Discrete Normalizing Flows as Variational Ansätze for Classical Statistical Mechanics

Notes

February 7, 2026

Abstract

We discuss the generalization of autoregressive variational ansätze for classical statistical mechanics—originally formulated over raw spin variables—to autoregressive models over *transformed tokens*. We identify discrete normalizing flows as the natural framework that preserves the key property of exact, tractable log-probabilities while allowing learned, expressive transformations of the configuration space. We describe the architecture, the joint training procedure, and discuss the physical content that the learned flow may reveal, including connections to the renormalization group, duality transformations, topological defect encoding, and computational complexity of phases.

Contents

1 Background: autoregressive variational free energy

Consider a classical spin system on a lattice of N sites with configuration $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$, where each $\sigma_i \in \{-1, +1\}$ (Ising case), and energy function $E(\boldsymbol{\sigma})$. The Boltzmann distribution at inverse temperature $\beta = 1/T$ is

$$p(\boldsymbol{\sigma}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{\sigma})}, \qquad Z = \sum_{\boldsymbol{\sigma}} e^{-\beta E(\boldsymbol{\sigma})}. \tag{1}$$

Wu et al. [1] proposed using autoregressive neural networks as variational ansätze. The variational distribution factorizes as

$$q_\theta(\boldsymbol{\sigma}) = \prod_{i=1}^{N} q_\theta(\sigma_i \mid \sigma_{<i}), \tag{2}$$

where each conditional is parameterized by a neural network (MADE, PixelCNN, or RNN). The variational free energy

$$F_{\text{var}}[q] = \langle E(\boldsymbol{\sigma}) \rangle_q + T \langle \ln q(\boldsymbol{\sigma}) \rangle_q \geq F_{\text{true}} \tag{3}$$

provides a rigorous upper bound on the true free energy $F_{\text{true}} = -T \ln Z$.

The crucial property of (2) is that $\ln q_\theta(\boldsymbol{\sigma}) = \sum_i \ln q_\theta(\sigma_i \mid \sigma_{<i})$ is *exact and tractable*, so both terms in (3) can be estimated unbiasedly by sampling from q_θ. Gradients are computed via the REINFORCE (policy gradient) estimator:

$$\nabla_\theta F_{\text{var}} = \mathbb{E}_{q_\theta}\Big[\big(E(\boldsymbol{\sigma}) + T \ln q_\theta(\boldsymbol{\sigma}) \big) \nabla_\theta \ln q_\theta(\boldsymbol{\sigma}) \Big]. \tag{4}$$

2 Motivation: autoregression over transformed tokens

The factorization (2) is tied to a particular ordering of the raw spin variables. For systems with complex correlation structures—especially near phase transitions, in frustrated systems, or in the presence of topological defects—this raw-spin factorization may be suboptimal. One expects that a more natural set of variables exists in which the autoregressive conditionals are simpler.

Several choices of transformed tokens are natural:

(i) **Block-spin tokens:** partition the lattice into blocks, label each block state as a single token.

(ii) **Fourier-mode tokens:** factorize autoregressively in momentum space, generating modes from low-k to high-k.

(iii) **Wavelet tokens:** use a multiscale decomposition, generating coarse scales before fine scales.

(iv) **Learned discrete tokens (VQ-VAE):** learn a discrete codebook via a vector-quantized autoencoder.

(v) **Domain-wall / defect tokens:** re-express configurations in terms of topological defects.

The central question is whether such generalizations can maintain the *rigorous variational bound* of (3).

3 The rigour problem with latent-variable models

The VQ-VAE approach defines a generative model with latent tokens $\boldsymbol{z} = (z_1, \ldots, z_K)$:

$$q(\boldsymbol{\sigma}) = \sum_{\boldsymbol{z}} p_\theta(\boldsymbol{z})\, p_\phi(\boldsymbol{\sigma} \mid \boldsymbol{z}), \tag{5}$$

where $p_\theta(\boldsymbol{z}) = \prod_k p_\theta(z_k \mid z_{<k})$ is an autoregressive prior and $p_\phi(\boldsymbol{\sigma} \mid \boldsymbol{z})$ is a decoder. The marginal (5) is *intractable*: evaluating $\ln q(\boldsymbol{\sigma})$ requires summing over all latent sequences. This breaks the rigorous bound.

3.1 Attempted fix: joint variational bound

Introducing a reference distribution $r(\boldsymbol{z})$ and using $\mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\sigma}, \boldsymbol{z}) \,\|\, p_{\mathrm{Boltz}}(\boldsymbol{\sigma}) \cdot r(\boldsymbol{z})) \geq 0$ with $r(\boldsymbol{z}) = p_\theta(\boldsymbol{z})$ yields:

$$\boxed{F_{\mathrm{true}} \leq \langle E(\boldsymbol{\sigma}) \rangle_q + T \langle \ln p_\phi(\boldsymbol{\sigma} \mid \boldsymbol{z}) \rangle_q} \tag{6}$$

This is a rigorous, tractable upper bound. However, the gap between (6) and the true variational free energy $F_{\mathrm{var}}[q]$ is

$$\Delta = T \cdot I_q(\boldsymbol{\sigma};\, \boldsymbol{z}), \tag{7}$$

the mutual information between spins and tokens under the joint model. For a good generative model this mutual information is large (potentially extensive in N), rendering the bound impractically loose.

