Topological Kolmogorov complexity and the Berezinskii-Kosterlitz-Thouless mechanism

Vittorio Vitale [1,2,3,*], Tiago Mendes-Santos,[4] Alex Rodriguez [1,5] and Marcello Dalmonte [1,2]

[1]*International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34151 Trieste, Italy*
[2]*SISSA, via Bonomea 265, 34136 Trieste, Italy*
[3]*Université Grenoble Alpes, CNRS, Laboratoire de Physique et Modélisation des Milieux Condensés (LPMMC), Grenoble 38000, France*
[4]*Theoretical Physics III, Center for Electronic Correlations and Magnetism, Institute of Physics,*
University of Augsburg, 86135 Augsburg, Germany
[5]*Dipartimento di Matematica e Geoscienze, Universitá degli Studi di Trieste, via Alfonso Valerio 12/1, 34127, Trieste, Italy*

Topology plays a fundamental role in our understanding of many-body physics, from vortices and solitons in classical field theory to phases and excitations in quantum matter. Topological phenomena are intimately connected to the distribution of information content that, differently from ordinary matter, is now governed by nonlocal degrees of freedom. However, a precise characterization of how topological effects govern the complexity of a many-body state, i.e., its partition function, is presently unclear. In this paper, we show how topology and complexity are directly intertwined concepts in the context of classical statistical mechanics. We concretely present a theory that shows how the *Kolmogorov complexity* of a classical partition function sampling carries unique, distinctive features depending on the presence of topological excitations in the system. We confront two-dimensional Ising, Heisenberg, and XY models on several topologies and study the corresponding samplings as high-dimensional manifolds in configuration space, quantifying their complexity via the intrinsic dimension. While for the Ising and Heisenberg models the intrinsic dimension is independent of the real-space topology, for the XY model it depends crucially on temperature: across the Berezkinskii-Kosterlitz-Thouless (BKT) transition, complexity becomes topology dependent. In the BKT phase, it displays a characteristic dependence on the homology of the real-space manifold, and, for *g*-torii, it follows a scaling that is solely genus dependent. We argue that this behavior is intimately connected to the emergence of an order parameter in data space, the conditional connectivity, which displays scaling behavior. Our approach paves the way for an understanding of topological phenomena emergent from many-body interactions from the perspective of Kolmogorov complexity.

Complexity and topology are two cornerstones in our interpretation and understanding of phenomena and mechanisms in diverse fields of science [1]. One particular domain where the two are at play is many-body physics: there, one is interested in how collective behavior emerges from the interplay of a large number of microscopic components, thus enabling effective descriptions that only use a finite number of variables. At the information-theoretic level, this implies a huge reduction of the Kolmogorov complexity corresponding to a given description of the physical system: a handful of parameters are sufficient to capture essential features of the corresponding physical phenomenon [2]. The prototypical example of such reduction of complexity is critical behavior, where correlation functions are precisely determined by a few parameters: the critical exponents [3].

Recently, thanks to dramatic advances in the field of nonparametric unsupervised learning methods [4–6], it has become possible to make quantitative statements about the complexity of many-body processes in the context of molecular science and, broadly speaking, quantum chemistry and physics [7–9]. In those domains, estimating complexity is particularly important to understand the interplay between the multitude of degrees of freedom, and, eventually, to help identify those that are most relevant for a physical phenomenon to occur. This is intimately related to feature selection [10–13] or dimensional reduction [14,15], i.e., discarding features that appear irrelevant or redundant or finding a representation of the data with few variables seen as complicated functions of the original ones. The same tools have been recently applied to critical phenomena [16–18], where they have been shown to lead to qualitative insights on how complexity is tied to critical behavior [19]. Instead, very little is known about how complexity is dictated by topological phenomena: Given a partition function, is it possible to infer that nonlocal degrees of freedom are at play by solely looking at its complexity? Beyond theoretical interest, addressing such a question can also be of experimental relevance to a variety of systems, including cold atom gases [20], arrays of Josephson junctions [21], and quantum annealing architectures [22], where topological phenomena have been recently demonstrated and where single-site resolution is often possible.

Here we show that, in classical statistical mechanics, the topological origin of the Berezinskii-Kosterlitz-Thouless [23]
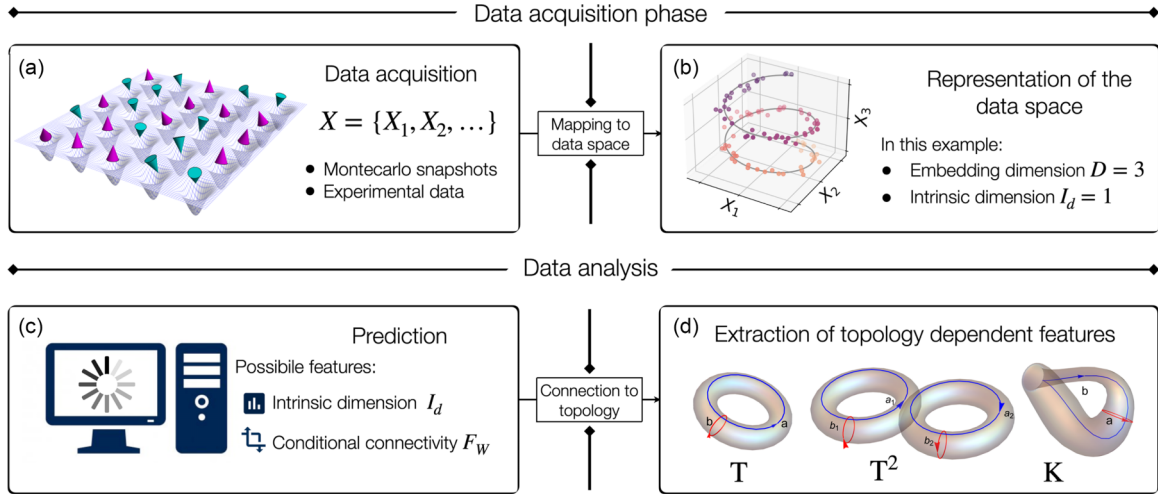
*vittorio.vitale@lpmmc.cnrs.fr

FIG. 1. Guideline to connect complexity of classical partition functions to topological phenomena. (a) Collected snapshots of the physical system—either Monte Carlo configurations or data from experiments. (b) Mapped configurations to a D-dimensional (embedding dimension) data space. (c) Extracted features of the data set, e.g., intrinsic dimension (I_d), conditional connectivity ($F_\mathbf{w}$). (d) Findings connected to the manifolds the system lives in. Here we represent a torus (T), a 2-torus (T^2), and a Klein bottle (K).

mechanism is tied to the emergence of a topological contribution to the Kolmogorov complexity characterizing the partition function. This relation is uniquely determined by the nature of the manifold upon which the statistical mechanics problem is defined, and is in sharp contrast to what happens for models governed by local order parameters, whose complexity is unrelated to the manifold topology.

First, we provide a theoretical framework that argues how complexity decreases as the dimension of the homology group of the real-space manifold increases: this reflects the fact that, the more complex the real-space manifold, the more constrained low-energy excitations of the system are. We then analyze the partition function of XY, Ising, and Heisenberg models on an array of two-dimensional manifolds—orientable as well as nonorientable. For each manifold, we estimate the Kolmogorov complexity utilizing the intrinsic dimension, a measure of the minimal number of degrees of freedom required to describe a data structure. We then define as topological Kolmogorov complexity of a manifold the difference between the complexity of the partition function on that manifold, and the one on the 1-torus (T^1).

In the BKT phase of the XY model, we observe that the complexity is strongly dependent on the manifold the model is defined on. In agreement with our theory, such complexity is dictated by the first homology group and satisfies a phenomenological finite-size scaling. Our results indicate that, while information is nonlocally encoded in real space (i.e., in winding numbers corresponding to vortex excitations), it is remarkably *local* in data space: this is shown by the emergence of a local order parameter—in data space—that we dub *conditional connectivity* and that displays scaling behavior across the BKT transition. The identification of such order parameters provides a key, intuitive connection between homology and complexity. In contrast, for the case of the Ising and Heisenberg model, we find no relation between complexity and real-space topology, again in agreement with our theory.

Our findings demonstrate a direct link between topology and complexity of a many-body description that parallels

recent efforts in both classical [24,25] and quantum [26–29] statistical mechanics in terms of entropy content. Based on the impact of complexity in other disciplines, this connection opens the door towards a transfer of concepts and methods between different fields.

