Machine-learning Iterative Calculation of Entropy for Physical Systems

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

1 The School of Physics and Astronomy, Tel Aviv University
2 The Center for Nanoscience and Nanotechnology, Tel Aviv University
3 Google Research, Tel-Aviv

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 either from a high computational cost, lack of generality or inaccuracy, and inability to treat complex, strongly interacting systems. In this paper, we present a novel method, termed 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 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 spheres. Lastly, 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 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:

\[ S = -k_B \sum p_i \log p_i, \]

where \( p_i \) is the probability that the system is in the \( i \)-th microstate, and \( k_B \) is the Boltzmann constant. For convenience, in what follows we work with units where \( k_B = 1 \).

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 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 evaluate 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 manipulate higher dimensional data into a 1D structure results in information loss. For example, it was recently demonstrated that compression-based algorithms miscalculate 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 MICE: Machine-learning Iterative Calculation of Entropy. 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 [8, 17, 18]). Lastly, it can be applied to both discrete and continuous distributions.

Below we test MICE on several canonical systems, including the Ising model on both square and triangular lattices, the XY model with and without an external magnetic field \( (H) \), and an athermal system of bi-disperse soft disks in 2D. We show that our approach provides state-of-the-art accuracy, and provides insightful infor-
A

\[ S(X) = S(X_1^1) + S(X_1^2) - M(X_1^1, X_1^2) \]

B

2D Ising, Ferromagnetic

C

2D Ising, Anti-Ferromagnetic

D

XY Model

FIG. 1. (A) Schematic illustration of MICE. By dividing into smaller subsystems and calculating the mutual information between them we reconstruct the entropy of the whole system. The entropy of the minimal subsystem \( S_k \) is calculated directly. Dashed red lines mark the length of interface \( (\ell_i) \) between two subsystems in the \( i \)th iteration. (B-D) The difference between estimations of \( s \) and known benchmarks. We present three estimation methods: MICE, naive extrapolation from a system of 16 spins (see text) and a compression-based method [8]. MICE shows superior performance in all cases. The three panels show results for (B) ferromagnetic Ising model on a square lattice (C) antiferromagnetic Ising model on a triangular lattice and (D) XY model on a square lattice. In panels B, C we benchmark against known analytical results for infinite systems [22] and [23], respectively. In panel D, we benchmark against the HOTRG calculation of [24].

In summary, the crux of our method is replacing the mutual information between two neighboring subsystems \( X_k \) 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 to two equal halves. If we deal with translationally invariant systems, as we will assume for the remainder of this work, the two halves are statistically indistinguishable, and we’ll denote both of them by \( X_1 \) (Fig. 1A). With this notation, Eq. (2) takes the form

\[ S(X_0) = 2S(X_1) - \mathcal{M}(X_1) \]

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

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

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

Eq. (4) decomposes the entropy \( S \) into contributions from different length scales. At very short scales, the iteration should only be carried out until \( X_k \) becomes small enough that its entropy can be directly calculated, either by brute-force enumeration or 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 calculating the mutual information between subsystems of varying sizes, cf. Fig. 1A. It is left to understand how to actually calculate the mutual information, which is the topic of the next section.
B. Estimating the Mutual Information

Recently, Belghazi et al. proposed a method to calculate the mutual information between high dimensional random variables with neural networks [25]. 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:

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

(5)

Here, \( F_{\theta} : A \times B \to \mathbb{R} \) is a family of functions parameterized by a vector of parameters \( \theta \); \( P_{A,B} \) 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 calculating \( \mathcal{M} \) to an optimization problem, which naturally suggests the prospect of using artificial neural networks (ANNs) to parameterize the function \( F_{\theta} \). 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.

Technically, the process is as follows: first, using standard methods, a sizable dataset of samples of the system is produced. Then, for a chosen 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 exponential moving average useful to reduce noise when estimating \( \mathcal{M} \) over the final training epochs (see supplementary information appendix Sec. 1C). 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 (see supplementary information appendix Sec. 1D).

