Unsupervised Learning Universal Critical Behavior via the Intrinsic Dimension

T. Mendes-Santos,[1,*] X. Turkeshi,[1,2,3,*] M. Dalmonte,[1,2] and Alex Rodriguez[1]

[1]*The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy*
[2]*SISSA, via Bonomea, 265, 34136 Trieste, Italy*
[3]*INFN, via Bonomea, 265, 34136 Trieste, Italy*

(Received 3 July 2020; revised 9 November 2020; accepted 15 December 2020; published 26 February 2021)

The identification of universal properties from minimally processed data sets is one goal of machine learning techniques applied to statistical physics. Here, we study how the minimum number of variables needed to accurately describe the important features of a data set—the intrinsic dimension (I_d)—behaves in the vicinity of phase transitions. We employ state-of-the-art nearest-neighbors-based I_d estimators to compute the I_d of raw Monte Carlo thermal configurations across different phase transitions: first-order, second-order, and Berezinskii-Kosterlitz-Thouless. For all the considered cases, we find that the I_d uniquely characterizes the transition regime. The finite-size analysis of the I_d allows us to not only identify critical points with an accuracy comparable to methods that rely on *a priori* identification of order parameters but also to determine the corresponding (critical) exponent ν in the case of continuous transitions. For the case of topological transitions, this analysis overcomes the reported limitations affecting other unsupervised learning methods. Our work reveals how raw data sets display unique signatures of universal behavior in the absence of any dimensional reduction scheme and suggest direct parallelism between conventional order parameters in real space and the intrinsic dimension in the data space.

DOI: 10.1103/PhysRevX.11.011040

Subject Areas: Condensed Matter Physics, Statistical Physics

I. INTRODUCTION

The growing field of machine learning (ML) is rapidly expanding our capabilities of analyzing and describing high-dimensional data sets [1–4]. With the increasing understanding of these methods, the community is becoming convinced that their outstanding performance is mostly due to the fact that this "high dimensionality" is applicable only to the embedding space, while the data sets lie in a manifold that can be twisted and topologically complex but whose intrinsic dimension I_d is typically much smaller than the large number of coordinates of the system [5,6] [see Fig. 1(a)]. The determination of this I_d is an active field of research [5,7,8] in unsupervised learning (UL), i.e., the branch of machine learning that aims to uncover the internal structure of a data set without the need for any label.

Recently, ML ideas have encountered fruitful applications in the context of statistical physics [9–11]. Such applications have ranged from the determination of physical properties [12–20] to the formulation of novel classes of variational ansätze [21–24]. These applications focused on analyzing and exploiting the results of dimensional reduction and using a variety of tools to analyze (or employ) the final representation (or truncation) obtained in this way. In various contexts, results obtained via these methods are, remarkably, already competitive with more traditional approaches [9].

Here, we pursue an alternative approach: Our main purpose is to show that, from a ML perspective, physically relevant and universal information can be gathered by analyzing the very same embedding procedure that carries out the dimensional reduction, rather than focusing on its final result. In particular, we show how the intrinsic dimension corresponding to the partition function of statistical mechanics models displays universal scaling behavior in the vicinity of phase transitions, and how it behaves as an order parameter for a corresponding structural transition in data space. In contrast to previous works [12,25–31], our approach focuses on data mining the data set as a whole; thus, it does not focus on any kind of projection. At the technical level, we employ a cutting-edge nearest-neighbor estimator of the I_d, which is suitably designed to deal with nonlinear data sets, i.e., data sets lying on nonlinear manifolds [8].

In order to access the complex data structure at phase boundaries, we numerically study instances of first-order,

[*]These authors contributed equally to this work.

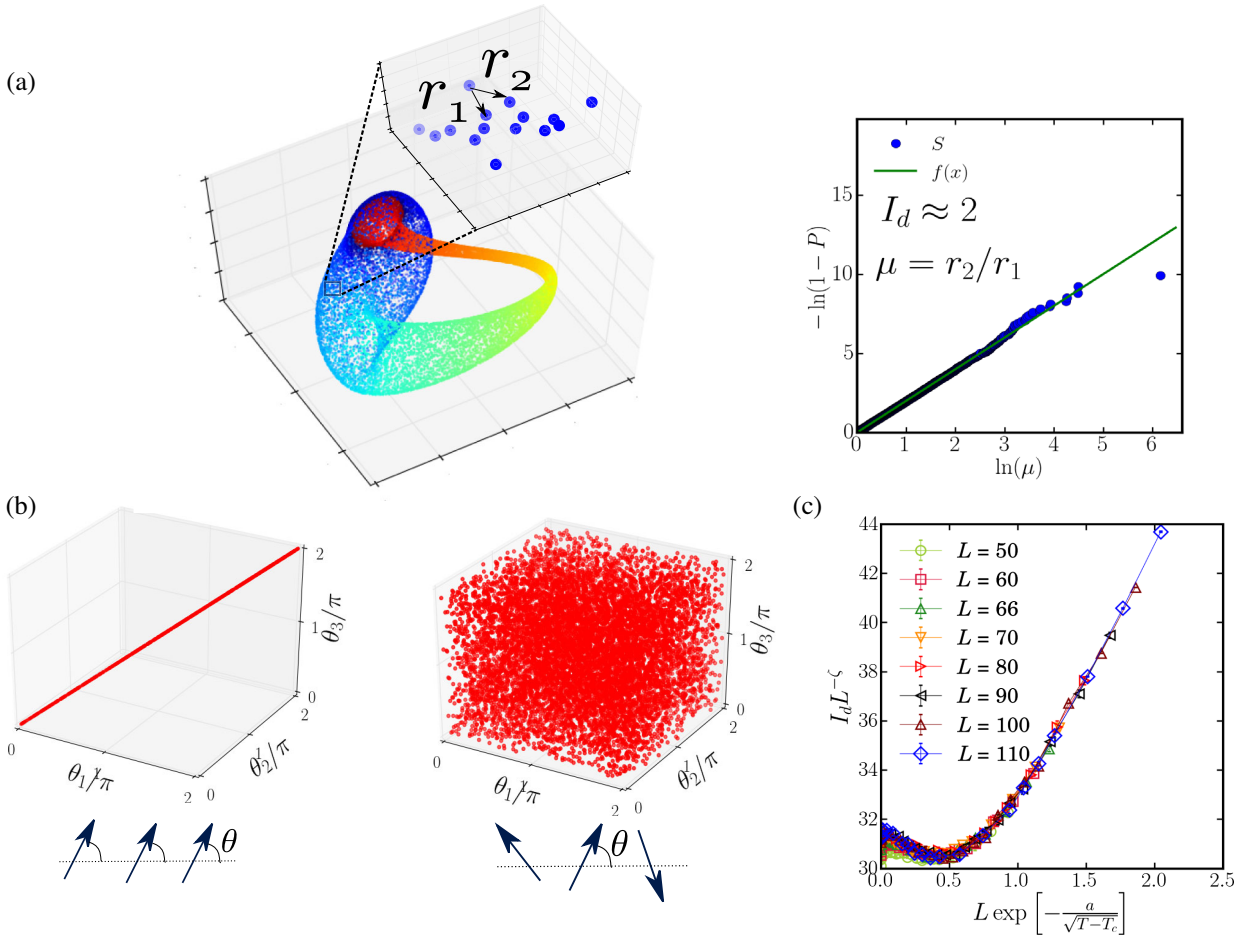

FIG. 1. (a) Schematics of the intrinsic dimension I_d. The important content of a data set typically lies within a manifold whose I_d is much lower than the number of coordinates. In the example, despite the fact that the synthetic data set (Klein's bottle-shaped data set) is embedded in a three-dimensional space, it can be effectively described by a twisted manifold whose $I_d = 2$. The key ingredients to compute the I_d are the first- and the second-nearest-neighbor distances, r_1 and r_2, of each point of the data set. The computation of the I_d is based on the fitting of the empirical, cumulative distribution function (CDF) of the ratio $\mu = r_2/r_1$, $P(\mu)$ [see text and Eq. (2)]. (b) Low- and high-temperature data sets of a three-site model in configuration space. The points represent the three-site XY model configurations: $\vec{\theta} = (\theta_1, \theta_2, \theta_3)$. The high- and zero-temperature cases show simple data structures: For $T = 100$, I_d is equal to the number of spins, while for $T = 0$, $I_d = 1$. (c) Intrinsic dimension in the vicinity of a phase transition. The I_d in the intermediate-temperature regime, representative of phase transitions in larger systems, is considerably more complex. The temperature dependence of I_d can be used to locate and characterize critical points. As an example, we show the universal data collapse of the I_d at the Berezinskii-Kosterlitz-Thouless transition described by the 2D XY model.

second-order (conformal), and Berezinskii-Kosterlitz-Thouless (BKT) transitions in two-dimensional (2D) classical spin systems. In all cases, I_d displays a universal scaling behavior corresponding to the transition properties of the underlying lattice model. (i) For first-order transitions, I_d peaks at the critical point due to the coexistence of different orders, and the finite-size corrections of the transition temperature are dictated by trivial scaling exponents. (ii) For both second-order and topological transitions, we observe universal scaling collapse, with transition temperature and critical exponents determined to the percent level. (iii) Most importantly, we provide compelling evidence that the I_d is an ideal tool to underpin topological transitions in an unsupervised fashion: As an example, we extract the critical temperature of the 2D XY model with 1% confidence even at modest system sizes.

We then develop a theoretical framework in support of the fact that I_d has characteristic features at transition points, which are governed by scaling theory. First, we show how several instances of the data set, in particular, the distribution of distances between sampled configurations, already reveal striking features about critical behavior for all classes of phase transitions. The basic idea is that the data space naturally clusters configurations characterized by similar physical properties (e.g., magnetization and winding number). The fact that the intrinsic dimension

has strong features at the phase transition then follows from its "local" nature (related to changes of scale in configuration space, as we specify below). Then, we discuss how, for the type of data sets we are interested in, the intrinsic dimension can be connected directly to a set of arbitrary many-body correlation functions, which, following the finite-size scaling hypothesis, justifies its scaling behavior in the vicinity of transition points. We then check *a posteriori* the validity of some of the assumptions at the basis of this framework.

Before diving into the main part of our paper, we provide a simplified picture that qualitatively captures how the intrinsic dimension is connected to the physical information obtained by sampling a partition function via Monte Carlo (MC) methods. The basic intuition behind the I_d of data sets generated in the low- and high-temperature regimes of a simple three-site XY model is shown in Fig. 1(b). At low temperature [left graphic of Fig. 1(b)], most of the spin configurations sampled during the Markov chain correspond to fully ferromagnetic spin arrangements [see the cartoon of Figs. 1(b) and 1(c), respectively]. In the limiting case $T = 0$, the ground states are given by XY ferromagnetic configurations, i.e., $\theta_1 = \theta_2 = \theta_3$, and the data set is described by a manifold that lies in a line ($I_d = 1$). In contrast, in the high-temperature regime, the data set is described by a manifold with $I_d = 3$: Each new Monte Carlo configuration corresponds to an arbitrary arrangement of the three spins, so the structure of the data set is that of a homogeneously occupied three-dimensional space. This simple example demonstrates how transitions in parameter space are accompanied by structural transitions in data space. Because of its collective origin, the transition region requires the computation of I_d in very-high-dimensional data space: In Fig. 1(c), we show a sample of our results, illustrating the scaling collapse of the intrinsic dimension corresponding to the 2D XY model in the vicinity of its BKT transition point.