3.2 Attempted fix: ELBO + importance sampling

Using an encoder $q_\psi(\boldsymbol{z} \mid \boldsymbol{\sigma})$, the standard ELBO gives a *lower* bound on $\ln q(\boldsymbol{\sigma})$:

$$\ln q(\boldsymbol{\sigma}) \geq \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{\sigma})}\big[\ln p_\phi(\boldsymbol{\sigma}|\boldsymbol{z}) + \ln p_\theta(\boldsymbol{z}) - \ln q_\psi(\boldsymbol{z}|\boldsymbol{\sigma})\big]. \tag{8}$$

Substituting this into (3) gives a quantity $\tilde{F} \leq F_{\mathrm{var}}[q]$, which is *not* a valid upper bound on F_{true}. One can evaluate $q(\boldsymbol{\sigma})$ via importance sampling after training, yielding a consistent estimator of $F_{\mathrm{var}}[q]$, but not a strict bound at finite sample size.

4 Discrete normalizing flows: the rigorous solution

The key insight is to replace the lossy VQ-VAE with a *bijective* discrete transformation. Since the map is a bijection on a finite set, no marginalization is needed and $\ln q(\boldsymbol{\sigma})$ remains exact.

4.1 Setup

Let $\boldsymbol{z} \in \{\pm 1\}^N$ denote latent spins and $\boldsymbol{\sigma} \in \{\pm 1\}^N$ denote physical spins. Define:

- A **base distribution** $p_\theta(\boldsymbol{z}) = \prod_k p_\theta(z_k \mid z_{<k})$, an autoregressive model with parameters θ.
- A **discrete flow** $f_\phi : \{\pm 1\}^N \to \{\pm 1\}^N$, a learnable bijection parameterized by ϕ.

The variational distribution over physical spins is

$$q(\boldsymbol{\sigma}) = p_\theta\big(f_\phi^{-1}(\boldsymbol{\sigma})\big), \tag{9}$$

and

$$\ln q(\boldsymbol{\sigma}) = \ln p_\theta(\boldsymbol{z}) = \sum_{k=1}^{N} \ln p_\theta(z_k \mid z_{<k}), \qquad \boldsymbol{z} = f_\phi^{-1}(\boldsymbol{\sigma}). \tag{10}$$

No Jacobian correction is needed: for a bijection on a finite set, the "Jacobian" is identically 1.

Proposition 1. *The variational free energy* $F_{\text{var}}[q] = \langle E(\boldsymbol{\sigma})\rangle_q + T\langle \ln q(\boldsymbol{\sigma})\rangle_q$ *with* q *defined by* (9) *provides a rigorous upper bound on* F_{true} *and can be evaluated exactly (up to sampling noise) via*

$$F_{\text{var}} = \mathbb{E}_{\boldsymbol{z}\sim p_\theta}\Big[E\big(f_\phi(\boldsymbol{z})\big) + T \ln p_\theta(\boldsymbol{z})\Big]. \tag{11}$$

4.2 Coupling layers for binary spins

We construct f_ϕ as a composition of L *discrete coupling layers*, each of which is bijective by construction.

Single layer. Partition the N sites into two groups (A, B) (e.g., a checkerboard partition on a square lattice). A coupling layer acts as:

$$\sigma_A = z_A, \qquad \sigma_B = z_B \odot m_\phi(z_A), \tag{12}$$

where $m_\phi : \{\pm 1\}^{|A|} \to \{\pm 1\}^{|B|}$ is a neural network that outputs a conditional *flip mask*, and $\odot$ denotes elementwise multiplication. Since $(\pm 1)^2 = +1$, the inverse is identical:

$$z_B = \sigma_B \odot m_\phi(\sigma_A), \qquad z_A = \sigma_A. \tag{13}$$

Composition. The full flow is

$$f_\phi = f^{(L)} \circ f^{(L-1)} \circ \cdots \circ f^{(1)}, \tag{14}$$

alternating which sites belong to group A and group B at each layer. Each $f^{(l)}$ has its own mask network with parameters $\phi^{(l)} \subset \phi$.

Algebraic interpretation. In the $\mathbb{F}_2$ representation ($\{0, 1\}$ with XOR), each coupling layer implements a conditional affine transformation: $\sigma_B = z_B \oplus m_\phi(z_A)$. The space of all such compositions forms a subgroup of $\mathrm{Aut}(\mathbb{F}_2^N)$. The flow searches for an element of this group such that the Boltzmann distribution, expressed in the coordinates $\boldsymbol{z} = f_\phi^{-1}(\boldsymbol{\sigma})$, admits the simplest autoregressive factorization.

4.3 Physical interpretation of a coupling layer

The algebraic definition (12) has a transparent physical meaning. Consider a square lattice with the checkerboard partition: group A is one sublattice (say the black squares) and group B is the other (white squares). The coupling layer examines *all* the black spins and, for each white spin independently, decides whether to flip it. The mask network $m_\phi(z_A)$ encodes this decision rule.

As a deterministic sublattice update. The simplest useful mask aligns each white spin with its local field from the black neighbours. If site $i \in B$ has local field $h_i = J \sum_{j \in A,\, j \sim i} z_j$, then

$$m_i = \mathrm{sign}(h_i) \tag{15}$$

flips $z_{B,i}$ to point along the local field. This is a *deterministic mean-field update on one sublattice*—one half-sweep of iterative conditional modes. The learned mask network can go well beyond this: it sees the *entire* A-sublattice configuration and can make non-local, context-dependent flip decisions.

Alternating refinement. Stacking layers with alternating partitions yields the structure shown in Table 1. Each layer refines one sublattice conditioned on the other—*deterministic alternating Gibbs sampling* with learned update rules.

