Correlation Lengths in the Language of Computable Information

Stefano Martiniani,[1,2,*] Yuval Lemberg,[3] Paul M. Chaikin,[2,†] and Dov Levine[3,‡]

[1]*Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota 55455, USA*
[2]*Center for Soft Matter Research, Department of Physics, New York University, New York 10003, USA*
[3]*Department of Physics, Technion—IIT, 32000 Haifa, Israel*

(Received 20 June 2020; accepted 17 August 2020; published 22 October 2020)

Computable information density (CID), the ratio of the length of a losslessly compressed data file to that of the uncompressed file, is a measure of order and correlation in both equilibrium and nonequilibrium systems. Here we show that correlation lengths can be obtained by decimation, thinning a configuration by sampling data at increasing intervals and recalculating the CID. When the sampling interval is larger than the system's correlation length, the data becomes incompressible. The correlation length and its critical exponents are thus accessible with no *a priori* knowledge of an order parameter or even the nature of the ordering. The correlation length measured in this way agrees well with that computed from the decay of two-point correlation functions $g_2(r)$ when they exist. But the CID reveals the correlation length and its scaling even when $g_2(r)$ has no structure, as we demonstrate by "cloaking" the data with a Rudin-Shapiro sequence.

DOI: 10.1103/PhysRevLett.125.170601

Physics, and indeed science in general, is a search to find and quantify correlations and order in nature. In many cases this organization is evident and quantifiable in terms of an order parameter that is identified with a broken symmetry. Such symmetry breaking is often associated with a phase transition at which the order parameter becomes finite, and a length scale for the persistence of the order which diverges as one approaches the transition. There are, however, systems in nature whose order we do not yet understand or for which we cannot define an order parameter in the conventional sense. Even in such cases, we may reasonably expect that there exist some as yet unidentified correlations, with associated length scales which may or may not diverge.

The basic idea we wish to exploit is the intimate connection between order and information: it takes less information to completely describe a system with correlations than an uncorrelated one. The basis for the quantification of these ideas can be found in information theory [1], in particular the Shannon entropy [2] and the Kolmogorov complexity [3,4]. In recent work [5] we have introduced a quantitative measure, the computable information density (CID), $\mathcal{K} \equiv \mathcal{L}(\mathbf{x})/L$, that is the binary code length, $\mathcal{L}(\mathbf{x})$, of a losslessly compressed file $\mathbf{x}$ (such as the microstate of a many-body system) divided by the uncompressed length L (the number of degrees of freedom) of $\mathbf{x}$ [6], which is closely related to the Shannon and Kolmogorov measures, and which is an excellent approximant of the thermodynamic entropy, S, for equilibrium systems. In what follows we estimate the CID using the unrestricted Lempel-Ziv string matching algorithm (LZ77) [7,8], a *universal* (i.e., requires no *a priori* knowledge of the nature of the ensemble) and asymptotically *optimal* code (i.e., $\lim_{L\to\infty} \mathcal{K} = S$) [1,5]. CID reveals the nature of phase transitions (first or second order), the position of critical points, and the exponent of critical slowing down, for both equilibrium and nonequilibrium phase transitions. Here we wish to explore whether CID can be used to determine correlation lengths for such systems [9].

The standard method for computing the correlation length ξ of a system is to calculate some correlation function, typically two-point, and see how it decays with distance. This presupposes that the order and proper correlation function is known. In this Letter, we propose a method that does not require this knowledge, which is based on the fundamental idea that correlations reduce the CID of a system.

If a system consists of uncorrelated elements, the CID takes its maximum value. To exploit this, we sample a system on various length scales Δ by culling out degrees of freedom on smaller scales. In Figs. 1(a) and 1(b) we consider a 1D model of randomly placed hard rods of length $\ell = 4$, while in Figs. 1(c)–1(d) we have a 1D Ising model [10] at finite temperature and zero applied magnetic field. The diagrams in Figs. 1(a) and 1(c) show a respective configuration from each of the models, which we sample on every fourth site ($\Delta = 4$). If $\Delta < \xi$, the remaining degrees of freedom still show correlations, albeit weakened, but if $\Delta > \xi$, all correlations are lost and the CID attains its maximal value. In the simplest cases, e.g., for the 1D hard-rod model, to estimate ξ we can simply look for the smallest value of Δ where the CID reaches its maximum. However, this is not always adequate, e.g., in the 1D Ising model the CID approaches its maximum exponentially, so we find that

