Interpreting machine learning of topological quantum phase transitions

Yi Zhang [1,2,*] Paul Ginsparg, [1,†] and Eun-Ah Kim [1,‡]

[1]Department of Physics, Cornell University, Ithaca, New York 14853, USA
[2]International Center for Quantum Materials, Peking University, Beijing 100871, China

There has been growing excitement over the possibility of employing artificial neural networks (ANNs) to gain new theoretical insight into the physics of quantum many-body problems. "Interpretability" remains a concern: can we understand the basis for the ANN's decision-making criteria in order to inform our theoretical understanding? "Interpretable" machine learning in quantum matter has to date been restricted to linear models, such as support vector machines, due to the greater difficulty of interpreting nonlinear ANNs. Here we consider topological quantum phase transitions in models of Chern insulator, $\mathbb{Z}_2$ topological insulator, and $\mathbb{Z}_2$ quantum spin liquid, each using a shallow fully connected feed-forward ANN. The use of quantum loop topography, a "domain knowledge"–guided approach to feature selection, facilitates the construction of faithful phase diagrams and semiquantitative interpretation of the criteria in certain cases. To identify the topological phases, the ANNs learn physically meaningful features, such as topological invariants and deconfinement of loops. The interpretability in these cases suggests hope for theoretical progress based on future uses of ANN-based machine learning on quantum many-body problems.

I. INTRODUCTION

There has been much recent activity in the quantum matter community applying ANN-based machine learning (ML) to synthetic [1–23] and experimental [24–27] quantum matter data. These efforts exploit the ability of ANN models to provide effective approximations of functions in high-dimensional spaces. ANNs can thus represent either many-body wave functions or complex mappings between many-body Hilbert space and associated emergent properties. Indeed ANN-based variational studies [1–7] and phase detection [8–21] have successfully reproduced known results, even for difficult cases such as topological characters and many-body localization. Even the simplest ANN with a single hidden layer can approximate any target function with a sufficiently wide hidden layer [28]; though in practice, such an architecture would be difficult to train efficiently, so multilayer architectures are preferred. Insofar as the goal of theoretical quantum matter physics is some fundamental understanding of its underlying properties, it would be disappointing if the ANN operated as a black box, giving little insight into the basis for its predictions or results. More generally, "interpretability" is an important challenge in many areas of algorithmic artificial intelligence, and is of particular importance in the use of ANNs as tools for scientific research.

The more expressive is a machine learning model, the harder it can be to interpret. It is difficult to characterize a function on a high-dimensional domain if it has an enormous number of parameters, and no obvious symmetries to permit easy visualization. If an ANN trained on a large number of labeled samples is able to predict with high precision the expected properties of new sample data, then we have certainly made progress. But if we are unable to extract from the ANN the specific features of the data, and combinations thereof, that it uses to make those predictions, then we have not achieved a deeper insight into the physics of the system. Our goal in physics is to develop some more compact formulation of the crucial degrees of freedom, derive from that some more general intuition into the system's behavior, and use that to develop analytic methods or physical laws that can be extended to a wide variety of related phenomena. A black box predictor that works only on a specific class of examples, and gives no insight into how it makes predictions, would be fundamentally unsatisfying as a tool for theoretical physics.

To date, there have been several preliminary endeavors on "interpretable" machine learning in problems of quantum matter. For instance, support vector machines have detected features of the order parameters in various spin models [29–31], the Hamiltonian constraints in gauge theories [29], and the level statistics in the many-body localization transition [32]. The support vector machines are intrinsically linear classifiers, based on finding a hyperplane to separate an already curated feature set, hence more easily interpreted. The more generally adopted ANN, on the other hand, can take an unmanageably large set of raw features, and transform and combine them in arbitrarily complex ways by way of millions or billions of learned parameter values. Its predictions are opaque unless we can determine how it has rearranged, amplified, and combined them into effective degrees of freedom

that govern the phenomena of interest. One viable interpretive method is Taylor expansion, which has succeeded in the extraction of certain order parameter expressions [33] and even laws of physics [34,35]; however, its complexity quickly evolves out of control as the ANN architecture or the target function becomes more complicated.

Here we consider three investigations of topological quantum phase transitions, each using a shallow fully connected feed-forward neural network. The quantum phase transitions are between topologically trivial states and three distinct topological phases in two spatial dimensions: a time-reversal symmetry breaking Chern insulator (CI), a time-reversal-invariant $\mathbb{Z}_2$ topological insulator (TI), and a $\mathbb{Z}_2$ quantum spin liquid (QSL). Of the three cases, the ANN-based phase diagram for a Chern insulator [11] and $\mathbb{Z}_2$ quantum spin liquid [12] have been previously obtained by two of us. The case of the ANN-based phase diagram for the $\mathbb{Z}_2$ TI, as far as we know, is first obtained in Sec. III here, although the model of Kane and Mele [36] is well known. In all three cases, the topological order is detected using only a simple shallow ANN, by using the physically motivated features introduced in Ref. [11], designated quantum loop topography (QLT). The QLT consists of a semilocal gauge-invariant product of two-point functions from (variational) Monte Carlo instances. The specific geometry of the QLT is guided by characteristics of the phase itself, i.e., it is based on "domain knowledge." Given the simplicity of our ANN, and its ability to interpolate between QLT and the topological phases of interest, it is plausible that insight into this physics can be derived by probing the "interior" of the ANN to illuminate properties of the function it has learned.

In this paper, we probe trained ANNs that yield correct topological quantum phase diagrams for the three cases of interest. We find interpretations of the "learning" of these ANNs fall into two classes: (1) a linear function corresponding to a topological invariant, and (2) nonlinear functions that build nonlocal and nonlinear observables from our QLT inputs. The CI case in Sec. II falls into class (1), and the $\mathbb{Z}_2$ TI and the $\mathbb{Z}_2$ QSL cases fall into class (2). The remainder of this paper is organized as follows. In Sec. II, we review the QLT and ANN-based phase detection for the Chern insulators, and interpret the trained ANN. In Sec. III, we obtain an ANN-based phase diagram for the $\mathbb{Z}_2$ TI using a new QLT, and again interpret what the trained ANN has learned about the system. In Sec. IV, we examine an ANN trained to detect the $\mathbb{Z}_2$ quantum spin liquid phase, and interpret the methodology it has learned. We close with a summary and concluding remarks in Sec. V.