II. RESULTS

To demonstrate the performance and versatility of MICE we chose four systems representing different classes of collective behavior: (a) 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; (b) the anti-ferromagnetic Ising model on a triangular lattice (\( J = -1 \)), a canonical example of a frustrated system with degenerate ground states [27]; (c) the continuous XY model on a square lattice, which has a continuous symmetry and features a topological phase-transition [27]; (d) lastly, we analyze an athermal system of a bidisperse mixture of elastic particles which undergoes a jamming transition when density increased [28]. In all of 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.

A. Spin models

All three spin models were simulated for a system of 64 \times 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] (see supplementery information Sec. 3).

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 \( F \) of Eq. (5): feed-forward convolutional nets [30]. We use 1–3 convolutional layers, each of 8–16 filters of size \( 3 \times 3 \), followed by 2 fully connected layers, using RELU activation, implemented in PyTorch [31]. Complete details about the hyperparameters for each model are given in Sec. 1 of the supplementary information. 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.

Our entropy estimations are shown in Fig. 1 and compared to known results [22–24]. In all three cases we see impressive quantitative agreement, with no fitting parameters. We 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 1-2 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, cf. Fig. 1C, 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 the compression algorithm method.
It is insightful to compare the performance against another very efficient, albeit naïve, 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 “naïve 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 $0.06 k_B$ per spin for all the systems and across all temperatures.

As presented above, our method requires training an ANN for every temperature. This is computationally costly. For example, a single training run for calculating $M$ between two $64 \times 32$ 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 a day or two. 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 ($\Theta$ 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” [32]. In our case it reduces the training time by 1-2 orders of magnitudes. For additional information see Sec. 1F of the supplementary information.

### B. Mutual Information as a probe

The main purpose of MICE is providing an accurate estimation of $S$. In addition, the byproducts of the calculation, namely the mutual information between systems at different sizes, 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. First, we look at $M$ between subsystems at various sizes for the ferromagnetic Ising model on a square lattice, plotted in Fig. 2. $M$ shows the familiar peak at the phase transition [34], a shadow of the divergence in the thermodynamic limit [35]. Indeed, $dM/dT$ peaks exactly at the theoretical critical temperature of infinite systems ($T_c = 2.269 J$), cf. Fig. 2B.

In addition, the accuracy of our calculation can be corroborated against known limits at both high and low temperatures. For $T < T_c$, 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, cf. Fig. 2.

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 the SI. However, the result is short and intuitive: the leading order behavior at high $T$ is

$$M = \frac{1}{2} \frac{\ell}{T^2}, \quad \text{for Ising model}$$

$$M = \frac{1}{4} \frac{\ell}{T^2}, \quad \text{for XY model}$$

where $\ell$ 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. 2C, 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 [36].

Indeed, for the 2D Ising ferromagnetic model, and for $T > T_c$ the mutual information per interface length is roughly constant, as expected. However, for $T < T_c$ the entropy is not extensive, and $M/\ell$ decays quickly with the size of the subsystem (Fig. 2D). This means that the summands in Eq. (4), which are $M$ normalized by the 2D volume (i.e. area), decay quickly for large subsystems. This is visualized in Fig. 2D. The figure 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 mu-
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 an \( T \)-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 \( H \), cf. Fig. 3C.

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, cf. Fig. 3B. 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 relatively small, because of the broken symmetry. In this all spins are mostly aligned with the field, even if it is

Lastly, 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, though the transition temperatures between the regimes are quite different due to the collective behavior of the spins, cf. Fig. 3D.

### III. 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, 37, 38]. 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 [39]. In these systems the entropy, which stems from steric interactions, is geometric in nature and measures the number of ways the 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 [39–41].

