://jfi.uchicago.edu/~leop/TALKS/Phase20TransitionsV2.4Dirac.pdf (accessed on 4 September 2019).
42. Wilson, K.G.; Kogut, J. The renormalization group and the epsilon expansion. *Phys. Rep.* **1974**, *12*, 75–200. [CrossRef]
43. Zinn-Justin, J. *Phase Transitions and Renormalization Group*; Oxford University Press: Oxford, UK, 2010.
44. Feynman, R. *QED. The Strange Theory of Light and Matter*; Princeton University Press: Princeton, NJ, USA, 1985.
45. Dirac, P. *Directions in Physics*; John Wiley & Sons: Chichester, UK, 1978.
46. Van der Waals, J.D. *Over de Continuiteit van den Gas- en Vloeistoftoestand*; Luitingh-Sijthoff: Amsterdam, The Netherlands, 1873.
47. Maxwell, J. Van der Waals on the Continuity of the Gaseous and Liquid States. *Nature* **1874**, *10*, 407.
48. Maxwell, J. On the dynamical evidence of the molecular constitution of bodies. *Nature* **1875**, *1*, 357–359. [CrossRef]
49. Ellerman, D. An introduction to partition logic. *Log. J. IGPL* **2014**, *22*, 94–125. [CrossRef]
50. Ellerman, D. The logic of partitions: introduction to the dual of the logic of subsets. In *The Review of Symbolic Logic*; Cambridge University Press: Cambridge, UK, 2010; Volume 3, pp. 287–350.
51. James, R.; Crutchfield, J. Multivariate Dependence beyond Shannon Information. *Entropy* **2017**, *19*, 531. [CrossRef]
52. Foster, D.; Foster, J.; Paczuski, M.; Grassberger, F. Communities, clustering phase transitions, and hysteresis: pitfalls in constructing network ensembles. *Phys. Rev. E* **2010**, *81*, 046115. [CrossRef]

53. 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*, 1236. [CrossRef] [PubMed]
54. Epstein, C.; Carlsson, G.; Edelsbrunner, H. Topological data analysis. *Inverse Probl.* **2011**, *27*, 120201.
55. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308. [CrossRef]
56. Niyogi, P.; Smale, S.; Weinberger, S. A Topological View of Unsupervised Learning from Noisy Data. *SIAM J. Comput.* **2011**, *20*, 646–663. [CrossRef]
57. Buchet, M.; Chazal, F.; Oudot, S.; Sheehy, D. Efficient and Robust Persistent Homology for Measures. *Comput. Geom.* **2016**, *58*, 70–96. [CrossRef]
58. Chintakunta, H.; Gentimis, T.; Gonzalez-Diaz, R.; Jimenez, M.J.; Krim, H. An entropy-based persistence barcode. *Pattern Recognit.* **2015**, *48*, 391–401. [CrossRef]
59. Merelli, E.; Rucco, M.; Sloot, P.; Tesei, L. Topological Characterization of Complex Systems: Using Persistent Entropy. *Entropy* **2015**, *17*, 6872–6892. [CrossRef]
60. 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. [CrossRef]
61. Maletic, S.; Rajkovic, M. Combinatorial Laplacian and entropy of simplicial complexes associated with complex networks. *Eur. Phys. J.* **2012**, *212*, 77–97. [CrossRef]
62. Maletic, S.; Zhao, Y. Multilevel Integration Entropies: The Case of Reconstruction of Structural Quasi-Stability in Building Complex Datasets. *Entropy* **2017**, *19*, 172. [CrossRef]
63. Baudot, P.; Bernardi, M. Information Cohomology Methods for Learning the Statistical Structures of Data. DS3 Data Science, Ecole Polytechnique. 2019. Available online: https://www.ds3-datascience-polytechnique.fr/wp-content/uploads/2019/06/DS3-426_2019_v2.pdf (accessed on 4 September 2019).
