SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area – PhD course in
Statistical Physics

Non-parametric learning in many-body physics.

Candidate:
Rajat K. Panda

Advisor:
Marcello Dalmonte

Academic Year 2022-23

To my loving parents...
Narmada and Rabindra

Preface

In this digital age, machine learning (a prominent element within the context of *artificial intelligence*) has become an integral part of modern life, influencing various aspects of our daily activities. The predictive power of these tools is increasing with the advancements in algorithms, increased computational power, and the availability of large data sets: one basic example of this is language-based models, which have revolutionized our capabilities of transferring and processing information. Data is the foundational element of machine learning, and the quality and relevance of the data sets have a profound impact on the performance and accuracy of machine learning models. Thus "*data is the new gold*", captures the immense value that data holds in our modern world. With the exponential growth in data, the increasing demand for better machine-learning tools comes as a natural consequence.

The field of physics is no exception to this, especially with the rise of big simulations and experiments generating huge amounts of data. Experiments at particle accelerators like the Large Hadron Collider(LHC) and Fermi lab, gravitational waves detector by LIGO and Virgo, climate models and weather simulations, Sloan Digital Sky Survey (SDSS), Cosmic Microwave Background(CMB) radiation providing insights into the early universe; all these studies generate massive data sets. In this context, developing more powerful, accurate, and specialized tools, is essential to meet the evolving needs in physics as well.

In this thesis, we aim to study different types of ML tools to characterize the many-body problem in the context of classical and quantum statistical mechanics, with a focus on unbiased and interpretable methods. We use techniques coming from so-called non-parametric learning (NPL), where the analysis is done without specifying a fixed set of parameters in an assumption-free

manner. NPL is a subset of the broad unsupervised learning category; the algorithm learns from the intrinsic structure of the data set. NPL is particularly valuable in scenarios where data assumptions are uncertain, data distributions are non-standard, or when dealing with complex data sets where relationships are difficult to predict beforehand. This works in our favor as the relation between data sets coming from many-body physics system and their effect on data structures when physical parameters change is not well understood - the simple reason being that the typical physicists mindset is very much focused on performing huge dimensional reduction from microscopic physical information (a canonical example of this being mean field theories, where the behavior of a large number of degrees of freedom is replaced by a single one).

The thesis is organized as follows. In Chapter 1, we discuss the theoretical framework behind the NPL methods used. We present basic ideas in our methodologies, which encompass the overarching principles guiding our approach. These include the linkage between statistical physics and data analysis, particularly in relation to intrinsic dimension (Id) and its significance in understanding critical characteristics. Additionally, we delve into the application of Principal Component Analysis (PCA) and introduce the notion of PCA entropy - a witness of information spreading under linear transformation, whose relevance to physical phenomena we start investigating here. In Chapter 2, we describe the data sets we work with; statistical mechanics inspired data sets, such as those generated via sampling classical Ising models through advanced Monte Carlo algorithms, and experimental data of a spinor Bose-Einstein condensate. The following chapters present the applications I worked on. In Chapter 3, we present and thoroughly assess a framework for understanding critical behavior in classical partition functions by employing NPL techniques on data sets containing thermal configurations; provide a robust analysis of phase transitions using different NPL methods. In Chapter 4, we present a theoretical framework for information extraction in synthetic quantum matter for the case of a quantum quench in a spinor Bose-Einstein

condensate experiment. In Chapter 5, we demonstrate how the field of network science can significantly improve and elevate our assumption-free characterization of various phases of matter. We present compelling evidence that network science offers a transparent and effective set of methods for investigating the properties of distinct phases and phase transitions, enabling us to identify critical points.

List of publications

The thesis is based on the following publications:

- Chapter 3: [1] Panda, R.K., Verdel, R., Rodriguez, A., Sun, H., Bianconi, G. and Dalmonte, M., 2023. Non-parametric learning critical behavior in Ising partition functions: PCA entropy and intrinsic dimension. arXiv:2308.13636.

- Chapter 4: [2] Verdel, R., Vitale, V., Panda, R.K., Donkor, E.D., Rodriguez, A., Lannig, S., Deller, Y., Strobel, H., Oberthaler, M.K. and Dalmonte, M., 2023. Data-driven discovery of relevant information in quantum simulators. arXiv:2307.10040.

- Chapter 5: [3] Sun, H., Panda, R.K., Verdel, R., Rodriguez, A., Dalmonte, M. and Bianconi, G., 2023. Network science Ising states of matter. arXiv:2308.13604.

Other publications:

- [4] Panda, R.K., Scardicchio, A., Schulz, M., Taylor, S.R. and Žnidarič, M., 2020. Can we study the many-body localisation transition?. Europhysics Letters, 128(6), p.67003.

Contents

Chapter 1

Non-paramatric Learning

Machine learning (ML) techniques can be broadly separated into two categories: parametric and non-parametric learning. Non-parametric methods are indeed a subset of the broader unsupervised learning category in machine learning. There are many parametric unsupervised learning methods for example, the Gaussian mixture model (GMM): commonly used in a parametric form, which significantly limits their capacity in fitting diverse multidimensional data distributions encountered in practice [5, 6]. In contrast to parametric methods, non-parametric learning (NPL) approaches do not make strong assumptions about the functional form of the relationship between the data set and output variable [7, 8, 9]. Non-parametric methods are valuable when dealing with complex high-dimensional data sets that don't follow a predefined model and learn from the intrinsic structures of the provided data set. The most popular examples of non-parametric methods include k-nearest neighbors (KNN) [10], kernel density estimation [11, 12], decision trees [13], random forests [14, 15], and various clustering algorithms [16]. Non-parametric methods allow for the model to be more flexible and adapt to the complexity of the data sets without specifying a fixed set of parameters: this makes them very promising as tools for the unbiased analysis of data sets generated from many-body physics [8].

In this chapter, we introduce the non-parametric methodologies used later in the thesis for analyzing large data sets relevant to statistical physics. We introduce key concepts, including the general philosophy of our approach,

the connection between statistical physics and data analysis through intrinsic dimension(I_d) and its relevance to critical properties, and the use of Principal Component Analysis (PCA) along with the introduction of a novel concept called PCA entropy. This chapter serves as an introduction to the tools and concepts that will be utilized in this thesis.

1.1 Intrinsic Dimension

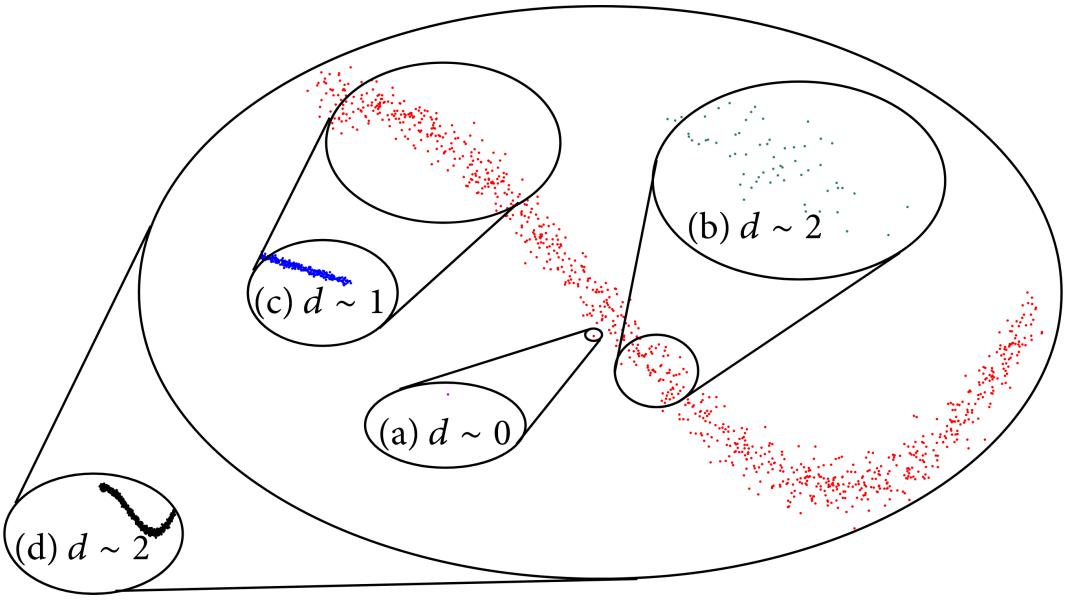

FIGURE 1.1: I_d with changing scales.(a) With smallest scale of interest $I_d \approx 0$, (b) With increased resolution we find $I_d \approx 2$, (c) $I_d \approx 2$, (d) When the entire system is considered $I_d \approx 2$. Figure adapted from [17].

In the age of *big data*, we deal with high-dimensional data sets very frequently. These data sets are often characterized by a multitude of features (higher dimensions), frequently numbering in the thousands. Thankfully, in many real-world data sets, these high-dimensional spaces tend to be sparsely populated, with data points effectively residing on lower-dimensional structures referred to as *data manifolds*. This characteristic is key to the success of

dimensionality reduction techniques. However, when dealing with data manifolds that have complex geometries, it is difficult to find a set of coordinates that can globally describe the manifold, traditional dimensionality reduction methods face challenges. In such cases, it becomes necessary to directly analyze the data manifold without the intermediate step of dimensionality reduction. The concept of intrinsic dimension (I_d) plays a very important role in characterizing these manifolds: it is defined as the minimum number of parameters needed to accurately describe the important characteristics of the data set itself [18]. Estimating the I_d of these high-dimensional data sets can, however, be very challenging. The data manifolds can be highly curved, twisted, or have inhomogeneities, making it hard to determine the true underlying dimensionality (see, e.g., an example shown in Fig 1.1). Finding accurate and reliable estimators of I_d is an active area of research [17, 19, 20, 21].

While trying to estimate the I_d of the system, we need to fix the scale of interest; as I_d can vary rapidly with the scale as shown in Fig 1.1 [17], which represents a curved line perturbed by noise.

- In Fig 1.1(a), at the smallest scale the space is poorly sampled and gives us an extreme situation resulting intrinsic dimension is 0.
- In the intermediate case (b), the noise dominates the signal which makes it necessary to have two independent directions to describe the dataset.
- For (c), we can recognize only the one-dimensional feature.
- In the case of (d) when the whole data space is considered, the intrinsic dimension is two.

Because of all the reasons mentioned above estimating I_d correctly poses a difficult task. However, the practical usefulness in dimensionality reduction [19, 22], feature selection [23], anomaly detection [24], data compression [25] of I_d makes it a useful measure for different reasons.

Over the years, multiple approaches have been developed to estimate the I_d [17]. We will briefly summarize some of the methods. Projective methods:

Through principal component analysis (PCA) [26], by minimizing a projection error for principle components (we will discuss more on this later in the chapter). Also Multidimensional Scaling (MDS)[27], ISOMAP [28], autoencoders [29]; all involving different kinds of dimensional reduction techniques. Most of the estimators discussed above exhibit three primary shortcomings. Firstly, they often struggle when dealing with varying data densities and curvatures. Secondly, there is a notable absence of a robust error-detection procedure. While the first issue has received attention in terms of mitigation, the second one has been largely overlooked. However, establishing the reliability of estimation is essential [19, 20], to circumvent these problems especially the last one, in Ref. [19], a new I_d estimator has been introduced called the two nearest-neighbor (two-NN) method which we will elaborate further in the following section. The method is relatively unaffected by the variations in density and curvatures; able to alert when the model fails to describe the data.

1.1.1 TWO-NN method

We present an I_d estimator for data sets and a relatively robust response to variations in data density and curvatures, the so-called two-NN method. Importantly, it possesses the capability to flag instances when the model fails to accurately describe the data, signaling unreliable measurements. This method is rooted in the computation of the probability distribution of volume shells within a uniform Poisson process, building upon concepts described in Ref [19, 17]. Instead of relying on maximum likelihood estimation, a fitting procedure is introduced that allows for immediate assessment of the correctness in I_d estimation.

Based on these foundational principles, the two-NN method is designed to assess the I_d of a data set by exclusively considering the first and second-nearest neighbor. It's important to note that the definition of "neighbor" in this context relies entirely on the metric used for distance measurements. For our calculations either the Hamming or the Euclidean distance can be used; in

practice, we use the Euclidean distance metric. But the two-NN method to predict the correct I_d, relies on two key assumptions; firstly, the points are generated independently; Secondly, there must be a local homogeneity of the points in the hyperspace, meaning the local density of the points must be constant [19, 20]. Let us imagine a data set which lives in $\mathbb{R}^N$ dimensional hyperspace, having N_r data points, with each point $\vec{x}_i$ with the first k nearest neighbor distances $r_1(\vec{x}_i), r_2(\vec{x}_i), \cdots r_k(\vec{x}_i)$. When the above-said assumptions are satisfied the distribution function of the ratio of the second nearest neighbor and the first $f(\mu)$, follows a Pareto distribution [19]:

$$f(\mu) = \frac{I_d}{\mu^{I_d+1}}, \tag{1.1}$$

where μ is

$$\mu_i = \frac{r_2(\vec{x}_i)}{r_1(\vec{x}_i)}. \tag{1.2}$$

Now let's derive this formula for the Pareto distribution with our given functional assumptions. The volume encaptured between two nearest neighbors $j-1$ and j is given by

$$\Delta v_j = V_0(r_j^d - r_{j-1}^d). \tag{1.3}$$

where V_0 represent the volume of the d dimensional unitary hypersphere. If we take a hypersphere of radius $r_2(\vec{x}_i)$ of volume V, with point $\vec{x}_i$ being at the center, the probability of finding another point within the sphere is given by Poisson distribution in accordance to our first assumption [30]:

$$P(V) = \rho V e^{-\rho V}, \tag{1.4}$$

where ρ is the density of the points, which is constant as per our second assumption. Using the same logic if we ask what is the probability that Δv belongs to a spherical hypershell

$$P(\Delta v \in [v, v + dv]) = \rho e^{-\rho v} dv. \tag{1.5}$$

Now if we consider two hypershells Δv_p and Δv_q and let $\Gamma = \frac{\Delta v_q}{\Delta v_p}$, substituting Γ in the Eq. 1.5 we get the exact probability distribution function (PDF) for Γ

$$\begin{aligned} P(\Gamma) &= \int_0^\infty dv_p \rho e^{-\rho v_p} \int_0^\infty dv_q \rho e^{-\rho v_q} \delta(\frac{v_q}{v_q} - \Gamma) \\ &= \int_0^\infty dv_p \int_0^\infty \rho e^{-\rho(v_p + v_q)} \\ &= \frac{1}{(1+\Gamma)^2}. \end{aligned} \tag{1.6}$$

Now we have to find the connection between Γ and μ by taking $p = 1$ and $q = 2$

$$\Gamma = \frac{\Delta v_2}{\Delta v_1} = \frac{V_0(r_2^d - r_1^d)}{V_0 r_1^d} = \left(\frac{r_2}{r_1}\right)^d - 1 = \mu^d - 1. \tag{1.7}$$

giving us the PDF $f(\mu)$

$$f(\mu) = \frac{d\mu^{d-1}}{\mu^{2d}} = \frac{d}{\mu^{d+1}}. \tag{1.8}$$

as you can see we arrive at eq. 1.1 and $I_d = d$.

1.1.2 Implementation of Two-NN method

For every point $\vec{x}_i$, the first two nearest neighbor distances $r_1(\vec{x}_i)$ and $r_2(\vec{x}_i)$ and the ratio $\mu_i = r_2(\vec{x}_i)/r_1(\vec{x}_i)$ are calculated. Under the condition that the data set is locally uniform in the range of next-nearest neighbors, as shown in equation 1.8, the distribution function of μ is given by

$$f(\mu) = I_d \mu^{-I_d - 1}. \tag{1.9}$$

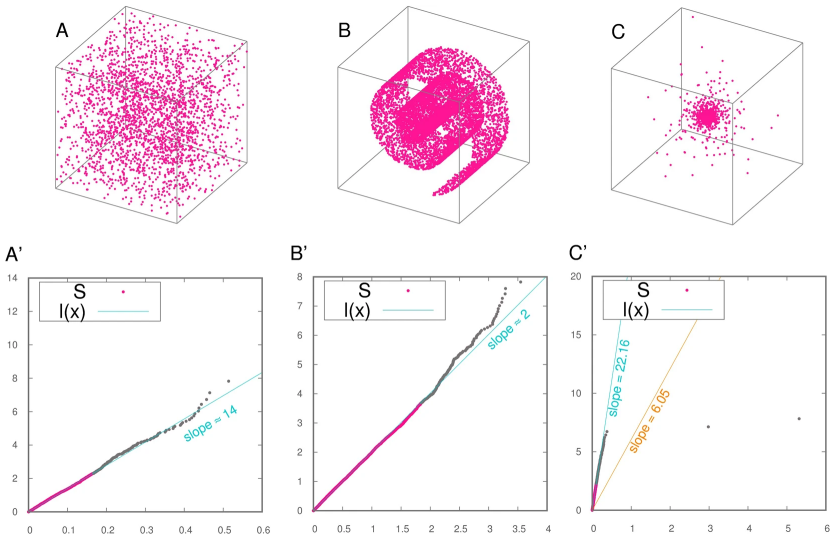

FIGURE 1.2: Different examples for calculating the I_d using different data set is illustrated, with each data set having 2500 points. The slope in the bottom panel of each column represents the I_d computed by the two-NN method. For A($A^{'}$): cube in dimension 14, the two-NN method retrieves the accurate I_d; B($B^{'}$): A data set drawn from a uniform distribution on a Swiss Roll embedded in a 3D space, where two-NN predicts the I_d, close to the theoretical one; C($C^{'}$): A Cauchy data set in $I_d = 20$ is analyzed where two-NN method fails to calculate the correct I_d, we elaborate further on this below. The figure is taken from ref. [19].

From the cumulative distribution function (CDF) of $f(\mu)$, denoted $P(\mu)$, we then obtain

$$I_d = -\frac{\ln\left[1 - P(\mu)\right]}{\ln(\mu)}. \tag{1.10}$$

In practice, one can use the empirical CDF, $P_{emp}(\mu)$, together with Eq. (1.10) to estimate the I_d by a linear fit of the points $\{(\ln(\mu), -\ln[1 - P_{emp}(\mu)])\}$, passing through the origin. This fit not only gives us the I_d but also a practical test for the reliability of the computation, if we diverge from the linear fit; it indicates the unstable nature of the calculation, in other words, tells us that the assumptions made for the two-NN method aren't fulfilled.

In Fig. 1.2, for the first two examples in (A) random points filling a cube and (B) uniform distribution on a 3D Swiss Roll, the two-NN method works

well. However, for the Cauchy data set the I_d calculated is strongly affected by very few points, because of the heavy-tailed nature of the data set which increases the probability of having $r_2 \ll r_1$, making the Pareto fit unstable. The resultant Pareto fit would have very few high values μ_i, so by discarding 10% of the highest μ_i, the I_d or the slope obtained is 22.16 which is more reliable and closer to the true I_d value.

In practice for the computation of I_d, we use the Euclidean distance metric. The distance calculations are done using *sklearn library* [31], but we observed that the distance calculation using the brute force method performed better than the ball-tree algorithm. We predict maybe at a larger dimension the ball-tree algorithm would outperform the brute force method.

The I_d is a scale-dependent quantity [17]. The scale of the data set, within the 2-NN method, is fixed by setting N_r (number of configurations taken) at temperature T as shown in fig. 1.3. We elaborate further in the following chapters. We note that, beyond 2-NN, it is possible to change scale by selecting a characteristic distance to associate a dimensionality to. In the context of this thesis, we are not interested in such a specific aspect, which might nevertheless also relate to physical properties corresponding to a dataset[32, 33, 34].

Implementation of Two-NN method:

- For each point $(\vec{x}_i)$ find the two nearest neighbor distances $r_1(\vec{x}_i)$ and $r_2(\vec{x}_i)$.
- Compute $\mu_i(\vec{x}_i) = \frac{r_2(\vec{x}_i)}{r_1(\vec{x}_i)}$.
- Compute the empirical cumulative distribution function $P_{emp}(\mu)$.
- Using eq. 3.5, calculate the slope for the plot $\{(\ln(\mu), -\ln[1 - P_{emp}(\mu)])\}$ which gives us the I_d.

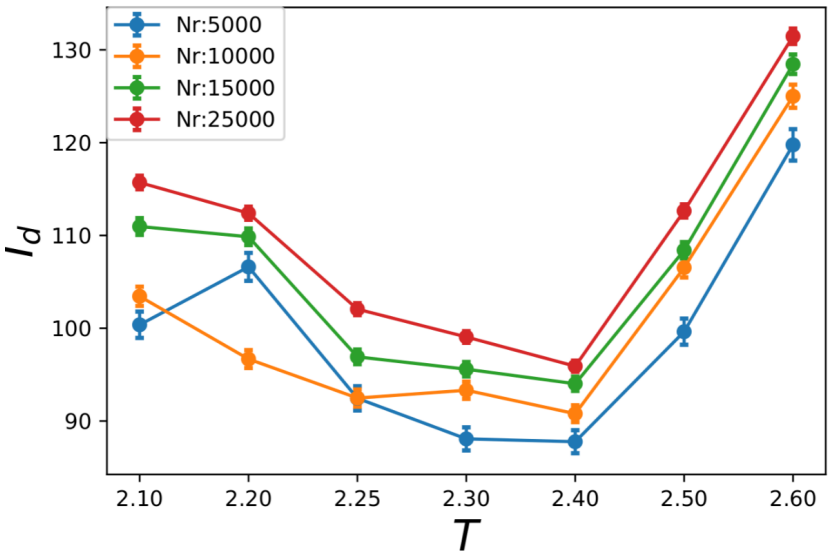

FIGURE 1.3: I_d calculation for 2D Ising with varying sample size N_r for system size $L = 40$.

1.1.3 Two-NN method applied in physics

In classical models

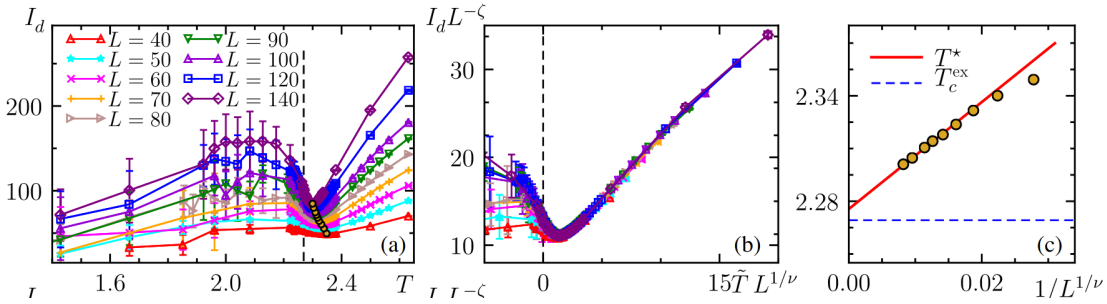

FIGURE 1.4: Second-order phase transition in 2D Ising model. (a) I_d as a function of temperature is plotted for different system sizes across the transition. (b) Data collapse of I_d on the finite-size scaling. (C) Finite-size scaling of the minimum temperature to extract T_c.Figure taken from ref. [32].

In classical physics, the *two-NN method* has been used in the context of

classical physics simulations and data analysis, particularly in areas like biophysics [35], molecular dynamics [36], statistical models [32, 37], and material science [38]. We will look closely at the work done in ref. [32], which is one of the first to study intrinsic dimension as the order parameter for different types of phase transition (first-order, second-order, and Berezinskii-Kosterlitz-Thouless) using the two-NN method for the raw configuration data generated from Monte-Carlo simulation at equilibrium. They find that I_d uniquely characterizes the transition regime for all considered cases. Through finite-size analysis of the I_d, they not only accurately identify critical points but also determine critical exponent ν for continuous transitions. Ref. [32] highlights how raw data sets exhibit distinctive signs of universal behavior without employing dimensionality reduction techniques, revealing a direct parallel between conventional order parameters in physical space and the intrinsic dimension in data space.

In fig. 1.4, one can clearly observe that the I_d shows a local minimum close to T_c indicating that the collective behavior of the system is effectively characterized by a small number of parameters. This observation implies that describing the system at the critical point requires significantly fewer details than describing it in its nearby regions, where additional information regarding the operators needed to deviate from criticality is necessary. The emergent simplicity at the critical point has implications for the structure of data set bring down I_d. This happens because of the approximate mapping between correlation function scaling and I_d close to criticality.

In quantum mechanical models

In the context of quantum many-body systems, the presence of intricate correlation patterns creates complex data structures. Typically, understanding the physical properties of different phases and phase transitions relies on correlation functions and their connection to observable response functions. The two-NN method has recently been employed to distinguish quantum phases of matter [39, 40]. In ref. [39], they have introduced a method to characterize

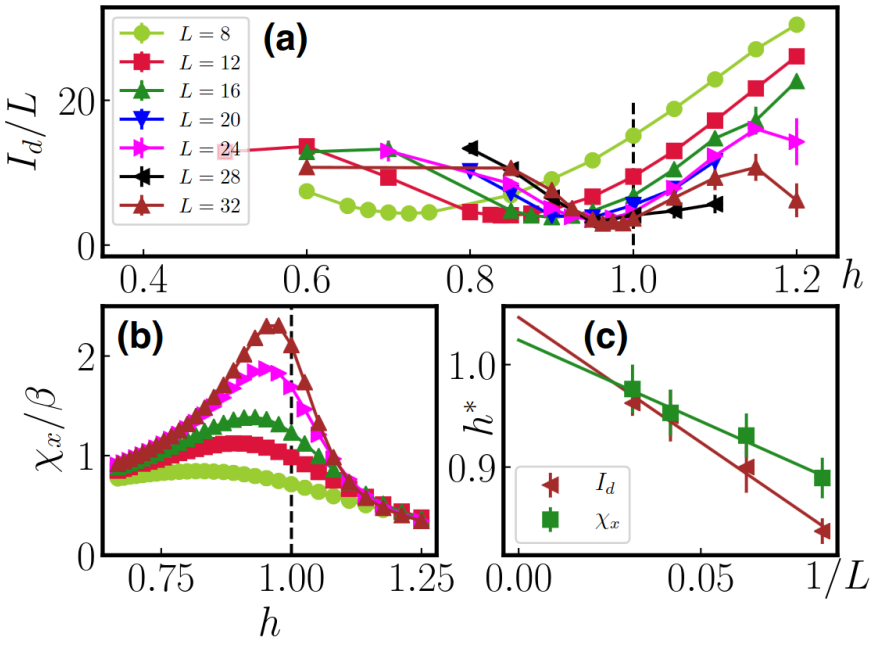

FIGURE 1.5: I_d for Quantum Ising model. (a) I_d as a function of h for varying system sizes. (b) Magnetization variance in the x direction as a function of the transverse field h. (C) Finite-size scaling for extraction of critical point h_c. Figure taken from ref. [39].

the correlation structure of quantum partition functions via the path integral manifold. To generate the data set for the transverse-field Ising model (TFIM), quantum Monte Carlo is being used. As we can observe in fig. 1.5, like the classical Isng case, there is a local minimum close to h_c implying an emergent simplification in the data structure.

1.2 Principal Component Analysis

With the rise of *big data*, the proliferation of large and complex data sets across various fields has prompted a growing need for robust tools to analyze high-dimensional data. Principal Component Analysis (PCA) is a dimensionality reduction method that stands out as one of the most popular techniques within the field of data science. At its core, PCA aims to transform a set of initially correlated variables from the original data set into a new set of uncorrelated variables termed "principal components" (PCs). This linear transformation is done with a primary objective: to retain as much information as possible from the original variables by maximizing the variance of the linear combination of the original variables. Following the PCA process, the resulting PCs are ordered, with the dominating few PCs capturing the bulk of the data set's inherent variations (statistical information) [41, 42]. In the following section, we will formally define PCA and how it can computed from an eigenvalue problem: from the singular value decomposition (SVD) of the (centered) data matrix.

1.2.1 Standard PCA

Let us now imagine a data matrix $\mathbf{X}[N_r, N]$, with each column N-dimensional vector, $\boldsymbol{x}_i \in \mathbb{R}^N$ representing the ith variable. We intend to find the linear combination of the vectors $\boldsymbol{x}_i$ such that maximizes the variance of

$$\sum_{i=1}^{N} w_i \mathbf{x}_i = \mathbf{X}\mathbf{w}, \tag{1.11}$$

with $\boldsymbol{w} = \{w_i\}_{i=1}^N$ are constants. But we know that variance of $\mathbf{X}\mathbf{w}$ can be written as

$$Var(\mathbf{X}\mathbf{w}) = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}, \tag{1.12}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix associated with the data set. Assuming that $\boldsymbol{w}$ is a unit-norm vector, the problem to maximize $\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}$ is very well defined;

we use a Lagrange multiplier λ, so now we have to maximize

$$\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} - \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1). \tag{1.13}$$

now when we differentiate with respect to $\boldsymbol{w}$, we get

$$\boldsymbol{\Sigma}\boldsymbol{w} - \lambda\boldsymbol{w} = \mathbf{0} \Leftrightarrow \boldsymbol{\Sigma}\boldsymbol{w} = \lambda\boldsymbol{w}. \tag{1.14}$$

we end up with the eigenvalue equation for the covariance matrix. As $\boldsymbol{w}_i$ are unit-normal vectors, so now we can write

$$Var(\boldsymbol{X}\boldsymbol{w}_i) = \boldsymbol{w}_i^T\boldsymbol{\Sigma}\boldsymbol{w}_i = \lambda_i\boldsymbol{w}_i^T\boldsymbol{w}_i = \lambda_i. \tag{1.15}$$

this tells us that the eigenvalues λ_i of the covariance matrix $\boldsymbol{\Sigma}$ are the variances for the corresponding eigenvector's linear combination $\boldsymbol{X}\boldsymbol{w}_i$ (also known as the principal components). The elements of PC are called the PC-score and the elements of $\boldsymbol{w}_i$ are called the PC-loading. The eigenvectors obtained from solving eq. 1.14, give us the orthonormal set of N vectors which successively maximize variance [41]. The most standard procedure for PCA is to work with the centered variables $\boldsymbol{x}_i^*$, given by

$$x_{ij}^* = x_{ij} - \bar{x}_i. \tag{1.16}$$

where $\bar{x}_i$ is the average of the variable $\boldsymbol{x}_i$, this gives us the column centered data matrix $\boldsymbol{X}^*$. Centering the data provides us with the geometric intuition for PCA, with the variances calculated from the *center of gravity* of the data set, as the origin is set at the middle of the data cloud in the hyperspace [43, 41]. On the other hand, centering is not mandatory; while working with unprocessed data is referred to as *uncentered PCA*. Uncentred principal components (PCs) are combinations of the original variables without centering that maximize non-central second moments, while ensuring that their crossed non-central second moments are equal to zero [43]. Though it might not be apparent both standard PCA and uncentred PCA behave similarly even when the $\bar{x}_i$

is large as observed in Ref. [43].

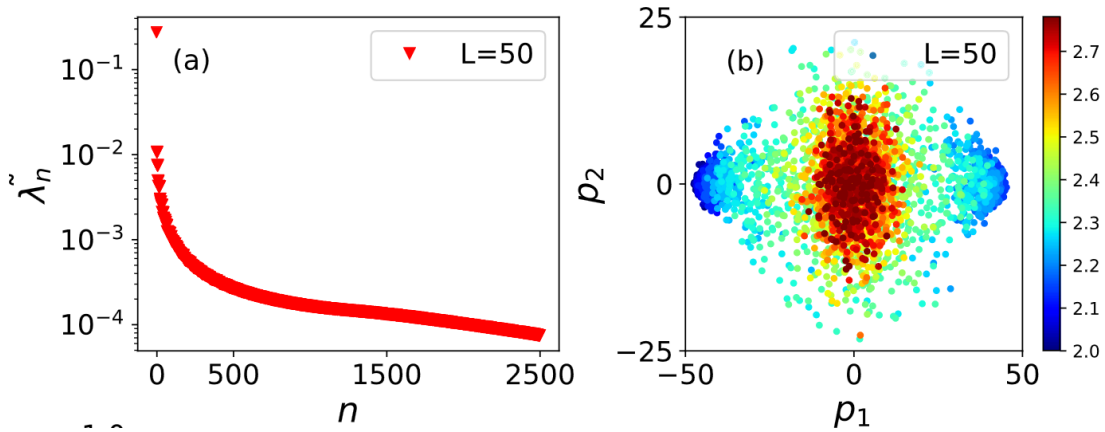

FIGURE 1.6: PCA findings for 2D Ising model.(a) The normalized eigenvalue spectrum of the covariance matrix. (b) The Ising configurations are mapped onto a two-dimensional plane defined by the two most significant principal components with the colored bar showing the temperature. Figure taken from ref. [44].

Once we have the processed data matrix $\boldsymbol{X}^*$, the covariance matrix can be written as [41, 42]

$$\boldsymbol{\Sigma} = \frac{1}{N_r - 1}\mathbf{X}^{*T}\mathbf{X}^* \tag{1.17}$$

as anticipated before, eq. 1.17 provides us the connection between the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$ with the *singular value decomposition* (SVD) of data matrix $\boldsymbol{X}^*$. Given any arbitrary matrix $\boldsymbol{M}[m,n]$, one can write the SVD decomposition as

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T. \tag{1.18}$$

where $\boldsymbol{U}[m,k]$ and $\boldsymbol{V}[n,k]$ are orthogonal matrices and $\boldsymbol{\Lambda}[k,k]$ is a diagonal matrix, with $k \leq min(m,n)$ being the rank of matrix $\boldsymbol{M}$ [45, 46]. The diagonal elements:$\{\lambda_i\}_{i=1}^k$ of $\boldsymbol{\Lambda}[k,k]$ are called the *singular values* of $\boldsymbol{M}$ and sorted in decreasing order. Now using eq. 1.18 in eq. 1.17, we get

$$\begin{aligned}\boldsymbol{\Sigma} &= \frac{1}{N_r - 1}\mathbf{X}^{*T}\mathbf{X}^{*} = (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T) \\ &\Rightarrow (N_r - 1)\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^T.\end{aligned} \tag{1.19}$$

as both $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices; $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \mathbb{1}$ and $\boldsymbol{\Lambda}^2$ representing the square of the diagonal elements which are nothing but the eigenvalues of the matrix $(N_r - 1)\boldsymbol{\Sigma}$, the λ_i from eq. 1.17. Therefore, PCA is analogous to performing an SVD on the centered data matrix $\boldsymbol{X}^*$. In fig. 1.6 example, we can see that the eigenvalue spectrum is sorted and the biggest eigenvalue is more than an order of magnitude larger than the next largest, indicating the effectiveness of the PCA-based dimension reduction which will be discussed further in the following section. Fig. 1.6(b) hints towards clustering in the data set at lower temperatures.

1.2.2 PCA-based dimension reduction

From the PCA, we obtain the eigenvalues for the covariance matrix $(N_r - 1)\boldsymbol{\Sigma}$ given by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0. \tag{1.20}$$

where k is the rank of $\mathbf{X}^*$, that is, $k \leq min(N_r, N)$. We then define the normalized eigenvalues, which is a standard measure to quantify the proportion of the total variance that is accounted for by the corresponding principal component. Namely,

$$\tilde{\lambda}_n = \frac{\lambda_n}{\sum_{n=1}^{k}\lambda_n}. \tag{1.21}$$

Till now we have not talked about the dimensionality reduction using PCA; to compute intrinsic dimension obtained by PCA (I_d^{PCA}), we use the normalized eigenvalues $(\tilde{\lambda})$. We define a an *ad hoc* cutoff $\epsilon \in (0, 1)$, then the I_d^{PCA} is defined as [26, 41, 32]

$$\sum_{n=1}^{I_d^{PCA}} \tilde{\lambda}_n \approx \epsilon. \tag{1.22}$$

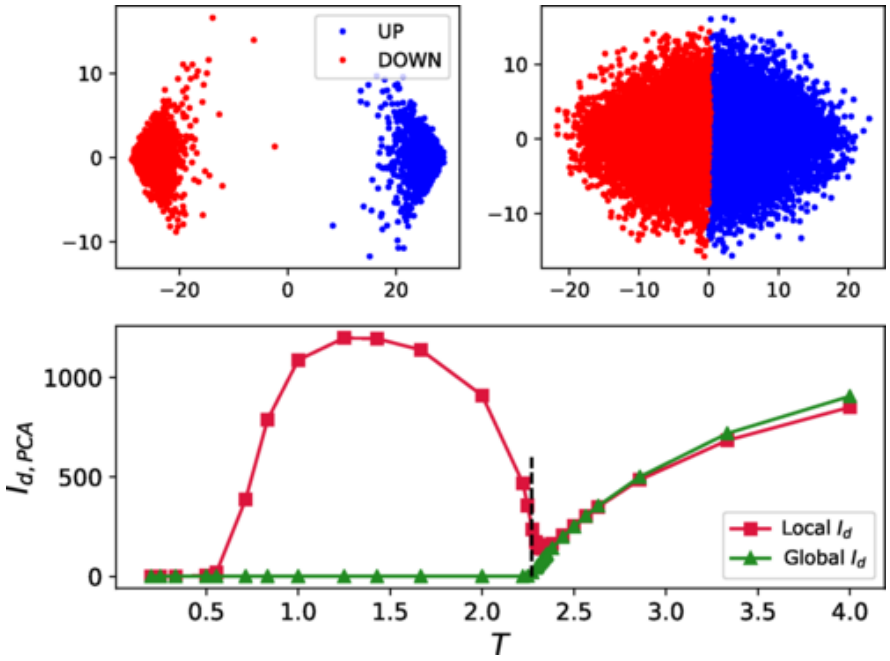

FIGURE 1.7: I_d^{PCA} for Ising model. (1)Upper panel: Clustering effects in the 2D Ising data set, show the projection of the two leading principal components for temperature below and above critical temperature (T_c) on the left and right panels respectively. For $T < T_c$, there are two clusters formed by configurations with positive(blue points) and negative total magnetization(red points) but the two clusters merge together when the system moves from the ordered phase to the disordered phase. (2) Lower panel: I_d^{PCA} considering the full Ising data set (global) and data set with configurations only positive magnetization(local). Figure taken from ref. [32].

