Supplemental Information for Machine learning glass caging order parameters with an artificial nested neural network

Kaihua Zhang,[1,*] Xinyang Li,[2,3,*] Yuliang Jin,[2,3,4,†] and Ying Jiang[1,5,‡]

[1] *School of Chemistry, Beihang University, Beijing 100191, China*
[2] *CAS Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China*
[3] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
[4] *Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, Zhejiang 325000, China*
[5] *Center of Soft Matter Physics and Its Applications, Beihang University, Beijing 100191, China*

CONTENTS

* Contributed equally to this work
† yuliangjin@mail.itp.ac.cn
‡ yjiang@buaa.edu.cn

S1. GLASS MODEL

The model [1–5] consists of $N = 125 - 8000$ polydisperse hard spheres (HSs), whose diameters are distributed according to a continuous function $P_D(D_{\rm min} \le D \le D_{\rm min}/0.45) \sim D^{-3}$. The volume of simulation box is V, and periodic boundary conditions are used. The system state is characterized by volume fraction φ and reduced temperature $\hat{T} = 1/\hat{P} = Nk_{\rm B}T/PV$, where P is the pressure, $\hat{P}$ the reduced pressure, $k_{\rm B} = 1$ the Boltzmann constant, and $T = 1$ the temperature. We set the mean diameter $D_{\rm mean}$ as unit length, and the particle mass m as unit mass. Crystallization is suppressed by polydispersity, and will not be discussed in this study.

The phase diagram of the model is presented in Fig. S1 [2]. Any state of the system is described by two thermodynamic parameters, the volume fraction φ and the reduced temperature $\hat{T}$. The Carnahan-Starling (CS) equation of state (EOS) [6] well captures the relationship between φ and $\hat{T}$ of liquid states [2]. The mode-coupling theory (MCT) transition point, $\hat{T}_{\rm MCT} \approx 0.044$ (or $\varphi_{\rm MCT} \approx 0.594$), was estimated in [2], below which equilibrium becomes difficult in ordinary molecular dynamics (MD) simulations. Following convention, the intersection between liquid and glass EOSs is defined as the glass transition point $\{\varphi_{\rm g}, \hat{T}_{\rm g}\}$. The location of glass transition point depends on the compression rate Γ, and therefore, not unique. As can be seen in Fig. S1, the EOS of an ultra-stable glass ($\hat{T}_{\rm g} \ll \hat{T}_{\rm MCT}$) displays enormous overshooting over the liquid EOS and the two EOSs are connected by an abrupt jump. In contrast, the EOS of a poorly annealed ordinary glass ($\hat{T}_{\rm g} \approx \hat{T}_{\rm MCT}$) merges smoothly to the liquid one.

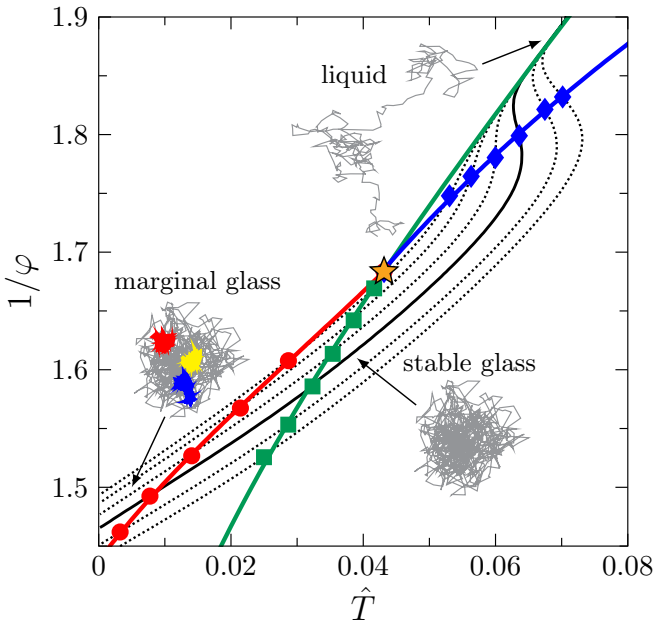

FIG. S1. Phase diagram of a hard sphere glass model [2]. The green, blue and red lines represent the CS liquid EOS, the melting line that separates liquid and glass phases, and the Gardner line that separates stable and marginally stable glass phases. The orange star represents the MCT transition point. The system is initially equilibrated at $\hat{T}_{\rm g}$ (green squares), and then evolves following the glass EOSs (dotted black lines) under compression or decompression. To study melting and Gardner transitions, we focus on the ultra-stable case of $\hat{T}_{\rm g} = 0.033$ (solid black line). Typical particle trajectories are plotted to show the diffusive motion in liquids, a confined cage in stable glasses, and the split of the cage into sub-cages in marginal glasses (the three sub-cages are visualized by trajectories of the same particle in three replicas, which are compressed from the same initial configuration at $\hat{T}_{\rm g}$).

To study melting and Gardner transitions, we consider ultra-stable glasses. Deeply supercooled liquid states are prepared by using an efficient swap Monte Carlo (MC) algorithm [1]. Once the initial states are obtained, we switch to regular MD (without swap) to simulate follow-up dynamics. These deeply supercooled liquid states have extraordinarily large structural relaxation (α-relaxation) time τ_α in the MD time unit, much larger than our MD simulation time window.

Two instabilities – the melting and the Gardner transitions – occur if one decompresses or compresses an ultra-stable HS glass. Under decompression, the glass is effectively "heated" up and eventually melts into a liquid at a melting temperature $\hat{T}_{\rm m}(\hat{T}_{\rm g}) > \hat{T}_{\rm g}$. Although glass melting is a non-equilibrium procedure by definition, previous experiments [7] and simulations [8] showed that, this procedure in ultra-stable glasses is very similar to the melting

of crystals, which is a first-order phase transition. In contrast, the melting of (poorly annealed) ordinary glasses is a smooth crossover without any discontinuous behavior.

On the other hand, a *Gardner transition* [9, 10] is expected to occur at $\hat{T}_{\rm G}(\hat{T}_{\rm g}) < \hat{T}_{\rm g}$ if an ultra-stable HS glass is compressed. The Gardner transition separates the *stable glass* (at $\hat{T} > \hat{T}_{\rm G}$) and the *marginally stable glass* (at $\hat{T} < \hat{T}_{\rm G}$) phases. It is predicted to be a second-order phase transition in large dimensions by the mean-field glass theory [10, 11]. Evidence of the Gardner transition in physical dimensions (2D and 3D) has been reported in a number of simulations [2, 4, 12] and experiments [13–15]. A fixed point is found by field-theory calculations, suggesting that the transition survives in low dimensions [16]. The existence of a Gardner transition in 3D ultra-stable HS glasses is supported by a recent numerical study, which combines finite-time-finite-size analyses with machine learning [17].

