

Surveying the Space of Descriptions of a Composite System with Machine Learning

Kieran A. Murphy¹, Yujing Zhang², and Dani S. Bassett^{1,3,4,5,6,7,8}

¹*Department of Bioengineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

²*School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907, USA*

³*Department of Electrical & Systems Engineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

⁴*Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

⁵*Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

⁶*Department of Physics & Astronomy, College of Arts & Sciences, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

⁷*The Santa Fe Institute, Santa Fe, New Mexico 87501, USA*

⁸*Montreal Neurological Institute, McGill University, Montreal, Quebec H3A 2B4, Canada*

 (Received 7 December 2024; revised 22 April 2025; accepted 29 May 2025; published 27 June 2025)

Multivariate information theory provides a general and principled framework for understanding how the components of a system are connected. Existing analyses are coarse in nature—built up from characterizations of discrete subsystems—and can be computationally prohibitive. In this work, we propose to study the continuous space of possible descriptions of a composite system as a window into its organizational structure. A description consists of specific information conveyed about each of the components, and the space of possible descriptions is equivalent to the space of lossy compression schemes of the components. We introduce a machine-learning framework to optimize descriptions that extremize key information theoretic quantities used to characterize organization, such as total correlation and O-information. Through case studies on spin systems, sudoku boards, and letter sequences from natural language, we identify extremal descriptions that reveal how system-wide variation emerges from individual components. By integrating machine learning into a fine-grained information theoretic analysis of composite random variables, our framework opens a new avenues for probing the structure of real-world complex systems.

DOI: [10.1103/gxrh-2xsv](https://doi.org/10.1103/gxrh-2xsv)

Multivariate information theory has emerged as a powerful lens for the understanding of complex systems, offering tools to uncover structure in the variation of multiple interacting components. From broad explorations of the nature of complexity [1–6] to detailed investigations of specific systems such as the brain [7–11], collective behavior in nature [12–15], gene regulatory networks [16], and toy models from condensed matter physics [17,18], information-theoretic approaches characterize organizational structure and reveal hidden interdependencies. These methods bridge disciplines, providing a domain-agnostic framework for quantifying how components interact to give rise to system-wide phenomena.

Prior work has predominantly focused on analyses of discrete subsystems in relation to the whole system, meaning that the variation of each component in the system is considered in its entirety. Here we propose to study the

space of partial entropy allocations, which we call “descriptions,” as a route to characterizing the interrelationships of a composite system. A description conveys partial information about each component and can be characterized by any of the commonly used summary quantities, such as total correlation [19] and O-information [5,20]. By formulating descriptions as a collection of communication channels, one per component [Fig. 1(a)], we can optimize descriptions with neural networks that maximize or minimize various information theoretic quantities.

We are interested in the space of possible descriptions for multiple reasons. First, we posit that the space of descriptions is relevant to the way that humans view complex systems: due to limited processing capacity [21], we focus on specific variation and ignore the rest, making the space of partial entropy allocations a natural object of study. Consequently, the range of possible descriptions might reasonably be related to the perception of complexity [1]. Second, extremal descriptions shed light on the system’s interrelationships by sifting noteworthy connections out of an abundance of variation [18]. For example, the description with maximal total correlation [19] reveals the specific variation among components that is most connected.

