Machine-learning iterative calculation of entropy for physical systems

Amit Nir (עמית ניר)[a,b], Eran Sela (ערן סלע)[a], Roy Beck[a,b,c,1], and Yohai Bar-Sinai (יוחאי בר-סיני)[a,b,d,1]

[a]The School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel; [b]The Center for Physics and Chemistry of Living Systems, Tel Aviv University, Tel Aviv 69978, Israel; [c]The Center for Nanoscience and Nanotechnology, Tel Aviv University, Tel Aviv 69978, Israel; and [d]Google Research, Tel-Aviv 67891, Israel

Edited by Pablo G. Debenedetti, Princeton University, Princeton, NJ, and approved October 16, 2020 (received for review August 11, 2020)

Characterizing the entropy of a system is a crucial, and often computationally costly, step in understanding its thermodynamics. It plays a key role in the study of phase transitions, pattern formation, protein folding, and more. Current methods for entropy estimation suffer from a high computational cost, lack of generality, or inaccuracy and inability to treat complex, strongly interacting systems. In this paper, we present a method, termed machine-learning iterative calculation of entropy (MICE), for calculating the entropy by iteratively dividing the system into smaller subsystems and estimating the mutual information between each pair of halves. The estimation is performed with a recently proposed machine-learning algorithm which works with arbitrary network architectures that can be chosen to fit the structure and symmetries of the system at hand. We show that our method can calculate the entropy of various systems, both thermal and athermal, with state-of-the-art accuracy. Specifically, we study various classical spin systems and identify the jamming point of a bidisperse mixture of soft disks. Finally, we suggest that besides its role in estimating the entropy, the mutual information itself can provide an insightful diagnostic tool in the study of physical systems.

entropy estimation | mutual information | machine learning | jamming

Entropy is a fundamental concept of statistical physics whose computation is crucial for a proper description of many phenomena, including phase transitions (1–3), pattern formation (4), self-assembly (5–7), protein folding (8–10), and many more. In the physical sciences, entropy is typically interpreted as quantifying the amount of disorder of a system or the level of quantum entanglement. Entropy is also a fundamental concept in other fields of thought—statistical learning, economy, inference, and cryptography, among others (11). There it is used to quantify the complexity of statistical distributions. Mathematically, entropy is defined as

$$S = -k_{\mathrm{B}} \sum_i p_i \log p_i, \qquad [1]$$

where p_i is the probability that the system is in the ith microstate, and k_{B} is the Boltzmann constant. For convenience, in what follows we work with units where $k_{\mathrm{B}} = 1$.

Analytic calculation of the entropy is achievable only for simple, weakly interacting systems. Experimentally, the entropy can be obtained, for example, by measuring the temperature (T) dependence of the specific heat down to low temperatures (12). Computationally, for all but the simplest systems, a direct calculation of the entropy is computationally infeasible, as it requires computational resources that grow exponentially with system size (13, 14). For example, a classical numerical approach involves integrating the specific heat, which is inferred from energy fluctuations, down to low temperatures (12). This method is computationally costly and can suffer from inaccuracies for systems with numerous ground states at low T. Other methods estimate directly the free energy (15) or embrace additional knowledge on the system, for example from experiment, to reduce the entropic contribution to a manageable computational task (16).

Recently, we and others have shown that using compression algorithms one can compute, to a good approximation, the entropy of fairly complex systems (8, 17, 18). This method is based on Kolmogorov's theorem that states that the optimal compression of data drawn from a distribution is bounded by the distribution's entropy (19, 20). The compression-based methods capitalize on decades of research in computer science, which resulted in fast and efficient compression algorithms, such as the Lempel–Ziv algorithm or variants of it (21) which are widely available. However, these algorithms treat data as a one-dimensional (1D) discrete string, and manipulating higher-dimensional data into a 1D structure results in information loss. For example, it was recently demonstrated that compression-based algorithms misestimate the entropy of systems with long-range correlations and fail to capture delicate transitions in complex systems (17).

Here, we introduce a generic approach which we term machine-learning iterative calculation of entropy (MICE). Our method improves on existing methods in a number of ways: First, it provides state-of-the-art accuracy. Second, it is scalable, in the sense that its computational cost grows logarithmically with system size. Third, it provides estimations of the actual entropy, with physical units, without additive or multiplicative corrections and with no fitting parameters. Fourth, since the underlying computations are performed with artificial neural nets, MICE can be naturally applied to various physical systems by adjusting the network architecture, rather than the digital representation of the system (e.g., flattening high-dimensional systems to one-dimensional byte arrays as in refs. 8, 17, and 18). Finally, it can be applied to both discrete and continuous distributions.

Below we test MICE on several canonical systems: the Ising model on both square and triangular lattices, the XY model

Significance

The calculation of entropy of a physical system is a fundamental step in learning its thermodynamic behavior. However, current methods to compute the entropy are often system specific and computationally costly. Here, we propose a method that is efficient, accurate, and general for computing the entropy of arbitrary physical systems. Our method is based on computing the mutual information between subsystems within the studied system, using a convolutional neural network. This iterative procedure provides accurate entropy evaluation for systems in and out of equilibrium.

Author contributions: R.B. and Y.B.-S. designed research; A.N., E.S., R.B., and Y.B.-S. performed research; A.N., R.B., and Y.B.-S. analyzed data; and A.N., E.S., R.B., and Y.B.-S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: ybarsinai@gmail.com or roy@tauex.tau.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2017042117/-/DCSupplemental.

First published November 19, 2020.

with and without an external magnetic field (H), and an athermal system of bidisperse soft disks in two dimensions (2D). We show that our approach provides state-of-the-art accuracy and provides insightful information about the physics as a byproduct.

The Method

Entropy and Mutual Information. In thermodynamics, entropy is considered to be an extensive quantity, i.e., a quantity that scales linearly with system size. This is only approximately true. In fact, the entropy is strictly subextensive. The quantity that measures the subextensiveness is called mutual information.

To be precise, the mutual information ($\mathcal{M}$) between two random variables A, B is defined by the relation (11)

$$S(A,B) = S(A) + S(B) - \mathcal{M}(A,B), \qquad [2]$$

where $S(A), S(B)$ are the entropies of A and B, respectively, and $S(A,B)$ is their joint entropy. It is easy to show that $\mathcal{M}(A,B)$ is strictly nonnegative (11). Therefore, if we think of A and B as two halves of a thermodynamical system, this equation tells us that the entropy of the joint system is smaller than the sum of the entropies of its components.

Eq. **2** is the basic relation on which our method relies. It allows calculation of the entropy of a large system by estimating the entropy of each of its halves and the mutual information between them. Since the computational cost of estimating the entropy grows exponentially with the system size, the latter might be a significantly easier problem than the former.

With this in mind, consider a large physical system X_0, of volume V_0, which we divide into two equal halves. If we deal with translationally invariant systems, as we assume for the remainder of this work, the two halves are statistically indistinguishable, and we denote both of them by X_1 (Fig. 1*A*). With this notation, Eq. **2** takes the form

$$S(X_0) = 2S(X_1) - \mathcal{M}(X_1), \qquad [3]$$

where $\mathcal{M}(X_k)$ is a shorthand notation for the mutual information between two neighboring subsystems X_k. Each of the halves can be further divided into two statistically indistinguishable halves, and this process can be iterated arbitrarily many times. After m iterations, we find that

$$s(X_0) \equiv \frac{S(X_0)}{V} = s_m - \frac{1}{2}\sum_{k=1}^{m} \frac{\mathcal{M}(X_k)}{V_k}, \qquad [4]$$

where $V_k = 2^{-k} V_0$ is the volume (or area in 2D) of the kth subsystem, and $s_m \equiv S(X_m)/V_m$ is the specific entropy of the mth subsystem.

Eq. **4** decomposes the entropy S into contributions from different length scales. At very short scales, the iteration should be carried out only until X_k becomes small enough that its entropy can by directly calculated, either by brute-force enumeration or by using other methods. Since V_k decreases exponentially with k, the number of needed iterations is logarithmic in the system size. In many cases the actual value of the first term in the right-hand side of Eq. **4**, i.e., the entropy of the smallest subsystem, is an uninteresting additive constant with no physical significance and can be ignored.

In summary, the crux of our method is replacing the problem of evaluating the entropy by that of calculating the mutual information between subsystems of varying sizes (Fig. 1*A*). It is left to understand how to actually calculate the mutual information, which is the topic of the next section.

Estimating the Mutual Information. Recently, Belghazi et al. (25) proposed a method to calculate the mutual information between high-dimensional random variables with neural networks. Their idea is simple and elegant: Following a theorem by Donsker and Varadhan (26), the mutual information between two variables, A and B, can be expressed as a solution to a maximization problem:

$$\mathcal{M} = \sup_{\theta\in\Theta} \left[\langle \mathcal{F}_\theta(A,B) \rangle_{P_{A,B}} - \log \left\langle e^{\mathcal{F}_\theta(A,B)} \right\rangle_{P_{A\times B}} \right]. \qquad [5]$$

Here, $\mathcal{F}_\Theta : A \times B \to \mathbb{R}$ is a family of functions parameterized by a vector of parameters θ, $P_{A,B}$ is the joint distribution of A and B, and $P_{A\times B}$ is product of their marginal distributions. In our case, since A and B are subsystems of a bigger system, $\langle\cdot\rangle_{P_{A,B}}$ means averaging over samples of A and B taken from the same sample of the bigger system, while $\langle\cdot\rangle_{P_{A\times B}}$ means averaging over samples of A and B taken independently. Heuristically, the reason that this representation works is that the mutual information measures how much the joint distribution differs from the product of marginal distributions. In fact, $\mathcal{M}(A,B)$ equals the Kubleck–Leibler divergence between these two distributions (11).

While there is much to be said about Eq. **5**, for the purpose of this work it suffices to note that it reduces the problem of

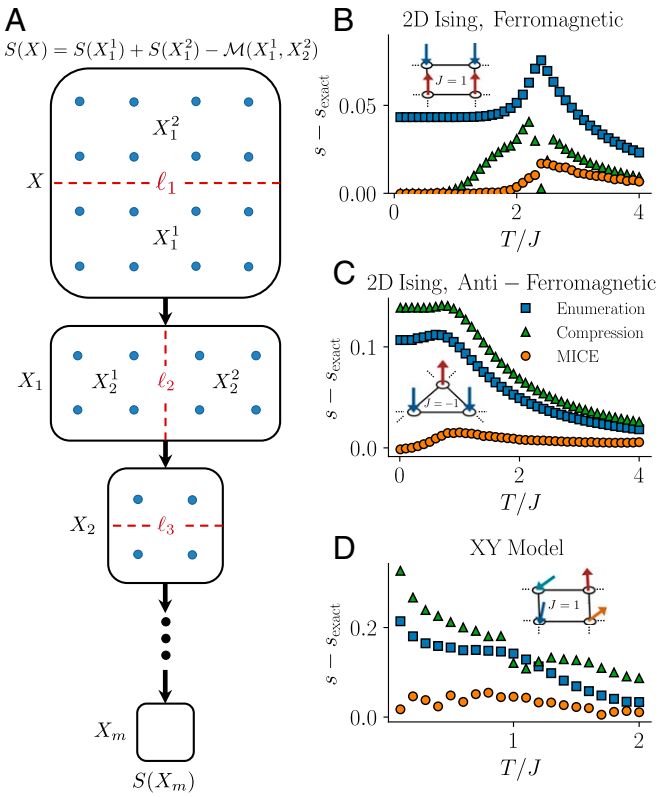

Fig. 1. (*A*) Schematic illustration of MICE. By dividing the system into smaller subsystems and calculating the mutual information between them we reconstruct the entropy of the whole system. The entropy of the smallest subsystem is calculated directly by enumeration. Dashed red lines mark the length of interface (ℓ_i) between two subsystems in the ith iteration. (*B–D*) The difference between MICE estimations of s and known benchmarks. Note that the units are chosen such that $k_B = 1$. We present three estimation methods: MICE, naive extrapolation from a system of 16 spins (main text), and a compression-based method (8). MICE shows superior performance in all cases. *B–D* show results for (*B*) the ferromagnetic Ising model on a square lattice, (*C*) the antiferromagnetic Ising model on a triangular lattice, and (*D*) the XY model on a square lattice. In *B* and *C* we benchmark against known analytical results for infinite systems (22, 23), respectively. In *D*, we benchmark against the calculation of ref. 24.

calculating $\mathcal{M}$ to an optimization problem, which naturally suggests the prospect of using artificial neural networks (ANNs) to parameterize the function $\mathcal{F}_\theta$. This is the core idea of Belghazi et al. (25), which we adopt. In machine-learning language, Eq. **5** is taken to be the loss function of the network.

