

Deep Generative Modeling of the Canonical Ensemble with Differentiable Thermal Properties

Shuo-Hui Li^{1,*}, Yao-Wen Zhang^{1,†} and Ding Pan^{1,2,‡}

¹Department of Physics, The Hong Kong University of Science and Technology, Hong Kong, China

²Department of Chemistry, The Hong Kong University of Science and Technology, Hong Kong, China

 (Received 13 July 2024; revised 15 March 2025; accepted 10 June 2025; published 8 July 2025)

It is a long-standing challenge to accurately and efficiently compute thermodynamic quantities of many-body systems at thermal equilibrium. The conventional methods, e.g., Markov chain Monte Carlo, require many steps to equilibrate. The recently developed deep learning methods can perform direct sampling, but only work at a single trained temperature point and risk biased sampling. Here, we propose a variational method for canonical ensembles with differentiable temperatures, which gives thermodynamic quantities as continuous functions of temperature akin to an analytical solution. The proposed method is a general framework that works with *any* tractable density generative model. At optimal, the model is theoretically guaranteed to be the *unbiased* Boltzmann distribution. We validated our method by calculating phase transitions in the Ising and XY models, demonstrating that our direct-sampling simulations are as accurate as Markov chain Monte Carlo methods but more efficient. Moreover, our differentiable free energy aligns closely with the exact one to the second-order derivative, indicating that the variational model captures the subtle thermal transitions at the phase transitions. This functional dependence on external parameters is a fundamental advancement in combining the exceptional fitting ability of deep learning with rigorous physical analysis.

DOI: [10.1103/8wx7-kyx8](https://doi.org/10.1103/8wx7-kyx8)

The canonical ensemble is for a system in thermal equilibrium with a heat bath, where the temperature, as a principal thermodynamic variable, determines the probability distribution of states. Calculating canonical ensemble averages gives thermodynamic properties of condensed-matter many-body systems, e.g., interacting magnetic spins and solvated proteins. This calculation is closely related to the partition function,

$$\mathcal{Z} = \int \exp[-\beta E(\mathbf{x})] d\mathbf{x}, \quad F = -T \log \mathcal{Z}, \quad (1)$$

where E is the energy function, β is the inverse of the temperature T , and F is the Helmholtz free energy. For example, after obtaining $\mathcal{Z}$, the mean energy and the heat capacity can be readily calculated as derivatives of the free energy,

$$\langle E \rangle = -\frac{\partial \log \mathcal{Z}}{\partial \beta}, \quad C_v = \beta^2 \frac{\partial^2 \log \mathcal{Z}}{\partial \beta^2}. \quad (2)$$

For most physical systems, it is impossible to analytically solve $\mathcal{Z}$: an exact enumeration belongs to the class of #P-hard [1]. In practical calculations, it is common to approximate $\mathcal{Z}$ numerically [2,3]. However, even numerical approximation remains highly challenging, especially for systems undergoing phase transitions (PTs), where the partition function must account for a large number of metastable states.

Because of the complexity of $\mathcal{Z}$, sample-based methods are widely used in numerical calculations, among which the Markov chain Monte Carlo (MCMC) method holds prominence. The MCMC method is an importance sampling technique giving the Boltzmann distribution of the canonical ensemble,

$$p(\mathbf{x}) = \frac{\exp[-\beta E(\mathbf{x})]}{\mathcal{Z}}. \quad (3)$$

It can provide the prevailing state configurations at a microscopic level, which are often unexplored in partition function approaches. This microscopic information is very valuable, e.g., in the identification of intermediate states in chemical reactions. Macroscopic properties, such as thermodynamic quantities, can also be estimated by averaging the samples [4], namely, statistical averaging. As MCMC has the

*Contact author: shuohuili@ust.hk

†These authors contributed equally to this work.

‡Contact author: dingpan@ust.hk

unbiased estimation guarantee, this statistical averaging is the ensemble averaging. An inherent limitation of MCMC is its efficiency. Because of the Markov chain structure of the rejection sampling, it takes multiple time steps to equilibrate and to produce independent and identically distributed samples, which prohibits the MCMC from performing direct sampling [5]. This autocorrelation problem becomes even worse near PTs [5,6]. Additionally, unlike the explicit temperature dependence in the partition function, the MCMC method typically requires more computational effort for results at multiple temperature points, further diminishing its overall efficiency.