64. Hopfield, J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [CrossRef]
65. Ackley, D.; Hinton, G.; Sejnowski, T.J. A Learning Algorithm for Boltzmann Machines. *Cogn. Sci.* **1985**, *9*, 147–169. [CrossRef]
66. Dayan, P.; Hinton, G.; Neal, R.; Zemel, R. The Helmholtz Machine. *Neural Comput.* **1995**, *7*, 889–904. [CrossRef]
67. Baudot, P. Elements of Consciousness and Cognition. Biology, Mathematic, Physics and Panpsychism: An Information Topology Perspective. *arXiv* **2018**, arXiv:1807.04520. Available online: https://arxiv.org/abs/1807.04520 (accessed on 4 September 2019).
68. Port, A.; Gheorghita, I.; Guth, D.; Clark, J.; Liang, C.; Dasu, S.; Marcolli, M. Persistent Topology of Syntax. In *Mathematics in Computer Science*; Springer: Berlin/Heidelberger, Germany, 2018; Volume 1, pp. 33–50.
69. Marr, D. *Vision*; MIT Press: Cambridge, MA, USA, 1982.
70. Poincare, H. Analysis Situs. *Journal de l'École Polytechnique* **1895**, *1*, 1–121.
71. Shannon, C.E. A Mathematical Theory of Communication. *Bell Labs Tech. J.* **1948**, *27*, 379–423. [CrossRef]
72. Andre, Y. *Symétries I. Idées Galoisiennes*; Chap 4 of Leçons de Mathématiques Contemporaines à l'IRCAM, Master; cel-01359200; IRCAM: Paris, France, 2009. Available online: https://cel.archives-ouvertes.fr/cel-01359200 (accessed on 5 september 2019).
73. Andre, Y. Ambiguity theory, old and new. *arXiv* **2008**, arXiv:0805.2568. Available online: https://arxiv.org/abs/0805.2568 (accessed on 4 September 2019).
74. Yeung, R. On Entropy, Information Inequalities, and Groups. In *Communications, Information and Network Security*; Springer, Boston, MA, USA, 2003; Volume 712, pp. 333–359.
75. Cathelineau, J. Sur l'homologie de sl2 a coefficients dans l'action adjointe. *Math. Scand.* **1988**, *63*, 51–86. [CrossRef]
76. Kontsevitch, M. The 11/2 Logarithm. 1995, Unpublished work.
77. Elbaz-Vincent, P.; Gangl, H. On poly(ana)logs I. *Compos. Math.* **2002**, *130*, 161–214. [CrossRef]
78. Elbaz-Vincent, P.; Gangl, H. Finite polylogarithms, their multiple analogues and the Shannon entropy. In *International Conference on Geometric Science of Information*; Springer: Berlin/Heidelberger, Germany, 2015.
79. Connes, A.; Consani, C. Characteristic 1, entropy and the absolute point. In *Noncommutative Geometry, Arithmetic, and Related Topics*; JHU Press: Baltimore, MD, USA, 2009.

80. Marcolli, M.; Tedeschi, R. Entropy algebras and Birkhoff factorization. *J. Geom. Phys.* **2018**, *97*, 243–265. [CrossRef]
81. Baez, J.; Fritz, T.; Leinster, T. A Characterization of Entropy in Terms of Information Loss. *Entropy* **2011**, *13*, 1945–1957. [CrossRef]
82. Baez, J.C.; Fritz, T. A Bayesian characterization of relative entropy. *Theory Appl. Categ.* **2014**, *29*, 422–456.
83. Boyom, M. Foliations-Webs-Hessian Geometry-Information Geometry-Entropy and Cohomology. *Entropy* **2016**, *18*, 433. [CrossRef]
84. Drummond-Cole, G.; Park, J.S.; Terilla, J. Homotopy probability theory I. *J. Homotopy Relat. Struct.* **2015**, *10*, 425–435. [CrossRef]
85. Drummond-Cole, G.; Park, J.S.; Terilla, J. Homotopy probabilty Theory II. *J. Homotopy Relat. Struct.* **2015**, *10*, 623–635. [CrossRef]
86. Park, J.S. Homotopy Theory of Probability Spaces I: Classical independence and homotopy Lie algebras. *aiXiv* **2015**, arXiv:1510.08289. Available online: https://arxiv.org/abs/1510.08289 (accessed on 4 September 2019).
