Clustering neural quantum states via diffusion maps

Yanting Teng, Subir Sachdev, and Mathias S. Scheurer

Phys. Rev. B **108**, 205152 — Published 27 November 2023

DOI: 10.1103/PhysRevB.108.205152

Clustering neural quantum states via diffusion maps

Yanting Teng,[1] Subir Sachdev,[1] and Mathias S. Scheurer[2,3]

[1] *Department of Physics, Harvard University, Cambridge MA 02138, USA*
[2] *Institut für Theoretische Physik, Universität Innsbruck, A-6020 Innsbruck, Austria*
[3] *Institute for Theoretical Physics III, University of Stuttgart, 70550 Stuttgart, Germany*

We discuss and demonstrate an unsupervised machine-learning procedure to detect topological order in quantum many-body systems. Using a restricted Boltzmann machine to define a variational ansatz for the low-energy spectrum, we sample wave functions with probability decaying exponentially with their variational energy; this defines our training dataset that we use as input to a diffusion map scheme. The diffusion map provides a low-dimensional embedding of the wave functions, revealing the presence or absence of superselection sectors and, thus, topological order. We show that for the diffusion map, the required similarity measure of quantum states can be defined in terms of the network parameters, allowing for an efficient evaluation within polynomial time. However, possible "gauge redundancies" have to be carefully taken into account. As an explicit example, we apply the method to the toric code.

I. INTRODUCTION

In the last few years, machine learning (ML) techniques have been very actively studied as novel tools in many-body physics [1–7]. A variety of valuable applications of ML has been established, such as ML-based variational ansätze for many-body wave functions, application of ML to experimental data to extract information about the underlying physics, ML methods for more efficient Monte-Carlo sampling, and employment of ML to detect phase transitions, to name a few. Regarding the latter type of applications, a particular focus has recently been on topological phase transitions [8–31]. This is motivated by the challenges associated with capturing topological phase transitions: by definition, topological features are related to the global connectivity of the dataset rather than local similarity of samples. Therefore, unless the dataset is sufficiently simple such that topologically connected pairs of samples also happen to be locally similar or features are used as input data that are closely related to the underlying topological invariant, the topological structure is hard to capture reliably with many standard ML techniques [11, 12].

In this regard, the ML approach proposed in Ref. 12, which is based on diffusion maps (DM) [32–35], is a particularly promising route to learn topological phase transitions; it allows to embed high-dimensional data in a low-dimensional subspace such that pairs of samples that are smoothly connected in the dataset will be mapped close to each other, while disconnected pairs will be mapped to distant points. As such, the method captures the central notion of topology. In combination with the fact that it is unsupervised and thus does not require *a priori* knowledge of the underlying topological invariants, it is ideally suited for the task of topological phase classification. As a result, there have been many recent efforts applying this approach to a variety of problems, such as different symmetry-protected, including non-Hermitian, topological systems [36–41], experimental data [39, 42], many-body localized states [43], and dynamics [44]; extensions based on combining DM with path finding [36] as well as with quantum computing schemes [45] for speed-up have also been studied.

As alluded to above, another very actively pursued application of ML in physics are neural network quantum states: as proposed in Ref. 46, neural networks can be used to efficiently parameterize and, in many cases, optimize variational descriptions of wave functions of quantum many-body systems [47–56]. In particular, restricted Boltzmann machines (RBMs) [4, 57] represent a very popular neural-network structure in this context. For instance, the ground states of the toric code model [58] can be exactly expressed with a *local* RBM ansatz [59], i.e., where only neighboring spins are connected to the same hidden neurons. When additional non-local extensions to the RBM ansatz of Ref. 59 are added, this has been shown to also provide a very accurate variational description of the toric code in the presence of a magnetic field [60].

In this work, we combine the DM approach of Ref. 12 with neural network quantum states with the goal of capturing topological order in an unsupervised way in interacting quantum many-body systems. We use a local network ansatz, with parameters Λ, as a variational description for the wave functions $|\Psi(\Lambda)\rangle$ of the low-energy subspace of a system with Hamiltonian $\hat{\mathcal{H}}$. While we also briefly mention other possible ways of generating ensembles of states, we primarily focus on an energetic principle: we sample wavefunctions such that the probability of $|\Psi(\Lambda)\rangle$ is proportional to $\exp(-\langle\hat{\mathcal{H}}\rangle_\Lambda/T)$ where $\langle\hat{\mathcal{H}}\rangle_\Lambda = \langle\Psi(\Lambda)|\,\hat{\mathcal{H}}\,|\Psi(\Lambda)\rangle$. As illustrated in Fig. 1(a), the presence of superselection sectors in the low-energy spectrum of $\hat{\mathcal{H}}$ implies that the ensemble of states decays into disconnected subsets of states for sufficiently small T (at least at fixed finite system size); these can be extracted, without need of prior labels, with dimensional reduction via DM (and subsequent k-means clustering), and thus allow to identify topological order. For sufficiently large T, more and more high-energy states are included and all sectors are connected, see Fig. 1(b), as can also be

readily revealed via DM-based embedding of the states.

Importantly, DM is a kernel technique in the sense that the input data x_l (in our case the states $|\Psi(\Lambda_l)\rangle$) does not directly enter as a high-dimensional vector but only via a similarity measure $S(x_l, x_{l'})$, comparing how "similar" two samples l and l' are. In the context of applying DM to the problem of topological classification, it defines what a smooth deformation ("homotopy") of samples is. We discuss two possible such measures. The first one is just the quantum mechanical overlap, $S_{\rm q}(\Lambda_l, \Lambda_{l'}) = |\langle\Psi(\Lambda_l)|\Psi(\Lambda_{l'})\rangle|^2$, of the wave functions. Although conceptually straightforward, its evaluation is computationally costly on a classical computer as it requires importance sampling. The local nature of our network ansatz allows us to also construct an alternative similarity measure that is expressed as a simple function of the network parameters Λ_l and $\Lambda_{l'}$ describing the two states to be compared. This can, however, lead to subtleties associated with the fact that two states with different Λ can correspond to the same wave functions (modulo global phase). We discuss how these "gauge redundancies" can be efficiently circumvented for generic states.

We illustrate these aspects and explicitly demonstrate the success of this approach using the toric code [58], a prototype model for topological order which has also been previously studied with other ML techniques with different focus [15–18, 59–61]. We show that the DM algorithm learns the underlying loop operators wrapping around the torus without prior knowledge; at low T, this leads to four clusters corresponding to the four ground states. At larger T, these clusters start to merge, as expected. Interestingly, the DM still uncovers the underlying structure of the dataset related to the expectation value of the loop operators. Finally, we also show that applying a magnetic field leads to the disappearance of clusters in the DM, capturing the transition from topological order to the confined phase.

The remainder of the paper is organized as follows. In Sec. II, we describe our ML approach in general terms, including the local network quantum state description we use, the ensemble generation, a brief review of the DM scheme of Ref. 12, and the similarity measure in terms of neural network parameters. Using the toric code model as an example, all of these general aspects are then discussed in detail and illustrated in Sec. III. Finally, explicit numerical results can be found in Sec. IV and a conclusion is provided in Sec. V.

II. GENERAL ALGORITHM

Here, we first present and discuss our algorithm [see Fig. 2(a)] in general terms before illustrating it using the toric code as an example in the subsequent sections. Consider a system of N qubits or spins, with associated operators $\{\hat{\boldsymbol{s}}\} = \{\hat{\boldsymbol{s}}_i, i = 1, \cdots, N\}$, $\hat{\boldsymbol{s}}_i = (\hat{s}_i^x, \hat{s}_i^y, \hat{s}_i^z)$, and interactions governed by a local, gapped Hamiltonian $\hat{\mathcal{H}} = \mathcal{H}(\{\hat{\boldsymbol{s}}\})$. We represent the states $|\Psi(\Lambda)\rangle$ of this system using neural network quantum states [46],

$$|\Psi(\Lambda)\rangle = \sum_{\boldsymbol{\sigma}} \psi(\boldsymbol{\sigma};\, \Lambda)\, |\boldsymbol{\sigma}\rangle\,, \tag{1}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, ..., \sigma_N | \sigma_i = \pm 1\}$ enumerates configurations of the physical spin variables in a local computational basis (e.g. s^z-basis) and Λ is the set of parameters that the network ψ depends on to output the wavefunction amplitude $\psi(\boldsymbol{\sigma};\, \Lambda) = \langle\boldsymbol{\sigma}|\Psi(\Lambda)\rangle$ for configuration $|\boldsymbol{\sigma}\rangle$. Because the physical Hilbert space scales exponentially with the system size, there is a trade-off between the expressivity versus efficiency when choosing a network architecture (or ansatz) ψ, so that the weights Λ can approximate the state $|\Psi(\Lambda)\rangle$ to a reasonable degree and can at the same time be an efficient representation (with minimal number of parameters Λ that scale as a polynomial in N). To reach the ground state or, more generally, the relevant low-energy sector of the Hamiltonian $\hat{\mathcal{H}}$ for the low-temperature physics, we minimize the energy in the variational subspace defined by Eq. (1) using gradient descent with a learning rate λ,

$$\Lambda \to \Lambda - \lambda\, \partial_\Lambda \langle\hat{\mathcal{H}}\rangle_\Lambda\,, \quad \langle\hat{\mathcal{H}}\rangle_\Lambda = \langle\Psi(\Lambda)|\, \hat{\mathcal{H}}\, |\Psi(\Lambda)\rangle\,. \tag{2}$$

Here, the quantum mechanical expectation value $\langle\hat{\mathcal{H}}\rangle_\Lambda$ is evaluated using importance sampling (see Appendix B).

While there are exponentially many states in the Hilbert space, the low-energy sector of a local Hamiltonian is expected to occupy a small subspace where states obey area law entanglement [62, 63] whereas a typical state obeys volume law [64, 65]. Motivated by these considerations, we consider a class of networks that naturally describe quantum states that obey area-law entanglement. Pictorially, in such networks, the connections from the hidden neurons (representing the weights Λ) to the physical spins are *quasi-local* [51, 53–55]. In that case, it holds

$$\psi(\boldsymbol{\sigma}, \Lambda) = \phi_1(\boldsymbol{\sigma}_1, \Lambda_1) \times \phi_2(\boldsymbol{\sigma}_2, \Lambda_2) \times \cdots\,, \tag{3}$$

where $\boldsymbol{\sigma}_\jmath = \{\sigma_k\}_{k\in\jmath}$ denote (overlapping) subsets of neighboring spins with $\cup_\jmath \boldsymbol{\sigma}_\jmath = \boldsymbol{\sigma}$ and $\Lambda_\jmath$ are the subsets of the network parameters (weights and biases) that are connected to the physical spins in $\jmath$.

Algorithm 1 Ensemble generation

procedure $(\{\Lambda\}_{n=1}^N)$
  init: optimized parameters Λ
  for k independent times **do**:
    for n sampling steps **do**:
      Propose new parameter $\Lambda_p = f(\Lambda_t)$
      Accept with probability determined by energy $\langle\hat{\mathcal{H}}\rangle_\Lambda$ and ensemble parameter T:
      $\Lambda_{t+1} = \mathbb{P}_{\rm accept}(\Lambda'|\Lambda; T)$
  return the last m states for each k: $\{\Lambda_i | i = n - m, ..., n\}_k$

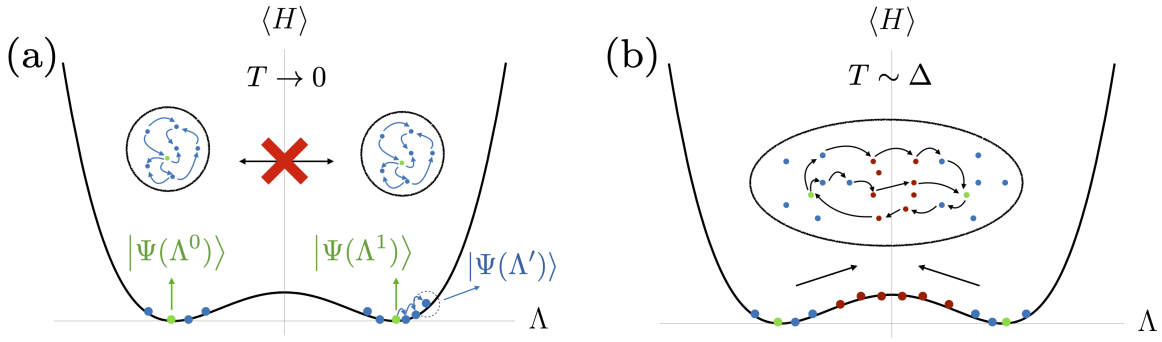

Figure 1. (a) An illustration of a "low-energy" ensemble. Two (or more) initial states, $|\Psi(\Lambda^0)\rangle$ and $|\Psi(\Lambda^1)\rangle$, from two distinct topological sectors are chosen as "seeds" (green dots). The dots denote the dataset (later fed into the DM), which are a set of quantum states labeled by network parameters Λ. This dataset is generated using the procedure outlined in Sec. II A and Algorithm. 1, where the next state Λ' (blue dots at each arrow) is proposed by a random local perturbation and accepted with probability based on the energy expectation $\langle H\rangle_{\Lambda'}$. In the small-$T$ regime, the full dataset is not inter-connected by such local perturbations and cluster among each topological sectors (at left and right valley). (b) An illustration of a "high-energy" ensemble. The states are generated using the same algorithm as before, however with large T (compared to the energy gap Δ). In this regime, the dataset include some of the low-energy states (blue dots), but also some high-energy states (red dots). Because the high-energy states are agnostic of the low-energy topological sectors, there exist paths (denoted by arrows among dots in the elliptical blob) such that the two initial seeds from distinct topological sectors effectively "diffuse" and form one connected cluster.

A. Dataset: network parameter ensembles

The dataset we use for unsupervised detection of topological order consists of an ensemble of wavefunctions $\{|\Psi(\Lambda)\rangle\}_l$, parameterized by the set of network parameters $\{\Lambda\}_l$. While, depending on the precise application, other choices are conceivable, we generate this ensemble such that the relative occurrence of a state $|\Psi(\Lambda)\rangle$ is given by $\rho_T(\Lambda) = \exp(-\langle\hat{\mathcal{H}}\rangle_\Lambda/T)/Z$, with appropriate normalization factor Z. As such, a small value of the "temperature-like" ensemble parameter T corresponds to a "low-energy" ensemble while large T parametrize "high-energy" ensembles.

