# Discrete Normalizing Flows as Variational Ansätze for Classical Statistical Mechanics

Notes

February 9, 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 2**

**2 Motivation: autoregression over transformed tokens 3**

**3 The rigour problem with latent-variable models 3**

3.1 Attempted fix: joint variational bound . . . 3

3.2 Attempted fix: ELBO + importance sampling . . . 4

**4 Discrete normalizing flows: the rigorous solution 4**

4.1 Setup . . . 4

4.2 Coupling layers for binary spins . . . 4

4.3 Enforcing $\mathbb{Z}_2$ spin-flip symmetry . . . 5

4.4 Physical interpretation of a coupling layer . . . 6

**5 Joint training procedure 7**

5.1 Gradient with respect to $\theta$ (base AR model) . . . 7

5.2 Gradient with respect to $\phi$ (flow parameters) . . . 7

5.3 Training algorithm . . . 9

**6 Physical content of the learned flow 9**

6.1 Emergent renormalization group . . . 9

6.2 Kramers–Wannier duality . . . 9

6.3 Topological defect encoding . . . 10

6.4 Flow depth as a probe of computational complexity . . . 10

6.5 Disentanglement of order parameter and fluctuations . . . 10

6.6 The flow as non-equilibrium dynamics . . . 10

**7 Case study: the 2D Ising model at criticality 11**
7.1 The challenge: no characteristic scale . . . . . 11
7.2 What the flow does: progressive smoothing of fractal domain walls . . . . . 11
7.3 Self-similarity of the flow at $T_c$ . . . . . 11
7.4 Division of labour between flow and base . . . . . 12
7.5 Physical meaning of the latent variables . . . . . 12
7.6 The conditional entropy profile as a diagnostic . . . . . 12
7.7 Connection to the CFT operator content . . . . . 13
7.8 What the latent configuration looks like . . . . . 13
7.9 Summary: the flow as a constructive RG . . . . . 13

**8 Preliminary numerical results 14**
8.1 Setup . . . . . 14
8.2 Results . . . . . 14
8.3 The entropy-preservation theorem for discrete flows . . . . . 14
8.4 REINFORCE mode collapse as the fundamental bottleneck . . . . . 15
8.5 Implications for the discrete flow framework . . . . . 15
8.6 Forward KL training: creating the right regime for the flow . . . . . 16

**9 Beyond bijective flows: relaxing the bijectivity constraint 18**
9.1 Markov-corrected variational ansatz . . . . . 18
9.2 Adding bijective layers: stochastic discrete normalizing flows . . . . . 24
9.3 $k$-to-1 surjections via symmetry . . . . . 24

**10 Comparison of approaches 25**

**11 Haar wavelet basis for the 2D Ising model 25**
11.1 Binary Haar wavelet transform . . . . . 26
11.2 Energy decomposition in the Haar basis . . . . . 27
11.3 Autoregressive factorisation in the Haar basis . . . . . 27
11.4 Implementation plan . . . . . 28

**12 Suggested numerical experiments 30**

**13 Related work 30**
13.1 Autoregressive variational ansätze for spin systems . . . . . 30
13.2 Normalizing flows for lattice field theory . . . . . 31
13.3 Neural network renormalization group . . . . . 31
13.4 Discrete normalizing flows in machine learning . . . . . 31
13.5 Boltzmann generators . . . . . 31
13.6 Discrete flow matching and diffusion for spin systems . . . . . 31
13.7 Positioning of the present proposal . . . . . 32

**14 Outlook 32**

# 1 Background: autoregressive variational free energy

Consider a classical spin system on a lattice of $N$ sites with configuration $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)$, 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

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

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

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 $q_\theta$. Gradients are computed via the REINFORCE (policy gradient) estimator:

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

## 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 $F_{\mathrm{true}}$. 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 $\theta$.
- A **discrete flow** $f_\phi : \{\pm 1\}^N \to \{\pm 1\}^N$, a learnable bijection parameterized by $\phi$.

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

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

### 4.2 Coupling layers for binary spins

We construct $f_\phi$ 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 Enforcing $\mathbb{Z}_2$ spin-flip symmetry

The Ising Hamiltonian is invariant under the global spin flip $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$, so the Boltzmann distribution satisfies $p(\boldsymbol{\sigma}) = p(-\boldsymbol{\sigma})$. A variational ansatz that respects this $\mathbb{Z}_2$ symmetry exactly has two advantages: (i) it halves the effective configuration space the model must learn, and (ii) it prevents mode collapse to a single magnetisation sector.

**$\mathbb{Z}_2$-symmetric base distribution.** Given a bare autoregressive model $p_\theta^{\mathrm{AR}}(\boldsymbol{z})$, we define the symmetrised base distribution

$$p_\theta(\boldsymbol{z}) = \tfrac{1}{2}\, p_\theta^{\mathrm{AR}}(\boldsymbol{z}) + \tfrac{1}{2}\, p_\theta^{\mathrm{AR}}(-\boldsymbol{z}), \tag{15}$$

which satisfies $p_\theta(\boldsymbol{z}) = p_\theta(-\boldsymbol{z})$ by construction. The log-probability is evaluated via the log-sum-exp identity:

$$\ln p_\theta(\boldsymbol{z}) = \mathrm{logsumexp}\big(\ln p_\theta^{\mathrm{AR}}(\boldsymbol{z}),\ \ln p_\theta^{\mathrm{AR}}(-\boldsymbol{z})\big) - \ln 2. \tag{16}$$

Sampling proceeds by drawing $\boldsymbol{z} \sim p_\theta^{\mathrm{AR}}$ and then flipping all spins with probability $\frac{1}{2}$.

**$\mathbb{Z}_2$-equivariant flow.** For the full variational distribution $q(\boldsymbol{\sigma}) = p_\theta(f_\phi^{-1}(\boldsymbol{\sigma}))$ to inherit the symmetry, the flow must be *equivariant*: $f_\phi(-\boldsymbol{z}) = -f_\phi(\boldsymbol{z})$. Combined with (15), this gives

$$q(-\boldsymbol{\sigma}) = p_\theta\big(f_\phi^{-1}(-\boldsymbol{\sigma})\big) = p_\theta\big(-f_\phi^{-1}(\boldsymbol{\sigma})\big) = p_\theta\big(f_\phi^{-1}(\boldsymbol{\sigma})\big) = q(\boldsymbol{\sigma}). \tag{17}$$

**Equivariant coupling layers.** Recall that a coupling layer acts as $\sigma_A = z_A$ and $\sigma_B = z_B \odot m_\phi(z_A)$. Under $\boldsymbol{z} \to -\boldsymbol{z}$:

- The $A$-sublattice maps to $-z_A$ (passthrough), giving $\sigma_A = -z_A$.
- The $B$-sublattice gives $\sigma_B = (-z_B) \odot m_\phi(-z_A)$.

For equivariance we need $\sigma_B = -z_B \odot m_\phi(z_A)$, which requires

$$m_\phi(-z_A) = m_\phi(z_A), \tag{18}$$

i.e., the mask network must be an *even function* of its input.

**Symmetrised mask network.** For any neural network $h_\phi$, define

$$g_\phi(x) = h_\phi(x) + h_\phi(-x). \tag{19}$$

Then $g_\phi(-x) = h_\phi(-x) + h_\phi(x) = g_\phi(x)$, so condition (18) is satisfied exactly. The mask is then $m_\phi(z_A) = \mathrm{sign}(g_\phi(z_A))$. This construction doubles the computational cost of the mask network (two forward passes per layer) but imposes no restriction on the network architecture $h_\phi$.

*Remark* 1. For ±1 spins, one might consider using $|z_A|$ or $z_A^2$ as input to enforce evenness. However, since $|z_i| = z_i^2 = 1$ for all $z_i \in \{\pm 1\}$, these are constant and carry no information. The symmetrisation (19) is the simplest approach that preserves the full expressiveness of the mask network.

### 4.4 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

$$m_i = \mathrm{sign}(h_i) \tag{20}$$

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 | $\sim 3$ lattice spacings |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $l$ | — | — | $\sim l$ lattice spacings |

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

$$\xi_{\text{eff}} \sim L_{\text{flow}}. \tag{21}$$

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 $f_\phi^{-1}$ 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 (20) 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 $\theta$ (base AR model) and $\phi$ (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 $\theta$ (base AR model)

Both the sampling distribution and the integrand depend on $\theta$. Applying the REINFORCE identity:

$$\nabla_\theta F_{\text{var}} = \mathbb{E}_{p_\theta}\Big[\Big(E\big(f_\phi(\boldsymbol{z})\big) + T\ln p_\theta(\boldsymbol{z})\Big)\,\nabla_\theta \ln p_\theta(\boldsymbol{z})\Big]. \tag{22}$$

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

$$\nabla_\theta F_{\text{var}} \approx \frac{1}{M}\sum_{m=1}^{M}\Big(R^{(m)} - b\Big)\,\nabla_\theta \ln p_\theta(\boldsymbol{z}^{(m)}), \qquad R^{(m)} = E(\boldsymbol{\sigma}^{(m)}) + T\ln p_\theta(\boldsymbol{z}^{(m)}). \tag{23}$$

## 5.2 Gradient with respect to $\phi$ (flow parameters)

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

$$F_{\text{var}} = \mathbb{E}_{\boldsymbol{z}\sim p_\theta}\Big[\underbrace{E\big(f_\phi(\boldsymbol{z})\big)}_{\text{depends on }\phi} + \underbrace{T\ln p_\theta(\boldsymbol{z})}_{\text{independent of }\phi}\Big]. \tag{24}$$

Two key observations:

1. The sampling distribution $p_\theta(\boldsymbol{z})$ does not depend on $\phi$, so we can move $\nabla_\phi$ inside the expectation without a REINFORCE correction.

2. The entropy term $T \ln p_\theta(\boldsymbol{z})$ does not depend on $\phi$ at all—the flow does not change which latent configuration was drawn, only where it lands in physical space.