As a representative example, we study here a bidisperse mixture of soft disks. This system exhibits a jamming transition at high densities [42]. 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 [43]; using static properties such as pair-correlations or fraction of jammed particles [42, 43]. Recently, Bupathi and collaborators tried to measure the entropy change of this system around the jamming transition, and have shown that a compression-based algorithm fails to extract a clear signature of it [17]. 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, cf. Fig. 4A. The
simulation is performed in a finite box with periodic boundary conditions. The area density of the particles, \( \phi \), determined by choosing the number of particles \( (N) \), is a control parameter. The system is expected to have a jamming transition at \( \phi_J \approx 0.841 \) [28, 43]. We denote the diameter of the smaller disks as \( \sigma \). Technical aspects of the calculation, as well as numerical parameters, are given in Materials and Methods.

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. Lastly, in the analysis of the spin models we employed MICE on subsystems of all sizes, between 1 spin and the whole system. However, in the jamming simulations the number of particles is so large (adequate resolution requires \( \sim 3 \times 10^6 \) pixels, as discussed below) that doing so will be both impractical and unnecessary. 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.

A. Continuous systems (differential entropy)

Since the system is continuous, the summation in Eq. (1) should be replaced by integration:

\[
\tilde{S} = - \int p(x) \log p(x) dx .
\] (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 non-multiplicative manner.

This non-multiplicative component, which depends logarithmically on the choice of units, is fundamentally related to the fact that the digital representation of the system is always discrete, and 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.

Conveniently, the representation of entropy 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, it is easy to show that \( \mathcal{M} \) becomes independent of the resolution in the limit that the resolution is fine enough [44]. 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

\[
\Delta \tilde{s} \equiv \tilde{s} \left( \frac{S(X_n)}{V_m} \right) \left( \frac{m}{2V_k} \mathcal{M}(X_k) \right)
\] (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. That is, 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, 37]. The result of the subtraction is commonly referred to as “excess entropy”.

B. 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 [45], see Fig. 4B. The spatial resolution of the image is a hyper-parameter of our method. We define the resolution as \( R = \sigma/p \), where \( p \) is the spatial extent of a pixel. Based on the discussion above, we expect the estimation of \( \mathcal{M} \) to converge to a constant value when \( R \) increased. This is indeed the case, as demonstrated in Fig. 4C. In what follows, we use a \( R = 9.5 \), for which \( \mathcal{M} \) is converged.

C. 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, cf. Fig. 2C. In precise terms, we expect

\[
\mathcal{M}(X_k) = \frac{\ell_k}{\ell_n} \mathcal{M}(X_n) .
\] (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. 4D shows that this happens for subsystems of length \( \sim 4\sigma \). In Fig. 4E 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 \( 4\sigma \), i.e. a 2D volume of \( A = 16\sigma^2 \).

D. 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 (derivation is detailed in Sec. 2 of the supporting information):

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

(10)

In Fig. 4F we present \( \Delta \tilde{s} \) as function of \( \phi \). To emphasize the phase transition, we also subtract a trivial linear dependence\[46\] of \( S \) on \( \phi \). \( \Delta \tilde{s} \) shows a clear indication for a phase transition at the expected jamming point \( \phi_J \approx 0.8415 \) (Fig. 4C). Importantly, we remind the reader that compression-based entropy estimations were unsuccessful in identifying this transition \[17\], demonstrating the effectiveness of MICE for out of equilibrium, off-lattice systems as well (see also supplementary Fig. S5).

**IV. 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 such as the 2D Ising model and the XY model. In addition, our method can be applied to out-of-equilibrium continuous systems. It successfully identified the jamming transition in a system of bi-disperse harmonic spheres, which was not possible using compression-based entropy estimation methods. Lastly, we demonstrated that MICE naturally allows 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 important open questions in condensed matter physics, such as the configurational entropy of amorphous solids \[47\], the entropy crisis of glassy systems \[40\], entropy of active matter \[38\], and more. 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 [48]. A relevant direction could be the extraction of entropy from quantum Monte Carlo simulations. These directions will be explored in future works.

a. Details about soft sphere simulations Soft spheres simulations 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 [49] 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.