I. KOLMOGOROV COMPLEXITY OF PARTITION FUNCTION SAMPLES

Kolmogorov complexity is the fundamental measure of algorithmic complexity and has found widespread applications in diverse fields of science (for an overview, see Ref. [2]). It corresponds to the length of the shortest computer program that produces a given output, that, in our case, corresponds to the data set built with the configurations sampled along the MC simulations. From an alternative viewpoint, it characterizes the capability of the output (the data set) to be compressed. In the following, our goal is to connect this information-theoretic viewpoint to topological phenomena. To establish the relation between topological features and complexity, we formulate a four-step procedure (summarized in Fig. 1) that encompasses (a) data acquisition, (b) proper representation, (c) data mining analysis in configuration space, and (d) repeated over different manifolds to single-out topological features.

(a) Data structures of partition functions. The first element of our discussion concerns the identification of the data set we target to describe a physical phenomenon. For an equilibrium classical statistical mechanics model, the starting point is very natural: the partition function. While the latter cannot be measured experimentally (nor, in many cases, directly computed), we focus on its sampling. In particular, labeling as X_i an element of the configuration space, our target data set is a collection of N_r such elements,

$$X = \{X_1, X_2, \ldots, X_{N_r}\}, \tag{1}$$

with $N_r \ll D_C$, where D_C is the total dimension of the configuration space (e.g., 2^N for a collection of N Ising spins).

We will discuss the case $X_i \neq X_j$, which, as we argue below, is typical of large many-body systems. The data space X contains information about arbitrary rank correlation functions, so it is expected to provide a full characterization of physical properties. Importantly, it is immediately available both in numerical (e.g., Monte Carlo) and real experiments.

(b) Intrinsic dimension and complexity. Once the data space is defined, it is important to find a suitable measure of information compression that serves as an estimator for the Kolmogorov complexity, $K(X)$. Commonly, data points in a data set are represented as points in a space whose dimension is the number of features needed to describe each sample, which is called the *embedding dimension*. The correlation between the data points determines the manifold in which the data lie, inducing a structure whose dimension is typically much smaller than the embedding one. This is known as *intrinsic dimension* and denoted by I_d. For data sets whose local density is approximately constant at the scale of nearest neighbors, the I_d corresponds roughly to the minimum number of variables needed to describe a data set, and is thus a direct estimator of Kolmogorov complexity. This connection has been shown in graph theory using one of the possible estimators of the I_d: the Hausdorff dimension [30,31].

The relation between intrinsic dimension and complexity has found applications in various fields. In molecular science, I_d is used to analyze the complexity of large data sets of molecular structures, such as protein structures, to gain insights into the underlying structural motifs that govern the molecule functions [32,33]. In computer vision, it is used to analyze images and videos, identifying the most important features and patterns relevant for tasks such as image classification or object detection [34,35]. In physics, I_d is used to characterize both quantum and classical systems [19,36–38], estimating the number of degrees of freedom required to describe a given physical system across phase transitions, and its universality has been numerically proved.

(c) Estimators for the intrinsic dimension. Different approaches have been proposed to estimate I_d [14,39–42]. The technique used here, the TWO-NN method [43], relies on the statistics of distances between nearest-neighbors elements in the data set. The assumption of such approaches is that nearest-neighbor points can be considered as uniformly drawn from I_d-dimensional hyperspheres. Hence, it is possible to set relations between the I_d and the statistics of neighboring distances and is particularly suitable for nonlinear manifolds. In the TWO-NN, one calculates the ratio $\mu_i = r_i^{(2)}/r_i^{(1)}$ for each point X_i in the data set, where $r_i^{(1)}$ and $r_i^{(2)}$ are the nearest-neighbor and the next-nearest-neighbor distances, respectively. Under the assumption above, it can be proved that μ_i is distributed according to

$$f(\mu) = I_d \mu^{-I_d - 1}. \tag{2}$$

Therefore, I_d can be estimated from the cumulative distribution function $P(\mu)$ as

$$I_d = -\frac{\ln[1 - P(\mu)]}{\ln \mu}. \tag{3}$$

An example is shown in Fig. 7 for three cases: the estimation of I_d corresponds to the slope of the straight lines. It is worth mentioning that I_d is a scale-dependent quantity. In particular,

changing the value of N_r, the number of configurations in the data set, amounts to changing the value of I_d but not its functional dependence. Also, as mentioned above, I_d has been utilized to investigate classical and quantum phase transitions because it shows a singular behavior in the vicinity of phase transition due to the change of correlation between the points in the data set. In essence, certain phases (e.g., the ferromagnetic phase in the Ising model or the BKT phase in the XY model [19]) are characterized by configurations that are strongly correlated, as they display the same physical properties. Conversely, outside these ordered phases, the data become lesser correlated the further we move away from the critical point (from a network theory viewpoint, they can be drawn from an Erdös-Renyi network). These effects reflect on the behavior of I_d in parameter space.

Finally, it is worth commenting on the definition of distance to be used in the TWO-NN method. The choice of a proper metric is guided by the fact that it is non-negative, equal to zero only for identical configurations, symmetric, and satisfying the triangular inequality. In this paper, we will use the Euclidean metric to estimate the distance between points as the most natural one to impose a geometry in data space. Other suitable distances may be utilized equivalently in the present case.

(d) Topological Kolmogorov complexity. We now need a criterion to determine whether complexity and topology talk to each other—that is, whether the complexity of a partition function can reveal the fact that information is carried by topological excitations or not. Inspired by similar reasoning at the basis of topological field theory, we use the following approach: Given a model, we input in two distinct (in a way we specify below) topologies and ask ourselves whether and how the complexity of partition function sampling has changed.

For simplicity, we consider a BKT phase (our arguments are immediately extended to cases with different topological defects, such as, e.g., liquid crystals [44]). Our starting point is the partition function of an XY model on a manifold with 1-torus topology, whose sampling and Kolmogorov complexity we denote as X^0 and $K(X^0)$, respectively. We then compare the latter with the one obtained from the sampling X of the partition function of the same model but on a manifold with a different topology.

The rationale behind this is the following [45]. When topological degrees of freedom carry the information in the system, i.e., at low temperatures in the XY model, a new set of constraints appears, such as the winding numbers (**w**) of these topological excitations. The paths on which independent winding numbers can be calculated, which also represent different and independent constraints on the partition function, are strictly related to the particular manifold taken into account. These paths are shown in Fig. 1(d), where different windings are represented on three manifolds: a 1-torus (T^1), a 2-torus (T^2), and a Klein bottle. The presence of different constraints must affect the complexity of the sampling of the partition function. Therefore, this effect is expected to be visible by studying its I_d.

To identify topology-related information content, we then posit the 1-torus as a reference manifold, and we identify as $K_{\text{topo}}^{(\mathcal{M})}$ the *topological Kolmogorov complexity* of the problem

defined on a manifold $\mathcal{M}$ as the difference of the intrinsic dimension computed from the partition function obtained on $\mathcal{M}$, and that on the 1-torus:

$$K_{\text{topo}}^{(\mathcal{M})} = K^{(\mathcal{M})} - K^{(T)}. \tag{4}$$

In the following, we show numerically that this quantity strikingly distinguishes phenomena which do have a topological origin, from those that do not. Most importantly, the former displays a universal scaling behavior for *g*-torii and a characteristic connection to the first homology groups for more complicated closed manifolds, demonstrating an unambiguous connection between topology and complexity.

II. MODELS

To illustrate our ideas above, we utilize three models of classical statistical mechanics: the Ising, Heisenberg, and XY models, as anticipated before. The 2D XY model displays a (BKT) topological phase transition associated with the emergence of topological defects as vortices, at temperature $T_{\text{BKT}} \simeq 0.8933$ [46–48]. The physical degrees of freedom are represented by rotors on the plane and the energy of the system can be written as follows:

$$E(\vec{\theta}) = -\sum_{\langle i,j \rangle} \vec{S}_i \vec{S}_j, \tag{5}$$

where $\vec{S}_i = (\cos\theta_i, \sin\theta_i)$ and $\theta \in [0, 2\pi]$. The configurations are represented by

$$\vec{\theta} = \{\cos\theta_1, \sin\theta_1, \ldots, \cos\theta_N, \sin\theta_N\}, \tag{6}$$

where N is the total size of the system.

To compare the topological character of excitations of the XY model with a case where such phenomenology is absent, we investigated the classical Ising model and the classical Heisenberg model in two dimensions.

The Ising model is described by

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j, \tag{7}$$

where the degrees of freedom are spins $s_i = \pm 1$ and the sum runs on the nearest-neighboring bonds of a lattice. The configuration states are defined as

$$\vec{s} = \{s_1, \ldots, s_N\}, \tag{8}$$

where N is the number of spins in the lattice. This model exhibits a second-order phase transition characterized by the breaking of $\mathbb{Z}_2$ symmetry at the critical temperature $T_c = 2/\ln(1 + \sqrt{2})$.

