Learning about learning by many-body systems

Weishun Zhong [*1] **Jacob M. Gold** [*1 2] **Sarah Marzen** [1 3] **Jeremy L. England** [1 4] **Nicole Yunger Halpern** [5 6 7]

Abstract

Many-body systems from soap bubbles to suspensions to polymers learn the drives that push them far from equilibrium. This learning has been detected with thermodynamic properties, such as work absorption and strain. We progress beyond these macroscopic properties that were first defined for equilibrium contexts: We quantify statistical mechanical learning with representation learning, a machine-learning model in which information squeezes through a bottleneck. We identify a structural parallel between representation learning and far-from-equilibrium statistical mechanics. As an application of this parallel we measure many-body systems' classification ability. Numerical simulations of a classical spin glass illustrate our technique. This toolkit exposes self-organization that eludes detection by thermodynamic measures. Our toolkit more reliably and more precisely detects and quantifies learning by matter.

Many-body systems can learn and remember patterns of drives that propel them far from equilibrium. Such behaviors have been predicted and observed in many settings, from charge-density waves (Coppersmith et al., 1997; Povinelli et al., 1999) to non-Brownian suspensions (Keim & Nagel, 2011; Keim et al., 2013; Paulsen et al., 2014), polymer networks (Majumdar et al., 2018), soap-bubble rafts (Mukherji et al., 2019), and macromolecules (Zhong et al., 2017). Such learning holds promise for engineering materials capable of memory and computation. This potential for applications, with experimental accessibility and ubiquity, have earned these classical nonequilibrium many-body systems much attention recently (Keim et al., 2019). We present a machine-learning toolkit for measuring the learning of drive patterns by many-body systems. Our toolkit detects and quantifies many-body learning more thoroughly and precisely than thermodynamic tools used to date.

A classical, randomly interacting spin glass exemplifies learning driven matter. Consider sequentially applying magnetic fields from a set $\{\vec{A}, \vec{B}, \vec{C}\}$, which we call a *drive*. The spins flip, absorbing work. In a certain parameter regime, the power absorbed shrinks adaptively: The spins migrate toward a corner of configuration space where their configuration approximately withstands the drive's insults. Consider then imposing fields absent from the original drive. Subsequent spin flips absorb more work than if the field belonged to $\{\vec{A}, \vec{B}, \vec{C}\}$.

A simple, low-dimensional property of the material—absorbed power—distinguishes drive inputs that fit a pattern from drive inputs that do not. This property reflects a structural change in the spin glass's configuration. The change is long-lived and not easily erased by a new drive (Gold & England, 2019). For these reasons, we say that the material has learned the drive.

Many-body learning has been quantified with properties commonplace in thermodynamics. Examples include power, as explained above, and strain in polymers that learn stress amplitudes. Such thermodynamic diagnoses have provided insights but suffer from two shortcomings. First, the thermodynamic properties vary from system to system. For example, work absorption characterizes the spin glass's learning; strain characterizes non-Brownian suspensions'. A more general approach would facilitate comparisons and standardize analyses. Second, thermodynamic properties were defined for macroscopic equilibrium states. Such properties do not necessarily describe far-from-equilibrium systems' learning optimally.

Separately from many-body systems' learning, machine learning has flourished over the past decade (Nielsen, 2015; Goodfellow et al., 2016). Machine learning has enhanced our understanding of how natural and artificial systems learn. We apply machine learning to measure learning by many-body systems.

We use *representation learning* (Bengio et al., 2012),

[*]Equal contribution [1]Physics of Living Systems, Department of Physics, MIT [2]Department of Mathematics, MIT [3]W. M. Keck Science Department, Claremont Colleges [4]GlaxoSmithKline AI/ML [5]ITAMP, Harvard-Smithsonian Center for Astrophysics [6]Department of Physics, Harvard University [7]Research Laboratory of Electronics, MIT. Correspondence to: Weishun Zhong <wszhong@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

where a bottleneck-shaped neural network receives a high-dimensional variable X, compresses relevant information into a low-dimensional *latent variable* Z, then decompresses Z into a prediction $\hat{Y}$ of a high-dimensional variable Y. The size of the bottleneck Z controls a tradeoff between the memory consumed and the prediction's accuracy.