Table 1: Alternating sublattice updates in a composition of coupling layers.

Layer	**Reads**	**Updates**	**Correlations resolved**
1	black sublattice	white sublattice	nearest-neighbour A–B
2	white (corrected)	black sublattice	next-nearest-neighbour A–A via B
3	black (corrected)	white sublattice	~ 3 lattice spacings
$\vdots$	$\vdots$	$\vdots$	$\vdots$
l	—	—	$\sim l$ lattice spacings

The correlation range captured by the flow grows linearly with depth:

$$\xi_{\text{eff}} \sim L_{\text{flow}}. \tag{16}$$

This is the direct origin of the scaling $L^* \sim \xi \sim |T - T_c|^{-\nu}$ discussed in Sec. 6: the flow must be at least as deep as the correlation length is long.

Action on defects. The inverse flow f_ϕ^{-1} maps physical configurations to the latent space. Its action on defects is instructive:

- **Smooth domains** (all spins aligned) pass through nearly unchanged—all masks are close to $+1$.
- **Domain walls** are progressively straightened, shrunk, or moved to a canonical position. Each layer peels away one "layer of roughness" from the wall.
- **Point defects** (a single flipped spin in a uniform background) are absorbed: the mask flips them back, mapping the configuration to the uniform state in $\boldsymbol{z}$-space.

The forward flow does the reverse: starting from a simple latent configuration, it *grows* defects layer by layer—first placing them roughly, then refining their shape and position.

Frustration increases per-layer difficulty. For an unfrustrated bipartite lattice (e.g. the ferromagnetic square Ising model), each B-spin wants to align with all its A-neighbours and there is no conflict; the local-field mask (15) already does a good job. For a frustrated system (triangular antiferromagnet, spin glass), each B-spin receives contradictory signals from its A-neighbours. The mask network must learn a complex, non-local compromise, requiring larger networks and deeper flows—reflecting the intrinsic computational hardness of the frustrated phase.

Summary. A discrete coupling layer is a *learned, deterministic, sublattice-conditional spin update*: the discrete analogue of one half-sweep of Gibbs sampling, optimised end-to-end so that the full composition maps a simple base distribution to the target Boltzmann distribution.

5 Joint training procedure

The trainable parameters are θ (base AR model) and ϕ (flow coupling networks). The objective (11) is minimized by gradient descent. The two parameter groups play very different roles, and their gradients have different structures.

5.1 Gradient with respect to θ (base AR model)

Both the sampling distribution and the integrand depend on θ. Applying the REINFORCE identity:

$$\nabla_\theta F_{\text{var}} = \mathbb{E}_{p_\theta}\Big[\Big(E\big(f_\phi(\boldsymbol{z})\big) + T\ln p_\theta(\boldsymbol{z})\Big)\,\nabla_\theta \ln p_\theta(\boldsymbol{z})\Big]. \tag{17}$$

This is the standard policy-gradient estimator from [1]. Variance reduction via a learned baseline $b(\boldsymbol{z}_{<k})$ is essential:

$$\nabla_\theta F_{\text{var}} \approx \frac{1}{M}\sum_{m=1}^{M}\Big(R^{(m)} - b\Big)\,\nabla_\theta \ln p_\theta(\boldsymbol{z}^{(m)}), \qquad R^{(m)} = E(\boldsymbol{\sigma}^{(m)}) + T\ln p_\theta(\boldsymbol{z}^{(m)}). \tag{18}$$

5.2 Gradient with respect to ϕ (flow parameters)

This gradient has a qualitatively different structure. Recall the objective pulled back to latent space:

$$F_{\text{var}} = \mathbb{E}_{\boldsymbol{z}\sim p_\theta}\Big[\underbrace{E\big(f_\phi(\boldsymbol{z})\big)}_{\text{depends on }\phi} + \underbrace{T\ln p_\theta(\boldsymbol{z})}_{\text{independent of }\phi}\Big]. \tag{19}$$

Two key observations:

1. The sampling distribution $p_\theta(\boldsymbol{z})$ does not depend on ϕ, so we can move ∇_ϕ inside the expectation without a REINFORCE correction.
2. The entropy term $T\ln p_\theta(\boldsymbol{z})$ does not depend on ϕ at all—the flow does not change which latent configuration was drawn, only where it lands in physical space.

Therefore:

$$\nabla_\phi F_{\text{var}} = \mathbb{E}_{p_\theta}\Big[\nabla_\phi E\big(f_\phi(\boldsymbol{z})\big)\Big]. \tag{20}$$

The entropy drops out entirely: *the flow is optimised purely by rearranging which latent configuration maps to which physical configuration, so as to lower the expected energy.*

The discrete barrier. Equation (20) looks like a standard backpropagation problem, but there is a fundamental obstacle. Inside each coupling layer, the mask is computed as

$$m_i = \operatorname{sign}\big(g_\phi(z_A)_i\big) \in \{\pm 1\}, \tag{21}$$

where g_ϕ is a neural network with continuous outputs. As ϕ varies continuously, g_ϕ changes smoothly, but $\operatorname{sign}(\cdot)$ snaps to ± 1. The composed map $f_\phi(\boldsymbol{z})$ is therefore a *piecewise-constant* function of ϕ: it takes discrete values and is flat almost everywhere, with discontinuous jumps at the boundaries where some $g_\phi(z_A)_i$ crosses zero. Consequently, $\nabla_\phi E(f_\phi(\boldsymbol{z})) = 0$ almost everywhere in the conventional sense—standard backpropagation yields no gradient signal.