The Heisenberg model is described by

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} \vec{S}_i \vec{S}_j, \tag{9}$$

where the sum runs on the nearest neighbors and $S_i = (\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, \cos\theta_i)$ is a three-component spin, with $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$. This statistical mechanics model displays no phase transition in temperature.

We utilize the Ising and Heisenberg models as a benchmark: since they do not display a topological phase transition, they can be exploited to verify that the behavior of the Kolmogorov complexity in the 2D XY model is due to the particular topological features of the latter, and not by some other, unwanted (e.g., finite volume) effect.

III. COMPLEXITY AND HOMOLOGY

A. Theory: Winding numbers and information compression

Before embarking on a quantitative determination of the intrinsic dimension in spin models, it is instructive to present a heuristic argument that anticipates the connection between topology and complexity.

Let us consider the data set X introduced in Eq. (1). The complexity of the data set X, i.e., $K(X)$, points out the information stored in the data, and hence how much information content can be encoded by the system. When constraints on the manifold are enforced, we expect it to be reflected in a decrease of complexity as one can attach to the snapshots nontrivial labels, that we denote with W_j, $j = 1, \ldots, N_r$. In formulas, $K(X)$ can be written as the difference between the complexity of the bare data set X^0 and the contribution due to the presence of labels W_j associated with each X_j:

$$K(X) = K(X^0) - K(W|X^0). \tag{10}$$

This observation is intimately related to the topology of the manifold the data set lives in. Let us first illustrate this consideration with a simple example. We consider a set of spin variables describing a strongly ordered Ising-ferromagnetic state on a square lattice defined on a 1-torus. The configurations of the spins will be everywhere analogous to the one obtained by a lattice with open boundary conditions, with the exception of the constraints at the boundaries of the plane. This implies that the additional condition of periodic boundary conditions will reduce the complexity of the data set (physically, this is a geometrical effect, related to the fact that the open boundary conditions make magnetic order weaker in a finite region of space [49]). However, when closed manifolds are considered, we do not expect any difference in the thermodynamic limit: the samples obtained from an Ising ferromagnet will have the same complexity regardless of the underlying manifold.

The situation is expected to be drastically different if the magnetic state has a BKT origin. There, low-energy excitations are vortices: those cannot be distinguished by local operators and are instead labeled by winding numbers. This implies that, at low temperatures, the space of configurations that is spanned by the partition function is strongly constrained by the fact that vortices must have specific windings. This is in strong contrast with what happens at higher temperatures, where vortices are unbound and free to propagate (and the concept of winding ultimately loses its significance as configurations with very strong local spin fluctuations start being present).

1. Low-temperature expansion

In the low temperature phase, it is possible to make concrete, quantitative predictions on the above mentioned picture. Let us consider the data set obtained by sampling the XY partition function for $T \ll T_c$, with $N_r \gg 1$, and on a

closed manifold with homology group $\mathcal{G} = \bigoplus_{j=1}^{k} \mathbb{Z}_{\alpha_j}$. In this regime, a large amount of configurations will have all spins aligned in a given direction—all of them in the extreme case $T = 0$. Such a cluster has intrinsic dimension ~ 1, and the average distance between states is $\propto \sqrt{N}/N_r$ [50]. This cluster, that we refer to as $T = 0$ cluster, has constant curvature: in fact, it can be (locally) thought of as a straight line in data space, due to the fact that all spins are described by a single angle θ.

Importantly, these are not the only configurations that are sampled. A finite fraction of them, representing low-energy excitations, is also present. They can be divided into two classes: (i) configurations that host local spin excitations, at energy difference $\delta\epsilon$ of order 1 (and, thus, vanishing energy density above the lowest energy states), and (ii) topological excitations, also at finite energy difference above the lowest energy state, and with a defined set of winding numbers. Note that the fact that topological excitations, despite having wildly different spin configurations with respect to the fixed direction case, still have vanishing energy density is particularly relevant here.

The overall data structure, and its complexity, are drastically affected by the sampling of excitations. We scrutinize them analyzing the various length scales at play in data space. To do that, let us first focus on excitations where only local changes of configurations are considered [case (i)]. Given that those only involve the change of a handful of spins, they describe a local modification of the one-dimensional manifold found at $T = 0$. The intrinsic dimension of the corresponding data set has been estimated following a maximum likelihood reasoning in Ref. [19], and scales as $I_d \propto \frac{N_r}{\sqrt{(N)}}$. The data set does not feature clustering and can be thought of as a (highly anisotropic) multidimensional cylinder of length $\propto \sqrt{N}$ (simply given by the $T = 0$ cluster length). We thus refer to the data set as the cylinder cluster. The width of the cylinder can be computed as follows. Given two configurations $\vec{\theta}^0$ and $\vec{\theta}^1$, the first representing a fully polarized state with $\theta_k^0 = \theta^0$ and the second a state with thermal excitations on top of that polarized state, one gets

$$d(\vec{\theta}^0, \vec{\theta}^1) = \left\{ 2 \sum_{k-1}^{N} \left(1 - \cos(\theta^0)\cos\left(\theta_k^1\right) - \sin(\theta^0)\sin\left(\theta_k^1\right) \right) \right.$$
$$= \left\{ 2 \left[N - \cos(\theta^0) \sum_k \cos\left(\theta_k^1\right) \right.\right.$$
$$\left.\left. - \sin(\theta^0) \sum_k \sin\left(\theta_k^1\right) \right] \right\}^{1/2} \quad (11)$$

Given that $\vec{\theta}^1$ only differ by $\vec{\theta}^0$ for few spins that are surrounded by spins that are aligned along θ^0, one can rewrite the expression above as

$$d(\vec{\theta}^0, \vec{\theta}^1) = \sqrt{2 \sum_k \left(E^0 - E_k^1 \right)}, \quad (12)$$

where E_0 is the average energy of $\vec{\theta}^0$, and E_k is the energy associated to the k vertex in $\vec{\theta}^1$. The formula above is

providing a direct link between energy of excitations and average distance within the cylinder cluster, that is $d_{\text{cyl}} \simeq \delta\epsilon = O(1)$.

In the absence of topological excitations, the above description will be sufficient to characterize the full data structure. Topological excitations change this situation rather drastically [case (ii)]. As we show below, analyzing the average distance between configurations, they generate clusters that are geometrically disconnected from the cylinder one, and whose number is given by the dimension of the homology group of the manifold the system is defined on. Such data structure leads to a characterization of the topology dependence of the intrinsic dimension, which we comment upon at the end of the section.

The key observation is that the average distance d_{topo} between the cylinder cluster and topological excitations at vanishing energy density is much larger than the estimated d_{cyl}. This is due to the fact that one cannot go from Eq. (12) to Eq. (13) as done previously; namely, in the case of topological excitations, each spin configuration varies smoothly with position, so it is not possible to assume that most neighbors have fixed magnetization. Instead, in this case, one gets, from Eq. (12), that

$$d_{\text{topo}} \simeq \sqrt{2 \sum_k \left[1 - \cos(\theta_k) - \sin(\theta_k) \right]} \simeq \sqrt{N} \quad (13)$$

due to the fact that the averages of $\cos(\theta_k)$ and $\sin(\theta_k)$ are negligible with respect to the constant term. From a different perspective, this can be understood as a consequence of the lack of long-range order: the latter fixes the average value of $\cos(\theta_k)$ to vanish at finite temperature. While this is, of course, not strictly true for arbitrary low-energy states, it gives us a rough estimate of the mismatch in scaling in Eq. (14).

The arguments above indicate that topological excitations lead to topological clusters in data space that are distinct from the main cylinder cluster. The number of such additional clusters is also important: in particular, this number scales, in general, with the number of possible low-energy excitations in the theory (due to the fact that the distance among these clusters is also typically of order L). For closed manifolds, such a number is the dimension of the first homology group of the manifold.

The effect of topological clusters on the intrinsic dimension of the manifold can be understood as follows. The intrinsic dimension estimation is still dominated by the cluster cylinder, since the latter contains most of the data points and, at the same time, the nearest-neighbor distance between points in that cluster is smaller than that in the topological ones [51].

Summarizing, the predictions of our theory are the following: (i) complexity is topology independent in the absence of topological excitations; (ii) in the presence of topological excitations, complexity decreases proportionally to the dimension of the first homology group of the real-space manifold; and (iii) the decrease is associated to clustering and to a different distribution of nearest-neighbor distances in the sample.

