Supplementary Material
Learning phase transitions from dynamics

Evert van Nieuwenburg,[1] Eyal Bairey,[2] and Gil Refael[1]

[1] *Institute for Quantum Information and Matter, Caltech, Pasadena, California 91125, USA*
[2] *Physics Department, Technion, 3200003, Haifa, Israel*

TRAINING LSTMS ON A SINGLE SPIN

To gain a better understanding of what the LSTM neurons are extracting from the magnetization input, we focus here on the case of a single spin. In this section, we consider the magnetization $m_i(t)$ of one spin close to the center of our 20 site chain, i.e. $i = 10$. We restrict ourselves to initial states for which the energy density $\varepsilon = 0.5$.

We start by training a single LSTM unit on this single spin data using the blanking setup described in the main text. After training, we analyzed the hidden state $h(t)$ and the LSTM output $y(t)$ on various magnetization curves. It is particularly instructive however to evaluate the neuron's behaviour on a handcrafted magnetization signal. Even though re-training this neuron (starting with random initial internal weights) results in different quantitative behaviour, it seems to always show qualitatively similar response. That is, the neuron output effectively functions as an integrator depending on the magnetization, as exemplified in Fig. 1. Intuitively, if the magnetization sticks to zero, the neuron slowly becomes more certain of ETH behaviour. However, having a magnetization that is constantly close to ± 1, builds confidence for the MBL classification. Based on these observations, the behaviour shown in Fig. 1 is for a handcrafted LSTM neuron designed to implement this integration behavior.

Training multiple LSTM units on a single spin is also beneficial for understanding, and shows that different LSTM units all essentially behave as similar integrators, but they integrate different parts of the signal. Combining the results of all these neurons together leads to a more robust prediction for the transition point w.r.t. W, as judged from the scaling analysis of the number of LSTM units (see main text).

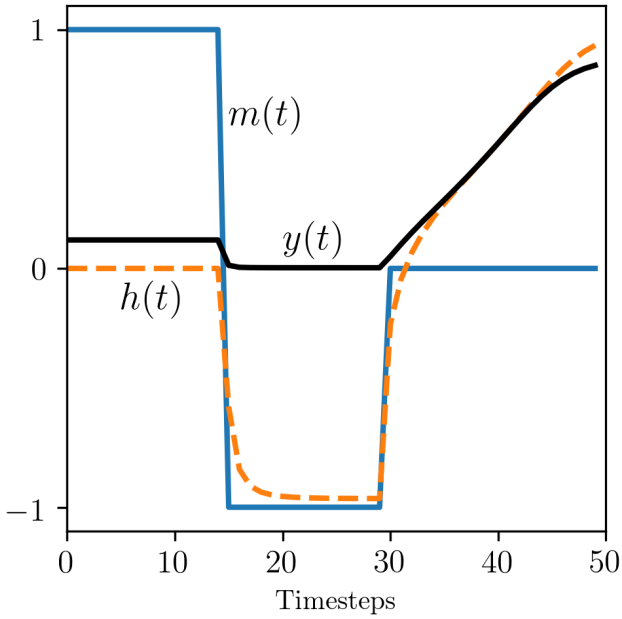

FIG. 1: A single handcrafted LSTM neuron, modeled after the observed behaviour of a trained LSTM neuron. To clearly demonstrate the various regimes of the neuron, we evaluate its hidden state $h(t)$ and the resulting prediction $y(t)$ (where $y(t) \approx 0$ corresponds to MBL and $y(t) \approx 1$ corresponds to thermalizing behaviour) for a specific magnetization signal $m(t)$. Whenever the magnetization is close to 0, the neuron builds confidence that the signal classifies as a thermalizing signal, whereas a magnetization that sticks around extreme values signals localization.

ROBUSTNESS OF BLANKING

In this section, we discuss several approaches to assess the validity of the network's predictions in the untrained regions of the phase space.

A. Blanking region dependence

First, we examine the sensitivity of the predicted phase transition to the size of the region in phase space where the network is trained on. On the one hand, one would like the network to rely on as little human input as possible; on the other hand, a network trained on very narrow regions in phase space might pick up features that are specific to that region rather than universal to the phase that region belongs to. We therefore repeat our training process for different blanked-out regions of the form $[W_B, 8 - W_B]$. For each blanked out region we train 10 networks and extract the W_c (from the peak in the confusion C) predicted by each. Figure 2 shows the mean W_c as well as the standard deviation as a function of the length of the training region W_B. We observe that the predictions for W_c converge to a narrow range when $W_B \gtrsim 0.75$ and is weakly dependent on W_B for larger values.