We employ one of the following relaxations to obtain a useful training signal:

Straight-through estimator (STE). In the forward pass, compute hard masks $m_i = \operatorname{sign}(g_\phi(z_A)_i)$. In the backward pass, replace sign' by the identity, so that gradients flow through the mask network as if $m_i \approx g_\phi(z_A)_i$:

$$\left.\frac{\partial m_i}{\partial g_i}\right|_{\text{STE}} = 1. \tag{22}$$

This is biased but empirically effective. Gradients backpropagate through the full composition $f^{(L)} \circ \cdots \circ f^{(1)}$.

Gumbel–softmax relaxation. During training, replace the hard mask with a continuous surrogate:

$$\tilde{m}_i = \tanh\left(\frac{g_\phi(z_A)_i + (\xi_1 - \xi_2)}{\tau}\right), \qquad \xi_{1,2} \sim \text{Gumbel}(0,1), \tag{23}$$

and anneal $\tau \to 0$ over training. At $\tau = 0$, the exact discrete flow is recovered.

Decoupling property. A key simplification: the gradient for θ does not require differentiating through the flow, and the gradient for ϕ does not require REINFORCE. The two parameter groups can be updated with different optimizers and learning rates.

5.3 Training algorithm

Algorithm 1 Joint training of autoregressive base + discrete flow

Require: Base AR network p_θ, flow layers $\{f_\phi^{(l)}\}_{l=1}^{L}$, temperature T, batch size M

for each training step **do**
  Sample $\boldsymbol{z}^{(m)} \sim p_\theta(\boldsymbol{z})$ for $m = 1, \ldots, M$ ▷ Autoregressive sampling
  Compute $\boldsymbol{\sigma}^{(m)} = f_\phi(\boldsymbol{z}^{(m)})$ ▷ Forward through flow (hard masks)
  Compute $\ln q^{(m)} = \sum_k \ln p_\theta(z_k^{(m)} \mid z_{<k}^{(m)})$ ▷ Exact log-probability
  Compute local free energy $R^{(m)} = E(\boldsymbol{\sigma}^{(m)}) + T \ln q^{(m)}$
  Update θ**:** $\theta \leftarrow \theta - \alpha_\theta \cdot \frac{1}{M} \sum_m (R^{(m)} - b)\, \nabla_\theta \ln p_\theta(\boldsymbol{z}^{(m)})$ ▷ REINFORCE
  Update ϕ**:** $\phi \leftarrow \phi - \alpha_\phi \cdot \frac{1}{M} \sum_m \nabla_\phi E(\tilde{f}_\phi(\boldsymbol{z}^{(m)}))$ ▷ STE backward pass
end for

6 Physical content of the learned flow

The learned bijection f_ϕ is an interpretable object: it defines a *change of variables* in configuration space that the model finds optimal for representing the Boltzmann distribution. We discuss several ways in which this transformation encodes non-trivial physics.

6.1 Emergent renormalization group

If the flow is structured hierarchically—e.g., layer 1 acts on nearest-neighbor pairs, layer 2 on 2×2 blocks, layer 3 on 4×4 blocks—then the composition implements a multiscale transformation. The base distribution captures an effective theory at the coarsest scale.

Diagnostic 1: self-similarity at criticality. At the RG fixed point $T = T_c$, the flow parameters should become approximately self-similar across scales: the mask networks at different layers should implement statistically similar transformations. Away from T_c, deep layers contribute little because correlations are short-range, and the flow effectively "terminates early."

Diagnostic 2: effective degrees of freedom. The base-distribution entropy $H[p_\theta]$ at each scale gives an estimate of the effective number of degrees of freedom, analogous to the c-function in the Zamolodchikov c-theorem. One can track how $H[p_\theta]$ varies with temperature and system size to extract scaling exponents.

6.2 Kramers–Wannier duality

The 2D Ising model on a square lattice possesses a duality transformation $\mathcal{D}$ that maps high-temperature configurations (spins) to low-temperature configurations (domain walls), with the self-dual point at T_c.

- A flow trained at T_c should approximate $\mathcal{D}$ (or its composition with a simple transformation), since the self-dual distribution has enhanced symmetry that the flow can exploit.
- One can test this by examining the flow's action on known configurations (all-up, checkerboard, single domain wall) and comparing with the analytical Kramers–Wannier map.
- More generally, for models with known dualities (Potts, gauge–Higgs), the learned flow may *rediscover* the duality transformation, providing a data-driven route to non-obvious dualities.

6.3 Topological defect encoding

In ordered phases, the dominant excitations are topological defects (domain walls in the Ising model, vortices in the XY model, monopoles in gauge theories). A well-trained flow should map configurations to a latent space where:

1. A small number of latent variables encode the *number and topology* of defects (their winding numbers, connectivity, etc.).
2. The remaining latent variables encode *smooth deformations* of defect positions and shapes.

This can be tested by feeding configurations with known defect content through f_ϕ^{-1} and examining the latent representation. Configurations with the same defect topology should map to nearby regions in $\boldsymbol{z}$-space.

6.4 Flow depth as a probe of computational complexity

The minimum flow depth L^* required to achieve a given free-energy accuracy is a measure of the *circuit complexity* of the Boltzmann distribution.

