HIERARCHICAL NORMALIZING FLOWS FOR INTERPRETABLE SCALE-WISE EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models of multiscale physical systems usually lack the inductive bias needed to generalize to larger systems than they were trained on. We study scale-wise extrapolation with a hierarchical generative model whose levels are the maps between adjacent scales. Because each level adds one scale, a model of a larger system can be constructed by adding a new level, often a copy of an existing one, rather than by retraining. This extrapolation is possible when the way the target distribution changes across scales has low-dimensional structure. The simplest case is self-similarity, which the renormalization group (RG) formalizes as a fixed point of a coarse-graining map. Our proposed model is a hierarchical normalizing flow whose levels are local, symmetry-equivariant continuous normalizing flows, trained with regularizers and a normalization that fix the reparameterization freedom, making the levels identifiable and interpretable. We characterize this freedom and give conditions under which the regularizers leave at most one optimal solution. We then test the method on two-dimensional lattice models, which are exactly self-similar at a phase transition and can be checked against exact results and Monte Carlo simulations. For the 2D Ising model, independently trained models learn the same levels, and across temperatures the learned levels trace a low-dimensional manifold reflecting the RG flow. At the phase transitions of the Ising and XY models, we show that a model trained at one lattice size extends to larger lattices by repeating one level, without new training. In both cases, the extrapolated models match models trained directly on larger systems, whereas a parameter-matched flow-matching model extrapolates poorly.

1 INTRODUCTION

Many complex physical systems, from critical phenomena to turbulence, are multiscale, with structure at every length from the microscopic scale to the system size and no separation of scales. They are where generative models are most needed and where they are most expensive, because a model trained at one size cannot be reused at another: the cost of training grows with the volume, and a convolutional generator, whose receptive field is fixed in lattice units, fails at sizes it was not trained on (He et al., 2024). We ask instead for a model organized by scale, whose levels each capture the dependence between two adjacent scales, so that a larger system is reached by adding a level. Such a model can be extended without new data whenever the dependence between scales has low-dimensional structure. The levels then trace a low-dimensional family of maps, and the level needed at a new scale can be taken from that family. The simplest case is self-similarity, where the map is the same at every scale. For multiscale distributions common in statistical physics, the renormalization group (RG) formalizes this as fixed points in an RG flow. In real space, a local rule merges neighboring degrees of freedom, and iterating it drives the effective couplings toward a fixed point for each phase, with an unstable fixed point at the transition between them (Kadanoff, 1966; Wilson, 1975; Niemeijer & van Leeuwen, 1974; Swendsen, 1979). Lattice models in two dimensions are the natural test case since their structure is well understood.

Hierarchical generative models have also long been read through the RG lens. Their layers have been interpreted as RG steps (Bény, 2013; Mehta & Schwab, 2014; Lin et al., 2017; Iso et al., 2018; De Mello Koch et al., 2020), the optimal coarse graining has been characterized in information-theoretic and optimal-transport terms (Koch-Janusz & Ringel, 2018; Lenggenhager et al., 2020; Gökmen et al., 2021a; Gordon et al., 2021; Kline & Palmer, 2022; Cotler & Rezchikov, 2023a),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and generative models have been built on the analogy as inverse-RG flows (Li & Wang, 2018; Hu et al., 2022), as fixed wavelet transforms with learned per-scale conditionals (Marchand et al., 2023; Lempereur & Mallat, 2026; Yu et al., 2020), as diffusion models read as inverse RG flows (Cotler & Rezchikov, 2023b; Masuki & Ashida, 2025; 2026), and as samplers for lattices larger than those simulated (Efthymiou et al., 2019; Shiina et al., 2021; Bachtis et al., 2022). Related work is discussed further in Appendix C.

The difficulty is the coarse-graining map itself. An expressive hierarchy of invertible maps realizes a given density in many ways, since a reparameterization inserted between two layers and undone by the next leaves the likelihood unchanged. The learned map therefore suffers from a representational degeneracy: it cannot be compared across runs or reused at a new scale, which is what extrapolation requires. Fixing the coarse-graining rule by hand, as in block-spin RG, removes the ambiguity but makes the coarse-graining a design choice. We instead construct a hierarchical normalizing flow that learns a flexible coarse-graining map which remains identifiable and interpretable. This then allows us to design extrapolation methods.

1.1 Contributions

A hierarchical normalizing flow with fixed reparameterization freedom. We propose a hierarchy of continuous normalizing flows on 2×2 cells, one level per scale, with a Haar split between levels and an equivariant vector field. A kinetic-energy term, a marginal term, and a normalization of the hidden variables fitted after training fix its reparameterization freedom.

Identifiability theory. We show that, in a hierarchy that fits the data exactly, the coarse value of each cell retains all the information the cell carries about its surroundings and, when the data density and the coarse-graining maps are analytic, is determined by the data up to a monotone map wherever the cell depends on its surroundings. Under a displacement and a regularity condition, the regularizers then leave at most one optimal hierarchy for given distributions of the fields at every scale.

Interpretable low-dimensional structure. We verify that independently trained models learn the same coarse-graining maps, so maps from different runs and scales can be compared. Across a temperature ladder for the 2D Ising model, we find that the maps lie on a low-dimensional manifold ordered by the correlation length, with the high- and zero-temperature RG fixed points at its ends and the critical fixed point where the two branches meet.

Scale-wise extrapolation in the self-similar case. At a phase transition, we show that the coarse-graining map is the same at every scale, so a model trained on the 2D Ising model on a 64×64 lattice can be extended to 128×128 by repeating one level, whereas a parameter-matched flow-matching model extrapolates poorly. Four repeated levels reach 1024×1024 with no significant loss of accuracy. At the Berezinskii–Kosterlitz–Thouless (BKT) transition of the 2D XY model, a 32×32 model extends to 64×64 the same way with similar results.

2 Method

We learn an exact-likelihood generative model whose levels perform real-space RG steps, each removing one octave of scale (a factor of two in length), following block-spin coarse graining (Kadanoff, 1966; Wilson, 1975) and the neural network RG of Li & Wang (2018). Each local map is a continuous normalizing flow (CNF) with the change of variables computed for its discretized map (Figure 1). Two regularizers together with a normalization of the hidden variables fitted after training single out at most one of the many flows that realize the same density, so that the learned levels can be compared across runs and lattice sizes (Appendix A). Section 3 states what each of them fixes.

Hierarchical block flow. Let $x \in \mathbb{R}^N$ be a real field on an $L \times L$ lattice with periodic boundary conditions ($N = L^2$ sites), drawn from a target $p(x)$. For the 2D Ising model, whose spins are discrete, we use the exact continuous relaxation of Zhang et al. (2012) as in Li & Wang (2018), a real field whose distribution determines the spin distribution exactly (Appendix A.1). For the 2D XY model, a planar unit vector per site, the same relaxation applied to each component gives a two-channel field (Appendix A.1). Before training, a fixed monotone map sends the distribution of a single site value to $\mathcal{N}(0, 1)$, or, for a vector, the distribution of its radius to that of the radius of a standard normal vector.

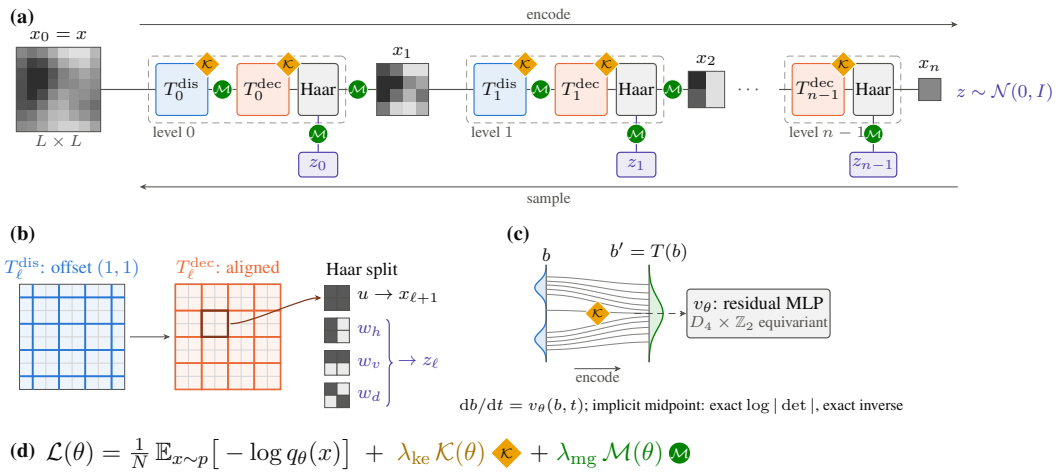


Figure 1: **An identifiable neural RG architecture.** (a) A level applies a shared *disentangler* flow T_ℓ^{dis} to the 2×2 cells at offset $(1, 1)$, a shared *decimator* flow T_ℓ^{dec} to the aligned cells, and a fixed Haar split that keeps the cell average as $x_{\ell+1}$ and sends the three differences to the latent variables z_ℓ. Encoding runs left to right and sampling in reverse. (b) The two partitions and the Haar split of one cell. (c) Each block map is a $D_4 \times \mathbb{Z}_2$-equivariant continuous normalizing flow on $\mathbb{R}^4$. (d) The objective, Equation (2): $\mathcal{K}$ acts inside every block map (diamonds) and $\mathcal{M}$ at every hidden field and at the latent variables (circles).

A *level* ℓ acts on a field $x_\ell \in \mathbb{R}^{N/4^\ell}$ ($x_0 = x$) in three steps, each acting on *cells*, 2×2 groups of sites. A shared bijector applied to every cell of a partition is a *block map*. A *disentangler*, which removes correlations across cell boundaries before decimation, applies one block $T_\ell^{\text{dis}} : \mathbb{R}^4 \to \mathbb{R}^4$ to every cell of the partition offset by one site in each direction and produces an intermediate field h_ℓ; a *decimator*, which decides what each cell passes to the next scale, applies its own block T_ℓ^{dec} to every aligned cell of h_ℓ; and a fixed orthogonal Haar transform (Haar, 1910) of each aligned cell forwards the normalized cell average to the *coarse variables* $x_{\ell+1}$ processed by the next level and sends the three difference coordinates to the *latent variables* z_ℓ. Together, the three steps make level ℓ a map $\mathcal{B}_\ell$ on the whole field x_ℓ. With an i.i.d. standard normal prior on $z = (z_0, \ldots, z_{n-1}, x_n)$, $n = \log_2 L$, the log-likelihood is

$$\log q_\theta(x) = \log \mathcal{N}(z; 0, I) + \sum_{\ell=0}^{n-1} \log \left|\det \nabla \mathcal{B}_\ell(x_\ell)\right| , \tag{1}$$

and sampling runs the same maps in reverse. Equation (1) factorizes as $q_\theta(x) = q(x_n) \prod_\ell q_\theta(x_\ell \mid x_{\ell+1})$, so all dependence between scales 2^ℓ and $2^{\ell+1}$ is created by level ℓ. With identity block maps the model reduces to a Gaussian wavelet pyramid (cf. Marchand et al., 2023).

Block CNF. Each block map is the time-one flow of a neural ODE (Chen et al., 2018; Grathwohl et al., 2019) on $b \in \mathbb{R}^4$, integrated by the implicit midpoint rule (Hairer et al., 2006), so that encoding and sampling are exact inverses and the discretized map's log-determinant is available in closed form (Appendix A.3). The vector field is a small residual MLP conditioned on time, with the lattice symmetry $D_4 \times \mathbb{Z}_2$ built in (Appendix A.2). For the XY model the blocks act on $\mathbb{R}^8$, the global $O(2)$ rotation of the spins replaces $\mathbb{Z}_2$, and each level has three disentanglers (Appendix A.5).

Objective. This architecture realizes a density in many ways, since an invertible map applied at each site of a hidden field (h_ℓ or x_ℓ, the interfaces between layers) and undone by the next layer leaves q_θ unchanged. We train by maximum likelihood with two regularizers that select among them,

$$\mathcal{L}(\theta) = \frac{1}{N}\,\mathbb{E}_{x\sim p}[-\log q_\theta(x)] + \lambda_{\text{ke}}\,\mathcal{K}(\theta) + \lambda_{\text{mg}}\,\mathcal{M}(\theta). \tag{2}$$

The kinetic energy $\mathcal{K}$ is the average over layers of $\mathbb{E}\int_0^1 \|v_\theta(b_t, t)\|^2\,\mathrm{d}t$, the Benamou–Brenier action (Benamou & Brenier, 2000) of the layer's block flow (cf. Finlay et al., 2020; Onken et al., 2021), which prefers, among block maps with the same input and output distributions, the one that moves its input least. The marginal term $\mathcal{M}$ is the empirical squared 2-Wasserstein distance between a

hidden field's pooled one-point marginal and $\mathcal{N}(0,1)$, averaged over the hidden fields h_ℓ and x_ℓ and over the latent variables z_ℓ. It fixes standard normal marginals at every scale, so that scale-to-scale change lives in the dependence structure, and, because the blocks are odd and equivariant, fixes the continuous part of the sitewise reparameterization freedom, leaving one sign per hidden field, which the kinetic energy fixes (Section 3).

Canonical normalization and block fingerprints. After training we fit, at each interface, an odd monotone quantile map Q carrying its marginal to $\mathcal{N}(0,1)$, the *canonical normalization*, and read every block as $\hat{T} = Q_{\text{out}} \circ T \circ Q_{\text{in}}^{-1}$. Adjacent normalizations cancel, so q_θ is unchanged, and the values entering every block and the coarse value leaving every decimator become standard normal (Appendix A.4). Blocks are compared by their *fingerprint*, the displacement field $\hat{T}(z_k) - z_k$ of the block run in the sampling direction on a fixed cloud of standard normal probe points $\{z_k\}$, and labeled by their effective correlation length $\xi_{\text{eff}} = \xi/2^\ell$, where ξ is the correlation length of the training data in lattice units (infinite at β_c).

Scaling readout. The variance v of the normalized 2×2 cell average of the field entering a level equals $2^{2-2\Delta}$ at a fixed point whose order parameter has scaling dimension Δ, the exponent that sets how the magnetization decreases with lattice size. We report the reparameterization-invariant $\Delta_{\max}(\ell) = 1 - \frac{1}{2}\log_2 v_{\max}(\ell)$, with $v_{\max}$ built from maximal correlations, the largest correlation of $\phi(x_i)$ and $\psi(x_j)$ over all functions ϕ, ψ, so that $\Delta_{\max}$ is a lower bound over sitewise reparameterizations. It characterizes the learned coarse graining rather than estimating the continuum exponent (Appendix A.6). It equals 0 for a fully aligned field and 1 for independent sites.

Growing the lattice by repeating a level. A model at $2L$ needs one more level than a model at L. At criticality every interior level should act as the same fixed-point map, so we keep the trained levels, insert a copy of the deepest interior level, the deepest level with both a disentangler and a decimator, and reattach the coarsest block, with every block expressed in its canonical normalization (Appendix A.7). We call the result the *grown* model. No parameter is trained at the larger size.

3 Reparameterization freedom and uniqueness

Many hierarchies realize the same density, the reparameterization freedom of Section 1, so before a learned block map can be interpreted or compared across models we need to know which of its features the training objective fixes. We characterize the reparameterizations that leave the likelihood unchanged (Theorem 1), show that in an exactly fitting hierarchy the coarse value of a decimator carries everything a cell says about the rest of the field (Theorem 2), and give conditions under which the regularizers leave at most one real-analytic hierarchy once the distributions of the fields h_ℓ are fixed (Theorem 3). For almost every weighting of the layers, uniqueness extends across those distributions under one further hypothesis (Corollary 4). Proofs are in Appendix D, and Appendix E summarizes the results and their limitations.

Setting. We write h_ℓ for the field between the disentangler and the decimator of level ℓ. The *hidden variables* are the values at the sites of the fields h_ℓ, $0 \le \ell \le n-2$, and of the coarse fields x_ℓ, $1 \le \ell \le n-1$, and the coarsest level, a single cell, has a decimator only. The *coarse value* of a cell of h_ℓ is the normalized average of its decimator's four outputs, which the Haar transform forwards to $x_{\ell+1}$, and its three *latent values* are the difference coordinates sent to z_ℓ. The same transform applied to the cell's own four values gives its normalized cell average and its *difference coordinates*. The *surroundings* of the cell are the values of h_ℓ in all other cells (Figure 1(b), and Figure D.1 in Appendix D.1). A hierarchy is *exactly fitting* if $q_\theta = p$, and *normalized* if every hidden variable has marginal $\mathcal{N}(0,1)$, which the canonical normalization imposes for hierarchies with the symmetries assumed in Theorem 3. The results concern a *hierarchy*, the collection of its block maps, one per layer, as C^1 diffeomorphisms of $\mathbb{R}^4$ rather than networks. It is *real-analytic* if all its block maps are. Levels with several disentanglers, and fields with several components such as the XY model, are treated in Appendices D.8 and D.9.

Reparameterization freedom. A *reparameterization* inserts a smooth invertible map after a layer and undoes it in the next layer, if there is one. It *keeps the architecture* if the modified layers still apply one shared block map to the same cells.

Theorem 1 (Reparameterization freedom)**.** *The reparameterizations that keep the architecture and leave the likelihood of every hierarchy unchanged are exactly the compositions of: (i) a monotone*

map of $\mathbb{R}$ applied at every site of h_ℓ or x_ℓ, where h_ℓ may have a different map at each of the four positions inside a cell and x_ℓ has a single map; and (ii) a latent reparameterization*: any map of the three latent values of a cell that preserves $\mathcal{N}(0, I_3)$, which may depend on the coarse value (at the coarsest level, whose output goes to the prior entirely, any map of its four outputs that preserves $\mathcal{N}(0, I_4)$).*

Family (i) changes the units of a hidden variable. Family (ii) is far larger, containing every rotation of the latent values that depends smoothly on the coarse value: the likelihood constrains the latent values only through their prior, and what it can fix in a decimator is the coarse value, up to the maps of family (i).

Sufficiency of the coarse value. A decimator *fits the level* if its latent values are i.i.d. $\mathcal{N}(0, I_3)$ over cells and independent of all coarse values, as the decimator of an exactly fitting hierarchy does.

Theorem 2 (Sufficiency of the coarse value)**.** *In an exactly fitting hierarchy: (a) a cell and its surroundings are independent given the coarse value, so the coarse value carries as much mutual information with the surroundings as the whole cell; (b) if p and the block maps are real-analytic and a cell of h_ℓ is not independent of its surroundings, any two decimators acting on the same field h_ℓ that fit the level have the same coarse value up to a monotone map of $\mathbb{R}$.*

By (a) no function of the cell has more mutual information with the surroundings than the coarse value, which is the optimum of the real-space mutual information criterion of Koch-Janusz & Ringel (2018) and Lenggenhager et al. (2020) (Appendix D.4). By (b), for real-analytic p and block maps, the data fix which cell values a decimator that fits the level treats as the same, except at a level whose cells are mutually independent, as at the disordered fixed point, where the likelihood does not fix the coarse value.

Selection by the regularizers. We consider block maps T that *admit a straight flow*, meaning that every intermediate map $b \mapsto (1-t)b + tT(b)$, $t \in [0,1]$, is a diffeomorphism, so that the least kinetic energy of a layer j is the mean squared displacement $\mathcal{K}_j = \mathbb{E}\|T(b) - b\|^2$ of its cells and $\mathcal{K}$ is its average over layers. The kinetic energy of the objective in Section 2 is at least this value, with equality for straight flows. Two conditions enter the uniqueness theorem, stated precisely in Appendix D.5. The *displacement condition* asks that a disentangler move its input less than half as far as any symmetry of its input other than the identity does, a symmetry being an odd, D_4-equivariant block map that, applied to every cell of the layer, preserves the distribution of the field entering it, such as the flip $b \mapsto -b$. For a decimator it asks that the cell move by less than the root-mean-square size of its normalized cell average. The *regularity condition* asks that the distribution of a cell's difference coordinates given the coarse value have a density, so that the optimal transport maps the kinetic energy selects for the decimators exist, and that these be smooth invertible block maps that admit straight flows.

Theorem 3 (Uniqueness given the hidden distributions)**.** *Let p be a positive real-analytic density with finite second moment, invariant under translations, the lattice symmetries and the flip $x \mapsto -x$. Consider the normalized, exactly fitting hierarchies with equivariant odd block maps that admit straight flows and have finite $\mathcal{K}$. If two real-analytic hierarchies of this class minimize $\mathcal{K}$ over the class, satisfy the displacement and regularity conditions, and give every h_ℓ the same distribution, they have the same block map in every layer.*

Normalization fixes the units of every hidden variable up to a sign. Under the displacement condition the kinetic energy fixes that sign and excludes every other symmetry of a disentangler's input. Under the regularity condition it fixes the latent values, the optimal transport of a cell's difference coordinates to the standard normal given the coarse value, and, where the likelihood does not fix the coarse value, and at the coarsest level, the whole decimator (Table E.1). The lattice symmetry is built into the block maps because a hierarchy and its mirror image under a lattice symmetry cost the same kinetic energy (Appendix D.6).

Layer weights and the no-tie assumption. Theorem 3 assumes that the two hierarchies give every hidden field h_ℓ the same distribution. Part of this assumption can be removed. Two hierarchies with the same total kinetic energy can distribute it differently over the layers, but if the layers are weighted, then for almost every choice of weights all minimizers have the same kinetic energy in every layer. What is left is the *no-tie assumption*: hierarchies with the same kinetic energy in every layer also give every h_ℓ the same distribution. We do not prove it, but empirically, independently trained models do agree (Section 4). Write $\mathcal{K}_w = \sum_j w_j \mathcal{K}_j$ for positive layer weights w_j.

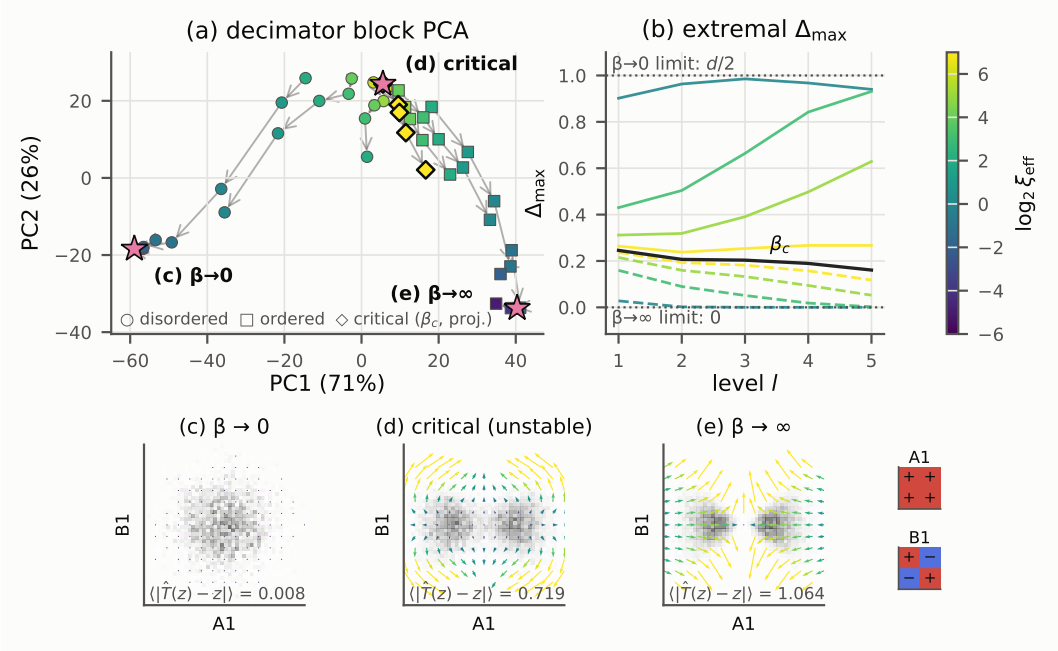


Figure 2: **The learned decimators trace the RG flow between three fixed points.** (a) Principal components of the fingerprints of the interior decimators of the eight temperature-ladder runs (seed 0) at $L = 128$ (circles: disordered, squares: ordered, color: $\log_2 \xi_{\text{eff}}$, arrows: increasing level), with the β_c run projected in (diamonds). The branches meet where ξ_{eff} is largest. (b) Scaling readout $\Delta_{\max}$ per interior level (solid: disordered, dashed: ordered, black: β_c, yellow: $\xi = L$). (c)–(e) The three decimators starred in (a), in the encoding direction, on the slice spanned by the uniform (A_1) and checkerboard (B_1) cell patterns: arrows: in-plane displacement where the data lie, colored by magnitude. Printed: mean four-coordinate displacement over held-out configurations (Appendix B.4).

Corollary 4 (Uniqueness across hidden distributions)**.** *Under the assumptions of Theorem 3, for almost every choice of positive layer weights w, all hierarchies of least $\mathcal{K}_w$ in the class have the same kinetic energy in every layer. If, in addition, hierarchies of least $\mathcal{K}_w$ that spend the same kinetic energy in every layer give every h_ℓ the same distribution, then, for these w, at most one real-analytic hierarchy that satisfies the displacement and regularity conditions has least $\mathcal{K}_w$.*

The second hypothesis is the no-tie assumption. Appendix D.7 discusses it together with the choice of layer weights.

