

# Learning disordered topological phases by statistical recovery of symmetry

Nobuyuki Yoshioka, Yutaka Akagi, and Hosho Katsura

*Department of Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*



(Received 3 October 2017; published 9 May 2018)

We apply the artificial neural network in a supervised manner to map out the quantum phase diagram of disordered topological superconductors in class DIII. Given the disorder that keeps the discrete symmetries of the ensemble as a whole, translational symmetry which is broken in the quasiparticle distribution individually is recovered statistically by taking an ensemble average. By using this, we classify the phases by the artificial neural network that learned the quasiparticle distribution in the clean limit and show that the result is totally consistent with the calculation by the transfer matrix method or noncommutative geometry approach. If all three phases, namely the  $\mathbb{Z}_2$ , trivial, and thermal metal phases, appear in the clean limit, the machine can classify them with high confidence over the entire phase diagram. If only the former two phases are present, we find that the machine remains confused in a certain region, leading us to conclude the detection of the unknown phase which is eventually identified as the thermal metal phase.

DOI: [10.1103/PhysRevB.97.205110](https://doi.org/10.1103/PhysRevB.97.205110)

## I. INTRODUCTION

Machine learning techniques construct and execute the computational algorithm which optimizes the quantified objective defined by gathered training data to make valuable predictions about previously unseen data. The surging development of new techniques has led to the recognition of its effectiveness in various research fields such as condensed matter physics. Examples of previous studies include the application of the restricted Boltzmann machine to compressed expression of quantum many-body states [1–10], acceleration of Monte Carlo simulation [11–14], and detection of phase transitions by unsupervised learning without teaching the notion of phases to the machine [15–19]. Among them, a problem that draws particular attention is the classification of various phases such as in topological systems [20–25], strongly correlated systems [26,27], and many-body localized systems [28–30].

In this paper, we investigate the quantum phase diagram of 2d noncentrosymmetric superconductors in class DIII with disorder motivated by the recent proposal of candidate materials such as  $\text{Cu}_x\text{Bi}_2\text{Se}_3$  [31,32] and  $\text{FeTe}_x\text{Se}_{1-x}$  [33–35]. The gapless excitation of topological superconductors including class DIII can be described by Majorana edge modes, which attract keen interest from the viewpoint of topological quantum computation [36,37]. While topological invariants in translationally invariant systems have been well studied including their concrete expressions and calculations [38–41], the theoretical understanding in disordered systems is far from complete. In particular, the formulation of Niu-Thouless-Wu, which is an extension to many-body systems and disordered systems [42], is known to break down for class DIII.

Our goal is to determine the phase diagram for finite disorder by applying the artificial neural network (ANN), given the information of phases only in the clean limit. There are two underlying key concepts. The first is the expressibility of the ANN. While the choice of the appropriate network architecture is a training-data-dependent problem [43], it is shown that, for arbitrary data groups  $\{\vec{x}_i, \mathbf{F}(\vec{x}_i)\}$  and arbitrary precision  $\epsilon > 0$ ,

an ANN can be constructed so that its prediction  $\tilde{\mathbf{F}}$  satisfies  $|\tilde{\mathbf{F}}(\vec{x}_i) - \mathbf{F}(\vec{x}_i)| < \epsilon$  [44–46]. The second is the recovery of the translational symmetry by ensemble average. While the translational symmetry is broken in a system with disorder such as a random potential [47], as an ensemble of disorder average the symmetry is *statistically recovered*. Our expectation is that an ANN learned from the data in the clean limit is capable of classifying such ensemble-averaged states. As we show later, the phase diagram obtained from our method is fully consistent with the results in both the transfer matrix (TM) method [48] and the calculation of a  $\mathbb{Z}_2$  index by noncommutative geometry which was recently proposed [49–52].

The rest of the paper is organized as follows. In Sec. II we provide the method to map out the phase diagram. The Hamiltonian for 2d noncentrosymmetric superconductors in class DIII with and without the disorder is introduced here. We also discuss the architecture, input, and output of the ANN which is employed as the classifier. In Sec. III we show the results obtained by performing both ternary and binary classification with the ANN, comparing to those by other two methods. Finally, the summary for the current work and the discussion on the future direction is given in Sec. IV. For completeness, we describe the two other methods to depict the phase diagram in Appendices A and B. Also we compare the result with and without the statistical recovery of symmetry in Appendix C.

## II. METHODS

### A. Bogoliubov–de Gennes Hamiltonian in real space

The Bogoliubov–de Gennes Hamiltonian for 2d noncentrosymmetric superconductors in class DIII in the clean limit is given in real space as [53]

$$H_0 = \sum_{\mathbf{r}} \sum_{k=1,2} \Psi_{\mathbf{r}}^\dagger t_k \Psi_{\mathbf{r}+\mathbf{e}_k} + \sum_{\mathbf{r}} \Psi_{\mathbf{r}}^\dagger v \Psi_{\mathbf{r}}, \quad (1)$$

$$t_1 = ts_3 \otimes \sigma_0 + \frac{i\Delta}{2} s_1 \otimes \sigma_3, \quad (2)$$

$$t_2 = ts_3 \otimes \sigma_0 + \frac{\Delta}{2} s_1 \otimes \sigma_3, \quad (3)$$

$$v = -\mu s_3 \otimes \sigma_0 - \Delta_2 s_2 \otimes \sigma_2. \quad (4)$$

For concreteness, our model is defined on a square lattice with cylindrical boundary conditions. Here,  $\Psi_{\mathbf{r}} = [c_{\mathbf{r}\uparrow}, c_{\mathbf{r}\downarrow}, c_{\mathbf{r}\uparrow}^\dagger, c_{\mathbf{r}\downarrow}^\dagger]^T$  is the Nambu operator with  $c_{\mathbf{r}\alpha}$  denoting an annihilation operator of an electron with spin  $\alpha$  at site  $\mathbf{r}$ ;  $t_{1(2)}$  and  $\mathbf{e}_{1(2)}$  are the hopping matrix and the primitive vector along the  $x$  ( $y$ ) direction with the transfer integral  $t$  and the helical  $p$ -wave coupling  $\Delta$ . The Pauli matrices  $s_k$  and  $\sigma_k$  ( $k = 0, 1, 2, 3$ ) operate on particle-hole and spin space, respectively. The on-site term,  $v$ , consists of the chemical potential  $\mu$  and the  $s$ -wave pairing  $\Delta_2$ . The mixture of the spin-singlet and the spin-triplet pairings are caused by the broken inversion symmetry [54]. Note that in the Hamiltonian the following symmetries are present: even particle-hole symmetry (PHS), odd time-reversal symmetry, and chiral symmetry. Thus the topological property is characterized by the  $\mathbb{Z}_2$  topological invariant [40,41,55–58]. For Eq. (1), we find that the system is in the  $\mathbb{Z}_2$  phase at  $2 - 2\sqrt{1 - (\Delta_2/\Delta)^2} < |\mu| < 2 + 2\sqrt{1 - (\Delta_2/\Delta)^2}$  if  $|\Delta_2/\Delta| < 1$  [59]. As the on-site disorder, we introduce a random potential with the amplitude distributed uniformly within the width  $W$ , namely,

$$H_W = \sum_{\mathbf{r}} \Psi_{\mathbf{r}}^\dagger (W_{\mathbf{r}} s_3 \otimes \sigma_0) \Psi_{\mathbf{r}} \quad \text{for } W_{\mathbf{r}} \in [-W/2, W/2]. \quad (5)$$

Consequently, the total Hamiltonian takes the form  $H = H_0 + H_W$ .