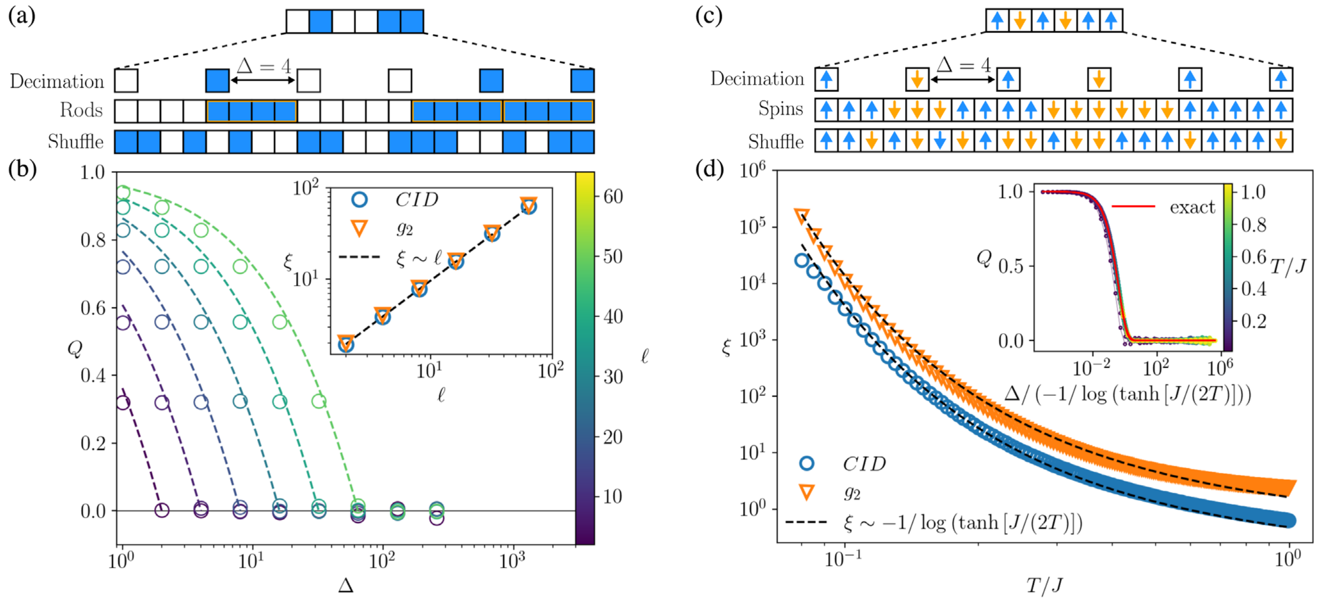

FIG. 1. Diagrams depict (a) 1D hard-rods of length $\ell = 4$ and (c) 1D Ising configurations, their shuffled counterparts, and the result of decimation when the sampling interval is $\Delta = 4$. (b) 1D hard rods, with lengths $\ell = 2^i$ $(1 \leq i \leq 6)$, randomly distributed on a grid of length $L = 2^{16+i}$, at fixed density $\rho \equiv N_r\ell/L = 1/4$. Upon decimation at intervals $\Delta > \ell$, the configurations reduce to a random sequence. The main panel shows $Q(\Delta, \ell)$, dashed lines are exact solution to leading order (see Eq. S5 in the Supplemental Material [11]). Inset shows ξ as computed from $Q(\Delta)$ and $g_2(r)$ taking ξ to be the value where $Q(\xi) = 0.025$ and where $g_2(\xi)$ is a its minimum (see the Supplemental Material [11]). (d) 1D Ising model of size $L = 2^{20}$ simulated by Wolff algorithm [24]. We performed the same analysis as for panel a but extracted ξ by fitting the curves to exponential functions of the form $g_2(r) = g_2(0)\exp(-r/\xi)$ and $Q(\Delta) = Q(1)\exp[-(\Delta - 1)/\xi]$. Data were averaged over 200 equilibrium configurations. The black dashed lines show the theoretical scaling for ξ with T. Inset shows the collapsed $Q(\Delta, T)$, along with the analytical solution Eq. (1) (red line).

in general it is better to study the way that the CID scales with Δ by collapsing the data. This procedure has the advantage of being independent of the system being analyzed.