4 Experiments

We study the 2D Ising model, ± 1 spins on a periodic $L \times L$ lattice at inverse temperature β, through the relaxation of Appendix A.1, and we refer to the resulting model as the neural RG (NRG). The XY model follows in Section 4.3. All models share one training setup (Appendix B.1). At the critical coupling $\beta_c = \frac{1}{2}\log(1+\sqrt{2})$ we train two seeds each at $L = 128$ and $L = 64$, and at $L = 128$ a temperature ladder with exact correlation lengths $\xi \in \{1, 8, 32, 128\}$ on the high-temperature (disordered) and low-temperature (ordered) sides of β_c. Samples are scored against held-out Monte Carlo configurations on the magnetization m, the mean spin of a configuration, its distribution $P(m)$ and mean modulus $\langle|m|\rangle$, the two-point function $C(r)$, the correlation of two spins a distance r apart, and the Binder cumulant U_4, a moment ratio of m scored against its exact critical value (Appendix B.3).

4.1 2D Ising model: Identifiability and interpretability

With the reparameterization freedom fixed, independently trained hierarchies learn the same block maps. The decimators of the two β_c seeds at $L = 128$ differ by 2 to 4% of their mean displacement

Table 1: **Extrapolation at the 2D Ising critical point** (β_c). Errors against held-out Monte Carlo configurations at the scored size, and of the Binder cumulant against its exact critical value (metrics and error convention in Appendix B.3, models in Table B.5, floors and sample counts in Appendix B.5). L: training size, then scored size. Two seeds per model, mean $\pm$ half range. Bold marks the better model within each group of rows. *Baseline* models are trained at the scored size, *repeat* grows the $L = 64$ hierarchy by one level per doubling, and *naive* samples the $L = 64$ flow-matching (FM) model at the larger size unchanged.

Model	L	TV_{32}	$\lvert\Delta\langle\lvert m\rvert\rangle\rvert$	$\overline{\lvert C/C_{\mathrm{ref}} - 1\rvert}$	$\lvert U_4 - U_4^*\rvert$
NRG baseline	128	0.057 ± 0.001	0.0046 ± 0.0010	0.033 ± 0.002	$\mathbf{0.0021 \pm 0.0007}$
FM baseline	128	$\mathbf{0.037 \pm 0.002}$	$\mathbf{0.0040 \pm 0.0003}$	$\mathbf{0.013 \pm 0.002}$	0.0050 ± 0.0006
NRG repeat	$64 \to 128$	$\mathbf{0.051 \pm 0.004}$	$\mathbf{0.0057 \pm 0.0026}$	$\mathbf{0.032 \pm 0.006}$	$\mathbf{0.0000 \pm 0.0014}$
FM naive	$64 \to 128$	0.199 ± 0.055	0.0667 ± 0.0264	0.178 ± 0.050	0.0550 ± 0.0321
NRG repeat	$64 \to 256$	$\mathbf{0.050 \pm 0.002}$	$\mathbf{0.0007 \pm 0.0021}$	$\mathbf{0.031 \pm 0.002}$	$\mathbf{0.0027 \pm 0.0013}$
FM naive	$64 \to 256$	0.410 ± 0.095	0.1385 ± 0.0455	0.383 ± 0.088	0.1151 ± 0.0623
NRG repeat	$64 \to 512$	$\mathbf{0.038 \pm 0.001}$	$\mathbf{0.0039 \pm 0.0017}$	$\mathbf{0.034 \pm 0.001}$	$\mathbf{0.0014 \pm 0.0014}$
FM naive	$64 \to 512$	0.604 ± 0.115	0.2013 ± 0.0536	0.560 ± 0.094	0.1878 ± 0.0842
NRG repeat	$64 \to 1024$	$\mathbf{0.046 \pm 0.001}$	$\mathbf{0.0018 \pm 0.0011}$	$\mathbf{0.036 \pm 0.001}$	$\mathbf{0.0011 \pm 0.0013}$
FM naive	$64 \to 1024$	0.758 ± 0.087	0.2552 ± 0.0480	0.685 ± 0.070	0.2673 ± 0.0843

at every level, the disentanglers by 2 to 7%, and on the probe cloud their fingerprints differ by about a quarter of the difference between adjacent levels of one run (Figure B.1). Across the temperature ladder the decimators then form a one-parameter family: two principal components of their fingerprints carry 97% of the variance, and although nothing in the training knows the temperature, the blocks lie along a single curve ordered by ξ_{eff} (Figure 2(a)). The curve has two branches, one for each side of the transition: their far ends are the attractive high- and zero-temperature fixed points of the block-spin RG, and the point where they meet, where the decimators of the independently trained β_c run land, is the unstable critical fixed point (Figure 2(c)–(e)). Within a run the levels move along the curve toward the identity end (disordered) or the aligned end (ordered), while the critical run stays at the meeting point, which is the RG flow of the 2D Ising model read off the learned blocks.

The scaling readout gives the same picture (Figure 2(b)). The β_c run drifts slowly, from $\Delta_{\max} = 0.25$ at level 1 to 0.16 at level 5, and the two $\xi = L$ runs stay within 0.12 to 0.27 at every interior level, while the runs with $\xi \leq 32$ cross over toward 1 (disordered) or 0 (ordered), the earlier the shorter ξ.

Shared multiscale structure across systems. Two microscopically different systems with the same large-scale behavior should share the same coarse graining beyond the finest levels, and the method captures this. The same hierarchy trained on the 2D lattice ϕ^4 theory, a continuous field in the Ising universality class (Appendix B.7), learns block maps that differ from the Ising maps at the finest levels and approach them with depth: the distance between the decimators of the two models at the same level and ξ_{eff} falls from 10 to 30 seed-to-seed gaps at level 1 to a few by levels 4 and 5, and on the ordered side reaches the seed-to-seed gap itself (Figures B.9, B.2 and B.3). The microscopic differences are absorbed by the first levels and shrink with each coarse-graining step, which is what universality asserts. The deeper levels are a structure the two models share.

4.2 2D Ising model: Extrapolation at the self-similar phase transition

A hierarchy trained at $L = 64$ and extended to $L = 128$ by repeating one level reproduces the larger lattice as well as a hierarchy trained there: on all four columns of Table 1 the repeated model is as close to the reference as the $L = 128$ baseline, within the spread between seeds, with no parameter trained at the larger size (Figure 3). At a fixed size the flow-matching model is the better sampler on $P(m)$ and $C(r)$ and comparable on $\langle|m|\rangle$, while both models trained at $L = 128$ hold the Binder cumulant within 0.006 of its critical value, so the hierarchy's advantage is not sample quality at the training size. It is extrapolation: the flow-matching model trained at $L = 64$ and sampled at $L = 128$ is four to six times further from the reference than the repeated hierarchy on the total variation and the correlation function, and more than ten times on $\langle|m|\rangle$ and the Binder cumulant, because its receptive field is fixed in lattice units, whereas the hierarchy's levels are indexed by scale and one more can be

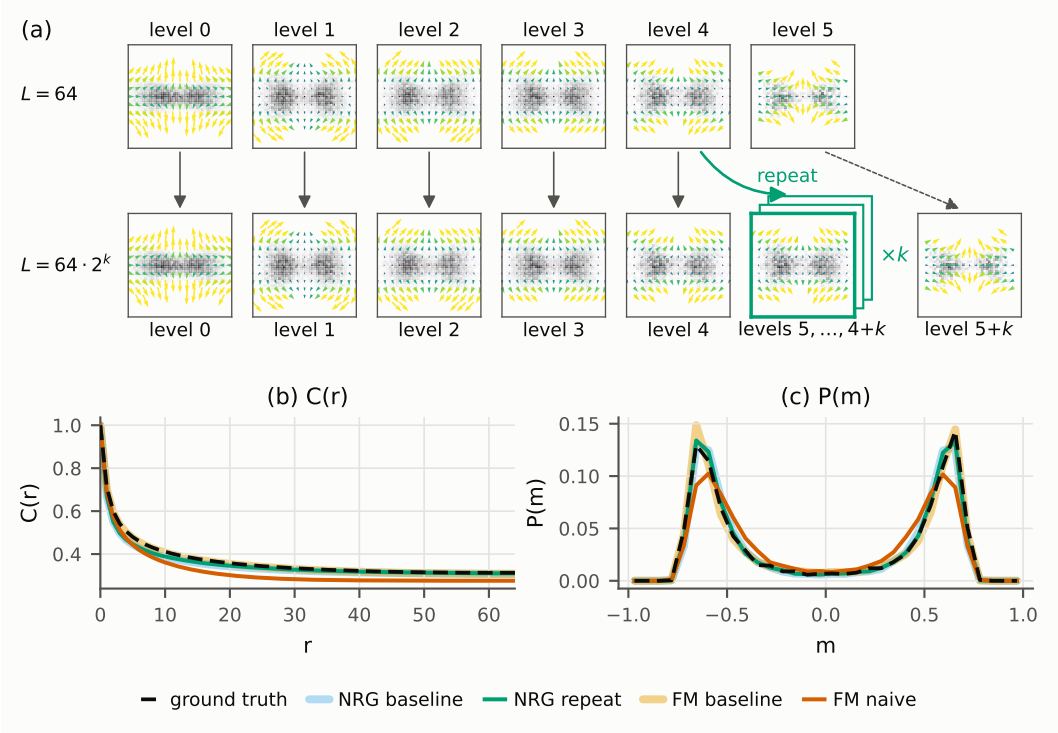


Figure 3: **Extrapolation from $L = 64$ to $L = 128$ at β_c by repeating a level.** (a) The model trained at $L = 64$ (top, with decimator maps as in Figure 2) is grown (bottom) by inserting k copies of level 4, its deepest interior level (green), and reattaching the coarsest decimator. The finest level is never repeated. $k = 1$ is the NRG repeat model of (b) and (c). (b) Two-point function and (c) magnetization histogram against held-out Monte Carlo configurations: cool colors the hierarchy, warm colors flow matching. Thick lines models trained at $L = 128$, thin lines models trained at $L = 64$ and applied at $L = 128$ (*NRG repeat*, *FM naive*).

Table 2: **Extrapolation at the 2D XY transition** (β_{BKT}), as Table 1, with *repeat* growing the $L = 32$ hierarchy, $|m|$ the modulus of the mean spin, U the $O(2)$ Binder cumulant and ρ_v the vortex density, all against held-out configurations (Appendices B.3 and B.5).

Model	L	TV_{32}	$\lvert\Delta\langle\lvert m\rvert\rangle\rvert$	$\overline{\lvert C/C_{\mathrm{ref}} - 1\rvert}$	$\lvert\Delta U\rvert$	$\lvert\Delta\rho_v\rvert$
NRG baseline	64	$\mathbf{0.030 \pm 0.005}$	$\mathbf{0.0014 \pm 0.0009}$	$\mathbf{0.005 \pm 0.003}$	0.0007 ± 0.0002	$\mathbf{0.0008 \pm 0.0001}$
FM baseline	64	0.119 ± 0.006	0.0102 ± 0.0001	0.033 ± 0.000	$\mathbf{0.0004 \pm 0.0001}$	0.0053 ± 0.0002
NRG repeat	$32 \to 64$	$\mathbf{0.029 \pm 0.002}$	$\mathbf{0.0025 \pm 0.0013}$	$\mathbf{0.013 \pm 0.002}$	$\mathbf{0.0014 \pm 0.0003}$	$\mathbf{0.0024 \pm 0.0001}$
FM naive	$32 \to 64$	0.114 ± 0.008	0.0029 ± 0.0064	0.035 ± 0.005	0.0017 ± 0.0010	0.0101 ± 0.0010
NRG repeat	$32 \to 128$	$\mathbf{0.075 \pm 0.014}$	$\mathbf{0.0049 \pm 0.0013}$	$\mathbf{0.020 \pm 0.001}$	$\mathbf{0.0001 \pm 0.0003}$	$\mathbf{0.0024 \pm 0.0001}$
FM naive	$32 \to 128$	0.145 ± 0.059	0.0082 ± 0.0090	0.065 ± 0.000	0.0007 ± 0.0013	0.0132 ± 0.0013
NRG repeat	$32 \to 256$	$\mathbf{0.138 \pm 0.018}$	$\mathbf{0.0091 \pm 0.0014}$	$\mathbf{0.033 \pm 0.003}$	$\mathbf{0.0003 \pm 0.0004}$	$\mathbf{0.0030 \pm 0.0001}$
FM naive	$32 \to 256$	0.196 ± 0.076	0.0107 ± 0.0055	0.088 ± 0.002	0.0041 ± 0.0025	0.0159 ± 0.0010

inserted. The grown and the trained hierarchy agree not only in distribution but as maps: decoding the same prior draw through both gives nearly the same configuration (Appendix B.6).

Extrapolation across four octaves. Repeating the same level k times extends the $L = 64$ models to $L = 64 \cdot 2^k$ (Figure 3(a)). Table 1 follows both seeds to $L = 1024$ without training a parameter beyond $L = 64$: every error stays at the level of the one-octave extrapolation, and the Binder cumulant stays within 0.003 of its critical value. The flow-matching model sampled at the same sizes has no level to repeat, and it drifts further from the reference with every doubling: by $L = 1024$ its total variation reaches 0.76 against a floor of 0.03, and its samples lack the larger domains that appear at the phase transition (Figures 4 and B.5).

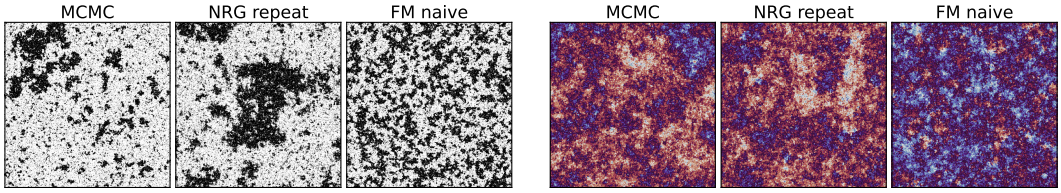


Figure 4: **Extrapolated samples.** The 2D Ising model at β_c, extrapolated from $L = 64$ to $L = 1024$ (left), and the 2D XY model at β_{BKT}, extrapolated from $L = 32$ to $L = 256$ (right). Each group shows a held-out Monte Carlo configuration (MCMC), a sample of the grown hierarchy (NRG repeat) and a sample of the flow-matching model trained at the smaller size (FM naive). XY color is the spin direction.

4.3 XY MODEL: EXTRAPOLATION AT THE BKT TRANSITION

The XY model, a planar unit spin per site, has no ordered phase: a Berezinskii–Kosterlitz–Thouless (BKT) transition at $\beta_{\mathrm{BKT}} = 1.120$ separates a disordered phase from a quasi-ordered phase in which correlations decay as a power law (Berezinskii, 1971; Kosterlitz & Thouless, 1973; Kosterlitz, 1974; Hasenbusch, 2005). We train at β_{BKT} with the blocks adapted to a two-component field with global $O(2)$ symmetry, two seeds at $L = 32$ and at $L = 64$ (Appendix B.2), and repeat the deepest interior level of the $L = 32$ models. The grown model is as close to the Monte Carlo reference as the model trained at $L = 64$ on the magnetization histogram, within a few times of it on the other metrics, and well ahead of the naive flow-matching model on the histogram, the correlation function and the vortex density (Table 2). The vortex density is the most demanding of these. A vortex is a topological defect, a plaquette around which the spin direction winds by a full turn, and its density depends on fine-scale structure rather than on two-point statistics, so a model can match the correlation function and still get it wrong. Learned models are known to struggle with vortices. Classifiers trained on raw XY configurations learn the magnetization rather than the vortices (Beach et al., 2018), and generative models of the XY model have only recently reproduced the vortex density, with a systematic bias near the transition (Zhang, 2025). Here the models differ most. At $L = 64$ the flow-matching model trained there over-produces vortices by 60% and the naive one by more than a factor of two, while the trained hierarchy is within 10% of the Monte Carlo density and the grown one within 30% (Figure B.6). Two and three repeats reach $L = 128$ and $L = 256$, where the grown model stays ahead of the flow-matching model on every metric (Table 2), and at $L = 256$ the flow-matching samples again lack large-scale structure (Figure 4).

5 DISCUSSION

In our construction, identifiability and interpretability are what make extrapolation possible. A level can be reused at a new scale because it is identifiable (Section 3) and interpretable as part of an RG flow. In particular, on the lattice models, the same block maps are learned by independently trained models (Figure B.1), and the block maps lie on the same low-dimensional manifold (Figure 2). Using standard dimensionality reduction, we can extract and interpret the scale-wise dynamics on this manifold as an RG flow and identify the block maps associated with features such as fixed points (Section 4.1). Furthermore, known physical properties, such as the universality of the Ising and ϕ^4 models, appear naturally as the convergence of their block maps with depth (Figure B.9). The result is a model that can be extended to lattice sizes it was never trained on (Figure 3, Tables 1 and 2), where a standard convolutional generative model fails to extrapolate the multiscale structure.

Limitations. In our theoretical analysis, the identifiability result rests on one key assumption that needs to be studied more carefully: that two hierarchies of least kinetic energy cannot spend the same kinetic energy in every layer while giving some hidden field a different distribution (Corollary 4, Appendix D.7). It is not proved, and the agreement between independent training runs is the evidence that it holds here. Growth is also tested only at known phase transitions (Sections 4.2 and 4.3), and lattice models are the test case in which the structure is exact. Extrapolation away from a transition would likely require interpolation between block maps along the curve of Figure 2, to the effective correlation length of the new scale. Future work would likely need to augment the current design to handle more complex systems, such as turbulent flows, where scaling is approximate but the dependence between scales is still expected to be low-dimensional.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and write code, to formulate mathematical claims and assist in writing proofs, to edit the paper to improve readability, and to identify relevant literature and format references. We have reviewed all AI-assisted work: the authors checked the proofs, tested the code, and verified the reported numbers and citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model, the objective and the canonical normalization are defined in Section 2 and Appendix A, which also gives in full the continuous relaxation of the Ising model and its exact normalization, the lattice symmetry, and the likelihood of the discretized flow. Appendix B lists every architectural and training hyperparameter, the datasets and how they were generated, the evaluation protocol of each figure and table, and the flow-matching baselines. The statements of Section 3 are proved in Appendix D, with every assumption stated. Code, configurations and the trained checkpoints will be released upon publication.

REFERENCES

Ryan Abbott, Michael S. Albergo, Aleksandar Botev, Denis Boyda, Kyle Cranmer, Daniel C. Hackett, Alexander G. D. G. Matthews, Sébastien Racanière, Ali Razavi, Danilo J. Rezende, Fernando Romero-López, Phiala E. Shanahan, and Julian M. Urban. Aspects of scaling and scalability for flow-based sampling of lattice QCD. *The European Physical Journal A*, 59(11):257, 2023. doi: 10.1140/epja/s10050-023-01154-w.

M. S. Albergo, G. Kanwar, and P. E. Shanahan. Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D*, 100(3):034515, 2019. doi: 10.1103/PhysRevD.100.034515.

Dimitrios Bachtis, Gert Aarts, Francesco Di Renzo, and Biagio Lucini. Inverse renormalization group in quantum field theory. *Physical Review Letters*, 128(8):081603, February 2022. ISSN 1079-7114. doi: 10.1103/physrevlett.128.081603. URL http://dx.doi.org/10.1103/PhysRevLett.128.081603.

Matthew J. S. Beach, Anna Golubeva, and Roger G. Melko. Machine learning vortices at the Kosterlitz–Thouless transition. *Physical Review B*, 97:045207, 2018. doi: 10.1103/PhysRevB.97.045207.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, January 2000. ISSN 0945-3245. doi: 10.1007/s002110050002. URL http://dx.doi.org/10.1007/s002110050002.

V. L. Berezinskii. Destruction of long-range order in one-dimensional and two-dimensional systems having a continuous symmetry group I. Classical systems. *Soviet Physics JETP*, 32:493–500, 1971.

David S Berman, Marc S Klinger, and Alexander G Stapleton. Bayesian renormalization. *Machine Learning: Science and Technology*, 4(4):045011, October 2023. ISSN 2632-2153. doi: 10.1088/2632-2153/ad0102. URL http://dx.doi.org/10.1088/2632-2153/ad0102.

Piotr Białas, Piotr Korcyl, and Tomasz Stebel. Hierarchical autoregressive neural networks for statistical systems. *Computer Physics Communications*, 281:108502, December 2022. ISSN 0010-4655. doi: 10.1016/j.cpc.2022.108502. URL http://dx.doi.org/10.1016/j.cpc.2022.108502.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax. Version 0.3.13.

Serena Bradde and William Bialek. PCA meets RG. *Journal of Statistical Physics*, 167(3-4): 462–475, March 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1770-6. URL http://dx.doi.org/10.1007/s10955-017-1770-6.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, June 1991. ISSN 1097-0312. doi: 10.1002/cpa.3160440402. URL http://dx.doi.org/10.1002/cpa.3160440402.

R. C. Brower and P. Tamayo. Embedded dynamics for ϕ^4 theory. *Physical Review Letters*, 62: 1087–1090, 1989. doi: 10.1103/PhysRevLett.62.1087.

Cédric Bény. Deep learning and the renormalization group. arXiv preprint arXiv:1301.3124, 2013. URL https://arxiv.org/abs/1301.3124.

Juan Carrasquilla and Roger G. Melko. Machine learning phases of matter. *Nature Physics*, 13(5): 431–434, February 2017. ISSN 1745-2481. doi: 10.1038/nphys4035. URL http://dx.doi.org/10.1038/nphys4035.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jui-Hui Chung and Ying-Jer Kao. Neural Monte Carlo renormalization group. *Physical Review Research*, 3(2):023230, 2021. doi: 10.1103/PhysRevResearch.3.023230.

R.L. Claypoole, G.M. Davis, W. Sweldens, and R.G. Baraniuk. Nonlinear wavelet transforms for image coding via lifting. *IEEE Transactions on Image Processing*, 12(12):1449–1459, December 2003. ISSN 1057-7149. doi: 10.1109/tip.2003.817237. URL http://dx.doi.org/10.1109/TIP.2003.817237.

Jordan Cotler and Semon Rezchikov. Renormalization group flow as optimal transport. *Physical Review D*, 108(2):025003, July 2023a. ISSN 2470-0029. doi: 10.1103/physrevd.108.025003. URL http://dx.doi.org/10.1103/PhysRevD.108.025003.

Jordan Cotler and Semon Rezchikov. Renormalizing diffusion models. arXiv preprint arXiv:2308.12355, 2023b. URL https://arxiv.org/abs/2308.12355.

Ingrid Daubechies and Wim Sweldens. Factoring wavelet transforms into lifting steps. *The Journal of Fourier Analysis and Applications*, 4(3):247–269, May 1998. ISSN 1531-5851. doi: 10.1007/bf02476026. URL http://dx.doi.org/10.1007/BF02476026.

Ellen De Mello Koch, Robert De Mello Koch, and Ling Cheng. Is deep learning a renormalization group flow? *IEEE Access*, 8:106487–106505, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.3000901. URL http://dx.doi.org/10.1109/ACCESS.2020.3000901.

Luigi Del Debbio, Joe Marsh Rossney, and Michael Wilson. Efficient modeling of trivializing maps for lattice ϕ^4 theory using normalizing flows: A first look at scalability. *Physical Review D*, 104 (9):094507, 2021. doi: 10.1103/PhysRevD.104.094507.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Stavros Efthymiou, Matthew J. S. Beach, and Roger G. Melko. Super-resolving the Ising model with convolutional neural networks. *Physical Review B*, 99(7):075113, February 2019. ISSN 2469-9969. doi: 10.1103/physrevb.99.075113. URL http://dx.doi.org/10.1103/PhysRevB.99.075113.

G. Evenbly and G. Vidal. Algorithms for entanglement renormalization. *Physical Review B*, 79 (14):144108, April 2009. ISSN 1550-235X. doi: 10.1103/physrevb.79.144108. URL http://dx.doi.org/10.1103/PhysRevB.79.144108.

Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, September 2003. ISBN 9780470013854. doi: 10.1002/0470013850. URL http://dx.doi.org/10.1002/0470013850.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3154–3164. PMLR, 2020. URL https://proceedings.mlr.press/v119/finlay20a.html.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, New York, 2nd edition, 1999. ISBN 9780471317166.

Mathis Gerdes, Pim de Haan, Corrado Rainone, Roberto Bondesan, and Miranda C. N. Cheng. Learning lattice quantum field theories with equivariant continuous flows. *SciPost Physics*, 15 (6):238, December 2023. ISSN 2542-4653. doi: 10.21468/scipostphys.15.6.238. URL http://dx.doi.org/10.21468/SciPostPhys.15.6.238.

Amit Gordon, Aditya Banerjee, Maciej Koch-Janusz, and Zohar Ringel. Relevance in the renormalization group and in information theory. *Physical Review Letters*, 126(24):240601, June 2021. ISSN 1079-7114. doi: 10.1103/physrevlett.126.240601. URL http://dx.doi.org/10.1103/PhysRevLett.126.240601.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Jonas Greitemann, Ke Liu, and Lode Pollet. Probing hidden spin order with interpretable machine learning. *Physical Review B*, 99(6):060404, February 2019. ISSN 2469-9969. doi: 10.1103/physrevb.99.060404. URL http://dx.doi.org/10.1103/PhysRevB.99.060404.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stéphane Mallat. Wavelet score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 478–491, 2022. doi: 10.52202/068431-0035.

Florentin Guth, Etienne Lempereur, Joan Bruna, and Stéphane Mallat. Conditionally strongly log-concave generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12224–12251. PMLR, 2023. URL https://proceedings.mlr.press/v202/guth23a.html.

