Many-body mobility edges in one and two dimensions revealed by convolutional neural networks

Anffany Chen *

Theoretical Physics Institute, University of Alberta, Edmonton, Alberta T6G 2E1, Canada
and Department of Physics, University of Alberta, Edmonton, Alberta T6G 2E1, Canada

(Received 9 January 2024; accepted 19 January 2024; published 12 February 2024)

We adapt a machine-learning approach to study the many-body localization transition in interacting fermionic systems on disordered one-dimensional (1D) and two-dimensional (2D) lattices. We perform supervised training of convolutional neural networks (CNNs) using labeled many-body wave functions at weak and strong disorder. In these limits, the average validation accuracy of the trained CNNs exceeds 99.95%. We use the disorder-averaged predictions of the CNNs to generate energy-resolved phase diagrams, which exhibit many-body mobility edges. We provide finite-size estimates of the critical disorder strengths at $W_c \sim 2.8$ and 9.8 for 1D and 2D systems of 16 sites, respectively. Our results agree with the analysis of energy-level statistics and inverse participation ratio. By examining the convolutional layer, we unveil its feature extraction mechanism which highlights the pronounced peaks in localized many-body wave functions while rendering delocalized wave functions nearly featureless.

DOI: 10.1103/PhysRevB.109.075124

I. INTRODUCTION

Artificial neural networks have proven to be valuable assets in tackling a wide range of problems in condensed matter physics [1,2]. With their remarkable ability to discern universal features from extensive datasets and generalize to unseen data, neural networks can be trained to perform quantum-state tomography [3], accelerate *ab initio* calculations [4–6], and classify various phases of matter based on numerical [7–23] and experimental data [24–27]. As universal function approximators [28,29], neural networks have also been utilized as variational *Ansätze* for many-body quantum states [30–34], achieving ground-state estimations on par with the state-of-the-art conventional methods.

One notable application of neural networks is in characterizing the many-body localization transition between an ergodic many-body quantum system, following the eigenstate thermalization hypothesis (ETH) [35–37], and a many-body localized (MBL) phase under strong disorder [38–40]. According to ETH, an isolated, quantum many-body system goes through quantum thermalization over time by acting as its own heat bath, with all local observables eventually assuming thermal expectation values. Introducing sufficiently strong disorder can induce a transition into the MBL phase, where all energy eigenstates become localized, rendering the system nonergodic and unable to self-thermalize. Therefore, the striking signature of the MBL phase is a partial retention of the initial condition over long times. This phenomenon has been experimentally observed in one-dimensional (1D) and two-dimensional (2D) ultracold gases [41,42] and may potentially serve as a mechanism for robust quantum memory.

To characterize the ETH-MBL transition, the conventional numerical approach is to perform finite-size scaling analyses on simulated data of observables (such as level statistics) over a range of system sizes [43–53]. This task is computationally demanding due to the exponential growth of Fock space with system size N. Furthermore, the ETH-MBL transition is known to suffer strongly from the finite-size effect, such that the apparent phase transition drifts toward strong disorder as N increases. Extrapolating finite-size results to the thermodynamic limit through data collapses is therefore subject to ambiguity, particularly as numerically accessible system sizes are limited to $N \sim O(10)$ sites [54]. Currently, a consensus on the scaling theory for this transition remains elusive.

Machine learning offers a promising alternative approach for characterizing the ETH-MBL transition. Authors of recent studies [8,16,18–23] have successfully automated the classification of the phases using data obtained from exact diagonalization of model Hamiltonians, most of which are 1D spin models. The types of data considered in these studies include many-body energy spectra, many-body wave functions, entanglement spectra of these wave functions, and other variants obtained through additional feature engineering. A range of learning algorithms, both supervised and unsupervised, has been implemented, including the use of support vector machines and various neural network architectures. Notably, results of these studies show that machine learning can provide a finite-size estimate of the phase transition using data from only a single system size. Without the need for scaling analysis to locate the transition point, phase diagrams can be efficiently generated.

In this paper, we employ the machine-learning approach to investigate the ETH-MBL transition in interacting fermionic systems on 1D and 2D disordered lattices, each consisting of 16 sites. Our method strategically pairs unprocessed many-body wave functions as input data with convolutional

*anffany@ualberta.ca

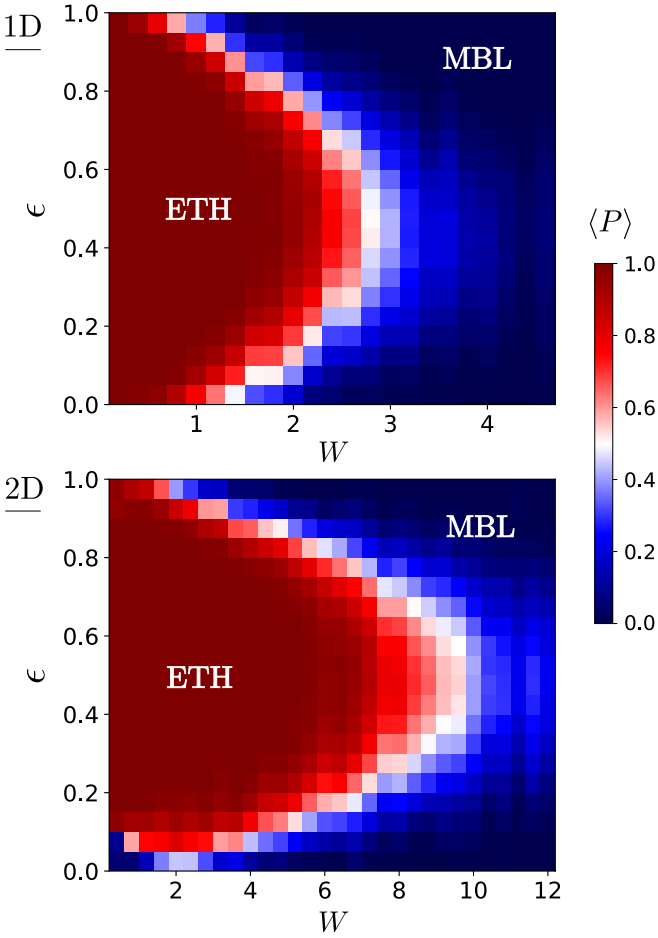

FIG. 1. Machine-predicted phase diagrams. Trained on many-body wave functions at the weak and strong disorder limits, our convolutional neural networks (CNNs) effectively generalize to classify wave functions near the transition region. We generate the phase diagrams of 16-site one-dimensional (1D) and two-dimensional (2D) disordered fermionic t-V models described by Eq. (1) using the disorder-averaged CNN prediction $\langle P \rangle$, representing the probability of the eigenstate thermalization hypothesis (ETH) phase. In both phase diagrams, the many-body mobility edge is clearly visible as the division between ETH (red) and many-body localized (MBL; blue) phases. The finite-size estimates of the critical disorder strengths are $W_c \sim 2.8$ in 1D and $W_c \sim 9.8$ in 2D.

