Scaling autoregressive models for lattice thermodynamics

Xiaochen Du
MIT ChemE/CSE
dux@mit.edu

Sulin Liu
MIT DMSE
sulinliu@mit.edu

Rafael Gómez-Bombarelli
MIT DMSE
rafagb@mit.edu

Abstract

Understanding the thermodynamics of solid state, crystalline materials is important for applications ranging from catalysis to electronics. Traditional sampling methods in the discrete spaces represented by periodic lattices are generally Markov-chain Monte Carlo-based, with limitations in speed and scalability. Autoregressive methods, used in text and image generation, have been adapted for sampling lattice thermodynamics. However, these methods rely on a fixed generation order and do not possess fast arbitrary marginal likelihood evaluation capabilities, making scaling lattice sizes challenging. Here, we develop and combine marginalization models with any-order inference and show the resulting models enable larger lattice generation in two ways: scaling training to larger lattice sizes and allowing for in-/out-painting from models trained on smaller lattices. We demonstrate our method using the Ising model and CuAu alloy.

1 Introduction

Understanding the equilibrium distribution of states is the first step towards generating realistic materials under experimental conditions. Materials exist at finite temperatures; thus it is not a single structure that defines a static material property. For example, molecules have a distribution of conformations, each with distinct properties (Noé et al., 2019) while catalyst surfaces evolve as a function of reactants, temperature (T), and external chemical potentials (μ) (Seh et al., 2017; Bruix et al., 2019). In the continuum space, for example, machine-learning-based optimal control methods have been used for enhanced sampling of molecular potential energy surfaces (Hua et al., 2024; Yuan et al., 2024).

In this work, we focus on periodic lattice systems that can be modeled using discrete states. Previously, Wu et al. (2019); Nicoli et al. (2020); Damewood et al. (2022) adapted autoregressive methods (ARMs) used in image and text generation to learn lattice thermodynamics. However, these ARMs are hampered in flexibility due to: (1) their fixed generation order and (2) requiring expensive evaluation of conditional probabilities over the entire sequence length, L, during training. Due to the former, arbitrary in-/out-painting tasks for conditional generation are out of scope for these fixed-order (FO) ARMs. Due to the latter, the computational graph of automatic differentiation scales as $\mathcal{O}(L^2)$, limiting training samples to a modest lattice size, thereby hampering the accuracy of thermodynamic observables.

We develop any-order (AO) ARMs that improve on FO-ARMs of previous works and marginalization models (MAMs) that scale training to larger lattice systems capable of sampling across T and μ

AI for Accelerated Materials Discovery (AI4Mat) Workshop, NeurIPS 2024, Vancouver.

(motivations described in Appendix A). We test our method on up to 20×20 Ising lattices and $4 \times 4 \times 8$ CuAu alloys, benchmarking on variational free energies, free energy estimations, and phase diagrams.

2 Methods

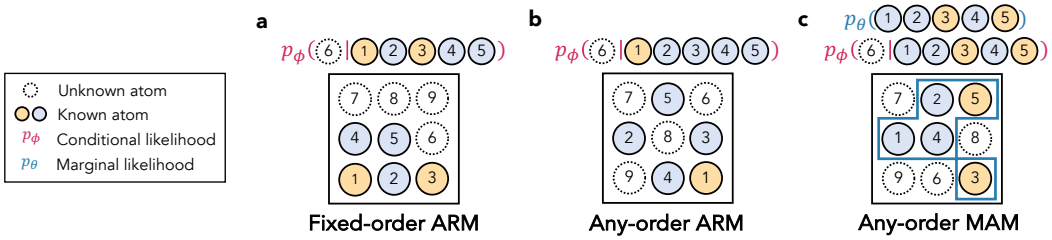

Figure 1: Autoregressive models investigated in this work.

Autoregressive models ARMs factor high-dimensional discrete distributions $p(\mathbf{x}; T, \mu_1, \mu_2, ..., \mu_n)$ at the specified temperature and chemical potentials (henceforth denoted as $p(\mathbf{x})$) into a sequence of conditional distributions: $\log p_\phi(\mathbf{x}) = \sum_{\ell=1}^{L} \log p_\phi(x_\ell | \mathbf{x}_{<\ell})$, where p_ϕ can be parameterized using a neural network and $\mathbf{x}_{<\ell} = \{x_1, ..., x_{\ell-1}\}$. In our case, examples are generated starting from the first lattice site, x_1, given by $p_\phi(x_1)$ until the last site given by $p_\phi(x_L | x_1, ..., x_{L-1})$, requiring L passes in total.

To model the Boltzmann distribution of our lattice system in the semigrand-canonical ensemble, we minimize the Kullback–Leibler (KL) divergence (Damewood et al., 2022; Liu et al., 2024):

$$\min_{\phi} D_{\mathrm{KL}} \left(p_\phi(\mathbf{x}) \middle\| \frac{f(\mathbf{x})}{Z} \right) \tag{1}$$

where the score $f(\mathbf{x}) = \exp\left(-\frac{E(\mathbf{x}) - \sum_{i=1}^{K} \mu_i N_i(\mathbf{x})}{k_{\mathrm{B}} T} \right)$ is the Boltzmann weight, $E(\mathbf{x})$ is the energy of the lattice, K is the total number of atom/spin types, $N_i(\mathbf{x})$ is the number of sites with identity i, k_{B} is the Boltzmann constant, T is the temperature, and Z the ensemble partition function. Equation 1 can be optimized with the REINFORCE algorithm (Williams, 1992) with samples drawn according to p_ϕ (Details in Appendix D).