Note that once the disorder is turned on, the wave number is no longer a good quantum number and thus the formula for the Kane-Mele invariant is no longer applicable. It is known that moderate randomness in spin rotational symmetry broken systems may cause destructive interference of time-reversal paths of the quasiparticle, suppressing the backscattering and thereby leading the system to show metallic behavior (weak antilocalization) in 2d [60–62]. In particular, the “insulator-metal” transition from the  $\mathbb{Z}_2$  phase, in which Majorana fermions pinned to the disorder percolate, gives rise to the so-called Majorana metal phase [63]. In 2d, the thermal conductivity grows logarithmically with the system size, which is understood as a consequence of the extended behavior of the quasiparticle over the whole system. Actually, the metallic property of thermal transport arises also when the bulk gap is closed in the clean limit. Thus, all of these will be collectively referred to as the thermal metal (ThM) phase in the following.

### B. Classification by artificial neural network

An artificial neural network (ANN) is a nonlinear function that takes an input  $\mathbf{x}$  to compute an output  $\mathbf{y}$  through sequential mappings by layers of “neurons” [46]. A neuron itself is a nonlinear function that applies the activation function to each element of the weighted input  $\mathbf{z} = W\mathbf{x}$ , and a set of neurons that share the identical weight matrix is called a layer. In the following, we denote the operations corresponding to activation and weight matrix for the  $i$ th layer as  $\mathcal{A}_i$  and  $\mathcal{W}_i$ , respectively. An ANN with layers that can be uniquely numbered according to the order of input and output and do not

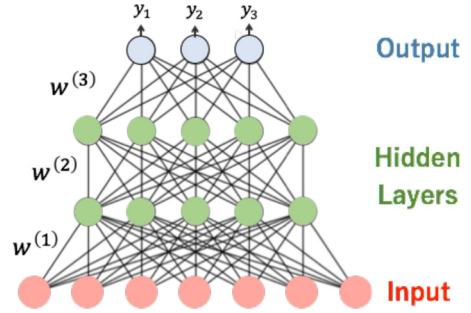


FIG. 1. The architecture of a feedforward artificial neural network with two hidden layers, at which the input data are compressed to extract some abstract feature for classification. The activation of the output layer is the softmax function so that the sum is unity, allowing us to interpret the confidence of the machine.

include any intralayer processing is referred to as a feedforward ANN. In this paper, we apply the feedforward ANN with two hidden layers. (See Fig. 1 for a graphical understanding of the architecture.) The output is calculated as

$$\mathbf{y} = \mathcal{A}_3(\mathcal{W}_3 \mathcal{A}_2(\mathcal{W}_2 \mathcal{A}_1(\mathcal{W}_1 \mathbf{x}))). \quad (6)$$

In our architecture, the activation functions of the hidden and output layers are the rectified linear unit (ReLU) and the softmax function, respectively. The definitions are given as

$$\text{ReLU}(z) = \max(0, z) \quad \text{for } \mathcal{A}_1, \mathcal{A}_2, \quad (7)$$

$$\text{softmax}(z_j; \mathbf{z}) = \exp(-z_j) / \sum_i \exp(-z_i) \quad \text{for } \mathcal{A}_3. \quad (8)$$

Next, we discuss the training process of the machine. The parameters  $\mathcal{W}$  are tuned via minimization of the cost function. This quantifies the performance of the machine which classifies the  $\mathbb{Z}_2$  and the trivial phases in the binary classification scheme, and additionally the ThM phase in the ternary classification scheme. In a classification problem with the current network architecture, the cross-entropy function is widely used [64] due to its convenience. In our paper, we also employ this function with the regularization term. Namely, the cost function is given as

$$\begin{aligned} \mathcal{L}(\mathcal{W}) = & - \sum_{j=1}^{(\#\text{data})} \sum_{k=1}^{(\#\text{class})} \hat{y}_j^{(k)} \log y_j^{(k)}(\mathbf{x}_j; \mathcal{W}) / (\#\text{data}) \\ & + \lambda \sum_{i=1}^{(\#\text{layers})} |\mathcal{W}^{(i)}|^2. \end{aligned} \quad (9)$$

Here,  $y_j^{(k)}$  is the output for the  $k$ th label by the ANN, or “the confidence of the machine,” for the  $j$ th input training or test data, which is modified by the optimization of  $\mathcal{L}$ . On the other hand,  $\hat{y}_j^{(k)} = \delta_{k, l_j}$  for the correct label  $l_j$  denotes the corresponding phase for the data, and hence is constant throughout the training. The second term, or the L2 regularization, suppresses the amplitude of the weight parameters, preventing the machine from overfitting to the training data. The parameters are updated by minibatch gradient descent with batch size 40 as  $\mathcal{W}_{j,k}^{(i)} \rightarrow \mathcal{W}_{j,k}^{(i)} - \eta(\partial \mathcal{L} / \partial \mathcal{W}_{j,k}^{(i)})$ , where  $\eta$  is the learning rate that is controlled by the AdaGrad method

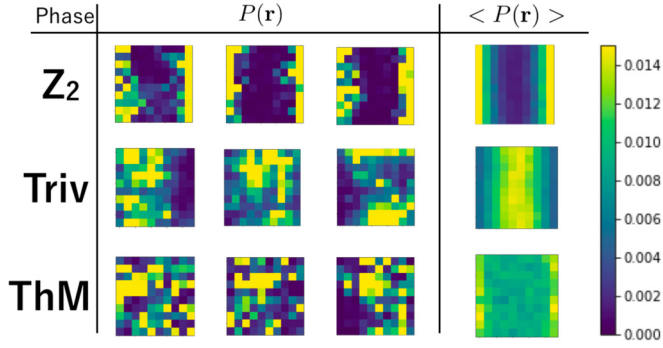


FIG. 2. Typical single-shot quasiparticle distribution of the first excited state,  $P(\mathbf{r})$ , and its disorder average,  $\langle P(\mathbf{r}) \rangle$ , over 500 realizations of random configurations. The parameters are taken from deep inside the phases as  $(\mu, W) = (2, 9), (6, 5), (2, 18)$  with  $\Delta = 3$  and  $\Delta_2 = 0$  from the top. The system size is taken as  $10 \times 10$ .

to efficiently reach the global minimum [65]. Furthermore, we apply the dropout method to avoid overfitting [66].

### C. Input data for machine

Adopted as the input data  $\mathbf{x}$  is the disorder average over  $N_r$  realizations of the spatial distribution of the quasiparticle,  $P(\mathbf{r})$ , corresponding to the first excited state. Our expectation is that the qualitatively different behavior of the quasiparticle gives the machine sufficient information to discriminate phases. The bulk-edge correspondence in the  $\mathbb{Z}_2$  phase, for instance, assures the robust edge localization of the low-lying states across the zero energy. Furthermore, the behaviors in other two phases, namely the bulk localization in the trivial phase and the delocalization of the quasiparticle over the system due to the bulk gap closing in the ThM phase, encourage us to consider  $P(\mathbf{r})$  for the lowest excitation as an appropriate input for the machine.