II. INTRINSIC DIMENSION

Before addressing the analysis of concrete statistical mechanics models, we present here a self-contained discussion on the intrinsic dimension and its state-of-the-art estimators. This section is propaedeutic to the critical identification of the best estimator to be used in our applications below.

The I_d is a concept that arises from the observation that, in natural data sets, the correlations between the input variables induce a structure, modifying the dimensionality of the manifold in which the data lie. In order to visualize this concept, one can imagine a data set with the Cartesian coordinates of points extracted from a circle. Although there are two input coordinates, they are strongly correlated, and the manifold in which the points lie has a $I_d = 1$. Therefore, in simple cases like this, or the one shown in Fig. 1(b), it roughly corresponds to the minimum number of variables needed to describe a data set [7,8].

Information about the I_d is important to determine if dimensional reduction of high-dimensional data sets results in information loss or not. Moreover, it can be used as a UL approach to characterize a system. Here, we mention a few examples: In biological physics, the I_d can be used to determine the number of independent directions a protein can have during a sequence evolution [32]; in image analysis, to distinguish between different kinds of image structures [33]; in astrophysics, to estimate the amount of information available in spectropolarimetric data [34]; in theoretical machine learning, to understand the properties of deep neural networks [35]; and in ecology, to characterize the minimum number of independent axes of variation that adequately describes the functional variation among plants [36].

Different approaches have been developed to estimate the I_d; see Ref. [7] for a review. For example, dimensional reduction techniques—such as principal component analysis (PCA) [37], multidimensional scaling [38], Isomap [39], locally linear embedding [40], autoencoders [41], t-distributed stochastic neighbor embedding (t-SNE) [42], or uniform manifold approximation and projection (UMAP) [43] to mention a few—search for a lower-dimensional space to project the data set by minimizing a projection error. The dimension of the identified subspace is viewed as an estimation of I_d. However, identifying this dimension is far from trivial. For instance, in the PCA case, one should take into consideration the spectrum of the eigenvalues of the covariance matrix and either look for a gap or choose *ad hoc* a cutoff parameter. It is worth saying that, for PCA, this strategy will not work if the manifold of lower dimensionality is curved. Furthermore, some of the above-mentioned methods, like t-SNE, are focused on visualization and assume that the dimension of the projection space is lower than the I_d. These projection algorithms aim to alleviate the problems that this dimension mismatch causes in the visualization and, therefore, are not well suited for I_d detection.

A closely related quantity is the fractal dimension [44], whose estimation relies on the scaling of the number of neighbors with the distance from a given point. This approach is largely employed in the study of percolation transitions [45], but it suffers from serious limitations when the density distribution of points is not uniform.

These limitations lead to the development of nearest-neighbor methods, in which it is assumed that nearest-neighborhood points can be considered as uniformly drawn from small enough I_d-dimensional hyperspheres (not all the data set) [5,7]. Indeed, the avoidance of any projection step and the relaxing on the condition of data uniformity (from the full data set to a small neighborhood around each point) are key features for obtaining good results in highly nonuniform, nonlinear data sets, even in really high

dimensions (a regime at which all the purely geometrical methods present a bias due to the curse of dimensionality).

The two nearest-neighbor (two-NN) method employed in this work belongs to this type of methods, with the particularity that by focusing only on the first two nearest neighbors [see Fig. 1(a)], the size of the I_d-dimensional hyperspheres at which the density is assumed to be constant is reduced to its minimum expression. The method is rooted in computing the distribution functions of neighborhood distances, which are functions of I_d. More specifically, for each point $\vec{x}$ in the data set, we consider its first and second nearest-neighbor distances $r_1(\vec{x})$ and $r_2(\vec{x})$, respectively. Under the condition that the data set is locally uniform in the range of second nearest neighbor, it has been shown in Ref. [8] that the distribution function of $\mu = r_2(\vec{x})/r_1(\vec{x})$ is

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

or, in terms of the cumulative distribution, $P(\mu)$,

$$I_d = -\frac{\ln[1 - P(\mu)]}{\ln(\mu)}, \tag{2}$$

which can be used to obtain I_d by fitting $S = \{(\ln(\mu), -\ln[1 - P^{\text{emp}}(\mu)]\}$ with a straight line passing through the origin. The function P^{emp} defines the empirical cumulate and is computed by sorting the values of μ in ascending order (see the Appendix A for more details). In Fig. 1(a), the steps for computing the I_d in a highly nonlinear manifold with complex topology (in this case, a Klein's bottle-shaped data set) are summarized: (1) Compute the distance from the first and second neighbors, (2) compute for each point μ and its empirical cumulate, and (3) fit S to a straight line.

We stress that this method is not free of drawbacks. As mentioned above, since it is a purely geometrical method, it is affected by the curse of dimensionality because the number of points needed to have an accurate measurement of the I_d grows exponentially with the I_d. Moreover, Eq. (2) was derived assuming a continuous real support. Therefore, applying it to data sets with a different support implies some degree of approximation that can fail in some limiting cases. For instance, this failure occurs when two or more configurations have the same coordinates. However, as we detail below, these drawbacks do not affect the results obtained in this work; in particular, these limitations do not kick in when investigating transitions, even when configuration spaces are composed of discrete variables such as Ising spins. These limitations only affect data sets corresponding to either very small system sizes or phases at extremely low temperatures where, during the MC sampling, configurations may be repeated, as the accessible configuration space is very limited.

III. MODELS

Our approach focuses on the high-dimensional data sets associated with the equilibrium configuration states of a partition function. Such states are sampled with Markov chain Monte Carlo simulations from the thermal weight $\rho(E) \sim e^{-E(\vec{x})/T}$, where $E(\vec{x})$ is the energy of an independent configuration $\vec{x}$ and T is the temperature. We employ Wolff's cluster algorithm [46,47], and for each data set, we consider N_r configurations.

We consider partition-function data sets of several models in the vicinity of various types of phase transitions [48,49]. The first example is the well-known Ising model in two dimensions:

$$E(\vec{s}) = -\sum_{\langle i,j \rangle} s_i s_j, \tag{3}$$

where the spin degrees of freedom are $s_i = \pm 1$, and $\langle i, j \rangle$ are the nearest-neighboring bonds of a square lattice, with $N_s = L \times L$ spins and periodic boundary condition. The Ising configuration states are defined as

$$\vec{s} = (s_1, s_2, \ldots, s_{N_s}). \tag{4}$$

This model describes a second-order phase transition characterized by the breaking of Z_2 symmetry at the critical temperature $T_c = 2/\ln(1 + \sqrt{2})$. In the vicinity of T_c, the spin correlation length diverges as $\xi \sim (T - T_c)^{-\nu}$, where the critical exponent is $\nu = 1$.

We also consider the first- and second-order phase transitions described by the q-state Potts model (qPM),

$$E(\vec{\sigma}) = -\sum_{\langle i,j \rangle} \delta_{\sigma_i, \sigma_j}, \tag{5}$$

where the spin $\sigma_i = 0, 1, 2, \ldots, q - 1$, and $\delta_{\sigma_i, \sigma_j}$ is the delta function. In particular, the $q = 2$ Potts model can be mapped into the Ising model. The Potts configuration states are defined by

$$\vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{N_s}). \tag{6}$$

The qPM is characterized by a discrete Z_q symmetry that is broken at the critical temperature $T_c = 1/\ln(1 + \sqrt{q})$. Importantly, this class of models displays a second-order phase transition for $q \leq 4$ and a first-order one for $q > 4$. We examine both of these regimes: the second-order transition described by the $q = 3$ PM (with correlation length critical exponent $\nu = 4/5$) and the first-order transition described by the $q = 8$ PM [50,51].

Finally, as a representative of the BKT universality class, we investigate the two-dimensional XY model [52–54]

$$E(\vec{\theta}) = -\sum_{\langle i,j \rangle} \vec{S}_i \vec{S}_j, \tag{7}$$

where $\vec{S}_i = (\cos(\theta_i), \sin(\theta_i))$, with $\cos(\theta_i)$ and $\sin(\theta_i)$ being the projection of the spin at site i in the x and y directions, respectively, and $\theta_i \in [0, 2\pi[$. The XY configurations are defined as

$$\vec{\theta} = [\cos(\theta_1), \sin(\theta_1), \ldots, \cos(\theta_{N_s}), \sin(\theta_{N_s})]. \tag{8}$$

This model is characterized by a continuous $U(1)$ symmetry, and it describes a phase transition between a high-temperature phase with exponentially decaying spin correlations and a low-temperature quasiordered phase characterized by power-law decaying correlations. The BKT critical temperature T_{BKT} is not known exactly; state-of-the-art estimations based on the analysis of the spin stiffness of lattices of order $O(10^6)$ spins give $T_{\mathrm{BKT}} = 0.8935(1)$ [55].

The detection of the BKT critical point is hindered by the fact that it cannot be characterized by conventional local order parameters, as in the examples discussed previously, and because of the exponential growth of the correlation length near T_{BKT}. Hence, the BKT transition represents a key challenge for any UL method.

A. How to characterize partition functions as data sets

Before proceeding to the discussion of the results, we point out some important aspects of the Ising, Potts, and XY data sets [see Eqs. (4), (6), and (8), respectively]. First, a crucial step to obtain the I_d [cf. Eq. (2)] is to consider a proper metric; the distance $r(\vec{x^i}, \vec{x^j})$ between two configuration states $\vec{x^i}$ and $\vec{x^j}$ must be non-negative, equal to zero only for identical configurations, and symmetric, and, it must also satisfy the triangular inequality.

For the XY data sets, the distance is defined as the Euclidean distance:

$$r(\vec{\theta^i}, \vec{\theta^j}) = \sqrt{2 \sum_{k=1}^{N_s} \left(1 - \vec{S}_k^i \vec{S}_k^j\right)}. \tag{9}$$

This distance properly takes into account the periodicity of the configuration states in the interval $\theta_i \in [0, 2\pi[$.

For both Ising and Potts configuration states, we consider the Hamming distance; i.e., $r(\vec{s^i}, \vec{s^j})$ [or $r(\vec{\sigma^i}, \vec{\sigma^j})$] is given by the number of positions in the state vectors ($\vec{s^i}$ and $\vec{s^j}$) for which the corresponding coordinates are different. The choice of the Hamming distance is motivated by the fact that the energy difference between two spins in the model of interest is given by a delta function.

As mentioned in the previous section, the two-NN method fails when two or more sampled configurations of the data set have identical coordinates. This issue typically occurs in the discrete-variable Ising and Potts data sets, when the total number of independent configuration states, N_c, is smaller or of the same order of the number of configurations used in the data set, N_r. For instance, for both Ising and Potts data sets, identical ferromagnetic configurations are sampled in most of the Monte Carlo steps when $T \ll T_c$. However, in the regime that we focus on here (i.e., T close to T_c and $L > 10$), as $N_c \gg N_r$, this issue is irrelevant, which we have explicitly checked in our data sets.