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 acknowledges support by the Israel Science Foundation (550/15, 154/19), the United StatesIsrael Binational Science Foundation (201696), and ARO (W911NF-20-1-0013). YBS also thanks his mother.

[1] M. Kardar, Statistical physics of fields (Cambridge University Press, 2007).
[2] P.-G. De Gennes and J. Prost, The physics of liquid crystals, Vol. 83 (Oxford university press, 1993).
[3] D. Frenkel, Physica A: statistical mechanics and its applications 263, 26 (1999).
[4] M. C. Cross and P. C. Hohenberg, Rev. Mod. Phys. 65, 851 (1993).
[5] R. Asor, O. Ben-nun Shaul, A. Oppenheim, and U. Raviv, ACS nano 11, 9814 (2017).
[6] Y.-S. Cho, G.-R. Yi, J.-M. Lim, S.-H. Kim, V. N. Manoharan, D. J. Pine, and S.-M. Yang, Journal of the American Chemical Society 127, 15968 (2005).
[7] A. Donev, I. Cisse, D. Sachs, E. A. Variano, F. H. Stillinger, R. Connelly, S. Torquato, and P. M. Chaikin, Science 303, 990 (2004).
[8] R. Aviniry, M. Kornreich, and R. Beck, Physical Review Letters 123 (2019), 10.1103/physrevlett.123.178102.
[9] M. C. Baxa, E. J. Haddadian, J. M. Jumper, K. F. Freed, and T. R. Somnick, Proceedings of the National Academy of Sciences 111, 15396 (2014).
[10] G. P. Brady and K. A. Sharp, Current opinion in structural biology 7, 215 (1997).
[11] D. J. MacKay, Information theory, inference and learning algorithms (Cambridge University Press, 2003).
[12] C. Kittel and H. Kroemer, Thermal physics (American Association of Physics Teachers, 1998).
[13] D. Frenkel, The European Physical Journal Plus 128, 10 (2013).
[14] N. Hansen and W. F. Van Gunsteren, Journal of chemical theory and computation 10, 2632 (2014).
[15] C. Jarzynski, Physical Review Letters 78, 2690 (1997).
[16] S. Piana, K. Lindorf-Larsen, and D. E. Shaw, Proceedings of the National Academy of Sciences 109, 17845 (2012).
[17] M. Zu, A. Bupathy, D. Frenkel, and S. Sastry, Journal of Statistical Mechanics: Theory and Experiment 2020, 023204 (2020).
[18] S. Martini, P. M. Chaikin, and D. Levine, Physical Review X 9, 011031 (2019).
[19] C. E. Shannon, Bell System Technical Journal 27, 623 (1948).
[20] A. Kolmogorov, Doklady of Russian Academy of Sciences 119 (5), 861 (1958).
[21] J. Ziv and A. Lempel, IEEE Transactions on Information Theory 23, 337 (1977).
[22] G. H. Wannier, Phys. Rev. 79, 357 (1950).
[23] L. Onsager, Physical Review 65, 117 (1944).
[24] J.-F. Yu, Z. Y. Xie, and T. Xiang, , 29004 (2013).
[25] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. C. Courville, CoRR abs/1801.04062 (2018), arXiv:1801.04062.
[26] M. D. Donsker and S. S. Varadhan, Communications on Pure and Applied Mathematics 36, 183 (1983).
[27] D. P. Landau and 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, and S. R. Nagel, Phys. Rev. E 68, 011306 (2003).
[29] J. Kent-Dobias and J. P. Sethna, Phys. Rev. E 98, 063306 (2018).
[30] Y. LeCun, Y. Bengio, and G. Hinton, Nature 521, 436 (2015).
[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in Advances in Neural Information Processing Systems 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. D’Alché Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
[32] S. J. Pan and Q. Yang, IEEE Transactions on Knowledge and Data Engineering 22, 1345 (2010).
[33] L. Onsager, Physical Review 65, 117 (1944).
[34] J. Wilms, M. Troyer, and F. Verstraete, Journal of Statistical Mechanics: Theory and Experiment 2011, P10011 (2011).
[35] J. Iaconis, S. Inglis, A. B. Kallin, and R. G. Melko, Physical Review B 87, 195134 (2013).
[36] M. M. Wolf, F. Verstraete, M. B. Hastings, and J. I. Cirac, Physical review letters 100, 070502 (2008).
[37] G. Ariel and H. Diamant, “Inferring entropy from structure,” (2020), arXiv:2004.10588 [cond-mat.stat-mech].
[38] C. Nardini, E. Fodor, E. Tjhung, F. van Wijland, J. Tailleur, and M. E. Cates, Phys. Rev. X 7, 021007 (2017).
[39] A. J. Liu and S. R. Nagel, Annu. Rev. Condens. Matter Phys. 1, 347 (2010).
Supplemental Materials