Any-order generation and in-/out-painting ARMs developed thus far assume a fixed ordering of lattice sites (FO-ARMs, Fig. 1(a)), but we train AO models to predict conditionals $p_\phi(x_{\sigma(\ell)} | \mathbf{x}_{\sigma(<\ell)})$ over an arbitrary ordering, σ. We train by uniformly sampling an ordering for each sample and following the prescribed order to predict p_ϕ at each step. With an AO model, we can arbitrarily fill empty lattice sites, which we show in our in-/out-painting procedure for a 2D lattice (Fig. 2, see Appendix B for additional use cases).

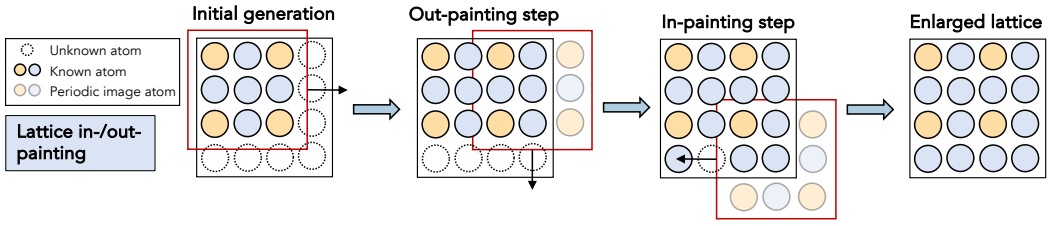

Figure 2: In-/out-painting procedure using any-order models.

Any-order marginalization models Liu et al. (2024) proposed AO-MAMs that are compatible with the energy-based training metric used in materials science settings (Equation 1). For a subset of indices, $\mathcal{S} \subseteq \{1, ..., L\}$ and its complement $\mathcal{S}'$, the marginal is defined as: $p(\mathbf{x}_{\mathcal{S}}) = \sum_{\mathbf{x}_{\mathcal{S}'}} p(\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\mathcal{S}'})$.

We can train a neural network to fit $p_\theta(\mathbf{x})$ to the target Boltzmann distribution $p(\mathbf{x}) = f(\mathbf{x})/Z$ and learn arbitrary marginals $p_\theta(\mathbf{x}_\mathcal{S})$ in the process. As such, p_θ can be used to approximate p_ϕ in Equation 1, enabling fast $\mathcal{O}(1)$ likelihood evaluation. Additionally, an AO-ARM is trained alongside a marginal network (Fig. 1(c)), enabling efficient generation of training samples (in principle, in $\mathcal{O}(1)$, instead of $\mathcal{O}(L)$ for ARMs) through AO block-wise Gibbs sampling of persistent Markov chains. The self-consistency between the marginal and conditional networks can be enforced with an additional term $\mathcal{L}^{\text{Mar}}_{\theta,\phi}(\mathbf{x}, \sigma, \ell) = \left[\log\left(p_\theta(\mathbf{x}_{\sigma(<\ell)})p_\phi(x_{\sigma(\ell)}|\mathbf{x}_{\sigma(<\ell)})\right) - \log p_\theta(\mathbf{x}_{\sigma(\leq\ell)})\right]^2$, giving rise to the objective:

$$\min_{\theta,\phi} D_{\text{KL}}(p_\theta(\mathbf{x}) \parallel p(\mathbf{x})) + \lambda \mathbb{E}_\mathbf{x} \mathbb{E}_\sigma \mathbb{E}_\ell \mathcal{L}^{\text{Mar}}_{\theta,\phi}(\mathbf{x}, \sigma, \ell) \tag{2}$$

where σ, ℓ are uniformly sampled from all possible elements.

Architecture We attempted standard multilayer perceptron (MLP) architectures for both the conditional net (p_ϕ) to predict an $L \times K$ matrix and the MLP marginal net (p_θ) that outputs a single scalar. Details in Appendices F & J. T and μ are treated as scalar inputs. The number of parameters in MLP models scales as $\mathcal{O}(L \cdot K)$, limiting its efficiency for higher-dimensional systems compared to Transformers that utilize a shared weight matrix.

Therefore, we developed a Transformer architecture to improve parameter efficiency and scale training to larger lattice sizes. T and μ are expanded to the same embedding dimension as $\mathbf{x}$ to increase model expressiveness. We additionally propose periodic sinusoidal functions across each dimension to account for lattice periodicity (details in Appendix E) on the basis of Rotary Positional Embeddings (RoPE) (Su et al., 2023), which already accounts for translational invariance in 2D for the Ising model and 3D for CuAu.