Recently, a type of deep neural network model was introduced to address the efficiency limitations of MCMC: the tractable density generative model [7–9]. As a generative model, it can directly produce one-shot independent and identically distributed samples from parametrized distributions, without requiring rejections. Additionally, as a tractable density, it provides the exact probability of the sample [7]. Some of the models can also provide free energy estimation [8,10], which is deemed challenging for MCMC. However, the limitation of temperature dependence persists, as separate models must be trained for each temperature point, forming an inefficient scheme similar to MCMC. Beyond the efficiency issue, this single-temperature training breaks the continuous temperature dependence of Boltzmann distribution. This is reflected in biases, making the sampled distribution favor a few high-probability states and ignoring the vast metastable states, a phenomenon also known as the model collapse. In the case of MCMC, despite inefficiency, its unbiased estimation promises accurate results. However, deep learning models, being variational, may exhibit biases and lack physical fidelity. When applying these models, one usually has to compensate with the aid of the MCMC rejection [8,11], annealing training [10], or sophisticated neural network architectures [12]. As proposed below, biases can be mitigated by employing a temperature-dependent model that accounts for thermal effects in the distribution.

Variational temperature-differentiable method—Here, we introduce the variational temperature-differentiable (VaTD) method as a general framework for modeling the canonical ensemble in a continuous temperature range; by a minimal adaption of incorporating temperature as an input parameter, one can turn *any* tractable density model into a VaTD model. After training with randomly sampled temperatures, this VaTD model can be an accurate direct-sampling model of the target Boltzmann distribution with temperature dependence. Moreover, as the model is differentiable with temperature, the estimates, including the free energy and the partition function, can be differentiated similarly to an analytical solution.

First, considering a single temperature, one can minimize the reverse Kullback-Leibler divergence [13] to optimize a tractable density generative model, i.e.,

$$D_{KL}(q_\theta||p) = \mathbb{E}_{\mathbf{x} \sim q_\theta} \log \frac{q_\theta(\mathbf{x})}{p(\mathbf{x})} = \log \mathcal{Z} + \mathbb{E}_{\mathbf{x} \sim q_\theta} [\log q_\theta(\mathbf{x}) + \beta E(\mathbf{x})], \quad (4)$$

where q_θ is the model distribution, and p is the Boltzmann distribution at this fixed temperature [Eq. (3)]. After optimization, the model q_θ is an approximation of p . Additional, because the $D_{KL} \geq 0$, the statistical average, $\mathbb{E}_{\mathbf{x} \sim q_\theta} [\log q_\theta(\mathbf{x}) + \beta E(\mathbf{x})]$, has a variational lower bound at $-\log \mathcal{Z}$, it can thus be used as an estimate of the partition function in Eq. (1) [8,10]. This has been successfully demonstrated in various applications [14–18]. However, since the actual Boltzmann distribution is temperature-dependent, training schemes that optimize for a single temperature may become biased toward the prevailing states at that temperature. Additionally, as shown in Eq. (4), the model is optimized based on its own samples, which can lead to persistent bias.

In the VaTD training, we propose to approximate the Boltzmann distribution by training in a continuous temperature range. Varying temperatures can reveal hidden local minima, a concept akin to parallel tempering in MCMC, where high temperatures are applied to surmount energy barriers. This motivates us to derive a temperature-dependent loss function from the variational definition of the Boltzmann distribution, i.e., the distribution minimizes free energy across a continuous temperature range,

$$\begin{aligned} \min_{q_\theta} \sum_{\{\mathbf{x}\}} [1/\beta q_\theta(\mathbf{x}, \beta) \log q_\theta(\mathbf{x}, \beta) + q_\theta(\mathbf{x}, \beta) E(\mathbf{x})], \\ \text{s.t. } \sum_{\{\mathbf{x}\}} q_\theta(\mathbf{x}, \beta) = 1, \quad \forall \beta. \end{aligned} \quad (5)$$

Using the Lagrange multiplier, one can see that the Boltzmann distribution in Eq. (3) is the solution [19].

