



OPEN ACCESS

RECEIVED

25 November 2019

REVISED

2 February 2020

ACCEPTED FOR PUBLICATION

18 February 2020

PUBLISHED

7 April 2020

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



PAPER

Unsupervised identification of topological phase transitions using predictive models

Eliska Greplova¹, Agnes Valenti¹, Gregor Boschung¹, Frank Schäfer², Niels Lörch² and Sebastian D Huber¹¹ Institute for Theoretical Physics, ETH Zurich, CH-8093, Switzerland² Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, SwitzerlandE-mail: geliska@phys.ethz.ch**Keywords:** topological phase transitions, unsupervised learning, quantum phase transitions, topological order, Ising gauge theory, toric code

Abstract

Machine-learning driven models have proven to be powerful tools for the identification of phases of matter. In particular, unsupervised methods hold the promise to help discover new phases of matter without the need for any prior theoretical knowledge. While for phases characterized by a broken symmetry, the use of unsupervised methods has proven to be successful, topological phases without a local order parameter seem to be much harder to identify without supervision. Here, we use an unsupervised approach to identify boundaries of the topological phases. We train artificial neural nets to relate configurational data or measurement outcomes to quantities like temperature or tuning parameters in the Hamiltonian. The accuracy of these predictive models can then serve as an indicator for phase transitions. We successfully illustrate this approach on both the classical Ising gauge theory as well as on the quantum ground state of a generalized toric code.

1. Introduction

Identifying phase transitions is one of the key questions in theoretical and experimental condensed matter physics alike. For the experimental characterization of thermodynamic phase transitions, there exists an excessive amount of possible tools, ranging from system specific, like the study of the conductivity in an electronic system, to very generic, like the specific heat. The latter is particularly appealing as it does not assume any prior knowledge: for example, structural transitions, the onset of magnetism, or the transition to superconductivity, all show up in this generic probe. The study of the specific heat is also a standard tool for the theoretician, especially given its generic power.

For quantum phase transitions [1], an equally generic tool as the specific heat for thermal transitions is the fidelity susceptibility. One investigates the derivative of the overlap $\partial_\beta \langle \psi(\beta + \epsilon) | \psi(\beta) \rangle$ [2] of two infinitesimally separated ground states $|\psi(\beta)\rangle$ as a function of some tuning parameter β . While this probe is in principle very powerful [3–6], it is typically hard to evaluate as one has rarely access to the full wave-function. At least not for most of the approximate numerical techniques and especially not in experimental studies. This raises the question if one can replace the fidelity susceptibility with a tool that is equally *unbiased*, *generic*, and *accessible* to typical numerical and experimental techniques.

In a recent publication some of the present authors introduced such an algorithmic method for classical systems with an order-parameter signaling an (arbitrary) symmetry breaking [7]. Here we demonstrate that one can successfully generalize this method to problems without a local order parameter, i.e. systems with a topological character. Moreover, we show that one can straightforwardly extend [7] to the quantum realm.

The method is based on the analysis of the accuracy of a predictive model. The central idea is to distill a predictive model that relates input data from numerical or experimental studies to the output in the form of a known tuning parameter such as the temperature or a parameter in the Hamiltonian β . Typically, one infers this predictive model via machine-learning techniques in the form of neural nets. The basic idea, however, is independent of the specific inference technique. In a next step, the accuracy of the predictive model is analyzed via the comparison of the predicted to the known value of the tuning parameter β . In particular, we show the

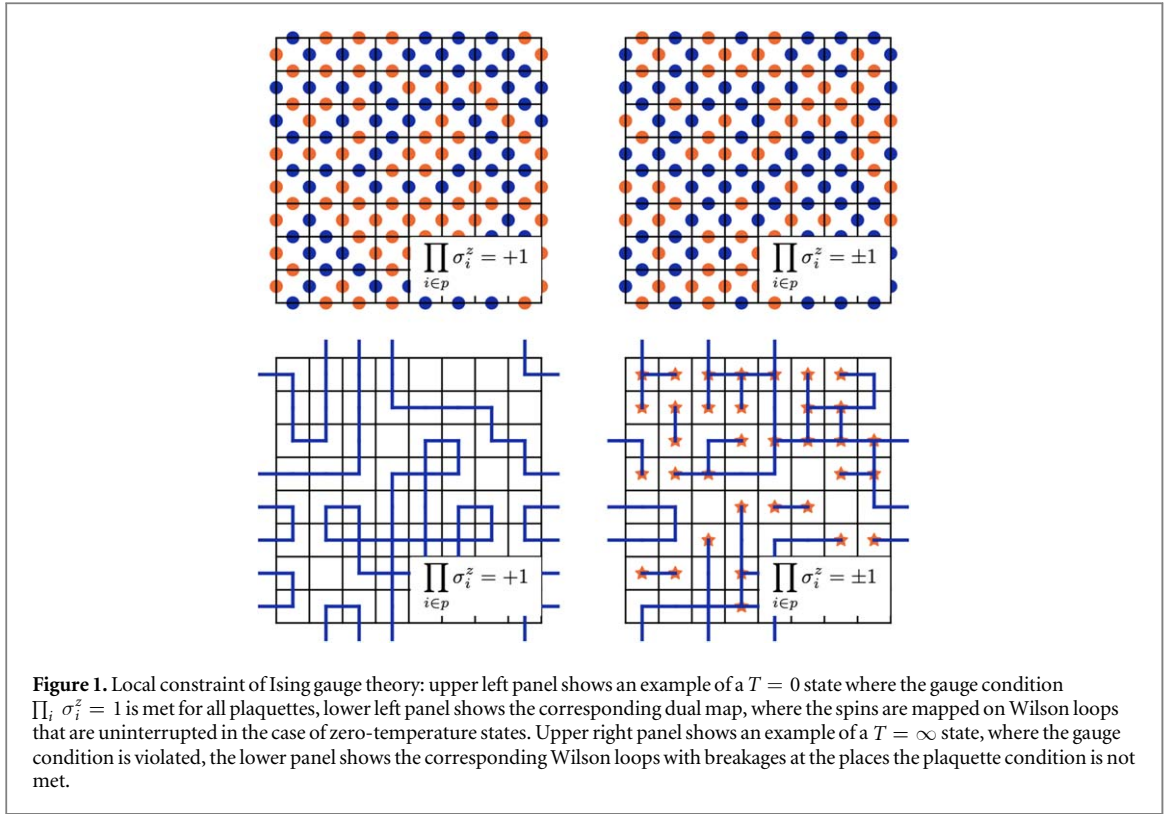


Figure 1. Local constraint of Ising gauge theory: upper left panel shows an example of a $T = 0$ state where the gauge condition $\prod_{i \in p} \sigma_i^z = 1$ is met for all plaquettes, lower left panel shows the corresponding dual map, where the spins are mapped on Wilson loops that are uninterrupted in the case of zero-temperature states. Upper right panel shows an example of a $T = \infty$ state, where the gauge condition is violated, the lower panel shows the corresponding Wilson loops with breakages at the places the plaquette condition is not met.

derivative of the prediction accuracy with respect to the tuning parameter to be an equally sensitive indicator of a phase transition as the fidelity susceptibility.

To illustrate our generalization of the methods of [7], we investigate two generic models hosting interesting thermodynamic phases without a local order parameter. First, we investigate the finite-temperature cross-over in Wegner's Ising gauge theory (IGT) [8–10] to show that we can analyze an interesting classical problem without a local order parameter. Second, we broaden the scope by taking the step from the IGT to a generalized toric code problem [11, 12] showcasing the applicability of the method to quantum problems.

2. The Ising gauge theory

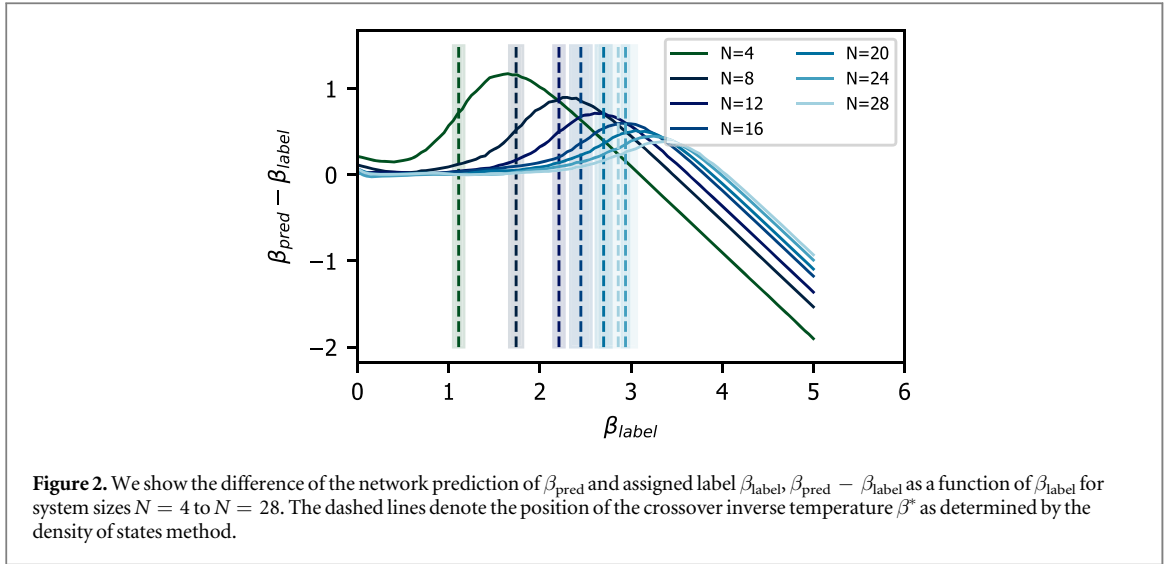
Wegner's Ising gauge theory (IGT) is a spin model defined on a $N \times N$ square lattice with spins placed on the lattice bonds [8–10, 13]. It is described by the Hamiltonian

$$H_{\text{IGT}} = -J \sum_p \prod_{i \in p} \sigma_i^z, \quad (1)$$

where J is a coupling constant, p refers to plaquettes on the lattice (see figure 1), and σ_i^z is the Pauli matrix describing a single spin-1/2. Periodic boundary conditions are imposed. The ground state of this Hamiltonian is a highly degenerate manifold, an arbitrary superposition of all states that meet the condition that the product of spins along each plaquette is equal to 1. At a finite temperature $T > 0$ the local constraints $\prod_{i \in p} \sigma_i^z = 1$ are violated (see figure 1). The IGT does not have a finite temperature phase transition. However, for finite system sizes one can find a crossover temperature, $T^* = 1/(k\beta^*)$ defined by the appearance of one plaquette with $\prod_{i \in p} \sigma_i^z = -1$, resulting in the scaling $T^* \sim 1/\ln(2N^2)$ [12, 14]. Matters are further complicated by the fact that the ground-state manifold cannot be characterized by a local order parameter [15, 10] owing to a local gauge degree of freedom. We come back to this point below.