Therefore:

$$\nabla_\phi F_{\text{var}} = \mathbb{E}_{p_\theta}\Big[\nabla_\phi E\big(f_\phi(\boldsymbol{z})\big)\Big]. \tag{25}$$

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 (25) looks like a standard backpropagation problem, but there is a fundamental obstacle. Inside each coupling layer, the mask is computed as

$$m_i = \operatorname{sign}\big(g_\phi(z_A)_i\big) \in \{\pm 1\}, \tag{26}$$

where $g_\phi$ is a neural network with continuous outputs. As $\phi$ varies continuously, $g_\phi$ changes smoothly, but $\operatorname{sign}(\cdot)$ snaps to $\pm 1$. The composed map $f_\phi(\boldsymbol{z})$ is therefore a *piecewise-constant* function of $\phi$: 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 $\operatorname{sign}'$ by the identity, so that gradients flow through the mask network as if $m_i \approx g_\phi(z_A)_i$:

$$\left.\frac{\partial m_i}{\partial g_i}\right|_{\text{STE}} = 1. \tag{27}$$

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:

$$\tilde{m}_i = \tanh\!\left(\frac{g_\phi(z_A)_i + (\xi_1 - \xi_2)}{\tau}\right), \qquad \xi_{1,2} \sim \text{Gumbel}(0,1), \tag{28}$$