Let us consider the eigenstate  $|\psi\rangle$  satisfying  $H|\psi\rangle = E_1|\psi\rangle$  with the lowest  $E_1 > 0$ . The degeneracy, if it exists, is lifted up to time-reversal symmetry, and the two states are identical in terms of the quasiparticle distribution, namely,

$$P(\mathbf{r}) = |\psi_{\uparrow}^e(\mathbf{r})|^2 + |\psi_{\downarrow}^e(\mathbf{r})|^2 + |\psi_{\uparrow}^h(\mathbf{r})|^2 + |\psi_{\downarrow}^h(\mathbf{r})|^2, \quad (10)$$

where the super(sub)script denotes the degree of freedom in the Nambu (spin) space. Some examples of single disorder realization  $P(\mathbf{r})$  and its disorder average  $\langle P(\mathbf{r}) \rangle$  for  $N_r = 500$  are shown in Fig. 2.

While it is difficult to find an evident pattern in respective  $P(\mathbf{r})$  due to the randomness, we expect that the translational symmetry is *statistically recovered* by taking the disorder average. For instance, the bulk-edge correspondence assures the Majorana edge mode in the  $\mathbb{Z}_2$  phase, which is robust against perturbation unless the bulk gap closes. The quasiparticle is localized at the edge although the amplitude of  $P(\mathbf{r})$  may become uneven along the circumference of the cylinder under spatial inhomogeneity. Such a fluctuation is eliminated by considering  $\langle P(\mathbf{r}) \rangle$ , which we confirm from the top row of Fig. 2. Furthermore, the localization in the bulk for the middle row indicates the thermal insulating property of the trivial phase, and the extension of the quasiparticle over the whole

system in the bottom row reflects the metallic behavior of the ThM.

We classify the phases by feeding  $\langle P(\mathbf{r}) \rangle$  to the ANN which learned the labels of  $P(\mathbf{r})$  in the clean limit. Both binary and ternary classification are considered.

## III. RESULTS

### A. Ternary classification

First, we carry out the ternary classification at finite  $\Delta_2$ . For  $|\Delta_2| < |\Delta|$ , the bulk gap is closed when (i)  $|\mu| < 2 - 2\sqrt{1 - (\Delta_2/\Delta)^2}$  or (ii)  $2 + 2\sqrt{1 - (\Delta_2/\Delta)^2} < |\mu| < 4\sqrt{1 - (\Delta_2/\Delta)^2}/2$ , and the system shows metallic behavior [53,59]. We focus on  $\Delta = 3, \Delta_2 = 2$  and feed three phases, namely the  $\mathbb{Z}_2$ , trivial, and ThM, to the machine, expecting to predict the whole phase with high confidence. Shown in Fig. 3(a) is the average output of 200 ANNs which takes  $\langle P(\mathbf{r}) \rangle$  with  $N_r = 500$  as the input. Each ANN is trained independently in a stochastic manner using the data from the clean limit indicated in Fig. 3(b). Only the region  $\mu \geq 0$  is shown since the phase diagram is symmetric with respect to  $\mu = 0$ . The black dots are the transition point obtained from the reliable TM method. (See Appendices A and B for details of other two methods.) Remarkably the machine has successfully learned their characteristics even in the vicinity of the phase boundaries and fully extended the phase diagram. As is obvious from Figs. 3(c) and 3(d), classifying  $X_1$  and  $X_2$ , or  $Y_1$  and  $Y_2$ , with a comparable precision is beyond our cognitive ability.

Next, let us focus along  $\mu = 2$ . In the clean limit, the system is in the  $\mathbb{Z}_2$  and enters the ThM and trivial phases sequentially by increasing the disorder, which is accurately captured by the ANN. The blurred output at  $W \sim 15$  between the ThM and trivial phases is attributed to the larger fluctuation of the data, which is suppressed by increasing  $N_r$ . Other  $\mathbb{Z}_2$ -ThM and ThM-trivial phase boundaries are nicely reproduced.

Furthermore, the weak disorder region between the  $\mathbb{Z}_2$  and the trivial phase at  $\mu \sim 3.5$  is unambiguously classified as the ThM. Let us emphasize again that this is attributed to the statistical recovery of the translational symmetry in the input data. Merely taking the average of the output is insufficient. (See Appendix C for further discussion.) Note that such close parallel boundaries require extra effort on the other two methods; determining the peak of the localization length, which diverges with the system size, by the TM becomes difficult due to the broadening by the finite-size effect, and that the noncommutative geometry approach does not work for critical phases.

### B. Binary classification

To examine the binary classification by the ANN, we consider  $\Delta_2 = 0$  at which the ThM phase is absent in the clean limit. Quasiparticle distributions are generated at  $\mu \in [0.5, 3.5]$  and  $[6.0, 10.0]$  for the  $\mathbb{Z}_2$  and trivial phases, respectively. The result is shown in Fig. 4. As is expected, the machine reproduces the  $\mathbb{Z}_2$ -trivial phase boundary not only in the clean limit, i.e., the transition point  $\mu = 4$ , but also at  $W > 0$ , which is obtained by the TM and the noncommutative geometry approach [67].

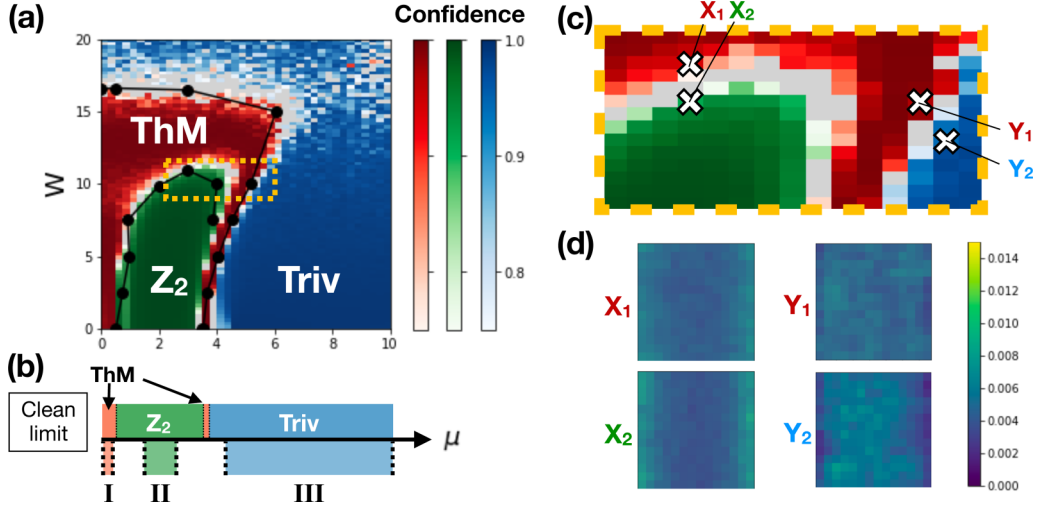


FIG. 3. (a) Average outputs of 200 ternary-classifying ANNs trained with the clean-limit data for  $t = 1, \Delta = 3, \Delta_2 = 2$ . The color of each point  $(\mu, W)$  indicates the confidence for the thermal metal (red),  $\mathbb{Z}_2$  (green), and trivial (blue) phase. The machine is highly confident of each phase but confused at the boundary. (b) The parameter  $\mu$  of 1000 training data with system size  $14 \times 14$  is uniformly distributed within I:  $[0.0, 0.3]$ , II:  $[1.0, 2.5]$ , and III:  $[4.0, 10.0]$ . During the training scheme, the network is tested by the data generated along  $\mu \in [0.0, 10.0]$  in the clean limit, resulting in accuracy over 90%. (c) Enlargement of the region surrounded by the orange dotted line in (a). (d) The averaged inputs  $\langle P(\mathbf{r}) \rangle$  for  $N_r = 500$  in the vicinity of the boundaries. The parameters  $(\mu, W)$  are given as  $X_1$ : (3, 11.5),  $X_2$ : (3, 10.75),  $Y_1$ : (5.25, 10.5),  $Y_2$ : (5.5, 10).