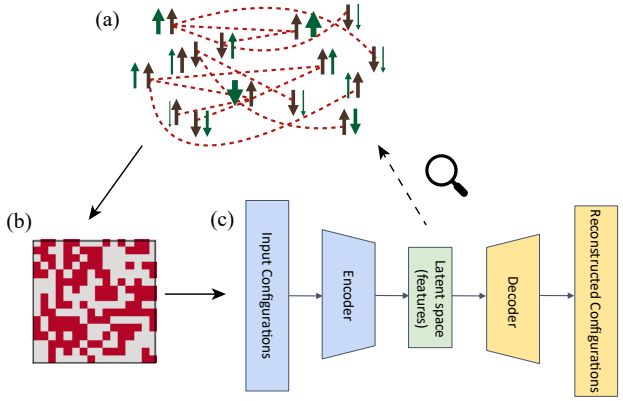

Figure 1. **(a) Schematic of driven spin-glass:** black arrows represent spins s_j, green arrows represent external magnetic fields $\vec{A}_j$, and red dashed lines represent disorder connections J_{jk}. **(b) Example configuration:** at some time-t, for visualization, we arrange configuration of the 256 spins on a 2-dimensional lattice (the real system lives on a mean-degree 4 Erdös-Rényi random graph). Red pixles represent 'up' spins $s_j = 1$, and gray pixels represent 'down' spins $s_j = -1$. **(c) Schematic of VAE:** our VAE consists of three parts, the encoder, latent space, and the decoder. We input spin configurations into the VAE and it outputs the reconstructed configurations. After training, we use the learned features in the latent space to investigate the learning behavior in the spin-glass system.

s

Our goal is to use representation learning to measure how effectively a far-from-equilibrium many-body system learns a drive. We illustrate with numerical simulations of the spin glass (Fig. 1(a)), whose learning has been detected with work absorption (Gold & England, 2019). However, our methods generalize to other platforms. Our measurement scheme offers three advantages:

1. Bottleneck neural networks register learning behaviors more thoroughly and precisely than work absorption.

2. Our framework applies to a wide class of strongly driven many-body systems. The framework does not rely on any particular thermodynamic property tailored to, e.g., spins.

3. Our approach unites a machine-learning sense of learning with the statistical mechanical sense. This union is conceptually satisfying.

Although our techniques can be applied to many facets of learning, including but not limited to: classification, memory capacity, discrimination ability, and novelty detection, in the following we focus on applying representation learning to measure the classification ability of the spin glass system.

Our measurement protocols share the following structure: The many-body system is trained with a drive (e.g., magnetic fields $\vec{A}$, $\vec{B}$, and $\vec{C}$). Then, the system is tested (e.g., with a field $\vec{D}$). Training and testing are repeated in many trials. Configurations realized by the many-body system at some time snapshot are used to train a bottleneck neural network via unsupervised learning. The neural network may then receive configurations from the testing of the many-body system. Finally, we analyze the neural network's bottleneck.

In this setup, we use configurations of the spin glass as inputs (Fig. 1(b)) to the neural network for unsupervised learning. Note that although we could have chosen supervised learning, it needs drive labels that are not directly available to the many-body system, and thus tends to pick up unphysical features in the latent space.

The neural network we use is a *variational autoencoder* (VAE), (Kingma & Welling, 2013; Jimenez Rezende et al., 2014; Doersch, 2016), a generative model that learns the target distribution via variational Bayesian inference, then generates new samples from the learned distribution.

Our VAE has five fully connected hidden layers, with neuron numbers 200-200-(number of Z neurons)-200-200 (Fig. 1(c)). We usually restrict the latent variable Z to 2-4 neurons. This choice facilitates the visualization of the latent space and suffices to quantify our spin glass's learning. Growing the number of degrees of freedom, and the number of drives, may require more dimensions. But our study suggests that the number of dimensions needed $\ll$ the system size.