One has to be very careful in choosing the cutoff ϵ, making sure that the estimated I_d^{PCA} is stable with respect to the changes ϵ. In practice, we find a range where I_d^{PCA} stabilizes to slight changes in ϵ which gives us the optimal ϵ.

We provide an example in fig. 1.7, taken from ref. [32], which studies how the I_d^{PCA} behaves across the second order phase transition for 2D Ising data set (generated by Monte-Carlo simulation). When the full data set is considered for the calculation of the I_d^{PCA} (global), below T_c; there is just one dominating PC, giving us $I_d^{PCA} \approx 1$ but as we cross T_c a sharp increase in I_d^{PCA} is observed, indicating the critical transition point.

The effectiveness of this approximation can be assessed by considering the variability linked to the retained by the PCs. Specifically, by comparing the sum of variances of the initial I_d^{PCA} variables with the trace (the sum of diagonal elements) of the covariance matrix. The information loss due to the dimensionality reduction is given by the *fidelity*

$$\mathcal{F}(I_d^{PCA}) = \frac{\sum_{i=1}^{I_d^{PCA}} \tilde{\lambda}_i}{Tr(\mathbf{\Sigma})} \tag{1.23}$$

We obtain fidelity $\mathcal{F}(I_d^{PCA}) \in [0, 1]$; when dimensionality reduction is exact, $\mathcal{F}$ goes to 1 [22].

Implementation for computing I_d^{PCA}:

- Compute the centered data matrix $\boldsymbol{X}^*$.
- Compute λ_i by performing SVD on the $\mathbf{X}^*$.
- Calculate the normalized eigenvalues $(\tilde{\lambda}_i)$ as per $\tilde{\lambda}_n = \frac{\lambda_n}{\sum_{n=1}^{k} \lambda_n}$.
- Define an *ad hoc* cutoff ϵ and using $\sum_{n=1}^{I_d^{PCA}} \tilde{\lambda}_n \approx \epsilon$ compute the I_d^{PCA}.

1.2.3 PCA entropy

The lack of a universally accepted criterion for selecting a cutoff ϵ in PCA dimensionality reduction can be unsatisfying. While we assert that our data-driven approach offers a way to impartially examine physical systems, it's worth noting that some prior knowledge, often derived from trial and error, is still required to make an informed choice regarding a suitable cutoff. This leads the way for us to work with PCA entropy S_{PCA} which does not require any external cutoff input and we work with the entire eigenvalue spectrum of the covariance matrix ensuring no loss of information. This measure is closely associated SVD entropy which has been applied in unsupervised methods for

feature selection in biology [47, 48]to quantify the complexity of ecological networks [49] and financial time series [50, 51, 52, 53, 54], and even in the characterization of the dimension of fractals [55].

We start with the eigenvalue decomposition of the sample covariance matrix. We can notice that the normalized eigenvalues $\tilde{\lambda}_n$ in Eq. (3.8) satisfy that (i) $\tilde{\lambda}_n \geq 0$ for all n , and (ii) $\sum_n \tilde{\lambda}_n = 1$ (by construction), we can follow Shannon's entropy formula [56] to define

$$S_{\text{PCA}} := -\frac{1}{\ln(k)} \sum_{n=1}^{k} \tilde{\lambda}_n \ln(\tilde{\lambda}_n). \tag{1.24}$$

The PCA entropy described in eq. (3.10) can serve as a measure of the correlations between the input variables in the data set. At one extreme, a fully correlated data set will have $S_{PCA} = 0$ as there will be only a single principle component ($\tilde{\lambda}_1 \approx 1$), and the rest will be negligible. The other extreme would be when the data set is random or completely uncorrelated then we will end up with a flat distribution for our eigenvalues implying maximum value for the S_{PCA}, for which $\tilde{\lambda}_n = 1/k$ for all n, we have $S_{\text{PCA}} = 1$.

Chapter 2

Many-Body Data Sets

In contemporary physics, data-driven approaches gaining popularity, steering into a new era of discovery and understanding. The rapid advancements in data collection and analysis techniques have led to the emergence of data-driven physics. From high-energy particle collisions at particle accelerators [57] to the study of cosmic phenomena through powerful telescopes [58], experiments by LIGO and VIRGO are generating huge data sets. In recent decades, with the rise of computing power with exascale machines as well as the refinement of algorithms, numerical techniques have gained widespread application within physics. These methods, alongside sophisticated data analysis, have significantly elevated the importance of numerical simulations. Our goal is to have an understanding of physical phenomena that relies on rigorous analysis of such huge data sets generated specifically for many-body physics in an assumption-free manner.

In this chapter, we will elaborate on how these data sets are generated concentrating on techniques that we will use later: Monte Carlo simulations, and tracking of configurations of experimental runs. Because of its versatility, the Monte Carlo method has gained immense popularity. Monte Carlo simulations are computational methods employed for simulating statistical systems, especially relevant for spin models. The primary objective of these simulations is to generate an ensemble of configurations that accurately represent the system, enabling the calculation of thermodynamic properties. Hence the name "*statistical sampling*" in the beginning, in 1949 Nicolas Metropolis coined the

term *Monte Carlo*. Monte Carlo(MC) simulations provide a practical alternative to solving systems analytically or exhaustively enumerating all possible configurations. [59, 60].

The capability of MC methods to deliver accurate results on equilibrium statistical mechanics crucially depends on its efficiency, which is strongly algorithmic dependent. A paradigmatic example of this are Markov chain Monte Carlo applied to spin systems. While simple MC had success in capturing ordered and disordered phases of Ising-like models, it was soon discovered that at continuous phase transitions (e.g. 2D Ising model on square a lattice), simulations suffer from severe critical slowing down (that is, the time to solution scales as a large power of the volume). The reason behind this phenomenon lies in its method of updating one site at a time, which proves effective for adjusting when the system has small correlation lengths but performs poorly with larger correlation lengths. As we approach the critical point, the correlation length diverges. Under such conditions, the algorithm tends to progress slowly through the configuration space of the model. As a consequence, this slow progression leads to high autocorrelations within the stochastic sequence of configurations, significantly decreasing the statistical reliability of the outcomes. To circumvent critical slowing down many cluster algorithms were proposed: Swensen-Wang [61], Worm algorithm [62, 63], and Wolff's cluster flip [64, 65]. In the coming sections, we delve into the fundamental principles that underlie both single and cluster flip MC simulations. These simulations will serve as our means of producing data sets.

The second route we employ is a statistical sampling of many-body experiments, from the lens of "in situ" imaging. The basic idea here is to collect large samples of snapshots, where each snapshot corresponds to a projective measurement of a large number of degrees of freedom. While such capabilities are by now established on several quantum simulation platforms, below, we will describe how this can be achieved in gases of bosonic particles, relevant to experiments with Bose-Einstein condensates.

2.1 Markov Chain Monte Carlo

In the realm of many-body theory in equilibrium, the partition function $\mathcal{Z}$ is arguably the most important and frequently employed entity for modeling a physical system, is given by

$$\mathcal{Z} = \sum_{\{x\}} e^{-\beta E(x)}, \tag{2.1}$$

where β is the inverse temperature $1/k_B T$ (k_B is Boltzmann constant), E is the energy of the configuration x and $\{x\}$ are the microscopic configurations of the system. It is the sum of all possible Boltzmann weights $P_B = e^{-\beta E(x)}$. But in statistical mechanics, we are usually interested in the thermodynamic average, so-called thermal expectation value, of some observable $\mathcal{O}$.

$$\langle \mathcal{O} \rangle = \frac{\sum_{\{x\}} \mathcal{O}_{\{x\}} e^{-\beta E_{\{x\}}}}{\mathcal{Z}} \tag{2.2}$$

but we know that the number of accessible states grows exponentially with the number of degrees of freedom in the system making this computation impractical. The solution to this problem comes in the form of *Markov process* and MC simulation done using the process is called Markov Chain Monte Carlo (MCMC).

A Markov process is a stochastic process in which the probability $P_x(t)$ follows the master equation

$$P_x(t+1) - P_x(t) = \sum_{x'} [P_x(t) W[x' \to x] - P_{x'}(t) W[x \to x']], \tag{2.3}$$

where $W[x \to x']$ is the transition probability. In eq. 2.3, on the right-hand side the first term describes incoming processes and the second part is the outgoing process. Markov chains have two properties:

1. Stationarity: Should not vary over time. For $t \gg 1$, we have $P_x(t) = P_x$

2. Memory-less evolution: Transition probability depends only on the last step.

this memory-less evolution is called Markovianity. We know that total probability is conserved which makes

$$\sum_{x} P_x(t) = 1 \tag{2.4}$$

2.1.1 Ergodicity and Detailed Balance

The motivation for us during the simulation is to generate the Markov chain iteratively such that the probability distribution approaches $P_B(x)$. That can only be achieved by imposing two more conditions on the Markov process, namely ergodicity and detailed balance. Ergodicity implies that given some configuration x with a finite number of iterations of the algorithm, any other configuration $x^{'}$ should be reachable. Essentially, the algorithm should never become trapped in a particular configuration, making it incapable of moving away from it. This signifies that the algorithm has the capacity to explore the entire configurational space.

After applying the stationarity condition in eq. 2.3, we get

$$\sum_{x^{'}} P_x(t) W[x^{'} \to x] = \sum_{x^{'}} P_{x^{'}}(t) W[x \to x^{'}] \tag{2.5}$$

which is called the *global balance condition*. An individual solution within the framework of global balance is to make each element of the sum equal, one at a time.

$$P_x(t) W[x^{'} \to x] = P_{x^{'}}(t) W[x \to x^{'}]. \tag{2.6}$$

we end up with so-called *detailed balanced condition*. Detailed balance and ergodicity together imply the convergence to the Boltzmann distribution. Detailed balance is not the necessary but a sufficient condition to ensure convergence to the Boltzmann distribution which gives us

$$\frac{W[x \to x^{'}]}{W[x^{'} \to x]} = \frac{e^{-\beta E(x^{'})}}{e^{-\beta E(x)}} = e^{-\beta \Delta E} \tag{2.7}$$

where ΔE is the change in the energy due to the update of the configuration.

2.1.2 Implementation of Metropolis algorithm

In ref. [66], first introduced an algorithm to make sure that the algorithm follows a detailed balance, which is called Metropolis' algorithm. One of the simplest way to enforce detailed balance is given by

$$W[x \to x^{'}] = \begin{cases} e^{-\beta \Delta E}, & \text{if } \Delta E > 0 \\ 1, & \text{if } \Delta E < 0 \end{cases} \tag{2.8}$$

Using the Ising model as an example, we provide the algorithm

Implementation of Metropolis algorithm:

- Pick a random site k.
- Compute the change in energy ΔE for flipping the spin at site k.
- Pick a random number with uniform dist. $r \in [0, 1]$
- Flip spin k only if $r < W[x \to x^{'}]$
- Go back to the first step.

The algorithm is simple and easy to implement. However, we must check the detailed balance for this case. In the case of $\Delta E > 0$, we have $W[x \to x^{'}] =$

$e^{-\beta\Delta E}$ but $W[x^{'} \to x] = 1$. Thus giving us

$$\frac{W[x \to x^{'}]}{W[x^{'} \to x]} = \frac{e^{-\beta\Delta E}}{1} = e^{-\beta\Delta E} \tag{2.9}$$

as required by the condition, it follows the detailed balance.

2.2 Wolff's algorithm

In the context of phase transitions within Ising models, we've pointed out a significant challenge that Monte Carlo (MC) simulations, known as "critical slowing down". This problem becomes harder as the system size grows, making it progressively challenging to generate statistically independent samples. In the subsequent discussion, we will discuss Wolff's algorithms that provide solutions to this issue [64, 65]. Although numerous algorithms are available for the 2D Ising model, we will concentrate on: the Wolff cluster algorithm. Ulli Wolff in 1989 generalized the notion of spin-flip for an arbitrary $O(n)$ σ model on the lattice and proposed a new Monte Carlo algorithm with the aim of solving the critical slowing down problem [65].

The implementation of Wolff's algorithm in the 2D Ising model

Implementation of Wolff's algorithm:

- Pick a random spin σ_k at site k, add it to the cluster $\mathcal{C}$.
- Whichever the neighbors of σ_k have the same sign as σ_k add them to the $\mathcal{C}$ with probability P_{add} (We will derive it in following section).
- Mark all the inspected sites to avoid visiting again.
- For every new spin added to $\mathcal{C}$, inspect the neighbors not in $\mathcal{C}$ and add them with probability P_{add}.
- Iterate until no new spin is added to the cluster.
- Once the cluster is built, flip all the spin in the cluster.

Now we must show the legitimacy of this Monte Carlo proposal, in order to prove that one has to show that the algorithm respects ergodicity and detailed balance. As single flip operations are allowed, one can imagine a situation where Wolff's algorithm mimics the Metropolis algorithm, therefore it has to be ergodic. We must bear in mind that the detailed balance is a sufficient condition in order to have a well-defined Monte Carlo integration procedure.

2.2.1 Detailed Balance

For our convenience let's take the example of the 2D Ising model. Let x and $x^{'}$ be two different configurations, have the transition probability $W[x \to x^{'}]$ then the detailed balance is given by

$$\frac{W[x \to x^{'}]}{W[x^{'} \to x]} = e^{-\beta \Delta E} \tag{2.10}$$

where ΔE is the change in energy after flipping the cluster. Let the border of the cluster has $m + n$ spins where m is the number of spins separating aligned spins and n separating the anti-aligned spins. The energy difference for $x \to x^{'}$

depends only on the boundary elements as that is the only place where changes from aligned to anti-aligned occur (and vice versa). For the Ising model

$$\Delta E = 2Jm - 2Jn = 2J(m-n). \tag{2.11}$$

In the same way for the cluster flip, the probability of not flipping the edges of the cluster

$$W[x \to x^{'}] = \prod_{\sigma \in \partial \mathcal{C}} (1 - p_{add}) = (1 - p_{add})^{m}. \tag{2.12}$$

then when we go from $x^{'}$ to X

$$W[x^{'} \to x] = \prod_{\sigma \in \partial \mathcal{C}} (1 - p_{add}) = (1 - p_{add})^{n}. \tag{2.13}$$

from this we ger

$$\frac{W[x \to x^{'}]}{W[x^{'} \to x]} = (1 - P_{add})^{(m-n)} \tag{2.14}$$

For detailed balance taking into account eq. 2.10 and eq. 2.11, we must have

$$P_{add} = 1 - e^{-2\beta J} \tag{2.15}$$

For the 2D Ising model, the detailed balance is satisfied for $P_{add} = 1 - e^{-2\beta J}$ making the Wolff cluster flips legitimate.

2.3 Spinor Bose-Einstein condensate: Data set

Quantum simulators provide potent tools for the exploration of strongly correlated quantum materials. Nevertheless, making sense of measurement results in such systems presents substantial difficulties. The motivation is to apply the non-parametric learning techniques that work efficiently in statistical models to characterize the data coming from different experiments in an unbiased manner.

To illustrate these ideas, we focus on a concrete example. We take the data

coming from ref. [67], where they present an approach for deducing the non-equilibrium quantum effective action using analog quantum simulators and validate the method through experimental demonstration, focusing on a quasi-one-dimensional spinor Bose gas out of equilibrium.

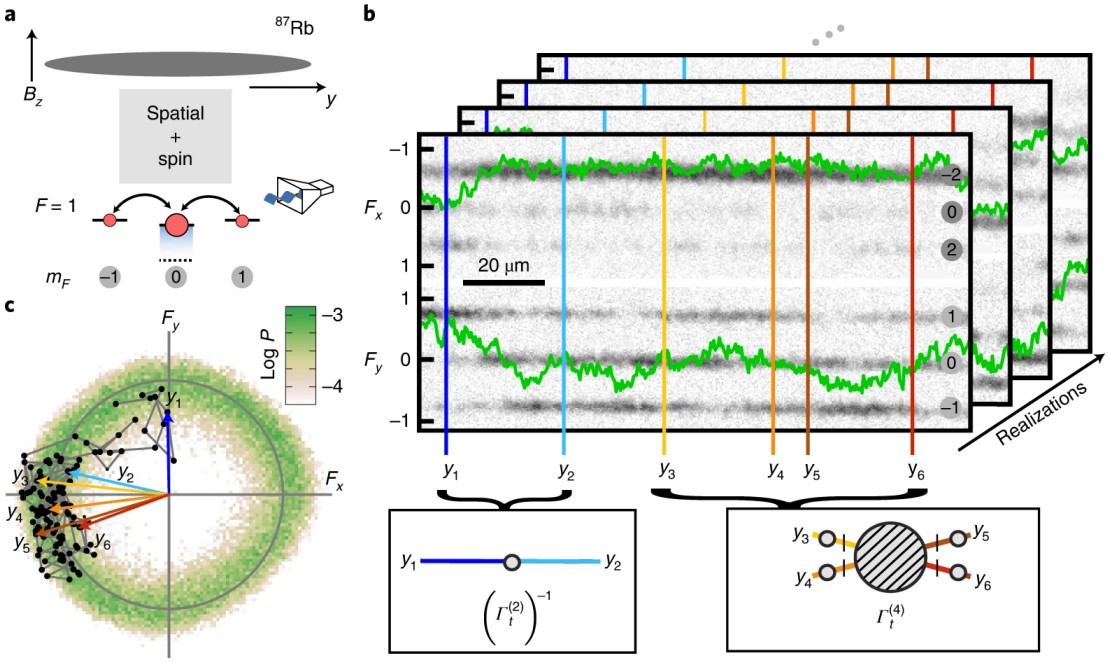

FIGURE 2.1: **Experimental platform and extraction of correlation functions.**. (a)^{87}Rb in magnetic field B_z. The spin dynamics take place in the $F = 1$ hyperfine manifold. (b) Spatially resolved snapshots of the dynamics are taken at different time intervals. From the atomic densities(grey shadings), F_x and F_y(Green lines) orthogonal spin projections were inferred as per eq. 2.16. (c) The distribution of $F_\perp$ from eq. 2.17, over 18s evolution time. Figure taken from ref. [67].

The quantum simulation is built on a spinor Bose–Einstein condensate (BEC) employing around 10^5 Rubidium (^{87}Rb) atoms in a quasi-one-dimensional elongated harmonic dipole trap [68]. (the data is taken from [67]; see this publication for further details on the experiment) This leads to spin dynamics in the $F = 1$ hyperfine manifold. The system is initialized with all atoms in the magnetic substate $m_F = 0$, but by changing a control parameter the spin-changing collision processes are brought into resonance. Because of this quench across

a quantum phase transition, the system is brought far from equilibrium. After the quench, at different times t the densities with spatial resolution along the longitudinal trap direction [69] n_{F,m_F} is measured in the state with hyperfine manifold F and magnetic sublevel m_F. With measured densities, one can simultaneously infer the two orthogonal spin projections F_x and F_y from the observed via

$$\begin{aligned} F_x &= (n_{1,+1} - n_{1,-1}) \,/\, (n_{1,+1} + n_{1,0} + n_{1,-1})\,, \\ F_y &= (n_{2,+2} - n_{2,-2}) \,/\, (n_{2,+2} + n_{2,0} + n_{2,-2})\,, \end{aligned} \tag{2.16}$$

At the final parameters of the quench, which places the system into the regime of the easy-plane ferromagnetic phase [70], these define the transverse spin field [71]

$$F_\perp = F_x + \mathrm{i}F_y \tag{2.17}$$

After approximately 1–3 s, the $F_\perp$ approaches the ground state distribution which is manifested as the formation of a ring in the transverse spin histogram. However, in this regime, the system is still highly excited and transverse spin phase excitations evolve dynamically in a self-similar fashion [72].

For the purpose of our investigation of the system, we use the data set formed by the bare densities n_α measured at different times t, with $\alpha = (F, m_F)$. For the measurement of every time slice, there are $N_r = 225$ independent realizations for each of the measured densities [see Eq. (2.16)]. Each density is linearly sampled at $N_s = 184$ spatial locations, which gives the number of *features* (entries) of each vector $\vec{\mathcal{N}}_i^\alpha(t)$. Combining all these into a data matrix would give us

$$\mathbf{M}^\alpha(t) = \{\vec{\mathcal{N}}_1^\alpha(t), \vec{\mathcal{N}}_2^\alpha(t), \ldots, \vec{\mathcal{N}}_{N_r}^\alpha(t)\}. \tag{2.18}$$

Therefore, for each time slice, we have a data matrix $\mathbf{M}^\alpha(t)[N_r, N_s]$. Further, we also consider *joint* data sets formed by concatenating horizontally subsets of the original data sets at a given time. More specifically, each row in a joint data set is formed by appending, one after the other(order of concatenation is irrelevant), single realizations of the observables of choice. Thus, for example,

if we combine all six densities, the resulting data sets will contain $N_s = 184 \cdot 6 = 1104$ columns and N_r rows. When needed, we will simply specify joint data sets by using the symbol of the corresponding observables joined by "$\|$".

Chapter 3

Non-parametric learning critical behavior in Ising partition functions: PCA entropy and intrinsic dimension

3.1 Introduction

Classical statistical mechanics revolves around the pivotal concept of the partition function [73]. Indeed, this quantity contains *all* relevant information about a statistical system at equilibrium, and thermodynamic quantities can be obtained from it. This quantity is defined as follows:

$$Z = \sum_{\{\mathcal{S}\}} e^{-\beta H(\mathcal{S})}, \tag{3.1}$$

where β is the inverse temperature, H is the Hamiltonian of the system and $\{\mathcal{S}\}$ are the microscopic configurations of the system. The probability that the system is in some particular state $\mathcal{S}$ is then given by Boltzmann law, namely

$$P(\mathcal{S}) = \frac{e^{-\beta H(\mathcal{S})}}{Z}. \tag{3.2}$$

A full knowledge of such probabilities would then allow us to compute any expectation value of physical quantities. However, a major problem in statistical mechanics is that, for a vast majority of cases, we only know the relative but not the absolute probability. In other words, we know $e^{-\beta H(\mathcal{S})}$ but not Z. Powerful computational techniques such as Monte Carlo methods [74, 75] and tensor networks [76, 77, 78], provide ways to circumvent this problem and allow for an efficient evaluation of expectation values of local observables such as two-point correlators, which are crucial to characterize phase transitions. Nonetheless, a significant part of the information encoded in the partition function may be left unexplored by such traditional approaches. This is an important point to consider, in particular, for systems that feature states of matter that cannot be described by 'typical' observables.

On the other hand, these methodologies—especially, Monte Carlo simulations,—by allowing a controlled generation of large volumes of synthetic microscopic snapshots of the systems of interest, offer a fresh perspective on many-body problems as *data structure* ones [32, 79, 44]. This has brought into play powerful tools from several fields such as high-dimensional statistics, inference, and machine learning, which are being adopted more and more in the physical sciences [80, 81, 82]. Among several methods that stem from these fields, *unsupervised learning* approaches have become prominent algorithms. Broadly speaking, such techniques aim at a characterization of the data through the understanding of underlying data relations. In condensed matter and statistical physics, such approaches have been mostly employed in the study of phase transitions and critical phenomena, including 2D [83, 44, 32, 84, 85] and 3D systems [86, 87, 88]. Additionally, there have been parallel efforts to estimate thermodynamic quantities such as the entropy—which are computationally very costly in traditional schemes—, using machine learning [89, 90] and information theoretic approaches [91], which work with reduced sampling. These previous works nonetheless call for methods that offer greater interpretability.

In this work, we put forward a theoretical approach for *learning* critical behavior in partition functions of classical spin models in an assumption-free manner. This is done by performing non-parametric statistical tests on large

data sets of many-body snapshots that are sampled according to a probability distribution as in Eq. (3.2). We showcase our approach by studying phase transitions in two- (2D) and three-dimensional (3D) Ising models.

In the first place, we study the *intrinsic dimension* (I_d) [17, 92, 19] of data sets in a range of temperatures across the featured phase transitions. This concept is relevant due to the following observations: (i) the points in a data set can normally be represented as points in a high-dimensional metric space, and (ii) such points may lie on a manifold, whose (intrinsic) dimension is lower than that of the embedding space, as correlations among input variables can induce a non-trivial structure on the data. Thus, the intrinsic dimension quantifies the minimum number of variables needed to faithfully describe the data. This concept has been widely used in data science for multiple applications, for example, in the fields of molecular science [19, 93, 94, 95] and image preprocessing [96, 97, 98, 99]. Recently, it has been realized that structural changes in the data associated with statistical mechanical problems can reveal critical phenomena. In terms of the data manifold, this can be unveiled as a reduction of the intrinsic dimension close to the critical point [32, 39]. In particular, Mendes-Santos et al. [32], showed this for the planar Ising model and other important 2D classical lattice models. Here, we extend the aforementioned work by considering the 3D Ising model, thereby analyzing the role of the physical dimensionality on the intrinsic dimension of the data manifolds. Concretely, we use two I_d-estimators: (i) the two nearest neighbor (TWO-NN) method [19]—a state-of-the-art estimator based on the distribution of the ratios between second- and first-nearest neighbor distances—, and (ii) a popular projection method known as principal component analysis (PCA) [41, 42]. We find that in general, it is harder to precisely determine the phase transition of the 3D Ising model through the intrinsic dimension, compared to the 2D case. We argue that this can be regarded as a non-trivial manifestation of the higher data dimensionality concomitant to the 3D model.

Motivated by these findings, we propose a second statistical test purely based on the eigendecomposition of the covariance matrix that is done within

PCA. More specifically, in analogy to Shannon's entropy [56], we define an entropy of the normalized eigenvalue spectrum of the covariance matrix. This quantity is dubbed *PCA entropy* (S_{PCA}). For the 2D case, we find a remarkable qualitative similarity to the exact thermodynamic entropy. In particular, S_{PCA} exhibits an inflection point close to the critical temperature. This is made more explicit by considering its derivative with respect to temperature, a quantity that resembles the heat capacity, which shows a clear divergence at the transition point. From the latter, we can perform a linear finite-size scaling analysis to estimate the critical temperature with less than 1% error. Similar results hold for the 3D model, using the same amount of data as for the intrinsic dimension estimation. Hence, the PCA entropy presents itself as a versatile tool to address higher-dimensional systems where a reliable intrinsic dimension estimation may become quite challenging. We note that similar spectral entropies have been introduced in the literature, mostly as unsupervised learning approaches for feature selection or as a measure of signal complexity. Applications range from biology [100, 47, 48, 101] and ecology [49] to stock market dynamics [50, 51, 52, 53, 54] and fractals [55]. Further, the PCA entropy has very recently been introduced as a theory-agnostic measure to rank operators in quantum simulators according to their relevance (information content) [2]. While some authors work with the spectrum of a covariance matrix as in the present work, some other authors define the entropy directly with the (normalized) singular values of the data matrix (the latter approach is sometimes dubbed *SVD entropy*). Our definition is based on the spectrum of the covariance matrix, since its normalized eigenvalues give the *proportion of total variance* accounted for by the principal components, and hence, a direct measure of the relevant information contained in the principal components [42, 80].

The chapter is organized as follows. In section 3.2, we provide a detailed description of our models and the corresponding data sets, as well as the methodology employed to create these data sets. Following that, in section 3.3, we focus on the intrinsic dimension estimation for the 3D Ising model. Subsections 3.3.1 and 3.3.2 delve into two different methods for estimating the intrinsic

dimension, namely, the TWO-NN method and PCA, respectively. We summarize the results and shortcomings of the methods to quantitatively capture the phase transition in the considered system. In section 3.4, we introduce the PCA entropy, S_{PCA}, and show its striking qualitative resemblance to the thermodynamic entropy of Ising models. This is exploited by extracting the transition point via a finite-size analysis of its numerical derivative with respect to temperature. Finally, we draw some conclusions and discuss further potential applications of our techniques in section 3.5.

3.2 Models and data sets

Before exploring the different tools considered in this work, we start by defining the models and the associated data sets that we consider for our study. In this work, we investigate the 2D and 3D classical Ising model with periodic boundary conditions having nearest neighbor interaction:

$$H = -\sum_{\langle i,j \rangle} S_i S_j, \tag{3.3}$$

where $S_i = \pm 1$ are the spin degrees of freedom defined on the sites of a square and cubic lattice for 2D and 3D, respectively [102, 103, 104]. The 2D Ising model is a paradigmatic model in statistical mechanics and beyond, and it is characterized by a second-order phase transition and $\mathbb{Z}_2$ spontaneous symmetry breaking. The system goes under an order-to-disorder phase transition at the critical temperature $T_c = 2/\ln(1+\sqrt{2}) \approx 2.269$ [102].

The exact solution of the Ising model on the simple cubic lattice is one of the long-standing open problems in rigorous statistical mechanics. The use of conformal bootstrap methods to calculate the critical exponents and critical point is still under active investigation [105, 106]. Nevertheless multiple numerical studies, especially Monte Carlo simulations, have been done to characterize the critical properties. Similar to the case of the 2D Ising model, the 3D system features a second-order phase transition, with the critical temperature predicted

at $T_c \approx 4.51$ [104, 107].

The data sets that we shall use for our subsequent analysis consist of equilibrium spin configurations of the systems introduced above. To form such data sets, we perform a stochastic sampling of the partition function of these models through Markov chain Monte Carlo (MC) simulations. Concretely, we use the Wolff cluster algorithm [65, 64], starting from the configuration with either all up spins or all down spins, chosen at random. Next, 30000 to 50000 'cluster flips' are performed for the system to equilibrate. After this, we collect $N_s = 10000$ state configurations, $\{\vec{x}^i \equiv (S_1^i, S_2^i, \ldots, S_N^i)\}_{i=1}^{N_s}$, where S_n^i is the spin variable at site n in the i-th realization, and $N = L^2$ for $D = 2$ or $N = L^3$ for $D = 3$, with L being the linear size of the system. Importantly, the collected configurations are separated by a number of cluster flips in the range of 1000 to 1500, so as to have as little correlation among them as possible (for a detailed discussion on decorrelation of sampled state configurations and autocorrelation times see Appendix 3.6.1).

For each temperature, we perform five independent MC simulations as described above, leading to an overall number of sampled configurations of 50000. This constitutes the total number of points in the data set at a given temperature. To perform statistics, we then use a subsampling algorithm [108, 109], wherein N_b 'batches' of data are formed by selecting, for each of them, $N_r = 10000$ configurations at random but without repetitions from the whole ensemble of 50000 sampled configurations. Each batch of data is then represented as a matrix with N_r rows (number of realizations) and N columns (number of degrees of freedom), that is, $\mathbf{X} = \{\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^{N_r}\}$. For more details on the subsampling technique, see Appendix 3.6.2.

3.3 Intrinsic dimension

High-dimensional data sets usually have hidden internal structures which essentially live on low-dimensional manifolds. Such manifolds can then be described—without losing relevant information—by a smaller number of features than the

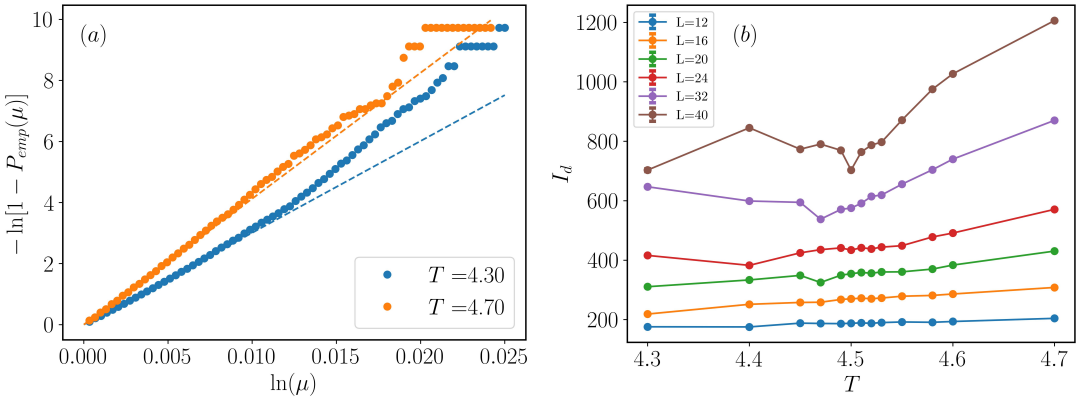

FIGURE 3.1: TWO-NN I_d estimation for the 3D Ising model.(a) Empirical cumulative distribution of data for $L = 20$ and temperature 4.30 and 4.70, where the dashed line shows the linear fit used to estimate I_d. With the scale used in this plot, the linear fit above corresponds to Pareto distribution. (b) I_d as the function of T for different L. While we expect the I_d to increase at high temperatures, and to drop as $T \to 0$, close to the transition point, this quantity features a local minimum, which becomes more apparent as the system size is increased.

embedding dimension. The reduced number of variables needed to describe the data is known as *intrinsic dimension*, I_d [17, 92]. This key observation is the reason for the great success of dimensional reduction algorithms. However, estimating the I_d of high-dimensional data sets is a problem that is far from trivial, since the corresponding data manifolds might be highly curved and twisted. Hence, this is an active field of research, with however some recent methodologies that have been shown to be able to mitigate the effects of curvature and inhomogeneities [19].

On the other hand, recent studies have shown the versatility and potential of the I_d as an unsupervised learning scheme to study critical phenomena in a variety of classical [32] and quantum [39] statistical mechanical models. Nonetheless, thus far, such efforts have been only carried out for low-dimensional systems. A systematic study of how volume can affect the estimation of the I_d of data sets associated to such many-body problems, hence,

remains an open question. In our work, we take a first step along this direction by systematically investigating the intrinsic dimension in the 3D Ising model. Specifically, we use two different methods to estimate the I_d of data sets, namely, TWO-NN [19] and PCA [41].

3.3.1 TWO-NN method

Although there are multiple ways to calculate the I_d, the two-NN method has recently gained popularity for its versatility in dealing with very high-dimensional data sets. This I_d-estimator only relies on the statistics of distances to each point's first two nearest neighbors. The method is rooted in computing the distribution function of neighborhood distances, which are functions of . For every point $\vec{x}_i$, the first two nearest neighbor distances $r_1(\vec{x}_i)$ and $r_2(\vec{x}_i)$ and the ratio $\mu_i = r_2(\vec{x}_i)/r_1(\vec{x}_i)$ are calculated. Under the condition that the data set is locally uniform in the range of next-nearest neighbors, it has been shown in Ref. [19] that the distribution function of μ is given by

$$f(\mu) = I_d \mu^{-I_d - 1}. \tag{3.4}$$

From the cumulative distribution (CDF) of $f(\mu)$, denoted $P(\mu)$, we then obtain

$$I_d = -\frac{\ln\left[1 - P(\mu)\right]}{\ln(\mu)}. \tag{3.5}$$

In practice, one can use the empirical CDF, $P_{emp}(\mu)$, together with Eq. (3.5) to estimate the I_d by a linear fit of the points $\{(\ln(\mu), -\ln[1 - P_{emp}(\mu)])\}$, passing through the origin as illustrated in Fig. 3.1(a).

In Fig. 3.1(b), we plot the estimated values of I_d as a function of temperature, for varying system size L, for the 3D Ising model. Though for small system sizes, there is no noticeable signal in the behavior of the I_d, as the system size is increased we observed that (i) at high temperatures the I_d monotonically increases as expected (since high-temperature Ising snapshots correspond to disordered (random) spin configurations), and (ii) most remarkable, around

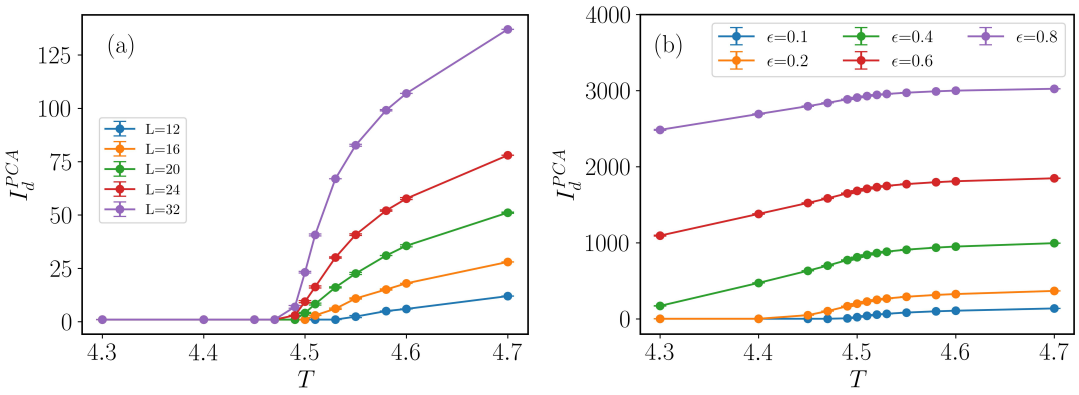

FIGURE 3.2: PCA-based I_d estimation for 3D Ising model.(a) I_d^{PCA} as a function of T for different system sizes with cutoff $\epsilon = 0.10$. For such an *ad hoc* cutoff, I_d^{PCA} abruptly drops to 1 below $T_c \approx 4.51$, while it rises above the transition. (b) I_d^{PCA} for $L = 32$, with varying cutoff ϵ [see Eq. (3.9)]. For sufficiently large values of ϵ, I_d^{PCA} does not drop to 1 just below the transition point. However, a signature of the transition can be observed as a visible change in the slope around $T_c \approx 4.51$.

the transition point, the I_d features a local minimum. As understood for 2D classical spin systems featuring continuous phase transitions, at the transition point the system becomes parametrically simpler due to universality, which in turn simplifies the concomitant data structure [32]. However, because of the higher dimensionality of the problem, unlike for the 2D Ising model, here we observe that the signal is weaker (for the accessed system sizes), making it significantly more challenging to get a reliable quantitative characterization of the phase transition through the I_d.