Having temperature as an input parameter, the VaTD models can be used as the variational $q_\theta(\mathbf{x}, \beta)$. This allows Eq. (5) to be formulated in a form that resembles Eq. (4), enabling straightforward deep learning training. As tractable density models, VaTD models are naturally normalized. We then turn the summation into estimation, and the minimum at each β into the minimum of an estimation over uniformly distributed β . Thus, the following loss function is equivalent to the free energy minimization in Eq. (5):

$$\mathcal{L} = \mathbb{E}_{\substack{\beta \sim \mathcal{U} \\ \mathbf{x} \sim q_\theta(\mathbf{x}, \beta)}} [\log q_\theta(\mathbf{x}, \beta) + \beta E(\mathbf{x})] = \mathbb{E}_{\beta \sim \mathcal{U}} [-\log \tilde{\mathcal{Z}}(\beta)]. \quad (6)$$

The stochastic gradient descent [35,36] can be performed to optimize it. Gradients can be obtained from the computational graph and automatic differentiation, which are widely used in deep learning [37]. As ensured in Eq. (5), optimizing this loss drives the model toward the unbiased Boltzmann distribution. Hence, after training, the model

can accurately learn the actual Boltzmann distribution along with its temperature dependence.

Besides its thermodynamic origin, this loss function also represents an integration of the reverse Kullback-Leibler divergence from Eq. (4) over a continuous temperature range. Consequently, this training does not require any dataset, as the sample is drawn from a generative model by direct sampling. Another benefit is that one gets a temperature-differentiable estimate of the free energy, $\log \tilde{Z}(\beta)$. So, its derivatives with respect to temperature, e.g., the mean energy and heat capacity in Eq. (2), can be readily calculated using automatic differentiation. As shown in Supplemental Material (SM) [19], this is more time efficient than the sample-based approach. Other thermodynamic quantities, e.g., magnetization, can be estimated by statistical averaging of directly sampled batches, and they are also differentiable functions of temperature.

One concern is that q_θ may not be flexible enough to cover the whole space of probability distribution, which may result in suboptimal solutions for Eq. (5) and miss the Boltzmann distribution. However, as guaranteed by the universal approximation theorem [38], a sufficiently large neural network model can represent any distribution. Since VaTD training is applicable to any tractable density model, it is easy to apply more advanced model structures. The VaTD training does not require a redesign of the model. Instead, it simply integrates temperature as an input, thus preserving the model's fitting capabilities. In contrast to previous works that involve constructing specialized models designed to consistently follow the Boltzmann distribution (e.g., [39]), our method serves as a general training framework for *any* tractable density model. Additionally, in the VaTD training, the model begins as a random distribution and is subsequently optimized to approximate the unbiased Boltzmann distribution.

In performing stochastic gradient descent, it is worth noting that the gradients of Eq. (6) are not trivial. As the model parameters θ are used in the expectation operator, the nice form of statistical averaging will be lost in the gradients. To overcome this, two numerical methods, namely reparametrization and REINFORCE estimation, have been proposed [20]. They enable the swapping of the differentiation operation with the expectation operation, thereby converting the derivative calculations of statistical averages into the statistical averaging of derivatives. The same problem reoccurs when calculating the derivative of the estimated thermal differentiable function. This can be solved again by using these two methods. Typically, only the first-order forms of these two methods are presented for training. We extend them to the second order in SM [19]. It is worth noting that higher-order generalizations are feasible as long as the neural network remains continuous.

Numerical experiments—In the numerical experiments, we validated our method using two famous statistical models: two-dimensional Ising and XY models. The former

one uses a discrete-variable model, PixelCNN [21], as the tractable density model used in the VaTD method, while the latter one uses a continuous-variable model, the normalizing flow (NF).

Our 2D Ising model is on a 16×16 square lattice with periodic boundary conditions (PBCs). The energy function is given by $E(\mathbf{x}) = -\sum_{\langle i,j \rangle} x_i x_j$, where $\langle i,j \rangle$ means the two spins are nearest neighbors on the lattice. Variables of Ising are discrete, with each variable x_i taking a value of either 1 or -1 , representing spin-up or spin-down, respectively.