Figure 2(a) depicts the latent space Z, where the neural network maps each inputted configuration to one latent-space dot. Close-together dots correspond to configurations produced by the same field, if the spin glass and neural network learn well. We illustrate this by coloring each dot according to the field that produced it. Furthermore, by color-coding the latent space with thermodynamic quantities, we are able to associate physical meanings to the latent space: one of the diagonal direction corresponds to absorbed-power (Fig. 2 (b)), and the radial direction corresponds to magnetization (Fig. 2 (c)).

Spin glass: A spin glass exemplifies the statistical mechanical learner (Gold & England, 2019). Simulations are of $N = 256$ classical spins. The j^{th} spin occupies one of two possible states: $s_j = \pm 1$.

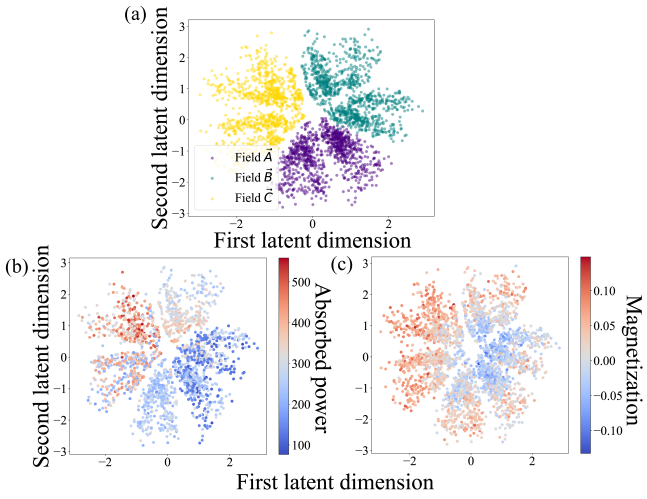

Figure 2. **(a) Visualization of latent space:** A VAE formed latent space while training on configurations assumed by a 256-spin glass during repeated exposure to three fields, A, B, and C. The neural network mapped each configuration to a dot in latent-space. We color each dot in accordance with the field that produced the configuration. **(b) Correspondence of absorbed power to a diagonal:** we color-coded the latent-space with the absorbed power, and find that the absorbed power grows from the bottom righthand corner to the upper lefthand corner. **(c) Correspondence of magnetization to the radial direction:** we color-coded the latent-space with magnitization, and find that the magnetization grows radially.

The spins couple together and experience an external magnetic field: Spin j evolves under a Hamiltonian

$$H_j(t) = \sum_{k \neq j} J_{jk} s_j s_k + A_j(t) s_j, \tag{1}$$

and the spin glass evolves under $H(t) = \frac{1}{2} \sum_{j=1}^{N} H_j(t)$, at time t. We call the first term in Eq. (1) the *interaction energy* and the second term the *field energy*. The couplings $J_{jk} = J_{kj}$ are defined in terms of an Erdös-Rényi random network: Spins j and k have some probability p of interacting, for all j and $k \neq j$. Each spin couples to four other spins, on average. The nonzero couplings J_{jk} are selected according to the standard normal distribution.

$A_j(t)$ denotes the magnitude and sign of the external field experienced by spin j at time t. The field always points along the same direction, the z-axis, so we omit the arrow from $\vec{A}_j(t)$. We will simplify the notation for the field from $\{A_j(t)\}_j$ to A (or B, etc.). Each A_j is selected according to a normal distribution with zero mean and standard deviation 3. The field changes every 100 seconds. To train the spin glass, we construct a drive by forming a set $\{A, B, \ldots\}$ of random fields. We randomly select a field from the set, then apply the field for 100 s. This selection-and-application process is performed 300 times.

The spin glass exchanges heat with a bath at a temperature $T = 1/\beta$. We set Boltzmann's constant to $k_B = 1$. Energies are measured in Kelvins (K). To flip, a spin must overcome a height-B energy barrier. Spin j tends to flip at a rate $\omega_j = e^{\beta[H_j(t) - B]}/(1 \text{ second})$. This rate has the form of Arrhenius's law and obeys detailed balance. The average spin flips once per 10^7 s. We model the evolution with discrete 100-s time intervals, using the Gillespie algorithm.