3 Experiments

3.1 Ising model

We describe Ising models with the score function $f(\mathbf{x}) = \exp\left(\frac{1}{2}(\mathbf{x}^\top \mathbf{J}\mathbf{x} + \mu\mathbf{x})/(k_\text{B}T)\right)$, where $\mathbf{x} \in \{-1, 1\}^L$ represents the spin state, $\mathbf{J}$ is a product of interaction strength ε and the adjacency matrix, and μ the field strength. Below the critical temperature for a particular lattice size, spins align with the magnetic field but become disordered above it. As smaller lattice sizes suffer from finite-size effects, there is a need to model larger lattices to get more accurate thermodynamic observables.

We trained various ARM and MAM models on different lattice sizes across $\mu \in [-0.4, +0.4]$ and $k_\text{B}T \in [1.5, 3.5]$ with $\varepsilon = 1$. Training details are provided in Appendix F. A results summary is presented in Table 1, where the average variational free energy normalized by temperature, $\langle \frac{F_q}{k_\text{B}T} \rangle = \langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ (the minimization objective in practice, details in Appendix C), was evaluated over 5 evenly-spaced values over each of μ and T for a total of 25 unique conditions. The best result has the lowest $\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$. See Table A1 for a breakdown of log scores ($\langle \log f(\mathbf{x}) \rangle$) and log-likelihoods ($\langle \log p(\mathbf{x}) \rangle$).

Table 1: Ising model generation results. The best result for each lattice size is bolded. FO models cannot perform arbitrary in-/out-painting tasks. 20×20 Ising ARM models exceed GPU capacity. $\langle \cdot \rangle$ is equivalent to $\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[\cdot]$.

	5×5	10×10	20×20	
Model	Trained	Trained	10×10 In-/out-painted	Trained
ARM		$\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ (↓)		
FO-MLP	-26.26	-104.69	-	-
AO-MLP	**-26.27**	**-104.77**	**-415.13**	-
MAM		$\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ (↓)		
AO-MLP	-26.22	-98.90	-401.31	Mode collapse
AO-Transformer	-26.25	-104.66	**-415.13**	-414.34

5×5 and 10×10 Ising models For both lattice sizes, AO-ARM MLP models train as well as FO-ARM MLP with the same architecture, while enabling in-/out-painting. AO-MAM MLP models do well for 5×5 lattices but struggle at the 10×10 size. Additional results for the 10×10 Ising models are presented in Fig. 3. In Fig. 3(a), the $\log f(\mathbf{x})$ distribution of samples across the same set of μ and T used for evaluating $\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ are plotted. The distribution of AO-MAM MLP samples deviate significantly from those of ARMs while the distribution of AO-MAM Transformer is much closer. Visually, in Fig. 3(c), AO-MAM 10×10 samples exhibit much finer grains compared with the larger clusters of AO-ARM and AO-MAM Transformer samples. In Fig. 3(b), we plot the free energy per site (method in Appendix G) as a function of $k_B T$ at various μ for the models studied and compare with the exact values at $\mu = 0$ (Beale, 1996) and the reference values obtained from the Wang-Landau method (Wang and Landau, 2001). We exclude AO-MAM MLP from comparison due to the poor results. See Appendix H for details on the reference values. Other than the excluded AO-MAM MLP model, all free energies obtained from the models differ slightly from the exact values at $\mu = 0.0$ and the Wang-Landau energies at other μ, especially at higher T. The AO-MAM Transformer model deviates only marginally more from reference values than ARM models.

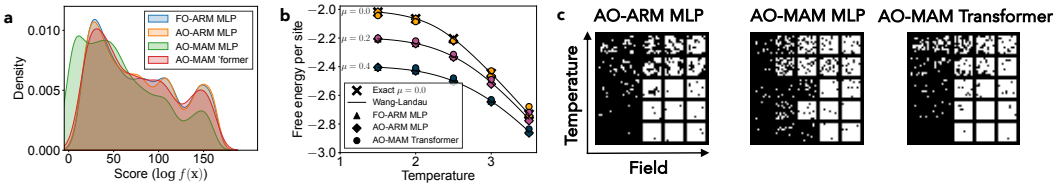

Figure 3: 10×10 Ising results. (a) Score distribution, (b) free energies, and (c) generated samples.

In-/out-painting and 20×20 Ising model Using the 10×10 AO models, we performed in-/out-painting (details in Appendix I) to obtain 20×20 Ising lattices. We also successfully trained a 20×20 model using MAM Transformer but the MAM MLP model suffered from mode collapse. 20×20 ARM models exceeded the allowed hardware memory, highlighting their limitations in scaling. 10×10 ARM MLP in-/out-painted and MAM in-/out-painted Transformer models perform equally well. The 20×20 MAM Transformer peforms almost as well as these two models. Additional results for the 20×20 Ising models are presented in Fig. A1. Qualitatively, 10×10 MAM MLP in-/out-painted samples (Fig. A1(a)) are similarly inferior to those of 10×10 ARM MLP in-/out-painted and 10×10 MAM Transformer in-/out-painted samples due to small cluster sizes at higher temperatures. For the 20×20 MAM Transformer model, the samples at higher temperatures are much closer to those of the 10×10 ARM in-/out-painted and 10×10 MAM in-/out-painted Transformer models, but seems to have issues at low temperatures and intermediate fields. Fig. A1(b) shows the $\log f(\mathbf{x})$ distribution of the samples. The lower score peaks (high T) for the 20×20 MAM Transformer are closer to those of 10×10 ARM MLP in-/out-painted and MAM in-/out-painted Transformer while the 10×10 MAM in-/out-painted MLP model is closer for the higher score (low T) peaks.