We do not study the melting and Gardner transitions in ordinary glasses: The melting of an ordinary glass is nearly reversible to the glass transition, and thus for our purpose, it is sufficient only to consider the latter. The Gardner transition in ordinary glasses is blurred by activated dynamics [2]; as a result, we do not expect to observe critical scalings.

S2. MOLECULAR SIMULATION METHODS

A. Molecular dynamics simulations

We use the Lubachevsky-Stillinger algorithm (event-driven MD) to simulate compression quench [18, 19]. During compression/decompression, the sizes of all particles are increased/decreased proportionally with a fixed rate $\Gamma = \frac{1}{2D}\frac{dD}{dt}$. The simulation time is expressed in units of $\sqrt{mD^2_{\rm mean}/(k_{\rm B}T)}$. We simulate each configuration and measure its $\hat{T}$ at a fixed packing fraction φ. Then we collect configurations at the desired $\hat{T}$, based on which physical quantities are computed. Additional simulation details can be found in Refs. [2, 17].

B. Swap algorithm

The swap algorithm simulates artificial dynamics that can efficiently accelerate reaching equilibrium [1]. At each swap MC step, two particles are randomly picked, and swapped if they do not overlap with neighbor particles at the new positions. While the dynamics are unrealistic, the final configurations in equilibrium are equivalent to those generated by standard MD and MC simulations.

S3. MACHINE LEARNING METHODS

A. Architecture of the artificial neural network

The nested neural network (NNN) comprises two levels of networks, which in general can have different structures (see Fig. 1). There are N duplicated small networks at the first level, each of which extracts the latent caging features of one single particle. The small network has only one hidden layer, besides the input and output layers. Both input and hidden layers have $M_{\rm r}$ neuron nodes, and the output layer has a single node. The $i^{\rm th}$ hidden node is connected by a single link to the $i^{\rm th}$ input node, and is activated by the exponential linear unit (ELU) function. We have checked that the machine learning results do not change with a different type of activation function, such as a tanh function, for both melting and Gardner transitions. The output node simply takes an average of $M_{\rm r}$ hidden nodes. The N small networks share the same parameters (weights and bias), and thus there are only $2M_{\rm r}$ free parameters at the first level. The number of parameters can be further reduced from $2M_{\rm r}$ to 2, considering the permutation symmetry of replicas. In principle, one could choose other architectures (e.g., fully connected feed-forward neural network (FNN)) for small networks, and set their parameters to be independent. In practice, however, we find that using a small number of free parameters at the first level can significantly increase the efficiency of the NNN model during training, without losing its compatibility and predictive power. Generally, the number of free parameters can be minimized by considering the symmetries of the physical system under consideration, while networks with redundant parameters can work equally well if proper regularization is imposed [20].

The N output nodes of the first-level small networks are considered as input nodes of the followed big FNN at the second-level. The FNN has one hidden layer of 128 nodes activated by ELU functions, and one output layer of two nodes that provide binary classifications through softmax activation functions.

B. Blanking window

For the supervised learning of phases, we need to label in advance to which phase a given configuration belongs, during training and validation. A blanking window $[\hat{T}_2, \hat{T}_1]$ is introduced to skip the vicinity of a (presumed) transition. Specifically, the following setup is used for the data presented in the main figures: for the melting transition, configurations at $\hat{T} > \hat{T}_1 = 0.083$ and $\hat{T} < \hat{T}_2 = 0.053$ are labeled as liquids and glasses respectively; for the Gardner transition, configurations at $\hat{T} > \hat{T}_1 = 0.011$ and $\hat{T} < \hat{T}_2 = 0.0045$ are labeled as stable and marginal glasses respectively. Note that, for both transitions, the machine learning results do not sensitively depend on the choice of blanking window (see Fig. 9, Secs. S6 B and S7 B). In contrast, the learning results of the glass transition correlates strongly to the blacking window.

C. Training the nested neural network and validating predictions

A cross-entropy cost function is minimized during training. The Adam optimizer [21] is used to implement the stochastic gradient descent method for updating the network parameters. To avoid overfitting, a dropout strategy [22] is used, which randomly skips 20% hidden nodes at each step. To augment the training data set, we perform $N_{\rm shuffle} = 20 - 200$ random shuffles of the elements in the input vector, which is equivalent to randomly ordering particle indexes. In this way, we expand the training data set to $N_{\rm s}^{\rm train} \times N_{\hat{T}} \times N_{\rm shuffle} \sim 10^5$ samples. The random shuffling apparently destroys spatial correlations (if there is any) between particles. However, we find that it does not modify the final predictions noticeably. Validation is performed after each training step, by calculating the cost function for the validation data set. The entire training procedure is terminated when the validation cost function reaches a minimum. Such an early stopping strategy can efficiently avoid overfitting.

D. Making predictions using the nested neural network

Once well trained, the NNN can make phase predictions for the samples in the testing data set. For each test sample at a temperature $\hat{T}$, the NNN provides an output value of 0 or 1. The arithmetic mean of the output over all test samples gives an estimation of the probability P of the system belonging to a specific phase, and $1 - P$ to the other.

To achieve a reliable prediction, we independently train our NNN for 10 times (runs) and calculate the mean and the statistical error of the predicted P. For each run, $N_{\rm s}^{\rm train}$ training samples and $N_{\rm s}^{\rm pred}$ prediction samples are randomly selected from $N_{\rm s}^{\rm total}$ samples, and the remaining $N_{\rm s}^{\rm valid}$ samples are used for validation.

E. Unsupervised classification using t-distributed stochastic neighbor embedding

We utilize the unsupervised t-distributed stochastic neighbour embedding (t-SNE) method [23] to group samples in the Gardner phase. The input data are the machine detected caging parameter $\{\tilde{q}_i\}$ of each sample, where $i = 1, 2, \ldots, N$. The algorithm conducts a nonlinear dimensionality reduction, which maps each vector $\{\tilde{q}_i\}$ to a point in two dimensions. In the two-dimensional space, data points are rearranged according to their similarities quantified by a t-distribution kernel function (see the inset of Fig. 2(d)).