In particular, we wish to study the quantity [25]

$$Q(\Delta) \equiv 1 - \frac{\mathcal{K}(\Delta)}{\mathcal{K}_{\text{shuf}}(\Delta)}, \tag{1}$$

where we denoted the CID as $\mathcal{K}$, and the subscript “shuf” refers to a configuration obtained by randomly shuffling all its degrees of freedom. Because a randomly shuffled configuration has no correlations, $\mathcal{K}_{\text{shuf}}(\Delta) \geq \mathcal{K}(\Delta)$, and $0 \leq Q \leq 1$.

The 1D hard-rod system consists of N_r rods, each occupying ℓ contiguous sites, randomly distributed on a lattice of length L sites. The fraction of occupied sites is $\rho = N_r\ell/L$. Configurations of the system are represented by strings $\{n_j\}$, where $j = 1, 2, \ldots, L$ and $n_j = 1$ if site j is occupied by a rod element, and $n_j = 0$ if it is not. The trivial correlation length is ℓ. Can we discover this by decimating configurations, computing their CID, and estimating ξ_{CID} from the value of Δ at which $Q(\Delta, \ell) \to 0$ for different values of ℓ?

The decimated configurations are obtained by retaining the occupancies $n_{j\cdot\Delta}$ (with $j = 1, 2, \ldots, L/\Delta$) of a configuration, deleting all the others, and then rescaling the system by a factor Δ. In Fig. 1(b) we plot $Q(\Delta)$ for several values of ℓ, with ξ_{CID} vs ℓ shown in the inset, along with the values of ξ_{g_2} obtained by finding the minimum of the two-point correlation function $g_2(r)$ of the undecimated configurations (see the Supplemental Material [11]). Both correlation lengths are close in value to ℓ but differ numerically by a small factor. The data collapse, along with the exact result, showing that $Q(\rho, \Delta, \ell) \to 0$ linearly with $\Delta - \ell$ as $\Delta \to \ell$, are given in the Supplemental Material [11].

We next consider the equilibrium 1D Ising model, which has a transition at $T = 0$. Both the entropy S and ξ may be solved for exactly [26], and give [27]

$$Q(\Delta, \xi) = \frac{e^{-\Delta/\xi}\coth^{-1}(e^{\Delta/\xi}) + \frac{1}{2}\log(1 - e^{-2\Delta/\xi})}{\log(2)}. \tag{2}$$

Here, $Q \to 0$ exponentially as $Q \sim e^{-2\Delta/\xi}$, making an extrapolation inadequate to determine ξ_{CID}. Rather, we generate equilibrium spin configurations for different temperatures, decimate these configurations, calculate the CID to find $Q(\Delta)$, and then collapse them to a universal curve [inset of Fig. 1(d)]. The collapse indicates that there is a single length scale ξ in the problem and yields its temperature dependence. An exponential fit to a single

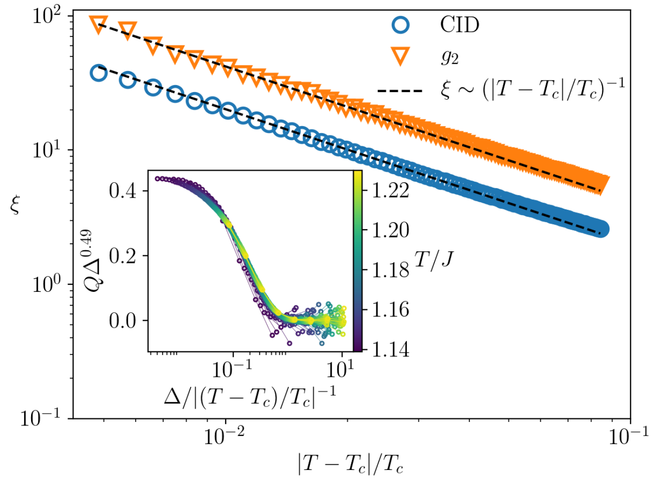

FIG. 2. 2D Ising model of size $L=2^{10}\times 2^{10}$ simulated by Wolff algorithm [24]. Correlation lengths ξ extracted by fitting the pair-correlation function to $g_2(r)=g_2(0)\exp(-r/\xi)/r^\eta$ and $Q(\Delta)=Q(1)\exp[-(\Delta-1)/\xi]/\Delta^\theta$. We find $\eta\approx 1/4$ and $\theta\approx 1/2$. The black dashed lines show the theoretical scaling $\xi\sim|T-T_c|/T_c$ with $T_c\approx 1.1345$. Data were averaged over 200 equilibrium configurations. Inset shows the collapse of the scaled $Q(\Delta,T)$.