87. Beilinson, A.; Goncharov, A.; Schechtman, V.; Varchenko, A. Aomoto dilogarithms, mixed Hodge structures and motivic cohomology of pairs of triangles on the plane. In *The Grothendieck Festschrift*; Birkhäuser: Boston, MA, USA, 1990; Volume 1, pp.135–172.
88. Aomoto, K. Addition theorem of Abel type for Hyper-logarithms. *Nagoya Math. J.* **1982**, *88*, 55–71. [CrossRef]
89. Goncharov, A. *Regulators*; Springer: Berlin/Heidelberger, Germany, 2005; pp. 297–324.
90. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
91. Hu, K.T. On the Amount of Information. *Theory Probab. Appl.* **1962**, *7*, 439–447.
92. Yeung, R. *Information Theory and Network Coding*; Springer: Berlin/Heidelberger, Germany, 2007.
93. McGill, W. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116. [CrossRef]
94. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–81. [CrossRef]
95. Tononi, G.; Edelman, G. Consciousness and Complexity. *Science* **1998**, *282*, 1846–1851. [CrossRef]
96. Studeny, M.; Vejnarova, J. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in Graphical Models*; MIT Press: Cambridge, MA, USA, 1999; pp. 261–296.
97. Margolin, A.; Wang, K.; Califano, A.; Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst. Biol.* **2010**, *4*, 428–440. [CrossRef] [PubMed]
98. Andrews, G. *The Theory of Partitions*; Cambridge University Press: Cambridge, UK, 1998.
99. Fresse, B. Koszul duality of operads and homology of partitionn posets. *Contemp. Math. Am. Math. Soc.* **2004**, *346*, 115–215.
100. Hochschild, G. On the cohomology groups of an associative algebra. *Ann. Math.* **1945**, *46*, 58–67. [CrossRef]
101. Weibel, C. *An Introduction to Homological Algebra*; Cambridge University Press: Cambridge, UK, 1995.
102. Kassel, C. Homology and Cohomology of Associative Algebras—A Concise Introduction to Cyclic Homology. Advanced Course on Non-Commutative Geometry. 2004. Available online: https://cel.archives-ouvertes.fr/cel-00119891/ (accessed on 4 September 2019).
103. Tate, J. *Galois Cohomology*; Springer Science & Business Media: Berlin/Heidelberger, Germany, 1991.
104. Cartan, H.; Eilenberg, S. *Homological Algebra*; Princeton University Press: Princeton, NJ, USA, 1956.
105. Mac Lane, S. *Homology*; Springer Science & Business Media: Berlin/Heidelberger, Germany, 1975.
106. Kendall, D. Functional Equations in Information Theory. *Probab. Theory Relat. Field* **1964**, *2*, 225–229. [CrossRef]
107. Lee, P. On the Axioms of Information Theory. *Ann. Math. Stat.* **1964**, *35*, 415–418. [CrossRef]
108. Baudot, P.; Tapia, M.; Goaillard, J. Topological Information Data Analysis: Poincare-Shannon Machine and Statistical Physic of Finite Heterogeneous Systems. Preprints 2018040157. 2018. Available online: https://www.preprints.org/manuscript/201804.0157/v1 (accessed on 5 September 2019).
109. Lamarche-Perrin, R.; Demazeau, Y.; Vincent, J. The Best-partitions Problem: How to Build Meaningful Aggregations? In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 17–20 November 2013; p. 18.