The drop of confidence along  $\mu = 0$  is also observed. This is understood as the  $\mathbb{Z}_2$ - $\mathbb{Z}_2$  transition line, which is consistent with the analysis of the staggered fermion model for class D [68]. Note that such a transition that lacks the change in the size dependence on the thermal conductivity or localization length is very difficult to detect even by the TM method.

The most remarkable confusion appears above the  $\mathbb{Z}_2$  phase, e.g.,  $\mu = 5$ , which clearly suggests phase transition. [See the gray region in Fig. 4(a).] While the output in the trivial phase at small disorder is close to unity, we observe that the confidence in the gray region is far below 1 regardless of the average number for input or the machine. Such a confusion implies the

qualitatively different feature from the trivial phase, namely, the consequence of entering a completely different phase. To reinforce this argument, we add two critical points, i.e.,  $\mu = 0$  for the  $\mathbb{Z}_2$ - $\mathbb{Z}_2$  and  $\mu = 4$  for the  $\mathbb{Z}_2$ -trivial transition points, as the third label. We observe in Fig. 4(b) that  $\mathbb{Z}_2$ - $\mathbb{Z}_2$  and  $\mathbb{Z}_2$ -trivial critical lines are present for finite  $W$ , and also that the previously confused region above the  $\mathbb{Z}_2$  phase exhibits the extended behavior of the quasiparticle by the ANN. Hence, this region is concluded as the thermal metal phase, which is also confirmed from the TM. We note that the  $\mathbb{Z}_2$ -ThM phase boundary predicted by the machine is quantitatively consistent with the result by the numerical calculation of TM.

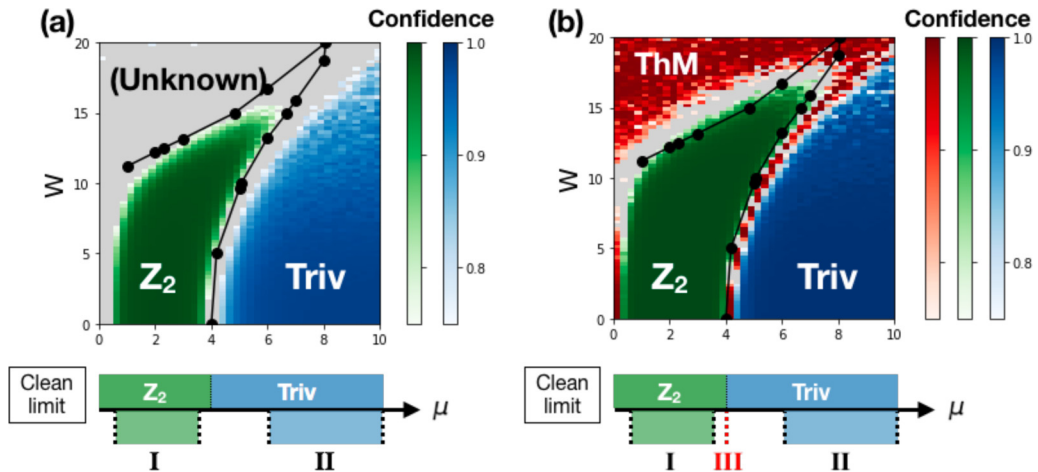


FIG. 4. Average output of 200 binary-classifying ANNs trained with the clean-limit data for  $t = 1, \Delta = 3, \Delta_2 = 0$ . The parameter  $\mu$  of 1000 training data with system size  $14 \times 14$  is uniformly distributed within (a) I:  $[0.5, 3.5]$  and II:  $[6.0, 10.0]$ , (b) I, II, and III:  $\mu = 0.0, 4.0$ , each corresponding to the  $\mathbb{Z}_2$  (green), the trivial (blue) phase, and the critical point (red). The performance of the machine is monitored with the test data generated at  $\mu \in [0.0, 10.0]$  in the clean limit, resulting in over 95% accuracy. The outputs above 0.75 for  $\langle P(\mathbf{r}) \rangle$  with  $N_r = 500$  are indicated by the depth of the color, and merely gray for below 0.75.



#### IV. DISCUSSION

In this work, the use of the ANN to the classify phases of 2d noncentrosymmetric superconductors in class DIII with the disorder is shown to be valid in the following two cases. One is the extension of the phase diagram of  $W = 0$  to  $W > 0$  when all possible phases are present in the clean limit. We have confirmed that the machine successfully learns the property of each phase from the *quasitranslational* symmetric  $\langle P(\mathbf{r}) \rangle$ . The confidence of the machine is high within the phases, which reflects the successful feature extraction. Another is the detection of the unlearned phase. A correctly optimized ANN judges a state with high confidence when the learned feature is present in the data, and vice versa. The new phase does not exhibit localization in either bulk or the edge, and thus the machine is confused. We confirmed that in both cases the consistency with other independent methods holds.

Let us note that although the analysis here is based on the first moment of the quasiparticle distribution, in general, higher moments may also play a crucial role. In such a case, we expect that by adding the appropriate higher moments the classification can be done in other random systems as well. Furthermore, we may consider alternative input to quasiparticle distribution for interacting systems with disorder; as long as the quantity contains information on the system and recovers the symmetry statistically, the validity of the proposed method is expected. For instance, learning the entanglement spectra with the ANN has been shown to be a valid idea [16,28–30]. Regarding the system without disorder, on the other hand, the auxiliary field configuration [26,27] and the equal-time two-point correlation function [21], both obtained by quantum Monte Carlo simulation, can be fed to the ANN to classify phases. We expect that such quantities are

capable of capturing the property even when the disorder is present.

#### ACKNOWLEDGMENTS

The authors acknowledge helpful discussions with Hideaki Obuse, Masatoshi Sato, Ryo Tamura, and Shu Tanaka. This work was supported by JSPS KAKENHI Grants No. JP15K17719, No. JP16H00985, No. JP17K14352, No. JP18H04478, and No. JP18H04220. N.Y. was supported by the JSPS through the Program for Leading Graduate Schools (ALPS) and a JSPS fellowship (JSPS KAKENHI Grant No. JP17J00743).