For the complete implementation details see *SI Appendix*, section 1. In broader strokes, the process is as follows: First, using standard methods, a sizable dataset of samples of the system is produced. Then, for each size of subsystem pair we generate two datasets: one in which the two subsystems are taken from the same larger sample (the "joint" dataset) and another in which each subsystem is sampled independently (the "product" dataset). Then, each of the datasets is fed to an ANN, the two averages in Eq. **5** are calculated, and the weights of the ANN are updated to maximize the loss. This process is repeated until the loss stops improving and $\mathcal{M}$ saturates. We found the exponential moving average useful to reduce noise when estimating $\mathcal{M}$ over the final training epochs. Finally, $\mathcal{M}$ is calculated from the trained ANN by averaging Eq. **5** over a separate dataset, different from the one used to train the network.

Results

To demonstrate the performance and versatility of MICE we chose four systems representing different classes of collective behavior: 1) the 2D ferromagnetic Ising model on a square lattice with coupling constant $J=1$, a canonical example of a system with a second-order phase transition; 2) the antiferromagnetic Ising model on a triangular lattice ($J=-1$), a canonical example of a frustrated system with degenerate ground states (27); 3) the continuous XY model on a square lattice, which has a continuous symmetry and features a topological phase transition (27); and 4) finally, an athermal system of a bidisperse mixture of elastic particles which undergoes a jamming transition when its density is increased above a certain threshold (28). For all these systems our method achieves state-of-the-art performance. In addition, in some cases it provides physical insights about the structure and scales of the emergent behavior, as discussed below.

Spin Models. All three spin models were simulated for a system of 64×64 spins with periodic boundary conditions. The distribution was sampled using standard, well-established methods: The Ising models were simulated using Metropolis Monte Carlo simulations as in ref. 8 and the XY model was simulated using the Wolff algorithm as in ref. 29 (*SI Appendix*, section 2).

Lattice systems can naturally be represented as 2D arrays (the triangular lattice can be represented on a square lattice with diagonal interactions) (27). This allows the usage of one of the most successful ANN architectures to parameterize $\mathcal{F}$ of Eq. **5**: feedforward convolutional nets (30, 31). We use one to three convolutional layers, each of 8 to 16 filters of size 3×3, followed by two fully connected layers, using RELU (rectified linear unit) activation, implemented in PyTorch (32). Complete details about the hyperparameters for each model are given in *SI Appendix*, section 1. We calculate $\mathcal{M}$ between subsystems of sizes ranging from a pair of spins to system size. The entropy of a single spin was trivially calculated using brute-force enumeration.

The deviations of our entropy estimations from known results (22–24) are shown in Fig. 1 *B–D*. In all three cases we see impressive quantitative agreement, to a fraction of k_B, with no fitting parameters. We also benchmark our results against the recently proposed compression-based algorithm (8). Relying on highly optimized code and treating the system as effectively 1D, the compression-based algorithm is obviously much faster, about one to two orders of magnitude in terms of runtime. However, while it captures the trend, it offers substantially inferior accuracy in some cases. For example, the low-temperature behavior of the antiferromagnetic Ising model (Fig. 1*C*) is governed by a thermodynamic number of ground states with long-range correlations. There, the error of MICE is smaller by an order of magnitude than that of the compression algorithm method.

It is insightful to compare the performance against another very efficient, albeit naive, estimation of s—calculating s for a small collection of spins by direct enumeration and neglecting the mutual information (i.e., the last term in Eq. **4**). In other words, this is assuming that S is extensive. This estimation, which we refer to as "naive extrapolation," provides only slightly worse accuracy than the compression method, as seen in Fig. 1. In all cases, MICE provides the most accurate calculation with a maximal error of $0.06\ k_\mathrm{B}$ per spin for all of the systems and across all temperatures. In *SI Appendix*, section 3 we also use MICE to estimate the heat capacity, showing it outperforms the standard method based on energy fluctuations, since the latter is hard to sample at low temperatures or near a phase transition.

As presented above, our method requires training an ANN for every temperature. This is computationally costly. For example, a single training run for calculating $\mathcal{M}$ between two 64×32 systems of the ferromagnetic Ising model takes several minutes on a standard personal computer. If we were to generate all points in Fig. 1 in this method, the computation time would reach 1 to 2 d. However, drastic improvements in the calculation time can be obtained by leveraging the similarity of the systems between different temperatures. This is done by using the weights (Θ in Eq. **5**) that were obtained by training for a given temperature as the initial conditions of the training process of a different temperature or size. This technique is ubiquitous in the field of machine learning, where it is called "transfer learning" (33). In our case it reduces the training time by one to two orders of magnitude. For additional information see *SI Appendix*, section 1F.

Mutual Information as a Probe. The main purpose of MICE is providing an accurate estimation of S. In addition, the byproduct of the calculation, namely the mutual information between systems at different sizes, which is essentially a decomposition of the entropy to contributions from different length scales, can be an interesting observable in its own right. Here we briefly discuss how it captures insightful aspects of the thermodynamics and can be used to assess the accuracy of the MICE against known limiting behaviors. In passing we note that the mutual information between different scales was shown to be informative in analysis of disordered systems (34, 35).

First, we look at $\mathcal{M}$ between subsystems at various sizes for the ferromagnetic Ising model on a square lattice, plotted in Fig. 2. $\mathcal{M}$ manifestly shows the phase transition (36, 37). Indeed, $d\mathcal{M}/dT$ peaks exactly at the theoretical infinite-system critical temperature $T_c=2.269J$ (Fig. 2*B*).*

In addition, the accuracy of our calculation can be corroborated against known limits at both high and low temperatures. For $T\ll T_c$, all spins essentially point in the same direction. To be precise, in the low-T limit the ground-state entropy of the whole system, or any subsystem, is exactly $\log(2)$. This implies that the mutual information between any two subsystems is also $\log(2)$ which we indeed observe for all subsystem sizes (Fig. 2*A*).

For $T\gg T_c$, the mutual information between two subsystems can be obtained by a rigorous high-T expansion. The calculation is straightforward but lengthy, and for the sake of brevity its details are given in *SI Appendix*, section 4A. However, the result is short and intuitive: The leading-order behavior at high T is

*In second-order phase transitions the entropy is continuous but its temperature derivative (which is proportional to the heat capacity) (1) diverges. Since S is a sum over $\mathcal{M}(X_i)$ (Eq. **4**), we expect $d\mathcal{M}/dT$ to diverge, rather than $\mathcal{M}$.

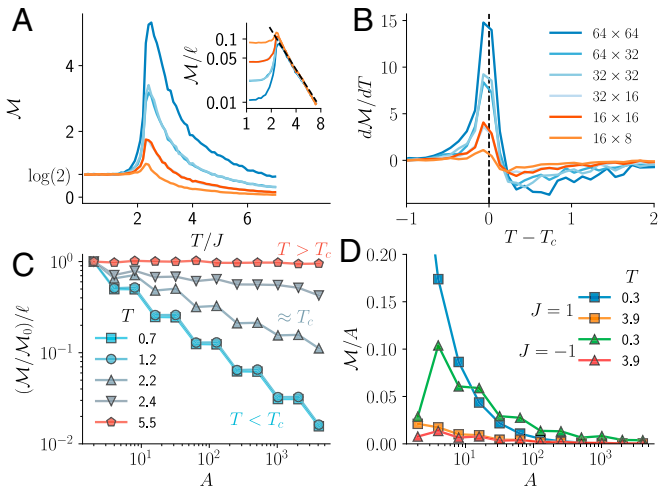

Fig. 2. Analyzing $\mathcal{M}$ for the 2D Ising model on a square lattice for different subsystem sizes. (*A*) $\mathcal{M}$ complies with two known limits: At low temperatures $\mathcal{M} = \log(2)$. At high temperatures $\mathcal{M}$ approaches the theoretical value of Eq. **6**, as shown in *Inset* (dashed line). (*B*) The derivative of the mutual information peaking at the theoretical value $T_c \approx 2.269J$ (23). (*C*) $\mathcal{M}$ normalized by the interface length for varying subsystem sizes (i.e., number of spins). For visual clarity, all curves are normalized to start at unity at zero area. (*D*) $\mathcal{M}$ per area as function of area for the ferromagnetic Ising model on a square lattice (squares) and the antiferromagnetic triangular lattice model (triangles) at various temperatures. $\mathcal{M}$ decays faster for the ferromagnetic model, as the correlation lengths are much shorter.

$$\mathcal{M} = \frac{1}{2}\frac{\ell}{T^2}, \qquad \text{for Ising model}$$
$$\mathcal{M} = \frac{1}{4}\frac{\ell}{T^2}, \qquad \text{for XY model,} \qquad \textbf{[6]}$$

where ℓ is the interface size between the subsystems, i.e., the number of spins in one system that directly interact with spins in the other. As seen in Fig. 2*A*, *Inset*, our method shows excellent agreement with this prediction, again with no fitting parameters. In passing we note that Eq. **6** is akin to the famous area law in quantum entanglement (38).

That is, when $T > T_c$, the mutual information per interface length is independent of the system size, as expected. However, for $T < T_c$ the entropy is not extensive, and $\mathcal{M}/\ell$ decays quickly with the size of the subsystem (Fig. 2*C*). This means that the summands in Eq. **4**, which are $\mathcal{M}$ normalized by the 2D volume (i.e., area), decay quickly for large subsystems. This is visualized in Fig. 2*D*. Fig. 2*D* also shows that in the antiferromagnetic model the summands decay more slowly, which is expected since it features long-range correlations.

Next, in Fig. 3 we examine the entropy and the mutual information in the XY model. At high temperatures $\mathcal{M}$ decays as described in Eq. **6**. Below the critical temperature, the famous Kosterlitz–Thouless transition temperature $T_{KT} = 0.8J$, $\mathcal{M}$ approaches a T-independent plateau for $H \neq 0$ and diverges logarithmically when $H = 0$. This divergence is due to the continuous degeneracy of the XY model, which is lifted in the presence of an external field. In the transition between these limits, $\mathcal{M}$ features a pronounced peak, which becomes smaller and shifts to higher temperatures with increasing H (Fig. 3*C*).