Let us anticipate some concrete expectations based on these predictions. For the case of g-torii, one expects the complexity $K(X)$ to decrease proportionally to $2g$. We note that similar scalings also apply to topological phases in
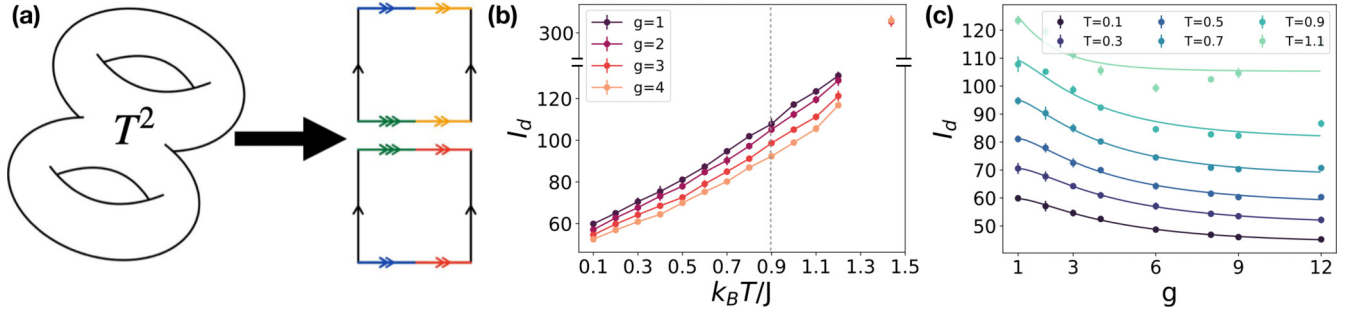
FIG. 2. Plot of I_d for g-torii. We consider the XY model on a square lattice. (a) Diagram of the construction of the 2-torus (T^2), edge with same color and arrows must be connected with periodic boundary conditions. (b) I_d of the two-dimensional XY model as a function of temperature T, for different increasing value of the genus g (top to bottom); (c) scaling of I_d as a function of the genus g, the lines correspond to Eq. (15), for increasing temperature going from bottom to top. In both panels, we consider a square lattice with linear dimension $L = 72$. We average our results over ten simulations with $N_r = 5000$ different configurations. The error bars are estimated as the standard deviation of the mean. The dashed vertical line marks the critical point.

quantum statistical mechanics (e.g., two-dimensional lattice gauge theories [26]). For the case of closed manifolds, we instead expect a dependence on the number of possible loops—that is, the dimension of the abelianization of the first homotopy group (not its Betti's number). For instance, we expect the same complexity for Klein bottles and 1-torii, despite the fact that their genus is different.

In conclusion, there are aspects of our theory that, at present, we do not know how to address. The first relevant one is, up to which temperatures such a picture holds. The second nontrivial aspect is the role of N_r: here, the main challenge is that maximum likelihood arguments (at the basis of the cylinder cluster scaling) are known to not always be accurate in terms of capturing the role played by the data set size. To clarify these points, and to corroborate our full theory picture, we rely on numerical experiments.

B. Numerical experiments

We now provide numerical simulations to support our theory. We consider a two-dimensional square lattice whose boundary conditions are controlled such that topologically different manifolds are realized. In Fig. 2(a), we illustrate how to build a 2-torus (T^2) with appropriate constraints on the boundaries of two square lattices. Equivalently, one could start from an octagon and apply periodic boundary conditions on opposite sides. The extension to generic g-torii is straightforward starting from regular polygons with $4g$ sides or using g square lattices. In practice, we start from a square lattice and divide it in vertical stripes instead of considering g square lattices. In Appendix C, we show numerical evidence of the equivalence of the two approaches. Nonorientable manifolds like the Klein bottle, the Moebius strip, and the real projective plane (RPP) can be realized with the appropriate boundary conditions on a single square lattice [see Fig. 3(a)].

To sample the configurations according to the partition function, we use Wolff's algorithm [52]. The data sets are defined as a collection of N_r points, such as in Eq. (1), each corresponding to a given configuration of spins as specified in the previous section. We take particular care of generating decorrelated configurations to avoid any possible bias of the sampling. To do so, we choose a suitable interval between

configurations that are saved during the Monte Carlo steps. We note that, to the best of our knowledge, for some of these manifolds, numerical simulations were not reported before, so we verified that autocorrelations are not significantly changed by the manifold topology.

1. Orientable closed manifolds

The first step of our analysis is the comparison of different g-torii. The number of nontrivial loops of a genus g torus (T^g) is $2g$, since its first-homotopy group is $\pi_1(T^g) = \bigoplus_{1=1}^{2g} \mathbb{Z}$. We expect, in a phase with no topological features, that the complexity should not show any dependence on g. Conversely, in the BKT phase, we expect it to be dependent on the particular manifold taken into account. Our theory predicts that the complexity $K(X, g)$ of the data set X as a function of the genus g decreases with g monotonically, up to finite volume effects. To cope with the latter, we utilize a phenomenological finite volume ansatz as

$$K(X, g) = K_0 - a_0 \cdot g e^{-a_1\sqrt{g}}, \tag{14}$$

which smooths out contributions at large g, since those are more sensitive to finite volume effects scaling roughly as $V \simeq g^2$ [53]. On the right-hand side, we include a topologically independent term K_0 and a topologically dependent term, that will capture the TKC.

In Fig. 2, we show the estimated value of the intrinsic dimension I_d as a function of temperature for different values of the genus g (b) and as a function of the genus g for different values of the temperature (c). The system taken into account is a square lattice with linear dimension $L = 72$ and the number of configurations of the data set is $N_r = 5000$. We observe that the curve are monotonically increasing with temperature as expected. The presence of the phase transition is not detected by the intrinsic dimension, differently from Ref. [19], because the number of configurations considered is lower; however, this fact is irrelevant to our analysis as we show below.

The most striking signal in Fig. 2(b) is that, below $T/J \lesssim 1.2/k_B$, the intrinsic dimension decreases as a function of the genus, and the distance among the curve remains constant, at least deep in the topological phase, where our reasoning applies. We scrutinize such genus dependence in Fig. 2(c).
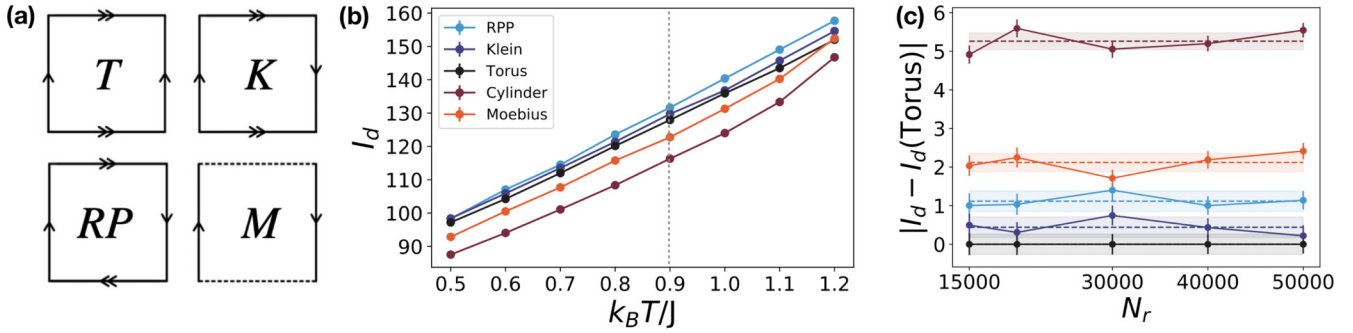
FIG. 3. Comparison between RPP, Klein, torus, cylinder, and Möbius topologies. We consider the XY model on a square lattice. (a) Diagram of the construction of torus (T), Klein bottle (*K*), real projective plane (*RP*), and Möbius strip (*M*). Edges with the same arrows must be connected with periodic boundary conditions. Edges with arrows in opposite directions denote antiperiodic boundary conditions. Dashed lines stand for open boundary conditions. (b) Results of I_d as a function of temperature across the BKT phase transition on a square lattice with linear dimension L; from top to bottom, we have results for RPP, Klein ~ torus, Möbius strip, cylinder. We average our results over ten simulations with $N_r = 5000$. The error bars are estimated as the standard deviation of the mean. The dashed vertical line marks the critical point. (c) $|I_d - I_d(\text{torus})|$ at $K_b T/J = 0.1$ as function of N_r for a square lattice with linear dimension $L = 48$. The dashed lines correspond to the average of I_d over N_r for the given manifold. Their error bars are represented by the shaded regions of the same colors. From top to bottom, we have results for cylinder, Möbius strip, RPP, Klein ~ torus.

The agreement between the points and the ansatz [curves obtained via a best fit based on Eq. (15)] is excellent for low temperatures. Oppositely, it worsens with increasing T and cannot describe I_d above the phase transition where the topological properties are not relevant any longer. We argue that in the thermodynamic limit, the curves for different values of the genus should overlap for all temperatures above the critical one $T > T_{\text{BKT}}$. To check this assumption, we calculate I_d for $T = 1.4 > T_{\text{BKT}}$ and observe that the values of the different curves are compatible within the error bars, as shown in Fig. 2(b).