S4. FINITE-SIZE SCALING FUNCTION OF THE SUSCEPTIBILITY

To determine the order of phase transition, a standard way is to apply a finite-size analysis of data obtained from experiments or simulations. For example, the fluctuation of order parameter, or the susceptibility, χ, follows a finite-size scaling function around a phase transition,

$$\chi/N^a = \mathcal{X}\left(|\hat{T} - \hat{T}_{\rm c}|N^b\right), \tag{S1}$$

where a, b are two exponents, $\mathcal{X}(x)$ a scaling function whose concrete form is not important in our discussion, and $\hat{T}_{\rm c}$ the transition temperature. The values of a and b depend on the nature of transition: (i) For a standard first-order phase transition without disorder, $a = 1$ and $b = 1$ [24]. An example is the first-order phase transition between positive and negative ferromagnetic phases in the Ising model under an external field. (ii) For a first-order phase

transition in the presence of disorder, such as the yielding transition [25] and the melting transition (see Fig. S2) in glasses, $a = 1$ and $b = 1/2$. Equation (S1) then results in two susceptibilities, a *disconnected* one, $\chi_{\rm dis} = \chi \sim N^a \sim N$, and a *connected* one $\chi_{\rm con} \sim d\mathcal{X}/d\hat{T} \sim N^b \sim N^{1/2}$. The two susceptibilities are related via, $\chi_{\rm dis} \sim \chi^2_{\rm con}$, a relation found in the random field Ising model [26, 27]. (iii) For a second-order phase transition, $a = b\gamma$ and $b = 1/d\nu$, where d is the dimensionality, and γ and ν are the critical exponents for the divergences of susceptibility and correlation length. A standard example is the second-order phase transition between paramagnetic and ferromagnetic phases in the Ising model without a field.

Let us examine the finite-size scaling of the density susceptibility at fixed $\hat{T}$, $\chi_\varphi = N\left[\overline{\varphi^2} - (\overline{\varphi})^2\right] / (\overline{\varphi})^2$, around the melting of ultra-stable glasses, where $\overline{\cdots}$ represents the average over samples. The density susceptibility displays a clear dependence on system size N, around the melting temperature $\hat{T}_{\rm m} \approx 0.064$ (see Fig. S2(a)). Its finite-size scaling satisfies Eq. (S1), with $a \approx 1$ and $b \approx 0.5$ (see Fig. S2(b)). The melting temperature $\hat{T}_{\rm m}(N)$ can be obtained from the peak position of $\chi_\varphi(\hat{T})$ (data plotted in Fig. 5(a)). Note that the caging susceptibility around the Gardner transition suffers from strong finite-size and finite-time effects simultaneously, making a direct analysis difficult (see Ref. [17]).

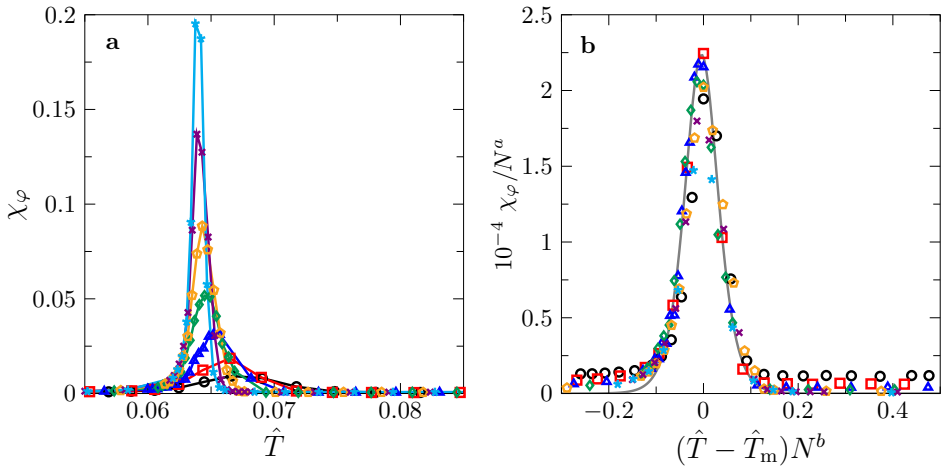

FIG. S2. Finite-size analysis of the density susceptibility around the melting transition. (a) Density susceptibility χ_φ as a function of $\hat{T}$, for a few different system sizes N. (b) Collapse of the χ_φ data according to Eq. (S1). Best collapsing is obtained by setting $a = 0.8$, $b = 0.5$ and $\hat{T}_{\rm c} = \hat{T}_{\rm m} = 0.064$.

S5. MACHINE LEARNING PHASE TRANSITIONS IN THE ISING MODEL

A. Machine learning algorithm

We study the Ising model in both two and three dimensions to validate the finite-size scaling function Eq. (2) in a standard equilibrium system. Machine learning is carried out using TensorFlow [28]. Following [29], we make use of a fully connected FNN, which comprises three layers (input, hidden and output) of nodes. The number of neurons in the input layer is equal to the number of spins $N = L^d$, where L is the linear size of the lattice and d is the dimensionality. The hidden layer is composed of 200 neurons activated by sigmoid functions, and the output layer has 2 neurons activated by a softmax function. During training, a cross-entropy cost function is minimized by means of a stochastic gradient descent method with an Adam optimizer [21]. In order to avoid overfitting, we adopt a dropout regularization [22], and an early stopping strategy.

The input data are spin configurations generated by the Wolff algorithm [30]. At a given magnetic field H and a given temperature T, we prepare $N_{\rm s} \approx 10,000$ samples, and use 80% of them for training, 10% for validation and 10% for prediction. Random shuffling is applied two to four times to make sure that there are sufficient samples at each combination of (H, T) during training.

B. Learning second-order phase transitions in two and three dimensions

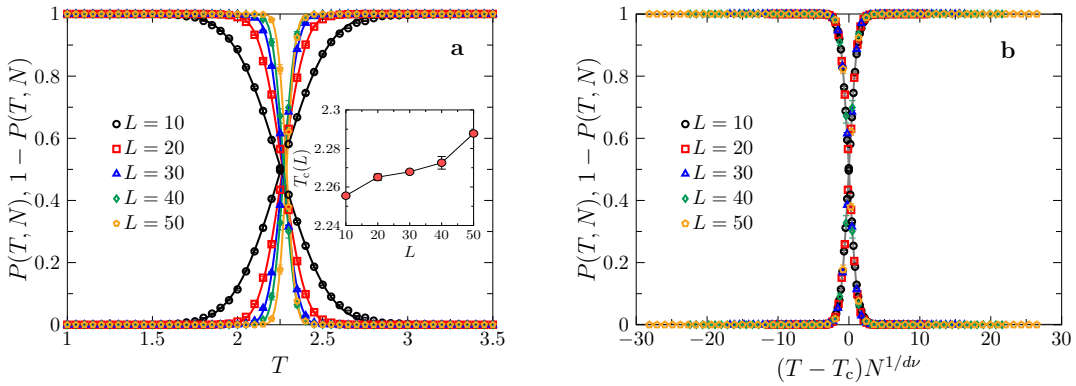

FIG. S3. Machine learning the second-order phase transition in the 2D Ising model. Output from machine leaning, $P(T, N)$ and $1 - P(T, N)$, as functions of (a) T and (b) $(T - T_c)N^{1/d\nu}$, for a few different $N = L^2$ at $H = 0$. The data points are fitted to Eq. (S2) (lines), where the fitting parameters T_c and w are presented in the inset of (a) and Fig. (S7). The exponent $b = 1/d\nu = 0.517(4)$ is obtained from fitting $w(N) = w_0 N^{-b}$, and is used in (b). Error bars represent the standard error of the mean in all figures.