To check whether a given spin state is in the IGT ground-state manifold, one has to verify that the condition $\prod_{i \in p} \sigma_i^z = 1$ is met for all plaquettes in the lattice. Equivalently, one can use the duality map to analyze the phase transition: we connect the edges of the lattice that contain spins with the same orientation and form loops. The IGT constrained phase then has the property that all these loops are closed. Whenever the constraint is violated it results in an open loop [10, 16, 13], see figure 1.

Distinguishing high and low temperature states of the model (1) is a well studied test case for machine learning recognition of phases of matter [14]. As one can see from figure 1, the IGT constitutes an interesting example where the phases are hard to distinguish visually without being *a priori* familiar with a local restrictions or the dual map. While an supervised approach is immediately successful at distinguishing the high and low



temperature phases [14], unsupervised approaches did not succeed without an explicit recipe what type of restriction to look at. There has been significant progress in this direction, but a fully general approach is yet to be found [17–24]. While methods like principal component analysis, clustering and variational auto-encoders have proven to be successful to determine the phase transitions in spin models possessing an order parameter [25], systems without order parameters still represent a challenge.

Here we show how the method introduced by Schäfer *et al* [7] can be generalized to systems without a local order parameter. One first pre-trains a neural network to relate a spin configuration $\{S\}_{\beta_{\text{label}}}$ to the (inverse) temperature β_{label} , at which the configuration was sampled. After this initial training, the performance of the estimator is assessed with respect to the true value. The derivative

$$\mathcal{D}(\beta_{\text{label}}) = \frac{\partial}{\partial \beta_{\text{label}}} \beta_{\text{pred}}(\{S\}_{\beta_{\text{label}}}) \quad (2)$$

is maximal where the estimator performs worst. In other words, a local maximum in $\mathcal{D}(\beta_{\text{label}})$ indicates a phase transition or cross-over temperature β_{label}^* . While this method does not in principle rely on a local order parameter, it has been shown that the network picks up on the magnetization pattern [7]. It was therefore unclear if one can generalize this strategy to the current problem. Here we show that this approach is valid even for phases of matter that do not contain an order parameter, or a finite temperature phase transition.

Our approach differs from prototypical unsupervised machine learning techniques, such as, e.g. principal component analysis, t-distributed stochastic neighbor embedding (t-SNE), or k -means clustering, since a fully supervised subroutine, namely a regression on the labeled system parameters, is employed. However, we intentionally refer to the approach as an unsupervised learning scheme, as the method aims ultimately to infer the phase diagram of the physical system and not its parameters and the algorithm thereby requires no prior knowledge of the phase labels, the number of different phases or character of the phase transition. In fact, the derivative (2) has generically a stronger signal when the parameters in the supervised part of the protocol are not learned up to high precision.

We create sample configurations of the IGT model and label them with $\beta = 1/(kT)$. We train a convolutional neural network to predict β given an IGT configuration as an input. Our neural network consists of 2 convolutional and 2 dense layers and was trained on 2×10^5 configurations for 100 different values of β (for details see appendix A).

In figure 2 we show how the difference between the true and predicted inverse temperatures $\beta_{\text{pred}} - \beta_{\text{label}}$ behaves as a function of the true β_{label} for seven different system sizes $N = 4, 8, 12, 16, 20, 24, 28$ (the total number of spins is $2N^2$). We see that the behavior of the prediction is not uniform for all inputs and, in fact, we observe that for all systems sizes there exists a different finite $\bar{\beta}$ above which the network has difficulties to identify the correct β_{label} . In figure 3 we show $\mathcal{D}(\beta_{\text{label}})$ which we evaluated as

$$\mathcal{D}(\beta_{\text{label}}) \approx \frac{\beta_{\text{pred}}(\{S\}_{\beta_{\text{label}}^{i+1}}) - \beta_{\text{pred}}(\{S\}_{\beta_{\text{label}}^{i-1}})}{\beta_{\text{label}}^{i+1} - \beta_{\text{label}}^{i-1}},$$

where sampled at discrete β_{label}^i . For all system sizes we observe a presence of a peak that indicates the position of the largest change in the difference between true and predicted β . The peak becomes less recognizable with increasing system size, which is consistent with the fact that the critical β^* keeps increasing with growing system size and in the infinite system limit the crossover behavior completely disappears.

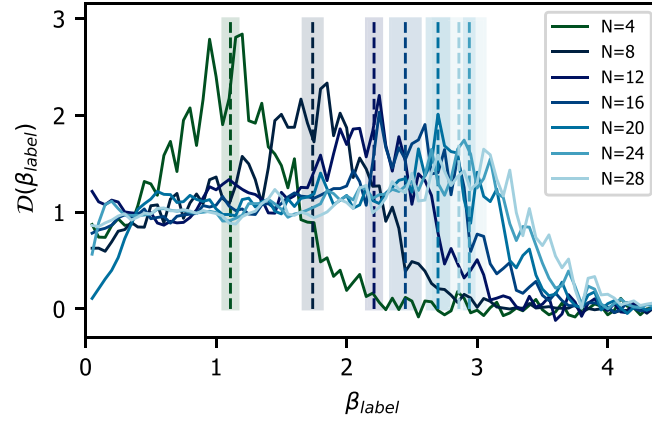


Figure 3. Derivative of the output of the predictive model, $\mathcal{D}(\beta_{\text{label}})$, as a function of assigned labels β_{label} for system sizes $N = 4$ to $N = 28$. The dashed lines denote the position of the crossover inverse temperature β^* as determined by the density of states method.

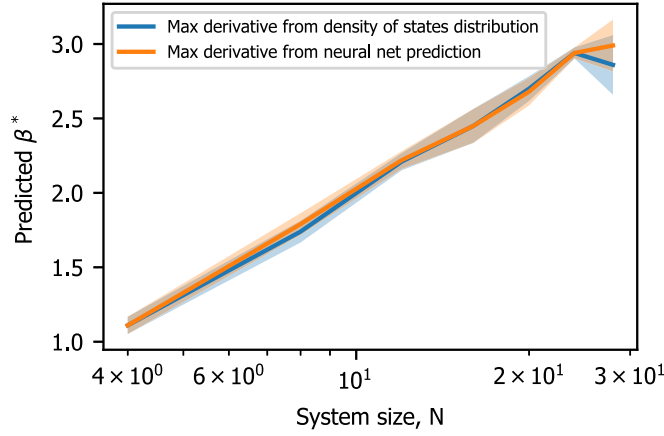


Figure 4. Positions of critical β^* as a function of a system size N . We show the scaling obtained from the unsupervised learning method and the scaling obtained from density of states in blue and orange respectively. The shaded areas represent the error bars. Error bars correspond to standard deviation from the mean $\tilde{\beta}^*$ evaluated by averaging over β^* predicted by five separately trained neural nets.

The neural network predicts a continuous parameter (inverse temperature) for our model and we observe a change of behavior at some critical value. We show in figure 4 the determined crossover temperature β^* as a function of system size. For the system sizes we were able to test numerically we recover logarithmic scaling as expected for the crossover temperature [14, 26].

To independently confirm the neural network predictions, we can analyze whether we can identify the physics of what the network is learning and reproduce its predictions by another physical model. From the training set, we can construct a density of states distribution, ϵ . In particular, the density of states can be written as a function of energy, E , and inverse temperature, β

$$\epsilon(\beta, E) = \frac{\sum_{n=1}^M \delta_{E, E_n} \delta_{\beta, \beta_n}}{\sum_{n=1}^M \delta_{\beta, \beta_n}}. \quad (3)$$

Here, $\delta_{a,b}$ is the Kronecker-delta symbol ($\delta_{a,b} = 1$ for $a = b$ and $\delta_{a,b} = 0$ for $a \neq b$), $E_n(\beta_n)$ is energy (label) of the n th configuration in the training set and M is the number of configurations in the training set. We show the distribution ϵ obtained for the system size $N = 8$ (128 spins) in figure 5.

We use the distribution (3) to calculate the most likely $\beta = \beta_{\text{pred}}$ for each configuration at a given energy, which immediately allows us to evaluate the relation between the assigned β and β_{pred} . Using the density of states we are able to reproduce the behavior in figure 3 (see appendix A). We show the detailed calculation and the dependencies of the predicted β_{pred} and its derivative $\mathcal{D}(\beta_{\text{label}})$ as a function of the true β_{label} in appendix A. This gives us a numerical evidence that the network is learning the density of states distribution shown in figure 5. We identify the logarithmic scaling (with system size) of the critical β^* predicted from the density of states (shown in blue in figure 4) analogously to the predictions obtained from the neural net model.

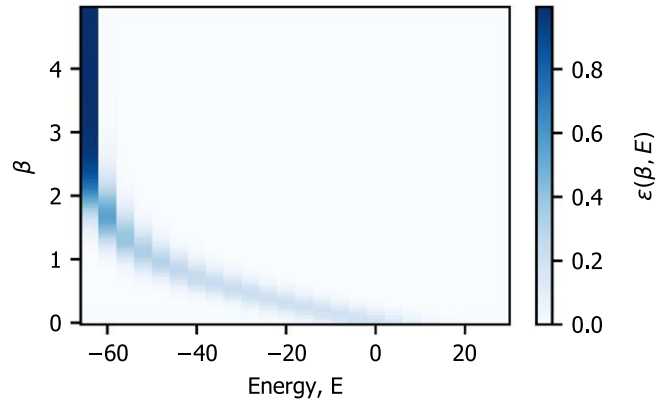


Figure 5. Density of states distribution $\epsilon(\beta, E)$ of the training set as a function of inverse temperature β and energy E . The plot above has been generated for system size $N = 8$.