For Ising, we incorporated the temperature as an input into the PixelCNN [21] as the VaTD model. PixelCNN is an autoregressive generative model for discrete-variable distributions. Our VaTD PixelCNN's probability distribution is decomposed into the product of a series of conditional probability distributions, i.e., $q_\theta(\mathbf{x}, \beta) = \prod_i q_\theta(x_i | x_0, \dots, x_{i-1}; \beta)$. The conditional distribution we chose is the Bernoulli distribution, $x_i \sim B(1, p)$, where p is the probability of x_i being in the spin-up state, as obtained from a residual network [22] given the preceding variables $\{x_0, \dots, x_{i-1}\}$ and the temperature parameter β . To have a unified input form for the residual network, we repeated β to match the size of $\{x_0, \dots, x_{i-1}\}$, and then concatenated them along the channel dimension. The PixelCNN model generates variables sequentially, one at a time. During each step, a variable is stochastically sampled from the Bernoulli distribution. The conditional causality is established through the dependence of the Bernoulli parameter p on the preceding variables. The model was trained in a temperature range of $\beta \in [0.05, 1.2]$. To eliminate the $\mathbb{Z}_2$ symmetry, we fixed the first spin to be spin-up.

After training, the VaTD PixelCNN model can perform direct sampling, and the sampled batches were used for statistical averaging. In Fig. 1(a), we estimated the free

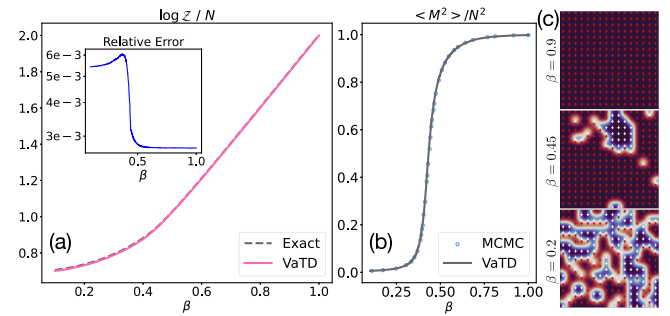


FIG. 1. (a) The exact and estimated free energy of the 2D Ising model on a 16×16 square lattice with PBCs, and the relative error between the two (inset). The temperature factor $(-T)$ is removed from the free energy $(-T \log \mathcal{Z})$ for better comparison. (b) The estimated squared magnetization of the Ising model. (c) State configurations directly sampled from the trained model at three temperatures, with background color as a continuous interpolation of the discrete site for better visualization.

energy as a continuous function of temperature using Eq. (6). As the 2D Ising can be analytically solved [40], we compared our numerical results with the exact free energy at each temperature point in Fig. 1(a), and found that the relative errors are small, with the maximum happening near the PT region where the system exhibits long-range correlations. Similarly, we also calculated the squared magnetization, which is in excellent agreement with the results obtained by the unbiased MCMC simulation, as shown in Fig. 1(b). This indicates that our statistical averages from direct sampling follow the actual ensemble averages closely, similar to MCMC but more efficient. The sampled configurations also provide valuable insights into the microscopic behaviors of the system, and demonstrate that the underlying physical transition is successfully captured. In Fig. 1(c), we observe that, at lower β , the uniform spin direction gradually collapses. Hence, our trained neural network model undergoes the same microscopic thermal transition as the target Ising model.

To further demonstrate that our model unbiasedly captures the subtle thermal transition effects near the PTs, we computed the mean energy and heat capacity, which are derivatives of the free energy as shown in Eq. (2), using automatic differentiation on the estimated free energy. We then compared with the unbiased MCMC results in Fig. 2. Overall, our results agree very well with those from the MCMC simulation, with only slight deviations observed near the PT region. This implies that the estimated free energy aligns closely with the exact one to the second-order derivative. Moreover, compared with statistical averaging using direct sampling, this free energy derivative estimation of mean energy and heat capacity has advantages in efficiency and accuracy (see SM [19] for detailed comparisons).