This rich behavior of $\mathcal{M}$ can be understood in simple terms. The high-temperature behavior is accurately described by Eq. **6**, which is a further corroboration of our method (Fig. 3*B*). The low-temperature behavior can be understood, much like in the case of the Ising model, in terms of collective behavior. For $H \neq 0$ and $T < T_{KT}$ all spins are mostly aligned with the field, even if it is relatively small, because of the broken symmetry. In this case, spins fluctuate mildly around their ground state and a harmonic approximation can be made. Within the harmonic approximation the mutual information, $\mathcal{M}_h$ (the subscript h stands for harmonic), can be obtained analytically in terms of block determinants of the Hamiltonian, a derivation which is given in detail in *SI Appendix*, section 4B. The results of this calculation are presented in Fig. 3*C* and show good quantitative agreement.

Finally, we remark that the generic behavior of $\mathcal{M}$—a T-independent plateau at low T followed by a peak and a power-law decay at large T—is also present in very small systems. In fact, even a system of two spins behaves in a qualitatively similar way, although the transition temperatures between the regimes are quite different due to the collective behavior of the spins (Fig. 3*D* and *SI Appendix*, section 5).

A Continuous, Out-of-Equilibrium System

One of the main advantages of MICE is that it is very versatile in terms of the systems it can operate on. As long as a well-defined distribution exists and samples can be drawn from it, and as long as the system can be digitally represented in a manner compatible with ANNs, MICE should be, at least potentially, applicable. In particular, the scheme presented above can be applied to out-of-equilibrium systems, whose entropy calculation is a challenge both technically and conceptually (8, 15, 17, 18, 39, 40). Clearly, the result of MICE will be an estimate of the entropy defined in Eq. **1**, which is the information-theoretic definition of entropy. Relating the result to other thermodynamic properties would depend on the details of the system, which is always the case in calculating thermodynamic properties of out-of-equilibrium systems.

Jammed solids are a prominent class of out-of-equilibrium systems whose entropy plays a crucial role in their dynamics (41). In these systems the entropy, which stems from steric interactions, is geometric in nature and measures the number of ways the

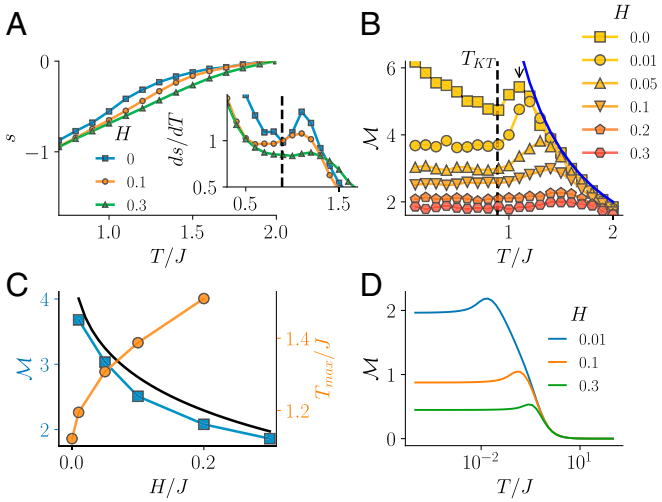

Fig. 3. Analysis of the XY model under the external field (H) using MICE. (*A*) Entropy as a function of temperature for various external fields. *Inset* shows ds/dT, and T_{KT} is marked with a dashed line. (*B*) Mutual information between two systems of size 32×16 spins, for varying fields. The arrow marks the peak in $\mathcal{M}(H = 0)$ at $T_{\max}$. The blue line is the high-temperature limit, Eq. **6**. (*C*) Two features of the curves in *B* are replotted: the low-T plateau value (evaluated at $T = 0.1J$), compared to the analytically calculated values at $T \to 0$ in the harmonic approximation, $\mathcal{M}_h$ (black line). $T_{\max}$ is plotted in orange circles. (*D*) Exact numerical calculation of $\mathcal{M}$ between two isolated spins for varying H, showing qualitatively similar behavior to that in *B* (although note that the temperature axis is logarithmic, unlike in *B*).

system's constituents can be ordered in space without overlap. When this depends sensitively on the density, jamming occurs. The jamming transition is also important as it is thought that understanding it would guide us in understanding one of the most important open problems in condensed-matter physics—the glass transition, which is also intimately related to entropic effects (41–43).

As a representative example, we study here a bidisperse mixture of soft disks. This system exhibits a jamming transition at high densities (44). Several works have attempted to identify the jamming transition of this system, using dynamic properties such as the jamming length scale or the effective viscosity (45) and using static properties such as pair correlations or fraction of jammed particles (44, 45). Recently, Zu et al. (17) tried to measure the entropic signature of the jamming transition and have shown that compression-based methods have failed to do so. The authors of ref. 17 have generously shared their dataset with us, to test our method on, which we do below.

The system is an equimolar bidisperse system of disks with one-sided harmonic interactions (Fig. 4*A*). The simulation is performed in a finite box with periodic boundary conditions. The area density of the particles, ϕ, is a control parameter which is changed by changing the number of particles, N. Further details about the simulation are given in *SI Appendix*, section 6. The system is expected to undergo a jamming transition at $\phi_J \approx 0.841$ (28, 45).

There are a few differences between this system and the spin models discussed above. First, it is not a lattice system with discrete states. Rather, here the state space is continuous, parameterized by the positions of the particles. This requires a careful treatment of the discretization scheme. The choice of discretization scheme, and specifically the spatial resolution of discretization, affects the results in a nontrivial manner. Finally, in the analysis of the spin models we employed MICE on subsystems of all sizes, between one spin and the whole system. However, the soft disk systems are so large that doing so will be both impractical and unnecessary (adequate resolution requires $\sim 3 \times 10^6$ pixels, as discussed below). Before describing the results, we briefly discuss how these challenges are resolved, since they are common to many physical systems of interest, both in and out of equilibrium.

Continuous Systems (Differential Entropy). Since the system is continuous, the summation in Eq. **1** should be replaced by integration:

$$\tilde{S} = -\int p(x) \log p(x) dx. \qquad [7]$$

This definition is known as differential entropy. Note that $\log p(x)$ is ill defined since it depends on the choice of units of x in a nonmultiplicative manner.

This nonmultiplicative component, which depends logarithmically on the length unit, is fundamentally related to the fact that the digital representation of the system is discrete and thus the differential entropy of Eq. **7** differs from the discrete entropy of Eq. **1** by a factor that diverges logarithmically with the resolution of the discretization. For a detailed derivation see *SI Appendix*, section 7.

Moreover, we also show there that, quite conveniently, the representation of S in terms of Eq. **4** offers a well-defined way to remove this divergence. While $\tilde{S}$ of a continuous system depends logarithmically on the resolution, $\mathcal{M}$ becomes independent of it

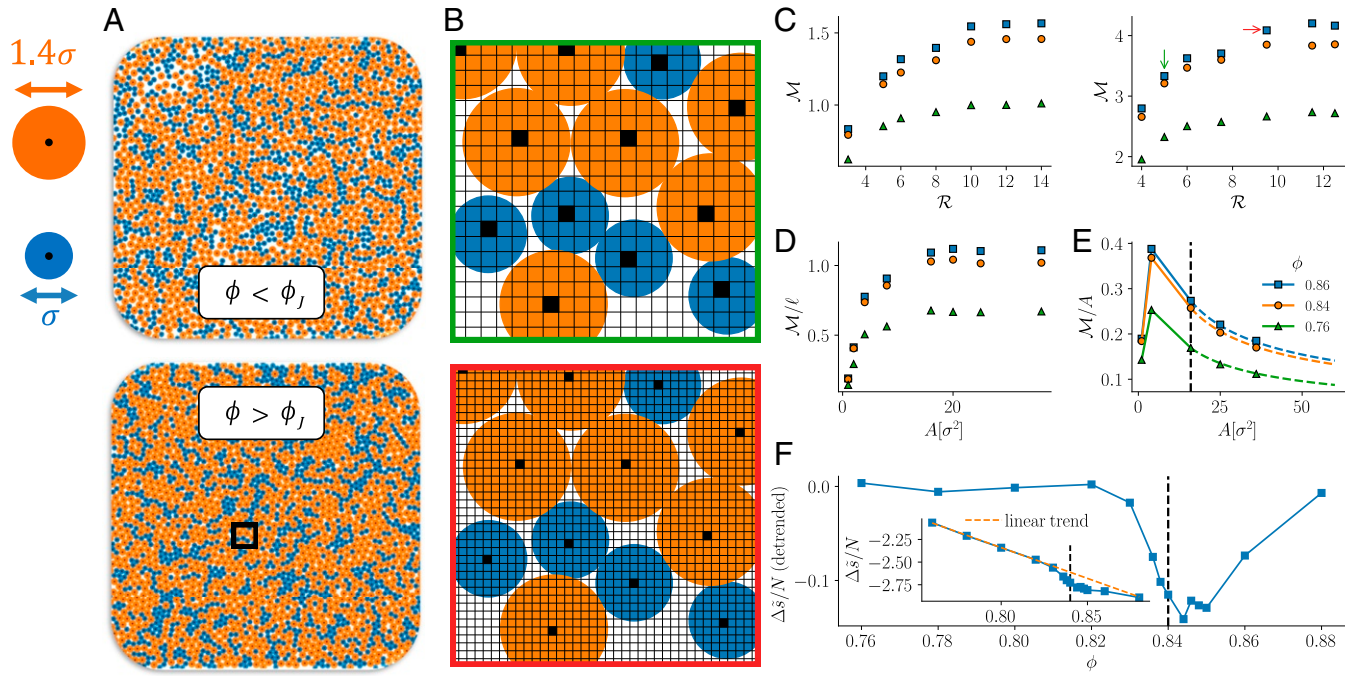

Fig. 4. (*A*) Snapshots from the bidisperse mixture simulation below and above the jamming transition density (ϕ_J). (*B*) A blowup of the marked region in *A*. We discretized the system (colored circles) as Boolean 2D images (black and white pixels). *Top* and *Bottom* show a spatial resolution of $\mathcal{R} = 5$ and $\mathcal{R} = 9.5$, respectively. The pixels are the input to MICE. (*C*) The effect of discretizing with various resolutions ($\mathcal{R}$) and various densities: $\mathcal{M}$ between two subsystems of size $2\sigma \times 1\sigma$ (*Left*) and $4\sigma \times 2\sigma$ (*Right*). At high resolutions, $\mathcal{M}$ becomes independent of $\mathcal{R}$. Green and red arrows indicate the resolutions represented in *B*, *Top* and *Bottom*, respectively. Different markers correspond to different densities; see key in *E*. (*D*) $\mathcal{M}/\ell$ as a function of the area of the subsystem (*A*) at various densities; see key in *E*. For large enough ℓ, $\mathcal{M}$ becomes linear in ℓ. (*E*) $\mathcal{M}/A$ as a function *A* at various densities. $\mathcal{M}$ becomes negligible for large subsystems. The dashed colored lines represent the extrapolation of Eq. **9**, based on the subsystem at the size represented by the black dashed line. (*F*) The density dependence of the excess entropy. *Inset* shows the results of MICE (blue) and the linear trend of $\tilde{s}/N$ at low densities (dashed orange line). For visual clarity, the linear trend in ϕ is subtracted in the main panel. The dashed black line represents the theoretical jamming transition point.

in the limit of very fine resolution. In fact, the necessary resolution is such that no physically relevant information is lost by the discretization, i.e., when all continuous configurations that map to the same discrete representation are equiprobable.