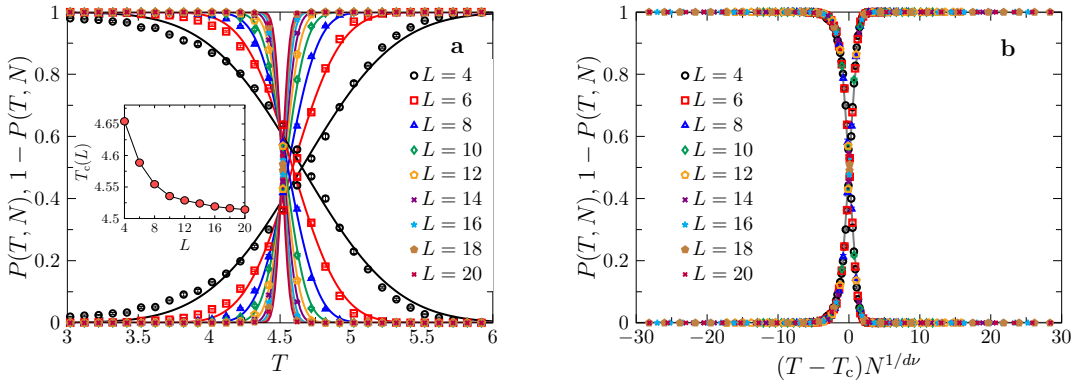

FIG. S4. Machine learning the second-order phase transition in the 3D Ising model. Output from machine leaning, $P(T, N)$ and $1 - P(T, N)$, as functions of (a) T and (b) $(T - T_c)N^{1/d\nu}$, for a few different $N = L^3$ at $H = 0$, where $b = 1/d\nu = 0.528(4)$. The inset of (a) shows $T_c(L)$.

In two and three dimensions, a second order phase transition occurs at T_c when the temperature is varied under the zero-field condition $H = 0$. Previous studies have established the values of T_c and ν (the critical exponent for the divergence of correlation length): $T_c = 2.26918531421\ldots$ [31] and $\nu = 1$ [32] in 2D; $T_c \approx 4.511528(6)$ [33] and $\nu \approx 0.63012(16)$ [34] in 3D. Supervised machine learning techniques have been well utilized to learn the continuous phase transition in the Ising model in both 2D [29] and 3D [35]. Here we reproduce these results using our algorithm. For this purpose, we generate zero-field ($H = 0$) input configurations around T_c at N_T different temperature points. The data points $P(T, N)$ obtained from machine learning are fitted to

$$P(T, N) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left\{\left[T - T_c(N)\right]/w(N)\right\}, \tag{S2}$$

where $N = L^d$, $\mathrm{erf}(x)$ is the error function, and $T_c(N)$ and $w(N)$ are two fitting parameters representing the critical temperature and the width of transition region (see Fig. S3 for 2D and Fig. S4 for 3D). The estimated critical temperatures agree with existing values (see insets of Figs. S3 and S4). Next, we examine the finite-size scaling function Eq. (2). The scaling function suggests that,

$$w(N) \sim N^{-b}, \tag{S3}$$

which is used to determine the critical exponent, $\nu = 0.967(4)$ (or $b = 1/d\nu = 0.517(4)$) in 2D and $\nu = 0.631(3)$ (or $b = 0.528(4)$) in 3D (see Fig. S7). These estimations are in a good agreement with the standard values, $\nu = 1$ in 2D [32] and $\nu \approx 0.63012$ in 3D [34]. Indeed, the data points of $P(T, L)$ for different sizes can be collapsed onto a single curve if they are plotted as a function of the rescaled quantity $(T - T_c)N^{1/d\nu}$ (see Figs. S3(b) and S4(b)).

C. Learning first-order phase transitions in two and three dimensions

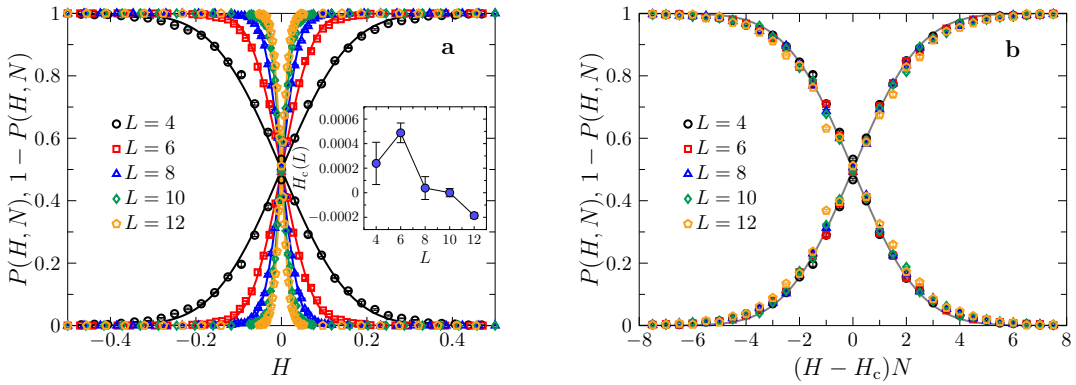

FIG. S5. Machine learning the first-order phase transition in the 2D Ising model. Output from machine leaning, $P(H, N)$ and $1 - P(H, N)$, as functions of (a) H and (b) $(H - H_c)N$, for a few different $N = L^2$. The inset of (a) shows $H_c(L)$.

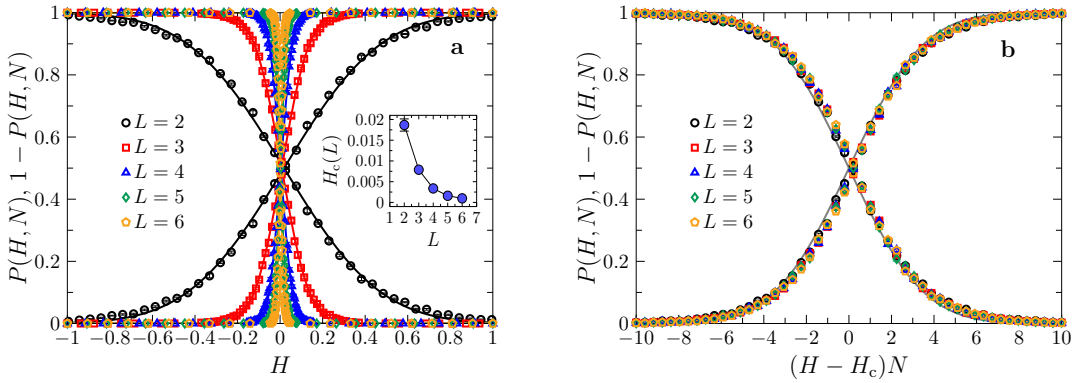

FIG. S6. Machine learning the first-order phase transition in the 3D Ising model. Output from machine leaning, $P(H, N)$ and $1 - P(H, N)$, as functions of (a) H and (b) $(H - H_c)N$, for a few different $N = L^3$. The inset of (a) shows $H_c(L)$.