In practice, to generate this ensemble, we here first optimize the parameters Λ via Eq. (2) to obtain wavefunctions with lowest energy expectation values. As Eq. (1) does not contain all possible states, this will, in general, only yield approximations to the exact low-energy eigenstates of $\hat{\mathcal{H}}$. However, as long as it is able to capture all superselection sectors of the system as well as (a subset of) higher energy states connecting these sectors, Eq. (1) will be sufficient for our purpose of detecting topological order or the absence thereof. We perform this optimization several times, $\Lambda \to \Lambda_l^0$, with different initial conditions, to obtain several "seeds", Λ_l^0; this is done to make sure we have a low-energy representative of all superselection sectors. Ideally the dataset is sampled directly from the the target probability distribution ρ_T, if for instance, one has access to an experimental system at finite temperature. Here, we adopt a Markov-chain-inspired procedure for generating the ensemble based on ρ_T for each of these seeds. Specifically, starting from a state Λ, we propose updates on a randomly chosen local block of parameters connected to the spins at sites $\jmath$,

$$\Lambda \to \Lambda' = \{\Lambda_1, \Lambda_2, \cdots, u(\Lambda_\jmath), \cdots, \Lambda_N\}, \tag{4}$$

where the update u only depends on $\Lambda_\jmath$. The proposed parameter Λ' given the current parameter Λ is accepted with probability

$$\mathbb{P}_{\text{accept}}(\Lambda'|\Lambda;T) = \min\Big(1,\, e^{-\frac{\langle\hat{\mathcal{H}}\rangle_{\Lambda'} - \langle\hat{\mathcal{H}}\rangle_\Lambda}{T}}\Big). \tag{5}$$

This means that if the proposed state $\Psi(\Lambda')$ has a lower energy expectation value than $\Psi(\Lambda)$, then the proposal will be accepted; otherwise, it will be accepted with a probability determined by the Boltzmann factor. The entire ensemble generation procedure is summarized in Algorithm 1.

B. Diffusion map

As proposed in Ref. 12, DM is ideally suited as an unsupervised ML algorithm to identify the presence and number of superselection sectors in a collection of states, such as $\{|\Psi(\Lambda)\rangle\}_l$ defined above. To briefly review the key idea of the DM algorithm [32–35] and introduce notation, assume we are given a dataset $X = \{x_l | l = 1, 2, ..., M\}$, consisting of M samples x_l. Below we will consider the cases $x_l = \Lambda_l$ and $x_l = |\Psi(\Lambda_l)\rangle$; in the first case, the samples are the network parameters parametrizing the wavefunction and, in the second, the samples are the wavefunctions themselves.

To understand DM intuitively, let us define a diffusion process among states $x_l \in X$. The probability of state x_l transitioning to $x_{l'}$ is defined by the Markov transition matrix element $p_{l,l'}$. To construct $p_{l,l'}$, we introduce a

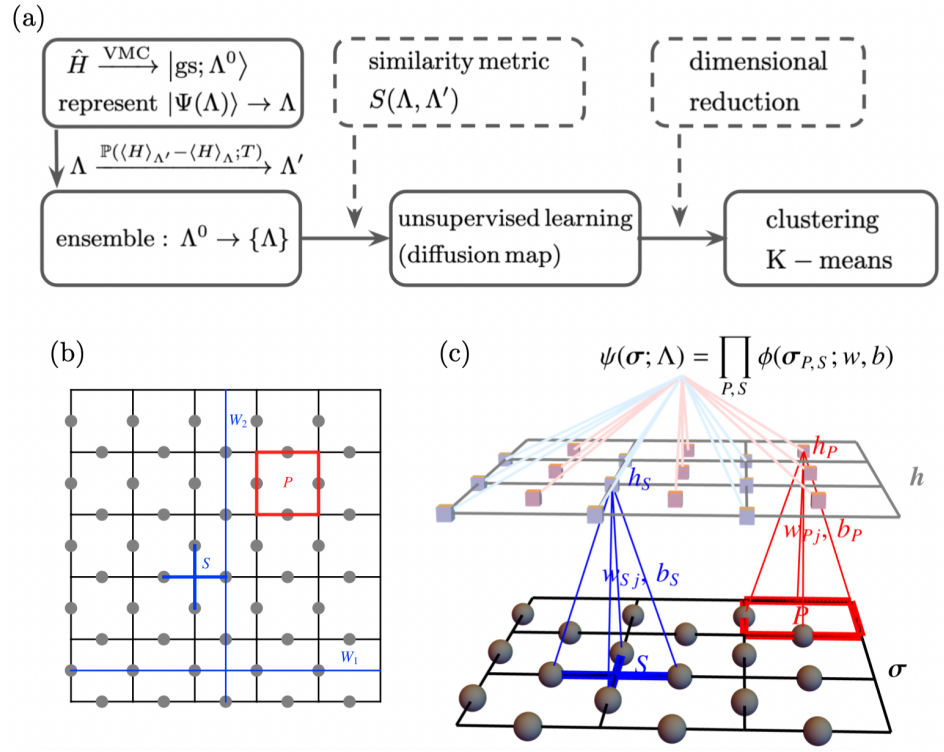

Figure 2. (a) Overview of the ML algorithm applied in this work: the "seeds" $\{\Lambda^0\}$ are computed using variational Monte Carlo (see Appendix B), a Markov-chain algorithm is used to generate the network parameter ensemble dataset (Sec. II A), then a similarity metric is used for the definition of kernels in the DM method (Sec. II B and Sec. II C), and finally k-means is applied to the low-dimensional embedding in the subspace provided by the dominant DM eigenvector components. (b) The square lattice geometry for the toric code model, where the qubits $\hat{s}_i$ are defined on the links of the lattice (grey dots). The Hamiltonian [given in Eq. (16)] is written in terms of the operators $\hat{\mathcal{P}}_P$ (supported by spins on plaquette P denoted by the red square) and star $\hat{\mathcal{S}}_S$ (supported by spins on star S denoted by the blue links). The two blue lines along $x(y)$ directions denote the Wilson loop operators $\hat{W}_{1,\bar{x}}(\hat{W}_{2,\bar{y}})$ along the straight paths $\bar{x}(\bar{y})$. (c) An illustration of the quasi-local ansatz in Eq. (17). The ansatz is a product over local function ϕ of spins in plaquette (or star), which depends on parameters $\{w_{Xj}, b_X\}$ for $X = P(S)$ being plaquette (or star).

symmetric and positive-definite kernel $k_\epsilon(x_l, x_{l'})$ between states x_l and $x_{l'}$. Then the transition probability matrix $p_{l,l'}$ is defined as

$$p_{l,l'} = \frac{k_\epsilon(x_l, x_{l'})}{z_l}, \quad z_l = \sum_{l'} k_\epsilon(x_l, x_{l'}), \tag{6}$$

where the factor z_l ensures probability conservation, $\sum_{l'} p_{l,l'} = 1\,\forall l$. Then spectral analysis on the transition probability matrix leads to information on the *global* connectivity of the dataset X, which, in our context of X containing low-energy states, allows to identify superselection sectors and, thus, topological order [12]. To quantify how strongly two samples x_l and $x_{l'}$ are connected, one introduces the $2t$-step diffusion distance [32–35],

$$D_{2t}(l, l') = \sum_{l''} \frac{1}{z_{l''}} [(p^t)_{l,l''} - (p^t)_{l',l''}]^2, \tag{7}$$

where p^t denotes the t-th matrix power of the transition probability matrix p. It was shown that D_{2t} can be computed from the eigenvalues λ_n and right eigenvectors ψ_n of the transition matrix p: with $\sum_{l'} p_{l,l'} (\psi_n)_{l'} = \lambda_n (\psi_n)_l$, and in descending ordering $\lambda_n > \lambda_{n+1}$, it follows

$$D_{2t}(l, l') = \sum_{n=1}^{M-1} \lambda_n^{2t} [(\psi_n)_l - (\psi_n)_{l'}]^2 \tag{8}$$

after straightforward algebra [35]. Geometrically, this means that the diffusion distance is represented as a Euclidean distance (weighted with λ_n) if we perform the non-linear coordinate transformation $x_l \to \{(\psi_n)_l, n = 0, \ldots M-1\}$. Furthermore, as the global connectivity is seen from the long-time limit, $t \to \infty$, of the diffusion distance, the largest eigenvalues are most important to describe the connectivity. To be more precise, let us choose a kernel k_ϵ of the form

$$k_\epsilon(x_l, x_{l'}) = \exp\left(-\frac{1 - S(x_l, x_{l'})}{\epsilon}\right), \tag{9}$$

where S is a *local similarity measure* which obeys $S \in [0, 1]$, $S(x_l, x_{l'}) = S(x_{l'}, x_l)$, and $S(x, x) = 1$. Here "local" means that $S(x_l, x_{l'}) = \sum_i \mathcal{S}_i(x_l, x_{l'})$ where $\mathcal{S}_i(x_l, x_{l'})$ only depend on the configuration of x_l and

$x_{l'}$ in the vicinity of site i. While we will discuss possible explicit forms of S for our quantum mechanical N spin/qubit system in Sec. II C below, a natural choice for a classical system of N spins, $x_l = \{\boldsymbol{S}_i^l, (\boldsymbol{S}_i^l)^2 = 1, i = 1, 2, \ldots, N\}$, is $S_{\mathrm{cl}}(x_l, x_{l'}) = \sum_i \boldsymbol{S}_i^l \cdot \boldsymbol{S}_i^{l'}/N$. In Eq. (9), ϵ plays the role of a "coarse graining" parameter that is necessary as we only deal with finite datasets X: for given X, we generically expect $k_\epsilon(x_l, x_{l'}) = p_{l,l'} = \delta_{l,l'}$ as $\epsilon \to 0$, i.e., all distinct samples are dissimilar if ϵ is sufficiently small and all eigenvalues λ_n approach 1 [66]. In turn, for $\epsilon \to \infty$ the coarse graining parameter is so large that all samples become connected, $k_\epsilon(x_l, x_{l'}) \to 1$; due to $\sum_{l'=1}^{M} p_{l,l'} = 1$, we have $p_{l,l'} \to 1/M$, which can be written as $p \to \hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^T$ with M-component unit vector $\hat{\boldsymbol{e}} = (1, 1, \ldots, 1)^T/\sqrt{M}$. Consequently, we will have $\lambda_{n>0} \to 0$ (with eigenvectors ψ_n perpendicular to $\hat{\boldsymbol{e}}$), while the largest eigenvalue λ_0 (with eigenvector $\hat{\boldsymbol{e}}$) is 1 as before (as a consequence of probability conservation). For values of ϵ in between these extreme limits, the DM spectrum contains information about X, including its topological structure: as shown in Ref. 12, the presence of $k \in \mathbb{N}$ distinct topological equivalence classes in X is manifested by a range of ϵ where $\lambda_1, \ldots \lambda_{k-1}$ are all exponentially close (in ϵ) to 1, with a clear gap to $\lambda_{n\geq k}$. Furthermore, the different samples l will cluster—with respect to the normal Euclidean measure, e.g., as can be captured with k-means—according to their topological equivalence class when plotted in the mapped $k-1$-dimensional space $\{(\psi_1)_l, (\psi_2)_l, \ldots, (\psi_{k-1})_l\}$. In the following, we will use this procedure to identify the superselection sectors in the ensemble of wave functions defined in Sec. II A. To this end, however, we first need to introduce a suitable similarity measure S, to be discussed next.

C. Local similarity measure

A natural generalization of the abovementioned classical similarity measure $S_{\mathrm{cl}} = \sum_i \boldsymbol{S}_i^l \cdot \boldsymbol{S}_i^{l'}/N$, which can be thought of as the (Euclidean) inner product in the classical configuration space, is to take the inner product in the Hilbert space of the quantum system,

$$S_{\mathrm{q}}(\Lambda_l, \Lambda_{l'}) = |\langle \Psi(\Lambda_l) | \Psi(\Lambda_{l'}) \rangle|^2. \tag{10}$$

While this or other related fidelity measures for low-rank quantum states could be estimated efficiently with quantum simulation and computing setups [67–70], estimating S_{q} is generally a computationally expensive task on a classical computer, as it requires sampling over spin configurations for our variation procedure. To make the evaluation of the similarity measure more efficient, we here propose an alternative route that takes advantage of the fact that we use a local ansatz for $\psi(\boldsymbol{\sigma}; \Lambda)$, see Eq. (3). Our goal is to express the similarity measure directly as

$$S_{\mathrm{n}}(\Lambda_l, \Lambda_{l'}) = \frac{1}{N_{\mathrm{J}}} \sum_{\mathrm{J}} f((\Lambda_l)_{\mathrm{J}}, (\Lambda_{l'})_{\mathrm{J}}), \tag{11}$$

where f only compares a local block of parameters denoted by J and is a function that can be quickly evaluated, without having to sample spin configurations. Furthermore, $S(x_l, x_{l'}) = S(x_{l'}, x_l)$ can be ensured by choosing a function f that is symmetric in its arguments and $S \in [0, 1]$ is also readily implemented by setting $N_{\mathrm{J}} = \sum_{\mathrm{J}}$ and appropriate rescaling of f such that $f \in [0, 1]$. The most subtle condition is

$$S_{\mathrm{n}}(\Lambda_l, \Lambda_{l'}) = 1 \quad \Longleftrightarrow \quad |\Psi(\Lambda_l)\rangle \propto |\Psi(\Lambda_l')\rangle\,, \tag{12}$$

since, depending on the precise network architecture used for $\psi(\boldsymbol{\sigma}; \Lambda)$, there are "gauge transformations" $g \in \mathcal{G}$ of the weights, $\Lambda_l \to g[\Lambda_l]$, with

$$|\Psi(\Lambda_l)\rangle = e^{i\vartheta_g} |\Psi(g[\Lambda_l])\rangle \tag{13}$$

for some global phase ϑ_g. We want to ensure that

$$S_{\mathrm{n}}(\Lambda_l, \Lambda_{l'}) = S_{\mathrm{n}}(\Lambda_l, g[\Lambda_{l'}]) = S_{\mathrm{n}}(g[\Lambda_l], \Lambda_{l'}) \tag{14}$$

for all such gauge transformations $g \in \mathcal{G}$. A general way to guarantee Eq. (14) proceeds by replacing,

$$S_{\mathrm{n}}(\Lambda_l, \Lambda_{l'}) \quad \longrightarrow \quad \max_{g, g' \in \mathcal{G}} S_{\mathrm{n}}(g[\Lambda_l], g'[\Lambda_{l'}]). \tag{15}$$

However, in practice, it might not be required to iterate over all possible gauge transformations in $\mathcal{G}$ due to the locality of the similarity measure. In the following, we will use the toric code and a specific RBM variational ansatz as an example to illustrate these gauge transformations and how an appropriate function f in Eq. (11) and gauge invariance (14) can be implemented efficiently.