Therefore, when we estimate S according to Eq. **4**, we can avoid the logarithmic divergence simply by omitting the first term in the right-hand side. That is, in what follows we do not present $\tilde{s}$ but rather

$$\Delta\tilde{s} \equiv \tilde{s} - \frac{S(X_m)}{V_m} = -\sum_{k=1}^{m} \frac{\mathcal{M}(X_k)}{2\,V_k}. \qquad [8]$$

As a side note, we remark that the omitted term, $S(X_m)/V_m$, is simply the entropy density of the smallest subsystem. It corresponds to the entropy of an "ideal gas" composed of copies of the smallest subsystem. Subtracting the entropy of an ideal gas is common in entropy calculations of thermodynamic systems (17, 39). The result of the subtraction is commonly referred to as "excess entropy."

Discretization. Since convolutional ANNs show state-of-the-art capabilities in extracting information from images, we discretize phase space by mapping a state of the system to a 2D image, whose pixels are black if they contain a center of a particle (Fig. 4*B*).[†] The spatial resolution of the image is a hyperparameter of our method. We measure the resolution with the dimensionless number $\mathcal{R} = \sigma/p$, where p is the spatial extent of a pixel and σ is the diameter of the smaller disk. Based on the discussion above, we expect the estimation of $\mathcal{M}$ to converge to a constant value when $\mathcal{R}$ is increased. This is indeed the case, as demonstrated in Fig. 4*C*. In what follows, we use $\mathcal{R} = 10$, for which $\mathcal{M}$ is converged. We note that in terms of resources, the computational cost of discretizing the system is negligible compared to simulating the system or training the ANN. In addition, as shown below, the ANN does not have to be applied on the whole system, so a fine discretization does not lead to a memory bottleneck, at least not in 2D.

Extrapolating the Mutual Information. The resolution required for convergence necessitates $\sim 10^6$ pixels to discretize the whole system. Feeding such a large image to an ANN might be possible, but requires unreasonable computational resources for the task at hand. Luckily, this is not necessary.

As discussed above, for large enough subsystems, that is, scales much larger than the longest correlation length of the system, we expect $\mathcal{M}$ to grow linearly with the interface length (Fig. 2*C*). In precise terms, we expect

$$\mathcal{M}(X_k) = \frac{\ell_k}{\ell_n}\mathcal{M}(X_n). \qquad [9]$$

If we assume this is obeyed for all systems larger than X_k, this relation can be used to replace the summands in Eq. **4**, and the summation can be done analytically without calculations on subsystems larger than X_k. Fig. 4*D* shows that this happens for subsystems of length $\sim 4\sigma$. In Fig. 4*E* we show that Eq. **9**, based on the values of $\mathcal{M}$ for this size, quantitatively reproduces the values of the summands of Eq. **4** for sizes larger than 4σ, i.e., a 2D volume of $A = 16\sigma^2$.

Results. We are now in position to calculate the entropy of the whole system for various densities. Assuming that Eq. **9** is satisfied for $n > m$, Eq. **4** can be analytically summed, yielding (*SI Appendix*, section 8)

$$s = s(x_m) - 2\frac{\mathcal{M}(X_m)}{V_m}. \qquad [10]$$

Fig. 4 *F*, *Inset* shows $\Delta\tilde{s}/N$ as a function of ϕ. It is seen that at low densities $\Delta\tilde{s}$ depends roughly linearly on the density (dashed orange line). To emphasize the phase transition, in the main panel of Fig. 4*F* we plot the same data with this linear trend subtracted. The change in the behavior of $\Delta\tilde{s}$ around the expected jamming point is evident. Importantly, we remind the reader that compression-based entropy estimations were less successful in showing this transition (section 3.5 of ref. 17). A more detailed comparison with the results of ref. 17 is given in *SI Appendix*, section 9.

Discussion and Conclusion

Machine-learning algorithms in general, and neural networks in particular, offer an effective tool to identify patterns in high-dimensional data with complex correlation structure. We have shown that these capabilities can be leveraged to tackle another important challenge—computing the entropy of physical systems.

The crux of the method is mapping the problem of entropy calculation to an iterative process of mutual information estimation. By doing so we were able to estimate the entropy of canonical statistical physics problems, both discrete and continuous, both in and out of equilibrium, outperforming compression-based entropy estimation methods. Finally, we demonstrated that MICE naturally allows us to decompose the entropy into contributions from different scales, providing an insightful diagnostic for the thermodynamics of physical systems.

We surmise that MICE could be a promising tool for the study of many important systems, such as the configurational entropy of amorphous solids (46), the entropy crisis of glassy systems (42), entropy of active matter (40), and more. The main limit of the proposed method would depend on the minimal system size for which Eq. **9** applies, which determines the largest input for which an ANN should be trained. This is the dominant factor in the computational cost of our method. In addition, we believe that with adequate modifications MICE could be used on quantum systems, for which the mutual information is fundamentally related to entanglement of quantum states (47). A relevant direction could be the extraction of entropy from quantum Monte Carlo simulations. These directions will be explored in future works.

Data Availability. All study data are included in this article and *SI Appendix*.

ACKNOWLEDGMENTS. We thank Daan Frenkel, Mengjie Zu, and Arunkumar Bupathy for fruitful discussions and for generously sharing their data and code. In addition, we thank Yuval Binyamini, Yakov Kantor, Haim Diamant, Gil Ariel, and Amit Moscovich-Eiger for fruitful discussions. We acknowledge support by the Israel Science Foundation (550/15, 154/19), the United States–Israel Binational Science Foundation (201696), and Army Research Office (W911NF-20-1-0013). Y.B.-S. also thanks his mother.

[†]Technically, pixels are black if they contain a center of one or more particles, although this never happens in the resolutions we work with.

1. M. Kardar, *Statistical Physics of Fields* (Cambridge University Press, 2007).
2. P. G. De Gennes, J. Prost, *The Physics of Liquid Crystals* (Oxford University Press, 1993), vol. 83.
3. D. Frenkel, Entropy-driven phase transitions. *Phys. Stat. Mech. Appl.* **263**, 26–38 (1999).
4. M. C. Cross, P. C. Hohenberg, Pattern formation outside of equilibrium. *Rev. Mod. Phys.* **65**, 851–1112 (1993).
5. R. Asor, O. Ben-nun Shaul, A. Oppenheim, U. Raviv, Crystallization, reentrant melting, and resolubilization of virus nanoparticles. *ACS Nano* **11**, 9814–9824 (2017).

6. Y. S. Cho *et al.*, Self-organization of bidisperse colloids in water droplets. *J. Am. Chem. Soc.* **127**, 15968–15975 (2005).
7. A. Donev *et al.*, Improving the density of jammed disordered packings using ellipsoids. *Science* **303**, 990–993 (2004).
8. R. Avinery, M. Kornreich, R. Beck, Universal and accessible entropy estimation using a compression algorithm. *Phys. Rev. Lett.* **123**, 178102 (2019).
9. M. C. Baxa, E. J. Haddadian, J. M. Jumper, K. F. Freed, T. R. Sosnick, Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15396–15401 (2014).
10. G. P. Brady, K. A. Sharp, Entropy in protein folding and in protein–protein interactions. *Curr. Opin. Struct. Biol.* **7**, 215–221 (1997).
11. D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
12. C. Kittel, H. Kroemer, *Thermal Physics* (American Association of Physics Teachers, 1998).
13. D. Frenkel, Simulations: The dark side. *Eur. Phys. J. Plus* **128**, 10 (2013).
14. N. Hansen, W. F. Van Gunsteren, Practical aspects of free-energy calculations: A review. *J. Chem. Theor. Comput.* **10**, 2632–2647 (2014).
15. C. Jarzynski, Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **78**, 2690 (1997).
16. S. Piana, K. Lindorff-Larsen, D. E. Shaw, Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17845–17850 (2012).
17. M. Zu, A. Bupathy, D. Frenkel, S. Sastry, Information density, structure and entropy in equilibrium and non-equilibrium systems. *J. Stat. Mech. Theor. Exp.* **2020**, 023204 (2020).
18. S. Martiniani, P. M. Chaikin, D. Levine, Quantifying hidden order out of equilibrium. *Phys. Rev. X* **9**, 011031 (2019).
19. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).
20. A. Kolmogorov, New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. *Doklady Russian Acad. Sci.* **119**, 861–864 (1958).
21. J. Ziv, A. Lempel, A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **23**, 337–343 (1977).
22. G. H. Wannier, Antiferromagnetism. The triangular Ising net. *Phys. Rev.* **79**, 357–364 (1950).
23. L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Phys. Rev.* **65**, 117–149 (1944).
24. J. F. Yu, Z. Y. Xie, T. Xiang, Two-dimensional classical XY model by HOTRG *Phys. Rev. E.* **89**, 013308. (2014).
25. I. Belghazi *et al.*, "Mine: Mutual information neural estimation" in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (PMLR, 2018), vol. 80, pp. 531–540.
26. M. D. Donsker, S. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. IV. *Commun. Pure Appl. Math.* **36**, 183–212 (1983).
27. D. P. Landau, K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2014).
28. C. S. O'Hern, L. E. Silbert, A. J. Liu, S. R. Nagel, Jamming at zero temperature and zero applied stress: The epitome of disorder. *Phys. Rev. E* **68**, 011306 (2003).
29. J. Kent-Dobias, J. P. Sethna, Cluster representations and the Wolff algorithm in arbitrary external fields. *Phys. Rev. E* **98**, 063306 (2018).
30. A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2012), pp. 1097–1105.
31. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
32. A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library" in *Advances in Neural Information Processing Systems 32*, H. Wallach *et al.*, Eds. (Curran Associates, Inc., 2019), pp. 8024–8035.
33. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
34. P. Ronhovde, Z. Nussinov, Multiresolution community detection for megascale networks by information-based replica correlations. *Phys. Rev. E* **80**, 016109 (2009).
35. Z. Nussinov *et al.*, "Inference of hidden structures in complex physical systems by multi-scale clustering" in *Information Science for Materials Discovery and Design*, T. Lookman, F. J. Alexander, K. Rajan, Eds. (Springer, 2016), pp. 115–138.
36. J. Wilms, M. Troyer, F. Verstraete, Mutual information in classical spin models. *J. Stat. Mech. Theory Exp.* **2011**, P10011 (2011).
37. J. Iaconis, S. Inglis, A. B. Kallin, R. G. Melko, Detecting classical phase transitions with Renyi mutual information. *Phys. Rev. B* **87**, 195134 (2013).
38. M. M. Wolf, F. Verstraete, M. B. Hastings, J. I. Cirac, Area laws in quantum systems: Mutual information and correlations. *Phys. Rev. Lett.* **100**, 070502 (2008).
39. G. Ariel, H. Diamant, Inferring entropy from structure. *Phys. Rev. E.* **102**, 022110 (2020).
40. C. Nardini *et al.*, Entropy production in field theories without time-reversal symmetry: Quantifying the non-equilibrium character of active matter. *Phys. Rev. X* **7**, 021007 (2017).
41. A. J. Liu, S. R. Nagel, The jamming transition and the marginally jammed solid. *Annu. Rev. Condens. Matter Phys.* **1**, 347–369 (2010).
42. A. Cavagna, Supercooled liquids for pedestrians. *Phys. Rep.* **476**, 51–124 (2009).
43. R. Monasson, Structural glass transition and the entropy of the metastable states. *Phys. Rev. Lett.* **75**, 2847–2850 (1995).
44. D. Koeze, D. Vågberg, B. Tjoa, B. Tighe, Mapping the jamming transition of bidisperse mixtures. *EPL* **113**, 54001 (2016).
45. D. Vågberg, D. Valdez-Balderas, M. A. Moore, P. Olsson, S. Teitel, Finite-size scaling at the jamming transition: Corrections to scaling and the correlation-length critical exponent. *Phys. Rev.* **83**, 030303 (2011).
46. E. Bouchbinder, J. Langer, I. Procaccia, Athermal shear-transformation-zone theory of amorphous plastic deformation. I. Basic principles. *Phys. Rev.* **75**, 036107 (2007).
47. L. Amico, R. Fazio, A. Osterloh, V. Vedral, Entanglement in many-body systems. *Rev. Mod. Phys.* **80**, 517 (2008).