Next, the XY model is more complex than the Ising model. As an example of the Berezinskii-Kosterlitz-Thouless transition [41,42], the XY model finds practical

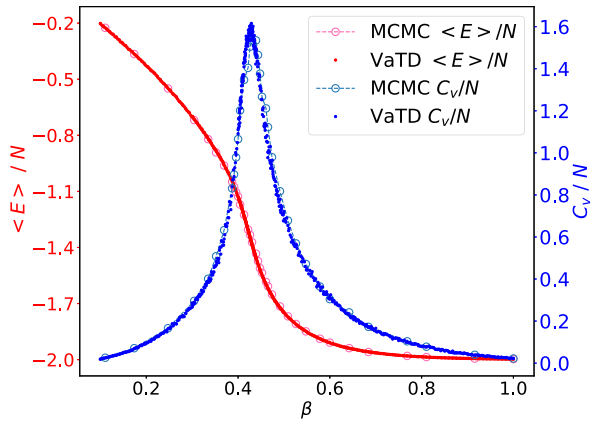


FIG. 2. The mean energy and heat capacity of the Ising model, estimated using the differentiation of the free energy, compared with the results obtained from the MCMC simulation.

applications in explaining various systems, including thin ^4He films [43], and quasi-2D layered superconductors [44]. The energy function of the XY model is $E(\mathbf{x}) = -\sum_{\langle i,j \rangle} \cos(x_i - x_j)$, where x_i and x_j are angles in the continuous range of $[-\pi, \pi]$. Similar to the Ising case, we also used a 2D 16×16 square lattice with PBCs.

For XY's continuous variables, we incorporated the temperature as an input into an NF model [23,45–48] as the VaTD model. In this VaTD NF model, the samples were initially drawn from a prior distribution, which is easy to sample from. We chose the von Mises distribution, $\text{VM}(\mu, \kappa)$, where μ controls the mean and κ controls the variance [49,50], as the prior distribution. To have a temperature-dependent prior distribution, we modified κ as $\kappa_\beta = \kappa\beta$. As VMs were sampled by an acceptance-rejection method, a special reparametrization should be used to optimize μ and κ [19,24] (see SM). The sample $\mathbf{z}$ from the VM prior distribution was then transformed using parametrized invertible transformations, i.e., $\mathbf{x} = f_\theta(\mathbf{z}, \beta)$. To incorporate temperature, we again repeated β and padded it with the sample variables $\mathbf{z}$ in the channel dimension, which is similar to the Ising case. The effects of these invertible transformations, besides changing the sample variables, also introduce probability changes to the VM prior distribution. These distribution changes are measured by the determinant of the Jacobian matrix [45]. The invertible transformation we chose is the piecewise cubic spline transformation [23]. This kind of transformation is designed for continuous variables in a closed range. Then, the final probability of a sample is the product of the prior probability and the determinant of the Jacobian, i.e., $q_\theta(\mathbf{x}, \beta) = \text{VM}(\mathbf{z}|\mu, \kappa_\beta) |\det(\partial \mathbf{z} / \partial \mathbf{x})|$, where $\mathbf{x} = f_\theta(\mathbf{z}, \beta)$. This model was trained in the temperature range of $\beta \in [0.4, 2.0]$. The U(1) symmetry was eliminated by pinning the first spin to 0.

After training, we performed direct sampling and used the obtained batches to perform statistical averaging. Figure 3(a) shows the estimated free energy as a continuous function of temperature. As XY is not exactly solvable, we calculated the standard deviation (STD) to evaluate the estimation quality. According to Eq. (3), a perfectly trained model would produce uniform free energy estimates regardless of $\mathbf{x}$, so the small STDs in Fig. 3(a) indicate accurate alignment with the target Boltzmann distribution. Then, to demonstrate our statistical averages accurately follow the ensemble averages, we compared our direct-sampling simulation results with the unbiased MCMC results; Fig. 3(b) shows the estimated squared magnetization, in excellent agreement with the MCMC results. To study the microscopic behavior, we plotted the sampled configurations at different temperatures in Fig. 3(c), which clearly shows the microscopic changes occurring near the PT. Specifically, with decreasing temperature, the vortex density decreases, and a quasi-long-range order emerges.

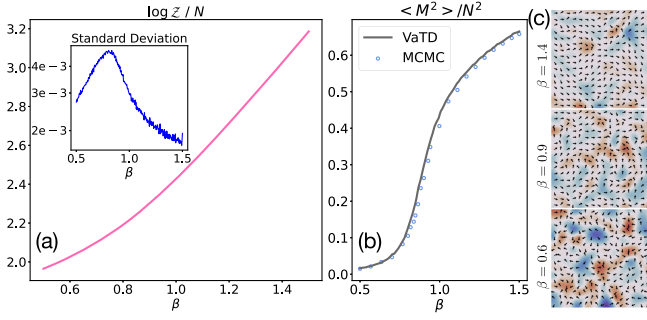


FIG. 3. (a) The estimated free energy and its standard deviation (inset) of the 2D XY model on a 16×16 square lattice with PBCs. The temperature factor ($-T$) is removed from the free energy ($-T \log \mathcal{Z}$) for better comparison. (b) The estimated squared magnetization of the XY model. (c) State configurations directly sampled from the trained model at three temperatures; the background color represents the vorticity.