Unfrustrated systems. For the ferromagnetic Ising model, we expect $L^* \sim \xi$, where $\xi \sim |T - T_c|^{-\nu}$ is the correlation length. At T_c, L^* diverges with system size as $L^* \sim N^{\nu/d}$.

Frustrated / glassy systems. For the Sherrington–Kirkpatrick model or other spin glasses, L^* may grow much faster—potentially exponentially in N—reflecting the NP-hardness of the ground-state problem. Mapping out $L^*(T, N)$ as a "computational phase diagram" could reveal transitions in computational complexity that coincide with (or differ from) thermodynamic phase boundaries.

6.5 Disentanglement of order parameter and fluctuations

Near a phase transition, the physically relevant decomposition is:

$$\boldsymbol{\sigma} = \underbrace{\bar{\boldsymbol{\sigma}}(\text{order parameter})}_{\text{first few } z_k} + \underbrace{\delta\boldsymbol{\sigma}(\text{fluctuations})}_{\text{remaining } z_k}. \tag{24}$$

If the base AR model generates the first few z_k first, one can check:

- Do the leading latent variables encode the magnetization?
- Is the conditional $p(z_{k+1}, \dots \mid z_1, \dots, z_k)$ approximately Gaussian for intermediate k (Landau–Ginzburg regime)?
- Does the mutual-information bottleneck—identifying which latent variables carry the most information about the energy—shift at T_c?

6.6 The flow as non-equilibrium dynamics

The flow f_ϕ defines a deterministic map from a simple (high-temperature-like) base distribution to the target Boltzmann distribution. Layer by layer, it "cools" the system. The intermediate distributions

$$q_l(\boldsymbol{\sigma}) = p_\theta\big((f^{(l)} \circ \cdots \circ f^{(1)})^{-1}(\boldsymbol{\sigma})\big), \qquad l = 1, \dots, L, \tag{25}$$

trace out a path in the space of probability distributions. One can measure the free-energy change at each layer,

$$\Delta F_l = F_{\text{var}}[q_l] - F_{\text{var}}[q_{l-1}], \tag{26}$$

and study whether the flow finds a quasi-static (reversible) path or a non-equilibrium shortcut. This has natural connections to Jarzynski's equality and optimal transport.

7 Case study: the 2D Ising model at criticality

We now specialise the general framework to the square-lattice Ising model at $T_c = 2J/\ln\big(1+\sqrt{2}\big) \approx 2.269\, J/k_B$ and develop concrete expectations for what the flow and the latent variables should learn. This system is exactly solvable and described by the $c = 1/2$ minimal-model conformal field theory (CFT), so sharp predictions are possible.

7.1 The challenge: no characteristic scale

Away from T_c, correlations decay as $\langle\sigma_i\sigma_j\rangle \sim e^{-|i-j|/\xi}$ with finite ξ, and a flow of depth $L \gtrsim \xi$ can capture all correlations. At T_c, correlations are power-law,

$$\langle\sigma_i\sigma_j\rangle \sim |i-j|^{-\eta}, \qquad \eta = \tfrac{1}{4}, \tag{27}$$

and $\xi \to \infty$. No finite-depth flow suffices: every layer matters, and the flow can never fully decorrelate the latent variables.

7.2 What the flow does: progressive smoothing of fractal domain walls

At T_c, domain walls are fractal curves described by SLE_3 (Schramm–Loewner evolution with $\kappa = 3$, fractal dimension $d_f = 11/8$). The inverse flow $f_\phi^{-1} : \boldsymbol{\sigma} \to \boldsymbol{z}$ acts as a progressive *simplifier*:

- **Layer 1** (reads black sublattice, flips white): removes roughness at the 1-lattice-spacing scale. Domain walls become slightly smoother.
- **Layer 2** (reads white, flips black): removes roughness at the 2-spacing scale.
- **Layer** l: removes fluctuations at scale $\sim l$.

After L layers, the latent configuration $\boldsymbol{z}$ looks like a *coarse-grained* version of $\boldsymbol{\sigma}$: the fractal domain walls have been straightened up to scale L, but their large-scale topology is preserved.

The forward flow $f_\phi : \boldsymbol{z} \to \boldsymbol{\sigma}$ does the reverse: starting from the smooth latent configuration, it progressively *roughens* the domain walls, adding fractal detail at finer and finer scales—a constructive, layer-by-layer assembly of the critical microstate.

7.3 Self-similarity of the flow at T_c

Because the critical Ising model is scale-invariant, the fluctuations at scale l are statistically identical to those at scale $l+1$ (up to rescaling). This implies:

1. The mask networks at different layers should learn **statistically similar transformations**—each layer does the "same job" at a different scale.
2. The free-energy reduction per layer,
$$\Delta F_l = F_{\mathrm{var}}[q_l] - F_{\mathrm{var}}[q_{l-1}], \tag{28}$$
should be approximately **constant** across layers (or decay as a slow power law $\sim l^{-(2-\eta)}$), reflecting the equal importance of all scales.
3. Away from T_c, ΔF_l drops sharply for $l > \xi$—deep layers become idle. At T_c, *no layer is idle.*

The constant ΔF_l at criticality is the flow-level signature of scale invariance.