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

**Decoupling property.** A key simplification: the gradient for $\theta$ does not require differentiating through the flow, and the gradient for $\phi$ 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 $p_\theta$, 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 $\theta$: $\theta \leftarrow \theta - \alpha_\theta \cdot \frac{1}{M} \sum_m (R^{(m)} - b) \nabla_\theta \ln p_\theta(\boldsymbol{z}^{(m)})$    ▷ REINFORCE
    Update $\phi$: $\phi \leftarrow \phi - \alpha_\phi \cdot \frac{1}{M} \sum_m \nabla_\phi E(\tilde{f}_\phi(\boldsymbol{z}^{(m)}))$    ▷ STE backward pass
end for
```

## 6 Physical content of the learned flow

The learned bijection $f_\phi$ 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 $2\times2$ blocks, layer 3 on $4 \times 4$ 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 $f_\phi^{-1}$ 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:

$$\boldsymbol{\sigma} = \underbrace{\bar{\boldsymbol{\sigma}}(\text{order parameter})}_{\text{first few } z_k} + \underbrace{\delta\boldsymbol{\sigma}(\text{fluctuations})}_{\text{remaining } z_k}. \tag{29}$$

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 $f_\phi$ 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

$$q_l(\boldsymbol{\sigma}) = p_\theta\big((f^{(l)} \circ \cdots \circ f^{(1)})^{-1}(\boldsymbol{\sigma})\big), \qquad l = 1, \dots, L, \tag{30}$$

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

$$\Delta F_l = F_{\text{var}}[q_l] - F_{\text{var}}[q_{l-1}], \tag{31}$$

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 $T_c = 2J/\ln\bigl(1+\sqrt{2}\bigr) \approx 2.269\,J/k_B$ 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 $\xi$, and a flow of depth $L \gtrsim \xi$ can capture all correlations. At $T_c$, correlations are power-law,

$$\langle\sigma_i\sigma_j\rangle \sim |i-j|^{-\eta}, \qquad \eta = \tfrac{1}{4}, \tag{32}$$

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,

   $$\Delta F_l = F_{\text{var}}[q_l] - F_{\text{var}}[q_{l-1}], \tag{33}$$

   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 $T_c$, $\Delta F_l$ drops sharply for $l > \xi$—deep layers become idle. At $T_c$, *no layer is idle.*

The constant $\Delta F_l$ 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**:

$$\underbrace{p_\theta(\boldsymbol{z})}_{\substack{\text{IR physics:}\\ \text{global topology,}\\ \text{order parameter}}} \quad \xrightarrow{f_\phi} \quad \underbrace{q(\boldsymbol{\sigma})}_{\text{full critical distribution}} \quad . \tag{34}$$

- The **flow** $f_\phi$ handles correlations at scales $\lesssim L_{\text{flow}}$: 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 $\gtrsim L_{\text{flow}}$: 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 |
| $\sigma$ | 1/16 | Spin (order parameter) |
| $\varepsilon$ | 1/2 | Energy density |

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

- **First latent variables**: amplitude of the spin field $\sigma$ 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 $\varepsilon$ 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

$$\underbrace{z_1}_{\mathbb{Z}_2\text{ sector}} \;\to\; \underbrace{z_2, \ldots, z_k}_{\text{large-scale topology}} \;\to\; \underbrace{z_{k+1}, \ldots, z_K}_{\text{medium-scale detail}} \;\to\; \underbrace{z_{K+1}, \ldots, z_N}_{\text{thermal noise}} \xrightarrow{f_\phi} \boldsymbol{\sigma} \tag{35}$$

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 Preliminary numerical results

We report results from a JAX/Flax implementation of the discrete flow framework applied to the 2D ferromagnetic Ising model on a $32 \times 32$ square lattice with periodic boundary conditions at $T_c \approx 2.269\, J/k_B$. All runs use $\mathbb{Z}_2$ symmetry (Sec. 4.3).

## 8.1 Setup

- **Base model (AR):** MADE with one hidden layer of 512 units ($\sim 1.05 \times 10^6$ parameters).
- **Flow (AR+flow):** 4 coupling layers with alternating checkerboard partitions; each mask network is an 8-hidden-layer ConvNet with 64 channels per layer and $3\times3$ kernels (receptive field $19\times19$ per layer; $\sim 1.04\times10^6$ flow parameters, matching the MADE parameter count).
- **Training:** REINFORCE for $\theta$, STE for $\phi$; Adam optimiser with $\alpha_\theta = \alpha_\phi = 10^{-3}$; batch size 500; 10 000 steps.

Exact reference values are computed via the Kaufman transfer-matrix formula for the $32 \times 32$ periodic lattice.

## 8.2 Results

Table 2: Variational free energy, energy, and entropy per site for the $32 \times 32$ Ising model at $T_c$, with $\mathbb{Z}_2$ symmetry. "AR" denotes the MADE-only baseline; "AR+flow" adds 4 discrete coupling layers.

| | $F_{\text{var}}/N$ | $\langle E\rangle/N$ | $S/N$ |
|---|---|---|---|
| Exact (Kaufman) | $-2.111$ | $-1.434$ | 0.298 |
| AR only | $-2.075$ | $-1.735$ | 0.150 |
| AR + flow | $-2.076$ | $-1.728$ | 0.153 |

Two observations stand out:

1. The gap to the exact solution, $\Delta F/N \approx 0.036$, is dominated by an *entropy deficit*: $S/N = 0.15$ versus the exact 0.30. The model oversamples ordered, low-energy configurations ($\langle E\rangle/N = -1.73$ versus $-1.43$) at the expense of the exponentially many disordered configurations that contribute to the entropy at $T_c$.
2. Adding discrete coupling layers with comparable parameter count does **not** improve the variational free energy.

## 8.3 The entropy-preservation theorem for discrete flows

The second observation has a simple structural explanation.

**Proposition 2** (Entropy preservation). *Let $f_\phi : \{\pm1\}^N \to \{\pm1\}^N$ be any bijection, and let $q(\boldsymbol{\sigma}) = p_\theta(f_\phi^{-1}(\boldsymbol{\sigma}))$. Then*

$$H[q] = H[p_\theta]. \tag{36}$$

*Proof.* By change of summation variable $\boldsymbol{z} = f_\phi^{-1}(\boldsymbol{\sigma})$:

$$H[q] = -\sum_{\boldsymbol{\sigma}} q(\boldsymbol{\sigma}) \ln q(\boldsymbol{\sigma}) = -\sum_{\boldsymbol{z}} p_\theta(\boldsymbol{z}) \ln p_\theta(\boldsymbol{z}) = H[p_\theta]. \tag{37}$$

□

This is obvious in retrospect—a bijection on a finite set is a permutation of probability masses, which cannot change the entropy—but its consequences for the variational framework are important. Decomposing the variational free energy:

$$F_{\text{var}} = \underbrace{\langle E(f_\phi(\boldsymbol{z}))\rangle_{p_\theta}}_{\text{flow can optimise}} - T \underbrace{H[p_\theta]}_{\text{flow cannot change}} . \tag{38}$$

The flow can only reduce $F_{\text{var}}$ by lowering the expected energy—it rearranges which latent configurations map to which physical configurations, routing high-probability $\boldsymbol{z}$'s to low-energy $\boldsymbol{\sigma}$'s. But it *cannot increase the entropy* of the variational distribution.

## 8.4 REINFORCE mode collapse as the fundamental bottleneck

The entropy deficit in Table 2 is a symptom of *mode collapse* in the REINFORCE training of the base distribution. The REINFORCE gradient estimator

$$\nabla_\theta F_{\text{var}} \approx \frac{1}{M} \sum_{m=1}^{M} \left(R^{(m)} - b\right) \nabla_\theta \ln p_\theta(\boldsymbol{z}^{(m)}) \tag{39}$$

creates a positive feedback loop:

1. Samples that happen to have low energy receive large negative advantage ⇒ REINFORCE increases their probability.
2. Higher probability for these configurations ⇒ they are sampled even more frequently.
3. Disordered configurations (which are exponentially more numerous but individually less probable) are undersampled ⇒ their gradient contribution vanishes.
4. The entropy collapses as the model concentrates on a shrinking set of ordered configurations.

The entropy term $T \ln q$ in the reward provides a self-correcting signal—as $q$ concentrates, $\ln q$ becomes less negative, worsening the reward. But with batch size $M = 500$ in a configuration space of size $2^{1024}$, the gradient variance is enormous, and this correction is overwhelmed by noise.

This failure mode is intrinsic to the reverse KL divergence optimised by the variational free energy: $\mathrm{D}_{\mathrm{KL}}(q \| p_{\mathrm{Boltz}})$ is mode-seeking, preferring to concentrate on a single mode rather than spread across all modes. It is particularly severe at $T_c$ where the Boltzmann distribution is supported on an exponentially large set of nearly degenerate domain configurations.

## 8.5 Implications for the discrete flow framework

The entropy-preservation theorem (36) reveals a fundamental division of labour:

- The **base distribution** $p_\theta$ controls the entropy budget—how much of configuration space the variational distribution covers.
- The **flow** $f_\phi$ controls the energy efficiency—given the entropy budget, it routes probability mass to the lowest-energy configurations.

The flow is beneficial when $p_\theta$ has high entropy but assigns probability to energetically suboptimal configurations. It is *useless* when $p_\theta$ has already mode-collapsed: rearranging probability mass within a low-entropy distribution cannot recover the missing entropy.

This motivates several directions for improving the framework:

1. **Entropy regularisation**: explicitly penalise low entropy in $p_\theta$ during training, ensuring the base maintains sufficient diversity for the flow to work with.

2. **Beta annealing**: start training at high temperature ($\beta \to 0$) where the entropy is naturally high, and gradually cool to the target $T_c$. Our experiments with beta annealing ($\beta_{\text{anneal}} = 0.998$) did yield higher entropy ($S/N = 0.21$) but at the cost of a larger overall gap ($\Delta F/N = 0.11$), suggesting the annealing schedule requires careful tuning.