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

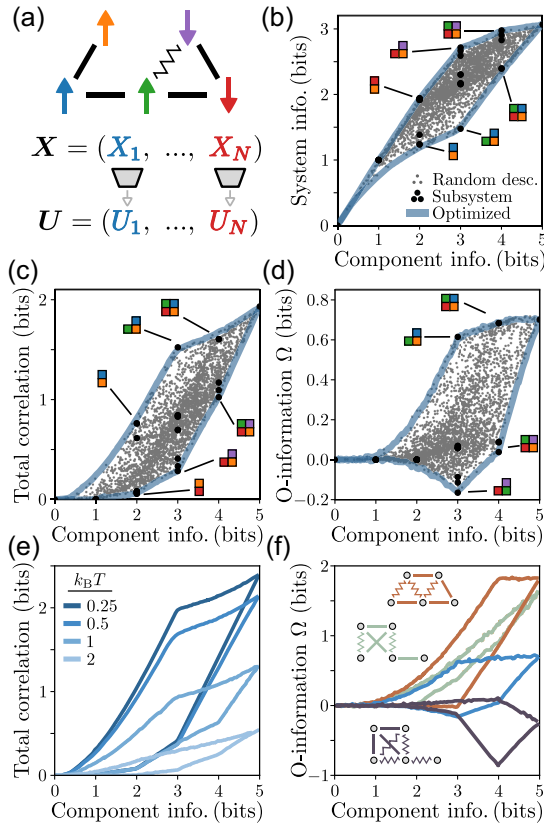


FIG. 1. Descriptions of a system. (a) The state $\mathbf{X}$ of a system of five interacting spins with ferromagnetic and antiferromagnetic couplings (straight and zigzag connectors, respectively) can be described by communicating information about each spin X_i . (b) The space of descriptions charted in terms of the total component information, $\sum_i I(X_i; U_i)$, and the system information, $I(\mathbf{X}; U)$. Possible descriptions include the accounting of discrete subsystems—i.e., subsets of components (black circles)—as well as a continuum of compression schemes for each component, which we randomly sample (gray dots) and optimize over (blue trace, standard error visualized). The space of possible descriptions can also be characterized by the total correlation (c), O-information (d), or other quantities from multivariate information theory. (e) Description space in terms of total correlation for the system in (a) at different temperatures. (f) Description space in O-information for various five-spin systems at $k_B T = 0.625$, with the blue trace the system in (a).

Finally, as we will show, the space of descriptions can be navigated with machine learning, offering practicality for real-world data and the potential to scale with continued advances in machine learning. By contrast, practicality is a significant issue for a popular framework for analyzing multivariate information content, partial information/entropy decomposition (PID/PED) [22,23]. The number of PID/PED terms to evaluate grows superexponentially with system size, rendering it impractical for more than around five components [24]. The terms generally require exhaustive calculation, though we note a recent exception that proposes to optimize a subset of terms with machine

learning [25]. With continuing advances in machine-learning methods to compress data for revealing important variation [26–30], there are new opportunities to study the structure of systems through the space of their descriptions.

Consider a system of N components whose states are represented with random variables X_i [Fig. 1(a)]. The full state of the system is represented with the random vector, $\mathbf{X} = (X_1, \dots, X_N)$. Let $\mathbf{U} = (U_1, \dots, U_N)$ be a *description* of the system state that conveys information about each component. Each U_i is generated from X_i via a probabilistic transformation, $U_i = f_i(X_i, \epsilon_i)$, where ϵ_i is an independent noise variable that introduces stochasticity into the transformation but carries no information about any component of $\mathbf{X}$. As a result, U_i is conditionally independent from all $X_{j \neq i}$ given X_i .

A description $\mathbf{U}$ is equivalent to a selection of entropy from each component. The mutual information $I(X_i; U_i)$ is the amount of entropy from X_i contained in U_i : $I(X_i; U_i) = H(X_i) - H(X_i|U_i)$, with $H(X_i)$ the Shannon entropy [31]. We can view the pieces of information in the description as a new composite system derived from measurements of the original one, and then characterize the new system’s information theoretic properties.

Among quantities that characterize the structure of entropy in a composite random variable, total correlation [19] measures the reduction in entropy when components are considered jointly versus independently, $\text{TC}(\mathbf{X}) = \sum_i H(X_i) - H(\mathbf{X})$. Because of the conditional independence of each U_i given X_i , we have $H(\mathbf{U}_S|\mathbf{X}_S) = \sum_{i \in S} H(U_i|X_i)$ for any subset of components S and can therefore evaluate the total correlation of the selected entropy in $\mathbf{U}$ in terms of transmitted information,

$$\text{TC}(\mathbf{U}) = \sum_i^N I(X_i; U_i) - I(\mathbf{X}; \mathbf{U}). \quad (1)$$