Finally, we mention that data sets generated by the XY configuration states typically lie in nonlinear manifolds, which can be noted by the fact that linear-dimension-reduction methods, such as PCA, fail to describe XY data sets; see Ref. [29]. In fact, even for the simple data set shown in Fig. 1(b), linear PCA fails in estimating the true I_d of the system when the proper distance between the configurations is taken into account; see Appendix A. This feature of the XY data sets reveals the necessity of using state-of-the-art I_d estimators (such as the two-NN method considered here) that properly takes into account nonlinearities.

IV. RESULTS

A. Second-order phase transitions

We start our discussion by considering second-order phase transitions (2PTs) described by the Ising and the three-state Potts models (3PMs); see Fig. 2. We consider data sets formed by $N_r = 5 \times 10^4$ configuration states. Overall, far from the transition, I_d is an increasing function of T. For low T, the computation of I_d is affected by the discreteness of the Potts (Ising) configurations, which reflects on the larger error bars. However, this issue is mitigated close to the critical point T_c. Remarkably, I_d exhibits a nonmonotonic behavior in the vicinity of the critical point [see Figs. 2(a1) and 2(b1)], which can be used to locate and characterize the transition point itself.

As conventionally done in the analyses of physical observables, e.g., magnetic susceptibility and heat capacity, we now consider a finite-size scaling (FSS) theory for I_d. First, based on the FSS hypothesis and postulating that I_d behaves as an order parameter for the transition, one has $I_d = L^{\zeta} f(\xi/L)$, where the correlation length diverges as $\xi \sim (T - T_c)^{-\nu}$, ν is a critical exponent, and ζ is a scaling exponent associated with the divergence of I_d at T_c. Figures 2(a2) and 2(b2) show the universal data collapse for the Ising and 3PM models, respectively. The values obtained for T_c and ν have a discrepancy with exact results of less than 0.5% and 4%, respectively. For the Ising model, we obtain $T_c = 2.283(2)$, $\nu = 1.02(2)$, and $\zeta = 0.410(5)$, while for the 3PM model, we get $T_c = 0.996(2)$,

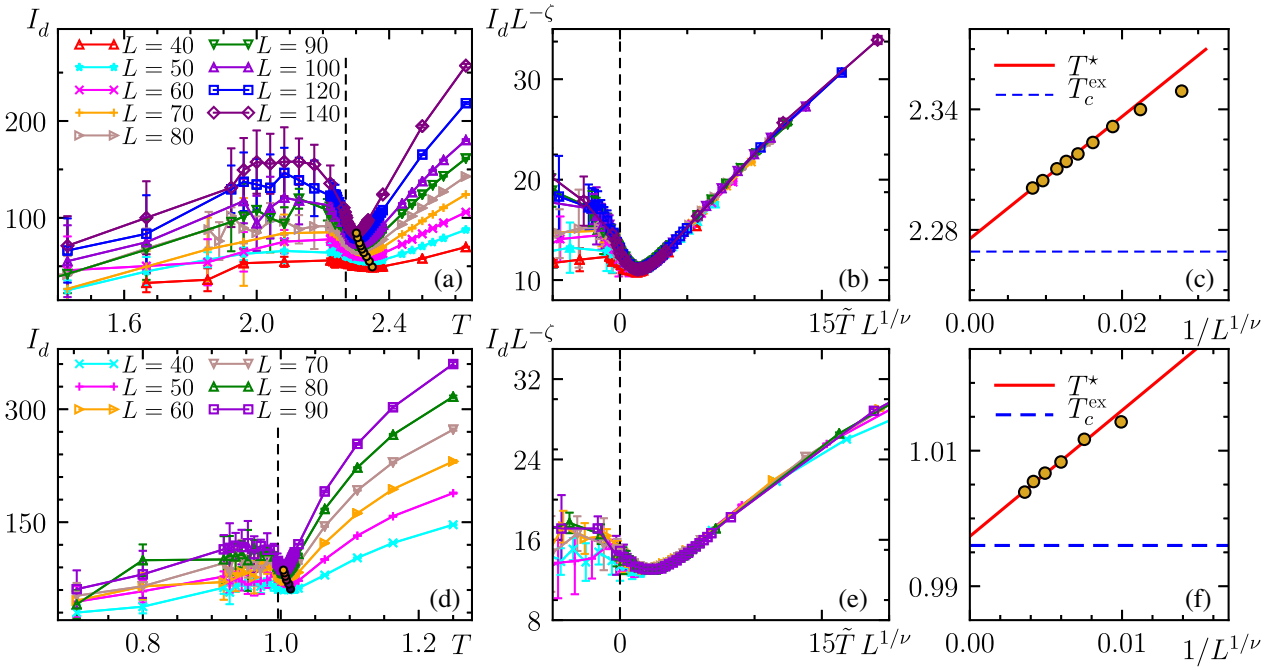

FIG. 2. Second-order phase transition. (a)–(c) Ising model. (d)–(f) $q = 3$ Potts model. (a,d) I_d as a function of T. Error bars are standard deviations associated with a distribution of n realizations of I_d (we typically consider $n \geq 5$). (b,e) Data collapse of I_d based on the finite-size scaling discussed in the main text. The best data collapse of the results gives $T_c = 2.283(2)$, $\nu = 1.02(2)$, and $\zeta = 0.410(5)$ for the case of the Ising model, and $T_c = 0.996(2)$, $\nu = 0.805(5)$, and $\zeta = 0.420(2)$ for the case of the Potts model. (c,f) finite-size scaling of the minimum temperature T^* (see text); the horizontal line is the exact result for T_c. The extrapolation returns $T_c = 2.2784(2)$ and $T_c = 0.9970(3)$ for the Ising and Potts cases, respectively.

$\nu = 0.805(5)$, and $\zeta = 0.420(2)$. See Appendix C, for a discussion about the details of the data-collapse procedure.

Furthermore, we consider the size scaling of the shift of the local minimum of $I_d(T)$ [i.e., the temperature $T^*(L)$],

$$T^*(L) - T_c \sim \frac{1}{L^{1/\nu}}. \tag{10}$$

We note that $T^*(L)$ is reminiscent of the universal scaling behavior of singular features of physical observables close to T_c (e.g., the peak of the magnetic susceptibility) [56]. In order to compute T_c, we employ the following procedure: (i) We obtain $T^*(L)$ by fitting the results in an interval close to $T^*(L)$ with a cubic function; the fitting is performed with a jackknife procedure, which allows us to establish an error bar for $T^*(L)$. (ii) We then consider the aforementioned FSS to compute T_c; the fitting is performed by considering different sets of points. This method provides a coherent error propagation for T_c. We obtain $T_c = 2.2784(2)$ for the Ising model and 0.9970(3) for the three-state Potts model. Their discrepancies with the exact values are, respectively, of order 0.4% and 0.2%.

The results of this analysis confirm the validity of our original assumption, that is, that the intrinsic dimension is a valid order parameter describing the transition in data space as a structural transition. We remark that this is validated in two steps—first, via the quality of the scaling collapse and, second, by the scaling of the transition temperature obtained by analysis a single feature of the I_d dependence with respect to the temperature. These steps represent two fundamental tests that any valid order parameter must satisfy at the transition points.

B. BKT phase transition

Unsupervised learning of phase transitions associated with symmetry breaking, as discussed in the previous section, can also be performed with other unsupervised methods, such as the PCA [12,27,57] and variational autoencoder (VAE) [25]. For example, the critical temperature of the Ising model can be obtained with an accuracy similar to the ones obtained here. Furthermore, the latent parameters of the VAE can be used to learn the local order parameter associated with both discrete and continuous symmetry-breaking transitions [25]. These methods are based on a dimension reduction and thus differ in a fundamental way from our unsupervised approach, which is based solely on the analysis of the I_d. As we discuss in this section, our approach can be extended to transitions that are characterized by nonlocal order parameters, such as the topological BKT phase transition, which are treated on the same footing as second-order transitions.

The difficulties in learning the BKT transition from raw XY configurations occur in both supervised [58] and

unsupervised [27] ML approaches. Recent progress based on diffusion maps [29,59] or topological data analysis [60] have been made to solve this problem, typically considering problem-specific insights (such as the structure of topological excitations). These approaches have shown how considerable qualitative insight can be gathered on the nature of the BKT transition. However, it is presently unclear if the raw data structure corresponding to topological transitions can exhibit universal features and, if so, whether unsupervised approaches can be used to detect the critical temperature with an accuracy that is comparable to conventional methods (that typically rely on the *a priori* knowledge of the order parameter).

In Fig. 3(a), we show the temperature dependence of I_d in the transition region. The intrinsic dimension clearly distinguishes the low-T regime, characterized by bound vortex-antivortex pairs, from the unbinding high-T regime. In the vicinity of the BKT critical point T_{BKT}, the behavior of I_d resembles the one observed for the second-order phase transitions; i.e., I_d exhibits a local minimum at $T^*(L)$ (observed for $L > 30$), which is a signature of the BKT transition. Note that the minimum is clearly visible already for lattices of order $L = 50$; at these sizes, the spin stiffness instead features a very smooth behavior, as considerably larger systems are required to appreciate a qualitative jump in the latter.

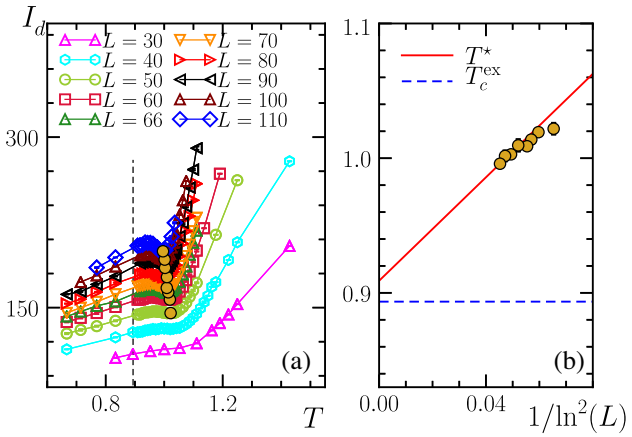

FIG. 3. Berezinskii-Kosterlitz-Thouless transition. Panel (a) shows the temperature dependence of I_d for different values of L. The dashed line indicates the value of the BKT critical temperature obtained in Ref. [55] using conventional methods, $T_{\mathrm{BKT}} = 0.8935(1)$. For each point, we harvested approximated ten instances of the data set and averaged the resulting estimates for the I_d. The error bars are the standard deviation of such results. The scaling collapse obtained with these data sets is depicted in Fig. 1(c). Panel (b) shows the finite-size scaling of T^* based on Eq. (11). Fitting the results for $L = 80$, 90, 100, and 110, we obtain $T_{\mathrm{BKT}} = 0.909 \pm 0.015$. In the text, we discuss how we obtain the local minimum of I_d, T^*. We compute the I_d of manifolds with $N_r = 5 \times 10^4$ configurations.

We consider the conventional FSS for the BKT transition, $I_d(T, L) = L^{\zeta} f(\xi(T)/L)$, where the singular value of the correlation length diverges exponentially, i.e., $\xi \sim \exp(a/\sqrt{T - T_c})$. In Fig. 1(c), we show the universal data collapse for different values of L, where a, T_{BKT}, and ζ are treated as free parameters in the collapse procedure; see Appendix C. The value obtained, $T_{\mathrm{BKT}} = 0.92(1)$, is in good agreement with estimations of T_{BKT} obtained in Ref. [56].