110. Pudlák, P.; Tůma, J. Every finite lattice can be embedded in a finite partition lattice. *Algebra Univ.* **1980**, *10*, 74–95. [CrossRef]

111. Gerstenhaber, M.; Schack, S. Simplicial cohomology is Hochschild Cohomology. *J. Pure Appl. Algebr.* **1983**, *30*, 143–156. [CrossRef]
112. Steenrod, N. Products of Cocycles and Extensions of Mapping. *Ann. Math.* **1947**, *48*, 290–320. [CrossRef]
113. Atiyah, M. Topological quantum field theory. *Publ. Math. IHÉS* **1988**, *68*, 175–186. [CrossRef]
114. Witten, E. Topological Quantum Field Theory. *Commun. Math. Phys.* **1988**, *117*, 353–386. [CrossRef]
115. Schwarz, A. Topological Quantum Field Theory. *arXiv* **2000**, arXiv:hep-th/0011260v1. Available online: https://arxiv.org/abs/hep-th/0011260 (accessed on 4 September 2019).
116. 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. [CrossRef]
117. Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, 864–871. [CrossRef]
118. Kohn, W.; Sham, L.J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, 1133–1138. [CrossRef]
119. Wheeler, J. Information, Physics, quantum: The search for the links. In Proceedings of the 3rd International Symposium on Foundations of Quantum Mechanics, Tokyo, Japan, 28–31 August 1989; pp. 354–368.
120. Von Bahr, B. On the central limit theorem in Rk. *Ark. Mat.* **1967**, *7*, 61–69. [CrossRef]
121. Ebrahimi, N.; Soofi, E.; Soyer, R. Multivariate maximum entropy identification, transformation, and dependence. *J. Multivar. Anal.* **2008**, *99*, 1217–1231. [CrossRef]
122. Conrad, K. Probability distributions and maximum entropy. *Entropy* **2005**, *6*, 10.
123. Adami, C.; Cerf, N. Prolegomena to a non-equilibrium quantum statistical mechanics. *Chaos Solition Fract.* **1999**, *10*, 1637–1650.
124. Kapranov, M. Thermodynamics and the moment map. *arXiv* **2011**, arXiv:1108.3472. Available online: https://arxiv.org/pdf/1108.3472.pdf (accessed on 4 September 2019).
125. Erdos, P. On the distribution function of additive functions. *Ann. Math.* **1946**, *47*, 1–20. [CrossRef]
126. Aczel, J.; Daroczy, Z. *On Measures of Information and Their Characterizations*; Academic Press: Cambridge, MA, USA, 1975.
127. Lifshitz, E.M.; Landau, L.D. *Statistical Physics (Course of Theoretical Physics, Volume 5)*; Butterworth-Heinemann: Oxford, UK, 1969.
128. Han, T.S. Linear dependence structure of the entropy space. *Inf. Control* **1975**, *29*, 337–368. [CrossRef]
129. Bjorner, A. Continuous partition lattice. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6327–6329. [CrossRef]
130. Postnikov, A. Permutohedra, Associahedra, and Beyond. *Int. Math. Res. Not.* **2009**, *2009*, 1026–1106. [CrossRef]
131. Matus, F. Conditional probabilities and permutahedron. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*; Elsevier: Amsterdam, The Netherlands, 2003.
132. Yeung, R. Facets of entropy. In *Communications in Information and Systems*; International Press of Boston: Somerville, MA, USA, 2015; Volume 15, pp. 87–117.
133. Yeung, R. A framework for linear information inequalities. *IEEE Trans. Inf. Theory* **1997**, *43*, 1924–1934. [CrossRef]
134. Zang, Z.; Yeung, R.W. On Characterization of Entropy Function via Information Inequalities. *IEEE Trans. Inf. Theory* **1997**, *44*, 1440–1452. [CrossRef]
135. Matúš, F. Infinitely Many Information Inequalities. *ISIT* **2007**, 41–47.
136. Takacs, D. *Stochastic Processes Problems and Solutions*; John Wiley & Sons: Chichester, UK, 1960.
137. Bourbaki, N. *Theory of Sets-Elements of Mathematic*; Addison Wesley: Boston, MA, USA, 1968.
138. Brillouin, L. *Scientific Uncertainty, and Information*; Academic Press: Cambridge, MA, USA, 2014.
139. Griffiths, R. Consistent Histories and the Interpretation of Quantum Mechanics. *J. Stat. Phys.* **1984**, *35*, 219. [CrossRef]
140. Omnes, R. Logical reformulation of quantum mechanics I. Foundations. *J. Stat. Phys.* **1988**, *53*, 893–932. [CrossRef]
141. Gell-Mann, M.; Hartle, J. Quantum mechanics in the light of quantum cosmology. In *Complexity, Entropy, and the Physics of Information*; Zurek, W., Ed.; Addison-Wesley: Boston, MA, USA, 1990; pp. 425–458.