7.4 Division of labour between flow and base

The flow and the base AR model split the representational work along a **scale axis**:

$$\underbrace{p_\theta(\boldsymbol{z})}_{\substack{\text{IR physics:}\\ \text{global topology,}\\ \text{order parameter}}} \xrightarrow{f_\phi} \underbrace{q(\boldsymbol{\sigma})}_{\text{full critical distribution}} . \tag{29}$$

- The **flow** f_ϕ handles correlations at scales $\lesssim L_{\mathrm{flow}}$: local spin alignment, domain-wall geometry, short-range order. This is the *UV* (ultraviolet) physics.
- The **base** $p_\theta(\boldsymbol{z})$ handles correlations at scales $\gtrsim L_{\mathrm{flow}}$: the global magnetisation sector, large-scale domain topology, long-range order-parameter correlations. This is the *IR* (infrared) physics.

The flow is a **constructive renormalisation group**: it runs the RG *backwards*, starting from the coarse (IR) description encoded in $\boldsymbol{z}$ and progressively adding fine (UV) detail.

7.5 Physical meaning of the latent variables

The autoregressive base generates $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ sequentially. The first variables condition everything that follows, so the model assigns them the most important, most informative features.

Early variables $(z_1, z_2, \ldots, z_{\text{few}})$: global structure.

- z_1 should encode the **global $\mathbb{Z}_2$ sector**—the overall sign of the magnetisation. At T_c, $p(z_1 = +1) \approx p(z_1 = -1) \approx 1/2$ (maximal uncertainty). This single bit is the most informative feature of a critical configuration.
- The next several z_k encode the **large-scale domain topology**: how many major domains exist, their rough spatial arrangement, the coarse shape of the dominant domain wall.

Middle variables: progressive refinement. These encode the domain-wall geometry at progressively finer scales—the "wavelet coefficients" of the magnetisation field at intermediate wavelengths. Each variable adds detail at a specific scale, like successive terms in a multiscale expansion.

Late variables $(z_{N-\text{few}}, \ldots, z_N)$: thermal noise. These are nearly independent of all preceding variables: $p(z_k \mid z_{<k}) \approx \text{Bernoulli}(1/2)$. They encode the finest-scale fluctuations that the flow could not fully decorrelate—the residual "thermal noise."

7.6 The conditional entropy profile as a diagnostic

The conditional entropy $H(z_k \mid z_{<k})$ measures how much new information each latent variable adds. Its profile is a sharp diagnostic of the phase:

At T_c (critical). $H(z_1) = \ln 2$ (the $\mathbb{Z}_2$ choice is maximally uncertain). $H(z_k \mid z_{<k})$ then decreases **slowly**—as a power law—because scale-invariant correlations mean there is always more structure to specify at the next scale. The information is spread across all scales.

Below T_c (ordered). $H(z_1) \ll \ln 2$ (the magnetisation is nearly determined). The remaining conditional entropies drop quickly to $\approx \ln 2$ as the variables become trivial fluctuations around the ordered state.

Above T_c (disordered). $H(z_k \mid z_{<k}) \approx \ln 2$ for all k—the variables are nearly independent. The base is close to a product distribution, and the flow does little.

The **slow power-law decay of $H(z_k \mid z_{<k})$ at T_c** is the autoregressive fingerprint of criticality.

7.7 Connection to the CFT operator content

The critical 2D Ising model is described by the $c = 1/2$ minimal-model CFT with primary operators:

Operator	Conformal dimension h	Physical meaning
$\mathbb{1}$	0	Identity
σ	1/16	Spin (order parameter)
ε	1/2	Energy density

The most relevant operator (lowest h) is the spin field σ. The latent variables should encode the amplitudes of these operators in a hierarchical way:

- **First latent variables**: amplitude of the spin field σ at the longest wavelength—this is the magnetisation. The low conformal dimension $h = 1/16$ means it is the most slowly decaying, most important mode.
- **Next variables**: amplitude of the energy density ε at long wavelengths, and the spin field at shorter wavelengths.
- **Deeper variables**: descendant operators (spatial derivatives of primaries), encoding finer spatial structure.

The autoregressive ordering effectively implements a **spectral decomposition** of the critical distribution, ordered by relevance (conformal dimension).

7.8 What the latent configuration looks like

If one visualises $\boldsymbol{z}$ on the lattice:

- **At** T_c: $\boldsymbol{z}$ should resemble a **smoothed, coarse-grained** version of $\boldsymbol{\sigma}$. The large-scale domain structure is visible, but the fractal roughness of domain walls has been removed. It looks like the output of a block-spin RG transformation applied $\sim L_{\text{flow}}$ times.
- **Below** T_c: $\boldsymbol{z}$ is nearly uniform (all $+1$ or all -1), since the flow can build the few thermal excitations from a simple base.
- **Above** T_c: $\boldsymbol{z} \approx \boldsymbol{\sigma}$—the flow has little to do, and the latent and physical configurations are nearly identical.

7.9 Summary: the flow as a constructive RG

$$\underbrace{z_1}_{\mathbb{Z}_2 \text{ sector}} \to \underbrace{z_2, \ldots, z_k}_{\text{large-scale topology}} \to \underbrace{z_{k+1}, \ldots, z_K}_{\text{medium-scale detail}} \to \underbrace{z_{K+1}, \ldots, z_N}_{\text{thermal noise}} \xrightarrow{f_\phi} \boldsymbol{\sigma} \tag{30}$$

The base AR model generates a coarse description of the critical configuration, ordered from the most relevant (IR) to the least relevant (UV) degrees of freedom. The flow then dresses this coarse description with the geometric detail at all scales up to the lattice spacing. At T_c, both components are maximally stressed: the base must capture power-law correlations, and the flow must build self-similar fractal structure at every scale.

8 Comparison of approaches

Table 2: Comparison of token-based variational ansätze.

