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

In this letter, we apply the artificial neural network in a supervised manner to map out the quantum phase diagram of disordered topological superconductor 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 the 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 the certain region, leading us to conclude the detection of the unknown phase which is eventually identified as the thermal metal phase. In our method, only the first moment of the quasiparticle distribution is used for input, but application to a wider variety of systems is expected by the inclusion of higher moments.

Introduction.— Machine learning is the construction and execution of the computational algorithm which optimizes the quantified objective. 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 [1–6], acceleration of Monte Carlo simulation [7], detection of phase transition by unsupervised learning without teaching the notion of phases to the machine [8–12]. Among them, a problem that draws particular attention is the classification of various phases such as in topological systems [13–17], strongly correlated systems [18, 19], and many-body localized systems [20].

In this letter, we investigate the quantum phase diagram of class DIII topological superconductor with disorder motivated by the recent proposal of the candidate materials such as $Cu_xBi_2Se_3$ [21, 22] and $FeTe_xSe_{1-x}$ [23–25]. 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 [26, 27]. While topological invariants in translational invariant systems have been well studied including their concrete expressions and calculations [28–31], 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 [32], 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. 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 $\widetilde{\mathbf{F}}$ satisfies $|\widetilde{\mathbf{F}}(\vec{x}_i) - \mathbf{F}(\vec{x}_i)| < \epsilon$ [33–35]. The second is the recovery of translational symmetry by ensemble average. While the translational symmetry is broken in a system with disorder such as a random potential [36], as an ensemble of disorder average the symmetry is *statistically recovered*. The ensemble averaged state is expected to be understood similarly as in the clean limit as long as the bulk gap is open, and hence we expect to understand the phase in disordered systems by an ANN which learned from the clean limit. As we show later, the phase diagram obtained from our method is fully consistent with the results in both the transfer matrix (TM) method [37] and the calculation of a $\mathbb{Z}_2$ index by noncommutative geometry which is recently proposed [38–41].

Hamiltonian.— The BdG Hamiltonian for 2d noncentrosymmetric class DIII topological superconductor in the clean limit is given in the real space as [42]

$$H_0 = \sum_{\mathbf{r}} \sum_{k=1,2} \Psi_{\mathbf{r}}^\dagger t_k \Psi_{\mathbf{r}+\mathbf{e}_\mathbf{k}} + \sum_{\mathbf{r}} \Psi_{\mathbf{r}}^\dagger v \Psi_{\mathbf{r}}, \tag{1}$$

$$t_1 = t s_3 \otimes \sigma_0 + \frac{i\Delta}{2} s_1 \otimes \sigma_3, \tag{2}$$

$$t_2 = t s_3 \otimes \sigma_0 + \frac{\Delta}{2} s_1 \otimes \sigma_3, \tag{3}$$

$$v = -\mu s_3 \otimes \sigma_0 - \Delta_2 s_2 \otimes \sigma_2. \tag{4}$$

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 and $c_{\mathbf{r}\alpha}$ is an annihilation operator of an electron with spin α 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 Δ. The Pauli matrices s_k and σ_k (k = 0,1,2,3) operate on particle-hole and spin space, respectively. The on-site term, v, consists of the chemical potential μ and the s-wave pairing Δ_2. The mixture of the spin-singlet and the spin-triplet pairings are caused by the broken inversion symmetry [43]. 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 [30, 31, 44–47]. 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 $|\Delta_2/\Delta| < 1$ [48]. As the on-site disorder, we introduce a random potential with the amplitude distributed uniformly with

width W, namely,

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

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 system may cause destructive interference of time-reversal paths of the quasiparticle, suppressing the back scattering and thereby lead the system to show metallic behavior (weak-antilocalization) in 2d [49–51]. In particular, "insulator-metal" transition from the $\mathbb{Z}_2$ phase, in which Majorana fermions pinned to the disorder percolates, gives rise to the so-called Majorana metal phase [52]. 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).

Classification by Artificial Neural Network.— An artificial neural network (ANN) is a nonlinear function that takes an input $\mathbf{x}$ to compute an output $\mathbf{y}$ though sequential mappings by layers of "neurons" [35]. 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 ith layer as $\mathcal{A}_i$ and $\mathcal{W}_i$, respectively. ANN that can uniquely number the layers according to the order of input and output and do not include any intralayer processing is referred to as a feedforward network. In this paper, we apply feedforward ANN with two hidden layers. [See Fig. 1 for a graphical understanding of the architecture.] The output is calculated as

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

In our architecture, the activation function of the hidden and output layers are Rectified Linear Unit (ReLU) and the Softmax function, respectively. The definitions are given as

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

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

Next, we discuss the training process of the machine. The parameters $\mathcal{W}$ are tuned via minimization of the cost function which quantify the performance of the machine. In a classification problem with the current network architecture, the cross-entropy function is widely used [53] due to its convenience. The function is given as,

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

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, $\hat{y}^{(k)} = \delta_{k,l_j}$ for the correct label l_j 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 mini-batch gradient descent with batch size 40 as $\mathcal{W}^{(i)}_{j,k} \to \mathcal{W}^{(i)}_{j,k} - \eta \left(\partial \mathcal{L} / \partial \mathcal{W}_{j,k} \right)$, where η is the learning rate that is controlled by AdaGrad method to efficiently reach the global minimum [54]. Furthermore, we apply the drop-out method to avoid overfitting [55].

