

# Analyzing non-equilibrium quantum states through snapshots with artificial neural networks

A. Bohrdt,<sup>1,2,3,4</sup> S. Kim,<sup>4</sup> A. Lukin,<sup>4</sup> M. Rispoli,<sup>4</sup> R. Schittko,<sup>4</sup> M. Knap,<sup>1,2</sup> M. Greiner,<sup>4</sup> and J. Léonard<sup>4</sup>

<sup>1</sup>*Department of Physics and Institute for Advanced Study,  
Technical University of Munich, 85748 Garching, Germany*

<sup>2</sup>*Munich Center for Quantum Science and Technology (MCQST), Schellingstr. 4, D-80799 München, Germany*

<sup>3</sup>*ITAMP, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*

<sup>4</sup>*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

(Dated: May 24, 2022)

Current quantum simulation experiments are starting to explore non-equilibrium many-body dynamics in previously inaccessible regimes in terms of system sizes and time scales. Therefore, the question emerges which observables are best suited to study the dynamics in such quantum many-body systems. Using machine learning techniques, we investigate the dynamics and in particular the thermalization behavior of an interacting quantum system which undergoes a non-equilibrium phase transition from an ergodic to a many-body localized phase. We employ supervised and unsupervised training methods to distinguish non-equilibrium from equilibrium data, using the network performance as a probe for the thermalization behavior of the system. We test our methods with experimental snapshots of ultracold atoms taken with a quantum gas microscope. Our results provide a path to analyze highly-entangled large-scale quantum states for system sizes where numerical calculations of conventional observables become challenging.

*Introduction.*— After a global quench in a thermalizing system, local observables approach a value which corresponds to their expectation value in a typical microcanonical many-body eigenstate of the system [1–3]. Depending on the properties of the system and the initial state, the path to thermal equilibrium can vary. For example, conserved quantities can slow down the equilibration process [4–6] or a quasi-stationary prethermal state can form, which exhibits properties different from the true thermal equilibrium state [7].

Quantum simulation experiments can enable the observation of the time-evolution of a quantum many-body system starting from a non-equilibrium state with almost perfect isolation from the environment. In the past decade, a variety of non-equilibrium phenomena has been observed with examples ranging from exotic phases realized through Floquet driving [8–10] to many-body localization [11] and prethermalization [12].

In many cases, theory can provide a clear prediction which observables should be studied, such as a given order parameter for a well-known phase transition. For some problems, however, it is not as clear which observable to look at, and by making a choice for one specific quantity, valuable information might be discarded. In many platforms with microscopic readout, Fock space snapshots of the quantum many-body state are the measured data set. Fock space snapshots provide a wealth of information about the quantum many-body state by providing access to both local observables and non-local, high-order correlations.

In order to address the challenge of finding suitable observables, artificial neural networks have recently emerged as a valuable tool in quantum many-body physics [13–17], and in nonequilibrium statistical me-

chanics [18]. Previous machine learning approaches to study non-equilibrium systems have focused on quantities such as the entanglement spectrum [19–21] or full eigenstates [22], which are, however, experimentally inaccessible.

In this work we study the dynamics of an interacting quantum many-body system in terms of experimental Fock space snapshots with the help of neural networks, Fig. 1a). We find this analysis to have two main advantages: (i) these snapshots are directly measured in many quantum simulation platforms, and large numbers of snapshots can be routinely obtained. (ii) Raw data is used, where no analysis for specific quantities has taken place and all available information can be used without any bias. We consider the one-dimensional Bose-Hubbard model

$$\hat{H} = \sum_i \left[ -J \left( \hat{a}_i^\dagger \hat{a}_{i+1} + \text{h.c.} \right) + \frac{U}{2} \hat{n}_i (\hat{n}_i - 1) + W h_i \hat{n}_i \right]. \quad (1)$$

Here,  $\hat{a}_i^{(\dagger)}$  annihilates (creates) a boson on site  $i$  and  $\hat{n}_i = \hat{a}_i^\dagger \hat{a}_i$  is the particle number operator. The first term corresponds to hopping between neighboring sites, the second term is the interaction, here fixed at  $U/J = 2.9$ , and the last term is the quasi-periodic potential mimicking on-site disorder with amplitude  $W$ , which can be created in a cold atom setup with an incommensurate lattice as  $h_i = \cos(2\pi\beta i + \phi)$ . In this work, we consider  $1/\beta = 1.618$ .

This system exhibits a many-body localized (MBL) phase, where thermalization breaks down as the disorder strength is increased beyond a critical value. The transition from an ergodic to a many-body localized phase is fundamentally different from the well-studied

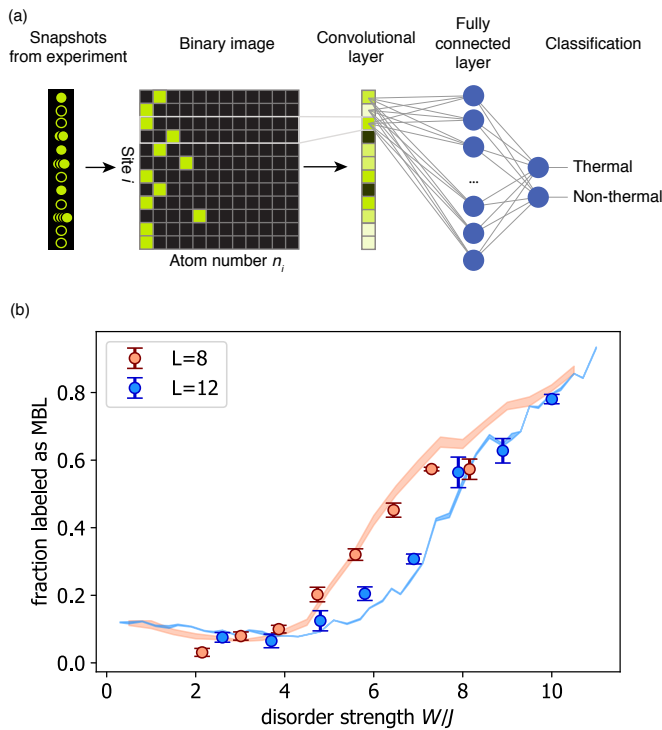


FIG. 1. **Machine learning many-body localization.** The Bose-Hubbard model with a quasi-periodic disorder potential exhibits a many-body localized (MBL) phase, where thermalization breaks down, as the disorder strength is increased beyond a critical value. a) We study the dynamics of the system after a quench for different disorder strengths by evaluating snapshots from a quantum gas microscope with neural networks. b) A neural network is trained to distinguish exact diagonalization snapshots at  $W/J = 0.3$  and  $W/J = 11$  for  $U/J = 2.9$  and a system with 8 and 12 sites at time  $tJ = 100$  after a global quench. After the training process is finished, snapshots at intermediate values of the disorder strength are used as input. The plot shows the resulting classification for numerical data (shaded band) as well as experimental snapshots (symbols). As the system size is increased, the fraction of snapshots classified as MBL begins to increase at larger values of  $W$ , indicating the transition in the finite size system. The accuracies are averaged over two independent runs and the errors denote one s.e.m.