Tobias Göbel, Julian R. Ebelt, Zier Mensch, Mathis Gerdes, and Miranda C. N. Cheng. Weight-space physics: Interpretable hypernetworks for lattice quantum field theories. arXiv preprint arXiv:2607.07127, 2026. URL https://arxiv.org/abs/2607.07127.

Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. Statistical physics through the lens of real-space mutual information. *Physical Review Letters*, 127(24):240603, December 2021a. ISSN 1079-7114. doi: 10.1103/physrevlett.127.240603. URL http://dx.doi.org/10.1103/PhysRevLett.127.240603.

Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. Symmetries and phase diagrams with real-space mutual information neural estimation. *Physical Review E*, 104(6):064106, December 2021b. ISSN 2470-0053. doi: 10.1103/physreve.104.064106. URL http://dx.doi.org/10.1103/PhysRevE.104.064106.

Alfred Haar. Zur Theorie der orthogonalen Funktionensysteme: Erste Mitteilung. *Mathematische Annalen*, 69(3):331–371, September 1910. ISSN 1432-1807. doi: 10.1007/bf01456326. URL http://dx.doi.org/10.1007/BF01456326.

Paul Hagemann, Sophie Mildenberger, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *SIAM Journal on Mathematics of Data Science*, 7(3):1337–1366, 2025. doi: 10.1137/23M1614092.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2 edition, 2006. ISBN 978-3-540-30663-4. doi: 10.1007/3-540-30666-8.

Martin Hasenbusch. The two-dimensional XY model at the transition temperature: a high-precision Monte Carlo study. *Journal of Physics A: Mathematical and General*, 38:5869–5883, 2005. doi: 10.1088/0305-4470/38/26/003.

Anna Hasenfratz, Ethan T. Neil, Letizia Parato, and Noah Schwartz. Renormalization-guided cascade upscaling for lattice field generation. arXiv preprint arXiv:2608.28581, 2026. URL https://arxiv.org/abs/2608.28581.

Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. ScaleCrafter: Tuning-free higher-resolution visual generation with diffusion models. In *International Conference on Learning Representations*, 2024.

Adriano Hernandez, Rumen Dangovski, Peter Y. Lu, and Marin Soljačić. Model stitching: Looking for functional similarity between representations. arXiv preprint arXiv:2303.11277, 2023. URL https://arxiv.org/abs/2303.11277.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Peter Holderrieth, Michael Samuel Albergo, and Tommi Jaakkola. LEAPS: A discrete neural sampler via locally equivariant networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23397–23416. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/holderrieth25a.html.

Wanda Hou and Yi-Zhuang You. Machine learning renormalization group for statistical physics. *Machine Learning: Science and Technology*, 4(4):045010, October 2023. ISSN 2632-2153. doi: 10.1088/2632-2153/ad0101. URL http://dx.doi.org/10.1088/2632-2153/ad0101.

Hong-Ye Hu, Shuo-Hui Li, Lei Wang, and Yi-Zhuang You. Machine learning holographic mapping by neural network renormalization group. *Physical Review Research*, 2(2):023369, June 2020. ISSN 2643-1564. doi: 10.1103/physrevresearch.2.023369. URL http://dx.doi.org/10.1103/PhysRevResearch.2.023369.

Hong-Ye Hu, Dian Wu, Yi-Zhuang You, Bruno Olshausen, and Yubei Chen. RG-Flow: a hierarchical and explainable flow model based on renormalization group and sparse prior. *Machine Learning: Science and Technology*, 3(3):035009, 2022. doi: 10.1088/2632-2153/ac8393.

J. Hubbard. Calculation of partition functions. *Physical Review Letters*, 3(2):77–78, July 1959. ISSN 0031-9007. doi: 10.1103/physrevlett.3.77. URL http://dx.doi.org/10.1103/PhysRevLett.3.77.

Satoshi Iso, Shotaro Shiba, and Sumito Yokoo. Scale-invariant feature extraction of neural network and renormalization group flow. *Physical Review E*, 97(5):053304, May 2018. ISSN 2470-0053. doi: 10.1103/physreve.97.053304. URL http://dx.doi.org/10.1103/PhysRevE.97.053304.

Ruoxi Jiang, Peter Y. Lu, Elena Orlova, and Rebecca Willett. Training neural operators to preserve invariant measures of chaotic attractors. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27645–27669, 2023. doi: 10.52202/075280-1202.

Leo P. Kadanoff. Scaling laws for Ising models near T_c. *Physics Physique Fizika*, 2(6):263–272, June 1966. ISSN 0554-128X. doi: 10.1103/physicsphysiquefizika.2.263. URL http://dx.doi.org/10.1103/PhysicsPhysiqueFizika.2.263.

Zahra Kadkhodaie, Florentin Guth, Stéphane Mallat, and Eero P Simoncelli. Learning multi-scale local conditional probability models of images. In *International Conference on Learning Representations*, 2023.

G. Kamieniarz and H. W. J. Blöte. Universal ratio of magnetization moments in two-dimensional Ising models. *Journal of Physics A: Mathematical and General*, 26(2):201–212, 1993. doi: 10.1088/0305-4470/26/2/009.

Gurtej Kanwar, Michael S. Albergo, Denis Boyda, Kyle Cranmer, Daniel C. Hackett, Sébastien Racanière, Danilo Jimenez Rezende, and Phiala E. Shanahan. Equivariant flow-based sampling for lattice gauge theory. *Physical Review Letters*, 125(12):121601, September 2020. ISSN 1079-7114. doi: 10.1103/physrevlett.125.121601. URL http://dx.doi.org/10.1103/PhysRevLett.125.121601.

Bruria Kaufman. Crystal statistics. II. Partition function evaluated by spinor analysis. *Physical Review*, 76(8):1232–1243, October 1949. ISSN 0031-899X. doi: 10.1103/physrev.76.1232. URL http://dx.doi.org/10.1103/PhysRev.76.1232.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations, 2021. URL https://arxiv.org/abs/2111.00254. Differentiable Programming workshop, NeurIPS 2021.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adam G Kline and Stephanie E Palmer. Gaussian information bottleneck and the non-perturbative renormalization group. *New Journal of Physics*, 24(3):033007, March 2022. ISSN 1367-2630. doi: 10.1088/1367-2630/ac395d. URL http://dx.doi.org/10.1088/1367-2630/ac395d.

Maciej Koch-Janusz and Zohar Ringel. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14(6):578–582, March 2018. ISSN 1745-2481. doi: 10.1038/s41567-018-0081-4. URL http://dx.doi.org/10.1038/s41567-018-0081-4.

J. M. Kosterlitz. The critical properties of the two-dimensional XY model. *Journal of Physics C: Solid State Physics*, 7:1046–1060, 1974. doi: 10.1088/0022-3719/7/6/005.

J. M. Kosterlitz and D. J. Thouless. Ordering, metastability and phase transitions in two-dimensional systems. *Journal of Physics C: Solid State Physics*, 6:1181–1203, 1973. doi: 10.1088/0022-3719/6/7/010.

H. A. Kramers and G. H. Wannier. Statistics of the two-dimensional ferromagnet. Part I. *Physical Review*, 60(3):252–262, August 1941. ISSN 0031-899X. doi: 10.1103/physrev.60.252. URL http://dx.doi.org/10.1103/PhysRev.60.252.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Boston, 2002. ISBN 9780817681340. doi: 10.1007/978-0-8176-8134-0. URL http://dx.doi.org/10.1007/978-0-8176-8134-0.

Etienne Lempereur and Stéphane Mallat. Hierarchic flows to estimate and sample high-dimensional probabilities. *Applied and Computational Harmonic Analysis*, 83:101854, March 2026. ISSN 1063-5203. doi: 10.1016/j.acha.2026.101854. URL http://dx.doi.org/10.1016/j.acha.2026.101854.

Patrick M. Lenggenhager, Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. Optimal renormalization group transformation from information theory. *Physical Review X*, 10(1):011037, February 2020. ISSN 2160-3308. doi: 10.1103/physrevx.10.011037. URL http://dx.doi.org/10.1103/PhysRevX.10.011037.

Shuo-Hui Li. Learning non-linear wavelet transformation via normalizing flow. arXiv preprint arXiv:2101.11306, 2021. URL https://arxiv.org/abs/2101.11306.

Shuo-Hui Li and Lei Wang. Neural network renormalization group. *Physical Review Letters*, 121 (26):260601, December 2018. ISSN 1079-7114. doi: 10.1103/physrevlett.121.260601. URL http://dx.doi.org/10.1103/PhysRevLett.121.260601.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025.

Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, July 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1836-5. URL http://dx.doi.org/10.1007/s10955-017-1836-5.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Peter Y. Lu, Samuel Kim, and Marin Soljačić. Extracting interpretable physical parameters from spatiotemporal systems using unsupervised learning. *Physical Review X*, 10(3):031056, September 2020. ISSN 2160-3308. doi: 10.1103/physrevx.10.031056. URL http://dx.doi.org/10.1103/PhysRevX.10.031056.

Tanguy Marchand, Misaki Ozawa, Giulio Biroli, and Stéphane Mallat. Multiscale data-driven energy estimation and generation. *Physical Review X*, 13(4):041038, November 2023. ISSN 2160-3308. doi: 10.1103/physrevx.13.041038. URL http://dx.doi.org/10.1103/PhysRevX.13.041038.

Kanta Masuki and Yuto Ashida. Generative diffusion model with inverse renormalization group flows. arXiv preprint arXiv:2501.09064, 2025. URL https://arxiv.org/abs/2501.09064.

Kanta Masuki and Yuto Ashida. Renormalization group flow matching for scalable local generative modeling. arXiv preprint arXiv:2608.23696, 2026. URL https://arxiv.org/abs/2608.23696.

Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–323, November 1995. ISSN 0012-7094. doi: 10.1215/s0012-7094-95-08013-2. URL http://dx.doi.org/10.1215/S0012-7094-95-08013-2.

Barry M. McCoy and Tai Tsun Wu. *The Two-Dimensional Ising Model*. Harvard University Press, December 1973. ISBN 9780674180734. doi: 10.4159/harvard.9780674180758. URL http://dx.doi.org/10.4159/harvard.9780674180758.

Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning. arXiv preprint arXiv:1410.3831, 2014. URL https://arxiv.org/abs/1410.3831.

Gabriel Melo, Leonardo Santiago, and Peter Y. Lu. Learning to emulate chaos: Adversarial optimal transport regularization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 1089-7690. doi: 10.1063/1.1699114. URL http://dx.doi.org/10.1063/1.1699114.

Cole Miles, Annabelle Bohrdt, Ruihan Wu, Christie Chiu, Muqing Xu, Geoffrey Ji, Markus Greiner, Kilian Q. Weinberger, Eugene Demler, and Eun-Ah Kim. Correlator convolutional neural networks as an interpretable architecture for image-like quantum matter data. *Nature Communications*, 12(1):3905, June 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23952-w. URL http://dx.doi.org/10.1038/s41467-021-23952-w.

Bálint Máté and François Fleuret. Multi-lattice sampling of quantum field theories via neural operator-based flows. *Machine Learning: Science and Technology*, 5(4):045053, December 2024. ISSN 2632-2153. doi: 10.1088/2632-2153/ad9707. URL http://dx.doi.org/10.1088/2632-2153/ad9707.

Kim A. Nicoli, Christopher J. Anders, Lena Funcke, Tobias Hartung, Karl Jansen, Pan Kessel, Shinichi Nakajima, and Paolo Stornati. Estimation of thermodynamic observables in lattice field theories with deep generative models. *Physical Review Letters*, 126(3):032001, January 2021. ISSN 1079-7114. doi: 10.1103/physrevlett.126.032001. URL http://dx.doi.org/10.1103/PhysRevLett.126.032001.

Th. Niemeijer and J. M. J. van Leeuwen. Wilson theory for 2-dimensional Ising spin systems. *Physica*, 71(1):17–40, January 1974. ISSN 0031-8914. doi: 10.1016/0031-8914(74)90044-5. URL http://dx.doi.org/10.1016/0031-8914(74)90044-5.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, September 2019. ISSN 1095-9203. doi: 10.1126/science.aaw1147. URL http://dx.doi.org/10.1126/science.aaw1147.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. OT-Flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021. doi: 10.1609/aaai.v35i10.17113.

Lars Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117–149, February 1944. ISSN 0031-899X. doi: 10.1103/physrev.65.117. URL http://dx.doi.org/10.1103/PhysRev.65.117.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, December 2024. ISBN 9781108832908. doi: 10.1017/9781108966351. URL http://dx.doi.org/10.1017/9781108966351.

Adam Rançon, Ulysse Rançon, Tomislav Ivek, and Ivan Balog. Dreaming up scale invariance via inverse renormalization group. *Physical Review E*, 113(5):055302, May 2026. ISSN 2470-0053. doi: 10.1103/3njc-5tlx. URL http://dx.doi.org/10.1103/3njc-5tlx.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970. ISBN 9781400873173. doi: 10.1515/9781400873173. URL http://dx.doi.org/10.1515/9781400873173.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976. ISBN 9780070542358.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 3rd edition, 1987.

David Schaich and Will Loinaz. An improved lattice measurement of the critical coupling in ϕ_2^4 theory. *Physical Review D*, 79:056008, 2009. doi: 10.1103/PhysRevD.79.056008.

Artan Sheshmani, Yi-Zhuang You, Baturalp Buyukates, Amir Ziashahabi, and Salman Avestimehr. Renormalization group flow, optimal transport, and diffusion-based generative model. *Physical Review E*, 111(1):015304, January 2025. ISSN 2470-0053. doi: 10.1103/physreve.111.015304. URL http://dx.doi.org/10.1103/PhysRevE.111.015304.

Kenta Shiina, Hiroyuki Mori, Yusuke Tomita, Hwee Kuan Lee, and Yutaka Okabe. Inverse renormalization group based on image super-resolution using deep convolutional networks. *Scientific Reports*, 11(1):9617, May 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-88605-w. URL http://dx.doi.org/10.1038/s41598-021-88605-w.

Ankur Singha, Elia Cellini, Kim A. Nicoli, Karl Jansen, Stefan Kühn, and Shinichi Nakajima. Multilevel generative samplers for investigating critical phenomena. In *International Conference on Learning Representations*, 2025.

Wim Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, January 1998. ISSN 1095-7154. doi: 10.1137/s0036141095289051. URL http://dx.doi.org/10.1137/S0036141095289051.

Robert H. Swendsen. Monte Carlo renormalization group. *Physical Review Letters*, 42(14):859–861, April 1979. ISSN 0031-9007. doi: 10.1103/physrevlett.42.859. URL http://dx.doi.org/10.1103/PhysRevLett.42.859.

Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86–88, 1987. doi: 10.1103/PhysRevLett.58.86.

Giacomo Torlai and Roger G. Melko. Learning thermodynamics with Boltzmann machines. *Physical Review B*, 94(16):165134, October 2016. ISSN 2469-9969. doi: 10.1103/physrevb.94.165134. URL http://dx.doi.org/10.1103/PhysRevB.94.165134.

Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Physics*, 13(5):435–439, February 2017. ISSN 1745-2481. doi: 10.1038/nphys4037. URL http://dx.doi.org/10.1038/nphys4037.

G. Vidal. Entanglement renormalization. *Physical Review Letters*, 99(22):220405, November 2007. ISSN 1079-7114. doi: 10.1103/physrevlett.99.220405. URL http://dx.doi.org/10.1103/PhysRevLett.99.220405.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. doi: 10.1090/gsm/058.

Lingxiao Wang, Gert Aarts, and Kai Zhou. Diffusion models as stochastic quantization in lattice field theory. *Journal of High Energy Physics*, 2024(5):060, 2024. doi: 10.1007/JHEP05(2024)060.

Sebastian J. Wetzel. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Physical Review E*, 96(2):022140, August 2017. ISSN 2470-0053. doi: 10.1103/physreve.96.022140. URL http://dx.doi.org/10.1103/PhysRevE.96.022140.

Kenneth G. Wilson. The renormalization group: Critical phenomena and the Kondo problem. *Reviews of Modern Physics*, 47(4):773–840, October 1975. ISSN 0034-6861. doi: 10.1103/revmodphys.47.773. URL http://dx.doi.org/10.1103/RevModPhys.47.773.

Ulli Wolff. Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361–364, January 1989. ISSN 0031-9007. doi: 10.1103/physrevlett.62.361. URL http://dx.doi.org/10.1103/PhysRevLett.62.361.

Dian Wu, Lei Wang, and Pan Zhang. Solving statistical mechanics using variational autoregressive networks. *Physical Review Letters*, 122(8):080602, February 2019. ISSN 1079-7114. doi: 10.1103/physrevlett.122.080602. URL http://dx.doi.org/10.1103/PhysRevLett.122.080602.

Yang yang Tan, Gert Aarts, Diaa E. Habibi, Biagio Lucini, and Lingxiao Wang. Diffusion models for sampling near criticality in lattice field theories. arXiv preprint arXiv:2607.08505, 2026. URL https://arxiv.org/abs/2607.08505.

Jason J. Yu, Konstantinos G. Derpanis, and Marcus A. Brubaker. Wavelet Flow: Fast training of high resolution normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6184–6196, 2020.

Kai Zhang. Learning thermodynamics and topological order of the two-dimensional XY model with generative real-valued restricted Boltzmann machines. *Physical Review B*, 111:024112, 2025. doi: 10.1103/PhysRevB.111.024112.

Yichuan Zhang, Zoubin Ghahramani, Amos Storkey, and Charles Sutton. Continuous relaxations for discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/c913303f392ffc643f7240b180602652-Abstract.html.

A Further Details on the Method

A.1 Targets, continuous relaxation and Gaussianization

Ising model. For spins $s \in \{-1, +1\}^N$ with $p(s) \propto \exp(\frac{\beta}{2} s^\top A s)$, where A is the nearest-neighbor adjacency matrix of the periodic square lattice, let $K = \beta(A + \alpha I)$ with $\alpha > 4$ so that K is positive definite. Since $s_i^2 = 1$, the shift changes only the normalization. The Gaussian integral $\exp(\frac{1}{2} s^\top K s) \propto \int \mathrm{d}x \, \exp(-\frac{1}{2} x^\top K^{-1} x + s^\top x)$ (Hubbard, 1959; Zhang et al., 2012) introduces a real field x with $p(x \mid s) = \mathcal{N}(Ks, K)$, whose marginal for ± 1 spins is the continuous target (Li & Wang, 2018)

$$\log w(x) = -\tfrac{1}{2} x^\top K^{-1} x + \sum_i \log 2 \cosh x_i, \qquad p(s_i = +1 \mid x) = \sigma(2x_i), \tag{3}$$

with σ the logistic function. Spins are recovered by independent draws from $\sigma(2x_i)$, and spin observables are estimated in the Rao–Blackwellized form $\langle s_i s_j \rangle = \langle \tanh x_i \tanh x_j \rangle$.

The normalization is $\log Z_x = \frac{N}{2} \log 2\pi + \frac{1}{2} \log\det K + \frac{N\beta\alpha}{2} + \log Z_{\mathrm{Ising}}(\beta)$, with $\log\det K$ computed spectrally and $\log Z_{\mathrm{Ising}}$ given by Kaufman's exact finite-lattice solution (Kaufman, 1949; Onsager, 1944), so likelihoods can be reported as absolute KL divergences per site. In terms of the Kramers–Wannier dual coupling $\beta^* = \frac{1}{2} \log\coth\beta$ (Kramers & Wannier, 1941), the exact correlation length is $1/\xi = 2(\beta^* - \beta)$ above T_c and $1/\xi = 4(\beta - \beta^*)$ below (McCoy & Wu, 1973).

ϕ^4 theory. The 2D lattice ϕ^4 theory has a real field ϕ_i per site and action $\sum_i \big[\frac{1}{2} \sum_\nu (\phi_{i+e_\nu} - \phi_i)^2 + \frac{1}{2} m^2 \phi_i^2 + \frac{1}{4} \lambda \phi_i^4\big]$, which for $\lambda > 0$ has a phase transition in the Ising universality class. We use $\lambda = 1$, for which Schaich & Loinaz (2009) locate the transition at $m_c^2 = -1.27251(16)$. The field is continuous, so no relaxation is needed and ϕ is the field the hierarchy models. Spin observables are computed on $\operatorname{sign}\phi$. The theory has no exact partition function.

XY model. With a unit vector $s_i \in \mathbb{R}^2$ per site and $p(s) \propto \exp(\beta \sum_{\langle ij \rangle} s_i \cdot s_j)$, the same Gaussian integral applied to each component with $K = \beta(A + \alpha I)$ and $\alpha = 5 > 2d$ gives a field $x \in \mathbb{R}^{L \times L \times 2}$ with

$$\log w(x) = -\tfrac{1}{2} x^\top K^{-1} x + \sum_i \log I_0(|x_i|), \qquad p(s_i \mid x) = \mathrm{vonMises}(x_i/|x_i|, \, |x_i|)\,, \tag{4}$$

where I_0 is the modified Bessel function and the conditional has direction $x_i/|x_i|$ and concentration $|x_i|$. Spins are recovered by one draw per site from it, and every XY observable is computed on those draws. The transition is the Berezinskii–Kosterlitz–Thouless point $\beta_{\mathrm{BKT}} = 1/0.8929 = 1.120$ (Hasenbusch, 2005), and the model has no exact partition function.

Gaussianization. Before training, the field is sent through a fixed sitewise map fitted on the training split. For the Ising and ϕ^4 fields it is a piecewise-linear interpolation between the empirical quantiles of the pooled training values and the standard normal quantiles, antisymmetrized so that it is exactly odd. For the XY field it acts on the radius, $x_i \mapsto f(|x_i|)\, x_i/|x_i|$, with f a monotone piecewise-linear map of the empirical radius quantiles to those of the Rayleigh distribution, the radius law of a standard normal 2-vector. A radial map is automatically odd and $O(2)$-equivariant. Equation (1) is written for the Gaussianized field, and the KL divergences we report are for the field of the relaxation, against its exact $\log Z_x$.

A.2 Exact lattice symmetry

For the Ising and ϕ^4 fields, the four Haar coordinates of a 2×2 cell transform under D_4 as listed in Table A.1, and the global spin flip negates all four. The (A_1, B_1) slices in the figures are the (u, w_d) planes.

The network receives the $D_4 \times \mathbb{Z}_2$ invariants u^2, w_d^2, $w_h^2 + w_v^2$, $w_h^2 w_v^2$, and $u\, w_h w_v w_d$, and its outputs are the coefficients of fixed covariant monomials,

$$\begin{aligned} v_u &= a_1\, u + a_2\, w_h w_v w_d, \qquad v_d = b_1\, w_d + b_2\, u\, w_h w_v, \\ (v_h, v_v) &= c_1\, (w_h, w_v) + c_2\, (u\, w_d w_v,\, u\, w_d w_h) + c_3\, (w_h^3, w_v^3), \end{aligned} \tag{5}$$

Table A.1: **Haar coordinates of a 2×2 cell** and the representations of D_4 under which they transform.

Coordinate	Pattern	Representation of D_4
u	normalized average (uniform)	trivial, A_1
(w_h, w_v)	horizontal and vertical differences	two-dimensional, E
w_d	diagonal difference (checkerboard)	one-dimensional sign representation, odd under reflections in the lattice axes and even under the diagonal ones, which we label B_1

where each monomial of degree k is normalized by $(1 + r^2)^{(k-1)/2}$, with r^2 the squared radius, so that the field grows at most linearly.

A lattice reflection or rotation about a cell center acts in two parts: it permutes the cells, which leaves the model invariant because one block map is shared across positions, and it acts within each cell. Cell equivariance therefore implies exact D_4 and spin-flip equivariance of the whole Ising or ϕ^4 hierarchy. Equivariance under translations that cross cell boundaries is only partially restored by the disentanglers.

XY model. The symmetry is $D_4 \times O(2)$, the global rotation or reflection of all spins replacing the flip. The block map acts on $b \in \mathbb{R}^{4\times 2}$, the four site vectors of a cell, and is built to be exactly equivariant under $O(2)$, rotations and reflections alike, and under D_4. One shared rule is evaluated at each corner u of the cell: with v the opposite corner and n_1, n_2 the two neighbors, it forms the even vectors (x_u, x_v, s), $s = (x_{n_1} + x_{n_2})/\sqrt{2}$, the odd vector $a = (x_{n_1} - x_{n_2})/\sqrt{2}$ and the odd scalars $c = (a \cdot x_u,\ a \cdot x_v,\ a \cdot s)$, feeds the same residual MLP as above (Table B.1) with 13 invariants built from the Gram matrix $bb^\top$ (seven quadratic entries and six quartic products of the c_k, passed through $\log(1 + \cdot)$ and $\operatorname{arcsinh}$), and emits coefficients e_k, o_k for

$$v_u = \sum_k e_k \, (x_u, x_v, s)_k + \frac{\sum_k o_k \, c_k}{1 + \operatorname{tr}(bb^\top)} \, a, \tag{6}$$

so that swapping n_1 and n_2 flips a and c together and leaves v_u unchanged. The Haar split acts per channel, so the coarse field carries one 2-vector per cell and the latent variables three, with an independent standard normal prior on every coordinate.

A.3 Likelihood of the discretized map

Each block map is the time-one flow $\mathrm{d}b/\mathrm{d}t = v_\theta(b, t)$, $t \in [0, 1]$, integrated forward to sample and backward to encode with the same vector field. The Jacobian $J = \partial_b v_\theta$ is computed exactly, since $b \in \mathbb{R}^4$ for the Ising and ϕ^4 models and $b \in \mathbb{R}^8$ for the XY model. The log-determinant of a CNF is usually obtained by integrating the divergence $\nabla \cdot v_\theta$ along with the state, which gives the log-determinant of the exact ODE flow, whereas the map actually applied is its discretization. With a fixed step count, maximum likelihood exploits the difference and reports held-out negative log-likelihoods below the exact bound set by $\log Z_x$, that is, negative KL divergences.