3.3.2 PCA-based I_d estimation

We now try a different approach to estimate the I_d of 3D Ising partition functions data sets. Specifically, we use the popular non-parametric technique known as PCA. The main idea behind PCA is that the essential information within a data set is contained in the variability of the data. Hence, one aims at finding the directions along which the data exhibit the highest variance. This

can be accomplished by means of a linear transformation of the set of coordinates [41]. The procedure to find such high-variance directions can be approached in different ways, for example, by diagonalizing the covariance matrix, or equivalently, by performing a singular value decomposition (SVD) of the data matrix [41, 42]. These approaches are briefly explained below. Each data set is represented by a rectangular matrix $\mathbf{X}\,[N_r, N]$, having the Monte Carlo snapshots as its rows, with N_r being the sample size. For convenience, we subtract the mean of each column from the entries of the columns to obtain the "centered data matrix", $\mathbf{X}^*$[44]. In this case, the sample covariance matrix can be estimated as [41, 42]

$$\mathbf{\Sigma} = \frac{1}{N_r - 1}\mathbf{X}^{*T}\mathbf{X}^*, \tag{3.6}$$

which is a $N \times N$ symmetric matrix ($\mathbf{X}^{*T}$ is the transpose of $\mathbf{X}^*$). It can be shown that the principal axis and their variance are defined, respectively, by the eigenvectors and eigenvalues of this matrix, which are obtained by solving the eigenvalue problem

$$\mathbf{\Sigma}\vec{w}_n = \lambda_n \vec{w}_n. \tag{3.7}$$

In practice, it is convenient to determine these quantities through an SVD of $\boldsymbol{X}^*$. In effect, one can readily show that the eigenvalues of $\mathbf{\Sigma}$ are proportional to the squared singular values of $\mathbf{X}^*$. Here we perform a full SVD on the matrix $\mathbf{X}^*$ using the package *scikit-learn*[31], which gives us $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$, where k is the rank of $\mathbf{X}^*$, that is, $k \leq min(N_r, N)$. We then define the normalized eigenvalues, which is a standard measure to quantify the proportion of the total variance that is accounted for by the corresponding principal component. Namely,

$$\tilde{\lambda}_n = \frac{\lambda_n}{\sum_{n=1}^{k} \lambda_n}. \tag{3.8}$$

The I_D^{PCA} can then be defined by choosing an *ad hoc* cutoff parameter ϵ for the integrated normalized spectrum of the covariance matrix [41]

$$\sum_{n=1}^{I_d^{PCA}} \tilde{\lambda}_n \approx \epsilon. \tag{3.9}$$

As discussed in recent works (see, in particular, Ref. [32]), the PCA-based I_d estimation differs from the TWO-NN one, in that the former can be regarded as a *global* estimator, while the latter is a *local* one. The implication of this fact is that rather than featuring a local minimum around the transition point, I_d^{PCA} drastically drops to 1 below the transition point [32]. We recover such behavior in the 3D case, too, specifically, for a value of the cutoff parameter of $\epsilon \sim 0.1$; see Fig. 3.2(a). We note that such a value is much smaller than the reported value in the case of 2D Ising [32]. We ascribe this as the clear signature of a non-trivial volume effect in the 3D case, which suppresses the dominance of the biggest contributing explained variance $\tilde{\lambda}_1$ (in data science, this is related to the so-called *curse of dimensionality* issue [110]). As we can observe in Fig. 3.2(b), for different values of the cutoff parameter ϵ, I_d^{pca} can vary significantly and require substantial fine-tuning to find the working window for the cutoff. Nevertheless, we note that even in those cases, a signature of the transition is still clearly visible through the form of a change in the slope of I_d^{PCA}.

In summary, the intrinsic dimension obtained via PCA can indeed host signatures of a phase transition, however, their visibility—and in fact, even their nature—is very sensitive to the choice of the cutoff parameter, signaling a degree of arbitrariness, and also making it challenging to obtain controlled estimates for the case of the 3D Ising model.

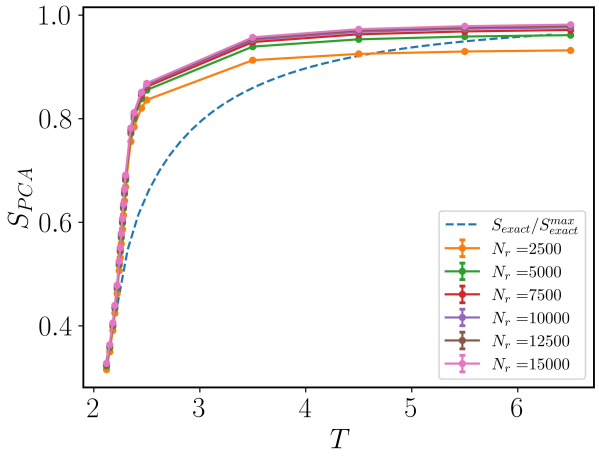

FIGURE 3.3: S_{PCA} comparison for varying samplesize. Comparison of S_{PCA}, for different sample sizes N_r, with the exact thermodynamic entropy per spin of the 2D Ising model with $L = 48$ as a function of temperature. Both entropies have been normalized such that their maximum possible value is 1 [see Eq. (3.10)].

3.4 PCA entropy

In order to circumvent the aforementioned difficulties in the unsupervised characterization of phase transitions in higher dimensional systems using I_d-based approaches, we now consider a complementary measure of data set complexity, namely, the *PCA entropy*, S_{PCA}. This quantity—and the closely related *SVD* entropy—has recently been employed in unsupervised schemes for feature selection in biology [47, 48, 101], to quantify the complexity of ecological networks [49] and financial time series [50, 51, 52, 53, 54], and even in the characterization of the dimension of fractals [55]. Further, very recently, this quantity has been employed as an unbiased metric to rank operators in quantum simulators based on their relevance (information content) [2]. It is one of the primary goals of this work to show that S_{PCA} can readily be used to characterize correlations and critical phenomena in other many-body problems as well. In the particular context of this work, we shall see that this quantity is less sensitive to volume effects, as opposed to the I_d estimators discussed

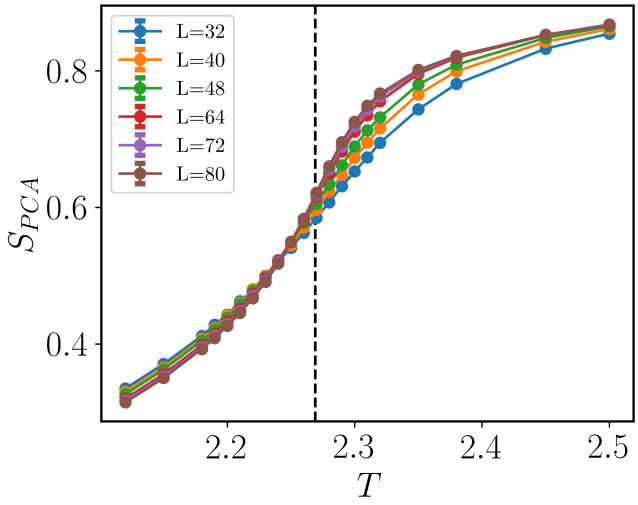

FIGURE 3.4: S_{PCA} as a function of temperature for different system sizes $L = 32 - 80$, for 2D Ising model. These plots exhibit a clear crossing point in the vicinity of the transition point, suggestive of a finite-size scaling analysis.

above. At the same time, as we will illustrate for the Ising models under consideration, this quantity bears a remarkable qualitative resemblance with the thermodynamic entropy. Importantly, the calculation of the PCA entropy is computationally very amenable.

The starting point to define the PCA entropy is the eigendecomposition of the sample covariance matrix. Specifically, by noticing that the normalized eigenvalues $\tilde{\lambda}_n$ in Eq. (3.8) satisfy that (i) $\tilde{\lambda}_n \geq 0$ for all n (as they are proportional to the squared singular values of $\mathbf{X}^*$), and (ii) $\sum_n \tilde{\lambda}_n = 1$ (by construction), we can follow Shannon's entropy formula [56] to define

$$S_{\text{PCA}} := -\frac{1}{\ln(k)} \sum_{n=1}^{k} \tilde{\lambda}_n \ln(\tilde{\lambda}_n). \tag{3.10}$$

In general, the PCA entropy in Eq. (3.10) can be used as a measure of the correlations among the input variables in the analyzed data set. Indeed, note that for an extremely 'correlated' data set, which under PCA can be fully described by a single principal component (i.e., $\tilde{\lambda}_1 \sim 1$, $\tilde{\lambda}_n \sim 0$, for $n \geq 2$), we get

$S_{\mathrm{PCA}} = 0$. Instead, for a fully 'uncorrelated' data set (e.g., a collection of independent random variables), for which $\tilde{\lambda}_n = 1/k$ for all n, we have $S_{\mathrm{PCA}} = 1$. Note that with the definition in Eq. (3.10), the maximum value that S_{PCA} can take is precisely 1.

Physically it is then clear that in the limits of $T \to 0$ and $T \to \infty$, for which the data sets are very 'ordered' and 'random-like', respectively, the behavior of S_{PCA} should, at least qualitatively, correspond to that of the thermodynamic entropy, that is, we expect S_{PCA} to vanish as $T \to 0$ and $S_{PCA} \sim 1$ as $T \to \infty$. That is exactly what we observe in Fig. 3.3, where we plot S_{PCA} for varying number of sample sizes (N_r), in the case of the 2D Ising model with $L = 48$. Furthermore, we compare those curves with the exact thermodynamic entropy per spin, which is computed using the explicit solution for finite square lattices with periodic boundary conditions (see, for instance, Refs. [90, 111, 112]). Note that the latter entropy is also normalized by its maximum possible value, in order to facilitate a direct comparison. This comparison suggests that S_{PCA} should asymptotically coincide with the thermodynamic entropy as $T \to \infty$ and $N_r \to \infty$. Apart from this limit, it is still quite remarkable the qualitative similarity between these two entropies, as already anticipated, even more so, as this is achieved even with reduced sampling, for example, $N_r = 2500$ in Fig. 3.3, which requires a very modest computational overhead.

In Fig. 3.4, we plot S_{PCA} for the 2D Ising model for different system sizes in a reduced range of temperatures around the transition point. In these and further calculations, we have fixed $N_r = 10000$. We note that S_{PCA} features a flex close to the transition point, which is immediately highlighted by the crossing of the curves when varying the system size L. This suggests a finite-size scaling analysis. To perform such an analysis in a more accurate way and allow for quantitative predictions, we compute the numerical derivative of S_{PCA}, which we approximate here by its symmetric difference quotient:

$$\frac{\delta S_{PCA}}{\delta T} := \frac{S_{PCA}(T + \Delta T) - S_{PCA}(T - \Delta T)}{2\Delta T}. \tag{3.11}$$

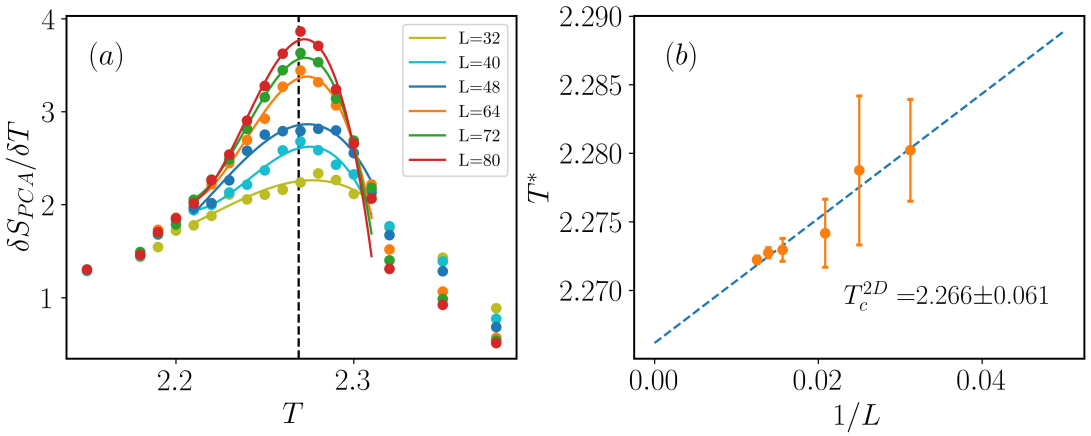

FIGURE 3.5: Calculating T_c^{2D} from S_{PCA}. (a) Plot of $\delta S_{PCA}/\delta T$ as a function of temperature for the 2D Ising model. The location of the flex in S_{PCA} is revealed by the peak in its derivative, occurring at $T^*(L)$. Solid lines show a smoothing curve of the data obtained via a standard smoothing spline function. (b) Linear finite-size scaling of the temperature where we get the maxima $T^*(L)$. This linear fit yields $T_c^{2D} = 2.266 \pm 0.061$.

This is shown in Fig. 3.5(a) for the 2D Ising model. We use a smooth spline approximation (using the function *splrep* from the package *scipy* [113]), to smooth out the curves and track the temperature at which they feature a local maxima, $T^*(L)$. The temperature window in which we perform the smoothing spline is $T \in [2.2, 2.31]$; solid lines in Fig. 3.5(a). This allows us to carry out a linear finite-size scaling analysis as shown in Fig. 3.5(b), which leads to an estimated critical temperature $T_c^{2D} = 2.266 \pm 0.061$, in excellent agreement with the exact value. In Fig. 3.5(b), we observe larger error bars for smaller system sizes. This is a consequence of the numerical derivative not having a prominent peak (and, in addition, the interpolation is more affected by fluctuations in the sampling when compared to larger lattices). For larger system sizes the peak in the derivatives is more pronounced and the calculation of $T^*(L)$ is less noisy; hence, we observe the sharp decrease in error bars. The error bars have been computed using the subsampling procedure explained in Sec. 3.2 and Appendix 3.6.2, averaging over 10 subsamples of data, each containing

$N_r = 10000$ data points. That is, for each batch of data we get a smooth spline approximation, and extract the corresponding local maxima T_i^*, we then compute the mean $\overline{T^*}$ and the subsampling error.

The corresponding results for the 3D Ising model are shown in Figs. 3.6 and 3.7. First, we note that, as opposed to the 2D case, the curves of S_{PCA} do not clearly cross as we vary L. This is most likely due to the fact that since we have fixed $N_r = 10000$, there will be a different normalization factor, $\ln(k)$, in the definition in Eq. (3.10) depending on whether L^3 is smaller of bigger than N_r. Indeed, if $N = L^3 < N_r$, then $k = N$, otherwise $k = N_r$. (For all the values of L considered in Fig. 3.4, it is always that case that $N = L^2 < N_r$, and hence, we always normalize the entropy by $\ln(N)$. Imposing a similar constraint in the 3D case would yield a bigger computational overhead or limit us in the system sizes that we can consider.) Yet, a similar analysis using the derivative of S_{PCA}, shown in Fig. 3.7(a), where we have used a temperature window $T \in [4.4 - 4.53]$ for the smoothing spline; solid lines in Fig. 3.7(a), allows us to perform a similar linear finite-size scaling analysis, shown in Fig. 3.7(b), yielding an

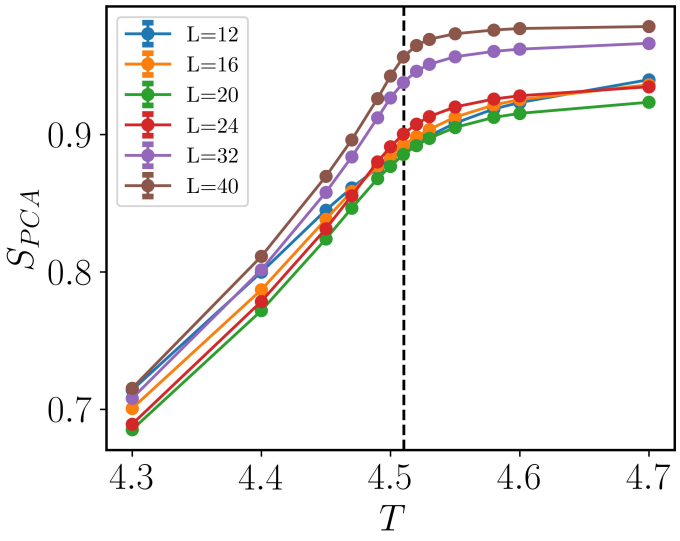

FIGURE 3.6: S_{PCA} as a function of temperature for different system sizes $L = 12 - 40$, for 3D Ising model.

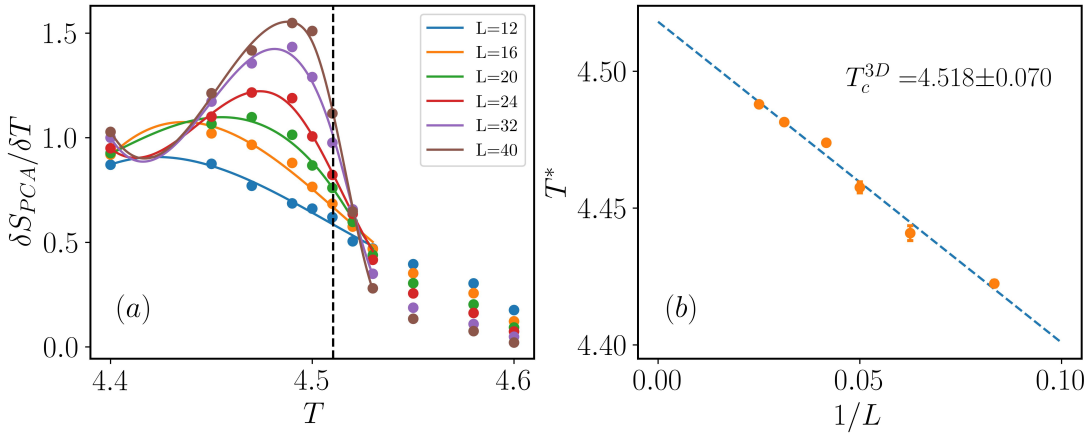

FIGURE 3.7: Extracting T_c^{3D} from S_{PCA}. (a) Plot of $\delta S_{PCA}/\delta T$ as a function of temperature for the 3D Ising model. The location of the flex in S_{PCA} is revealed by the peak in its derivative, occurring at $T^*(L)$. Solid lines show a smoothing curve of the data obtained via a standard smoothing spline function. (b) Linear finite-size scaling of the temperature where we get the maxima $T^*(L)$. This linear fit yields $T_c^{3D} = 4.518 \pm 0.070$.

estimated critical temperature $T_c^{3D} = 4.518 \pm 0.070$, once again in very good agreement with the reported value in the literature.

Finally, we note that the smoothing splines shown in Figs. 3.5(a) and 3.7(a), were done using a smoothing condition parameter s [113], so that $\sum_i (g_i - y_i)^2 \leq s$, where $g(x)$ is the smoothed interpolation of (x, y). In practice, we found that setting s to less than 1% of the maximum of y gives stable results, and concretely, we set $s = 0.005$.

3.5 Conclusions and outlook

In summary, we have introduced a theoretical framework to learn critical behavior in partition functions of classical systems using non-parametric unsupervised approaches. We have showcased our methods by studying phase transitions in classical Ising models in 2D and 3D rectangular lattices, harvesting thermal configurations from MC simulations. In the first place, we have

unveiled the role of volume in the estimation of the intrinsic dimension of data sets of thermal MC configurations. The intrinsic dimension is widely used in machine learning and has recently been applied in unsupervised studies of critical phenomena in 2D classical systems. We explored this property for the first time in 3D systems. We found that, while it is still possible to detect the transition point with reasonable accuracy through changes in the behavior of this quantity as a function of temperature, in general, its estimation becomes much more challenging than in the 2D case. The latter holds when using both local and global estimators such as the TWO-NN method and PCA, respectively. Further, this observation is very likely a direct manifestation of what in data science is known as the *curse of dimensionality* [110]. In the quest to overcome this difficulty, we have then introduced the concept of *PCA entropy*—a "Shannon entropy" of the normalized spectrum of the covariance matrix. This and related spectral entropies are widely used in unsupervised approaches for feature selection tasks as well as a measure of signal complexity. Here, we have applied this quantity for the first time to data sets of statistical mechanics systems and found a striking qualitative similarity with the thermodynamic entropy of the Ising model, exhibiting in particular a flex around the transition point, both for the 2D and 3D cases. This allows for a very accurate estimation of the critical temperature (with less than 1% error) by a conventional finite-size scaling analysis. Further, we have argued how the PCA entropy can asymptotically recover the thermodynamic entropy while being computationally efficient and interpretable—as opposed to other machine learning approaches—due to its own definition.

Several interesting questions remain as directions for future research. In particular, it would be very interesting to see the scope of the PCA entropy in the study of different kinds of phase transitions such as Berezinskii-Kosterlitz-Thouless (BKT) transitions. In this respect, analyzing whether and how the PCA entropy is sensible to the effects of topology is a question well deserving of attention. Besides, characterizing the limitations of the intrinsic dimension due to volume effects in different systems is another critical question to be explored. Additionally, our methods can readily be applied to learn path

integrals of quantum statistical systems, thereby complementing and extending previous theoretical works [39]. Finally, the analysis of experimental data sets associated with many-body problems is also immediately within reach, as already exemplified in recent related works [114, 2]. Along a separate route, it is essential to mention that the dimensional analysis performed here indicates that manifolds describing partition functions are in fact very rich and correlated: a very promising route to unfold such correlations is provided by network theory—that we are illustrating, in the context of Ising partition functions, in a parallel work [115].

3.6 Appendix

3.6.1 Analysis of the decorrelation of state configurations via I_d and S_{PCA}

In this appendix, we elaborate on how we minimize the correlation between the configuration extracted from the Monte Carlo simulation. In order to make sure that we have attained the desired data set with decorrelated configurations, we study I_d and S_{PCA} as the function of sampling interval d_s, the number of Wolff's cluster flips between two consecutive configurations saved. For all the calculations below we have $N_r = 5000$ with the configurations taken from the same Monte Carlo simulation and averaged over 5 realizations. The system sizes are fixed, $L = 48$ for 2D and $L = 24$ for 3D Ising.

In Fig. 3.8, we can observe that after an initial increase in I_d with d_s, the I_d value saturates and stabilizes within error bars for increments of d_s. The point after which the I_d saturates indicates the minimum value of d_s required to build the uncorrelated data set. We call such a value *decorrelation time* and denoted $d_s^\star$. This decorrelation time increases with temperature: for 2D at $T = 2.27$, $d_s^\star \simeq 10$ seems to be enough to decorrelate the configurations, but for $T = 2.35$ we need $d_s^\star \simeq 40$. We observe a similar trend for the 3D case with increasing temperature requiring higher times: at $T = 4.60$, we get $d_s^\star \simeq 500$. In practice,

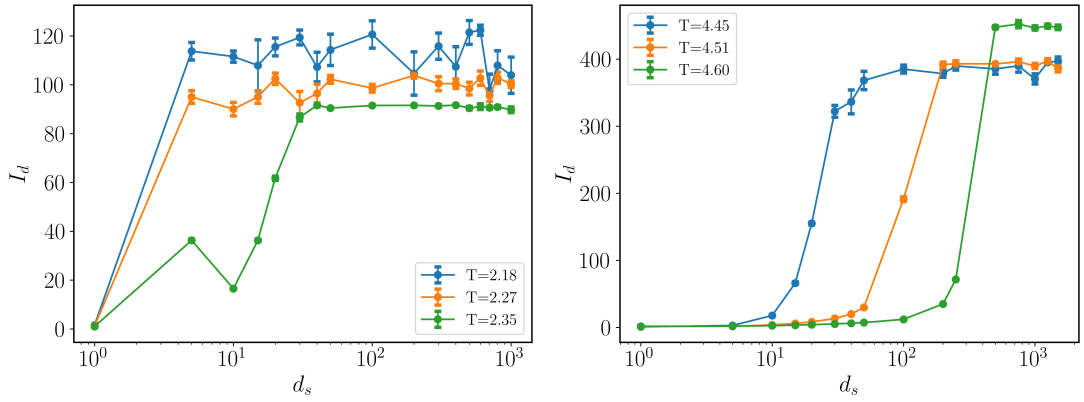

FIGURE 3.8: I_d as a function of the sampling interval d_s, for 2D and 3D Ising partition function data sets, respectively, at different temperatures. After some transient behavior, the I_d saturates at some given value and does not change further. This insensitivity with respect to the sampling interval signals the point after which configurations sampled during the MC simulations are essentially uncorrelated from each other. This defines the *decorrelation time* $d_s^\star$ (see main text).

however, we set a final sampling interval to be at least two or three times $d_s^\star$, which for the latter case, for example, corresponds to $1000 - 1500$ cluster flips in between sampled state configurations.

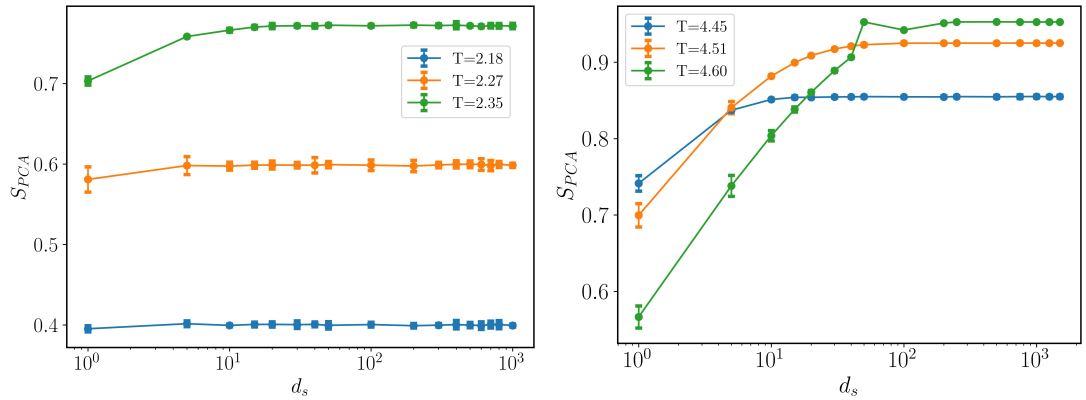

FIGURE 3.9: S_{PCA} as a function of the sampling interval d_s for 2D and 3D Ising partition function data sets, respectively, at different temperatures. We observe compatible values of the $d_s^\star$, with those estimated from the I_d analysis; c.f. Fig. 3.8.

In Fig. 3.9 what we observe for S_{PCA} complements the findings from I_d, with S_{PCA} rising and eventually saturating at some given value as d_s is increased. We note that, in general, the values of $d_s^\star$ that can be read out from the latter plots are compatible with those estimated using the I_d, both in 2D and 3D.

We note that the decorrelation time $d_s^\star$ is an intrinsic property of the specific algorithm utilized to carry out the MC sampling. This is analogous to the *autocorrelation time*, which is a paramount quantity to analyze in any MC simulation. Indeed, the key difficulty in (dynamic) MC is that the successive states in the underlying Markov chain are correlated, naturally increasing the error of estimates [116]. For some given observable, for example, the magnetization M, the autocorrelation function $C(t)$ as a function of the MC time t is given by

$$C(t) = \frac{\langle M_j M_{j+t} \rangle - \langle M \rangle^2}{\langle M^2 \rangle - \langle M \rangle^2}, \tag{3.12}$$

with j denoting some reference time, which we can choose arbitrarily since at equilibrium time translational invariance holds. In the above definition, we use

$$\langle M^\alpha \rangle = \frac{1}{N_t} \sum_{i=1}^{N_t} M_i^\alpha, \quad \langle M_j M_{j+t} \rangle = \frac{1}{N_t - t} \sum_{i=1}^{N_t - t} M_i M_{i+t}, \tag{3.13}$$

where N_t is the total number of MC steps.

For well-formulated algorithms, it is typically expected that the autocorrelation function introduced above will decay exponentially with t, that is,

$$C(t) \simeq \exp(\tau / t), \tag{3.14}$$

where τ is the autocorrelation time of the observable in the given algorithm. To be more precise, this time is called the exponential autocorrelation time τ_{exp}. A second autocorrelation time is so-called *integrated autocorrelation time*

(IAT), τ_{int}, which determines the statistical errors in the MC estimates of observables [116]. The latter can be estimated as follows

$$\tau_{int}(W) = \frac{1}{2}\sum_{t=1}^{W-1} C(t) + R(W), \tag{3.15}$$

with

$$R(W) = \frac{C(W)}{1 - \frac{C(W)}{C(W-1)}}, \tag{3.16}$$

that shall converge fast for $W \gg 1$.

We computed the IAT for the temperatures above for the 2D and 3D systems, using 10000 successive configurations after the equilibration. We found $\tau_{int} \simeq 33, 40$, and 48 for $T = 2.18, 2.27$, and 2.35 respectively in the 2D case. For the 3D case we get $\tau_{int} \simeq 46, 51$, and 54 for $T = 4.45, 4.51$, and 4.60 respectively.

Whether or not the autocorrelation time of observables and the decorrelation time estimated via the I_d and S_{PCA} analyses can be related to each other is a question well deserving a more in-depth exploration, which however we leave for future research. Nevertheless, we should mention that the used decorrelation times, as defined above, are a crucial piece of information to ensure the reproducibility of the results discussed in this work.

3.6.2 Subsampling

In this appendix, we describe the subsampling algorithm used to perform statistics on the collected data and establish the corresponding error bars.

Given a data set with a total number of points N_T: $\mathbf{X} \equiv \{\vec{x}^1, \ldots, \vec{x}^{N_T}\}$, we repeatedly compute a quantity of interest ϕ on N_b 'batches' (subsamples) of data, which are obtained by randomly drawing samples of size N_r *without* replacement from the finite population $\{\vec{x}^1, \ldots, \vec{x}^{N_T}\}$. We denote such estimates as $\phi(\mathbf{X_{fi}})$, with $\beta = 1, \ldots, N_b$. From these estimates, we can compute the sample mean:

$$\overline{\phi} = \frac{1}{N_b}\sum_{\beta=1}^{N_b} \phi(\mathbf{X_{fi}}). \tag{3.17}$$

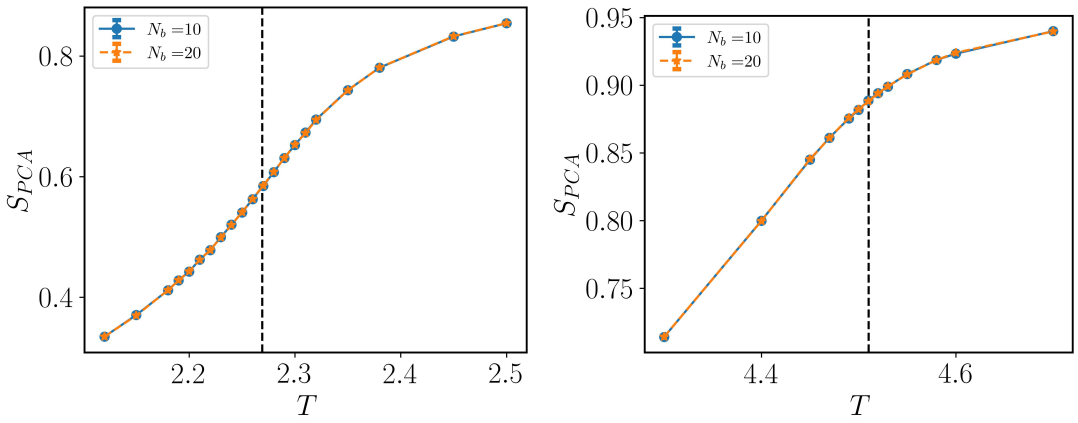

FIGURE 3.10: S_{PCA} as a functions of temperature different value of N_b, for 2D and 3D Ising partition function data sets. The system size for 2D is $L = 32$ and for 3D $L = 12$.

The associated standard error can be estimated as [108, 109]

$$SE \approx \sqrt{\frac{N_r}{N_T - N_r}} \times \sqrt{\frac{1}{N_b} \sum_{\beta} (\phi(\mathbf{X}_\beta) - \overline{\phi})^2}. \tag{3.18}$$

This formula is known as the (stochastic) delete-d Jackknife standard error estimator (with $d = N_T - N_r$), which is usually employed within subsampling schemes [108]. We note that this method is also related to the bootstrap method [109], with the main difference that samples are drawn without replacement. The latter fact is crucial, for example, when estimating the I_d through the TWO-NN algorithm, which works under the assumption of no repetitions in the considered data points (if repetitions occur, different estimators based on discrete distances can be employed [117]).

Finally, it can be shown that under adequate conditions the distribution of $\phi(\mathbf{X}_\beta)$ will converge to the sampling distribution of ϕ. In particular it is required that $N_r \to \infty$ as $N_T \to \infty$, but with $N_r/N_T \to 0$. In practice, the choice of the parameters above is data-dependent. Here, we have $N_T = 50000$, and found consistent results with the choices $N_b = 10$ and $N_r = 10000$ (unless otherwise specified), as mentioned in the main text.

In Fig. 3.10, we check the effects on S_{PCA} while changing N_b for some fixed system size. We find negligible change in S_{PCA} value for changing $N_b = 10$ to $N_b = 20$ in the case of both 2D and 3D Ising partition function data sets. We can observe that in Fig. 3.10 the S_{PCA} values lay on top of each other for varying N_b.

Chapter 4

Data-driven discovery of relevant information in quantum simulators

4.1 Introduction

Recent remarkable advances in highly controlled synthetic quantum devices have revolutionized the study of strongly correlated systems [118, 119, 120, 121, 122, 123]. A key element of many of such platforms is their capacity to produce large data sets of many-body snapshots, for example, via generalized projective measurements of the entire wave function [69]. However, the analysis of such outcome poses in general serious challenges, which typically force us to rely on assumptions for certain quantities, disregarding part of the information content of the generated data—in data science language, a dimensional reduction with an uncontrolled loss of information. A particularly important problem is the identification of the most informative observables to describe such quantum many-body systems—a paramount task at the core of quantum field theory [124, 125], that is even more daunting for systems driven out of equilibrium. To address this, one needs to develop methods to process the maximum amount of information in quantum simulator output, which are able to identify relevant features—and thus degrees of freedom—emerging from the underlying physical system, without making any assumption nor uncontrolled dimensional reduction.

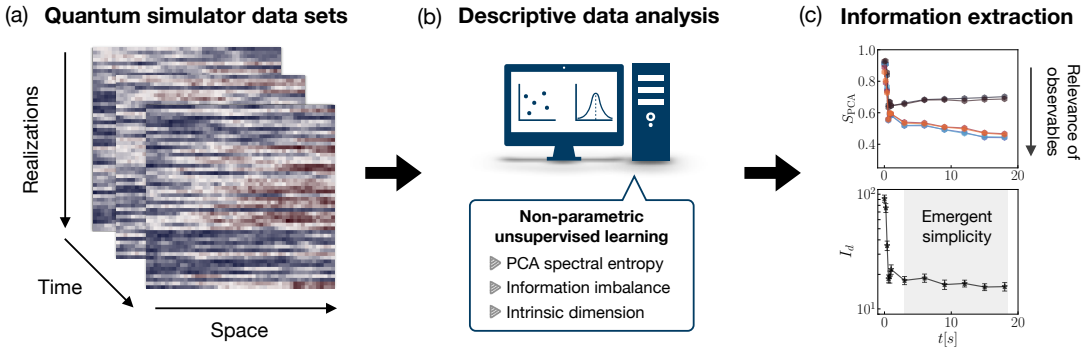

FIGURE 4.1: Assumption-free unveiling of relevant information in quantum simulation. (a) We start from snapshots of a many-body system, which are represented as 2D arrays at different times. At a fixed time, each row corresponds to a different realization, while each column is a different data feature, e.g., the atomic density in a given magnetic substate at a given spatial location. (b) Using non-parametric unsupervised learning tools, we perform an exploratory analysis to uncover interesting features of the data, without making any assumptions. (c) From this description, we infer relevant properties of the physical system: (Upper panel) By quantifying the information content of different observables, the spectral entropy of the sample covariance matrix provides a metric to rank the latter according to the strength of the correlations captured by them, thence guiding the identification of relevant (i.e., most informative) degrees of freedom. (Lower panel) After a quick fall to relatively small values, the intrinsic dimension of data sets features a long, stable plateau as a function of time (shaded region), providing a lower bound for the timescale after which the dynamics may become simpler and be captured by universal scaling.

In this work, we introduce a theoretical framework for data-driven information discovery in quantum simulation, which is schematically illustrated in Fig. 4.1. We start by considering collections of independent quantum simulator snapshots, which resolve, for example, the dynamics of a many-body system in space and time [Fig. 4.1(a)]. Such data sets are characterized using non-parametric unsupervised learning methods [Fig. 4.1(b)]. Finally, from this system-agnostic and unsupervised description of the data, we infer relevant information for the physical system under study [Fig. 4.1(c)].