Finally, note that, while we focus on applying DM in this work, a similarity measure in terms of neural network parameters can also be used for other kernel techniques such as kernel PCA. Depending on the structure of the underlying dataset, DM has clear advantage over kernel PCA: the former really captures the global connectivity of the dataset rather than the subspace with most variance that is extracted by the latter. This is why kernel PCA fails when identifying, e.g., winding numbers, in general datasets where DM still works well [12]. Specifically for our case study of the toric code below, we find that kernel PCA can also identify topological sectors for small T and without magnetic field, $h = 0$, as a result of the simple data structure; however, only DM works well when h is turned on, as we discuss below.

III. EXAMPLE: TORIC CODE

Now we illustrate our DM-based ML algorithm using the toric code model [58], defined on an $L_x \times L_y$ square

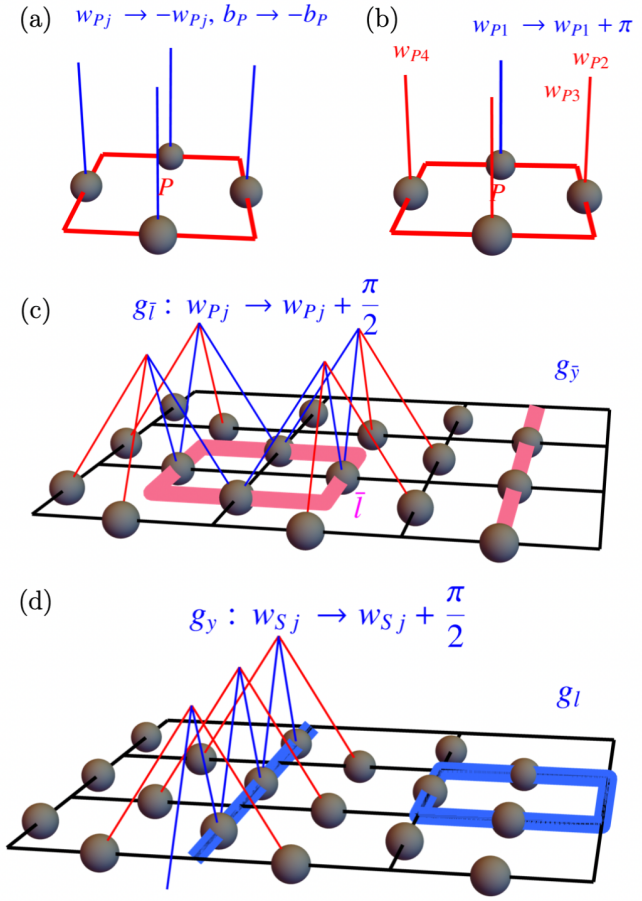

Figure 3. Gauge freedom of RBM ansatz in Eq. (17). The following transformations only lead to a global phase: (a) Multiplying all the parameters of a plaquette (or star, not shown) by a minus sign, see Eq. (18a); (b) A π shift of a single parameter, see Eqs. (18b) and (18c); (c) A $\pi/2$ shift to the weights crossed by a string $\bar{l}$, defined by $g_{\bar{l}}$ in Eq. (18e). The straight pink line represents the transformation on a non-contractible loop denoted by g_y; (d) Same as (c) but for loops on the direct lattice and g_l and $g_{\bar{y}}$, cf. Eq. (18d).

lattice with spin-1/2 operators or qubits on every bond, see Fig. 2(b), leading to a total of $N = 2L_xL_y$ spins; throughout this work, we will assume periodic boundary conditions. Referring to all four spins on the edges of an elementary square (vertex) of the lattice as plaquette P (star S), the plaquette and star operators are defined as $\hat{\mathcal{P}}_P = \prod_{i\in P} \hat{s}_i^z$ and $\hat{\mathcal{S}}_S = \prod_{i\in S} \hat{s}_i^x$, respectively. The toric code Hamiltonian then reads as

$$\hat{H}_{\rm tc} = -J_P \sum_P \hat{\mathcal{P}}_P - J_S \sum_S \hat{\mathcal{S}}_S, \tag{16}$$

where the sums are over all plaquettes and stars of the lattice. All "stabilizers" $\hat{\mathcal{P}}_P$, $\hat{\mathcal{S}}_S$ commute among each other and with the Hamiltonian. Focusing on $J_P, J_S > 0$, the ground states are obtained as the eigenstates with eigenvalue $+1$ under all stabilizers. A counting argument, taking into account the constraint $\prod_S \hat{\mathcal{S}}_S = \prod_P \hat{\mathcal{P}}_P = \mathbb{1}$, reveals that there are four, exactly degenerate ground states for periodic boundary conditions.

To describe the ground-states and low-energy subspace of the toric code model (16) variationally, we parameterize $\psi(\boldsymbol{\sigma}; \Lambda)$ in Eq. (1) using the ansatz

$$\psi_{\rm rbm}(\boldsymbol{\sigma}; \Lambda) = \prod_P \cos(b_P + \sum_{j\in P} w_{Pj}\sigma_j) \times \prod_S \cos(b_S + \sum_{j\in S} w_{Sj}\sigma_j), \tag{17}$$

proposed in Ref. 59, where every plaquette P (star S) is associated with a "bias" b_P (b_S) and four weights $w_{P,j}$ ($w_{S,j}$), all of which are chosen to be real here, i.e., $\Lambda = \{b_P, b_S, w_{P,j}, w_{S,j}\}$. This ansatz can be thought of as an RBM [46] (see Appendix A), as illustrated in Fig. 2(c), with the same geometric properties as the underlying toric code model. It is clear that Eq. (17) defines a quasi-local ansatz as it is of the form of Eq. (3), with $\jmath$ enumerating all plaquettes and stars (and thus $N_\jmath = 2N$). For this specific ansatz, the gauge transformations $g \in \mathcal{G}$, as introduced in Sec. II C above, are generated by the following set of operations on the parameters b_P, b_S, $w_{P,j}$, and $w_{S,j}$:

1. For X being any plaquette or star, multiplying all biases and weights of that plaquette or star by -1 [see Fig. 3(a)],

$$g_{X,-} : b_X \to -b_X,\ w_{Xj} \to -w_{Xj}, \tag{18a}$$

leaves the wave function invariant [$\vartheta_g = 0$ in Eq. (13)].

2. Adding π to either the bias or any of the weights associated with the plaquette or star X [see Fig. 3(b)],

$$g_{X,\pi,b} : b_X \to b_X + \pi, \tag{18b}$$
$$g_{X,\pi,j} : w_{Xj} \to w_{Xj} + \pi, \quad j \in X, \tag{18c}$$

leads to an overall minus sign [$\vartheta_g = \pi$ in Eq. (13)].

3. For any closed loop ℓ (or $\bar{\ell}$) on the direct (or dual lattice), adding $\frac{\pi}{2}$ to all weights of the stars (plaquettes) that are connected to the spins crossed by the string [see Fig. 3(c-d)],

$$g_\ell : w_{Sj} \to w_{Sj} + \frac{\pi}{2}, \quad Sj \in \ell, \tag{18d}$$
$$g_{\bar{\ell}} : w_{Pj} \to w_{Pj} + \frac{\pi}{2}, \quad Pj \in \bar{\ell}, \tag{18e}$$

leads to $\vartheta_g = 0$ or π in Eq. (13) depending on the length of the string. Note that any loop configuration $\mathcal{L}$, which can contain an arbitrary number of loops, can be generated by the set $\{g_S, g_P, g_{x,y}, g_{\bar{x},\bar{y}}\}$, where g_S (g_P) creates an elementary loop on the dual (direct) lattice encircling the star S (plaquette P), see Fig. 3(c,d), and $g_{x,y}$ ($g_{\bar{x},\bar{y}}$) creates a non-contractible loop on the direct (dual) lattice along the x, y direction. Since the length of any contractible loop is even, $\vartheta_g = 0$ for any string transformations generated by g_S and g_P. Meanwhile, on an odd lattice, the gauge transformations $g_{x,y}(g_{\bar{x},\bar{y}})$ involve an odd number of sites and thus lead to $\vartheta_g = \pi$.

A highly inefficient way of dealing with this gauge redundancy would be to use a choice of S_n in Eq. (11) which is not invariant under any of the transformations in Eq. (18); this would, for instance, be the case by just taking the Euclidean distance of the weights,

$$S_{\rm eu}(\Lambda_l,\Lambda_{l'}) \propto ||\Lambda_l - \Lambda_{l'}||^2 = \sum_X \Big[(b^l_X - b^{l'}_X)^2 + \sum_{j\in X}(w^l_{Xj} - w^{l'}_{Xj})^2\Big],$$

where the sum over X involves all plaquettes and stars. Naively going through all possible gauge transformations to find the maximum in Eq. (15) would in principle rectify the lack of gauge invariance. However, since the number of gauge transformations scales exponentially with system size N (holds for each of the three classes, 1.-3., of transformations defined above), such an approach would become very expensive for large N. Luckily, locality of the ansatz and of the similarity measure allows us to construct similarity measures that can be evaluated much faster: as an example, consider

$$S_n(\Lambda_l,\Lambda_{l'}) = \frac{1}{2} + \frac{1}{10N}\sum_X \max_{\tau_X=\pm}\Big[\sum_{j\in X}\cos 2(\tau_X w^l_{Xj} - w^{l'}_{Xj}) + \cos 2(\tau_X b^l_X - b^{l'}_X)\Big], \tag{19}$$

which clearly obeys $S_n(\Lambda_l,\Lambda_{l'}) = S_n(\Lambda_{l'},\Lambda_l)$, $S_n(\Lambda_l,\Lambda_{l'}) \in [0,1]$, and locality [it is of the form of Eq. (11) with $\jmath$ enumerating all X]. Concerning gauge invariance, first note that the choice of $\cos(\cdot)$ immediately leads to invariance under Eq. (18a). Second, for each X we only have to maximize over two values (τ_X) to enforce invariance under Eqs. (18b) and (18c), i.e., the maximization only doubles the computational cost.

The "string" redundancy, see Eqs. (18d) and (18e), however, is not yet taken into account in Eq. (19). It can be formally taken care of by maximizing over all possible loop configurations, denoted by $\mathcal{L}$,

$$S_{\rm str}(\Lambda_l,\Lambda_{l'}) = \frac{1}{2} + \frac{1}{10N}\max_{\mathcal{L}}\Big\{\sum_X \max_{\tau_X=\pm}\Big[\sum_{j\in X}\mu^{\mathcal{L}}_{Xj}\cos 2(\tau_X w^l_{Xj} - w^{l'}_{Xj}) + \cos 2(\tau_X b^l_X - b^{l'}_X)\Big]\Big\}, \tag{20}$$

where $\mu^{\mathcal{L}}_{Xj} = -1$ if Xj lives on a loop contained in $\mathcal{L}$ and $\mu^{\mathcal{L}}_{Xj} = 1$ otherwise. While there is an exponential number of such strings, Ref. 12 has proposed an algorithm to efficiently find an approximate maximum value. In our case, this algorithm amounts to randomly choosing a plaquette P or a star S or a direction $d = x, y$ and then applying g_S or g_P or $g_{d=x,y}$ to Λ_l in Eq. (19). If this does not decrease the similarity, keep that transformation; if it decreases the similarity, discard the gauge transformation. Repeat this procedure N_g times. In Ref. 12, N_g between 10^3 and 10^4 was found to be enough for a large system consisting 18×18 square-lattice sites (total of $N = 2 \times 18^2$ Ising spins). On top of this, g_S and g_P are local and, hence, the evaluation of the *change* of the local similarity in Eq. (19) under any given g_S or g_P only requires evaluation of $O(N^0)$ terms in (19), i.e., independent of system size.

In the numerical simulations below, using Eq. (19) without sampling over loop configurations $\mathcal{L}$ turned out to be sufficient. The reason is that, for our Markov-chain-inspired sampling procedure of Λ_l (see Appendix C), updates that correspond to these loop transformations happen very infrequently. Furthermore, even if a few pairs of samples are incorrectly classified as distinct due to the string redundancy, the DM will still correctly capture the global connectivity and, hence, absence or presence of topological sectors.