A more accurate estimation of T_{BKT} is based on the finite-size scaling of $T^*(L)$. This approach relies on the computation of $T^*(L)$, which is performed with the same procedure described in the previous section, and the finite-size scaling ansatz [56]:

$$T^*(L) - T_{\mathrm{BKT}} \sim \frac{1}{\ln^2 L}. \tag{11}$$

As discussed before, this procedure allows us to establish an error bar for the calculated T_{BKT}. We obtain $T_{\mathrm{BKT}} = 0.909 \pm 0.015$, which is compatible, within error bars, with Ref. [55], where simulations with up to $O(10^6)$ spins were carried out. For comparison, the best alternative method [29] utilizing unsupervised learning techniques reported relative errors of the order of 5%.

Conventionally, T_{BKT} is obtained with the aid of the so-called Nelson-Kosterlitz universal jump of the spin wave stiffness [61], which allows us to determine the finite-site critical temperature $T^*_{\mathrm{BKT}}(L)$. The FSS [Eq. (11)] is then used to determine the BKT critical point at the thermodynamic limit. Remarkably, here we observe that the intrinsic dimensions of raw XY data sets exhibit a clear signature of the finite-site T_{BKT}, even for the moderate system sizes we have considered.

C. First-order phase transitions

Finally, we consider an example of a first-order phase transition (1PT): the eight-state Potts model (8PM). As is typical of 1PT, the system exhibits a finite-size correlation length at T_c, $\xi_8 = 23.9$ [62]. For $L > \xi_8$, the transition can be described by trivial and generic critical exponents, e.g., $\nu = 1/d$, with d being the system dimension [51,63,64]. Furthermore, the finite-size shift of the critical temperature, $T_c(L)$—conventionally detected, for example, by the maximum value of the magnetic susceptibility—scales as $T_c(L) - T_c \sim 1/L^d$.

Figure 4 shows that I_d also exhibits a clear signature of the 1PT, featured by a peak at T_c for $L \gg \xi_8$. For $L \approx \xi_8$, the temperature dependence of I_d resembles the one observed for 2PTs in Fig. 2; i.e., I_d exhibits a local minimum at a temperature T^*. Interesting, the FSS of T^* is in agreement with first-order transitions [see Fig. 4(b)] [63,64]; the discrepancy of the calculated $T_c = 0.7448(1)$ from the exact value is less the 0.05%.

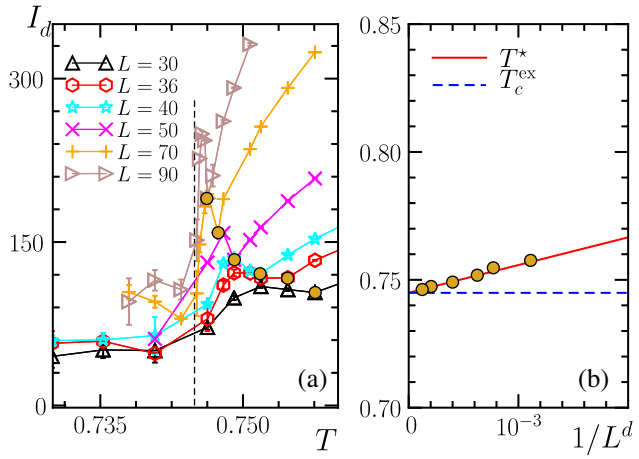

FIG. 4. First-order phase transition. Panel (a) shows I_d as a function of T for the eight-state Potts model. Panel (b) shows the finite-size scaling of T^*, where $d = 2$ (see text).

V. DISCUSSION

So far, our results support the fact that, in the vicinity of a phase transition, the intrinsic dimension displays universal behavior at first-order, second-order, and BKT transitions, and works as an order parameter signaling a transition between different data structures in configuration space. Within this framework, the position of the transition is always identified with the scaling of the minimum of the intrinsic dimension.

In continuous phase transitions, the collective behavior is captured by only a handful of parameters, which suggests that the amount of information required to describe the system is parametrically simpler at the critical point when compared to its vicinity, as the latter region requires additional information on the operators required to perturb away from criticality. This emergent simplicity may have several consequences at the data structure level. The most direct consequence is that one expects a simplified data structure to be described by a minimum of the intrinsic dimension at the transition point, which is exactly what we have observed at both second-order and BKT transitions. We note that this expectation is not related to the number of states sampled by the partition function (this number is, in our case, fixed by N_r, and configurations are never repeated). The discussion of how our results change with N_r is reported in Appendix B.

For first-order transitions, the above reasoning is not applicable as it relies on universal behavior and, thus, the existence of a continuum limit. In these cases, one expects that the data space in the vicinity of the transition point will feature two separate regions, each of them composed of states representing the two phases meeting at T_c. Exactly at the transition point, one expects an abrupt change in the data structure: Indeed, the MC sampling will access a large number of configurations corresponding to both phases (in analogy to metastability) and thus display a sharp increase [see Fig. 4(a)].

The arguments above serve as a qualitative guideline behind the basic picture we put forward: The simplified field theory description applicable at transition points reflects directly into the data structure of the problem. We now provide a data-driven discussion in support of this picture, specifically emphasizing the connection between the data set and correlations in the system via the (generic) definition of distance that we employ. For the sake of concreteness, we first elaborate on the presence of distinctive features in the vicinity of T_c and then make the connection to universal scaling.

A. Why the I_d exhibits a singular behavior in the vicinity of T_c

The I_d is a scale-dependent quantity [7], which can be intuitively understood by looking at the example depicted in Fig. 5, where an approximately one-dimensional object appears as two dimensional when looking at a different scale by zooming in. The scale of the data set, as estimated with the two-NN model, is fixed by N_r for a given T since it fixes the actual meaning of the first and second nearest neighbors [8]; here, we always consider $N_r = 5 \times 10^4$. In the following, we will show how changes in the scale of the data (configuration) space appear when there is a phase transition, leading to the emergence of features in the I_d.

A first test is to check that these changes in the scale effectively occur. To this end, we analyze the statistics of r_1 and r_2. For example, the distribution function of the first neighbor distances, $f(r_1)$, changes for both Ising and BKT critical points; see Figs. 6(a1) and 6(b1). The position of the peak of $f(r_1)$ sharply decreases as one crosses the transition, and the variance associated with $f(r_1)$, Δr_1, has

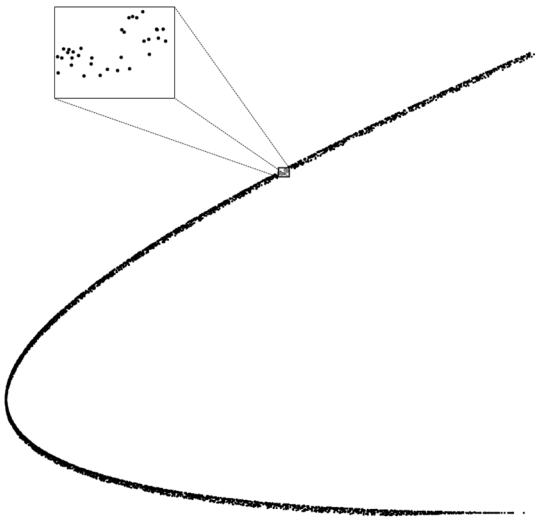

FIG. 5. Scale dependence of I_d. The data set shown presents an $I_d = 1$ or $I_d = 2$ depending on the scale that is considered.

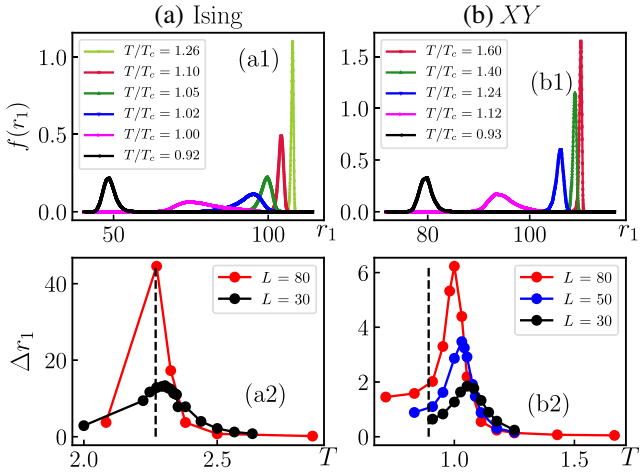

FIG. 6. Statistics of first nearest-neighbor distances, r_1. In panels (a1) and (b1), we show the kernel density estimation of the probability density function of r_1 for the Ising and XY data sets, respectively. All the results have $L = 80$. Panels (b1) and (b2) show the temperature dependence of the variance associated with the distribution functions from panel (a) for different system sizes. Results for the second nearest-neighbor distances, r_2, are qualitatively the same.

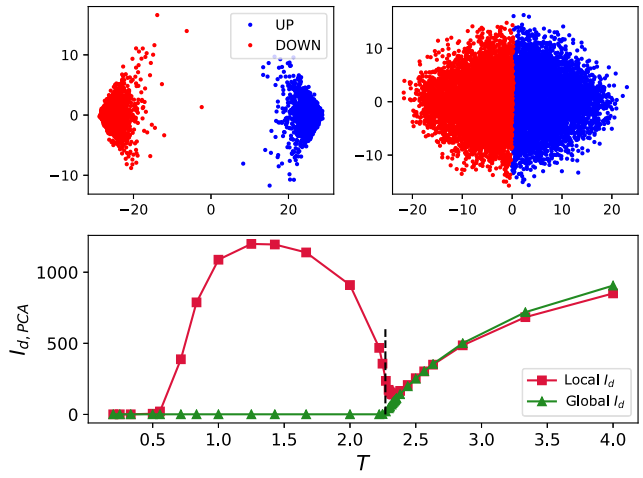

FIG. 7. I_d obtained with PCA. Panels (a1) and (a2) show the projection of the Ising data set in the two leading principal components for $T/T_c \approx 0.94$ and $T/T_c \approx 1.10$, respectively. Configurations with total magnetization $M > 0$ are represented by the blue points, while ones with $M \leq 0$ are shown by the red points. Panel (b) shows the PCA estimation of the I_d considering the full Ising data set (global I_d) and the data set generated by configurations with total magnetization $M > 0$ ("local" I_d). For all the results, $L = 60$.

a peak close to the transition; see Figs. 6(a2) and 6(b2). Interestingly, our results indicate that the quantity Δr_1 also exhibits universal scaling behavior at Ising and BKT critical points. Moreover, the results for both the Ising and XY data sets are qualitatively the same, highlighting the fact that data-wise, symmetry-breaking, and topological transitions are treated on the same footing. However, it is important to stress that, contrary to what happens in the case of I_d, the peak in Δr_1 is not expected to present features when the data sets are not homogeneous in density since the relevant distances will also be affected due to these inhomogeneities (see Ref. [8] for a further discussion on why the two-NN method is only mildly affected by this problem). In this sense, the intrinsic dimension, being solely sensitive to changes of scale and not local density features, provides a considerably more reliable probe for phase transitions.

In order to understand the underlying cause of this change of scale, we first focus on discrete symmetry-breaking transitions. In those cases, PCA provides an understanding of the data structure emerging at critical points [12,27]. For instance, the Ising data set features clusters characterized by configurations with positive, $+M$, and negative, $-M$, total magnetization for $T < T_c$. In contrast, a single cluster is formed for $T > T_c$; see Figs. 7(a1) and 7(a2). This clustering structure allows us to understand the connectivity between neighboring configurations in the Ising data set. For $T > T_c$, the magnetization of neighbors is completely random. In contrast, configurations connect to first and second neighbors with the same magnetization sign for $T < T_c$; see Fig. 8(a).