Another unsupervised approach that has proven to be successful for both classical and quantum systems is the confusion scheme introduced in [18]. We compare both approaches in appendix C and show that the confusion scheme is not suitable for the example of IGT studied here.

3. The toric code and its generalizations

So far we have analyzed the performance of our method on the cross-over of a *classical* spin-1/2 model. When going to *quantum* models, two complications arise, related to the *input* and *output* of our predictive model. For classical systems, simple spin configurations are the natural input. For quantum systems, generically entanglement in the form of non-classically correlated configurations plays a key role. Consequently, the choice of training data needs to either reflect some prior knowledge of the system, or one has to sample over various classical projections of the entangled wave function. On the output side, one can either target a finite-temperature transition, or investigate a quantum phase transition at zero temperature. In the former, the output of the predictive model stays the same: β_{pred} , the inverse temperature. For zero temperature transitions, one can still investigate a single-parameter family of Hamiltonians $H(\beta)$. The obvious prediction task is then to reproduce the tuning parameter β , rather than the temperature.

We now turn to a concrete model of a quantum phase transition in a system without a local order parameter. The obvious generalization of (1) is the application of a transverse field [9, 13, 27]

$$H_{\text{TR}} = -\sum_p \prod_{i \in p} \sigma_i^z - g \sum_l \sigma_l^x. \quad (4)$$

The model above is very well studied, has a confinement-deconfinement transition at a critical g^* , and is a working horse for the study of \mathbb{Z}_2 spin-liquids. Instead of directly working with this simple model we go beyond (4) in two ways: (i) We restrict ourselves to a subset of gauge-invariant ground states by moving to the toric code [11]. (ii) We generalize the transverse field to allow for an exact solution. We detail both steps in the following.

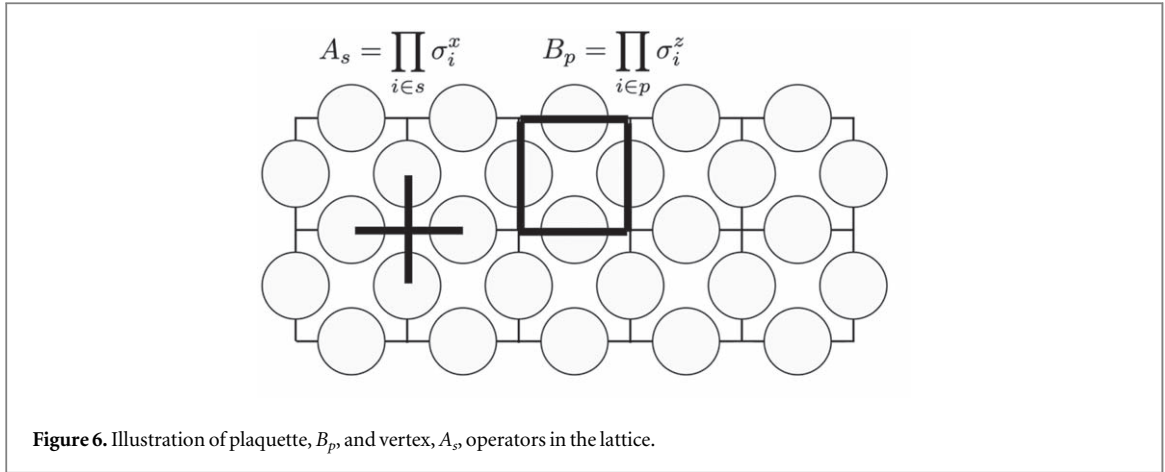
The IGT of equation (4) has a local \mathbb{Z}_2 gauge degree of freedom. The generators of this gauge transformation are the vertex operators

$$A_s = \prod_{i \in s} \sigma_i^x, \quad (5)$$

that consist of a product of σ_x operators along a vertex, s , of the lattice. The geometry of the vertex operator is illustrated in figure 6. The operators A_s commute with the Hamiltonian, i.e. $[H_{\text{TR}}, A_s] = 0$ for all vertices s . In other words, one can obtain an eigenstate by changing the sign of the classical σ_z -variables of another eigenstate, as long as one does so for all spins connected to one vertex. The toric code Hamiltonian

$$H_{\text{TC}} = -\sum_s A_s + H_{\text{IGT}} = -\sum_s A_s - \sum_p B_p, \quad (6)$$

elevates the generators of the gauge transformation to a term in the Hamiltonian. As a consequence, the ground states of the toric code correspond to the gauge-invariant ground states of H_{TR} [27]. For our numerical purposes below, we largely benefit from the exact solution of the above Hamiltonian: we can write one of the four (un-normalized) ground states as [28]



$$|\text{TC}\rangle = \frac{1}{2} \prod_s (1 + B_p) |0_x\rangle, \quad (7)$$

where $|0_x\rangle$ is a reference state with all spins up in the σ^x basis. Then, applying products of Pauli z -matrices along the two non-contractible loops yields the other three orthogonal ground states. We can easily see that the ground states are indeed gauge invariant by applying gauge transformations, obtaining $A_s |\text{TC}\rangle = |\text{TC}\rangle$.

Applying a transverse field a spin-model typically excludes an exact solution. The present case is no difference. However, in a recent publication, Chamon and Castelnovo introduced the following generalization of the toric code [12, 26, 29, 30]

$$H = H_{\text{TC}} + \sum_s e^{-\beta \sum_{i \in p} \lambda_i \sigma_i^x} = -\sum_s A_s + \sum_p \left(-B_p + \sum_s e^{-\beta \sum_{i \in p} \lambda_i \sigma_i^x} \right), \quad (8)$$

where $\lambda_i \in [-1, 1]$ describes the particular configuration of added background fields and $\beta > 0$ characterizes their amplitude. A transition to a topologically trivial phase occurs at a critical value of the field strength β_c . The field configuration λ_i influences the critical value β_c . A detailed analysis of this phase transition has been provided in [30].

To finish our discussion of these exactly solvable models we write the ground state of (8)

$$|\Psi\rangle = \frac{1}{\sqrt{Z}} e^{\beta \sum_i \lambda_i \sigma_i^x} |\text{TC}\rangle = \frac{1}{\sqrt{Z}} \sum_{h \in H} e^{\beta \sum_i \lambda_i \sigma_i^x(h)} h |0_x\rangle. \quad (9)$$

This ground state is four-fold degenerate when periodic boundary conditions are considered [28]. We denote with H the abelian group whose elements h are all possible operations defined by the action of products of plaquette operators on an initial (reference) spin-configuration $|0_x\rangle$. By $\sigma_i^x(h)$ we denote the eigenvalue of the operator σ_i^x on the eigenstate $h|0_x\rangle$. As a consequence, the term $\sigma_i^x(h)$ can take the values ± 1 . The normalization factor, Z corresponds to the partition function for this ground state and is given by

$$Z := \sum_{h \in H} e^{\beta \sum_i \lambda_i \sigma_i^x(h)}.$$

With these considerations we are now in the position to show that the analysis of the predictive model can point out the topological phase transition of this quantum model as well. Unlike in the IGT, discussed in the previous section, the highly entangled ground states of the modified toric code model (8) are not fully characterized by a spin configuration alone. On the other hand, equation (9) provides a closed analytical form for the ground states of the family of the Hamiltonians (8). In addition to that, these ground states are only four-fold degenerate in the topological phase. We take advantage of the knowledge of the modified toric code ground states and show this to be sufficient for identification of the phase transition from the predictive model.

3.1. Projection onto spin configurations

We consider a projection of the ground states of the Hamiltonian (8) onto the σ_x and σ_z bases. These two types of projections correspond to experimentally accessible measurements and we show that both allow us to detect the topological phase transition of the full quantum model. As for the IGT cross-over analyzed previously, we are yet again in the situation where we are able to input a configuration into the predictive model and ask it to predict a continuous parameter.

There are two crucial differences here: first, we are considering a zero temperature topological phase transition that is driven by the applied field strength β . The second difference lies in the behavior of the projected spin configurations in the two phases. In particular, we are able to draw parallels to phase transitions of classical

spin models. As we elaborate below, choosing a basis to project on corresponds to mapping the phases of the quantum model to phases of a specific classical spin model.

3.1.1. The σ_x -projection

Let us first consider the projection onto the σ_x basis. We notice that the ground state (9) represents a superposition of x -spin-configurations $|S_h\rangle := h|0\rangle_x$ for all elements of the group H . All states $|S_h\rangle$ fulfill the so-called closed loop condition

$$A_s|S_h\rangle = |S_h\rangle \quad (10)$$

for all values of β . In connection to the IGT, this corresponds to the condition of gauge invariance. More concretely, local constraints are imposed, that the product of σ_x eigenvalues around a vertex is equal to one. The value of the field strength, β , influences the weight of a given spin configuration (see equation (9)). Therefore, the probability to obtain a particular configuration $|S_h\rangle$ after projection onto σ_x -basis is given by

$$p(S_h) = |\langle S_h|\Psi(\beta)\rangle|^2 = \frac{e^{\beta \sum_i \lambda_i \sigma_i^x(h)}}{\sum_{\tilde{h} \in H} e^{\beta \sum_i \lambda_i \sigma_i^x(\tilde{h})}}. \quad (11)$$

We can understand the physics of the σ_x -projected ground state by first considering limiting cases of the field strength β . When $\beta \rightarrow 0$, the ground state (9) corresponds to the ground state of the pure toric code Hamiltonian (6). Therefore, when projected onto the σ_x basis, all possible $|S_h\rangle$ are equally likely (since all $|S_h\rangle$ are weighted equally in the full eigenstate). When $\beta \rightarrow \infty$, on the other hand, all configurations but $|S\rangle = |0\rangle$ are exponentially suppressed and hence, the projected spin configurations are always ordered.

Thus, what used to be a topological phase transition of the full quantum state is now a transition from disordered spin-configurations (β small) to an ordered spin-configuration (β large, all spins up). We observe that, provided there is a finite β at which the transition between ordered and disordered configurations manifests itself, we obtained a phase transition that shows resemblance to the phase transition of the 2D Ising model. We show that indeed the 2D Ising model and its phase transition can be recovered by a simple change of variables, see appendix C.