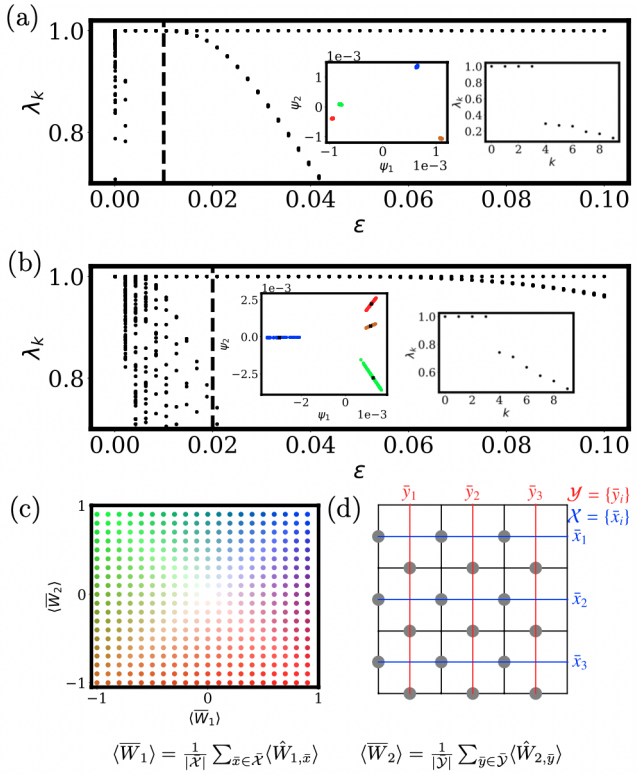

Figure 4. (a) DM spectrum for topological phase at $h = 0$ and $T = 0.1$ using the neutral network similarity measure in Eq. (19). Inset left: associated leading DM components; color represents the loop observable expectations values defined in (c-d), for ϵ indicated by the dashed line. Inset right: DM spectrum in descending order at $\epsilon = 0.01$ indicated by the dashed line. (b) Same as (a), but using exact overlaps $S_{\rm q}$ in Eq. (10) as metric. (c) Color map for the non-local loop values $\langle\overline{W}_1\rangle, \langle\overline{W}_2\rangle$ in the left insets of (a) and (b). (d) Different straight Wilson loops $\hat{W}_{1,\bar{x}_i}$ ($\hat{W}_{2,\bar{y}_i}$) along x (y) direction, denoted by blue (red) lines. The loop values in the color map in (c) are spatial averages over all straight-loop expectation values (as in the equations for $\langle\overline{W}_1\rangle, \langle\overline{W}_2\rangle$).

IV. NUMERICAL RESULTS

We next demonstrate explicitly how the general procedure outlined above can be used to probe and analyze topological order in the toric code. We start from the pure toric code Hamiltonian defined in Eq. (16) using the variational RBM ansatz in Eq. (17). An ensemble of network parameters is generated by applying the procedure of Sec. II A (see also Algorithm 1) for a system size of $N = 18$ spins; more details on ensemble generation, including the form of u in Eq. (4), are given in Appendix C. From now on, we measure all energies in units of J_P and set $J_S = J_P = 1$.

Let us first focus on the low-energy ensemble and choose $T = 0.1$ in Eq. (5). For the simple similarity measure in Eq. (19), that can be exactly evaluated at a time linear in system size N, we find the DM spectrum shown in Fig. 4(a) as a function of ϵ in Eq. (9). We observe the hallmark feature of four superselection sectors [12]: there is a finite range of ϵ where there are four eigenvalues exponentially close to 1. The association of samples (in our case states) and these four sectors is thus expected to be visible in a scatter plot of a projected subspace spanned by the first three non-trivial eigenvectors $\psi_{1,2,3}$ [12]; note the zeroth eigenvector $(\psi_0)_l = C$ is always constant with eigenvalue $\lambda = 1$ from probability conservation. In fact, we can see these clusters already in the first two components, see left inset in Fig. 4(a). Then a standard k-means algorithm is applied onto this projected subspace to identify the cluster number for each data point. Here, we use the standard algorithm in sklearn [71] to find the clusters given by diffusion map eigenvalues. To verify that the ML algorithm has correctly clustered the states according to the four physical sectors, we compute the expectation value for each state of the string operators,

$$\hat{W}_{1,\bar{x}} = \prod_{i\in\bar{x}} \hat{s}_i^x, \quad \hat{W}_{2,\bar{y}} = \prod_{i\in\bar{y}} \hat{s}_i^x, \tag{21}$$

where $\bar{x}(\bar{y})$ are loops defined on the dual lattice winding along the $x(y)$ direction, shown as blue lines in Fig. 2(b). We quantify the association of a state to physical sectors by the average of a set of straight loops $\mathcal{X}(\mathcal{Y})$ winding around the $x(y)$ direction, shown as blue (red) lines in Fig. 4(d). Indicating this averaged expectation value $\langle\overline{W}_1\rangle, \langle\overline{W}_2\rangle$ in the inset of Fig. 4(a) using the color code defined in Fig. 4(c), we indeed see that the clustering is done correctly.

To demonstrate that this is not a special feature of the similarity measure in Eq. (19), we have done the same analysis, with result shown in Fig. 4(b), using the full quantum mechanical overlap measure in Eq. (10). Quantitative details change but, as expected, four superselection sectors are clearly identified and the clustering is done correctly. We reiterate that the evaluation of the neural-network similarity measure in Eq. (19) [exact evaluation $\mathcal{O}(N)$] is much fast than that in Eq. (10) (even when it is computed with importance sampling). Note, however, that once S_n is computed for all samples, the actual DM-based clustering takes the same amount of computational time for both approaches. Consequently, suppose there is a quantum simulator that can measure the quantum overlap in Eq. (10) or any other viable similarity measure for that matter, then we can equivalently use the "measured" similarity for an efficient clustering of the superselection sectors via the DM scheme. As a next step, we demonstrate that the superselection sectors are eventually connected if we take into account states with sufficiently high energy. To this end, we repeat the same analysis but for an ensemble with $T = 1$. As can be seen in the resulting DM spectrum in Fig. 5(a), there is no value of ϵ where more than one eigenvalue is (exponentially) close to 1 and separated from the rest of the spectrum by a clear gap. Here we used again the simplified measure in Eq. (19), but have checked nothing changes qualitatively when using the overlap measure. To verify that this is the correct answer for the given dataset, we again computed the expectation value of the loop operators in Eq. (21) for each state in the ensemble. This is shown in Fig. 5(b), where we also use color to indicate the energy expectation value for each state. We

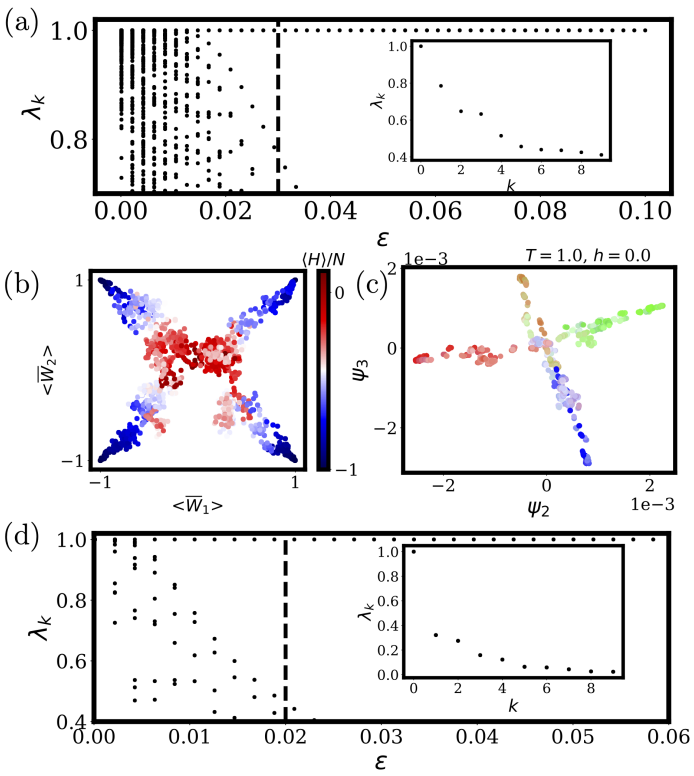

Figure 5. (a) DM spectrum for the high-energy ensemble at $h = 0$ and $T = 1$. The inset is the spectrum at $\epsilon = 0.03$ indicated by the dashed line in the main panel; (b) Spatially averaged straight Wilson loops $\langle\overline{W}_{1(2)}\rangle$ [see Fig. 4(c-d)] along two directions for the states in (a), where the color encodes energy density $\langle H\rangle/N$; (c) Leading DM components where the color of the dots encodes $\langle\overline{W}_{1(2)}\rangle$ using the color map in Fig. 4(c); (d) DM spectrum for the trivial phase at $h = 1.0$ and $T = 0.1$ using the quantum metric S_q.

can clearly see the four low-energy (blue) sectors (with $|W_{1,2}| \simeq 1$) are connected via high-energy (red) states (with $|W_{1,2}| \ll 1$). This agrees with the DM result that all states are connected within the ensemble (topological order is lost). We can nonetheless investigate the clustering in the leading three non-trivial DM components $\psi_{1,2,3}$. Focusing on a 2D projection in Fig. 5(c) for simplicity of the presentation, we can see that the DM reveals very interesting structure in the data: the four lobes roughly correspond to the four colors blue, red, orange, and green associated with the four superselection sectors and the states closer to $|W_{1,2}| = 1$ (darker color) appear closer to the tips. Finally, note that the colors are arranged such that the red and green [orange and blue] lobes are on opposite ends, as expected since they correspond to $(W_1, W_2) \simeq (1, -1)$ and $(-1, 1)$ [$(-1, -1)$ and $(1, 1)$].

Another route to destroying topological order proceeds via application of a magnetic field. To study this, we extend the toric code Hamiltonian according to

$$\hat{H}'_{\rm tc} = \hat{H}_{\rm tc} - h \sum_i \hat{s}_i^z. \tag{22}$$

Clearly, in the limit of $h \to \infty$, the ground state is just a state where all spins are polarized along $\hat{s}^z$ and topological order is lost. Starting from the pure toric model ($h = 0$) and turning on h reduces the gap of the "charge excitations" defined by flipping $\hat{\mathcal{S}}_S$ from $+1$ in the toric code groundstate to -1. Their condensation leads to a second-order quantum phase transition [72–75].

Before addressing the transition, let us study the large-h limit. We first note that our ansatz in Eq. (17) can capture the polarized phase as well. For instance, denoting the "northmost" (and "southmost") spin of the plaquette P (and star S) by $j_0(P)$ (and $j_0(S)$), respectively, the spin polarized state is realized for [see also Fig. 8(a) in the Appendix]

$$b_P = b_S = -\frac{\pi}{4}, \quad w_{Xj} = \begin{cases} \frac{\pi}{4}, & j = j_0(X), \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

In fact, the spin polarized state has many representations within our RBM ansatz in Eq. (17), including representations that are not just related by the gauge transformations in Eq. (18). For instance, the association $j \to j_0(X)$ of a spin to a plaquette and star can be changed, e.g., by using the "easternmost" spin. As discussed in more detail in Appendix A 2, this redundancy is a consequence of the product from of $\psi_{\rm rbm}(\boldsymbol{\sigma})$ in Eq. (17) and the fact that $\psi_{\rm rbm}(\boldsymbol{\sigma})$ is *exactly* zero if there is a single j with $\sigma_j = -1$; consequently, it is a special feature of the simple product nature of the spin-polarized ground state. While in general there can still be additional redundancies besides the aforementioned gauge transformations, we do not expect such a structured set of redundancy to hold for generic states. There are various ways of resolving this issue. The most straightforward one is to replace the simple overlap measure $S_{\rm n}$ in Eq. (11) by the direct overlap $S_{\rm q}$ in Eq. (10) for a certain fraction of pairs of samples l and l'. If this fraction is large enough, the DM algorithm will be able recognize that clusters of network parameters that might be distinct according to $S_{\rm n}$ actually correspond to identical wave functions. We refer to Appendix A 3 where this is explicitly demonstrated. We note, however, that kernel PCA will not work anymore in this case; it will incorrectly classify connected samples

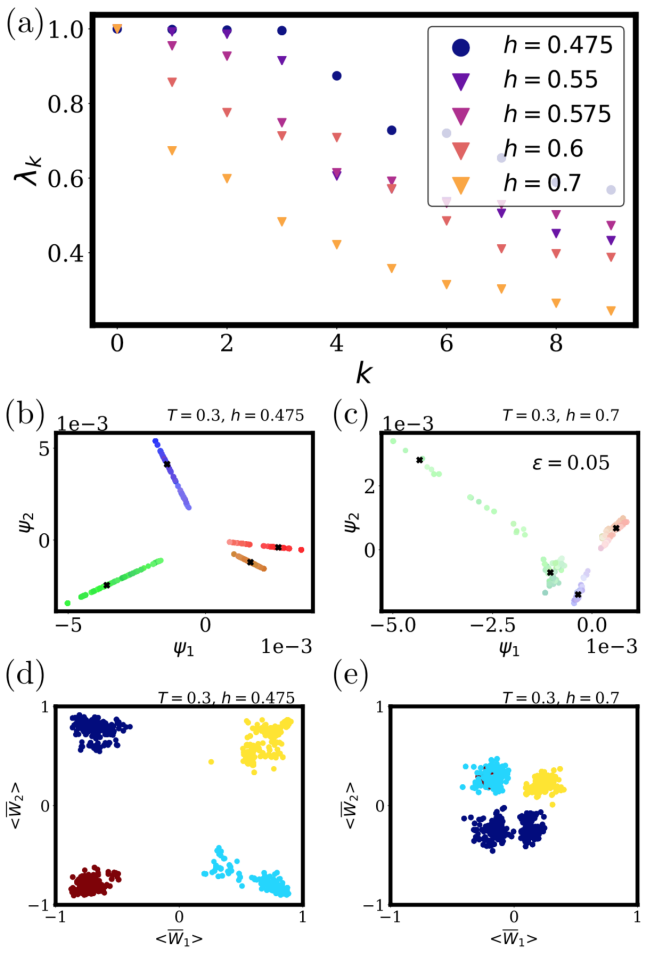

Figure 6. DM spectra for low-energy ensembles with $T = 0.3$ at finite field h. (a) First 10 eigenvalues for various field values $h = 0.475, 0.55, 0.575, 0.6, 0.7$ at $\epsilon = 0.05$. The dot marker ($h = 0.475$) shows that the eigenvalue spectra have four-fold degeneracy, indicating signature for topological order. In comparison, for spectra marked by the triangular markers ($h \geq 0.55$), such degeneracy is absent. A transition field value $h_t \simeq 0.55$ is identified by observing that a gap opens in the degenerate eigenvalue spectra. This is consistent with what we have observed in the fidelity using the same dataset [see Appendix B 1]. (b) Projected eigenvectors onto the first two components for $h = 0.475$. The color encodes $\langle \overline{W}_{1(2)} \rangle$ with the color scheme of Fig. 4(c). The black cross marks the k-means centers. (c) Same as (b) for $h = 0.7$. (d) Expectation for averaged straight Wilson loops $\langle \overline{W}_{1(2)} \rangle$ along two directions for the states in (b). The color encodes the clustering results from k-means in the projected subspace of the eigenvectors shown in (b). (e) Same as (d) for ensemble shown in (c).

as distinct as it's based on the variance of the data rather than connectivity. For simplicity of the presentation, we use S_q for all states in the main text and focus on DM.