neural networks (CNNs) for classification. Designed for image recognition, CNNs are expected to be equally suited for learning the local features in many-body wave functions. We train CNNs with a simple architecture to differentiate wave functions sampled from deep within the ETH and MBL phases, each labeled according to its respective phase. The trained CNNs are then tasked with classifying wave functions from the intermediate region. The disorder-averaged predictions of the CNNs are used to construct phase diagrams over energy density ϵ and disorder strength W, which clearly display the many-body mobility edges in 1D and 2D, as shown in Fig. 1.

In the following, we first introduce the fermionic t-V model on 1D and 2D disordered lattices (Sec. II A) and detail our procedure for collecting eigenstate samples via exact diagonalization (Sec. II B). We delve into the supervised training of our neural-network phase classifiers, describing the input data (Sec. III A), convolutional network architecture (Sec. III B), and training techniques (Sec. III C). We then present the energy-resolved phase diagrams based on the predictions of trained CNNs (Sec. IV A), compare them with the transition behaviors of energy-level statistics and inverse participation ratio (IPR; Sec. IV B), and interpret the decision-making mechanism of our trained CNNs (Sec. IV C).

II. FERMIONIC t-V MODELS WITH DISORDER

A. Model construction

We consider repulsive spinless fermions hopping on 1D and 2D lattices with random on-site potentials. The Hamiltonian is given by

$$H = \sum_{\langle i,j \rangle} \left[-t(c_i^\dagger c_j + c_j^\dagger c_i) + V\left(n_i - \frac{1}{2}\right)\left(n_j - \frac{1}{2}\right) \right] + \sum_{i=1}^{N} u_i \left(n_i - \frac{1}{2}\right), \tag{1}$$

where $c_i^\dagger$ creates a spinless fermion at site i, $n_i = c_i^\dagger c_i$ is the number operator, $\langle i, j \rangle$ goes over combinations of nearest neighbors, N is the system size, t is the hopping amplitude, V is the strength of the nearest-neighbor repulsive interaction, and u_i are on-site potentials randomly drawn from a uniform distribution in the range $[-W, W]$. For $t = \frac{1}{2}$ and $V = 1$, Eq. (1) on a 1D chain can be exactly mapped via Jordan-Wigner transformation to a spin-$\frac{1}{2}$ antiferromagnetic Heisenberg chain subject to a random field in the z direction [55,56]:

$$H = \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j + \sum_{i=1}^{N} u_i S_i^z, \tag{2}$$

This 1D spin/fermionic model has been well studied, with its critical disorder estimated to be $W_c \sim 3.5$ at zero total magnetization, $\sum_{i=1}^{N} S_i^z = 0$, corresponding to the half-filling sector in the fermionic picture [44,46,47,57,58]. The many-body mobility edge has been demonstrated through finite-size scaling analyses of various observables [46,58] and machine-learning technique [8], which we will use to benchmark our phase diagram in the 1D case. From here onward, we set $t = \frac{1}{2}$ and $V = 1$, and focus on half-filling for both the 1D and 2D systems.

The lattice geometries considered here are 1D chain and 2D square lattices, both given periodic boundary conditions to prevent localization by the boundaries. Each lattice consists of $N = 16$ sites, with the 2D lattice arranged as 4×4. To construct the many-body Hamiltonians, we start by defining the creation operators in the occupancy number basis of the 2^{16}-dimensional Fock space. The basis states are ordered by the total number of particles N_f, such that a particle-number conserving term like $c_i^\dagger c_j$ would be block-diagonal with each block corresponding to a specific N_f. Using these creation operators, we construct the many-body Hamiltonian as per Eq. (1) and the specified lattice geometries. In the following analysis, we focus on the half-filling sector,

$N_f = 8$, described by the $\mathcal{D} \times \mathcal{D}$ diagonal block with $\mathcal{D} = N$ choose $N_f = 12\,870$.

B. Exact diagonalization

For training neural networks to differentiate ETH and MBL wave functions, we select representative disorder strengths deep within each phase: $(W_{\text{ETH}}, W_{\text{MBL}}) = (0.2, 12)$ for 1D and $(0.4, 24)$ for 2D. Choosing these values does not require knowledge of the critical disorder W_c because one can always verify that conventional observables follow the expected ETH/MBL behaviors at these values. For the phase diagrams, we define suitable grids of W values in the intermediate regions: $W \in [0.2, 4.6]$ for 1D and $[0.4, 12]$ for 2D. At each selected W value, we implement 50 random disorder realizations and perform exact diagonalization of the Hamiltonians. Each energy spectrum is normalized as $\epsilon = (E - E_{\text{min}})/(E_{\text{max}} - E_{\text{min}})$, where ϵ is the energy density and $E_{\text{min}}/E_{\text{max}}$ is the smallest/largest eigenvalue of the spectrum. For the phase diagram in the 2D case, we use 50 additional disorder realizations for each W value in the range [8.4, 10].

Eigenstates from each disorder realization are binned into 20 equal energy intervals between $\epsilon = 0$ and 1. In each bin, we discard all but the 50 eigenstates with energy densities closest to the center of the bin, greatly reducing data storage and computational demands during subsequent analysis. We do, however, keep all the eigenvalues for computing the energy-level statistics later. We observe that using such a small sample of eigenstates does not significantly affect the disorder-averaged values of IPR and machine predictions. Note that, due to the low density-of-states near $\epsilon = 0$ and 1, bins in these regions contain <50 eigenstates per disorder realization.