This framework is based on three techniques: (i) spectral entropies calculated from a principal component analysis (PCA) of the data, (ii) the information imbalance between a subset and the full set of observations, and (iii) the intrinsic dimension of the concomitant data manifolds. These tools which quantify—from different angles—the information content and correlations in the data, have found several successful applications in various fields, such as chemical and biomolecular science [47, 48, 101, 126, 127, 128, 93, 94, 21], ecology [49], stock market dynamics [50, 51, 52, 129, 53, 54], and image analysis [96, 98, 99, 130].

To demonstrate the capabilities of our approach we apply it to experimental data of a spinor Bose-Einstein condensate (BEC) [131]: we evaluate the full set of experimentally measured densities without knowledge of the post-processing steps which are necessary in order to infer the relevant spin variables from them. Our main results are as follows: (I) PCA spectral entropies and information imbalance allow for a theory-agnostic determination of the most informative measured observables. The predictive power of these methods is demonstrated by showing that they can also unveil combinations of the measured densities, which are key to describe the spin structure of the system [72, 131, 70]. (II) The behavior of the intrinsic dimension as a function of time, displays a rapid decay to significantly smaller values, after which it features very long, stable plateaus. As argued below, this observation is in strong agreement with the formation of spin structure and the emergence of self-similar dynamics [132, 72, 133, 131, 134].

4.2 Quantum simulation on a spinor BEC and data sets

We consider the dynamics realized by a BEC of ^{87}Rb in the $F = 1$ hyperfine spin ground state manifold confined in a quasi-one-dimensional elongated harmonic dipole trap (the data evaluated here are taken from [131]; see this publication for further details on the experiment). The system is initialized with all atoms in the magnetic substate $m_\mathrm{F} = 0$. By instantaneously changing a control parameter we tune spin-changing collision processes into resonance. This procedure implements a quench across a quantum phase transition which brings the system far from equilibrium. For different times t after the quench, we simultaneously infer the two orthogonal spin projections F_x and F_y from the observed densities with spatial resolution along the longitudinal trap direction [69] via

$$\begin{aligned} F_x &= (n_{1,+1} - n_{1,-1}) \,/\, (n_{1,+1} + n_{1,0} + n_{1,-1})\,, \\ F_y &= (n_{2,+2} - n_{2,-2}) \,/\, (n_{2,+2} + n_{2,0} + n_{2,-2})\,, \end{aligned} \tag{4.1}$$

where n_{F,m_F} is the density in the state with hyperfine manifold F and magnetic sublevel m_F. At the final parameters of the quench, which places the system into the regime of the easy-plane ferromagnetic phase [70], these define the transverse spin field $F_\perp = F_x + \mathrm{i}F_y$ [71]. Here, the interplay between energy offsets and spin interactions favor a finite transverse magnetization. During the dynamics, the transverse spin field approaches its ground state distribution, which manifests itself in the formation of a ring in the transverse spin histogram after approximately 1–3 s, as shown in Fig. 4.2(b). Nevertheless, in this regime, the system is still highly excited and transverse spin phase excitations evolve dynamically in a self-similar fashion [72]. Such relaxation dynamics is quite rich but complex, making a controlled microscopic characterization extremely challenging. In fact, the interpretation above is motivated by heuristic arguments. The key point we are interested in here is to lay such description

based solely on experimental observations, analyzed in a blind-folded manner. That is, we wish to extract essential descriptive elements (most important operators and complexity of the dynamics) without relying on any assumption.

At each evolution time, we obtain a data set with $N_r = 225$ independent realizations for each of the measured densities [see Eq. (4.1)]. Such data sets are denoted by a matrix $\mathbf{M}^{\alpha}(t) = \{\vec{\mathcal{N}}_1^{\alpha}(t), \vec{\mathcal{N}}_2^{\alpha}(t), \ldots, \vec{\mathcal{N}}_{N_r}^{\alpha}(t)\}$, where each row $\vec{\mathcal{N}}_i^{\alpha}(t)$ contains the spatial density profile of a single realization in one of the considered internal states, succinctly labeled here by $\alpha = (F, m_F)$. Each density is linearly sampled at 184 spatial locations, which gives the number of *features* (entries) of each vector $\vec{\mathcal{N}}_i^{\alpha}(t)$. Further, we also consider *joint* data sets formed by concatenating horizontally subsets of the original data sets at a given time. More specifically, each row in a joint data set is formed by appending, one after the other, single realizations of the observables of choice. Thus, for example, if we combine all six densities, the resulting data sets will contain $184 \cdot 6 = 1104$ columns and N_r rows. The particular order in which we concatenate the combined observables is irrelevant for our methods. When needed, we will simply specify joint data sets by using the symbol of the corresponding observables joined by "$\|$".

4.3 Assumption-free identification of relevant observables

We now perform a descriptive analysis of the data sets above, with the task of identifying the most informative observables. We start by introducing a PCA-based spectral entropy. PCA is a non-parametric approach that uses an orthogonal transformation to seek for the directions along which the data exhibit more variation [136, 80]. This problem can equivalently be posed in terms of a singular value decomposition (SVD) of the column-centred data matrix $\mathbf{M}^{\star}$ (where the mean value of each column is subtracted from the entries in the column). (For simplicity, we have omitted the observable and time indices of our

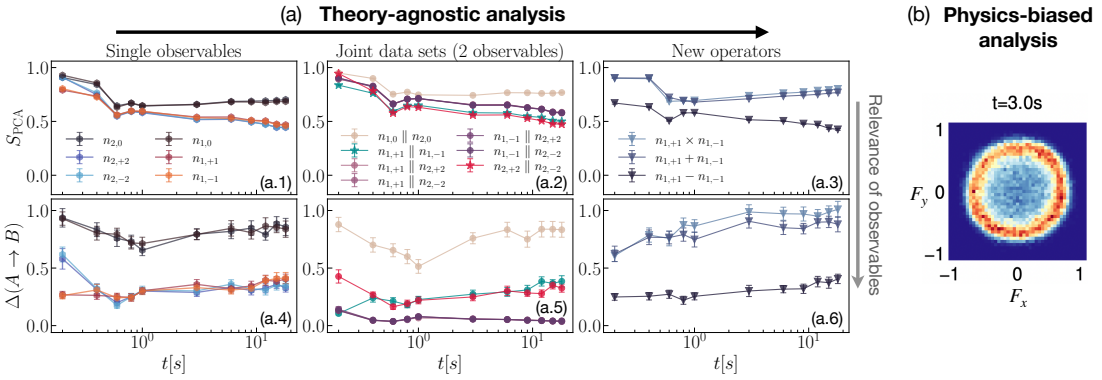

FIGURE 4.2: Assumption-free identification of relevant observables. (a) PCA entropy, S_{PCA}, and information imbalance, $\Delta(A \to B)$, as metrics of relevance: lower values of S_{PCA} signal stronger correlations within a data set, while lower values of $\Delta(A \to B)$ indicate that more information—from the *full* space of observations—is retained by the features of a given data set. Both metrics clearly show that $n_{1,\pm1}$ and $n_{2,\pm2}$ are more relevant (in the sense above) over the full evolution [panels (a.1) and (a.4)]. Identification of relevant groups is also possible by analysing joint data sets: for pairs of observables, $n_{1,+1} \parallel n_{1,-1}$ and $n_{2,+2} \parallel n_{2,-2}$, have a lower S_{PCA} [$\star$ markers in panel (a.2)], and are hence more relevant than any other pair (see ranking of all possible pairs in Ref. [135]). Information imbalance provides a complementary view on the relevance of such pairs, namely, the full space of measured observables is better described if features from both relevant pairs are considered simultaneously [points with $\Delta(A \to B) \sim 0$ in panel (a.5), corresponding to the joint data sets $n_{1,\pm1} \parallel n_{2,\pm2}$. Note that the points of the latter four data sets lie on top of each other—they are equally informative—, both in panel (a.2) and (a.5)]. Finally, we can rank new operators defined from the identified relevant pairs, as illustrated here for a few combinations of $n_{1,+1}$ and $n_{1,-1}$, with their difference $n_{1,+1} - n_{1,-1}$ being more relevant [panels (a.3) and (a.6)]. (b) Histogram of the transverse spin variable in the $F_x - F_y$ plane at $t = 3s$, featuring a ring-like structure. Based on physical arguments [72, 131, 70], this variable is the relevant field to describe the quenched system. Our theory-agnostic approach identifies the relevant observables from which this variable is inferred [see Eq. (4.1)], hence cross-validating the latter analysis.

previous notation). This allows to find the eigenvalues of the covariance matrix $\mathbf{\Sigma} = \boldsymbol{M}^{\star T}\boldsymbol{M}^{\star}/(N_r - 1)$, that is, $\mathbf{\Sigma}\vec{w}_k = \lambda_k \vec{w}_k$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R \geq 0$

(R is the rank of $\mathbf{M}^\star$). The eigenvalues give a measure for the variance of the data along the principal axes, determined by the corresponding eigenvectors. We then define the normalized eigenvalues $\tilde{\lambda}_k := \lambda_k / \sum_k \lambda_k$ [137]. By construction, $\tilde{\lambda}_k \geq 0$ and $\sum_{k=1}^{R} \tilde{\lambda}_k = 1$. The *PCA spectral entropy* can be thus defined in analogy to Shannon's entropy [56], namely,

$$S_{\mathrm{PCA}} := -\frac{1}{\ln(R)} \sum_{k=1}^{R} \tilde{\lambda}_k \ln(\tilde{\lambda}_k), \tag{4.2}$$

where for convenience we have normalized by its maximum possible value, $\ln(R)$, which is obtained for a flat eigenvalue distribution. Thus, for a perfectly uncorrelated process (e.g. white noise), $S_{\mathrm{PCA}} = 1$. Instead, for data with correlations among the variables $S_{\mathrm{PCA}} < 1$. This quantity can hence be used as a measure for ranking observables according to their information content (strength of their correlations).

As a second way to quantify the information content of observations, we compute their *information imbalance* in terms of distance ranks (for details see Refs. [126, 127, 135]). As opposed to PCA, this method is not based on a linear transformation; instead, it quantifies the relative information content between different distance measures. Concretely, given a data set with N_r points and $\mathcal{F}$ features, one defines distance measures A and B between data points on two subsets of the feature space and computes the corresponding rank matrices $R_{ij}^{A/B}$, such that $R_{ij}^{A/B} = m$ if j is the m-th nearest neighbor of i (according to A/B). The information imbalance—from A to B—can then be estimated as the average rank in B restricted to nearest neighbors in A, that is

$$\Delta(A \to B) \approx \frac{2}{N_r^2} \sum_{i,j:R_{ij}^A=1} R_{ij}^B. \tag{4.3}$$

With this definition, one can show that when the ranks computed with the two metrics are consistent with each other $\Delta(A \to B) \sim 0$ ('A captures B reliably'), whereas when they are fully uncorrelated $\Delta(A \to B) \sim 1$ ('A cannot predict B') [126, 127, 135]. Here, we consider the joint data sets of the six measured

observables (data points have $184 \cdot 6 = 1104$ coordinates). Space A is formed by the subset features corresponding to, for example, a single observable or a pair of observables (in which cases, we compute distances using only 184 or $184 \cdot 2 = 368$ coordinates, respectively), while B is chosen as the full space of features. This way, the information imbalance from A to B gives us a direct measure of the relevance of observables and their combinations to describe the full space of observations. Throughout this work, we use the Euclidean metric to compute distances between data points.

Our main results are shown in Fig. 4.2(a). Let us first analyze the results for the PCA entropy [panels (a.1)–(a.3)]. A clear separation between two groups of observables is noted as the system evolves [panel (a.1)], with $n_{1,\pm1}$ and $n_{2,\pm2}$ yielding lower values of S_{PCA}. According to the interpretation above, we conclude that these observables capture stronger correlations and are hence more relevant. Next, we consider joint data sets of two different observables [panel (a.2)]. The most relevant pairs according to this analysis are $n_{1,+1} \parallel n_{1,-1}$ and $n_{2,+2} \parallel n_{2,-2}$. The latter result is in excellent agreement with the physics-motivated analysis, in which such observables play a key role in the definition of the transverse spin [see Eq. (4.1) and Fig. 4.3(b)]. (The densities $n_{1,0}$ and $n_{2,0}$, are only important for normalization.) Going one step further, in panel (a.3) we explore the relevance of concrete functional combinations of the pair $\{n_{1,+1}, n_{1,-1}\}$ (similar results hold for $\{n_{2,+2}, n_{2,-2}\}$). We find that the difference $n_{1,+1} - n_{1,-1}$ has the lowest S_{PCA}, once again in remarkable agreement with the physics-motivated *ansatz* [138].

We now turn our attention to the information imbalance analysis [panels (a.4)–(a.6)]. These results are completely consistent with the analysis above based on S_{PCA}, and provide a complementary view on the relevance of observables. In fact, in panel (a.4), we can see that the observables with a lower PCA entropy have also a lower information imbalance; thus, they retain more information of the full space of features. Interestingly, in the analysis of pairs of observables [panel (a.5)], we note that in order to capture the full space of observations, one needs to consider features from both relevant pairs. Indeed, we see that the joint data sets $n_{1,\pm1} \parallel n_{2,\pm2}$ have $\Delta(A \to B) \sim 0$ for the full

evolution [note that the points of those data sets are on top of each other in panel (a.5) of Fig. 4.2]. The new operator $n_{1,+1} - n_{1,-1}$ has also a significantly lower information imbalance, compared to other combinations [panel (a.6)].

In Ref. [135], we show results for other possible joint data sets. We also include details on the subsampling analysis [108, 109] used for the estimation of error bars.

4.4 Complexity evolution of data sets

We now provide a further characterization of the data sets. Specifically, we study their *intrinsic dimension* I_d, at the considered evolution times. The I_d quantifies the minimum number of variables needed to describe the data [139, 92, 128], thereby providing a measure of their (Kolmogorov) complexity [140, 141, 37, 114]. Here we use a distance-based I_d estimator, namely, the TWO-NN algorithm [128], which works as follows. For each point $\vec{\mathcal{N}}_i$ in a generic data set, we compute the distance to its first and second nearest neighbors, denoted by r_{i_1}, r_{i_2}. Next, we define the ratio $\mu_i := r_{i_2}/r_{i_1}$. For data points that are locally uniformly distributed on a I_d-dimensional hypersphere, the distribution of μ is given by $f(\mu) = I_d \mu^{-I_d - 1}$ [128]. The cumulative distribution function $F(\mu)$, then satisfies

$$-\ln[1 - F(\mu)] = I_d \ln(\mu), \tag{4.4}$$

which is used to estimate I_d through a linear fit of the points $\{(\ln(\mu), -\ln[1 - F_{\text{emp}}(\mu)])\}$, where $F_{\text{emp}}(\mu)$ is the empirical cumulate [135].

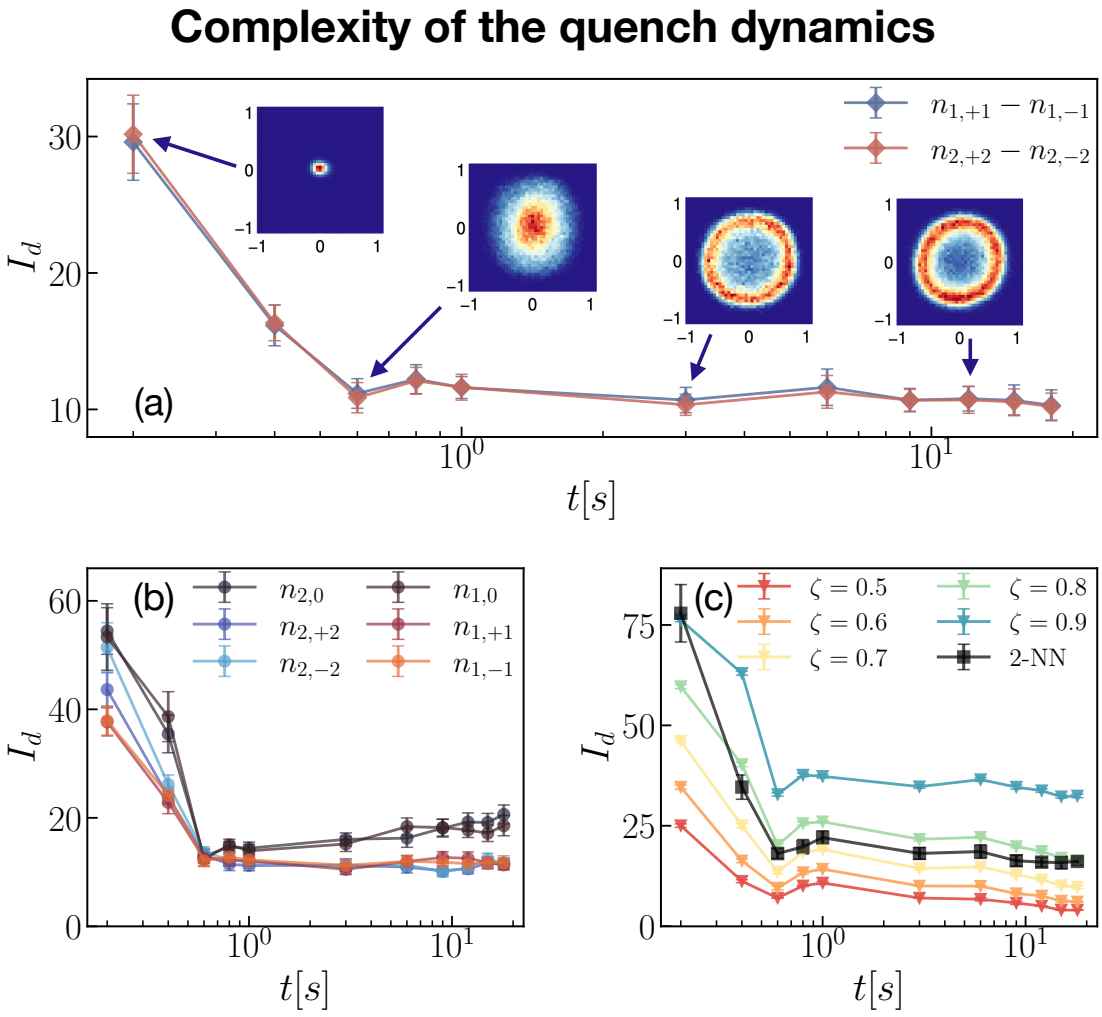

FIGURE 4.3: Complexity of the dynamics via intrinsic dimension. Intrinsic dimension as a function of time for (a) the relevant observables, $n_{1,+1} - n_{1,-1}$ and $n_{2,+2} - n_{2,-2}$, (b) all measured densities individually, and (c) joint data sets of all six observables together. In all instances, an initially large I_d quickly decays to smaller values (around $t = 0.6s$), subsequently exhibiting long plateaus. The insets in panel (a) show histograms of the transverse spin in the $F_x - F_y$ plane at selected times. The first drop in the I_d is associated to a grow in the spin length, which remains approximately constant for $t1s$. In the latter regime a ring-like structure is then observed (illustrated here at $t = 3s, 12s$). Spatial correlations of spin phase excitations exhibit self-similar dynamics in a regime that starts around $t \sim 3s$ [131]. The observed structural simplification of the data strongly correlates with such universal dynamics. Hence, the plateaus in the plot of I_d provide a lower bound for the onset of simpler dynamics and universal scaling. Panel (c) also shows the I_d estimate based on PCA for various values of the cutoff ζ (see main text).

Shown in Fig. 4.3(a) is the plot of I_d as a function of time, of the data sets corresponding to the identified relevant operators, $n_{1,+1} - n_{1,-1}$ and $n_{2,+2} - n_{2,-2}$. We observe the same trend in both instances: a quick decay of I_d to considerably smaller values, subsequently displaying long, stable plateaus. The reduction of the I_d signals a simplification of the data structure due to the buildup of correlations among the input variables. The latter is a direct manifestation of the correlations among the elementary constituents of the system. From the physical viewpoint, the post-quench correlations are associated with the formation of a ring-like structure, with an approximately constant radius, in the transverse component of the collective spin degree of freedom (insets) [72, 131]. In turn, spatial correlations of the spin phase excitations exhibit universal scaling dynamics [72]. In the present experiment, the universal scaling regime starts approximately at $t \sim 5s$ [131]. The physical basis for such scaling evolution is a dynamical reduction of the relevant parameters in the system. This is strongly consistent with the observed structural simplification of the data, as also observed in recent studies of critical behavior—in and out of equilibrium—in classical and quantum statistical mechanics systems [142, 143, 144, 114]. Therefore, in the present case, the observed I_d plateau provides a theory-agnostic lower bound for the timescale after which the dynamics may become simpler, allowing for the emergence of self-similar behavior.

Importantly, this prediction can be made by directly studying the I_d of data sets of all measured densities, as shown in Fig. 4.3(b), where we observe an overall similar trend. We note, however, that the "irrelevant" observables $n_{1,0}$ and $n_{2,0}$, have a growing I_d, rather than a plateau. This further confirms the relevance predictions based on PCA entropy and information imbalance. Further, in Fig. 4.3(c) we plot the I_d of joint data sets of the six measured observables together, showing once again the noted trend. In this plot, we also show an I_d estimation based on PCA, which is defined by choosing an *ad hoc* cutoff parameter ζ, for the integrated spectrum of the covariance matrix [136, 142], i.e., $\sum_{k=1}^{I_d} \tilde{\lambda}_k \approx \zeta$. We find that for all considered values of ζ, we recover the same qualitative features as the TWO-NN I_d-estimate. A quantitative agreement can also be achieved for a suitable choice of ζ in the range $0.7 \leq \zeta \leq 0.9$, at the

different evolution times. This agreement further confirms the applicability of PCA in our previous analysis and implies that curvature effects of the data manifold are negligible.

4.5 Conclusions

We have introduced an assumption-free method to diagnose and rank relevant correlations in the dynamics of out-of-equilibrium quantum systems. The method exploits the full spectrum of principal components, as well as recently developed techniques based on information imbalance. We have successfully identified the most relevant operators describing the dynamics of Bose-Einstein condensates, confirming previous heuristic approaches (and thus, validating the physical relevance based solely on experimental observations). Utilizing manifold characterization methods, we have also found stable plateaus of the intrinsic dimension of the data sets corresponding to different times, thus providing bounds on the time frame realizing universal quantum dynamics. Our approach is immediately extended to other classes of quantum simulators—including fermion gases and lattice spin models—providing a flexible, assumption-free framework to discover physical phenomena, as well as to validate their functioning. Our work complements recent theoretical approaches with similar goals regarding the identification of relevant observables [145, 146] and characterizing the complexity of quantum dynamics [114, 147, 148].

4.6 Supplemental Material

4.6.1 More on information imbalance

In the following, we summarize the main ideas that lead to Eq. (4.3) in the main text, but the reader is referred to Ref. [126] for a more detailed explanation. The notion of *information imbalance* refers to a statistical test aimed at

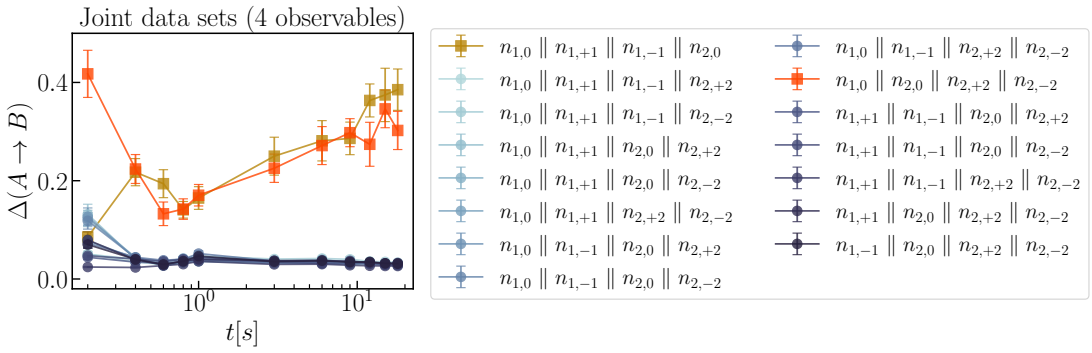

FIGURE 4.4: Information imbalance of joint data sets of four observables. All joint data sets that combine features from the two relevant pairs $\{n_{1,+1}, n_{1,-1}\}$ and $\{n_{2,+2}, n_{2,-2}\}$ have a very small information imbalance. Instead, the joint data sets that only use features from one of the relevant pairs cannot predict so well the full space of features. The latter are indicated with square markers in this plot.

assessing the relative information content of different distance measures defined on the same data space [126, 127]. This is done by analyzing the *ranks* of the first nearest neighbors of each point. More specifically, given a finite data set $\mathbf{M} = \{\vec{\mathcal{N}}_i\}_{i=1}^{N_r}$ and two distance measures $D_A(\vec{\mathcal{N}}_i, \vec{\mathcal{N}}_j)$ and $D_B(\vec{\mathcal{N}}_i, \vec{\mathcal{N}}_j)$, we can rank the neighbors of a point $\vec{\mathcal{N}}_i$, in the two spaces under consideration, by sorting, from smallest to largest, the pairwise distances between such a point and the rest of points using the corresponding metric. These rankings are encoded in the so-called rank matrices $R_{i,j}^{A/B}$. Thus, for example, $R_{i,j}^{A} = 1$ if $\vec{\mathcal{N}}_j$ is the 1st nearest neighbor of $\vec{\mathcal{N}}_i$ in space A, $R_{i,j}^{A} = 2$ if $\vec{\mathcal{N}}_j$ is the 2nd nearest neighbor of $\vec{\mathcal{N}}_i$, etc. The key insight of this method is the fact that the full correlation structure between the two metrics under study is essentially captured by the conditional rank distribution $p(R^B|R^A = 1)$, that is, the probability distribution of the ranks R_{ij}^B in space B restricted to those pairs (i, j) for which $R_{ij}^A = 1$ (i.e., the pairs of nearest neighbors in A). Then, the closer this distribution is to a delta function peaked at 1, the more information (at the level of local neighborhoods) about space B is contained within space A. The deviation of $p(R^B|R^A = 1)$ from such a delta function is quantified by the conditional expectation $\langle R^B|R^A = 1\rangle$, which is used to define the information

imbalance from space A to space B, namely

$$\Delta(A \to B) = \frac{2}{N_r}\langle R^B | R^A = 1\rangle = \frac{2}{N_r^2}\sum_{i,j:R^A_{ij}=1} R^B_{ij}. \tag{4.5}$$

In the limit case where the two spaces are completely equivalent (meaning that first nearest neighbors in A are exactly the same as those in B), we have that $\sum_{j:R^A_{ij}=1} R^B_{ij} = 1$ (for a given i), and hence $\sum_{i,j:R^A_{ij}=1} R^B_{ij} = N_r$. Therefore, the information imbalance in Eq. (4.5), $\Delta(A \to B)$ vanishes as $1/N_r$. A vanishing information imbalance indicates that A can fully predict B in the sense specified above. In the extremely opposite case in which the two spaces are completely independent, we have that $\sum_{j:R^A_{ij}=1} R^B_{ij} = \frac{1}{N_r-1}\cdot\frac{1}{2}N_r(N_r-1) = \frac{N_r}{2}$ and hence $\sum_{i,j:R^A_{ij}=1} R^B_{ij} = \frac{N_r^2}{2}$. Therefore, the resulting information imbalance is $\Delta(A \to B) = 1$, implying that A cannot predict B. Finally, we note that, due to its own definition, the information imbalance is asymmetric; therefore, one can study predictability between the two considered distances in both directions.

4.6.2 Relevance ranking via S_{PCA} and information imbalance: further combinations

In this section, we show the ranking of further combinations of operators. As explained in the main text, we analyse joint data sets formed by concatenating horizontally measured data sets. Thus, for example, $n_{1,+1} \,\|\, n_{1,-1}$ refers to a data set for which we have concatenated horizontally the data sets corresponding to the observables $n_{1,+1}$ and $n_{1,-1}$. Complementing Fig. 4.2 of the main text, in Figs. 4.6 and 4.5, we show the ranking of all joint data sets formed by two observables, according to S_{PCA} and $\Delta(A \to B)$, respectively. As mentioned in the main text, the joint data sets with a lower value of S_{PCA}, throughout almost the whole evolution, are $n_{1,+1} \,\|\, n_{1,-1}$ and $n_{2,+2} \,\|\, n_{2,-2}$ ($\star$ markers in Fig. 4.6). These pairs of observables are indeed the most relevant ones to describe the physics of the quenched system. We also note that the joint data set $n_{1,0} \,\|\, n_{2,0}$ is the one with the largest PCA entropy, and hence is the "least" informative

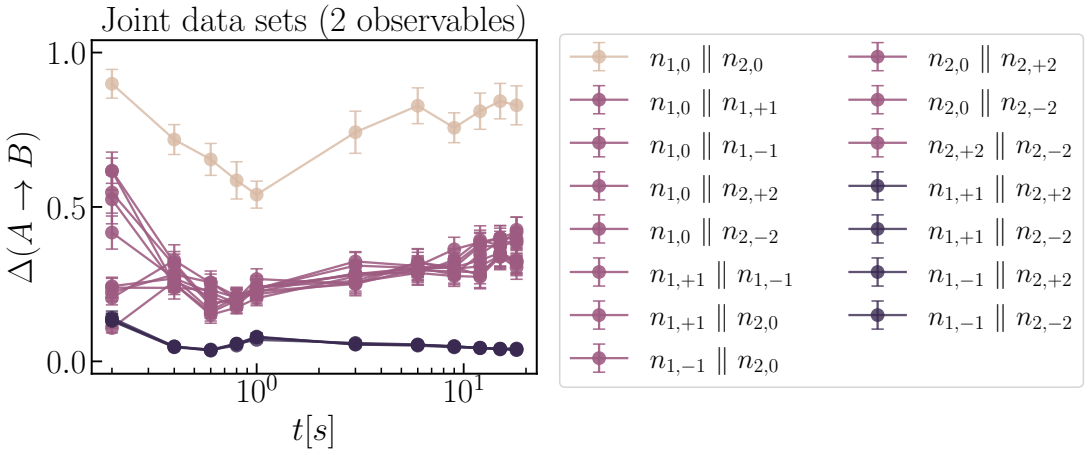

FIGURE 4.5: Information imbalance of joint data sets of two observables. This plot shows the results for all joint data sets of two observables (cf. Fig. 4.2(a.5) of the main text). Note that the points corresponding to the joint data sets $n_{1,+1} \parallel n_{2,+2}$, $n_{1,+1} \parallel n_{2,-2}$, $n_{1,-1} \parallel n_{2,+2}$, and $n_{1,-1} \parallel n_{2,-2}$, lie basically on top of each other (they are equally informative), with an information imbalance $\Delta(A \to B) \sim 0$. These four data sets combine features of the two relevant pairs $\{n_{1,+1}, n_{1,-1}\}$ and $\{n_{2,+2}, n_{2,-2}\}$.

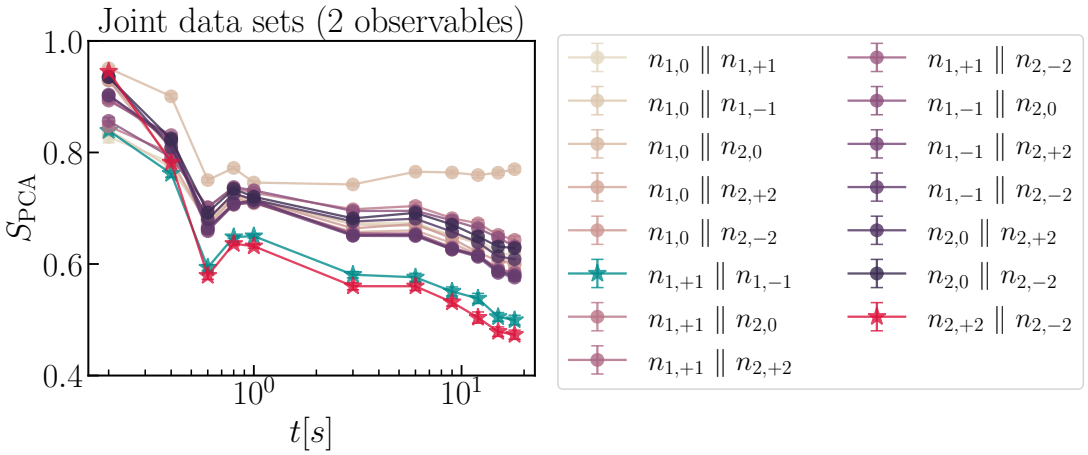

FIGURE 4.6: PCA entropy of joint data sets of two observables. This plot shows the results for all joint data sets of two observables (cf. Fig. 4.2(a.2) of the main text).

one. This is also in agreement with physical considerations as the populations in the respective substates are only important to normalize properly the collective transverse spin variable (see Eq. (4.1) in the main text). Information imbalance—which does not rely on a linear transformation as does PCA, but on ranks of nearest neighbors—provides a complementary view on the relevance of the analysed observations. Let us recall that $\Delta(A \to B) \sim 0$ means that A can reliably predict B, whereas $\Delta(A \to B) \sim 1$ implies that A cannot predict B. Here, we measure the information imbalance from a space A of features corresponding to joint data sets of two observables, to the space B of features corresponding to the joint data set of all six observables together. In Fig. 4.5, we observe that the joint data set with the largest information imbalance (smallest information content) is $n_{1,0} \parallel n_{2,0}$, in full agreement with the ranks obtained using the PCA entropy. Next, we observe that the joint data sets with the smallest information imbalance are not $n_{1,+1} \parallel n_{1,-1}$ and $n_{2,+2} \parallel n_{2,-2}$, but those where features from these two relevant pairs are combined, namely: $n_{1,+1} \parallel n_{2,+2}$, $n_{1,+1} \parallel n_{2,-2}$, $n_{1,-1} \parallel n_{2,+2}$, and $n_{1,-1} \parallel n_{2,-2}$ (markers with darker color in Fig. 4.5). In fact, any of these four joint data sets can predict almost entirely the full space of features, since $\Delta(A \to B) \sim 0$ for such data sets. The meaning of this observation is that in order to predict better the full space of features (given by the coordinates of the joint data set of the measured six observables), one needs to consider features from both of the identified relevant pairs $n_{1,+1} \parallel n_{1,-1}$ and $n_{2,+2} \parallel n_{2,-2}$.

Similar conclusions apply if one considers groups of more than two observables. This is illustrated here for groups of four observables (quadruplets) in Figs. 4.7 and 4.4. In terms of PCA entropy the most relevant joint data set is the one that combines the features of the two relevant pairs, namely, $n_{1,+1} \parallel n_{1,-1} \parallel n_{2,+2} \parallel n_{2,-2}$ ($\star$ markers in Fig. 4.7). We note however that in this case the relative difference in PCA entropy is not as pronounced as in the case of single or two observables. Regarding information imbalance, we observe once again those joint data sets that combine features from the two relevant pairs can predict almost entirely the full space of features. There are indeed only two combinations that only involve features from a single

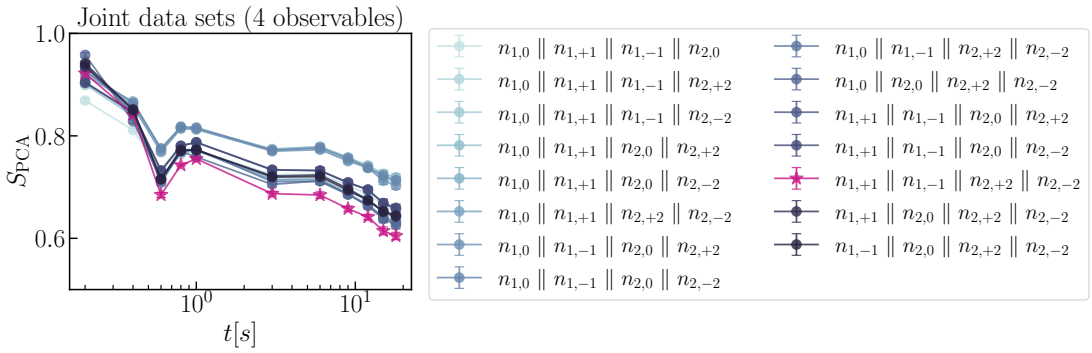

FIGURE 4.7: PCA entropy of joint data sets of four observables. The joint data set with the smallest PCA entropy is the one that combines the two relevant pairs ($\star$ markers).

relevant pair (plus the two "irrelevant" observables $n_{1,0}$ and $n_{2,0}$), namely, $n_{1,0} \parallel n_{1,+1} \parallel n_{1,+1} \parallel n_{2,0}$ and $n_{1,0} \parallel n_{2,0} \parallel n_{2,+2} \parallel n_{2,-2}$ (square markers in Fig. 4.4), which clearly have a significantly larger information imbalance compared to the rest.

4.6.3 Linear fit to estimate I_d from the empirical cumulates in the TWO-NN method

In this section, we show examples of the linear fitting procedure used to estimate the value of the intrinsic dimension in the TWO-NN method; see Eq. (4.4) in the main text. In Fig. 4.8, we show the empirical cumulative distributions of the ratios $\mu_i = r_{i_2}/r_{i_1}$, sorted in ascending order, for the observable $n_{2,+2} - n_{2,-2}$, at all evolution times. If the condition of constant density in the range of first two nearest neighbors holds, a plot of the resulting points $\{\ln(\mu), -\ln[1 - F_{\mathrm{emp}}(\mu)]\}$ will be a line that passes through the origin and whose slope gives the estimated value of I_d. Verifying that the empirical cumulates are indeed consistent with a Pareto distribution as described above, is the first step to guarantee the applicability of the TWO-NN method. Besides, on its own, this kind of plot is also very informative about the local structure of complicated data manifolds.