We wish to remark here finally, that both the l_2 reg-

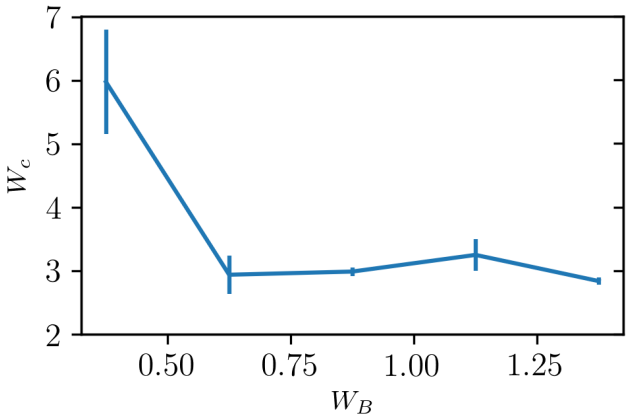

FIG. 2: The predicted W_c after 25 epochs of training, averaged over 10 re-trained networks with the same parameters as in the main text, as extracted from the peaks in the confusion C (see main text). The x-axis shows W_B representing the regions of the phase space, namely $[W_B, 8 - W_B]$, that were used to train the network. Importantly, from $W_B \gtrsim 0.75$ the predicted W_c shows little variation and reasonably small variance, indicating that the extracted phase boundary is insensitive to the amount of data included (symmetrically) from both ends of the phase space.

ularization and the confidence enhancement affect the shape of the learned MBL phase diagram. In their absence, we observe a weaker dependence of the critical disorder strength on the energy density.

Calibration

We also examine the network's calibration, namely how well its confidence corresponds to its accuracy [2]. We train a network on the random-field Heisenberg model at energy density $\epsilon = 0.5$ in the regions $W \in [0, 0.5]$ and $W \in [7.5, 8]$ (as in the main text), and then analyze its predictions on a calibration set containing the test set (different disorder realizations taken from the training region) as well as the untrained regions $W \in [0.5, 1]$ and $W \in [7, 7.5]$. To collect a smooth histogram, we repeat this process over 100 networks. Fig. 3 shows the probability that an instance belongs to the ergodic phase as a function of the output of its corresponding network (top panel), as well as a histogram of the outputs of the networks (bottom panel). We observe that in these regions of phase space the networks are mostly underconfident, assigning a lower confidence value than their actual accuracy. This is especially true for the MBL phase, where the networks are on average less confident than for the ergodic phase.

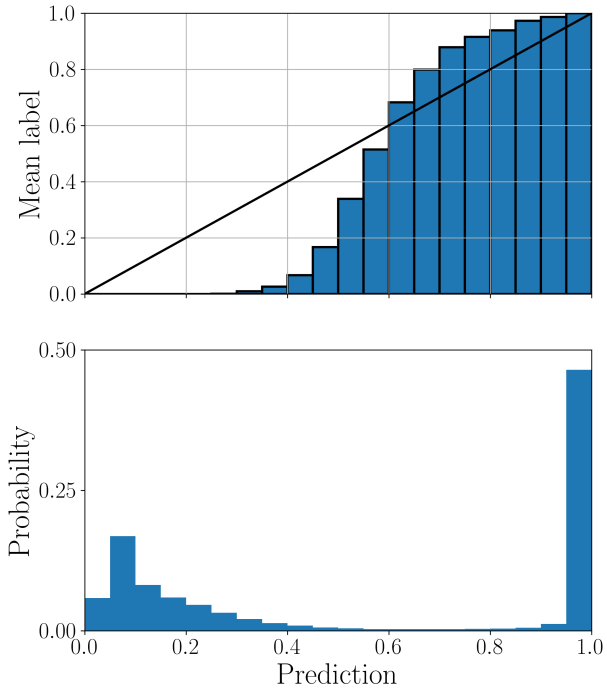

FIG. 3: A closer look at network performance in a region where the correct phase is known. We train 100 networks on the regions $W \in [0, 0.5]$ and $W \in [7.5, 8]$ in the random-field Heisenberg model at energy density $\epsilon = 0.5$, and evaluate their performance on additional samples from these parameter ranges as well as the $W \in [0, 0.5]$ and $W \in [7.5, 8]$. Top panel shows the mean label (0 corresponds to MBL, 1 to ergodic) as a function of network prediction, averaged over the all the predictions of the different networks in equal intervals of length 0.05. Bottom panel shows the probability for each prediction to occur in each of these intervals. We observe a higher mean confidence for ergodic samples in this regime compared to the MBL samples.