3. **Larger batch sizes**: reducing REINFORCE variance to allow the entropy self-correction signal to compete with the energy-lowering signal.

4. **Forward KL training**: replacing the mode-seeking reverse KL with the mode-covering forward KL, using MCMC samples from the Boltzmann distribution as training data (see Sec. 8.6).

## 8.6 Forward KL training: creating the right regime for the flow

The fundamental bottleneck identified above is that REINFORCE-based reverse KL training collapses the entropy of the base distribution, leaving the flow with nothing to improve. Forward KL training offers a qualitatively different approach that creates exactly the conditions under which the discrete flow becomes beneficial.

**Forward KL objective.** Instead of minimising $\mathrm{D_{KL}}(q \| p_{\text{Boltz}})$ (reverse KL), we minimise the forward KL divergence

$$\mathrm{D_{KL}}(p_{\text{Boltz}} \| q) = \sum_{\boldsymbol{\sigma}} p_{\text{Boltz}}(\boldsymbol{\sigma}) \ln \frac{p_{\text{Boltz}}(\boldsymbol{\sigma})}{q(\boldsymbol{\sigma})}, \tag{40}$$

which, up to the constant $-H[p_{\text{Boltz}}]$, is equivalent to minimising the negative log-likelihood (NLL) on samples drawn from the Boltzmann distribution:

$$\mathcal{L}_{\text{NLL}}(\theta, \phi) = -\mathbb{E}_{\boldsymbol{\sigma} \sim p_{\text{Boltz}}}\big[\ln q(\boldsymbol{\sigma})\big] = -\mathbb{E}_{\boldsymbol{\sigma} \sim p_{\text{Boltz}}}\big[\ln p_\theta\big(f_\phi^{-1}(\boldsymbol{\sigma})\big)\big]. \tag{41}$$

Training samples $\{\boldsymbol{\sigma}^{(m)}\}$ are obtained from MCMC (e.g. Wolff cluster algorithm) and the gradients are computed by standard backpropagation through the exact log-probability $\ln p_\theta(f_\phi^{-1}(\boldsymbol{\sigma}))$—no REINFORCE is needed.

**Key advantages.**

1. **Mode-covering behaviour.** The forward KL $\mathrm{D_{KL}}(p_{\text{Boltz}} \| q)$ is *mode-covering*: it penalises $q$ wherever $p_{\text{Boltz}}$ has support, forcing the model to spread probability across all relevant configurations rather than collapse onto a single mode. This directly prevents the entropy collapse that plagues REINFORCE training.

2. **Low-variance gradients.** Maximum likelihood estimation on MCMC samples produces gradients with much lower variance than REINFORCE. The gradient $\nabla_\theta \ln p_\theta(\boldsymbol{z})$ is a simple score function evaluated on a fixed sample—no reward signal, no baseline subtraction, no advantage estimation.

3. **Correct entropy by construction.** Since the training data comes from the Boltzmann distribution, which has entropy $S_{\text{exact}}$, the model naturally learns to match this entropy level. The resulting base distribution $p_\theta$ will have high entropy and assign probability to energetically suboptimal configurations.

**The ideal regime for the flow.** Forward KL training creates exactly the conditions under which the discrete flow becomes essential. After training with NLL on MCMC samples, the base distribution $p_\theta$ has:

- **High entropy**: $H[p_\theta] \approx S_{\text{exact}}$, covering the full support of the Boltzmann distribution.
- **Suboptimal energy**: the autoregressive factorisation cannot perfectly capture the spatial correlations in $p_{\text{Boltz}}$, so the energy $\langle E(f_\phi(\boldsymbol{z}))\rangle$ is higher than optimal.

This is precisely the regime where the entropy-preservation theorem (36) is not a limitation. The flow's role is to *rearrange* the high-entropy probability mass, routing it toward lower-energy configurations without compressing it—lowering $\langle E\rangle$ while preserving the entropy budget.

**Empirical evidence: Tran et al. on the Potts model.** The prediction that forward KL training enables the flow to improve over the autoregressive baseline is supported by the experiments of Tran et al. [12], who trained discrete autoregressive flows on the 2D $q$-state Potts model using maximum likelihood on Metropolis–Hastings samples. On small lattices ($3 \times 3$ and $4 \times 4$, with $q = 3, 4, 5$), the discrete flow consistently matched or improved over the autoregressive baseline in terms of negative log-likelihood (NLL). The gains were largest at weak coupling ($J = 0.1$), where the base distribution has high entropy but poor spatial correlations—precisely the regime identified above. At the strongest coupling ($J = 0.5$, $3 \times 3$, $q = 3$), the flow provided no improvement, consistent with the base already capturing the correlations of this small, strongly ordered system.

These results provide a useful contrast with our reverse KL experiments in Table 2. The key difference is the training objective: Tran et al. train on MCMC samples (forward KL), which preserves entropy and creates the conditions for the flow to help. Our REINFORCE-based training (reverse KL) collapses the entropy first, leaving the flow with nothing to improve. However, the Tran et al. experiments have significant limitations: the systems are tiny (9–16 spins), the flow network is a lookup table that cannot scale, no comparison to exact partition functions is provided, and—crucially—no variational free-energy bound is produced. The hybrid approach proposed below addresses these gaps.

**Hybrid training: Forward KL → Reverse KL.** A practical training protocol combines both objectives:

1. **Phase 1 (Forward KL):** Train both $\theta$ and $\phi$ on MCMC samples via the NLL objective (41). This establishes the correct entropy level and teaches the flow to improve spatial correlations.
2. **Phase 2 (Reverse KL):** Fine-tune by minimising the variational free energy $F_{\text{var}}$ via REINFORCE (for $\theta$) and STE (for $\phi$). Starting from the high-entropy initialisation of Phase 1, the entropy self-correction signal in the REINFORCE reward is now strong enough to prevent collapse, and the variational bound is tightened.

Phase 1 solves the cold-start problem: instead of REINFORCE exploring a $2^N$-dimensional space from scratch, it begins with a model that already assigns reasonable probability to the correct configurations. Phase 2 recovers the rigorous variational bound, which is not guaranteed by the forward KL alone.

**Applicability and limitations.** Forward KL training requires an efficient MCMC sampler for the target distribution. For the 2D ferromagnetic Ising model, the Wolff cluster algorithm has a dynamic critical exponent $z \approx 0.25$, making MCMC nearly free even at $T_c$. More generally, forward KL is applicable whenever:

- Local or cluster MCMC updates can decorrelate configurations in polynomial time (unfrustrated spin systems, some lattice field theories).
- Parallel tempering or other enhanced sampling methods provide sufficiently diverse samples.

For *frustrated* or *glassy* systems—where MCMC itself suffers from exponential autocorrelation times—forward KL training is not viable, and the pure reverse KL approach (with improvements such as entropy regularisation and annealing) remains the only option. This creates an interesting dichotomy: the discrete flow framework is most straightforwardly useful (via forward KL) for systems where MCMC works but the variational ansatz benefits from a learned change of variables, and most needed but hardest to train (via reverse KL) for systems where MCMC fails.

# 9 Beyond bijective flows: relaxing the bijectivity constraint

The entropy-preservation theorem (Proposition 2) identifies a fundamental limitation of bijective discrete flows: the flow cannot increase the entropy of the variational distribution, so the entire entropy budget must come from the base $p_\theta$. A natural question is whether the bijectivity constraint can be relaxed while maintaining the rigorous variational bound on $F_{\text{true}}$.

The standard relaxations from the latent-variable literature fail: the ELBO provides a *lower* bound on $\ln q(\boldsymbol{\sigma})$, which, when substituted into (3), yields a quantity *below* $F_{\text{var}}$—not a valid upper bound on $F_{\text{true}}$. The joint bound (Sec. 3) is rigorous but loose by an extensive amount. We describe three approaches that succeed, each with different tradeoffs.

## 9.1 Markov-corrected variational ansatz

The simplest extension appends a *fixed* Markov transition kernel to the autoregressive base, with no bijective flow. The generative process is

$$\boldsymbol{z} \sim p_\theta(\boldsymbol{z}), \qquad \boldsymbol{\sigma} \sim T(\cdot \mid \boldsymbol{z}), \tag{42}$$

where $T$ is a fixed, analytically known transition kernel (e.g., a Gibbs sweep targeting the Boltzmann distribution). The variational distribution is the marginal

$$q(\boldsymbol{\sigma}) = \sum_{\boldsymbol{z}} p_\theta(\boldsymbol{z})\, T(\boldsymbol{\sigma} \mid \boldsymbol{z}), \tag{43}$$

which is intractable—evaluating $\ln q(\boldsymbol{\sigma})$ requires summing over all $\boldsymbol{z}$. Nevertheless, a rigorous variational bound can be obtained without evaluating $\ln q$.