Another quantity, O-information Ω [5,20], characterizes the interactions between components as dominated by redundancy ($\Omega > 0$) or synergy ($\Omega < 0$). Redundant information is available from individual components, whereas synergistic information emerges only from their combinations [9]. For a description $\mathbf{U}$ of a system $\mathbf{X}$, the O-information is equal to

$$\Omega(\mathbf{U}) = (N - 2)I(\mathbf{U}; \mathbf{X}) + \sum_i^N [I(U_i; X_i) - I(\mathbf{U}_{/i}; X_{/i})], \quad (2)$$

with the notation $\mathbf{U}_{/i}$ indicating the set that excludes index i , $\{U_j; j \neq i\}$. We seek to extremize these and other quantities that are composed entirely of entropy measurements of subsets of components.

To find descriptions that extremize a summary quantity like total correlation, we devised a deep learning setup based on constrained communication. Information about a component X_i was transmitted to a representation U_i by encoding each outcome x_i to a probability distribution in latent space, $p(u_i|x_i)$. The sum total information about all components $\sum_i I(U_i; X_i) := I_{\text{in}}$ was set to a desired value, $\hat{I}_{\text{in}}$, through a loss proportional to $|I_{\text{in}} - \hat{I}_{\text{in}}|$ [32]. We estimated $I(U_i; X_i)$ with a lower bound based on likelihood ratios computed from a batch of data [33], and then sampled latent points for the description components $u_i \sim p(u_i|x_i)$ in the remaining loss calculations.

The remaining mutual information terms include subsets with multiple components and were optimized with InfoNCE [34], a variant of noise contrastive estimation (NCE) common in representation learning [35]. InfoNCE approximates the mutual information between two variables by contrasting positive and negative pairs of their outcomes. This process employs two neural networks: one to encode the first variable and another to encode the second, mapping them into a shared representation space. The InfoNCE loss for a batch of B samples, indexed by superscripts α and β , is given by

$$\mathcal{L}_{\text{NCE}}(U; X) = -\frac{1}{B} \sum_{\alpha=1}^B \log \frac{\exp[s(u^\alpha, x^\alpha)]}{\sum_{\beta=1}^B \exp[s(u^\alpha, x^\beta)]}, \quad (3)$$

where $s(u^\alpha, x^\beta)$ measures the similarity between the representations of u^α and x^β in the shared space, taken to be the squared Euclidean distance in this work. Positive pairs (u^α, x^α) are contrasted against negative pairs (u^α, x^β) sampled within the batch.

To summarize, the training loss combines (i) a sum total information constraint on the pieces of the description with (ii) any remaining terms in the summary quantity that we wish to extremize. For example, to minimize total correlation would require minimizing $\mathcal{L} = \gamma|I_{\text{in}} - \hat{I}_{\text{in}}| + \mathcal{L}_{\text{NCE}}(U; X)$. We note that this formulation closely resembles a distributed information bottleneck (IB) scenario with the joint variable X serving as the relevance variable [23,29].

To maximize total correlation requires minimizing $I(U; X)$, and InfoNCE, as a lower bound, cannot be directly used. For any such minimization term, we employed an adversarial setup where auxiliary networks were trained to maximize mutual information via InfoNCE, and then the loss was negated and applied to the description encodings.

For evaluation, mutual information terms were estimated via Monte Carlo sampling using likelihood ratios. Standard error is reported and lies within plot markers for all results presented in this work.

We examined the space of descriptions for three systems. A spin system with $N = 5$ sites has ferromagnetic and antiferromagnetic couplings [Fig. 1(a)]. The probability distribution over states $\mathbf{x} = (x_1, \dots, x_N)$ is given by

$$p(\mathbf{x}) = \frac{1}{Z} \exp[-\mathcal{H}(\mathbf{x})/k_B T]. \quad (4)$$

Z is a normalization constant called the partition function and $\mathcal{H}(\mathbf{x}) = -\sum_{\langle ij \rangle} J_{ij} x_i x_j$ is the energy of the state $\mathbf{x}$. $J_{ij} = 1, -1$ for the ferromagnetic and antiferromagnetic couplings, and 0 otherwise; we set $k_B T = 0.625$.