A. Data preprocessing and augmentation

Input features were normalized between values -1 and 1. For the jammed 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.

B. Network Architecture

Our method was implemented using the PyTorch library [31]. For subsystems of input size larger than $32 \times 32$ we used three convolutional layers with 16 filters of size $3 \times 3$ each and a rectified linear unit (ReLU) activation. For smaller subsystems, we use only two convolutional layers. For subsystems of size $4 \times 4$ or smaller, only one convolutional layer is used. The last convolutional layer is followed by fully connected layer of size $M \times \frac{k}{2}$, ReLU layer and another fully connected layer of size $\frac{k}{2} \times 1$. Here, $k$ is the number of neurons in the last convolutional layer, see Fig. S1. The batch size for training was 128.

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:

$$
\langle M\rangle_{i+1} = \langle M\rangle_i + \alpha \left( M_{i+1} - \langle M\rangle_i \right).
$$

(S1)

where $M_j$ is the output of the network after $j$ optimization iterations, and $\langle M\rangle_i$ 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 $M$ 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 $M$ estimation on the training set stopped increasing. $M$ was estimated over the independent validation set as well, and this value was used for subsequent calculations. By comparing the estimation of $M$ over the training and validation sets, one can verify that the network did not overfit the data.
FIG. S1. The flow of MICE. A marginal and joint data batches are constructed from our dataset. An estimation of the mutual information is performed by our network. The shown architecture was used for subsystems larger than $32 \times 32$. Smaller subsystems used $1 - 2$ convolutional layers.

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 [25], a moving average with a window size of 100 iterations (orange).

E. Dataset size

For the spin models we used a dataset of 5000 samples of a $64 \times 64$ system. An exception is the XY model with an external field where we used 2000 simulations. For the jammed system we used a set of 100 simulations.

F. Transfer Learning

For random initial network weights the estimation of $M$ gives 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 temperature or size of the subsystem. In the main text we 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. S3. This is an interesting direction for further research, which we did not further...
explore. Transfer learning across subsystem size was not used in the main text.

**S-I. 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 sub-

$$S(X_{k-1}) = 2S(X_k) - \mathcal{M}(X_k) \quad \Rightarrow \quad s(X_{k-1}) \equiv \frac{S(X_{k-1})}{V_{k-1}} = s(X_k) - \frac{\mathcal{M}(X_k)}{2V_k}, \quad (S2)$$