case of equilibrium phase transitions, as it describes a non-equilibrium setting [23–31]. Finding the transition point is numerically challenging, because it is usually obtained from entanglement properties or the level statistics, which can only be obtained for small system sizes where full diagonalization of the Hamiltonian is possible. Here, we focus on Fock space snapshots of the many-body quantum state as input data, which are the direct output of quantum gas microscopy experiments and thus experimentally readily accessible for the systems of interest. This approach has the advantage that significantly bigger system sizes can be reached experimentally. We consider the dynamics of two one-dimensional sys-

tems of 8 and 12 sites, which are initialized in a Mott-insulating state with exactly one particle per site. In Fig. 1, we first train the network to distinguish snapshots of the many-body quantum state, obtained from exact diagonalization calculations, for low ( $W/J = 0.3$ ) and high ( $W/J = 11.0$ ) disorder strength for an interaction strength of  $U/J = 2.9$  in the comparatively long-time limit at time  $tJ = 100$ . We average over ten different disorder realizations, obtained by varying the phase  $\phi$  in the potential. After the network has learned to label the extremal cases correctly with sufficiently high accuracy ( $> 90\%$ ), we input snapshots for intermediate values of the disorder strength. After training the neural network on numerically simulated snapshots, we use experimental data as input, where each snapshot stems from a different disorder realization. As output, for each disorder strength we obtain the fraction of snapshots labeled as *many-body localized* and *thermalizing*, see Fig. 1. Based on these results, we conclude that the many-body localization transition is located within the range of  $W/J \approx 4 - 8$  with strong finite-size drifts. This result is in agreement with previous experiments [32, 33], which considered conventional observables such as the local entropy. Notably, the local entropy exhibits volume law scaling both in the thermal and the MBL phase and is thus by itself not sufficient to locate the transition without exact numerics [32]. Our results, in contrast, are able to distinguish the two phases without any theoretical input, which suggests that the network learned a more suitable observable to distinguish the two phases. In [34], we show the level statistics for system sizes  $L = 6, 7, 8$  for comparison. Similar to the machine learning analysis of a disordered spin chain based on the entanglement spectrum in [19], the transition found by the neural network is as sharp as the level statistics, but exhibits a small shift to larger disorder strengths.

While we have only compared two extremal disorder strengths in the long-time limit, the full dynamics of the system contain much more information. We proceed by analyzing the time- and disorder-strength dependence of the system after the global quench.

*Learning thermalization.*— We now investigate the system’s approach to thermal equilibrium by comparing each time step to a thermal state of the same Hamiltonian. The performance of the network in distinguishing dynamics from equilibrium can then be used as a probe of thermalization.

In order to compare the time evolved state to thermal equilibrium, all conserved quantities of the model should be considered [3]. In our experiment, both the energy density and the particle number are conserved during the many-body evolution. The energy density of the initial state is matched by choosing the temperature of the thermal state accordingly. We take the conservation of the total particle number into account by calculating

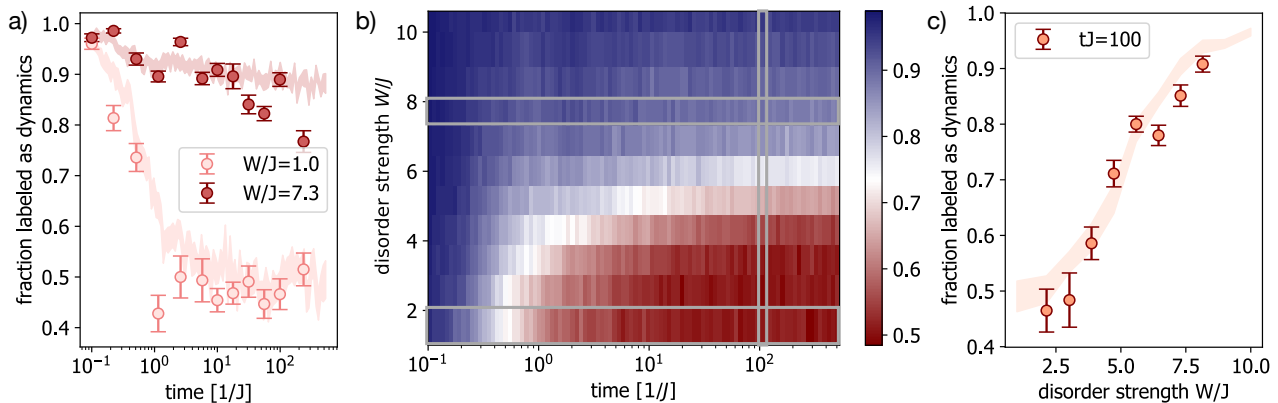


FIG. 2. **Learning thermalization.** A system with 8 sites and  $U/J = 2.9$  is initialized in a Mott-insulating state of one particle per site and the ensuing time evolution is investigated. In each time step, the neural network is trained to distinguish snapshots from the current time step from snapshots from a thermal state with the same energy density, both obtained from exact diagonalization. A high accuracy indicates that the current time step can be easily distinguished from the thermal state. a) The resulting classification as *dynamics* versus *equilibrium* for  $W/J = 1.0$  and  $W/J = 7.3$ , averaged over 12 different disorder realizations (shaded line). Experimental data from the dynamics after the quench is used as input at selected time steps (symbols). b) Exact diagonalization results for disorder strengths between  $W/J = 1$  and  $W/J = 10$  for the full dynamics. c) Classification as *dynamics* versus *equilibrium* at time  $tJ = 100$  for disorder strengths between  $W/J = 1$  and  $W/J = 10$ . The results are averaged over 10 independent runs and the error bars correspond to the s.e.m.

the thermal state within a fixed particle number sector. We numerically generate snapshots from such a state in thermal equilibrium as well as from the time-evolved state for each time step under consideration.

For each time step, we train the network to label the snapshots from the thermal equilibrium distribution as *equilibrium*, and the snapshots from the numerically time-evolved initial state as *dynamics*. The neural network parameters optimized for each time step separately. We then test the network's performance by inputting experimental data with different evolution times. In Fig. 2a) the resulting classification into the categories *dynamics* versus *equilibrium* is shown as a function of time. Here, we average over 12 different disorder realizations and take snapshots at the corresponding effective temperatures.

For small  $W/J$ , the system thermalizes comparably fast: for times  $tJ > 10$ , the network reaches an accuracy of 50%, equivalent to guessing between the two classes. This means the network fails to distinguish snapshots from the time-evolved state from the corresponding thermal state. For high values of  $W/J$  the system fails to thermalize on the time-scales accessed here, and the network is able to distinguish the current timestep from the thermal equilibrium state with a high accuracy. Using an interpretable network architecture [45], we find that for intermediate disorder strengths, higher order correlations play a role in the classification task, see [34]. We study the long time limit at  $tJ = 100$  for a range of values of the disorder strength. As shown in Fig. 2c), the fraction of snapshots classified as *dynamics* rises strongly between  $W/J \approx 4$  and  $W/J \approx 8$  and reaches

values close to 1, indicating that the system has not reached thermal equilibrium.

We benchmark our experimental results by testing the network with theoretical snapshots not used during training and find good agreement throughout the range of the covered parameters.

This procedure has the advantage that the features used to make the classification can vary for different time steps and the network specifically searches for differences between the current time and thermal equilibrium. It is therefore in principle capable of identifying specific observables that have not yet reached their thermal equilibrium value and thus find, for example, (almost-) conserved quantities. Indeed, with this method we find deviations from thermal equilibrium already in the range of  $W/J \approx 2 - 5$ , in contrast to the results from the classification scheme in Fig. 1b). This indicates an improved sensitivity of our method. Here we consider a system which exhibits a transition from thermalizing behavior to many-body localization, which constitutes a canonical example in the study of non-equilibrium phenomena. Note, however, that our scheme is not limited to the system considered here and can be applied to a variety of models. This method also allows to detect, for example, prethermal behavior and the existence of conserved quantities that keep their value during the dynamics and therefore never reach a generic thermal equilibrium value. Another canonical model to study equilibration behavior is the transverse field Ising model, which has an extensive number of conserved quantities. In [34], we show that a neural network performs significantly worse in distinguishing the time-evolved state from an

approximative generalized Gibbs ensemble, where a few conserved quantities are taken into account, than the simple thermal state discussed above, where only the energy density is considered. This highlights the capability of our approach to identify conserved quantities, which can drastically alter the thermalization process. Our method comes at the expense that one needs snapshots from the thermal density matrix for training, which – especially in the case of a non-thermalizing phase such as MBL – may need to be generated numerically. In the following, we overcome this limitation by analyzing the transition in the dynamics with an unsupervised scheme that, in principle, does not rely on theory data.