The DM spectrum for large magnetic field, $h = 1$, and low temperatures, $T = 0.1$, is shown in Fig. 5(d). Clearly, there is no value of ϵ for which there is more than one eigenvalue close to 1 while exhibiting a gap to the rest of the spectrum. This shows that, as expected, the magnetic field h has lead to the loss of topological order.

To study with our DM algorithm the associated phase transition induced by h, we repeat the same procedure for various different values of h. The resulting spectra for selected h are shown in Fig. 6(a). We see that there are still four sectors for $h = 0.55$ in the data that are absent for $h = 0.575$ and larger values. While the associated critical value of h is larger than expected [72–74], this is not a shortcoming of the DM algorithm but rather a consequence of our simple local variational ansatz in Eq. (17). In particular, one can analytically show that the simple RBM anstaz is able to capture the exact Toric code state as well as exact polarized state (Appendix A), while not faithfully capture states around the critical field. By computing the fidelity as well as loop-operator expectation values, we can see that a critical value around $h = 0.55$ is the expected answer for our dataset (see Appendix B 1). More sophisticated ansätze for the wavefunction are expected to yield better values, but this is not the main focus of this work. More importantly, we see in Fig. 6(b) that the DM clustering of the states correctly reproduces the clustering according to the averaged loop operator expectation values $\langle \overline{W}_j \rangle$ (again indicated with color). Note that we have further projected the states onto a two-dimensional subspace of the three-dimensional subspace identified by diffusion maps [76]. Alternatively, this can be seen in Fig. 6(d) where $\langle \overline{W}_j \rangle$ is indicated for the individual samples. Using four different colors for the four different clusters identified by the DM, we see that all states are clustered correctly. As expected based on the eigenvalues, there are no clear clusters anymore for larger h, Fig. 6(c); nonetheless, naively applying k-means clustering in $\psi_{1,2,3}$ manages to discover some residual structure of the wavefunctions related to $\langle \overline{W}_j \rangle$ as demonstrated in Fig. 6(e). Note that while in Fig. 6(c) we have plotted the two-dimensional subspace, diffusion maps predict the clustering by projecting down to a $1d$ subspace of the leading eigenvector component ψ_1.

V. SUMMARY AND DISCUSSION

In this work, we have described an unsupervised ML algorithm for quantum phases with topological order. We use neural network parameters to efficiently represent an ensemble of quantum states, which are sampled according to their energy expectation values. To uncover the structure of the superselection sectors in the quantum states, we used the dimensional reduction technique of diffusion map and provided a kernel defined in terms of network parameters. As opposed to a kernel based on the overlap of wavefunctions (or other quantum mechanical similarity measures of states for that matter), this metric can be evaluated efficiently (within polynomial time) on a classical computer.

We illustrated our general algorithm using a quasi-local restricted Boltzmann machine (RBM) and the toric code model in an external field; the choice of network ansatz was inspired by previous works [59, 60] showing the existence of efficient representations of the low-energy spectrum in terms of RBMs. Allowing for spatially inhomogeneous RBM networks, we identified the "gauge symmetries" of the ansatz, i.e., the set of changes in the network parameters that do not change the wavefunction, apart from trivial global phase factors. We carefully designed a similarity measure that is gauge invariant—a key property as, otherwise, identical wavefunctions represented in different gauges would be falsely identified as being distinct. We showed that the resultant unsupervised diffusion-map-based embedding of the wavefunctions is consistent with the expectation values of loop operators; it correctly captures the presence of superselection sectors and topological order at low energies and fields, as well as the lack thereof when higher-energy states are involved and/or the magnetic field is increased. We also verified our results using the full quantum mechanical overlap. Since our procedure is general, natural follow-up works would explore applications to other models and using other variational *ansätze*. In that regard, models where the ideal variational ansatz (RBM or beyond) is not known, seem particularly interesting.

On a more general level, our analysis highlights the importance of the following two key properties of diffusion maps: first, in the presence of different topological sectors, the leading eigenvectors of diffusion maps capture the connectivity rather than, e.g., the variance as is the case for PCA. For this reason, the clustering is still done correctly even if a fraction of pairs of wavefunctions are incorrectly classified as being distinct due to the usage of an approximate similarity measure. This is why complementing the neural-network similarity measure, which has additional, state-specific redundancies in the large-field limit, by direct quantum mechanical overlaps for a certain fraction of pairs of states is sufficient to yield the correct classification. The second key property is that diffusion map is a kernel technique. This means that the actual machine learning procedure does not require the full wavefunctions as input; instead, only (some measure of) the kernel of all pairs of wavefunctions in the dataset is required. We have used this to effectively remove the gauge redundancy in the RBM parametrization of the states by proper definition of the network similarity measure in Eq. (20). Since the evaluation of full quantum mechanical similarity measures, like the wavefunction overlap, are very expensive on classical computers, an interesting future direction would be to use the emerging quantum-computing resources to evaluate a similarity

measure quantum mechanically. This could then be used as input for a diffusion-map-based clustering.

We finally point out that the ensemble of states we used in this work, which was based on sampling states according to their energy with respect to a Hamiltonian, is only one of many possibilities. The proposed technique of applying diffusion map clustering using a gauge-invariant kernel in terms of network parameters of a variational description of quantum many-body wavefunctions can be applied more generally, in principle, to any ensemble of interest. For instance, to consider arbitrary local perturbations, one could generate an ensemble using finite depth local unitary circuits. Alternatively, one could generate an ensemble based on (Lindbladian) time-evolution to probe the stability of topological order against time-dependent perturbations or the coupling to a bath. We leave the investigation of such possibilities for future works.

VI. CODE AND DATA AVAILABILITY

The Monte Carlo simulations in this work were implemented in JAX [77]. Python code and data will be available at https://github.com/teng10/ml_toric_code/.

ACKNOWLEDGEMENTS

Y.T. acknowledges useful discussions with Dmitrii Kochkov, Juan Carrasquilla, Khadijeh Sona Najafi, Maine Christos and Rhine Samajdar. Y.T. and S.S. acknowledge funding by the U.S. Department of Energy under Grant DE-SC0019030. M.S.S. thanks Joaquin F. Rodriguez-Nieva for a previous collaboration on DM [12]. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Appendix A: Variational Ansatz: Restricted Boltzmann Machine

The variational ansatz in Eq. (17) is a *further-restricted* restricted Boltzmann machine (RBM), first introduced by Ref. 59. RBM is a restricted class of Boltzmann machine with an "energy" function $E_{\mathrm{RBM}}(\boldsymbol{\sigma}, \boldsymbol{h}; \Lambda)$ dependent on the network parameters Λ, where $\boldsymbol{\sigma}$ are physical spins and $\boldsymbol{h} = \{h_1, h_2, \cdots, h_N | \, h_i = \pm 1\}$ are hidden spins (or hidden neurons) that are Ising variables. The parameters Λ define the coupling strength among the physical and hidden spins. The restriction in RBM is that the couplings are only between the physical spin σ_i and hidden spin h_j with strength $-w_{ij}$, so that the "energy" function takes the form $E_{\mathrm{RBM}}(\boldsymbol{\sigma}, \boldsymbol{h}; \Lambda) = -\sum_i a_i \sigma_i - \sum_i b_i h_i - \sum_{ij} w_{ij} \sigma_i h_j$. It is a generative neural network that aims to model a probability distribution $\mathbb{P}$ based on the Boltzmann factor,

$$\mathbb{P}(\boldsymbol{\sigma}; \Lambda) = \frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E_{\mathrm{RBM}}(\boldsymbol{\sigma}, \boldsymbol{h}; \Lambda)}, \tag{A1a}$$

$$\text{normalization} \quad Z = \sum_{\boldsymbol{\sigma}, \boldsymbol{h}} e^{-E_{\mathrm{RBM}}(\boldsymbol{\sigma}, \boldsymbol{h}; \Lambda)}. \tag{A1b}$$

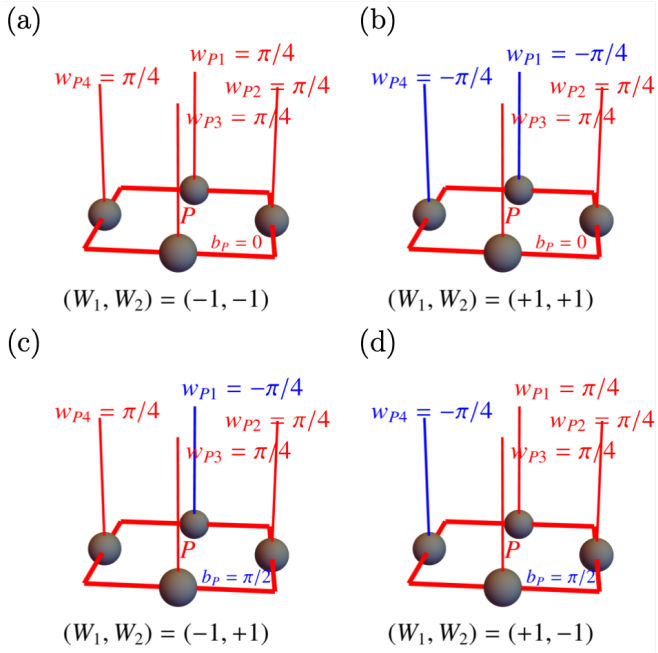

Figure 7. RBM representations of the four toric code ground states in the eigenbasis [Eq. (A4)] of loop operators $\hat{W}_1, \hat{W}_2$ in Eq. (A3a).

For the task of modeling a quantum wavefunction amplitude $\psi(\boldsymbol{\sigma}; \Lambda)$, RBMs can be used as a variational ansatz by extending the parameters Λ to complex numbers.

Further restricting parameters to the interlayer connections to the plaquette and star geometry in the toric code model [cf. Fig. 2(c)] and taking all parameters Λ to be purely imaginary, we recover the ansatz in Eq. (17) (up to normalization factor $\widetilde{Z}$),

$$\begin{aligned} \psi(\boldsymbol{\sigma}; \Lambda) &= \frac{1}{\widetilde{Z}} \sum_{X=P,S} \sum_{h_X = \pm 1} e^{-i \sum_X (w_{Xj} \sigma_j + b_X) h_X}, \\ &= \frac{1}{\widetilde{Z}} \prod_{X=P,S} \cos\Big(\sum_{j \in X} w_{Xj} \sigma_j + b_X\Big). \end{aligned} \tag{A2}$$

The $\cos(\cdot)$ factors come from summing over the hidden neurons and the ansatz factorizes into the product of individual plaquette (star) terms because of the restricted connections. The estimation of physical observables of a wave function based on the RBM ansatz requires Monte Carlo sampling procedure which we discuss in Appendix B.

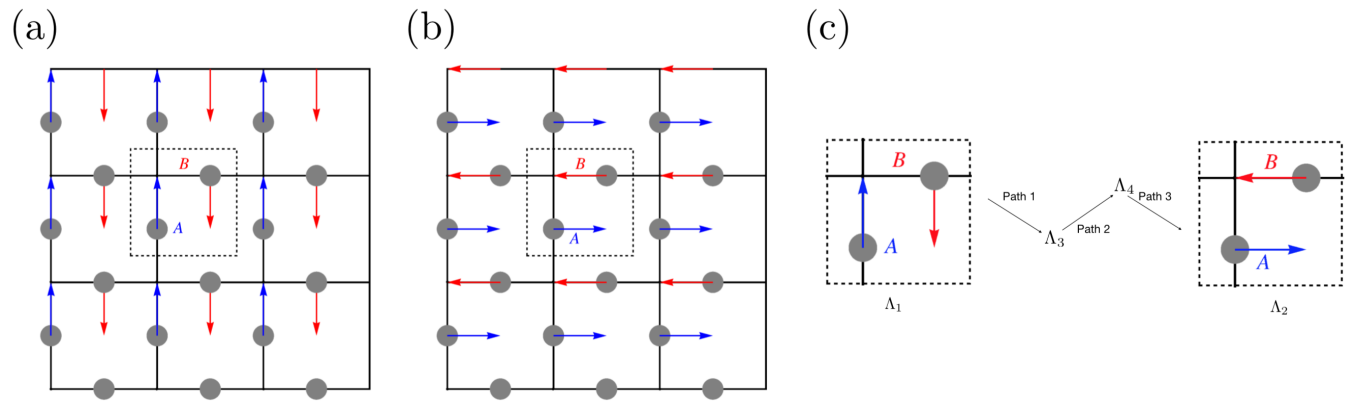

Figure 8. (a-b) Two RBM representations Eq. (A8) of the polarized state. (c) A path that connects the presentation for two spins in (a-b), which is explicitly shown in Table. I.

1. Ground states representation in different topological sectors

Placing the toric code model in Eq. (16) on the torus geometry, it is useful to define the loop operators,

$$\hat{W}_1 = \prod_{i\in\bar{l}_x} \hat{s}_i^x, \quad \hat{W}_2 = \prod_{i\in\bar{l}_y} \hat{s}_i^x, \tag{A3a}$$

$$\hat{V}_1 = \prod_{i\in l_x} \hat{s}_i^z, \quad \hat{V}_2 = \prod_{i\in l_y} \hat{s}_i^z, \tag{A3b}$$

where $l_{x,y}$ is a non-contractible loop along x, y direction, and $\bar{l}_{x,y}$ is similar on the dual lattice. Note the loop operators along two directions do not commute with each other as $\left[\hat{W}_1, \hat{V}_2\right] \neq 0$ and $\left[\hat{W}_2, \hat{V}_1\right] \neq 0$. However, since the hamiltonian commute with these loop operators $\left[\hat{W}_{1,2}, \hat{H}_{\rm tc}\right] = \left[\hat{V}_{1,2}, \hat{H}_{\rm tc}\right] = 0$, it follows that the ground state subspace is four-fold degenerate and spanned by the eigenvectors of the loop operators.