#### APPENDIX A: TRANSFER MATRIX

In this appendix, we introduce the transfer matrix (TM) method for quasi-one-dimensional disordered systems [48,69,70]. A metal-insulator transition such as the Anderson transition can be understood from the size dependence of the localization length  $\lambda$  and (thermal) conductivity  $g$ , which is easily computed by the TM method. Let us consider a quasi-one-dimensional system with the length  $L_x$  and the width  $S$ . We assume that the time-independent Schrödinger equation is given as follows,

$$L_{i-1}^\dagger \psi_{i-1} + H_i \psi_i + L_i \psi_{i+1} = E \psi_i. \quad (\text{A1})$$

Here,  $H_i = H_i^\dagger$  is the Hamiltonian restricted on the  $i$ th block and  $E$  is the eigenenergy. We may simply consider the slice of the rectangle as a block when only the nearest-neighbor hopping is present, but otherwise not necessarily a geometrical intersection.  $\psi_i$  is the  $2S$ -dimensional wave function of the

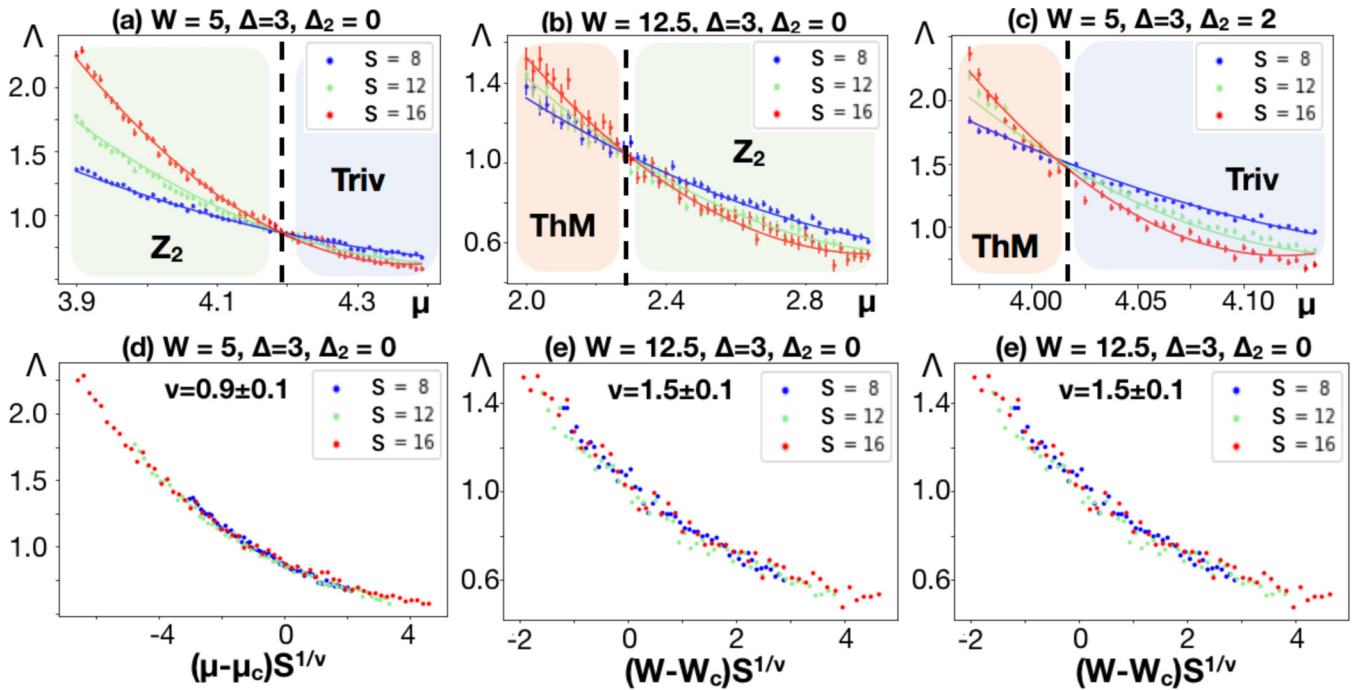


FIG. 5. Finite-size scaling of the dimensionless localization length,  $\Lambda$ , by the transfer matrix method for the Hamiltonian defined by Eqs. (1)–(5) in the main text is shown in (a)–(c). The parameters are given as  $(W, \Delta, \Delta_2) =$  (a)  $(5, 3, 0)$  under OBC, (b)  $(12.5, 3, 0)$  under PBC, and (c)  $(5, 3, 2)$  under OBC. Presented in (d)–(f) are the corresponding data collapses. The system width varies as  $S = 8, 12, 16$ .

(quasi)particle on the  $i$ th block and  $L_i$  is the hopping matrix from the  $i$ th to  $(i + 1)$ th block. Assuming  $\det|L_i| \neq 0$ , Eq. (A1) is rewritten as

$$\begin{pmatrix} \psi_{i+1} \\ \psi_i \end{pmatrix} = \begin{pmatrix} L_i^{-1}(E - H_i) & -L_i \\ I_{2S \times 2S} & 0 \end{pmatrix} \begin{pmatrix} \psi_i \\ \psi_{i-1} \end{pmatrix} \\ := \hat{T}_i(E) \begin{pmatrix} \psi_i \\ \psi_{i-1} \end{pmatrix}, \quad (\text{A2})$$

where the above-defined  $\hat{T}_i(E)$  is referred to as one-step TM. The wave function at the edge and the total TM,  $\hat{T}(E)$ , is obtained as follows,

$$\begin{pmatrix} \psi_{L_x} \\ \psi_{L_x-1} \end{pmatrix} = \left( \prod_{i=0}^{L_x-1} \hat{T}_i \right) \begin{pmatrix} \psi_1 \\ \psi_0 \end{pmatrix} := \hat{T}(E) \begin{pmatrix} \psi_1 \\ \psi_0 \end{pmatrix}. \quad (\text{A3})$$

In the limit of  $L_x \rightarrow \infty$ , we consider the positive-definite operator,  $\hat{\Gamma} = \lim_{S \rightarrow \infty} (\hat{T} \hat{T}^\dagger)^{1/2S}$ , to introduce

$$\lambda_j = \frac{1}{\ln \gamma_j}, \quad (\text{A4})$$

where  $\gamma_j$  is the  $j$ th eigenvalue, which is positive and finite, of  $\hat{\Gamma}$ . The corresponding eigenfunction behaves as  $\exp(\pm x/\lambda_j)$ , with the sign denoting the direction of the decay, and therefore

$\lambda_j$  can be understood as the localization length. We set the length of the system from  $10^4$  to  $10^5$  so that the statistical error is small enough.