III. NEURAL-NETWORK PHASE CLASSIFIER

Neural networks are complex, nonlinear functions consisting of alternating layers of linear and nonlinear maps. The linear maps are defined by a large number of adjustable parameters, *weights* and *biases*. The nonlinear maps are *activation functions* that mimic the behavior of biological neurons, which produce an output only when the input exceeds a certain threshold. A given neural network can be trained to approximate any function to a certain degree of accuracy. In *supervised learning*, the model is provided with a training dataset consisting of input-output pairs, and the weights and biases are adjusted to minimize a *loss function*, which quantifies the difference between the predictions of the model and the correct outputs. We refer to Refs. [59,60] for comprehensive introductions to neural networks and machine learning.

Our objective is to train a neural network to approximate the hypothetical function which maps a many-body wave function to the correct binary classification, ETH (labeled 1) or MBL (labeled 0). Our neural network would merely be an approximation to this function, so its output would not be binary but rather a continuous real number P ranging from 0 to 1, representing the probability that the input wave function belongs to the ETH phase. Upon disorder averaging, the prediction of the model can be regarded as an order parameter $\langle P \rangle$, transitioning from 1 in the ETH phase to 0 in the MBL phase. Assuming that the trained model is not biased toward one phase over the other, the point $\langle P \rangle = 0.5$ can be interpreted as the critical point for the $N = 16$ systems considered here.

A. Input data

For the input data, we use the probability densities $|\Psi_j|^2$ of the many-body wave functions, where j labels the occupancy number basis. This choice is economical because (i) each wave function serves as a data sample, unlike using the energy spectrum as the input which requires one exact diagonalization per sample, and (ii) it avoids additional feature engineering, such as calculating the entanglement spectra of the wave functions, which increases the computational costs. Moreover, this approach does not assume *a priori* wave function behaviors in either phase, leading to data-driven results.

Every time we train a model, we prepare a set of labeled data by randomly selecting 10 000 ETH and 10 000 MBL wave functions with energy densities $0.15 < \epsilon < 0.85$ collected at disorder strengths W_{ETH} and W_{MBL}, as discussed in Sec. II B, and pair them with outputs of 1 and 0, respectively. We observe that the MBL wave functions are typically localized on a subset of basis states, displaying a few highly pronounced peaks, while the ETH wave functions are distributed across all basis states. This visible difference motivates our choice of using CNNs to classify the wave functions, as CNNs are designed for image recognition. Our input probability densities can be viewed as 1 by 12 870 grayscale images, with each pixel representing the probability at a specific basis state.

B. CNN

As shown in Fig. 2, our simple CNN consists of (i) a convolutional layer followed by a max-pooling layer for feature extraction, (ii) a dense layer with dropout regularization for classification, and (iii) an output layer of one sigmoid neuron for prediction. These layers are built in Python using the TensorFlow package [61]. In the following, we describe the operations in each layer and their purposes.

1. Convolutional layer

The linear map in this layer is the convolution operation between *kernels* (or *filters*) and the input data. The total number of kernels m and the length of each kernel l are *hyperparameters* which are fixed prior to the training process. Having multiple kernels are crucial for detecting different local features in the input data. Each kernel slides across the input with a stride of 1, computing dot products between its array of l weights and the corresponding segment of the input data it covers at each position. Each dot product, added with the bias of the kernel, is passed through a nonlinear rectified linear unit (ReLU) activation function defined as

$$\text{ReLU}(x) = \max(0, x), \tag{3}$$

which sets negative values to zero. The outputs of the ReLU neurons form a feature map, so m kernels give rise to m feature

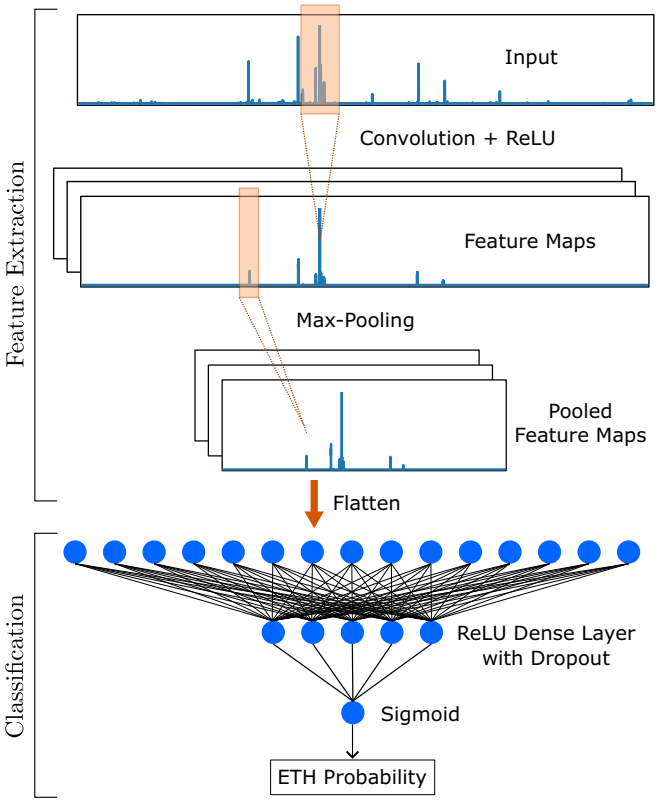

FIG. 2. Convolutional neural network (CNN) architecture. For the task of classifying many-body wave functions, we design a simple CNN composed of (i) a convolutional layer followed by max pooling for feature extraction, (ii) a dense layer with dropout regularization for classification, and (iii) an output layer consisting of a single sigmoid neuron for predicting the probability for the input wave function to be in the eigenstate thermalization hypothesis (ETH) phase. All activation functions are chosen to be rectified linear units (ReLUs) except in the output layer.

maps. The weights and biases in the kernels are initialized randomly and optimized during the training process.