Equivalent reasoning based on PCA is applicable to the Potts data sets.

To illustrate how the locality (and the connectivity between neighboring configurations) affects the behavior of the I_d, we consider two estimates of the I_d provided by PCA. In the first case, we employ all the configurations of the Ising data set, $I_{d,\mathrm{PCA}}$ (global), while, in the second, we consider just configurations with $M > 0$, $I_{d,\mathrm{PCA}}$ (local);

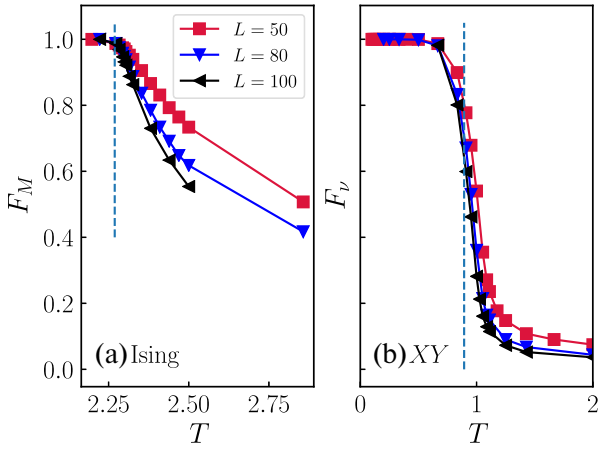

FIG. 8. Connectivity between neighboring points. Panel (a) shows the fraction of points in the Ising data set whose first two nearest neighbors have the same magnetization sign. Similarly, in panel (b), we show the fraction of points in the XY data set whose first two nearest neighbors have the same winding number (see text).

see Appendix A for more details. The latter quantity provides a local estimate (within the range scale of a single cluster) for $T < T_c$, which is analogous to the local measure of the I_d provided by the two-NN method. As shown in Fig. 7(b), the global $I_{d,\mathrm{PCA}}$ sharply goes to 1 below T_c, while the local $I_{d,\mathrm{PCA}}$ exhibits the same nonmonotonic behavior close to T_c that we observe in Fig. 2. This result highlights that the locality of the I_d is the crucial element to understand its nonmonotonic behavior close to T_c.

The connectivity between neighboring configurations is also related to the physical properties of the BKT transition. In this case, the most suitable quantity to characterize configurations and the corresponding clustering structure in phase space is the winding number w (see Appendix D), as excitations have topological (global) nature [29]. Above the BKT transition, vortex-antivortex pairs are unbounded. Therefore, the MC simulation samples configurations with different w. By contrast, in the quasi-long-range-order regime ($T < T_{\mathrm{BKT}}$), most of the configurations have $w = (0, 0)$. This feature of the BKT transition affects the connectivity between neighboring configurations. In particular, the fraction of configurations whose first two neighbors are connected to points with the same w, F_w, is negligible for $T > T_{\mathrm{BKT}}$, but it is equal to 1 in the topological phase, which is illustrated in Fig. 8(b). Thus, for $T < T_{\mathrm{BKT}}$, the I_d is a property of the manifold exclusively formed by configurations connected to neighbors with the same winding number.

In a nutshell, the underlying cause for the sensitivity of the I_d to phase transitions is that both the symmetry-breaking and topological transitions affect the neighboring configurations' connectivity. The key aspect is that nearest-neighbor configurations have identical physical properties (order parameter and winding number) when the system is in the ordered phases (symmetry-broken or quasi-long-range-order types). By contrast, in the disordered phase, the first and second neighbors' physical properties are entirely random. At the data structure level, the phase transition represents a change of scale between those regimes.

B. Why I_d exhibit universal scaling behavior

Based on the fact that I_d shows a characteristic minimum feature in the vicinity of phase transitions at $T_c(L)$, we now provide an argument in support of universal scaling of the latter temperature against the system size. The key aspect of our argument is that the distances r_1 and r_2 are related with many-body correlation functions in the system, computed at equilibrium.

We analyze the curve $\ln(1 - P_i)$ versus $\ln \mu_i$ close to the origin. From Eq. (2), the slope of this curve is proportional to I_d. The curve starts at the origin; we assume that its slope can be correctly determined by sampling the first point of the curve several times (e.g., by sampling several independent Markov chains); this seems very well satisfied based on our earlier numerical observations (fluctuations and deviations from linear behavior typically appear only for very large values of μ). Within this assumption, one obtains the following estimate for I_d:

$$I_d = -\frac{\ln(1 - 1/N_r)}{\ln[r_2(1)] - \ln[r_1(1)]}. \tag{12}$$

At the end of this section, we give an alternative justification for such a scaling behavior. From now on, we specifically consider the Euclidian distance function [see Eq. (9)]; using the Hamming distance will not affect the substance of our reasoning, but it will change some of the details.

For the sake of simplicity, we can assume that the reference configuration $i = 1$ corresponds to the lowest energy state. This second assumption relies on the fact that such a state is the one that has a higher probability of being sampled at any temperature, and, at least at sufficiently low temperatures, it is very likely to be the state with the lowest value of μ, as low-lying excitations typically differ from the lowest energy states by a low amount of spin flips (representative of spin waves), when compared to the average distance between states. Within this approximation, we can fix the coordinates of the reference configuration: $s_j = s \ \ \forall\, j$ for the Ising data set and $\theta_j = \theta \ \ \forall\, j$ for the XY data set.

We now proceed by analyzing the denominator of Eq. (12). We define

$$\alpha_{0f} = \sum_{j=1}^{N_s} S_0 S_{j,f}, \tag{13}$$

where S_0 represents the coordinates of the reference configuration; see Fig. 9(c). Thus, the distance between two configurations reads

$$r_f(1) = \sqrt{2N_s}\sqrt{1 - \frac{\alpha_{0f}}{N_s}}. \tag{14}$$

We then get

$$\ln[r_2(1)] - \ln[r_1(1)] = \frac{\ln(1 - \frac{\alpha_{01}}{N_s}) - \ln(1 - \frac{\alpha_{02}}{N_s})}{2}. \tag{15}$$

For $T \gg T_c$, the coordinates of neighboring configurations are expected to be completely random compared to S_0. Thus, it is reasonable to expect that $\alpha_{0f} \ll N_s$ (we will come back to this point below). By analyzing the transition from the disordered phase, we can expand the logarithms up to second order in α_{0f}/N_s and get

$$\ln[r_2(1)] - \ln[r_1(1)] = \mathcal{F}_2 + \mathcal{F}_4 + \cdots, \tag{16}$$

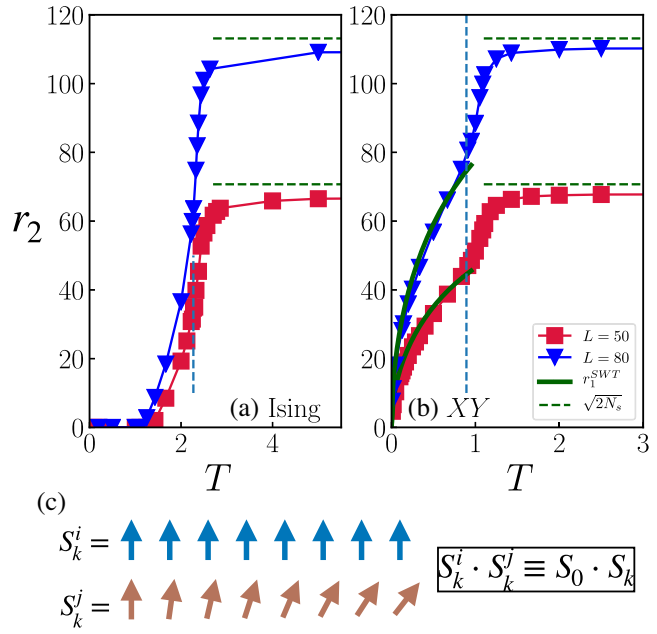

FIG. 9. Correlation functions and distance between neighboring configurations. Temperature dependence of the smallest r_2 (this distance is representative of the lowest value of μ, see text) for the (a) Ising and (b) XY models. The lines (solid and dashed) represent the predictions for r_2 based on Eq. (14) and the corresponding expressions of the asymptotic formula for the correlation functions in the high- and low-temperature regimes (see text). Panel (c) illustrates the basic assumption behind Eq. (14); i.e., correlations between configurations are equivalent to correlation functions (see text).

where the function $\mathcal{F}_p$ contains all p-spin correlation functions, $S_{j_1,f}S_{j_2,f}\ldots S_{j_p,f}$, taken over the single states 1 and 2. In principle, one should also retain other orders; in fact, the difference of the two distances depends parametrically on arbitrary-body correlation functions.

Now, we make a third assumption—that the correlations contained in $\mathcal{F}_p$ can be replaced by the corresponding thermal averages. The rationale behind this is that, based on our first assumption above, we are actually considering the states that have the highest weight in the partition function—so the ones that contribute the most to the computation of the correlation function. Here, temperature plays a clear physical role: Higher temperatures let us sample states that are (on average) at a larger distance from the lowest-energy state when compare to lower temperatures. One can reformulate the above as follows. For any given Markov chain, we have a given $\mathcal{F}_p^{(k)}$, which depends on correlations on a single pair of configurations. Then, averaging over the various Markov chains gives us an averaged value that depends on the average of correlations over the various configurations. This last formulation is closer to the numerical recipe that we utilize to estimate I_d, where, in fact, we obtain the latter from averaging the I_d resulting from several distinct simulations.

Now, since we are dealing with thermal averages, we can recall the finite-size scaling hypothesis. This hypothesis tells us that, if a quantity develops a singular behavior at the transition point (not necessarily a divergence), the temperature corresponding to such a feature, T_{feat}, will be shifted according to FSS theory as (for second-order phase transitions)

$$(T_{\text{feat}} - T) \propto \frac{1}{L^{\nu}}. \tag{17}$$

Thus, we are in a position to make a statement: If any of the arbitrary-body correlation functions contained in the definition of our distance displays singular behavior at the transition point, they will dictate the scaling of the position of the minimum of I_d according to FSS and reveal the critical exponent ν (or, in the case of BKT, they will be consistent with the logarithmic scaling expected there). The behavior of all other correlations is not expected to affect this scaling behavior at all, as they do not display any nonsingular feature, by definition.

We note that our reasoning and the validity of its assumptions can be *a posteriori* verified by noticing that we imply that features in the distribution of the distances r_1 will also be related to critical behavior. In particular, we consider the pivotal assumption that the correlation contained in Eq. (14) can be replaced by the corresponding thermal averages, and we compare the predictions for $r_f(1)$ with our numerical results [65]. Figure 9 shows this comparison for both the Ising and XY models. (i) In the disordered phases, $\alpha_{0,f} \ll 1$ because of the exponential decay of the correlations, and thus $r_f \approx \sqrt{2N_s}$. (ii) On the other hand, in the symmetry-broken phases, $\alpha_{0,f} \approx O(N_s)$, given the long-range nature of the correlations, which implies that $r_f \ll 1$. (iii) Finally, in the XY model's critical phase, the temperature dependence of the correlations is given by $S_0 \cdot S_{j,f} \sim |j|^{-T/2\pi}$ (where |j| represents the spatial distance from a reference site). By computing the corresponding $\alpha_{0,f}$, one can obtain the temperature dependence of r_f. The numerical results display very good agreement with our predictions; see Figs. 9(a) and 9(b). It is worth noting that, while our argument justifies critical scaling for $T_c(L)$ and does not justify the full collapse scaling observed for I_d, it is still directly informative about both critical temperature and the critical exponent ν.