The spins absorb work when the field changes, as from $\{A_j(t)\}$ to $\{A'_j(t)\}$. The change in the spin glass's energy equals the work absorbed by the spin glass: $W := \sum_{j=1}^{N} \left[A'_j(t) - A_j(t)\right] s_j$. Absorbed power is defined as $W/(100 \text{ s})$. The spin glass dissipates heat by losing energy as spins flip.

The spin glass is initialized in a uniformly random configuration C. Then, the spins relax in the absence of any field for 100,000 seconds. The spin glass navigates to near a local energy minimum. If a protocol is repeated in multiple trials, all the trials begin with the same C.

In a certain parameter regime, the spin glass learns its drive effectively, even according to the absorbed power (Gold & England, 2019). Consider training the spin glass on a drive $\{A, B, C\}$. The spin glass absorbs much work initially. If the spin glass learns the drive, the absorbed power declines. If a dissimilar field D is then applied, the absorbed power spikes. The spin glass learns effectively in the Goldilocks regime $\beta = 3 \text{ K}^{-1}$ and $B = 4.5$ K (Gold & England, 2019): The temperature is high enough, and the barriers are low enough, that the spin glass can explore its configuration space. But T is low enough, and the barriers are high enough, that the spin glass is not hopelessly peripatetic. Robust learning is distinguished from superficially similar behaviors in in (Zhong et al., 2020).

A system *classifies* a drive when identifying the drive as one of many possibilities. A VAE, we find, reflects more of a spin glass's classification ability than absorbed power does: We generated random fields A, B, C, D, and E. From 4 of the fields, we formed the drive $\mathcal{D}_1 := \{A, B, C, D\}$. On the drive, we trained the spin glass in each of 1,000 trials. In each of 1,000 other trials, we trained a fresh spin glass on a drive $\mathcal{D}_2 := \{A, B, C, E\}$. We repeated this process for each of the 5 possible 4-field drives. 90% of the trials were randomly selected for training our neural network. The rest 10% were used for testing.

At each *physical time-t* ($0 \leq t \leq 300$) as measured by the time-evolution in the Gillespie algorithm, we train a fresh VAE with the corresponding configurations from all of the trials in the training set, and treat the corresponding trained latent space latent space (as in Fig. 2) as empirical probability distributions.

We inputted into the neural network a time-t configura-

tion from a test trial. The neural network compressed the configuration into a latent-space point. We use maximum-likelihood estimation to find out which drive most likely, according to the empirical probability density calculated from the trained latent space, generated the test latent-space point. (see (Bishop, 2006) and (Zhong et al., 2020)). We performed this testing and estimation for each trial in the test data. The fraction of trials in which the estimation succeeded constitutes the *score*. The score is plotted against physical time-t (not training epochs) in Fig. 3 (blue, upper curve).

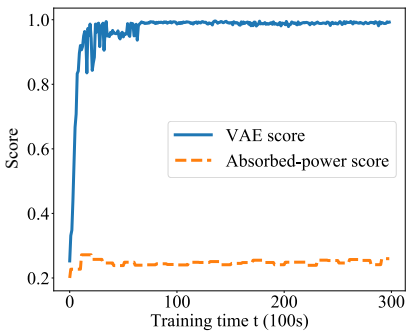

Figure 3. **Quantification of a many-body system's classification ability:** A spin glass classified a drive as one of five possibilities. We define the system's classification ability as the score of maximum-likelihood estimation performed with a VAE (blue, upper curve). We compare with the score of maximum-likelihood estimation performed with absorbed power (orange, lower curve). The variational-autoencoder score rises to near the maximum, 1.00. The thermodynamic score exceeds the random-guessing score, $1/5$, slightly. The neural network detects more of the spins' classification ability.

We compare with the classification ability attributed to the spin glass by the absorbed power: For each drive and each time t, we histogrammed the power absorbed while that drive was applied at t in a neural-network-training trial. Then, we took a trial from the test set and identified the power absorbed at t. We inferred which drive most likely, according to the histograms, produced that power. The guess's score appears as the orange, lower curve in Fig. 3.