curve Q vs Δ/ξ then gives the value of ξ_{CID}. $\xi_{\mathrm{CID}}(T)$ and $\xi_{g_2}(T)$ are shown in Fig. 1(d); both show the same T dependence as the analytic $\xi(T)$.

Results in 2D for the q-state Potts models ($2\le q\le 8$) [28,29] are shown in the Supplemental Material [11]. For $q=2$, this is the Ising model. Figure 2 (inset) shows the collapse of Q obtained by scaling the axes; this allows us to determine the critical exponent to be $\nu=1$, where $\xi(T)\propto(T-T_c)^{-\nu}$. Fitting $Q(\Delta,T)$ at a single temperature gives us the numerical value of ξ_{CID}, which is plotted alongside the value obtained from $g_2(r)$ in the main panel of Fig. 2. Notice that while decimating by Δ correctly yields configurations with correlation length ξ/Δ, these are not equilibrium configurations with the same correlation length. This can be seen for instance from the fact that magnetization is invariant under decimation. In the Supplemental Material [11], we consider an alternative blocking transformation, known as "majority rule" [30], that yields valid equilibrium configurations and for which we can derive an exact expression for 2D Ising analogous to Eq. (2), and verify that there is good agreement between theory and numerical results in 2D.

We now consider the conserved lattice gas (CLG), a dynamical nonequilibrium lattice model of the conserved directed percolation class [31]. In the CLG [as illustrated in Fig. 3(a)] an occupied site is considered "active" if one of the nearest neighbors is also occupied (orange circles). Sites have a maximum occupancy of 1 particle. At each time step, active sites are emptied stochastically by moving the particle to one of the empty neighboring sites (black arrows). The model has a continuous phase transition from a low density absorbing phase (where all sites are inactive) to a high density active phase where the dynamics persist forever. Configurations at the critical point are hyperuniform [32,33]. In 1D the critical density $\rho_c=1/2$ corresponds to a periodic arrangement where every other site is occupied (i.e., 101010, ...). In Fig. 3(b) we show $\xi(\rho)$ as obtained both from CID and $g_2(r)$, as well as the scaling collapse of $Q(\Delta,\rho)$ for different densities (inset). We find that $\xi(\rho)$ diverges with the exponent $\nu=2$ for both measures as $\rho\to\rho_c$.