Supplementary Information for

Machine-learning Iterative Calculation of Entropy for Physical Systems

Amit Nir, Eran Sela Roy Beck, and Yohai Bar Sinai

Corresponding Authors
Roy Beck, roy@tauex.tau.ac.il
Yohai Bar-Sinai, ybarsinai@gmail.com

This PDF file includes:

Supporting Information Text

1. *MICE* implementation details

A. Data preprocessing and augmentation. Input features were normalized between values -1 and 1. For the soft disk system, this means that empty pixels are set to -1 and pixels which contain a particle center are set to 1. Since all our systems are symmetric under reflections, we performed data augmentation by reflecting both vertically and horizontally. In the data of the XY model without an external field, a global random phase was also used for data augmentation. In addition, due to translational symmetry one can sample subsystems anywhere within the larger system. Combining all these, a single snapshot of 64x64 spins can generate about 15,000 training samples.

B. Network Architecture. Our method was implemented using the PyTorch library (1). For subsystems of input size larger than 32×32 we used three convolutional layers with 16 filters of size 3×3 each and a rectified linear unit (ReLU) activation. For smaller subsystems, we use only two convolutional layers. For subsystems of size 4×4 or smaller, only one convolutional layer is used. The convolutional layers are followed by two fully connected layers, with $\frac{k}{2}$ and 1 output neurons, respectively, where k is the number of output neurons in the last convolutional layer. The batch size for training was 128.

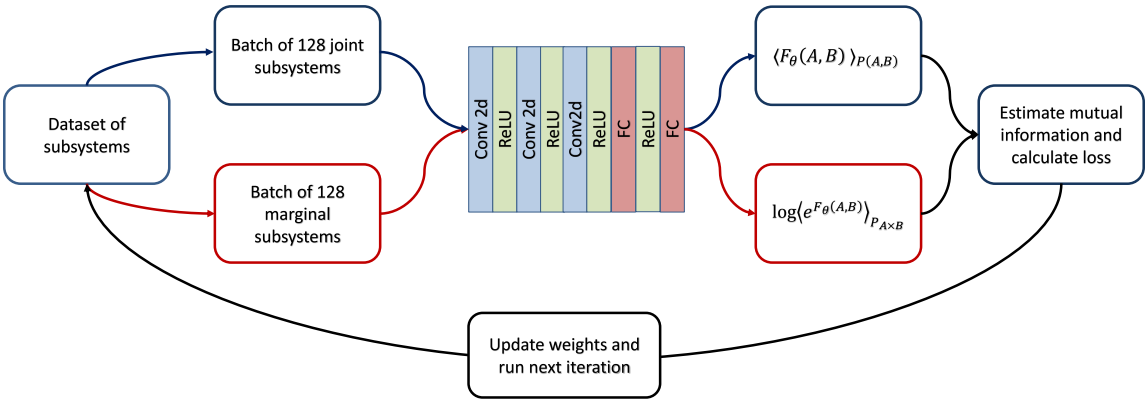

Fig. S1. The flow of *MICE*. The simulations are used to generate a marginal dataset and a joint dataset (see main text for definition) dataset. The specific architecture of the ANN shown here was used for subsystem pars larger than 32×32. Smaller subsystems used $1-2$ convolutional layers, as detailed in Sec. 1B.

C. Noise Reduction. The output of the neural network (ANN) is averaged over the marginal and joint distributions to give a bound on the mutual information (see Eq. (5) of the main text). As the network learning process progresses, the bound becomes tighter. However, at each iteration the averaging is performed over a small batch of 128 samples. Therefore, the network's output is extremely noisy. To smooth the results we use a moving exponential average:

$$\langle \mathcal{M} \rangle_{i+1} = \langle \mathcal{M} \rangle_i + \alpha \Big(\mathcal{M}_{i+1} - \langle \mathcal{M} \rangle_i \Big). \qquad [S1]$$

where $\mathcal{M}_j$ is the output of the network after j optimization iterations, and $\langle \mathcal{M} \rangle_i$ is our averaged estimation after i iterations, see Fig. S2. Throughout the manuscript we used the exponential averaging with $\alpha = 10^{-3}$.

D. Validation. For estimating $\mathcal{M}$ we implemented the standard scheme of using a validation set. Two independent datasets with ratio of 80-20 were created before training. The network was trained over the large (training dataset), and the training phase was terminated when the $\mathcal{M}$ estimation on the training set stopped increasing. $\mathcal{M}$ was estimated over the independent validation set as well, and this value was used for subsequent calculations. By comparing the estimation of $\mathcal{M}$ over the training and validation sets, one can verify that the network did not overfit the data.

E. Dataset size. For the spin models we used a dataset of 5000 samples of a 64×64 system. An exception is the XY model with an external field where we used 2000 simulations. For the soft disk system we used a set of 100 simulations. In general, for the systems considered in the manuscript we typically needed about $10^4 - 10^5$ samples (obtained from the the $\sim 10^3$ actual samples by data augmentation, see A above) to achieve reasonable convergence.

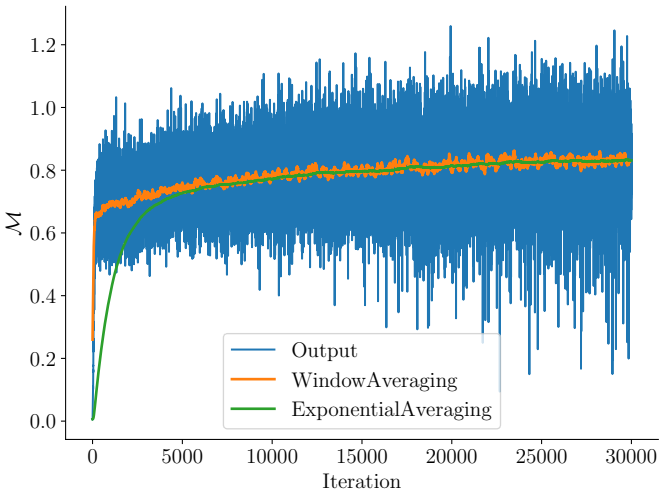

Fig. S2. Noise reduction. The raw output of the network (blue) and an exponential average with $\alpha = 10^{-3}$ (green) are shown during a typical training loop. In addition, we demonstrate another noise reduction method, used by the original authors of (2), a moving average with a window size of 100 iterations (orange).

F. Transfer Learning. When initiating the network weights at random the resulting estimation of $\mathcal{M}$ is roughly zero. During training it increases until a plateau is reached. For our choice of hyperparameters this can take a few thousand training iterations, cf. Fig. S3. This process can be expedited if the network is not initialized at a random initial condition but instead the weights of a network that was trained for a different system are used, a technique called "Transfer Learning"

This can be done in a number of ways - e.g. transfer learning across temperatures or the sizes of the subsystem. In the main text we only used transfer learning across different temperatures. In Fig. S3 we show the result of training with and without transfer learning, which can reduce training time by 1-2 orders of magnitude. We note that transfer learning works better when we first train on high T and then transfer to lower T, similar to simulated annealing strategy in optimization.

We note that transfer learning across subsystem size is slightly more tricky since the input size to the ANN is different. One simple-minded way to overcome this is to pad the smaller subsystems with zeros, which gives reasonable results, cf. Fig. S3B. This is an interesting direction for future research, which we did not further explore. Transfer learning across subsystem size was not used in the main text.

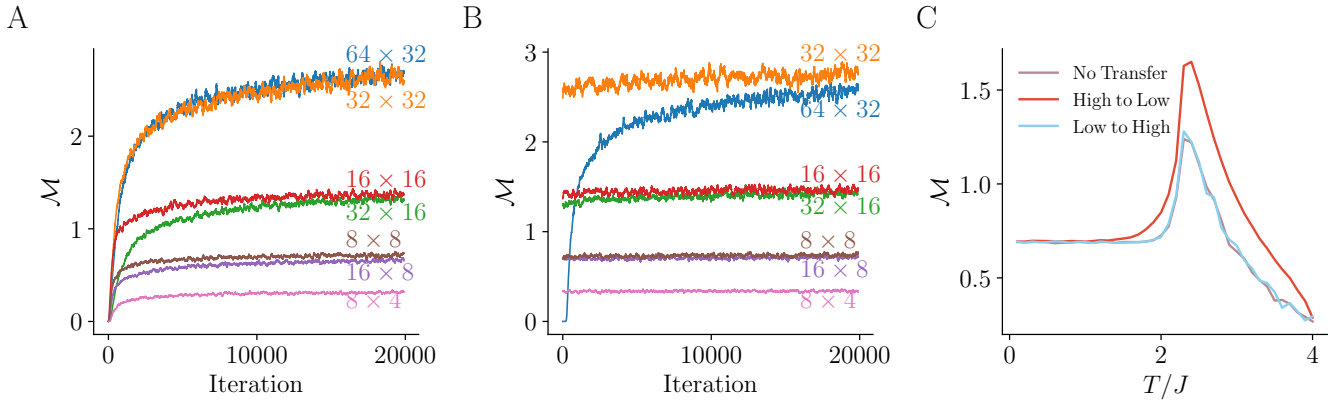

Fig. S3. Effect of transfer learning. (A)-(B) Learning process as function of iteration for various subsystem sizes. (A) Without transfer learning (i.e. random initial weights for each ANN). (B) With transfer learning from one subsystem size to another. $\mathcal{M}$ plateaus at the same level with or without transfer learning, but the number of iterations needed to reach the plateau changes drastically. (C) $\mathcal{M}$ as function of temperature for 16×16 subsystem of the 2d ferromagnetic Ising model. Adding transfer learning from high to low temperature improves the results dramatically while transfer learning in the opposite direction is not effective. All trainings were done for 3000 iterations at every temperature.

2. Spin Model Simulations

Sampling the distribution of the Ising systems was preformed using standard Monte-Carlo sampling.