Input Data.— Adopted as input data $\mathbf{x}$ is the disorder average over N_r realizations of the spacial distribution of quasiparticle, $P(\mathbf{r})$, corresponding to the first excited state. Namely, for the eigenstate $|\psi\rangle$ satisfying $H|\psi\rangle = E_1|\psi\rangle$ with the lowest $E_1 > 0$, we compute

$$P(\mathbf{r}) = |\psi^e_\uparrow(\mathbf{r})|^2 + |\psi^e_\downarrow(\mathbf{r})|^2 + |\psi^h_\uparrow(\mathbf{r})|^2 + |\psi^h_\downarrow(\mathbf{r})|^2, \quad (10)$$

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 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

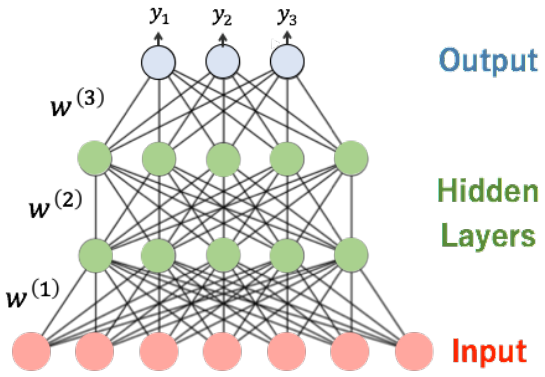

FIG. 1. (Color online) The architecture of a feedforward-type artificial neural network with a single hidden layer, at which the input data is compressed to extract some abstract feature for classification. The activation of the output layer is the Softmax function so that the sum is unity, allowing us to interpret as the confidence of the machine.

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 spacial 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 phase by feeding $\langle P(\mathbf{r})\rangle$ to the ANN which learned the labels of $P(\mathbf{r})$ in the clean limit. Here, we consider both binary and ternary classification. The machine learns the $\mathbb{Z}_2$ and trivial phases in the former, and additionally the ThM phase in the latter.

Ternary classification.— First, we carry out the ternary classification at finite Δ_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 [42, 48]. 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 is the average output of 200 independently trained machines for input $\langle P(\mathbf{r})\rangle$ with N_r =500. Only $\mu \geq 0$ 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 [56]. Remarkably the machine has successfully learned their characteristics and fully extended the phase diagram by its generalizing skill.

First, let us focus along $\mu = 2$. In the clean limit, the system is in the $\mathbb{Z}_2$ and enters the ThM and the trivial phase by increasing the disorder. The topologically nontrivial phase experiences the weak-antilocalization and the Anderson localization of the quasiparticle, 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

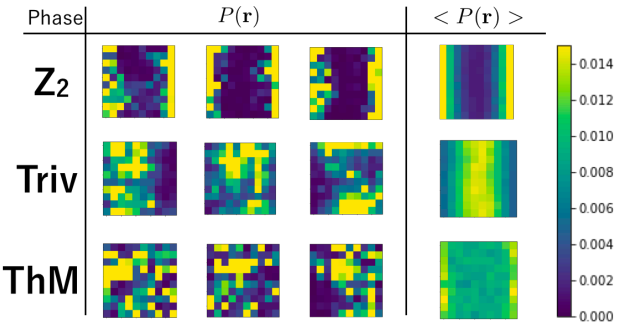

FIG. 2. (Color online) The single-shot quasiparticle distribution of the first excited state, $P(\mathbf{r})$, and its disorder average, $\langle P(\mathbf{r})\rangle$, over 500 realization of random configurations. The parameters are taken as $(\mu, W) = (2, 9), (6, 5), (2, 18)$ with $\Delta = 3, \Delta_2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