For a fixed temperature $T < T_c$, a first-order phase transition occurs at $H_c = 0$ when H is varied. To our knowledge, the finite-size scaling Eq. (2) of the first-order phase transition in the Ising model has not been systematically studied yet within the machine learning framework. In this study, we show that the expected scaling Eq. (2) is fully consistent with our data. We set $k_B T/J = 2.1$ for the 2D model and $k_B T/J = 4.0$ for the 3D model, where J is the interaction constant. Configurations are generated at N_H different external fields around $H_c = 0$, with positive and negative fields evenly divided. The data points $P(H, N)$ obtained from machine learning are fitted to

$$P(H,N) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left\{\left[H - H_c(N)\right]/w(N)\right\}. \tag{S4}$$

As shown in Figs. S5 and S6, the estimated transition field is close to $H_c = 0$. Furthermore, we obtain $b \approx 0.96(2)$ for the 2D model, and $b \approx 1.012(4)$ for the 3D model, which are consistent with the expected value $b = 1$ [24].

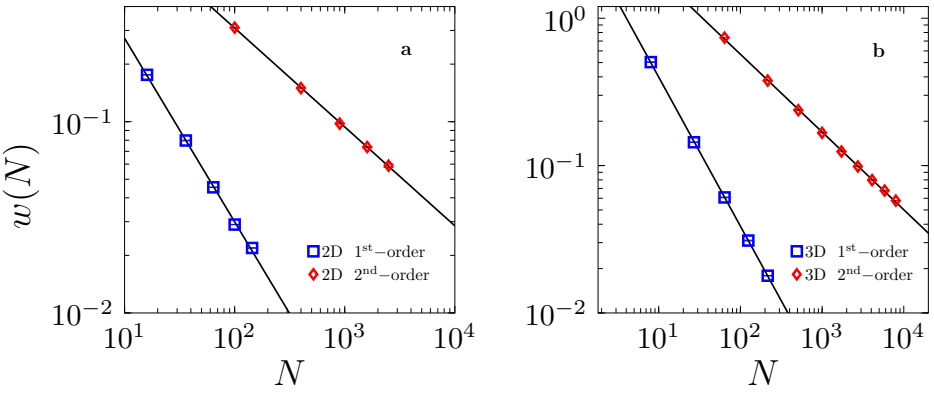

FIG. S7. Comparing the finite size scalings of first- and second-order phase transitions in the Ising model. The data of transition width $w(N)$ are obtained from Figs. (S3-S6). The exponent b is obtained from fitting $w(N) = w_0 N^{-b}$ (lines). (a) In the 2D Ising model, we obtain $b = 0.96(2)$ for the first-order phase transition, and $b = 1/d\nu = 0.517(4)$ (i.e., $\nu = 0.967(4)$) for the second-order phase transition. (b) In the 3D Ising model, we obtain $b = 1.012(4)$ for the first-order phase transition, and $b = 0.528(4)$ (i.e., $\nu = 0.631(3)$) for the second-order phase transition.

D. Distinguishing between first- and second-order phase transitions

Based on above analyses, we confirm that, by utilizing the scaling function Eq. (2), the original machine learning approach proposed in [29] can be generalized to identify both first- and second-order phase transitions, in the standard Ising model. Very importantly, the order of phase transition can be identified because the finite-size exponents b in Eq. (2) are distinguishable within the numerical accuracy for first- and second-order phase transitions. As shown in Fig. S7, $b = 1$ for first-order phase transitions (without considering the effect of disorder), and $b = 1/d\nu$ for second-order phase transitions. While the phase transitions in the Ising model are in equilibrium, we show that the approach can be further generalized to non-equilibrium first-order (melting transition) and second-order (Gardner transition) phase transitions in disordered systems such as glasses (see Fig. 5).

S6. ADDITIONAL RESULTS FOR THE MELTING TRANSITION

A. Dependence of the melting temperature on the decompression rate

The $\hat{T} - \varphi$ EOSs of ultra-stable glasses in Fig. S8 show that the melting transition temperature $\hat{T}_{\rm m}$ decreases with slower decompression. It is expected that, in the limit $\Gamma \to 0$, the hysteresis in EOS will disappear and the glass melting will become a continuous crossover. However, for the range of Γ relevant to this study, the discontinuous feature remains. In the main text, we examine the finite-size effect for a fixed decompression rate $\Gamma = -10^{-4}$, and do not further discuss the rate-dependence.

B. Independence of learning results on the blanking window

During training, the samples at $\hat{T} > \hat{T}_1$ and $\hat{T} < \hat{T}_2$ are labeled as in the liquid and glass phases respectively. The samples in the blanking window $[\hat{T}_2, \hat{T}_1]$ are not used. Figure S9 shows that the machine predicted melting temperature $\hat{T}_{\rm m}$ and transition width w are independent of the blanking window (more specifically, the center of window $\hat{T}_{\rm center} = (\hat{T}_1 + \hat{T}_2)/2$ and the width of window $\Delta\hat{T} = \hat{T}_1 - \hat{T}_2$). Note that, obviously we should require $\hat{T}_{\rm m}$ to be inside of the blanking window, i.e., $\hat{T}_2 < \hat{T}_{\rm m} < \hat{T}_1$. With this restriction, the choice of blanking window is flexible.

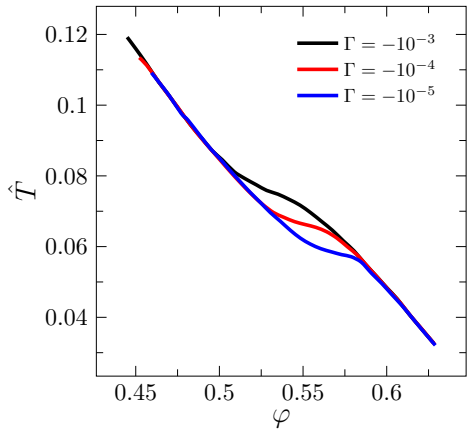

FIG. S8. Evolution of temperature $\hat{T}$ as a function of volume fraction φ under decompression, for a few different decompression rate Γ. The system is composed of $N = 125$ particles, and is decompressed from ultra-stable states at $\{\varphi_g = 0.63, \hat{T}_g = 0.033\}$. The liquid and glass EOSs are connected around the melting temperature $\hat{T}_m$, which decreases with slower decompression.