Sampling the distribution of the XY simulation was performed using the Wolff algorithm implemented in the c++ library provided in Ref. (3). To generate uncorrelated samples the mean cluster size at each temperature, c, was calculated and the simulation was sampled every $2/c$ steps. That is, each spin was flipped twice on average between two subsequent samples at all temperatures.

3. Specific Heat Estimation Using *MICE*

A standard method of estimating the entropy of thermodynamic systems is to integrate the specific heat from low temperatures. This method relies on the relations

$$c_V = T\frac{dS}{dT}\ , \qquad \text{and} \tag{S2}$$

$$c_V = \frac{\langle E^2\rangle - \langle E\rangle^2}{T^2}\ , \tag{S3}$$

where c_V is the heat capacity, E is the energy and $\langle\cdot\rangle$ denotes thermal averaging. $S(T)$ can be calculated using Eq. (S3) and integrating the energy fluctuations from zero temperature to T.

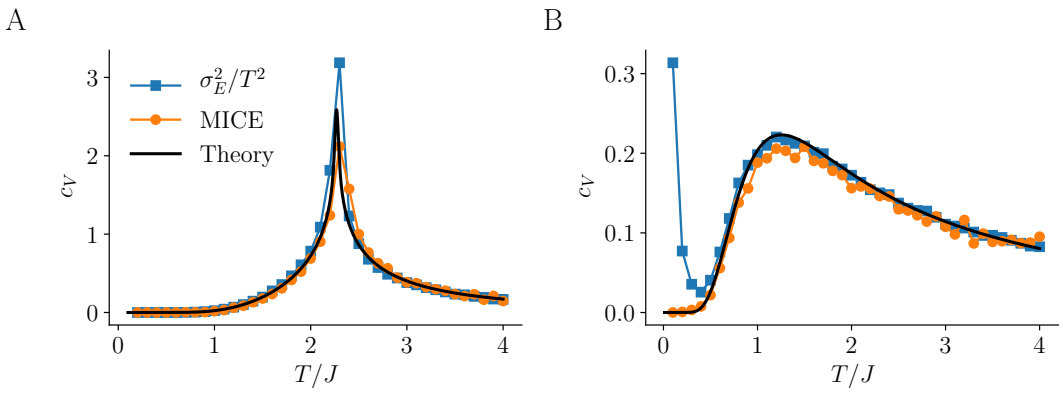

Fig. S4. Estimating c_V using energy fluctuation estimation, (Eq. (S3), blue), and *MICE* (Eq. (S2), orange), compared to the theoretical value (black). (A) 2D Ising model (B) 2D anti-ferromagnetic Ising model.

Alternatively, one take the inverse direction: using the estimation of S, as calculated by *MICE*, together with Eq. (S2) to estimate c_V. In Fig. S4 we compare this estimation of c_V (orange) to the estimation of c_V using energy fluctuations (Eq. (S3), blue). It is evident that the energy fluctuations overestimate c_V in the 2D ferromagnetic Ising model near the phase transition, and at low temperatures in the anti-ferromagnetic triangular lattice, which features high degeneracy of low energy states.

4. Analytic Calculation of $\mathcal{M}$ at high and low temperature limit for spin models

A. High temperature. Here we derive Eq. (6) of the main text by a rigorous high-T expansion of the partition function and marginal probabilities. Physically this expansion relies on the fact that at high temperatures correlations become local. At high temperature we explicitly obtain the area law, $\mathcal{M}(A,B) \propto \ell$, stating that $\mathcal{M}$ is proportional to the area ℓ (or length in two dimensions) of the interface between regions A and B, rather their volume.

The mutual information between subsystems A and B, whose union is $A \cup B = X$, is defined as:

$$\mathcal{M}(A,B) = S(A) + S(B) - S(X), \tag{S4}$$

where the entropy of a subsystem A is given in terms of the marginal probability:

$$S(A) = -\sum_{\alpha} P_A(\alpha)\log P_A(\alpha). \tag{S5}$$

Here, α labels microstates of A. For the spins models, the microstates are given in terms of the configurations of spins $z_a, a \in A$. We assume that the entire system X under consideration is described by an equilibrium distribution:

$$P_X(\boldsymbol{z}) = \frac{e^{-\beta E(\boldsymbol{z})}}{Z}\ , \qquad Z = \sum_{\{z_i=\pm 1\}} e^{-\beta E(\boldsymbol{z})}\ . \tag{S6}$$

Here and in what follows boldface letters (e.g. $\boldsymbol{z}$) denote vectors. The marginal distribution of subsystem A is obtained by tracing out the spins in its complement, $P_A = \mathrm{Tr}_B\, P_X$.

We proceed by an explicit evaluation of $\mathcal{M}$ at high temperature for the Ising model:

$$E_{\text{Ising}}(\boldsymbol{z}) = -J\sum_{\langle i,j\rangle} z_i z_j - H\sum_i z_i\ , \qquad z_i = \pm 1\ . \tag{S7}$$

The expansion of the partition function in powers of β up to second order is

$$Z = \sum_{\{z_i=\pm 1\}} \left(1 - \beta E(\boldsymbol{z}) + \frac{1}{2}\beta^2 E(\boldsymbol{z})^2 + \dots\right) = 2^N + \frac{1}{2}\beta^2 \left[J^2 \left(\sum_{\langle i,j\rangle} 1\right) 2^N + H^2 \left(\sum_i 1\right) 2^N\right] + \mathcal{O}\left(\beta^3\right)$$
$$= 2^N \left[1 + \frac{1}{2}\beta^2 \left(J^2 N_{\text{links}} + H^2 N\right)\right] + \mathcal{O}\left(\beta^3\right), \quad \text{[S8]}$$

where $\sum_{\langle i,j\rangle} 1 = N_{\text{links}}$ is the total number of links and $\sum_i 1 = N$ is the number of sites. In what follows we omit the external field (H) for clarity and conciseness of presentation, and only mention its effect in the end result.

Next, we perform a high temperature expansion up to order β^2 of the marginal probability

$$P_A(\boldsymbol{z}_A) = \sum_{\boldsymbol{z}_b} P(\boldsymbol{z}_A, \boldsymbol{z}_B) = \sum_{\boldsymbol{z}_B} \frac{1 - \beta E(\boldsymbol{z}_A, \boldsymbol{z}_B) + \frac{1}{2}\beta^2 E^2(\boldsymbol{z}_A, \boldsymbol{z}_B)}{Z} + \mathcal{O}\left(\beta^3\right) . \quad \text{[S9]}$$

Here $\boldsymbol{z}_A$ is fixed and spins $\boldsymbol{z}_B$ in B act like an environment for A and are traced out.

Tracing out the first order term in the numerator of Eq. (S9) annihilates any terms that involve at least one spin in B. Therefore, the first order term yields simply the energy of subsystem A,

$$E_A(\boldsymbol{z}_A) = -J \sum_{\langle a,a'\rangle \in A} z_a z_{a'} . \quad \text{[S10]}$$

Tracing over the second order term in the numerator of Eq. (S9) involves a double sum over neighbors $\sum_{\langle ij\rangle} \sum_{\langle i'j'\rangle} z_i z_j z_{i'} z_{j'}$. The only combinations of i, j, i', j' that are not annihilated by tracing out are:

1. $i, j, i', j' \in A$. Summation over these quadruplets yields $E_A(\boldsymbol{z}_A)^2$.
2. $i, j, i', j' \in B$. Summation over these quadruplets yields $J^2 N^B_{\text{links}}$ where N^B_{links} is the number of links in B.
3. $i \in A, j \in B$ and $\langle i, j\rangle = \langle i', j'\rangle$. Summation over these quadruplets yields $J^2\ell$ where ℓ is the number of links between A and B.
4. In the triangular lattice there's a fourth option where there exist two distinct spins $i, i' \in A$ which have a common neighbor $j \in B$. The sum over such pairs of spins in A is denoted $\sum'_{aa'}$.

Therefore, the numerator of Eq. (S9) yields, to second order in β,

$$P_A(\boldsymbol{z}_a) = 2^{N_B} \frac{1 - \beta E_A(\boldsymbol{z}_A) + \frac{1}{2}\beta^2 \left(E_A(\boldsymbol{z}_a)^2 + J^2 N^B_{\text{links}} + J^2\ell + J^2 \sum'_{aa'} z_a z_{a'}\right)}{Z} + \mathcal{O}\left(\beta^3\right) . \quad \text{[S11]}$$

Proceeding with the expansion, plugging in Eq. (S8) and using $N^A_{\text{links}} + N^B_{\text{links}} + \ell = N_{\text{links}}$, we get

$$P_A(\boldsymbol{z}_a) = \frac{1 - \beta E_A(\boldsymbol{z}_a) + \frac{1}{2}\beta^2 E_A(\boldsymbol{z}_a)^2 + \frac{1}{2}\beta^2 J^2 \sum'_{aa'} z_a z_{a'}}{Z_A} + \mathcal{O}\left(\beta^3\right) \text{ , with} \quad \text{[S12]}$$

$$Z_A = 2^{N^A} \left[1 + \frac{1}{2}\beta^2 J^2 N^A_{\text{links}}\right] + \mathcal{O}\left(\beta^3\right) . \quad \text{[S13]}$$

Eq. (S12) has the form of a Boltzmann distribution (note the similarity of Eq. (S13) to Eq. (S8)) derived from the Hamiltonian E_A, with extra couplings generated by the tracing out of B (the last term in the numerator of Eq. (S12)). A straightforward but tedious calculation, which will not be detailed here, shows that up to quadratic order in β these couplings do not affect the entropy. That is, while they do clearly affect the probabilities of individual states (as explicitly shown in Eq. (S12)) their combined contribution to S cancels out to quadratic order when summed over all states. Therefore, as far as entropy calculations are concerned we can write

$$P_A(\boldsymbol{z}_A) \approx \frac{e^{-\beta E_A(S_a)}}{Z_A} + \mathcal{O}\left(\beta^3\right) , \qquad Z_A = \sum_{\boldsymbol{z}_A} e^{-\beta E_A(\boldsymbol{z}_A)} + \mathcal{O}\left(\beta^3\right) , \quad \text{[S14]}$$

and treat P_A as a standard Boltzmann distribution, for which we have $S = \partial_T (T \log Z)$. Plugging this into Eq. (S4) gives

$$\mathcal{M}(A, B) = \partial_T \left(T \log \frac{Z_A Z_B}{Z_X}\right) + \mathcal{O}\left(\beta^3\right) . \quad \text{[S15]}$$

Physically the numerator $(Z_A Z_B)$ is the partition function for all the spins in X without the interactions through links connecting A and B. Finally, using Eq. (S8) and Eq. (S13) we obtain the result

$$\mathcal{M}_{\text{Ising}}(A, B) = \frac{1}{2}\left(\frac{J}{T}\right)^2 \ell + \mathcal{O}\left(\beta^3\right) . \quad \text{[S16]}$$

Note that neither the sign of J nor the lattice symmetry (square versus triangular) influence the answer to order β^2 – the only relevant parameters are the number of links connecting the two subsystems ℓ and the coupling constant J. Also, up to this order the magnetic field H does not contribute. A very similar calculation leads to the same form for the XY model, with only a change in the prefactor:

$$\mathcal{M}_{\mathrm{XY}}(A,B) = \frac{1}{4}\left(\frac{J}{T}\right)^2 \ell + \mathcal{O}\left(\beta^3\right) \ . \tag{S17}$$

Both Eq. (S16) and Eq. (S17) are valid also when A and B do not compose the whole system, but are a part of a larger system.