II. INTERPRETING LINEAR ML: CHERN INSULATORS

QLT for CI assigns a $D(d_c)$-dimensional vector of complex numbers to each lattice site j, thereby forming a quasi-two-dimensional "image." The elements of the vector associated to the site j are chained products such as

$$\tilde{P}_{jk}|_\alpha \tilde{P}_{kl}|_\beta \tilde{P}_{lj}|_\gamma \text{ (QLT for CI)}, \tag{1}$$

where k and l are two sites that form a triangle with site j. In Eq. (1), each $\tilde{P}_{jk}|_\alpha \equiv \langle c_j^\dagger c_k \rangle_\alpha$ is a variational Monte Carlo sample of the two-point correlations associated with sites j, k
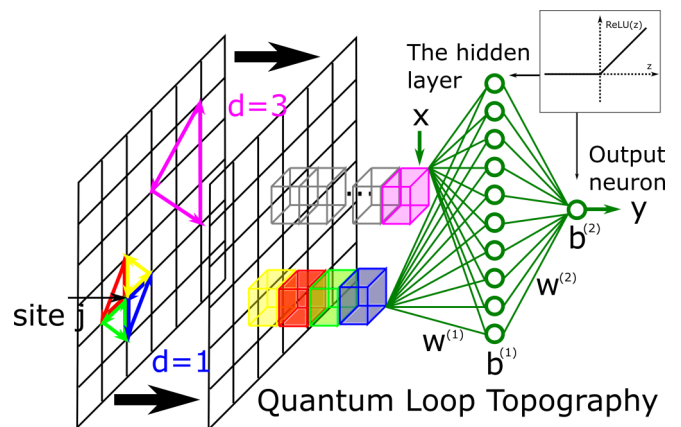

FIG. 1. Our supervised machine learning architecture for a topological quantum Hall insulator consists of a quantum loop topography feature-selection layer and an ANN with a single hidden layer. The neurons are rectified linear units with $\text{RELU}(z) = \max(z, 0)$ as the nonlinear activation function, as illustrated in the inset.

evaluated at Monte Carlo step α, and β, γ label different Monte Carlo steps. The length $D(d_c)$ of the vector is set by the total number of triangles anchored at the site j with lateral distance $d \leqslant d_c$, where d_c is the cutoff scale that can remain close to the lattice constant for a gapped system (see Fig. 1).

The above choice of QLT for CI is motivated by the characteristic response function that defines CI, the Hall conductivity for free fermion systems [11]:

$$\sigma_{xy} = \frac{e^2}{h} \frac{1}{N} \sum_{\triangle jkl} 4\pi i P_{jk} P_{kl} P_{lj} S_{\triangle jkl}, \tag{2}$$

where $P_{ij} \equiv \langle c_i^\dagger c_j \rangle$ is the equal-time two-point correlation between sites i and j, $S_{\triangle jkl}$ is the signed area of the triangle jkl, and N is the total number of sites [37,38]. Hence QLT in Eq. (1) provides input that could contribute to the Hall conductivity, albeit with noisy single Monte Carlo instance data $\tilde{P}_{jk}|_\alpha \equiv \langle c_j^\dagger c_k \rangle_\alpha$. More importantly, since it is constructed from loops, QLT provides only gauge-invariant data to the ANN.

We now train an ANN using QLT from a model that exhibits a topological quantum phase transition (TQPT) between trivial insulator and Chern insulators [39], and probe the ANN to interpret how it has learned. The model Hamiltonian is a tight-binding model on a two-dimensional square lattice [11]:

$$H(\kappa) = \sum_{\vec{r}} (-1)^y c_{\vec{r}+\hat{x}}^\dagger c_{\vec{r}} + [1 + (-1)^y (1-\kappa)] c_{\vec{r}+\hat{y}}^\dagger c_{\vec{r}}$$

$$+ (-1)^y \frac{i\kappa}{2} [c_{\vec{r}+\hat{x}+\hat{y}}^\dagger c_{\vec{r}} + c_{\vec{r}+\hat{x}-\hat{y}}^\dagger c_{\vec{r}}] + \text{H.c.}, \tag{3}$$

where $\vec{r} = (x, y)$ and $0 \leqslant \kappa \leqslant 1$ is a tuning parameter. The $\kappa = 1$ limit is the π-flux square lattice model for a quantum Hall insulator with Chern number $C = 1$, while the $\kappa = 0$ limit reduces to decoupled two-leg ladders. $H(\kappa)$ interpolates between the quantum Hall insulator and the normal insulator with a TQPT at $\kappa = 0.5$. We study a system of size 12×12 lattice spacings.

Figure 1 shows the architecture of our single hidden layer ANN. The trained ANN with weights $w^{(1)}, w^{(2)}$ and biases
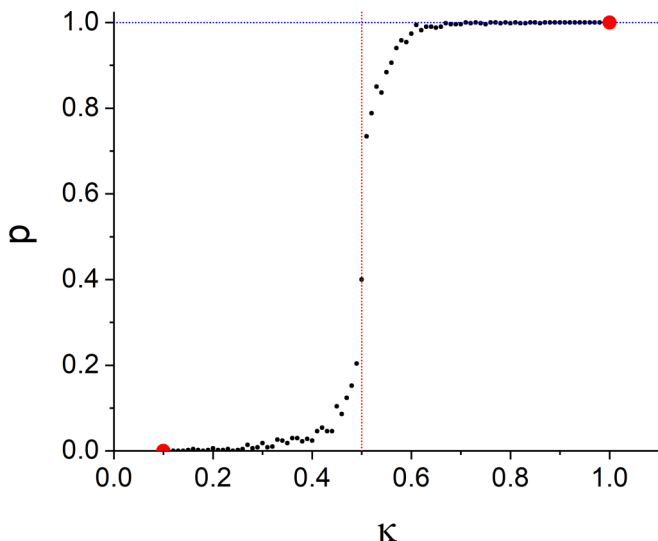
FIG. 2. After successful training on samples from both the quantum Hall insulator at $\kappa = 1.0$ and trivial insulator phases at $\kappa = 0.1$ [the filled circles in red at $(0.1, 0)$ and $(1, 1)$], the RELU ANN reliably provides consistent determination of the phase. p is the probability that the ANN assigns a sample state to be in the quantum Hall phase. The vertical red line is the critical value $\kappa = 0.5$ between the two phases.

$b^{(1)}, b^{(2)}$ transforms a QLT input vector x into output y given by

$$y = \sigma \left[\sum_j w_j^{(2)} \sigma \left(\sum_i w_{ji}^{(1)} x_i + b_j^{(1)} \right) + b^{(2)} \right]. \quad (4)$$

$\sigma(z)$ is the nonlinear activation function applied to the neurons, here taken as rectified linear units (RELU), i.e., with

$$\sigma(z) \equiv \mathrm{RELU}(z) = \max(z, 0). \quad (5)$$

[The RELU choice has proven more efficient to train in the context of deep networks, and as well affords a simpler interpretation than the sigmoid function $S(z) = [1 + \exp(z)]^{-1}$. used in Ref. [11].]

Figure 2 shows the typical phase recognition by a successfully trained ANN. Here, the network confidence p of the ground state being a Chern insulator at the given model parameter κ is assessed by taking the average of the neural network outputs for 500 independent input samples at the given parameter value. An ANN trained at the two marked training points reliably detects the transition from trivial insulator to CI at $\kappa = 0.5$.