142. Lieb, E.H.; Yngvason, J. A Guide to Entropy and the Second Law of Thermodynamics. In *Statistical Mechanics*; Springer: Berlin/Heidelberger, Germany, 1998; Volume 45, pp. 571–581.
143. Feynman, R. Space-Time Approach to Non-Relativistic Quantum Mechanics. *Rev. Mod. Phys.* **1948**, *20*, 367–387. [CrossRef]

144. Merkh, T.; Montufar, G. Factorized Mutual Information Maximization. *arXiv* **2019**, arXiv:1906.05460. Available online: https://arxiv.org/abs/1906.05460 (accessed on 4 September 2019).
145. Weiss, P. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.* **1907**, *6*, 661–690. [CrossRef]
146. Parsegian, V. *Van der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists*; Cambridge University Press: Cambridge, UK, 2006.
147. 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. [CrossRef]
148. Hà, H.T.; Van Tuyl, A. Resolutions of square-free monomial ideals via facet ideals: A survey. *arXiv* **2006**, arXiv:math/0604301. Available online: https://arxiv.org/abs/math/0604301 (accessed on 4 September 2019).
149. Newman, M.E.J. Complex Systems: A Survey. *arXiv* **2011**, arXiv:1112.1440v1. Available online: http://arxiv.org/abs/1112.1440v1 (accessed on 4 September 2019).
150. Mezard, M.; Montanari, A. *Information, Physics, and Computation*; Oxford University Press: Oxford, UK, 2009.
151. Vannimenus, J.; Toulouse, G. Theory of the frustration effect. II. Ising spins on a square lattice. *J. Phys. Condens. Matter* **1977**, *10*, 115. [CrossRef]
152. Rovelli, C. Notes for a brief history of quantum gravity. *arXiv* **2008**, arXiv:gr-qc/0006061v3. Available online: https://arxiv.org/pdf/gr-qc/0006061.pdf (accessed on 4 September 2019).
153. Sorkin, R. Finitary Substitute for Continuous Topology. *Int. J. Theor. Phys.* **1991**, *30*, 923–947. [CrossRef]
154. Strong, S.P.; Van Steveninck, R.D.R.; Bialek, W.; Koberle, R. On the application of information theory to neural spike trains. In Proceedings of the Pacific Symposium on Biocomputing, Maui, HI, USA, 4–9 January 1998; pp. 621–632.
155. Niven, R. Non-asymptotic thermodynamic ensembles. *EPL* **2009**, *86*, 1–6. [CrossRef]
156. Niven, R. Combinatorial entropies and statistics. *Eur. Phys. J.* **2009**, *70*, 49–63. [CrossRef]
157. Niven, R. Exact Maxwell–Boltzmann, Bose–Einstein and Fermi–Dirac statistics. *Phys. Lett. A* **2005**, *342*, 286–293. [CrossRef]
158. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938. [CrossRef]
159. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. *Physica A* **2001**, *302*, 89–99. [CrossRef]
160. Tsallis, C. Entropic Nonextensivity: A possible measure of Complexity. *Chaos Solition Fract.* **2002**, *13*, 371–391. [CrossRef]
161. Ritort, F. Nonequilibrium fluctuations in small systems: From physics to biology. *Adv. Chem. Phys.* **2008**, *137*, 31–123.
162. Artin, M.; Grothendieck, A.; Verdier, J. *Theorie des Topos et Cohomologie Etale des Schemas—(SGA 4) Vol I, II, III*; Springer: Berlin/Heidelberger, Germany, 1972.