Before ending the section, we present a different approach to determining the dependence between I_d and the smallest value of μ as per Eq. (2). An alternative way to qualitatively estimate I_d from data distributed according to Eq. (1) is to apply the maximum likelihood criterion. Utilizing the commonly used log-likelihood function $\ell_k = \log[f(\mu_k)]$, one obtains that, for data sets where $I_d \gg 1$, one has $I_D \simeq N_r/\ln(\mu_1)$. The scaling with N_r is different with respect to Eq. (12), which is not unexpected due to the fact that (1) we are considering sampling a few

configurations in the previous approximation and (2) maximum likelihood does not necessarily capture the correct scaling with the number of points in the set (as one may expect, many of those do not contribute to the determination of the minimum). Nevertheless, this difference is irrelevant for the sake of our argument above, as we are not immediately interested in the N_r scaling. What is important is that the maximum likelihood method returns exactly the same functional dependence on μ_1, thus providing a data-driven justification of the first assumption presented above.

VI. CONCLUSIONS

We have shown that phase transitions can be learned through a single property of raw data sets of configurations—the intrinsic dimension—without any need to perform dimensional reduction. The key observation made here is that, in analogy to physical observables, the intrinsic dimension exhibits universal scaling behavior close to different classes of transitions: first-order, second-order, and BKT types. This observation indicates how the intrinsic dimension, in the vicinity of critical points, behaves as an order parameter in data space, showing how the latter undergoes a structural transition that parallels the phase transition identified by conventional order parameters.

At the practical level, we have shown that the finite-size analysis of the intrinsic dimension allows us not just to detect but also to characterize critical points in an unsupervised manner. In particular, we have shown that the intrinsic dimension allows one to estimate transition temperatures and (critical) exponents of both first- and second-order transitions with accuracies ranging from 1% to 0.1% at very modest system sizes. In addition, the method is equally applicable to topological transitions, where we have demonstrated an accurate (with 1% confidence) estimate of the location of the BKT topological transition competitive with more traditional methods at the same system sizes. This latter result suggests that the lack of any dimensional reduction allows us to retain topological information in the vicinity of the phase transition, which may be lost otherwise [27,29].

A fundamental aspect of our approach is that it is based on a I_d-estimation method suitable to learn complex manifolds, such as the twisted XY manifold emerging at the BKT critical point. The results demonstrate the potential of state-of-the-art I_d-estimator [8,66] methods to tackle many-body problems, and they motivate an even stronger methodological connection between data-mining techniques and many-body physics. We also note that, in comparison with previous applications in other fields [32,35,67], the values of the intrinsic dimension reported here are considerably larger. For future applications, like combining our analysis with clustering methods that do not rely on dimension reduction [68], it may be interesting to develop novel estimators that focus on large values of I_d, potentially trading absolute accuracy with numerical efficiency (in the spirit of Ref. [66]). Another interesting question to address in the future is to assess the possibility of data lying in submanifolds with different I_d [69] and how it affects our method. This scenario is plausible in cases where phases coexist.

Some of the methods presented here may be applied to quantum mechanical objects, such as quantum partition functions, density matrices, and wave functions. It is an open challenge to determine whether the data mining of quantum objects can provide an informative perspective on the latter, such as, e.g., accessing entanglement or other more challenging forms of quantum correlations. Finally, while we focus on configuration generated by Monte Carlo sampling, our approach is equally applicable to experimentally generated data; it may be interesting to apply it to settings where raw data configurations are available, such as, e.g., quantum gas microscope experiments [18,70,71].

ACKNOWLEDGMENTS

We acknowledge useful discussions with R. Ben Ali Zinati, R. Fazio, A. Laio, and R. T. Scalettar. The work of T. M. S., X. T., and M. D. is partly supported by the ERC under Grant No. 758329 (AGEnTh) and by the Quantera program QTFLAG; they have received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 817482. This work has been carried out within the activities of Trieste Institute for the Theory of quantum technologies. T. M. S. and X. T. acknowledge computing resources at Cineca Supercomputing Centre through the Italian SuperComputing Resource Allocation via the ISCRA grants ICT20_CMSP and MLforPT.

APPENDIX A: TWO-NN METHOD AND COMPARISON WITH PCA

In this Appendix, we provide more details about the two-NN method. As described in Ref. [8], the intrinsic dimension I_d can be obtained through the following steps:

(1) For each point i of the data set ($i = 1, 2, \ldots, N_r$), compute its first- and second-nearest neighbors, $r_1(i)$ and $r_2(i)$, respectively.
(2) For each point i, compute the ratio $\mu_i = r_2(i)/r_1(i)$.
(3) The empirical cumulate is defined as $P^{\mathrm{emp}}(\mu) = i/N_r$, while the values of μ_i are sorted in ascending order through a permutation, i.e., $(\mu_1, \mu_2, \ldots \mu_{N_r})$, where $\mu_i < \mu_j$, for $i < j$.
(iv) Finally, the resulting $S = \{(\ln(\mu), -\ln[1 - P^{\mathrm{emp}}(\mu)]\}$ are fitted with a straight line passing through the origin. The slope of this line is equal to I_d [see Eq. (2)].

Figure 10 shows the plot of S for the basic three-site XY example presented in Fig. 1(b). It is worth mentioning that, while we depict the configurations $\vec{\theta} = (\theta_1, \theta_2, \theta_3)$ for clarity of illustration in Fig. 1(b), in our calculations,

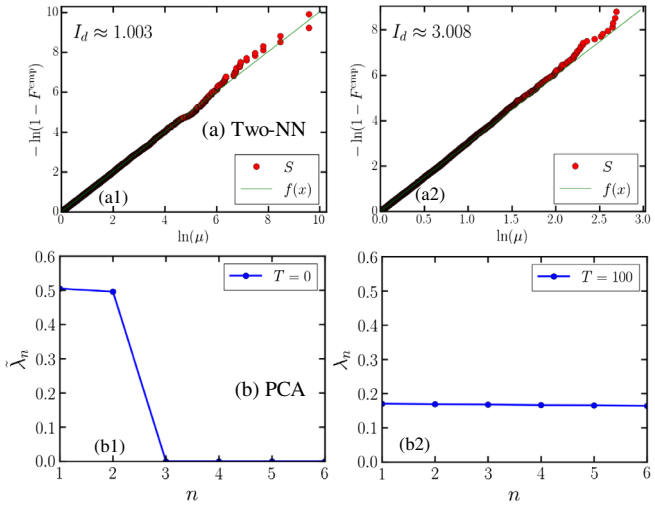

FIG. 10. Three-site XY model. Panels (a1) and (a2) show results of the two-NN method: fitting of the data points S for (a1) $T = 0$ and (a2) $T = 100$. The data set has $N_r = 10^3$ configurations. We obtain (a1) $I_d = 1$ and (a2) $I_d = 3$. Panels (b1) and (b2) show results of the PCA method: normalized eigenvalues of the covariance matrix, $\tilde{\lambda_n}$, obtained from the raw XY configurations. Here, we use the same notation as Ref. [27]. For $n > I_d^{\mathrm{PCA}}$, $\tilde{\lambda_n} \to 0$. PCA predicts (b1) $I_d^{\mathrm{PCA}} = 2$ and (b2) $I_d^{\mathrm{PCA}} = 6$, which is not in agreement with the exact results (see text).

$\vec{\theta}$ is defined as in Eq. (8). In this way, the distance between two configurations $\vec{\theta^i}$ and $\vec{\theta^j}$, $r(\vec{\theta^i}, \vec{\theta^j}) = \sqrt{2\sum_{k=1}^{N_s}(1 - \vec{S}_k^i \vec{S}_k^j)}$, properly takes into account the periodicity of the variables θ_k^i. Another important technical aspect is that the fit of S is unstable for larger values of μ. As is considered in Ref. [8], we discard the 10% of points characterized by the highest values of μ. Based on this approach, we obtain $I_d \approx 1$ and $I_d \approx 3$ for the zero- and high-temperature regimes, respectively, which is consistent with the value expected from physically reasonable assumptions [see Figs. 10(a1) and 10(a2)].

We also show some examples of the plot of S for data sets generated in the vicinity of the critical points of the Ising and 2D XY models; see Figs. 11(a) and 11(b), respectively. In both cases, the points S are well fitted by a straight line passing through the origin. We obtain similar results for the other system sizes and values of T considered in this work.

In contrast, simple linear-dimension-reduction methods, such as PCA, fail to describe the I_d of the XY data sets. To illustrate this point, we employ linear PCA in the same collection of configurations considered in the last paragraph. As can be seen from Figs. 10(b1) and 10(b2), even for this simple example, PCA fails to obtain the true I_d; for $T = 0$, $I_d^{\mathrm{PCA}} = 2$, while for $T = 100$, $I_d^{\mathrm{PCA}} = 6$. This failure is related to the fact that the XY manifolds are curved [29].

However, PCA can describe the main features of the Ising data set. Based on this, we consider the PCA

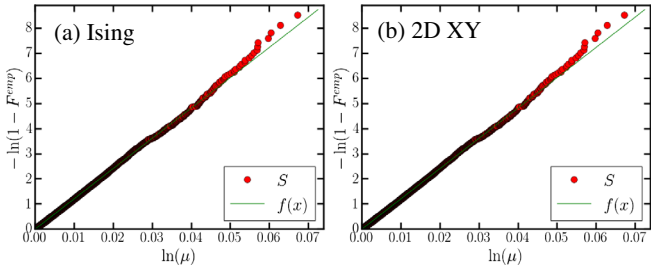

FIG. 11. Results of the two-NN method: fitting of the data points S for (a) Ising and (b) 2D XY thermal data sets generated close to the critical point [in panel (a), $T \approx 2.2$, while in panel (b), $T \approx 0.89$]; in both cases, we consider $L = 40$, and the data set has $N_r = 10^4$ configurations. We obtain $I_d \approx 47$ and $I_d \approx 120$, respectively.

estimation of the I_d in Fig. 7 of the main text. Here, we give more details about the computation of $I_{d,\mathrm{PCA}}$. First, we consider the eigenvalues of the covariance matrix $\mathbf{X^T X w_n} = \lambda_\mathbf{n} \mathbf{w_n}$ (we use the same notation of Ref. [27]). We then define the normalized eigenvalues, $\tilde{\lambda_n} = (\lambda_n / \sum_i^{N_r} \lambda_i)$. The I_d is defined by choosing an *ad hoc* cutoff parameter for the integrated spectrum of the covariance matrix, i.e.,

$$\sum_{n=1}^{I_{d,\mathrm{PCA}}} \tilde{\lambda}_n \approx f, \tag{A1}$$

where f represents a fraction of the eigenvalues of the covariance matrix. In Fig. 7(c), we consider $f = 0.6$. The value of $I_{d,\mathrm{PCA}}$ depends on f. However, we observe that the qualitative behavior of the function $I_{d,\mathrm{PCA}}(T)$ is not affected by the value of f (as long as $f \geq 0.5$). In particular, for all the values of f that we consider (i.e., $f = 0.5$, 0.6, 0.7, 0.8, and 0.9), the global $I_{d,\mathrm{PCA}}(T)$ goes to 1 immediately below T_c, and the local $I_{d,\mathrm{PCA}}(T)$ exhibit a nonmonotonic behavior.