Our interpretation of a trained ANN begins with inspection of the final weights $w_j^{(2)}$ for each neuron j. Despite initialization with random fluctuations [Fig. 3(a)], we often find the $w_j^{(2)}$ to be largely concentrated on a single hidden layer neuron, which we label $j_{\max}$ [see Fig. 3(b)]. With the ANN output determined by this single neuron, we can ignore the rest of the hidden layer neurons, and trace the firing of the output neuron back to the firing condition of $j_{\max}$, in turn encoded in the weights $w_{j_{\max}, i}^{(1)}$ and bias $b_{j_{\max}}^{(1)}$. Here $i = (\vec{r}, t)$ labels the inputs according to their lattice sites $\vec{r}$ and the triangles $t \in [1, D(d_c)]$. The real and imaginary parts are also treated

as separate inputs. Figures 3(c) and 3(d) show the distribution of selected $w_{j_{\max}, i}^{(1)}$ for each lattice site $\vec{r}$ at random initialization and after the training. By inspection of $w_{j_{\max}, i}^{(1)}$ for all i's we find that the neuron $j_{\max}$ ended up weighting as most significant the *imaginary parts* of $\tilde{P}_{jk}|_\alpha \tilde{P}_{kl}|_\beta \tilde{P}_{lj}|_\gamma$ coming from the smallest triangles jkl, namely, $j = \vec{r}$, $k = \vec{r} \pm \hat{x}$, and $j = \vec{r} \pm \hat{y}$. Moreover, $w_{j_{\max}, i}^{(1)}$ is approximately evenly distributed across all four $d = 1$ triangles and across all real-space positions $\vec{r}$, as shown by the four colors for the four triangles in Fig. 3(d). For all of the other inputs, including all real parts, the associated weights $w_{j_{\max}, i}^{(1)}$ are close to zero, as shown in magenta in Fig. 3(d).

The observation of significant separation between the weights $w_{j_{\max}, i}^{(1)}$ of the smallest triangles and the rest implies that we can approximate the firing condition by concentrating on those smallest triangles. Moreover, since the weights for the smallest triangles cornered at each site are roughly the same, we can approximate the firing condition for the Chern Hall insulator by the following criterion:

$$w_{j_{\max}}^{(2)} \max \left[\bar{w}_{j_{\max}}^{(1)} \sum_{i = \vec{r}, \pm\hat{x}, \pm\hat{y}} \mathrm{Im} P_{\vec{r}\pm\hat{y},\vec{r}} P_{\vec{r},\vec{r}\pm\hat{x}} P_{\vec{r}\pm\hat{x},\vec{r}\pm\hat{y}} + b_{j_{\max}}^{(1)}, 0 \right]$$
$$+ b^{(2)} > 0. \quad (6)$$

In the above, we have replaced the $w_{j_{\max}, i}^{(1)}$ over all positions $\vec{r}$ and four triangles (labeled by the relative position of the other two vertices $\pm\hat{x}$ and $\pm\hat{y}$) with their average $\bar{w}_{j_{\max}}^{(1)}$. Reading off the weights and biases from a learned ANN and inserting their values, $\bar{w}_{j_{\max}}^{(1)} = -0.208$, $w_{j_{\max}}^{(2)} \sim -4.84$, $b_{j_{\max}}^{(1)} \sim 3.73$, and $b^{(2)} \sim 9.03$ into Eq. (6) yields

$$-4.84 \max \left[-0.208 \sum_{d_{\triangle jkl}=1} \mathrm{Im} P_{jk} P_{kl} P_{lj} + 3.73, 0 \right] + 9.03 > 0$$
$$\Leftrightarrow \frac{4\pi}{N} \sum_{d_{\triangle jkl}=1} -\mathrm{Im} P_{jk} P_{kl} P_{lj}/2 > 0.4, \quad (7)$$

where $N = L^2 = 144$. Considering $S_{\triangle jkl} = 1/2$ for the $d_{\triangle jkl} = 1$ triangles, the above criterion implies that the ANN relied on the imaginary part of the QLT input associated with the smallest triangles, and diagnosed the system to be a Chern insulator when their contribution to the Chern number was substantial. This is a reasonable and efficient diagnosis, given the exact formula Eq. (2) for the invariant in the position basis.

Our successful interpretation of the ANN's learning in the case of the Chern insulator was thus enabled by two aspects of our approach: (i) the QLT was effective in providing the relevant features, and (ii) the exact topological invariant was known in the local basis and so guided our interpretation. The effectiveness of the QLT is reflected in the ANN learning a linear function based on a single $j_{\max}$ neuron.