Approach	**Rigorous bound on F?**	**Tractable $\ln q(\boldsymbol{\sigma})$?**	**Expressiveness**
Raw-spin AR [1]	Yes	Exact	Limited by ordering
Block-spin AR (lossless)	Yes	Exact	Limited by block partiti
VQ-VAE + AR prior	Only via loose joint bound	No	High
ELBO + importance sampling	Asymptotically	Estimated	High
Discrete NF + AR base	**Yes**	**Exact**	**High (learned bijectio**

9 Suggested numerical experiments

As a concrete starting point, we propose the following experiments on the **2D Ising model** on an $L \times L$ square lattice with periodic boundary conditions:

$$E(\boldsymbol{\sigma}) = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j. \tag{31}$$

Architecture.

- **Base model:** MADE network over $\boldsymbol{z} \in \{\pm 1\}^N$ with raster-scan ordering, hidden layers of width $4N$.
- **Flow:** $L_{\text{flow}} = 4$–8 coupling layers with alternating checkerboard partitions; each mask network is a small ConvNet (2–3 layers, 3×3 kernels).
- **System sizes:** 8×8, 16×16, 32×32.

Training. STE for flow gradients, REINFORCE with learned baseline for base-model gradients. Sweep temperatures across $T_c \approx 2.269\, J/k_B$.

Diagnostics.

1. Compare $F_{\text{var}}[q]$ against exact results (transfer matrix for small L, Monte Carlo for larger L).
2. Plot base-distribution entropy $H[p_\theta]$ and layer-by-layer free-energy reduction ΔF_l as functions of T.
3. Feed configurations with known defect content (single domain wall, vortex pair) through f_ϕ^{-1}; visualize the latent representation.
4. Measure the minimum flow depth L^* needed for a target accuracy $|F_{\text{var}} - F_{\text{true}}|/N < \epsilon$ and plot $L^*(T)$.
5. At T_c, compare the learned flow with the Kramers–Wannier transformation on test configurations.

10 Related work

The present proposal sits at the intersection of three lines of work: autoregressive variational methods for statistical mechanics, normalizing flows for lattice field theory, and discrete generative models in machine learning. We survey each in turn and identify the gap that the discrete flow framework fills.

10.1 Autoregressive variational ansätze for spin systems

Wu, Wang, and Zhang [1] introduced the use of autoregressive neural networks (PixelCNN, MADE) as variational ansätze for classical statistical mechanics, with exact log-probabilities enabling a rigorous variational free-energy bound. Subsequent work has refined the autoregressive architecture: Biazzo, Wu, and Carleo [2] proposed the TwoBo architecture, incorporating knowledge of the two-body interaction structure into sparse autoregressive networks for frustrated systems with >1000 spins. Bialas, Korcyl, and Stebel [3] introduced hierarchical associations between spins and neurons, achieving scaling with the linear extent L rather than the

total number of spins. Pan and Zhang [4] augmented variational autoregressive networks with message passing to better capture spin-spin interactions. All of these works use *purely autoregressive* distributions without flow layers; the expressiveness is limited by the autoregressive factorization itself.

10.2 Normalizing flows for lattice field theory

Albergo, Kanwar, and Shanahan [5] pioneered the use of normalizing flows (specifically RealNVP affine coupling layers with checkerboard masking) for lattice field theory, demonstrating the method on ϕ^4 scalar theory in 2D with *continuous* field variables. Kanwar et al. [6] extended this to gauge-equivariant flows for U(1) lattice gauge theory. Nicoli et al. [7, 8] showed that normalizing-flow samplers can estimate absolute free energies (not just differences) for continuous-field theories. Comprehensive reviews of this programme are given in Refs. [9, 10]. A key limitation of this entire line of work is its restriction to *continuous* degrees of freedom; discrete spin models are not directly addressed.

10.3 Neural network renormalization group

Li and Wang [11] proposed the Neural Network Renormalization Group (NeuralRG), which uses normalizing flows in a hierarchical, multiscale architecture motivated by the RG. The variational free energy provides a rigorous upper bound. Applied to the 2D Ising model, the method uses RealNVP coupling layers and therefore operates with a *continuous relaxation* of the binary spin variables, rather than natively discrete transformations.

10.4 Discrete normalizing flows in machine learning

Tran et al. [12] introduced discrete normalizing flows with two architectures: discrete autoregressive flows and discrete bipartite flows. The bipartite flow uses a modular location-scale transform that reduces to XOR for binary variables—precisely the coupling layer (12) used here. They tested on Potts models as a benchmark, evaluating negative log-likelihood as a function of coupling strength. However, their work frames the problem as *density estimation* (maximum likelihood on samples), not as a variational statistical-mechanics problem: there is no free-energy objective, no rigorous bound on F, and no physical analysis of the learned flow. Hoogeboom et al. [13] proposed Integer Discrete Flows for lossless compression, using additive coupling with rounding for ordinal data, but did not target spin systems.

10.5 Boltzmann generators

Noé et al. [14] introduced Boltzmann generators—normalizing flows trained on the energy function to sample Boltzmann distributions for molecular systems. This work operates entirely in continuous configuration space (molecular coordinates) and does not address discrete degrees of freedom.

10.6 Discrete flow matching and diffusion for spin systems

More recently, Tuo et al. [15] applied discrete flow matching to the Ising model, learning continuous-time transport maps from noisy distributions to the Boltzmann distribution. This is fundamentally different from the bijective coupling-layer approach: flow matching does not use invertible transformations and does not provide the same type of exact variational bounds. Ghio et al. [16] provided a theoretical analysis of flows, diffusion, and autoregressive methods for spin glasses, showing that these methods can encounter first-order phase transitions along the generative path that impede sampling—an important caveat for any flow-based approach in the glassy regime.