**Derivation of the bound.** Define a *reverse kernel* $\tilde{T}(\boldsymbol{z} \mid \boldsymbol{\sigma})$—an auxiliary distribution that approximates the posterior over $\boldsymbol{z}$ given $\boldsymbol{\sigma}$. Consider the joint distribution of the forward process, $P_{\text{fwd}}(\boldsymbol{z}, \boldsymbol{\sigma}) = p_\theta(\boldsymbol{z})\, T(\boldsymbol{\sigma} \mid \boldsymbol{z})$, and the "target joint" $P_{\text{target}}(\boldsymbol{z}, \boldsymbol{\sigma}) = p_{\text{Boltz}}(\boldsymbol{\sigma})\, \tilde{T}(\boldsymbol{z} \mid \boldsymbol{\sigma})$. Since $\mathrm{D_{KL}}(P_{\text{fwd}} \| P_{\text{target}}) \geq 0$:

$$\begin{aligned} 0 &\leq \mathbb{E}_{P_{\text{fwd}}}\Big[\ln \frac{p_\theta(\boldsymbol{z})\, T(\boldsymbol{\sigma}|\boldsymbol{z})}{p_{\text{Boltz}}(\boldsymbol{\sigma})\, \tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})}\Big] \\ &= \mathbb{E}\big[\ln p_\theta(\boldsymbol{z}) + \ln T(\boldsymbol{\sigma}|\boldsymbol{z}) - \ln \tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma}) + \beta E(\boldsymbol{\sigma}) + \ln Z\big], \end{aligned} \tag{44}$$

which rearranges to

$$\boxed{F_{\text{true}} \leq \mathbb{E}\Big[E(\boldsymbol{\sigma}) + T \ln p_\theta(\boldsymbol{z}) + T \ln \frac{T(\boldsymbol{\sigma}|\boldsymbol{z})}{\tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})}\Big],} \tag{45}$$

where the expectation is over $\boldsymbol{z} \sim p_\theta$, $\boldsymbol{\sigma} \sim T(\cdot|\boldsymbol{z})$. This is a *rigorous upper bound* on $F_{\text{true}}$ for *any* choice of $T$ and $\tilde{T}$.

**Structure of the bound.** Equation (45) decomposes into three interpretable terms:

$$\tilde{F} = \underbrace{\mathbb{E}[E(\boldsymbol{\sigma})]}_{\text{energy}} \underbrace{-T\,H[p_\theta]}_{\text{base entropy}} \underbrace{+T\mathbb{E}\Big[\ln \frac{T(\boldsymbol{\sigma}|\boldsymbol{z})}{\tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})}\Big]}_{\text{Markov correction}}. \tag{46}$$

The first two terms are the standard variational free energy of $p_\theta$, as if the configurations were $\boldsymbol{z}$ rather than $\boldsymbol{\sigma}$. The third term is the cost of the stochastic correction—the average log-ratio of forward to reverse transitions.

**Gap analysis.** The gap between (45) and the true variational free energy $F_{\text{var}}[q]$ of the marginal (43) is

$$\tilde{F} - F_{\text{var}}[q] = T\mathbb{E}_{q(\boldsymbol{\sigma})}\big[\mathrm{D}_{\mathrm{KL}}\big(p(\boldsymbol{z}|\boldsymbol{\sigma}) \,\big\|\, \tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})\big)\big] \geq 0, \tag{47}$$

where $p(\boldsymbol{z}|\boldsymbol{\sigma}) = p_\theta(\boldsymbol{z})\,T(\boldsymbol{\sigma}|\boldsymbol{z})/q(\boldsymbol{\sigma})$ is the true posterior. The bound is tight when $\tilde{T}$ matches the posterior exactly.

**Specialisation to Gibbs sweeps.** For the Ising model, the natural choice is a systematic-scan Gibbs sweep targeting the Boltzmann distribution. Here $\boldsymbol{z}$ is the *starting* (old) configuration produced by the AR base, and $\boldsymbol{\sigma}$ is the *output* (new) configuration on which the energy is evaluated. The sweep updates spins sequentially in order $i = 1, \ldots, N$: at step $i$, sites $j < i$ already hold their new values $\sigma_j$, while sites $j > i$ still hold their old values $z_j$. The transition probability factorises as

$$T(\boldsymbol{\sigma} \mid \boldsymbol{z}) = \prod_{i=1}^{N} p_{\text{Boltz}}(\sigma_i \mid \sigma_{<i},\, z_{>i}), \tag{48}$$

where each factor is the single-site Gibbs conditional

$$p_{\text{Boltz}}(\sigma_i \mid \text{neighbours}) = \frac{e^{\beta\sigma_i h_i}}{2\cosh\beta h_i}, \qquad h_i = J\sum_{j\sim i} s_j, \qquad s_j = \begin{cases} \sigma_j & \text{if } j < i \text{ (already updated)}, \\ z_j & \text{if } j > i \text{ (not yet updated)}. \end{cases} \tag{49}$$

On a 2D lattice, site $i$ typically has neighbours with both lower and higher indices, so the local field $h_i$ genuinely mixes old ($z$) and new ($\sigma$) spin values. Each factor is analytically known, so $\ln T(\boldsymbol{\sigma}|\boldsymbol{z})$ is a sum of $N$ tractable terms.

For the reverse kernel, a Gibbs sweep in the *reverse* order $i = N, N-1, \ldots, 1$ is a natural, parameter-free choice:

$$\tilde{T}(\boldsymbol{z} \mid \boldsymbol{\sigma}) = \prod_{i=N}^{1} p_{\text{Boltz}}(z_i \mid z_{>i},\, \sigma_{<i}). \tag{50}$$

Because the forward and reverse sweeps visit sites in opposite order, they see different mixtures of old and new spins at each step, so $\ln T(\boldsymbol{\sigma}|\boldsymbol{z}) \neq \ln \tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})$ in general. Both $\ln T$ and $\ln \tilde{T}$ are sums of $N$ local Boltzmann conditionals—no neural networks, no learning, no STE.

**The telescoping problem.** Despite the apparent asymmetry between forward and reverse sweeps, the Markov correction in (45) *trivialises* when both kernels target the Boltzmann distribution at the same temperature. At step $i$, both the forward and reverse sweeps condition on the same background $(\sigma_{<i},\, z_{>i})$, so the local field $h_i$ is identical. The per-site log-ratio is therefore

$$\ln p_{\text{Boltz}}(\sigma_i \mid \sigma_{<i}, z_{>i}) - \ln p_{\text{Boltz}}(z_i \mid \sigma_{<i}, z_{>i}) = \beta(\sigma_i - z_i)\,h_i. \tag{51}$$

Summing over all sites, this is a telescoping sum of single-spin energy changes. Define the sequence of partially updated configurations $\boldsymbol{c}^{(i)} = (\sigma_1, \ldots, \sigma_i, z_{i+1}, \ldots, z_N)$, so that $\boldsymbol{c}^{(0)} = \boldsymbol{z}$ and $\boldsymbol{c}^{(N)} = \boldsymbol{\sigma}$. The energy change at step $i$ is

$$E(\boldsymbol{c}^{(i)}) - E(\boldsymbol{c}^{(i-1)}) = -(\sigma_i - z_i)\,h_i, \tag{52}$$

so the total log-ratio telescopes:

$$\ln \frac{T(\boldsymbol{\sigma}|\boldsymbol{z})}{\tilde{T}(\boldsymbol{z}|\boldsymbol{\sigma})} = \beta \sum_{i=1}^{N} (\sigma_i - z_i)\, h_i = -\beta \sum_{i=1}^{N} \big[E(\boldsymbol{c}^{(i)}) - E(\boldsymbol{c}^{(i-1)})\big] = \beta\big[E(\boldsymbol{z}) - E(\boldsymbol{\sigma})\big]. \tag{53}$$

Substituting into the bound (45):

$$\tilde{F} = \mathbb{E}\big[E(\boldsymbol{\sigma}) + T \ln p_\theta(\boldsymbol{z}) + E(\boldsymbol{z}) - E(\boldsymbol{\sigma})\big] = \mathbb{E}\big[E(\boldsymbol{z}) + T \ln p_\theta(\boldsymbol{z})\big] = F_{\text{var}}[p_\theta]. \tag{54}$$

*The Gibbs sweep drops out entirely.* The bound reduces to the variational free energy of the base distribution, as if no MCMC step had been applied. This is not a coincidence: any transition kernel $T$ that satisfies detailed balance with respect to $p_{\text{Boltz}}$, paired with the natural reverse $\tilde{T} = T$ or any $\tilde{T}$ whose log-ratio with $T$ equals $\beta \Delta E$, produces the same cancellation. In particular, the Wolff cluster algorithm, random-scan Gibbs, and Swendsen–Wang all trivialise the bound for the same reason.

The physical interpretation is clear: a single MCMC step at the target temperature is "too well matched" to the Boltzmann distribution. The importance weight exactly compensates for any energy change the kernel produces, yielding zero net improvement.

**The invisible improvement.** The trivialisation of the bound does *not* mean the Gibbs sweep is useless—it means the bound cannot see the improvement. The variational distribution after the sweep,

$$q(\boldsymbol{\sigma}) = \sum_{\boldsymbol{z}} p_\theta(\boldsymbol{z})\, T(\boldsymbol{\sigma} \mid \boldsymbol{z}), \tag{55}$$

is genuinely different from $p_\theta$ and provably better. Since $T$ leaves $p_{\text{Boltz}}$ invariant, the data processing inequality for Markov chains guarantees contraction of the KL divergence:

$$\mathrm{D}_{\mathrm{KL}}(q \| p_{\text{Boltz}}) \leq \mathrm{D}_{\mathrm{KL}}(p_\theta \| p_{\text{Boltz}}), \tag{56}$$

which is equivalent to

$$F_{\text{var}}[q] \leq F_{\text{var}}[p_\theta]. \tag{57}$$

The Gibbs sweep is a *provable improvement operator* on the variational distribution: it increases entropy (by convolving $p_\theta$ with the stochastic kernel) and adjusts the energy, with the net effect always reducing the variational free energy.

After $k$ sweeps, the contraction compounds:

$$\mathrm{D}_{\mathrm{KL}}(q_k \| p_{\text{Boltz}}) \leq (1 - \gamma)^k \, \mathrm{D}_{\mathrm{KL}}(p_\theta \| p_{\text{Boltz}}), \tag{58}$$

where $\gamma$ is the spectral gap of the transition matrix. Near $T_c$, $\gamma \sim L^{-z}$ with dynamic critical exponent $z \approx 2$ for single-site Gibbs and $z \approx 0.25$ for Wolff, so the contraction per sweep is slow for Gibbs but fast for Wolff.

The problem is that $F_{\text{var}}[q]$ is *not computable*: evaluating $\ln q(\boldsymbol{\sigma})$ requires the intractable sum over all $\boldsymbol{z}$. The hierarchy of bounds is therefore

$$F_{\text{true}} \;\leq\; F_{\text{var}}[q] \;\leq\; F_{\text{var}}[p_\theta] \;=\; \tilde{F}, \tag{59}$$

where the gap $F_{\text{var}}[p_\theta] - F_{\text{var}}[q] > 0$ is a real improvement that the evaluable bound $\tilde{F}$ cannot capture. The role of $\beta$-annealing, introduced next, is precisely to make this hidden improvement visible.

**Resolution: $\beta$-annealing.** To obtain a non-trivial bound, the transition kernels must target distributions at *intermediate* temperatures, not the physical temperature. This is the standard annealed importance sampling (AIS) construction [17], here combined with a learned base.

Define a temperature schedule $0 < \beta_1 < \beta_2 < \cdots < \beta_K = \beta$ and intermediate (unnormalised) distributions $\tilde{\pi}_k(\boldsymbol{x}) = e^{-\beta_k E(\boldsymbol{x})}$. Set $\tilde{\pi}_0(\boldsymbol{x}) = p_\theta(\boldsymbol{x})$ (the learned base, which is normalised). The generative process is:

1. Sample $\boldsymbol{x}^{(0)} \sim p_\theta$.
2. For $k = 1, \ldots, K$: apply an MCMC kernel $T_k$ that leaves $\pi_k \propto \tilde{\pi}_k$ invariant, producing $\boldsymbol{x}^{(k)} \sim T_k(\cdot \mid \boldsymbol{x}^{(k-1)})$.
3. Output $\boldsymbol{\sigma} = \boldsymbol{x}^{(K)}$.

The AIS log-weight is