We therefore define the likelihood of the discretized map itself. For the implicit midpoint step $m = b + \frac{h}{2} v_\theta(m, t + \frac{h}{2})$, $b' = 2m - b$, the implicit function theorem gives $\partial m/\partial b = (I - \frac{h}{2} J(m))^{-1}$, so the step Jacobian is $(I - \frac{h}{2} J)^{-1}(I + \frac{h}{2} J)$ and a step contributes $\log|\det(I + \frac{h}{2} J(m))| - \log|\det(I - \frac{h}{2} J(m))|$ to the log-determinant, exactly at any step count. Reversing the step exchanges numerator and denominator, so the log-determinants of encoding and sampling cancel exactly.

The midpoint equation has a unique solution if $\frac{|h|}{2}\|J\|_2 < 1$ holds over the domain. This contraction factor, evaluated at the solution as the local proxy, is monitored per layer. The step count is increased when it approaches one, and a model that exceeded it would be excluded from density-based comparisons. None of the runs reported here did.

A.4 Canonical normalization

For each hidden interface of an Ising or ϕ^4 hierarchy, an odd, monotone, piecewise-linear quantile map Q to $\mathcal{N}(0, 1)$ is fitted to the block inputs recorded while encoding the training split. Oddness

ensures that the normalization does not introduce a $\mathbb{Z}_2$ asymmetry and that a nonzero mean or skew of a hidden marginal remains visible after canonicalization. For the XY model the map acts on the radius of each 2-vector and leaves its direction unchanged, which plays the same role for $O(2)$ (Table A.2).

A block's output normalization is the same object as its successor's input normalization, so within one model the normalizations cancel. A decimator's output normalization is conjugated by the Haar transform onto the coarse component only. The latent variables z_ℓ, the finest input, and the top coarse field are left unmapped. When blocks from two models are composed, as in the growth of Appendix A.7, the normalizations at their interface no longer cancel, and the composition $Q^{\text{in}}_{\text{new}} \circ (Q^{\text{in}}_{\text{old}})^{-1}$ remains at the junction and re-expresses the output of one block in the coordinates the next block was trained on.

Two blocks are compared by the root-mean-square difference of their fingerprints on the shared probe cloud. A principal component analysis of the fingerprints of a family of runs is unsupervised and labeled by ξ_{eff} only afterwards.

A.5 The Ising and XY hierarchies side by side

The ϕ^4 theory uses the Ising hierarchy unchanged, since its field is a real scalar with the same $D_4 \times \mathbb{Z}_2$ symmetry. Table A.2 lists what the XY hierarchy does differently because its symmetry is $D_4 \times O(2)$. Everything not listed, including the objective, the optimizer, the step-count controller, the fingerprints and the growth procedure, is as above and in Appendix B.1.

Table A.2: **The XY hierarchy against the Ising hierarchy.** Only the rows listed differ. Here $r = |x_i|$ is the radius of a site vector and the Rayleigh distribution is the radius law of a standard normal 2-vector.

Component	Ising	XY
Field and symmetry	One real value per site, with $D_4 \times \mathbb{Z}_2$ and the flip negating the field	One 2-vector per site, with $D_4 \times O(2)$ and the rotation or reflection acting on every site vector
Block input	$b \in \mathbb{R}^4$	$b \in \mathbb{R}^{4\times 2}$
Vector field	Invariants and covariants of Appendix A.2	Gram-matrix invariants and the corner rule of Appendix A.2; same trunk, width and conditioning
Layers per level	One disentangler at offset $(1, 1)$, then the decimator	Three disentanglers at offsets $(1, 1)$, $(0, 0)$, $(1, 1)$, then the decimator (the theory covers an odd number, Appendices D.8 and D.9)
Gaussianization of the data	Odd piecewise-linear quantile map to $\mathcal{N}(0, 1)$	Radial map to the Rayleigh distribution, direction kept (Appendix A.1)
Marginal term $\mathcal{M}$	W_2^2 between the pooled one-point marginal and $\mathcal{N}(0, 1)$	W_2^2 between the sorted radii and Rayleigh plotting positions. A hidden field pools all sites, a latent field pools the two stripe details together and the diagonal detail separately
Kinetic energy $\mathcal{K}$	Per-block action divided by 4	Divided by 8
Canonical normalization	Odd monotone quantile map of each interface to $\mathcal{N}(0, 1)$	Monotone map of the radius to the Rayleigh distribution at each interface, direction unchanged. It fixes the continuous radial freedom but not the sign $x \mapsto -x$, the only sitewise orthogonal map that commutes with $O(2)$, which the kinetic energy fixes as in the Ising case
Parameters	2.11×10^6 ($L = 64$)	4.04×10^6 ($L = 64$), 3.27×10^6 ($L = 32$): 1.93×10^5 per layer
Solver at end of training	4 to 128 steps, contraction 0.43 to 0.67	4 steps on every layer but the top decimator (5 to 19); contraction 0.32 to 0.58
Hardware and wall time	See Table B.1	One H200 GPU; about 12 hours at $L = 64$ and 5 hours at $L = 32$

A.6 Extremal scaling readout

$\Delta_{\max}$, reported for the Ising runs, characterizes the coarse-graining scheme rather than the continuum exponent. Under the unit-marginal convention each site of the field entering a level has unit variance,

so the variance of the normalized 2×2 cell average is $v = 1 + 2\rho_{\mathrm{nn}} + \rho_{\mathrm{d}}$, with ρ_{nn} the nearest-neighbor and ρ_{d} the diagonal in-cell correlation, and $v = 2^{2-2\Delta}$ at a fixed point with order-parameter dimension Δ.

To make the reading independent of the sitewise reparameterization freedom, we replace each correlation by the maximal correlation over sitewise reparameterizations, $\rho_{\max} = \sup_{\phi,\psi} \mathrm{Corr}(\phi(x_i), \psi(x_j))$, estimated on the rank scores $\Phi^{-1}((\mathrm{rank} + \frac{1}{2})/n)$, with Φ the standard normal distribution function and n the number of values, with a shuffle-based noise floor, and define

$$\Delta_{\max}(\ell) = 1 - \tfrac{1}{2} \log_2 \big(1 + 2\rho_{\max,\mathrm{nn}} + \rho_{\max,\mathrm{d}}\big). \tag{7}$$

This quantity is invariant under any strictly monotone sitewise map. It is computed on the encoded held-out configurations, on the field x_ℓ entering each level, before that level's disentangler, which is the renormalized field that $\xi_{\mathrm{eff}} = \xi/2^\ell$ labels throughout. At level 0 it is the Gaussianized data itself, so the level-0 reading is a model-independent reference.

Since the extremal function pairs for the two correlations differ in general, $v_{\max}$ bounds the cell variance over sitewise reparameterizations from above rather than being the variance in any one of them, and $\Delta_{\max}$ is correspondingly a lower bound on the scaling dimension that any sitewise reparameterization would report. Its two limits are exact: in the fully ordered limit $\beta \to \infty$ all correlations equal one and $\Delta_{\max} = 0$, and in the fully disordered limit $\beta \to 0$ the sites are independent and $\Delta_{\max} = 1$. Off criticality $\Delta_{\max}(\ell)$ crosses over from its near-critical value toward 1 (disordered side) or 0 (ordered side) at the level where the cell scale reaches ξ.

A.7 Growth by repeating a level

An Ising model trained at $L = 64$ has six levels, $\ell = 0, \ldots, 5$. Levels 0 to 4 have a disentangler and a decimator each, and the coarsest level 5 has a single decimator acting on the last 2×2 field. The model at $L = 128$ has seven. We keep levels 0 to 4 in place, insert a copy of level 4, the deepest level with both layers, as the new level 5, and reattach the coarsest decimator as level 6. Level 4 of the $L = 64$ model and level 5 of the $L = 128$ model act at the same extent, $L/2^\ell = 4$ cells, and at the same ξ_{eff}, which is why the deepest level is the one to copy. The XY model at $L = 32$ has five levels, and its level 3 is copied in the same way, once per doubling.

The copied layers are inserted together with their canonical normalizations, so the only new map is the junction normalization of Appendix A.4 between the output of level 4 and the input of its copy, the new level 5. The finest level, which is specific to the Gaussianization and the microscopic lattice, is never copied.

B Experimental Details

B.1 Architecture and training

Table B.1 lists the settings of the hierarchy for the Ising model. The adaLN-Zero gates are zero-initialized, so that the hierarchy is the identity at initialization. The two regularizer weights were chosen from paired-seed Ising runs at $L = 128$ on the critical and the $\xi = 1$ ordered systems, as the cell of a grid over both weights with the smallest seed-to-seed block gap and the best hidden marginals, after exploratory grids on smaller systems that fixed only the direction of the search. They were then held fixed for every run. The ϕ^4 runs use this recipe unchanged, and the XY runs differ from it only as listed in Table A.2.

All reported numbers use the parameters at the end of training.

B.2 Runs and data

Table B.2 lists every training run of the paper. Each dataset contains 40,000 configurations, 36,000 for training and 4,000 held out, and the Gaussianizing map of Appendix A.1 is fitted on the training split only. The Ising couplings of the temperature ladder are obtained by inverting the exact $\xi(\beta)$ of Appendix A.1, and the ϕ^4 ladder is matched to it as described in Appendix B.7. The Ising and XY configurations are generated by the Wolff cluster algorithm (Wolff, 1989) in parallel chains, the XY

Table B.1: **Architecture, solver, objective, optimizer, and compute** of the Ising hierarchy.

Blocks per level	One disentangler and one decimator on 2×2 cells, followed by the Haar split
Coarsest level	A single block, and therefore only a decimator
Depth	$\log_2 L$; blocks are not shared across levels
Vector field	Residual MLP of width 128 and depth 2, SiLU activations
Time embedding	32 fixed Gaussian Fourier features of scale 16 and a two-layer MLP
Conditioning	adaLN-Zero scale, shift, and zero-initialized gate on each residual block (Peebles & Xie, 2023)
Parameters	2.49×10^6 at $L = 128$, 2.11×10^6 at $L = 64$
Integrator	Implicit midpoint rule, four Newton iterations per step
Steps per layer	16 initially, then between 4 and 128, set at run time by a controller that keeps the contraction factor of Appendix A.3 near 0.75: raised by a factor $2^{1/4}$ when it is exceeded, lowered when the factor stays below $0.75/2^{1/4}$ for 200 steps
Contraction at end of training	0.43 to 0.67 in the reported Ising runs
Objective	Equation (2) on the Gaussianized training split; every term is intensive
Negative log-likelihood	Per site
$\mathcal{K}$	Average over layers of the per-block action divided by the block's input dimension 4
$\mathcal{M}$	Averaged over the hidden fields h_ℓ and x_ℓ and over the latent variables z_ℓ
λ_{mg}	3×10^{-3}, constant after a 200-step linear warm-up
λ_{ke}	Cosine decay from 3×10^{-4} to 3×10^{-5} after the same warm-up
Optimizer	AdamW, weight decay 10^{-4}
Learning rate	10^{-3}, 200 warm-up steps followed by cosine decay
Batch size, steps	128, 4000
Gradient normalization	Division by a running average (decay 0.99) of the global gradient norm, clipping at ten times that average
Precision	Single precision, reduced-precision matrix multiplication disabled
Hardware and wall time	One H200 GPU per run; 1.4 to 1.6 hours at $L = 64$ and 4.3 to 8.3 hours at $L = 128$, depending on the step-count controller
Software	JAX with Equinox (Bradbury et al., 2018; Kidger & Garcia, 2021)

ones in 16 chains per split with alternating random and ordered starts. The ϕ^4 configurations are generated by checkerboard Metropolis sweeps alternated with embedded-Ising cluster flips of the sign field at fixed $|\phi|$ (Brower & Tamayo, 1989), in 16 parallel chains. Thermalization and thinning of each dataset are set from a pilot chain's integrated autocorrelation times of the energy and of $|m|$. Each Ising configuration is lifted to a field by one draw from $\mathcal{N}(Ks, K)$ with $\alpha = 5$, and each XY configuration by one such draw per component. The held-out configurations used as references at lattice sizes where no model was trained are listed with the scoring in Table B.6.

Table B.2: **Training runs.** Every run has two independently trained seeds and its own dataset of 40,000 configurations.

Model	Coupling	L	Used in
Ising	$\beta_c = \frac{1}{2}\log(1+\sqrt{2})$ (Onsager, 1944)	128, 64	Figures 2, 3, B.1; Table 1
Ising ladder	$\xi \in \{1, 8, 32, 128\}$ on each side of β_c	128	Figures 2, B.9
ϕ^4	$m^2 = m_c^2$, $\lambda = 1$	128, 64	Figures B.1, B.9; Table B.7
ϕ^4 ladder	$\xi \in \{8, 32, 128\}$ on each side of m_c^2	128	Figure B.9
XY	$\beta_{\mathrm{BKT}} = 1.120$	64, 32	Table 2

B.3 Observables and metrics

Every quantity reported in the paper is listed in Tables B.3 and B.4. Spin observables are computed on model samples after mapping the sampled fields back through the inverse Gaussianizing map. Ising spins are drawn from $p(s_i = +1 \mid x) = \sigma(2x_i)$, and where the observable allows it the draw is integrated out (Rao–Blackwellized) by replacing s_i with $\tanh x_i$. The ϕ^4 observables use $\operatorname{sign}\phi$, and XY spins are drawn from the von Mises conditional (Appendix A.1). References are the held-out Monte Carlo configurations of the same size and, for the Ising Binder cumulant, its exact critical value. The ϕ^4 theory and the XY model have no exact reference and every one of their errors is against the held-out configurations. An error against a reference is the absolute difference between the seed mean and the reference, quoted with half the range of the seeds' own values.

Table B.3: **Observables.** $N = L^2$ sites, β the inverse temperature, $\langle\cdot\rangle$ the mean over samples.

Quantity	Definition	Estimator
m	Magnetization $N^{-1}\sum_i s_i$	Sampled spins
$\langle\lvert m\rvert\rangle$	Mean absolute magnetization	Sampled spins
$P(m)$	Histogram of m on 32 equal bins of $[-1, 1]$ (for the XY model, of $\lvert m\rvert$ on 32 equal bins of $[0, 1]$)	Sampled spins
U_4	Binder cumulant $1 - \langle m^4\rangle/(3\langle m^2\rangle^2)$; its exact critical value for the periodic square lattice is $U_4^* = 0.6107$ (Kamieniarz & Blöte, 1993)	Sampled spins
$C(r)$	Two-point function $\langle s_i s_{i+r}\rangle$ at separation r along a lattice axis, averaged over both axes and all sites	$\langle\tanh x_i \tanh x_{i+r}\rangle$
KL/N	Forward divergence $N^{-1}\mathbb{E}_p[\log p - \log q_\theta]$ on held-out configurations, with the exact $\log Z_x$ of Appendix A.1	Standard error over configurations
$\Delta_{\max}$	Extremal scaling readout	Appendix A.6
$\lvert m\rvert$ (XY)	Modulus of the mean spin vector $N^{-1}\sum_i s_i$, $s_i \in \mathbb{R}^2$	Spins drawn from the von Mises conditional (Appendix A.1)
U (XY)	$O(2)$ Binder cumulant $1 - \langle\lvert m\rvert^4\rangle/(2\langle\lvert m\rvert^2\rangle^2)$	Sampled spins
ρ_v (XY)	Vortex density: plaquettes around which the spin angle winds by $\pm 2\pi$, of either sign, per plaquette	Sampled spins
$C(r)$ (XY)	$\langle s_i \cdot s_{i+r}\rangle$ at separation r along a lattice axis, averaged over both axes and all sites	Sampled spins

Errors are quoted as follows. Tables 1, 2 and B.7 give the mean over two seeds $\pm$ half their range. With two seeds per model the half range indicates the seed-to-seed variation, not a confidence interval, and a half range printed as 0.000 is below 0.0005.

Table B.4: **Comparison metrics.** A subscript ref denotes the reference value.

Metric	Definition
TV_{32}	Total variation $\frac{1}{2}\sum_b \lvert P_b - P_b^{\mathrm{ref}}\rvert$ between the 32-bin histograms of m. Its sampling floor is the value a perfect sampler would show, $\frac{1}{2}\sum_b \sqrt{2P_b(1-P_b)/(\pi n_{\mathrm{eff}})}$ with $n_{\mathrm{eff}}^{-1} = n_{\mathrm{ref}}^{-1} + n_{\mathrm{model}}^{-1}$ and P_b the pooled histogram (Table B.6)
$\lvert\langle\lvert m\rvert\rangle - \mathrm{ref}\rvert$	Absolute error of the mean absolute magnetization
$\lvert U_4 - U_4^*\rvert$	Distance of the Binder cumulant from its exact critical value
$\lvert U - U_{\mathrm{ref}}\rvert$, $\lvert\rho_v - \rho_{v,\mathrm{ref}}\rvert$ (XY)	Absolute errors of the $O(2)$ Binder cumulant and of the vortex density against the held-out configurations of the same size
$\overline{\lvert C/C_{\mathrm{ref}} - 1\rvert}$	Mean relative error of $C(r)$: over every $1 \le r \le 64$ at $L = 128$ (Table 1), over $r \ge 1$ on a grid of every $r \le 16$ and about 24 separations per octave up to $L/2$ in the rows scored beyond $L = 128$, over every $1 \le r \le L/2$ for the XY model, and over $r \le L/2$ for the ϕ^4 theory
Fingerprint difference	Root-mean-square difference, per component, of the displacement fields of two blocks on the shared probe cloud (Appendix B.4)
Map agreement	Pearson correlation between the fields two models decode from the same prior draw, and the fraction of sites whose spins agree, averaged over the shared draws; mismatched draws pair different prior draws

B.4 Block characterization

This section gives the protocol behind Figure 2 and the flow-map figures below, which concern the Ising and ϕ^4 runs at $L = 128$.

Canonical normalizations. They are fitted on the full 36,000-configuration training split of each run and enter every quantity below.

Fingerprints. The fingerprint of a block is its displacement field on one fixed set of 8,192 standard normal points in $\mathbb{R}^4$, the probe cloud of Section 2, pushed through the block in the sampling direction, where the block's input distribution is standard normal. The principal component analysis of Figure 2(a) uses the fingerprints of the interior decimators, those of every level except the finest and the coarsest, of the eight Ising ladder runs at seed 0. The seed gaps of Figure B.9 use both seeds of every rung.

Representative decimators. The three decimators of Figure 2(c)–(e), all from Ising runs, are the first interior block (level 1) of the β_c run, where ξ is infinite (critical) and the deepest interior block of the $\xi = 1$ run on each side of the transition (disordered for $\beta \to 0$, ordered for $\beta \to \infty$).

Flow maps. Unlike the fingerprints, the flow maps of Figure 2(c)–(e) are tied to the data rather than to the probe cloud: the arrows are the displacement of the block in the encoding direction, under the canonical normalization, evaluated on a grid in the (A_1, B_1) slice of the cell (Appendix A.2) and masked to the region occupied by the 4,000 encoded held-out configurations of the corresponding run, and the printed number is the mean displacement norm over those configurations. The two directions are used deliberately: the input distribution of a block is standard normal only in the sampling direction, so fingerprints are comparable across runs there, while the structure of the data is visible only in the encoding direction. Figures B.2 and B.3 show the encode-direction maps of every decimator along three trajectories, and Figure B.1 those of the two critical seeds of each model with their difference.

Scaling readout. The $\Delta_{\max}$ profiles of Figure 2(b) are computed for the Ising runs on the 4,000 held-out configurations of each run and shown for the interior levels 1–5 only: level 0 is specific to the non-universal input transformation, and at level 6 the field has only 2×2 sites, so correlations measured across its periodic boundary are biased downward.

B.5 Extrapolation

Table B.5 lists the Ising and XY models of Tables 1 and 2 and Figure 3, and Table B.6 the sample counts, references and total-variation floors at every scored size. The observables and metrics are those of Appendix B.3.

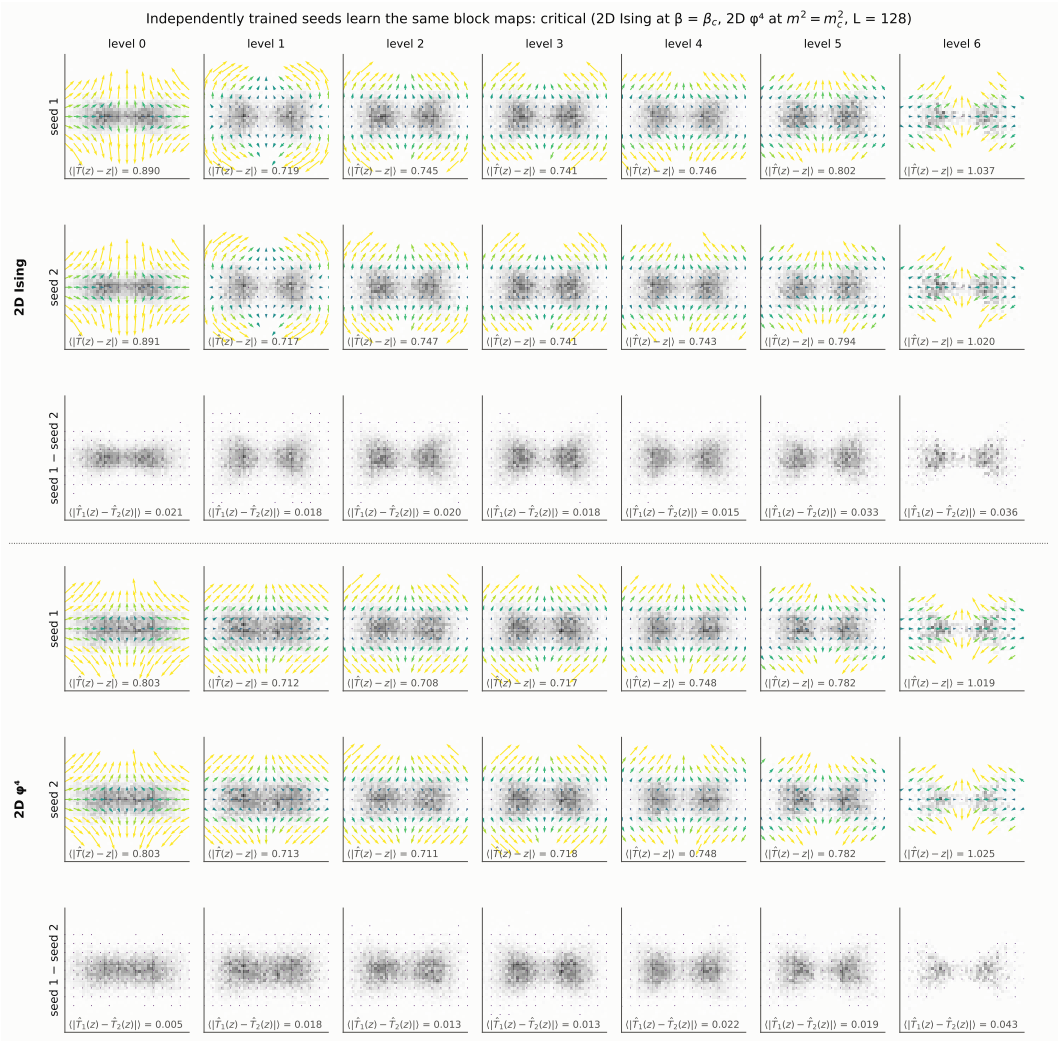


Figure B.1: **Independently trained seeds learn the same block maps.** Encode-direction decimator maps at every level of the two critical runs at $L = 128$ of the 2D Ising model (top) and of the 2D ϕ^4 theory (bottom), on the (A_1, B_1) slice as in Figure 2(c)–(e). The third row of each system is the difference of the two seeds' maps, $\hat{T}_1(z) - \hat{T}_2(z)$, drawn on the same arrow and color scale as the maps. Its printed number is the mean four-coordinate difference of the two maps over the same encoded held-out points, to be read against the displacement printed in the seed 1 panel above it. The seeds differ by 2 to 4% of the displacement at every level (0.6% at the first ϕ^4 level), and the Ising disentanglers, not shown, by 2 to 7%. On the shared probe cloud the Ising pair's root-mean-square fingerprint difference is 0.008 to 0.018 per component at the interior levels, against 0.029 to 0.068 between adjacent levels of one run, about a quarter of one level step. The seed gaps of both models are the floor of Figure B.9.