B. Low-temperature expansion - XY model in a magnetic field. Statistical mechanics problems of continuous variables can be treated at low temperatures via an harmonic treatment of the interactions, i.e. a mapping to a system of coupled harmonic oscillators. This technique can be applied to compute $\mathcal{M}$ too (4), yielding closed-form formulas. Here we apply this method to the XY model in an external magnetic field (H) in the zero-temperature limit.

The XY model in a magnetic field is defined by the partition function

$$Z = \int_0^{2\pi} d\boldsymbol{\theta} e^{-\beta E(\boldsymbol{\theta})}, \qquad E(\boldsymbol{\theta}) = -J\sum_{\langle i,j\rangle}\cos(\theta_i-\theta_j) - H\sum_i \cos\theta_i. \tag{S18}$$

At low temperature $T \ll J, H$ the variables $\boldsymbol{\theta}$ explore only the vicinity of the minimum of the external potential $-H\cos\theta_i$, and since we consider a frustration-free lattice (square lattice), also the differences $\theta_i - \theta_j$ on neighboring links $\langle i,j\rangle$ will be located near the minima of $-J\cos(\theta_i-\theta_j)$. Performing a harmonic approximation of the overall potential we get:

$$Z_0 = \int_{-\infty}^{\infty} d\boldsymbol{\theta} e^{-\beta E_0(\boldsymbol{\theta})} \ , \qquad E_0(\boldsymbol{\theta}) = \frac{J}{2}\sum_{\langle i,j\rangle}(\theta_i-\theta_j)^2 + \frac{H}{2}\sum_i \theta_i^2 + \text{const} \ . \tag{S19}$$

Here, we extended the variables θ_i from being angles to unconstrained real numbers. Accordingly, microstates of the full system X satisfy a multivariate normal distribution

$$p(\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}\boldsymbol{\theta}^T M \boldsymbol{\theta}}}{Z_0} \ , \qquad \text{with} \qquad M_{ij} = \frac{H+zJ}{T}\delta_{ij} - \frac{J}{T}\delta_{\langle i,j\rangle} \ . \tag{S20}$$

M is the system's Hessian, a $N\times N$ matrix where N is the number of sites in the system X. Here z is the coordination number ($z=4$ for a square lattice) and $\delta_{\langle i,j\rangle}=1$ if i and j are neighbors and 0 otherwise. The entropy of a multivariate Gaussian is well known:

$$S(X) = \frac{N}{2}\log 2\pi e - \frac{1}{2}\log\det M. \tag{S21}$$

For a single spin in a magnetic field, for example, this gives $S = \log\left(\sqrt{2\pi eT/H}\right)$ which is valid as long as the variance of θ, $(T/H)^2$, is sufficiently small compared to $(2\pi)^2$.

The key object required for the calculation of the $\mathcal{M}$ is the marginal probability for a subsystem A. It is obtained by integrating $p(\boldsymbol{\theta})$ over all degrees of freedom $\boldsymbol{\theta}_B \in B$,

$$p_A(\boldsymbol{\theta}_A) = \frac{1}{Z}\int_0^{2\pi} d\boldsymbol{\theta}_B e^{-\beta E(\boldsymbol{\theta}_A,\boldsymbol{\theta}_B)}. \tag{S22}$$

To perform the Gaussian integral we decompose the matrix M as

$$M = \begin{pmatrix} M_{AA} & M_{AB} \\ M_{BA} & M_{BB} \end{pmatrix}, \tag{S23}$$

where, M_{AA} is an $N^A\times N^A$ matrix acting only on the N^A degrees of freedom in A, and similarly for M_{BB}. The off-diagonal blocks $M_{AB} = M_{BA}^T$ couple the two subsystems. Thus,

$$p_A(\boldsymbol{\theta}_A) = e^{-\frac{1}{2}\boldsymbol{\theta}_A^T M_{AA}\boldsymbol{\theta}_A}\int d\theta_B \exp\left[-\frac{1}{2}\boldsymbol{\theta}_B^T M_{BB}\boldsymbol{\theta}_B - \boldsymbol{\theta}_A^T M_{AB}\boldsymbol{\theta}_B\right] \ . \tag{S24}$$

Performing the Gaussian integral over $\boldsymbol{\theta}_B$ gives

$$P(\boldsymbol{\theta}_A) = \left((2\pi)^{N_B}\det M_{BB}\right)^{1/2}\exp\left[-\frac{1}{2}\boldsymbol{\theta}_A^T M_{AA}\boldsymbol{\theta}_A - \frac{1}{2}\boldsymbol{\theta}_A^T\left(M_{AB}M_{BB}^{-1}M_{BA}\right)\boldsymbol{\theta}_A\right] \ . \tag{S25}$$

Since the marginal distribution is also Gaussian, its entropy is given by Eq. (S21), with the effective Hessian (covariance matrix) of A given by Eq. (S24),

$$M_A^{\mathrm{eff}} = M_{AA} - M_{AB}M_{BB}^{-1}M_{BA} \ . \tag{S26}$$

M_A^{eff} contains direct interactions inside A, as well as new interactions $M_{AB}M_{BB}^{-1}M_{BA}$ generated by tracing out the environment B. We thus have

$$\mathcal{M} = \frac{1}{2}\log\frac{\det M_X}{\det M_A^{\text{eff}}\det M_B^{\text{eff}}}. \tag{S27}$$

Note that this expression gives the $T \to 0$ limit of $\mathcal{M}$ and is independent of T. Finite temperature corrections are not present in the harmonic approximation and start to appear when the variance of spins becomes of order 2π and deviations from the Gaussian distribution are sampled.

For the system described in the main text $\mathcal{M}$ was computed by evaluating the determinant in Eq. (S21) numerically using the effective covariance matrix Eq. (S26).

5. $\mathcal{M}$ between two XY-spins in a magnetic field

It is instructive to contrast the result in the main text for the $\mathcal{M}$ of the XY model with that for a system consisting of only two spins. This can be calculated exactly, and is shown in Fig. S5. At high temperature $\mathcal{M}$ decreases like $\mathcal{M} \to \frac{1}{4}\left(\frac{J}{T}\right)^2$, indicated by a dashed line in the right panel, as predicted by Eq. (S17). As $T \to 0$, we can see in the central panel a logarithmic divergence with T which is cut-of when $T \approx H$.

Indeed it is easy to derive from Eqs. (S20), (S26) and (S27) the zero temperature limit of $\mathcal{M}$,

$$\lim_{T\to 0}\mathcal{M}_{\text{two spins}} = \log\frac{H+J}{\sqrt{H(H+J)}}. \tag{S28}$$

As H increases, the cutoff of the logarithmic divergence occurs at higher temperatures, and the peak thus shifts to higher temperatures. Thus,the peak itself, as well as its H-dependence features, are already present in a two-spin system.

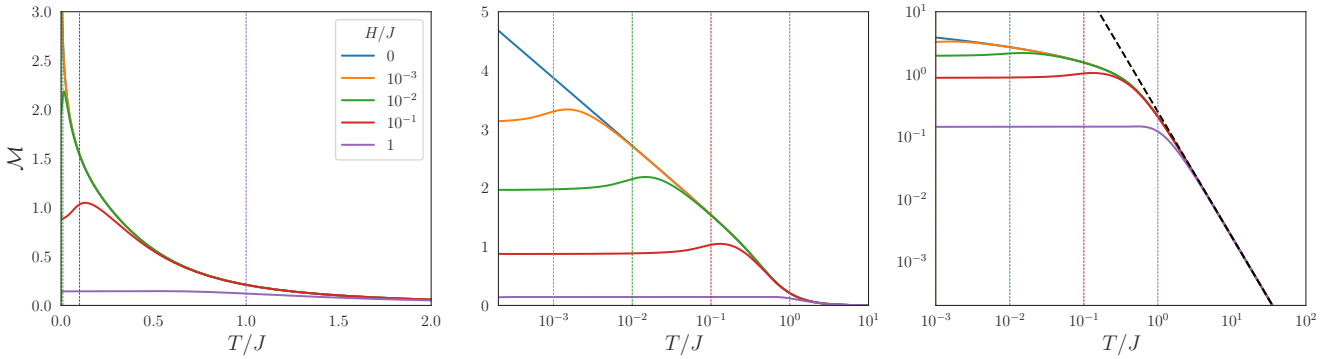

Fig. S5. Exact calculation of $\mathcal{M}$ for two XY spins ($J = 1$) in the presence of external field (H). The same data are shown in linear-linear, log-linear and log-log scales (some data of the middle panel appears also in the main text). Colored vertical dashed lines show $T = H$ with the color code corresponding to H as in the legend. The dashed black line in the right panel is the high temperature expansion limit of Eq. (S17).

6. Simulations of the soft sphere system

The system is an equimolar system of larger and smaller spheres. We choose the units such that the diameter of the smaller sphere is unity, and the radius of the larger one is 1.4. The dynamics were simulated using a fast inertial relaxation engine algorithm (5) in a square box of size 150 with periodic boundary conditions. 100 realizations were generated for each ϕ, ranging between 14,000 to 17,000 particles.

7. Discrete vs. differential entropy

As discussed in the main text, the system of bidisprese sphere is a continuous system, parameterized by a continuous vector $\boldsymbol{x} \in \mathbb{R}^{2N}$ where N is the number of particles in the system. However, the state of the system is provided to the ANN as a binary image, which is a discrete variable. Here we discuss the subtleties of comparing the discrete and continuous definitions of entropy (Eq. (1) and (7) of the main text, respectively).

Let us denote $p(\boldsymbol{x})$ the probability density of observing the configuration $\boldsymbol{x}$. The discretization is a mapping of the continuous vector $\boldsymbol{x}$ to an image $I(\boldsymbol{x})$ where I takes one of a finite set of values which we denote $I_1, I_2, \ldots$. Each I_i is associated with its pre-image Ω_i, observation probability p_i and phase-space volume v_i, defined as follows:

$$\Omega_i \equiv \{\boldsymbol{x} \mid I(\boldsymbol{x}) = I_i\}\ , \qquad p_i \equiv \int_{\Omega_i} p(\boldsymbol{x})d\boldsymbol{x}\ , \qquad v_i \equiv \int_{\Omega_i} 1\, d\boldsymbol{x}\ . \tag{S29}$$

In the limit of very fine discretization, i.e. $\max_i\{v_i\} \to 0$, and assuming $p(x)$ is not ill-behaved, the second definition can be approximated as

$$p_i \approx p(\boldsymbol{x}_i) v_i \ , \tag{S30}$$

where $\boldsymbol{x}_i$ is any point in Ω_i. This approximation is accurate when the discretization is fine enough such that p doesn't change considerably across Ω_i, i.e. when all configurations that are mapped to the same image are roughly equiprobable. When this happens, the differential entropy $\tilde{S}$ can be approximated by a Riemman sum:

$$\begin{aligned} \tilde{S} &= -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \approx -\sum_i \Big(p(\boldsymbol{x}_i) \log p(\boldsymbol{x}_i) \Big) \cdot v_i \\ &\approx -\sum_i \left(\frac{p_i}{v_i} \log \left(\frac{p_i}{v_i} \right) \right) \cdot v_i = \sum_i \left(-p_i \log p_i + p_i \log v_i \right) = S + \sum_i p_i \log v_i \ . \end{aligned} \tag{S31}$$

We see that $\tilde{S}$ differs from S by a term logarithmic in the resolution size. This term, however, cancels out when computing $\mathcal{M}$ rather than S.