2. Max-pooling layer

This layer performs a down-sampling operation dictated by a hyperparameter p. It slides a window of size p (with a stride of p) across each feature map and selects the maximum value within each window. This process results in m pooled feature maps with length reduced by a factor of p.

3. Dense layer

The pooled feature maps are flattened into a single 1D vector v of length L, which is then fed into a dense (or fully connected) layer. A hyperparameter q dictates the number of ReLU neurons in this layer. The linear operation:

$$f(v) = Av + b, \tag{4}$$

where A is the $q \times L$ weight matrix and b is the bias vector, maps v to a vector of length q. The ReLU activation function is then applied to the resulting vector element wise. The weights and biases in A and b are optimized during training.

4. Dropout layer

A dropout layer following the dense layer randomly deactivates a fraction d of the neurons in the dense layer during training by setting their activation functions to zero. This is a regularization technique designed to prevent overfitting to nonuniversal features specific to the training data. The introduced randomness prevents any single neuron in the dense layer from becoming too specialized to certain patterns from the training data, encouraging the model to learn more robust and universal feature. The dropout layer only operates during training.

5. Output layer

In the final layer, the output of the dense layer is multiplied by a $1 \times q$ weight matrix and then combined with a bias; as before, the weights and bias are optimized during training. The resulting value is passed through a sigmoid activation function given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{5}$$

which maps any real number into a value $P \in [0, 1]$. For the task of binary classification, P can be interpreted as the probability for the 1 class. In this case, it represents the probability for the input wave function to be in the ETH phase.

C. Supervised training

1. Loss function

Supervised training of our CNN amounts to tuning the weights and biases to minimize the difference between model predictions and the correct labels provided in the training dataset. Here, the difference is measured by the *binary cross - entropy*, which is a common choice of loss function for binary classification:

$$\text{Loss} = -[y \ln(P) + (1 - y) \ln(1 - P)], \tag{6}$$

where y is the correct label, and P is the model prediction. When the label is 0, the first term in Eq. (6) vanishes and the loss function is approximately P for $P \sim 0$. Similarly, when the label is 1, the loss function approximately measures how far P strays from 1.

2. Gradient descent

The dataset prepared in Sec. III A is randomly split 50/50 into a training set and a test set for evaluating the performance of the model on unseen data. The training process is organized into *epochs*; each is a complete pass through the entire training set. During each epoch, the training set is randomly divided into smaller *batches* for mini-batch gradient descent. We typically set the batch size to be 50 samples. Each batch goes through the layers of the model, a process known as *feedforward*. For each sample in the batch, a loss is computed according to Eq. (6). Then the *backpropagation* algorithm calculates the gradient of the loss with respect to the model parameters. This gradient is averaged over all samples in the batch. The optimization algorithm uses this average gradient to update the parameters, moving them in the direction of the steepest descent (opposite to the gradient), with the magnitude

of change controlled by the *learning rate*. As opposed to fixing the same learning rate for all parameters, we have opted to use TensorFlow's Adam optimizer, which adjusts the learning rate for each parameter throughout the training process.

3. Training history

At the end of each epoch, the average loss over the test set is computed and recorded, known as the validation loss. A successful training is marked by a decreasing trend in both training loss (average loss over training data) and validation loss as training progresses over successive epochs, which indicates that the model is learning and generalizing well to unseen data. However, an increase in validation loss coupled with a decrease in training loss can signal potential overfitting, where the model memorizes the training data instead of learning generalizable features. To prevent significant overfitting, we implement *early stopping*, which halts training when validation loss stops decreasing for a few consecutive epochs.

4. Cross-validation

Due to the inherent randomness in the training process, including weight/bias initialization and data sampling, model training can vary with each iteration, often converging to different local minima in the loss function landscape. To evaluate model performance reliably, we perform k-fold cross-validation, training the model $k = 20$ times with different training and test sets. Examining all training histories together enables accurate assessment of the performance of the model under a given set of hyperparameters, which facilitates hyperparameter tuning.

5. Hyperparameters

Through experimentation, we determined that the following set of hyperparameters yields optimal validation loss at the end of training—$m = 16$, $l = 10$, $p = 2$, $q = 60$, and $d = 0.2$—for models trained with wave functions of the 1D system. This set of hyperparameters leads to 99.99% validation accuracy (averaged over k folds of training). Validation accuracy is defined as the percentage of correct predictions on the test data, considering a prediction correct if its rounded integer value matches the label. Minor variations in these hyperparameters do not significantly affect performance. We thus fix the hyperparameters at these values when training CNN models for the 1D system.

However, we find that this set of hyperparameters is suboptimal for training with wave functions of the 2D system, resulting in an average test accuracy of 99.74%. Further experimentation reveals that increasing the kernel size l from 10 to 100 and the dropout rate d from 0.2 to 0.5 significantly improves test accuracy, achieving an average of 99.97%. The improved performance due to a larger kernel suggests that the important local features in wave functions of the 2D system likely span a wider range of basis states. Accordingly, we fix the hyperparameters at $m = 16$, $l = 100$, $p = 2$, $q = 60$, and $d = 0.5$ for the 2D case.

IV. RESULTS AND DISCUSSION

A. Machine-predicted phase diagrams

To generate the energy-resolved phase diagram of the 1D fermionic chain, we trained 20 CNN models, with architecture described in Sec. III B and hyperparameters in Sec. III C, using wave functions at disorder strengths $W_{\mathrm{ETH}} = 0.2$ and $W_{\mathrm{MBL}} = 12$ (see Secs. II B and III A for details on training data). Similarly, in the 2D case, we trained another 20 CNN models, this time using wave functions of the 2D fermionic system at $W_{\mathrm{ETH}} = 0.4$ and $W_{\mathrm{MBL}} = 24$. Our trained models demonstrate >99.95% validation accuracy in predicting the correct phases of the test wave functions.

Exploiting the generalization capacity of neural networks, we input wave functions from the intermediate regions. At every pair of discretized energy density ϵ and disorder strength W, we first averaged the prediction of one trained CNN over wave functions belonging to one disorder realization. These averages were then further averaged over all disorder realizations and 20 CNNs. The disorder- and training-averaged probability for the ETH phase forms the phase diagrams shown in Fig. 1. Note that the models trained on wave functions from the 1D and 2D systems were specifically used to produce their respective 1D and 2D phase diagrams. We further note that we did not follow the common practice of truncating the phase diagrams at low and high energies since the average prediction of our CNNs appears convergent despite the scarcity of data in these regions.