III. NONLINEAR ML: THE TWO-DIMENSIONAL $\mathbb{Z}_2$ TOPOLOGICAL INSULATORS

We now turn to a topological insulator with no known expression for the topological invariant in the position basis: the time-reversal-invariant $\mathbb{Z}_2$ TI in two dimensions, also known
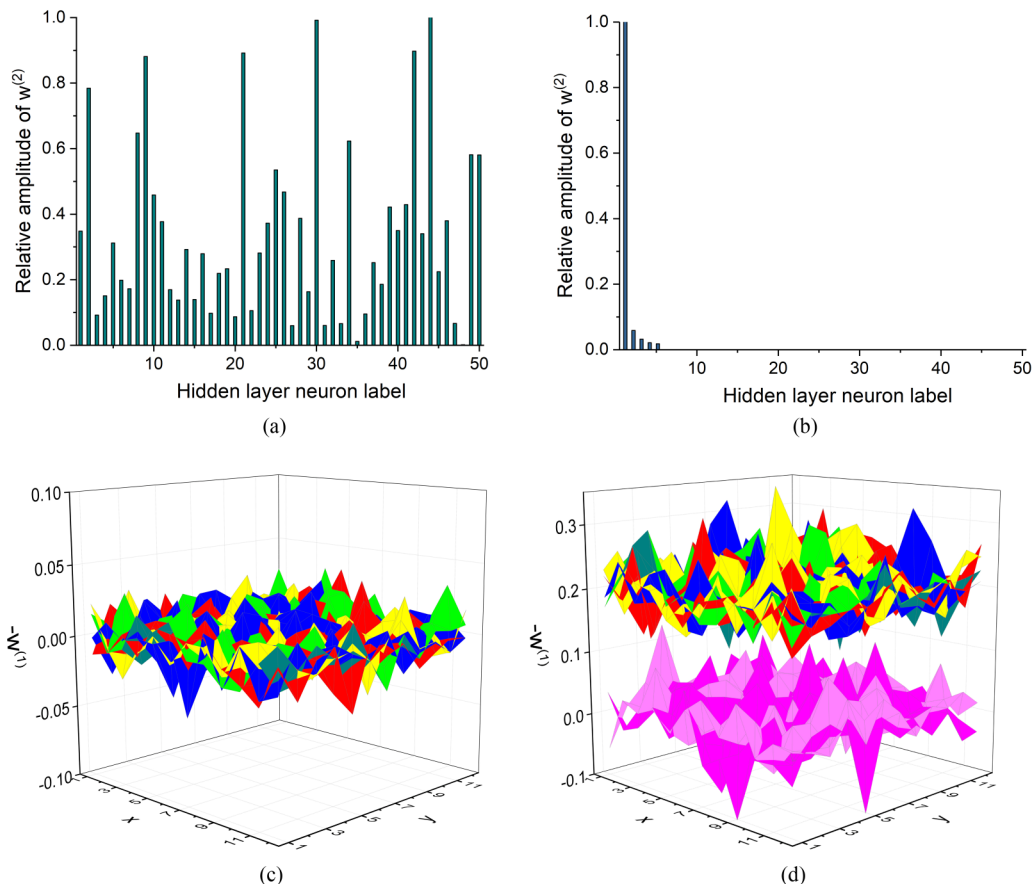
FIG. 3. (a), (b) The distribution of the (absolute values of) weights $w_j^{(2)}$ between the hidden layer neurons and the output neuron in an ANN with RELU neurons, (a) at random initialization before, and (b) after training. In the case shown, $j_{\max}$ is the very first neuron with $j = 1$. (c), (d) The distribution of weights $w_{j_{\max},i}^{(1)}$ associated with the input i for the dominant neuron $j_{\max}$, (c) at random initialization before, and (d) after training. Here i labels the type of triangle, the real and imaginary parts of the contributions, and the lattice position $\vec{r} = (x, y)$. The major contribution to the output comes from the *imaginary parts* of the correlations from the four $d = 1$ triangles (see Fig. 1), whose weights are depicted in red, yellow, green, and blue, respectively. In comparison, the weights associated with other inputs are much closer to zero even after training, as illustrated by the $w_{j_{\max},i}^{(1)}$ distributions in magenta.

as the quantum spin-Hall insulator, with a quantized spin-Hall conductance and a vanishing charge Hall conductance. The characteristic $\mathbb{Z}_2$ topological invariant is only known as a loop integral

$$I = \frac{1}{2\pi i} \oint d\vec{k} \cdot \partial_{\vec{k}} \ln[\mathrm{Pf}(\vec{k}) + i\delta], \qquad (8)$$

of the phase winding of the Pfaffian $\mathrm{Pf}(\vec{k}) = \mathrm{Pf}[\langle u_n(\vec{k})|\Theta|u_m(\vec{k})\rangle]$ over a contour in momentum space enclosing half the Brillouin zone [36], where n and m are band labels, and Θ is the time-reversal operator. A position-basis expression for the $\mathbb{Z}_2$ index I, the counterpart of Eq. (2) for the Chern number, is not known in general.

In the presence of spin s_z conservation, a $\mathbb{Z}_2$ TI is equivalent to two copies of Chern insulators, with $\sigma_{xy}^{\uparrow} = 1$ for the spin-up ($s_z = \uparrow$) electrons and an antichiral quantum Hall insulator $\sigma_{xy}^{\downarrow} = -1$ for the spin-down electrons ($s_z = \downarrow$). Hence Eq. (2) for each spin component will serve as a position-basis expression for the $\mathbb{Z}_2$ index. It is known, however, that the $\mathbb{Z}_2$ TI state is well defined through the momentum space

expression for the $\mathbb{Z}_2$ index Eq. (8), even when the Rashba spin-orbit coupling breaks s_z conservation.

We turn to the physical response of an effective spin current to a transverse electric field: the spin-Hall conductivity. Shi *et al.* [40] introduced the effective spin current in spin-orbit coupled systems, in the absence of average torque, as a time derivative of the spin-displacement operator: $\hat{\mathcal{J}}_s = \frac{d\hat{r}\hat{s}_z}{dt} = 1/\hbar[H, \hat{r}\hat{s}_z]$. The flat-band Hamiltonian is defined as $\hat{H}' = 1 - \hat{P}$, where $\hat{P}$ is the projection operator onto the valence band of H. $\hat{H}'$ is adiabatically connected to the model Hamiltonian H, and thus shares the topological properties such as the spin-Hall conductivity and the $\mathbb{Z}_2$ topological index. According to the Kubo formula, the spin-Hall conductivity of H' is

$$\mathrm{tr}P[P, x\hat{s}_z][P, y]$$
$$= \sum_{\triangle jkl, s_j^z, s_k^z, s_l^z} P_{j,s_j^z;k,s_k^z} P_{k,s_k^z;l,s_l^z} P_{l,s_l^z;j,s_j^z} \left(\hat{s}_j^z x_j - \hat{s}_k^z x_k\right)(y_k - y_l),$$
$$(9)$$

where $P_{j,s_j^z;k,s_k^z} \equiv \langle c_{j,s_j^z}^{\dagger} c_{k,s_k^z} \rangle$ are the two-point correlators of H and the summation is over all triangles, with vertices j,
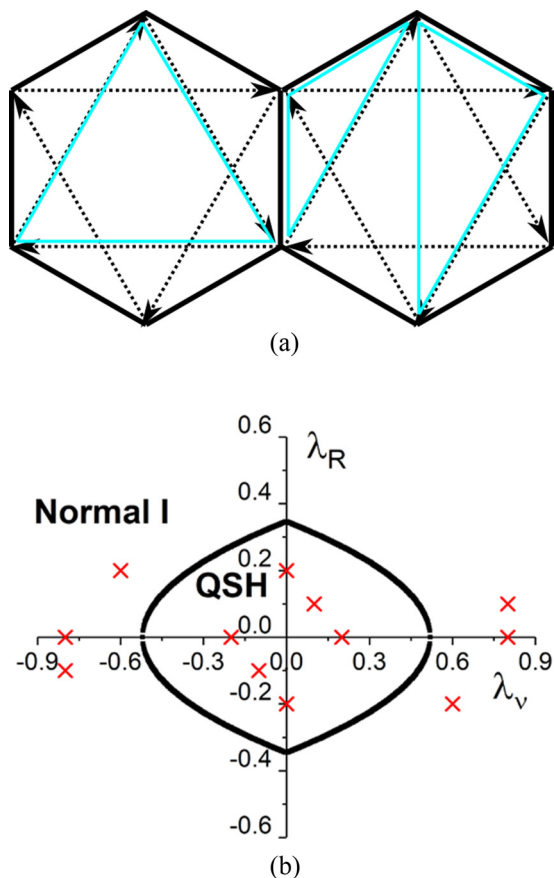
(a)

(b)