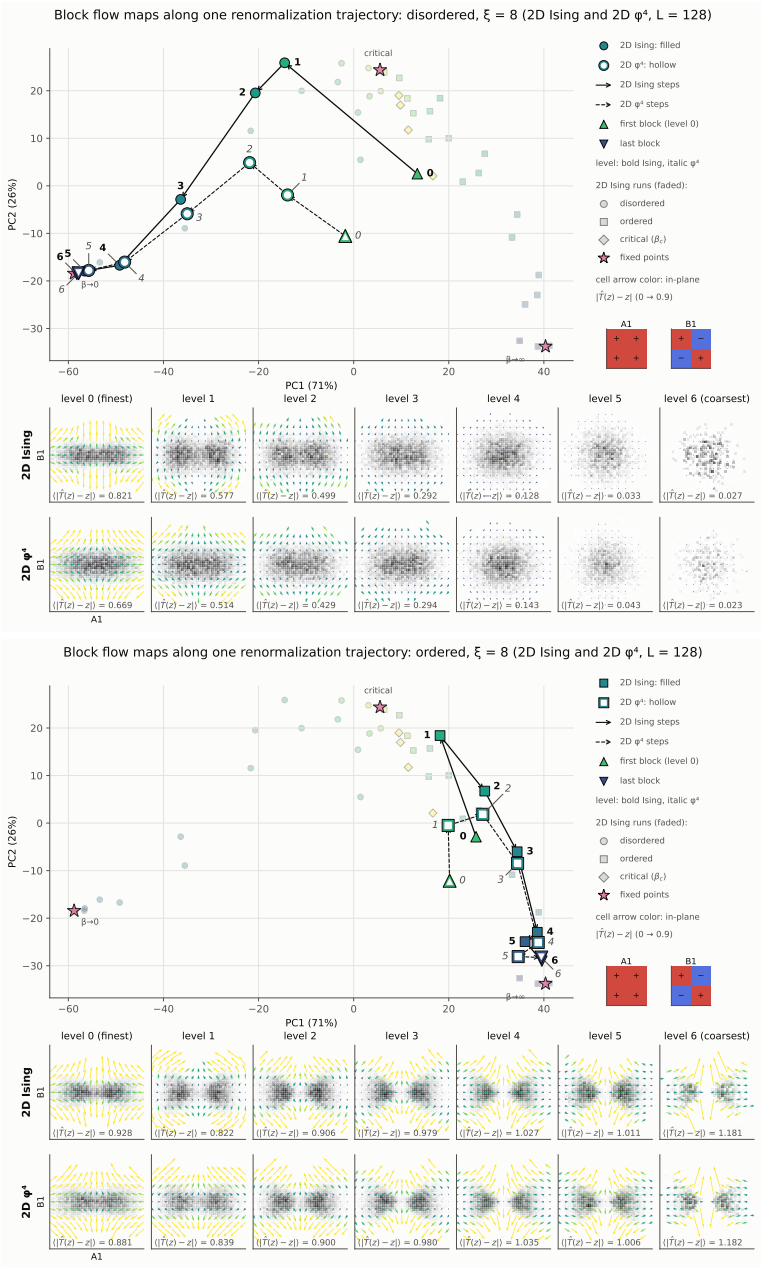


Figure B.2: **Block maps along two off-critical trajectories** (top: disordered side, bottom: ordered side, both at $\xi = 8$), for the 2D Ising run and its matched 2D ϕ^4 run (Appendix B.7). In each panel the chart locates the decimators of both runs in the principal-component plane of Figure 2(a), which is fitted on the Ising ladder only (filled markers and solid arrows: Ising; hollow markers and dashed arrows: ϕ^4), and the panels show the exact block map of every level, the Ising row above the ϕ^4 row, as in Figure 2(c)–(e): in-plane arrows on a grid in the (A_1, B_1) slice, masked to the region the data occupy, over the empirical block-input distribution, with the mean four-coordinate displacement over encoded held-out configurations printed in each panel. The first block is specific to the non-universal input transformation and the last to the finite lattice, which is why neither enters the principal-component fit, although both are marked in the chart.

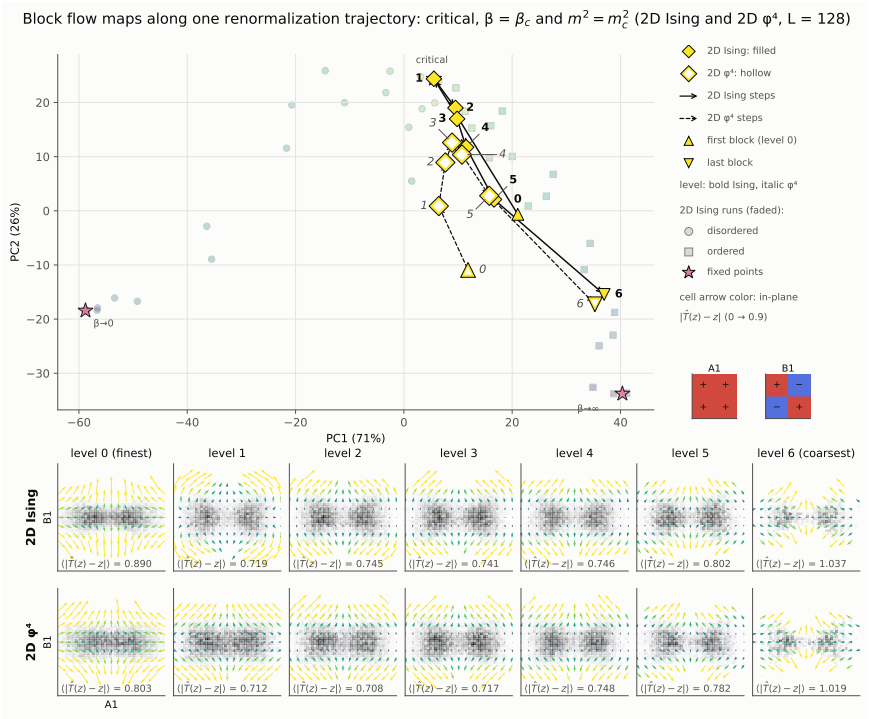


Figure B.3: **Block maps along the critical trajectory** ($L = 128$; 2D Ising at β_c and 2D ϕ^4 at $m^2 = m_c^2$), in the same format as Figure B.2.

Table B.5: **Models compared in the extrapolation experiments.** Each row of Tables 1 and 2 is the mean over two independently trained seeds.

Model	Trained at	Applied at	Construction
NRG baseline	$L = 128$	$L = 128$	Ising hierarchy trained at β_c on the $L = 128$ data
NRG repeat	$L = 64$	$L = 128$ to 1024	$L = 64$ critical run grown by repeating its deepest interior level once per doubling (Appendix A.7)
FM baseline	$L = 128$	$L = 128$	U-Net of base width 17 (2.51×10^6 parameters)
FM naive	$L = 64$	$L = 128$ to 1024	U-Net of base width 16 (2.23×10^6 parameters), sampled with the same weights, which the fully convolutional architecture permits
NRG baseline (XY)	$L = 64$	$L = 64$	XY hierarchy trained at β_{BKT} on the $L = 64$ data (4.04×10^6 parameters, Table A.2)
NRG repeat (XY)	$L = 32$	$L = 64$ to 256	$L = 32$ run at β_{BKT} (3.27×10^6 parameters) grown by repeating its deepest interior level once per doubling
FM baseline (XY)	$L = 64$	$L = 64$	U-Net of base width 22 with two channels (4.20×10^6 parameters)
FM naive (XY)	$L = 32$	$L = 64$ to 256	U-Net of base width 19 with two channels (3.14×10^6 parameters), sampled with the same weights

Table B.6: **Scoring of Tables 1, 2 and B.7.** Samples drawn per seed, held-out Monte Carlo configurations used as the reference, and the resulting sampling floor of TV_{32} (Table B.4). No model is trained at $L \geq 256$, so those rows have no trained reference model.

Model	Scored at L	Samples per seed	Held-out configurations	TV_{32} floor
Ising	128	32,768	4,000	0.028
Ising	256	4,096	4,000	0.036
Ising	512, 1024	4,096	3,686	0.035, 0.033
ϕ^4	128, 64	32,768	4,000	0.027, 0.029
XY	64	32,768	4,000	0.015
XY	128	4,096	4,000	0.020
XY	256	4,096	3,686	0.019

Flow-matching models. The flow-matching models use a convolutional U-Net (Ho et al., 2020; Dhariwal & Nichol, 2021) without attention blocks, trained by flow matching (Lipman et al., 2023). It has six resolutions, from L down to $L/32$ at the training size, with channel multipliers $(1, 1, 2, 2, 4, 4)$, three residual blocks per resolution, periodic convolutions, one channel for the Ising model and two for the XY model, and exact symmetry under the global spin flip by antisymmetrization of the output. For the XY model the antisymmetrized output is exactly equivariant only under the rotation by π, and the rest of $O(2)$ enters through random global rotations and reflections of each training pair, whereas the hierarchy's vector fields are $O(2)$-equivariant by construction. Training uses conditional flow matching on the linear interpolant, on the same datasets as the hierarchies, with the same optimizer, batch size, and step budget, two seeds per model, and base widths chosen to match the hierarchies' parameter counts (Table B.5). Samples are drawn from the parameters at the end of training, as for the hierarchies, by a 64-step midpoint integration.

Growth beyond one octave. The Ising rows of Table 1 scored beyond $L = 128$ repeat level 4 of the same two $L = 64$ models $k = 2, 3, 4$ times, and the XY rows of Table 2 beyond $L = 64$ repeat level 3 of the two $L = 32$ models twice and three times. The FM naive rows sample the same flow-matching models at each size unchanged. These rows are scored on fewer samples, so their total-variation floor is higher (Table B.6). The Ising one-octave step $k = 1$, scored the same way, agrees with its 32,768-sample row within 1.7 standard errors. Figure B.5 shows Ising samples of both models at every size beside Monte Carlo configurations, and Figure B.7 the same for the XY model.

Likelihoods. The likelihood of a flow-matching model is available only through the probability-flow ODE, with a divergence estimate in place of the exact log-determinant of the discretized sampler, so the models are compared on observables. The KL per site of the Ising hierarchies, evaluated on the same held-out configurations against the exact normalization of Appendix A.1, is 0.00789 for the trained $L = 128$ models and 0.00825 to 0.00833 for the repeated models (standard error 0.00002, both seeds). It is not used to rank the models, since the flow-matching models have no exact likelihood. The ϕ^4 theory and the XY model have no exact partition function, so no KL divergence is reported for them.

Failure mode of the naive flow-matching model. Applied at $L = 128$, the Ising network trained at $L = 64$ produces samples whose $C(r)$ is too low at every separation and whose histogram fills the region between the two peaks, because its receptive field is fixed in lattice units and nothing in it corresponds to the added octave of scale, whereas the hierarchy's levels are indexed by scale and one more can be inserted. The XY network shows the same failure in its vortex density, which it over-produces at every size (Figures B.6 and B.7). Figure B.4 shows Ising samples of the four models beside held-out configurations, and Figure B.6 XY samples.

B.6 Map agreement between the grown and the trained model

For the Ising model at β_c, the grown and the trained model agree not only in distribution but as maps. Decoding the same prior draw through the model trained at $L = 128$ and through the model grown from $L = 64$ gives fields with correlation 0.995 (worst draw 0.963) and 98.1% spin agreement over 16 shared draws, against a mean absolute correlation of 0.022 for mismatched draws (independent fields of a $\mathbb{Z}_2$-symmetric model have 50% spin agreement) (Figure B.8).

The two models were trained independently, at different lattice sizes, so this is a statement about the training objective rather than about one run: the hierarchy with its reparameterization freedom fixed

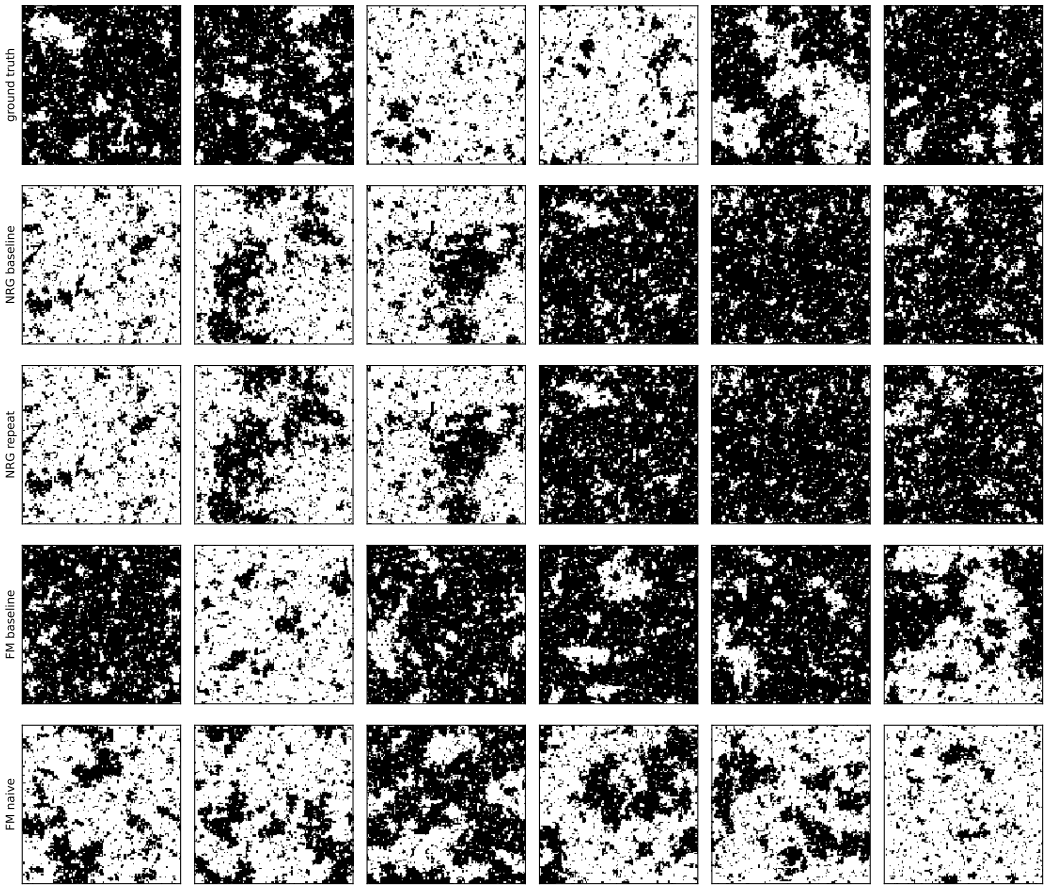


Figure B.4: **Ising samples at** $L = 128$**.** Rows: ground truth, NRG baseline, NRG repeat, FM baseline, FM naive. Each panel shows the sign of a sampled field mapped back through the inverse quantile transform. The two NRG rows decode the same prior draw in each column, so their panels are aligned by construction. The ground-truth and FM panels are independent draws. The global sign is a symmetry.

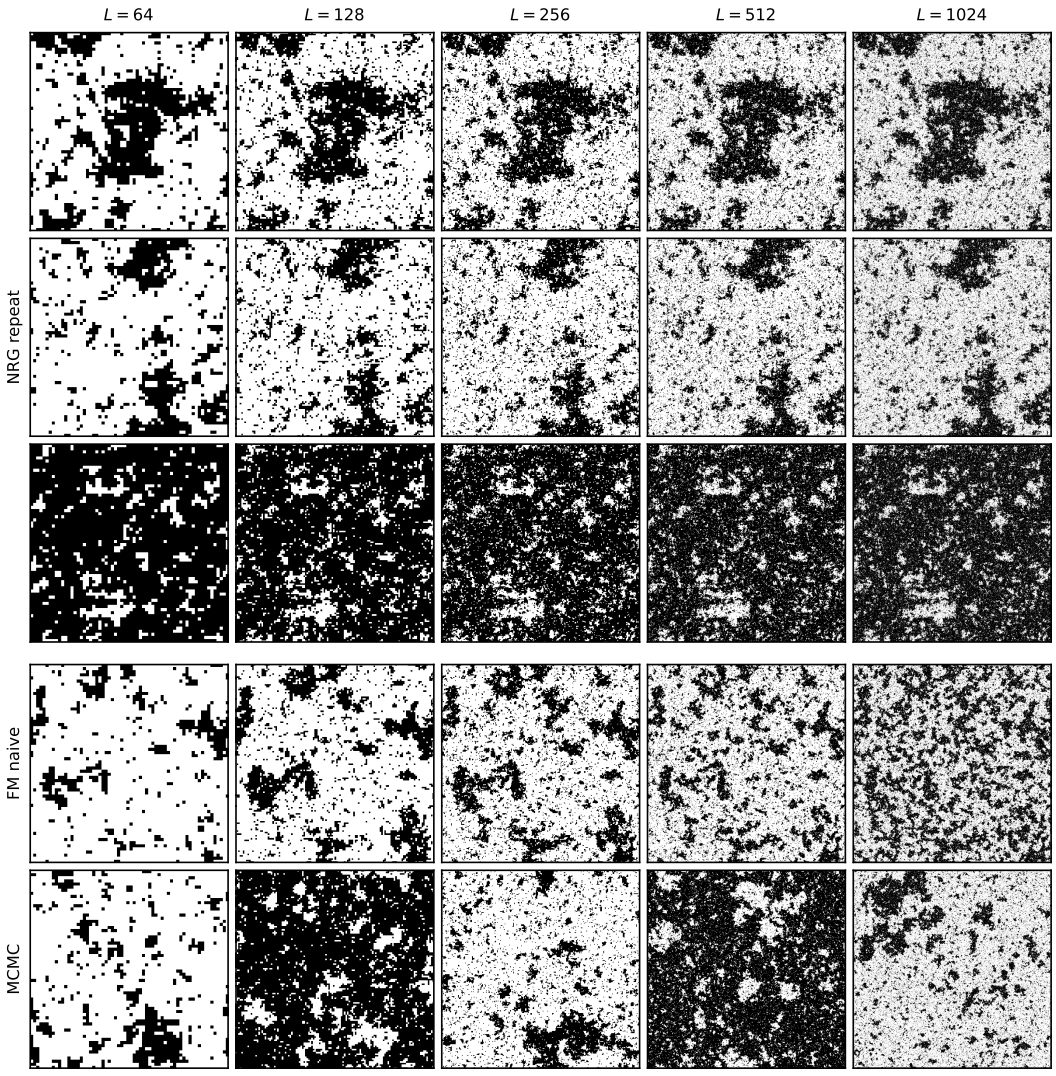


Figure B.5: **Ising samples across four octaves.** Columns are $L = 64$ to 1024. *NRG repeat*: three latent draws decoded by the $L = 64$ model (seed 0) grown to every size. One prior draw is made for the largest lattice, and each smaller lattice decodes the same top field x_n and the same coarsest latent variables z_ℓ, so each row is one configuration to which every doubling adds one finer octave of detail. The draws sit at the 10th, 28th and 75th percentiles of $|m|$ at $L = 128$, and the third is shown globally sign-flipped, which is the decode of the negated latent and an equally likely draw. *FM naive*: the $L = 64$ flow-matching model integrated from one noise field drawn at $L = 1024$ and Haar-averaged down to each smaller size, which couples the sizes as the shared latent does. These samples are for display only, and the rows of Table 1 are scored on independent samples. *MCMC*: at each size the held-out configuration whose $|m|$ is closest to the median of its split, with the sign alternating across columns. The hierarchy keeps one configuration across sizes at ever finer lattice spacing. The flow-matching model shares some large-scale structure up to $L = 512$ and loses its large domains at $L = 1024$. Spins are drawn per site from $p(s_i = +1 \mid x)$, and at $L = 512$ and 1024 the panels have more sites than printed pixels.

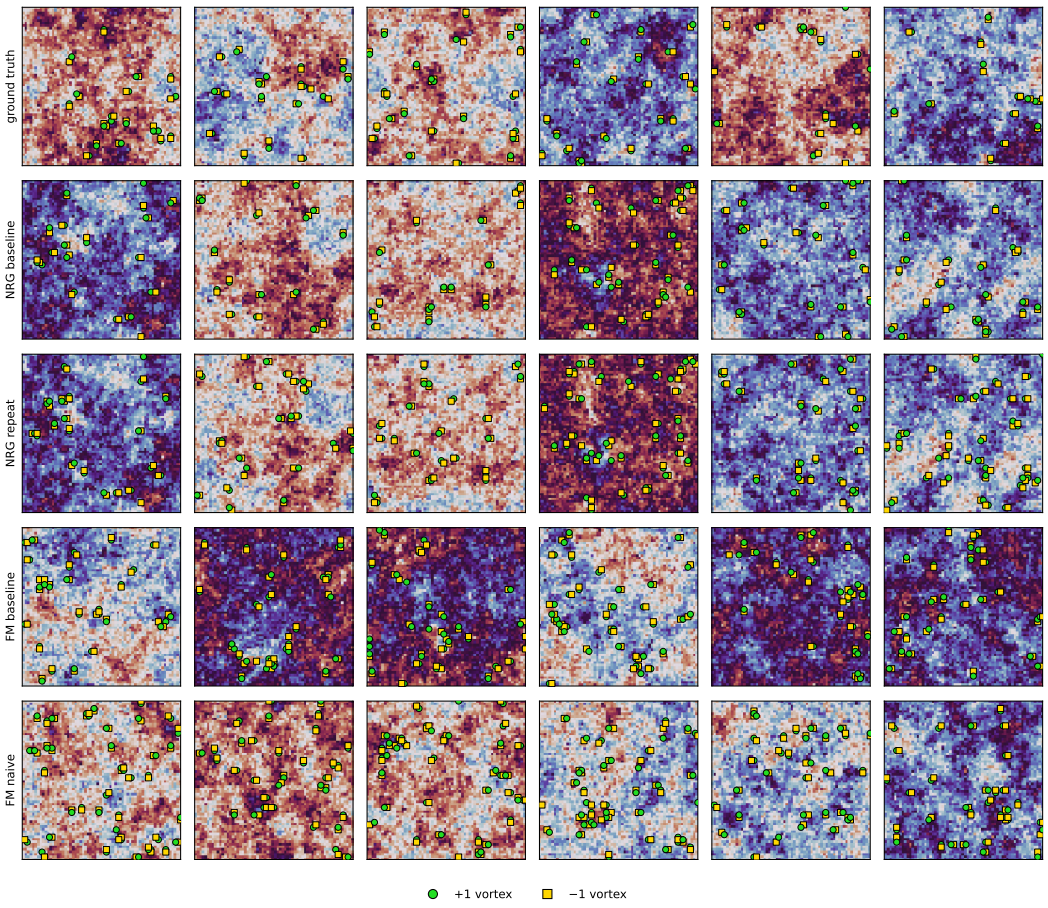


Figure B.6: **XY samples at $L = 64$, β_{BKT}.** Rows: held-out Monte Carlo configurations, NRG baseline, NRG repeat, FM baseline, FM naive. Color is the spin direction, and green circles and yellow squares mark vortices of charge $+1$ and -1. Model rows show the spins drawn from the von Mises conditional that the scores use. The two NRG rows decode the same prior draw in each column. The other rows are independent samples. The flow-matching models over-produce vortices.

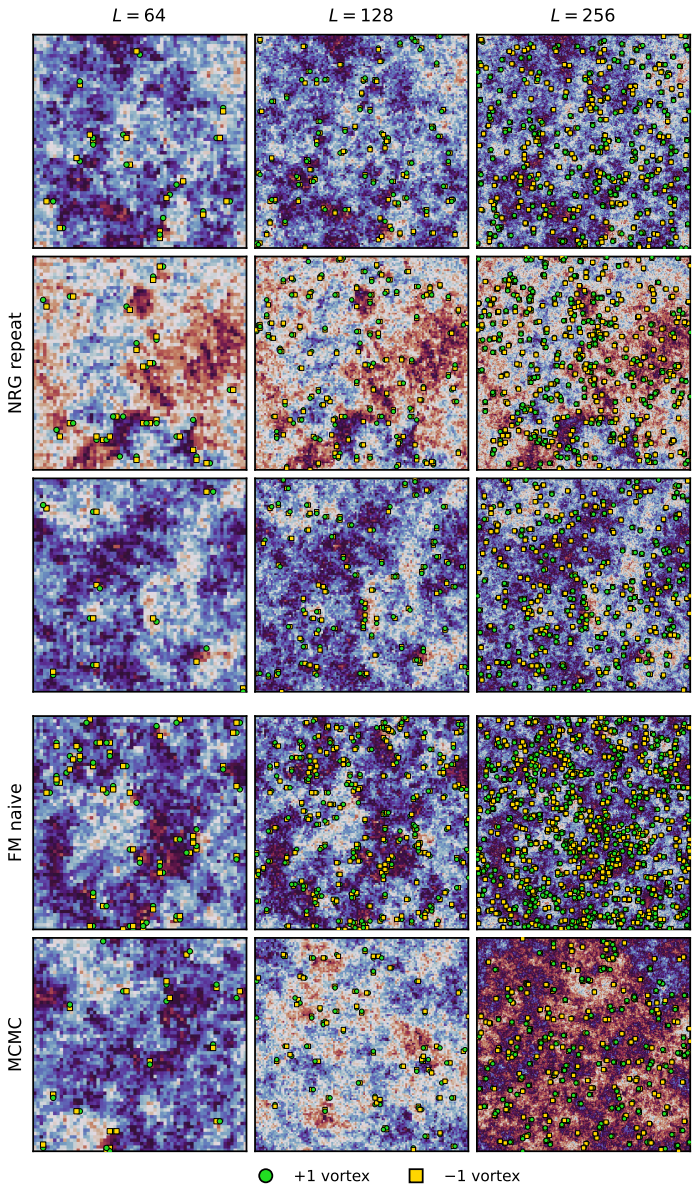


Figure B.7: **XY samples across three octaves, β_{BKT}.** As Figure B.5: three prior draws through the $L = 32$ model with its deepest interior level repeated once, twice and three times, each draw made for the largest lattice and decoded from the top so that every size shares the same coarse structure; one sample of the $L = 32$ flow-matching model coupled across sizes for display only; and at each size the held-out configuration whose $|m|$ is nearest the median. Color is the spin direction, model panels show the spins drawn from the von Mises conditional, and green circles and yellow squares mark vortices of charge $+1$ and -1. The naive flow-matching model produces far more vortices at every size than the grown hierarchy or the Monte Carlo configurations.

as in Section 2 converged to the same multiscale factorization of the critical Ising measure, which admits a reading as a learned nonlinear wavelet transform (Appendix C).