In both phase diagrams, the mobility edge is clearly visible as the division between ETH (red) and MBL (blue) phases. For the 1D system, the mobility edge agrees with previous studies [8,46], exhibiting a characteristic bell shape with the tip dropping slightly below $\epsilon = 0.5$. We estimate the critical disorder to be $W_c \sim 2.8$, agreeing with a previous machine-predicted estimate for the $N = 16$ chain [8]. Our finite-size estimate of W_c is smaller than the thermodynamic limit $W_c \sim 3.5$ determined through finite-size scaling analyses [44,46,47,57,58]. This difference is expected due to the strong finite-size effect at the ETH-MBL transition.

In comparison, the phase diagram of the 2D system exhibits notable differences. The tip of the mobility edge is more aligned to $\epsilon = 0.5$. Moreover, the eigenstates at small W are predicted to have high probability of localization near $\epsilon = 0$, in contrast with the 1D case where no states are localized at small W. The estimated critical disorder is $W_c \sim 9.8$, significantly greater than the 1D case. The increase in critical disorder with higher spatial dimension is a well-known phenomenon in the noninteracting limit and can be understood in terms of classical random walks on lattices [62].

B. Comparison with conventional observables

To verify the machine-predicted phase diagrams, we analyze two conventional observables across the transition. The first observable is the energy-level statistics based on the gap ratios in the many-body energy spectrum [43]:

$$r_\alpha = \frac{\min\{\epsilon_{\alpha+1} - \epsilon_\alpha, \epsilon_\alpha - \epsilon_{\alpha-1}\}}{\max\{\epsilon_{\alpha+1} - \epsilon_\alpha, \epsilon_\alpha - \epsilon_{\alpha-1}\}}, \tag{7}$$

where ϵ_α's are the sorted energy densities of a single disorder realization. Using the full energy spectrum of each disorder realization, we compute the averaged gap ratio within each of the 20 energy intervals between $\epsilon = 0$ and 1. These values are then averaged over all disorder realizations. In the ETH phase, the expected value of the disorder-averaged gap ratio $\langle r \rangle$ is $\langle r \rangle_{\text{GOE}} = 4 - 2\sqrt{3} = 0.536$ [63], corresponding to the Wigner surmise of the Gaussian orthogonal ensemble (GOE). In the MBL phase, the level statistics is described by the Poisson distribution, averaging to $\langle r \rangle_{\text{P}} = 2\ln 2 - 1 = 0.386$.

The second observable is the IPR, defined as

$$\text{IPR}(\Psi) = \sum_{j=1}^{\mathcal{D}} |\Psi_j|^4, \tag{8}$$

where Ψ is a many-body wave function and j goes over the occupancy number basis in the half-filling sector. The inverse of IPR quantifies the support of Ψ in our choice of basis. For each disorder realization, we average the IPR of up to 50 eigenstates per energy interval (see Sec. II B for details on data sampling). These values are then averaged over all disorder realizations. In contrast with the Anderson localization transition, where $\langle \text{IPR} \rangle$ defined in the real-space position basis increases from $1/N$ to 1 toward the strong disorder limit, the $\langle \text{IPR} \rangle$ in the Fock-space basis increases more gradually. Unlike the former, it does not approach 1 because localized many-body wave functions have nonzero support over many basis states even at strong disorder. Thus, we compute the disorder-average of log(IPR) to highlight the transition from ETH to MBL.

In Fig. 3, we plot $\langle r \rangle$ and $\langle \ln(\text{IPR}) \rangle$ as functions of ϵ and W. For these conventional observables, pinpointing the transition boundary involves analyses over various system sizes and finite-size scaling. Thus, we focus on the contours, which are lines of equal value. In both the 1D and 2D cases, the $\langle \ln(\text{IPR}) \rangle$ contour at around -3.5 is very similar to the mobility edge in Fig. 1. On the other hand, the contours of $\langle r \rangle$, while consistent with the mobility edges, have less pointy profiles with weaker curvature. Note that our data obtained from 50 disorder realizations are insufficient for converging $\langle r \rangle$ near $\epsilon = 0$ and 1 where eigenvalues are scarce.

C. Model interpretation

To understand the decision making of our trained CNNs, we examine the kernel weights and biases and the feature maps generated from the input data. Prior to classification by the dense layer, the input probability densities $|\Psi|^2$ undergo a series of operations: convolution with the kernels plus biases (resulting in convolution feature maps), ReLU activation (feature maps), and max pooling (pooled feature maps). For both CNNs trained on the 1D and 2D systems, the kernels generally have highly fluctuating weights between -1 and 1, along with small negative biases [see Fig. 4(a) for examples in the 1D case]. During convolution, these weights effectively scale down $|\Psi|^2$, which is then shifted downward by the negative biases. The negative values in the resulting convolution feature maps, due to the negative biases, are truncated by the ReLU activation. Lastly, the max-pooling layer, with a minimal pool size of 2, down-samples the feature maps by a factor of two without significantly altering the extracted features.