2. Nonorientable and open manifolds

While for orientable manifolds we gave our predictions based on the number of independent nontrivial loops, in the case of nonorientable ones we need to properly extend our diagnostic. This can be done by looking at the first homology group of the manifold, i.e., the Abelianization of their first homotopy group. We expect that (i) data sets defined on spaces with the same dimension of the homology group will be characterized by the same Kolmogorov complexity irrespectively of their genus and their orientability; (ii) the complexity may be strongly affected by the number of open boundaries in finite-size settings; and (iii) in the phase with no topological properties, the complexity shall show no scaling with the topological features of the manifolds.

Homology and complexity.. In Fig. 3(b), we plot I_d of the XY model, on a square lattice with linear dimension $L = 72$, calculated from $N_r = 5000$ configurations of spins as in Eq. (1). We consider the cases of spins arranged on a RPP, cylinder, torus, Klein bottle, and Möbius strip [schematically depicted in Fig. 3(a)]. The corresponding homology groups are written in Table I.

This choice of topologies allows us to consider the influence of homotopy groups and boundaries separately. According to our definition of TKC, we utilize the 1-torus as a reference: the cylinder allows us to investigate the role

of boundaries in the estimation of I_d, while the others are nonorientable, closed manifolds which are the main focus of the present section.

From Fig. 3(b), few qualitative features are appreciable. First, the presence of boundaries in the manifold decreases the value of I_d consistently. This is clear from the result of the Möbius strip (one boundary) and cylinder (two boundaries) that have the same homology group. This can be explained as due to the constraint enforced by the presence of an edge that amounts to adding more labels (W_j) to the data. Second, we see that the torus and Klein bottle have compatible I_d for the whole temperature range investigated, in agreement with our reasoning in the previous section (Sec. III A). To quantify such a connection, in Fig. 3(c), we plot the absolute value of the topological Kolmogorov complexity as a function of N_r for a square lattice with $L = 48$ at low temperature $k_b T/J = 0.1$. For each topology, we also estimate an average value of the TKC across our range of N_r: this is indicated by dashed lines, and the corresponding estimated error is indicated by a shaded area.

Two main observations are in order. The first is about the relative behavior of the closed manifolds: There is a sharp difference between the RPP and 1-torus. In particular, their TKC difference is compatible with 1 regardless of the sampling. This is remarkable as, e.g., for $N_r = 50\,000$, the overall intrinsic dimension exceeds 100. Oppositely, the difference between the Klein bottle and 1-torus is always small and is compatible with 0 for three values of the sampling. Thus, we can state that the TKC of a partition function supporting topological excitations depends on the homology of the manifold

TABLE I. Homology groups of the topologies considered.

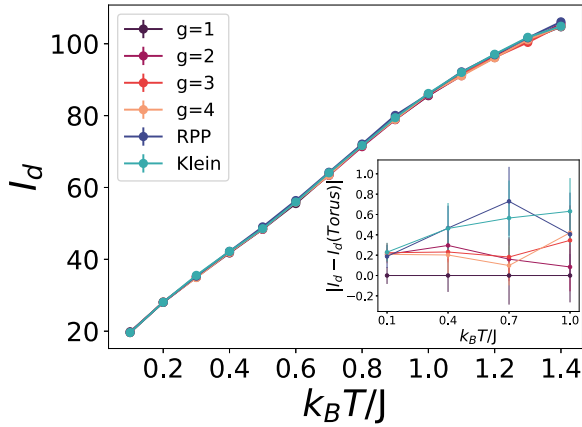|	Closed manifolds		Open manifolds	
RPP	Torus	Klein bottle	Cylinder	Möbius strip
$\mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}_2$	$\mathbb{Z}$	$\mathbb{Z}$

FIG. 4. Plot of I_d for the Heisenberg model. We consider a square lattice with linear dimension $L = 72$. We consider $N_r = 5000$, and I_d is estimated as average over ten simulations. The error bars are estimated as the standard deviation of the mean. The absence of any topological features is evident. In the inset, we show the difference of I_d of all the topologies with respect to the reference case of the torus. We observe the values are compatible within one standard deviation of the mean.

(and, e.g., not only on its genus). The second observation is that, as expected, the presence of edges is strongly affecting the TKC. This is evident from the fact that the Möbius strip, cylinder, and RPP have very distinct TKCs, despite their homology groups having the same dimension or being exactly the same.

We now contrast the XY results with what we observe for the case of the Heisenberg model, depicted in Fig. 4. In that case, it is known that no transition is present as a function of temperature. We observe that the values of I_d are compatible for any topology. We attribute this to the fact that the model is topologically trivial and shows a clearly different behavior with respect to the XY model. In the inset of Fig. 4, we plot $|I_d - I_d(\text{torus})|$ for a few values of temperature to highlight that the curves are compatible within error bars. In Appendix A, we also show the results for the Ising model. The conclusions we can draw are similar. There is not a clear trend in the value of the intrinsic dimension with respect to topology. Indeed, all the values—for different topologies—are compatible within error bars since both phases of the Ising model are topologically trivial.

3. Conditional connectivity: A local order parameter in data space

To validate the argument above, we need to identify testable predictions that connect topological excitations in real space with the data structure and that can be independently verified when investigating partition function samples. Informed by our analysis above, we focus on the role of winding numbers in the data structure.

The winding number w_Γ is a key quantity for studying topological systems. In a spin system, it is an estimate of the number of times the spins wind following a given path Γ along the manifold. In the case of the XY model, where the ith spin

is represented by a single angle θ_i, it can be written as follows:

$$w_\Gamma = \frac{1}{2\pi} \sum_{i=1}^{L_\Gamma} \Delta\theta_{i,i+1}, \qquad (15)$$

where Γ is comprised of L_Γ sites, and $\Delta\theta_{i,i+1} = \theta_{i+1} - \theta_i$ is rescaled into the range $(-\pi, \pi]$. As shown in Fig. 1(d), the possible independent paths on which winding numbers can be calculated depend on the manifolds themselves. We define as $\mathbf{w} = \{w_{\Gamma_1}, \ldots, w_{\Gamma_k}\}$ the collection of w'_Γs calculated on the paths $\Gamma_1, \ldots, \Gamma_k$. For instance, in the case of a torus (T) and a 2-torus (T^2), one has $\mathbf{w}(T) = \{w_a, w_b\}$ and $\mathbf{w}(T^2) = \{w_{a_1}, w_{b_1}, w_{a_2}, w_{b_2}\}$, respectively [see Fig 1(d)]. Remarkably, $\mathbf{w}$ is a key metric for understanding the clustering structure of configurations in phase space because winding numbers capture the topological nature of excitations, namely, above the BKT transition, vortex-antivortex pairs are unbounded, resulting in a wide range of different $\mathbf{w}$ values among sampled configurations. However, in the quasi-long-range-order regime (below T_{BKT}), the majority of configurations have the same $\mathbf{w}$, i.e., for a torus topology $\mathbf{w} = (0, 0)$.

According to the rationale above, the physical properties of the BKT phase transition strongly affect the connectivity between neighboring configurations. In particular, the fraction of points whose first and second neighbors have the same $\mathbf{w}$, that we dub *conditional connectivity* $F_{\mathbf{w}} = N_{\mathbf{w}}/N_r$, must be informative of the transition. Here N_r is the total number of configurations sampled and $N_{\mathbf{w}}$ is the number of configurations whose first two neighbors have the same $\mathbf{w}$. Evidently, $F_{\mathbf{w}}$ is close to zero above the BKT transition (where configurations exhibit a wide range of $\mathbf{w}$) but is equal to 1 in the topological phase (where the values of $\mathbf{w}$ are constrained due to the topological nature of the excitations and the configurations appear as clusters in data space).