*Confusion learning.*– Several unsupervised learning schemes that use the network performance to probe whether and where a phase transition or more general, a qualitative change in the data, exists have been proposed [37–39]. Here, we adapt a scheme termed “confusion learning” introduced in Ref. [37]. In brief, the scheme works as follows: We have a dataset of snapshots for values of the disorder strength  $0.3 \leq W/J \leq 11.0$ . The goal is to test whether a value  $W^*$  exists at which the data changes qualitatively. We start with a guess for  $W^*$  and label all snapshots for  $W \leq W^*$  as *phase A* and correspondingly all snapshots with  $W > W^*$  as *phase B*. Assuming the snapshots are qualitatively different for  $W \leq W^*$  as compared to  $W > W^*$ , the network should achieve a high accuracy in assigning the correct labels. However, if there is no qualitative change at the  $W^*$  under consideration, there will be confusion about the correct labels and the accuracy will thus be lower. Therefore, if there is a qualitative change in the data, the accuracy as a function of  $W^*$  will be maximal if  $W^*$  corresponds to the transition point. Trivially, the test accuracy is expected to approach unity when the guessed  $W^*$  corresponds to the minimum or maximum value of  $W$ , because all data are labelled equally and no confusion occurs. In total, the presence of a critical point is therefore signalled by a characteristic *W*-shape of the test accuracy as a function of the control parameter.

We train the neural network with numerical snapshots in the long-time limit ( $tJ = 100$ ) in order to test for the presence of a phase transition. Subsequently, we use experimental data as input to the network, Fig. 3a). The data shows the onset of a maximum around  $W^*/J = 7$ , indicating the presence of a critical point in agreement with Fig. 1b). The contrast in the *W*-shape achieved here is comparable to the signal seen for a spin model in [37], where instead of snapshots the entanglement spectrum is used as input to the neural network. In order to isolate the signal of the phase transition from the trivial part of the *W*-shape, we subtract the accuracy obtained when training on randomly labeled data. The resulting difference, shown in the inset of Fig. 3a), exhibits a clear peak at  $W^*/J = 7$ , that indicates the transition

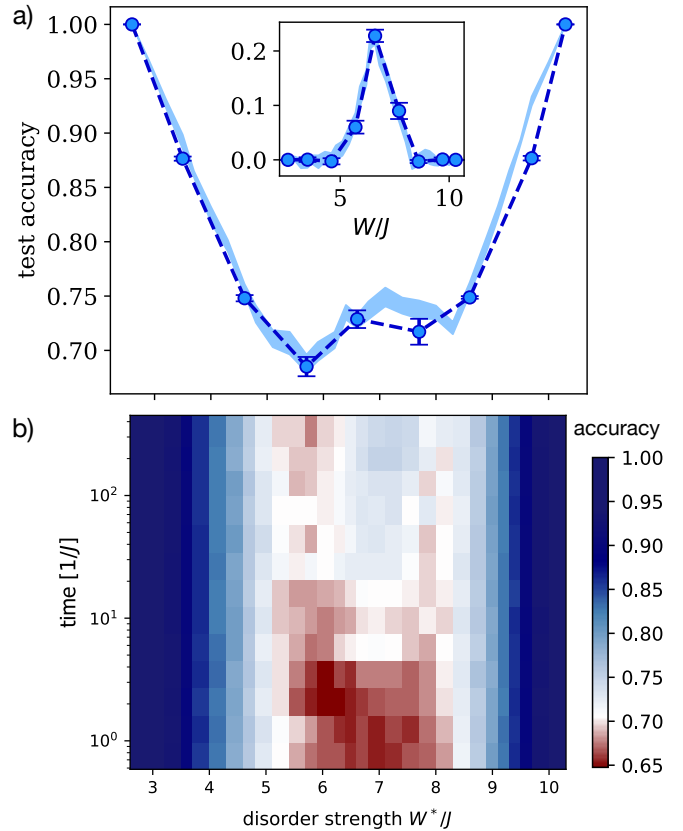


FIG. 3. **Confusion learning.** Snapshots of the many-body quantum state of a system with 12 sites,  $U/J = 2.9$ , and various disorder strengths  $W/J$  are analyzed using the confusion learning scheme. A neural network is trained to label all snapshots with  $W < W^*$  as *phase A* and the remainder as *phase B*. If a qualitative change in the data occurs, the accuracy will peak at an intermediate value of  $W^*$ . a) The resulting accuracy at time  $tJ = 100$  after the global quench for training on numerically simulated data (shaded line) and sorting experimental data (symbols). Inset: same data after subtracting the accuracy for randomly labeled data. b) The accuracy for repeating the training process for different time points during the dynamics after the quench using numerically simulated data. The results are averaged over 10 independent runs and the error bars correspond to the error based on one s.e.m.

between the different dynamical phases. We also check with theoretical snapshots not used during training and find qualitatively similar behaviour. We attribute the slight deviation in the maximum to the coarse resolution in the disorder strength for the experimental data. Since differences in the thermalization behavior only present themselves in the course of the dynamics, we expect the phase transition to remain hidden at short evolution times. In order to reveal this effect, we perform the same method with theoretical snapshots at different evolution times. In Fig. 3b), the resulting accuracy achieved by the network is shown as a function of  $W^*$ . These results have several advantages compared to the

previous methods: as opposed to Fig. 1b), we do not a priori assume that there is a transition. Moreover, we specifically train the network to find differences between the snapshots at all available values of the disorder strength, thus avoiding bias from the choice of training data.

*Summary and Outlook.*— In this work, we used machine learning techniques to study the non-equilibrium dynamics after a global quench in the one-dimensional Bose-Hubbard model with a quasi-periodic disorder potential. We used supervised as well as unsupervised machine learning methods to probe for a qualitative change in experimental snapshots as the disorder strength is tuned. Comparing the results for systems with 8 and 12 sites, we find that the critical value of the disorder strength increases with the system size, proving the need for methods applicable in large – experimentally accessible – systems. In contrast to standard tools to locate the MBL transition, the methods used here can be directly applied to experimental data taken with a quantum gas microscope and are not limited to small system sizes. We furthermore studied the approach to thermal equilibrium – or lack thereof – by training a neural network to distinguish snapshots from the current time step from snapshots from a thermal ensemble at the same energy and particle density. The accuracy achieved by the network indicates how non-thermal the time-dependent quantum many-body state is.

An exciting future research direction consists of applying the same scheme to identify conserved or almost-conserved quantities in experimentally accessible data, for example by using a generalized Gibbs ensemble for comparison. Apart from the concrete system studied here, it would be interesting to consider other models and phenomena, for example quantum scars [40, 41] and Hilbert space fragmentation [42–44]. In order to gain additional physical insights, interpretability is an extremely important direction for future work and it would be interesting to study which observables the network uses to make the classifications considered here [45], and how those observables change during the time evolution of the many-body system.