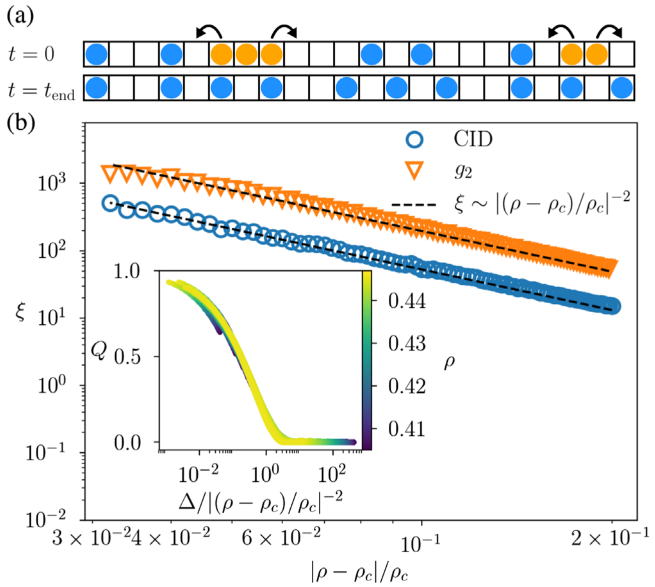

FIG. 3. 1D conserved lattice gas. (a) At time $t=0$ the system is in an active randomly sampled state (active sites in orange) and the possible moves prescribed by the dynamics are indicated by the arrows. When the density $\rho\le\rho_c$ the system relaxes to an absorbing state with no active sites. (b) Correlation lengths ξ for a system of size $L=2^{17}$ starting from randomly sampled states, extracted from $Q(\Delta)$ and $|g_2(r)|$ by taking ξ to be the value where $Q(\xi)=0.05$ and $|g_2(\xi)|=0.05$. The black dashed lines show the fitted scaling $\xi\sim|(\rho-\rho_c)/\rho_c|^{-2}$, where $\rho_c=0.5$. Inset shows the collapse of the scaled $Q(\Delta,\rho)$. Data were averaged over 15 independently generated configurations.

We now want to see whether CID decimation can measure correlation lengths in systems with no two-point correlations. To this end, we will "cloak" strings in two ways that destroy their two-point correlations. To do this, we multiply 1D Ising configurations by (i) a random Bernoulli sequence (RBS) of equal numbers of ±1, and (ii) the deterministic Rudin-Shapiro sequence [34,35] (RS). Notice that this cloaking is exactly equivalent to studying two variants of the "Mattis glass," with RS and RBS ground states [36]. Both of these sequences have $g_2(j,k)=\delta_{jk}$, but while for RS the Kolmogorov complexity and CID tend to zero as the sequence length increases, they are maximal for RBS [5].

Multiplying a sequence with structure in its $g_2(r)$ [or, equivalently, its structure factor $S(k)$] by RBS will produce

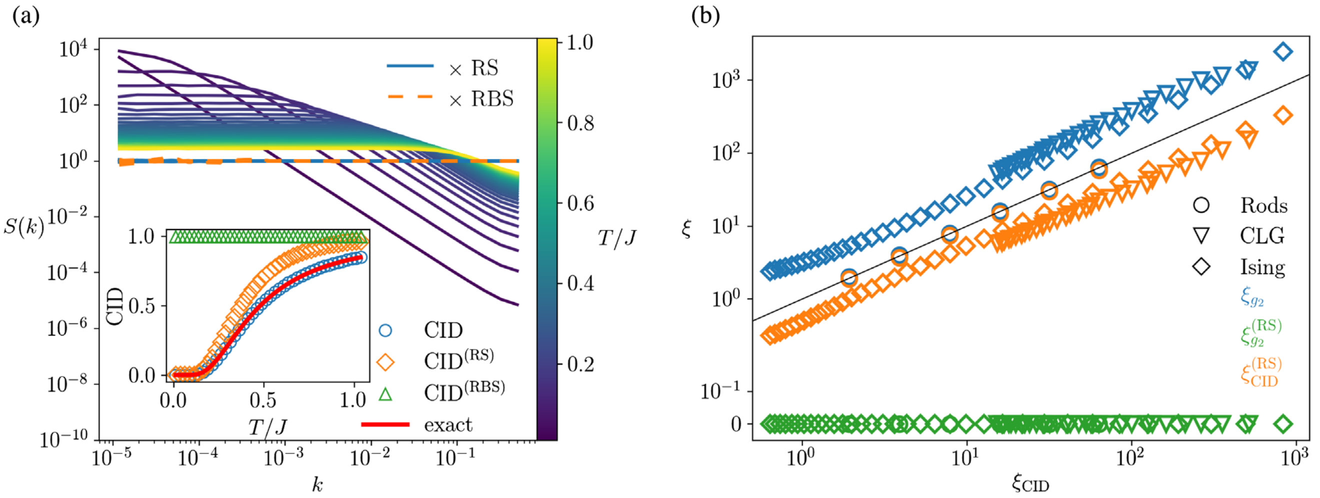

FIG. 4. (a) Structure factor $S(k)$ for the 1D Ising model of size $L = 2^{20}$. $S(k)$ is shown for the unaltered system, as well as for the Rudin-Shapiro, RS-cloaked, and Random Bernoulli sequence (RBS) randomly-cloaked sequences, neither of which show any structure, indicating that all pair correlations have been destroyed. Inset: the CID as a function of T for the uncloaked and cloaked systems: the RS-cloaked system has nontrivial CID, but the randomly cloaked system has maximum CID and is incompressible. Data were averaged over 200 equilibrium configurations. (b) Correlation lengths for the hard-rod, CLG, and 1D Ising systems, uncloaked [extracted from $g_2(r)$, in blue] and cloaked by RS [extracted from $g_2(r)$ and $Q(\Delta)$, in green and orange, respectively]. The data is plotted against $\xi_{\rm CID}$, as measured by CID decimation of the uncloaked systems, so that a slope of 1 (solid line) means identical scaling for the RS-cloaked and uncloaked systems.