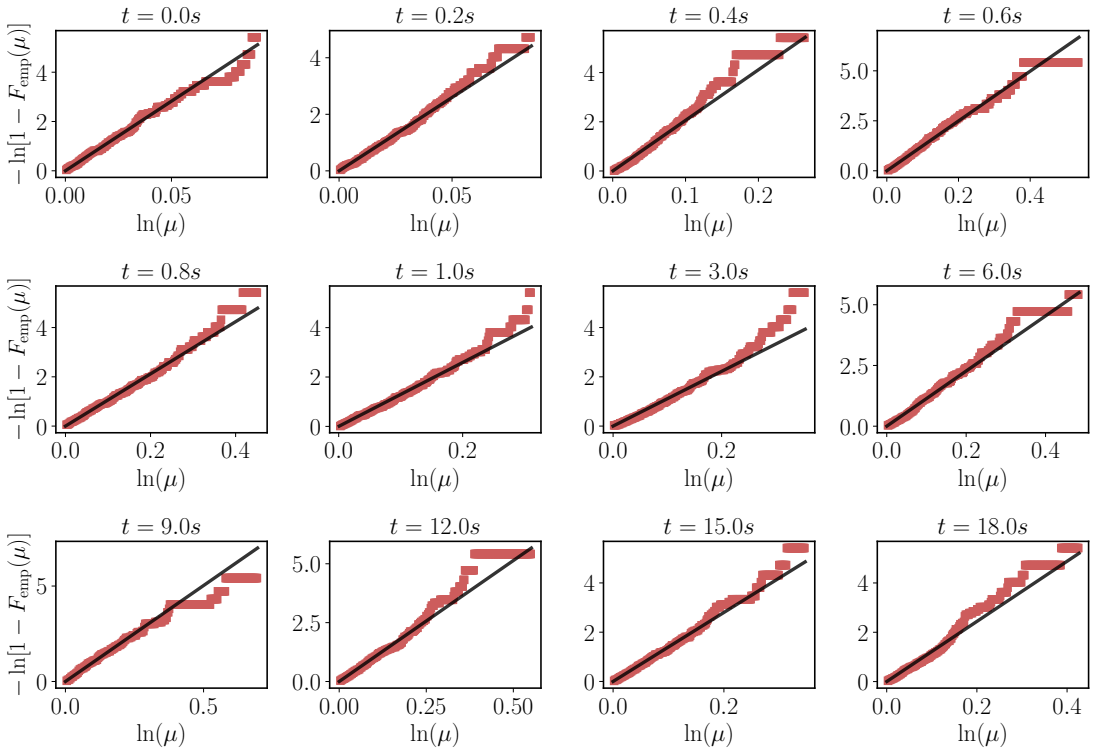

FIGURE 4.8: Empirical cumulative distributions at all evolution times for the data sets corresponding to the relevant observable $n_{2,+2} - n_{2,-2}$. The black curve show the linear fit according to Eq. (4.4) in the main text, whose slope gives the estimated value of I_d. This procedure is valid as long as the empirical cumulative distribution function is consist with a Pareto distribution, at least over a significant range of values of $\ln(\mu)$, as is the case here.

4.6.4 Subsampling error estimation

Due to the limited number of realizations used in the present analysis, we opted for using a technique known as subsampling [108, 109] to have a sensible estimation of the statistical errors. The subsampling algorithm is described below.

At a given time and for a given measured observable, we have $N_r = 225$ independent realizations forming our data set, that is, $\mathbf{M} = \{\vec{\mathcal{N}}_i\}_{i=1}^{N_r}$, where for

simplicity we have omitted the indices labeling the observable and the evolution time. Using these data we compute a certain numerical statistic ϑ. Given two preset integers b and $q < N_r$, the subsampling analysis proceeds as follows:

1. Form b random 'batches' (subsamples) of data by drawing $q < N_r$ points at random but *without* replacement from the data set $\{\vec{\mathcal{N}}_1, \vec{\mathcal{N}}_2, \ldots, \vec{\mathcal{N}}_{N_r}\}$.

2. Estimate the statistic of interest on each subsample, that is, ϑ_i, for $i \in \{1, \ldots, b\}$.

3. Compute the subsample mean $\overline{\vartheta} = \frac{1}{b}\sum_{i=1}^{b} \vartheta_i$. The standard error can then be estimated as follows

$$\mathrm{SE} \approx \sqrt{\frac{q}{N_r - q}} \cdot \sqrt{\frac{1}{b}\sum_{i=1}^{b}(\vartheta_i - \overline{\vartheta})^2}, \tag{4.6}$$

This formula is known as the delete-d Jackknife standard error estimator (with stochastic subsampling) [108, 109].

While formally this method requires $q/N_r \to 0$, $q \to \infty$, and $b \to \infty$ as $N_r \to \infty$ (so that the distribution of the ϑ_i converges to the sampling distribution of ϑ), in practice the choice of these parameters is problem specific. In our analysis, we did not find significant changes for $b \geq 30$. Hence, we fixed $b = 30$. Furthermore, to compute a meaningful statistic on each subsample, we set $q = 100$.

We used this method as sampling is performed *without* replacement. This is important as the TWO-NN algorithm used to estimate the I_d works under the assumption of no repetitions among the data points.

Chapter 5

Network science Ising states of matter

5.1 Introduction

Networks [149, 150, 151, 152, 153] encode the information present in a large variety of natural and artificial interacting systems by representing them as graphs, i.e. a set of nodes (representing the element of the system) connected by links or edges (representing typically the interactions). In particular, the underlying architecture of complex systems is encoded in networks that strongly deviate from random graphs whose information content can be mined by exploiting their statistical, combinatorial as well as the geometrical and topological nature [154]. Networks are hence a simple yet very powerful framework that has been able in the last twenty years to transform our understanding of complex systems. These complex networks obey relevant organization principles while retaining also a stochastic nature.

Recently great attention has been addressed to formulate unsupervised machine learning algorithms to detect different phases of matter [80, 83, 44, 86, 155, 142]. The core of such approaches is that learning methods - in particular, unsupervised - shall be capable of revealing hidden structures in data sets, that correspond to physical information (such as, e.g., the existence of order parameters). However the very advanced, complementary tools of network science

to extract information from complex data of interacting systems have not yet been systematically employed for this task.

Here we want to show how network science can enrich and enhance our unsupervised characterization of phases of matter. Indeed we will show evidence that network science provides a very transparent set of methods to investigate the characteristics of different phases across critical phase transitions and allows the determination of unsupervised indicators of their critical points.

Historically networks have been used in condensed matter for describing physical interactions existing among the elements of a system, as well as structured in configuration and Hilbert space. In principle, they can be used as well to represent abstract data structures coming from numerical simulations or directly from experiments, giving access to a whole new toolbox to define and interpret many-body correlations (of arbitrary order, if expressed in terms of local observables).

Reflecting this two-fold possible application of network science, recently several works have explored the use of networks to model or represent interactions in condensed matter systems. For instance, networks can be considered to define Hamiltonians whose interaction terms are determined by a complex network rather than by a lattice. In particular networks can be used to define different critical phenomena including the Ising model [156, 157, 158] and the inverse Ising model [159], as well as quantum critical phenomena including the transverse Ising model [160, 161, 162], the Bose-Hubbard model [163], the Jaynes-Cummings-Hubbard model [164], in addition to others classical collective phenomena and inverse problems [165, 166, 167, 168]. Alternatively, networks can define quantum environments [169, 170] or even multilayer couplings between interdependent superconductor networks [171]. Networks also can be used to represent correlations in complex and financial networks [172, 173] as well as in quantum systems. In particular recently network structures have been shown to encode the quantum long-range mutual information (and other measures of quantum correlations) existing among the nodes of quantum lattice models in one [174, 175, 176] and two dimensions [177], providing in some cases indicators for quantum critical points.

As we will see in this work, weighted networks [178] are amenable to be analysed and treated with Topological Data Analysis (TDA) [179, 180, 181, 182] which provides a very efficient way to probe topological, large scale and global network properties. TDA, although until now only applied to point clouds, is raising significant interest to do unsupervised inference of phase transitions [183, 184, 185, 186, 187, 85, 188], and has wide applications, including the characterization of universal dynamics in quantum gases [189] and of confinement in lattice field theory [190, 191, 192].

Finally and most relevantly for our work, networks have been proposed to capture the underlying structure of quantum spin systems as revealed by wave function snapshots that can be probed experimentally as well as sampled from Monte Carlo simulations [193]. However, despite these very pioneering works [174, 193], little attention has been so far addressed to study phases of matter using network science.

Here we launch a large scale systematic study of the phase of matter based on network science. We leverage a multiplicity of tools developed in network science and we reveal the combinatorial, statistical, geometrical and topological network representation of different phases of matter. We provide an in-depth characterization of the networks generated from single snapshots of spin system configurations.

In Ref. [193] it was shown that wave function networks constructed starting from quantum wave function snapshots are strongly deviating from random graphs and for a wide range of values of the threshold distances they give rise to networks with very broad degree distribution. An open question is whether the complex properties of these networks are inherently quantum effects or they can be observed in classical systems as well.

In this work we consider 2D Ising snapshot networks (IsingNets) following a construction proposed in Ref. [193] applied to classical 2D Ising model snapshots and we characterize their structure using advanced statistical, combinatorial, geometrical and topological tools of network theory. In order to provide an in-depth analysis of the IsingNet, across different phases we focus our attention on IsingNets obtained starting from Monte Carlo simulations

of the 2D Ising model performed across the phase transition. IsingNets are obtained from a sample of state configurations of the 2D Ising model which constitute the set of nodes of the IsingNets. Each pair of nodes of an IsingNet is associated with a distance, here taken to be the Euclidean distance between the configuration snapshots. IsingNets are constructed starting from the fully connected distance matrix between the nodes by connecting only the nodes whose distance is smaller than a threshold value of the distance.

Our in-depth network analysis of the IsingNets will allow us to go well beyond the characterization of these networks based solely on the degree distribution. Possibly in the future, this in-depth analysis can be conducted also on networks built from quantum wave function snapshots in order to assess which are the properties inherently quantum in the latter networks.

Our analysis is conducted following two main directions whose goal is different but complementary. First, we will perform an analysis of the IsingNets that is agnostic about the choice of the distance threshold. In particular, we will study network properties as a function of the distance threshold. These include percolation properties, persistence homology, network embedding and statistical characterization of the distance matrices. Secondly, we will consider specific choices of the distance threshold and we will characterize the statistical, combinatorial and geometrical/spectral properties of the IsingNets, showing the important roles of degree-degree and weight-degree correlations in these systems.

Anticipating our main results we have found that IsingNets reflect the symmetry of the configuration space of the 2D Ising model in a prominent way. In particular, the percolation properties of the IsingNets strongly deviate from the percolation properties of networks in which the same distances among the nodes are distributed randomly. In fact, below the critical temperature of the 2D Ising model, the IsingNets are characterized by two giant components whereas their randomized counterpart displays only a single giant component. When the threshold distance defining the IsingNets is raised significantly these two giant clusters merge, but interestingly they keep a rather compact structure as revealed by our persistent homology results highlighting that

the Betti number of their clique complex are strongly suppressed with respect to their random counterpart. Our statistical and combinatorial analysis of the IsingNets strongly demonstrates the complex organization of these networks which display strong heterogeneity in both their topological (degree, clustering coefficient, degree correlations) and their weighted network properties. In particular nodes of higher-degree are characterized by having neighbours connected by stronger affinity weights (smaller distances). Finally, the IsingNets display very interesting spectral properties indicating a significant geometrical organization of these networks at criticality as it is revealed by their spectral properties.

While our analysis is performed on data coming from Monte Carlo simulations the approach can be readily applied also to experimental data.

5.2 The 2D Ising model Monte Carlo simulations

We consider a square 2D dimensional lattice of dimension $L \times L$ where on each site n is located the spin $S_n \in \{-1, 1\}$. The nearest neighbour spins are interacting through the Hamiltonian

$$H = -\sum_{\langle n,m \rangle} S_n S_m. \tag{5.1}$$

The $2D$ Ising model is characterized by $\mathbb{Z}_2$ spontaneous symmetry breaking and undergoes a second-order phase transition at $T_c = 2/\ln(1+\sqrt{2}) \approx 2.269$ [102]. Starting from Markov Chain Monte Carlo simulations of this model, for each temperature single snapshots $\vec{x}_i = \{S_1, S_2, \ldots, S_{L^2}\}$ of the spin system are sampled at equilibrium [1]. More specifically, we use the Wolff cluster algorithm [194, 195], starting from the configuration with either all up spins or all down spins, chosen at random. Next, 30000 to 50000 'cluster flips' are performed for the system to equilibrate. After this, we collect snapshots every 1000 to 1500 cluster flips to ensure that the collected state configurations are

as uncorrelated as possible [1]. In total, we gather 10000 snapshots during a Monte Carlo run.

For each temperature, five independent Monte Carlo simulations are performed as prescribed. By combining the sampled configuration snapshots of these runs, we thus obtain a data set with $N_r = 50000$ independent thermal configurations $\{\vec{x}_i\}_{i=1}^{N_r}$. The starting point to construct the IsingNets is a set of N configuration snapshots $i \in \{1, 2, \ldots, N\}$ randomly selected from the data set described above, and the fully connected distance matrix $\mathbf{d}$ of elements d_{ij} between these states. Here the distance d_{ij} between two generic snapshots $\vec{x}_i$ and $\vec{x}_j$ is taken to be their Euclidean distance. As discussed in previous works, such manifolds are typically living in very high dimensional subspaces [142, 1], so that simple dimensional reductions are not applicable, and a full-fledged network analysis is needed.

5.3 Network characterization across the distance filtration

5.3.1 Weight filtration

As anticipated in the introduction, in this first Section, our analysis focuses on the properties of IsingNets observed as a function of the distance filtration. We consider IsingNets which are graphs $G = (V, E)$ formed by a set of N nodes V and a set of links E with $(i, j) \in E$ only if the nodes (state configurations) i and j have distance $d_{ij} < r$. Here r determines the distance filtration and indicates a tunable parameter that ranges from the minimum of the distances $r_{min} = \min_{i,j} d_{ij}$ between the N nodes to their maximum distance $r_{max} = \max_{i,j} d_{ij}$. We are thus here completely agnostic about the best choice of r as we study the properties of IsingNets across all possible choices of the distance threshold r.

In particular as a function of r we will explore the percolation properties of the IsingNets which define an agglomeration from N disconnected nodes to a single connected component as r is raised from r_{min} to r_{max} and we will

compare this process with the corresponding null model obtained from the same process applied to a randomized distance $\mathbf{d}^{rand}$ matrix. The randomized distance matrix $\mathbf{d}^{rand}$ is constructed by randomly reshuffling the upper triangular elements of d and subsequently symmetrizing the matrix. Therefore the null model networks display the same distribution of "distances" as the true IsingNet while being completely randomized. Note however that one of the main differences between the distance matrix $\mathbf{d}$ and the randomized distance matrix $\mathbf{d}^{rand}$ is that the entries of $\mathbf{d}^{rand}$ are not proper distances as they do not obey the triangular inequality.

In this section, we will use a combination of tools coming from network science to analyse the IsingNets described above. In particular, anticipating the detailed description of the methods used to perform this analysis in the following paragraphs, we will use persistent homology [182, 180, 181] to further characterize topologically the mentioned aggregation process. In this way, we will show that persistent homology is able to detect the position of the critical temperature of the 2D Ising model under study. This analysis will be accompanied by the visualization of the network using Minimum Spanning Trees [196, 173], the results of the network embedding conducted using the UMAP (Uniform manifold approximation and projection for dimension reduction) algorithm [197], and the statistical characterization of the distance matrix as a function of the temperature conducted using the closeness centrality [198] distribution.

5.3.2 Percolation process

We start our investigation exploring the percolation process [199, 150, 200, 201, 151] monitoring the connected component of the network as a function of the filtration parameter r. To contrast the behavior of the percolation process of IsingNets with a null-hypothesis percolation process, we consider the process defined on the actual IsingNet distance matrix $\mathbf{d}$ with the same process defined starting from a matrix $\mathbf{d}^{rand}$ obtained by randomly permuting distances among pair of nodes.

The percolation process of IsingNets reveals a major difference with the percolation process on the randomized distance matrix: mainly the IsingNets obtained for the 2D Ising model below the phase transition, i.e. $T < T_c$ display for a very significant range of values of the filtration parameter r, two giant components while the randomized process only displays one giant component. This phenomenon is evident from the plot in Figures 5.1 and 5.2 showing the relative size of the largest component R and the second largest component R_2, which are both giant, i.e. extensive for a wide range of r values. Indeed below the critical temperature T_c, the IsingNets display two transitions as the value of the filtration parameter is raised (see Figures 5.1 and 5.2). The first transition is characterized by the emergence of the two equal size giant components corresponding to the symmetry of the configuration snapshots of the 2D Ising model for $T < T_c$ and the second one is characterized by the merging of these two giant components for very large values of r, characterized by the disappearance of a significant second largest connected component (orange line in Figure 5.1 (a) and Figure 5.2 (a). This phenomenology is dramatically different from the percolation obtained in the randomized null model where the giant component is unique for every value of r (see Figures 5.1 and 5.2). For temperatures above the critical one (see Figure 5.3), instead, only one giant component is observed corresponding to the paramagnetic state of the 2D Ising model. In order to further characterize the percolation process we also monitor as a function of r the average size of finite components $\langle \bar{s} \rangle$, the number of components $n_{\bar{s}}$ and the inverse participation ratio Y whose inverse determines the number of typical clusters. Specifically the inverse participation ratio Y is defined as

$$Y = \sum_p \left(\frac{\bar{s}_p}{\sum_q \bar{s}_q} \right)^2 = \frac{1}{N^2} \sum_p \bar{s}_p^2. \tag{5.2}$$

where $\bar{s}_p$ indicates the size of the p-th largest component. The abrupt increase of Y at large r further indicates the merging of two giant components (Figure 5.1 (g) and Figure 5.2 (g)).

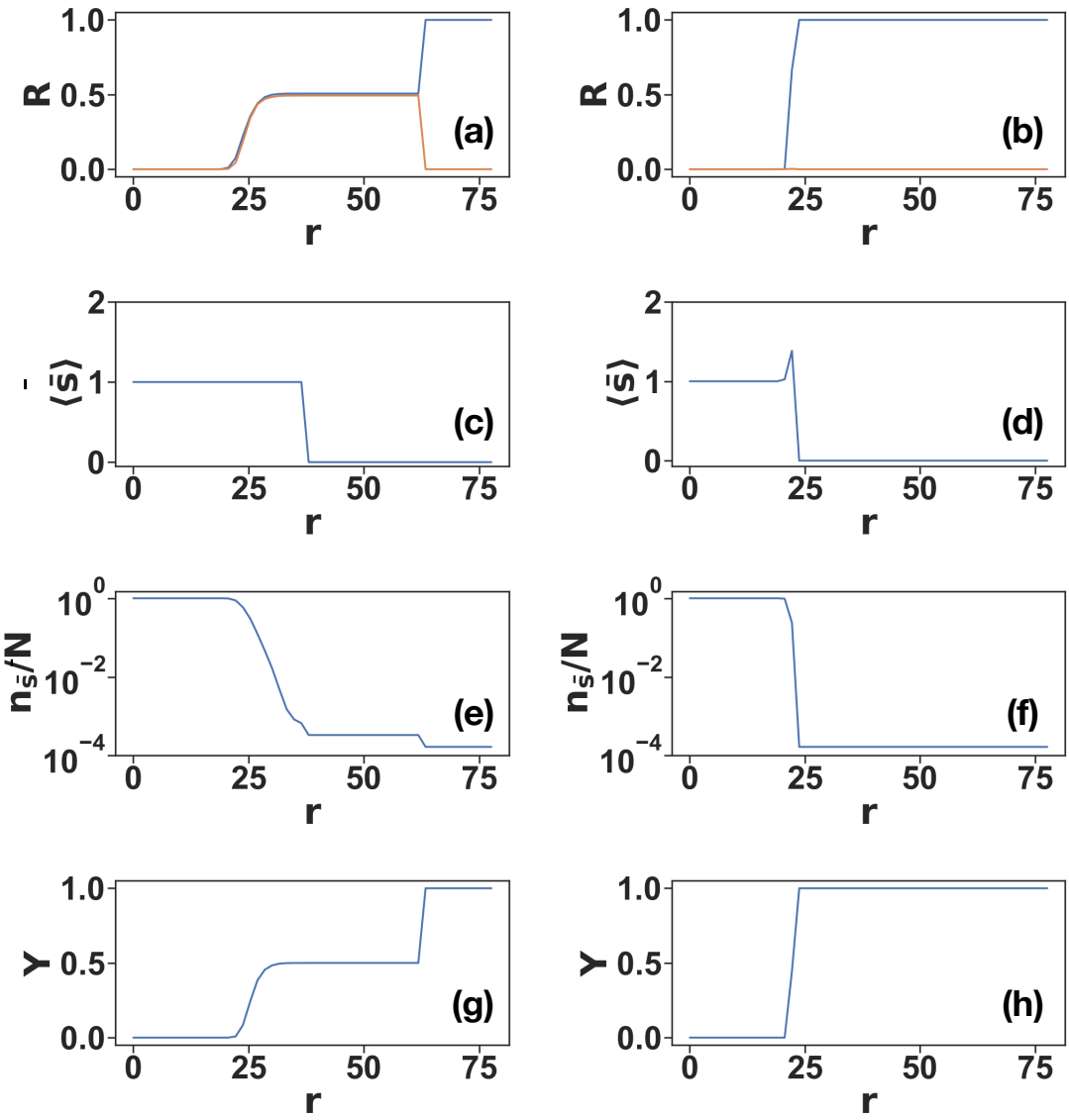

FIGURE 5.1: The percolation properties of the IsingNet (left panels) generated from 2D Ising model Monte Carlo simulations on spin systems of linear size $L = 40$ at temperature $T = 2.12 < T_c$ are shown as a function of filtration parameter r. Nodes are connected if their distance is less than r. Five quantities are measured: the fraction of nodes in the largest connected component (the first row, blue line) and the fraction of nodes in the second largest connected component (the first row, orange line), the average size of components that are smaller than the second largest component $\langle s \rangle$ (the second row), the number of components n_s (the third row) and the inverse participation ratio Y (the fourth row). The results are compared with these quantities obtained from corresponding percolation properties obtained from a randomly permuted distance matrix (right panels). The number of nodes of the IsingNets is $N = 6000$.

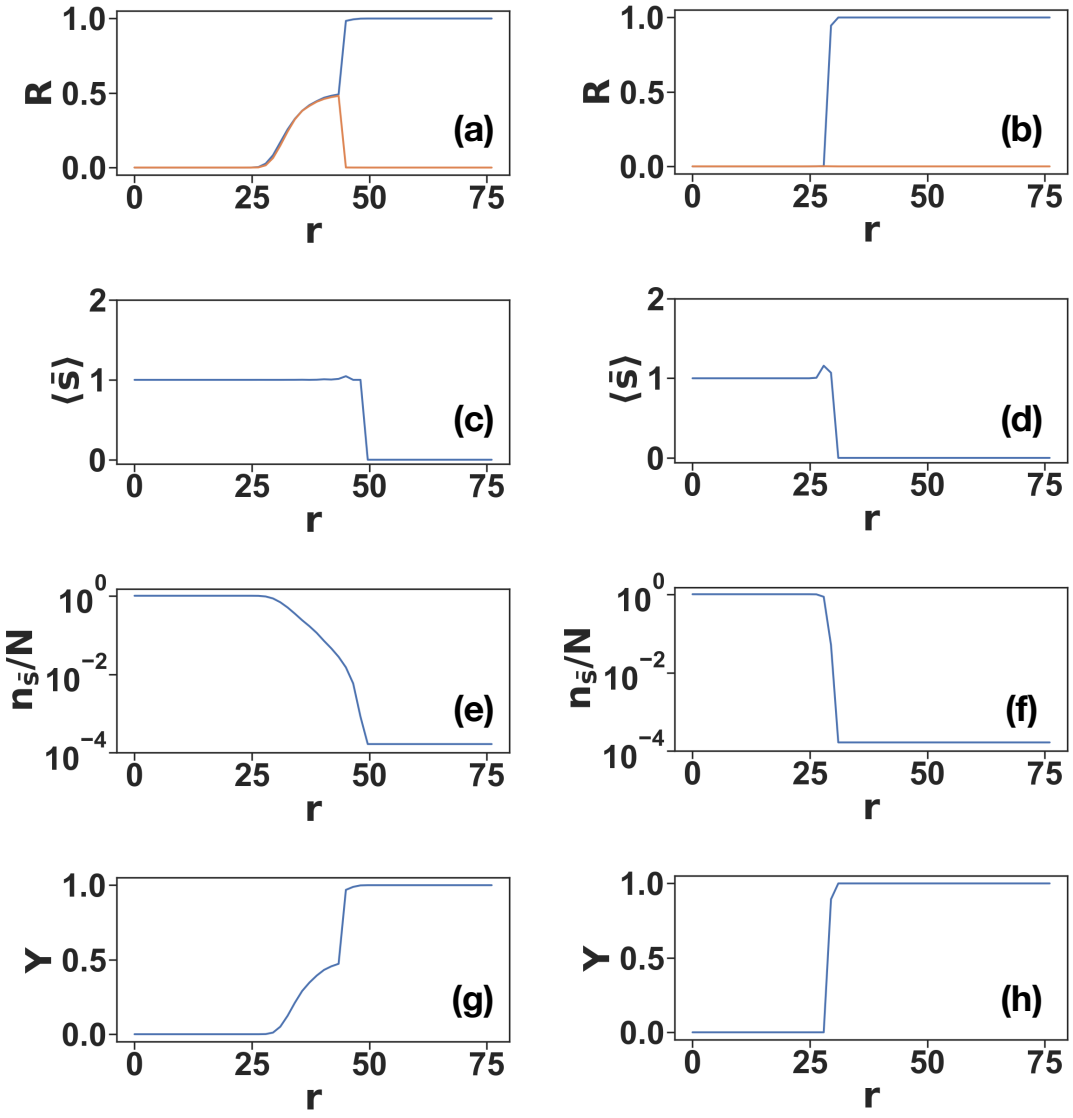

FIGURE 5.2: Same as Figure 5.1 but with IsingNets obtained from 2D Ising model simulations at $T = 2.35$.

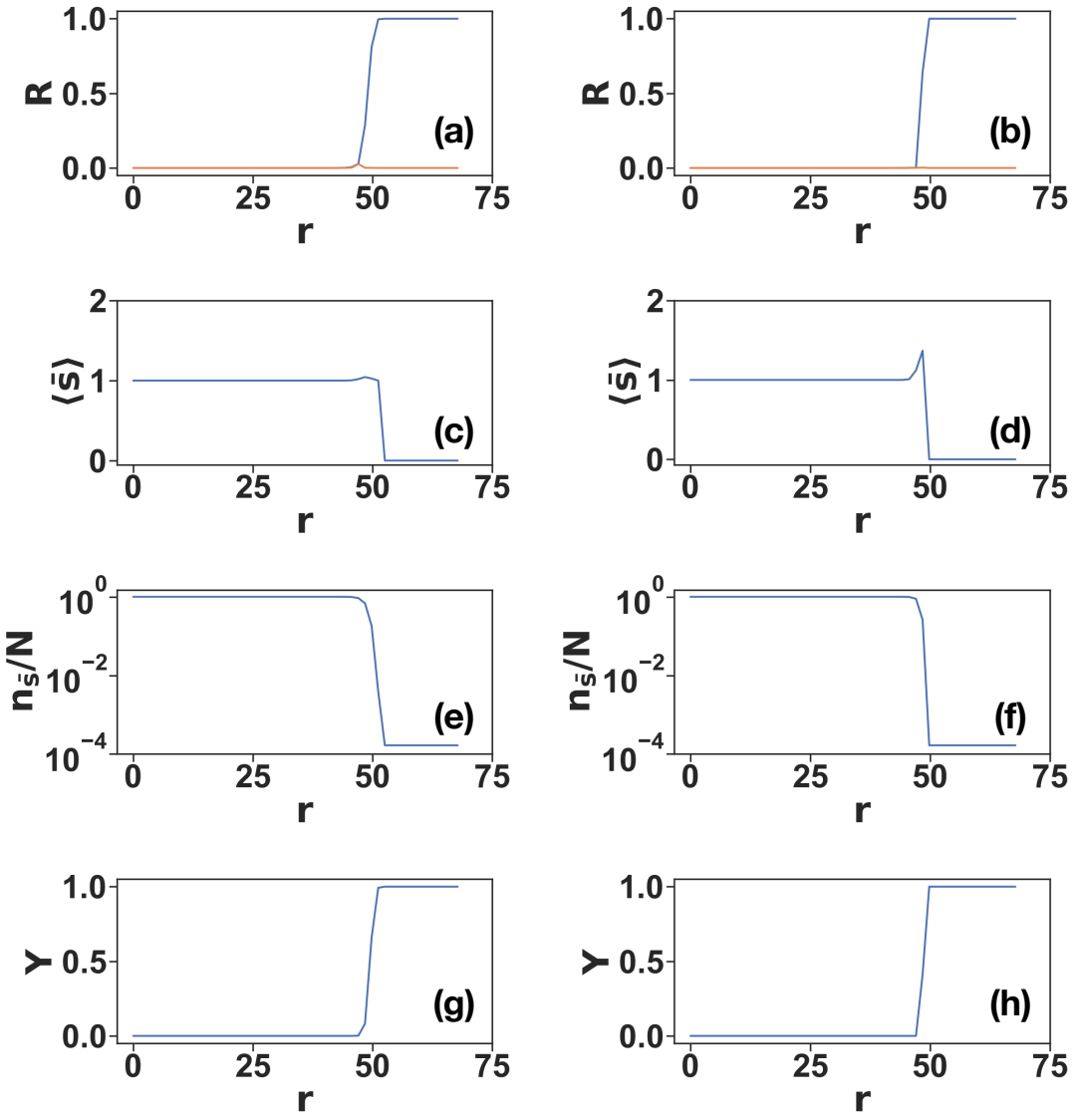

FIGURE 5.3: Same as Figure 5.1 but with IsingNets obtained from 2D Ising model simulations at $T = 2.50$.

On an Erdös-Renyi random graph the average size of finite component $\langle \bar{s} \rangle$ plays the role of the percolation susceptibility [199], diverging in correspondence with the emergence of the (single) giant component. In the IsingNets

(see Figures 5.1-5.3 the average size of finite component $\langle \bar{s} \rangle$ has a very suppressed maximum with respect to the same quantity measured on the considered null model, indicating that the agglomeration of the two giant clusters proceeds by subsequent agglomeration of very small components and isolated nodes rather than by the agglomeration of finite clusters of diverging average size as in a random graph. This is also confirmed by the behavior of the number of clusters $n_{\bar{s}}$ as a function of the filtration parameter r which for the IsingNets decays less steeply than in the randomized null model. Finally, the inverse participation ratio Y for low temperatures reveals a significant plateau at $Y = 1/2$ indicating the existence of two giant components of approximately equal size (see Figure 5.1). Above the critical temperature, for $T > T_c$ as the filtration parameter is raised only one giant component emerges and the difference with respect to the randomized null model is reduced (see Fig. 5.3).

5.3.3 Persistent homology

Topology is the study of shapes and their invariant properties under continuous deformations (see for an introduction [154, 179]). Major examples of topological invariants are the Betti numbers. The Betti number β_0 indicates the number of connected components, the Betti number β_1 indicates the number of one-dimensional holes, the number of β_2 indicates the number of two-dimensional holes, etc. For instance, a point has Betti numbers $\beta_0 = 1$ and $\beta_n = 0$ for any other value of n, a circle has non-zero Betti numbers $\beta_0 = \beta_1 = 1$ and a sphere has non-zero Betti numbers $\beta_0 = \beta_2 = 1$. An important result of algebraic topology [154, 179, 182, 180, 181] is that the $\bar{n}$-dimensional Betti number $\beta_{\bar{n}}$ is the rank of the $\bar{n}$-dimensional homology group of the considered topological space.

In the discrete setting, the Betti numbers are defined in general for simplicial complexes. Simplicial complexes are a type of higher-order network formed by a set of simplices such as nodes, links, triangles, tetrahedra, etc. They have the additional property of being closed under the inclusion of faces

of each simplex. This last property implies that if a triangle belongs to the simplicial complexes also all its links and nodes belong to the simplicial complex.

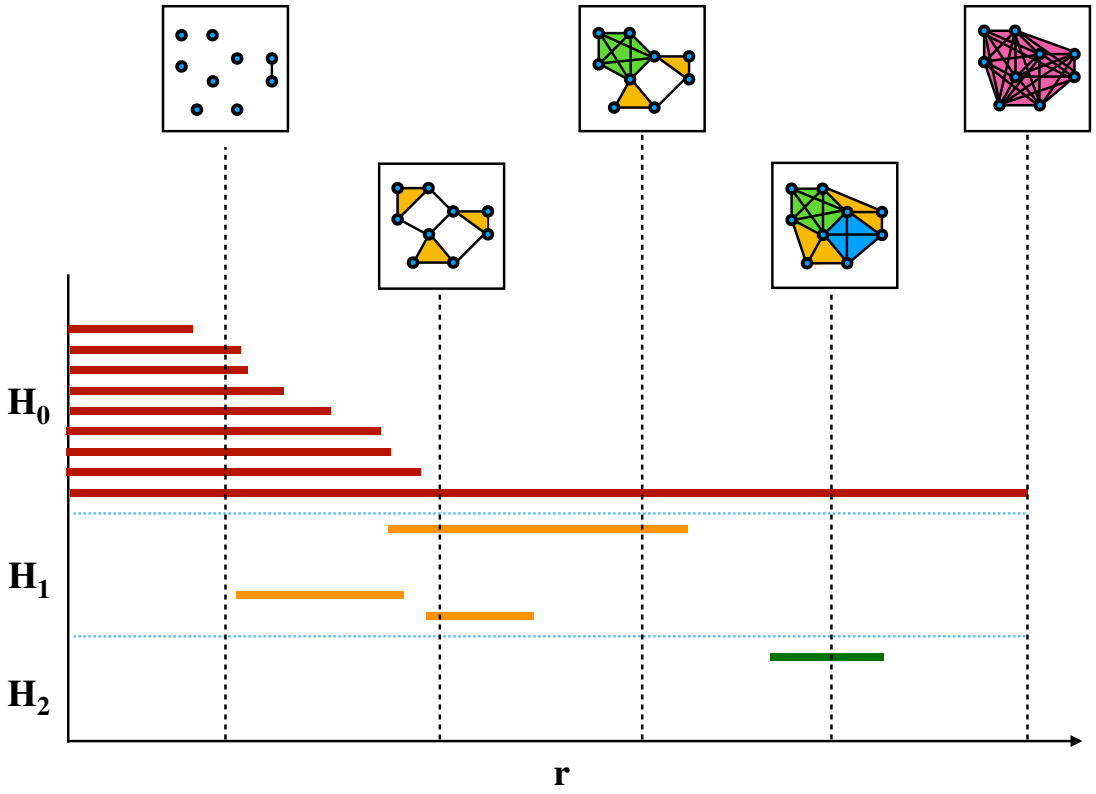

FIGURE 5.4: A schematic illustration of the filtration process. The barcodes are used to show the appearance and disappearance of topological features corresponding to different homology classes as the filtration parameter r is increased. The filtration process ends at $r = r_{\max}$ when all N nodes are fully connected and form a N-simplex. Simplices of different dimensions are indicated by different colors.

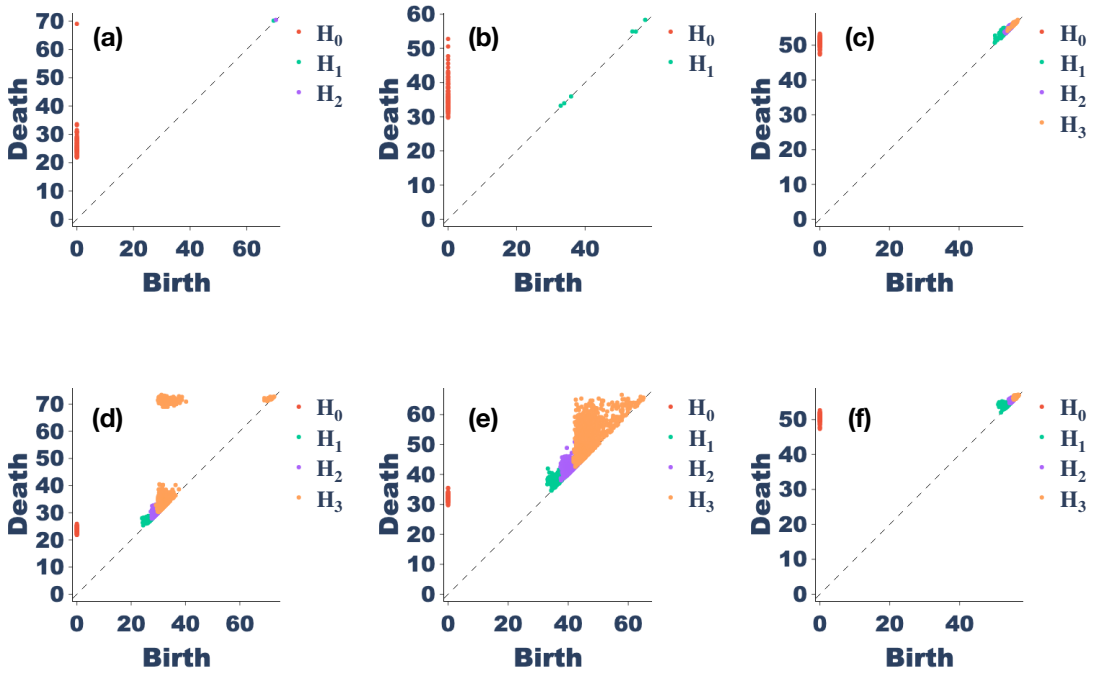

FIGURE 5.5: The persistent diagram corresponding to homology classes in H_0, H_1, H_2, and H_3 of the IsingNet clique complexes are plotted as a function of the filtration parameter r. Panels (a), (b), and (c) show the persistent diagrams of IsingNets obtained from the spin system of linear size $L = 40$ at $T = 2.12$ (a), $T = 2.25$ (b), and $T = 2.50$ (c). Panels (d), (e), and (f) show the persistent diagram of corresponding randomized null models obtained at $T = 2.12$ (d), $T = 2.25$ (e), and $T = 2.50$ (f). The networks are formed by $N = 100$ nodes.