*Acknowledgements.*— We would like to thank Eugene Demler, Fabian Grusdt, Florian Kotthoff, Cole Miles, and Frank Pollmann for fruitful discussions. We acknowledge support from the Technical University of Munich - Institute for Advanced Study, funded by the German Excellence Initiative and the European Union FP7 under grant agreement 291763, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy–EXC–2111–390814868, TRR80, DFG grant No. KN1254/2-1, No. KN1254/1-2, from the European Research Council (ERC) under the European Unions

Horizon 2020 research and innovation programme (grant agreement No. 851161), from the NSF Graduate Research Fellowship Program (S.K.), from the NSF grant PHY-1734011, NSF grant PHY-1806604I, NSF grant OAC-1934598, the Gordon and Betty Moore Foundations EPiQS Initiative, the Vannevar Bush Award, and the Swiss National Science Foundation (J. L.).



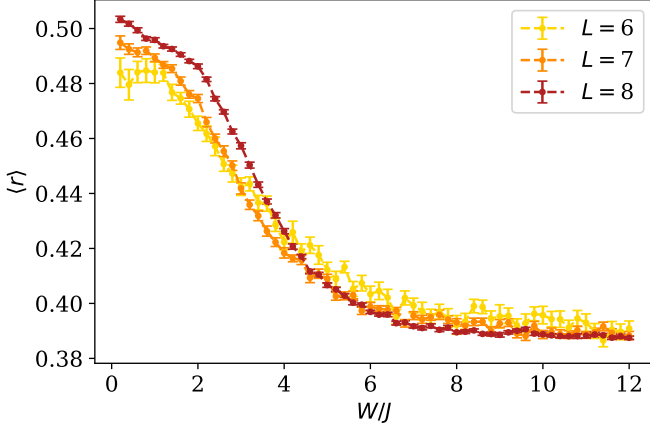


FIG. 4. **Level statistics.** Average value of the ratio of adjacent energy gaps for different system sizes  $L = 6, 7, 8$  at a density of one particle per site as a function of disorder strength.

## SUPPLEMENTARY INFORMATION

### TRANSITION

Upon increasing the disorder strength, the level statistics of the Hamiltonian evolves from the Gaussian-orthogonal ensemble to Poisson statistics as the system enters the MBL phase [25, 26]. We consider the level spacings

$$\delta_\phi^{(n)} = |E_\phi^{(n)} - E_\phi^{(n-1)}|, \quad (2)$$

where  $E_\phi^{(n)}$  is the  $n$ -th eigenenergy of Hamiltonian (1) in the main text with disorder realization given by  $\phi$ . The ratio of adjacent gaps is then given as

$$r_\phi^{(n)} = \min(\delta_\phi^{(n)}, \delta_\phi^{(n+1)}) / \max(\delta_\phi^{(n)}, \delta_\phi^{(n+1)}). \quad (3)$$

For a given system size  $L$ , we fix the particle density to one particle per site and for each disorder strength  $W/J$  obtain the average value of this ratio over 30 disorder realizations, i.e. different values of  $\phi$ . In Fig. 4, the resulting average value of the ratio of adjacent energy gaps  $\langle r \rangle$  is shown as a function of the disorder strength for system sizes  $L = 6, 7, 8$ . The shift of the drop in  $\langle r \rangle$  to larger values of  $W/J$  agrees well with the results presented in Fig.1b) of the main text.

In order to relate to previous work, in particular Ref. [33], we directly evaluate observables from the snapshots and calculate the transport distance  $\Delta x$ , defined as

$$\Delta x = 2 \sum_d |d| \cdot \left\langle G_c^{(2)}(i, i+d) \right\rangle_i \quad (4)$$

with

$$G_c^{(2)}(i, i+d) = \langle \hat{n}_i \hat{n}_{i+d} \rangle - \langle \hat{n}_i \rangle \langle \hat{n}_{i+d} \rangle, \quad (5)$$

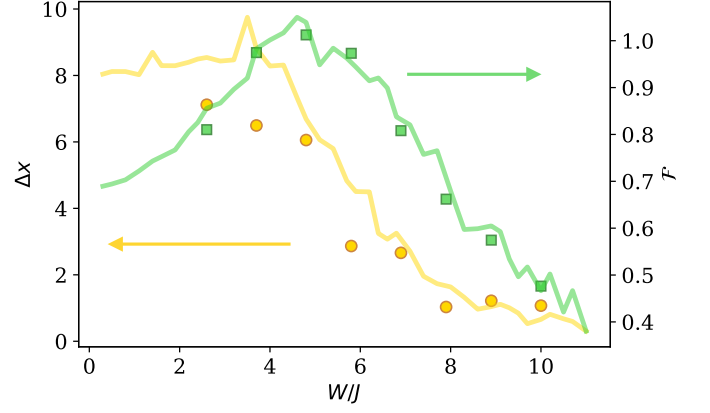


FIG. 5. **Observables.** The transport distance  $\Delta x$  and on-site fluctuations  $\mathcal{F}$ , see text, are evaluated from the same snapshots. Shaded bands correspond to exact diagonalization snapshots, symbols are based on experimental data. We simultaneously evaluate snapshots from ten different disorder realizations in the numerical data. In the experimental data, each snapshot is from a different disorder realization. The errors denote one s.e.m.

and the on-site fluctuations  $\mathcal{F}$ , defined as

$$\mathcal{F} = G_c^{(2)}(d=0), \quad (6)$$

in Fig. 5. Comparing the output of the neural network in Fig.1b) of the main text with the transport distance  $\Delta x$  shows a similar behavior, from which one might conjecture that the network uses a similar observable to make the distinction. Note that with this approach, we are able to make a quantitative prediction on the basis of single or few snapshots, for which the observables shown in Fig. 5 are not clearly converged to their average value.

### Confusion learning: experimental data

In Fig. 6, we show the same analysis as presented in Fig.3 of the main text, but using a network that has been solely trained on experimental data. A clear “W”-shape does not emerge. The experimental result agrees qualitatively with the result using exact numerical data, where we only used data at the same values of disorder strength as available from the experiment. We hence attribute the lack of a clear “W”-shape to the rough grid of values of the disorder strength – consisting of eight values between  $W/J = 2$  and  $W/J = 10$ .

### Unsupervised learning of the transition

We use the unsupervised scheme introduced in [38, 39] to locate the transition to the many-body localized phase as a function of the disorder strength  $W$ . In Fig. 7, we

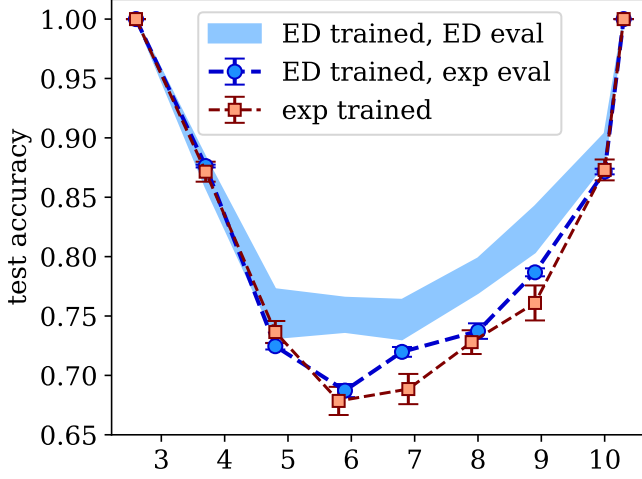


FIG. 6. **Confusion learning - experimental data.** Same analysis as shown in Fig.3 of the main text with training on experimental data (red symbols). The same analysis is performed for numerical data from exact diagonalization, using the same values of the disorder strength.