As pointed out by MacKinnon and Kramer, the finite-size scaling of the maximum localization length,  $\lambda_{\max} := \lambda$ , is equivalent to the scaling theory of conductance  $g$  [48,60]. The dimensionless localization length in the vicinity of the metal-insulator transition point is assumed to be expressed by one-parameter scaling. Namely, by writing the parameter related to the transition as  $q$  (e.g., chemical potential  $\mu$  and the amplitude of the Anderson potential  $W$  in our work),

$$\Lambda(q) := \frac{\lambda(q)}{S} = \Lambda_c + \sum_{n=1}^N a_n (q - q_c)^n S^{n/\nu} \\ + \sum_{n=1}^{N'} b_n (q - q_c)^n S^{n/\nu+y}, \quad (\text{A5})$$

where the subscripts  $c$  denote the value at the critical point,  $a_n$  and  $b_n$  are the expansion coefficients, and  $\nu$  is the critical exponent for the localization length. The third term is the irrelevant length scale collection by the boundary, whose size dependence is corrected by  $y < 0$ . Finite integers  $N$  and  $N'$  denote the number of the fitting parameters, which is taken

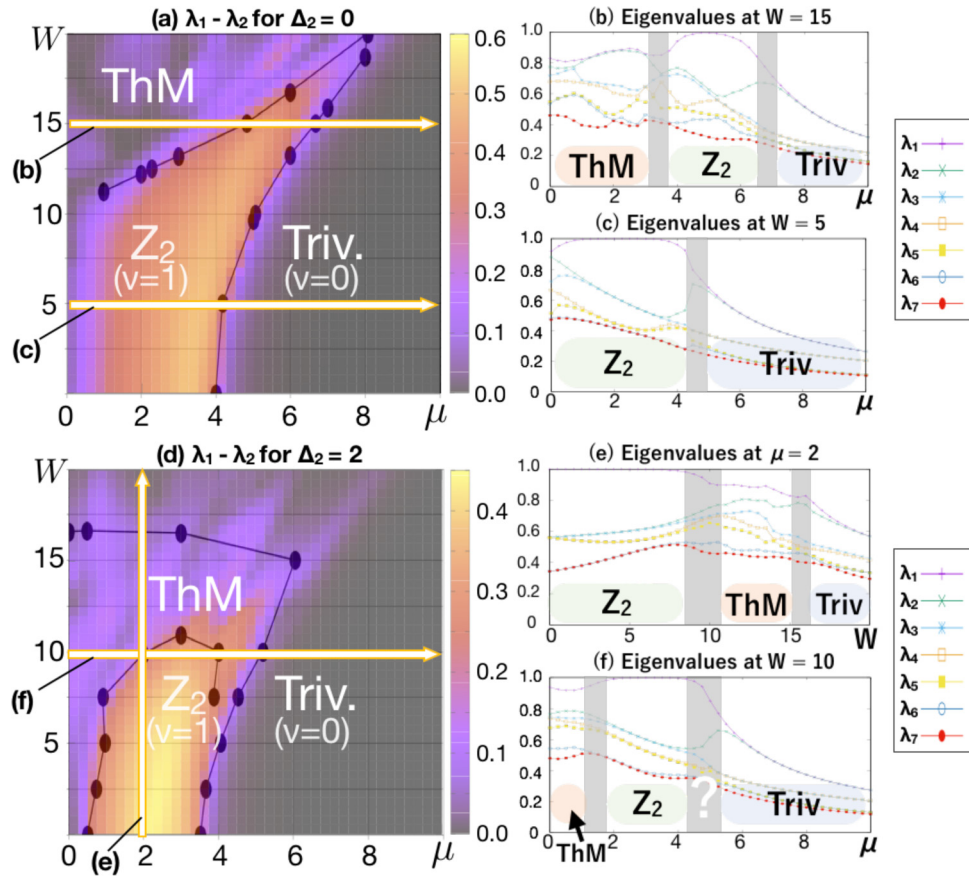


FIG. 6. (a) and (d)  $\lambda_1 - \lambda_2$  as a function of the chemical potential  $\mu$  and the disorder strength  $W$  for class DIII Hamiltonian given by Eq. (1)–(5) of the main text. The parameters are taken as  $(t, \Delta, \Delta_2) =$  (a) (1, 3, 0) and (d) (1, 3, 2), respectively, and the system size is  $20 \times 20$ . (b), (c), (e), and (f) show  $\mu$  and  $W$  dependencies of the eigenvalues  $\lambda_i$ ,  $i = 1, 2, \dots, 7$ , of the operator  $\mathcal{A}$  for the system size  $20 \times 20$ . The parameters are taken as  $(t, \Delta, \Delta_2, W) =$  (b) (1, 3, 0, 15), (c) (1, 3, 0, 5), (f) (1, 3, 2, 10), and (e)  $(t, \Delta, \Delta_2, \mu) =$  (1, 3, 2, 2), respectively. The gray bars in (b), (c), (e), and (f) denote marginal areas.

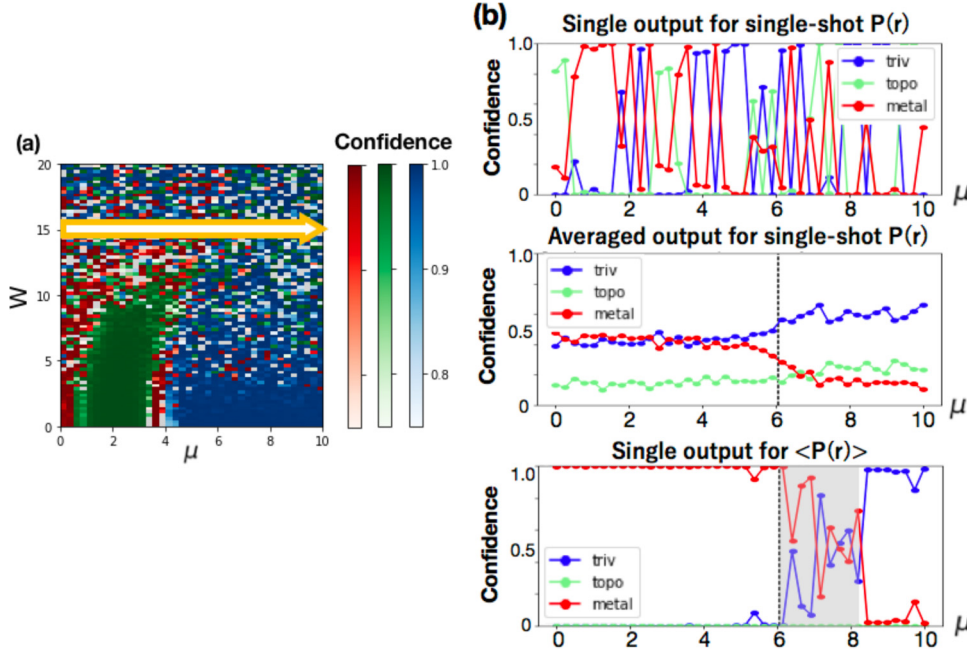


FIG. 7. (a) The output of ANN for single-shot  $P(\mathbf{r})$  for  $\Delta = 3, \Delta_2 = 2$ . Boundaries between phases are hardly recognizable. (b) The single output for  $P(\mathbf{r})$ , the average of 200 outputs for independently generated  $P(\mathbf{r})$ , and the single output for  $\langle P(\mathbf{r}) \rangle$  with  $N_r = 500$  from the top. The amplitude of the random potential is fixed as  $W = 15$ .

as  $N = 2, N' = 0$  in this work. Examples of  $\mathbb{Z}_2$ -trivial and  $\mathbb{Z}_2$ -thermal metal (ThM) phase transitions are shown in Fig. 5, in which the rising (falling) of  $\Lambda$  for extended (localized) states is indeed observed.