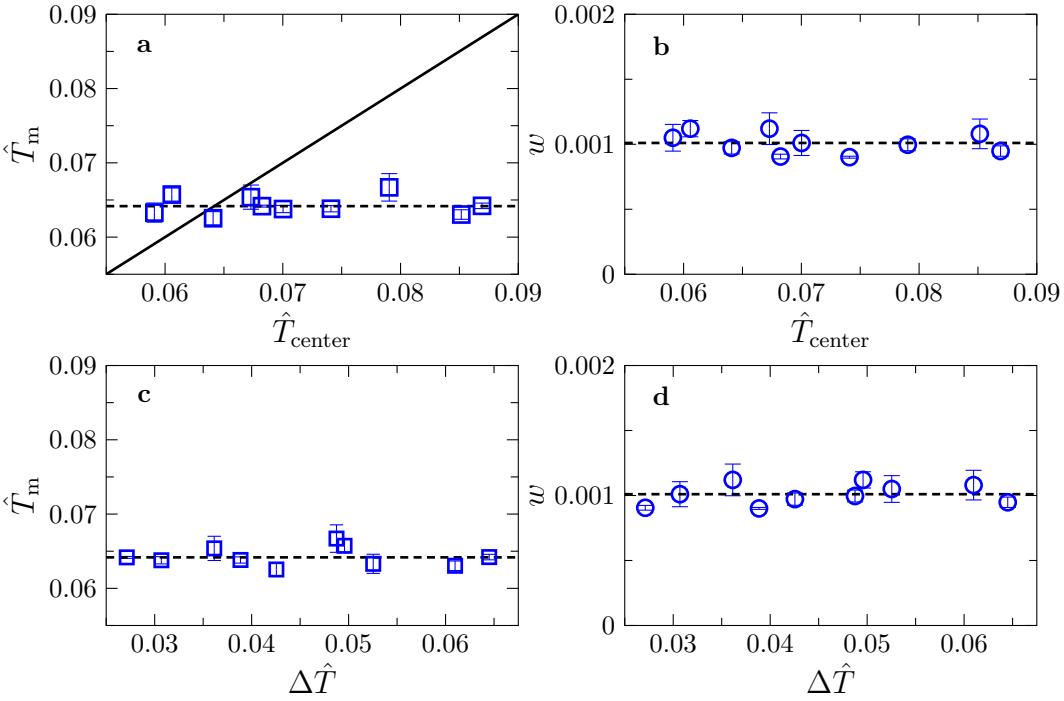

FIG. S9. Independence of $\hat{T}_m$ and w on the blanking window $[\hat{T}_2, \hat{T}_1]$, for the melting transition. The two-level NNN is trained using a few different combinations of $\hat{T}_1$ and $\hat{T}_2$, for $N = 2000$ and $\Gamma = 10^{-4}$. The predicted $\hat{T}_m$ and w are plotted as functions of $\hat{T}_{center}$ and $\Delta\hat{T}$. The horizontal dashed lines are $\hat{T}_m(N = 2000) = 0.064$ and $w(N = 2000) = 0.001$ obtained for $T_1 = 0.083$ and $T_2 = 0.053$. The solid line in (a) represents $\hat{T}_m = \hat{T}_{center}$. The same data in (a) are plotted in Fig. 9(a).

S7. ADDITIONAL RESULTS FOR THE GARDNER TRANSITION

A. Choice of input data

According to the predictions from the mean-field glass theory [11, 36], the features of stable and marginally stable phases are encoded in $\{\Delta_i\}$, and thus in principle one should be able to use $\{|\mathbf{r}_i^{AB}|^2\}$ as input data to train networks. However, in practice, the network fails to correctly identify both phases, when $\{|\mathbf{r}_i^{AB}|^2\}$ are used as input. We find that (see Fig. S10), as $\hat{T} \to 0$, the predicted probability $P \approx 0.5$, while physically we expect $P \approx 1$ (the samples should

all belong to the marginally stable phase at sufficiently low $\hat{T}$). On the other hand, correct and robust predictions are obtained when $\{u_i^{AB} = \frac{|\mathbf{r}_i^{AB}|^2}{\Delta_i} - 1\}$ are used as input. Indeed, the cage size $\Delta_i = \langle |\mathbf{r}_i^{AB}|^2 \rangle_{\rm r}$ generally becomes smaller with decreasing $\hat{T}$, but this effect is independent of the physics of Gardner transition. The purpose of normalization is to remove this effect.

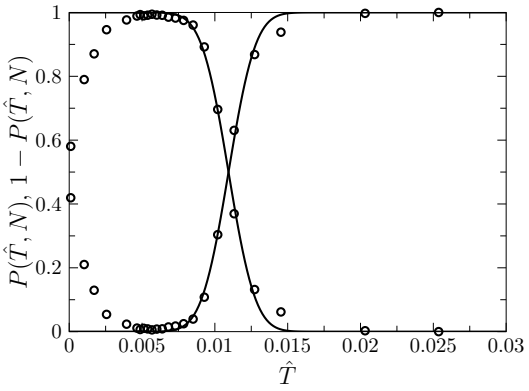

FIG. S10. Failure to learn the Gardner transition using $\{|\mathbf{r}_i^{AB}|^2\}$ as input data. The test is performed for $N = 2000$ systems.

B. Independence of learning results on the blanking window

Figure S11 shows that the machine predicted Gardner transition temperature $\hat{T}_{\rm G}$ and the transition width w are independent of the choice of blanking window.