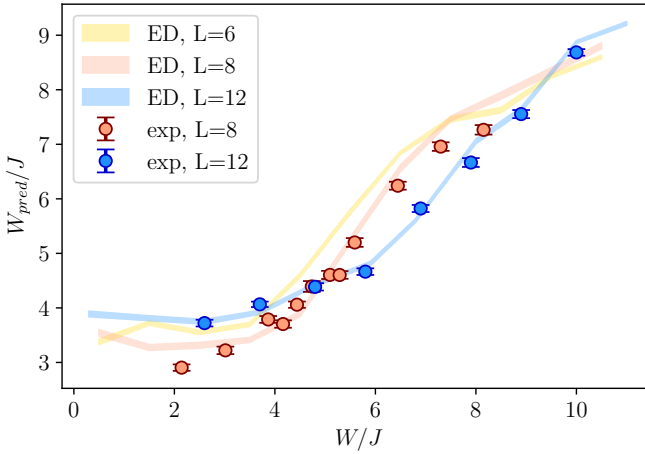


FIG. 7. **Unsupervised learning of a phase transition.** Snapshots of the many-body quantum state of a system with 6, 8, and 12 sites,  $U/J = 2.9$ , and various disorder strengths  $W/J$  are analyzed using the unsupervised learning scheme introduced in [38, 39]. A neural network is trained to label a given snapshot with the corresponding value of  $W/J$ . If a qualitative change in the data occurs, the derivative  $\delta W_{\text{pred}}/\delta W_{\text{label}}$  will exhibit a maximum. The plot shows  $W_{\text{pred}}$  as a function of  $W_{\text{label}}$ . Shaded bands are exact diagonalization and symbols experimental data. The results are averaged over 25 independent runs and the error bars corresponds to one s.e.m. .

show the predicted values of the disorder strength as a function of the actual values,  $W_{\text{label}}$ . The experimental data agrees well with numerics. The steepest slope, indicating the transition, shows a similar shift as observed in the supervised learning scheme used in Fig.1 of the main

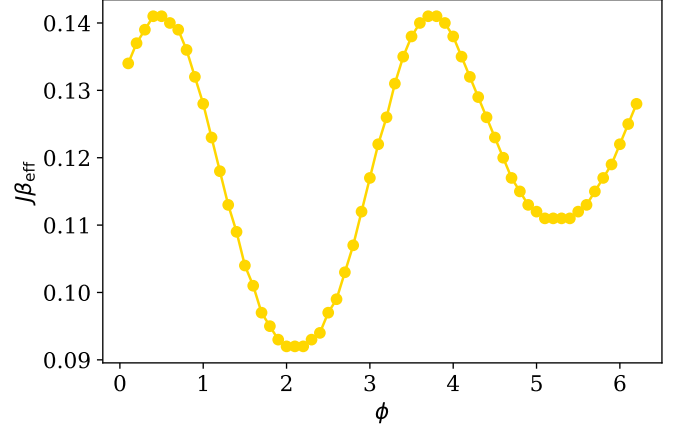


FIG. 8. **Effective temperatures.** Effective inverse temperatures  $\beta_{\text{eff}}$  for  $U/J = 2.9$ ,  $W/J = 4.5$  as a function of  $\phi$  for a system with  $L = 6$  sites and a density of one particle per site.

text.

## LEARNING THERMALIZATION

In order to compare the time evolved state to thermal equilibrium, all conserved quantities of the model should be taken into account. The energy density of the initial state can be matched by choosing the temperature of the thermal state accordingly. In particular, the energy density of the initial state  $|\psi_0\rangle$  is given by  $E_i = \langle \psi_0 | \hat{H} | \psi_0 \rangle$ . The effective temperature  $T_{\text{eff}}$  is then determined such that the density matrix of the system,  $\hat{\rho}_\beta = \frac{1}{Z} \exp(-\beta_{\text{eff}} \hat{H})$ , with the inverse temperature  $\beta_{\text{eff}} = 1/T_{\text{eff}}$  and  $Z = \text{tr}(\exp(-\beta_{\text{eff}} \hat{H}))$  fulfills

$$E_i = \text{tr}(\hat{H} \hat{\rho}_\beta). \quad (7)$$

The energy density  $E(\beta) = \text{tr}(\hat{H} \hat{\rho}_\beta)$  is calculated for a range of values  $\beta$  until the effective temperature is determined such that Eq. 7 is fulfilled. Due to the disorder potential, this effective temperature varies for different values of  $\phi$ , where  $\phi$  determines the disorder realization. In Fig. 8, the effective inverse temperature is shown as a function of  $\phi$  for a system with  $L = 6$  sites at unity filling for interaction strength  $U/J = 2.9$  and disorder strength  $W/J = 4.5$ .

### Interpretability: higher-order correlation functions

Following the same approach as in main text Fig. 2, but using the correlator convolutional neural network (CCNN) architecture, we can gain insights into the information used to solve the classification task. In particular, the order of correlation considered enters as a

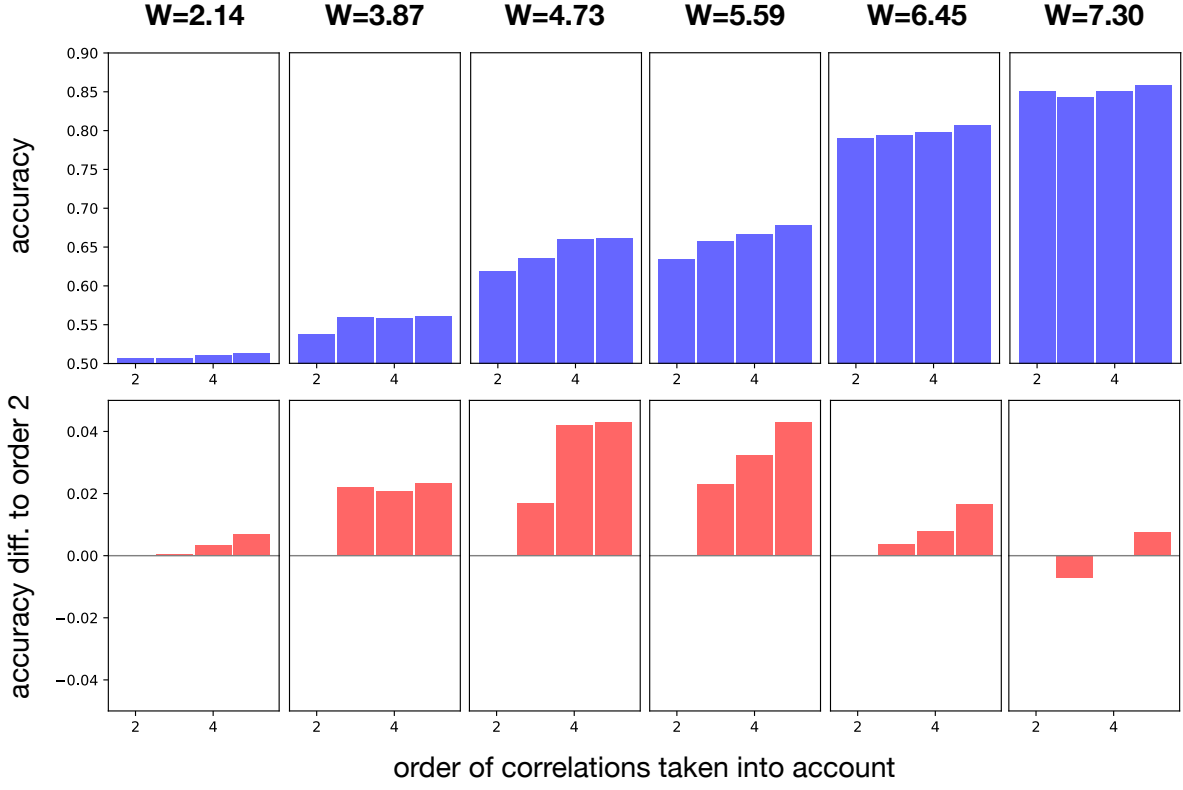


FIG. 9. **Accuracy as a function of order of correlations.** Accuracy obtained for comparing snapshots from time step  $tJ = 7.3$  to a thermal ensemble using the correlator convolutional neural network (CCNN) architecture [45]. The order of correlation functions used by the neural network serves as a hyper parameter in the CCNN architecture. The top plot shows the obtained accuracy for different disorder strengths for correlation orders 2, 3, 4, and 5. The bottom plot shows the difference to the accuracy obtained when taking second order correlation functions into account to allow for a better comparison of the results between the different disorder strengths.