FIG. 4. Upper panel: the honeycomb lattice for the model Hamiltonian in Eq. (11). The sign ν_{ij} of the spin-orbit interaction between the next-nearest neighbors are positive along the arrows and negative against the arrows. The three smallest types of triangles are illustrated in cyan. Lower panel: the phase diagram of the tight-binding model in Eq. (11) with $t = 1$ and $\lambda_{\rm SO} = 0.1$. The red crosses indicate the parameters used for the supervised machine learning training set.

k, and l. Since there is no spin-quantization direction in the presence of Rashba spin-orbit coupling, we propose to use the following QLT to probe the quantum spin-Hall (QSH) effect:

$$
\text{(QLT for QSH):} \quad
\begin{aligned}
s_j^x \tilde{P}_{jk}|_\alpha \tilde{P}_{kl}|_\beta \tilde{P}_{lj}|_\gamma, \\
s_j^y \tilde{P}_{jk}|_\alpha \tilde{P}_{kl}|_\beta \tilde{P}_{lj}|_\gamma, \\
s_j^z \tilde{P}_{jk}|_\alpha \tilde{P}_{kl}|_\beta \tilde{P}_{lj}|_\gamma.
\end{aligned}
\quad (10)
$$

As in the last section, $\tilde{P}$'s are to be evaluated at independent Monte Carlo steps, and we focus only on the smallest triangles jkl, which should account for the major contributions, due to the exponentially decaying correlations in a gapped system.

We now employ the QLT in Eq. (10) to train an ANN to recognize the $\mathbb{Z}_2$ TI phase in the two-parameter phase space of the Kane-Mele model of Ref. [36]:

$$
\begin{aligned}
H = t \sum_{\langle ij \rangle} c_i^\dagger c_j + i\lambda_{\rm SO} \sum_{\langle\langle ij \rangle\rangle} \nu_{ij} c_i^\dagger S^z c_j \\
+ i\lambda_R \sum_{\langle ij \rangle} c_i^\dagger (\vec{s} \times \hat{d}_{ij})_z c_j + \lambda_\nu \sum_{\langle i \rangle} \xi_i c_i^\dagger c_i,
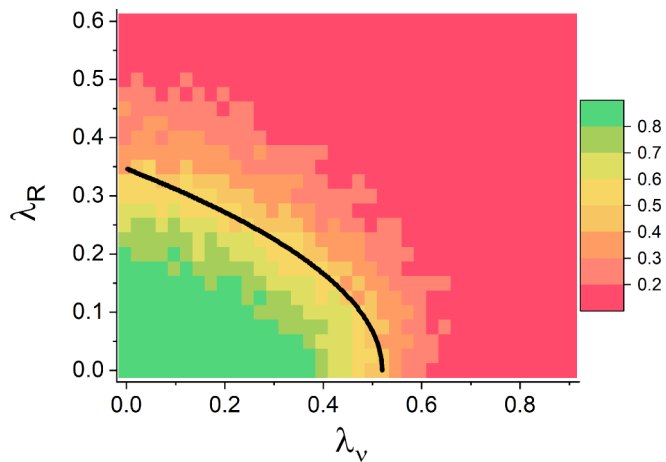\end{aligned}
\quad (11)
$$

FIG. 5. After supervised machine learning, the ANN predicted likelihood of the Hamiltonian in Eq. (11) being a $\mathbb{Z}_2$ TI in the $[\lambda_\nu, \lambda_R]$ parameter space, the upper right quadrant of the phase diagram in Fig. 4 lower panel.

where $t = 1$ is the nearest-neighbor hopping amplitude, $\lambda_{\rm SO} = 0.1$ is the spin-orbit coupling between the next-nearest neighbors, the sign $\nu_{ij} = \pm 1$ depends on whether the direction is along or against the arrow (see Fig. 4 upper panel), λ_ν is a staggered potential, and λ_R is the Rashba term. For $\lambda_R = 0$, s^z is a good quantum number, and the model reduces to two independent copies of quantum Hall insulators [39]. In the presence of a finite λ_R Rashba term, however, s^z is no longer a good quantum number, and there is no longer a conserved spin. For small ratios of $\lambda_R/\lambda_{\rm SO}$ and $\lambda_\nu/\lambda_{\rm SO}$, the model Eq. (11) is known to realize a two-dimensional $\mathbb{Z}_2$ TI, and otherwise a normal insulator [36]. The exact phase diagram in Fig. 4, lower panel, is obtained through an explicit evaluation of the $\mathbb{Z}_2$ index given in Eq. (8). Here we attempt to reproduce the phase diagram using QLT-based machine learning.

We consider variational Monte Carlo samples of $s_j^x P_{jk} P_{kl} P_{lj}$, $s_j^y P_{jk} P_{kl} P_{lj}$, and $s_j^z P_{jk} P_{kl} P_{lj}$ over the three smallest types of triangles (see Fig. 4, upper panel) and feed the corresponding QLT inputs into an ANN with a single hidden layer of RELU, as in the previous section. The training set consists of QLTs obtained from exact wave functions at the training points marked in Fig. 4, lower panel. To avoid approximate conservation of spin s^z, we randomly cycle s^x, s^y, and s^z of the training samples during supervised machine learning. The trained ANN is then applied to QLT samples obtained from the phase space between $\lambda_\nu \in [0, 0.9]$ and $\lambda_R \in [0, 0.6]$ to assess the likelihood that the input belongs to a $\mathbb{Z}_2$ TI.

The resulting phase diagram obtained by the ANN is shown in Fig. 5, and is consistent with the exact phase diagram. The benefit of using the QLT-based local input is that this approach allows the investigation of systems with disorder. With the original definition of the $\mathbb{Z}_2$ invariant requiring a momentum space integral, there was no framework to assess whether a realistic system with disorder is a $\mathbb{Z}_2$ TI (although due to its topological nature one would expect the $\mathbb{Z}_2$ TI to be immune to small disorder). Our success in using the QLT-based machine learning to recognize the $\mathbb{Z}_2$ TI paves
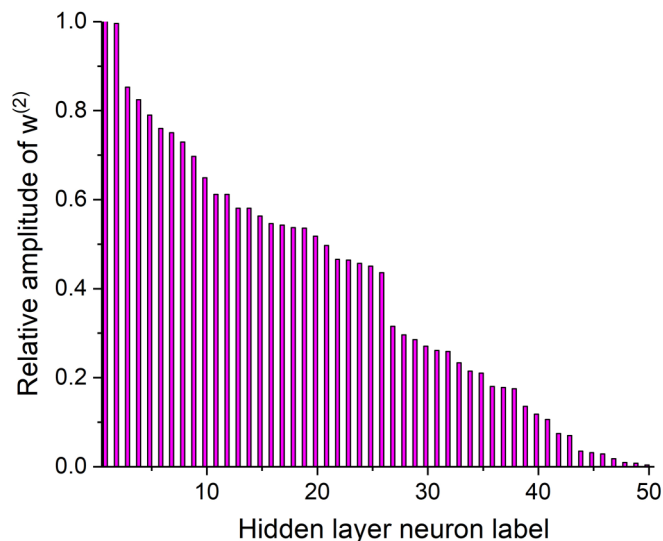
FIG. 6. The distribution of the (absolute values of) weights $w_j^{(2)}$ between the hidden layer neurons and the output neuron in an ANN with RELUS. The decision boundary for the $\mathbb{Z}_2$ TI involves the interplay between multiple hidden layer neurons.

the way for studying realistic models with disorder to discover more $\mathbb{Z}_2$ TI in nature.