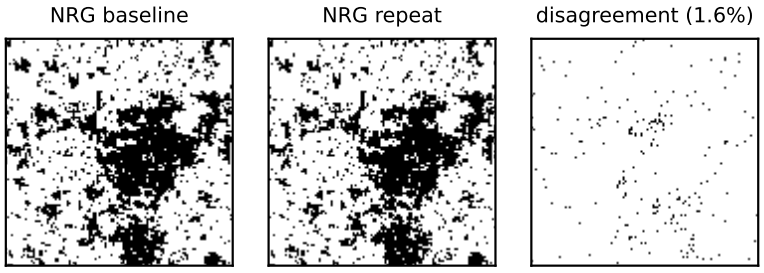


Figure B.8: **Same prior draw through two decoders at $L = 128$.** Spin configurations from the model trained at $L = 128$ (NRG baseline), the model grown from $L = 64$ (NRG repeat), and their disagreement (1.6% of sites for the displayed draw). Over 16 shared draws the baseline and repeat fields have correlation 0.995 and spin agreement 98.1%. Mismatched draws have a mean absolute correlation of 0.022.

B.7 Shared multiscale structure: the ϕ^4 theory against the Ising model

The ϕ^4 runs mirror the Ising ladder (Table B.2): at $L = 128$ the critical point and $\xi \in \{8, 32, 128\}$ on each side of it, and at $L = 64$ the critical point. The off-critical runs are matched to the Ising ladder by nominal ξ, the critical and $\xi = 128$ runs by matching U_4 at $L = 128$. Table B.7 scores the critical models as Table 1 does, and Figure B.9 compares the decimators of the matched runs level by level, in units of the seed-to-seed gap of Figure B.1. The block maps themselves are shown in Figures B.2 and B.3.

Table B.7: **2D ϕ^4 at its critical point** ($\lambda = 1$, $m^2 = m_c^2 = -1.27251$), scored as in Table 1, whose 2D Ising baseline row is repeated for comparison. m is the mean of $\operatorname{sign}\phi$ over sites and $C(r)$ the correlation of $\operatorname{sign}\phi$, averaged over $r \le L/2$. The references are $\langle|m|\rangle = 0.501$ at $L = 128$ and 0.545 at $L = 64$ (floors in Table B.6). Each entry is the mean over two seeds $\pm$ half their range. The $L = 64$ row is scored at $L = 64$.

Model	Training L	Scored L	$\mathrm{TV}_{32}[P(m)]$	$\lvert\langle\lvert m\rvert\rangle - \mathrm{ref}\rvert$	$\overline{\lvert C/C_{\mathrm{ref}} - 1\rvert}$
NRG, 2D Ising (Table 1)	128	128	0.057 ± 0.001	0.0046 ± 0.0010	0.033 ± 0.002
NRG, 2D ϕ^4	128	128	0.038 ± 0.002	0.0040 ± 0.0002	0.025 ± 0.000
NRG, 2D ϕ^4	64	64	0.028 ± 0.000	0.0010 ± 0.0005	0.015 ± 0.000

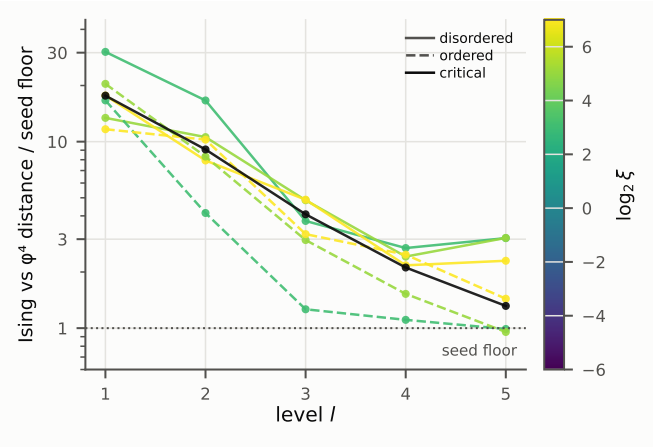


Figure B.9: **Microscopic differences decay under the learned flow.** Root-mean-square fingerprint distance between the Ising decimator and the matched ϕ^4 decimator at each interior level, divided by the seed floor (the mean of the two models' seed-to-seed gaps at that level), for the critical run and the $\xi \in \{8, 32, 128\}$ runs on both sides at $L = 128$. The distance falls from 10 to 30 times the floor at level 1 to a few times it by levels 4 and 5. The ordered runs reach the floor, the critical run reaches 2.1 times it at level 4 and 1.3 at level 5, and the disordered runs at $\xi = 8$ and 32 remain at 2.4 to 3 times it.

C Related Work

The renormalization group and deep learning. The correspondence between the layers of a deep network and RG steps was proposed by Bény (2013) and made explicit for restricted Boltzmann machines by Mehta & Schwab (2014). Lin et al. (2017), Iso et al. (2018), and De Mello Koch et al. (2020) examine how far the analogy holds. A second line constructs the coarse-graining map from an information criterion: the real-space mutual information of Koch-Janusz & Ringel (2018) retains the operators of lowest scaling dimension (Gordon et al., 2021) and, when optimal, keeps the renormalized interactions short-ranged (Lenggenhager et al., 2020), it has been related to a Gaussian information bottleneck (Kline & Palmer, 2022) and used to read off phases and order parameters (Gökmen et al., 2021a;b), and the exact RG itself has been written as an optimal-transport gradient flow (Cotler & Rezchikov, 2023a). These constructions produce interpretable maps rather than samplers of the target ensemble. Our decimators are learned as part of an exact-likelihood generative model and are interpreted afterwards. Learned real-space RG transformations for the Ising model, in the tradition of Monte Carlo RG (Swendsen, 1979), have also been used to extract fixed points and exponents (Chung & Kao, 2021; Hou & You, 2023). Our extremal readout differs in being invariant under the sitewise reparameterization of the hidden fields.

Generative models built on the RG. Neural network RG (Li & Wang, 2018) composes invertible maps into an RG-structured latent space, and RG-Flow (Hu et al., 2022) organizes disentangler and decimator bijectors on 4×4 kernels into a hierarchical flow with a sparse prior. Hu et al. (2020) use the same construction to learn a holographic mapping. Our model is the exact-likelihood, symmetry-exact descendant of this family, and it differs in the questions asked of it: we fix the reparameterization of the block maps and compare them across runs, temperatures, and lattice sizes. The disentangler of our hierarchy plays the role of its namesake in entanglement renormalization (Vidal, 2007; Evenbly & Vidal, 2009), which likewise applies a map across cell boundaries before each coarse-graining step.

Inverse-RG samplers learn the fine-graining direction from configuration pairs and generate lattices larger than any simulated (Efthymiou et al., 2019; Shiina et al., 2021; Bachtis et al., 2022), and Rançon et al. (2026) show that very few parameters suffice to regenerate critical configurations by inverting a majority rule. Diffusion and flow-matching models have been interpreted as inverse RG flows, with noise level in the role of scale (Cotler & Rezchikov, 2023b; Sheshmani et al., 2025), an identification that Bayesian renormalization refines by replacing noise level with an information-geometric scale (Berman et al., 2023) and that also appears in the stochastic-quantization reading of diffusion models (Wang et al., 2024). Masuki & Ashida (2025) use the exact RG flow as the noising process, and Masuki & Ashida (2026) obtain from it a local flow-matching velocity field. yang Tan et al. (2026) apply a shared-weight convolutional diffusion sampler at lattice sizes above those trained on. In these models scale is a continuous time variable of the sampler. In ours it is the level index of the architecture, which is what allows a level to be repeated.

Fixed wavelets, learned conditionals, and lifting schemes. Wavelet conditional RG (Marchand et al., 2023) and its successors (Lempereur & Mallat, 2026; Guth et al., 2022; Kadkhodaie et al., 2023) keep the wavelet transform fixed and learn the conditional distribution of the detail coefficients at each scale given the coarser field, exploiting the fact that these conditionals are local and well conditioned even when the joint density is not (Guth et al., 2023). Wavelet Flow (Yu et al., 2020) is the normalizing-flow version of the same design on a Haar split. Our model is the converse: the Haar split is fixed and orthogonal, but the invertible nonlinearity applied to each cell before the split is learned, so the whole transform is a learned nonlinear wavelet transform in the sense of Li (2021), who identified coupling layers with the lifting steps of Sweldens (1998); Daubechies & Sweldens (1998) and their nonlinear variants (Claypoole et al., 2003). The result of Section 4.1, that independently trained models converge to the same factorization, indicates that for the critical Ising measure this learned wavelet is the same across the trainings we performed once the reparameterization freedom is fixed.

Multiscale flows and extrapolation across sizes. The multiscale factor-out of RealNVP and Glow (Dinh et al., 2017; Kingma & Dhariwal, 2018), which sends half of the variables to the prior at each level, is replaced here by the Haar split, so that the factored variables are the difference coordinates of a cell and the retained one is its average. Flow-based Boltzmann samplers (Noé et al., 2019) and their lattice-field-theory counterparts (Albergo et al., 2019; Kanwar et al., 2020; Nicoli et al.,

2021) face a training cost that grows steeply with the lattice volume (Debbio et al., 2021), and at fixed capacity their sample quality degrades exponentially in the volume (Abbott et al., 2023). Trained weights have been carried to a larger volume as the initialization for further training there, for translation-equivariant continuous flows (Gerdes et al., 2023) and neural-operator flows (Máté & Fleuret, 2024), and by small-lattice pretraining (Abbott et al., 2023), with Máté & Fleuret (2024) reporting that zero-shot transfer to much larger lattices fails. These methods transfer weights across volumes, whereas we transfer a level indexed by scale.

Multilevel samplers organize generation coarse to fine for the Ising model near criticality (Singha et al., 2025), and hierarchical domain decomposition serves the same purpose for autoregressive samplers (Białas et al., 2022). Concurrently with this work, Hasenfratz et al. (2026) grow critical ϕ^4 configurations from $L \leq 32$ to $L = 2048$ by repeatedly inverting a separately tuned linear blocking with a conditional flow followed by rethermalization. Their construction and ours are complementary: they tune the blocking once and learn the conditional, we learn the blocking with an exact likelihood and reuse it without retraining or rethermalization, and their rethermalization could be applied to our samples where the finite-size mismatch of a grown model matters.

On the image side, convolutional diffusion models are known to fail at resolutions above the training resolution because their receptive field is fixed in pixels (He et al., 2024), the same failure as our naive flow-matching model, and function-space formulations (Lim et al., 2025; Hagemann et al., 2025) restore resolution independence at the cost of the exact likelihood.

Regularization and identifiability. Kinetic-energy regularization of neural ODEs (Finlay et al., 2020; Onken et al., 2021) selects the least-action, near-optimal-transport vector field among those with the same endpoint distributions, following the Benamou–Brenier formulation (Benamou & Brenier, 2000). Optimal-transport objectives have likewise been used to constrain learned dynamical models to reproduce the invariant measure of a physical system (Jiang et al., 2023; Melo et al., 2026). Whether independently trained networks learn interchangeable representations has been tested by model stitching (Hernandez et al., 2023). The same-draw agreement of our grown and trained models is a statement of this kind for a generative model, made possible because the canonical normalization removes the remaining coordinate freedom. Bradde & Bialek (2017) relate principal component analysis to the RG. Our principal component analysis is of the learned maps themselves rather than of the data.

Machine learning for the Ising model. Cluster Monte Carlo (Swendsen & Wang, 1987; Wolff, 1989), which largely removes the critical slowing down that afflicts local Metropolis updates (Metropolis et al., 1953), remains the reference sampler and supplies our ground truth. Boltzmann machines were among the first learned models of thermodynamic states (Torlai & Melko, 2016). Learned samplers aim at efficiency near criticality, from variational autoregressive networks (Wu et al., 2019) to locally equivariant samplers (Holderrieth et al., 2025), and learned representations of spin configurations have been used to detect phases (Carrasquilla & Melko, 2017; van Nieuwenburg et al., 2017; Wetzel, 2017), to identify order parameters and hidden spin order with interpretable architectures (Greitemann et al., 2019), and to extract multi-point correlators from experimental snapshots (Miles et al., 2021). Outside spin models, physical parameters have been read off the latent variables of models of spatiotemporal systems (Lu et al., 2020) and phase structure off the weights of trained lattice flows (Göbel et al., 2026).

These two lines have largely diverged: the samplers aimed at efficiency treat the trained model as a black box that reproduces one thermodynamic state, while the interpretable models extract physical content but do not sample. This work sits on the interpretability side and connects the two: the object we interpret is the coarse-graining map of an exact-likelihood sampler, and the payoff of interpreting it is that it can be reused at a lattice size that was never trained on.

D Proofs for Section 3

This appendix proves the results of Section 3. Appendix D.1 fixes the notation and the standing assumptions, and Appendix D.2 collects three lemmas used throughout. Appendices D.3 to D.6 prove Theorems 1 to 3 in turn. Proposition D.15, which records what the two regularizers select, is proved in Appendix D.5, and Appendix D.6 also proves Proposition D.21. Appendix D.7 proves Corollary 4 and discusses its no-tie assumption, Appendix D.8 extends the results to levels with several disentan-

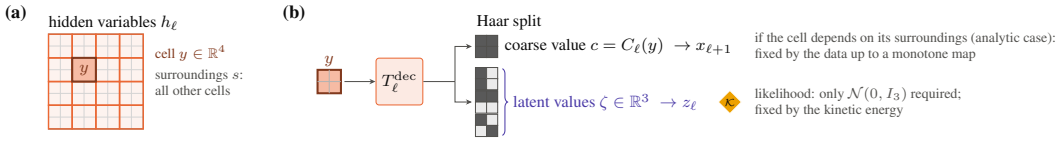


Figure D.1: (a) A cell y of the hidden variables h_ℓ, the input of one decimator block map, and its surroundings s. (b) A decimator keeps one coarse value per cell and discards three latent values. At a level where a cell depends on its surroundings, and for real-analytic p and block maps, the coarse value is fixed by the data up to a monotone map (Theorem 2). The latent values are constrained by the likelihood only through their prior (Theorem 1) and are selected by the kinetic energy under the conditions of Theorem 3.

glers (Proposition D.22) and Appendix D.9 to fields with several components (Proposition D.25). Appendix E summarizes the selection and lists the limitations.

D.1 Notation and setting

A *hierarchy* M is the collection of its block maps, and the density it realizes is the q_θ of the paper. We write $\mathcal{C}$ and $\mathcal{H}$ for classes of hierarchies, C for the coarse-value map of a decimator and c for a coarse value, H for the Haar matrix, $\gamma_m = \mathcal{N}(0, I_m)$ with density φ_m, dropping the subscript when the context fixes it, and $F_\# \mu$ for the image of a distribution μ under a map F.

Layout. The level-ℓ lattice is $\Lambda_\ell = (\mathbb{Z}/L_\ell\mathbb{Z})^2$, $L_\ell = L/2^\ell$, $|\Lambda_\ell| = L_\ell^2$, with fields $V_\ell = \mathbb{R}^{\Lambda_\ell}$, for $0 \le \ell \le n$ and $n = \log_2 L$, Λ_n being a single site. A *cell* is a set of four sites $o + r$, $r \in \{0,1\}^2$, and r is the *position* of the site in it. The *aligned cells* of Λ_ℓ are the cells with corner $o = 2\mathbf{j}$ and the *offset cells* those with corner $o = 2\mathbf{j} + (1,1)$, where $\mathbf{j} \in (\mathbb{Z}/L_{\ell+1}\mathbb{Z})^2$ is the index of the cell, with components j_a. Each family partitions Λ_ℓ and splits the sites into the same four *position classes*, the parity classes of the site coordinates. For $L_\ell \ge 4$ an aligned cell and an offset cell that meet share one site. For either partition $\mathcal{P}$ and a map T of $\mathbb{R}^4$, the *cell layer* $\mathcal{L}_{\mathcal{P},T}$ applies T to the four values of every cell of $\mathcal{P}$, in the order of the positions. A map of V_ℓ is *sitewise* if it applies a map of $\mathbb{R}$ at every site, and $g^\otimes$ applies the same g at all sites.

Levels and the pyramid. Level $\ell \le n-2$ consists of the disentangler $\mathcal{L}_\ell^{\mathrm{dis}}$, the cell layer on the offset cells with block map T_ℓ^{dis}, and the decimator $\mathcal{L}_\ell^{\mathrm{dec}}$, the cell layer on the aligned cells with block map T_ℓ^{dec}; $\mathcal{B}_\ell = \mathcal{L}_\ell^{\mathrm{dec}} \circ \mathcal{L}_\ell^{\mathrm{dis}}$. The lattice Λ_{n-1} is one cell, and level $n-1$ has a decimator only, with $\mathcal{B}_{n-1} = \mathcal{L}_{n-1}^{\mathrm{dec}}$.

The Haar matrix of a cell is the orthogonal matrix

$$H = H_1 \otimes H_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \qquad H_1 = 2^{-1/2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The orthogonal split S_ℓ applies H to every aligned cell and collects the first components into $x_{\ell+1} \in V_{\ell+1}$, the other three into z_ℓ. With $x_0 = x$ and

$$(x_{\ell+1}, z_\ell) = S_\ell \mathcal{B}_\ell(x_\ell),$$

the *pyramid* $(z_0, \dots, z_{n-1}, x_n)$ is a bijective function of x, and (1) is the density of x when the pyramid is standard normal.

Hidden variables. The *fields of hidden variables* are $h_\ell = \mathcal{L}_\ell^{\mathrm{dis}}(x_\ell)$ for $\ell \le n-2$ and x_ℓ for $1 \le \ell \le n-1$. In Haar coordinates the decimator of level ℓ is $D_\ell = H\, T_\ell^{\mathrm{dec}} = (C_\ell, Z_\ell) : \mathbb{R}^4 \to \mathbb{R} \times \mathbb{R}^3$, with *coarse value* C_ℓ, the first Haar coordinate u of its output, in the coordinates (u, w_h, w_v, w_d) of Appendix A.2, and *latent values* Z_ℓ, the other three.

Classes. $\mathcal{C}$ is the set of hierarchies whose block maps are C^1 diffeomorphisms of $\mathbb{R}^4$ onto $\mathbb{R}^4$. The group D_4 acts on the positions of a cell, hence on $\mathbb{R}^4$ by permutation matrices P_σ. $\mathcal{C}_{\mathrm{sym}} \subset \mathcal{C}$ is the set of hierarchies whose block maps commute with every P_σ (are *equivariant*) and are odd.

Assumptions on the data. (D1) p is a positive continuous density on V_0 with finite second moment. (D2) p is invariant under all translations of Λ_0, under the lattice symmetries generated by the

reflections $i_a \mapsto L-1-i_a$ of a coordinate and the transpose, and under $x \mapsto -x$. (D1) holds throughout, (D2) where stated. Theorem 3 assumes four things, to which we refer by name: the *symmetry assumption*, that is, (D2) with hierarchies in $\mathcal{C}_{\mathrm{sym}}$; *analyticity* of p and of the block maps; the *displacement condition*; and the *regularity condition*, the last two defined in Appendix D.5. Analyticity enters through Theorem 2(b): every level either pins the coarse value or has independent cells. The symmetry assumption makes all cells of a layer share one input distribution, so that one shared block map can be optimal for all of them, reduces the units of a hidden variable to one sign per field, and makes flipping a hidden field leave every distribution unchanged, which is why the kinetic energy rather than the likelihood must fix that sign. All hierarchies considered have finite $\mathcal{K}$ and, where the kinetic energy is used, block maps that admit straight flows.

Exactly fitting and normalized. A hierarchy is *exactly fitting* if $q_\theta = p$, that is, if the pyramid of $x \sim p$ is standard normal, and *normalized* if every hidden variable, that is, every site of every field of hidden variables, has marginal $\mathcal{N}(0,1)$. Under (D2) and for hierarchies in $\mathcal{C}_{\mathrm{sym}}$ all sites of such a field have the same marginal (Lemma D.2), so an exactly fitting hierarchy in $\mathcal{C}_{\mathrm{sym}}$ is normalized if and only if the marginal term $\mathcal{M}$ of Section 2 vanishes.

Kinetic energy. A block map T *admits a straight flow* if $b \mapsto (1-t)b + tT(b)$ is a diffeomorphism of $\mathbb{R}^4$ for every $t \in [0,1]$. Its least kinetic energy is then the *displacement* $\mathbb{E}\|T(b)-b\|^2$, and for a map that does not admit a straight flow every flow realizing it still costs at least its displacement (Lemma D.3).

We write $\mathcal{K}_j$ for the least kinetic energy of layer j, averaged over its cells, and $\mathcal{K}$ for the average of the $\mathcal{K}_j$ over the $2n-1$ layers, the normalization of our implementation.

In Appendices D.5 and D.6 $\mathcal{K}_j$ is therefore the displacement of layer j, and every lower bound on a displacement below also bounds the kinetic energy of a competitor whose block maps do not admit straight flows. There $\delta = \|T(b)-b\|$ denotes the L^2 norm of a block map's displacement field, so that its displacement is δ^2. When its argument is a random vector X, $\|X\|$ denotes the L^2 norm $(\mathbb{E}\|X\|^2)^{1/2}$. Inside an expectation $\|\cdot\|$ is the Euclidean norm. By Lemma D.2 every cell of a layer contributes equally when the hierarchy is in $\mathcal{C}_{\mathrm{sym}}$ and (D2) holds.

D.2 Preliminary lemmas

Three lemmas are used throughout. Lemma D.1 records what exact fit says about the hidden variables: the latent variables discarded at a level are standard normal and independent of everything kept, and every field of hidden variables has a positive density, real-analytic when p and the block maps are. Lemma D.2 transfers the symmetries of p to the hidden variables. Its consequence used most often is that all cells of a layer have the same input distribution, invariant under the symmetries of a cell and under the flip, so that one cell stands for the whole layer. Lemma D.3 gives the least kinetic energy of a block map and says when the straight flow attains it: for diffeomorphisms that are gradients of convex functions and for maps whose Jacobian stays uniformly within distance less than one of the identity.

Lemma D.1 (Densities of the hidden variables)**.** *(i) In a hierarchy in $\mathcal{C}$, if x_ℓ has a positive continuous density then so does its image under any C^1 diffeomorphism of V_ℓ, in particular h_ℓ and $S_\ell\mathcal{B}_\ell(x_\ell)$. (ii) If the hierarchy is exactly fitting, then for every $\ell \le n-2$ the latent variables z_ℓ are i.i.d. γ_3 over cells and independent of $x_{\ell+1}$, $(z_0,\dots,z_{\ell-1})$ is independent of x_ℓ, the output of the coarsest decimator is γ_4, and every field of hidden variables has a positive continuous density. (iii) If in addition p and all block maps are real-analytic, every field of hidden variables has a positive real-analytic density.*

Proof. (i) If F is a C^1 diffeomorphism of V_ℓ and x_ℓ has the positive continuous density ρ, then $F(x_\ell)$ has the density $\rho \circ F^{-1}\,|\det \nabla F^{-1}|$, again positive and continuous (Folland, 1999).

(ii) The encoder of levels $\ell+1,\dots,n-1$ maps $x_{\ell+1}$ bijectively to $(z_{\ell+1},\dots,z_{n-1},x_n)$. The pyramid is standard normal, so z_ℓ is i.i.d. γ_3 and independent of $(z_{\ell+1},\dots,x_n)$, hence of $x_{\ell+1}$, and $(z_0,\dots,z_{\ell-1})$ is independent of $(z_\ell,\dots,x_n)$, hence of x_ℓ. As Λ_{n-1} is one cell, $(x_n, z_{n-1}) = H\,T^{\mathrm{dec}}_{n-1}(x_{n-1})$ is four coordinates of the pyramid, hence γ_4.

The densities follow by induction on ℓ. The field x_0 has the positive continuous density p, and if x_ℓ has one, so have h_ℓ and $(x_{\ell+1}, z_\ell) = S_\ell\mathcal{B}_\ell(x_\ell)$ by (i). By the independence just shown, the density

ϖ of the latter factorizes as $\varpi(u,\zeta) = \rho_{\ell+1}(u)\varphi(\zeta)$, so the density $\rho_{\ell+1}$ of $x_{\ell+1}$, with ϖ taken in its continuous version, is $\rho_{\ell+1}(u) = \varpi(u,0)/\varphi(0)$, positive and continuous.