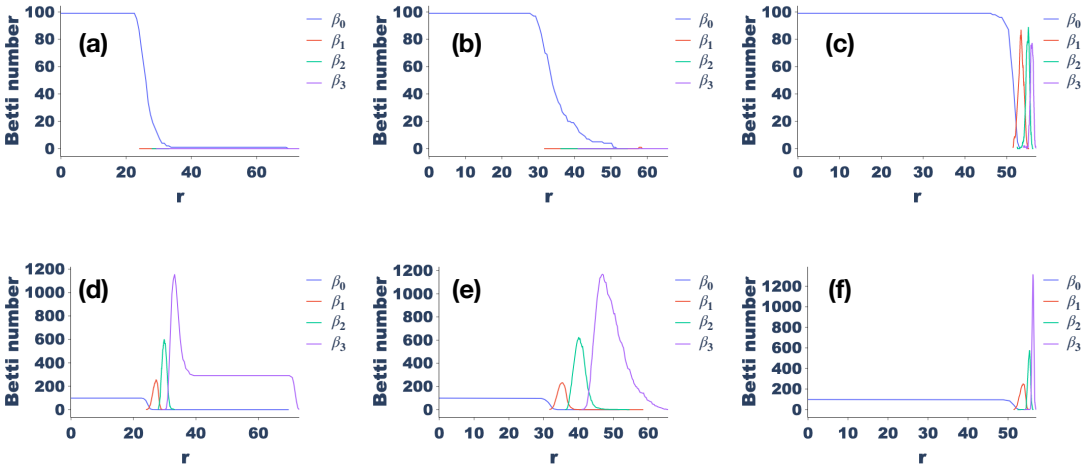

FIGURE 5.6: The Betti numbers β_0, β_1, β_2, and β_3 of the IsingNet clique complexes are plotted as a function of the filtration parameter r. Panels (a), (b), and (c) show the persistent diagrams of IsingNets obtained from the spin system of linear size $L = 40$ at $T = 2.12$ (a), $T = 2.25$ (b), and $T = 2.50$ (c). Panels (d), (e), and (f) show the persistent diagram of corresponding randomized null models obtained at $T = 2.12$ (d), $T = 2.25$ (e), and $T = 2.50$ (f). The networks are formed by $N = 100$ nodes.

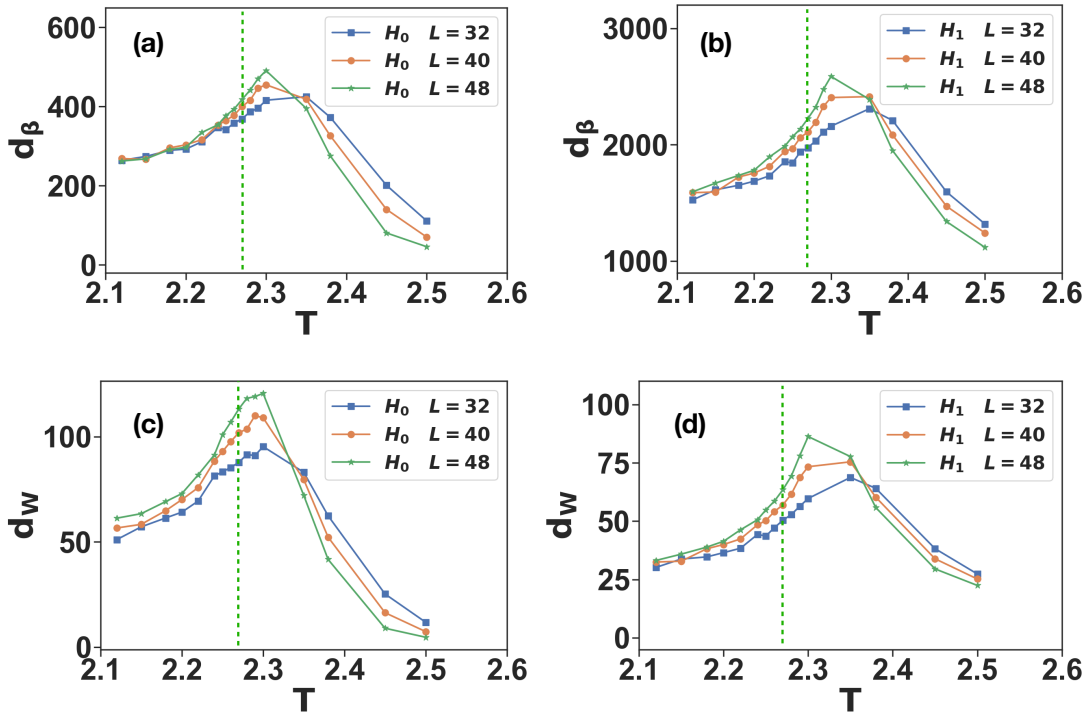

FIGURE 5.7: The Betti distance d_β (panel (a) and (b)) and the Wasserstein distance d_W (panel (c) and (d)) between the persistent diagrams of the IsingNets and its randomized null models are plotted as a function of the temperature T for homology H_0 and H_1. The plot shows the finite size scaling of the distances on systems of size $L = 32$, $L = 40$, and $L = 48$. The dashed line indicates the critical temperature T_c. The IsingNets on which the persistent diagrams have been calculated have $N = 300$ nodes.

Topological data analysis [179, 180, 181, 182, 178, 202] and in particular persistent homology allows to characterize the topological properties of data as a function of a filtration parameter and is becoming increasingly important in network and data science with applications ranging from the study of gene-expression to the investigation of brain networks. As mentioned in the introduction TDA is recently becoming a very popular computational tool to study phase transitions as well [189, 190, 191, 192, 183, 184, 185, 186, 187, 85, 188]. However, in this context, most of the TDA so far are performed on point clouds rather than on networks. In our setting, when each pair of nodes (i, j)

is assigned a distance d_{ij}, persistent homology characterizes the topology of data by forming a simplicial complex representation of the data and characterizing its homology, i.e. the connected components (H_0 homology classes); the independent cycles -one-dimensional holes- (H_1 homology classes); the two-dimensional holes (H_2 homology classes) etc. The simplicial complexes [154, 178, 180, 181] that we use to perform the topological data analysis are the so-called Vietoris-Rips complexes of the network generated by filling all the simplices having all links at distance $d_{ij} < r$. Thus as a function of r the set of simplicial complexes forms a filtration.

As the filtration parameter r is raised, first each node belongs to a different connected component, then connected components merge progressively as described also by the previous paragraph (see Figure 5.4 for a schematic description of the filtration). However, as r increases there is the potential also for one-dimensional holes (or network cycles) to emerge with each independent cycle represented by a different H_1 homological class. Eventually, as r increases these cycles will become filled and thus disappear. Moreover, also higher-dimensional holes represented by higher dimensional homological classes $H_{\bar{n}}$might arise and eventually coalesce as r is further increased giving rise to barcodes representing the topology of the data.

By monitoring the homology classes as a function of r, the results of persistent homology are typically summarized by barcodes where each homology class is represented as a bar that extends through the corresponding range of values of r for which the homology class is observed (see Figure 5.4). The barcodes are then represented by persistent diagrams where for each homology class the value of r where the homology class is disappearing (death) is plotted versus the value of r corresponding to the first appearance of the homology class (birth), see for instance Figure 5.5. Significant homology classes are the ones represented by points far from the diagonal which last for a large interval of values of the filtration parameter r.

The investigation of the persistent diagrams for the IsingNets can further characterize these discrete structures by revealing important properties about

how the clusters merge as a function of r. In particular, the persistent diagrams of the IsingNets allow us to show that clusters remain compact with a suppression of the Betti numbers with respect to the corresponding persistent diagram of the randomized null models. This result is evident from Figure 5.5 and Figure 5.6 where we plot the persistent diagram (for homology $H_{\bar{n}}$ with $\bar{n} \in \{0, 1, 2, 3\}$ and the Betti numbers $\beta_{\bar{n}}$ with $\bar{n} \in \{0, 1, 2, 3\}$ as a function of the filtration parameter r for IsingNets and their randomized null models as a function of the temperature T. Indeed the persistent diagrams of IsingNets corresponding to homology $H_{\bar{n}}$ with $\bar{n} \in \{0, 1, 2, 3\}$ (see Figure 5.5) reveal that homology classes for the IsingNets are less persistent than the homology classes of the null model as they are represented in the diagram by points closer to the diagonal than in the null model. Moreover also the Betti numbers $\beta_{\bar{n}}$ with $\bar{n} \in \{0, 1, 2, 3\}$ are strongly suppressed with respect to their null model counterpart.

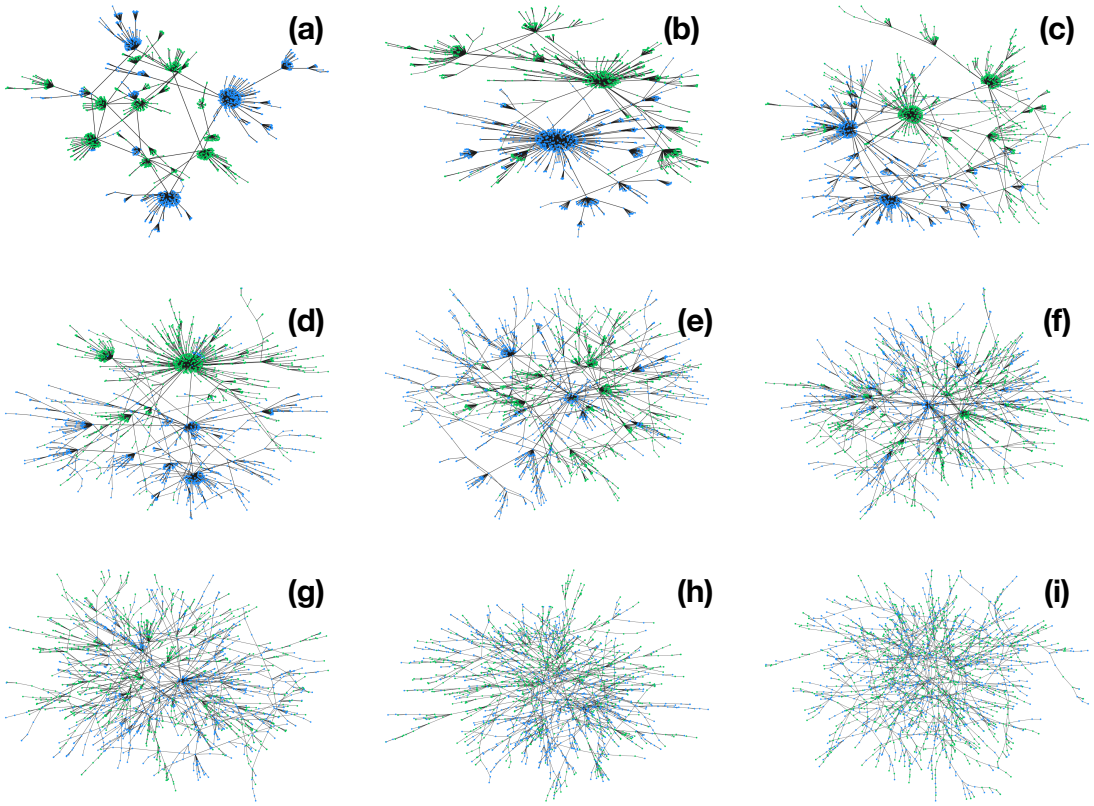

FIGURE 5.8: Minimum spanning trees (MST) of the IsingNets obtained from the spin system of linear size $L = 40$ are plotted for different temperatures T above and below the phase transition at $T_c = 2.269\ldots$. The nodes are colored using a K-means clustering algorithm with $K = 2$. The number of nodes of the MSTs is $N = 2000$. The MSTs are generated at $T = 2.12$ (a), $T = 2.20$ (b), $T = 2.26$ (c), $T = 2.28$ (d), $T = 2.30$ (e), $T = 2.35$ (f), $T = 2.38$ (g), $T = 2.50$ (h), $T = 3.50$ (i).

The persistent diagrams of IsingNets can be compared to the persistent diagrams of their randomized null model as a function of the temperature. This comparison can be performed by considering different measures of distances between persistent diagrams.

Here we show both the Weisserstein distance and the Betti distance among the persistent diagrams of the IsingNets and their corresponding randomized null model. The Weissestein distance captures the minimum distance over all

perfect matchings between points in two persistent diagrams, i.e.,

$$d_W = \left[\inf_{\eta: I \to N} \sum_{x \in I} ||x - \eta(x)||_\infty^2 \right]^{1/2} \tag{5.3}$$

where $x \in I$ indicates a point $x = (b, d)$ with birth b and dead d in the persistent diagram of the IsingNet, while $\eta(x) \in N$ indicates the matched point in the persistent diagram of the corresponding null model. The map η denotes any bijection between I and N. The Betti distance computes the L_2 distance between Betti curves of two persistent diagrams.

$$d_\beta = \left[\sum_r \left(\beta^{[I]}(r) - \beta^{[N]}(r) \right)^2 \right]^{1/2} \tag{5.4}$$

where $\beta^{[I]}(r)$ and $\beta^{[N]}(r)$ indicate the Betti number of the IsingNet and the corresponding null model with filtration parameter r.

These distance measures provide good indicators of the critical points as they display a maximum approaching the critical temperature of the 2D Ising model as the size L of the 2D lattice increases (see Figure 5.7). Note that here, due to the computational cost of calculating the persistent diagrams corresponding to higher-order homological classes, we focus here only on homology classes H_0 and H_1.

This is rather clear evidence that the IsingNets have a topology that encodes fundamental properties of the underlying spin system.

5.3.4 Minimum Spanning Tree Visualization

Our unsupervised analysis of the IsingNets includes their visualization which reveals their highly heterogeneous structure below the critical temperature T_c and at criticality. A very efficient way to visualize the IsingNets is by plotting their Minimum Spanning Trees (MST) [196, 173]. The MST is the subtree of the network whose sum of the distances between the connected nodes is minimal,

and its topology reveals important properties of the fully connected weighted distance matrix of the IsingNets. In particular, as shown in Figure 5.8 for T deep in the ferromagnetic phase the topology of the MST of the IsingNets is dominated by very relevant hub nodes, and the network displays a clear partition between the two clusters detected by the K-means algorithm with $K = 2$ indicated in the figure by two different colors of the nodes. As the temperature is raised to the critical region, $T \simeq T_c$ the hubs of MST become less dominant. Above the phase transition, the MST becomes clearly more random with a suppression of the hubs in the MST.

5.3.5 Network embedding

Our analysis of the IsingNets is here enriched by considering the UMAP 2-dimensional embedding of the fully connected network endowed with the distance matrix $\mathbf{d}$. UMAP is a widely used embedding algorithm exploiting dimension reduction (for more detail see for instance Ref.[203]). The embedding is here performed as a function of the temperature T and the nodes are colored according to their clustering in two groups performed using K-means (see Figure 5.9).

The UMAP embedding provides a clear visualization of the two main clusters of the nodes of the IsingNets present for $T < T_c$ and corresponding to the symmetry among configuration snapshots and of their merging as the temperature T is raised above the critical temperature.

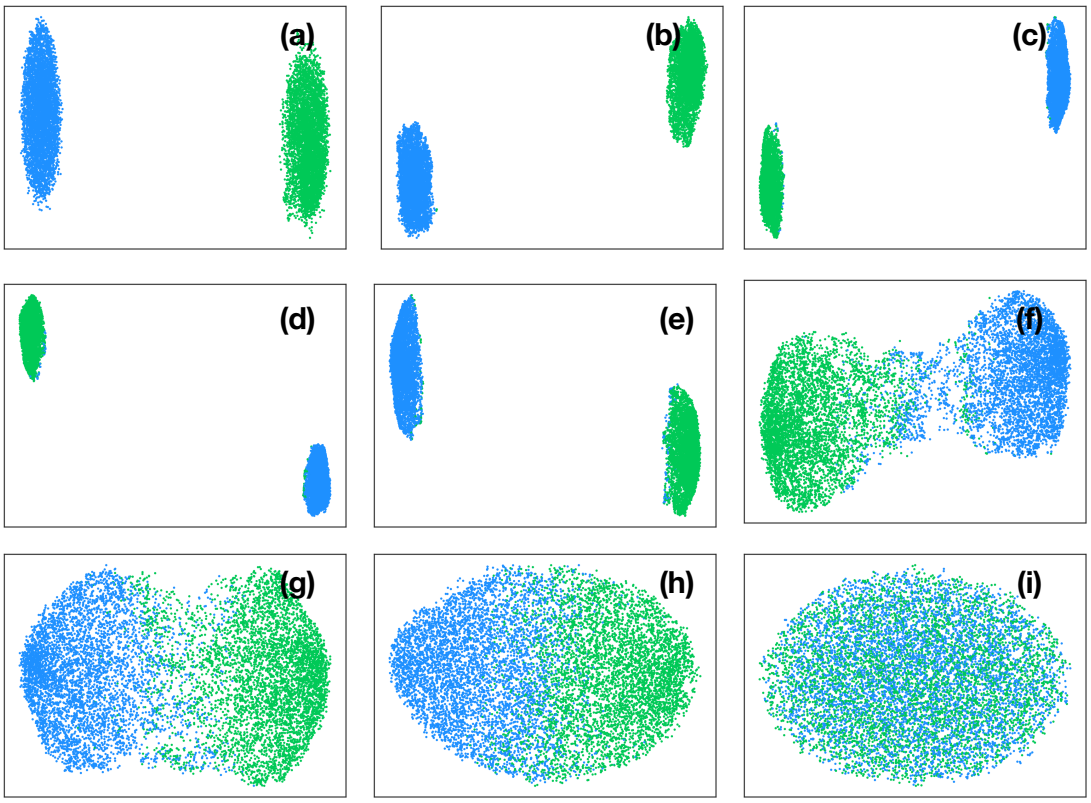

FIGURE 5.9: IsingNets obtained from the spin system of linear size $L = 40$ at different temperatures T are embedded into a two-dimensional space using Uniform Manifold Approximation and Projection (UMAP) embedding algorithm. The nodes are colored using a K-means clustering algorithm. The number of nodes of the IsingNets is $N = 10^4$. The IsingNets are generated at $T = 2.12$ (a), $T = 2.20$ (b), $T = 2.26$ (c), $T = 2.28$ (d), $T = 2.30$ (e), $T = 2.35$ (f), $T = 2.38$ (g), $T = 2.50$ (h), $T = 3.50$ (i).

5.3.6 Metric-based centrality measures

To conclude our characterization of the IsingNets without imposing a fixed value of the filtration parameter, we present here the statistical properties of some important geometrical aspects of the IsingNet captured by the closeness centrality of the nodes.

The closeness centrality Cl_i of a node i measures how close is the node to the other nodes of the network, and is defined as the inverse of the average distance of node i to the other nodes of the network, i.e.

$$CL_i = \frac{N-1}{\sum_{j \neq i} d_{ij}} \tag{5.5}$$

We investigate the statistical properties (the mean $\langle Cl \rangle$, the standard deviation $\sigma(Cl)$, the skewness $Sk(Cl)$ and the kurtosis $Ku(Cl)$) of the distribution of the closeness centrality in IsingNets as a function of the temperature T. In Figure 5.10 we show that the average closeness centrality decreases as a function of the temperature, demonstrating that on average the distances between the nodes are higher at higher temperatures. The higher-order statistics of the closeness centrality distribution are even more revealing of the IsingNets organization and the skewness and kurtosis of the closeness distribution provide a good unsupervised indicator of the critical point (see Figure 5.10). Indeed the standard deviation of the closeness centrality displays a maximum for temperatures close to the critical temperature; the skewness of the closeness centrality is negative below the critical temperature and positive above the critical temperature and the kurtosis becomes negative above the critical temperature, while at very high temperature is is strongly affected by noise.

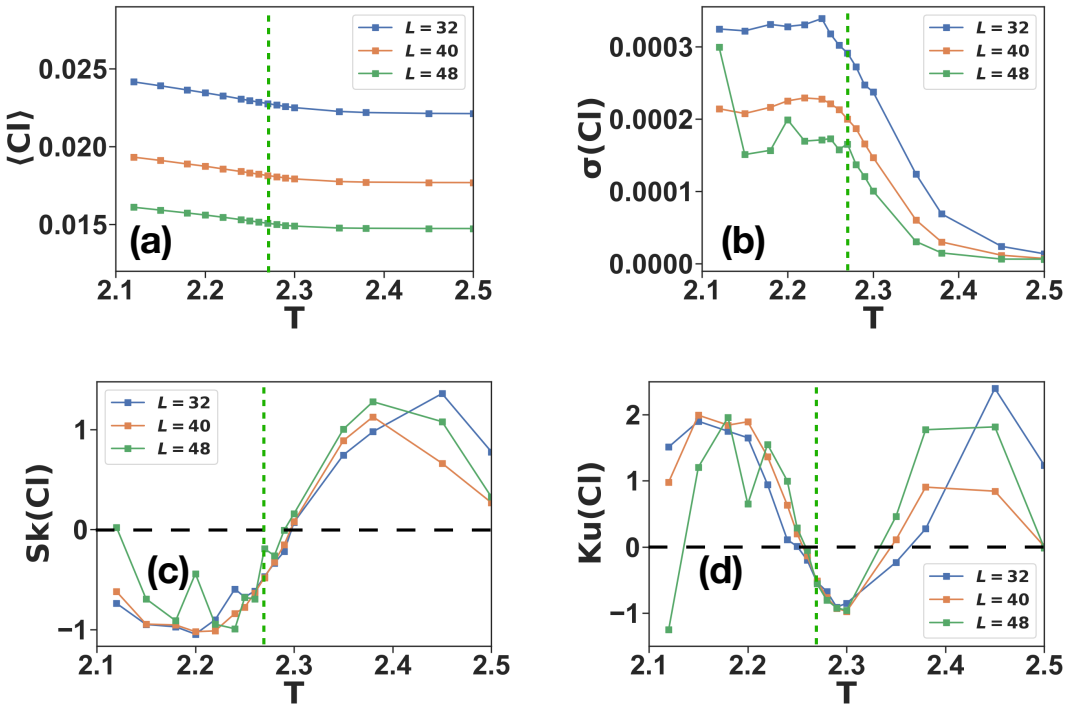

FIGURE 5.10: The mean $\langle Cl \rangle$, standard deviation $\sigma(Cl)$, skewness $Sk(Cl)$, and kurtosis $Ku(Cl)$ of closeness distribution at different temperatures T and sizes L are plotted on the IsingNets where all the distance between each pair of nodes is retained. The closeness is calculated via Eq. 5.5. The corresponding random networks are formed by randomly permuting the distances between node pairs. The average closeness centrality of the IsingNets coincides by construction with the average of the null model (not shown) however the higher-order statistics strongly depart from the null model behavior. The vertical green dashed lines indicate the critical temperature T_c the horizontal black dashed lines indicate $Sk(Cl) = 0$ in panel (c) and $Ku(Cl) = 0$ in panel (d).

5.4 Network characterization of IsingNets at a given value of the threshold distance r

5.4.1 IsingNets at given threshold distance r

In this section, we study the properties of IsingNets where we fix a given choice of filtration parameter r. In particular we will consider the IsingNets, whose $N \times N$ adjacency matrix $\mathbf{A}$ has elements

$$A_{ij} = \theta(r - d_{ij}), \tag{5.6}$$

with $\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ otherwise. In the following, we will indicate with $i \sim j$ two neighbour nodes for which $A_{ij} = 1$.

We adopt the same choice of the parameter r proposed in Ref. [193] where it was shown that for a wide range of choices of r the IsingNets are scale-free presenting often power-law exponents less than two which are known to occur in a variety of context [204, 205, 206, 207]. In particular, here we take r equal to the average distance of the 5^{th} nearest neighbour.

Similarly, the randomized network forming our null model in this section is performed by thresholding the randomized distance matrix $\mathbf{d}^{rand}$ with the same threshold used for the corresponding IsingNet.

We consider the statistical and combinatorial properties of these networks where we assign to each link (i, j) a weight w_{ij} equal to the inverse of the distance d_{ij}, provided this distance is smaller than r, i.e.

$$w_{ij} = \frac{1}{d_{ij}} A_{ij} \tag{5.7}$$

We provide an in-depth network analysis of these networks investigating their degree and strength distribution, and going beyond these statistical properties providing evidence of the presence of degree correlations, of a nontrivial k-core structure and of weight degree correlations. Moreover, the investigation of the spectral properties of the IsingNets will demonstrate non-trivial

signatures of criticality. This analysis provides clear evidence that not only the degree distribution of IsingNets is strongly different from the one of an Erdös-Renyi random graph, but the network also obeys important higher-order correlations reflecting the correlations existing in the spin system configuration snapshots.

5.4.2 Degree and strength distribution

One of the most simple yet fundamental property of a network is its degree distribution $P(k)$ which characterize globally the network starting from the knowledge of the node's degrees where the degree k_i of the node i indicates the sum of the links incident to it, i.e.

$$k_i = \sum_{j=1}^{N} A_{ij}. \tag{5.8}$$

Thus while the degree is a local property of the nodes the degree distribution $P(k)$ is able to characterize the global properties of the networks. In particular scale-free [208, 209, 210] and in general degree distribution with second $\langle k^2 \rangle$ (and eventually first $\langle k \rangle$) moment diverging with the network size have been shown to have a significant role in determining the response of the network to random damage and the critical behavior of the dynamics defined on top of these networks, such as epidemic spreading and the Ising model [209, 149].

For weighted networks it is also possible to define weighted degree also called *strength* s_i of the generic node i [211] given by the sum of the weights of its incident links, i.e.

$$s_i = \sum_{j=1}^{N} w_{ij}. \tag{5.9}$$

From the sequence of the node's strengths, it is possible to extract also the strength distribution $P(s/s_0)$ where s_0 is the minimal strength of the links. This distribution has been shown to display broad distributions in a number of

real weighted networks, such as collaboration networks and airport networks [212].

In Figure 5.11 we plot the degree $P(k)$ and strength $P(s/s_0)$ distribution for temperatures below and above the phase transition demonstrating that these distributions are broad. We note that below the critical temperature T_c, the networks are not only broad but also dense, i.e. having an average degree growing with the network size (for models of these networks see [204, 205, 206, 207]).

To quantify the scale-free nature of these distributions we plot the ratio between the second moment, the average degree $\langle k \rangle$ and the average strength $\langle s \rangle$ as a function of the temperature, showing that that $\langle k \rangle$, and $\langle k^2 \rangle / \langle k \rangle$ are good indicators of the critical point displaying a maximum for $T = T_c$ without noticeable finite size effects for the sizes investigated in this work (see Figure 5.12).

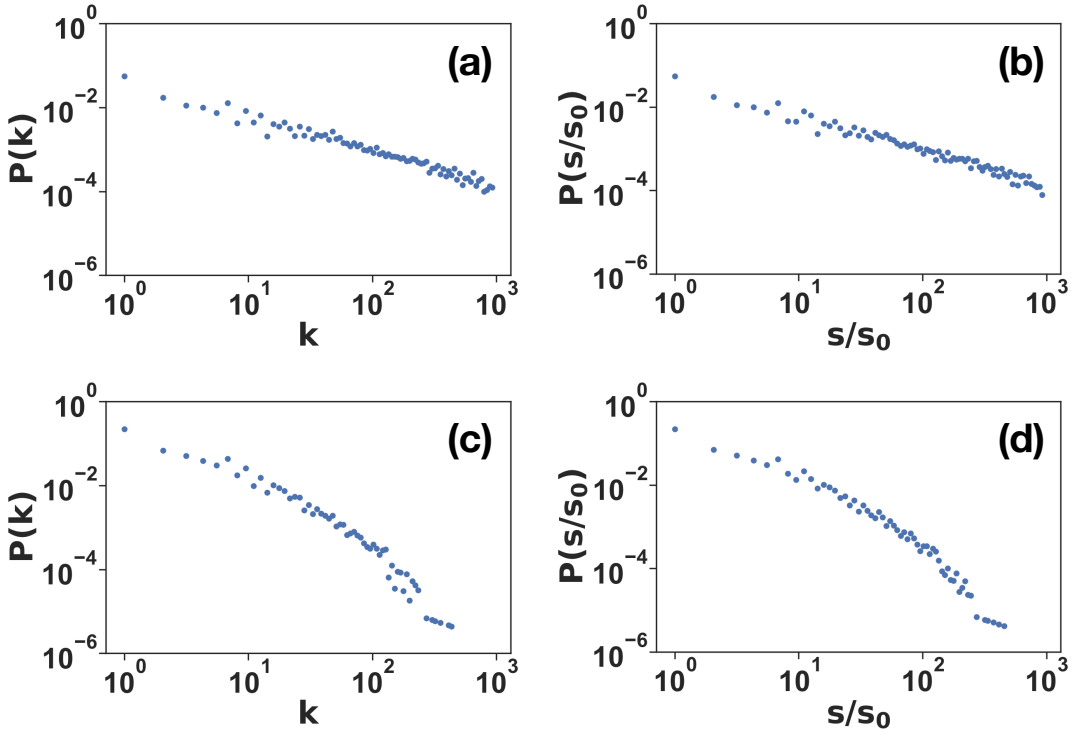

FIGURE 5.11: Degree distribution $P(k)$ and strength distribution $P(s/s_0)$ of IsingNets obtained from the spin system of linear size $L = 40$ formed by $N = 10^4$ nodes at temperature $T = 2.12$ and $T = 2.50$. Panel (a) shows the degree distribution of IsingNet at $T = 2.12$. Panel (b) shows the strength distribution of the same network as panel (a). Panel (c) is the same as panel (a) but obtained at $T = 2.50$. Panel (d) is the same as panel (b) but obtained at $T = 2.50$.

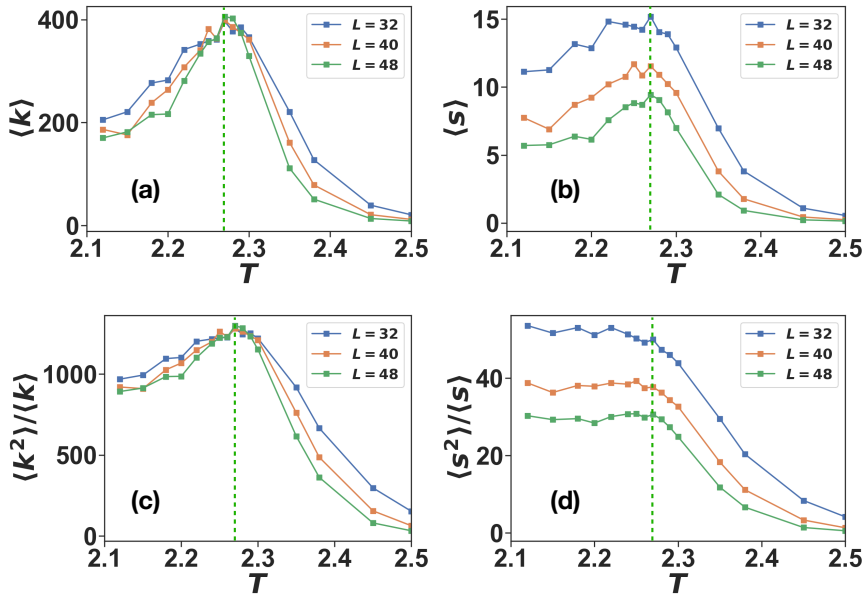

FIGURE 5.12: The average degree $\langle k \rangle$ (panel (a)) and the average strength $\langle s \rangle$ (panel (b)) are plotted together with the ratio $\langle k^2 \rangle / \langle k \rangle$ (panel (c)) and $\langle s^2 \rangle / \langle s \rangle$ (panel (d)) as a function of temperature T. The dashed line indicates the critical temperature T_c. The IsingNets are formed by $N = 10^4$ nodes.

5.4.3 Degree correlations

In order to go beyond single node statistics and to explore how far IsingNets are from random networks with a given degree distribution, in this section we characterize the degree correlations [150, 151] of the IsingNets. Degree-degree correlations measure to what extent the degree of two end nodes of the same link are correlated. Degree correlations are typically classified as either assortative or disassortative. Assortative degree correlations imply that highly connected nodes are more likely to be connected to highly connected nodes while low degree nodes are more likely to be connected to low degree nodes than in a maximally random network with the same degree distribution. Conversely, disassortative networks are networks in which highly connected nodes

are more likely to be connected to low degree nodes than in the maximally random network with the same degree distribution. Examples of assortative networks are social networks while examples of disassortative networks include the Internet and the protein interaction networks. The degree-degree correlations can be quantified by considering the average degree of the neighbour of a node $k_{nn}(i)$ defined as

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j \sim i} k_j \tag{5.10}$$

When $k_{nn}(i)$ tends to be higher for nodes of higher degree k_i the network is classified as assortative. Instead when $k_{nn}(i)$ is typically lower for nodes of higher degree k_i then the network is classified as disassortative. The IsingNets are clearly displaying a disassortative behavior for $T < T_c$ that deviates strongly from the behavior of the null model in which the distance matrix $\mathbf{d}$ is reshuffled (see Figure 5.13). On the contrary, for $T > T_c$ the trend of k_{nn} versus k does not appear to be fully monotonic, while nodes of larger degrees remain more likely to connect to low degree nodes.

Interestingly the degree correlations are also affecting the average clustering coefficient $C(k)$ [213, 151, 214] of nodes of degree k which display a decay as a function of k typical of networks with disassortative degree correlations (see Figure 5.13).

Degree-degree correlations can also be detected by the Pearson correlation coefficient $\bar{r}$ [150] between degrees of linked nodes which is plotted as a function of the temperature in Fig. 5.14(a) showing a clear negative (disassortative) correlations for low temperatures which strongly deviates from the null model. In Fig. 5.14(b) we also report the average clustering coefficient C as a function of the temperature showing that IsingNets display a much larger average clustering coefficient than the null model counterpart and that this average clustering coefficient is higher deep in the ferromagnetic phase (lower temperatures).

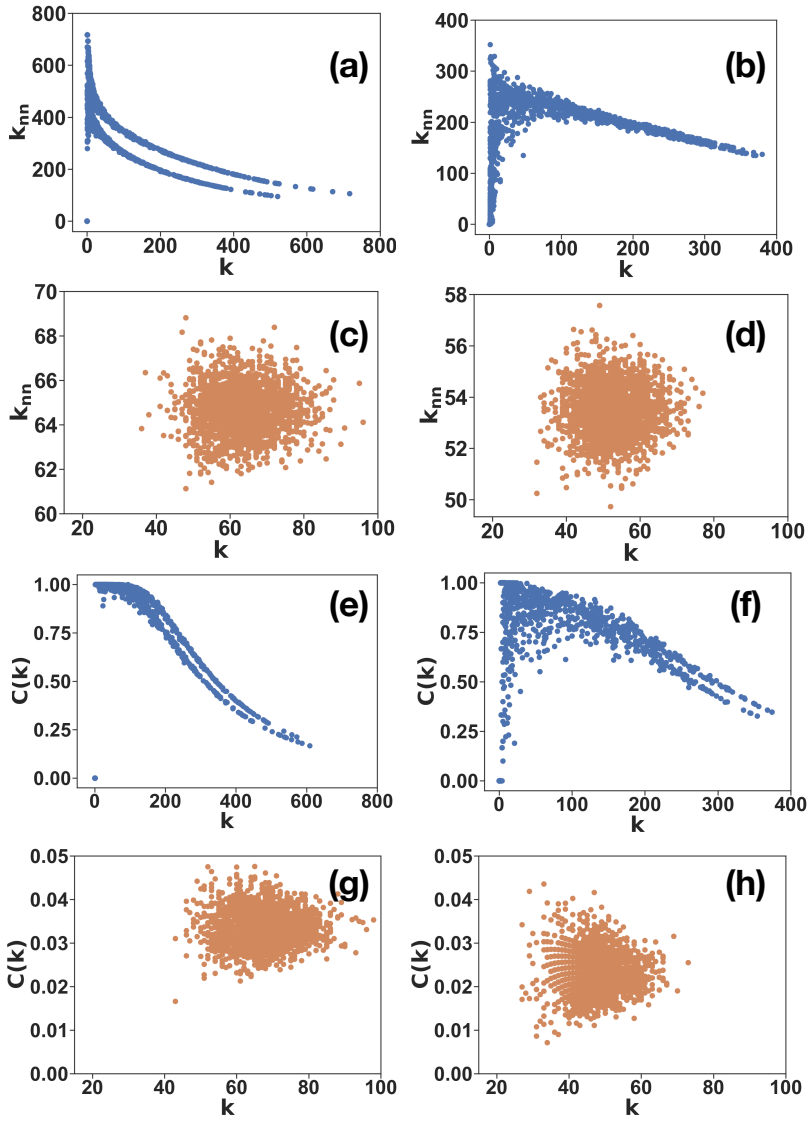

FIGURE 5.13: Average nearest neighbour degree k_{nn} and clustering coefficient $C(k)$ on IsingNets obtained from Monte Carlo simulations of the spin system of linear size $L = 40$ and corresponding null models with $N = 2000$ nodes. The threshold distance of connecting two nodes is the 5th nearest neighbour average distance. The random network is obtained by randomly permuting distances between node pairs and nodes are connected with the same threshold. The average neighbour degree *knn* is shown as a function of degree k on IsingNets and corresponding null models. Panel (a) and (b) show *knn* obtained at $T = 2.12$ (a) and $T = 2.35$ (b). Panel (c) and (d) show *knn* of the corresponding null model at $T = 2.12$ (c) and $T = 2.35$ (d). Panel (e) and (f) show $C(k)$ obtained at $T = 2.12$ (e) and $T = 2.35$ (f). Panel (g) and (h) show $C(k)$ of the corresponding null model at $T = 2.12$ (g) and $T = 2.35$ (h).