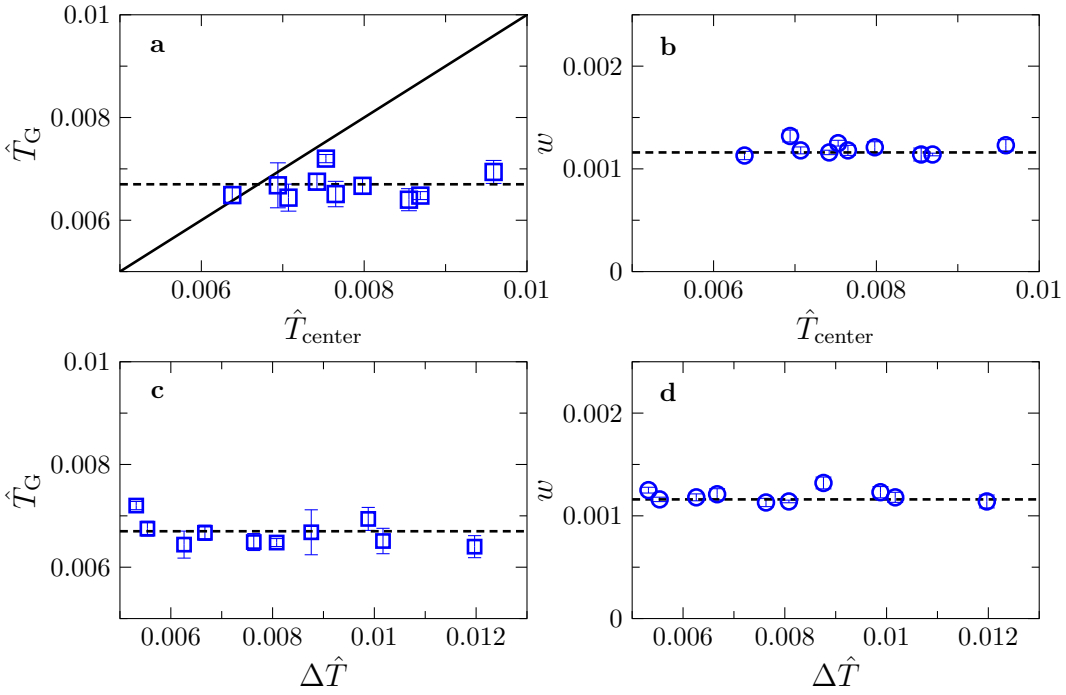

FIG. S11. Independence of $\hat{T}_{\rm G}$ and w on the blanking window $[\hat{T}_2, \hat{T}_1]$, for the Gardner transition. The predicted $\hat{T}_{\rm G}$ and w are plotted as function of $\hat{T}_{\rm center}$ and $\Delta\hat{T}$ (for $N = 8000$ systems). The horizontal dashed lines are $\hat{T}_{\rm G}(N = 8000) = 0.0067$ and $w(N = 8000) = 0.0012$ obtained for $T_1 = 0.011$ and $T_2 = 0.0045$. The solid line in (a) represents $\hat{T}_{\rm G} = \hat{T}_{\rm center}$. The same data in (a) are plotted in Fig. 9(b).

S8. ADDITIONAL DYNAMICAL DATA FOR THE GLASS TRANSITION

In Fig. S12, we plot the data of mean-squared displacement (MSD),

$$\delta r^2(t) = \frac{1}{N}\overline{\sum_{i=1}^{N} |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2}, \tag{S5}$$

around the liquid to ordinary glass transition, where $\mathbf{r}_i(0)$ is the position of particle i right after compression with a rate $\Gamma = 10^{-3}$, and $\mathbf{r}_i(t)$ is the position at time t (we set $t = 0$ and $\Gamma = 0$ after compression). The dynamics clearly slow down with decreasing $\hat{T}$, but activated processes are non-negligible since the MSD is not completely flat even at low $\hat{T}$. Figure S12(b) shows that the average cage size Δ changes smoothly with $\hat{T}$ and depends sensitively on the measurement time t. For a comparison, see the MSD of ultra-stable glasses in Fig. 2 of Ref. [2].

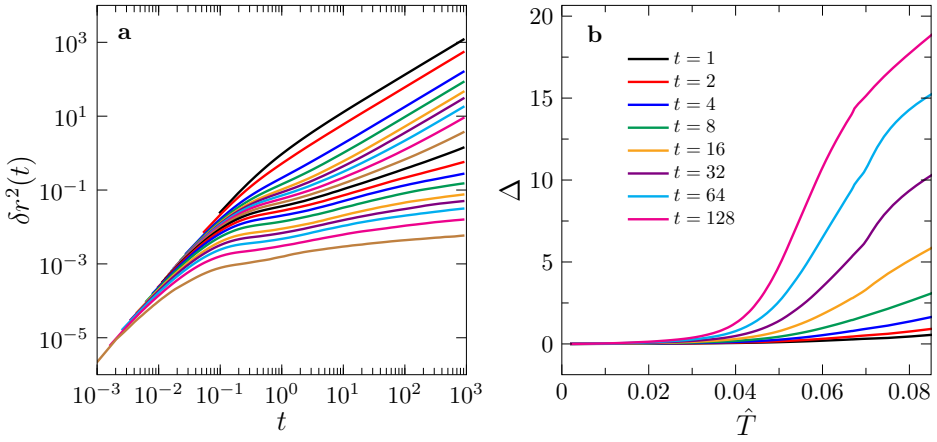

FIG. S12. Dynamical data of the glass transition. (a) MSD data at different densities ($N = 500$): from top to bottom, $\varphi =$ 0.3, 0.4, 0.5, 0.53, 0.55, 0.56, 0.57, 0.58, 0.59, 0.6, 0.61, 0.62, 0.63, 0.64, 0.645, 0.65, 0.655, 0.66 ($\hat{T} = 0.27$, 0.16, 0.084, 0.069, 0.059, 0.054, 0.050, 0.045, 0.040, 0.036, 0.030, 0.025, 0.019, 0.014, 0.011, 0.0078, 0.0049, 0.0018). (b) Average cage size Δ as a function of $\hat{T}$ for a few different measurement time t.

S9. LEARNING A TOY MODEL

We artificially construct a toy model consisting of two phases, which are represented by two distribution functions respectively: a single Gaussian distribution $p_{\rm I}(x) = p_{\rm G}(x;\mu,\sigma)$ for phase I, and a two-Gaussian distribution $p_{\rm II}(x) = \frac{1}{2}p_{\rm G}(x;\mu_1,\sigma_1) + \frac{1}{2}p_{\rm G}(x;\mu_2,\sigma_2)$ for phase II, where $p_{\rm G}(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the standard Gaussian (normal) distribution of mean μ and variance σ^2. We choose the parameters such that the means and variances are identical for the two distributions, but the skewnesses $\langle\left(\frac{x-\mu}{\sigma}\right)^3\rangle$ are different (see Table S1): $\mu = 0, \sigma = 1.768$ in $p_{\rm I}(x)$, and $\mu_1 = -1, \sigma_1 = 2, \mu_2 = 1, \sigma_2 = 0.5$ in $p_{\rm II}(x)$.

The setup for the machine learning algorithm is similar to that illustrated in Fig. 1. We use $N = 1$ "particle", whose feature is described by an input vector $\mathbf{V}$ of $M_{\rm r} = 1000$ random numbers drawn from $p_{\rm I}(x)$ or $p_{\rm II}(x)$. Thus $\mathbf{V}$ is a representation of the probability distribution function for each phase. The front-connected small network maps the distribution $p_{\rm I}(x)$ or $p_{\rm II}(x)$ to a single scalar "order parameter" q. The two-level NNN is trained by $N_{\rm s}^{\rm train} = 15000$ samples, and makes the phase prediction for $N_{\rm s}^{\rm pred} = 2500$ samples. The rest of the method is equivalent to that for learning the glass model.

The Pearson correlation coefficients r between the first three moments and the parameter q are computed, taking account of both phases (see Table S1). The predicted q is strongly correlated to the skewness ($r \approx 1$), while its correlation to the mean or the variance is negligible ($r \approx 0$). This exercise shows that our method can correctly extract higher-order statistical moments from the input data when the mean and variance are trivial.