Suppose we work in the eigenbasis of $\hat{W}_{1,2}$; we define the four orthogonal ground states $|\psi_i\rangle$ $(i = 0, 1, 2, 3)$ that span $\mathcal{L}$ as,

$$\hat{W}_1 |\psi_0\rangle = -|\psi_0\rangle, \quad \hat{W}_2 |\psi_0\rangle = -|\psi_0\rangle, \tag{A4a}$$

$$\hat{W}_1 |\psi_1\rangle = |\psi_1\rangle, \quad \hat{W}_2 |\psi_1\rangle = |\psi_1\rangle, \tag{A4b}$$

$$\hat{W}_1 |\psi_2\rangle = -|\psi_2\rangle, \quad \hat{W}_2 |\psi_2\rangle = |\psi_2\rangle, \tag{A4c}$$

$$\hat{W}_1 |\psi_3\rangle = |\psi_3\rangle, \quad \hat{W}_2 |\psi_3\rangle = -|\psi_3\rangle. \tag{A4d}$$

The RBM ansatz in Eq. (A2) can represent eigenstates of $\hat{W}_{1,2}$ with eigenvalues $(W_1, W_2) = (\pm 1, \pm 1)$. Ref. [59] gave an representation of $|\psi_0\rangle$ with parameters,

$$w_{Pj} = \frac{\pi}{4}, \quad b_P = 0, \quad w_{Sj} = \frac{\pi}{2}, \quad b_S = 0. \tag{A5a}$$

On a system with odd number of sites along x and y direction, the other three degenerate states can be realized analogously by fixing the weights associated to stars to be $w_{Sj} = 0, b_S = 0$. Then the four states can be chosen by changing the w_{Pj} and b_P as shown in Fig. 7.

2. Network parameter redundancies in polarized phase

In Sec. III, we identified a set of gauge transformations Eq. (18) that leave a generic wavefunction parameterized by the RBM ansatz invariant up to a global phase. Such gauge transformations should be taken into consideration when evaluating the similarity measure S_n. Moreover, we have numerically verified that for states generated close to the exact toric code wave functions, S_n is a good proxy for the quantum measure $S_{\rm q}$ after explicit removals of such redundancies via S_n in Eq. (19). However, as alluded to in the discussions of the large-h limit, there are state-specific redundancies that are generally not related by the gauge transformations in Eq. (18).

Let us illustrate such redundancies here for the polarized state $|\Psi\rangle = |1, \cdots, 1\rangle_z$ which has all up spins in z-basis. Notice that there is the same number of $\cos(\cdot)$ factors in the wavefunction ansatz as the number of spins. As a result, we can define a "*covering*" by assigning each individual spin to a single factor, and choosing the weights to ensure all spins are up. Any such "covering" is a valid representation of the polarized state. For example, one representation is,

$$b_P = b_S = -\frac{\pi}{4}, \quad w_{Sj} = \begin{cases} \frac{\pi}{4}, & j = j_s(S), \\ 0, & \text{otherwise}, \end{cases} \quad \text{and} \quad w_{Pj} = \begin{cases} \frac{\pi}{4}, & j = j_n(P), \\ 0, & \text{otherwise}. \end{cases} \tag{A6}$$

where $j_s(S)$ denotes the "southmost" spin in the star S and $j_n(P)$ denotes the "northmost" spin in the plaquette P [see Fig. 8(a)]. Any such coverings of the spins will correspond to a polarized state. For example, performing a "rotation" leads to a different covering in Fig. 8(b). Actually, because most amplitudes in local-z basis are 0 so there are so few constraints in the wave function amplitudes, a continuous set of weights exist to represent the polarized state, so there are an infinite amount of redundancies for completely polarized state.

To illustrate this, let us consider an example of just two spins [the boxed region in Fig. 8(c)] with the same RBM ansatz, which can be easily generalized to more spins. For two spins, such ansatz is given by,

$$\psi_\Lambda(\sigma_A,\sigma_B)=\cos(b_S+w_{SA}\sigma_A+w_{SB}\sigma_B)\cos(b_P+w_{PA}\sigma_A+w_{PB}\sigma_B), \tag{A7}$$

where the weights $\Lambda=\{\Lambda_S=\{b_S,w_{SA},w_{SB}\},\Lambda_P=\{b_P,w_{PA},w_{PB}\}\}$ with $\Lambda_{Xj}\in[0,\pi)$ for $X=S$ or P fully determine the two-qubits state. For example, the following two choices of weights [Λ_1 and Λ_2 pictorially in Fig. 8(c)] both parametrize the polarized state:

$$\Lambda_1=\{b_S=-\frac{\pi}{4},w_{SA}=0,w_{SB}=\frac{\pi}{4},b_P=-\frac{\pi}{4},w_{PA}=\frac{\pi}{4},w_{PB}=0\}, \tag{A8a}$$

$$\Lambda_2=\{b_S=-\frac{\pi}{4},w_{SA}=\frac{\pi}{4},w_{SB}=0,b_P=-\frac{\pi}{4},w_{PA}=0,w_{PB}=\frac{\pi}{4}\}, \tag{A8b}$$

$$\psi_{\Lambda_{1,2}}=\begin{cases}1, & \sigma_A=\sigma_B=1,\\ 0, & \text{otherwise.}\end{cases} \tag{A8c}$$

Now to illustrate the continuous redundancies, we construct a path in the parameter space to go from Λ_1 to Λ_2. The path is composed of three steps [Fig. 8(c)],

$$\Lambda_1\xrightarrow{\text{path 1}}\Lambda_3\xrightarrow{\text{path 2}}\Lambda_4\xrightarrow{\text{path 3}}\Lambda_2, \tag{A9}$$

where the intermediate parameters are given by,

$$\Lambda_3=\{b_S=0,w_{SA}=\frac{\pi}{4},w_{SB}=-\frac{\pi}{4},b_P=-\frac{\pi}{4},w_{PA}=\frac{\pi}{4},w_{PB}=0\}, \tag{A10}$$

$$\Lambda_4=\{b_S=0,w_{SA}=\frac{\pi}{4},w_{SB}=-\frac{\pi}{4},b_P=-\frac{\pi}{4},w_{PA}=0,w_{PB}=\frac{\pi}{4}\}. \tag{A11}$$

Along each path component, referred to as path 1 through 3 in Table I, the parameters of S (or P) are varied and the

Path 1 $\Lambda_1\to\Lambda_3$	$w_{SB}=b_S+w_{SA}-\frac{\pi}{2}$ $w_{SA}:[0,\frac{\pi}{4}),w_{SB}:[\frac{\pi}{4},-\frac{\pi}{4}),b_S:[-\frac{\pi}{4},0)$	Λ_P fixed $w_{PA}=\frac{\pi}{4},w_{PB}=0,b_P=-\frac{\pi}{4}$	product $\psi=\psi_S\times\psi_P$
$\cos(b_X+w_{XA}+w_{XB})$	$\neq 0$ if $b_S+w_{SA}\neq\frac{n}{2}\pi,n\in\mathbb{Z}\to 0\to 1$	1	$\to 0\to 1$
$\cos(b_X+w_{XA}-w_{XB})$	0		0 ✓
$\cos(b_X-w_{XA}+w_{XB})$	$\cos(2b_S-\frac{\pi}{2})\to 0$	0	0 ✓
$\cos(b_X-w_{XA}-w_{XB})$		0	0 ✓
Path 2 $\Lambda_3\to\Lambda_4$	Λ_S fixed $w_{SA}=\frac{\pi}{4},w_{SB}=-\frac{\pi}{4},b_S=0$	$w_{PB}=b_P-w_{PA}+\frac{\pi}{2}$ $w_{PA}:[\frac{\pi}{4},0],w_{PB}:[0,\frac{\pi}{4}],b_P=-\frac{\pi}{4}$	
$\cos(b_X+w_{XA}+w_{XB})$	1	1	1
$\cos(b_X+w_{XA}-w_{XB})$	0	$\cos(2w_{PA}-\frac{\pi}{2})\to 0$	0 ✓
$\cos(b_X-w_{XA}+w_{XB})$	0		0 ✓
$\cos(b_X-w_{XA}-w_{XB})$		0	0 ✓
Path 3 $\Lambda_4\to\Lambda_2$	$w_{SB}=-b_S+w_{SA}+\frac{\pi}{2}$ $w_{SA}=\frac{\pi}{4},w_{SB}:(-\frac{\pi}{4},0],b_S:(0,-\frac{\pi}{4}]$	Λ_P fixed $w_{PA}=0,w_{PB}=\frac{\pi}{4},b_P=-\frac{\pi}{4}$	
$\cos(b_X+w_{XA}+w_{XB})$	1	1	1
$\cos(b_X+w_{XA}-w_{XB})$		0	0 ✓
$\cos(b_X-w_{XA}+w_{XB})$	0		0 ✓
$\cos(b_X-w_{XA}-w_{XB})$		0	0 ✓

Table I. A path going from Λ_1 to Λ_2 is composed of three steps. Path 1 ($\Lambda_1\to\Lambda_3$) is smooth except at the point $w_{SA}=\frac{\pi}{4},w_{SB}=-\frac{\pi}{4},b_S=0$, where the wavefunction vanishes. This is denoted by the red arrows in the first row. Path 2 and 3 are both smooth. The last column illustrates that the wavefunction ψ remains in the polarized state along the path.

other held fixed, while remaining in the exactly polarized state. The path is continuous except at a singular point on path 1 where the wave function vanishes at $\Lambda_{\text{singular}} = \{b_S = 0, w_{SA} = \frac{\pi}{4}, w_{SB} = -\frac{\pi}{4}, b_P = -\frac{\pi}{4}, w_{PA} = \frac{\pi}{4}, w_{PB} = 0\}$.

3. Resolving the special redundancies

In Appendix A 2, we explicitly showed that there can be a large set of redundancies given a polarized state. Hence, for simplicity in the main text, we have used the direct overlap S_{q} in Eq. (10) as the relevant measure at finite field values. As discussed in the main text, a straightforward way to alleviate the redundancies in the similarity measure S_n in Eq. (19) of the network parameters is to complement it with the direct overlap. By using a combination of both measures, we are able to reduce the amount of computational cost of the direct overlap by a fraction as the similarity is easy to compute. More specifically, we define a mixed measure S_m by replacing a random fraction (given by f) of the similarity measure pairs $\{l, l'\}$ by a rescaled overlap measure $\widetilde{S}_q$ such that,

$$S_m(l,l') = \begin{cases} \widetilde{S}_q(l,l') & \text{with probability } f, \\ S_n(l,l') & \text{with probability } 1-f. \end{cases} \tag{A12}$$

The following rescaling of the overlap measure S_q is necessary as we want to include the two measures on an equal-footing given by,

$$\widetilde{S}_q = \frac{S_q - n_q}{m_q - n_q} \cdot (m_n - n_n) + n_n, \tag{A13a}$$

$$m_q = \max(S_q), \quad n_q = \min(S_q), \tag{A13b}$$

$$m_n = \max(S_n), \quad n_n = \min(S_n). \tag{A13c}$$

For example, we see that the minimum of the rescaled overlap is the same as the minimum of the similarity $\min(\widetilde{S}_q) = \min(S_n)$.

In Fig. 9, we demonstrate that by using a mixed measure with a fraction of $f = 0.4$ replacement, our algorithm with DM is able to identify the presence (indicated by the shaded blue region for smaller field values $h = 0.475$ and $h = 0.55$) and absence ($h = 0.7$) of superselection sectors across various field values, consistent with the predictions of the algorithm using direct overlap (shown in Fig. 6). We note that in the case with a mixed measure, DM is a natural technique as the algorithm looks for connectivity; whereas kernel PCA would fail to identify such transition (since a fraction of pairs of wave functions are incorrectly considered to be dissimilar by S_n, the leading kernel PCA components still show four separated clusters up to the largest magnetic field, $h = 1$).

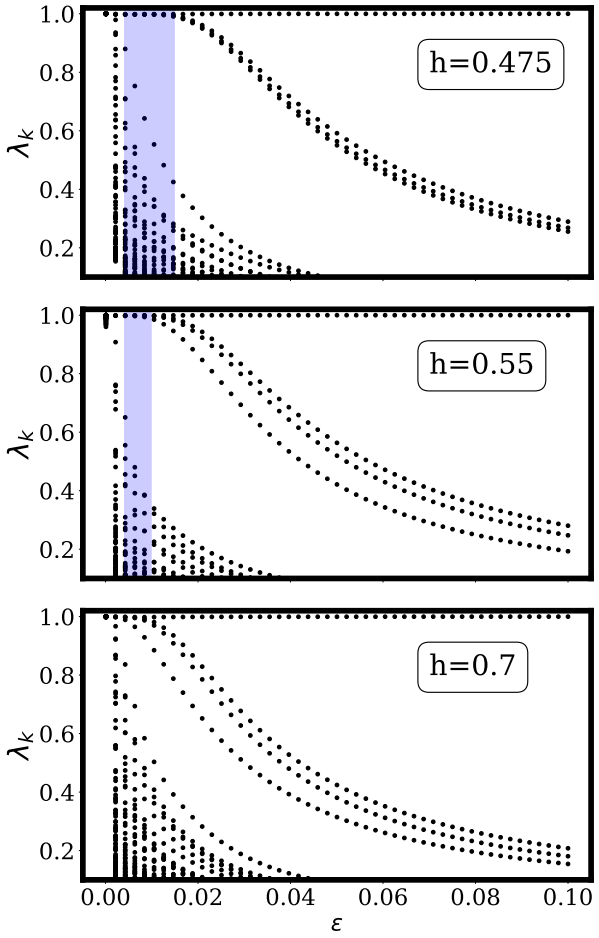

Figure 9. DM spectra for different field values $h = 0.475, 0.55, 0.7$ at $T = 0.3$ using a mixed similarity measure S_m with a fraction $f = 0.4$ in Eq. (A12). The blue shaded regions highlight the existence of a range of ϵ with spectral gap between the degenerate eigenvalues and the decaying eigenvalues, indicating underlying superselection sectors. As the field value approaches the transition field h_c, the range of such region shrinks and disappears at high field $h = 0.7$, indicating the absence of sectors.