Future work could involve new architectures or training strategies to improve the performance of 20×20 and larger MAM models.

3.2 CuAu

CuAu alloy generation is harder than Ising lattice generation due to the multiple low-temperature phases at different compositions. At higher temperatures, the alloy becomes disordered. Similar to the Ising model, the thermodynamics depends on the lattice size, hence a need to sample larger lattices to obtain more realistic phase diagrams.

The energy model is parametrized by a cluster expansion as in Damewood et al. (2022). We trained MAM models for the $2 \times 2 \times 4$ case and show an in-/out-painting application to $4 \times 4 \times 8$, details in Appendices J & L. In Table 2, $\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ was evaluated over 8 uniformly-distributed $\mu \in [-0.12, +0.12]$ and 4 uniformly-distributed $T \in [200\text{ K}, 1200\text{ K}]$ for a total of 32 unique conditions. See Table A2 for a breakdown of $\langle \log f(\mathbf{x}) \rangle$ and $\langle \log p(\mathbf{x}) \rangle$.

For the $2 \times 2 \times 4$ models, we again demonstrate that the AO-ARM MLP model performs as well as the FO-ARM model, with the benefit of AO generation. In terms of $\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$, the Transformer model performs better than the AO-MAM MLP model, and is close to the AO-ARM

MLP performance. For the $4 \times 4 \times 8$ samples generated using in-/out-painting, AO-MAM MLP seems to perform the best but the $\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ across all models are not ~ 8 times the respective values of $2 \times 2 \times 4$ lattices. We attribute the discrepancy to the limitations of a small starting lattice size such that when in-painting the edges of the 4×4 cross section, we cannot include the periodic image atoms in our marginal computation.

Table 2: CuAu model generation results with the best for each lattice size bolded.

	$2 \times 2 \times 4$	$4 \times 4 \times 8$
Model	Trained	$2 \times 2 \times 4$ Inpainted
ARM	$\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ (↓)	
FO-MLP	**-32.92**	-
AO-MLP	**-32.92**	-215.29
MAM	$\langle \log p(\mathbf{x}) - \log f(\mathbf{x}) \rangle$ (↓)	
AO-MLP	-31.87	**-223.16**
AO-Transformer	-32.84	-217.06

Fig. 4 shows the phase diagrams for the $2 \times 2 \times 4$ CuAu models (with the exception of FO-ARM MLP), each of which plots the observed samples by the sampling temperature and the sample Au concentration. Each phase diagram was generated with 2048 samples uniformly spread across 32 evenly-spaced bins in each of $\mu \in [-0.12, +0.12]$ and $T \in [200\text{ K}, 1200\text{ K}]$. The reference two-phase equilibria line was obtained from metadynamics, which shows the separation between low-temperature ordered phases and high-temperature disordered solid solution. See Appendix K for information on the metadynamics procedure and Fig. A2 for the metadynamics phase diagram. The AO-MAM Transformer and AO-ARM MLP models perform much better than the AO-MAM MLP model in showing phase transition between the low and high temperatures, but some samples do appear beneath the metadynamics equilibria line. We note the metadynamics "ground-truth" equilibria line is for the $2 \times 2 \times 4$ supercell, and is notably different from that of the original $4 \times 4 \times 8$ supercell in Damewood et al. (2022), which was a better fit to the phase diagram of the ARM MLP model in that work.

Further studies may include training both ARM and MAM models at the $4 \times 4 \times 8$ size to have a more conclusive comparison with metadynamics sampling and in-/out-painting to or training MAM on even larger lattices.

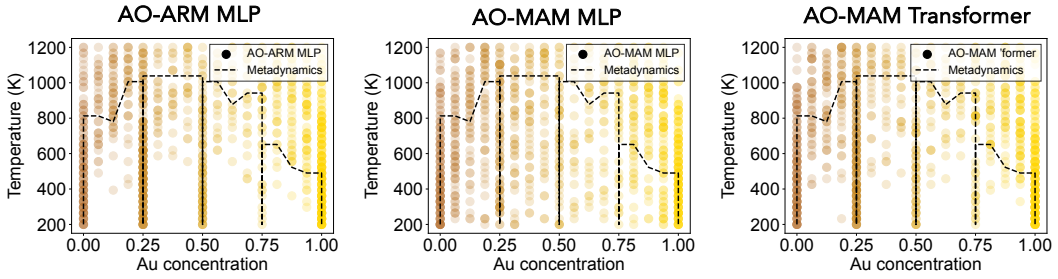

Figure 4: Temperature-concentration phase diagrams for $2 \times 2 \times 4$ CuAu with two-phase equilibria lines obtained from metadynamics.

4 Acknowledgements

The authors thank Hoje Chun, James Damewood, Alexander Hoffman, and Juno Nam for helpful discussions and review of the work. X.D. acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No. 2141064. The authors are grateful for computation time allocated on the NERSC Perlmutter cluster under Project No. m4737 and on the MIT SuperCloud cluster.