Let us now explore the topological phase transition in the toric code model using the unsupervised learning method we introduced above. We train a neural network on the projected σ_x configurations labeled with the field strength, β . We used a network consisting of two convolutional (100 filters, kernel size 3 and 2), one dense layer with 100 neurons and one dropout layer with dropout rate 0.15. We train the neural network on 59950 configurations containing 100 different values of β between 0 and 1. All the simulations were performed for the system size $N = 20$ (800 spins). Once the model is trained we apply it on 2000 new configurations for 30 different values of β and evaluate the derivative $\mathcal{D}(\beta_{\text{label}})$ (2) of the outcome, see figure 7. We show $\mathcal{D}(\beta_{\text{label}})$ for six example field configurations. The field configurations correspond to different distributions $\{\lambda_i\}$ and are detailed in appendix C. As shown in [30] the position and the existence of the phase transition is strongly dependent on the distribution $\{\lambda_i\}$ of added fields.

It was shown in [26] that the topological phase transition of the generalized toric code model can be determined from the behavior of the fidelity between two ground states with slightly varied field strengths ($\delta\beta \rightarrow 0$)

$$F_\beta = \langle \Psi(\beta) | \Psi(\beta + \delta\beta) \rangle. \quad (12)$$

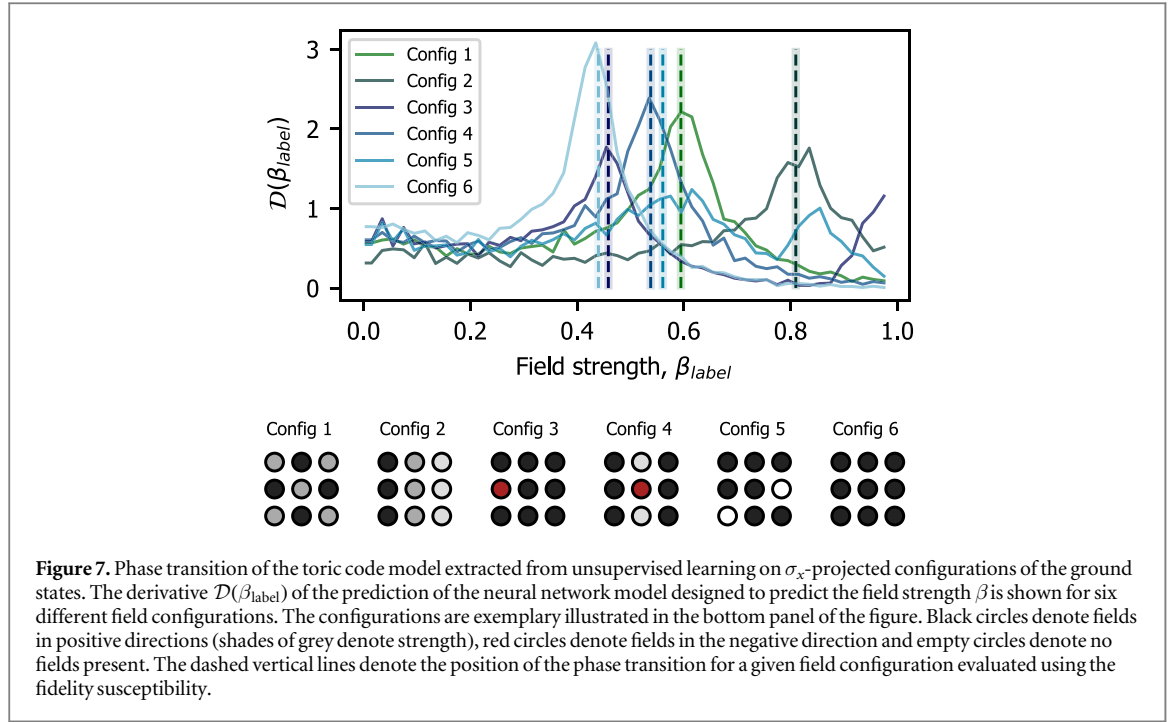
In other words, we calculate the overlap of two ground state wave functions with applied fields whose magnitudes are very close to each other. We can indeed observe a change in the behavior of the overlap in the neighborhood of the phase transition. The rate of this change is better analyzed by studying the derivative of the quantity in equation (12), the so-called fidelity susceptibility

$$\chi_F = - \left. \frac{\partial^2 \ln F}{\partial^2(\delta\beta)} \right|_{\delta\beta=0}. \quad (13)$$

We observe in figure 7 that the dashed lines determined from fidelity susceptibility calculation are in good agreement with the maximum of the peaks of the derivative $\mathcal{D}(\beta_{\text{label}})$ of the predictive model. We show details of the fidelity susceptibility calculation in appendix C.

3.1.2. The σ_z -projection

We can ask whether a particular projection is necessary to determine the topological phase transition from the spin configurations alone. Let us consider measuring the ground state in the σ_z basis instead of σ_x . In order to simplify mathematical expressions let us without loss of generality choose a different state from the ground state manifold



$$|\Psi_z\rangle = \frac{1}{\sqrt{Z_z}} e^{\beta \sum_i \lambda_i \sigma_i^x} \sum_{g \in G} g |0_z\rangle. \quad (14)$$

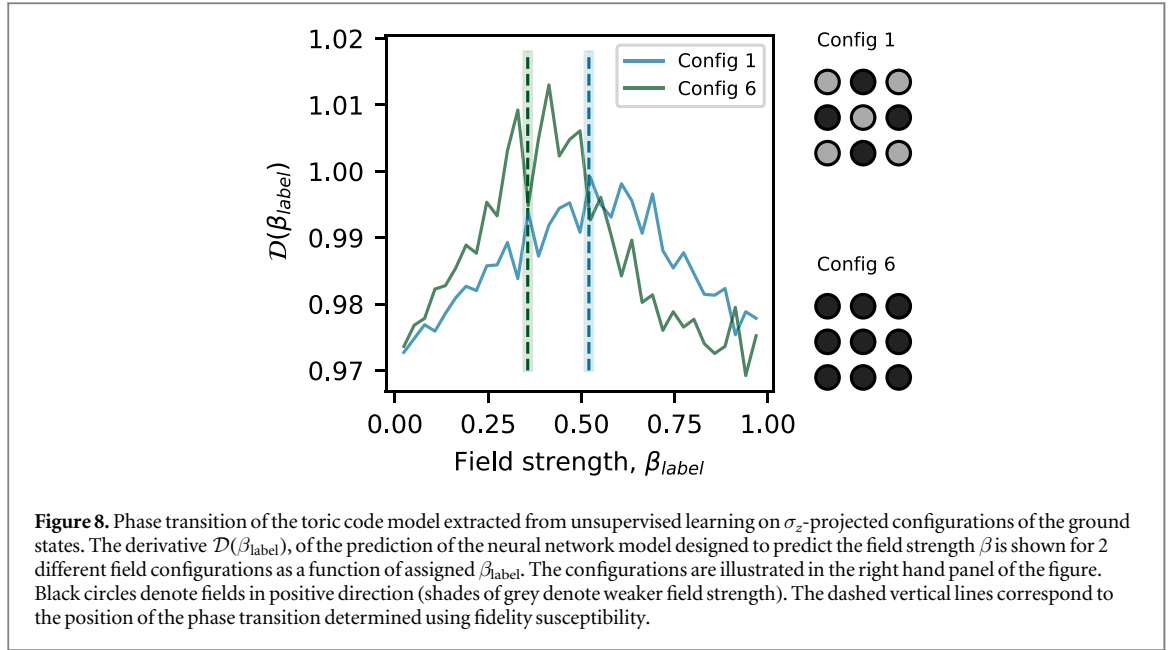
Here, analogously to equation (9), G is the abelian group of possible products of vertex operators and $|0_z\rangle$ is the reference state. Note that we chose a different reference state. As opposed to equation (9), all spins of the reference state are aligned in an eigenstate of σ_z instead of σ_x . The normalization is denoted with Z_z and not elaborated on further here.

Let us again examine the limiting behavior of β if σ_z was measured on every spin of the state (14). If $\beta \rightarrow 0$ we obtain the exact toric code ground state. Projective measurement of σ_z on this ground state then results in the configuration $S_g = g|0_z\rangle$, hence the closed loop (plaquette) conditions $B_p|S_g\rangle = |S_g\rangle$ are fulfilled. Every configuration $g|0_z\rangle$ fulfilling these constraints is obtained with equal probability. We note here, that the local plaquette constraints are in exact correspondence to the IGT local constraints fulfilled in the zero temperature phase.

Applying the same logic as in the case of the σ_x projection, we can conclude that in the case $\beta \rightarrow \infty$ we arrive at a completely polarized state, where all spins are aligned in the x -direction. If we now project onto a σ_z eigenstate, the plaquette constraints will not stay preserved. In fact, any configuration in σ_z basis will be obtained with equal probability. Hence, we find that in the σ_z projection the phase transition arises from a quite different process than we observed before: for small β the system would be in the state where loop conditions are preserved, while for large β they are violated. While in the case of σ_x the phase transition simply changes the weight for some states from the set preserving loop condition, in the case of σ_z projection we transition from the state where all the states preserving loop condition are weighted equally to the phase where the loop constraints are completely violated. We can therefore draw parallels to the previously examined IGT transition at finite temperature. In particular, in both cases we observe phases that can be distinguished by checking for a violation of the local closed loop constraints. However, there is a crucial difference between these two transitions. IGT exhibits a finite temperature cross-over and the violation of local constraints is a result of thermal excitations. Here, we consider a quantum phase transition at zero temperature, where the local constraints are violated due to the interplay with added perturbations. In particular, for IGT in the thermodynamic limit there is only a transition at infinite inverse temperature, β , whereas the quantum phase transition we consider here occurs at a finite field strength β in the thermodynamic limit as well.

We employ the unsupervised learning technique on the σ_z projection of the modified toric code ground state (14) with the strength of the background field β as a label for the supervised part of the protocol. This time our neural net model consists of two convolutional layers (with 128 filters and kernel size 2) and three dense layers (with 100, 100 and 50 neurons, respectively).