Appendix B: Optimization with Variational Monte Carlo

To find the ground state $|\Psi(\Lambda^0)\rangle \propto \sum_{\boldsymbol{\sigma}} \psi(\boldsymbol{\sigma}; \Lambda^0) |\boldsymbol{\sigma}\rangle$, we wish to minimize the energy expectation $\langle E \rangle = \langle \Psi | \hat{H} | \Psi \rangle / \langle \Psi | \Psi \rangle$ (omitting the variational parameters Λ^0 in this section), which is bounded by the ground state energy by the variational principle. An exact computation $\langle E \rangle_{\text{exact}}$ is costly as the summation enumerates over exponentially many spin configurations $\boldsymbol{\sigma}$ as the system size increases. Here we use variational Monte Carlo (VMC) importance sampling algorithm to estimate such expectation values. The idea is to compute relative probability between different configurations and sample from the true wavefunction probability density $|\psi(\boldsymbol{\sigma})|^2$, without having to compute $|\psi(\boldsymbol{\sigma})|^2$ for all $\boldsymbol{\sigma}$. To perform this algorithm, we initialize M random configurations

$\{\boldsymbol{\sigma}_i\}_{i=1}^M$ and continue each with random walks based on previous configurations, hence forming M Markov chains.

In particular, the Metropolis–Rosenbluth algorithm [78] is used to propose the next configuration $\boldsymbol{\sigma}_i'$ that is locally connected to c_i according to function $g(\boldsymbol{\sigma}'|\boldsymbol{\sigma})$. For the toric code model, we use two types of proposals: spin flips and vertex flips. Here, we will assume a probability of p for proposing spin flips and analogously $1-p$ for vertex flips that are equally likely at all sites:

$$g(\boldsymbol{\sigma}'|\boldsymbol{\sigma}) = \begin{cases} \frac{p}{n_s}, & \text{for spin flips} \\ \frac{1-p}{n_v}, & \text{for vertex flips} \end{cases} \tag{B1}$$

where n_s and n_v are the number of all possible spin and vertex flips. The acceptance of $\boldsymbol{\sigma}'$ is determined by a probability,

$$\mathbb{P}_{\text{accept}}(\boldsymbol{\sigma} \to \boldsymbol{\sigma}') = \min\left(|\frac{\psi(\boldsymbol{\sigma}')}{\psi(\boldsymbol{\sigma})}|^2, 1\right). \tag{B2}$$

The random walks will be repeated long enough so that the final configurations at the tail of the chains $\Sigma_{\text{MC}} = \{\boldsymbol{\sigma}_f\}_{i=b}^M$ approximate samples drawn from the probability distribution $|\psi(\boldsymbol{\sigma})|^2$. A certain number b of walkers in each chain are discarded to reduce the biases from initialization of the chains. Then the expectation of an observable $\hat{O}$ is given by,

$$\langle \hat{O} \rangle_{\text{MC}} = \frac{\sum_{\boldsymbol{\sigma}} \psi(\boldsymbol{\sigma})^* \langle \boldsymbol{\sigma}|\hat{O}|\Psi\rangle}{\sum_{\boldsymbol{\sigma}} |\psi(\boldsymbol{\sigma})|^2}, \tag{B3a}$$

$$= \frac{\sum_{\boldsymbol{\sigma}} |\psi(\boldsymbol{\sigma})|^2 \frac{\langle \boldsymbol{\sigma}|\hat{O}|\Psi\rangle}{\psi(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}} |\psi(\boldsymbol{\sigma})|^2}, \tag{B3b}$$

$$= \frac{1}{M} \sum_{\boldsymbol{\sigma} \in \Sigma_{\text{MC}}} \frac{\langle \boldsymbol{\sigma}|\hat{O}|\Psi\rangle}{\psi(\boldsymbol{\sigma})}. \tag{B3c}$$

Defining a local value of the operator $\hat{O}$ as,

$$O_{\text{loc}} = \frac{\langle \boldsymbol{\sigma}|\hat{O}|\Psi\rangle}{\psi(\boldsymbol{\sigma})}, \tag{B4}$$

then the Monte Carlo estimation is the average of the local values in the Markov chain: $\langle \hat{O} \rangle_{\text{MC}} = \frac{1}{M} \sum_{\boldsymbol{\sigma} \in \Sigma_{\text{MC}}} O_{\text{loc}}$.

Next, to minimize $\langle E \rangle$, we can compute its gradient with respect to the weights Λ^0 in terms of the local energy E_{loc} and wavefunction amplitude derivative D_i:

$$\partial_{\Lambda_i} \langle E \rangle = \langle E_{\text{loc}} D_i \rangle - \langle E_{\text{loc}} \rangle \langle D_i \rangle \tag{B5a}$$

$$E_{\text{loc}} = \frac{\langle \boldsymbol{\sigma}| H |\Psi\rangle}{\psi(\boldsymbol{\sigma})}, \quad D_i = \frac{\partial_{\Lambda_i} \psi(\boldsymbol{\sigma})}{\psi(\boldsymbol{\sigma})} \tag{B5b}$$

Finally, we use gradient descent with learning rate λ,

$$\Lambda_i \to \Lambda_i - \lambda \partial_{\Lambda_i} \langle E \rangle, \tag{B6}$$

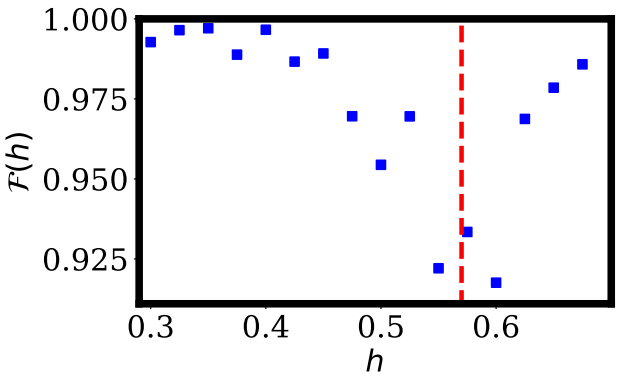

Figure 10. Fidelity $\mathcal{F}$ as a function of field h. The red dashed line is drawn to guide the eye, where the dip in fidelity indicates the critical field value $h_c \simeq 0.57$. This field value roughly agrees with what the DM algorithm identifies in Fig. 6.

to minimize the energy expectation value. The gradient descent is performed by using an adaptive Adam optimizer [79]. We repeat this training step until empirical convergence.

Note that the RBM ansatz can get stuck in local minima. To find the toric code ground state, we initialize the network parameters close to the analytic solutions in Eq. (A5).

1. Fidelity

To find the approximate ground states at finite field values h with step size Δh, we initialize the weights to be those from the previous field value $h - \Delta h$, and then use the current optimized weights as the initialization for the next step $h + \Delta h$. A good indication of a quantum phase transition is by inspecting the fidelity $\mathcal{F}(h)$ defined as,

$$\mathcal{F}(h) = |\langle \psi(h) | \psi(h + \Delta h) \rangle|^2. \tag{B7}$$

The critical field h_c is identified as a dip in the fidelity, indicating an abrupt change in the ground state wavefunction. A field value of $h_c \simeq 0.57$ (at dashed line in Fig. 10) is found for the RBM ansatz. Note that one can get more accurate field value by including loop expectations in the ansatz as done in Ref. 60.

Appendix C: Ensemble generation

Using the algorithm outlined under Algorithm 1, we can generate physical ensembles characterized by parameter $T = 0.1, 0.3, 1$, starting from the initial seeds via VMC optimization in Sec. B. The other choices of parameters for the ensembles are number of independent chains $k = 2$, length of each chain $n = 250$, and number

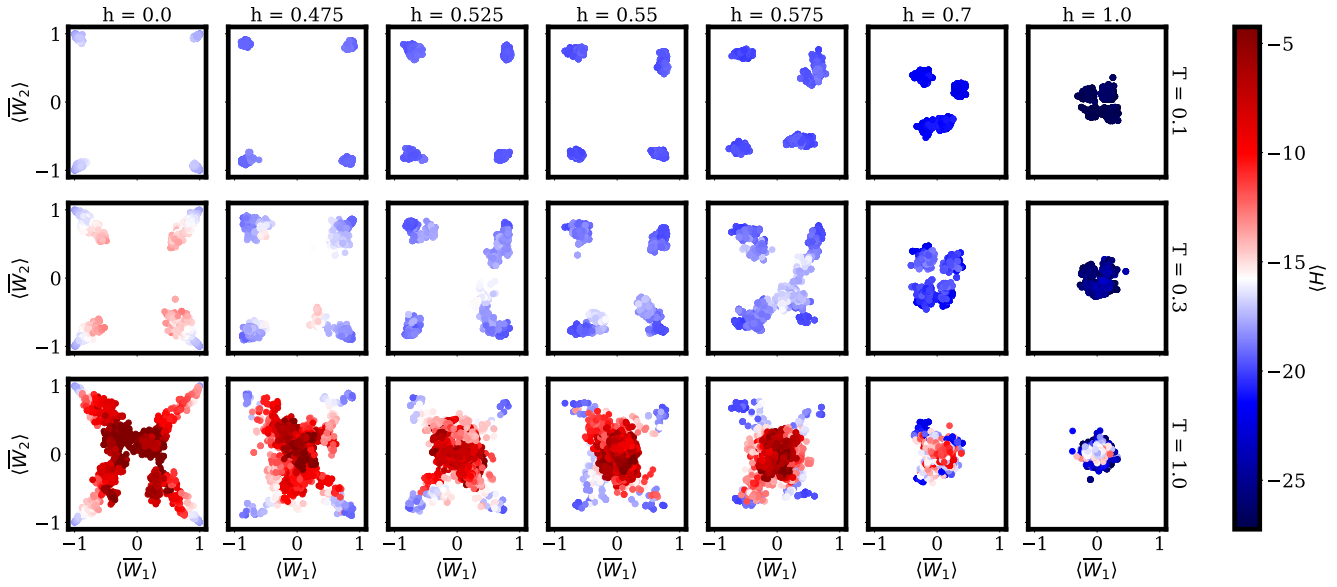

Figure 11. Illustration of the diffusion processes for different ensemble parameter T and field h at $N = 18$ spins. The loop expectation values $\langle \overline{W}_{1,2} \rangle$ form four distinct clusters in the two-dimensional plane for small T and h. For large $T = 1.$ at all fields and intermediate $T = 0.3$ at higher fields $h > 0.57$, the clusters "diffuse" and topological order is lost. Such "diffusion" process can be visualized by color coding the energy expectation $\langle H \rangle$.

of samples kept $m = n$. The parameter proposal function we use consists of with probability p_m randomly apply minus sign or randomly adding local noise at a single spin site $\jmath$. More precisely,

$$f(\Lambda, \xi) = \begin{cases} f_{-,\jmath}, & \text{with probability} : p_m, \\ f_{\text{local},\jmath}, & \text{with probability} : 1 - p_m, \end{cases} \tag{C1a}$$

$$f_{-,\jmath} = \begin{cases} -(\Lambda)_i, & i \in \jmath \\ (\Lambda)_i, & i \notin \jmath \end{cases} \tag{C1b}$$

$$f_{\text{local},\jmath} = \begin{cases} \text{uniform}(0, \xi) + (\Lambda)_{\text{i}}, & \text{i} \in \jmath \\ (\Lambda)_i, & i \notin \jmath \end{cases} \tag{C1c}$$

In the exact toric code state, $f_{-,\jmath}$ corresponds to act σ_x operator at site $\jmath$ to create a pair of m-particles. In the trivial phase, depending on the parametrization of the state, $f_{-,\jmath}$ could correspond to a single spin flip at site $\jmath$. The hyperparameters are chosen to be $p_m = 0.3$ and $\xi = 0.2$. In Fig. 11, we visualize the ensembles by computing their loop expectations $\langle \overline{W}_j \rangle$ at different field values.

[1] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab, "A high-bias, low-variance introduction to Machine Learning for physicists," Physics Reports A High-Bias, Low-Variance Introduction to Machine Learning for Physicists, **810**, 1–124 (2019).

[2] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová, "Machine learning and the physical sciences," Rev. Mod. Phys. **91**, 045002 (2019).

[3] Sankar Das Sarma, Dong-Ling Deng, and Lu-Ming Duan, "Machine learning meets quantum physics," Physics Today **72**, 48–54 (2019), arXiv:1903.03516 [physics.pop-ph].

[4] Roger G. Melko, Giuseppe Carleo, Juan Carrasquilla, and J. Ignacio Cirac, "Restricted boltzmann machines in quantum physics," Nature Physics **15**, 887–892 (2019).

[5] Juan Carrasquilla, "Machine learning for quantum matter," Advances in Physics: X **5**, 1797528 (2020).

[6] Juan Carrasquilla and Giacomo Torlai, "How To Use Neural Networks To Investigate Quantum Many-Body Physics," PRX Quantum **2**, 040201 (2021).

[7] Anna Dawid, Julian Arnold, Borja Requena, Alexander Gresch, Marcin Płodzień, Kaelan Donatella, Kim A. Nicoli, Paolo Stornati, Rouven Koch, Miriam Büttner, Robert Okuła, Gorka Muñoz-Gil, Rodrigo A. Vargas-Hernández, Alba Cervera-Lierta, Juan Carrasquilla, Vedran Dunjko, Marylou Gabrié, Patrick Huembeli, Evert van Nieuwenburg, Filippo Vicentini, Lei Wang, Sebastian J. Wetzel, Giuseppe Carleo, Eliška Greplová, Roman Krems, Florian Marquardt, Michał Tomza, Maciej Lewenstein, and Alexandre Dauphin, "Modern applications of machine learning in quantum sciences," (2022), arXiv:2204.04198 [cond-mat, physics:quant-ph].

[8] Juan Carrasquilla and Roger G. Melko, "Machine learning phases of matter," Nature Physics **13**, 431–434 (2017).

[9] Pengfei Zhang, Huitao Shen, and Hui Zhai, "Machine learning topological invariants with neural networks," Phys. Rev. Lett. **120**, 066401 (2018).