To interpret the ANN's learning from the QLT for the $\mathbb{Z}_2$ TI Eq. (10), we again plot the modulus of $w_j^{(2)}$ between the hidden layer RELUS and the output neuron. We now find all the neurons to have substantial weight in the trained ANNs (see Fig. 6). Contrasting the weight distribution in Fig. 6 for $\mathbb{Z}_2$ TI with that for the CI (see the right column of Fig. 3), we see that the ANN for the QSH learns a more complex function of the QLT input. The earlier concentration of $w_j^{(2)}$ to one neuron $j_{\max}$ for the CI means that the ANN effectively expressed a linear function of the QLT as the threshold for that neuron, because the QLT features were already so effective for the CI. That the ANN forms a nonlinear function with the QLT inputs for the QSH implies that the position-based expression for the $\mathbb{Z}_2$ invariant is not a simple linear combination of the QLT Eq. (10). This further motivates the use of the ANN-based approach for the QSH. It is moreover plausible that we can gain insight into the presently unknown position-based expression for the $\mathbb{Z}_2$ invariant, by studying the weights and biases of the function learned by the ANN [41]. It might also be possible to adapt conventional machine learning methods such as the Lasso [42], or more general L^1 regularizations, to enforce sparsity as in Fig. 3(b), and thereby facilitate interpretability.

IV. INTERPRETING NONLINEAR ML: THE $\mathbb{Z}_2$ LATTICE GAUGE THEORY

In this section, we turn to strongly interacting models of lattice gauge theory. Previously, ANN-based machine learning has been used to improve the efficiency of lattice QCD simulations [43] and detected phase boundaries of topological quantum phase transitions [8,12,29], using the mapping between the $T = 0$ quantum problem in two dimensions and the lattice gauge theory in three dimensions. Here we revisit
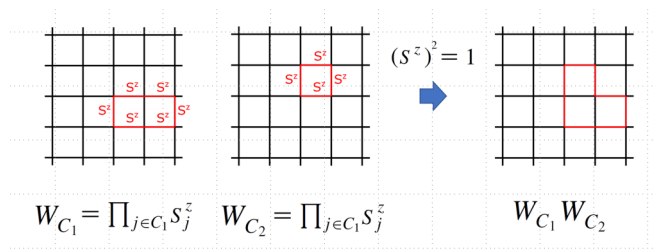

FIG. 7. Wilson loops in the $\mathbb{Z}_2$ lattice gauge theory correspond to the product of spin s^z operators around a closed string. The product of smaller loops can give rise to larger loops.

the ANN-based phase diagram of Ref. [12], where two of us successfully trained an ANN to recognize the $\mathbb{Z}_2$ quantum spin liquid phase, the deconfined phase for the corresponding three-dimensional $\mathbb{Z}_2$ lattice gauge theory. The goal is to interpret what the ANN learns from the semilocal QLT training data. Specifically, the question is how the ANN, using the behavior of small loops as features, can detect a confining phase, ordinarily signaled by the area vs length behavior of large loops, as explained below [44].

For simplicity, we consider the $\mathbb{Z}_2$ lattice gauge theory given by the Hamiltonian on a three-dimensional cubic lattice:

$$\beta H_{3D} = -\lambda_b \sum_j S_j - \lambda_p \sum_p \prod_{j \in p} S_j, \qquad (12)$$

where $S_j = \pm 1$ lives on the bonds of the cubic lattice, and p denotes the square plaquettes. For small values of λ_b, the system has a phase transition at the critical value $\lambda_p \sim 0.76$ between a deconfined phase at large λ_p and a confined phase at small λ_p.

An important physical concept in the lattice gauge theory is the Wilson loop: the path-ordered gauge field integrated around a closed loop. Since the Wilson loops are gauge invariant, they provide meaningful measurements of the lattice gauge theory and a natural dataset for machine learning [43]. For the $\mathbb{Z}_2$ lattice gauge theory, the Wilson loop around a given loop C is defined as

$$W_C = \prod_{j \in C} S_j. \qquad (13)$$

The deconfined phase of the lattice gauge theory is not distinguished by any broken symmetries, but rather by the limiting behavior of the Wilson loop in the thermodynamic limit: in the confined phase, the expectation value of the Wilson loop decays exponentially $\langle W_C \rangle \propto \exp(-A_C)$ as a function of the area A_C enclosed by the loop C, whereas in the deconfined phase, $\langle W_C \rangle \propto \exp(-l_C)$ decays only as the length l_C of the loop C [45].

In what follows, we will take advantage of the Abelian nature of $\mathbb{Z}_2$, which permits small Wilson loops to fuse in a unique channel and form larger ones, consisting of the products of the smaller ones, as in Fig. 7. In a single Monte Carlo snapshot, the sampled values of all Wilson loops can thus be obtained from those of the smallest Wilson loops, $W_p = \prod_{j \in p} S_j$, around the square lattice plaquettes. In accord with this observation, we use the classical Monte Carlo
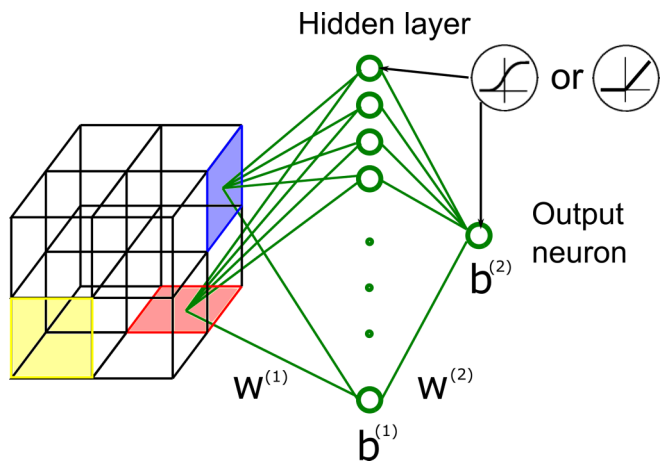
FIG. 8. The application of supervised machine learning on the $\mathbb{Z}_2$ lattice gauge theory uses Monte Carlo samples of the smallest Wilson loops, the plaquettes, as inputs to the ANN. The ANN consists of neurons with RELU activation functions.

samples of the smallest Wilson loops W_p on the $L = 12$ dual lattice as the inputs to the ANN (see Fig. 8).