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

$$s(X_1) = s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0}$$

$$s(X_2) = s(X_1) - \frac{\mathcal{M}(X_1)}{2V_1} = s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0} - \frac{\mathcal{M}(X_1)}{4V_0}$$

$$s(X_3) = s(X_2) - \frac{\mathcal{M}(X_2)}{2V_2} = s(X_0) - \frac{\mathcal{M}(X_0)}{2V_0} - \frac{\mathcal{M}(X_1)}{4V_0} - \frac{\mathcal{M}(X_2)}{8V_0}$$

$$\vdots$$

$$s(X_m) = s(X_0) - \frac{1}{2V_0} \sum_{k=0}^{m-1} \frac{\mathcal{M}(X_k)}{2^k}$$

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 $\ell_{-k}/\ell_0 = 2^\lfloor k/2 \rfloor$, 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

$$S(X_m) = s(X_0) - \frac{\mathcal{M}_0}{2V_0} \sum_{k=0}^{m-1} 2^{\lfloor k/2 \rfloor} - k. \quad (S4)$$

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

$$s(X_m) = s(X_0) - 2\frac{\mathcal{M}_0}{V_0}. \quad (S5)$$

**S-II. SPIN MODEL SIMULATIONS**

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

Sampling the distribution of the XY simulation was performed using the c++ library provided in Ref. [29]. To generate uncorrelated samples the mean cluster size at each temperature, $c_s$, was calculated and the simulation was sampled every $2/c_s$ steps. That is, each spin was flipped twice on average between two subsequent samples at all temperatures.

**S-III. 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 $\ell$ (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:

$$\mathcal{M}(A,B) = S(A) + S(B) - S(X), \quad (S6)$$

where the entropy of a subsystem $A$ is given in terms of
the marginal probability:
\[ S(A) = - \sum_{\alpha} P_A(\alpha) \log P_A(\alpha). \] (S7)

Here, \( \alpha \) 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:
\[ P_X(z) = \frac{e^{-\beta E(z)}}{Z}. \] (S8)

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

We proceed by an explicit evaluation of \( M \) at high temperature for the Ising model:
\[ E_{\text{Ising}}(z) = -J \sum_{(i,j)} z_i z_j - H \sum_i z_i, \quad z_i = \pm 1. \] (S9)

The expansion of the partition function in powers of \( \beta \) up to second order is
\[ Z = \sum_{\{z_i = \pm 1\}} (1 - \beta E(z) + \frac{1}{2} \beta^2 E(z)^2 + \ldots) \] (S10)
\[ = 2^N + \frac{1}{2} \beta^2 \left[ J^2 \left( \sum_{(i,j)} 1 \right) 2^N + H^2 \left( \sum_i 1 \right) 2^N \right] + O(\beta^3) \]
\[ = 2^N \left[ 1 + \frac{1}{2} \beta^2 \left( J^2 N_{\text{links}} + H^2 N \right) \right] + O(\beta^3), \] (S11)

where \( \sum_{(i,j)} 1 = N_{\text{links}} \) 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.

Therefore, the numerator of Eq. (S12) yields, to second order in \( \beta \),
\[ P_A(z_a) = 2^{N_B} \frac{1 - \beta E_A(z_A) + \frac{1}{2} \beta^2 (E_A(z_a))^2 + J^2 N_{\text{links}}^B + J^2 \ell + J^2 \sum_{aa'} z_aoz_{a'}}{Z} + O(\beta^3). \] (S14)

Proceeding with the expansion, plugging in Eq. (S11) and using \( N_{\text{links}}^A + N_{\text{links}}^B + \ell = N_{\text{links}} \), we get
\[ P_A(z_a) = \frac{1 - \beta E_A(z_a) + \frac{1}{2} \beta^2 (E_A(z_a))^2 + \frac{1}{2} \beta^2 J^2 \sum_{aa'} z_aoz_{a'}}{Z_A} + O(\beta^3), \] (S15)
\[ Z_A = 2^{N_A} \left[ 1 + \frac{1}{2} \beta^2 J^2 N_{\text{links}}^A \right] + O(\beta^3). \] (S16)