This demonstrates an accurate capture of the underlying physical transition.

Furthermore, in Fig. 4, we compared the mean energy and heat capacity obtained by the MCMC simulation and the differentiation of the estimated free energy. In comparison to the MCMC result, we observed only minor deviations around the PT, indicating a good fit of the free energy to the second-order derivative and a good capture of the subtle PT transition. Numerical results in SM also show that this free energy derivative estimation of mean energy and heat capacity gives better convergence than the statistical averaging with direct sampling [19]. Moreover, the derivative approach expresses heat capacity as a mean value over a batch of independent estimates, as shown in Eq. (2). This results in faster convergence than the energy deviations used in sample-based approaches; we have a detailed theoretical and numerical analysis in SM [19]. In contrast to the one-by-one generation process of PixelCNN, the NF model is significantly faster due to its all-to-all

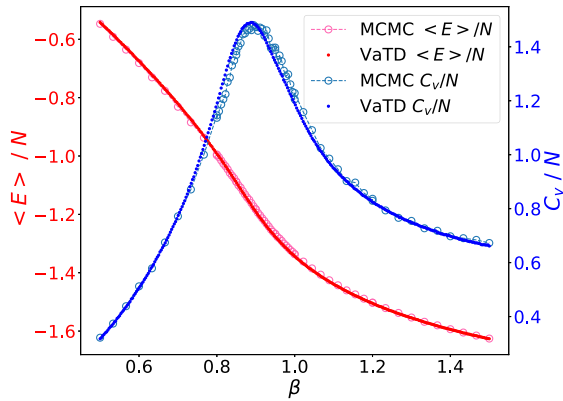


FIG. 4. The mean energy and heat capacity of the XY model, estimated using the differentiation of the free energy, compared with the results obtained from the MCMC simulation.

style of generation. Including training time, the NF model matches Hamiltonian Monte Carlo in estimating 18 temperatures at a convergence accuracy of 10^{-3} , and it becomes faster as the number of temperatures increases. Additionally, the low STD allows us to estimate the heat capacity using the first-order derivative, which offers a significant speedup and reduces the memory requirement. We give a detailed description of this in SM [19]. The proposed method is also effective for the XY model with lattice sizes up to 258×258 . Larger lattices can be accommodated with increased GPU memory or additional GPU cards.

Conclusion and outlook—Here, we present a novel method for variational modeling of the canonical ensemble in a continuous temperature range. In the proposed method, any tractable density generative model can be turned into a VaTD model by directly including temperature as a part of the input. By converting the variational definition of Boltzmann distribution into a temperature-dependent loss, we propose the VaTD training framework to unbiasedly optimize the model. This allows thermodynamic quantities, such as free energy and partition function, to be differentiable functions of temperature. Resembling analytic solutions, this differentiable free energy gives a concise way of expressing thermodynamic quantities as its derivatives. Our method is as accurate as the unbiased MCMC method but more efficient. Moreover, as a single model with temperature dependence, its direct-sampling estimation remains valid across a continuous temperature range. In contrast to previous deep generative models that optimize at a single temperature, our method preserves the essential temperature dependence. As a result, there is no need for training additional models at discrete temperature points, and a single model can reliably capture subtle thermal effects, even at PTs.

As a general training framework, with minor modification, we expect to have immediate practical applications by adapting our method to these popular variational deep generative models at a single temperature [14–18, 25, 51, 52]. Moreover, our proposed method can be seen as a generalization of the variational mean-field method, which consists of multiple conditional parameters. While our current implementation focuses on temperature, it opens up intriguing possibilities for exploring other parameters, such as pressure and external magnetic fields. In numerical experiments, we mainly applied our method to lattice models, but it is also viable to extend its application to more realistic systems with many metastable states, e.g., atomistic simulations.