In Figs. 1(b)–1(d), we surveyed the space of descriptions for the 5-spin Ising system in several ways. First, descriptions were randomly sampled by creating a binary symmetric channel [31] for each spin with random noise (gray dots). Second, we formed descriptions corresponding to complete information about each possible discrete subsystem (black circles). Third, we probed the boundaries by extremizing total correlation or O-information (light blue curve). The descriptions that extremize total correlation also extremize system information [Fig. 1(b)]. The method successfully found descriptions that closely trace the bounds of the randomly sampled descriptions, which can be densely sampled for this small system.

We interpret the space of descriptions *globally* through its overall shape when charted in terms of different quantities, and *locally* through specific extremal descriptions. From a global perspective, the space of descriptions can tell of the nature of interactions between components across different levels of information. The spin system in Fig. 1(a) has three-bit descriptions that are either redundant or synergistic, even though the full state description is redundant [Fig. 1(d)]. Intuitively, the range of descriptions narrows and drops in total correlation as the temperature of the system grows [Fig. 1(e)]. With different connections between the spins, description space boundaries can vary dramatically and highlight qualitatively distinct multivariate relationships [Fig. 1(f)].

From a local perspective, the extremal descriptions reveal subsystems of interest. Among three-bit descriptions, the maximal total correlation for the five spin system in Fig. 1(a) links the ferromagnetic chain (orange, blue, green) [Fig. 1(c)]. The minimal total correlation connects the most distant spins (orange and red), then incorporates a second spin (purple) from the frustrated triangle—i.e., three spins connected by couplings for which there must be at least one inconsistent edge. In terms of O-information, maximal redundancy lies in the ferromagnetic chain while maximal synergy resides in the frustrated triangle ([Fig. 1(d)]. For the case of binary variables, broadening the space of descriptions from discrete subsystems to include partial information merely fills in the space between subsystem descriptions and reveals nothing new, though it facilitates optimization.

The second system is a 4×4 sudoku board (Fig. 2), where every square contains a digit from one to four and no digit can be repeated in a row, column, or quadrant. Sudoku is representative of constraint satisfaction problems, which are rich in structure and central to a wide variety of practical

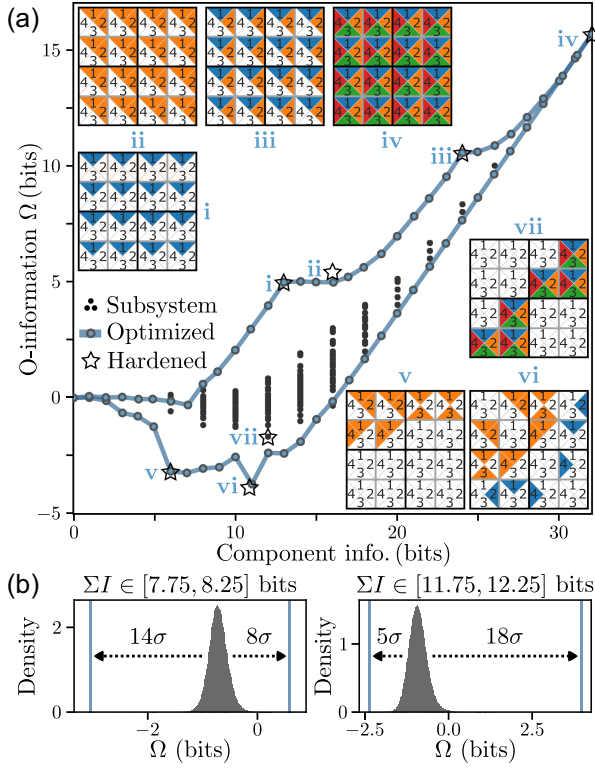


FIG. 2. The space of descriptions of a 4×4 sudoku board. (a) Discrete subsets of squares (black circles) and machine-learning-optimized boundaries (blue circles), in terms of O-information. Optimized soft compression schemes are converted to hard compression schemes (black stars) and visualized according to the corresponding Roman numerals. The hard compression scheme for each square in a board is displayed by coloring numbers according to groupings. For example, if one number in a square is blue and the rest are white, the blue number is distinguishable from the remaining three, and the three are indistinguishable from each other. (b) We randomly sampled 10^6 hard descriptions within the information range at the top of each plot. The optimized descriptions have O-information values (blue vertical lines) far from the distribution of randomly sampled schemes (gray).