Unseen phases

Finally, it is interesting to examine how the network reacts to a phase it had not encountered during training. Focusing now on our time-dependent model, we train a network for the binary classification task of distinguishing between the Floquet-MBL and the time-crystal phases, without exposing it to any data from the ergodic phase during training. Rather than confidently misclassifying this unseen phase, the network assigns a highly confused output to most of the ergodic phase (Fig. 4). This suggests that the network is robust to adversarial examples; moreover, it suggests that our method can be used not only for detecting phase transitions between

known phases, but also in a semi-unsupervised matter for finding new phases.

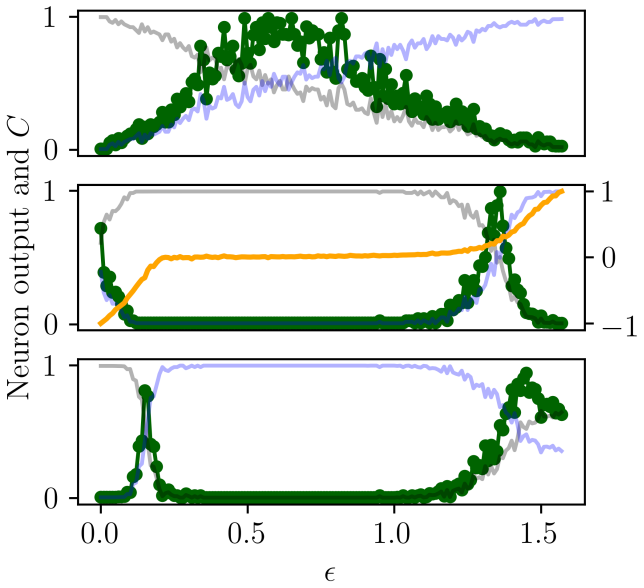

FIG. 4: Networks assign a confused output to phases they had not encountered during training. We train networks to distinguish between two of the three phases of the model featured in Fig. 3 in the main text. In the top panel, a network is trained on the two extremes of the parameter range (time-crystal at $\epsilon = 0$ and Floquet-MBL at $\epsilon = \pi$). When evaluated, the network does not assign a confident prediction to any of these two phases throughout most of the parameter range, indicating the existance of a third, unseen phase - the ergodic phase. Similarly, in the middle panel a network is trained only on the ergodic and Floquet-MBL phases, and in the bottom panel on the time-crystal and ergodic phases; each network is baffled by the phase it was not trained on. The middle panel also shows the averaged final imbalance throughout the parameter range (yellow line) as a rough indicator of the correct phase.

COMPARISON WITH IMBALANCE

To illustrate the ambiguity in using the imbalance $\mathcal{I}$ as a transition indicator, consider the MBL system discussed in the main text. In the left panel of Fig. 5 we show the traces of a single spin in the middle of the sample (for a single disorder realization), at energy density $\varepsilon = 0.5$ and at two different disorder strengths. For weak disorder, the spin quickly relaxes whereas for strong disorder the spin maintains a value close to its initial one for the entire simulated time t_{final}. Superimposed on top of these example curves is the disorder averaged imbalance $\langle \mathcal{I}(t) \rangle$.

Clearly, the disorder averaged imbalance at the final simulation time $\langle \mathcal{I}(t_{\text{final}}) \rangle$ easily distinguishes between the extreme regimes of weak and strong disorder. We know that in those regimes the model is ergodic and non-ergodic, respectively, so we might naively use this quantity to compute a phase diagram as in the rightmost panel of Fig. 5. Although suggestive, a criterion for the phase boundary from $\mathcal{I}(t)$ (for $t = 500$ in this case) is non-trivial. If we had access to infinite times, the remaining imbalance would be a clear criterion; but for finite time data as in an experiment, one would need to extrapolate. What is more, is that a finite size scaling attempt on the imbalance does not show a crossing point [1]. Rather, here we suggest a method that, given the physics of the extreme regimes, consistently finds such a threshold between the strong and weak disorder regimes from the data only.