Last but not least, let us note that the appropriate boundary condition must be applied to detect the transition from or into the  $\mathbb{Z}_2$  phase [70]. In two-dimensional systems, we have two options: the open or periodic boundary condition (OBC or PBC) along the direction perpendicular to the transferred direction. The edge state appears along the transferred direction with the OBC, while the state is merely localized in the first or the last block with the PBC. Thus, to determine the  $\mathbb{Z}_2$ -trivial ( $\mathbb{Z}_2$ -ThM) phase boundary, we must consider OBC (PBC) systems. Note that the trivial-ThM boundary is detected in either way.

## APPENDIX B: $\mathbb{Z}_2$ INDEX FOR 2D CLASS DIII SYSTEMS WITH DISORDER

In this appendix, we introduce the noncommutative geometry approach to map out the phase diagram of 2d class DIII systems. The  $\mathbb{Z}_2$  index derived in previous works [49,50] is numerically advantageous since it can be determined from the discrete spectrum of a certain compact operator without taking the disorder average. See Ref. [51] for detailed numerical implementation. The definition of the  $\mathbb{Z}_2$  index of 2d class DIII systems is given as

$$\nu = \ker \dim [\mathcal{A} - 1] \text{ modulo } 2, \quad (\text{B1})$$

where  $\nu = 0$  and 1 correspond to the trivial and the  $\mathbb{Z}_2$  phases, respectively. The operator  $\mathcal{A}$  measures the difference between two projections,

$$\mathcal{A} = P_F - \mathcal{D}_a^* P_F \mathcal{D}_a. \quad (\text{B2})$$

Here,  $P_F$  is the projection operator onto the quasiparticle states below zero energy. The Dirac operator  $\mathcal{D}_a$  is defined by

$$\mathcal{D}_a(\mathbf{r}) := \frac{r_1 + ir_2 - (a_1 + ia_2)}{|r_1 + ir_2 - (a_1 + ia_2)|}, \quad (\text{B3})$$

where  $\mathbf{r} = (r_1, r_2) \in \mathbb{Z}^2$  denotes the position operator of a square lattice and  $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2 \setminus \mathbb{Z}^2$  is a vector off the lattice points. The operator  $\mathcal{D}_a^*$  is the adjoint of the Dirac operator  $\mathcal{D}_a$ . Hereafter, we regard  $\lambda_i$  as the  $i$ th eigenvalue of the operator  $\mathcal{A}$  in descending order including multiplicity.

Shown in Fig. 6(a) is  $\lambda_1 - \lambda_2$  as a function of the chemical potential  $\mu$  and the disorder amplitude  $W$  with the pairings fixed as  $\Delta = 3$  and  $\Delta_2 = 0$ . The orange-colored region denotes the  $\mathbb{Z}_2$  phase since  $\lambda_1 \sim 1$  [see, for instance, Fig. 6(c)] and  $\lambda_1 - \lambda_2 \neq 0$  evidently hold and thus imply  $\nu = 1$ . In Fig. 6(a) we see that the numerical result is in good agreement with the boundary obtained by the TM. The two black areas above and to the right of the  $\mathbb{Z}_2$  phase are identified as the ThM and the trivial phases, respectively. This is done in the following way. When the spectral gap is open (= trivial or  $\mathbb{Z}_2$  phase), the eigenvalues below unity always come in pairs [see Figs. 6(b) and 6(c)] owing to the two symmetries: the time-reversal symmetry of the Hamiltonian and the supersymmetric structure of the operator  $\mathcal{A}$ . However, the doublet structure is not guaranteed when the spectral or the mobility gap vanishes (ThM phase), and in fact, each eigenvalue shows no such a specific structure in the leftmost region of Fig. 6(b).

The difference between the first and second eigenvalues for  $\Delta_2 = 2$  is also given in Fig. 6(d). In the orange region,  $\lambda_1 \sim 1$  [see, for instance, Fig. 6(e)] and  $\lambda_1 - \lambda_2 \neq 0$ , and hence  $\nu = 1$  which corresponds to the  $\mathbb{Z}_2$  phase. The black region denotes the trivial phase with  $\nu = 0$  because there is no  $\lambda_1 \sim 1$ . [See

Fig. 6(e).] While the boundary of the  $\mathbb{Z}_2$  phase is consistent with the TM, detection of the phase boundary between the ThM and the trivial phase requires deep consideration in some situations. In Fig. 6(e), the two phases are distinguishable by the presence of the doublet structure, whereas in Fig. 6(f), it is hard to tell whether the intermediate region between the  $\mathbb{Z}_2$  and the trivial phase is a finite window of the ThM. As is seen in Fig. 3 of the main text, this is indeed a small window of ThM, which is unambiguously captured by the ANN.

### APPENDIX C: SINGLE-SHOT AND AVERAGED DATA

In the following, we see that the success by the ANN is attributed to the recovery of symmetry, but not merely by the law of large numbers. Taking the disorder average of the input data corresponds to an appropriate feature selection, which

is crucial in training our machine. Since the ANN is a totally nonlinear function, this is not the case for averaging the output.

As is shown in Fig. 7(a), classification of  $P(\mathbf{r})$ , i.e., the single-shot realization, results in total meaninglessness, particularly in the strong disorder region. For the sake of simplicity, let us restrict the amplitude as  $W = 15$  in the following. Shown at the top of Fig. 7(b) is the output for single-shot. The random values reflect the fact that the ANN is confused by the translational-symmetry-broken behavior of the quasiparticle. We see in the middle that averaging such outputs in a brute-force manner does not improve the situation at all. Although the faint slope around the boundary seems to capture the phase transition, the output converges far below unity. It is questionable whether we can determine the phase in general. Shown at the bottom is the appropriate classification for  $\langle P(\mathbf{r}) \rangle$  with  $N_r = 500$ , in which the features of the quasitranslational states are detected appropriately.