$$\ln W = \sum_{k=1}^{K} \ln \frac{\tilde{\pi}_k(\boldsymbol{x}^{(k-1)})}{\tilde{\pi}_{k-1}(\boldsymbol{x}^{(k-1)})} = \ln \frac{\tilde{\pi}_1(\boldsymbol{x}^{(0)})}{p_\theta(\boldsymbol{x}^{(0)})} + \sum_{k=2}^{K} \big[ -(\beta_k - \beta_{k-1})\, E(\boldsymbol{x}^{(k-1)}) \big]. \tag{60}$$

The *MCMC kernels do not appear in the weight*—only the energies at intermediate configurations and the base log-probability. This means *any* mixing kernel can be used at each step: Gibbs sweeps, Wolff cluster updates, or Swendsen–Wang, without needing to evaluate their (potentially intractable) transition probabilities.

By Jensen's inequality applied to the identity $\mathbb{E}[W] = Z/Z_0$ [17, 19], the bound is

$$\boxed{F_{\text{true}} \le -T\,\mathbb{E}[\ln W] = \mathbb{E}\Big[ -T \ln \frac{\tilde{\pi}_1(\boldsymbol{x}^{(0)})}{p_\theta(\boldsymbol{x}^{(0)})} + \sum_{k=2}^{K} T(\beta_k - \beta_{k-1})\, E(\boldsymbol{x}^{(k-1)}) \Big].} \tag{61}$$

This is a rigorous upper bound on $F_{\text{true}}$ for any number of annealing steps $K$, any temperature schedule, and any choice of MCMC kernels. Crucially, the AIS bound sits between the invisible improvement and the trivialised bound, refining the hierarchy (59):

$$F_{\text{true}} \;\le\; F_{\text{var}}[q_K] \;\le\; \tilde{F}_{\text{AIS}}(K) \;\le\; \cdots \;\le\; \tilde{F}_{\text{AIS}}(1) \;\le\; F_{\text{var}}[p_\theta]. \tag{62}$$

The AIS bound makes the hidden improvement of the MCMC steps (57) partially visible, with the gap $\tilde{F}_{\text{AIS}}(K) - F_{\text{var}}[q_K]$ controlled by the variance of $\ln W$.

**Role of the learned base.** In standard AIS, the base distribution is simple (e.g., uniform, with $\tilde{\pi}_0 = 1$ and $Z_0 = 2^N$), requiring many temperature steps to bridge the gap to $p_{\text{Boltz}}$. Replacing the uniform base with a learned autoregressive model $p_\theta$ that already approximates $p_{\text{Boltz}}$ dramatically reduces the number of annealing steps needed. If $p_\theta$ is perfect, $K = 0$ suffices and the bound reduces to $F_{\text{var}}[p_\theta] = F_{\text{true}}$. In practice, a moderate $K$ (a few Wolff sweeps across a short temperature range) suffices to correct the entropy deficit that $p_\theta$ alone cannot resolve.

**Choice of MCMC kernel.** Since the kernel does not appear in the weight, the only criterion is *mixing efficiency.* Near $T_c$, the Wolff cluster algorithm has dynamic critical exponent $z \approx 0.25$, compared to $z \approx 2$ for single-site Gibbs. This makes Wolff the natural choice: each annealing step thermalises the system efficiently, keeping the variance of $\ln W$ low with fewer steps.

**Training.** The bound (61) is minimised with respect to $\theta$ alone. Since the sampling distribution depends on $\theta$ through $p_\theta$ (and indirectly through the MCMC chain), the gradient is a REINFORCE estimator:

$$\nabla_\theta \tilde{F} = \mathbb{E}\Big[\big(\tilde{F}_{\text{local}} - b\big)\, \nabla_\theta \ln p_\theta(\boldsymbol{x}^{(0)})\Big], \tag{63}$$

where $\tilde{F}_{\text{local}} = -T \ln W$ is the per-sample bound estimand and $b$ is a learned baseline. This is identical in form to (4). The MCMC kernels are not differentiated through; they are treated as fixed, non-trainable layers.

**Entropy increase.** The variational distribution $q(\boldsymbol{\sigma})$ is the marginal of the AIS chain, which is a convolution of $p_\theta$ with $K$ MCMC kernels at progressively higher $\beta$. Each kernel broadens the distribution, so

$$H[q] \geq H[p_\theta], \tag{64}$$

with the gap growing with $K$. The $\beta$-annealing injects entropy *gradually*: at low $\beta_k$ (high temperature), the MCMC kernels are highly stochastic and inject substantial entropy; at high $\beta_k$ (near the physical temperature), they make only small, targeted corrections. This graduated approach avoids the all-or-nothing character of a single MCMC step at the target temperature.

**Interpolation between variational and MCMC.** The construction provides a smooth interpolation:

- $K = 0$: pure variational ansatz, $\tilde{F} = F_{\text{var}}[p_\theta]$.
- $K$ moderate: learned base + short AIS correction.
- $K \to \infty$ with mixing: $\tilde{F} \to F_{\text{true}}$ regardless of $p_\theta$.

The learned base reduces the number of annealing steps needed for a given accuracy, while the AIS correction patches the entropy deficit that the base alone cannot resolve. This is the key advantage over both pure variational methods (which are entropy-limited) and pure AIS (which requires many steps from a simple base).

**Effect on training dynamics.** A crucial consequence of the $\beta$-annealing is that it changes the *optimisation landscape* for $\theta$. The AIS objective $\tilde{F}_{\text{AIS}}(\theta) = \mathbb{E}[-T \ln W]$ is a different function of $\theta$ than $F_{\text{var}}[p_\theta]$, and minimising it yields a different optimal $\theta^*$.

In the pure variational setting ($K = 0$), the REINFORCE reward for a sample $\boldsymbol{z}$ is $r(\boldsymbol{z}) = E(\boldsymbol{z}) + T \ln p_\theta(\boldsymbol{z})$. A sample that happens to land on a low-energy configuration receives a large negative reward, gets upweighted, and the positive feedback loop of Sec. 8.4 drives mode collapse. With $\beta$-annealing, the reward becomes

$$r_{\text{AIS}}(\boldsymbol{z}) = -T \ln W = -T \ln p_\theta(\boldsymbol{z}) + \sum_k T \Delta\beta_k\, E(\boldsymbol{x}^{(k-1)}), \tag{65}$$

which depends on the *entire annealing trajectory*, not just $\boldsymbol{z}$. A "bad" starting configuration (high energy, wrong structure) can still produce a good trajectory if the MCMC steps correct it. The reward is smoothed by the chain.

This changes what $p_\theta$ needs to learn. In the pure variational setting, $p_\theta$ must simultaneously capture both the energy and the entropy of $p_{\text{Boltz}}$—a task at which REINFORCE fails because the mode-seeking reverse KL sacrifices entropy. In the AIS setting, $p_\theta$ need only provide a good *proposal distribution* for the annealing chain: the MCMC steps handle the energy correction. A high-entropy $p_\theta$ that roughly captures the correlation structure is rewarded, because diverse starting points lead to better chain coverage.

The mechanism is variance reduction. The per-sample reward variance $\mathrm{Var}[\ln W]$ decreases with $K$: more annealing steps thermalise bad samples, reducing the spread of rewards. Lower variance produces a more uniform gradient signal across samples, preventing $p_\theta$ from over-concentrating on a few low-energy modes. The number of annealing steps $K$ thus serves as a tunable knob that trades compute for training stability.

In summary, the $\beta$-annealing breaks the REINFORCE mode collapse feedback loop by changing the role of $p_\theta$ from "approximate $p_{\mathrm{Boltz}}$ alone" to "provide a good starting point for the annealing chain"—a fundamentally easier optimisation problem.

**Alternative base: tensor train / matrix product state.** The preceding discussion raises the question of what parameterisation is best suited for a high-entropy base distribution. A natural candidate is the *tensor train* (TT) decomposition, known in physics as a *matrix product state* (MPS). For $N$ binary spins, the TT represents the distribution as

$$p(\boldsymbol{\sigma}) = \mathrm{Tr}\big[A_1(\sigma_1)\, A_2(\sigma_2) \cdots A_N(\sigma_N)\big], \tag{66}$$

where each $A_i(\sigma_i)$ is a $D \times D$ matrix and $D$ is the *bond dimension*. Like the autoregressive network, the TT provides exact log-probabilities (computable in $O(ND^2)$ by matrix multiplication) and exact sampling (via a left-to-right sweep using the chain rule). In fact, the TT *is* an autoregressive model: the conditional $p(\sigma_i \mid \sigma_{<i})$ is determined by the left environment contracted with $A_i$ and a right normalisation factor.

The key advantage of the TT as a base for the $\beta$-annealed framework is threefold:

1. **Implicit entropy regularisation.** The mutual information across any bond of the TT is bounded by $\ln D$. A mode-collapsed distribution—concentrated on a few configurations—requires high mutual information across cuts, which demands large $D$. At moderate $D$, the TT *cannot* mode-collapse: the bond dimension imposes a structural upper bound on how peaked the distribution can be. In contrast, an autoregressive neural network is a universal approximator that can represent arbitrarily concentrated distributions, with no built-in safeguard against mode collapse.

2. **DMRG training.** The TT admits optimisation by the density matrix renormalisation group (DMRG), which sweeps through the chain optimising one tensor core at a time via a local least-squares or eigenvalue problem. This is *not* REINFORCE: there is no high-variance score-function estimator, no positive feedback loop, and no mode collapse. The convergence is stable and well understood. (If the TT were trained with REINFORCE, it would mode-collapse just like an AR network—the advantage is that the TT admits a better optimiser.)

3. **Bond dimension controls the effective temperature.** At $D = 1$, the TT is a product distribution (independent spins), which has maximum entropy $N \ln 2$. As $D$ increases, the TT can represent progressively longer-range correlations, up to a correlation length $\xi_{\mathrm{MPS}} \sim \ln D$ (for a 2D system mapped to 1D). Choosing $D$ such that $\xi_{\mathrm{MPS}} \approx \xi(T_{\mathrm{eff}})$ gives a base distribution that naturally approximates the Boltzmann distribution at an effective temperature $T_{\mathrm{eff}} > T_c$. The $\beta$-annealing chain then cools from $T_{\mathrm{eff}}$ to the physical temperature, building the long-range correlations that $D$ cannot capture.