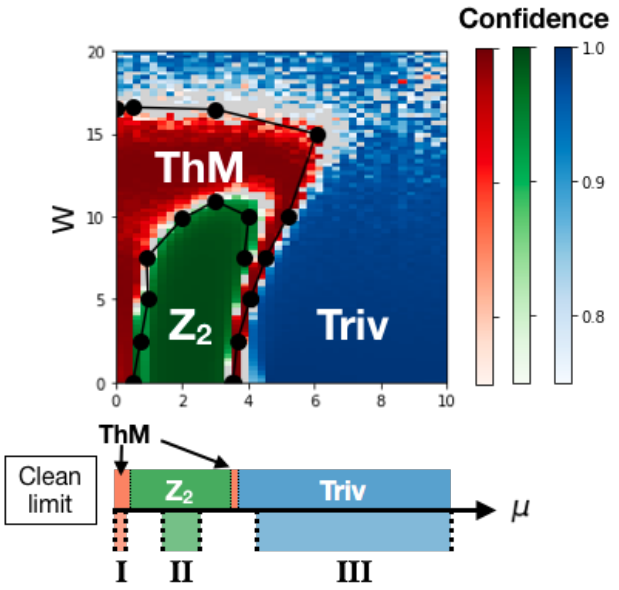

FIG. 3. Average outputs of 200 ternary classifying ANN trained in the clean limit for $t = 1, \Delta = 3, \Delta_2 = 2$. The parameter μ of 1000 training data with system size 14×14 is uniformly distributed within I: [0.0, 0.3], II: [1.0, 2.5], and III: [4.0, 10.0], each corresponding to the thermal metal (red), $\mathbb{Z}_2$ (green), and trivial (blue) phase. The depth of the colors denotes the confidence of the machine. 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%. The machine is highly confident of each phase but confused at the boundary.

data, which is suppressed by increasing N_r. Other $\mathbb{Z}_2$-ThM and ThM-trivial phase boundaries are nicely reproduced. Furthermore, in the weak disorder region at $\mu \sim 3.5$, or the region between the $\mathbb{Z}_2$ and the trivial phase, is unambiguously classified as ThM. While the detection of close parallel boundaries requires extra numerical effort on the TM or the noncommutative geometry approach due to the finite-size effect, the ANN, with moderate computational cost, gives quantitative prediction that is consistent with the other two methods as well as the analytical result in the clean limit [57].

Binary classification.— Next, 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 phase, 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 [58].

The drop of confidence along $\mu = 0$ is also observed. This is understood as $\mathbb{Z}_2 - \mathbb{Z}_2$ transition line, which is consistent with the analysis of the staggered fermion model for class D [59]. Note that, such transition that lacks change in size dependence on thermal conductivity or localization length is very difficult

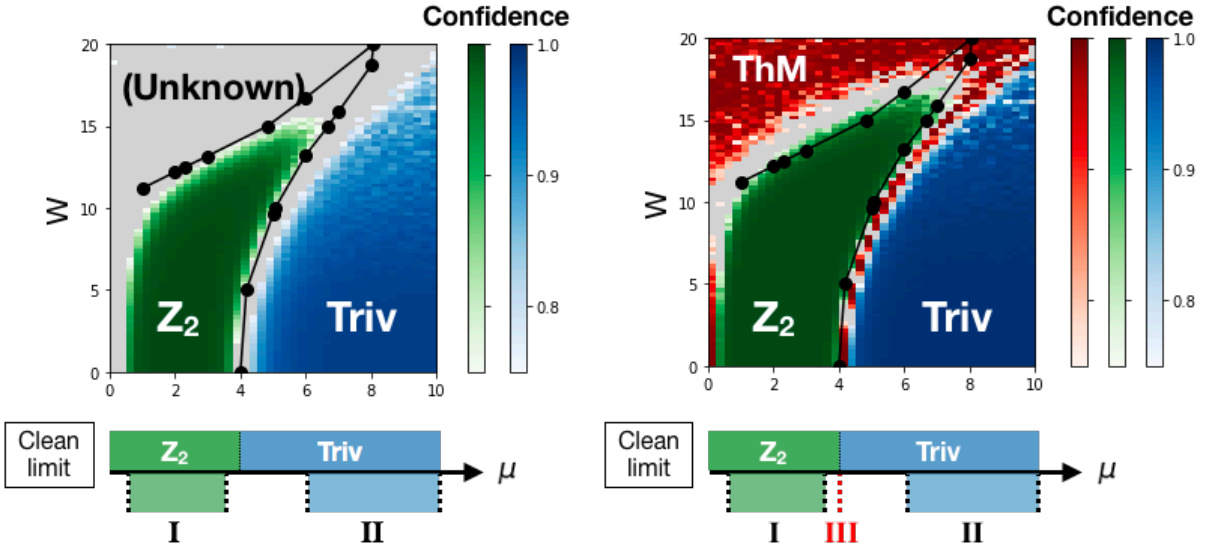

FIG. 4. Average output of binary classifying ANN trained at the clean limit for $t = 1, \Delta = 3, \Delta_2 = 0$. The parameter μ of 1000 training data with system size 10×10 is uniformly distributed within (a) I: [0.5, 3.5] and II: [6.0, 10.0], (b) I, II, and III: $\mu = 0.0, 4.0$, each corresponding 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.