Eq. (S15) has the form of a Boltzmann distribution (note the similarity of Eq. (S16) to Eq. (S11)) derived from the Hamiltonian \( E_A \), with extra couplings generated by the tracing out of \( B \) (the last term in the numerator of Eq. (S15)). A straightforward but tedious calculation, which will not be detailed here, shows that up to

Next, we perform a high temperature expansion up to order \( \beta^2 \) of the marginal probability
\[ P_A(z_A) = \sum_{z_b} P(z_A, z_B) \]
\[ = \sum_{z_b} \frac{1 - \beta E(z_A, z_B) + \frac{1}{2} \beta^2 E^2(z_A, z_B)}{Z} + O(\beta^3). \]

Here \( z_A \) is fixed and spins \( z_B \) in \( B \) act like an environment for \( A \) and are traced out.

Tracing out the first order term in the numerator of Eq. (S12) annihilates any terms that involve at least one spin in \( B \). Therefore, the first order term yields simply the energy of subsystem \( A \),
\[ E_A(z_A) = -J \sum_{(a,a')} z_aoz_{a'}. \] (S13)

Tracing over the second order term in the numerator of Eq. (S12) involves a double sum over neighbors \( \sum_{(i,j)} \sum_{(i',j')} z_i z_j z_i' z_j' \). 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 \( E_A(z_A)^2 \).
2. \( i, j, i', j' \in B \). Summation over these quadruplets yields \( J^2 N_{\text{links}}^B \) where \( N_{\text{links}}^B \) is the number of links in \( B \).
3. \( i \in A, j \in B \) and \( (i, j) = (i', j') \). Summation over these quadruplets yields \( J^2 \ell \) where \( \ell \) 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 \( \sum_{aa'} \).
quadratic order in $\beta$ these couplings do not affect the entropy – while they do clearly affect the probabilities of individual states, as shown in Eq. (S15), 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

$$P_A(z_A) \approx \frac{e^{-\beta E_A(S_A)}}{Z_A} + \mathcal{O}(\beta^3),$$

$$Z_A = \sum_{z_A} e^{-\beta E_A(z_A)} + \mathcal{O}(\beta^3),$$  \hspace{1cm} (S17)

and treat $P_A$ as a standard Boltzmann distribution, for which we have $S = \partial_T (T \log Z)$. Plugging this into Eq. (S6) gives

$$\mathcal{M}(A, B) = \partial_T \left( T \log \frac{Z_A Z_B}{Z_X} \right) + \mathcal{O}(\beta^3).$$  \hspace{1cm} (S18)

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. (S11) and Eq. (S16) we obtain the result

$$\mathcal{M}_{\text{Ising}}(A, B) = \frac{1}{2} \left( \frac{J}{T} \right)^2 \ell + \mathcal{O}(\beta^3).$$  \hspace{1cm} (S19)

Note that neither the sign of $J$ nor the lattice symmetry (square versus triangular) influence the answer to order $\beta^2$ – the only relevant parameters are the number of links connecting the two subsystems $\ell$ 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:

$$\mathcal{M}_{\text{XY}}(A, B) = \frac{1}{4} \left( \frac{J}{T} \right)^2 \ell + \mathcal{O}(\beta^3).$$  \hspace{1cm} (S20)

Both Eq. (S19) and Eq. (S20) 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 [50], 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

$$Z = \int_0^{2\pi} d\theta e^{-\beta E(\theta)},$$

$$E(\theta) = -J \sum_{(i,j)} \cos(\theta_i - \theta_j) - H \sum_{i} \cos \theta_i.$$  \hspace{1cm} (S21)

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

$$Z_0 = \int_{-\infty}^{\infty} d\theta e^{-\beta E_0(\theta)},$$

$$E_0(\theta) = \frac{J}{2} \sum_{(i, j)} (\theta_i - \theta_j)^2 + \frac{H}{2} \sum_{i} \theta_i^2 + \text{const}.$$  \hspace{1cm} (S22)