The main limitation is the 1D structure: for a 2D $L \times L$ lattice mapped to 1D via a snake ordering, the entanglement entropy across a cut scales as $O(L)$, requiring $D \sim e^{O(L)}$ for exact representation at $T_c$. At moderate $D$, the TT captures only short-range correlations— but this is precisely the regime where it serves as a high-temperature base for the AIS chain. The bond dimension $D$ plays the same role as the effective temperature $T_{\mathrm{eff}}$: it determines how much of the physics $p_\theta$ captures versus how much the annealing chain must supply.

### 9.2 Adding bijective layers: stochastic discrete normalizing flows

The $\beta$-annealed ansatz of Sec. 9.1 can be strengthened by inserting bijective coupling layers between the MCMC steps. Adapting the continuous stochastic normalizing flows of Wu et al. [18], the generative process becomes

$$\boldsymbol{z} \sim p_\theta(\boldsymbol{z}), \qquad \boldsymbol{\sigma}' = f_\phi(\boldsymbol{z}), \qquad \boldsymbol{\sigma}' \xrightarrow{\text{AIS}} \boldsymbol{\sigma}, \tag{67}$$

where the AIS chain starts from the flow output $\boldsymbol{\sigma}'$ rather than directly from $\boldsymbol{z}$. The bijective flow contributes no additional terms to the AIS log-weight—its "Jacobian" is 1 on a finite set—but it rearranges probability mass *deterministically*, lowering $\mathbb{E}[E(\boldsymbol{\sigma}')]$ at zero variance cost.

More generally, bijective coupling layers and $\beta$-annealed MCMC steps can be freely interleaved:

$$\boldsymbol{z} \xrightarrow{f_\phi} \boldsymbol{x}^{(0)} \xrightarrow{T_1 \text{ at } \beta_1} \boldsymbol{x}^{(1)} \xrightarrow{g_\phi^{(1)}} \boldsymbol{x}^{(1)} \xrightarrow{T_2 \text{ at } \beta_2} \cdots \xrightarrow{T_K \text{ at } \beta} \boldsymbol{\sigma}, \tag{68}$$

with the AIS log-weight (60) receiving contributions only from the $\beta$-annealing ratios; the bijective layers are transparent.

The three components play distinct roles in the bound:

$$\tilde{F} = \underbrace{\mathbb{E}[E(\boldsymbol{\sigma})]}_{\substack{\text{flow lowers,} \\ \text{AIS corrects}}} \; \underbrace{-T\,H[p_\theta]}_{\substack{\text{AR base} \\ \text{sets entropy}}} \; \underbrace{+T\,\mathbb{E}[\text{AIS overhead}]}_{\substack{\text{annealing injects entropy} \\ \text{at variance cost}}}. \tag{69}$$

The bijective flow reduces the first term without affecting the other two. The $\beta$-annealing reduces the gap between the base entropy and the target entropy, at the cost of increasing the third term. The optimal architecture uses the flow to handle what it can (energy rearrangement) and the annealing to handle what only it can (entropy injection).

**Training.** The gradients decompose cleanly across the parameter groups:

- $\nabla_\theta$ (AR base): REINFORCE, as in (63).
- $\nabla_\phi$ (bijective flow): STE through the coupling layers, as in (25). The flow gradient does not involve REINFORCE and does not depend on the MCMC kernels.

Only $\theta$ and $\phi$ are trained—the MCMC kernels and $\beta$-schedule are fixed, non-trainable components that inject entropy for free.

### 9.3 $k$-to-1 surjections via symmetry

A complementary approach exploits the known symmetries of the target distribution to construct surjective maps with small, tractable fibers.

**General construction.** Let $G$ be a finite symmetry group of the Hamiltonian: $E(g \cdot \boldsymbol{\sigma}) = E(\boldsymbol{\sigma})$ for all $g \in G$, so that $p_{\text{Boltz}}(g \cdot \boldsymbol{\sigma}) = p_{\text{Boltz}}(\boldsymbol{\sigma})$. Given a bare autoregressive model $p_\theta^{\text{AR}}(\boldsymbol{z})$ and a $G$-equivariant flow $f_\phi$, define the $G$-symmetrised distribution

$$q(\boldsymbol{\sigma}) = \frac{1}{|G|} \sum_{g \in G} p_\theta^{\text{AR}}\big(f_\phi^{-1}(g \cdot \boldsymbol{\sigma})\big). \tag{70}$$

This is a $|G|$-to-1 surjection from the "labeled" space (where the orientation of $\boldsymbol{\sigma}$ relative to the lattice matters) to the "unlabeled" quotient. The log-probability is evaluated exactly via

$$\ln q(\boldsymbol{\sigma}) = \text{logsumexp}_{g \in G}\big(\ln p_\theta^{\text{AR}}(f_\phi^{-1}(g \cdot \boldsymbol{\sigma}))\big) - \ln |G|. \tag{71}$$

Since each term in the logsumexp is an exact autoregressive log-probability, $\ln q(\boldsymbol{\sigma})$ is exact, and the variational bound $F_{\text{var}}[q] \geq F_{\text{true}}$ is rigorous.

**The $\mathbb{Z}_2$ case.** The symmetrisation (15) already used for the spin-flip symmetry is the special case $G = \mathbb{Z}_2 = \{e, -\mathbb{1}\}$ with $|G| = 2$. The logsumexp is over 2 terms and is trivially cheap.

**Lattice symmetries.** For the square-lattice Ising model with periodic boundary conditions, the full space group is

$$G = \underbrace{\mathbb{Z}_L \times \mathbb{Z}_L}_{\text{translations}} \rtimes \underbrace{D_4}_{\text{point group}} \times \underbrace{\mathbb{Z}_2}_{\text{spin flip}}, \tag{72}$$

with $|G| = L^2 \times 8 \times 2 = 16L^2$. For an $L = 32$ lattice, $|G| = 16{,}384$. Evaluating (71) requires $|G|$ forward passes of the inverse flow and the autoregressive network, which is expensive but embarrassingly parallel. In practice, one may use a subgroup (e.g., translations only, $|G| = L^2 = 1024$) to balance cost and benefit.

**Entropy gain.** Unlike the bijective flow alone, the $G$-symmetrised distribution (70) has entropy

$$H[q] = H[p_\theta^{\mathrm{AR}}] + \underbrace{\Delta H_G}_{\geq 0,\, \leq \ln|G|}, \tag{73}$$

where $\Delta H_G$ is the entropy gained from symmetrisation. The gain is maximal ($\Delta H_G = \ln|G|$) when the $|G|$ copies of $\boldsymbol{z}$-space are disjoint in probability (i.e., $p_\theta^{\mathrm{AR}}$ spontaneously breaks the symmetry), and zero when $p_\theta^{\mathrm{AR}}$ is already $G$-invariant. At $T_c$, where the Boltzmann distribution is exactly $G$-invariant, the symmetrisation allows $p_\theta^{\mathrm{AR}}$ to concentrate on one symmetry sector while the sum restores the full symmetry, potentially yielding up to $\ln|G|$ bits of additional entropy.

**Combining both approaches.** The stochastic layers (Sec. 9.2) and the symmetry surjection are complementary: the former injects entropy through noise, the latter through symmetry averaging. They can be composed: use a $G$-symmetrised stochastic normalizing flow, gaining entropy from both sources while maintaining a rigorous bound.

## 10 Comparison of approaches

Table 3: Comparison of token-based variational ansätze.

| **Approach** | **Rigorous bound on $F$?** | **Tractable $\ln q(\boldsymbol{\sigma})$?** | **Expressivene** |
|---|---|---|---|
| Raw-spin AR [1] | Yes | Exact | Limited by orde |
| Block-spin AR (lossless) | Yes | Exact | Limited by block pa |
| 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 bij** |
| **AR + $\beta$-annealing (AIS)** | **Yes** | **Via AIS weight** | **High (AR + MC** |
| **NF + $\beta$-annealing** | **Yes** | **Via AIS weight** | **Higher (bijection +** |
| **$G$-symmetrised NF** | **Yes** | **Exact (logsumexp)** | **High + symme** |

## 11 Haar wavelet basis for the 2D Ising model

The autoregressive models of the preceding sections factorise the distribution over a 1D ordering of the lattice sites. Any such ordering maps the 2D lattice to a 1D chain, introducing long-range dependencies: nearest neighbours on the lattice can be $O(L)$ apart in the 1D sequence. A more natural approach is to work in a *multiscale* basis that respects the 2D structure. The Haar wavelet provides a concrete, lossless change of variables from spins to hierarchical coarse/detail coefficients, with clean physical properties.

### 11.1 Binary Haar wavelet transform

The standard Haar wavelet acts on real-valued signals. For binary spins $\sigma \in \{+1, -1\}$, the natural discrete analogue replaces addition/subtraction with multiplication.

**One level, 2×2 block.** Consider a $2 \times 2$ block of spins:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \longleftrightarrow \begin{cases} c = \sigma_{11} & \text{(coarse: anchor spin)}, \\ d_1 = \sigma_{11}\,\sigma_{12} & \text{(horizontal detail)}, \\ d_2 = \sigma_{11}\,\sigma_{21} & \text{(vertical detail)}, \\ d_3 = \sigma_{11}\,\sigma_{22} & \text{(diagonal detail)}. \end{cases} \tag{74}$$

All four variables are binary $\{+1, -1\}$, and the map is lossless: $\sigma_{12} = c\,d_1$, $\sigma_{21} = c\,d_2$, $\sigma_{22} = c\,d_3$. The detail variables $d_1, d_2, d_3$ are *bond variables*: $d_i = +1$ when the two spins are aligned, $d_i = -1$ when anti-aligned.

**Recursive application on a $2^n \times 2^n$ lattice.** Apply the $2 \times 2$ transform to the full lattice, grouping spins into non-overlapping blocks. The coarse variables $c$ form a $2^{n-1} \times 2^{n-1}$ lattice; the detail variables $(d_1, d_2, d_3)$ form a $3 \times 2^{n-1} \times 2^{n-1}$ array. Apply the same transform recursively to the coarse lattice. After $n$ levels:

- Level $n$ (coarsest): 1 coarse variable $c^{(n)}$.
- Level $k$ ($0 \le k \le n-1$): $3 \times 4^{n-1-k}$ detail variables $\boldsymbol{d}^{(k)}$, capturing the internal structure of each $2 \times 2$ block at scale $k$.

The total variable count is $1 + 3\sum_{j=0}^{n-1} 4^j = 4^n = N$. The transform is a bijection $\{+1, -1\}^N \to \{+1, -1\}^N$.

**Forward transform (spins → wavelet coefficients).**