and theoretical challenges, from scheduling to combinatorial optimization [36–38]. We took the probability of states $p(\mathbf{x})$ to be uniform across valid boards and zero otherwise. The dense constraints severely restrict the number of valid boards from 4^{16} ($\approx 10^9$) to 288, suggesting an intricate organizational structure linking the states of the squares. For this and remaining analyses, we found it necessary to run a different optimization per value of $\hat{l}_{\text{in}}$, in contrast to the spin system where a single optimization spanned the full range of description information $\hat{l}_{\text{in}}$. For each optimization, the coefficient of the transmitted information loss, γ , was constant for the first half of training, and then increased exponentially over the second half.

We focused on O-information to highlight modes of information sharing [Fig. 2(a)]; analysis using total

correlation, more revealing of component (in)dependence, is in the Supplemental Material [39]. All descriptions become highly redundant after around 16 bits, reflective of the dense constraints between squares. The discrete subsystems of a board state have minimal O-information (maximal synergy) at six squares and include the distantly connected triplets of squares shown in scheme vii.

In contrast to the spin system, the machine-learning-optimized boundaries for the sudoku board reach beyond the descriptions of discrete subsystems. The optimized compression schemes are soft, meaning that the information conveyed about each square resides in the layout of distributions $p(u_i|x_i)$ in latent space and communicates partial distinguishability between possible outcomes. For ease of interpretation, we converted each square’s compression scheme to a hard clustering of outcomes. After optimization, gradient descent was performed directly on the compression scheme to drive the statistical distinguishability between probabilistic representations to perfect distinguishability and indistinguishability, as measured by the Bhattacharyya coefficient [28,40,41].

The hardened compression schemes show which digits for each square were clustered together [having the same color in the insets to Fig. 2(a)], and are intuitive for the maximal O-information (redundant) descriptions. The same partial information is communicated about every square, from 0.8 bits per square to distinguishing one number from the other three (descr. i) to distinguishing all four digits (descr. iv). The descriptions with minimal O-information (synergistic) (v–vii) are less comprehensible. Description v communicates one bit three different ways, spread out across a row and a quadrant. The most synergistic description found for any level of communicated information, vi, is an intricate pattern of partial information, with a symmetry across the diagonal that was also present in the subsystem with minimal O-information (descr. vii).

The space of hard compression schemes of a discrete random variable with m outcomes is equivalent to the partitions of a set of m elements. There are 15 possible hard compression schemes for a variable with four outcomes, and 16 squares, making $15^{16} \approx 10^{18}$ different hard descriptions of a 4×4 sudoku board. Without accounting for symmetries, a manual search through these descriptions is impractical; by contrast, the machine-learning approach identified extremal descriptions on the order of a few minutes per optimization. To ground the performance of the extremized descriptions, we employed rejection sampling to obtain 10^6 random hard descriptions for two ranges of component information [Fig. 2(b)]. The extremized descriptions are several standard deviations away from the mean O-information of the sampled descriptions. Importantly, while there is no guarantee of global optimality, extremized descriptions bound the optimum, can be improved through hyperparameter tuning or repeated runs, and nevertheless highlight notable entropy allocations.