We use normalized Monte Carlo samples of the set of W_p for the two distinctive phases above ($\lambda_p = 0.83$) and below ($\lambda_p = 0.68$) the critical value as the training set, with $\lambda_b = 0.1$, to perform the supervised learning. In Ref. [12], the phase diagram mapped out for the λ_b/λ_p plane by the optimized ANN was found to be in good agreement with the value of T_c determined by finite-size scaling of Monte Carlo data, giving confidence in that phase diagram. As in the earlier sections, we use a shallow ANN with RELU neurons (see Fig. 8). An inspection of an ANN trained to correctly distinguish phases of the $\mathbb{Z}_2$ lattice gauge theory ($\mathbb{Z}_2$ topological order) shows multiple hidden layer neurons to participate. This is shown in the distribution of weights $w_j^{(2)}$ in Fig. 9. As in the case of the $\mathbb{Z}_2$ TI (see Sec. III), this implies that the decision boundary criterion for the $\mathbb{Z}_2$ lattice gauge theory is also a nontrivial function of the inputs.

To gain insight into the function that the ANN learns, we introduce nonlinearity within the preprocessing so that the target function can be represented approximately linearly. For this, we include asymptotically higher-order terms of the normalized inputs x_i as new inputs, in exchange for reducing the width of the hidden layer. This way, we aim to determine the nature of the nonlinearity captured in Fig. 9. In particular, we can include higher-order terms $x_i x_j$ in the input for further training steps when the inputs x_i and x_j show a strong correlation at the current step, for instance $y(x_i, x_j) + y(x_j, x_i) - 2y(x_i/2 + x_j/2, x_i/2 + x_j/2)$ with all other inputs omitted in the expression. For simplicity, we limit ourselves to the quadratic order of the original inputs, related to W_p for the study of the $\mathbb{Z}_2$ lattice gauge theory. In the meantime, we gradually reduce the hidden layer width of the RELU neural network to approach the linear limit, and eliminate newly included inputs that do not contribute significantly to the output, reducing the width while maintaining performance (see Fig. 10). When the supervised machine learning of the RELU neural network finally converges with a small hidden layer width, the ANN can be interpreted as in the
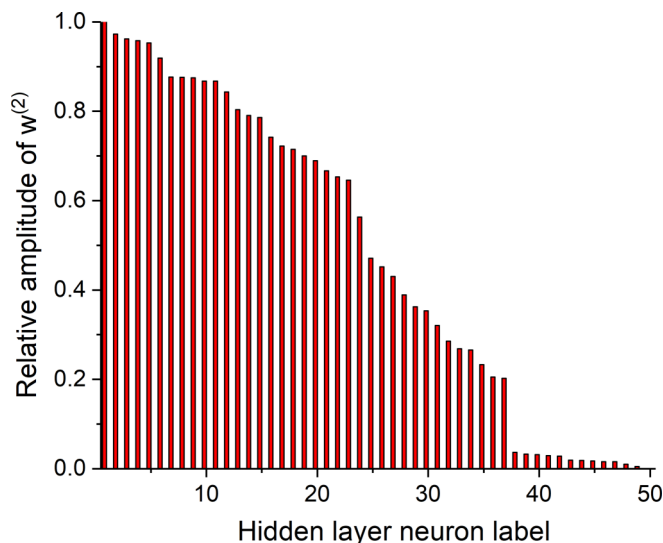

FIG. 9. The distribution of the absolute values of the weights between the hidden layer neurons and the output neuron. The decision boundary for the $\mathbb{Z}_2$ lattice gauge theory ($\mathbb{Z}_2$ topological order) involves the interplay between multiple hidden layer neurons.

linear function formalism of Sec. II. In comparison with Taylor expansion [33,34], our anatomy of the ANN is both automated and relatively cost-efficient, as only higher-order contributions deemed relevant are retained in the end. Also, our algorithm mainly offers a preprocessing of the inputs, which can complement Taylor expansion of the ANN for a nonlinear decision function in terms of the inputs.

The above iterative approach singles out the products of neighboring inputs (see Fig. 11) as new inputs that simultaneously permit the hidden layer width to shrink to as narrow as three neurons wide and contribute with the most weight. Here we averaged over the weights for the inputs of identical geometry to obtain the relative contributions $\bar{w}_{j_{\max}}^{(1)}$ for each type of higher-order inputs. The selection of the higher-order inputs and their weights offers much insight into the learning of the ANN. Firstly, it is notable that products of more distant inputs are left out. The new higher-order inputs are exclusively those that combine two smaller loops from the initial input to form a larger loop. This is a feature that is commensurate with expectation for the Abelian gauge theory, for which small Wilson loops fuse to form larger Wilson loops. Secondly, the fact that the new larger loops acquire larger weight compared to the original small loop indicates that the ANN's criteria are consistent with expectation for a deconfinement transition of the $\mathbb{Z}_2$ lattice gauge theory. For a rigorous identification
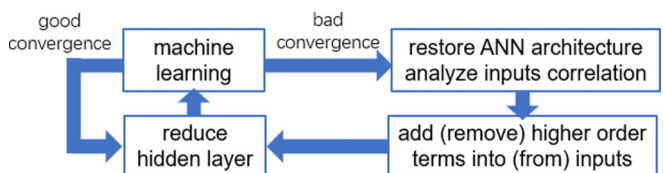

FIG. 10. A supervised machine learning framework that progressively includes higher-order terms of the original inputs to handle the nonlinearity, and reduces the hidden layer width for interpretability.
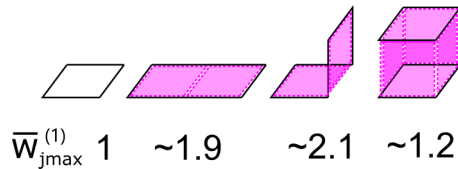
FIG. 11. Higher-order terms of the original inputs showing strong correlations in the machine learning of the deconfinement of the $\mathbb{Z}_2$ lattice gauge theory are mostly the products of local plaquettes that give rise to larger Wilson loops. Such quadratic order terms are included as new inputs, progressively to preprocess the nonlinearity, so that the complexity of the RELU ANN can be reduced for interpretation. The numbers below are the relative $w^{(1)}$ weights averaged over equivalent inputs under translations and rotations.

of the transition, the expectation value of the Wilson loop needs to be calculated for a large loop in the thermodynamic limit, a challenging task for any computational approach. But we have demonstrated that the ANN can discover the deconfinement transition by using small loops, together with slightly larger loops obtained through multiplication, to arrive at a rudimentary yet physically sound judgment.