References

P. D. Beale. Exact Distribution of Energies in the Two-Dimensional Ising Model. *Physical Review Letters*, 76(1):78–81, Jan. 1996. doi: 10.1103/PhysRevLett.76.78. URL https://link.aps.org/doi/10.1103/PhysRevLett.76.78. Publisher: American Physical Society.

A. Bruix, J. T. Margraf, M. Andersen, and K. Reuter. First-principles-based multiscale modelling of heterogeneous catalysis. *Nature Catalysis*, 2(8):659–670, Aug. 2019. ISSN 2520-1158. doi: 10.1038/s41929-019-0298-3. URL https://www.nature.com/articles/s41929-019-0298-3. Number: 8 Publisher: Nature Publishing Group.

J. H. Chang, D. Kleiven, M. Melander, J. Akola, J. M. Garcia-Lastra, and T. Vegge. CLEASE: a versatile and user-friendly implementation of cluster expansion method. *Journal of Physics: Condensed Matter*, 31(32):325901, May 2019. ISSN 0953-8984. doi: 10.1088/1361-648X/ab1bbc. URL https://dx.doi.org/10.1088/1361-648X/ab1bbc. Publisher: IOP Publishing.

H. Chun, J. Lunger, J. K. Kang, R. Gómez-Bombarelli, and B. Han. Active learning accelerated exploration of single-atom local environments in multimetallic systems for oxygen electrocatalysis, Mar. 2024. URL https://chemrxiv.org/engage/chemrxiv/article-details/65f09f1b9138d2316153e3f4.

J. Damewood, D. Schwalbe-Koda, and R. Gómez-Bombarelli. Sampling lattices in semi-grand canonical ensemble with autoregressive machine learning. *npj Computational Materials*, 8(1):1–10, Apr. 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00736-4. URL https://www.nature.com/articles/s41524-022-00736-4. Number: 1 Publisher: Nature Publishing Group.

X. Du, J. K. Damewood, J. R. Lunger, R. Millan, B. Yildiz, L. Li, and R. Gómez-Bombarelli. Machine-learning-accelerated simulations to enable automatic surface reconstruction. *Nature Computational Science*, pages 1–11, Dec. 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00571-7. URL https://www.nature.com/articles/s43588-023-00571-7. Publisher: Nature Publishing Group.

X. Hua, R. Ahmad, J. Blanchet, and W. Cai. Accelerated Sampling of Rare Events using a Neural Network Bias Potential, Jan. 2024. URL http://arxiv.org/abs/2401.06936. arXiv:2401.06936.

P. Li and F. Ding. Origin of the herringbone reconstruction of Au(111) surface at the atomic scale. *Science Advances*, 8(40):eabq2900, Oct. 2022. doi: 10.1126/sciadv.abq2900. URL https://www.science.org/doi/10.1126/sciadv.abq2900. Publisher: American Association for the Advancement of Science.

S. Liu, P. Ramadge, and R. P. Adams. Generative Marginalization Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 31773–31807. PMLR, July 2024. URL https://proceedings.mlr.press/v235/liu24az.html. ISSN: 2640-3498.

J. Low, J. Yu, M. Jaroniec, S. Wageh, and A. A. Al-Ghamdi. Heterojunction Photocatalysts. *Advanced Materials*, 29(20):1601694, 2017. ISSN 1521-4095. doi: 10.1002/adma.201601694. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.201601694. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201601694.

K. A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller, and P. Kessel. Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2):023304, Feb. 2020. doi: 10.1103/PhysRevE.101.023304. URL https://link.aps.org/doi/10.1103/PhysRevE.101.023304. Publisher: American Physical Society.

F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, Sept. 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaw1147. URL https://www.science.org/doi/10.1126/science.aaw1147.

C. J. Owen, Y. Xie, A. Johansson, L. Sun, and B. Kozinsky. Low-index mesoscopic surface reconstructions of Au surfaces using Bayesian force fields. *Nature Communications*, 15(1):3790, May 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-48192-6. URL https://www.nature.com/articles/s41467-024-48192-6. Publisher: Nature Publishing Group.

Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov, and T. F. Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321): eaad4998, Jan. 2017. doi: 10.1126/science.aad4998. URL https://www.science.org/doi/10.1126/science.aad4998. Publisher: American Association for the Advancement of Science.

Y. Shen, S. I. Morozov, K. Luo, Q. An, and W. A. Goddard III. Deciphering the Atomistic Mechanism of Si(111)-7 × 7 Surface Reconstruction Using a Machine-Learning Force Field. *Journal of the American Chemical Society*, 145(37):20511–20520, Sept. 2023. ISSN 0002-7863. doi: 10.1021/jacs.3c06540. URL https://doi.org/10.1021/jacs.3c06540. Publisher: American Chemical Society.

J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, Nov. 2023. URL http://arxiv.org/abs/2104.09864. arXiv:2104.09864 [cs].

F. Wang and D. P. Landau. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters*, 86(10):2050–2053, Mar. 2001. doi: 10.1103/PhysRevLett.86.2050. URL https://link.aps.org/doi/10.1103/PhysRevLett.86.2050. Publisher: American Physical Society.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

D. Wu, L. Wang, and P. Zhang. Solving Statistical Mechanics Using Variational Autoregressive Networks. *Physical Review Letters*, 122(8):080602, Feb. 2019. doi: 10.1103/PhysRevLett.122.080602. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.080602. Publisher: American Physical Society.