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

$$p(\theta) = \frac{e^{-\frac{1}{2} \theta^T M \theta}}{Z_0}, \quad \text{with}$$

$$M_{ij} = \frac{H + z J}{T} \delta_{ij} - \frac{J}{T} \delta(i, j).$$  \hspace{1cm} (S23)

$M$ 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 $\delta_{ij} = 1$ if $i$ and $j$ are neighbors and 0 otherwise. The entropy of a multivariate Gaussian is well known:

$$S(X) = \frac{N}{2} \log 2\pi e - \frac{1}{2} \log \det M.$$  \hspace{1cm} (S24)

For a single spin in a magnetic field, for example, this gives $S = \log \left( \sqrt{2\pi e T/H} \right)$ which is valid as long as the variance of $\theta$, $(T/H)^2$, 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 $p(\theta)$ over all degrees of freedom $\theta_B \in B$,

$$p_A(\theta_A) = \frac{1}{Z} \int_0^{2\pi} d\theta_B e^{-\beta E(\theta_A, \theta_B)}.$$  \hspace{1cm} (S25)

To perform the Gaussian integral we decompose the matrix $M$ as

$$M = \begin{pmatrix} M_{AA} & M_{AB} \\ M_{BA} & M_{BB} \end{pmatrix},$$  \hspace{1cm} (S26)

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 $M_{AB} = M_{BA}$ couple the two subsystems. Thus,
FIG. S4. 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. (S20).

$$P_A(\theta_A) = e^{-\frac{1}{2}\theta_A^T M_{AA} \theta_A} \int d\theta_B \exp \left[ -\frac{1}{2}\theta_B^T M_{BB} \theta_B - \theta_A^T M_{AB} \theta_B \right].$$

(S27)

Performing the Gaussian integral over $\theta_B$ gives

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

(S28)

Since the marginal distribution is also Gaussian, its entropy is given by Eq. (S24), with the effective Hessian (covariance matrix) of $A$ given by Eq. (S27),

$$M_{eff}^A = M_{AA} - M_{AB} M_{BB}^{-1} M_{BA}. \quad \text{(S29)}$$

$M_{eff}^A$ 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

$$\mathcal{M} = \frac{1}{2} \log \frac{\det M_X}{\det M_{eff}^A \det M_{eff}^B}. \quad \text{(S30)}$$

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 $2\pi$ 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. (S24) numerically using the effective covariance matrix Eq. (S29).

and is shown in Fig. S4. At high temperature $\mathcal{M}$ decreases like $\mathcal{M} \to \frac{1}{4} \left( \frac{H}{T} \right)^2$, indicated by a dashed line in the right panel, as predicted by Eq. (S20). As $T \to 0$, we can see in the central panel a logarithmic divergence with $T$ which is cut-off when $T \approx H$.

Indeed it is easy to derive from Eqs. (S23), (S29) and (S30) the zero temperature limit of $\mathcal{M}$,

$$\lim_{T \to 0} \mathcal{M}_{\text{two spins}} = \log \frac{H + J}{\sqrt{H(H + 2J)}}. \quad \text{(S31)}$$

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.

**S-IV. $\mathcal{M}$ 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,
FIG. S5. Entropy estimation of the bidisperse soft sphere mixture, using three different methods. For clarity, a linear trend in $\phi$ was subtracted from each curve, according to $\Delta \tilde{s} = A + B(\Delta s - C\phi)$. The values of $A, B, C$ were chosen such that all curves have the same values at $\phi = 0.76$ and $\phi = 0.88$. In blue, the results of MICE ($A = 0, B = 0, C = -104197$). In orange, the results a compression algorithm (the method presented in [8]) over 100,000 subsystems of size $6\sigma$ and the same spatial resolution used by MICE ($A = 83.29, B = 93.25, C = 0.1$). In green, results reported in [17] ($A = -0.84, B = 1.96, C = -0.96$).