Finally, we analyzed the statistics of 4-grams in the English language based on data from Wikipedia. Letters are combined with numerous soft constraints that facilitate learning and error correction and have long been used as an object of study in information theory [42] and statistical mechanics [43]. We surveyed the space of descriptions of 4-grams that are themselves 4-letter words and the first or second half of 8-letter words. From a global perspective, total correlation varied less across the different 4-grams than did O-information [Figs. 3(a) and 3(b)]. The space of descriptions is almost entirely synergistic for the first four letters of 8-letter words, whereas 4-letter words contain variation that is more redundant. A typical O-information assessment, without the notion of a description and corresponding to full information about each letter, would also be negative for the former and positive for the latter (rightmost points). The most negative O-information occurs around 10–12 bits of component information—well beyond the full-information value. Discrete subsystems, shown for the 4-letter words in Figs. 3(a) and 3(b)

(squares), are far too coarse to capture the shape of the space of descriptions.

We hardened optimized descriptions [stars in Fig. 3(a)] and focused on maximal total correlation here due to its relative interpretability (other hardened descriptions are given in the Supplemental Material [39]). After hardening, each 4-gram cluster corresponds to a particular code u and contributes $\Delta TC(u) := p(u)tc(u)$, where $tc(u)$ is the local (pointwise) total correlation [44] and $TC(U) = \sum_u \Delta TC(u)$ with $u \in \mathcal{U}$ the set of possible codes. Evidently, to maximize total correlation, it is important to group 4-letter words starting with **th**, and the second half of 8-letter words that end in **ed** and **ing**. These schemes provide a starting point for a deeper linguistic analysis, serving as a sieve on the space of groupings of letters.

In this work, we introduced a machine-learning framework to study the space of partial descriptions of composite systems—a space too vast to adequately explore, even by random sampling, for all but the simplest cases. Crucially, each learned mapping from X to U defines a valid description of the system, so the method does not depend on convergence in the conventional sense: even suboptimal solutions yield interpretable and potentially informative points in the space of descriptions. Additionally, though our focus was on discrete variables, the approach extends naturally to continuous ones. Finally, the framework is flexible enough to extremize a broad class of quantities, such as the binding entropy [20], S-information [5,9,20] and ΔI [45], Tononi-Sporns-Edelman complexity [1], and specific atoms of PED [23], with each offering a different view of how entropy is distributed across components. Altogether, the framework opens new avenues for exploring the structural underpinnings of complex systems, offering both theoretical insights and practical tools for uncovering a scaffold of interconnected variation across components.

Acknowledgments—This work was supported by the Army Research Office under Grant No. W911NF-24-1-0228 and by the Center for Curiosity.

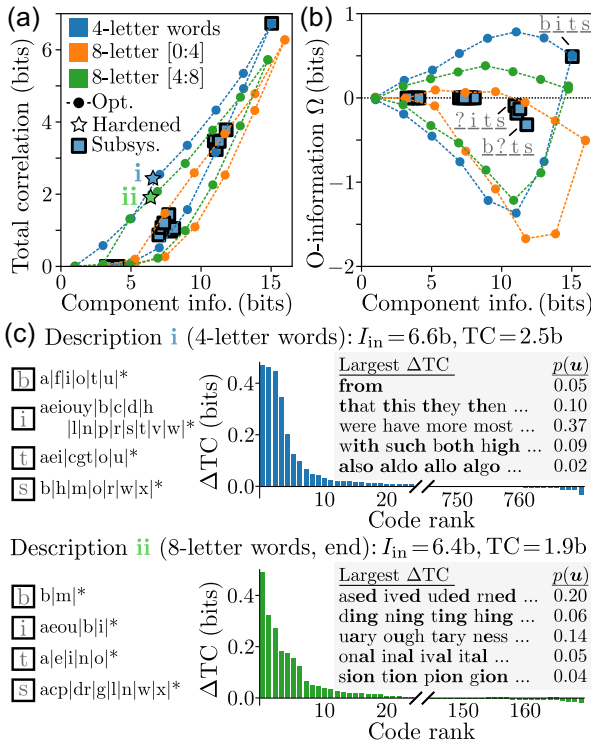


FIG. 3. Statistical structure in 4-letter sequences. The space of descriptions for 4-grams taken from 4- and 8-letter words, plotted in terms of (a) total correlation and (b) O-information. (c) The hardened descriptions for the maximal total correlation points marked as stars in (a), where groupings of letters are separated by vertical bars, and the group with all remaining letters of the alphabet is represented by an asterisk (*). The top contributions ΔTC to total correlation are shown at right, with the most probable n-grams inside each grouping shown. Letters are bolded to highlight recognizable letter patterns central to each grouping.