We show in figure 8 the results for $N = 4$ (32 spins) and two different field configurations of the 6 field configurations defined in appendix C and previously studied on a larger lattice for the x -projections. The reduction of the system size and number of field configurations presented here are a consequence of



constructing projections onto the σ_z -basis from the ground state containing σ_x fields: mixed $\sigma_z\sigma_x$ terms make Monte Carlo update computationally significantly more expensive (for details see appendix C).

We note here, that for both σ_x and σ_z projections we limit ourselves to a single topological sector with the choice of the ground state in equation (9). Since the other topological sectors exhibit qualitatively the same phase transition at the same transition point the discussion above can be extended to any ground state within the topological sector.

3.2. Phase transition determination from the stabilizer expectation values

Finally, we discuss on how to obtain the topological phase transition in the toric code model by extracting necessary information by measurements that can be readily performed on the quantum state at hand and do not require projections onto the spin configurations. It was shown in [30] that the behavior of the expectation values of the stabilizer operators are intimately related to the position of the topological phase transition in the toric code model. We use our predictive model to evaluate the position of the phase transition from the expectation value of the stabilizer operators to offer an alternative method to determine the position of the phase transition.

As in the previous sections, we train a neural network to predict the value of the field strength amplitude, β . This time we use as an input the expectation value of the plaquette operator, $\langle B_p \rangle$ (with β as a label). Then we use the network to predict the field strength β for the expectation values of B_p evaluated with respect to the new set of quantum states. We use a neural network with two dense layers (with 20 neurons each). The derivative $\mathcal{D}(\beta_{\text{label}})$ of the predictive model is shown as a function of field strength for six distinctive field configurations in figure 9. We again compare to the position of the phase transition obtained by fidelity susceptibility method (dashed lines) and observe an excellent agreement. As in the case of configurational data, we used the topological sector defined by the ground state in equation (9). Our result is again independent of the state in the ground state manifold, as all ground states are locally indistinguishable in the topological phase. As a consequence, the local expectation value $\langle B_p \rangle$ does not depend on the topological sector examined.

While the connection between expectation values of stabilizer operators and the position of the phase transition have not been shown analytically, another numerical evidence was provided in [31]. The authors examine direct detection of anyons, a process that can be mapped onto the expectation values which we investigated here. The presence of anyons is then immediately tied to the existence of topological order. We elaborate on the connection to the present work in appendix C.

4. Discussion

Unsupervised machine learning techniques for phase classification in condensed matter physics are potentially powerful tools for the discovery of new quantum phases. Due to the lack of local order parameters, phases exhibiting topological order present a challenging task for unsupervised methods. In this work, we have shown that a novel unsupervised method, namely the analysis of predictive neural network models [7], can reliably detect the violations of topological order, or a topological phase transition should it exist.

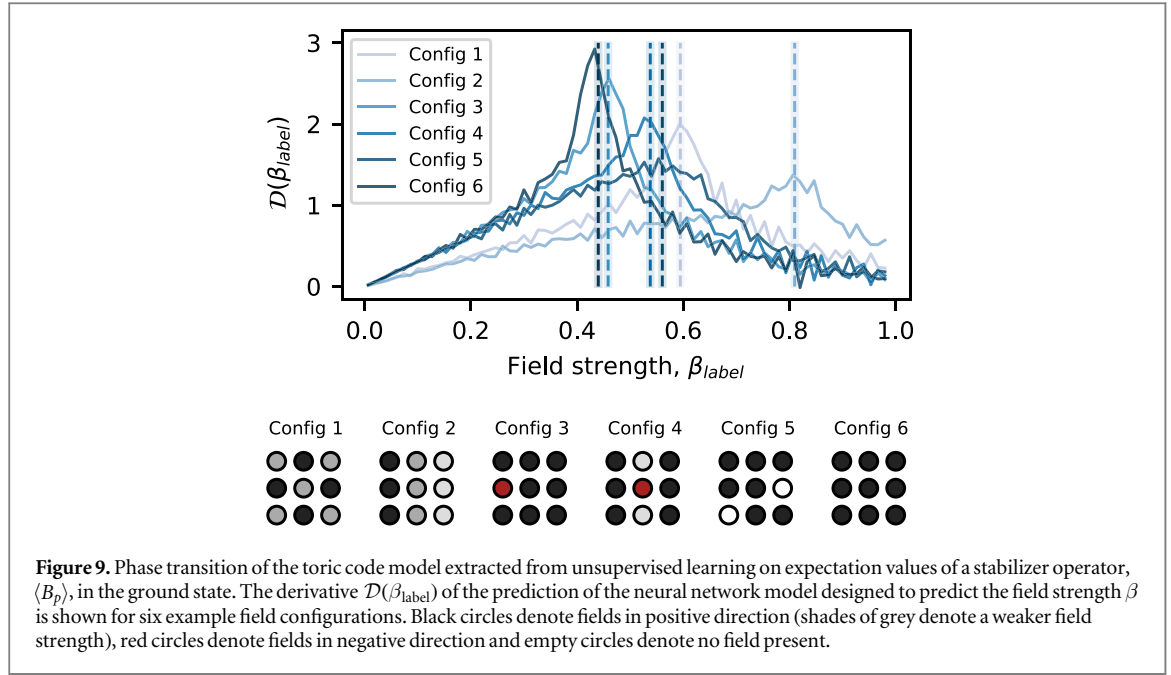


Figure 9. Phase transition of the toric code model extracted from unsupervised learning on expectation values of a stabilizer operator, $\langle B_p \rangle$, in the ground state. The derivative $\mathcal{D}(\beta_{\text{label}})$ of the prediction of the neural network model designed to predict the field strength β is shown for six example field configurations. Black circles denote fields in positive direction (shades of grey denote a weaker field strength), red circles denote fields in negative direction and empty circles denote no field present.

Topologically ordered states have been particularly challenging for unsupervised learning techniques, because the quantity characterizing topological order is inherently non-local and hard to identify from raw data. In the method presented here, we trained the network on an arbitrary continuous parameter associated to the state and then analyzed the errors in the network predictions. We presented numerical evidence that these prediction errors are signatures of a phase transition. We showed that this conclusion was independent of the particular type of phase transition present in the system and the type of the input data. To determine which type of phase transition is present, applying our method in conjunction with principal component analysis [17], variational auto-encoders [25], or confusion schemes [18, 21] that all succeed in determination of phase transition governed by local order parameter can be used as a guideline.

Providing the resolution to the problem of finding the cross-over temperature in the IGT and its generalizations in an unsupervised manner is the first step towards developing reliable techniques that can be applied to study the models whose phase diagrams are not yet fully understood.

Acknowledgments

We thank Juan Carrasquilla for fruitful discussions. We acknowledge Mark H Fischer for contributions to early versions of the code. We are grateful for financial support from the Swiss National Science Foundation and the NCCR QSIT. This work has received funding from the European Research Council under grant agreement no. 771503.

Appendix A. IGT: predictive model

We created the samples used for training (like those shown in figure 1) of our model using Monte Carlo simulations. We created data for system sizes $N \times N \times 2$ with $N \in \{4, 8, 12, 16, 20, 24, 28\}$. For each system size we created 100 different values of $\beta \in [0, 5]$. We generated 20 000 configurations for each pair $[\beta, N]$. The neural net we used consists of 2 convolutional (128 filters, kernel size 3) and 2 dense layers (300 and 100 neurons, respectively). We observed that our method is flexible with respect to the hyper parameters of the neural network. However, too shallow networks that predict the same average β for all states inside and outside of the topological sector should be avoided.

We trained the network by minimizing the mean-squared-error loss function

$$L^{\text{mse}}(\beta_{\text{pred}} - \beta_{\text{label}}) = \frac{1}{n} \sum_n (\beta_{\text{pred}} - \beta_{\text{label}})^2, \quad (\text{A1})$$

where β_{pred} is the β determined by the network and β_{label} is the label of the given input sample, n is the batch size. The predictions of β by the network and their divergences are shown in figures 2 and 3.

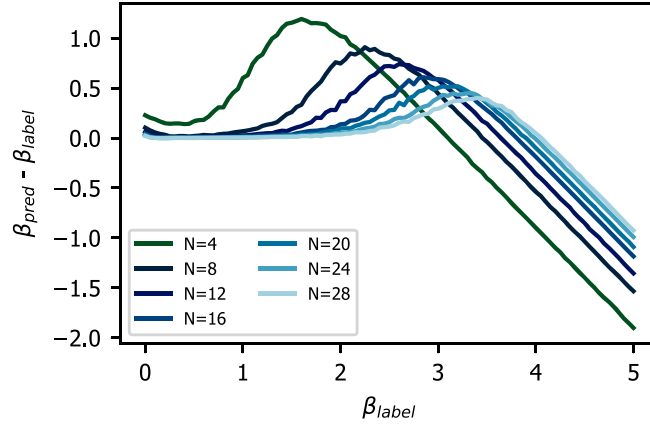


Figure A1. Density of states based prediction of β . We plot the difference between true and predicted β , $\beta_{pred} - \beta_{label}$, as a function of β_{label} .

In order to evaluate the error bars of the neural net predictions, we repeated the training procedure outlined above for 5 separate models (identical construction, separately generated training sets). Then we evaluated standard deviation of the critical β^* .

We can replicate the predictions achieved by a neural network using a density of states based model as explained in the main text. Let us consider lattice configurations (training samples) X_n with their assigned inverse temperature labels $\beta_n = \beta_{label}(X_n)$. We can evaluate an energy, E_n of each of these configurations using the formula

$$E = -J \sum_p \prod_{i \in p} \sigma_i^z, \quad (A2)$$

where the first summation is over all plaquettes, p , whereas the second summation is over spins within each plaquette. For convenience we choose $J = 1$. Then we can construct the density of states distribution of the training set

$$\epsilon(\beta, E) = \frac{\sum_{n=1}^N \delta_{E, E_n} \delta_{\beta, \beta_n}}{\sum_{n=1}^N \delta_{\beta, \beta_n}}. \quad (A3)$$

Here, $\delta_{a,b}$ is the Kronecker delta symbol ($\delta_{a,b} = 1$ for $a = b$ and $\delta_{a,b} = 0$ for $a \neq b$), E_n is energy for the configuration X_n evaluated using formula A2 and N is number of configurations X_n in the training set. We can write the energy distribution in the form above because the energy of the lattice configuration, E is discrete by construction and β is discretized in steps as explained above. An example of this distribution is shown in figure 5 for the system size $N = 8$.