- 
- [1] G. Carleo and M. Troyer, *Science* **355**, 602 (2017).
  - [2] G. Torlai and R. G. Melko, *Phys. Rev. B* **94**, 165134 (2016).
  - [3] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. G. Melko, and G. Carleo, *Nat. Phys.* **14**, 447 (2018).
  - [4] D. L. Deng, X. Li, and S. Das Sarma, *Phys. Rev. B* **96**, 195145 (2017).
  - [5] X. Gao and L.-M. Duan, *Nat. Commun.* **8**, 662 (2017).
  - [6] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, *Phys. Rev. B* **96**, 205152 (2017).
  - [7] G. Carleo, Y. Nomura, and M. Imada, *arXiv:1802.09558*.
  - [8] N. Freitas, G. Morigi, and V. Dunjko, *arXiv:1803.02118*.
  - [9] H. Saito, *J. Phys. Soc. Jpn.* **86**, 093001 (2017).
  - [10] H. Saito and M. Kato, *J. Phys. Soc. Jpn.* **87**, 014001 (2018).
  - [11] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, *Phys. Rev. B* **95**, 041101 (2017).
  - [12] L. Huang and L. Wang, *Phys. Rev. B* **95**, 035105 (2017).
  - [13] L. Wang, *Phys. Rev. E* **96**, 051301 (2017).
  - [14] H. Shen, J. Liu, and L. Fu, *arXiv:1801.01127*.
  - [15] L. Wang, *Phys. Rev. B* **94**, 195105 (2016).
  - [16] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, *Nat. Phys.* **13**, 435 (2017).
  - [17] Y.-H. Liu and E. P. L. van Nieuwenburg, *Phys. Rev. Lett.* **120**, 176401 (2018).
  - [18] P. Broecker, F. F. Assaad, and S. Trebst, *arXiv:1707.00663*.
  - [19] W. Hu, R. R. P. Singh, and R. T. Scalettar, *Phys. Rev. E* **95**, 062122 (2017).
  - [20] J. Carrasquilla and R. G. Melko, *Nat. Phys.* **13**, 431 (2017).
  - [21] Y. Zhang and E.-A. Kim, *Phys. Rev. Lett.* **118**, 216401 (2017).
  - [22] T. Ohtsuki and T. Ohtsuki, *J. Phys. Soc. Jpn.* **85**, 123706 (2016).
  - [23] T. Ohtsuki and T. Ohtsuki, *J. Phys. Soc. Jpn.* **86**, 044708 (2017).
  - [24] P. Zhang, H. Shen, and H. Zhai, *Phys. Rev. Lett.* **120**, 066401 (2018).
  - [25] D. Carvalho, N. A. García-Martínez, J. L. Lado, and J. Fernández-Rossier, *Phys. Rev. B* **97**, 115453 (2018).
  - [26] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, *Phys. Rev. X* **7**, 031038 (2017).
  - [27] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, *Sci. Rep.* **7**, 8823 (2017).
  - [28] F. Schindler, N. Regnault, and T. Neupert, *Phys. Rev. B* **95**, 245134 (2017).
  - [29] J. Venderley, V. Khemani, and E. Kim, *arXiv:1711.00020*.
  - [30] E. P. L. van Nieuwenburg, E. Bairey, and G. Rafael, *arXiv:1712.00450*.
  - [31] L. A. Wray, S.-Y. Xu, Y. Xia, Y. S. Hor, D. Qian, A. V. Fedorov, H. Lin, A. Bansil, R. J. Cava, and M. Z. Hasan, *Nat. Phys.* **6**, 855 (2010).
  - [32] L. Fu and E. Berg, *Phys. Rev. Lett.* **105**, 097001 (2010).
  - [33] Z. Wang, P. Zhang, G. Xu, L. K. Zeng, H. Miao, X. Xu, T. Qian, H. Weng, P. Richard, A. V. Fedorov, H. Ding, X. Dai, and Z. Fang, *Phys. Rev. B* **92**, 115119 (2015).
  - [34] Z. F. Wang, H. Zhang, D. Liu, C. Liu, C. Tang, C. Song, Y. Zhong, J. Peng, F. Li, C. Nie, L. Wang, X. J. Zhou, X. Ma, Q. K. Xue, and F. Liu, *Nat. Mater.* **15**, 968 (2016).
  - [35] P. Zhang, K. Yaji, T. Hashimoto, Y. Ota, T. Kondo, K. Okazaki, Z. Wang, J. Wen, G. D. Gu, H. Ding, and S. Shin, *Science* **360**, 182 (2018).
  - [36] A. Kitaev, *Ann. Phys. (N.Y.)* **303**, 2 (2003).
  - [37] S. B. Bravyi and A. Y. Kitaev, *Ann. Phys. (N.Y.)* **298**, 210 (2002).
  - [38] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, *Phys. Rev. Lett.* **49**, 405 (1982).
  - [39] M. Kohmoto, *Ann. Phys. (N.Y.)* **160**, 343 (1985).
  - [40] C. L. Kane and E. J. Mele, *Phys. Rev. Lett.* **95**, 146802 (2005).
  - [41] C. L. Kane and E. J. Mele, *Phys. Rev. Lett.* **95**, 226801 (2005).
  - [42] Q. Niu, D. J. Thouless, and Y. S. Wu, *Phys. Rev. B* **31**, 3372 (1985).
  - [43] Another well-known issue is the rise in the numerical cost. Together with the necessity of avoiding overfitting, this problem implies that excessive pursuit in decreasing  $\epsilon$  is undesirable in practical situations.
  - [44] G. Cybenko, *Math. Control Signals Systems* **2**, 303 (1989).
  - [45] K. Hornik, M. Stinchcombe, and H. White, *Neural Networks* **2**, 359 (1989).
  - [46] M. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).
  - [47] P. Anderson, *Phys. Rev.* **109**, 1492 (1958).
  - [48] A. MacKinnon and B. Kramer, *Z. Phys. B* **53**, 1 (1983).



- [49] H. Katsura and T. Koma, *J. Math. Phys.* **57**, 021903 (2016).
- [50] H. Katsura and T. Koma, *J. Math. Phys.* **59**, 031903 (2018).
- [51] Y. Akagi, H. Katsura, and T. Koma, *J. Phys. Soc. Jpn.* **86**, 123710 (2017).
- [52] Considering the Chalker-Coddington model, e.g., I. C. Fulga, A. R. Akhmerov, J. Tworzydło, B. Béri, and C. W. J. Beenakker, *Phys. Rev. B* **86**, 054505 (2012), or the scattering matrix theory as in Ref. [59] are other options.
- [53] M. Sato and S. Fujimoto, *Phys. Rev. B* **79**, 094504 (2009).
- [54] Although the square lattice is considered in the effective model, i.e., Eq. (1), we assume that the underlying crystal structure of the original Hamiltonian breaks inversion symmetry.
- [55] L. Fu and C. L. Kane, *Phys. Rev. B* **76**, 045302 (2007).
- [56] A. P. Schnyder, S. Ryu, A. Furusaki, and A. W. W. Ludwig, *Phys. Rev. B* **78**, 195125 (2008).
- [57] A. Y. Kitaev, *AIP Conf. Proc.* **1134**, 22 (2009).
- [58] X.-L. Qi, T. L. Hughes, and S.-C. Zhang, *Phys. Rev. B* **81**, 134508 (2010).
- [59] M. Diez, I. C. Fulga, D. I. Pikulin, J. Tworzydło, and C. W. J. Beenakker, *New J. Phys.* **16**, 063049 (2014).
- [60] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, *Phys. Rev. Lett.* **42**, 673 (1979).
- [61] S. Hikami, *Phys. Rev. B* **24**, 2671 (1981).
- [62] F. Evers and A. D. Mirlin, *Rev. Mod. Phys.* **80**, 1355 (2008).
- [63] T. Senthil and M. P. A. Fisher, *Phys. Rev. B* **62**, 7850 (2000).
- [64] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
- [65] J. Duchi, E. Hazan, and Y. Singer, *J. Mach. Learn. Res.* **12**, 2121 (2011).
- [66] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *J. Mach. Learn. Res.* **15**, 1929 (2014).
- [67] At larger disorder, the phase boundaries approach each other. Therefore, the machine is confused, i.e., the output remains far below 1, by the finite-size effect, resulting in the small estimation of the  $\mathbb{Z}_2$  phase.
- [68] M. V. Medvedyeva, J. Tworzydło, and C. W. J. Beenakker, *Phys. Rev. B* **81**, 214203 (2010).
- [69] L. Molinari, *J. Phys. A: Math. Gen.* **30**, 983 (1997).
- [70] A. Yamakage, K. Nomura, K. I. Imura, and Y. Kuramoto, *Phys. Rev. B* **87**, 205141 (2013).