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 number of average for input or the machine. Such 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 the critical points, i.e., $\mu = 0, 4$ 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 is judged to be in the "phase" of the critical point, implying the detection of the extended behavior of the quasiparticle by the ANN. Hence, this region is concluded as the thermal metal phase, which is also confirmed from TM and the noncommutative geometry approach. 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.

Discussion.— In this work, the classification of phases with ANN in DIII topological superconductor 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 *quasi*-translational 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. Correctly optimized ANN judge 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 confirm 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 play a crucial role. In such a case, we simply generate multichannel input data and train the ANN similarly. It is natural to expect that by adding the appropriate higher moments the classification can be done in other random system as well.

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

[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, and M. Troyer, arXiv:1703.05334 (2017).
[4] D. L. Deng, X. Li, and S. D. Sarma, arXiv:1609.09060 (2016).
[5] X. Gao and L. M. Duan, preprint arXiv:1701.05039 (2017).

[6] Y. Nomura, A. Darmawan, Y. Yamaji, and M. Imada, arXiv:1709.06475 (2017).
[7] J. Liu, Y. Qi, Z. Y. Meng, and L. Fu, Phys. Rev. B **95**, 041101 (2017).
[8] L. Wang, Phys. Rev. B **94**, 195105 (2016).
[9] E. P. L. Nieuwenburg, Y.-H. Liu, and S. D. Huber, Nat. Phys. **13**, 435 (2017).
[10] Y.-H. Liu and E. P. L. Nieuwenburg, arXiv:1706.08111 (2017).
[11] P. Broecker, F. F. Assaad, and S. Trebst, arXiv:1707.00663 (2017).
[12] W. Hu, R. R. P. Singh, and R. T. Scalettar, Phys. Rev. E **95**, 062122 (2017).
[13] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).
[14] Y. Zhang and E. A. Kim, Phys. Rev. Lett. **118**, 216401 (2017).
[15] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **85**, 123706 (2016).
[16] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **86**, 044708 (2017).
[17] P. Zhang, H. Shen, and H. Zhai, arXiv:1708.09401 (2017).
[18] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Phys. Rev. X **7**, 031038 (2017).
[19] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Scientific Reports **7** (2017).
[20] F. Schindler, N. Regnault, and T. Neupert, Phys. Rev. B **95**, 245134 (2017).
[21] 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).
[22] L. Fu and E. Berg, Phys. Rev. Lett. **105**, 097001 (2010).
[23] 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).
[24] 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).
[25] P. Zhang, K. Yaji, T. Hashimoto, Y. Ota, T. Kondo, K. Okazaki, Z. Wang, J. Wen, G. D. Gu, H. Ding, and S. Shin, arXiv:1706.05163 (2017).
[26] A. Kitaev, Ann. Phys. (NY) **303**, 2 (2003).
[27] S. B. Bravyi and A. Y. Kitaev, Ann. Phys. (NY) **298**, 210 (2002).
[28] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, Phys. Rev. Lett. **49**, 405 (1982).
[29] M. Kohmoto, Ann. Phys. (NY) **160**, 343 (1985).
[30] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 146802 (2005).
[31] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 226801 (2005).
[32] Q. Niu, D. J. Thouless, and Y. S. Wu, Phys. Rev. B **31**, 3372 (1985).
[33] G. Cybenko, Math. Control Signals Systems **2**, 303 (1989).
[34] K. Hornik, M. Stinchcombe, and H. White, Neural Networks **2**, 359 (1989).
[35] M. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).
[36] P. Anderson, Phys. Rev. **109**, 1492 (1958).
[37] A. MacKinnon and B. Kramer, Z. Phys. B **53**, 1 (1983).
[38] H. Katsura and T. Koma, J. Math. Phys. **57**, 021903 (2016).
[39] H. Katsura and T. Koma, arXiv:1611.01928 (2016).
[40] Y. Akagi, H. Katsura, and T. Koma, arXiv:1709.05853 (2017).
[41] Considering Chalker-Coddington model, e.g. I.C. Fulga, A. R. Akhmerov, J. Tworzydlo, B. Béri, and C. W. J. Beenakker, Phys. Rev. B **86**, 054505 (2012), or the scattering matrix theory as in Ref. [48] are other options.
[42] M. Sato and S. Fujimoto, Phys. Rev. B **79**, 094504 (2009).
[43] 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.
[44] L. Fu and C. L. Kane, Phys. Rev. B **76**, 045302 (2007).
[45] A. P. Schnyder, S. Ryu, A. Furusaki, and A. W. W. Ludwig, Phys. Rev. B **78**, 195125 (2008).
[46] A. Y. Kitaev, AIP Conf. Proc. **1134**, 22 (2009).
[47] X.-L. Qi, T. L. Hughes, and S.-C. Zhang, Phys. Rev. B **81**, 134508 (2010).
[48] M. Diez, I. C. Fulga, D. I. Pikulin, J. Tworzydło, and C. W. J. Beenakker, New J. Phys. **16**, 063049 (2014).
[49] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).
[50] S. Hikami, Phys. Rev. B **24**, 2671 (1981).
[51] F. Evers and A. D. Mirlin, Rev. Mod. Phys. **80**, 1355 (2008).
[52] T. Senthil and M. P. A. Fisher, Phys. Rev. B **62**, 7850 (2000).
[53] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, Nj, USA, 2006).
[54] J. Duchi, E. Hazan, and Y. Singer, J. Mach. Learn. Res. **12**, 2121 (2011).
[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, J. Mach. Learn. Res. **15**, 1929 (2014).
[56] For the detailed information for the TM and the noncommutative geometry approach, see Supplemental Material.
[57] 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 Supplemental Material for further discussion.
[58] At larger disorder, the phase boundary approach each other. Therefore, 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.
[59] M. V. Medvedyeva, J. Tworzydło, and C. W. J. Beenakker, Phys. Rev. B **81**, 214203 (2010).