[10] Yi Zhang and Eun-Ah Kim, "Quantum Loop Topography for Machine Learning," Phys. Rev. Lett. **118**, 216401 (2017).

[11] Matthew J. S. Beach, Anna Golubeva, and Roger G. Melko, "Machine learning vortices at the kosterlitz-thouless transition," Phys. Rev. B **97**, 045207 (2018).

[12] Joaquin F. Rodriguez-Nieva and Mathias S. Scheurer, "Identifying topological order through unsupervised machine learning," Nature Physics **15**, 790–795 (2019).

[13] Japneet Singh, Mathias S. Scheurer, and Vipul Arora, "Conditional generative models for sampling and phase transition indication in spin systems," SciPost Phys. **11**, 043 (2021).

[14] Y.-H. Tseng and F.-J. Jiang, "Berezinskii–kosterlitz–thouless transition – a universal neural network study with benchmarks," Results in Physics **33**, 105134 (2022).

[15] Eliska Greplova, Agnes Valenti, Gregor Boschung, Frank Schäfer, Niels Lörch, and Sebastian D Huber, "Unsupervised identification of topological phase transitions using predictive models," New J. Phys. **22**, 045003 (2020).

[16] Yi Zhang, Roger G. Melko, and Eun-Ah Kim, "Machine learning F_2 quantum spin liquids with quasiparticle statistics," Phys. Rev. B **96**, 245119 (2017).

[17] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill, "Provably efficient machine learning for quantum many-body problems," Science **377**, eabk3333 (2022).

[18] Nicolas Sadoune, Giuliano Giudici, Ke Liu, and Lode Pollet, "Unsupervised Interpretable Learning of Phases From Many-Qubit Systems," (2022), arXiv:2208.08850 [cond-mat, physics:quant-ph].

[19] Alex Cole, Gregory J. Loges, and Gary Shiu, "Interpretable Phase Detection and Classification with Persistent Homology," arXiv e-prints , arXiv:2012.00783 (2020), arXiv:2012.00783 [cond-mat.stat-mech].

[20] Dan Sehayek and Roger G. Melko, "Persistent Homology of $\mathbb{Z}_2$ Gauge Theories," Phys. Rev. B **106**, 085111 (2022), arXiv:2201.09856 [cond-mat, physics:hep-th].

[21] Niklas Käming, Anna Dawid, Korbinian Kottmann, Maciej Lewenstein, Klaus Sengstock, Alexandre Dauphin, and Christof Weitenberg, "Unsupervised machine learning of topological phase transitions from experimental data," Machine Learning: Science and Technology **2**, 035037 (2021).

[22] Chi-Ting Ho and Daw-Wei Wang, "Robust identification of topological phase transition by self-supervised machine learning approach," New Journal of Physics **23**, 083021 (2021).

[23] Min-Ruei Lin, Wan-Ju Li, and Shin-Ming Huang, "Quaternion-based machine learning on topological quantum systems," arXiv e-prints (2022), arXiv:2209.14551 [quant-ph].

[24] Gilad Margalit, Omri Lesser, T. Pereg-Barnea, and Yuval Oreg, "Renormalization-group-inspired neural networks for computing topological invariants," Phys. Rev. B **105**, 205139 (2022).

[25] Sungjoon Park, Yoonseok Hwang, and Bohm-Jung Yang, "Unsupervised learning of topological phase diagram using topological data analysis," Phys. Rev. B **105**, 195115 (2022).

[26] Ming-Chiang Chung, Tsung-Pao Cheng, Guang-Yu Huang, and Yuan-Hong Tsai, "Deep learning of topological phase transitions from the point of view of entanglement for two-dimensional chiral p-wave superconductors," Phys. Rev. B **104**, 024506 (2021).

[27] Yuan-Hong Tsai, Kuo-Feng Chiu, Yong-Cheng Lai, Kuan-Jung Su, Tzu-Pei Yang, Tsung-Pao Cheng, Guang-Yu Huang, and Ming-Chiang Chung, "Deep learning of topological phase transitions from entanglement aspects: An unsupervised way," Phys. Rev. B **104**, 165108 (2021).

[28] Alejandro José Uría-Álvarez, Daniel Molpeceres-Mingo, and Juan José Palacios, "Deep learning for disordered topological insulators through entanglement spectrum," arXiv e-prints , arXiv:2201.13306 (2022), arXiv:2201.13306 [cond-mat.dis-nn].

[29] Paolo Molignini, Antonio Zegarra, Evert van Nieuwenburg, R. Chitra, and Wei Chen, "A supervised learning algorithm for interacting topological insulators based on local curvature," SciPost Phys. **11**, 073 (2021).

[30] Simone Tibaldi, Giuseppe Magnifico, Davide Vodola, and Elisa Ercolessi, "Unsupervised and supervised learning of interacting topological phases from single-particle correlation functions," arXiv e-prints (2022), arXiv:2202.09281 [cond-mat.supr-con].

[31] Andrea Tirelli and Natanael C. Costa, "Learning quantum phase transitions through topological data analysis," Phys. Rev. B **104**, 235146 (2021).

[32] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," Proceedings of the National Academy of Sciences **102**, 7426–7431 (2005).

[33] Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald Coifman, "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators," in *Advances in Neural Information Processing Systems*, Vol. 18 (MIT Press, 2005).

[34] Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," Applied and Computational Harmonic Analysis **21**, 113–127 (2006).

[35] Ronald R. Coifman and Stéphane Lafon, "Diffusion

maps," Applied and Computational Harmonic Analysis **21**, 5–30 (2006).
[36] Mathias S. Scheurer and Robert-Jan Slager, "Unsupervised Machine Learning and Band Topology," Phys. Rev. Lett. **124**, 226401 (2020).
[37] Yang Long, Jie Ren, and Hong Chen, "Unsupervised Manifold Clustering of Topological Phononics," Phys. Rev. Lett. **124**, 185501 (2020).
[38] Li-Wei Yu and Dong-Ling Deng, "Unsupervised learning of non-hermitian topological phases," Phys. Rev. Lett. **126**, 240402 (2021).
[39] Yefei Yu, Li-Wei Yu, Wengang Zhang, Huili Zhang, Xiaolong Ouyang, Yanqing Liu, Dong-Ling Deng, and L. M. Duan, "Experimental unsupervised learning of non-Hermitian knotted phases with solid-state spins," npj Quantum Information **8**, 116 (2022), arXiv:2112.13785 [quant-ph].
[40] Yanming Che, Clemens Gneiting, Tao Liu, and Franco Nori, "Topological quantum phase transitions retrieved through unsupervised machine learning," Phys. Rev. B **102**, 134213 (2020).
[41] En-Jui Kuo and Hossein Dehghani, "Unsupervised Learning of Symmetry Protected Topological Phase Transitions," arXiv:2111.08747 [cond-mat, physics:quant-ph] (2021), arXiv:2111.08747 [cond-mat, physics:quant-ph].
[42] Eran Lustig, Or Yair, Ronen Talmon, and Mordechai Segev, "Identifying Topological Phase Transitions in Experiments Using Manifold Learning," Phys. Rev. Lett. **125**, 127401 (2020).
[43] Alexander Lidiak and Zhexuan Gong, "Unsupervised Machine Learning of Quantum Phase Transitions Using Diffusion Maps," Phys. Rev. Lett. **125**, 225701 (2020).
[44] Gaurav Gyawali, Mabrur Ahmed, Eric Aspling, Luke Ellert-Beck, and Michael J. Lawler, "Revealing microcanonical phase diagrams of strongly correlated systems via time-averaged classical shadows," (2022), arXiv:2211.01259 [cond-mat, physics:quant-ph].
[45] Apimuk Sornsaeng, Ninnat Dangniam, Pantita Palittapongarnpim, and Thiparat Chotibut, "Quantum diffusion map for nonlinear dimensionality reduction," Phys. Rev. A **104**, 052410 (2021).
[46] Giuseppe Carleo and Matthias Troyer, "Solving the quantum many-body problem with artificial neural networks," Science **355**, 602–606 (2017).
[47] Xun Gao and Lu-Ming Duan, "Efficient representation of quantum many-body states with deep neural networks," Nature Communications **8**, 662 (2017).
[48] Giuseppe Carleo, Yusuke Nomura, and Masatoshi Imada, "Constructing exact representations of quantum many-body systems with deep neural networks," Nat Commun **9**, 5322 (2018).
[49] Sirui Lu, Xun Gao, and L.-M. Duan, "Efficient representation of topologically ordered states with restricted Boltzmann machines," Phys. Rev. B **99**, 155136 (2019).
[50] Or Sharir, Amnon Shashua, and Giuseppe Carleo, "Neural tensor contractions and the expressive power of deep neural quantum states," (2021), 10.48550/arXiv.2103.10293.
[51] Jing Chen, Song Cheng, Haidong Xie, Lei Wang, and Tao Xiang, "Equivalence of restricted Boltzmann machines and tensor network states," Phys. Rev. B **97**, 085104 (2018).
[52] Yusuke Nomura, "Investigating Network Parameters in Neural-Network Quantum States," (2022), arXiv:2202.01704 [cond-mat, physics:physics, physics:quant-ph].
[53] Dong-Ling Deng, Xiaopeng Li, and S Das Sarma, "Quantum Entanglement in Neural Network States," , 17 (2017).
[54] Zhih-Ahn Jia, Lu Wei, Yu-Chun Wu, Guang-Can Guo, and Guo-Ping Guo, "Entanglement area law for shallow and deep quantum neural network states," New J. Phys. **22**, 053022 (2020).
[55] Song Cheng, Jing Chen, and Lei Wang, *Information Perspective to Probabilistic Modeling: Boltzmann Machines versus Born Machines*, Tech. Rep. (2017) arXiv:1712.04144 [cond-mat, physics:physics, physics:quant-ph, stat].
[56] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo, "Neural-network quantum state tomography," Nature Physics **14**, 447–450 (2018).
[57] David E Rumelhart, James L McClelland, PDP Research Group, *et al.*, "Parallel distributed processing," Foundations **1** (1988).
[58] A. Yu Kitaev, "Fault-tolerant quantum computation by anyons," Annals of Physics **303**, 2–30 (2003), arXiv:quant-ph/9707021.
[59] Dong-Ling Deng, Xiaopeng Li, and S. Das Sarma, "Machine learning topological states," Phys. Rev. B **96**, 195145 (2017).
[60] Agnes Valenti, Eliska Greplova, Netanel H. Lindner, and Sebastian D. Huber, "Correlation-Enhanced Neural Networks as Interpretable Variational Quantum States," (2021), arXiv:2103.05017 [cond-mat, physics:quant-ph].
[61] Ze-Pei Cian, Mohammad Hafezi, and Maissam Barkeshli, "Extracting Wilson loop operators and fractional statistics from a single bulk ground state," arXiv e-prints (2022), arXiv:2209.14302 [cond-mat.str-el].
[62] M. B. Hastings, "An area law for one-dimensional quantum systems," J. Stat. Mech. **2007**, P08024–P08024 (2007).
[63] F. Verstraete, M. M. Wolf, D. Perez-Garcia, and J. I. Cirac, "Criticality, the Area Law, and the Computational Power of Projected Entangled Pair States," Phys. Rev. Lett. **96**, 220601 (2006).
[64] Michael M. Wolf, Frank Verstraete, Matthew B. Hastings, and J. Ignacio Cirac, "Area Laws in Quantum Systems: Mutual Information and Correlations," Physical Review Letters **100** (2008), 10.1103/PhysRevLett.100.070502.
[65] J. Eisert, M. Cramer, and M. B. Plenio, "*Colloquium* : Area laws for the entanglement entropy," Rev. Mod. Phys. **82**, 277–306 (2010).
[66] Note that if there are identical samples in the dataset, then there will also be additional 0 eigenvalues in the limit $\epsilon \to 0$.
[67] Jing-Ling Chen, Libin Fu, Abraham A. Ungar, and Xian-Geng Zhao, "Alternative fidelity measure between two states of an N -state quantum system," Phys. Rev. A **65**, 054304 (2002).
[68] Paulo E. M. F. Mendonça, Reginaldo d. J. Napolitano, Marcelo A. Marchiolli, Christopher J. Foster, and Yeong-Cherng Liang, "Alternative fidelity measure between quantum states," Phys. Rev. A **78**, 052330 (2008).
[69] Zbigniew Puchała and Jarosław Adam Miszczak, "Bound on trace distance based on superfidelity," Phys. Rev. A

79, 024302 (2009).
[70] J. A. Miszczak, Z. Puchała, P. Horodecki, A. Uhlmann, and K. Zyczkowski, "Sub- and super-fidelity as bounds for quantum fidelity," (2008), arXiv:0805.2037 [quant-ph].
[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research **12**, 2825–2830 (2011).
[72] I. S. Tupitsyn, A. Kitaev, N. V. Prokof'ev, and P. C. E. Stamp, "Topological multicritical point in the phase diagram of the toric code model and three-dimensional lattice gauge Higgs model," Phys. Rev. B **82**, 085114 (2010).
[73] Simon Trebst, Philipp Werner, Matthias Troyer, Kirill Shtengel, and Chetan Nayak, "Breakdown of a Topological Phase: Quantum Phase Transition in a Loop Gas Model with Tension," Phys. Rev. Lett. **98**, 070602 (2007).
[74] Fengcheng Wu, Youjin Deng, and Nikolay Prokof'ev, "Phase diagram of the toric code model in a parallel magnetic field," Phys. Rev. B **85**, 195104 (2012).
[75] Michael Schuler, Seth Whitsitt, Louis-Paul Henry, Subir Sachdev, and Andreas M. Läuchli, "Universal Signatures of Quantum Critical Points from Finite-Size Torus Spectra: A Window into the Operator Content of Higher-Dimensional Conformal Field Theories," Phys. Rev. Lett. **117**, 210401 (2016), arXiv:1603.03042 [cond-mat.str-el].
[76] To separate $2(n)$ clusters, one need to look at $1(n-1)$-dimensional subspace.
[77] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang, "JAX: composable transformations of Python+NumPy programs," (2018).
[78] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller, "Equation of State Calculations by Fast Computing Machines," J. Chem. Phys. **21**, 1087–1092 (1953).
[79] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).