APPENDIX B: SCALING OF I_d WITH THE NUMBER OF CONFIGURATIONS

In this section, we discuss the scaling of the I_d with the number of configurations, N_r, considered in the data set; for all the results shown in the main text, $N_r = 5 \times 10^4$. The first important aspect to consider is that the two-NN is a scale-dependent method. In other words, the estimation of the I_d is performed on a length scale that is related to the first- and second-neighbor distances of each point. Thus, by varying N_r, one is probing a different neighborhood size, i.e., estimating I_d at different scales [8].

To illustrate how this change in scale affects the I_d of the thermal data sets considered here, we first consider the three-site XY model in Fig. 12(a). For $T = 1$, the I_d converges to three as expected for the high-temperature regime of this model. In the low-temperature regime, $T \approx 10^{-6}$; however, $I_d(N_r)$ exhibits a plateau at $I_d = 1$

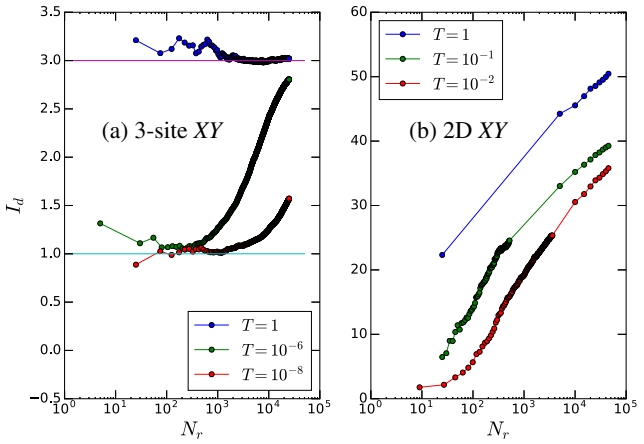

FIG. 12. Scaling of the I_d with the number of configurations in the data set. In panel (a), we show the results for the three-site XY model, while in panel (b), we show results for the 2D XY model with $L = 10$.

for $N_r \in [10^0, 10^3[$. As illustrated in Fig. 1(b), this simple data set is well described by a one-dimensional manifold. This plateau in $I_d(N_r)$ is a signature of this soft direction [8]. Nevertheless, by further increasing N_r, the I_d increases ($I_d \to 3$ in this case), as an effect of the decrease of the scale in which I_d is estimated. In this scalar regime, the number of soft directions cannot be determined. We stress that the computation of the I_d of the high-dimensional data sets considered here is always performed in this regime. In this case, the I_d exhibit an exponential scaling with N_r, as exemplified in Fig. 12(b).

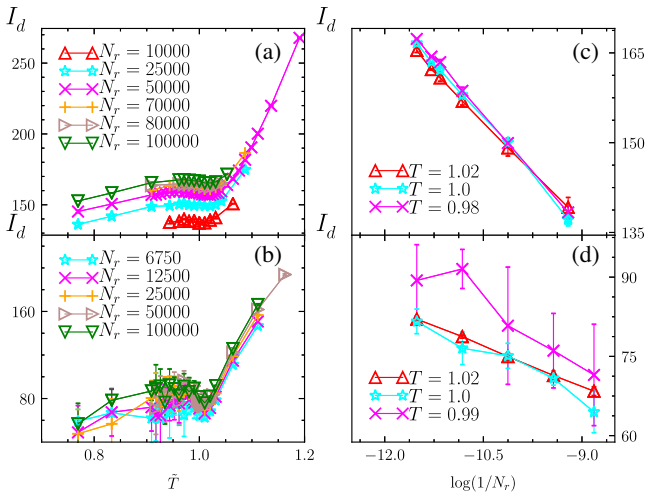

FIG. 13. Temperature dependence of the I_d for different values of N_r for the (a) 2D XY model and (b) three-state Potts model; in both cases, $L = 60$. For each point, we harvest approximately ten instances of the data set and average the resulting estimates for the I_d. The error bars are the standard deviation of such results. The scaling of I_d with N_r for certain values of T for: (c) the 2D XY model, (d) the three-state Potts model.

We now discuss how the temperature dependence of I_d is affected by the change in N_r. Figures 13(a1) and 13(b1) show $I_d(T)$ for different values of N_r for the Potts and 2D XY data sets, respectively. Despite the change of the absolute value of I_d with N_r, the qualitative behavior of $I_d(T)$ is not modified. Most importantly, we observe that the position of the local minimum at T^* does not shift with N_r for $N_r > 10^4$. Furthermore, as expected, the scaling of I_d with N_r is exponential [see Figs. 13(a2) and 13(b2)], at least in the vicinity of the phase transition. Similar results are obtained for other system sizes and for the Ising model. Summing up, our results indicate that, as long as $N_r > 10^4$, the universal scaling behavior exhibited by the I_d is not affected by the scale at which the I_d is measured.

APPENDIX C: DATA COLLAPSE

In this section, we discuss the finite-size analysis employed for the estimates of the critical temperature and exponents presented in Sec. IV. Our procedure is a standard search of the minimal least-square difference fit between our data and an appropriately chosen scaling function hypothesis. Let us first focus on the second-order phase transitions, concerning the Ising and three-state Potts models. The method is divided into four steps.

(1) First we choose a suitable mesh for the parameter ranges for T_c, ν, and α.

(2) From our data, we compute the scaling variables $x_{\rm dat}(T_c, \nu) = (T - T_c)L^{1/\nu}$ and $y_{\rm dat}(\alpha) = I_d(T)L^{-\alpha}$ for a different R-range of system sizes $\{L_1, L_2, \ldots, L_R\}$.

(iii) We choose a parametric functional hypothesis $f(x; \{a\})$.

(iv) For each $[x_{\rm dat}(T_c, \nu), y_{\rm dat}(\alpha)]$, we compute the best fit of the hypothesis function $\{a^\star\}$ through the Levenberg-Marquardt algorithm. We store the residuals as

$$\epsilon(T_c, \nu, \alpha) = \frac{||f(x_{\rm dat}(T_c, \nu); \{a^\star\}) - y_{\rm dat}(\alpha)||}{||y_{\rm dat}(\alpha)||}. \tag{C1}$$

The optimal set of parameters for each set $\{L_1, L_2, \ldots, L_R\}$ is located in the minimum $\epsilon(T_c, \nu, \alpha)$. In order to keep a low bias on the hypothesis function $f(x, \{a\})$, we choose various k-degree polynomials $Q_k(x; a_0, a_1, \ldots, a_k)$. Thus, we obtain a set of optimal $\{T_c^\star\}$, $\{\nu^\star\}$, and $\{\alpha^\star\}$ for each choice of degree k and each set of system sizes $\{L_k\}$. Our estimates and errors for the critical temperature and critical exponents are then estimated as the average and standard deviation of these sets, respectively.

The analysis for the BKT transition (the XY model) is performed in a similar fashion. The only difference is the choice of scaling variable, which, for this case, is

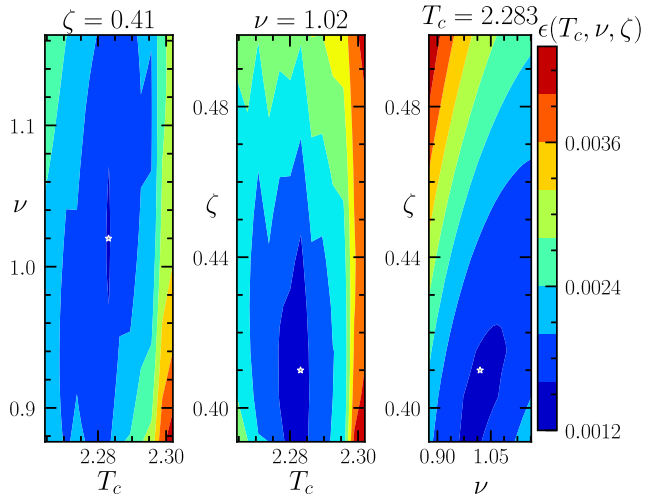

FIG. 14. Contour plot of the average residuals projected on the direction of the optimal parameters for the Ising model. The white star points to the optimal parameters for our finite-size analysis.

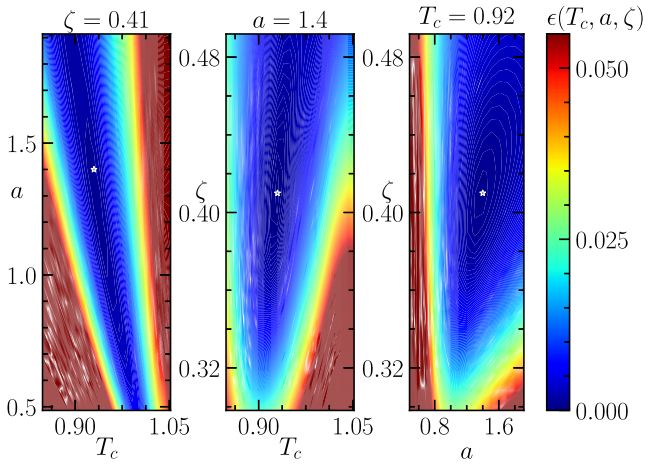

FIG. 15. Contour plot of the average residuals projected in the direction of the optimal parameters for the XY model. The white star points to the optimal parameters for our finite-size analysis.

$$x_{\mathrm{dat}}(T_c, a) = L \exp\left[-\frac{a}{\sqrt{T - T_c}}\right]. \tag{C2}$$

For the Ising, three-state Potts, and XY models, we select polynomials of degrees $k \in \{5, 6, 7, 8\}$ and different sets of system sizes among the $L \geq 70$ ones. For the XY model, we perform the data collapse within the range $T = [0.91, 1.10]$ and use a bin of $\Delta T \approx 0.005$. Our estimations for the XY model are $T_c = 0.92(1)$, $a = 1.4(1)$, and $\zeta = 0.40(1)$.

We visualize the resulting residuals for both the Ising and the XY models; see Figs. 14 and 15, respectively. Since the parameter space is 3D, for convenience, we plot the projected directions along with the optimal critical parameters.

APPENDIX D: DEFINITION OF THE WINDING NUMBER

We now discuss the definition of the winding number mentioned in Fig. 8(b) of the main text. We consider a closed path along the x and y directions of the square lattice and define

$$w_x = \frac{1}{2\pi} \sum_{i=1}^{L_x} \Delta\theta_{(i,y=1)} \tag{D1}$$

and

$$w_y = \frac{1}{2\pi} \sum_{i=1}^{L_y} \Delta\theta_{(x=1,i)}, \tag{D2}$$

where the angle difference is $\Delta\theta_{(x=1,i)} = \theta_{(i+1,y=1)} - \theta_{(i,y=1)}$; $\Delta\theta$ is rescaled into the range $(-\pi, \pi]$. We compute the $w = (w_x, w_y)$ for each configuration of the data set. We then define the total number of configurations whose first two nearest neighbors have the same w, N_w. Figure 8(b) shows the fraction of those points, $F_w = N_w/N_r$, as a function of T.