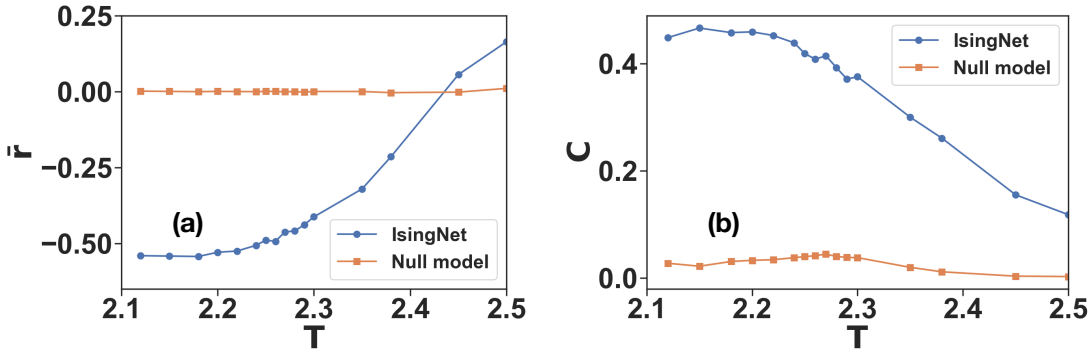

FIGURE 5.14: Pearson correlation coefficient $\bar{r}$ (a) and average clustering coefficient C (b) on IsingNets and corresponding null models formed at different temperatures. The Pearson correlation coefficient r is calculated on networks with 10^4 nodes and the average clustering coefficient C is calculated for IsingNet with $N = 2000$ nodes coming from Monte Carlo simulation with linear size $L = 40$. The threshold distance of connecting two nodes is the 5th nearest neighbour average distance. The random network is obtained by randomly permuting distances between node pairs and nodes are connected with the same threshold.

5.4.4 K-core structure

Networks can be decomposed in nested K-cores characterizing their core-periphery structure [215, 216, 217]. A K-core is a subgraph of the network formed by a set of $M(K)$ nodes each having at least K connections with the other nodes of the set. Power-law networks with exponent $\gamma \in (2, 3]$ display a significant K-core structure with the maximum K diverging with the network size and a power-law decay of $M(K)$ as a function of K. On the contrary Erdös and Renyi networks have a finite number of K-cores also when the average degree diverges. Here we show that IsingNets have a very rich K-core structure having statistical properties that change below and above the critical temperature (see Figure 5.15). Indeed above the critical temperature, we observe a behavior similar to the expected behavior for sparse scale-free networks with power-law exponent between two and three presenting a broad (seemingly power-law straight line on a log-log plot) decay of $M(K)$ versus K. However, below the critical temperature where the average degree diverges, the K-cores include more nodes

while the decay of $M(K)$ versus K is better approximated by an exponential (straight line in a log-linear plot) rather than by a power-law.

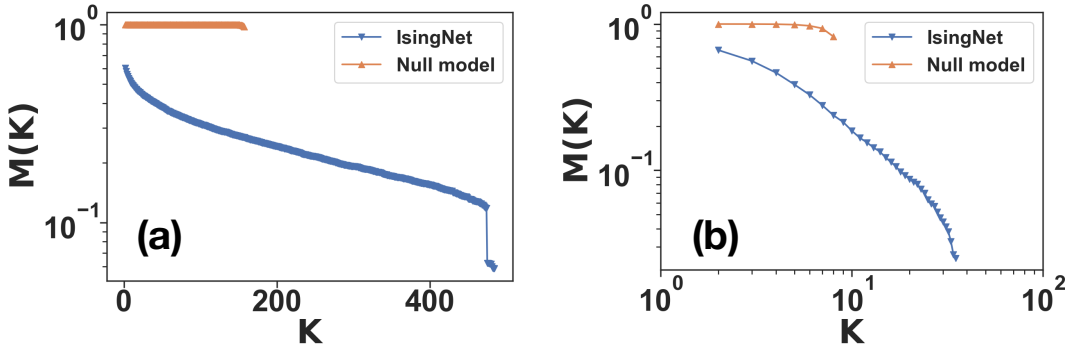

FIGURE 5.15: Fraction of nodes in the $K-$core $M(K)$ as a function of the core size K on IsingNet and null model with $N = 10^4$ nodes at $T = 2.12$ (a) and $T = 2.50$ (b).The IsingNets are generated from 2D Ising model Monte Carlo simulations of the spin system of linear size $L = 40$. Panel (a) is shown with a linear-log scale and panel (b) is shown with a log-log scale.

5.4.5 Weight-topology correlations

Interestingly in weighted networks not only the network topology can reveal relevant degree correlations showing that the networks deviate from maximally random networks, but also the weights can be distributed in a non-random way. In particular, there are two main network analyses that are able to detect weight-topology correlations. The first analysis [211] involves studying the normalized strength s/k versus the degree k for each node of the network. If the weights are distributed randomly and independently on the degree of the two end nodes there should not be any significant dependence of s/k with k. Conversely if s/k increases with k it implies that nodes with higher degrees are incident in average to links with larger weights. The second analysis [218] investigates the weight-topology correlations aiming at revealing the weight heterogeneity among links that connect to the same node. This heterogeneity if any can be quantified by calculating the inverse participation ratio Y for the

weights of the links ending to node i, defined as

$$Y_i = \sum_{j \sim i} \left(\frac{w_{ij}}{s_i} \right)^2 . \tag{5.11}$$

If the weights w_{ij} of the links (i, j) incident to node i are homogeneous, $Y_i \sim 1/k_i$. If the weights are highly heterogeneous, Y_i^{-1} indicates the effective number of links with significant weight. Interestingly when we measure Y_i for the IsingNets we observe that this second type of weight heterogeneity is missing in the data and that $Y_i \sim 1/k_i$ indicating that for each node i the weights of the links incident to it have all weights of comparable order or magnitude (see figure 5.16).

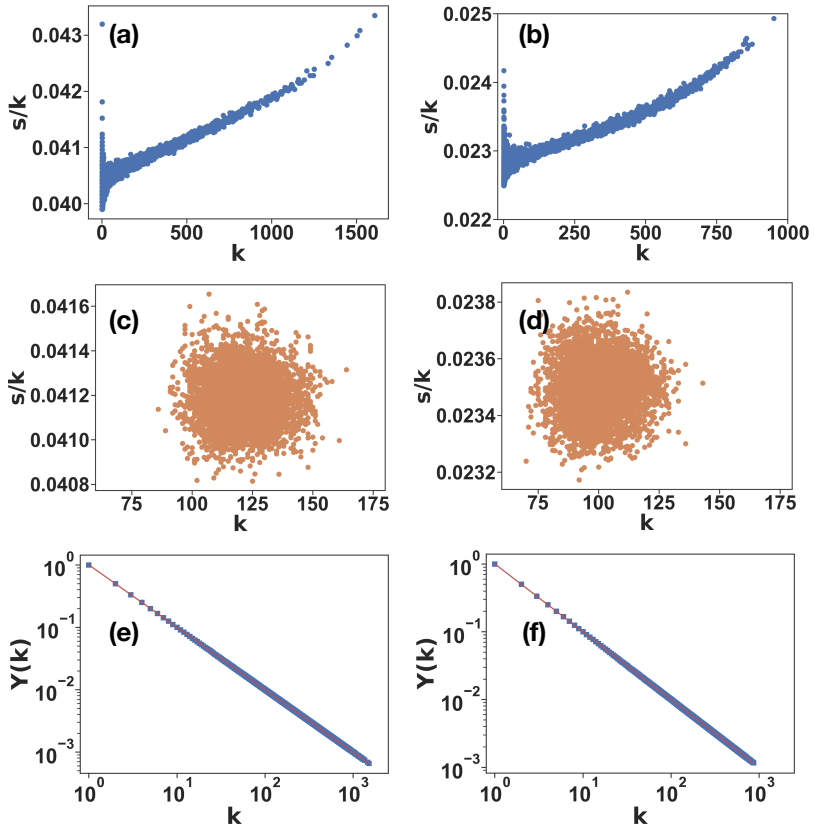

FIGURE 5.16: The ratio of strength and degree and the inverse participation ratio on IsingNets obtained from Monte Carlo simulations of the spin system of linear size $L = 40$ are shown versus degree on networks formed at different temperatures. The networks are formed by $N = 5000$ samples. The threshold distance of connecting two nodes is the 5th nearest neighbour average distance. The random networks are obtained by randomly permuting distances between node pairs and nodes are connected with the same threshold. The first row shows the strength degree ratio s/k versus degree k on IsingNets and the second row shows which on corresponding random networks. The third row shows the inverse participation ratio $Y(k)$ versus degree k. The networks are formed by simulation obtained at temperature $T = 2.12$ (left column), $T = 2.35$ (right column).

5.4.6 Spectral properties of Ising networks

The IsingNets do not only have very interesting combinatorial and statistical properties encoded in their highly correlated structure but display also relevant geometrical properties reflected in their spectrum. In particular, the critical IsingNets display non-trivial spectral properties characterized by a power-law scaling close to criticality and a highly degenerate spectral gap. The spectral properties of networks are usually probed by considering the spectrum of the graph Laplacian describing diffusion processes. The graph Laplacian $\mathbf{\Delta}$ is defined as $\mathbf{\Delta} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal matrix whose diagonal elements are the degrees of the nodes (i.e. $D_{ii} = k_i$), and $\mathbf{A}$ is the adjacency matrix of the network. The graph Laplacian $\mathbf{\Delta}$ is semi-definite positive and the spectrum always includes a zero eigenvalue with degeneracy given by the number of connected components of the network, i.e. given by the Betti number β_0. The smallest non-zero eigenvalue of the graph Laplacian of a network is also called the Fiedler eigenvalue and is typically indicated as λ_2 (as it is the second smallest eigenvalue in a connected network). In the literature, often one distinguishes between network models displaying a finite Fiedler eigenvalue $\lambda_2 \to \lambda_2^\star > 0$ in the limit $N \to \infty$ and network models in which $\lambda_2 \to 0$ as $N \to \infty$. In the first case, we say that the networks display a spectral gap whereas in the latter case we say that the "spectral gap closes". Examples of networks with finite spectral dimension are random graphs above the percolation threshold and examples of networks in which the spectral gap closes are finite dimensional lattices.

In several networks in which the spectral gap closes, it is possible to observe the spectral dimension d_S [154, 219]. The spectral dimension d_S is the dimension perceived by diffusion processes on the networks encoded in the graph Laplacian. On a lattice, the spectral dimension coincides with the Euclidean dimension d of the lattice while on general network topology, the spectral dimension can be distinct from the Hausdorff dimension of the network. Interestingly also small world networks with infinite Hausdorff dimension can

have a finite spectral dimension $d_S \geq 2$ [154, 220]. The spectral dimension determines the scaling of the cumulative density of the eigenvalues $\rho_c(\lambda)$ of the graph Laplacian for $\lambda \ll 1$ in networks where the spectral gap closes. In particular we have that networks with a spectral dimension d_S have a cumulative distribution $\rho_c(\lambda)$ that obeys for $\lambda \ll 1$

$$\rho_c(\lambda) \simeq C\lambda^{d_S/2}, \tag{5.12}$$

where C is a constant. In Figure 5.17 we show that the IsingNets display nontrivial spectral properties that have very peculiar characteristics strongly deviating from their corresponding null model. Particularly noticeable are the spectral properties of IsingNets close to the critical point where one observes the coexistence of a highly degenerate finite Fiedler eigenvalue λ_2 with a power-law scaling of the cumulative distribution

$$\rho_c(\lambda) \simeq C\lambda^{\hat{d}/2}, \tag{5.13}$$

for $\lambda > \lambda_2$ with a exponent given by $\hat{d} \simeq 0.78 \pm 0.04$ for $L = 40$. Above the critical temperature, the degeneracy of the Fiedler eigenvalue is reduced and one observes a nontrivial spectrum reminiscent of the scale-dependent spectral dimension discussed in Refs. [221] within the critical region that at higher temperatures converges with the spectrum of the null model. Below the critical dimension, the spectral gap remains highly degenerate while the rest of the spectrum remains broadly distributed.

The spectrum of the graph Laplacian is also key to characterize the network von Neumann entropy S_{VN} [222, 223, 224] defined as

$$S_{VN} = -\sum_{\lambda} \frac{\lambda}{\langle k \rangle N} \ln\left(\frac{\lambda}{\langle k \rangle N}\right). \tag{5.14}$$

The von Neumann entropy strongly departs from the von Neumann entropy of the null model for low temperatures displaying a local maximum for $T = T_c$ (see Figure 5.18).

An interesting open question that will be addressed in the following works is the relation between these spectral properties of the IsingNets graph Laplacians and the intrinsic dimension and the entropy measures that have been recently proposed starting from the unsupervised PCA analysis of spin systems [142, 1].

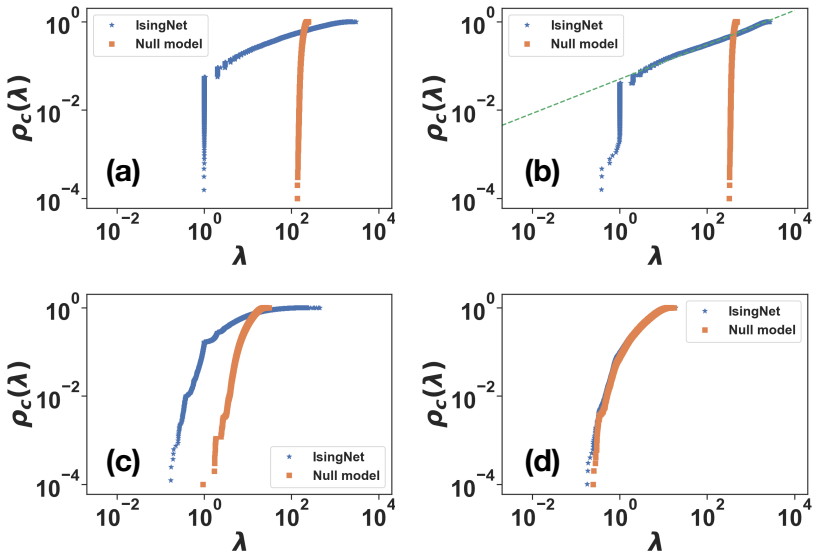

FIGURE 5.17: The cumulative distribution $\rho_c(\lambda)$ of the eigenvalues λ of graph Laplacian $\mathbf{\Delta}$ of IsingNets and corresponding null models at different temperatures T. The distribution is shown at $T = 2.12$ (a), $T = 2.27$ (b), $T = 2.50$ (c), and $T = 3.50$ (d). Data are shown for Isingnets generated from Monte Carlo simulations of spin system of linear size $L = 40$. In panel (b), the dashed lines indicate a power-law growth shown in Eq. 5.13 with exponent $\hat{d} = 0.78 \pm 0.04$.

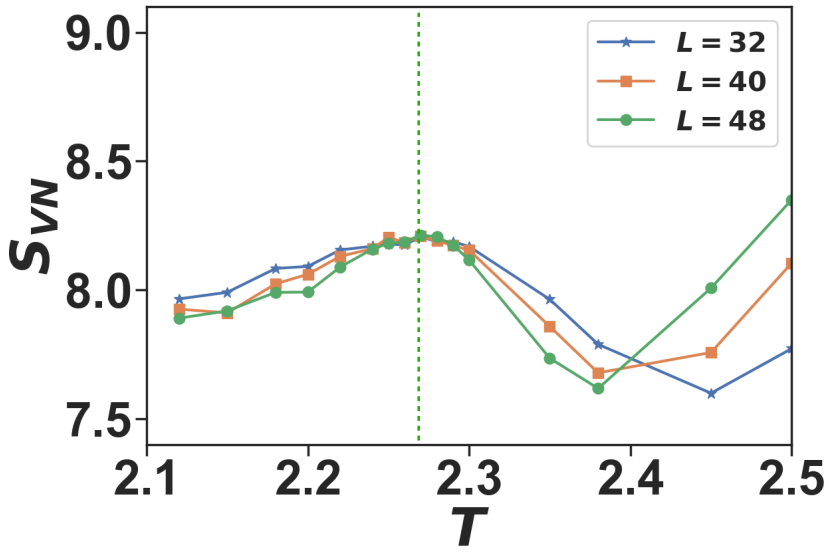

FIGURE 5.18: The von Neumann entropy S_{VN} of different system size L is shown versus the temperature T. The IsingNets are formed by $N = 10^4$ nodes. The dashed line indicates the critical temperature T_c.

5.5 Conclusions

In this work, we have launched a systematic network analysis of unsupervised learning of different states of matter. The analyzed IsingNets obtained from Monte Carlo simulations of the 2D Ising model are shown to reveal the statistical, combinatorial, geometrical and topological organization of these networks. Through the paper, we have shown that true IsingNets are highly non-random by comparing their structural properties with the structural properties of the randomized counterparts. Importantly, we have also identified several indicators of the phase transition.

We have addressed the characterization of IsingNets following two different approaches. In the first approach, we have studied the structure of IsingNets as the filtration parameter r is increased, which enforces an effective percolation process in which nodes are subsequently aggregated by considering connected pairs of nodes at increasing distances. This percolation process

reveals the presence of two giant components in IsingNets in the ferromagnetic state, each one corresponding to configurations with different magnetization. The same filtration scheme has also been used here to study the topology of the data by constructing the clique complexes of the IsingNets and calculating their persistent diagram. Interestingly, the persistent diagrams reveal that real IsingNets are formed by compact clusters as the Betti numbers of their clique complex are strongly suppressed with respect to the clique complex of the corresponding randomized null models. This network analysis conducted across the filtration is also enriched by effective visualization of the network embedding conducted using MST and the UMAP embedding and by the statistical characterization of the distribution of closeness centralities.

Secondly, our investigation of IsingNets has been conducted by considering only links at a distance less than a threshold value r taken to be the average distance of the $5th$ nearest neighbours nodes according to the (fully connected) distance matrix. These networks display a broad degree and strength distribution but their complexity extends well beyond the degree and strength distribution because IsingNets have strong degree-degree correlations and weights-degree correlations, and a rich core structure that changes significantly across the phase transition reflecting the highly nontrivial structure of the spin system that they describe.

This work opens new perspectives for the unsupervised characterization of the study of phases of matter using the tools of network science. This work can be extended in different directions. On one side, the analysis performed here for the $2D$ Ising model can be extended to the study of other classical critical phenomena with the goal of characterizing the possible presence or the lack of universalities among the networks constructed from spin system configuration snapshots. Similarly, the same toolbox can be utilized to attack out-of-equilibrium critical behavior.

Another natural extension is to consider path integrals of quantum systems [225]. While the data structures of such objects might feature anisotropies due to the different roles of space and imaginary time correlations, they shall be equally amenable to the analysis discussed here. Moreover, single-sliced

path integrals can also be represented as networks, as discussed in Ref. [193]: this last route provides a very promising venue for future investigation is the application of the proposed network science tools to study directly experimental data of many-body wave function snapshots.

Chapter 6

Outlook

In summary, our study presents theoretical frameworks for understanding the critical behavior in partition functions of classical systems using non-parametric unsupervised methods. We illustrate our approach by investigating phase transitions in classical Ising models within 2D and 3D lattices, utilizing thermal configurations obtained through Monte Carlo (MC) simulations. We compare the strengths and weaknesses of the two-NN method against PCA entropy. We find that I_d study for 3D proves to be notably more challenging than in 2D due to the *curse of dimensionality* [110]. To overcome this challenge, we have introduced the concept of PCA entropy, which is a novel application in the context of statistical mechanics. This entropy metric exhibits striking qualitative similarities with the thermodynamic entropy of the Ising model, particularly around the transition point in both 2D and 3D cases, enabling precise estimation of the critical temperature through conventional finite-size scaling analysis. Using PCA entropy and information imbalance, we are able to rank relevant correlations in the dynamics of out-of-equilibrium quantum systems in an assumption-free manner. We have successfully identified the most relevant operators describing the dynamics of Bose-Einstein condensates. From the intrinsic dimension characterization, we are able to find the time bound of universal dynamics.

Further, we systematically analyze networks generated from unsupervised learning of different states of matter, with a focus on IsingNets derived from 2D Ising model Monte Carlo simulations. Our analysis uncovers the statistical,

combinatorial, geometrical, and topological aspects of these networks. To characterize IsingNets, we employ two distinct approaches. First, we investigate their structure as we increase the filtration parameter r. This process enforces an effective percolation mechanism that aggregates nodes based on connected pairs at increasing distances, revealing the presence of two substantial components in the ferromagnetic state, each corresponding to configurations with different magnetization. We also study network topology by constructing clique complexes and calculating their persistent diagrams. Interestingly, these diagrams reveal that real IsingNets are composed of compact clusters, with significantly suppressed Betti numbers compared to randomized null models. Our network analysis, conducted across the filtration process, is further enhanced by effective network visualization techniques, including Minimum Spanning Tree (MST) and UMAP embedding. We also provide statistical characterizations of closeness centralities. In a second approach, we examine IsingNets by considering links at a distance less than a threshold value r. typically set as the average distance to the 5th nearest neighbor nodes based on the fully connected distance matrix. These networks exhibit diverse degree and strength distributions, reflecting their complexity, which extends beyond these distributions. Notably, IsingNets feature strong degree-degree correlations, weights-degree correlations, and a dynamic core structure that undergoes significant changes across phase transitions, highlighting the intricate structure of the spin system they represent.

In the thesis, we employ the Ising model as a test system, but our findings have broader applicability. The results have the potential to pave the path for new avenues of exploration in the fields of Statistical Mechanics and Quantum Field Theory. A crucial area for investigation involves the characterization of constraints on intrinsic dimension due to volume-related effects near the critical point. The natural extension would be to analyze various types of phase transitions, such as Berezinskii-Kosterlitz-Thouless (BKT) transitions especially in the scope of PCA entropy and network theory. For quantum systems, our methods can be readily used to learn path integrals of quantum statistical systems, extending previous works [39]. Data sets generated from

many-body experiments would be of great interest as shown in recent works [2, 193].

Another possible future research direction concerns out-of-equilibrium statistical mechanics. Since we have worked with systems in equilibrium it would be interesting to study how these tools perform in characterizing out-of-equilibrium dynamics for problems such as directed percolation [226, 227]. In this context, to apply these techniques for spin systems with disorder driven out of equilibrium, systems having a (de)-localization transitions would be an interesting direction for future research.

Bibliography

[1] PANDA, R. K. *et al.*, Non-parametric learning critical behavior in Ising partition functions: PCA entropy and intrinsic dimension, *arXiv e-prints*, arXiv:2308.13636, arXiv:2308.13636, Aug. 2023. DOI: 10.48550/arXiv.2308.13636. arXiv: 2308.13636 [cond-mat.stat-mech].

[2] VERDEL, R. *et al.*, Data-driven discovery of relevant information in quantum simulators, *arXiv e-prints*, arXiv:2307.10040, arXiv:2307.10040, Jul. 2023. DOI: 10.48550/arXiv.2307.10040. arXiv: 2307.10040.

[3] SUN, H. *et al.*, Network science Ising states of matter, *arXiv e-prints*, arXiv:2308.13604, arXiv:2308.13604, Aug. 2023. DOI: 10.48550/arXiv.2308.13604. arXiv: 2308.13604 [cond-mat.dis-nn].

[4] PANDA, R. K.; SCARDICCHIO, A.; SCHULZ, M.; TAYLOR, S. R.; Ž NIDARIČ, M., Can we study the many-body localisation transition? *EPL (Europhysics Letters)*, vol. 128, no. 6, p. 67 003, 2020. DOI: 10.1209/0295-5075/128/67003. [Online]. Available: https://doi.org/10.1209%2F0295-5075%2F128%2F67003.

[5] REYNOLDS, D. A. *et al.*, Gaussian mixture models. *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[6] MALLAPRAGADA, P. K.; JIN, R.; JAIN, A., Non-parametric mixture models for clustering, in *Structural, Syntactic, and Statistical Pattern Recognition*, HANCOCK, E. R.; WILSON, R. C.; WINDEATT, T.; ULUSOY, I.; ESCOLANO, F., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 334–343, ISBN: 978-3-642-14980-1.

[7] GYÖRFI, L., *Principles of nonparametric learning*. Springer, 2002, vol. 434.

[8] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J., *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer New York Inc., 2001.

[9] JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R., *et al.*, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[10] MACK, Y.; ROSENBLATT, M, Multivariate k-nearest neighbor density estimates, *Journal of Multivariate Analysis*, vol. 9, no. 1, pp. 1–15, 1979, ISSN: 0047-259X. DOI: https://doi.org/10.1016/0047-259X(79)90065-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0047259X79900654.

[11] NADARAYA, E. A., On non-parametric estimates of density functions and regression curves, *Theory of Probability & Its Applications*, vol. 10, no. 1, pp. 186–190, 1965. DOI: 10.1137/1110024. eprint: https://doi.org/10.1137/1110024. [Online]. Available: https://doi.org/10.1137/1110024.

[12] WATSON, G. S.; LEADBETTER, M. R., On the Estimation of the Probability Density, I, *The Annals of Mathematical Statistics*, vol. 34, no. 2, pp. 480–491, 1963. DOI: 10.1214/aoms/1177704159. [Online]. Available: https://doi.org/10.1214/aoms/1177704159.

[13] QUINLAN, J. R., Induction of decision trees, *Machine learning*, vol. 1, pp. 81–106, 1986.

[14] LOUPPE, G., *Understanding random forests: From theory to practice*, 2015. arXiv: 1407.7502 [stat.ML].

[15] BREIMAN, L., Random forests, *Machine learning*, vol. 45, pp. 5–32, 2001.

[16] XU, R.; WUNSCH, D., Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005. DOI: 10.1109/TNN.2005.845141.

[17] CAMPADELLI, P.; CASIRAGHI, E.; CERUTI, C.; ROZZA, A., Intrinsic dimension estimation: Relevant techniques and a benchmark framework, *Mathematical Problems in Engineering*, vol. 2015, pp. 1–21, 2015.

[18] BISHOP, C. M., *Neural networks for pattern recognition*. Oxford university press, 1995.

[19] FACCO, E.; D'ERRICO, M.; RODRIGUEZ, A.; LAIO, A., Estimating the intrinsic dimension of datasets by a minimal neighborhood information, *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[20] FACCO, E. *et al.*, The intrinsic dimension of biological data landscapes, 2017.

[21] GLIELMO, A. *et al.*, Unsupervised learning methods for molecular simulation data, *Chemical Reviews*, vol. 121, no. 16, pp. 9722–9758, Aug. 2021. DOI: 10.1021/acs.chemrev.0c01195. [Online]. Available: https://doi.org/10.1021/acs.chemrev.0c01195.

[22] FAN, M.; GU, N.; QIAO, H.; ZHANG, B., *Intrinsic dimension estimation of data by principal component analysis*, 2010. arXiv: 1002.2050 [cs.CV].

[23] GOLAY, J.; LEUENBERGER, M.; KANEVSKI, M., Feature selection for regression problems based on the morisita estimator of intrinsic dimension, *Pattern Recognition*, vol. 70, pp. 126–138, 2017, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.05.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317301905.

[24] STOLZ, B. J.; TANNER, J.; HARRINGTON, H. A.; NANDA, V., Geometric anomaly detection in data, *Proceedings of the National Academy of Sciences*, vol. 117, no. 33, pp. 19 664–19 669, 2020. DOI: 10.1073/pnas.2001741117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2001741117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2001741117.

[25] ANSUINI, A.; LAIO, A.; MACKE, J. H.; ZOCCOLAN, D., Intrinsic dimension of data representations in deep neural networks, in *Advances in Neural Information Processing Systems*, WALLACH, H. *et al.*, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/cfcce0621b49c983991ead4c3d4d3b6b-Paper.pdf.

[26] WOLD, S.; ESBENSEN, K.; GELADI, P., Principal component analysis, *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[27] KRUSKAL, J. B.; WISH, M., *Multidimensional scaling*. Sage, 1978.

[28] BALASUBRAMANIAN, M.; SCHWARTZ, E. L., The isomap algorithm and topological stability, *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[29] KRAMER, M. A., Nonlinear principal component analysis using autoassociative neural networks, *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[30] MOLTCHANOV, D., Distance distributions in random networks, *Ad Hoc Networks*, vol. 10, no. 6, pp. 1146–1166, 2012, ISSN: 1570-8705. DOI: https://doi.org/10.1016/j.adhoc.2012.02.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870512000224.

[31] PEDREGOSA, F. *et al.*, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] MENDES-SANTOS, T.; TURKESHI, X.; DALMONTE, M.; RODRIGUEZ, A., Unsupervised learning universal critical behavior via the intrinsic dimension, *Phys. Rev. X*, vol. 11, p. 011 040, 1 2021. DOI: 10.1103/PhysRevX.11.011040. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.11.011040.

[33] ALLEGRA, M.; FACCO, E.; DENTI, F.; LAIO, A.; MIRA, A., Data segmentation based on the local intrinsic dimension, *Scientific Reports*, vol. 10, no. 1, p. 16 449, 2020.

[34] MACOCCO, I.; GLIELMO, A.; GRILLI, J.; LAIO, A., Intrinsic dimension estimation for discrete metrics, *Physical Review Letters*, vol. 130, no. 6, p. 067 401, 2023.

[35] FACCO, E.; PAGNANI, A.; RUSSO, E. T.; LAIO, A., The intrinsic dimension of protein sequence evolution, *PLoS computational biology*, vol. 15, no. 4, e1006767, 2019.

[36] PINO, S. D. *et al.*, *Zundeig: The structure of the proton in liquid water from unsupervised learning*, 2023. arXiv: 2308.15319 [cond-mat.mtrl-sci].

[37] VITALE, V.; MENDES-SANTOS, T.; RODRIGUEZ, A.; DALMONTE, M., *Topological kolmogorov complexity and the Berezinskii-Kosterlitz-Thouless mechanism*, 2023. arXiv: 2305.05396 [cond-mat.stat-mech].

[38] OFFEI-DANSO, A.; HASSANALI, A.; RODRIGUEZ, A., *High dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning*, 2022. arXiv: 2112.11894 [cond-mat.soft].

[39] MENDES-SANTOS, T.; ANGELONE, A.; RODRIGUEZ, A.; FAZIO, R.; DALMONTE, M., Intrinsic dimension of path integrals: Data-mining quantum criticality and emergent simplicity, *PRX Quantum*, vol. 2, p. 030 332, 3 2021. DOI: 10.1103/PRXQuantum.2.030332. [Online]. Available: https://link.aps.org/doi/10.1103/PRXQuantum.2.030332.

[40] MENDES-SANTOS, T. *et al.*, *Wave function network description and kolmogorov complexity of quantum many-body systems*, 2023. arXiv: 2301.13216 [cond-mat.quant-gas].

[41] JOLLIFFE, I., Principal component analysis, in *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005, ISBN: 9780470013199. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa501.

[42] JOLLIFFE, I. T.; CADIMA, J., Principal component analysis: A review and recent developments, *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, 2016.

[43] CADIMA, J.; JOLLIFFE, I. T., On relationships between uncentred and column-centred principal component analysis, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:9451668.

[44] HU, W.; SINGH, R. R. P.; SCALETTAR, R. T., Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination, *Phys. Rev. E*, vol. 95, p. 062 122, 6 2017. DOI: 10.1103/PhysRevE.95.062122. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.95.062122.

[45] GOLUB, G. H.; REINSCH, C., Singular value decomposition and least squares solutions, in *Handbook for Automatic Computation: Volume II: Linear Algebra*, Springer, 1971, pp. 134–151.

[46] GOLUB, G. H.; VAN LOAN, C. F., *Matrix computations*. JHU press, 2013.

[47] ALTER, O.; BROWN, P. O.; BOTSTEIN, D., Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, 2000. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.97.18.10101.

[48] VARSHAVSKY, R.; GOTTLIEB, A.; LINIAL, M.; HORN, D., Novel Unsupervised Feature Filtering of Biological Data, *Bioinformatics*, vol. 22, no. 14, e507–e513, Jul. 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl214. eprint: https://academic.oup.com/bioinformatics/article-pdf/22/14/e507/614890/btl214.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btl214.

[49] STRYDOM, T.; DALLA RIVA, G. V.; POISOT, T., Svd entropy reveals the high complexity of ecological networks, *Frontiers in Ecology and Evolution*, vol. 9, 2021, ISSN: 2296-701X. DOI: 10.3389/fevo.2021.623141. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fevo.2021.623141.

[50] CARAIANI, P., The predictive power of singular value decomposition entropy for stock market dynamics, *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 571–578, 2014, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2013.08.071. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437113008212.

[51] GU, R.; XIONG, W.; LI, X., Does the singular value decomposition entropy have predictive power for stock market? — evidence from the shenzhen stock market, *Physica A: Statistical Mechanics and its Applications*, vol. 439, pp. 103–113, 2015, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2015.07.028. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037843711500638X.

[52] GU, R.; SHAO, Y., How long the singular value decomposed entropy predicts the stock market? — evidence from the dow jones industrial average index, *Physica A: Statistical Mechanics and its Applications*, vol. 453, pp. 150–161, 2016, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2016.02.030. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437116001965.

[53] ALVAREZ-RAMIREZ, J.; RODRIGUEZ, E., A singular value decomposition entropy approach for testing stock market efficiency, *Physica A: Statistical Mechanics and its Applications*, vol. 583, p. 126 337, 2021, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2021.126337. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437121006105.

[54] ESPINOSA-PAREDES, G.; RODRIGUEZ, E.; ALVAREZ-RAMIREZ, J., A singular value decomposition entropy approach to assess the impact of Covid-19 on the informational efficiency of the WTI crude oil market, *Chaos, Solitons & Fractals*, vol. 160, p. 112 238, 2022, ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2022.112238. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077922004489.

[55] WENG, X.; PERRY, A.; MAROUN, M.; VUONG, L. T., Singular Value Decomposition and Entropy Dimension of Fractals, *arXiv e-prints*, arXiv:2211.12338, arXiv:2211.12338, Nov. 2022. DOI: 10.48550/arXiv.2211.12338. arXiv: 2211.12338 [cond-mat.stat-mech].

[56] SHANNON, C. E., A mathematical theory of communication, *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[57] BRITTON, D; LLOYD, S. L., How to deal with petabytes of data: The lhc grid project, *Reports on Progress in Physics*, vol. 77, no. 6, p. 065 902, 2014. DOI: 10.1088/0034-4885/77/6/065902. [Online]. Available: https://dx.doi.org/10.1088/0034-4885/77/6/065902.

[58] SCAIFE, A., Big telescope, big data: Towards exascale with the square kilometre array, *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20 190 060, 2020.

[59] NEWMAN, M. E.; BARKEMA, G. T., *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.

[60] WALTER, J.-C.; BARKEMA, G., An introduction to monte carlo methods, *Physica A: Statistical Mechanics and its Applications*, vol. 418, pp. 78–87, 2015, Proceedings of the 13th International Summer School on Fundamental Problems in Statistical Physics, ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2014.06.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437114004798.

[61] SWENDSEN, R. H.; WANG, J.-S., Nonuniversal critical dynamics in monte carlo simulations, *Physical review letters*, vol. 58, no. 2, p. 86, 1987.

[62] PROKOF'EV, N.; SVISTUNOV, B.; TUPITSYN, I., "worm" algorithm in quantum monte carlo simulations, *Physics Letters A*, vol. 238, no. 4, pp. 253–257, 1998, ISSN: 0375-9601. DOI: https://doi.org/10.1016/S0375-9601(97)00957-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0375960197009572.

[63] PROKOF'EV, N.; SVISTUNOV, B., Worm algorithms for classical statistical models, *Phys. Rev. Lett.*, vol. 87, p. 160 601, 16 2001. DOI: 10.1103/PhysRevLett.87.160601. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.87.160601.

[64] WOLFF, U., Comparison between cluster monte carlo algorithms in the ising model, *Physics Letters B*, vol. 228, no. 3, pp. 379–382, 1989, ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(89)91563-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0370269389915633.

[65] WOLFF, U., Collective monte carlo updating for spin systems, *Phys. Rev. Lett.*, vol. 62, pp. 361–364, 4 1989. DOI: 10.1103/PhysRevLett.62.361. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.62.361.

[66] METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E., Equation of state calculations by fast computing machines, *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[67] PRÜFER, M. *et al.*, Experimental extraction of the quantum effective action for a non-equilibrium many-body system, *Nature Physics*, vol. 16, no. 10, pp. 1012–1016, 2020. DOI: 10.1038/s41567-020-0933-6. [Online]. Available: https://doi.org/10.1038%2Fs41567-020-0933-6.