- [1] G. Tononi, O. Sporns, and G. M. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5033 (1994).
- [2] G. Nicolis and C. Nicolis, *Foundations of Complex Systems: Emergence, Information and Prediction* (World Scientific, Singapore, 2012).
- [3] J. Ladyman, J. Lambert, and K. Wiesner, What is a complex system?, *Eur. J. Philos. Sci.* **3**, 33 (2013).
- [4] B. C. Daniels, C. J. Ellison, D. C. Krakauer, and J. C. Flack, Quantifying collectivity, *Curr. Opin. Neurobiol.* **37**, 106 (2016).
- [5] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, Quantifying high-order interdependencies via multivariate

- extensions of the mutual information, *Phys. Rev. E* **100**, 032305 (2019).
- [6] D. A. Ehrlich, A. C. Schneider, V. Priesemann, M. Wibral, and A. Makkeh, A measure of the complexity of neural representations based on partial information decomposition, *Trans. Mach. Learn. Res.* (2023), <https://openreview.net/forum?id=R8TU3pfzFr>.
- [7] L. Martignon, G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia, Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies, *Neural Comput.* **12**, 2621 (2000).
- [8] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, Quantifying information modification in developing neural networks via partial information decomposition, *Entropy* **19**, 494 (2017).
- [9] T. F. Varley, M. Pope, M. Grazia, Joshua, and O. Sporns, Partial entropy decomposition reveals higher-order information structures in human brain activity, *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2300888120 (2023).
- [10] A. I. Luppi, F. E. Rosas, P. A. Mediano, D. K. Menon, and E. A. Stamatakis, Information decomposition and the informational architecture of the brain, *Trends Cognit. Sci.* **28**, 352 (2024).
- [11] M. Pope, T. F. Varley, and O. Sporns, Time-varying synergy/redundancy dominance in the human cerebral cortex, *J. Phys.* **6**, 015015 (2025).
- [12] J. M. Miller, X. R. Wang, J. T. Lizier, M. Prokopenko, and L. F. Rossi, Measuring information dynamics in swarms, in *Guided Self-Organization: Inception* (Springer, New York, 2014), pp. 343–364.
- [13] K. Pilkiewicz, B. Lemasson, M. Rowland, A. Hein, J. Sun, A. Berdahl, M. Mayo, J. Moehlis, M. Porfiri, E. Fernández-Juricic *et al.*, Decoding collective communications using information theory tools, *J. R. Soc. Interface* **17**, 20190563 (2020).
- [14] C. R. Twomey, A. T. Hartnett, M. M. Sosna, and P. Romanczuk, Searching for structure in collective systems, *Theor. Biosci.* **140**, 361 (2021).
- [15] A. L. Burns, T. M. Schaerf, J. Lizier, S. Kawaguchi, M. Cox, R. King, J. Krause, and A. J. Ward, Self-organization and information transfer in Antarctic krill swarms, *Proc. R. Soc. B* **289**, 20212361 (2022).
- [16] T. E. Chan, M. P. Stumpf, and A. C. Babbie, Gene regulatory network inference from single-cell data using multivariate information measures, *Cell syst.* **5**, 251 (2017).
- [17] S. Sootla, D. O. Theis, and R. Vicente, Analyzing information distribution in complex systems, *Entropy* **19**, 636 (2017).
- [18] T. Scagliarini, D. Nuzzi, Y. Antonacci, L. Faes, F. E. Rosas, D. Marinazzo, and S. Stramaglia, Gradients of o-information: Low-order descriptors of high-order dependencies, *Phys. Rev. Res.* **5**, 013025 (2023).
- [19] S. Watanabe, Information theoretical analysis of multivariate correlation, *IBM J. Res. Dev.* **4**, 66 (1960).
- [20] R. G. James, C. J. Ellison, and J. P. Crutchfield, Anatomy of a bit: Information in a time series observation, *Chaos* **21**, 037109 (2011).
- [21] R. Marois and J. Ivanoff, Capacity limits of information processing in the brain, *Trends Cognit. Sci.* **9**, 296 (2005).