(iii) A real-analytic bijection with everywhere invertible differential has a real-analytic inverse (Krantz & Parks, 2002), and $\det \nabla F^{-1}$ is real-analytic and nowhere zero, hence of constant sign, so its absolute value is real-analytic, and the induction in the proof of (ii) gives real-analytic densities. The induction avoids marginal densities, which need not be real-analytic when the joint density is: here the joint density factorizes, so the density of $x_{\ell+1}$ is its restriction to $\zeta = 0$, rescaled, and restrictions of real-analytic functions are real-analytic (Krantz & Parks, 2002). □

Lemma D.2 (Symmetry of the hidden variables)**.** *Assume (D2) and let the hierarchy be in $\mathcal{C}_{\mathrm{sym}}$. Then the distribution of x_ℓ is invariant under all translations and lattice symmetries of Λ_ℓ and under the flip, and the distribution of h_ℓ is invariant under translations by two sites, the lattice symmetries and the flip. Consequently the four position classes of a field of hidden variables have the same one-site distribution, which is symmetric; all cells of a layer have the same input distribution; and that distribution is invariant under every P_σ and under the flip. The statements about translations hold for every hierarchy in $\mathcal{C}$ when p is translation invariant.*

Proof. The flip, the lattice symmetries and translations by two sites map aligned cells to aligned cells and offset cells to offset cells, permuting the positions inside a cell. Equivariance and oddness let them pass through every cell layer, and the permutation-invariant first row of H carries them to the next level, where a translation by two sites becomes a unit translation, so induction over the levels gives the invariances. Since L_ℓ is even, the reflection $i_a \mapsto L_\ell - 1 - i_a$ sends the aligned cell with corner $2j_a$ along that axis to the aligned cell with corner $L_\ell - 2 - 2j_a$ and the offset cell with corner $2j_a + 1$ to the offset cell with corner $L_\ell - 3 - 2j_a$, and inside each cell it exchanges the two positions along the reflected axis. The transpose maps aligned cells to aligned cells and offset cells to offset cells and exchanges positions $(0,1)$ and $(1,0)$. A translation by two sites maps cells to cells of the same partition and fixes positions.

For such a map g with position permutation σ,

$$g \circ \mathcal{L}_{\mathcal{P},T} = \mathcal{L}_{\mathcal{P},P_\sigma T P_\sigma^{-1}} \circ g = \mathcal{L}_{\mathcal{P},T} \circ g, \tag{D.1}$$

because T commutes with P_σ. For translations σ is the identity and no symmetry of T is needed, and the flip commutes with the layer because T is odd.

The first row of H is invariant under permutations of the cell's four values, so $x_{\ell+1}(g\,v) = g'\,x_{\ell+1}(v)$ with g' the same type of symmetry on $\Lambda_{\ell+1}$ (the flip for the flip), and a translation of Λ_ℓ by two sites becomes a unit translation of $\Lambda_{\ell+1}$. Induction on ℓ from (D2) gives the invariances. Unit translations generate all translations of $\Lambda_{\ell+1}$.

A reflection changes the parity of the reflected coordinate, so the lattice symmetries act transitively on the four position classes; the cells of one partition are related by translations by two sites; and composing a lattice symmetry with such a translation gives a symmetry that maps a given cell to itself and realizes σ on its positions. This gives the three consequences. □

Lemma D.3 (Displacement and kinetic energy)**.** *Let a block map T be the time-one map of a flow $\dot b_t = v(b_t,t)$ and let $b_0 \sim \mu$. Then $\mathbb{E}\int_0^1 \|v(b_t,t)\|^2 \mathrm{d}t \geq \mathbb{E}\|T(b_0) - b_0\|^2$, with equality if and only if almost every trajectory is a straight line traversed at constant speed. A flow with equality exists if $X_t = (1-t)\,\mathrm{id} + tT$ is a diffeomorphism of $\mathbb{R}^4$ for every $t \in [0,1]$. This holds in either of two cases: T is a diffeomorphism and the gradient of a convex function; or $\sup_b \|\nabla T(b) - I\| \leq \lambda < 1$ in operator norm.*

Proof. By the Cauchy–Schwarz inequality $\|\int_0^1 \dot b_t\,\mathrm{d}t\|^2 \leq \int_0^1 \|\dot b_t\|^2 \mathrm{d}t$, with equality if and only if $\dot b_t$ is constant.

If every X_t is a diffeomorphism, the field $v(\cdot,t) = (T - \mathrm{id}) \circ X_t^{-1}$ generates the flow X_t, whose trajectories are straight.

If $T = \nabla\psi$ with ψ convex, then for $t < 1$ the map X_t is the gradient of the $(1-t)$-strongly convex function $(1-t)\|b\|^2/2 + t\psi$, whose Hessian $(1-t)I + t\nabla T$ is at least $(1-t)I$ in the order on symmetric matrices; ψ is C^2 because $T = \nabla\psi$ is C^1. It is injective by strict convexity and onto because $b \mapsto (1-t)\|b\|^2/2 + t\psi(b) - q \cdot b$ is strongly convex for each $q \in \mathbb{R}^4$, hence attains a

minimum, at which the gradient of $(1-t)\|b\|^2/2 + t\psi$ equals q; and $\nabla X_t \succeq (1-t)I$ is invertible. So X_t is a diffeomorphism, and $X_1 = T$.

If $\sup_b \|\nabla T(b) - I\| \le \lambda < 1$ then $X_t - \mathrm{id} = t(T - \mathrm{id})$ is λ-Lipschitz, so $X_t(b) = q$ exactly when b is a fixed point of $b \mapsto q - t(T(b) - b)$, a contraction with constant at most λ, which has exactly one fixed point (Rudin, 1976). So X_t is a bijection, and $\nabla X_t = I + t(\nabla T - I)$ is invertible. In both cases the straight flow exists and attains the bound. □

D.3 Reparameterization freedom

Theorem 1 asks which maps inserted after a layer, and undone in the next layer if there is one, keep the architecture and leave the likelihood unchanged. At a disentangler output the decimator undoes the inserted map completely, so the likelihood cannot see it. At a decimator output the next level can undo only the action on the coarse values, while the latent values go to the prior, and this is where the likelihood constrains the inserted map. The theorem concerns changes that preserve the likelihood whatever the other block maps are. A particular hierarchy can admit more, which Theorem 3 deals with. Definition D.4 makes the two requirements precise. At a disentangler output the likelihood imposes nothing, because the inserted map and its inverse cancel (Lemma D.5), and the requirement that both modified layers remain cell layers forces the map to act site by site (Lemma D.6). At a decimator output the likelihood does constrain the map, through the identity (D.2), which must hold for every density that the later levels can assign to the coarse variables. Those densities are rich enough to tell apart any two fields that are not permutations of each other (Lemma D.8), and this forces the coarse part of the inserted map to depend on the coarse variables alone and its latent part to preserve the prior (Lemma D.9). The proof of the theorem assembles these pieces.

Definition D.4 (Reparameterizations)**.** *(a)* At a disentangler output h_ℓ, a reparameterization is a C^1 diffeomorphism g of V_ℓ onto itself. It replaces $\mathcal{L}_\ell^{\mathrm{dis}}$ by $g \circ \mathcal{L}_\ell^{\mathrm{dis}}$ and $\mathcal{L}_\ell^{\mathrm{dec}}$ by $\mathcal{L}_\ell^{\mathrm{dec}} \circ g^{-1}$. *(b)* At the output of the decimator of level $\ell \le n-2$, it is a pair (Ξ, G) of C^1 diffeomorphisms, Ξ of $V_{\ell+1} \times \mathbb{R}^{3|\Lambda_{\ell+1}|}$ onto itself and G of $V_{\ell+1}$ onto itself. It replaces $\mathcal{L}_\ell^{\mathrm{dec}}$ by $S_\ell^\top \Xi S_\ell \circ \mathcal{L}_\ell^{\mathrm{dec}}$ and the first layer $\mathcal{L}$ of level $\ell+1$ by $\mathcal{L} \circ G^{-1}$. *(c)* At the output of the coarsest decimator it is a C^1 diffeomorphism Ξ of $\mathbb{R}^4$ onto itself applied to that output.

A reparameterization is *admissible* if, for every hierarchy in $\mathcal{C}$, the modified layers are cell layers on the same partitions with block maps that are C^1 diffeomorphisms, and *likelihood-preserving* if $\log q_{\theta'}(x) = \log q_\theta(x)$ for every x and every hierarchy in $\mathcal{C}$, where θ' is the modified hierarchy.

A reparameterization of the coarse variables by g is the case $\Xi = g \times \mathrm{id}$, $G = g$ of (b). We let the map G that the next level undoes differ a priori from the coarse part of Ξ. Likelihood preservation and admissibility together force them to agree. Because both requirements must hold for every hierarchy in $\mathcal{C}$, the proofs may test them on convenient hierarchies, such as the one with identity block maps.

D.3.1 Disentangler outputs

Lemma D.5 (Cancellation)**.** *Every reparameterization of type (a), and every reparameterization of type (b) with $\Xi = g \times \mathrm{id}$ and $G = g$, is likelihood-preserving.*

Proof. The composition of the two modified layers equals that of the original ones, so the pyramid is unchanged as a function of x, and the log-determinants of the modified layers add $\log|\det \nabla g| - \log|\det \nabla g| = 0$ at corresponding points. □

Admissibility alone determines the form of the map at a disentangler output: for $L_\ell \ge 4$ an offset cell and an aligned cell that meet share one site, so a map that is a cell layer on both partitions can act on that site only through its own value.

Lemma D.6 (Admissible maps between two layers are sitewise)**.** *Let $L_\ell \ge 4$. A C^1 diffeomorphism g of V_ℓ is an admissible reparameterization of h_ℓ if and only if it is sitewise with one C^1 diffeomorphism of $\mathbb{R}$ per position class.*

Proof. Admissibility for the hierarchy with identity block maps shows that g is a cell layer on the offset cells, and that g^{-1} is a cell layer on the aligned cells with an invertible block map, hence so is

g: with K_{b} and K_{c} the block maps of g on the offset cells and on the aligned cells, $g(v)|_\beta = K_{\mathrm{b}}(v|_\beta)$ for every offset cell β and $g(v)|_\kappa = K_{\mathrm{c}}(v|_\kappa)$ for every aligned cell κ.

Let i be a site with offset cell β and aligned cell κ, so that $\beta \cap \kappa = \{i\}$. If v and v' agree at i, choose v'' equal to v on β and to v' on κ. Then $g(v)_i = K_{\mathrm{b}}(v''|_\beta)_i = g(v'')_i = K_{\mathrm{c}}(v''|_\kappa)_i = g(v')_i$. So $g(v)_i$ is a function g_i of v_i alone, and since K_{b} is the same for all offset cells, g_i depends only on the position class of i. As g is the product of the g_i and is a C^1 diffeomorphism, so is every g_i.

Conversely, a sitewise map with one map per position class is a cell layer on both partitions, and compositions of cell layers on one partition are cell layers on it, so it is admissible. □

D.3.2 Decimator outputs

Two facts about the standard normal distribution are used repeatedly: the density identity that characterizes the maps preserving it, and that the only monotone homeomorphisms of $\mathbb{R}$ preserving γ_1 are $\pm\mathrm{id}$.

Lemma D.7 (Maps that preserve the standard normal)**.** *(a) A C^1 diffeomorphism R of $\mathbb{R}^m$ satisfies $\varphi_m(R(\zeta))|\det \nabla R(\zeta)| = \varphi_m(\zeta)$ for all ζ if and only if R preserves γ_m. (b) A monotone homeomorphism of $\mathbb{R}$ that preserves γ_1 is* id *or* $-\mathrm{id}$.

Proof. (a) The image of γ_m under R has the density $\varphi_m(R^{-1}(\zeta'))|\det \nabla R^{-1}(\zeta')|$, which equals $\varphi_m(\zeta')$ for all ζ' if and only if, with $\zeta' = R(\zeta)$, the stated identity holds.

(b) Let Φ be the standard normal distribution function. If g is increasing, the distribution function of the image is $\Phi \circ g^{-1} = \Phi$, so $g = \mathrm{id}$. If g is decreasing it is $1 - \Phi \circ g^{-1} = \Phi$, so $g^{-1} = -\mathrm{id}$. □

The likelihood at a decimator output depends on the density that the later levels assign to the coarse variables. The next lemma shows that this family of densities is large: it contains the product over sites of any one positive smooth density and the product over aligned cells of any one centred normal density, so that ratios of such densities distinguish two fields unless one is a permutation of the other, and a map that preserves all of them and is a cell layer on the offset cells is the identity. Part (b) is used in Lemma D.9 to show that the coarse part of an inserted map cannot depend on the latent values, and part (c) in the proof of Theorem 1 to show that the map undone by the next level equals the coarse part of the inserted map.

Fix $\ell \le n-2$. A density ρ on $V_{\ell+1}$ is *realizable* if some choice of levels $\ell+1, \dots, n-1$ in $\mathcal{C}$, with encoder E, assigns it to $x_{\ell+1}$, that is, $\rho(v) = \varphi(E(v))|\det \nabla E(v)|$.

Lemma D.8 (Realizable densities separate points)**.** *(a) For every positive smooth density ρ_1 on $\mathbb{R}$ the product $\prod_i \rho_1(v_i)$ is realizable. If $\ell+1 \le n-2$, for every positive definite 4×4 matrix Σ the product over the aligned cells of $\Lambda_{\ell+1}$ of centred normal densities with covariance Σ is realizable. If $\ell+1 = n-1$, every centred normal density on $\mathbb{R}^4$ is realizable. (b) If the multisets of coordinates of v and v' differ, there are realizable $\rho, \tilde\rho$ with $\rho(v)/\rho(v') \ne \tilde\rho(v)/\tilde\rho(v')$. (c) Let Ψ be a C^1 diffeomorphism of $V_{\ell+1}$ that preserves every realizable density and is a cell layer on the offset cells of $\Lambda_{\ell+1}$ if $\ell+1 \le n-2$. Then $\Psi = \mathrm{id}$.*

Proof. (a) Let F be the distribution function of ρ_1 and $g = \Phi^{-1} \circ F$, a smooth increasing diffeomorphism of $\mathbb{R}$ onto $\mathbb{R}$ with $\varphi_1(g(t))g'(t) = \rho_1(t)$. Give the first layer of level $\ell+1$ the block map $g \times g \times g \times g$ and all other layers the identity. The encoder is then an orthogonal map applied to $g^{\otimes}(v)$, so the assigned density is $\prod_i \varphi_1(g(v_i))g'(v_i) = \prod_i \rho_1(v_i)$. For the normal densities give the decimator of level $\ell+1$ the linear block map $\Sigma^{-1/2}$ and all other layers the identity.

(b) Some a occurs m times among the coordinates of v and $m' \ne m$ times among those of v'. For $\sigma > 0$ let $\eta_\sigma(t) = \exp(-(t-a)^2/\sigma^2)$ and $\tilde\rho_1 = \rho_1(1+\eta_\sigma)/c_\sigma$ with c_σ the normalizing constant. The constants cancel in the ratio, and $\tilde\rho(v)/\tilde\rho(v')$ equals $\rho(v)/\rho(v')$ times $\prod_i(1+\eta_\sigma(v_i))/\prod_i(1+\eta_\sigma(v'_i))$, which tends to $2^{m-m'} \ne 1$ as $\sigma \to 0$.

(c) We test Ψ against two families of realizable densities: product densities with a narrow bump show that its block map permutes the four coordinates, and cell-wise Gaussians on the aligned cells with distinct correlations show that the permutation fixes every position. Suppose first $\ell+1 \le n-2$, so

that $L_{\ell+1} \geq 4$, and let e be the block map of Ψ. Since Ψ preserves $\prod_i \rho_1(v_i)$, the ratio

$$A(b) = \prod_{j=1}^{4} \rho_1(e_j(b)) \, |\det \nabla e(b)| \Big/ \prod_j \rho_1(b_j)$$

satisfies $\prod_\beta A(v|_\beta) = 1$ for all v. Taking v with the same values b on every offset cell gives $A(b)^{n_\mathrm{b}} = 1$, with n_b the number of offset cells, and $A(b) > 0$, so $A \equiv 1$.

Dividing the identity $A \equiv 1$ for the density $\rho_1(1 + \vartheta\eta_\sigma)/c_\vartheta$, with $\vartheta \in (0, 1]$ and c_ϑ its normalizing constant, by the one for ρ_1 gives

$$\prod_j \big(1 + \vartheta\eta_\sigma(b_j)\big) = \prod_j \big(1 + \vartheta\eta_\sigma(e_j(b))\big) \qquad \text{for all } \vartheta \in (0, 1].$$

Both sides are polynomials in ϑ, so the coefficients of ϑ agree: $\sum_j \eta_\sigma(b_j) = \sum_j \eta_\sigma(e_j(b))$ for all a and σ. Letting $\sigma \to 0$ with a fixed shows that every a occurs equally often among the b_j and among the $e_j(b)$.

Hence on the open dense set where the b_j are distinct, $e(b)$ is a permutation of the coordinates of b. By continuity the permutation is constant on each connected component of that set, so ∇e is a permutation matrix there.

The complement is the union of the hyperplanes $\{b_j = b_{j'}\}$, which we call walls. Two components that share a piece of a wall carry the same permutation matrix, because ∇e is continuous across that piece and takes the constant value of each component on either side of it. Any two components are joined by a chain in which consecutive members share a wall, since crossing one wall exchanges two coordinates. So $\nabla e = P_\pi$ for one permutation π of the positions, and $e = P_\pi$ with no additive constant, since $e(b)$ is a permutation of the coordinates of b on a dense set. Thus Ψ permutes sites.

Let Σ_0 be symmetric with zero diagonal and distinct nonzero off-diagonal entries, and

$$\Sigma = I + \vartheta\Sigma_0,$$

positive definite for small $\vartheta > 0$.

Under the product over aligned cells of $\mathcal{N}(0, \Sigma)$, two sites in one aligned cell have a covariance that determines their unordered pair of positions in the cell, and two sites in different aligned cells have covariance 0. So Ψ preserves this distribution only if it maps every aligned cell onto an aligned cell and preserves the unordered pair of positions of any two sites of an aligned cell, hence the position of every site in its aligned cell, which is the parity class of the site; $L_{\ell+1}$ is even, so reduction modulo $L_{\ell+1}$ does not change parities.

The site at position r of an offset cell is moved to position $\pi(r)$ of the same offset cell, and since $r, \pi(r) \in \{0, 1\}^2$ it keeps its parity class only if $\pi(r) = r$. So $\pi = \mathrm{id}$ and $\Psi = \mathrm{id}$.

If $\ell + 1 = n - 1$ the same argument shows that Ψ is a coordinate permutation of $\mathbb{R}^4$, and it preserves $\mathcal{N}(0, \Sigma)$ for this Σ only if it is the identity. □

At a decimator output, with c the coarse variables and $\zeta \in \mathbb{R}^m$ the latent variables of the whole level, a likelihood-preserving map Ξ has to satisfy the identity (D.2) below for every realizable ρ. The next lemma solves it: the level sets of the coarse part of Ξ turn out to be the fibres $\{c\} \times \mathbb{R}^m$, so the coarse part is a function of c alone, and the latent part then has to preserve γ_m for every c.

Lemma D.9 (Form of a likelihood-preserving map at a decimator output)**.** *Let $\Xi = (\Xi_c, \Xi_\zeta)$ be a C^1 diffeomorphism of $V_{\ell+1} \times \mathbb{R}^m$ onto itself such that for every realizable ρ there is a density ρ' with*

$$\varphi_m(\Xi_\zeta(c, \zeta)) \, \rho'(\Xi_c(c, \zeta)) \, |\det \nabla\Xi(c, \zeta)| = \varphi_m(\zeta) \, \rho(c) \qquad \textit{for all } (c, \zeta). \tag{D.2}$$

Then $\Xi(c, \zeta) = (\tilde{g}(c), R(c, \zeta))$ where $\tilde{g}$ is a C^1 diffeomorphism of $V_{\ell+1}$, $R(c, \cdot)$ is a C^1 diffeomorphism of $\mathbb{R}^m$ that preserves γ_m for every c, and $\rho' = \tilde{g}_\# \rho$. Conversely every such Ξ satisfies (D.2) *for every ρ.*

Proof. For $\hat{c} \in V_{\ell+1}$ the level set $\Gamma_{\hat{c}} = \{\Xi_c = \hat{c}\} = \Xi^{-1}(\{\hat{c}\} \times \mathbb{R}^m)$ is connected and closed. Let $(c, \zeta), (c', \zeta') \in \Gamma_{\hat{c}}$. Dividing (D.2) at the two points cancels $\rho'(\hat{c})$ and shows that $\rho(c)/\rho(c')$ is the same for all realizable ρ, so c' is a coordinate permutation of c by Lemma D.8(b). Thus $\Gamma_{\hat{c}}$ lies in the

union of the finitely many disjoint closed sets $\{c'\} \times \mathbb{R}^m$, c' a permutation of c, and being connected it lies in one of them, $\{c\} \times \mathbb{R}^m$.

Since $\nabla\Xi$ is invertible, the differential of Ξ_c is onto at every point, so $\Gamma_{\hat{c}}$ is a C^1 submanifold of dimension m. Such a submanifold contained in the m-dimensional affine space $\{c\} \times \mathbb{R}^m$ is an open subset of it, being near each of its points the image of an injective C^1 map of $\mathbb{R}^m$ with invertible differential, which is open by the inverse function theorem (Rudin, 1976). It is also closed, and $\{c\} \times \mathbb{R}^m$ is connected, so $\Gamma_{\hat{c}} = \{c\} \times \mathbb{R}^m$.

Every point lies in some $\Gamma_{\hat{c}}$, so $\Xi_c(c, \zeta) = \tilde{g}(c)$. The map $\tilde{g}$ is C^1, injective because distinct sets $\{c\} \times \mathbb{R}^m$ are distinct level sets, surjective because Ξ is, and $\nabla\tilde{g}$ is invertible because $\nabla\Xi_c = [\nabla\tilde{g} \;\; 0]$ has full rank.

Write $R = \Xi_\zeta$. For fixed c, $R(c, \cdot)$ is injective, $\nabla_\zeta R$ is invertible because $\nabla\Xi$ is block triangular, and $R(c, \cdot)$ is onto: a point $(\tilde{g}(c), \zeta')$ has a preimage (c'', ζ) with $\tilde{g}(c'') = \tilde{g}(c)$, so $c'' = c$.

With $\Theta(c, \zeta) = \varphi_m(R(c, \zeta))|\det \nabla_\zeta R(c, \zeta)|/\varphi_m(\zeta)$, Equation (D.2) reads $\Theta(c, \zeta) = \rho(c)/(\rho'(\tilde{g}(c))|\det \nabla\tilde{g}(c)|)$, which does not depend on ζ. Integrating $\Theta(c, \zeta)\varphi_m(\zeta)$ over ζ with the substitution $\zeta' = R(c, \zeta)$ gives 1, so $\Theta \equiv 1$. Hence $R(c, \cdot)$ preserves γ_m by Lemma D.7(a), and $\rho'(\tilde{g}(c))|\det \nabla\tilde{g}(c)| = \rho(c)$, that is, $\rho' = \tilde{g}_\# \rho$. The converse is the same computation read backwards. □

D.3.3 Proof of Theorem 1

With these definitions Theorem 1 takes the following form.

Theorem D.10 (Theorem 1, precise form)**.** *A reparameterization is admissible and likelihood-preserving if and only if it is* (a) *at a disentangler output, a sitewise map with one monotone C^1 diffeomorphism of $\mathbb{R}$ per position class;* (b) *at the output of a decimator of level $\ell \leq n - 2$, a pair (Ξ, G) with $G = g^{\otimes}$ for a monotone C^1 diffeomorphism g of $\mathbb{R}$ and $\Xi(c, \zeta) = (g^{\otimes}(c), R(c, \zeta))$, where R acts cell by cell through one map $R_1(c_\kappa, \zeta_\kappa)$ and $R_1(c, \cdot)$ preserves γ_3 for every $c \in \mathbb{R}$, which we call a* latent reparameterization*; or* (c) *at the coarsest output, any Ξ that preserves γ_4.*

Proof of Theorem 1. At a disentangler output. By Lemma D.5 every reparameterization of type (a) preserves the likelihood, and it is admissible if and only if it is sitewise with one diffeomorphism of $\mathbb{R}$, which is monotone, per position class (Lemma D.6, which applies because h_ℓ exists only for $\ell \leq n - 2$, where $L_\ell \geq 4$). This is case (a).

At a decimator output, $\ell \leq n - 2$. Since the reparameterization must preserve the likelihood of every hierarchy, we test it on convenient ones: identity block maps up to level ℓ let (c, ζ) range over the whole space, and arbitrary later levels let ρ range over all realizable densities. Lemma D.9 then gives the form of Ξ, admissibility makes it act cell by cell with one map, and Lemma D.8(c) identifies G. Let ρ be the density that the unmodified levels $\ell + 1, \ldots, n - 1$, with encoder E, assign to $x_{\ell+1}$. The modified levels have encoder $E \circ G^{-1}$ and assign to the new coarse variables the density $\rho' = G_\# \rho$. With $m = 3|\Lambda_{\ell+1}|$, the log-density of the modified hierarchy is

$$\log q_{\theta'}(x) = \log q_\theta(x) + \log \frac{\varphi_m(\Xi_\zeta(c, \zeta))\, \rho'(\Xi_c(c, \zeta))\, |\det \nabla\Xi(c, \zeta)|}{\varphi_m(\zeta)\, \rho(c)}, \qquad (c, \zeta) = S_\ell \mathcal{B}_\ell(x_\ell),$$

the logarithm of the ratio of the two sides of (D.2).

With identity block maps at levels $0, \ldots, \ell$ the map $x \mapsto x_\ell$ is a composition of orthogonal splits and coordinate selections, hence linear and onto, and S_ℓ is orthogonal, so the point (c, ζ) takes every value in $V_{\ell+1} \times \mathbb{R}^m$ and an identity of log-densities for all x is one for all (c, ζ).

As the levels from $\ell + 1$ on range over $\mathcal{C}$, ρ ranges over all realizable densities. So (Ξ, G) is likelihood-preserving if and only if (D.2) holds with $\rho' = G_\# \rho$ for every realizable ρ.