[1] M. I. Jordan and T. M. Mitchell, *Machine Learning: Trends, Perspectives, and Prospects*, Science **349**, 255 (2015).
[2] Y. LeCun, Y. Bengio, and G. Hinton, *Deep Learning*, Nature (London) **521**, 436 (2015).
[3] P. Domingos, *A Few Useful Things to Know about Machine Learning*, Commun. ACM **55**, 78 (2012).
[4] 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).
[5] E. Levina and P. J. Bickel, *Maximum Likelihood Estimation of Intrinsic Dimension*, edited by L. K. Saul, Y. Weiss, and L. Bottou (MIT Press, Cambridge, MA, 2004).
[6] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, *Modelling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model*, Phys. Rev. X **10**, 041044 (2020).
[7] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, *Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework*, Math. Probl. Eng. **2015**, 759567 (2015).
[8] 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).
[9] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *Machine Learning and the Physical Sciences*, Rev. Mod. Phys. **91**, 045002 (2019).
[10] 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).

[11] J. Carrasquilla, *Machine Learning for Quantum Matter*, Adv. Phys. X **5**, 1 (2020).

[12] L. Wang, *Discovering Phase Transitions with Unsupervised Learning*, Phys. Rev. B **94**, 195105 (2016).

[13] J. Carrasquilla and R. G. Melko, *Machine Learning Phases of Matter*, Nat. Phys. **13**, 431 (2017).

[14] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, *Learning Phase Transitions by Confusion*, Nat. Phys. **13**, 435 (2017).

[15] Y. Zhang and E.-A. Kim, *Quantum Loop Tomography for Machine Learning*, Phys. Rev. Lett. **118**, 216401 (2017).

[16] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, *Machine Learning Phases of Strongly Correlated Fermions*, Phys. Rev. X **7**, 031038 (2017).

[17] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, *Identifying Quantum Phase Transitions Using Artificial Neural Networks on Experimental Data*, Nat. Phys. **15**, 917 (2019).

[18] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap, *Classifying Snapshots of Doped Hubbard Model with Machine Learning*, Nat. Phys. **15**, 921 (2019).

[19] Y. Zhang, A. Mesaros, K. Fujita, S. D. Edkins, M. H. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. C. S. Davis, E. Khatami, and E.-A. Kim, *Machine Learning in Electronic-Quantum-Matter Imaging Experiments*, Nature (London) **570**, 484 (2019).

[20] D. Bachtis, G. Aarts, and B. Lucini, *Extending Machine Learning Classification Capabilities with Histogram Reweighting*, Phys. Rev. E **102**, 033303 (2020).

[21] G. Carleo and M. Troyer, *Solving the Quantum Many-Body Problem with Artificial Neural Networks*, Science **355**, 602 (2017).

[22] D.-L. Deng, X. Li, and S. Das Sarma, *Quantum Entanglement in Neural Network States*, Phys. Rev. X **7**, 021021 (2017).

[23] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, *Quantum Boltzmann Machine*, Phys. Rev. X **8**, 021050 (2018).

[24] M. Schmitt and M. Heyl, *Quantum Many-Body Dynamics in Two Dimensions with Artificial Neural Networks*, Phys. Rev. Lett. **125**, 100503 (2020).

[25] S. J. Wetzel, *Unsupervised Learning of Phase Transitions: From Principal Component Analysis to Variational Autoencoders*, Phys. Rev. E **96**, 022140 (2017).

[26] K. Ch'ng, N. Vazquez, and E. Khatami, *Unsupervised Machine Learning Account of Magnetic Transitions in the Hubbard Model*, Phys. Rev. E **97**, 013306 (2018).

[27] 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).

[28] N. C. Costa, W. Hu, Z. J. Bai, R. T. Scalettar, and R. R. P. Singh, *Principal Component Analysis for Fermionic Critical Points*, Phys. Rev. B **96**, 195138 (2017).

[29] J. F. Rodriguez-Nieva and M. S. Scheurer, *Identifying Topological Order through Unsupervised Machine Learning*, Nat. Phys. **15**, 790 (2019).

[30] Y. Long, J. Ren, and H. Chen, *Unsupervised Manifold Clustering of Topological Phononics*, Phys. Rev. Lett. **124**, 185501 (2020).

[31] E. A.-C. E. Lopez, A. Scheuer, and F. Chinesta, *On the Effect of Phase Transition on the Manifold Dimensionality: Application to the Ising Model*, MEMOCS **6**, 251 (2018), https://msp.org/memocs/2018/6-3/p05.xhtml.

[32] E. Facco, A. Pagnani, E. T. Russo, and A. Laio, *The Intrinsic Dimension of Protein Sequence Evolution*, PLoS Comput. Biol. **15**, e1006767 (2019).

[33] N. Krueger and M. Felsberg, *A Continuous Formulation of Intrinsic Dimension*, edited by R. Harvey and A. Bangham (BMVA Press, Norwich, 2003).

[34] A. A. Ramos, H. Socas-Navarro, A. L. Ariste, and M. M. González, *The Intrinsic Dimensionality of Spectropolarimetric Data*, Astrophys. J. **660**, 1690 (2007).

[35] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, *Intrinsic Dimension of Data Representations in Deep Neural Networks*, in NeurIPS, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (MIT Press, Cambridge, MA, 2019).

[36] D. C. Laughlin, *The Intrinsic Dimensionality of Plant Traits and Its Relevance to Community Assembly*, J. Ecol. **102**, 186 (2014).

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

[38] I. Borg and P. J. Groenen, *Modern Multidimensional Scaling, Modern Multidimensional Scaling: Theory and Applications* (Springer-Verlag, New York, 2005).

[39] M. Balasubramanian and E. L. Schwartz, *The Isomap Algorithm and Topological Stability*, Science **295**, 7 (2002).

[40] S. T. Roweis and L. K. Saul, *Nonlinear Dimensionality Reduction by Locally Linear Embedding*, Science **290**, 2323 (2000).

[41] M. A. Kramer, *Nonlinear Principal Component Analysis Using Autoassociative Neural Networks*, AIChE J. **37**, 233 (1991).

[42] L. van der Maaten and G. Hinton, *Nonlinear Principal Component Analysis Using Autoassociative Neural Networks*, J. Mach. Learn. Res. **9**, 2579 (2008), https://www.jmlr.org/papers/v9/vandermaaten08a.html.

[43] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426.

[44] F. Camastra and A. Vinciarelli, *Estimating the Intrinsic Dimension of Data with a Fractal-Based Method*, IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1404 (2002).

[45] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor and Francis, London, 1991).

[46] U. Wolff, *Collective Monte Carlo Updating for Spin Systems*, Phys. Rev. Lett. **62**, 361 (1989).

[47] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, England, 2005).

[48] P. Di Francesco, P. Mathieu, and D. Sénéchal, *Conformal Field Theory* (Springer, New York, 1997).

[49] M. Henkel, *Conformal Invariance and Critical Phenomena* (Springer, Berlin, Heidelberg, 1999).

[50] F. Y. Wu, *The Potts Model*, Rev. Mod. Phys. **54**, 235 (1982).

[51] S. Iino, S. Morita, N. Kawashima, and A. W. Sandvik, *Detecting Signals of Weakly First-Order Phase Transitions in Two-Dimensional Potts Models*, J. Phys. Soc. Jpn. **88**, 034006 (2019).

[52] J. M. Kosterlitz and D. J. Thouless, *Ordering, Metastability and Phase Transitions in Two-Dimensional Systems*, J. Phys. C **6**, 1181 (1973).

[53] R. Gupta, J. DeLapp, G. G. Batrouni, G. C. Fox, C. F. Baillie, and J. Apostolakis, *Phase Transition in the 2D XY Model*, Phys. Rev. Lett. **61**, 1996 (1988).

[54] Y. Che, C. Gneiting, T. Liu, and F. Nori, *Topological Quantum Phase Transitions Retrieved through Unsupervised Machine Learning*, Phys. Rev. B **102**, 134213 (2020).

[55] Y.-D. Hsieh, Y.-J. Kao, and A. W. Sandvik, *Finite-Size Scaling Method for the Berezinskii-Kosterlitz-Thouless Transition*, J. Stat. Mech. (2013) P09001.

[56] A. W. Sandvik, *Computational Studies of Quantum Spin Systems*, AIP Conf. Proc. **1297**, 135 (2010).

[57] C. Wang and H. Zhai, *Machine Learning of Frustrated Classical Spin Models. I. Principal Component Analysis*, Phys. Rev. B **96**, 144432 (2017).

[58] M. J. S. Beach, A. Golubeva, and R. G. Melko, *Machine Learning Vortices at the Kosterlitz-Thouless Transition*, Phys. Rev. B **97**, 045207 (2018).

[59] A. Lidiak and Z. Gong, *Unsupervised Machine Learning of Quantum Phase Transitions Using Diffusion Maps*, Phys. Rev. Lett. **125**, 225701 (2020).

[60] Q. H. Tran, M. Chen, and Y. Hasegawa, *Topological Persistence Machine of Phase Transitions*, arXiv:2004.03169.

[61] D. R. Nelson and J. M. Kosterlitz, *Universal Jump in the Superfluid Density of Two-Dimensional Superfluids*, Phys. Rev. Lett. **39**, 1201 (1977).

[62] E. Buddenoir and S. Wallon, *The Correlation Length of the Potts Model at the First-Order Transition Point*, J. Phys. A **26**, 3045 (1993).

[63] M. E. Fisher and A. N. Berker, *Scaling for First-Order Phase Transitions in Thermodynamic and Finite Systems*, Phys. Rev. B **26**, 2507 (1982).

[64] K. Binder and D. P. Landau, *Finite-Size Scaling at First-Order Phase Transitions*, Phys. Rev. B **30**, 1477 (1984).

[65] We note that, by verifying this assumption, we indirectly check the ones taken before as well.

[66] V. Erba, M. Gherardi, and P. Rotondo, *Intrinsic Dimension Estimation for Locally Undersampled Data*, Sci. Rep. **9**, 17133 (2019).

[67] A. Rodriguez, M. d'Errico, E. Facco, and A. Laio, *Computing the Free Energy without Collective Variables*, J. Chem. Theory Comput. **14**, 1206 (2018).

[68] M. d'Errico, E. Facco, A. Laio, and A. Rodriguez, *Automatic Topography of High-Dimensional Data Sets by Non-parametric Density Peak Clustering*, arXiv:1802.10549.

[69] M. Allegra, E. Facco, F. Denti, A. Laio, and A. Mira, *Data Segmentation Based on the Local Intrinsic Dimension*, Sci. Rep. **10**, 16449 (2020).

[70] C. Gross and I. Bloch, *Microscopy of Many-Body States in Optical Lattices*, in *Annual Review of Cold Atoms and Molecules*, edited by K. W. Madison, K. Bongs, L. D. Carr, A. M. Rey, and H. Zhai (World Scientific, Singapore, 2015).

[71] E. Khatami, E. Guardado-Sanchez, B. M. Spar, J. F. Carrasquilla, W. S. Bakr, and R. T. Scalettar, *Visualizing Strange Metallic Correlations in the Two-Dimensional Fermi-Hubbard Model with Artificial Intelligence*, Phys. Rev. A **102**, 033326 (2020).