Supplemental Materials for *"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

I. TRANSFER MATRIX

In this section, we introduce the transfer matrix (TM) method for quasi-one-dimensional disordered system [S1–S3]. Metal-insulator transition such as Anderson transition can be understood from the size dependence of the localization length λ 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,

$$L_{i-1}^{\dagger}\psi_{i-1} + H_i\psi_i + L_i\psi_{i+1} = E\psi_i. \tag{S1}$$

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 if only nearest neighbor hopping is present. If further hopping is concerned, it is not necessarily a geometrical intersection. ψ_i is the $2S$-dimensional wave function of the (quasi)particle on the i-th block, L_i is the hopping matrix from ith to $(i+1)$-th block. Assuming $\det|L_i| \neq 0$, Eq. (S1) is rewritten as

$$\begin{pmatrix} \psi_{i+1} \\ \psi_i \end{pmatrix} = \begin{pmatrix} L_i^{-1}(E - H_i) & -L_i \\ I_{2S\times 2S} & 0 \end{pmatrix}\begin{pmatrix} \psi_i \\ \psi_{i-1} \end{pmatrix} := \hat{T}_i(E)\begin{pmatrix} \psi_i \\ \psi_{i-1} \end{pmatrix}. \tag{S2}$$

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

$$\begin{pmatrix} \psi_{L_x} \\ \psi_{L_x-1} \end{pmatrix} = \left(\prod_{i=0}^{L_x-1}\hat{T}_i\right)\begin{pmatrix} \psi_1 \\ \psi_0 \end{pmatrix} := \hat{T}(E)\begin{pmatrix} \psi_1 \\ \psi_0 \end{pmatrix}. \tag{S3}$$

In the limit of $L_x \to \infty$, the Oseledec theorem [S4] on random matrix products assures the positive definiteness of $\hat{T}$, implying the localization of the free fermion in quasi-one-dimensional system. That is, by introducing

$$\lambda_i = \frac{1}{\ln \gamma_i}, \tag{S4}$$

where γ_i is the eigenvalue of $\hat{T}$, the eigenfunction behaves as $\psi_i \sim \exp(\pm x/\lambda_i)$. Here, λ_i can be understood as the localization length since γ_i is positive and finite, with the sign denoting the direction of the decay. In reality, we set the length of the system from 10^4 to 10^5 so that the statistical error is small enough.