hyper parameter of the network architecture. The network can thus be trained to distinguish dynamics from thermal equilibrium taking into account for example only correlations up to second order. Note that given the Fock space snapshots we consider here, all correlations considered are density-density correlations. In Fig. 9, we compare the accuracies obtained by the network when taking into account correlations up to second, third, fourth, and fifth order for different disorder strength, for comparing the time step  $tJ = 7.3$  to thermal equilibrium. Since the overall scale increases significantly with increasing disorder strength, as the state becomes less and less thermal, we show in Fig. 9 bottom the accuracies obtained for correlations of order 3 – 5 with the accuracy for order 2 subtracted. For low and high disorder, taking into account correlation functions of order higher than two does not significantly increase the accuracy obtained by the CCNN. However, for intermediate disorder strengths, higher order correlations play a role in the classification task and the accuracy improves by up to 5% when considering higher order correlators. This result is in accordance with Ref. [33], which showed sizable higher order

correlations in the critical regime at intermediate disorder strength.

### Transverse field Ising model and generalized Gibbs ensemble

The transverse field Ising model,

$$\hat{H} = -J \sum_i (\hat{\sigma}_i^z \hat{\sigma}_{i+1}^z + h \hat{\sigma}_i^x), \quad (8)$$

has an extensive number of local conservation laws, which are known to be [35, 36]

$$I^{(n,+)} = -J \sum_j (S_{j,j+n}^{xx} + S_{j,j+n+2}^{yy}) + h(S_{j,j+n-1}^{xx} + S_{j,j+n-1}^{yy}) \quad (9)$$

and

$$I^{(n,-)} = -J \sum_j (S_{j,j+n}^{xy} - S_{j,j+n}^{yx}) \quad (10)$$



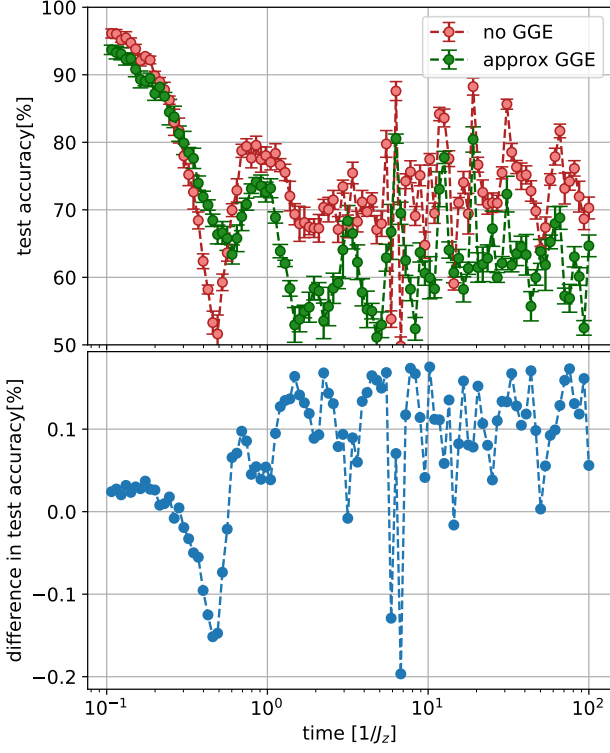


FIG. 10. **Thermalization and conserved quantities.** We consider the dynamics after initializing the system in a product state in the transverse field Ising model (TFIM). In the TFIM, there is an extensive number of conserved quantities, which have to be considered to correctly describe the long-time limit. Top: If only the conserved energy is taken into account (red), the long-time limit is still comparably easy to distinguish from the thermal density matrix. The performance of the network drops significantly if the first (i.e., most local) two conserved quantities are additionally taken into account in an approximation of the generalized Gibbs ensemble (green). Bottom: Difference between the accuracies shown in the top plot.

with  $I^{(1,+)} = H$  and

$$S_{j,j+l}^{\alpha\beta} = \sigma_j^\alpha \left[ \prod_{k=1}^{l-1} \sigma_{j+k}^z \right] \sigma_{j+l}^\beta, \quad S_{j,j}^{yy} = -\sigma_j^z. \quad (11)$$

We numerically simulate the dynamics for  $J = 1$ ,  $h = 2$  starting from the initial product state

$$|\psi_0\rangle = \prod_i (\sin \theta/2 |\uparrow\rangle_i \cos \theta/2 |\downarrow\rangle_i) \quad (12)$$

with  $\theta = 3.0$  using exact diagonalization for a system of size  $L = 12$ . We then train a neural network to distinguish snapshots sampled from the time-evolved state

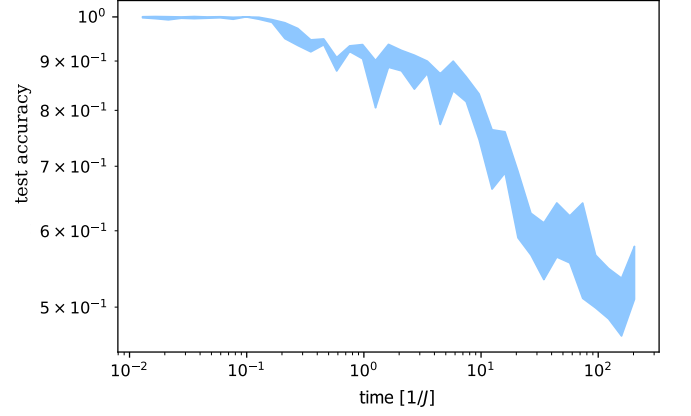


FIG. 11. **Learning thermalization.** The system is initialized in a uniform state ( $|111111111111\rangle$ ) and the ensuing time evolution is investigated. In each time step, the neural network is trained to distinguish snapshots from the current time step from snapshots from the long-time limit. The plots show the resulting accuracy for  $W/J = 6.4$ ,  $U/J = 2.9$ .

from a thermal ensemble. For the thermal ensemble, we consider two different choices:

- a thermal ensemble  $\rho = 1/Z \exp(-\beta H)$ , where the inverse temperature  $\beta$  is determined by the energy  $\langle \psi_0 | H | \psi_0 \rangle$  as discussed above
- an approximation to the generalized Gibbs ensemble,  $\rho_{\text{GGE}} = 1/Z_{\text{GGE}} \exp(-\beta H - \lambda_2 I^{2,+} - \lambda_3 I^{3,+})$  with Lagrange multipliers  $\lambda_{2,3}$  determined in the same way for the conserved quantities defined above, with  $Z_{\text{GGE}}$  the corresponding partition sum for the approximation of the GGE, such that  $\text{tr} \rho_{\text{GGE}} = 1$ . We determine the parameters to match the expectation values in the initial state, yielding  $\beta = 0.304$ ,  $\lambda_2 = 0.336$ , and  $\lambda_3 = 0.1$ .