Y. Xiao, Y. Wang, S.-H. Bo, J. C. Kim, L. J. Miara, and G. Ceder. Understanding interface stability in solid-state batteries. *Nature Reviews Materials*, 5(2):105–126, Feb. 2020. ISSN 2058-8437. doi: 10.1038/s41578-019-0157-5. URL https://www.nature.com/articles/s41578-019-0157-5. Number: 2 Publisher: Nature Publishing Group.

J. Yuan, A. Shah, C. Bentz, and M. Cameron. Optimal control for sampling the transition path process and estimating rates. *Communications in Nonlinear Science and Numerical Simulation*, 129:107701, Feb. 2024. ISSN 1007-5704. doi: 10.1016/j.cnsns.2023.107701. URL https://www.sciencedirect.com/science/article/pii/S1007570423006226.

A Motivations for scaling lattice autoregressive models

In materials science, some phenomena are only observed in a sufficiently large supercell. For example in surface reconstruction, the Au(111) herringbone reconstruction requiring a periodic length of ~ 30 nm (Li and Ding, 2022; Owen et al., 2024) and Si(111) 7x7 surface reconstructions (Shen et al., 2023; Du et al., 2023). As such, scaling the training lattice size of autoregressive models is important to uncover these phenomena and sample the correct thermodynamics.

In cases where scaling training to these large sizes is impractical, we can train the model on smaller lattice sizes with the goal of observing the same large-scale thermodynamics by in-/out-painting using the trained model. This is akin to training a machine learning force field, where the training data consists of smaller lattice sizes amenable to first-principles calculations but the trained force field is applied to larger supercells during production runs (Owen et al., 2024).

B Additional use cases for in-/out-painting

Additionally, in-/out-painting can be used for conditional generation and constrained design, which is important for materials science applications such as in designing catalysts (Chun et al., 2024), battery interfaces (Xiao et al., 2020), semiconductor heterojunctions (Low et al., 2017), or understanding surface reconstructions (Du et al., 2023) given specific thermodynamic conditions. Specifically, we want to predict arbitrary unknown lattice sites which can only be done with an any-order model, e.g., generating the remainder of a catalyst given the support and parts of the active site or surface reconstructions given partially occluded experimental observations. This cannot be done with a fixed-order autoregressive generation method.

C Variational free energy

The KL divergence can be expanded as such (Wu et al., 2019; Nicoli et al., 2020; Damewood et al., 2022):

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\phi(\mathbf{x})||p(\mathbf{x})\right) &= D_{\mathrm{KL}}\left(p_\phi(\mathbf{x})\middle\|\frac{f(\mathbf{x})}{Z}\right) \\ &= \mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\left[\log p_\phi(\mathbf{x}) - \log f(\mathbf{x}) + \log Z\right] \\ &= \frac{1}{k_{\mathrm{B}}T}\mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\left[E(\mathbf{x}) - \sum_{i=1}^{K}\mu_i N_i(\mathbf{x}) + k_{\mathrm{B}}T\log p_\phi(\mathbf{x}) + k_{\mathrm{B}}T\log Z\right] \\ &= \frac{F_q - F}{k_{\mathrm{B}}T} \geq 0 \end{aligned}$$

where the variational free energy $F_q = \mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\left[E(\mathbf{x}) - \sum_{i=1}^{K}\mu_i N_i(\mathbf{x}) + k_{\mathrm{B}}T\log p_\phi(\mathbf{x})\right]$ is an upper bound to the true free energy of the system $F = -k_{\mathrm{B}}T\log(Z)$. Thus in practice, we minimized $\frac{F_q}{k_{\mathrm{B}}T} = \mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\left[\log p_\phi(\mathbf{x}) - \log f(\mathbf{x})\right] = D_{\mathrm{KL}}\left(p_\phi(\mathbf{x})||f(\mathbf{x})\right)$ as our learning objective.

D Training by REINFORCE

We drew a batch of samples, $\mathbf{x}$, using an autoregressive model (ARM or MAM) and estimated the gradient of the KL divergence with REINFORCE (Williams, 1992; Wu et al., 2019):

$$\begin{aligned} \nabla_\phi D_{\mathrm{KL}}\left(p_\phi(\mathbf{x})\middle\|\frac{f(\mathbf{x})}{Z}\right) &= \nabla_\phi D_{\mathrm{KL}}\left(p_\phi(\mathbf{x})\,\|f(\mathbf{x})\right) \text{ (from Appendix C)} \\ &= \mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\left[\nabla_\phi \log p_\phi(\mathbf{x})\left(\log p_\phi(\mathbf{x}) - \log f(\mathbf{x})\right)\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\nabla_\phi \log p_\phi(\mathbf{x}^{(i)})\left(\log p_\phi(\mathbf{x}^{(i)}) - \log f(\mathbf{x}^{(i)})\right) \end{aligned}$$

E Periodic sinusoidal positional embedding vectors

We set values for ω_i, the frequency of each dimension in the RoPE embedding vector, subject to periodicity constraints across each system dimension:

$$\sin(\omega_i x) = \sin(\omega_i(x + L))\forall x, L \in \mathbb{R}^+$$
$$\cos(\omega_i x) = \cos(\omega_i(x + L))\forall x, L \in \mathbb{R}^+$$
$$L\omega_i = 2n\pi, n \in \mathbb{Z} \Rightarrow \omega_i = \frac{2n\pi}{L}$$

where x is the position in each dimension.