10.7 Positioning of the present proposal

Table 3 summarises the landscape. Each existing approach addresses a subset of the desiderata; the present proposal combines discrete bijectivity (from [12]), the variational free-energy framework (from [1]), and the physical interpretability of normalizing flows (from [11, 5]). What is new is the synthesis: discrete coupling layers operating *natively* on binary spins, composed with an autoregressive base, trained via the variational free energy with rigorous bounds, and analysed for physical content (flow depth scaling, duality, defect encoding).

Table 3: Positioning of the discrete NF + AR base proposal relative to existing work.

	Discrete spins	**Bijective flow**	**Rigorous F bound**	**Learned transform**
Wu et al. [1]	✓	—	✓	—
Albergo et al. [5]	—	✓	✓	✓
Li & Wang [11]	relaxed	✓	✓	✓
Tran et al. [12]	✓	✓	—	✓
Tuo et al. [15]	✓	—	—	✓
This proposal	✓	✓	✓	✓

11 Outlook

Several extensions are immediate:

- **q-state Potts model.** The coupling layers generalize to conditional permutations of $\{1, \ldots, q\}$, parameterized by a network that outputs a permutation matrix (or its Gumbel–Sinkhorn relaxation).
- **Continuous spins (XY, Heisenberg).** Standard continuous normalizing flows apply directly, with the full machinery of coupling layers and affine transformations.
- **Lattice gauge theories.** The coupling layers can be adapted to respect gauge symmetry by acting on gauge-invariant combinations (plaquettes, Wilson loops), following the approach of Albergo et al.
- **Quantum systems.** The variational free energy can be replaced by the variational energy of a quantum Hamiltonian, with the autoregressive model representing an amplitude (autoregressive neural quantum state).

The discrete normalizing flow framework provides a principled, rigorous, and physically interpretable generalization of autoregressive variational methods. The learned bijection is not merely a computational device—it is a window into the structure of the Boltzmann distribution, encoding renormalization, duality, and the organization of topological defects.

References

[1] D. Wu, L. Wang, and P. Zhang, "Solving statistical mechanics using variational autoregressive networks," *Phys. Rev. Lett.* **122**, 080602 (2019). [arXiv:1809.10606]

[2] I. Biazzo, D. Wu, and G. Carleo, "Sparse autoregressive neural networks for classical spin systems," *Mach. Learn.: Sci. Technol.* **5**, 045001 (2024). [arXiv:2402.16579]

[3] P. Bialas, P. Korcyl, and T. Stebel, "Hierarchical autoregressive neural networks for statistical systems," *Comput. Phys. Commun.* **281**, 108502 (2022).

[4] F. Pan and P. Zhang, "Message passing variational autoregressive network for solving intractable Ising models," *Commun. Phys.* **7**, 205 (2024).

[5] M. S. Albergo, G. Kanwar, and P. E. Shanahan, "Flow-based generative models for Markov chain Monte Carlo in lattice field theory," *Phys. Rev. D* **100**, 034515 (2019). [arXiv:1904.12072]

[6] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, "Equivariant flow-based sampling for lattice gauge theory," *Phys. Rev. Lett.* **125**, 121601 (2020). [arXiv:2003.06413]

[7] 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). [arXiv:1910.13496]

[8] K. A. Nicoli, C. J. Anders, L. Funcke, T. Hartung, K. Jansen, S. Kessel, S. Nakajima, and P. Stornati, "Estimation of thermodynamic observables in lattice field theories with deep generative models," *Phys. Rev. Lett.* **126**, 032001 (2021). [arXiv:2007.07115]

[9] M. S. Albergo, D. Boyda, D. C. Hackett, G. Kanwar, K. Cranmer, S. Racanière, D. J. Rezende, and P. E. Shanahan, "Introduction to normalizing flows for lattice field theory," arXiv:2101.08176 (2021).

[10] K. Cranmer, G. Kanwar, S. Racanière, D. J. Rezende, and P. E. Shanahan, "Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics," *Nat. Rev. Phys.* **5**, 526–535 (2023).

[11] S.-H. Li and L. Wang, "Neural network renormalization group," *Phys. Rev. Lett.* **121**, 260601 (2018). [arXiv:1802.02840]

[12] D. Tran, K. Vafa, K. K. Agrawal, L. Dinh, and B. Poole, "Discrete flows: Invertible generative models of discrete data," in *Advances in Neural Information Processing Systems 32* (NeurIPS 2019). [arXiv:1905.10347]

[13] E. Hoogeboom, J. W. T. Peters, R. van den Berg, and M. Welling, "Integer discrete flows and lossless compression," in *Advances in Neural Information Processing Systems 32* (NeurIPS 2019). [arXiv:1905.07376]

[14] 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). [arXiv:1812.01729]

[15] H. Tuo, H. Zeng, Y. Chen, and L. Cheng, "Scalable multitemperature free energy sampling of classical Ising spin states," arXiv:2503.08063 (2025).

[16] D. Ghio, Y. M. Dandi, F. Krzakala, and L. Zdeborová, "Sampling with flows, diffusion, and autoregressive neural networks: A spin-glass perspective," *Proc. Natl. Acad. Sci. USA* **121**, e2311810121 (2024). [arXiv:2308.14085]