To see this, let's say $\boldsymbol{x}$ and $\boldsymbol{y}$ are random variables, with the joint probability density $p(\boldsymbol{x}, \boldsymbol{y})$ and marginal densities $p^x(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y}$ and $p^y(\boldsymbol{y}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x}$. In addition, we have two discretization schemes $I^x(\boldsymbol{x})$ and $I^y(\boldsymbol{y})$ that map each observation to some finite set. We define, in analogy to Eq. (S29),

$$\begin{aligned} &\Omega_{ij} \equiv \left\{ (\boldsymbol{x}, \boldsymbol{y}) \mid I^x(\boldsymbol{x}) = I_i^x \text{ and } I^y(\boldsymbol{x}) = I_j^y \right\} \ , && p_{ij} \equiv \int_{\Omega_{ij}} p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} \ , && v_{ij} \equiv \int_{\Omega_{ij}} 1 \, d\boldsymbol{x} d\boldsymbol{y} \ , \\ &\Omega_i^x \equiv \left\{ \boldsymbol{x} \mid I^x(\boldsymbol{x}) = I_i^x \right\} \ , && p_i^x \equiv \int_{\Omega_i^x} p^x(\boldsymbol{x}) d\boldsymbol{x} \ , && v_i^x \equiv \int_{\Omega_i^x} 1 \, d\boldsymbol{x} \ , \\ &\Omega_j^y \equiv \left\{ \boldsymbol{y} \mid I^y(\boldsymbol{y}) = I_j^y \right\} \ , && p_j^y \equiv \int_{\Omega_j^y} p^y(\boldsymbol{y}) d\boldsymbol{y} \ , && v_j^y \equiv \int_{\Omega_j^y} 1 \, d\boldsymbol{y} \ . \end{aligned}$$

Eqs. (1)-(2) of the main text can be combined to represent the mutual information as

$$\mathcal{M} = \int p(\boldsymbol{x}, \boldsymbol{y}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p^x(\boldsymbol{x}) p^y(\boldsymbol{y})} \right) d\boldsymbol{x} \, d\boldsymbol{y} \tag{S32}$$

Since $v_{ij} = v_i^x v_j^y$, the analog of Eq. (S30) is

$$p_{ij} \approx p(\boldsymbol{x}_i, \boldsymbol{y}_j) v_i^x v_j^y \ , \qquad p_i^x \approx p^x(\boldsymbol{x}_i) v_i^x \ , \qquad p_j^y \approx p^y(\boldsymbol{y}_j) v_j^y \ . \tag{S33}$$

Finally, combining all the above we get

$$\mathcal{M} \approx \sum_{i,j} p(\boldsymbol{x}_i, \boldsymbol{y}_j) \log \left(\frac{p(\boldsymbol{x}_i, \boldsymbol{y}_j)}{p^x(\boldsymbol{x}_i) p^y(\boldsymbol{y}_j)} \right) v_i^x v_j^y \approx \sum_{i,j} p_{ij} \log \left(\frac{p_{ij}}{p_i^x p_j^y} \right) \ , \tag{S34}$$

which identifies with the discrete definition of $\mathcal{M}$.

As an aside, we note that Eq. (S31) has an intuitive interpretation: $\log v_i$ is exactly the entropy of a uniform distribution over Ω_i (whose probability density is $p = 1/v_i$). Therefore, the differential entropy $\tilde{S}$ measures the uncertainty (=entropy) associated with knowing in which Ω_i the observation $\boldsymbol{x}$ resides, plus the average uncertainty (=entropy) associated with knowing where does $\boldsymbol{x}_i$ resides within Ω_i. The latter cancels out when computing $\mathcal{M}$.

8. Derivation of Eq. (10) of the main text

Eq. (4) of the main text starts with a system X_0 of a given volume V_0 and looks at smaller and smaller subsystems (i.e. larger m). For the purposes of Eq. (10) of the main text we want to explore the other direction – assuming that X_0 is by itself a part of a much larger system and extrapolating from X_0 to the system size. To comply with the notation of the main text, where larger m's correspond to smaller subsystems X_m, we consider subsystems which are formally indexed by negative integers. Also, it will be useful to consider Eq. (3) of the main text normalized per unit volume. For any k we have

$$S(X_{k-1}) = 2S(X_k) - \mathcal{M}(X_k) \qquad \Rightarrow \qquad s(X_{k-1}) \equiv \frac{S(X_{k-1})}{V_{k-1}} = s(X_k) - \frac{\mathcal{M}(X_k)}{2V_k} \ , \tag{S35}$$

where we used the fact that $V_{k-1} = 2V_k$. Using this relation recursively we get

$$\begin{aligned}
s(X_{-1}) &= s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0} \\
s(X_{-2}) = s(X_{-1}) - \frac{\mathcal{M}(X_{-1})}{2V_{-1}} &= s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0} - \frac{\mathcal{M}(X_{-1})}{4V_0} \\
s(X_{-3}) = s(X_{-2}) - \frac{\mathcal{M}(X_{-2})}{2V_{-2}} &= s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0} - \frac{\mathcal{M}(X_{-1})}{4V_0} - \frac{\mathcal{M}(X_{-2})}{8V_0} \\
&\vdots \\
s(X_{-m}) &= s(X_0) - \frac{1}{2V_0} \sum_{k=0}^{m-1} \frac{\mathcal{M}(X_{-k})}{2^k}
\end{aligned} \tag{S36}$$

We now assume that for subsystems larger than X_0 the mutual information is extensive, so by Eq. (9) of the main text we have $\mathcal{M}(X_{-k}) = (\ell_{-k}/\ell_0)\mathcal{M}(X_0)$. For our choice of selecting subsystems, we also have $\ell_{-k}/\ell_0 = 2^{\lfloor \frac{k+1}{2} \rfloor}$, where $\lfloor \cdot \rfloor$ is the floor function. We assume that X_0 is a square subsystem (subsystems alternate between square and rectangular, cf. Fig. 1 of the main text). Putting all this together we get

$$S(X_{-m}) = s(X_0) - \frac{\mathcal{M}_0}{2V_0} \sum_{k=0}^{m-1} 2^{\lfloor \frac{k+1}{2} \rfloor - k} \ . \tag{S37}$$

One can easily verify that in the limit $m \to \infty$ the sum in the last equation approaches 4. We conclude that

$$s(X_{-m}) = s(X_0) - 2\frac{\mathcal{M}_0}{V_0} \ . \tag{S38}$$

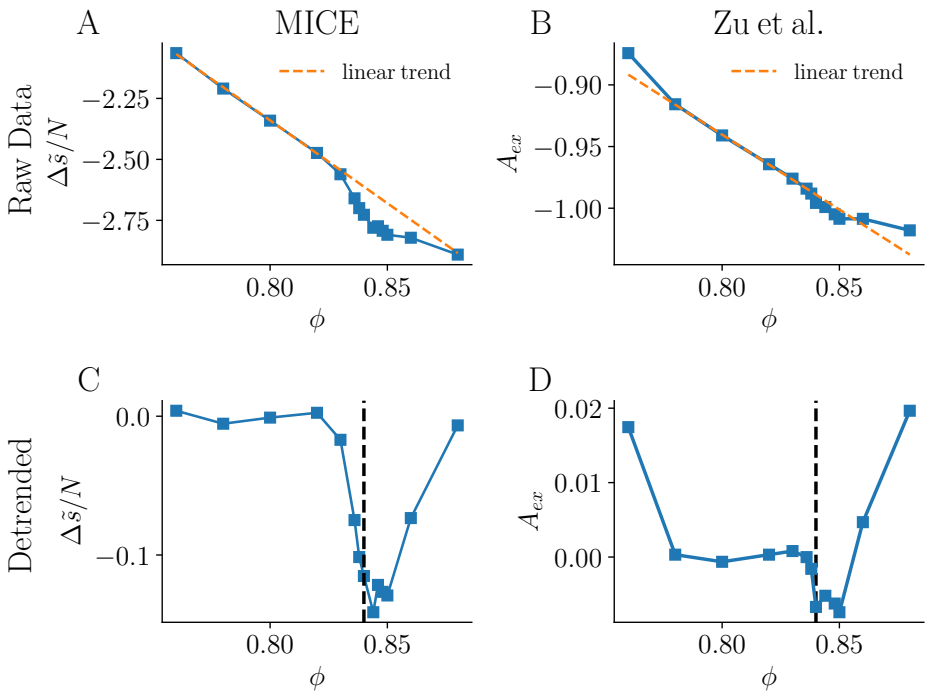

Fig. S6. Entropy estimation of the bidisperse soft sphere mixture, using two different methods, see text of Sec. 9 for a description. The dashed black line represents the theoretical jamming transition point.

9. Comparing *MICE* and the results of Zu et. al.

In the main text we claimed that *MICE* outperforms the compression method used by Zu et. al. (6) in detecting the jamming point. This was based on their statements that their Computable Information Density (CID) estimates do not show a minimum near the jamming point (see Sec. 3.5 of their paper).

In Fig. S6 we show a direct comparison between our data (left column, the same data appear as Fig. 4F and its inset in the main text) and theirs (right column, taken from Figure 6A of (6)).

The top row shows the estimation of the "excess entropy", i.e. the difference in entropy from some baseline behavior: With *MICE* this is achieved by omitting the entropic contribution of the smallest scales, cf. Eq. (8) of the main text. Zu et. al. do this by subtracting the information density of an ideal gas (see Sec. 2.3.2 of Zu et. al. (6)). These two baselines are conceptually similar but quantitatively different and therefore the absolute numbers differ somewhat between the methods. The trend, however, is informative.

To better visualize the signature of the transition, in the bottom row we plot the same data as the top row, with a linear trend (shown in dashed orange in the top row) subtracted. It is seen that the deviations from linearity are very pronounced when measured with *MICE*, but the CID estimation shows small deviations compared to the overall effect.

References

1. Paszke A, et al. (2019) Pytorch: An imperative style, high-performance deep learning library in *Advances in Neural Information Processing Systems 32*, eds. Wallach H, et al. (Curran Associates, Inc.), pp. 8024–8035.
2. Belghazi I, Rajeswar S, Baratin A, Hjelm RD, Courville AC (2018) MINE: mutual information neural estimation. *CoRR* abs/1801.04062.
3. Kent-Dobias J, Sethna JP (2018) Cluster representations and the wolff algorithm in arbitrary external fields. *Phys. Rev. E* 98(6):063306.
4. Katsinis D, Pastras G (2020) An inverse mass expansion for the mutual information in free scalar qft at finite temperature. *Journal of High Energy Physics* 2020(2):1–60.
5. Bitzek E, Koskinen P, Gähler F, Moseler M, Gumbsch P (2006) Structural relaxation made simple. *Phys. Rev. Lett.* 97(17):170201.
6. Zu M, Bupathy A, Frenkel D, Sastry S (2020) Information density, structure and entropy in equilibrium and non-equilibrium systems. *Journal of Statistical Mechanics: Theory and Experiment* 2020:023204.