F Ising model training details

All MLP models used 4 hidden layers with width 512. MAM models are trained with an additional marginal network with the same hidden sizes and consistency loss parameter $\lambda = 100$. 5×5 Ising models used 5 Gibbs sampling steps while 10×10 and 20×20 models used 10 Gibbs sampling steps during training. The 5×5 Transformer model had 3 Transformer blocks with width 256 for both the embedding dimension and hidden layer, while the larger lattice size models had 512 for the corresponding dimensions. λ for Transformers was set to 10.

During training, a batch of 1250 samples were sampled at each iteration across a set of 5 temperatures and 5 fields, for a total of 25 unique conditions. The conditions were uniformly distributed across $\mu \in [-0.4, +0.4]$ and $k_\mathrm{B}T \in [1.5, 3.5]$ at the start and the change in $\Delta\mu, \Delta k_\mathrm{B}T$ per epoch were sampled according to $\Delta\mu \sim \mathcal{N}(0, 0.05^2), \Delta k_\mathrm{B}T \sim \mathcal{N}(0, 0.2^2)$. One epoch elapsed after 100,000 training samples were sampled. All models were trained for 500 epochs.

G Ising model free energy estimation with importance sampling

While the variational free energy F_q (Appendix C) can be used as the free energy estimate, it can be biased (Nicoli et al., 2020). The partition function, $\hat{Z}$, and hence the lattice free energy can be estimated using an asymptotically-unbiased method (Nicoli et al., 2020) as follows:

$$\begin{aligned} Z &= \sum_{\text{all } \mathbf{x}} f(\mathbf{x}) \\ &= \sum_{\text{all } \mathbf{x}} p(\mathbf{x}) \frac{f(\mathbf{x})}{p(\mathbf{x})} \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mathbf{x})}{p(\mathbf{x})} \\ &= \hat{Z}_N \end{aligned}$$

where we obtained $N = 10,000$ samples at each set of μ and $k_\mathrm{B}T$ and $\hat{A} = k_\mathrm{B}T \ln \hat{Z}_N$. $\hat{A}$ is then normalized by the number of sites to obtain the free energy per site.

H Ising model ground truth free energies

The exact values for the free energies at $\mu = 0$ and $k_\mathrm{B}T \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$ were obtained from a series expansion (Beale, 1996; Damewood et al., 2022).

The Wang-Landau algorithm (Wang and Landau, 2001) is a Markov-chain Monte Carlo (MCMC) method used to obtain density-of-state estimates, $S(\mathbf{x})$, for each state, $\mathbf{x}$. After which, the lattice partition function, $Z = \sum_{\text{all } \mathbf{x}} S(\mathbf{x}) \cdot f(\mathbf{x})$, was constructed and used to estimate the free energy per site at each temperature and μ. The results using the Wang-Landau method were obtained from Damewood et al. (2022), where the algorithm was run for 10^{11} steps at each $\mu \in \{0.0, +0.2, +0.4\}$ and then each set of results was individually applied to $k_\mathrm{B}T \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$.

I 20 x 20 Ising model in-/out-painting

A batch of 1250 20×20 empty Ising lattices were initialized at 5 temperatures and 5 fields uniformly distributed across $\mu \in [-0.4, +0.4]$ and $k_\mathrm{B}T \in [1.5, 3.5]$, for a total of 25 unique conditions. Following the procedure in Fig. 2, a 10×10 section was sampled using the AO models. Afterwards, out-painting steps were performed in the first dimension with a stride of 3 until the frame matched the periodic image sites. Then the frame was reset to the original 10×10 section and shifted in the second dimension also with a stride of 3. The process was repeated until we performed the last in-painting step to obtain fully filled 20×20 lattices.

The conditional probabilities for each site was similarly estimated in 10×10 sections in a random order conditional on any existing log-likelihoods for individual sites. The $\log p(\mathbf{x})$ for each lattice was then obtained as a sum over individual log-likelihoods.

J CuAu training details

All MLP models used 4 hidden layers with width 512. MAM models were trained with an additional marginal network with the same hidden sizes and consistency loss parameter $\lambda = 100$. 4 Gibbs sampling steps were used to train the $2 \times 2 \times 4$ CuAu model. The Transformer model had 3 Transformer blocks with width 512 for both the embedding dimension and hidden layer. λ for Transformers was set to 10.