Having access to the energy E_n of a given configuration X_n , we can then evaluate the average β of all states with energy E , which we denote by β^{av} for a configuration X_n in the training set

$$\beta^{av}(E) = \frac{\sum_{n=1}^N \delta_{E, E_n} \beta_n}{\sum_{n=1}^N \delta_{E, E_n}}. \quad (A4)$$

The function above predicts the value of β which is most likely for a given energy, E , given the energy distribution of the training set. We can use the function (A4) to determine the relation between assigned labels, β_n and values of β predicted by our model

$$\beta^{est}(\beta) = \frac{\sum_{m=1}^M \beta^{av}(E_m) \delta_{\beta, \beta_m}}{\sum_{m=1}^M \delta_{\beta, \beta_m}}, \quad (A5)$$

where M is the number of configurations X_m in an arbitrarily chosen test set. Using equation (A5) we can predict the estimated β for a range of true labels. In figure A1 we show the difference between true and predicted β as a function of true β . In figure A2 we show the derivative of the estimated β as a function of true β . Comparing with figure 3 we see that our model based on the density of states in the training set is reproducing well the actions of the neural net model we introduced in the main text. We have used maxima determined by the density of states as a dashed-line reference for the position of the transition in the main text.

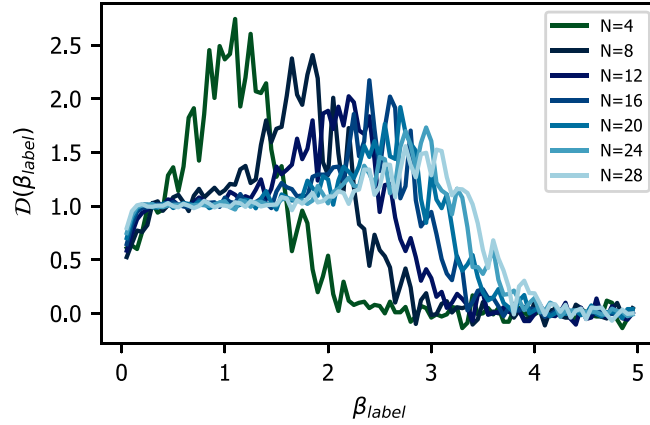


Figure A2. We show the derivative of the density of states based prediction, $\mathcal{D}(\beta_{\text{label}})$, as a function of β_{label} .

Appendix B. Toric code: predictive model

B.1. Mapping to Ising model

The projection of the modified toric code ground state on the σ_x basis can be understood by mapping to a classical Ising model. Let us examine the ground state of the toric code with fields (9)

$$|\Psi\rangle = \frac{1}{\sqrt{Z}} \sum_{h \in H} e^{\beta \sum_i \lambda_i \sigma_i^x(h)} h |0_x\rangle. \quad (\text{B1})$$

We perform a projection of $|\Psi\rangle$ onto the σ_x basis. As stated in the main text, the outcome of the projection are the configurations $|S_h\rangle$ fulfilling the closed-loop condition $A_s |S_h\rangle = |S_h\rangle$. In addition, the probability to obtain $|S_h\rangle$ after projection is given by

$$p(S_h) = |\langle S_h | \Psi \rangle|^2 = \frac{e^{\beta \sum_i \lambda_i \sigma_i^x(h)}}{\sum_{\tilde{h} \in H} e^{\beta \sum_i \lambda_i \sigma_i^x(\tilde{h})}}. \quad (\text{B2})$$

The system can be mapped to the classical Ising model as follows [26]. First we notice, that every group element h uniquely determines the configuration $|S_h\rangle$ by applying h to a reference spin configuration, which we choose to be $|0_x\rangle$ (all spins up in x -basis). Then, $|S_h\rangle = h|0_x\rangle$. In addition, h corresponds to the product of a set I_h of plaquette operators $h = \prod_{p \in I_h} B_p$. Every such set of plaquette operators corresponding to a spin configuration $|S_h\rangle$ can be mapped to the following pseudo-spin configuration: artificial degrees of freedom (pseudo-spins) $\theta_p \in \{-1, 1\}$ are introduced on every plaquette. The value of the pseudo-spin θ_p is determined by I_h : if $B_p \in I_h$ (plaquette flipped) it is equal to -1 , else it is equal to one. As a consequence, the original spin-configuration $\{\sigma_i^x(h)\}$ (corresponding to $|S_h\rangle$) can be deduced from the pseudo-spin configuration θ_p by applying the rule $\sigma_i^x(h) = \theta_p \theta_{p'}$. Here, p and p' are the two adjacent plaquettes to spin i . The geometry of the mapping is illustrated in figure B1.

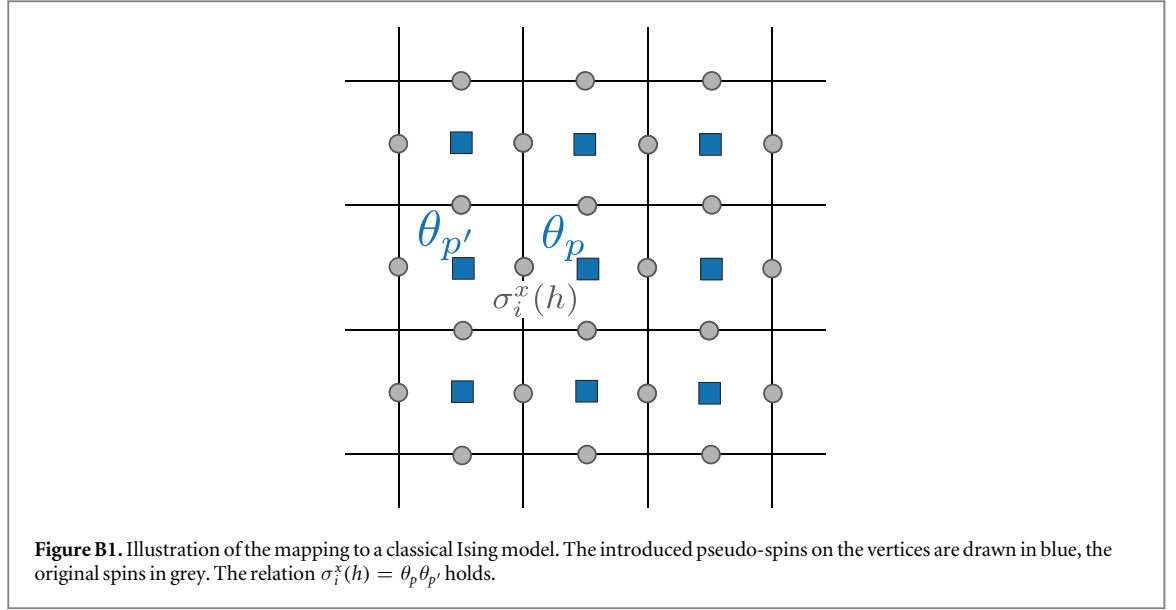
Let us translate the mapping into the calculation of the probability $p(S_h)$. Inserting the rule $\sigma_i^x(h) = \theta_p \theta_{p'}$ to equation (B2) yields

$$p(S_h) = p(\{\theta^h\}) = \frac{e^{\beta \sum_{\langle p, p' \rangle} J_{p, p'} \theta_p^h \theta_{p'}^h}}{\sum_{\{\theta\}} e^{\beta \sum_{\langle p, p' \rangle} J_{p, p'} \theta_p \theta_{p'}}}, \quad (\text{B3})$$

with $J_{p, p'} = \lambda_i$ for the plaquettes p, p' adjacent to edge i and summing over nearest-neighbors $\langle p, p' \rangle$. The pseudo-spin configuration obtained from the group element h by applying the explained mapping is denoted by the parameters $\{\theta^h\}$. In contrast, the sum over $\{\theta\}$ represents a sum over all possible pseudo-spin configurations. We recognize the expression (B3) as Boltzmann weight for an Ising model with bond strengths $J_{p, p'}$ at temperature $T = 1/(k_B \beta)$. The topological phase transition undergone by the studied perturbed toric code model hence shows the behavior of an Ising phase transition from disordered pseudo-spin configurations to ordered spin configurations after projecting onto the σ_x basis.

B.2. Calculation of the fidelity susceptibility

We compare the position of the phase transition found by the neural network to the transition indicated by the fidelity susceptibility [2]. The fidelity susceptibility is defined as



$$\chi_F = - \left. \frac{\partial^2 \ln \langle \Psi(\beta) | \Psi(\beta + \Delta\beta) \rangle}{\partial (\Delta\beta)^2} \right|_{\Delta\beta=0}, \quad (\text{B4})$$

where the state $|\Psi(\beta)\rangle$ is a ground state of a given Hamiltonian with respect to the parameter β . For our particular model, $|\Psi(\beta)\rangle$ is given in equation (9). It has been shown, that a divergence or maximum of the fidelity susceptibility χ_F indicates a second-order symmetry-breaking quantum phase transition [3–5]. Numerical evidence suggests that topological phase transitions are indicated in the same way [6]. We can calculate the fidelity susceptibility for the introduced disordered toric model as

$$\chi_F = \frac{1}{4} \frac{\sum_{h \in H} \left(\sum_i \lambda_i \sigma_i^x(h) \right)^2 e^{\beta \sum_i \lambda_i \sigma_i^x(h)} \cdot Z}{Z^2}, \quad (\text{B5})$$