In Fig. 5(a), we show $F_{\mathbf{w}}$ in the case where only two paths on the lattice are taken into account for each manifold, which is to be compared with Fig. 5(b) where all the possible nontrivial paths are considered. We observe that the two figures display the same qualitative behavior, showing a transition near the critical temperature, but they are quantitatively different. This is expected since the larger the number of possible independent windings, the more constrained the system in the BKT phase. In Fig. 5(c), we show $F_{\mathbf{w}}$ for the case discussed in Fig. 5(a) but for different system sizes. In the inset, we perform a collapse of the curve with estimated parameters $c = 0.72(1)$ and $a = 1.76(7)$. To do so, we use a suitable mesh for the parameters, fixing the critical temperature $k_B T_c/J = 0.8933$ [48]. Then we calculate the scaling variables

$$x = \ln L - at^{-1/2}, \quad y = F_{\mathbf{w}} \ln(L/c) \qquad (16)$$

and choose a parametric function hypothesis $f(x, a, c)$ to be compared with y. Here $t = (T - T_c)/T_c$. We use the Levenberg-Marquardt algorithm to compute the best fit for each choice of a, c, and f, where f is considered to be a polynomial with varying degree $k = 5, 6, 7, 8$. Our estimates for a and c are calculated as the average of all the best fits and their errors are the standard deviation of the latter. We
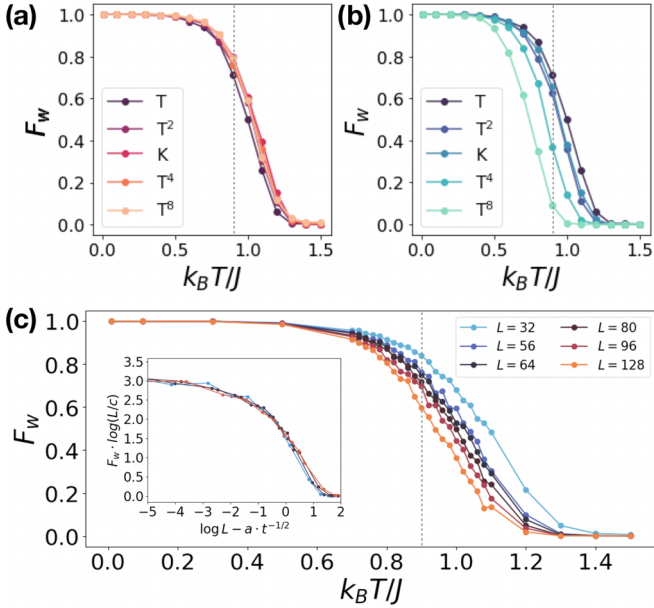
FIG. 5. Fraction of points in data space whose first two nearest neighbors have the same winding number $F_{\mathbf{w}}$. In (a), we consider only two nontrivial loops. In (b), we consider every possible nontrivial loop on the manifold. The crossover point moves (from right to left) depending on the topology according to the legend. The square lattice has linear dimension $L = 72$. In (c), we consider $F_{\mathbf{w}}$ for the 1-torus at different system sizes L^2, going from $L = 32$ (top curve) to $L = 128$ (bottom curve). These results indicate how, below the critical temperature (labeled by the dashed vertical lines), the *local* (connectivity) properties of the data set become severely constrained by *global* properties (winding numbers) of the configurations sampled. This is a local order parameter in data space that has a universal behavior at the transition, as shown in the inset. Here $t = K_b(T - T_c)/J$, with T_c value set at $K_b T_c/J = 0.8933$, and $a = 1.767 \pm 0.003$ and $c = 0.721 \pm 0.003$, estimated through a data collapse.

observe that the conditional connectivity $F_{\mathbf{w}}$ displays a universal behavior. It can be considered an order parameter in the data space and, as argued before, is closely related to the fact that the Kolmogorov complexity displays a topological contribution in the BKT phase. In fact, the local connectivity (as estimated by I_d) is strongly constrained by the global properties of the manifold, which are expressed by the behavior of $F_{\mathbf{w}}$. This discussion allowed us to make two clear statements that we want to summarize here: (i) The connectivity between neighboring configurations is strongly related to the physical properties of the BKT transition. Remarkably, nearest-neighbor configurations have identical physical properties (winding numbers) when the system is in the ordered (topological) phase. Conversely, in the disordered phase, the first and second neighbors' physical properties are entirely random as witnessed by $F_{\mathbf{w}}$ approaching zero. (ii) Changing the number of w's calculated and compared, the transition point ($F_{\mathbf{w}}$ going to 1) moves as more constraints are applied on the configurations. Hence, it should be expected to have different behaviors for the Kolmogorov complexity $K(X)$ on different manifolds, as has clearly been shown before.

IV. DISCUSSION AND CONCLUSIONS

We have proposed a theory that connects topological properties of classical statistical mechanics to the Kolmogorov complexity of the corresponding statistical sampling of the many-body state according to the problem partition function. The original approach we proposed is inspired by topological field theory: take a manifold, sample the partition function on the model defined on the manifold itself, and get an invariant. By gauging such a procedure on a 1-torus, we have defined a topological Kolmogorov complexity associated to a given problem, on a given manifold. Based on the connection, we establish between topological properties in real space and data structures in configuration space, the overall conclusion of our theory is that the more constrained a many-body system is by topological effects, the lower the Kolmogorov complexity of its sampling will be.

We have then addressed our theory in the framework of classical statistical mechanics models: Ising, Heisenberg, and XY defined on square lattices. In the Berezinskii-Kosterlitz-Thouless phase of the XY model, the complexity of the partition function strongly depends on the homology of the target manifold, demonstrating a strong, qualitative connection between topology and complexity. Most importantly, the topological Kolmogorov complexity attains universal (in the sense of sampling-independent) values in the BKT phase that are suggestive of its direct link to the winding of spins in the manifold. While some of these values appear compatible with integer TKC, we believe a more extended study is required here, that should address larger data sets from the ones we consider here. This will require the development of unique methods for intrinsic dimension estimation, in particular, with respect to errors. In our numerical experiments, we have also observed a strong dependence on complexity with respect to boundaries. Finally, at high temperatures, and in the Ising and Heisenberg models, we observe no connection between topology and complexity, as expected.

Within the XY model, our theory allows us to track the origin of the relation between complexity and topology by analyzing the interplay between metric distances and winding numbers. In particular, we identify an order parameter—conditional connectivity—that, while fully local in data space, correctly predicts the topological transition taking place in real space. The order parameter satisfied phenomenological scaling relations, further proof of its relation with the BKT transition. This interpretation might be useful in understanding how learning methods such as dynamical maps and persistent homology have found very convincing applications on the XY model [54–57] by specifically exploiting information about winding numbers. In fact, these methods might have indirectly detected the simple order parameters we proposed. It is worth commenting on the fact that, while exotic topologies might seem hard to realize, in general, they are an experimental reality in quantum simulators. In fact, those can be achieved in a variety of manners, including synthetic dimensions [58,59], atom configurations, and atom rearrangements in optical tweezers, the latter already being experimentally demonstrated [60].

Summing up, our theory and numerical experiments support the fact that real-space topological information becomes

local in data space. This last conceptual insight can serve as a starting point for the exploration of the interplay of complexity and topology beyond equilibrium classical statistical mechanics. One clear avenue is off-equilibrium physics, where the same toolbox we used here shall be, in principle, directly applicable. Another horizon is quantum equilibrium systems. In that context, the connection between information and topology has found profound applications in topological matter, e.g., the formulation of topological entanglement entropy [26]. However, such quantities are typically not amenable to experimental tests at large scale, as they require the knowledge of the full system wave function. A statistical approach like the one pursued here could reveal signatures of topological order and/or topological effects even in the presence of limited, but statistically significant, sampling, as long as experiments are able to dynamically tune topology (as demonstrated, e.g., in Refs. [61,62]). It is, however, important to stress that, while our paper provides basic tools in this direction (for instance, the partition functions we study can be thought of as path integrals of one-dimensional fluids with fractionalized excitations), attacking the quantum regime necessarily requires a more convoluted approach. This is not only due to the fact that topology has different incarnations in that setting (e.g., topological order, band structures, and critical phenomena) but, most importantly, to the fact that quantum systems provide data sets in multiple basis. It is an open question to determine how a clever combination of those, possibly together with randomized sampling [63], would lead to a direct connection between topology and complexity.