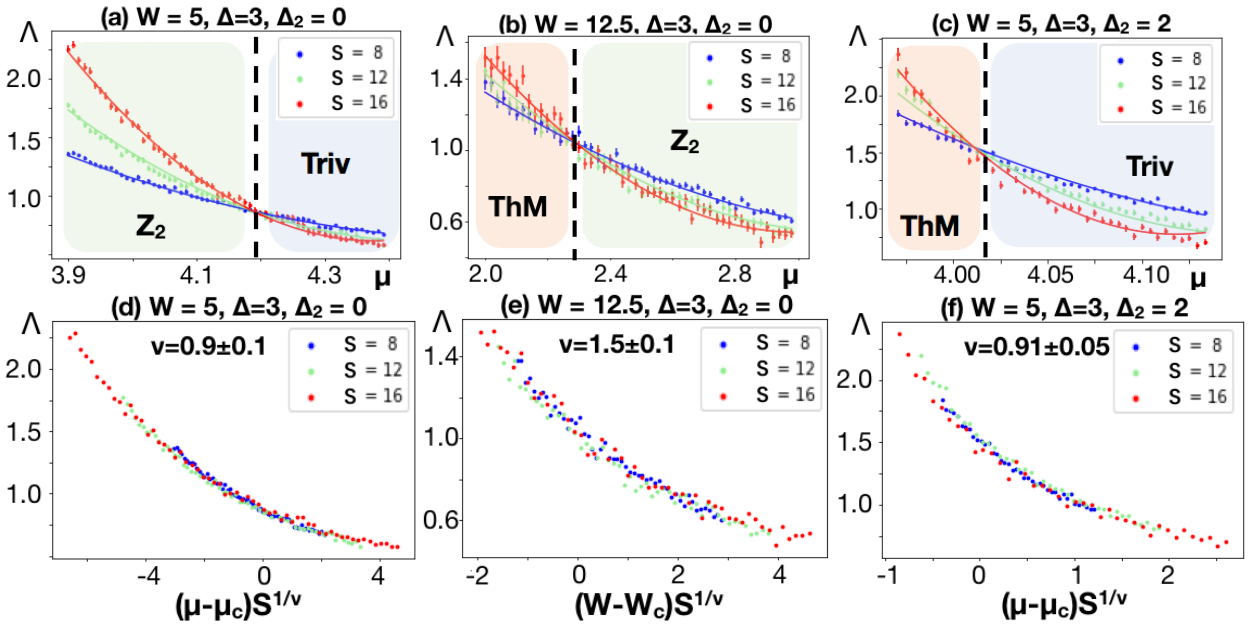

FIG. S1: Finite-size scaling of the dimensionless localization length, Λ, by the transfer matrix method for the Hamiltonian defined in Eqs. (1) - (5) is shown in (a)-(c). The parameters are given as (W, Δ, Δ_2) = (a) $(5, 3, 0)$ under OBC, (b) $(12.5, 3, 0)$ under PBC, and (c) $(5, 3, 2)$ under OBC. Presented in (d)-(f) are the corresponding data collapses. The system width varies as $S = 8, 12, 16$.

MacKinnon pointed out in his paper that the finite-size scaling of the maximum localization length, $\lambda_{\max} := \lambda$, is equivalent to the scaling theory of conductance g, which was established by " the gang of four" [S1, S5]. We assume that the dimensionless localization length in the vicinity of the metal-insulator transition point is expressed by one-parameter scaling. Namely, by writing the parameter related to the transition as q (e.g. chemical potential μ and the amplitude of the Anderson potential W in our work),

$$\Lambda(q) := \frac{\lambda(q)}{S} = \Lambda_c + \sum_{n=1}^{N} a_n (q - q_c)^n S^{n/\nu} + \sum_{n=1}^{N'} b_n (q - q_c)^n S^{n/\nu + y}, \tag{S5}$$

where the subscripts c denote the value at the critical point, a_n and b_n are the expansion coefficients, and ν is the critical exponent for the localization length. The third term is the irrelevant length scale collection by the boundary, whose size dependence corrected by $y < 0$. Finite integers N and N' denote the number of the fitting parameters, which is taken as $N = 2, N' = 0$ in this work. Examples for $\mathbb{Z}_2$-trivial and $\mathbb{Z}_2$-thermal metal (ThM) phase transitions are shown in Fig. S1, in which the rising(falling) of Λ for extended(localized) states are indeed observed.

Last but not least, let us note that appropriate boundary condition must be applied to detect the transition from or into the $\mathbb{Z}_2$ phase [S3]. In two-dimensional systems, we have two options: open and periodic boundary condition (OBC or PBC) along the direction perpendicular to the transferred direction. With the

OBC, the edge state appears along the transferred direction, and hence metallic. However, with the PBC, the state is merely localized in the first or the last block, which is similar to insulating state. Thus, to determine $\mathbb{Z}_2$-trivial ($\mathbb{Z}_2$-ThM) phase boundary, we must consider OBC (PBC) system. Note that, trivial-ThM boundary is detected in the either way.

II. NONCOMMUTATIVE GEOMETRY

In this section, we introduce the noncommutative geometry approach to map out the phase diagram of 2d class DIII system. The $\mathbb{Z}_2$ index derived in Refs. [S6, S7] is numerically advantageous since it can be determined from the discrete spectrum of a certain compact operator without taking the disorder average. See Ref. [S8] for detailed numerical implementation. The definition of the $\mathbb{Z}_2$ index of 2d class DIII system is given as follows,

$$\nu = \text{ker dim}\,[\mathcal{A} - 1] \ \text{ modulo 2}, \tag{S6}$$

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,

$$\mathcal{A} = P_{\mathrm{F}} - \mathcal{D}_a^* P_{\mathrm{F}} \mathcal{D}_a. \tag{S7}$$

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

$$\mathcal{D}_a(\boldsymbol{r}) := \frac{r_1 + ir_2 - (a_1 + ia_2)}{|r_1 + ir_2 - (a_1 + ia_2)|}, \tag{S8}$$

where $\boldsymbol{r} = (r_1, r_2) \in \mathbb{Z}^2$ denotes the position operator of a square lattice and $\boldsymbol{a} = (a_1, a_2) \in \mathbb{R}^2 \backslash \mathbb{Z}^2$ 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 λ_i as the i-th eigenvalue of the operator $\mathcal{A}$ in descending order including multiplicity.