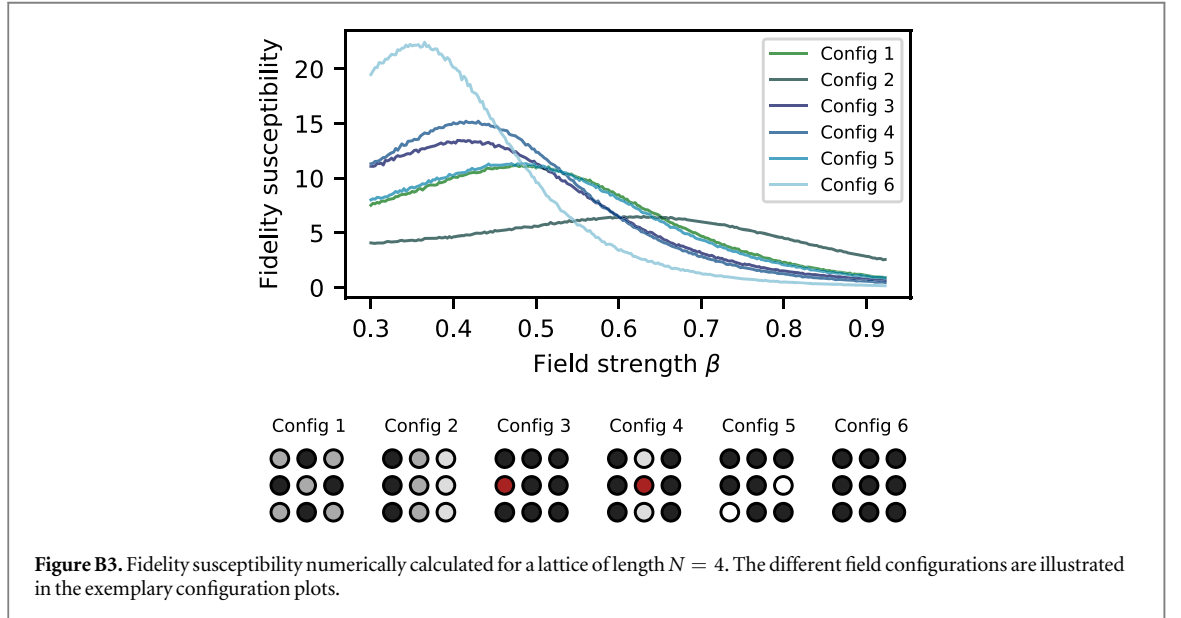
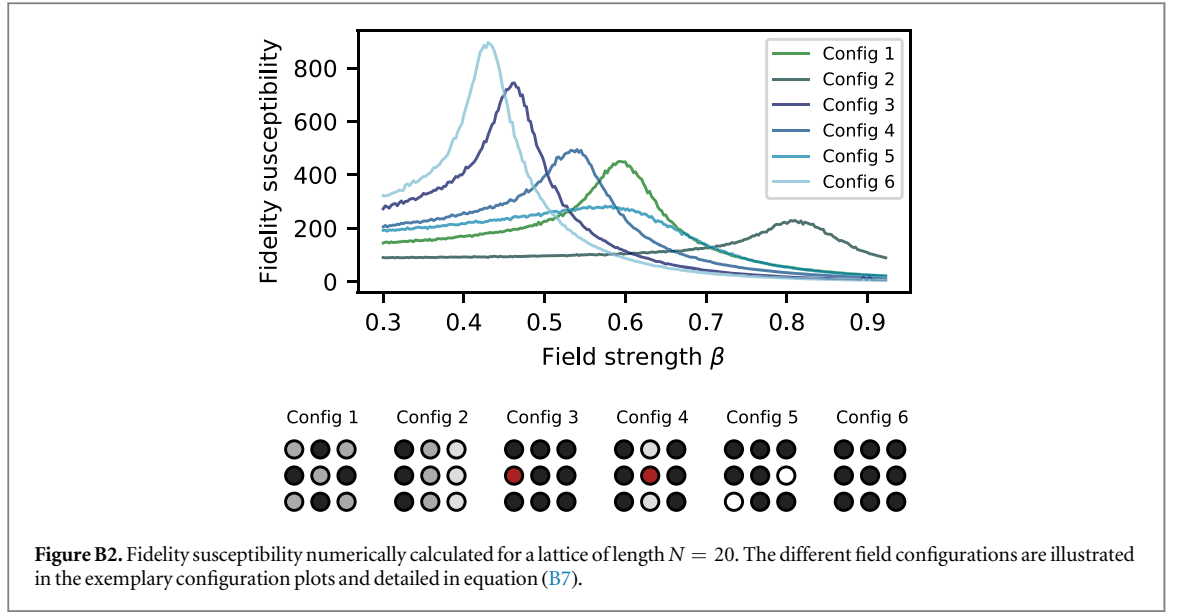
$$- \frac{1}{4} \frac{\left(\sum_{h \in H} \left(\sum_i \lambda_i \sigma_i^x(h) \right)^2 e^{\beta \sum_i \lambda_i \sigma_i^x(h)} \right)^2}{Z^2}. \quad (\text{B6})$$

We numerically evaluate the expression via Monte Carlo sampling for the different field configurations examined throughout this work and compare the position of the maximum with the position of the phase transition found by the neural network. In particular, we calculate the fidelity susceptibility for the following 6 different field configurations:

$$\begin{aligned} \text{'Config 1': } \lambda_i &= \begin{cases} 0.5 & \text{if } i \bmod 2 = 0, \\ 1 & \text{else,} \end{cases} \\ \text{'Config 2': } \lambda_i &= \begin{cases} 0.5 & \text{if } i \bmod 3 = 1, \\ 0.25 & \text{if } i \bmod 3 = 2, \\ 1 & \text{else,} \end{cases} \\ \text{'Config 3': } \lambda_i &= \begin{cases} -0.5 & \text{if } i \bmod 10 = 9, \\ 1 & \text{else,} \end{cases} \\ \text{'Config 4': } \lambda_i &= \begin{cases} -0.5 & \text{if } i \bmod 20 = 19, \\ 0.25 & \text{if } i \bmod 5 = 4, \\ 1 & \text{else,} \end{cases} \\ \text{'Config 5': } \lambda_i &= \begin{cases} 0 & \text{if } i \bmod 5 = 0, \\ 1 & \text{else,} \end{cases} \\ \text{'Config 6': } \lambda_i &= 1 \quad \forall i. \end{aligned} \quad (\text{B7})$$

The chosen configurations constitute representative examples and incorporate different distributions of positive, negative and zero fields. The numerical simulations of the fidelity susceptibility for the 6 different field configurations on lattices of lengths examined throughout this work ($N = 20$ and $N = 4$) are shown in figures B2 and B3.

The fidelity susceptibility can be connected to the heat capacity of the classical Ising model explained in the previous subsection, as elaborated in [6] and [30].



B.3. Numerical simulation of projections

The projection of the ground state of the perturbed toric model onto the σ_x or σ_z basis is in both cases simulated via Monte Carlo sampling. To project on the σ_x basis, we aim to obtain a configuration S_h sampled from the probability distribution $p(S_h)$ (B3). Such a configuration is reached via a Markov chain. More concretely, we start with a lattice with all spins up in x basis and construct the Markov chain as follows: in each step (with given spin configuration S_{h_i}), a random plaquette p is picked. The decision, whether the plaquette should be flipped and $S_{h_i} \rightarrow S_{h_{i+1}}$ is made via a Metropolis-Hastings test. The four spins around the chosen plaquette are flipped with probability

$$\frac{p(S_{h_{i+1}})}{p(S_{h_i})} = e^{\beta \sum_{i \in p} \lambda_i (\sigma_i^x(h_{i+1}) - \sigma_i^x(h_i))}. \quad (\text{B8})$$

After thermalization time, the spin configuration S_h is obtained with probability $p(S_h)$ and a projection is simulated.

Projecting on the σ_z basis follows the same principle with the caveat that the spin-flip probability is computationally expensive to calculate. We start from a state in the ground state manifold

$$|\Psi_z\rangle = \frac{1}{\sqrt{Z_z}} e^{\beta \sum_i \lambda_i \sigma_i^x} \sum_{g \in G} g|0_z\rangle \quad (\text{B9})$$

and project on the σ_z basis. In particular, let us examine the probability to obtain the spin configuration

$$|z_M\rangle = \prod_{i \in M} \sigma_i^x |0_z\rangle, \quad (\text{B10})$$

where M is a set of spins that are flipped in the configuration $|z_M\rangle$ with respect to the initial state $|0_z\rangle$. Then, the probability to project on $|z_M\rangle$ is given by

$$p(z_M) = |\langle z_M | \Psi_z \rangle|^2 = \left(\frac{\sum_C \prod_{j \notin C_M} \cosh(\frac{\beta}{2} \lambda_j) \prod_{i \in C_M} \sinh(\frac{\beta}{2} \lambda_j)}{\sum_C \prod_{j \notin C} \cosh(\beta \lambda_j) \prod_{i \in C} \sinh(\beta \lambda_j)} \right)^2, \quad (\text{B11})$$

where the closed loops C correspond to the set of spins that are flipped when applying a product of vertex operators to the initial state $\prod A_s |0_z\rangle = \prod_{i \in C} \sigma_i^x |0_z\rangle$. The sum is over all possible closed loops, hence over all possible products of vertex operators. Similarly, the set C_M can be constructed from the closed loop C by flipping the spins in M

$$\prod_{i \in M} \sigma_i^x \prod_{i \in C} \sigma_i^x |0_z\rangle = \prod_{i \in C_M} \sigma_i^x |0_z\rangle. \quad (\text{B12})$$

In order to obtain a spin configuration sampled from the distribution p defined in equation (B11), we construct a Markov chain by starting with a spin configuration $|0_z\rangle$ in the σ_z basis. In each step, a random spin is chosen and flipped (updating the spin configuration $|z_i\rangle$ to $|z_{i+1}\rangle$) with probability $p(z_{i+1})/p(z_i)$. As the computation of the spin flip probability is expensive, we simulate z -projections only for $2 \times 4 \times 4 = 32$ spins.

The neural network for the predictive model on the lattice with length $N = 4$ (outcomes shown in figure 8) consists of two convolutional layers with 128 neurons each and three dense layers with 100, 100 and 50 neurons. Training was conducted on a set of 157960 examples in total, 144 values of β between 0 and 1. For evaluation of the trained model to predict the field parameter, the values for β were chosen to be 72 discrete steps between 0 and 1. A total of 100800 evaluation examples was generated, data augmentation (rotations, translations, mirror) led to an additional factor of 100.