A score maximizes at 1.00 if the drive is always guessed accurately. The score is lower-bounded by the random-guessing value $1/(\text{number of drives}) = 1/5$. In Fig. 3, each score grows over tens of field switches. The absorbed-power score begins at[1] 0.20 and comes to fluctuate around 0.25. The neural network's score comes to fluctuate slightly below 1.00. Hence the neural network detects more of the spin glass's classification ability than the absorbed power does, in addition to suggesting a means of quantifying the classification ability rigorously.

Discussion: We have detected and quantified a many-body system's learning of its drive, using representation learning, with greater sensitivity than absorbed power affords. We illustrated by quantifying a many-body system's ability to classify drives, with the score of maximum-likelihood estimates calculated from a VAE's latent space. Our toolkit extends to quantifying memory capacity, discrimination, and novelty detection (Zhong et al., 2020). Our toolkit also has wide applicability, not depending on whether the system exhibits magnetization or strain or another thermodynamic response. Furthermore, our representation-learning toolkit signals many-body learning more sensitively than does the seemingly best-suited thermodynamic tool. This work engenders several opportunities:

(i) *Interpreting latent space:*

Convention biases thermodynamicists toward measuring volume, magnetization, heat, work, etc. The neural network might identify new macroscopic variables better-suited to far-from-equilibrium statistical mechanics, or hidden nonlinear relationships amongst thermodynamic variables. In fact, by decoding configurations from the latent space then computing values of the corresponding thermodynamic variables, we are able to associate physical meanings in two latent-space directions: One of the diagonal direction corresponds to increasing absorbed power, and the radial direction corresponds to increasing magnitization. The two directions we found are nonorthogonal, suggesting that the neural network has identified a nonlinear combination of existing thermodynamic variables that is more successful in detecting many-body learning than any of the known thermodynamic variables.

(ii) Resolving open problems in statistical mechanical learning: Our toolkit is well-suited to answering open problems about many-body learners, such as soap bubble rafts that learns the amplitude of the stress that it is trained with (Mukherji et al., 2019; Miller, 2019). Bottleneck neural networks may reveal what microscopic properties distinguish trained from untrained rafts.

(iii) Learning about representation learning: Representation learning, we argue, shares its structure with problems in which a strong drive forces a many-body system. The system's microstate, like X, occupies a high-dimensional space. A macrostate synopsizes the microstate in a few numbers, such as particle number and magnetization. This synopsis parallels the concept of low-dimensional latent variables[2].

[1] The neural network's score begins a short distance from 0.20. The distance, we surmise, comes from stochasticity of three types: the spin glass's initial configuration, the maximum-likelihood estimation, and stochastic gradient descent. Stochasticity of only the first two types affects the absorbed-power score.

[2] See (Alemi & Fischer, 2018) for a formal parallel between

This parallel enables us to use representation learning to gain insight into statistical mechanics. Recent developments in information-theoretic far-from-equilibrium statistical mechanics (e.g., (Still et al., 2012; Parrondo et al., 2015; Crutchfield, 2017; Kolchinsky & Wolpert, 2017)) might, in turn, shed new light on representation learning.

References

Alemi, A. A. and Fischer, I. TherML: Thermodynamics of Machine Learning. *arXiv:1807.04162*, 2018.

Bengio, Y., Courville, A., and Vincent, P. Representation Learning: A Review and New Perspectives. *arXiv:arXiv:1206.5538*, 2012.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Coppersmith, S. N., Jones, T. C., Kadanoff, L. P., Levine, A., McCarten, J. P., Nagel, S. R., Venkataramani, S. C., and Wu, X. Self-organized short-term memories. *Phys. Rev. Lett.*, 78:3983–3986, May 1997. doi: 10.1103/PhysRevLett.78.3983. URL https://link.aps.org/doi/10.1103/PhysRevLett.78.3983.

Crutchfield, J. P. The origins of computational mechanics: A brief intellectual history and several clarifications. *arXiv:1710.06832*, 2017.