Acknowledgments—We thank Lei Wang, Wen-Qing Xie, Chu Li, Tao Li, and Jun-Ting Yu for many useful discussions. S.-H. L. and D. P. acknowledge support from Hong Kong Research Grants Council (RGC) (GRF-16302423, GRF-16301723). D. P. acknowledges support

from the Croucher Foundation through the Croucher Innovation Award, National Natural Science Foundation of China (NSFC) through the Excellent Young Scientists Fund (22022310), and the NSFC/RGC Joint Research Scheme (N_HKUST664/24). Part of this work was carried out using computational resources from the National Supercomputer Center in Guangzhou, China, and the X-GPU cluster supported by the RGC Collaborative Research Fund C6021-19EF.

Data availability—The code implementation is publicly available at [53].

-
- [1] A. Bulatov and M. Grohe, The complexity of partition functions, *Theor. Comput. Sci.* **348**, 148 (2005).
 - [2] M. Levin and C. P. Nave, Tensor renormalization group approach to two-dimensional classical lattice models, *Phys. Rev. Lett.* **99**, 120601 (2007).
 - [3] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, *Mach. Learn.* **37**, 183 (1999).
 - [4] R. K. Pathria, *Statistical Mechanics* (Elsevier, New York, 2016).
 - [5] M. E. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, New York, 1999).
 - [6] R. M. Neal, Monte Carlo implementation, in *Bayesian Learning for Neural Networks* (Springer, New York, 1996), pp. 55–98.
 - [7] I. Goodfellow, NIPS 2016 tutorial: Generative adversarial networks, [arXiv:1701.00160](https://arxiv.org/abs/1701.00160).
 - [8] S.-H. Li and L. Wang, Neural network renormalization group, *Phys. Rev. Lett.* **121**, 260601 (2018).
 - [9] F. Noé, S. Olsson, J. Köhler, and H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, *Science* **365**, eaaw1147 (2019).
 - [10] D. Wu, L. Wang, and P. Zhang, Solving statistical mechanics using variational autoregressive networks, *Phys. Rev. Lett.* **122**, 080602 (2019).
 - [11] K. A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller, and P. Kessel, Asymptotically unbiased estimation of physical observables with neural samplers, *Phys. Rev. E* **101**, 023304 (2020).
 - [12] P. Białas, P. Czarnota, P. Korcyl, and T. Stebel, Simulating first-order phase transition with hierarchical autoregressive networks, *Phys. Rev. E* **107**, 054127 (2023).
 - [13] S. Kullback and R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.* **22**, 79 (1951).
 - [14] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racaniere, D. J. Rezende, and P. E. Shanahan, Equivariant flow-based sampling for lattice gauge theory, *Phys. Rev. Lett.* **125**, 121601 (2020).
 - [15] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel, S. Nakajima, and P. Stornati, Estimation of thermodynamic observables in lattice field theories with deep generative models, *Phys. Rev. Lett.* **126**, 032001 (2021).
 - [16] P. Wirnsberger, G. Papamakarios, B. Ibarz, S. Racaniere, A. J. Ballard, A. Pritzel, and C. Blundell, Normalizing flows for atomic solids, *Mach. Learn.: Sci. Technol.* **3**, 025009 (2022).
 - [17] R. Ahmad and W. Cai, Free energy calculation of crystalline solids using normalizing flows, *Model. Simul. Mater. Sci. Eng.* **30**, 065007 (2022).
 - [18] H. Xie, Z.-H. Li, H. Wang, L. Zhang, and L. Wang, Deep variational free energy approach to dense hydrogen, *Phys. Rev. Lett.* **131**, 126501 (2023).
 - [19] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/8wx7-kyx8> for details of (1) a general introduction of the tractable density generative model, (2) the derivation of temperature-dependent Boltzmann distribution using the variational minimization of free energy, (3) the derivation of the first- and second-order derivatives of the expectation operator over parametrized distributions, (4) the derivation of mean energy and heat capacity from the differentiation of the partition function, and convergence time comparisons, and (5) implementation details and supplementary results of the Ising and XY experiments, which includes Refs. [8,11,20–34].
 - [20] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, Monte Carlo gradient estimation in machine learning, *J. Mach. Learn. Res.* **21**, 5183 (2020), <https://www.jmlr.org/papers/v21/19-346.html>.
 - [21] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, Pixel recurrent neural networks, in *International Conference on Machine Learning* (PMLR, New York, 2016), pp. 1747–1756.
 - [22] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Las Vegas, 2016), pp. 770–778.
 - [23] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Cubic-spline flows, [arXiv:1906.02145](https://arxiv.org/abs/1906.02145).
 - [24] C. Naesseth, F. Ruiz, S. Linderman, and D. Blei, Reparameterization gradients through acceptance-rejection sampling algorithms, in *Artificial Intelligence and Statistics* (PMLR, 2017), pp. 489–498.
 - [25] H.-Y. Hu, S.-H. Li, L. Wang, and Y.-Z. You, Machine learning holographic mapping by neural network renormalization group, *Phys. Rev. Res.* **2**, 023369 (2020).
 - [26] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, [arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
 - [27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, [arXiv:1406.2661](https://arxiv.org/abs/1406.2661).
 - [28] S. P. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, England, 2004).
 - [29] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
 - [30] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, Neural ordinary differential equations, [arXiv:1806.07366](https://arxiv.org/abs/1806.07366).
 - [31] B. Efron, Bootstrap methods: Another look at the Jackknife, *Ann. Stat.* **7**, 1 (1979).
 - [32] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs), [arXiv:1511.07289](https://arxiv.org/abs/1511.07289).
 - [33] A. F. Agarap, Deep learning using rectified linear units (ReLU), [arXiv:1803.08375](https://arxiv.org/abs/1803.08375).