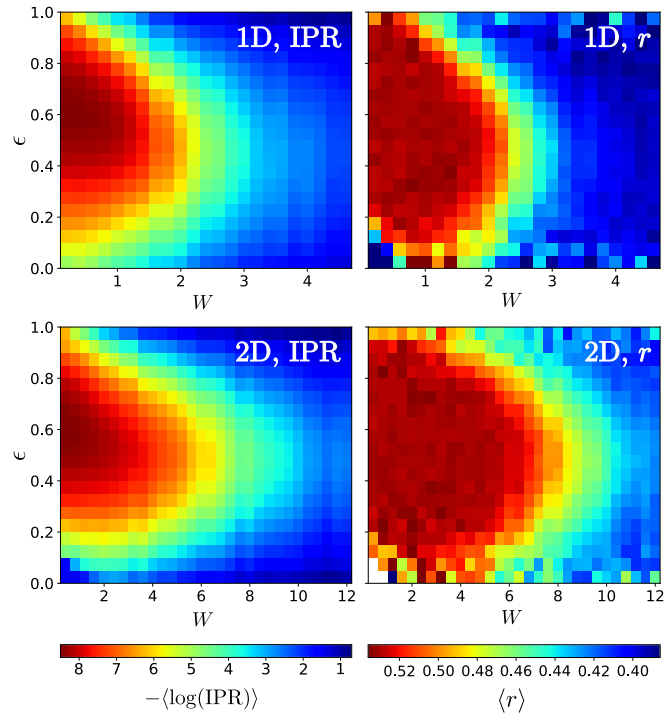

FIG. 3. Inverse participation ratio (IPR) and energy-level statistics. For the 16-site one-dimensional (1D) and two-dimensional (2D) systems, disorder-averaged $\langle \log(\text{IPR}) \rangle$ and gap ratio $\langle r \rangle$ as functions of energy density ϵ and disorder strength W bear similar qualitative features as the machine-predicted phase diagrams (Fig. 1). For $\langle r \rangle$, the max and min values of the colorbar correspond to $\langle r \rangle_{\text{GOE}}$ and $\langle r \rangle_{\text{P}}$, respectively, and the noise is due to insufficient eigenvalue data near $\epsilon = 0$ and 1.

Figure 4(b) shows the feature extraction process by a typical kernel applied to an ETH/MBL wave function of the 1D system. The ETH wave function, characterized by its low probability density at almost all basis states, becomes nearly featureless after the application of the negative bias followed by ReLU activation. In contrast, the same operation accentuates the pronounced peaks in the MBL wave function, reducing smaller signals to zero while preserving the more significant ones. This explains how our convolutional layer effectively highlights the key differences between MBL and ETH wave functions, thereby simplifying the classification task for the dense layer.

V. CONCLUSIONS

In this paper, we investigated interacting spinless fermionic systems on 1D and 2D lattices with random on-site potentials, focusing on systems of 16 sites. Through exact diagonalization, we collected many-body wave functions from various disorder realizations. We then conducted supervised training of neural networks using wave functions at weak and strong disorder, labeled as ETH and MBL, respectively. We specifically chose CNNs for their capability in local pattern recognition. Utilizing effective training techniques, including dropout regularization and cross-validation, our CNNs

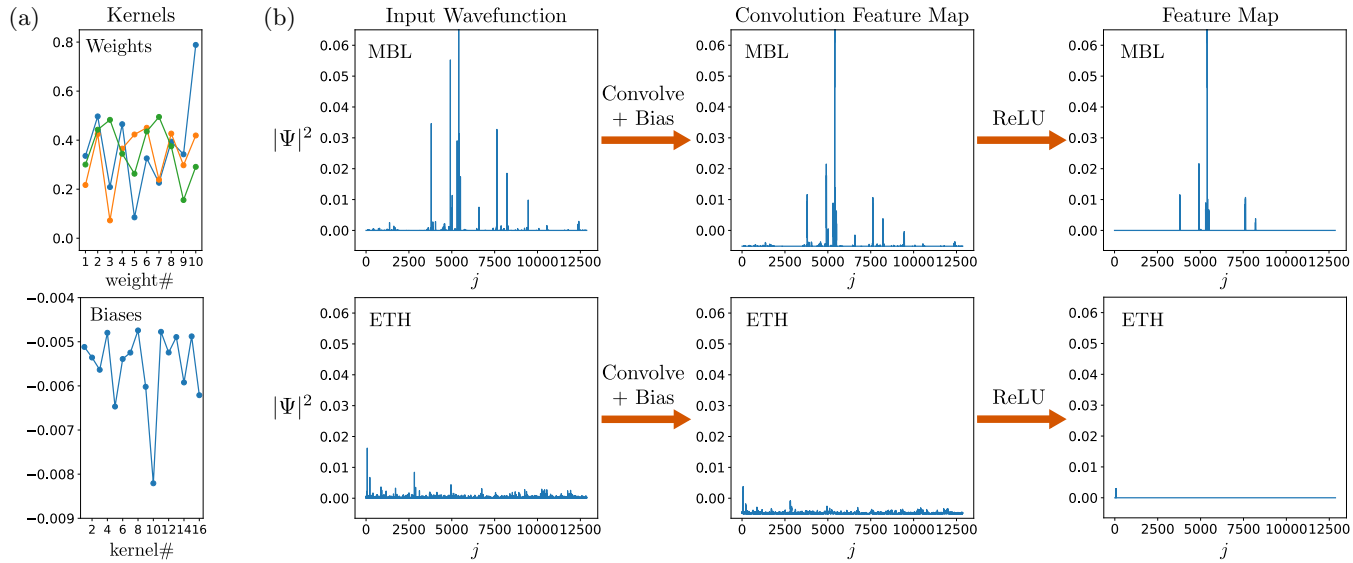

FIG. 4. Interpretation of the convolutional layer. (a) (top panel) The kernel weights in our trained convolutional neural networks (CNNs) are highly fluctuating values between -1 and 1. Here, we show the weights of three kernels (each with $l = 10$ weights) belonging to a CNN trained on the wave functions of the one-dimensional (1D) system. (bottom panel) The kernel biases (one bias per kernel) of the same CNN are small negative numbers. (b) By learning to apply a small negative bias right before rectified linear unit (ReLU) activation, a typical kernel in our trained CNNs truncates small values in the input probability density $|\Psi|^2$, which renders an eigenstate thermalization hypothesis (ETH) wave function nearly featureless and accentuates the pronounced profile of a many-body localized (MBL) wave function. This feature extraction mechanism, demonstrated here with wave functions of the 1D system, is observed in both CNNs trained on 1D and two-dimensional (2D) systems.

achieved >99.95% accuracy on test data, successfully classifying wave functions deep in the ETH and MBL phases.