Doersch, C. Tutorial on Variational Autoencoders. *arXiv:1606.05908*, 2016.

Gold, J. M. and England, J. L. Self-organized novelty detection in driven spin glasses. *arXiv:1911.07216*, 2019.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Jimenez Rezende, D., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proc. 31st Int. Conf. on Machine Learning*, 2014.

Keim, N. C. and Nagel, S. R. Generic transient memory formation in disordered systems with noise. *Phys. Rev. Lett.*, 107:010603, Jun 2011. doi: 10.1103/PhysRevLett.107.010603. URL https://link.aps.org/doi/10.1103/PhysRevLett.107.010603.

Keim, N. C., Paulsen, J. D., and Nagel, S. R. Multiple transient memories in sheared suspensions: Robustness, structure, and routes to plasticity. *Phys. Rev. E*, 88:032306, Sep 2013. doi: 10.1103/PhysRevE.88.032306. URL https://link.aps.org/doi/10.1103/PhysRevE.88.032306.

representation learning and equilibrium thermodynamics.

Keim, N. C., Paulsen, J. D., Zeravcic, Z., Sastry, S., and Nagel, S. R. Memory formation in matter. *Rev. Mod. Phys.*, 91:035002, Jul 2019. doi: 10.1103/RevModPhys.91.035002. URL https://link.aps.org/doi/10.1103/RevModPhys.91.035002.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114*, 2013.

Kolchinsky, A. and Wolpert, D. H. Dependence of dissipation on the initial distribution over states. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083202, aug 2017. doi: 10.1088/1742-5468/aa7ee1. URL https://doi.org/10.1088\%2F1742-5468\%2Faa7ee1.

Majumdar, S., Foucard, L. C., Levine, A. J., and Gardel, M. L. Mechanical hysteresis in actin networks. *Soft Matter*, 14:2052–2058, 2018. doi: 10.1039/C7SM01948C. URL http://dx.doi.org/10.1039/C7SM01948C.

Miller, J. A raft of soap bubbles remembers its past. *Physics Today*, 2019.

Mukherji, S., Kandula, N., Sood, A. K., and Ganapathy, R. Strength of mechanical memories is maximal at the yield point of a soft glass. *Phys. Rev. Lett.*, 122:158001, Apr 2019. doi: 10.1103/PhysRevLett.122.158001. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.158001.

Nielsen, M. *Neural Networks and Deep Learning*. Determination Press, 2015.

Parrondo, J. M. R., Horowitz, J. M., and Sagawa, T. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015. ISSN 1745-2481. doi: 10.1038/nphys3230. URL https://doi.org/10.1038/nphys3230.

Paulsen, J. D., Keim, N. C., and Nagel, S. R. Multiple transient memories in experiments on sheared non-brownian suspensions. *Phys. Rev. Lett.*, 113:068301, Aug 2014. doi: 10.1103/PhysRevLett.113.068301. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.068301.

Povinelli, M. L., Coppersmith, S. N., Kadanoff, L. P., Nagel, S. R., and Venkataramani, S. C. Noise stabilization of self-organized memories. *Phys. Rev. E*, 59:4970–4982, May 1999. doi: 10.1103/PhysRevE.59.4970. URL https://link.aps.org/doi/10.1103/PhysRevE.59.4970.

Still, S., Sivak, D. A., Bell, A. J., and Crooks, G. E. Thermodynamics of prediction. *Phys. Rev. Lett.*, 109:120604, Sep 2012. doi: 10.1103/PhysRevLett.109.120604. URL https://link.aps.org/doi/10.1103/PhysRevLett.109.120604.

Zhong, W., Schwab, D. J., and Murugan, A. Associative pattern recognition through macro-molecular self-assembly. *Journal of Statistical Physics*, 167(3):806–826, May 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1774-2. URL https://doi.org/10.1007/s10955-017-1774-2.

Zhong, W., Gold, J. M., Marzen, S., England, J. L., and Halpern, N. Y. Quantifying many-body learning far from equilibrium with representation learning, 2020.