APPENDIX A: ISING MODEL RESULTS

We here contrast the XY results with what we observe for the case of the Ising model. In particular, we show I_d as a function of the temperature for Torii of different genus g [Fig. 6(a)], and other topologies [Fig. 6(b)]. We consider the Ising model on a square lattice with L = 72. We take $N_r = 10^4$
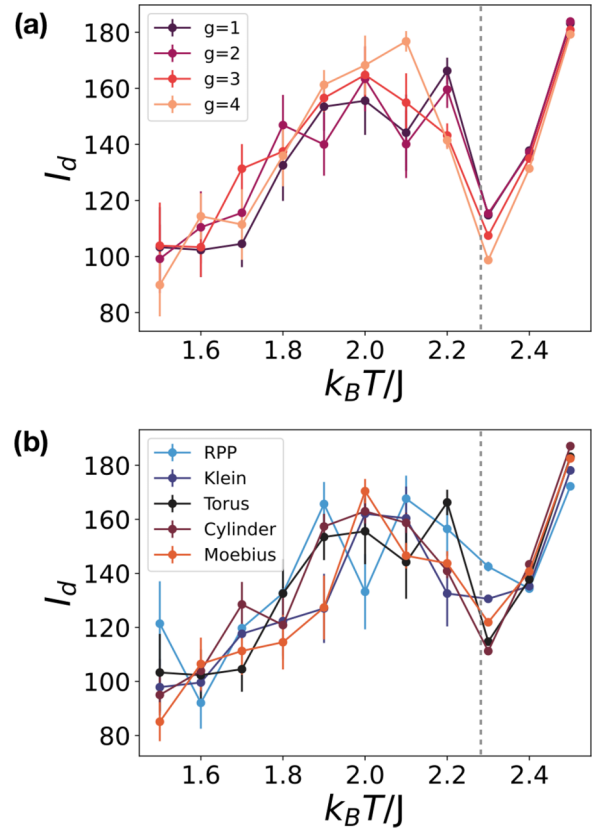

FIG. 6. Plot of I_d for the Ising model. We consider a square lattice with linear dimension $L = 72$. We consider $N_r = 10^4$ and I_d is estimated as average over ten simulations. Value of I_d for different topologies. For both (a) and (b), we observe all the values are compatible within the error bars as both phases of the Ising model are topologically trivial. The critical point is denoted with a dashed vertical line.

snapshots and I_d is estimated as an average over ten simulations. We observe that I_d clearly shows a local minimum near the critical temperature as expected but does not show dependence on the topology of the manifold the systems lie on. We observe all the values are compatible within the error
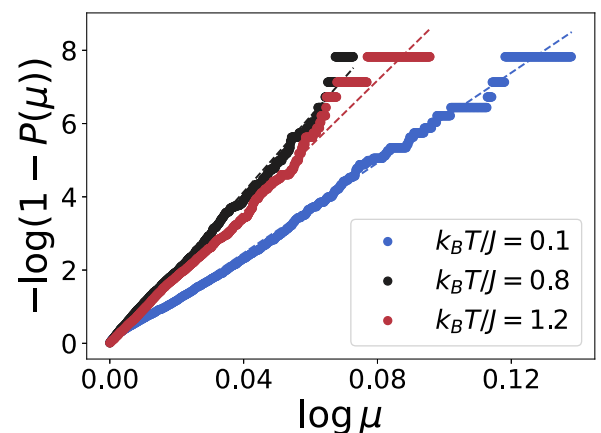

FIG. 7. Check of the Pareto distribution. Here we consider a square lattice with $L = 72$ and $N_r = 2500$. Here $t = k_b T / J$.
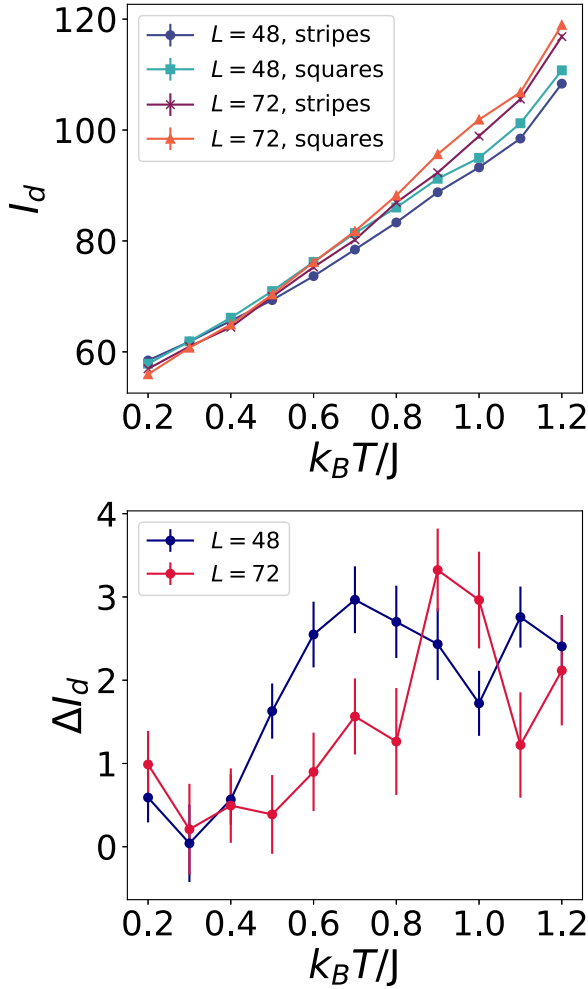
FIG. 8. Comparison between the different procedures for getting a 4-torus. Comparison between four squares lattices with $L = 24$ and rectangular lattices with linear dimensions 48 and 12; and between square lattices with $L = 32$ and rectangular lattices with linear dimensions 72 and 18.

bars as both phases of the Ising model are topologically trivial. Moreover, finite-size effects here play a strong role at the transition point, where most configurations differ in a way that is distinct from the XY case. Here, the complexity of Klein and torus—that are within error bars for the XY case given their homotopy groups—widely differ.

APPENDIX B: PARETO DISTRIBUTION CHECK

The starting assumption of the TWO-NN method is that the ratio $\mu_i = r_i^{(2)}/r_i^{(1)}$, of the next-nearest and nearest neighbors is Pareto distributed, namely, that $f(\mu) = I_d \mu^{-I_d-1}$. Here we check this assumption for a few cases of the XY model; we fix temperature $k_b T/J = 0.1$, $k_b T/J = 0.8$, and $k_b T/J = 1.2$, and compute I_d. Then we compare the cumulative distribution corresponding to the estimated I_d and the one computed from the distribution of distances in the data set. We obtain the plots in Fig. 7. Perfect agreement between the curve (exact cumulative distribution function) and the points (cumulative distribution obtained from the samples configurations) is evident.

APPENDIX C: STRINGS VS SQUARES

In the main text, we described the method we use for computing I_d. In a nutshell, we sample configurations via Wolff's algorithm and realize the different manifolds appropriately tuning the boundary conditions.

We observe that in the case of the g-torii, the easiest way to do so, from a computational point of view, is to consider a single square lattice, divide it in g stripes, and apply the necessary periodic boundary conditions, instead of considering g square lattices. This allows us to easily consider the same number of lattice sites, i.e., the same lattice volume, for each genus g and to also realize odd-g torii. In this section, we present the comparison, for $g = 4$, between I_d calculated using four square lattices or a single square lattice divided into four stripes. In the upper panel of Fig. 8, we present the comparison between 48^2 sites and 72^2 sites for the two different geometries for the XY model. We observe that even though they differ around the critical temperature (which is shifted towards $k_b T/J = 1$ due to finite size corrections), increasing the system size entails that the agreement between the two geometries grows. This is due to the fact that the correlation length grows at the transition point such that finite-size effects are much more relevant. However, we observe that these effects at the transition do not matter in view of the purpose of the main text since our understanding applies at the topological phase. In the lower panel of Fig. 8, we show explicitly the difference in I_d of the two approaches ΔI_d. We observe ΔI_d starting to grow later in the case of larger system sizes, testifying that $\Delta I_d > 0$ is a finite-size effect.

[1] J. Ladyman, J. Lambert, and K. Wiesner, What is a complex system? Eur. J. Philos. Sci. **3**, 33 (2013).

[2] M. Li, P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications* (Springer, New York, 2008), Vol. 3.

[3] S. Carroll, The biggest ideas in the universe | 23. criticality and complexity.

[4] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv:1802.03426.

[5] M. Kaya and H. Ş. Bilge, Deep metric learning: A survey, Symmetry **11**, 1066 (2019).

[6] B. Kulis *et al.*, Metric learning: A survey, Found. Trends Mach. Learn. **5**, 287 (2012).

[7] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physics sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[8] J. Carrasquilla, Machine learning for quantum matter, Adv. Phys.: X **5**, 1797528 (2020).

[9] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).

[10] A. Jović, K. Brkić, and N. Bogunović, A review of feature selection methods with applications, in *Proceedings of the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (IEEE, Opatija, Croatia 2015), pp. 1200–1205.

[11] J. Cai, J. Luo, S. Wang, and S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing **300**, 70 (2018).

[12] X. Deng, Y. Li, J. Weng, and J. Zhang, Feature selection for text classification: A review, Multimedia Tools Appl. **78**, 3797 (2019).

[13] A. Glielmo, C. Zeni, B. Cheng, G. Csányi, and A. Laio, Ranking the information content of distance measures, PNAS Nexus **1**, pgac039 (2022).

[14] L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. **9** (2008).

[15] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[16] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).

[17] W. Hu, R. R. P. Singh, and R. T. Scalettar, Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination, Phys. Rev. E **95**, 062122 (2017).

[18] S. J. Wetzel, Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders, Phys. Rev. E **96**, 022140 (2017).