163. Kolmogorov, A.N. *Grundbegriffe der Wahrscheinlichkeitsrechnung*; Springer: Berlin/Heidelberger, Germany, 1933.
164. 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.* **2014**, *10*, e1003408. [CrossRef] [PubMed]
165. Vigneaux, J. Information theory with finite vector spaces. *IEEE Trans. Inf. Theory* **2019**, *99*, 1. [CrossRef]
166. Wada, T.; Suyari, H. The k-Generalizations of Stirling Approximation and Multinominal Coefficients. *Entropy* **2013**, *15*, 5144–5153. [CrossRef]
167. Cerf, N.; Adami, C. Negative entropy and information in quantum mechanic. *Phys. Rev. Lett.* **1997**, *79*, 5194. [CrossRef]
168. Cerf, N.; Adami, C. Entropic Bell Inequalities. *Phys. Rev. A* **1997**, *55*, 3371. [CrossRef]
169. Oudot, S. *Persistence Theory: From Quiver Representations to Data Analysis*; American Mathematical Society: Providence, RI, USA, 2015; Volume 209.
170. Schneider, G. Two visual systems. *Science* **1969**, *163*, 895–902. [CrossRef] [PubMed]
171. Goodale, M.; Milner, A. Separate visual pathways for perception and action. *Trends Neurosci.* **1992**, *15*, 20–25. [CrossRef]
172. Kelley, H. Gradient theory of optimal flight paths. *ARS* **1960**, *30*, 947–954. [CrossRef]

173. Dreyfus, S. The numerical solution of variational problems. *J. Math. Anal. Appl.* **1962**, *5*, 30–45. [CrossRef]
174. Rumelhart, D.; Hinton, G.; Williams, R. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
175. Montufar, G.; Morton, J. Discrete Restricted Boltzmann Machines. *J. Mach. Learn. Res.* **2015**, *21*, 653–672.
176. Amari, S. Neural learning in structured parameter spaces—Natural Riemannian gradient. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; pp. 127–133.
177. Martens, J. New Insights and Perspectives on the Natural Gradient Method. *arXiv* **2017**, arXiv:1412.1193. Available online: https://arxiv.org/abs/1412.1193 (accessed on 4 September 2019).
178. Bengtsson, I.; Zyczkowski, K. *Geometry of Quantum States: An Introduction to Quantum Entanglement*; Cambridge University Press: Cambridge, UK, 2006.
179. Pascanu, R.; Bengio, Y. Revisiting Natural Gradient for Deep Networks. *arXiv* **2014**, arXiv:1301.3584. Available online: https://arxiv.org/abs/1301.3584 (accessed on 4 September 2019).
180. Waddington, C.H. *The Strategy of the Genes*; Routledge: Abingdon-on-Thames, UK, 1957.
181. Teschendorff, A.; Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **2017**, *8*, 15599. [CrossRef]
182. 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**, *34*, 2077–2086. [CrossRef] [PubMed]
183. Thom, R. Stabilite struturelle et morphogenese. *Poetics* **1974**, *3*, 7–19. [CrossRef]
184. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105–117. [CrossRef]
185. Nadal, J.P.; Parga, N. Sensory coding: information maximization and redundancy reduction. In Proceedings of the Neuronal Information Processing, Cargèse, France, 30 June–12 July 1997; World Scientific: Singapore, 1999; Volume 7, pp. 164–171.
186. Bell, A.J.; Sejnowski, T. An information maximisation approach to blind separation and blind deconvolution. *Neural Comput.* **1995**, *7*, 1129–1159. [CrossRef] [PubMed]
187. Chen, N.; Glazier, J.; Izaguirre, J.; Alber, M. A parallel implementation of the Cellular Potts Model for simulation of cell-based morphogenesis. *Comput. Phys. Commun.* **2007**, *176*, 670–681. [CrossRef]
188. Galvan, A. Neural plasticity of development and learning. *Hum. Brain Mapp.* **2010**, *31*, 879–890. [CrossRef]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).