Shown in Fig. S2(a) is $\lambda_1 - \lambda_2$ as a function of the chemical potential μ 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. S2(c)] and $\lambda_1 - \lambda_2 \neq 0$ evidently hold and thus imply $\nu = 1$. In Fig. S2(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 Fig. S2(b)(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

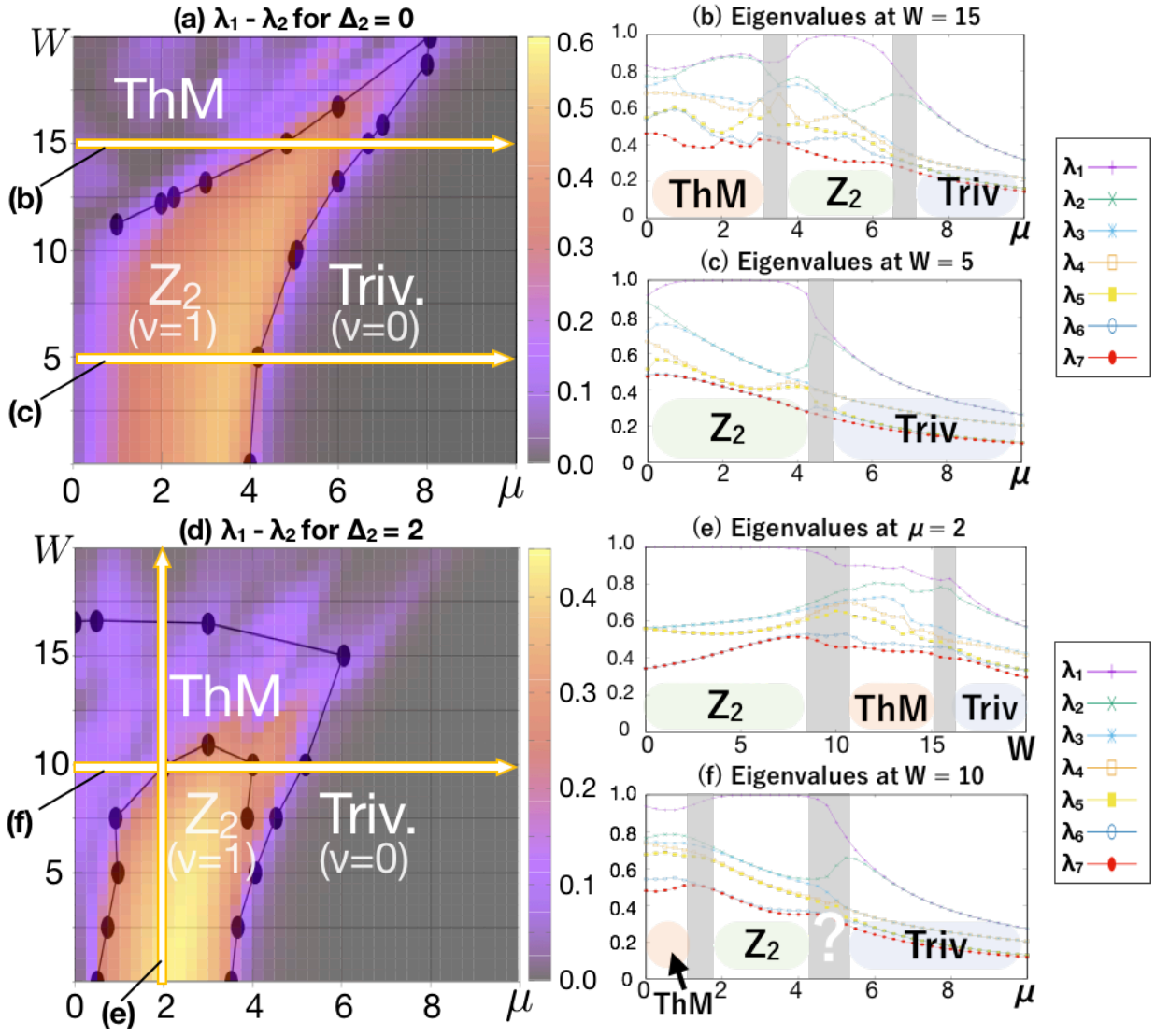

FIG. S2: (a) and (d) $\lambda_1 - \lambda_2$ as a function of the chemical potential μ and the disorder strength W for class DIII Hamiltonian given by Eq. (1)-(5) of the main text. The parameters are taken as (t, Δ, Δ_2) = (a) (1.0, 3.0, 0.0) and (d) (1.0, 3.0, 2.0), respectively, and the system size is 20×20. (b), (c), (e), and (f) show μ and W dependences of the eigenvalues λ_i, $i = 1, 2, \ldots, 7$, of the operator $\mathcal{A}$ for the system size 20×20. The parameters are taken as (t, Δ, Δ_2, W) = (b) (1.0, 3.0, 0.0, 15.0), (c) (1.0, 3.0, 0.0, 5.0), (f) (1.0, 3.0, 2.0, 10.0), and (e) $(t, \Delta, \Delta_2, \mu)$ = (1.0, 3.0, 2.0, 2.0), respectively. The gray bars in (b), (c), (e), and (f) denote marginal areas.