TRAINING PROCEDURE

For the detection of the MBL transition in the main text, the training was performed as mentioned for $W \in 0.125, 0.25, 0.375, 0.5, 7.625, 7.5, 7.875, 8$. At $\epsilon = 0.5$, we generated 100 realizations for each disorder strength, 60 for training and 40 for test. Each realization randomized both disorder and initial spin configuration. The training and test sets therefore consisted of 480 and 320 samples respectively, though effectively the 5600 samples at $0.625 \leq W \leq 7.5$ are utilized during training due to the confidence enhancement [3]. Note that within the blanking framework we can only include samples from the labeled regions in the test set, since we cannot assess the network's accuracy in regions where the correct label is unknown. The assessment of overfitting is therefore limited to the training region, and needs to be interpreted slightly differently from the usual case [4]. Namely, the generalization we wish to test is that of whether the network managed to extract the right physical model from the training regions only. A perfect fit of the data is hence not necessarily a bad thing, if indeed one can show that the data in these regions is in principle enough to extract the correct model. Such statements might be testable using adversarial perturbations, but we have not investigated this direction. Regardless, within the training regions we observed that the training-set and test-set accuracies increased with the number of LSTM units in the range we examined. For 2, 4, 16 and 32 units respectively, averaged over 10 retrainings, we obtained final training accuracies 0.94, 0.99, 1, 1 and corresponding test-set accuracies 0.89, 0.97, 0.99, 0.99.

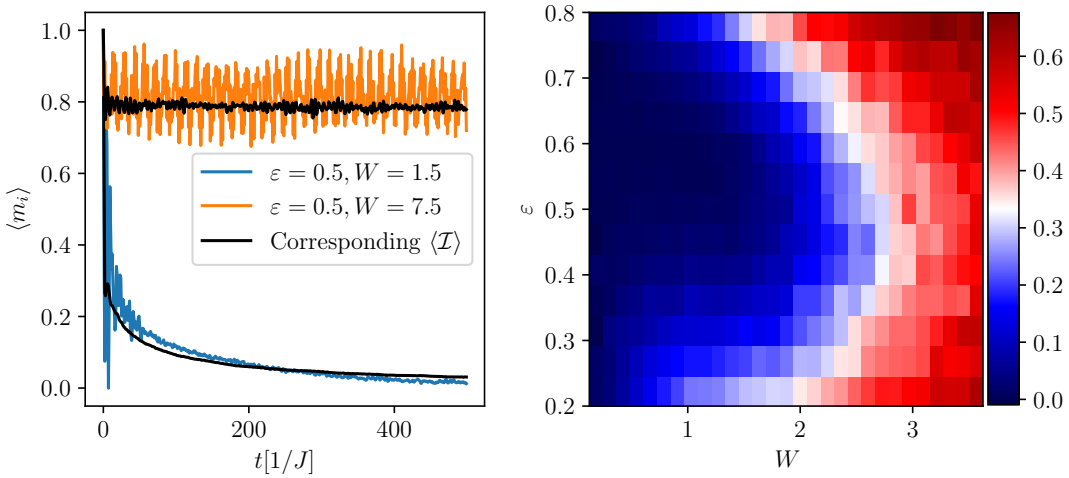

FIG. 5: The left panel shows time traces of the magnetization of a central spin ($i = 8$) in a 20-site spin chain, for a given disorder realization and parameters indicated. For weak disorder ($W = 1.5$) the spin quickly relaxes, whereas for strong disorder it mostly retains its initial value. The disorder averaged imbalance $\mathcal{I}(t)$ for the corresponding parameter regimes is shown superimposed, indicating that these behaviours are typical for the spins and given regimes. The right panel shows the imbalance at the final time $t_{\text{final}} = 500$ as a function of energy density and disorder strength. The resulting phase diagram is suggestive, but putting a phase boundary via a threshold on $\mathcal{I}$ is ambiguous.

[1] S. Iyer, V. Oganesyan, G. Refael, and D. A. Huse, Phys. Rev. B **87**, 134202 (2013)
[2] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, ArXiv e-prints (2017), arXiv:1706.04599
[3] F. Schindler, N. Regnault, and T. Neupert, Physical Review B **95**, 245134 (2017)
[4] C. Zhang, S. Bengio, M. Hardt B. Recht, and O. Vinyals, ArXiv e-prints (2016), arXiv:1611.03530