As shown in Fig. 10, the performance of the neural network in distinguishing dynamics from thermal/generalized Gibbs ensemble is significantly worse in the latter case, indicating that the long-time dynamics is better described by the GGE. This emphasizes the importance in taking conserved quantities into account in order to correctly describe the equilibration behavior, and in particular the ability of a neural network to capture the differences between the time evolved state and the standard thermal ensemble.

### Distinguish from long-time limit

In order to study the dynamics of the quantum many-body system, we here compare snapshots from the current time step to the long-time limit. In a thermalizing system, the long-time limit corresponds to a thermal

equilibrium state and the scheme is thus basically the same as the thermalization learning scheme introduced in the main text. This is, however, not the case for the MBL phase. In Fig. 11, the accuracy achieved on a test set not used during training is shown as a function of time. In each time step, the neural network parameters are optimized to enable the classification of snapshots into the categories *current timestep* versus *long-time limit*. This procedure has the advantage that the features used to make the classification can vary for different time steps and in particular, the network specifically searches for differences between the current time and the long-time limit. It is therefore in principle capable of identifying specific observables that have not yet reached their long-time value.

In Fig. 12, the accuracy achieved on a test set not used during training is shown as a function of time when starting from the product state

$$|\psi_0\rangle = |20202020\rangle \quad (13)$$

for  $W/J = 0, 4.8, 8.9$ . We compare the resulting accuracy to the imbalance, defined as

$$\mathcal{I} = \frac{1}{L \cdot N_s} \sum_s \sum_i (n_i^s - n_i^{\text{ref}}), \quad (14)$$

where  $N_s$  is the number of snapshots, the first sum runs over all snapshots,  $n_i^s$  is the occupation of site  $i$  in snapshot  $s$ , and  $n_i^{\text{ref}} = 0(2)$  for  $i$  even (odd).

In all three cases, the first tunneling events cause a sharp decay in the imbalance on a time scale of one hopping time. The accuracy with which the neural network can distinguish the current time step from the long-time limit is always larger than the difference of the imbalance to its long-time limit.

In the many-body localized case, Fig. 12c), the accuracy shows qualitatively the same behavior as the imbalance. However, in the critical phase, Fig. 12b), there is no fast initial decay in the accuracy and instead, it is still higher than 50%, which corresponds to its lower limit, after several hundred hopping times. Without disorder, Fig. 12a), the imbalance has almost reached its long-time value after one hopping time. The accuracy with which the network can distinguish the current time step from the long-time limit decays on a slower time-scale of about ten hopping times.

- 
- [1] J. M. Deutsch. Quantum statistical mechanics in a closed system. *Phys. Rev. A*, 43:2046–2049, Feb 1991.  
[2] Mark Srednicki. Chaos and quantum thermalization. *Physical Review E*, 50(2):888–901, Aug 1994.  
[3] Marcos Rigol, Vanja Dunjko, and Maxim Olshanii. Thermalization and its mechanism for generic isolated quantum systems. *Nature (London)*, 452(7189):854–858, apr 2008.

- [4] Jonathan Lux, Jan Müller, Aditi Mitra, and Achim Rosch. Hydrodynamic long-time tails after a quantum quench. *Phys. Rev. A*, 89:053608, May 2014.  
[5] Subroto Mukerjee, Vadim Oganesyan, and David Huse. Statistical theory of transport by strongly interacting lattice fermions. *Phys. Rev. B*, 73:035113, Jan 2006.  
[6] A Bohrdt, C B Mendl, M Endres, and M Knap. Scrambling and thermalization in a diffusive quantum many-body system. *New Journal of Physics*, 19(6):063001, jun 2017.  
[7] J. Berges, Sz. Borsányi, and C. Wetterich. Prethermalization. *Phys. Rev. Lett.*, 93:142002, Sep 2004.  
[8] M. Aidelsburger, M. Atala, M. Lohse, J. T. Barreiro, B. Paredes, and I. Bloch. Realization of the hofstadter hamiltonian with ultracold atoms in optical lattices. *Physical Review Letters*, 111:185301, 2013.  
[9] Hirokazu Miyake, Georgios A. Siviloglou, Colin J. Kennedy, William Cody Burton, and Wolfgang Ketterle. Realizing the harper hamiltonian with laser-assisted tunneling in optical lattices. *Phys. Rev. Lett.*, 111:185302, Oct 2013.  
[10] Gregor Jotzu, Michael Messer, Rémi Desbuquois, Martin Lebrat, Thomas Uehlinger, Daniel Greif, and Tilman Esslinger. Experimental realization of the topological haldane model with ultracold fermions. *Nature*, 515(7526):237–240, Nov 2014.  
[11] M. Schreiber, S. S. Hodgman, P. Bordia, H. P. Luschen, 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(6250):842–845, Jul 2015.  
[12] M. Gring, M. Kuhnert, T. Langen, T. Kitagawa, B. Rauer, M. Schreitl, I. Mazets, D. A. Smith, E. Demler, and J. Schmiedmayer. Relaxation and prethermalization in an isolated quantum system. *Science*, 337(6100):1318–1322, Sep 2012.  
[13] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, Feb 2018.  
[14] Giuseppe Carleo, Kenny Choo, Damian Hofmann, James E.T. Smith, Tom Westerhout, Fabien Alet, Emily J. Davis, Stavros Efthymiou, Ivan Glasser, Sheng-Hsuan Lin, and et al. Netket: A machine learning toolkit for many-body quantum systems. *SoftwareX*, 10:100311, Jul 2019.  
[15] Benno S. Rem, Niklas Käming, Matthias Tarnowski, Luca Asteria, Nick Fläschner, Christoph Becker, Klaus Sengstock, and Christof Weitenberg. Identifying quantum phase transitions using artificial neural networks on experimental data. *Nature Physics*, 15(9):917–920, Jul 2019.  
[16] Yi Zhang, A. Mesaros, K. Fujita, S. D. Edkins, M. H. Hamidian, K. Ch’ng, H. Eisaki, S. Uchida, J. C. Séamus Davis, Ehsan Khatami, and et al. Machine learning in electronic-quantum-matter imaging experiments. *Nature*, 570(7762):484–490, Jun 2019.  
[17] Annabelle Bohrdt, Christie S. Chiu, Geoffrey Ji, Muqing Xu, Daniel Greif, Markus Greiner, Eugene Demler, Fabian Grusdt, and Michael Knap. Classifying snapshots of the doped hubbard model with machine learning. *Nature Physics*, 15(9):921–924, 2019.  
[18] Alireza Seif, Mohammad Hafezi, and Christopher Jarzynski. Machine learning the thermodynamic arrow of time.

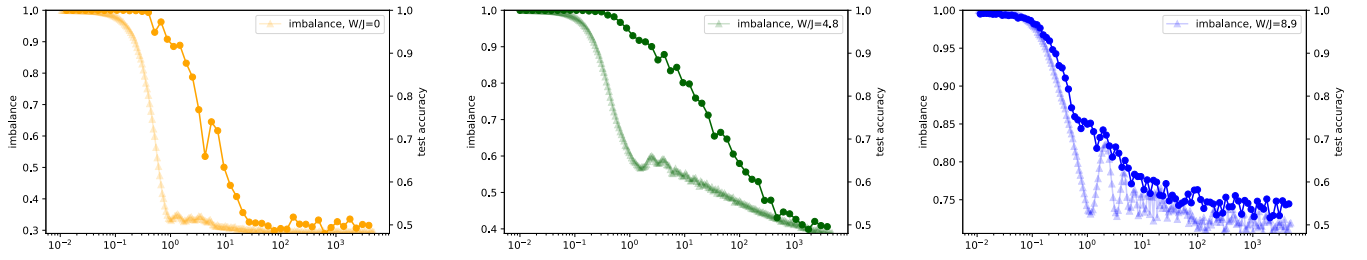


FIG. 12. **Learning thermalization – 2.** The system is initialized in a charge density wave state and the ensuing time evolution is investigated. In each time step, the neural network is trained to distinguish snapshots from the current time step from snapshots from the long-time limit. The plots show the resulting accuracy for a)  $W/J = 0$ , b)  $W/J = 4.8$  and c)  $W/J = 8.9$ . A high accuracy indicates that the current time step can be easily distinguished from the long-time limit. The imbalance is shown for comparison.