```
1. Set $\boldsymbol{s}^{(0)} \leftarrow \boldsymbol{\sigma}$ (the $2^n \times 2^n$ spin array).
2. For $k = 0, 1, \dots, n-1$:
    (a) Partition $\boldsymbol{s}^{(k)}$ into non-overlapping $2 \times 2$ blocks.
    (b) For each block, compute $(c, d_1, d_2, d_3)$ via (74).
    (c) Store $\boldsymbol{d}^{(k)}$ (the detail array at level $k$).
    (d) Set $\boldsymbol{s}^{(k+1)} \leftarrow$ the array of coarse variables $c$.
3. Output $c^{(n)} = \boldsymbol{s}^{(n)}$ and $\{\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(n-1)}\}$.
```

**Inverse transform (wavelet coefficients → spins).**

```
1. Set $\boldsymbol{s}^{(n)} \leftarrow c^{(n)}$.
2. For $k = n-1, n-2, \dots, 0$:
    (a) For each block at level $k$, recover the four spins from the coarse variable $c$ (from $\boldsymbol{s}^{(k+1)}$)
        and the detail variables $(d_1, d_2, d_3)$ (from $\boldsymbol{d}^{(k)}$): $\sigma_{\mathrm{TL}} = c$, $\sigma_{\mathrm{TR}} = c\,d_1$, $\sigma_{\mathrm{BL}} = c\,d_2$,
        $\sigma_{\mathrm{BR}} = c\,d_3$.
```

(b) Assemble into $\boldsymbol{s}^{(k)}$.

3. Output $\boldsymbol{\sigma} = \boldsymbol{s}^{(0)}$.

### 11.2 Energy decomposition in the Haar basis

The Ising energy $E = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j$ decomposes into contributions from each level of the hierarchy.

**Intra-block energy.** For a $2 \times 2$ block at level $k$, the four nearest-neighbour bonds *within* the block are

$$\begin{aligned} \sigma_{11}\sigma_{12} &= d_1, & \sigma_{11}\sigma_{21} &= d_2, \\ \sigma_{12}\sigma_{22} &= d_1 d_3, & \sigma_{21}\sigma_{22} &= d_2 d_3. \end{aligned} \tag{75}$$

The intra-block energy is

$$E_{\text{intra}} = -J\big(d_1 + d_2 + d_1 d_3 + d_2 d_3\big) = -J(d_1 + d_2)(1 + d_3). \tag{76}$$

The coarse variable $c$ *does not appear*: the intra-block energy depends only on the detail variables. This is a consequence of the $\mathbb{Z}_2$ symmetry—the energy is invariant under global spin flip, and $c$ encodes the absolute orientation that the energy cannot see.

**Inter-block energy.** Consider two horizontally adjacent blocks $A$ and $B$ at level $k$, with coarse variables $c^A$ and $c^B$ and detail variables $(d_1^A, d_2^A, d_3^A)$ and $(d_1^B, d_2^B, d_3^B)$. The two inter-block bonds (connecting the right column of $A$ to the left column of $B$) are

$$\begin{aligned} \sigma_{12}^A \sigma_{11}^B &= c^A d_1^A \cdot c^B = (c^A c^B)\, d_1^A, \\ \sigma_{22}^A \sigma_{21}^B &= c^A d_3^A \cdot c^B d_2^B = (c^A c^B)\, d_3^A d_2^B. \end{aligned} \tag{77}$$

The product $c^A c^B$ is a *coarse-level bond variable*: at the next level of the hierarchy, it equals the detail variable connecting the two parent blocks. The inter-block energy thus couples detail variables at level $k$ to coarse variables at level $k+1$, but the coupling is *local* (involves only adjacent blocks) and *known analytically.*

**Total energy.** The full Ising energy decomposes as

$$E(\boldsymbol{\sigma}) = \sum_{k=0}^{n-1} \Big[ \sum_{\text{blocks at level } k} E_{\text{intra}}^{(k)}(\boldsymbol{d}^{(k)}) + \sum_{\text{adjacent pairs}} E_{\text{inter}}^{(k)}(\boldsymbol{d}^{(k)},\, \boldsymbol{c}^{(k+1)}) \Big], \tag{78}$$

where each term is local and analytically known. The coarsest variable $c^{(n)}$ does not appear in the energy at all.

### 11.3 Autoregressive factorisation in the Haar basis

The Haar wavelet defines a natural coarse-to-fine autoregressive factorisation:

$$p(\boldsymbol{\sigma}) = p(c^{(n)}) \times \prod_{k=n-1}^{0} p\big(\boldsymbol{d}^{(k)} \,\big|\, c^{(n)},\, \boldsymbol{d}^{(n-1)}, \ldots, \boldsymbol{d}^{(k+1)}\big). \tag{79}$$

Each factor conditions on all coarser levels. The log-probability is

$$\ln p(\boldsymbol{\sigma}) = \ln p(c^{(n)}) + \sum_{k=n-1}^{0} \ln p\big(\boldsymbol{d}^{(k)} \,\big|\, \text{coarser levels}\big), \tag{80}$$

which is exact and tractable (a sum of $n$ terms).

**Structure of the conditionals.** At level $k$, the conditional $p(\boldsymbol{d}^{(k)} \mid \text{coarser levels})$ is a distribution over $3 \times 2^{n-k-1} \times 2^{n-k-1}$ binary variables. The energy decomposition (78) shows that this conditional depends on:

1. The *intra-block* energy at level $k$: depends only on $\boldsymbol{d}^{(k)}$, local within each block.
2. The *inter-block* energy at level $k$: couples detail variables in adjacent blocks, with coefficients determined by the coarse variables $\boldsymbol{c}^{(k+1)}$ (which are known from the coarser levels).

Crucially, the inter-block coupling is *short-range*: it connects only nearest-neighbour blocks. Conditioned on the coarse variables, the detail variables at level $k$ form a 2D system with local interactions— a much simpler object than the original $N$-spin Ising model.

**The coarsest level.** For a $\mathbb{Z}_2$-symmetric system, $p(c^{(n)}) = \frac{1}{2}$ (uniform over $\{+1, -1\}$), contributing exactly $\ln 2$ to the entropy. This single bit of entropy is *structurally guaranteed*—it cannot be lost to mode collapse.

**Within-block factorisation.** The three detail variables $(d_1, d_2, d_3)$ within each block are correlated (through the intra-block energy (76)). They can be factorised autoregressively:

$$p(d_1, d_2, d_3 \mid \text{context}) = p(d_1 \mid \text{ctx})\, p(d_2 \mid d_1, \text{ctx})\, p(d_3 \mid d_1, d_2, \text{ctx}), \tag{81}$$

where "context" denotes the coarse variables and detail variables from adjacent blocks. Each factor is a single binary conditional, parameterised by a neural network.

## 11.4 Implementation plan

We describe a concrete implementation for the $2^n \times 2^n$ Ising model.

**Architecture: scale-equivariant ConvNet.** At each level $k$, the conditional $p(\boldsymbol{d}^{(k)} \mid \text{coarser levels})$ is parameterised by a convolutional neural network $f_\theta^{(k)}$ that takes as input the coarse-level information and outputs the conditional probabilities for the detail variables.

At the critical point, the RG fixed-point structure implies that the conditional has the *same functional form* at every level (up to rescaling). This motivates *weight sharing*: a single ConvNet $f_\theta$ is used at all levels, with the scale index $k$ provided as an additional input (e.g., via a scale embedding added to the feature maps). This reduces the parameter count from $O(n)$ networks to $O(1)$.

The ConvNet operates on the $2^{n-k-1} \times 2^{n-k-1}$ grid of blocks at level $k$. Its input channels are:

- The coarse variables $\boldsymbol{c}^{(k+1)}$ (1 channel, $2^{n-k-1} \times 2^{n-k-1}$).
- Upsampled coarser-level information (optional, for capturing longer-range context).

Its output channels are the parameters of the within-block autoregressive conditionals (81): 3 logits per block (one for each of $d_1$, $d_2$, $d_3$, with $d_2$ and $d_3$ conditioned on earlier detail variables via masking).

**Sampling algorithm.**

1. Sample $c^{(n)} \sim \text{Uniform}\{+1, -1\}$.
2. For $k = n-1, n-2, \ldots, 0$:

(a) Compute the coarse array $\boldsymbol{c}^{(k+1)}$ from the already- generated coarser levels (via the inverse transform applied down to level $k+1$).

(b) Run the ConvNet $f_\theta$ on $\boldsymbol{c}^{(k+1)}$ (with scale embedding $k$) to obtain logits for $\boldsymbol{d}^{(k)}$.

(c) Sample $\boldsymbol{d}^{(k)}$ autoregressively within each block: $d_1 \to d_2 \to d_3$.

3. Apply the full inverse Haar transform to recover $\boldsymbol{\sigma} \in \{+1, -1\}^N$.

The total cost is $n$ ConvNet evaluations (one per level), each on a grid of decreasing size. The dominant cost is level 0 (finest), which operates on a $2^{n-1} \times 2^{n-1}$ grid.

**Log-probability computation.** Given a spin configuration $\boldsymbol{\sigma}$:

1. Apply the forward Haar transform to obtain $\{c^{(n)}, \boldsymbol{d}^{(n-1)}, \ldots, \boldsymbol{d}^{(0)}\}$.
2. Compute $\ln p(c^{(n)}) = -\ln 2$.
3. For $k = n-1, \ldots, 0$: run the ConvNet on $\boldsymbol{c}^{(k+1)}$ to obtain logits, then evaluate $\ln p(\boldsymbol{d}^{(k)} \mid \text{coarser levels})$ as a sum of per-block, per-variable log-probabilities.
4. Sum: $\ln p(\boldsymbol{\sigma}) = -\ln 2 + \sum_k \ln p(\boldsymbol{d}^{(k)} \mid \text{coarser levels})$.

This gives an *exact* log-probability, suitable for the variational free energy bound or the AIS weight.

**Training.** The model can be trained by any of the methods discussed in this paper:

- **Reverse KL (REINFORCE):** minimise $F_{\text{var}}[p_\theta] = \mathbb{E}[E(\boldsymbol{\sigma}) + T \ln p_\theta(\boldsymbol{\sigma})]$. The hierarchical structure mitigates mode collapse: the coarsest level has fixed entropy $\ln 2$, and the shared ConvNet prevents any single level from collapsing independently.
- **Forward KL (MLE on MCMC samples):** minimise $-\mathbb{E}_{p_{\text{Boltz}}}[\ln p_\theta(\boldsymbol{\sigma})]$. The coarse-to-fine structure means the network at each level sees *local* detail variables conditioned on coarser context—a simpler learning problem than the flat AR model.
- **$\beta$-annealing (AIS):** use the Haar-basis AR model as the base $p_\theta$ in the AIS framework of Sec. 9.1. The hierarchical base captures short-range correlations; the AIS chain (with Wolff updates) builds the long-range correlations.

**Complexity comparison.**

| | **Snake-order AR** | **Haar-basis AR** |
|---|---|---|
| Autoregressive depth | $N = L^2$ | $n = \log_2 L$ levels |
| Conditional range | $O(L)$ | $O(1)$ (local) |
| $\mathbb{Z}_2$ symmetry | Must be learned | Exact ($c^{(n)}$ decouples) |
| At $T_c$ | Global dependence | Self-similar across levels |
| Network evaluations | $N$ (or 1 masked) | $n$ ConvNets |
| Weight sharing | None | Across all $n$ levels |

**Relation to the coupling layers.** The checkerboard partition used in the discrete coupling layers (Sec. 4.2) is precisely one level of the Haar decomposition: the $A$ sublattice plays the role of the coarse variables, and the $B$ sublattice plays the role of the detail variables. Stacking $n$ coupling layers with progressively coarser checkerboard partitions recovers the full Haar hierarchy. The Haar-basis AR model makes this multiscale structure explicit and allows weight sharing across levels, which the flat coupling-layer stack does not.

**Practical assessment: concentration at the finest level.** The token count at each level is $1 + 3 + 12 + \cdots + 3(L/2)^2 = L^2$. The finest level alone contributes $3(L/2)^2 = 3L^2/4$, i.e. *75% of all degrees of freedom*, independent of $L$. The coarser levels $(s_0, d_1, \ldots, d_{n-2})$ together account for only $L^2/4$ variables. Consequently, the multi-scale decomposition does not reduce the dimensionality of the hard modelling problem—it reorganises it.

The value of the Haar basis is therefore not in simplifying the finest level, but in providing *structured conditioning context*: when modelling $p(\boldsymbol{d}^{(0)} \mid \text{coarser levels})$, the coarse variables supply a compressed summary of the global domain structure. For small lattices ($L \leq 32$), a flat MADE with sufficient hidden units can capture the same correlations; the Haar basis becomes advantageous at large $L$ where the correlation length at $T_c$ exceeds the effective receptive field of a flat autoregressive model.

The multi-scale decomposition is also valuable as a *diagnostic*: the scale-by-scale conditional entropy $H(\boldsymbol{d}^{(k)} \mid \text{coarser levels})$ reveals which length scales carry the most uncertainty, providing a direct probe of the RG structure that flat models do not offer.

A reference implementation of the bijective Haar transform for $\{+1, -1\}$ spins is provided in `dsflow_ising/multiscale.py`.

**Connection to multi-scale autoregressive models in vision.** The coarse-to-fine autoregressive factorisation (79) is closely related to the Visual Autoregressive Modeling (VAR) framework [?], which replaces next-token prediction with next-scale prediction for image generation. VAR uses a multi-scale VQVAE tokeniser with additive residuals in continuous feature space; the Haar basis provides the discrete analogue, with multiplicative residuals that keep all variables in $\{+1, -1\}$. The Variational Lossy Autoencoder [?] provides a complementary perspective: by limiting the autoregressive decoder's receptive field, it forces global information into the latent code—analogous to how the Haar hierarchy forces long-range structure into the coarse levels.

## 12 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:

$$E(\boldsymbol{\sigma}) = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j. \tag{82}$$

**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 $\Delta F_l$ as functions of $T$.

3. Feed configurations with known defect content (single domain wall, vortex pair) through $f_\phi^{-1}$; 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.

# 13 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.

## 13.1 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.

## 13.2 Normalizing flows for lattice field theory

Albergo, Kanwar, and Shanahan [5] pioneered the use of normalizing flows (specifically Real-NVP affine coupling layers with checkerboard masking) for lattice field theory, demonstrating the method on $\phi^4$ 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.

## 13.3 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.

### 13.4 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 $y_d = (\mu_d + \sigma_d x_d) \bmod K$ that reduces to XOR for binary variables—precisely the coupling layer (12) used here. They tested discrete autoregressive flows on the 2D $q$-state Potts model ($q = 3, 4, 5$; lattices of $3 \times 3$ and $4 \times 4$ spins), training by maximum likelihood on Metropolis–Hastings samples. The flow improved NLL over the autoregressive baseline in most settings, with the largest gains at weak coupling ($J = 0.1$) and larger systems, consistent with the entropy-preservation analysis of Sec. 8.6: forward KL training maintains high entropy in the base, creating the conditions under which the flow can rearrange probability mass to lower energy. However, their work frames the problem as *density estimation*, not variational inference: there is no free-energy objective, no rigorous bound on $F$, no comparison to exact partition functions, and no physical analysis of the learned flow. The systems studied are also very small (9–16 spins), and the flow network is a lookup table that cannot scale to larger lattices. 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.

### 13.5 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.

### 13.6 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.

### 13.7 Positioning of the present proposal

Table 4 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).

## 14 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).

Table 4: 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** | ✓ | ✓ | ✓ | ✓ |

- **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]

[17] R. M. Neal, "Annealed importance sampling," *Stat. Comput.* **11**, 125–139 (2001). [arXiv:physics/9803008]

[18] H. Wu, J. Köhler, and F. Noé, "Stochastic normalizing flows," in *Advances in Neural Information Processing Systems 33* (NeurIPS 2020), pp. 5933–5944. [arXiv:2002.06707]

[19] C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.* **78**, 2690–2693 (1997). [arXiv:cond-mat/9610209]

[20] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," in *Advances in Neural Information Processing Systems 37* (NeurIPS 2024). [arXiv:2404.02905]

[21] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *International Conference on Learning Representations* (ICLR 2017). [arXiv:1611.02731]