[68] STAMPER-KURN, D. M.; UEDA, M., *Spinor bose gases: Explorations of symmetries, magnetism and quantum dynamics*, 2012. arXiv: 1205.1888 [cond-mat.quant-gas].

[69] KUNKEL, P. *et al.*, Simultaneous readout of noncommuting collective spin observables beyond the standard quantum limit, *Phys. Rev. Lett.*, vol. 123, p. 063 603, 6 2019. DOI: 10.1103/PhysRevLett.123.063603. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.123.063603.

[70] KAWAGUCHI, Y.; UEDA, M., Spinor bose–einstein condensates, *Physics Reports*, vol. 520, no. 5, pp. 253–381, 2012, Spinor Bose–Einstein condensates, ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.

2012.07.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157312002098.

[71] SADLER, L. E.; HIGBIE, J. M.; LESLIE, S. R.; VENGALATTORE, M.; STAMPER-KURN, D. M., Spontaneous symmetry breaking in a quenched ferromagnetic spinor bose–einstein condensate, *Nature*, vol. 443, no. 7109, pp. 312–315, 2006. DOI: 10.1038/nature05094. [Online]. Available: https://doi.org/10.1038/nature05094.

[72] PRÜFER, M. *et al.*, Observation of universal dynamics in a spinor bose gas far from equilibrium, *Nature*, vol. 563, no. 7730, pp. 217–220, 2018. DOI: 10.1038/s41586-018-0659-0. [Online]. Available: https://doi.org/10.1038/s41586-018-0659-0.

[73] FEYNMANN, R. P., *Statistical mechanics: a set of lectures*. Reading, Mass. : W. A. Benjamin, 1972.

[74] LANDAU, D. P.; BINDER, K., *A Guide to Monte Carlo Simulations in Statistical Physics*, 4th ed. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139696463.

[75] KRAUTH, W., *Statistical Mechanics Algorithms and Computations*. Oxford: Oxford University Press, 2006, ISBN: 9781429459501 1429459506.

[76] KRAMERS, H. A.; WANNIER, G. H., Statistics of the two-dimensional ferromagnet. Part II, *Phys. Rev.*, vol. 60, pp. 263–276, 3 1941. DOI: 10.1103/PhysRev.60.263. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.60.263.

[77] BAXTER, R. J., Exactly solved models in statistical mechanics, in *Integrable Systems in Statistical Mechanics*, pp. 5–63. DOI: 10.1142/9789814415255_0002. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9789814415255_0002. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789814415255_0002.

[78] NISHINO, T., Density matrix renormalization group method for 2D classical models, *Journal of the Physical Society of Japan*, vol. 64, no. 10, pp. 3598–3601, 1995. DOI: 10.1143/JPSJ.64.3598. eprint: https://doi.org/10.1143/JPSJ.64.3598. [Online]. Available: https://doi.org/10.1143/JPSJ.64.3598.

[79] CARRASQUILLA, J.; MELKO, R. G., Machine learning phases of matter, *Nature Physics*, vol. 13, no. 5, pp. 431–434, 2017. DOI: 10.1038/nphys4035. [Online]. Available: https://doi.org/10.1038/nphys4035.

[80] MEHTA, P. *et al.*, A high-bias, low-variance introduction to machine learning for physicists, *Physics Reports*, vol. 810, pp. 1–124, 2019, A high-bias, low-variance introduction to Machine Learning for physicists, ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2019.03.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157319300766.

[81] CARLEO, G. *et al.*, Machine learning and the physical sciences, *Rev. Mod. Phys.*, vol. 91, p. 045 002, 4 2019. DOI: 10.1103/RevModPhys.91.045002. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.91.045002.

[82] BEDOLLA, E.; PADIERNA, L. C.; CASTAÑEDA-PRIEGO, R., Machine learning for condensed matter physics, *Journal of Physics: Condensed Matter*, vol. 33, no. 5, p. 053 001, 2020. DOI: 10.1088/1361-648X/abb895. [Online]. Available: https://dx.doi.org/10.1088/1361-648X/abb895.

[83] WANG, L., Discovering phase transitions with unsupervised learning, *Phys. Rev. B*, vol. 94, p. 195 105, 19 2016. DOI: 10.1103/PhysRevB.94.195105. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.94.195105.

[84] WANG, C.; ZHAI, H., Machine learning of frustrated classical spin models. I. Principal component analysis, *Phys. Rev. B*, vol. 96, p. 144 432, 14 2017. DOI: 10.1103/PhysRevB.96.144432. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.96.144432.

[85] SALE, N.; GIANSIRACUSA, J.; LUCINI, B., Quantitative analysis of phase transitions in two-dimensional XY models using persistent homology, *Phys. Rev. E*, vol. 105, p. 024 121, 2 2022. DOI: 10.1103/PhysRevE.105.024121. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.105.024121.

[86] WETZEL, S. J., Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders, *Phys. Rev. E*, vol. 96, p. 022 140, 2 2017. DOI: 10.1103/PhysRevE.96.022140. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.96.022140.

[87] CH'NG, K.; VAZQUEZ, N.; KHATAMI, E., Unsupervised machine learning account of magnetic transitions in the Hubbard model, *Phys. Rev. E*, vol. 97, p. 013 306, 1 2018. DOI: 10.1103/PhysRevE.97.013306. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.97.013306.

[88] ZHANG, R.; WEI, B.; ZHANG, D.; ZHU, J.-J.; CHANG, K., Few-shot machine learning in the three-dimensional Ising model, *Phys. Rev. B*, vol. 99, p. 094 427, 9 2019. DOI: 10.1103/PhysRevB.99.094427. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.99.094427.

[89] NIR, A.; SELA, E.; BECK, R.; BAR-SINAI, Y., Machine-learning iterative calculation of entropy for physical systems, *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 234–30 240, 2020. DOI: 10.1073/pnas.2017042117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2017042117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2017042117.

[90] JANIK, R. A., Entropy from machine learning, *arXiv preprint arXiv:1909.10831*, 2019.

[91] AVINERY, R.; KORNREICH, M.; BECK, R., Universal and accessible entropy estimation using a compression algorithm, *Phys. Rev. Lett.*, vol. 123, p. 178 102, 17 2019. DOI: 10.1103/PhysRevLett.123.178102. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.123.178102.

[92] CAMASTRA, F.; STAIANO, A., Intrinsic dimension estimation: Advances and open problems, *Information Sciences*, vol. 328, pp. 26–41, 2016, ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2015.08.029. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025515006179.

[93] FACCO, E.; PAGNANI, A.; RUSSO, E. T.; LAIO, A., The intrinsic dimension of protein sequence evolution, *PLOS Computational Biology*, vol. 15, no. 4, pp. 1–16, Apr. 2019. DOI: 10.1371/journal.pcbi.1006767. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006767.

[94] ALLEGRA, M.; FACCO, E.; DENTI, F.; LAIO, A.; MIRA, A., Data segmentation based on the local intrinsic dimension, *Scientific Reports*, vol. 10, no. 1, p. 16 449, 2020. DOI: 10.1038/s41598-020-72222-0. [Online]. Available: https://doi.org/10.1038/s41598-020-72222-0.

[95] VALERIANI, L. *et al.*, The geometry of hidden representations of large transformer models, *arXiv preprint arXiv:2302.00294*, 2023.

[96] LEVINA, E.; BICKEL, P., Maximum likelihood estimation of intrinsic dimension, in *Advances in Neural Information Processing Systems*, SAUL, L.; WEISS, Y.; BOTTOU, L., Eds., vol. 17, MIT Press, 2004. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf.

[97] KRÜGER, N.; FELSBERG, M., A continuous formulation of intrinsic dimension, in *British Machine Vision Conference*, 2003.

[98] GONG, S.; BODDETI, V.; JAIN, A. K., On the intrinsic dimensionality of image representations, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 3982–3991. DOI: 10.1109/CVPR.2019.00411. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00411.

[99] POPE, P.; ZHU, C.; ABDELKADER, A.; GOLDBLUM, M.; GOLDSTEIN, T., The intrinsic dimension of images and its impact on learning, in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XJk19XzGq2J.

[100] SABATINI, A. M., Analysis of postural sway using entropy measures of signal complexity, *Medical and Biological Engineering and Computing*, vol. 38, no. 6, pp. 617–624, 2000. DOI: 10.1007/BF02344866. [Online]. Available: https://doi.org/10.1007/BF02344866.

[101] VARSHAVSKY, R.; GOTTLIEB, A.; HORN, D.; LINIAL, M., Unsupervised feature selection under perturbations: meeting the challenges of biological data, *Bioinformatics*, vol. 23, no. 24, pp. 3343–3349, Nov. 2007, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm528. eprint: https://academic.oup.com/bioinformatics/article-pdf/23/24/3343/16861997/btm528.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btm528.

[102] ONSAGER, L., Crystal statistics. i. a two-dimensional model with an order-disorder transition, *Physical Review*, vol. 65, no. 3-4, p. 117, 1944.

[103] FRANCESCO, P.; MATHIEU, P.; SÉNÉCHAL, D., *Conformal field theory*. Springer Science & Business Media, 2012.

[104] PREIS, T.; VIRNAU, P.; PAUL, W.; SCHNEIDER, J. J., GPU accelerated Monte Carlo simulation of the 2D and 3D Ising model, *Journal of Computational Physics*, vol. 228, no. 12, pp. 4468–4477, 2009.

[105] BILLO, M. *et al.*, Line defects in the 3d Ising model, *Journal of High Energy Physics*, vol. 2013, no. 7, pp. 1–23, 2013.

[106] EL-SHOWK, S. *et al.*, Solving the 3d ising model with the conformal bootstrap, *Phys. Rev. D*, vol. 86, p. 025 022, 2 2012. DOI: 10.1103/PhysRevD.86.025022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.86.025022.

[107] FISHER, M. E., The theory of equilibrium critical phenomena, *Reports on Progress in Physics*, vol. 30, no. 2, p. 615, 1967. DOI: 10.1088/0034-4885/30/2/306. [Online]. Available: https://dx.doi.org/10.1088/0034-4885/30/2/306.

[108] POLITIS, D. N.; ROMANO, J. P.; WOLF, M., *Subsampling*, 1st ed. Springer New York, 1999. DOI: https://doi.org/10.1007/978-1-4612-1554-7.

[109] SHAO, J.; TU, D., *The Jackknife and Bootstrap*, 1st ed. Springer New York, 1995. DOI: https://doi.org/10.1007/978-1-4612-0795-5.

[110] CHEN, L., Curse of dimensionality, in *Encyclopedia of Database Systems*, LIU, L.; ÖZSU, M. T., Eds. Boston, MA: Springer US, 2009, pp. 545–546, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_133. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_133.

[111] KAUFMAN, B., Crystal statistics. II. Partition function evaluated by spinor analysis, *Phys. Rev.*, vol. 76, pp. 1232–1243, 8 1949. DOI: 10.1103/PhysRev.76.1232. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.76.1232.

[112] DENBLEYKER, A. *et al.*, Controlling sign problems in spin models using tensor renormalization, *Phys. Rev. D*, vol. 89, p. 016 008, 1 2014. DOI: 10.1103/PhysRevD.89.016008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.89.016008.

[113] VIRTANEN, P. *et al.*, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[114] MENDES-SANTOS, T. *et al.*, Wave function network description and Kolmogorov complexity of quantum many-body systems, *arXiv e-prints*, arXiv:2301.13216, arXiv:2301.13216, Jan. 2023. DOI: 10.48550/arXiv.2301.13216. arXiv: 2301.13216 [cond-mat.quant-gas].

[115] SUN, H. *et al.*, Network science Ising states of matter, *arXiv e-prints*, arXiv:2308.13604, arXiv:2308.13604, Aug. 2023. DOI: 10.48550/arXiv.2308.13604. arXiv: 2308.13604 [cond-mat.dis-nn].

[116] SANDVIK, A. W., Computational Studies of Quantum Spin Systems, *AIP Conference Proceedings*, vol. 1297, no. 1, pp. 135–338, Nov. 2010, ISSN: 0094-243X. DOI: 10.1063/1.3518900. eprint: https://pubs.aip.org/aip/acp/article-pdf/1297/1/135/11407753/135_1_online.pdf. [Online]. Available: https://doi.org/10.1063/1.3518900.

[117] MACOCCO, I.; GLIELMO, A.; GRILLI, J.; LAIO, A., Intrinsic dimension estimation for discrete metrics, *Phys. Rev. Lett.*, vol. 130, p. 067 401, 6 2023. DOI: 10.1103/PhysRevLett.130.067401. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.130.067401.

[118] BLOCH, I.; DALIBARD, J.; ZWERGER, W., Many-body physics with ultracold gases, *Rev. Mod. Phys.*, vol. 80, pp. 885–964, 3 2008. DOI: 10.1103/RevModPhys.80.885. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.80.885.

[119] LEWENSTEIN, M.; SANPERA, A.; AHUFINGER, V., *Ultracold Atoms in Optical Lattices: Simulating quantum many-body systems*. Oxford University Press, Mar. 2012, ISBN: 9780199573127. DOI: 10.1093/acprof:oso/9780199573127.001.0001. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199573127.001.0001.

[120] GEORGESCU, I. M.; ASHHAB, S.; NORI, F., Quantum simulation, *Rev. Mod. Phys.*, vol. 86, pp. 153–185, 1 2014. DOI: 10.1103/RevModPhys.86.153. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.86.153.

[121] GROSS, C.; BLOCH, I., Quantum simulations with ultracold atoms in optical lattices, *Science*, vol. 357, no. 6355, pp. 995–1001, 2017. DOI: 10.1126/science.aal3837. eprint: https://www.science.org/doi/pdf/10.1126/science.aal3837. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aal3837.

[122] DALEY, A. J. *et al.*, Practical quantum advantage in quantum simulation, *Nature*, vol. 607, no. 7920, pp. 667–676, 2022. DOI: 10.1038/s41586-022-04940-6. [Online]. Available: https://doi.org/10.1038/s41586-022-04940-6.

[123] CAO, Y. *et al.*, Quantum chemistry in the age of quantum computing, *Chemical Reviews*, vol. 119, no. 19, pp. 10 856–10 915, Oct. 2019. DOI: 10.1021/acs.chemrev.8b00803. [Online]. Available: https://doi.org/10.1021/acs.chemrev.8b00803.

[124] TSVELIK, A. M., *Quantum Field Theory in Condensed Matter Physics*, 2nd ed. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511615832.

[125] KLEINERT, H., *Collective Classical and Quantum Fields*. WORLD SCIENTIFIC, 2018. DOI: 10.1142/10545. eprint: https://www.worldscientific.com/doi/pdf/10.1142/10545. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/10545.

[126] GLIELMO, A.; ZENI, C.; CHENG, B.; CSÁNYI, G.; LAIO, A., Ranking the information content of distance measures, *PNAS Nexus*, vol. 1, no. 2, Apr. 2022, ISSN: 2752-6542. [Online]. Available: https://doi.org/10.1093/pnasnexus/pgac039.

[127] DONKOR, E. D.; LAIO, A.; HASSANALI, A., Do machine-learning atomic descriptors and order parameters tell the same story? the case of liquid water, *Journal of Chemical Theory and Computation*, vol. 0, no. 0, null, 0, PMID: 36920997. DOI: 10.1021/acs.jctc.2c01205. eprint: https://doi.org/10.1021/acs.jctc.2c01205. [Online]. Available: https://doi.org/10.1021/acs.jctc.2c01205.

[128] FACCO, E.; D'ERRICO, M.; RODRIGUEZ, A.; LAIO, A., Estimating the intrinsic dimension of datasets by a minimal neighborhood information, *Scientific Reports*, vol. 7, no. 1, p. 12 140, 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-11873-y.

[129] CHAKRABORTI, A.; HRISHIDEV; SHARMA, K.; PHARASI, H. K., Phase separation and scaling in correlation structures of financial markets, *Journal of Physics: Complexity*, vol. 2, no. 1, p. 015 002, 2020. DOI: 10.1088/2632-072X/abbed1. [Online]. Available: https://dx.doi.org/10.1088/2632-072X/abbed1.

[130] WENG, X.; PERRY, A.; MAROUN, M.; VUONG, L. T., Singular value decomposition and entropy dimension of fractals, in *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 2022, pp. 427–431. DOI: 10.1109/ICICML57342.2022.10009680.

[131] PRÜFER, M. *et al.*, Experimental extraction of the quantum effective action for a non-equilibrium many-body system, *Nature Physics*, vol. 16, no. 10, pp. 1012–1016, 2020. DOI: 10.1038/s41567-020-0933-6. [Online]. Available: https://doi.org/10.1038/s41567-020-0933-6.

[132] BERGES, J.; BOGUSLAVSKI, K.; SCHLICHTING, S.; VENUGOPALAN, R., Universality far from equilibrium: From superfluid bose gases to heavy-ion collisions, *Phys. Rev. Lett.*, vol. 114, p. 061 601, 6 2015. DOI: 10.1103/PhysRevLett.114.061601. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.114.061601.

[133] ERNE, S.; BÜCKER, R.; GASENZER, T.; BERGES, J.; SCHMIEDMAYER, J., Universal dynamics in an isolated one-dimensional bose gas far from equilibrium, *Nature*, vol. 563, no. 7730, pp. 225–229, 2018. DOI: 10.1038/s41586-018-0667-0. [Online]. Available: https://doi.org/10.1038/s41586-018-0667-0.

[134] GLIDDEN, J. A. P. *et al.*, Bidirectional dynamic scaling in an isolated bose gas far from equilibrium, *Nature Physics*, vol. 17, no. 4, pp. 457–461, 2021. DOI: 10.1038/s41567-020-01114-x. [Online]. Available: https://doi.org/10.1038/s41567-020-01114-x.

[135] Supplemental material for details on the information imbalance method, the relevance ranking of further combinations of observables, the linear

fit used to estimate the intrinsic dimension with the TWO-NN method, and the subsampling technique used to estimate error bars.

[136] JOLLIFFE, I., Principal component analysis, in *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005, ISBN: 9780470013199. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa501.

[137] The normalized eigenvalue $\tilde{\lambda}_n$ is interpreted as the *proportion of total variance* that is accounted for by the n-th principal component.

[138] Here we only consider the simplest combinations of the relevant observables $n_{1,+1}$ and $n_{1,-1}$. However, we note that, in principle, we cannot disregard that other, more complicated, combinations may yield even lower values of S_{PCA}.

[139] CAMPADELLI, P.; CASIRAGHI, E.; CERUTI, C.; ROZZA, A., Intrinsic dimension estimation: Relevant techniques and a benchmark framework, *Mathematical Problems in Engineering*, vol. 2015, p. 759 567, 2015. DOI: 10.1155/2015/759567. [Online]. Available: https://doi.org/10.1155/2015/759567.

[140] LI, M.; VITÁNYI, P., *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. Springer New York, 2009. DOI: https://doi.org/10.1007/978-0-387-49820-1.

[141] STAIGER, L., Kolmogorov complexity and hausdorff dimension, *Information and Computation*, vol. 103, no. 2, pp. 159–194, 1993, ISSN: 0890-5401. DOI: https://doi.org/10.1006/inco.1993.1017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0890540183710175.

[142] MENDES-SANTOS, T.; TURKESHI, X.; DALMONTE, M.; RODRIGUEZ, A., Unsupervised learning universal critical behavior via the intrinsic dimension, *Phys. Rev. X*, vol. 11, p. 011 040, 1 2021. DOI: 10.1103/PhysRevX.11.011040. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.11.011040.

[143] MENDES-SANTOS, T.; ANGELONE, A.; RODRIGUEZ, A.; FAZIO, R.; DALMONTE, M., Intrinsic dimension of path integrals: Data-mining quantum criticality and emergent simplicity, *PRX Quantum*, vol. 2, p. 030 332, 3 2021. DOI: 10.1103/PRXQuantum.2.030332. [Online]. Available: https://link.aps.org/doi/10.1103/PRXQuantum.2.030332.

[144] TURKESHI, X., Measurement-induced criticality as a data-structure transition, *Phys. Rev. B*, vol. 106, p. 144 313, 14 2022. DOI: 10.1103/PhysRevB.106.144313. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.144313.

[145] MILES, C. *et al.*, Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data, *Nature Communications*, vol. 12, no. 1, p. 3905, 2021. DOI: 10.1038/s41467-021-23952-w. [Online]. Available: https://doi.org/10.1038/s41467-021-23952-w.

[146] BOHRDT, A. *et al.*, Analyzing nonequilibrium quantum states through snapshots with artificial neural networks, *Phys. Rev. Lett.*, vol. 127, p. 150 504, 15 2021. DOI: 10.1103/PhysRevLett.127.150504. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.127.150504.

[147] SCHMITT, M., From observations to complexity of quantum states via unsupervised learning, *Phys. Rev. B*, vol. 106, p. L041110, 4 2022. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[148] CASAGRANDE, H. P. *et al.*, Complexity of spin configurations dynamics due to unitary evolution and periodic projective measurements, *arXiv e-prints*, vol. 106, arXiv:2305.03334, arXiv:2305.03334, 4 May 2023. DOI: 10.48550/arXiv.2305.03334. arXiv: 2305.03334 [cond-mat.stat-mech]. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[149] BARABASI, A.-L., *Network Science*. Cambridge University Press, 2016, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online].

Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[150] NEWMAN, M., *Networks*. Oxford University Press, 2018, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[151] BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A., *Dynamical processes on complex networks*. Cambridge University Press, 2008, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[152] ESTRADA, E., *The structure of complex networks: theory and applications*. Oxford University Press, USA, 2012, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[153] DOROGOVTSEV, S. N.; MENDES, J. F., *The nature of complex networks*. Oxford University Press, 2022, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[154] BIANCONI, G., *Higher-order networks*. Cambridge University Press, 2021, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[155] RODRIGUEZ-NIEVA, J. F.; SCHEURER, M. S., Identifying topological order through unsupervised machine learning, *Nature Physics*, vol. 15, no. 8, pp. 790–795, 4 2019. DOI: 10.1038/s41567-019-0512-x. [Online]. Available: https://doi.org/10.1038/s41567-019-0512-x.

[156] DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. F., Ising model on networks with an arbitrary distribution of connections, *Physical Review E*, vol. 66, no. 1, p. 016 104, 4 2002. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[157] BIANCONI, G., Mean field solution of the Ising model on a barabási–albert network, *Physics Letters A*, vol. 303, no. 2-3, pp. 166–168, 4 2002. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[158] LEONE, M; VÁZQUEZ, A; VESPIGNANI, A; ZECCHINA, R., Ferromagnetic ordering in graphs with arbitrary degree distribution, *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 28, pp. 191–197, 4 2002. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[159] NGUYEN, H. C.; ZECCHINA, R.; BERG, J., Inverse statistical problems: From the inverse Ising problem to data science, *Advances in Physics*, vol. 66, no. 3, pp. 197–261, 4 2017. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[160] BIANCONI, G., Superconductor-insulator transition on annealed complex networks, *Physical Review E*, vol. 85, no. 6, p. 061 113, 4 2012. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[161] BIANCONI, G., Superconductor-insulator transition in a network of 2d percolation clusters, *Europhysics Letters*, vol. 101, no. 2, p. 26 003, 4 2013. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[162] CHEPURI, R.; KOVÁCS, I. A., Complex quantum network models from spin clusters, *arXiv preprint arXiv:2210.15838*, vol. 106, p. L041110, 4 2022. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[163] HALU, A.; FERRETTI, L.; VEZZANI, A.; BIANCONI, G., Phase diagram of the bose-hubbard model on complex networks, *Europhysics Letters*, vol. 99, no. 1, p. 18 001, 4 2012. DOI: 10.1103/PhysRevB.106.L041110.

[Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[164] HALU, A.; GARNERONE, S.; VEZZANI, A.; BIANCONI, G., Phase transition of light on complex quantum networks, *Physical Review E*, vol. 87, no. 2, p. 022 104, 4 2013. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[165] VICSEK, T.; CZIRÓK, A.; BEN-JACOB, E.; COHEN, I.; SHOCHET, O., Novel type of phase transition in a system of self-driven particles, *Physical Review Letters*, vol. 75, no. 6, p. 1226, 4 1995. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[166] BALLERINI, M. *et al.*, Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study, *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1232–1237, 4 2008. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[167] MORCOS, F. *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, E1293–E1301, 4 2011. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[168] MORA, T.; BIALEK, W., Are biological systems poised at criticality? *Journal of Statistical Physics*, vol. 144, pp. 268–302, 4 2011. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[169] NOKKALA, J.; GALVE, F.; ZAMBRINI, R.; MANISCALCO, S.; PIILO, J., Complex quantum networks as structured environments: Engineering and probing, *Scientific Reports*, vol. 6, no. 1, p. 26 861, 4 2016. DOI: 10.1103/

PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[170] NOKKALA, J. *et al.*, Reconfigurable optical implementation of quantum complex networks, *New Journal of Physics*, vol. 20, no. 5, p. 053 024, 4 2018. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[171] BONAMASSA, I. *et al.*, Interdependent superconducting networks, *Nature Physics*, vol. 106, pp. 1–8, 4 2023. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[172] TUMMINELLO, M.; ASTE, T.; DI MATTEO, T.; MANTEGNA, R. N., A tool for filtering information in complex systems, *Proceedings of the National Academy of Sciences*, vol. 102, no. 30, pp. 10 421–10 426, 4 2005. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[173] BONANNO, G.; CALDARELLI, G.; LILLO, F.; MANTEGNA, R. N., Topology of correlation-based minimal spanning trees in real and model markets, *Physical Review E*, vol. 68, no. 4, p. 046 130, 4 2003. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[174] VALDEZ, M. A.; JASCHKE, D.; VARGAS, D. L.; CARR, L. D., Quantifying complexity in quantum phase transitions via mutual information complex networks, *Physical Review Letters*, vol. 119, no. 22, p. 225 301, 4 2017. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[175] SUNDAR, B.; VALDEZ, M. A.; CARR, L. D.; HAZZARD, K. R. A., Complex-network description of thermal quantum states in the Ising spin chain, *Physical Review A*, vol. 97, p. 052 320, 5 2018. DOI: 10.1103/PhysRevA.97.052320. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.97.052320.

[176] SOKOLOV, B.; ROSSI, M. A. C.; GARCÍA-PÉREZ, G.; MANISCALCO, S., Emergent entanglement structures and self-similarity in quantum spin chains, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 380, no. 2227, p. 20 200 421, 4 2022. DOI: 10.1098/rsta.2020.0421. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[177] BAGROV, A. A. *et al.*, Detecting quantum critical points in the $t - t'$ fermi-hubbard model via complex network theory, *Scientific Reports*, vol. 10, no. 1, p. 20 470, 4 2020. DOI: 10.1038/s41598-020-77513-0. [Online]. Available: https://doi.org/10.1038/s41598-020-77513-0.

[178] PETRI, G.; SCOLAMIERO, M.; DONATO, I.; VACCARINO, F., Topological strata of weighted complex networks, *PloS one*, vol. 8, no. 6, e66506, 4 2013. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[179] EDELSBRUNNER, H.; HARER, J. L., *Computational topology: an introduction*. American Mathematical Society, 2022, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[180] OTTER, N.; PORTER, M. A.; TILLMANN, U.; GRINDROD, P.; HARRINGTON, H. A., A roadmap for the computation of persistent homology, *EPJ Data Science*, vol. 6, pp. 1–38, 4 2017. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[181] VACCARINO, F.; FUGACCI, U.; SCARAMUCCIA, S., Persistent homology: A topological tool for higher-interaction systems, in *Higher-Order Systems*, vol. 106, Springer, 2022, pp. 97–139. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[182] GHRIST, R., Barcodes: The persistent topology of data, *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 4 2008. DOI: 10.

1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[183] DONATO, I. *et al.*, Persistent homology analysis of phase transitions, *Physical Review E*, vol. 93, p. 052 138, 5 2016. DOI: 10.1103/PhysRevE.93.052138. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.93.052138.

[184] OLSTHOORN, B.; HELLSVIK, J.; BALATSKY, A. V., Finding hidden order in spin models with persistent homology, *Physical Review Res.*, vol. 2, p. 043 308, 4 2020. DOI: 10.1103/PhysRevResearch.2.043308. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.2.043308.

[185] COLE, A.; LOGES, G. J.; SHIU, G., Quantitative and interpretable order parameters for phase transitions from persistent homology, *Physical Review B*, vol. 104, p. 104 426, 10 2021. DOI: 10.1103/PhysRevB.104.104426. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.104.104426.

[186] TRAN, Q. H.; CHEN, M.; HASEGAWA, Y., Topological persistence machine of phase transitions, *Physical Review E*, vol. 103, p. 052 127, 5 2021. DOI: 10.1103/PhysRevE.103.052127. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.103.052127.

[187] TIRELLI, A.; COSTA, N. C., Learning quantum phase transitions through topological data analysis, *Physical Review B*, vol. 104, p. 235 146, 23 2021. DOI: 10.1103/PhysRevB.104.235146. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.104.235146.

[188] TIRELLI, ANDREA *et al.*, Unsupervised machine learning approaches to the q-state potts model, *Eur. Phys. J. B*, vol. 95, no. 11, p. 189, 4 2022. DOI: 10.1140/epjb/s10051-022-00453-3. [Online]. Available: https://doi.org/10.1140/epjb/s10051-022-00453-3.

[189] SPITZ, D.; BERGES, J.; OBERTHALER, M.; WIENHARD, A., Finding self-similar behavior in quantum many-body dynamics via persistent homology, *SciPost Phys.*, vol. 11, p. 060, 4 2021. DOI: 10.21468/SciPostPhys.11.3.060. [Online]. Available: https://scipost.org/10.21468/SciPostPhys.11.3.060.

[190] SEHAYEK, D.; MELKO, R. G., Persistent homology of $\mathbb{Z}_2$ gauge theories, *Physical Review B*, vol. 106, p. 085 111, 8 2022. DOI: 10.1103/PhysRevB.106.085111. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.085111.

[191] SALE, N.; LUCINI, B.; GIANSIRACUSA, J., Probing center vortices and deconfinement in su(2) lattice gauge theory with persistent homology, *Physical Review D*, vol. 107, p. 034 501, 3 2023. DOI: 10.1103/PhysRevD.107.034501. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.107.034501.

[192] SPITZ, D.; URBAN, J. M.; PAWLOWSKI, J. M., Confinement in non-abelian lattice gauge theory via persistent homology, *Physical Review D*, vol. 107, p. 034 506, 3 2023. DOI: 10.1103/PhysRevD.107.034506. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.107.034506.

[193] MENDES-SANTOS, T *et al.*, Wave function network description and kolmogorov complexity of quantum many-body systems, *arXiv preprint arXiv:2301.13216*, vol. 106, p. L041110, 4 2023. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[194] WOLFF, U., Collective Monte Carlo updating for spin systems, *Physical Review Letters*, vol. 62, no. 4, p. 361, 4 1989. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[195] WOLFF, U., Comparison between cluster Monte Carlo algorithms in the Ising model, *Physics Letters B*, vol. 228, no. 3, pp. 379–382, 4 1989. DOI:

10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[196] GRAHAM, R. L.; HELL, P., On the history of the minimum spanning tree problem, *Annals of the History of Computing*, vol. 7, no. 1, pp. 43–57, 4 1985. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[197] MCINNES, L.; HEALY, J.; MELVILLE, J., Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426*, vol. 106, p. L041110, 4 2018. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[198] BAVELAS, A., Communication patterns in task-oriented groups, *The journal of the acoustical society of America*, vol. 22, no. 6, pp. 725–730, 4 1950. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[199] DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F., Critical phenomena in complex networks, *Reviews of Modern Physics*, vol. 80, no. 4, p. 1275, 4 2008. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[200] LI, M. *et al.*, Percolation on complex networks: Theory and application, *Physics Reports*, vol. 907, pp. 1–68, 4 2021. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[201] COHEN, R.; EREZ, K.; BEN-AVRAHAM, D.; HAVLIN, S., Resilience of the internet to random breakdowns, *Physical Review Letters*, vol. 85, no. 21, p. 4626, 4 2000. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[202] BOBROWSKI, O.; SKRABA, P., A universal null-distribution for topological data analysis, *Scientific Reports*, vol. 13, no. 1, p. 12 274, 4 2023. DOI:

10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[203] BAPTISTA, A.; SÁNCHEZ-GARCÍA, R. J.; BAUDOT, A.; BIANCONI, G., Zoo guide to network embedding, *arXiv preprint arXiv:2305.03474*, vol. 106, p. L041110, 4 2023. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[204] SEYED-ALLAEI, H.; BIANCONI, G.; MARSILI, M., Scale-free networks with an exponent less than two, *Physical Review E*, vol. 73, no. 4, p. 046 113, 4 2006. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[205] COURTNEY, O. T.; BIANCONI, G., Dense power-law networks and simplicial complexes, *Physical Review E*, vol. 97, no. 5, p. 052 303, 4 2018. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[206] CARON, F.; FOX, E. B., Sparse graphs using exchangeable random measures, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 5, pp. 1295–1366, 4 2017. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[207] TIMAR, G.; DOROGOVTSEV, S. N.; MENDES, J. F. F., Scale-free networks with exponent one, *Physical Review E*, vol. 94, no. 2, p. 022 302, 4 2016. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[208] BARABÁSI, A.-L.; ALBERT, R., Emergence of scaling in random networks, *Science*, vol. 286, no. 5439, pp. 509–512, 4 1999. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[209] ALBERT, R.; BARABÁSI, A.-L., Statistical mechanics of complex networks, *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 4 2002. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[210] VOITALOV, I.; VAN DER HOORN, P.; VAN DER HOFSTAD, R.; KRIOUKOV, D., Scale-free networks well done, *Physical Review Research*, vol. 1, no. 3, p. 033 034, 4 2019. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[211] BARTHÉLEMY, M.; BARRAT, A.; PASTOR-SATORRAS, R.; VESPIGNANI, A., Characterization and modeling of weighted networks, *Physica a: Statistical mechanics and its applications*, vol. 346, no. 1-2, pp. 34–43, 4 2005. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[212] BARRAT, A.; BARTHELEMY, M.; PASTOR-SATORRAS, R.; VESPIGNANI, A., The architecture of complex weighted networks, *Proceedings of the national academy of sciences*, vol. 101, no. 11, pp. 3747–3752, 4 2004. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[213] WATTS, D. J.; STROGATZ, S. H., Collective dynamics of 'small-world'networks, *Nature*, vol. 393, no. 6684, pp. 440–442, 4 1998. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[214] BIANCONI, G.; MARSILI, M., Effect of degree correlations on the loop structure of scale-free networks, *Physical Review E*, vol. 73, no. 6, p. 066 127, 4 2006. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[215] ALVAREZ-HAMELIN, J; DALL'ASTA, L.; BARRAT, A.; VESPIGNANI, A., Large scale networks fingerprinting and visualization using the k-core decomposition, *Advances in neural information processing systems*, vol. 18,

p. L041110, 4 2005. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[216] CARMI, S.; HAVLIN, S.; KIRKPATRICK, S.; SHAVITT, Y.; SHIR, E., A model of internet topology using k-shell decomposition, *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11 150–11 154, 4 2007. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[217] DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. F., K-core organization of complex networks, *Physical Review Letters*, vol. 96, no. 4, p. 040 601, 4 2006. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[218] ALMAAS, E.; KOVACS, B; VICSEK, T; OLTVAI, Z.; BARABÁSI, A.-L., Global organization of metabolic fluxes in the bacterium escherichia coli, *Nature*, vol. 427, no. 6977, pp. 839–843, 4 2004. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[219] BURIONI, R.; CASSI, D., Random walks on graphs: Ideas, techniques and results, *Journal of Physics A: Mathematical and General*, vol. 38, no. 8, R45, 4 2005. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[220] CORREIA, J. D.; WHEATER, J. F., The spectral dimension of non-generic branched polymer ensembles, *Physics Letters B*, vol. 422, no. 1-4, pp. 76–81, 4 1998. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[221] AMBJØRN, J.; JURKIEWICZ, J.; LOLL, R., The spectral dimension of the universe is scale dependent, *Physical Review Letters*, vol. 95, no. 17, p. 171 301, 4 2005. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[222] ANAND, K.; BIANCONI, G., Entropy measures for networks: Toward an information theory of complex topologies, *Physical Review E*, vol. 80, no. 4, p. 045 102, 4 2009. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[223] ANAND, K.; BIANCONI, G.; SEVERINI, S., Shannon and von Neumann entropy of random networks with heterogeneous expected degree, *Physical Review E*, vol. 83, no. 3, p. 036 109, 4 2011. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[224] DE DOMENICO, M.; NICOSIA, V.; ARENAS, A.; LATORA, V., Structural reducibility of multilayer networks, *Nature Communications*, vol. 6, no. 1, p. 6864, 4 2015. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[225] MENDES-SANTOS, T.; ANGELONE, A.; RODRIGUEZ, A.; FAZIO, R.; DALMONTE, M., Intrinsic dimension of path integrals: Data-mining quantum criticality and emergent simplicity, *PRX Quantum*, vol. 2, no. 3, p. 030 332, 4 2021. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[226] HENKEL, M.; HINRICHSEN, H.; LÜBECK, S., *Non-equilibrium phase transitions: volume 1: absorbing phase transitions*. Springer Science & Business Media, 2008, vol. 106, p. L041110. DOI: 10.1103/PhysRevB.106.L041110. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.106.L041110.

[227] HINRICHSEN, H., Non-equilibrium critical phenomena and phase transitions into absorbing states, *Advances in Physics*, vol. 49, no. 7, pp. 815–958, 4 2000. DOI: 10.1080/00018730050198152. [Online]. Available: https://doi.org/10.1080%2F00018730050198152.