B.4. Detection of quasiparticles

We elaborated in the main text, that measuring a stabilizer expectation value contains sufficient information to indicate the position of the topological phase transition. This behavior can be related to a detection of the topological phase transition by measuring quasiparticles. More concretely, numerical evidence has been presented in [31], that a topological phase transition can be indicated by a detection of quasiparticles. For the toric model, the toric Hamiltonian can be modified such that the ground state contains a pair of quasiparticles

$$H_m = - \sum_{p \neq p_1, p_2} B_p - \sum_s A_s + \sum_{p=p_1, p_2} B_p. \quad (\text{B13})$$

At the plaquettes p_1 and p_2 , the expectation value $\langle B_p \rangle_m = -1$ measured on the ground state shows the existence of a quasiparticle. Here, the subscript m denotes that the expectation value is taken with respect to the ground state of the modified toric code. When adding a phase-transition driving perturbation parametrized by a field β , the position of the phase transition is indicated by a divergence in the derivative $\partial_\beta \langle B_p \rangle_m$ with $p \in \{p_1, p_2\}$. If the added perturbation is of the form

$$H_m \rightarrow H_m + \sum_p e^{-\beta \sum_{i \in p} \lambda_i \sigma_i^x}, \quad (\text{B14})$$

the following relation holds

$$\langle B_p \rangle = -\langle B_{p_1} \rangle_m. \quad (\text{B15})$$

Here, the expectation value $\langle B_p \rangle$ is evaluated on the ground state of the model without quasiparticles (9) examined throughout this work. We conclude that the divergence in the derivative of $\langle B_p \rangle$ indicates the position of the phase transition. We therefore understand, that the predictive model is able to reconstruct this behavior as the accuracy of the predicted field strength depends on the slope of the expectation value $\langle B_p \rangle$.

Appendix C. Comparison with the confusion scheme

We compare the introduced unsupervised approach with the similar confusion scheme developed in [18]. We give here a quick summary of the scheme, for further details we refer to the work of the original authors. As starting point, one is given (uniformly sampled) data in the range (β_a, β_b) and an unknown critical value β_c separating two phases, with $\beta_a < \beta_c < \beta_b$. The critical point can be estimated by systematically ‘guessing’. In

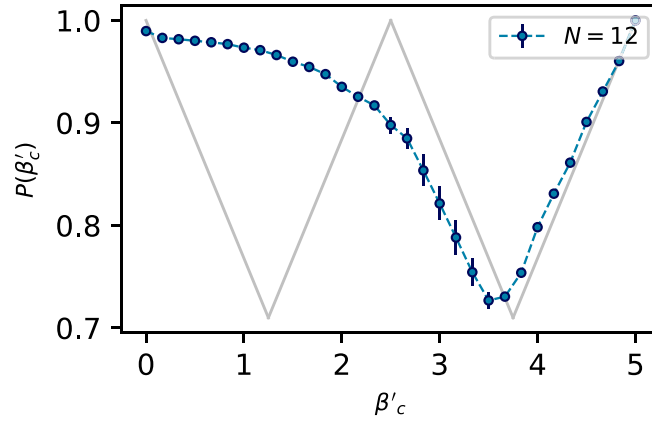


Figure C1. Accuracy $P(\beta'_c)$ of the confusion scheme on the IGT problem with $N = 12$ (blue dots). In grey, the ideal W-shape indicating a phase transition is shown as comparison. The crossover from ground state to non ground state is at around $\beta_c = \beta'_c = 2.5$. The error bars are obtained by averaging over ten different and independent Monte Carlo runs for obtaining the data.

particular, one chooses a value β'_c as guess for the critical point and separates the dataset in two parts: all values smaller than β'_c are labeled with 0, all values larger than β'_c are labeled with 1. A neural network is trained to reproduce the labels. This procedure is repeated for all values β'_c in the interval (β_a, β_b) , and the performance $P(\beta'_c)$ of the trained networks is plotted and analyzed. The main idea is, that the network performs best when β'_c is chosen to be the critical point. More concretely, the performance has a W-shape, if a phase transition is recognized (see figure C1). As a consequence, the main difference between our method and the confusion scheme is that we use a single network and analyze its prediction, while the method of van Nieuwenburg *et al* requires separate networks for each point of the phase diagram they wish to check (which might make it less suitable for high dimensional parameter spaces).

The confusion scheme has proven to perform well on a variety of phase transitions in classical systems and quantum systems. We employ the confusion scheme on the IGT crossover analyzed in section 2. The network performance $P(\beta'_c)$ is shown in figure C1 for the system size $N = 12$. As the typical W-shape is not reproduced, the position of the phase transition is not recognized. Instead, we obtain a shifted V-shape. We can understand the result in the following way. The confusion method can easily distinguish between all states with $\beta < \beta_c$ on one side of the transition, but it cannot distinguish at the states in the ordered sector. More concretely, all samples in the ordered sector ($\beta > \beta_c$) are in the ground state (no local constraints are violated) and thus indistinguishable to the networks. In the disordered section, the configurations at different β'_c are distinguishable by different numbers of local plaquette constraint violations. Smaller values of β lead to a larger number of frustrated plaquettes. As a consequence, the network is able to distinguish all the states in the disordered phase.

For this example, a relatively small network architecture consisting of a convolutional layer ($12 \times 12 = 144$ neurons, 5 filters and kernel size 3) and a dense layer with one output neuron was chosen. Some testing with larger networks showed a similar accuracy curve. Tests on smaller networks showed a different V-shape. Specifically, the obtained V-shape was not shifted to the disordered phase. Henceforth, the networks were not able to distinguish the configurations in none of the phases and had to guess randomly. We conclude, that the application of the confusion scheme to IGT is not straightforward.

References

- [1] Sachdev S 2011 *Quantum Phase Transitions* (Cambridge: Cambridge University Press)
- [2] You W-L, Li Y-W and Gu S-J 2007 Fidelity, dynamic structure factor, and susceptibility in critical phenomena *Phys. Rev. E* **76** 022101
- [3] Venuti L C and Zanardi P 2007 Quantum critical scaling of the geometric tensors *Phys. Rev. Lett.* **99** 095701
- [4] Gu S-J and Lin H-Q 2009 Scaling dimension of fidelity susceptibility in quantum phase transitions *Europhys. Lett.* **87** 10003
- [5] Zanardi P and Paunković N 2006 Ground state overlap and quantum phase transitions *Phys. Rev. E* **74** 031123
- [6] Abasto D F, Hamma A and Zanardi P 2008 Fidelity analysis of topological quantum phase transitions *Phys. Rev. A* **78** 010301
- [7] Schäfer F and Lörch N 2019 Vector field divergence of predictive model output as indication of phase transitions *Phys. Rev. E* **99** 062107
- [8] Wegner F J 1971 Duality in generalized ising models and phase transitions without local order parameters *J. Math. Phys.* **12** 2259–72
- [9] Fradkin E and Susskind L 1978 Order and disorder in gauge systems and magnets *Phys. Rev. D* **17** 2637
- [10] Kogut J B 1979 An introduction to lattice gauge theory and spin systems *Rev. Mod. Phys.* **51** 659
- [11] Kitaev A 2006 Anyons in an exactly solved model and beyond *Ann. Phys.* **321** 2
- [12] Castelnovo C and Chamon C 2007 Entanglement and topological entropy of the toric code at finite temperature *Phys. Rev. B* **76** 184442
- [13] Sachdev S 2018 Topological order, emergent gauge fields, and fermi surface reconstruction *Rep. Prog. Phys.* **82** 014001
- [14] Carrasquilla J and Melko R G 2017 Machine learning phases of matter *Nat. Phys.* **13** 431

- [15] Elitzur S 1975 Impossibility of spontaneously breaking local symmetries *Phys. Rev. D* **12** 3978
- [16] Giles R 1981 Reconstruction of gauge potentials from wilson loops *Phys. Rev. D* **24** 2160
- [17] Wang L 2016 Discovering phase transitions with unsupervised learning *Phys. Rev. B* **94** 195105
- [18] Van Nieuwenburg E P L, Liu Y-H and Huber S D 2017 Learning phase transitions by confusion *Nat. Phys.* **13** 435
- [19] Rodriguez-Nieva J F and Scheurer M S 2019 Identifying topological order via unsupervised machine learning *Nat. Phys.* **15** 790–5
- [20] Zhao K-W, Kao W-H, Wu K-H and Kao Y-J 2019 Generation of ice states through deep reinforcement learning *Phys. Rev. E* **99** 062106
- [21] Broecker P, Assaad F F and Trebst S 2017 Quantum phase recognition via unsupervised machine learning arXiv:1707.00663
- [22] Shirinyan A A, Kozin V K, Hellsvik J, Pereiro M, Eriksson O and Yudin D 2019 Self-organizing maps as a method for detecting phase transitions and phase identification *Phys. Rev. B* **99** 041108
- [23] Ponte P and Melko R G 2017 Kernel methods for interpretable machine learning of order parameters *Phys. Rev. B* **96** 205146
- [24] Valletti S M P, Vlcek L, Ziatdinov M, Vasudevan R K and Kalinin S V 2020 Reconstruction of the lattice hamiltonian models from the observations of microscopic degrees of freedom in the presence of competing interactions arXiv:2001.06854
- [25] Wetzel S J 2017 Unsupervised learning of phase transitions: from principal component analysis to variational autoencoders *Phys. Rev. E* **96** 022140
- [26] Castelnovo C and Chamon C 2008 Quantum topological phase transition at the microscopic level *Phys. Rev. B* **77** 054433
- [27] Fradkin E 2013 *Field Theories of Condensed Matter Physics* (Cambridge: Cambridge University Press)
- [28] Kitaev A Y 2003 Fault-tolerant quantum computation by anyons *Ann. Phys.* **303** 2–30
- [29] Tsomokos D I, Osborne T J and Castelnovo C 2011 Interplay of topological order and spin glassiness in the toric code under random magnetic fields *Phys. Rev. B* **83** 075124
- [30] Valenti A, van Nieuwenburg E, Huber S and Greplova E 2019 Hamiltonian learning for quantum error correction *Phys. Rev. Res.* **1** 033092
- [31] Manna S, Srivatsa N S, Wildeboer J and Nielsen A E B 2019 Quasiparticles as detector of topological quantum phase transitions arXiv:1909.02046