TABLE S1. Machine learning results for the toy model. Presented are the first three moments (mean, variance and skewness) of probability distribution functions $p_{\rm I}(x)$ and $p_{\rm II}(x)$, and the Pearson correlation coefficient r between the moment and the parameter q learned by the small network.

	$p_{\rm I}(x)$	$p_{\rm II}(x)$	r
mean	0	0	-0.018
variance	3.125	3.125	0.109
skewness	0	-5.625	0.957

[1] Ludovic Berthier, Daniele Coslovich, Andrea Ninarello, and Misaki Ozawa, "Equilibrium sampling of hard spheres up to the jamming density and beyond," Physical Review Letters **116**, 238002 (2016).

[2] Ludovic Berthier, Patrick Charbonneau, Yuliang Jin, Giorgio Parisi, Beatriz Seoane, and Francesco Zamponi, "Growing timescales and lengthscales characterizing vibrations of amorphous solids," Proceedings of the National Academy of Sciences **113**, 8397–8401 (2016).

[3] Yuliang Jin, Pierfrancesco Urbani, Francesco Zamponi, and Hajime Yoshino, "A stability-reversibility map unifies elasticity, plasticity, yielding, and jamming in hard sphere glasses," Science Advances **4**, eaat6387 (2018).

[4] Yuliang Jin and Hajime Yoshino, "Exploring the complex free-energy landscape of the simplest glass by rheology," Nature Communications **8**, 14935 (2017).

[5] Beatriz Seoane and Francesco Zamponi, "Spin-glass-like aging in colloidal and granular glasses," Soft Matter **14**, 5222–5234 (2018).

[6] Tomáš Boublík, "Hard-sphere equation of state," The Journal of Chemical Physics **53**, 471–472 (1970).

[7] Stephen F Swallen, Katherine Traynor, Robert J McMahon, MD Ediger, and Thomas E Mates, "Stable glass transformation to supercooled liquid via surface-initiated growth front," Physical Review Letters **102**, 065503 (2009).

[8] Robert L Jack and Ludovic Berthier, "The melting of stable glasses is governed by nucleation-and-growth dynamics," The Journal of Chemical Physics **144**, 244506 (2016).

[9] Elisabeth Gardner, "Spin glasses with p-spin interactions," Nuclear Physics B **257**, 747–765 (1985).

[10] Patrick Charbonneau, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, "Fractal free energy landscapes in structural glasses," Nature Communications **5**, 3725 (2014).

[11] Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, *Theory of Simple Glasses: Exact Solutions in Infinite Dimensions* (Cambridge University Press, 2020).

[12] Qinyi Liao and Ludovic Berthier, "Hierarchical landscape of hard disk glasses," Physical Review X **9**, 011049 (2019).

[13] Antoine Seguin and Olivier Dauchot, "Experimental evidence of the gardner phase in a granular glass," Physical Review Letters **117**, 228001 (2016).

[14] Korbinian Geirhos, Peter Lunkenheimer, and Alois Loidl, "Johari-goldstein relaxation far below t g: Experimental evidence for the gardner transition in structural glasses?" Physical Review Letters **120**, 085705 (2018).

[15] Andrew P. Hammond and Eric I. Corwin, "Experimental observation of the marginal glass phase in a colloidal glass," Proceedings of the National Academy of Sciences **117**, 5714–5718 (2020).

[16] Patrick Charbonneau and Sho Yaida, "Nontrivial critical fixed point for replica-symmetry-breaking transitions," Physical Review Letters **118**, 215701 (2017).

[17] Huaping Li, Yuliang Jin, Ying Jiang, and Jeff ZY Chen, "Determining the nonequilibrium criticality of a gardner transition via a hybrid study of molecular simulations and machine learning," Proceedings of the National Academy of Sciences **118**, e2017392118 (2021).

[18] Boris D Lubachevsky and Frank H Stillinger, "Geometric properties of random disk packings," Journal of Statistical Physics **60**, 561–583 (1990).

[19] Monica Skoge, Aleksandar Donev, Frank H Stillinger, and Salvatore Torquato, "Packing hyperspheres in high-dimensional euclidean spaces," Physical Review E **74**, 041127 (2006).

[20] Dongkyu Kim and Dong-Hee Kim, "Smallest neural network to learn the ising criticality," Physical Review E **98**, 022138 (2018).

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations (2015).

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of Machine Learning Research **15**, 1929–1958 (2014).

[23] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," Journal of Machine Learning Research **9**, 2579–2605 (2008).

[24] Kurt Binder, "Theory of first-order phase transitions," Reports on progress in physics **50**, 783 (1987).

[25] Misaki Ozawa, Ludovic Berthier, Giulio Biroli, Alberto Rosso, and Gilles Tarjus, "Random critical point separates brittle and ductile yielding transitions in amorphous materials," Proceedings of the National Academy of Sciences **115**, 6656–6661 (2018).

[26] Michael Gofman, Joan Adler, Amnon Aharony, A Brooks Harris, and Moshe Schwartz, "Evidence for two exponent scaling in the random field ising model," Physical Review Letters **71**, 1569 (1993).
[27] Thomas Nattermann, "Theory of the random field ising model," in *Spin glasses and random fields* (World Scientific, 1998) pp. 277–298.
[28] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (2016) pp. 265–283.
[29] Juan Carrasquilla and Roger G. Melko, "Machine learning phases of matter," Nature Physics **13**, 431 (2017).
[30] M. E. J. Newman and G. T. Barkema, *Monte Carlo methods in statistical physics* (Clarendon Press, Oxford, 1999).
[31] Lars Onsager, "Crystal statistics. i. a two-dimensional model with an order-disorder transition," Physical Review **65**, 117 (1944).
[32] Michael E Fisher and Robert J Burford, "Theory of critical-point scattering and correlations. i. the ising model," Physical Review **156**, 583 (1967).
[33] AL Talapov and HWJ Blöte, "The magnetization of the 3d ising model," Journal of Physics A: Mathematical and General **29**, 5727 (1996).
[34] Massimo Campostrini, Andrea Pelissetto, Paolo Rossi, and Ettore Vicari, "25th-order high-temperature expansion results for three-dimensional ising-like systems on the simple-cubic lattice," Physical Review E **65**, 066127 (2002).
[35] Rui Zhang, Bin Wei, Dong Zhang, Jia-Ji Zhu, and Kai Chang, "Few-shot machine learning in the three-dimensional ising model," Physical Review B **99**, 094427 (2019).
[36] Patrick Charbonneau, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi, "Glass and jamming transitions: From exact results to finite-dimensional descriptions," Annual Review of Condensed Matter Physics **8**, 265–288 (2017).