a maximally random sequence with $g_2(r) = 0$ for $r \neq 0$ [and $S(k) = 1$ [37]]. Moreover, this will increase its CID [38], and cause all information about the original configuration to be lost (unless decoded by the identical random sequence). A similar multiplication by RS will remove all two-point correlations, also giving $g_2(r) = 0$ (for $r \neq 0$) and $S(k) = 1$, but will not appreciably change the Kolmogorov complexity of the original, since the RS itself has negligible Kolmogorov complexity. In this sense, “cloaking” by RS makes the sequence look random as far as $g_2(r)$ and $S(k)$ are concerned, while still retaining all the original information and order, although in a different form. We therefore expect that we should be able to recover the correlation length of the original, uncloaked system.

In Fig. 4(a) (inset) we show the CID for 1D Ising configurations at different temperatures, and for the same configurations cloaked by RS and by RBS. RBS-cloaked 1D Ising has a flat CID $= 1$ indicating a correlationless, maximally disordered system, but RS-cloaked 1D Ising retains much of its correlations. In the main panel of Fig. 4(a) we graph $S(k)$ for 1D Ising configurations, this shows increased correlations as T is lowered. Multiplying any configuration by RBS or RS gives $S(k) = 1$. We now perform the decimation procedure to determine the correlation length of the cloaked configurations. In Fig. 4(b) we plot the correlation lengths for the RS-cloaked configurations as determined from CID decimation and from $g_2(r)$, vs $\xi_{\rm CID}$ for the uncloaked 1D Ising configurations. For random cloaking, all information is lost, with no temperature dependence. However, for RS cloaking, although $\xi_{g_2}^{(RS)}(r) = 0$, $\xi_{\rm CID}^{(RS)}$ agrees well with the correlation length of the uncloaked 1D Ising system. Figure 4(b) shows that similar conclusions hold also for the hard rod and 1D CLG systems when cloaked by RS (see also the Supplemental Material [11]).

We note that RBS cloaking is analogous to the situation encountered in the analysis of static configurations of a square-lattice spin glass with random quenched disorder. In this case, without knowledge of the random couplings, the CID (or any other static estimator) would not be able to detect the order of an individual configuration because (like for RBS) individual configurations exhibit no correlations (and therefore are not compressible). It might, however, be possible to extract a correlation length by CID when considering the dynamics of the systems, viz. whole trajectories rather than individual configurations. We also argue that because structural glasses lack quenched disorder [39], the CID may be an effective tool for the analysis of the glass transition, e.g., in soft sphere systems, given a sufficiently accurate CID estimator for continuum two- and three-dimensional systems.

CID decimation presents a simple and general method for finding the correlation length of equilibrium and nonequilibrium systems, or in fact of any temporal or spatial array (e.g., a sequence or an image), with no *a priori* knowledge of a possible order parameter, as well as in systems where two-point correlations are uninformative. We expect that this technique may lead to the discovery of order and aid in the quantification of correlation lengths in a wealth of new systems.

We particularly thank Yariv Kafri for crucial discussions at many stages of this project. This work was primarily supported by the National Science Foundation Physics of Living Systems Grant No. 1504867. D. L. thanks the U.S.-Israel Binational Science Foundation (Grant No. 2014713), the Israel Science Foundation (Grant No. 1866/16). P. M. C. was supported partially by the Materials Research Science and Engineering Center (MRSEC) Program of the National Science Foundation under Award No. DMR-1420073.

*mart5523@umn.edu
†chaikin@nyu.edu
‡dovlevine19@gmail.com

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

[2] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[3] A. N. Kolmogorov, Int. J. Comput. Math. **2**, 157 (1968).

[4] G. J. Chaitin, J. ACM **13**, 547 (1966).

[5] S. Martiniani, P. M. Chaikin, and D. Levine, Phys. Rev. X **9**, 011031 (2019).