Leveraging the generalization ability of the neural network, we provided the CNNs with wave functions near the transition region and used the disorder-averaged prediction $\langle P \rangle$, representing the probability for the ETH phase, to construct phase diagrams. The energy-resolved phase diagrams over energy density ϵ and disorder strength W precisely locate the many-body mobility edges in both 1D and 2D systems. We estimated the critical disorder strengths to be $W_c \sim 2.8$ for 1D and $W_c \sim 9.8$ for 2D, applicable to finite-sized systems of 16 sites.

Our analysis of energy-level statistics and IPR corroborates our phase diagrams by showing similar qualitative features. We further examined the weights, biases, and feature maps of the CNN, gaining insights into its feature extraction mechanism. We found that the convolutional layer has learned to truncate small values in the input probability densities through negative biases and ReLU activation, effectively retaining only the strong input signals for classification. This mechanism was observed in both CNNs trained on 1D and 2D systems, demonstrating its applicability across different dimensions. In future studies, one could investigate its connection to conventional observables or potentially formulate order parameters inspired by the learned mechanism.

The ultimate success of the machine-learning approach for characterizing the ETH-MBL phase boundary hinges on precise quantification of the predictions of the machine. This includes quantifying the uncertainties in the predictions and conducting finite-size scaling analysis to extend finite-size results to the thermodynamic limit. Assessing the effectiveness of transfer learning, particularly by applying CNNs trained on 1D systems to classify wave functions in 2D systems and vice versa, could reveal whether the machine-based order parameter P is universal, independent of lattice configurations and spatial dimensions. Lastly, one could experiment with neural networks with more advanced architectures, which may generalize better to the transition region and lead to more precise determination of the phase boundary.

ACKNOWLEDGMENTS

This paper benefited greatly from discussions with I. Boettcher, J. Maciejko, C. Sun, S. Dey, and D. Noby Joseph. The numerical computation was enabled in part by support provided by Compute Ontario [64] and the Digital Research Alliance of Canada [65]. The author gratefully acknowledges the support of Natural Sciences and Engineering Research Council of Canada Discovery Grant No. RGPIN-2020-06999, Avadh Bhatia Fellowship, startup fund UOFAB Startup Boettcher, and the Faculty of Science at the University of Alberta.

[1] S. D. Sarma, D.-L. Deng, and L.-M. Duan, Machine learning meets quantum physics, Phys. Today **72**(3), 48 (2019).

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine

learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).
[3] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).
[4] R. M. Balabin and E. I. Lomakina, Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies, J. Chem. Phys. **131**, 074104 (2009).
[5] K. Ryczko, D. A. Strubbe, and I. Tamblyn, Deep learning and density-functional theory, Phys. Rev. A **100**, 022512 (2019).
[6] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, and Y. Xu, Deep-learning density functional theory Hamiltonian for efficient *ab initio* electronic-structure calculation, Nat. Comput. Sci. **2**, 367 (2022).
[7] T. Ohtsuki and T. Ohtsuki, Deep learning the quantum phase transitions in random two-dimensional electron systems, J. Phys. Soc. Jpn. **85**, 123706 (2016).
[8] F. Schindler, N. Regnault, and T. Neupert, Probing many-body localization with neural networks, Phys. Rev. B **95**, 245134 (2017).
[9] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).
[10] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).
[11] Y. Zhang and E.-A. Kim, Quantum loop topography for machine learning, Phys. Rev. Lett. **118**, 216401 (2017).
[12] W. Hu, R. R. P. Singh, and R. T. Scalettar, Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination, Phys. Rev. E **95**, 062122 (2017).
[13] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Machine learning quantum phases of matter beyond the fermion sign problem, Sci. Rep. **7**, 8823 (2017).
[14] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Machine learning phases of strongly correlated fermions, Phys. Rev. X **7**, 031038 (2017).
[15] Y.-H. Liu and E. P. L. van Nieuwenburg, Discriminative cooperative networks for detecting phase transitions, Phys. Rev. Lett. **120**, 176401 (2018).
[16] E. V. H. Doggen, F. Schindler, K. S. Tikhonov, A. D. Mirlin, T. Neupert, D. G. Polyakov, and I. V. Gornyi, Many-body localization and delocalization in large quantum chains, Phys. Rev. B **98**, 174202 (2018).
[17] M. Matty, Y. Zhang, Z. Papić, and E.-A. Kim, Multifaceted machine learning of competing orders in disordered interacting systems, Phys. Rev. B **100**, 155141 (2019).
[18] W. Zhang, L. Wang, and Z. Wang, Interpretable machine learning study of the many-body localization transition in disordered quantum Ising spin chains, Phys. Rev. B **99**, 054208 (2019).
[19] H. Théveniaut and F. Alet, Neural network setups for a precise detection of the many-body localization transition: Finite-size scaling and limitations, Phys. Rev. B **100**, 224202 (2019).
[20] P. Huembeli, A. Dauphin, P. Wittek, and C. Gogolin, Automated discovery of characteristic features of phase transitions in many-body localization, Phys. Rev. B **99**, 104106 (2019).
[21] W.-J. Rao, Machine learning for many-body localization transition, Chin. Phys. Lett. **37**, 080501 (2020).
[22] R. Kausar, W.-J. Rao, and X. Wan, Learning what a machine learns in a many-body localization transition, J. Phys.: Condens. Matter **32**, 415605 (2020).
[23] H. Théveniaut, Z. Lan, G. Meyer, and F. Alet, Transition to a many-body localized regime in a two-dimensional disordered quantum dimer model, Phys. Rev. Res. **2**, 033154 (2020).
[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 *et al.*, Machine learning in electronic-quantum-matter imaging experiments, Nature (London) **570**, 484 (2019).
[25] B. S. Rem, N. Käming, M. Tarnowski, L. Asteria, N. Fläschner, C. Becker, K. Sengstock, and C. Weitenberg, Identifying quantum phase transitions using artificial neural networks on experimental data, Nat. Phys. **15**, 917 (2019).
[26] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap, Classifying snapshots of the doped Hubbard model with machine learning, Nat. Phys. **15**, 921 (2019).
[27] S. Ghosh, M. Matty, R. Baumbach, E. D. Bauer, K. A. Modic, A. Shekhter, J. A. Mydosh, E.-A. Kim, and B. J. Ramshaw, One-component order parameter in URu_2Si_2 uncovered by resonant ultrasound spectroscopy and machine learning, Sci. Adv. **6**, eaaz4074 (2020).
[28] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, Neural Networks **2**, 359 (1989).
[29] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks **4**, 251 (1991).
[30] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).
[31] D.-L. Deng, X. Li, and S. Das Sarma, Machine learning topological states, Phys. Rev. B **96**, 195145 (2017).
[32] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, Phys. Rev. B **96**, 205152 (2017).
[33] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Symmetries and many-body excitations with neural-network quantum states, Phys. Rev. Lett. **121**, 167204 (2018).
[34] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, Phys. Rev. Lett. **124**, 020503 (2020).
[35] J. M. Deutsch, Quantum statistical mechanics in a closed system, Phys. Rev. A **43**, 2046 (1991).
[36] M. Srednicki, Chaos and quantum thermalization, Phys. Rev. E **50**, 888 (1994).
[37] M. Rigol, V. Dunjko, and M. Olshanii, Thermalization and its mechanism for generic isolated quantum systems, Nature (London) **452**, 854 (2008).
[38] D. Basko, I. Aleiner, and B. Altshuler, Metal-insulator transition in a weakly interacting many-electron system with localized single-particle states, Ann. Phys. **321**, 1126 (2006).
[39] R. Nandkishore and D. A. Huse, Many-body localization and thermalization in quantum statistical mechanics, Annu. Rev. Condens. Matter Phys. **6**, 15 (2015).
[40] D. A. Abanin, E. Altman, I. Bloch, and M. Serbyn, Colloquium: Many-body localization, thermalization, and entanglement, Rev. Mod. Phys. **91**, 021001 (2019).
[41] M. Schreiber, S. S. Hodgman, P. Bordia, H. P. Lüschen, M. H. Fischer, R. Vosk, E. Altman, U. Schneider, and I. Bloch, Observation of many-body localization of interacting fermions in a quasirandom optical lattice, Science **349**, 842 (2015).