- [34] M. Girolami and B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, *J. R. Stat. Soc. Ser. B* **73**, 123 (2011).
- [35] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.* **22**, 400 (1951), <https://www.jstor.org/stable/2236626>.
- [36] B. T. Polyak and A. B. Juditsky, Acceleration of stochastic approximation by averaging, *SIAM J. Control Optim.* **30**, 838 (1992).
- [37] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, Automatic differentiation in machine learning: A survey, *J. Mach. Learn. Res.* **18**, 1 (2018).
- [38] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, *Neural Netw.* **2**, 359 (1989).
- [39] M. Dibak, L. Klein, A. Krämer, and F. Noé, Temperature steerable flows and Boltzmann generators, *Phys. Rev. Res.* **4**, L042005 (2022).
- [40] L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition, *Phys. Rev.* **65**, 117 (1944).
- [41] J. M. Kosterlitz and D. J. Thouless, Ordering, metastability and phase transitions in two-dimensional systems, *J. Phys. C* **6**, 1181 (1973).
- [42] V. Berezinskii, Destruction of long-range order in one-dimensional and two-dimensional systems having a continuous symmetry group I. Classical systems, *Sov. Phys. JETP* **32**, 493 (1971).
- [43] D. J. Bishop and J. D. Reppy, Study of the superfluid transition in two-dimensional ^4He films, *Phys. Rev. Lett.* **40**, 1727 (1978).
- [44] Z. A. Xu, N. P. Ong, Y. Wang, T. Kakeshita, and S. Uchida, Vortex-like excitations and the onset of superconducting phase fluctuation in underdoped $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$, *Nature (London)* **406**, 486 (2000).
- [45] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* **22**, 2617 (2021), <https://www.jmlr.org/papers/v22/19-1028.html>.
- [46] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *International Conference on Machine Learning* (PMLR, 2015), pp. 1530–1538.
- [47] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, Neural importance sampling, *ACM Trans. Graph.* **38**, 1 (2019).
- [48] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, *Adv. Neural Inf. Process. Syst.* **32** (2019), <https://proceedings.neurips.cc/paper/2019/hash/7ac71d433f282034e088473244df8c02-Abstract.html>.
- [49] N. I. Fisher, *Statistical Analysis of Circular Data* (Cambridge University Press, Cambridge, England, 1993).
- [50] D. Best and N. I. Fisher, Efficient simulation of the von Mises distribution, *J. R. Stat. Soc. C* **28**, 152 (1979).
- [51] H. Xie, L. Zhang, and L. Wang, Ab-initio study of interacting fermions at finite temperature with neural canonical transformation, *J. Mach. Learn.* **1**, 38 (2022).
- [52] H. Xie, L. Zhang, and L. Wang, m^* of two-dimensional electron gas: A neural canonical transformation study, *SciPost Phys.* **14**, 154 (2023).
- [53] <https://github.com/li012589/vatd> for code implementation.