[6] Notice that the CID is not the same as the compression ratio (or compressibility) ϱ of the sequence, in fact $\mathrm{CID} = \varrho \log_2 |\alpha|$, where $|\alpha|$ is the dictionary size of the sequence [40].

[7] J. Ziv and A. Lempel, IEEE Trans. Inf. Theory **23**, 337 (1977).

[8] P. C. Shields, IEEE Trans. Inf. Theory **45**, 1283 (1999).

[9] The use of information to find what might be regarded as a correlation length for written English was the subject of Ref. [41] in the context of predicting letters in a string, where correlations of up to eight letters were inferred. A similar idea was discussed in Ref. [42] where patch entropy, calculated with different block sizes, was employed to discuss correlation lengths in physical systems.

[10] E. Ising, Z. Phys. **31**, 253 (1925).

[11] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.125.170601 for a definition of the models, exact solutions, supplementary data, and implementation details, which include Refs. [12–23].

[12] L. Onsager, Phys. Rev. **65**, 117 (1944).

[13] A. Codello, V. Drach, and A. Hietanen, J. Stat. Mech. (2015) P11008.

[14] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford University Press, Oxford, 1992).

[15] M. J. E. Golay, J. Opt. Soc. Am. **39**, 437 (1949).

[16] M. J. E. Golay, J. Opt. Soc. Am. **41**, 468 (1951).

[17] W. Rudin, Proc. Am. Math. Soc. **10**, 855 (1959).

[18] S. Constantinescu and L. Ilie, SIAM J. Discrete Math. **21**, 466 (2007).

[19] S. Martiniani, Sweetsourcod, https://github.com/smcantab/sweetsourcod (2018).

[20] J. Kärkkäinen, D. Kempa, and S. J. Puglisi, in *Annual Symposium on Combinatorial Pattern Matching* (Springer, New York, 2013), pp. 189–200.

[21] J. Kärkkäinen, D. Kempa, and S. J. Puglisi, Lz77 factorization algorithms, https://www.cs.helsinki.fi/group/pads/lz77.html (2013).

[22] J. Skilling, AIP Conf. Proc. **707**, 381 (2004).

[23] G. Altay, hilbert_curve, https://github.com/galtay/hilbert_curve (2015).

[24] U. Wolff, Phys. Rev. Lett. **62**, 361 (1989).

[25] In principle, Q depends on control parameters such as temperature or density, which are suppressed here; they will be indicated where relevant.

[26] S. Salinas, *Introduction to Statistical Physics* (Springer Science & Business Media, New York, 2001).

[27] Where we use the analytic value for the entropy in place of the CID in Eq. (1). See the Supplemental Material [11] for a full derivation.

[28] R. B. Potts, in *Mathematical Proceedings of the Cambridge Philosophical Society* (Cambridge University Press, Cambridge, England, 1952), Vol. 48, pp. 106–109.

[29] F.-Y. Wu, Rev. Mod. Phys. **54**, 235 (1982).

[30] L. P. Kadanoff, Phys. Phys. Fiz. **2**, 263 (1966).

[31] H. Hinrichsen, Adv. Phys. **49**, 815 (2000).

[32] D. Hexner and D. Levine, Phys. Rev. Lett. **114**, 110602 (2015).

[33] S. Torquato and F. H. Stillinger, Phys. Rev. E **68**, 041113 (2003).

[34] H. S. Shapiro, Extremal problems for polynomials and power series, Ph.D. thesis, Massachusetts Institute of Technology, 1952.

[35] J. P. Allouche and J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations* (Cambridge University Press, Cambridge, England, 2003).

[36] D. C. Mattis, Phys. Lett. **56A**, 421 (1976).

[37] There may also be a delta function at $q = 0$.

[38] Unless the sequence is itself already random, in which case there is no change.

[39] S. Karmakar and G. Parisi, Proc. Natl. Acad. Sci. U.S.A. **110**, 2752 (2013).

[40] J. Ziv and A. Lempel, IEEE Trans. Inf. Theory **24**, 530 (1978).

[41] C. E. Shannon, Bell Syst. Tech. J. **30**, 50 (1951).

[42] J. Kurchan and D. Levine, J. Phys. A **44**, 035001 (2011).