[42] J.-Y. Choi, S. Hild, J. Zeiher, P. Schauß, A. Rubio-Abadal, T. Yefsah, V. Khemani, D. A. Huse, I. Bloch, and C. Gross, Exploring the many-body localization transition in two dimensions, Science **352**, 1547 (2016).

[43] V. Oganesyan and D. A. Huse, Localization of interacting fermions at high temperature, Phys. Rev. B **75**, 155111 (2007).

[44] A. Pal and D. A. Huse, Many-body localization phase transition, Phys. Rev. B **82**, 174411 (2010).

[45] S. Bera, H. Schomerus, F. Heidrich-Meisner, and J. H. Bardarson, Many-body localization characterized from a one-particle perspective, Phys. Rev. Lett. **115**, 046603 (2015).

[46] D. J. Luitz, N. Laflorencie, and F. Alet, Many-body localization edge in the random-field Heisenberg chain, Phys. Rev. B **91**, 081103(R) (2015).

[47] V. Khemani, S. P. Lim, D. N. Sheng, and D. A. Huse, Critical properties of the many-body localization transition, Phys. Rev. X **7**, 021013 (2017).

[48] R. K. Panda, A. Scardicchio, M. Schulz, S. R. Taylor, and M. Ånidari, Can we study the many-body localisation transition? Europhys. Lett. **128**, 67003 (2020).

[49] P. Sierant, D. Delande, and J. Zakrzewski, Thouless time analysis of Anderson and many-body localization transitions, Phys. Rev. Lett. **124**, 186601 (2020).

[50] J. Šuntajs, J. Bonča, T. Prosen, and L. Vidmar, Quantum chaos challenges many-body localization, Phys. Rev. E **102**, 062144 (2020).

[51] D. Abanin, J. Bardarson, G. De Tomasi, S. Gopalakrishnan, V. Khemani, S. Parameswaran, F. Pollmann, A. Potter, M. Serbyn, and R. Vasseur, Distinguishing localization from chaos: Challenges in finite-size systems, Ann. Phys. **427**, 168415 (2021).

[52] A. S. Aramthottil, T. Chanda, P. Sierant, and J. Zakrzewski, Finite-size scaling analysis of the many-body localization transition in quasiperiodic spin chains, Phys. Rev. B **104**, 214201 (2021).

[53] A. Morningstar, L. Colmenarez, V. Khemani, D. J. Luitz, and D. A. Huse, Avalanches and many-body resonances in many-body localized systems, Phys. Rev. B **105**, 174205 (2022).

[54] F. Pietracaprina, N. Macé, D. J. Luitz, and F. Alet, Shift-invert diagonalization of large many-body localizing spin chains, SciPost Phys. **5**, 045 (2018).

[55] I. Affleck, Field theory methods and quantum critical phenomena, in *Fields, Strings and Critical Phenomena*, edited by E. Brézin and J. Zinn-Justin (North Holland, Amsterdam, 1989), pp. 563–640.

[56] O. Derzhko, Jordan-Wigner fermionization for spin-$\frac{1}{2}$ systems in two dimensions: A brief review, J. Phys. Stud. **5**, 49 (2001).

[57] F. Alet and N. Laflorencie, Many-body localization: An introduction and selected topics, C. R. Phys. **19**, 498 (2018).

[58] T. Orito and K.-I. Imura, Multifractality and Fock-space localization in many-body localized states: One-particle density matrix perspective, Phys. Rev. B **103**, 214206 (2021).

[59] M. A. Nielsen, *Neural Networks and Deep Learning* (Determination Press, San Francisco, 2015).

[60] T. Neupert, M. H. Fischer, E. Greplova, K. Choo, and M. M. Denner, Introduction to machine learning for the sciences, arXiv:2102.04883.

[61] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv:1603.04467.

[62] A. Chen, J. Maciejko, and I. Boettcher, Anderson localization transition in disordered hyperbolic lattices, arXiv:2310.07978.

[63] Y. Y. Atas, E. Bogomolny, O. Giraud, and G. Roux, Distribution of the ratio of consecutive level spacings in random matrix ensembles, Phys. Rev. Lett. **110**, 084101 (2013).

[64] computeontario.ca.

[65] alliancecan.ca.