- Nature Physics*, 2020.
- [19] Frank Schindler, Nicolas Regnault, and Titus Neupert. Probing many-body localization with neural networks. *Phys. Rev. B*, 95:245134, Jun 2017.
  - [20] Jordan Venderley, Vedika Khemani, and Eun-Ah Kim. Machine learning out-of-equilibrium phases of matter. *Phys. Rev. Lett.*, 120:257204, Jun 2018.
  - [21] Yi-Ting Hsu, Xiao Li, Dong-Ling Deng, and S. Das Sarma. Machine learning many-body localization: Search for the elusive nonergodic metal. *Phys. Rev. Lett.*, 121:245701, Dec 2018.
  - [22] Wei Zhang, Lei Wang, and Ziqiang Wang. Interpretable machine learning study of the many-body localization transition in disordered quantum ising spin chains. *Physical Review B*, 99(5), Feb 2019.
  - [23] D.M. Basko, I.L. Aleiner, and B.L. Altshuler. Metal-insulator transition in a weakly interacting many-electron system with localized single-particle states. *Annals of Physics*, 321:1126–1205, 2006.
  - [24] I. Gornyi, A. Mirlin, and D. Polyakov. Interacting electrons in disordered wires: Anderson localization and low- $T$  transport. *Phys. Rev. Lett.*, 95:206603, Nov 2005.
  - [25] Vadim Oganesyan and David A. Huse. Localization of interacting fermions at high temperature. *Phys. Rev. B*, 75:155111, Apr 2007.
  - [26] Arijeet Pal and David A. Huse. Many-body localization phase transition. *Phys. Rev. B*, 82:174411, Nov 2010.
  - [27] Maksym Serbyn, Z. Papić, and Dmitry A. Abanin. Local conservation laws and the structure of the many-body localized states. *Phys. Rev. Lett.*, 111:127201, Sep 2013.
  - [28] David A. Huse, Rahul Nandkishore, and Vadim Oganesyan. Phenomenology of fully many-body-localized systems. *Phys. Rev. B*, 90(17):174202, 2014.
  - [29] M. Serbyn, M. Knap, S. Gopalakrishnan, Z. Papić, N. Y. Yao, C. R. Laumann, D. A. Abanin, M. D. Lukin, and E. A. Demler. Interferometric probes of many-body localization. *Phys. Rev. Lett.*, 113:147204, Oct 2014.
  - [30] Dmitry A. Abanin, Ehud Altman, Immanuel Bloch, and Maksym Serbyn. Colloquium: Many-body localization, thermalization, and entanglement. *Reviews of Modern Physics*, 91(2), May 2019.
  - [31] B. Chiaro, C. Neill, A. Bohrdt, M. Filippone, F. Arute, K. Arya, R. Babbush, D. Bacon, J. Bardin, R. Barends, S. Boixo, D. Buell, B. Burkett, Y. Chen, Z. Chen, R. Collins, A. Dunsworth, E. Farhi, A. Fowler, B. Foxen, C. Gidney, M. Giustina, M. Harrigan, T. Huang, S. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, E. Lucero, J. McClean, X. Mi, A. Megrant, M. Mohseni, J. Mutus, M. McEwen, O. Naaman, M. Neeley, M. Niu, A. Petukhov, C. Quintana, N. Rubin, D. Sank, K. Satzinger, A. Vainsencher, T. White, Z. Yao, P. Yeh, A. Zalcman, V. Smelyanskiy, H. Neven, S. Gopalakrishnan, D. Abanin, M. Knap, J. Martinis, and P. Roushan. Direct measurement of non-local interactions in the many-body localized phase, 2020.
  - [32] Alexander Lukin, Matthew Rispoli, Robert Schittko, M. Eric Tai, Adam M. Kaufman, Soonwon Choi, Vedika Khemani, Julian Léonard, and Markus Greiner. Probing entanglement in a many-body-localized system. *Science*, 364(6437):256–260, 2019.
  - [33] Matthew Rispoli, Alexander Lukin, Robert Schittko, Sooshin Kim, M. Eric Tai, Julian Léonard, and Markus Greiner. Quantum critical behaviour at the many-body localization transition. *Nature*, 573(7774):385–389, 2019.
  - [34] See supplementary online material for more details, in particular results on interpretability and the transverse field Ising model, taking into account conserved quantities of the model, including Refs. [35, 36].
  - [35] Michael Grady. Infinite set of conserved charges in the ising model. *Phys. Rev. D*, 25:1103–1113, Feb 1982.
  - [36] Tomaz Prosen. A new class of completely integrable quantum spin chains. *Journal of Physics A: Mathematical and General*, 31(21):L397–L403, may 1998.
  - [37] Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Physics*, 13:435–439, 02 2017.
  - [38] Frank Schäfer and Niels Lörch. Vector field divergence of predictive model output as indication of phase transitions. *Phys. Rev. E*, 99:062107, Jun 2019.
  - [39] Eliska Greplova, Agnes Valenti, Gregor Boschung, Frank Schäfer, Niels Lörch, and Sebastian D Huber. Unsupervised identification of topological phase transitions using predictive models. *New Journal of Physics*, 22(4):045003, Apr 2020.
  - [40] Hannes Bernien, Sylvain Schwartz, Alexander Keesling, Harry Levine, Ahmed Omran, Hannes Pichler, Soonwon Choi, Alexander S. Zibrov, Manuel Endres, Markus Greiner, and et al. Probing many-body dynamics on a 51-atom quantum simulator. *Nature*, 551(7682):579–584, Nov 2017.

- [41] C. J. Turner, A. A. Michailidis, D. A. Abanin, M. Serbyn, and Z. Papić. Weak ergodicity breaking from quantum many-body scars. *Nature Physics*, 14(7):745–749, 2018.
- [42] Pablo Sala, Tibor Rakovszky, Ruben Verresen, Michael Knap, and Frank Pollmann. Ergodicity breaking arising from hilbert space fragmentation in dipole-conserving hamiltonians. *Phys. Rev. X*, 10:011047, Feb 2020.
- [43] Vedika Khemani, Michael Hermele, and Rahul Nandkishore. Localization from hilbert space shattering: From theory to physical realizations. *Phys. Rev. B*, 101:174204, May 2020.
- [44] Sebastian Scherg, Thomas Kohlert, Pablo Sala, Frank Pollmann, Bharath H. M., Immanuel Bloch, and Monika Aidelsburger. Observing non-ergodicity due to kinetic constraints in tilted fermi-hubbard chains, 2020.
- [45] Cole Miles, Annabelle Bohrdt, Ruihan Wu, Christie Chiu, Muqing Xu, Geoffrey Ji, Markus Greiner, Kilian Q. Weinberger, Eugene Demler, and Eun-Ah Kim. Correlator convolutional neural networks: An interpretable architecture for image-like quantum matter data. *arXiv:2011.03474*, 2020.