During training, a batch of 2048 samples were sampled at each iteration across a set of 4 temperatures and 4 chemical potentials, for a total of 16 unique conditions. The conditions were uniformly distributed across $\mu \in [-0.12, +0.12]$ and $T \in$ [200 K, 1200 K] at the start and the change in $\Delta\mu, \Delta T$ per epoch were sampled according to $\Delta\mu \sim \mathcal{N}(0, 0.02^2), \Delta T \sim \mathcal{N}(0, 50^2)$. One epoch elapsed after 100,000 training samples were sampled. All models were trained for 400 epochs

K 2 x 2 x 4 CuAu phase diagram metadynamics

Metadynamics was performed following the procedure of Damewood et al. (2022) using CLEASE (Chang et al., 2019). The collective variable for the metadynamics runs is the concentration of Au, with 17 bins, one for each possible concentration of Au. Each new sample was proposed using semi-grand canonical Monte Carlo and accepted according to the Metropolis-Hastings algorithm. Metadynamics was run at 32 temperatures uniformly spread across $T \in$ [200 K, 1200 K]. At each temperature, each metadynamics step consists of running the sampling for a maximum of 20,000 sweeps until the convergence criterion was met: when the most infrequent bin was visited at least 80% of the average. The artificial potential (V) is initially modified by 10,000 $k_\mathrm{B}T$ when the sampler visits a bin, and is halved at each metadynamics step until 0.0001 to ensure convergence of the free energy curve.

Following the metadynamics runs, the points on the convex hull of V across all bins at each temperature were taken to construct the reference phase diagram in Fig. A2. The two-phase equilibria boundary was then extracted from the metadynamics phase diagram by linearly connecting the sampled points with the lowest temperature at each Au concentration.

L 4 x 4 x 8 CuAu in-/out-painting

A batch of 2048 $4 \times 4 \times 8$ empty CuAu lattices were initialized at 4 temperatures and 4 fields uniformly distributed across $\mu \in [-0.12, +0.12]$ and $T \in$ [200 K, 1200 K], for a total of 16 unique conditions. Following the procedure in Fig. 2, a $2 \times 2 \times 4$ section was sampled using the AO models. Similar to Appendix I, in-/out-painting steps were performed in each of the 3 dimensions with a single stride per step until the lattices were fully filled.

The conditional probabilities for each site was similarly estimated in $2 \times 2 \times 4$ sections in a random order conditional on any existing log-likelihoods for individual sites. The $\log p(\mathbf{x})$ for each lattice was then obtained as a sum over individual log-likelihoods.

M Additional tables

Table A1: Ising model generation results in terms of $-\langle \log f(\mathbf{x}) \rangle$ and $\langle \log p(\mathbf{x}) \rangle$

	5×5	10×10	20×20	
Model	Trained	Trained	10×10 In-/out-painted	Trained
ARM	$-\langle \log f(\mathbf{x}) \rangle$ / $\langle \log p(\mathbf{x}) \rangle$			
FO-MLP	-19.22 / -7.04	-77.70 / -27.01	-	-
AO-MLP	-19.29 / -6.98	-76.91 / -27.93	-302.99 / -112.14	-
MAM	$-\langle \log f(\mathbf{x}) \rangle$ / $\langle \log p(\mathbf{x}) \rangle$			
AO-MLP	-19.55 / -6.69	-58.54 / -40.44	-278.71 / -122.60	Mode collapse
AO-Transformer	-19.14 / -7.19	-75.26 / -29.41	-305.16 / -109.97	-267.14 / -147.20

Table A2: CuAu model generation results in terms of $-\langle \log f(\mathbf{x}) \rangle$ and $\langle \log p(\mathbf{x}) \rangle$

	$2 \times 2 \times 4$	$4 \times 4 \times 8$
Model	Trained	$2 \times 2 \times 4$ Inpainted
ARM	$-\langle \log f(\mathbf{x}) \rangle$ / $\langle \log p(\mathbf{x}) \rangle$	
FO-MLP	-29.16 / -3.76	-
AO-MLP	-29.06 / -3.86	-188.35 / -26.94
MAM	$-\langle \log f(\mathbf{x}) \rangle$ / $\langle \log p(\mathbf{x}) \rangle$	
AO-MLP	-27.00 / -4.86	-189.17 / -33.99
AO-Transformer	-28.77 / -4.08	-189.02 / -28.04

N Additional figures

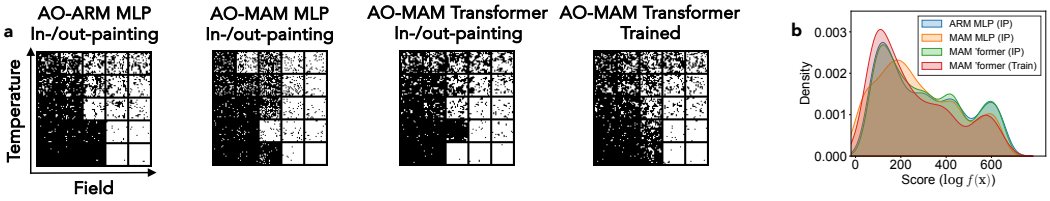

Figure A1: 20×20 Ising results. (a) Generated samples and (b) score distribution. IP denotes in-/out-painted.

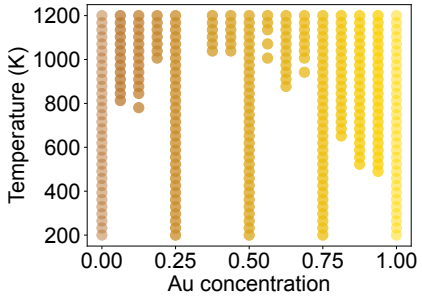

Figure A2: Metadynamics reference temperature-concentration phase diagram for $2 \times 2 \times 4$ CuAu.