V. CONCLUSIONS

To summarize, we studied the interpretability of machine learning in the context of three distinct topological quantum phase transitions learned by shallow, fully connected feed-forward ANNs. The quantum phase transitions of interest are between topologically trivial states and a CI, TI, and QSL, respectively. As is well known in the machine learning literature [46], the more expressive the machine learning architectures are, the more opaque are their decision-making criteria. The relatively simple and minimally nonlinear structure of our ANN was able to learn topological phases, aided by QLT: a physically motivated feature selection scheme. For the CI, the criteria the ANN learned amounted to evaluating a noisy version of the relevant topological invariant, which was linear

in the QLT inputs. For the TI and the QSL, on the other hand, the ANNs based their criteria on nonlinear functions of the inputs. For QSL, we determined that the ANN fused the neighboring Wilson loops of the corresponding $\mathbb{Z}_2$ gauge theory in the QLT to form larger Wilson loops.

Our successful interpretation of QLT-based machine learning gives us confidence in the ANN's phase detection by confirming that its decision criterion is guided by key physical properties of the target phases. Our results should serve to encourage wider application of QLT-based machine learning. We note the important role of physical insight that guided the design of QLT feature selection which, in turn, enabled interpretable machine learning. The shallow ANN depth, combined with physical insight, powered the QLT-enabled interpretation of the learning for the CI. An understanding of deconfinement in the thermodynamic limit was critical to the successful interpretation of fully nonlinear machine learning of the QSL. Our approach to interpreting the learning of QSL bears similarity to the variational autoencoder [47,48], and further investigating this similarity could be an interesting future direction. These results also provide hope that interpretable machine learning in the future can instead *inform* our physical insight, when our prior understanding is not sufficient to craft the necessary informative features. We could imagine instead such informative composite features emerging in later layers of a deep neural network fed only naive features, and whose interpretation would then lead to a better theoretical understanding of the underlying physics.

[1] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[2] D.-L. Deng, X. Li, and S. Das Sarma, Phys. Rev. X **7**, 021021 (2017).

[3] X. Gao and L.-M. Duan, Nat. Commun. **8**, 662 (2017).

[4] J. Liu, H. Shen, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 241104(R) (2017).

[5] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nat. Phys. **14**, 447 (2018).

[6] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Phys. Rev. X **8**, 011006 (2018).

[7] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Phys. Rev. Lett. **121**, 167204 (2018).

[8] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).

[9] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Phys. Rev. X **7**, 031038 (2017).

[10] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Sci. Rep. **7**, 8823 (2017).

[11] Y. Zhang and E.-A. Kim, Phys. Rev. Lett. **118**, 216401 (2017).

[12] Y. Zhang, R. G. Melko, and E.-A. Kim, Phys. Rev. B **96**, 245119 (2017).

[13] Y.-T. Hsu, X. Li, D.-L. Deng, and S. Das Sarma, Phys. Rev. Lett. **121**, 245701 (2018).

[14] P. Zhang, H. Shen, and H. Zhai, Phys. Rev. Lett. **120**, 066401 (2018).

[15] N. Sun, J. Yi, P. Zhang, H. Shen, and H. Zhai, Phys. Rev. B **98**, 085402 (2018).

[16] E. Greplova, A. Valenti, G. Boschung, F. Schäfer, N. Lörch, and S. Huber, New J. Phys. **22**, 045003 (2020).

[17] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **85**, 123706 (2016).

[18] O. Balabanov and M. Granath, Phys. Rev. Res. **2**, 013354 (2020).

[19] Y.-H. Tsai, M.-Z. Yu, Y.-H. Hsu, and M.-C. Chung, arXiv:1909.04784.

[20] M. D. Caio, M. Caccin, P. Baireuther, T. Hyart, and M. Fruchart, arXiv:1901.03346.

[21] J. F. Rodriguez-Nieva and M. S. Scheurer, Nat. Phys. **15**, 790 (2019).

[22] P. Mehta and D. J. Schwab, arXiv:1410.3831.

[23] A. Seif, M. Hafezi, and C. Jarzynski, arXiv:1909.12380.

[24] Y. Zhang, A. Mesaros, K. Fujita, S. D. Edkins, M. H. Hamidian, K. Ch'ng, H. Eisaki, S. Uchida, J. C. S. Davis, E. Khatami, and E.-A. Kim, Nature (London) **570**, 484 (2019).

[25] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap, Nat. Phys. **15**, 921 (2019).

[26] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, Nat. Phys. **15**, 917 (2019).

[27] J. B. Goetz, Y. Zhang, and M. J. Lawler, arXiv:1901.11042.

[28] G. Cybenko, Math. Control Signal Syst. **2**, 303 (1989).

[29] P. Ponte and R. G. Melko, Phys. Rev. B **96**, 205146 (2017).

[30] K. Liu, J. Greitemann, and L. Pollet, Phys. Rev. B **99**, 104410 (2019).

[31] J. Greitemann, K. Liu, and L. Pollet, Phys. Rev. B **99**, 060404(R) (2019).

[32] W. Zhang, L. Wang, and Z. Wang, Phys. Rev. B **99**, 054208 (2019).

[33] S. J. Wetzel and M. Scherzer, Phys. Rev. B **96**, 184410 (2017).

[34] C. Wang, H. Zhai, and Y.-Z. You, Sci. Bull. **64**, 1228 (2019).

[35] Y. Wu, P. Zhang, H. Shen, and H. Zhai, Phys. Rev. A **98**, 010701(R) (2018).

[36] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 146802 (2005).

[37] A. Kitaev, Ann. Phys. **321**, 2 (2006).

[38] R. Bianco and R. Resta, Phys. Rev. B **84**, 241106(R) (2011).

[39] F. D. M. Haldane, Phys. Rev. Lett. **61**, 2015 (1988).

[40] J. Shi, P. Zhang, D. Xiao, and Q. Niu, Phys. Rev. Lett. **96**, 076604 (2006).

[41] Please contact the authors for preliminary results.

[42] R. Tibshirani, J. R. Stat. Soc. B **58**, 267 (1996).

[43] P. E. Shanahan, D. Trewartha, and W. Detmold, Phys. Rev. D **97**, 094506 (2018).

[44] We thank Eduardo Fradkin for first raising this question to us.

[45] E. Fradkin and S. H. Shenker, Phys. Rev. D **19**, 3682 (1979).

[46] M. Kuhn and K. Johnson, *Applied Predictive Modeling* (Springer, New York, 2013).

[47] D. P. Kingma and M. Welling, in *Proceedings of the International Conference on Learning Representations* (ICLR.cc, 2014).

[48] D. Jimenez Rezende, S. Mohamed, and D. Wierstra, in *Proceedings of the 31st International Conference on International Conference on Machine Learning* (JMLR.org, 2014), Vol. 32, p. II1278.