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. S2(b).

The difference between the first and second eigenvalues for $\Delta_2 = 2$ is also given in Fig. S2(d). In the orange-colored region, $\lambda_1 \sim 1$ [see, for instance, Fig. S2(e)] and $\lambda_1 - \lambda_2 \neq 0$, and hence $\nu = 1$ which corresponds to the $\mathbb{Z}_2$ phase. The black-colored region denotes the trivial phase with $\nu = 0$ because there is

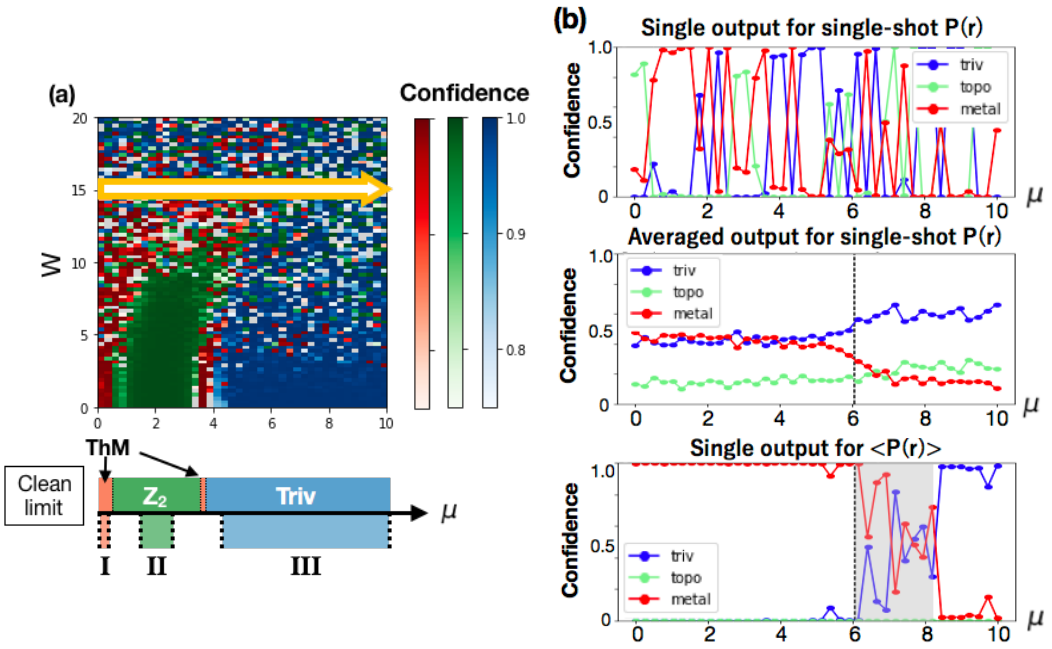

FIG. S3: (a) The output of ANN for single-shot $P(\mathbf{r})$ for $\Delta = 3, \Delta_2 = 2$. Boundary between phases are hardly recognizable. (b) The single output for $P(\mathbf{r})$, the average of 200 outputs for independently generated $P(\mathbf{r})$, and the single output for $\langle P(\mathbf{r})\rangle$ with $N_r = 500$ from the top. The amplitude of the random potential is fixed as $W = 15$.

no $\lambda_1 \sim 1$ [see Fig. S2(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. S2(e), the two phases are distinguishable by the presence of the doublet structure, whereas in Fig. S2(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.

III. 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 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. S3(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. S3(b) is the output for single-shot. The totally randomness reflects 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 the 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 feature of the quasi-translational states are detected appropriately.

[S1] A. MacKinnon and B. Kramer, Z. Phys. B **53**, 1 (1983).

[S2] L. Molinari, J. Phys. A: Math. Gen. **30**, 983 (1997).

[S3] A. Yamakage, K. Nomura, K. Imura, and Y. Kuramoto, Phys. Rev. B **87**, 205141 (2013).

[S4] V. I. Oseledec, Trans. Moscow Math. Soc. **19**, 197 (1968).

[S5] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).

[S6] H. Katsura and T. Koma, J. Math. Phys. **57**, 021903 (2016).

[S7] H. Katsura and T. Koma, arXiv:1611.01928 (2016).

[S8] Y. Akagi, H. Katsura, and T. Koma, arXiv:1709.05853 (2017).