[19] T. Mendes-Santos, X. Turkeshi, M. Dalmonte, and A. Rodriguez, Unsupervised learning universal critical behavior via the intrinsic dimension, Phys. Rev. X **11**, 011040 (2021).

[20] Z. Hadzibabic, P. Krüger, M. Cheneau, B. Battelier, and J. Dalibard, Berezinskii-Kosterlitz–Thouless crossover in a trapped atomic gas, Nature (London) **441**, 1118 (2006).

[21] Z. Han, A. Allain, H. Arjmandi-Tash, K. Tikhonov, M. Feigel'Man, B. Sacépé, and V. Bouchiat, Collapse of superconductivity in a hybrid tin–graphene Josephson junction array, Nat. Phys. **10**, 380 (2014).

[22] A. D. King, J. Carrasquilla, J. Raymond, I. Ozfidan, E. Andriyash, A. Berkley, M. Reis, T. Lanting, R. Harris, F. Altomare *et al.*, Observation of topological phenomena in a programmable lattice of 1,800 qubits, Nature (London) **560**, 456 (2018).

[23] V. Drouin-Touchette, The Kosterlitz-Thouless phase transition: An introduction for the intrepid student, arXiv:2207.13748.

[24] H.-H. Tu, Universal entropy of conformal critical theories on a Klein bottle, Phys. Rev. Lett. **119**, 261603 (2017).

[25] W. Tang, L. Chen, W. Li, X. C. Xie, H.-H. Tu, and L. Wang, Universal boundary entropies in conformal field theory: A quantum Monte Carlo study, Phys. Rev. B **96**, 115136 (2017).

[26] A. Kitaev and J. Preskill, Topological entanglement entropy, Phys. Rev. Lett. **96**, 110404 (2006).

[27] M. Levin and X.-G. Wen, Detecting topological order in a ground state wave function, Phys. Rev. Lett. **96**, 110405 (2006).

[28] P. Fromholz, G. Magnifico, V. Vitale, T. Mendes-Santos, and M. Dalmonte, Entanglement topological invariants for one-dimensional topological superconductors, Phys. Rev. B **101**, 085136 (2020).

[29] T. Micallo, V. Vitale, M. Dalmonte, and P. Fromholz, Topological entanglement properties of disconnected partitions in the Su-Schrieffer-Heeger model, SciPost Phys. Core **3**, 012 (2020).

[30] T. Mendes-Santos, M. Schmitt, A. Angelone, A. Rodriguez, P. Scholl, H. Williams, D. Barredo, T. Lahaye, A. Browaeys, M. Heyl *et al.*, Wave function network description and Kolmogorov complexity of quantum many-body systems, arXiv:2301.13216.

[31] L. Staiger, Kolmogorov complexity and Hausdorff dimension, Inf. Comput. **103**, 159 (1993).

[32] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature (London) **559**, 547 (2018).

[33] E. Facco, A. Pagnani, E. T. Russo, and A. Laio, The intrinsic dimension of protein sequence evolution, PLoS Comput. Biol. **15**, e1006767 (2019).

[34] S. Gong, V. N. Boddeti, and A. K. Jain, On the intrinsic dimensionality of image representations, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA, 2019), pp. 3987–3996.

[35] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, The intrinsic dimension of images and its impact on learning, arXiv:2104.08894.

[36] R. Verdel, V. Vitale, R. K. Panda, E. D. Donkor, A. Rodriguez, S. Lannig, Y. Deller, H. Strobel, M. K. Oberthaler, and M. Dalmonte, Data-driven discovery of relevant information in quantum simulators, Phys. Rev. B **109**, 075152 (2024).

[37] T. Mendes-Santos, A. Angelone, A. Rodriguez, R. Fazio, and M. Dalmonte, Intrinsic dimension of path integrals: Data-mining quantum criticality and emergent simplicity, PRX Quantum **2**, 030332 (2021).

[38] X. Turkeshi, Measurement-induced criticality as a data-structure transition, Phys. Rev. B **106**, 144313 (2022).

[39] F. Camastra and A. Staiano, Intrinsic dimension estimation: Advances and open problems, Inf. Sci. **328**, 26 (2016).

[40] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, Intrinsic dimension estimation: Relevant techniques and a benchmark framework, Math. Probl. Eng. **2015**, 1 (2015).

[41] M. A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE J. **37**, 233 (1991).

[42] S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. **2**, 37 (1987).

[43] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, Estimating the intrinsic dimension of datasets by a minimal neighborhood information, Sci. Rep. **7**, 12140 (2017).

[44] M. Nakahara, *Geometry, Topology and Physics* (CRC Press London, 2018).

[45] We note, en passant, that this reasoning is loosely inspired by the one used in Ref. [27] to identify a topological contribution to entropy in quantum mechanical systems, with the key difference that here we are not referring to partitions but, rather, different manifolds.

[46] V. Berezinsky, Destruction of long range order in one-dimensional and two-dimensional systems having a continuous

symmetry group. I. Classical systems, Zh. Eksp. Teor. Fiz. **59**, 907 (1971) [Sov. Phys. JETP 32, 493 (1971)].

[47] J. M. Kosterlitz and D. J. Thouless, Ordering, metastability and phase transitions in two-dimensional systems, J. Phys. C **6**, 1181 (1973).

[48] A. W. Sandvik, Computational studies of quantum spin systems, in *AIP Conference Proceedings* (American Institute of Physics Vietri sul Mare, (Italy), 2010), Vol. 1297, pp. 135–338.

[49] M. Henkel, *Conformal Invariance and Critical Phenomena* (Springer Science & Business Media, Heidelberg, 1999).

[50] Let us remind the reader that the Euclidean distance between two configurations $\vec{\theta}^i, \vec{\theta}^j$ reads [19]

$$d(\vec{\theta}^i, \vec{\theta}^j) = \sqrt{2\sum_{k=1}^{N}\left(1 - \vec{S}_k^i \vec{S}_k^j\right)}. \qquad \text{(C1)}$$

.

[51] We note that this picture might be an interesting starting point to investigate the manifold beyond the 2-NN estimator we utilize, e.g., considering fixed-length scale estimators. I. Macocco, A. Glielmo, J. Grilli, and A. Laio, Intrinsic dimension estimation for discrete metrics, Phys. Rev. Lett. **130**, 067401 (2023).

[52] U. Wolff, Collective Monte Carlo updating for spin systems, Phys. Rev. Lett. **62**, 361 (1989).

[53] It might be possible to refine this ansatz incorporating eventual anisotropies in the geometry of the system. However, for the system sizes we are interested in, it would be challenging to make concrete claims (due to the challenging scaling of our I_d estimators versus the embedding dimension), so we leave this aspect to future investigations.

[54] J. F. Rodriguez-Nieva and M. S. Scheurer, Identifying topological order through unsupervised machine learning, Nat. Phys. **15**, 790 (2019).

[55] Q. H. Tran, M. Chen, and Y. Hasegawa, Topological persistence machine of phase transitions, Phys. Rev. E **103**, 052127 (2021).

[56] N. Sale, J. Giansiracusa, and B. Lucini, Quantitative analysis of phase transitions in two-dimensional *xy* models using persistent homology, Phys. Rev. E **105**, 024121 (2022).

[57] H. Otsuka, K. Shiina, and Y. Okabe, Comprehensive studies on the universality of BKT transitions–machine-learning study, Monte Carlo simulation, and level-spectroscopy method, J. Phys. A: Math. Theor. **56**, 235001 (2023).

[58] O. Boada, A. Celi, J. I. Latorre, and M. Lewenstein, Quantum simulation of an extra dimension, Phys. Rev. Lett. **108**, 133001 (2012).

[59] O. Boada, A. Celi, J. Rodríguez-Laguna, J. I. Latorre, and M. Lewenstein, Quantum simulation of non-trivial topology, New J. Phys. **17**, 045007 (2015).

[60] D. Bluvstein, H. Levine, G. Semeghini, T. T. Wang, S. Ebadi, M. Kalinowski, A. Keesling, N. Maskara, H. Pichler, M. Greiner *et al.*, A quantum processor based on coherent transport of entangled atom arrays, Nature (London) **604**, 451 (2022).

[61] F. Nogrette, H. Labuhn, S. Ravets, D. Barredo, L. Béguin, A. Vernier, T. Lahaye, and A. Browaeys, Single-atom trapping in holographic 2D arrays of microtraps with arbitrary geometries, Phys. Rev. X **4**, 021034 (2014).

[62] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J.-G. Liu, R. Samajdar *et al.*, Quantum optimization of maximum independent set using Rydberg atom arrays, Science **376**, 1209 (2022).

[63] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, Nat. Rev. Phys. **5**, 9 (2023).