- [22] P. L. Williams and R. D. Beer, Nonnegative decomposition of multivariate information, [arXiv:1004.2515](https://arxiv.org/abs/1004.2515).
- [23] R. A. Ince, The partial entropy decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal, [arXiv:1702.01591](https://arxiv.org/abs/1702.01591).
- [24] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, Synergy, redundancy, and multivariate information measures: An experimentalist's perspective, *J. Comput. Neurosci.* **36**, 119 (2014).
- [25] A. Kolchinsky, Partial information decomposition: Redundancy as information bottleneck, *Entropy* **26**, 546 (2024).
- [26] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, Deep variational information bottleneck, in *International Conference on Learning Representations* (2017), <https://openreview.net/forum?id=HyxQzBceg>.
- [27] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, *Nat. Phys.* **14**, 578 (2018).
- [28] K. A. Murphy and D. S. Bassett, Interpretability with full complexity by constraining feature information, in *International Conference on Learning Representations (ICLR)* (2023), https://openreview.net/forum?id=R_OL5mLhsv.
- [29] K. A. Murphy and D. S. Bassett, Information decomposition in complex systems via machine learning, *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2312988121 (2024).
- [30] K. A. Murphy and D. S. Bassett, Machine-learning optimized measurements of chaotic dynamical systems via the information bottleneck, *Phys. Rev. Lett.* **132**, 197201 (2024).
- [31] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).
- [32] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, Understanding disentangling in β -VAE, [arXiv:1804.03599](https://arxiv.org/abs/1804.03599).
- [33] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, On variational bounds of mutual information, in *International Conference on Machine Learning* (PMLR, 2019), pp. 5171–5180.
- [34] A. v. d. Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, [arXiv:1807.03748](https://arxiv.org/abs/1807.03748).
- [35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *International Conference on Machine Learning* (PMLR, 2020), pp. 1597–1607.
- [36] C. P. Gomes, B. Selman, N. Crato, and H. Kautz, Heavy-tailed phenomena in satisfiability and constraint satisfaction problems, *J. Autom. Reasoning* **24**, 67 (2000).
- [37] M. Ercsey-Ravasz and Z. Toroczkai, The chaos within sudoku, *Sci. Rep.* **2**, 725 (2012).
- [38] M. Varga, R. Sumi, Z. Toroczkai, and M. Ercsey-Ravasz, Order-to-chaos transition in the hardness of random Boolean satisfiability problems, *Phys. Rev. E* **93**, 052211 (2016).
- [39] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/gxrh-2xsv> for implementation specifics and additional results for the sudoku and 4-gram systems.
- [40] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans. Commun. Tech.* **15**, 52 (1967).

- [41] K. A. Murphy, S. Dillavou, and D. S. Bassett, Comparing the information content of probabilistic representation spaces, *Trans. Mach. Learn. Res.* (2025), <https://openreview.net/forum?id=adhsMqURI1>.
- [42] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 379 (1948).
- [43] G. J. Stephens and W. Bialek, Statistical mechanics of letters in words, *Phys. Rev. E* **81**, 066119 (2010).
- [44] T. Scagliarini, D. Marinazzo, Y. Guo, S. Stramaglia, and F. E. Rosas, Quantifying high-order interdependencies on individual patterns via the local o-information: Theory and applications to music analysis, *Phys. Rev. Res.* **4**, 013184 (2022).
- [45] S. Nirenberg, S. M. Carcieri, A. L. Jacobs, and P. E. Latham, Retinal ganglion cells act largely as independent encoders, *Nature (London)* **411**, 698 (2001).