By Lemma D.9, $\Xi = (\tilde{g}, R)$ and $G_\# \rho = \tilde{g}_\# \rho$, so $\Psi = \tilde{g}^{-1} \circ G$ preserves every realizable ρ. Admissibility of Ξ for the identity hierarchy means that $S_\ell^\top \Xi S_\ell$ is a cell layer on the aligned cells, so Ξ acts cell by cell with one map: $\tilde{g} = g^{\otimes}$ for a diffeomorphism g of $\mathbb{R}$, and $R(c, \zeta)|_\kappa = R_1(c_\kappa, \zeta_\kappa)$. For fixed c the map $R(c, \cdot)$ is the product of the $R_1(c_\kappa, \cdot)$ and preserves γ_m, so, reading off the marginal distribution of cell κ, $R_1(c_\kappa, \cdot)$ preserves γ_3. Every real number occurs as an entry c_κ, so $R_1(c, \cdot)$ preserves γ_3 for every $c \in \mathbb{R}$.

Admissibility of G means that G is a cell layer on the offset cells of $\Lambda_{\ell+1}$ when $\ell+1 \le n-2$, and imposes nothing when $\ell+1 = n-1$. In both cases $\Psi = (g^{\otimes})^{-1} \circ G$ satisfies the hypotheses of Lemma D.8(c), so $G = g^{\otimes}$. Conversely such pairs are admissible and satisfy (D.2) by the converse part of Lemma D.9. This is case (b).

At the coarsest output. There is no later layer, so the log-density changes by $\log \varphi_4(\Xi(t)) + \log|\det \nabla \Xi(t)| - \log \varphi_4(t)$ at the output t, which takes every value in $\mathbb{R}^4$. By Lemma D.7(a) this vanishes identically if and only if Ξ preserves γ_4, and every Ξ is admissible since Λ_{n-1} is one cell. This is case (c). □

Remark D.11 (Size of the latent family). For any C^1 map $O : \mathbb{R} \to SO(3)$, $R_1(c, \zeta) = O(c)\zeta$ is a latent reparameterization, so that family is infinite-dimensional. Block maps realized by flows have positive Jacobian determinant, so an even number of the four monotone maps at a disentangler output must be decreasing, and a decreasing map of the coarse variables must come with a latent reparameterization of negative determinant. In $\mathcal{C}_{\mathrm{sym}}$ the four maps coincide and are odd, so orientation imposes nothing there. The normalization is what reduces them to a sign (Proposition D.15(a)).

D.4 Sufficiency of the coarse value

Fix a level of an exactly fitting hierarchy. Two values of a cell are called equivalent if the surroundings cannot tell them apart. Theorem 2 rests on Lemma D.13: when the decimator fits the level, two cell values with the same coarse value are equivalent, because the latent values are independent of everything else. Part (a) follows at once, and part (b) from the separating–independent dichotomy (Lemma D.14), where analyticity enters, together with the identifiability argument of Appendix D.4.3.

D.4.1 Definitions and the precise statement

Fix a level $\ell \le n-2$ of an exactly fitting hierarchy and let p_h be the positive continuous density of $h = h_\ell$ (Lemma D.1). For a cell κ write $y = h|_\kappa \in \mathbb{R}^4$ and $s = h|_{\Lambda_\ell \setminus \kappa}$ for its *surroundings*, so that $p_h = p_h(y, s)$.

Two cell values are *equivalent* if

$$y \sim y' \quad \Longleftrightarrow \quad s \mapsto p_h(y, s)/p_h(y', s) \text{ is constant.}$$

This is an equivalence relation, and for almost every pair it says that the surroundings have the same conditional distribution given y as given y'.

All cell values are equivalent if and only if $p_h(y, s) = \lambda(y)\, p_h(y_0, s)$, in which case both factors are integrable and the cell is *independent of its surroundings*. The level is *separating* at the cell κ if no nonempty open subset of $\mathbb{R}^4$ consists of mutually equivalent values, that is, if every nonempty open set of cell values contains two values that the surroundings tell apart, and *separating* if this holds for some cell. If p is translation invariant it then holds for every cell (Lemma D.2). Open sets are the right notion because a coarse value that varied along a level set of another would sweep out an open set of equivalent values (Appendix D.4.3).

A decimator $D = (C, Z)$, a C^1 diffeomorphism of $\mathbb{R}^4$ onto $\mathbb{R} \times \mathbb{R}^3$, *fits the level* if $(Z(h|_\kappa))_\kappa$ is i.i.d. γ_3 and independent of the whole family $(C(h|_\kappa))_\kappa$. By Lemma D.1 the decimator of an exactly fitting hierarchy fits its level.

Theorem D.12 (Theorem 2, precise form)**.** *With this notation:* (a′) *if D fits the level, a cell and its surroundings are independent given $C(y)$;* (b′) *if p is translation invariant and p_h is, up to positive factors depending on the cell alone and on the surroundings alone, a real-analytic function of homeomorphic images of the cell and of the surroundings (the hypothesis of Lemma D.14, which holds under analyticity), every level is either separating or has mutually independent cells;* (c′) *if $D = (C, Z)$ and $D' = (C', Z')$ both fit a separating level, then $C' = \phi \circ C$ for a monotone C^1 diffeomorphism ϕ of $\mathbb{R}$, and $D' \circ D^{-1}(c, \zeta) = (\phi(c), R(c, \zeta))$ with $R(c, \cdot)$ preserving γ_3 for every c.*

Part (a′) is proved in Appendix D.4.2 as a consequence of Lemma D.13, part (b′) there after Lemma D.14, and part (c′) in Appendix D.4.3. Part (b) of Theorem 2 follows from Lemma D.14 and (c′): under analyticity the hypothesis of the lemma holds at every level (proof of (b′)), so a level with a cell that is not independent of its surroundings is separating, and (c′) applies. Translation

invariance is not needed. The level set aside after Theorem 2 in Section 3, where no cell depends on its surroundings, has mutually independent cells by the last paragraph of the proof of (b′).

D.4.2 SUFFICIENCY AND THE DICHOTOMY

Lemma D.13 (Sufficiency). *If D fits the level, then $C(y) = C(y')$ implies $y \sim y'$.*

Proof. D is a bijection on every cell, so s is a bijective function of the coarse and latent values of the cells other than κ. Since D fits the level, $Z(y)$ has distribution γ_3 and is independent of all coarse values and of the latent values of the other cells, hence of $(C(y), s)$.

The density of $(c, \zeta, s) = (D(y), s)$ is $\tilde{p}(c, \zeta, s) = p_h(D^{-1}(c, \zeta), s)\, J_D(c, \zeta)$ with $J_D = |\det \nabla D^{-1}|$, positive and continuous. Independence gives $\tilde{p}(c, \zeta, s) = \varphi_3(\zeta)\, \omega(c, s)$ for almost every (c, ζ, s), with ω the density of (c, s), hence $\tilde{p}(c, \zeta, s)\, \varphi_3(\zeta') = \tilde{p}(c, \zeta', s)\, \varphi_3(\zeta)$ for almost every (c, ζ, ζ', s), and then at every point, both sides being continuous.

With $y = D^{-1}(c, \zeta)$ and $y' = D^{-1}(c, \zeta')$ this reads

$$\frac{p_h(y, s)}{p_h(y', s)} = \frac{\varphi_3(\zeta)\, J_D(c, \zeta')}{\varphi_3(\zeta')\, J_D(c, \zeta)},$$

which does not depend on s. So $y \sim y'$. □

Part (a′) follows from the first paragraph of the proof of Lemma D.13: $Z(y)$ is γ_3 and independent of $(C(y), s)$, so given $C(y) = c$ the cell $y = D^{-1}(c, Z(y))$ is a function of a variable independent of s. The lemma itself, which uses neither symmetry nor weight sharing across cells, is what part (c′) uses. By the chain rule of mutual information, $I(y; s) = I(C(y); s) + I(y; s \mid C(y))$, valid in $[0, \infty]$ (Polyanskiy & Wu, 2024), part (a) of Theorem 2 is equivalent to $I(y; s \mid C(y)) = 0$. It implies $I(y; s) = I(C(y); s)$, and is equivalent to it when $I(y; s) < \infty$.

Hence the scalar $C(y)$ attains $\sup_f I(f(y); s) = I(y; s)$ over measurable functions f of the cell. The criterion of Koch-Janusz & Ringel (2018) and Lenggenhager et al. (2020) is this supremum with three restrictions: the surroundings are replaced by a region separated from the cell by a buffer, the number of coarse variables is fixed in advance, and f ranges over a parametric family of stochastic maps. Since such a region is a function of s and the data-processing inequality holds for stochastic f (Polyanskiy & Wu, 2024), part (a′) shows that $C(y)$ attains the supremum of their criterion over all stochastic maps of the cell, with the cell taken in the hidden variables h_ℓ.

The hypothesis of the next lemma allows homeomorphisms α, χ and positive factors f_1, f_2 so that it is unchanged by the sitewise reparameterizations of Theorem 1, which need not be analytic. It is then a property of the level rather than of the units chosen for h_ℓ.

Lemma D.14 (Separating–independent dichotomy). *Suppose that $p_h(y, s) = P(\alpha(y), \chi(s))\, f_1(y)\, f_2(s)$, where α and χ are homeomorphisms of $\mathbb{R}^4$ and of $\mathbb{R}^{|\Lambda_\ell|-4}$ onto open sets $\mathcal{U}$ and $\mathcal{T}$, P is positive and real-analytic on $\mathcal{U} \times \mathcal{T}$, and f_1, f_2 are positive. Then either the level is separating or the cell is independent of its surroundings.*

Proof. If all values in a nonempty open set U_0 are equivalent, then $P(\xi, t)P(\xi', t') - P(\xi, t')P(\xi', t) = 0$ for $\xi, \xi' \in \alpha(U_0)$, which is open, and all $t, t' \in \mathcal{T}$.

As $\mathcal{U}$ and $\mathcal{T}$ are connected, the left side is real-analytic on the open connected set $\mathcal{U}^2 \times \mathcal{T}^2$ and vanishes on a nonempty open subset, hence identically, by the identity theorem for real-analytic functions (Krantz & Parks, 2002). Then $P(\xi, t) = P(\xi, t_0)P(\xi_0, t)/P(\xi_0, t_0)$, and $p_h(y, s)$ is a product of a function of y and a function of s, so the cell is independent of its surroundings. □

Proof of part (b′). By Lemma D.1(iii) the hypothesis of Lemma D.14 holds at every level, with α, χ the identity, when p and the block maps are real-analytic, and it survives sitewise reparameterizations of h, which only compose α, χ with homeomorphisms and multiply f_1, f_2 by positive factors.

If p is translation invariant, the distribution of h_ℓ is invariant under translations by two sites (Lemma D.2), which act transitively on the cells, so the alternative of Lemma D.14 is the same for every cell.

In the second case every cell is independent of all others jointly, and the cells are then mutually independent: split off the first cell, note that the second is independent of the remaining ones because it is independent of all others jointly, and repeat. □

D.4.3 IDENTIFIABILITY OF THE COARSE VALUE

Proof of part (c′). Let $D = (C, Z)$ and $D' = (C', Z')$ fit a separating level. Level sets of C and of C' each consist of mutually equivalent cell values (Lemma D.13). If a level set of C met a continuum of level sets of C', their union would contain an open set of mutually equivalent values, which separation forbids. A level set $\Gamma = \{C = c_0\}$ is the image of $\{c_0\} \times \mathbb{R}^3$ under D^{-1}, hence connected, so $C'(\Gamma)$ is an interval, and all points of Γ are equivalent by Lemma D.13 for D.

Suppose $C'(\Gamma)$ contains an open interval I_0. The set $U = C'^{-1}(I_0)$ is open and nonempty, and every $y \in U$ has $C'(y) = C'(y_1)$ for some $y_1 \in \Gamma$, so $y \sim y_1$ by Lemma D.13 for D'. Then all points of U are equivalent, contradicting that the level is separating. So $C'(\Gamma)$ is an interval containing no open interval, hence a single point.

So C' is constant on every level set of C, that is, $C' = \phi \circ C$ with $\phi(c) = C'(D^{-1}(c, 0))$, a C^1 function. From $\nabla C' = \phi'(C)\nabla C$ and $\nabla C'$ vanishing nowhere, ϕ' vanishes nowhere, and ϕ is onto because C' is. So ϕ is a monotone C^1 diffeomorphism of $\mathbb{R}$, and $D' \circ D^{-1}$ has the form $(c, \zeta) \mapsto (\phi(c), R(c, \zeta))$, where $R(c, \cdot)$ is a C^1 diffeomorphism of $\mathbb{R}^3$ for every c as in the proof of Lemma D.9.

Given $C(y) = c$, equivalently $C'(y) = \phi(c)$, both ζ and $R(c, \zeta)$ have distribution γ_3, so $R(c, \cdot)$ preserves γ_3 for almost every c under the distribution of $C(y)$, whose density $\int \tilde{p}_y(c, \zeta)\, \mathrm{d}\zeta$ is positive, $\tilde{p}_y$ being the density of $D(y)$. So the exceptional set is Lebesgue null.

For bounded continuous f the continuous map $c \mapsto \int f(R(c, \zeta))\varphi_3(\zeta)\, \mathrm{d}\zeta$ equals $\int f\, \mathrm{d}\gamma_3$ off a null set, hence everywhere, and $R(c, \cdot)$ preserves γ_3 for every c. □

Without assuming that the level is separating, the same argument shows that C' is constant on every level set of C whose class of equivalent cell values contains no nonempty open set.

D.5 SELECTION BY THE REGULARIZERS

Proposition D.15 states what the two regularizers select among exactly fitting hierarchies, one kind of freedom at a time: the units of the hidden variables (a); the sign of a block map and, more generally, its composition with a symmetry of its input (b); the latent values of a decimator (c); and the whole decimator at the levels where the likelihood does not fix the coarse value and at the coarsest level (d). The proof occupies the next three subsubsections, the last of which treats parts (c) and (d). The subsection closes by showing that the selected maps are symmetric.

A *symmetry of the input* of a disentangler is a block map k that is a C^1 diffeomorphism, commutes with every P_σ and is odd, and whose cell layer on the offset cells preserves the distribution of x_ℓ. This is stronger than preserving the distribution of one offset cell's four values. The flip $k = -\mathrm{id}$ is always one.

A hierarchy satisfies the *displacement condition* if each of its disentanglers T satisfies $\|T(b) - b\| < \frac{1}{2}\|k(b) - b\|$ for every symmetry $k \neq \mathrm{id}$ of its input, and each of its decimators satisfies $\|T^{\mathrm{dec}}(y) - y\| < \|\bar{y}\|$, with $\bar{y}$ the normalized cell average. These are the two inequalities of Proposition D.15(b) below.

Write $Hy = (\bar{y}, y_\Delta)$ for a cell y in Haar coordinates, and ν_c for the conditional distribution of y_Δ given that the coarse value is $C(y) = c$. A hierarchy satisfies the *regularity condition* if, at every level, ν_c is absolutely continuous for every c; the optimal transport maps $\nabla\psi_c^*$ from ν_c to γ_3 (the star is part of the name, not a convex conjugate), at separating levels, and $\nabla\psi^*$ from the distribution of Hy to γ_4, at levels with independent cells and at the coarsest level, are C^1 diffeomorphisms, the former jointly C^1 in (c, y_Δ); and the decimator block map $y \mapsto H^\top(C(y), \nabla\psi^*_{C(y)}(y_\Delta))$ admits a straight flow. These are the maps R^* and $\nabla\psi^*$ selected in the proofs of (c) and (d) below. In case (d) the selected block map is the gradient of a convex function and so admits a straight flow (Lemma D.3). The first clause makes the optimal transport map from ν_c unique. The second makes the selected

maps C^1 and, by Lemma D.19, odd and equivariant. The third makes the selected decimator block map a diffeomorphism, so that it stays in the class, whose kinetic energy equals its displacement. The condition concerns the hierarchy's own coarse value through ν_c.

Proposition D.15 (Selection)**.** *Under the symmetry assumption, consider normalized, exactly fitting hierarchies of finite $\mathcal{K}$ whose block maps admit straight flows.*

(a) Units. *A sitewise reparameterization of a field of hidden variables that keeps the hierarchy normalized and equivariant is $\pm\mathrm{id}$. The canonical normalization is unchanged by any odd increasing sitewise reparameterization.*

(b) Signs and symmetries. *Let $k \neq \mathrm{id}$ be a symmetry of the input of a disentangler T. If $\|T(b)-b\| < \frac{1}{2}\|k(b)-b\|$ then $T\circ k$ has larger kinetic energy than T. If a decimator satisfies $\|T^{\mathrm{dec}}(y)-y\| < \|\bar{y}\|$, with $\bar{y}$ the normalized cell average, no decimator of at most the same kinetic energy has the opposite coarse value.*

(c) Latent values. *Under the regularity condition, for a given coarse value map C, $\mathcal{K}$ has exactly one minimizer over the latent reparameterizations $(C, Z) \mapsto (C, R(C, Z))$: the latent values are the images of the cell's difference coordinates under the optimal transport map to γ_3, conditionally on the coarse value.*

(d) Independent cells. *Under the regularity condition, at a level whose cells are independent, and at the coarsest level, $\mathcal{K}$ has exactly one minimizer among decimators that fit the level with normalized coarse values: the optimal transport map from the distribution of the cell to γ_4, which is the identity if the cell is already γ_4.*

Throughout this subsection the hierarchies are in $\mathcal{C}_{\mathrm{sym}}$, are exactly fitting, (D2) holds, and all fields of hidden variables have finite second moments, which follows from (D1) when $\mathcal{K} < \infty$.

Part (a) follows from Lemma D.7 alone (Appendix D.5.1). Part (b) compares displacements (Appendix D.5.2): Lemma D.16 bounds from below the displacement of a block map composed with a symmetry of its input, and Lemma D.17 shows that this composition is a legitimate competitor, that is, that exchanging the two block maps keeps the hierarchy exactly fitting and normalized. Parts (c) and (d) rest on Lemma D.18 (Appendix D.5.3), which identifies the measure-preserving map closest to a given map F as an optimal transport map composed with F. Lemma D.19, after the proofs, shows that the selected maps inherit the symmetries of their input.

D.5.1 Units

Proof of Proposition D.15(a). In a normalized hierarchy every site of a field of hidden variables has distribution γ_1. A sitewise reparameterization $(g_r)_r$ keeps the block maps on both sides in $\mathcal{C}_{\mathrm{sym}}$ only if it commutes with the position permutations, which act transitively, and with the flip. So $g_r = g$ for all r, and g is odd.

If the new hierarchy is normalized then g preserves γ_1, so $g = \pm\mathrm{id}$ by Lemma D.7(b).

Let μ be the one-site distribution of a field of hidden variables, with distribution function F_μ, and $Q_\mu = \Phi^{-1} \circ F_\mu$, increasing and odd because μ is symmetric. Reparameterizing the field by an odd increasing g makes its one-site distribution $g_{\#}\mu$, so $F_{g_{\#}\mu} = F_\mu \circ g^{-1}$ and $Q_{g_{\#}\mu} \circ g = Q_\mu$: the normalized field, and with it every normalized block map, is unchanged. For $g = -\mathrm{id}$ the normalized field changes sign. □

D.5.2 Signs and symmetries of the input

Recall that $\delta = \|T(b) - b\|$ is the L^2 norm of a block map's displacement field under its input distribution μ. A symmetry k of the input preserves μ, and we write $m_k = \|k(b) - b\|$. Lemma D.16 is a nearest-point argument: if T moves its input by less than half the distance that a symmetry k of the input moves it, T is strictly nearer the identity than $T \circ k$, in the same way that a point within half the spacing of a lattice has a unique nearest lattice point.

Lemma D.16 (Displacement cost of a symmetry)**.** *Let $b \sim \mu$ on $\mathbb{R}^d$ with finite second moment, let T be a measurable map with $\delta = \|T(b) - b\| < \infty$, and let k be a measurable map that preserves μ, with $m_k = \|k(b) - b\|$. Then $\|T(k(b)) - b\| \geq m_k - \delta$. In particular $\|T(k(b)) - b\| > \delta$ whenever $\delta < m_k/2$. For $k = -\mathrm{id}$ and symmetric μ, $m_k = 2\|b\|$.*

Proof. Split $T(k(b)) - b = (k(b) - b) + (T(k(b)) - k(b))$, where $\|T(k(b)) - k(b)\| = \delta$ because $k(b) \sim \mu$, and apply the triangle inequality. □

The flip itself violates the condition: with T and k both the flip, $T \circ k$ is the identity, which is cheaper.

The disentangler comparison in part (b), and in the disentangler case of the proof of Theorem 3, is between a disentangler and its composition with a symmetry of its input. The next lemma shows that every second disentangler with the same input and output distributions is such a composition, and that exchanging the two keeps the hierarchy exactly fitting and normalized and leaves the kinetic energy of every other layer unchanged.

Lemma D.17 (Disentanglers with the same input and output distributions)**.** *If two disentanglers with block maps T, T' map the same distribution of x_ℓ to the same distribution of h_ℓ, then $T' = T \circ k$ for a symmetry k of the input. Replacing one by the other in an exactly fitting normalized hierarchy gives an exactly fitting normalized hierarchy with the same kinetic energy in all other layers, whose block maps admit straight flows if those of the original and the new block map do.*

Proof. $k = T^{-1} \circ T'$ is a C^1 diffeomorphism that commutes with every P_σ and is odd, and its layer maps the distribution of x_ℓ to itself because the two disentangler layers are bijections of V_ℓ that map it to the same distribution.

After the replacement the distribution of h_ℓ, hence of every later field of hidden variables, is unchanged, and the displacement of a layer depends only on its block map and the distribution of its input.

The pyramid is standard normal before the replacement, so $(z_0, \dots, z_{\ell-1})$, unchanged as a function of x, is independent of x_ℓ (Lemma D.1). The remaining components are a function of x_ℓ alone before and after, with unchanged distribution because that of h_ℓ is unchanged. Hence the joint distribution of the pyramid is unchanged, and the modified hierarchy is exactly fitting and normalized. □

Part (b) now follows.

Proof of Proposition D.15(b). By Lemma D.16 applied to the four values of one offset cell, $T \circ k$ has a strictly larger displacement than T when $\delta < m_k/2$.

For the flip, replacing the disentangler T by $-T = T \circ (-\mathrm{id})$, which is T composed with the flip since T is odd, while leaving the decimator fixed, changes the sign of h_ℓ, leaves the distribution of h_ℓ unchanged because it is symmetric, and leaves the displacement of the decimator unchanged, because $T^{\mathrm{dec}}(-h) - (-h) = -(T^{\mathrm{dec}}(h) - h)$ for an odd block map. This is the case $k = -\mathrm{id}$, where the condition reads $\|T(b) - b\| < \|b\|$.

For a decimator $D = (C, Z)$ with input y, let $\bar{y} = (Hy)_1$. If decimators with coarse values C and $-C$ both move their input by at most δ in L^2, then, H being orthogonal, $\|C(y) - \bar{y}\| \le \delta$ and $\|-C(y) - \bar{y}\| \le \delta$, so $\|\bar{y}\| \le \delta$. A decimator with $\|T^{\mathrm{dec}}(y) - y\| < \|\bar{y}\|$ therefore has no competitor of at most the same kinetic energy with the opposite coarse value. □

Verifiability on trained models. The inequalities for the flip and for a decimator's coarse sign involve only the displacements and second moments of the block maps. The inequality for every other symmetry of the input requires knowing every equivariant odd block map that preserves the distribution of x_ℓ, and $\|k(b) - b\|$ is not a function of second moments. If that distribution has such symmetries arbitrarily close to the identity, the condition holds only for the identity block map. An i.i.d. standard normal field is an example. Take a smooth vector field that preserves γ_4, supported in a ball whose images under D_4 and the flip are disjoint, and average it over those images. The time-t maps of its flow, applied to every offset cell, are symmetries of the input for small t. There the hierarchy with identity block maps is exactly fitting and normalized, because the Haar split is orthogonal, has $\mathcal{K} = 0$, and is the unique minimizer, because $\mathcal{K} = 0$ forces every block map to equal the identity almost everywhere, hence everywhere by continuity. So the displacement condition is a real restriction: it fails for every disentangler other than the identity whose input has symmetries arbitrarily close to the identity, harmlessly in this example, where the identity is the minimizer and satisfies it.

D.5.3 LATENT VALUES AND UNDETERMINED COARSE VALUES

In parts (c) and (d) the competitors of a block map are its compositions with maps that preserve a standard normal distribution, and the question is which of these measure-preserving maps lies closest, in L^2, to a fixed map F determined by the block map. The answer is the optimal transport map from the distribution of F to the standard normal, composed with F: each measure-preserving map ς, paired with F, gives a coupling of the two distributions, and when the distribution of F is absolutely continuous the optimal coupling is unique and given by a map. We write W_2 for the 2-Wasserstein distance and $\nabla\psi^*$ for an optimal transport map, the gradient of a convex function. In Lemma D.18, μ is the distribution that the competitors preserve.

Lemma D.18 (Closest measure-preserving map)**.** *Let μ be a distribution on $\mathbb{R}^m$ with finite second moment and $F \in L^2(\mu;\mathbb{R}^m)$ such that $\nu = F_{\#}\mu$ is absolutely continuous. Among measurable maps ς that preserve μ, $\|\varsigma - F\|_{L^2(\mu)}$ has a unique minimizer up to μ-null sets, namely $\varsigma = \nabla\psi^* \circ F$, where $\nabla\psi^*$ is the optimal transport map from ν to μ for the quadratic cost.*

Proof. Since $F \in L^2(\mu)$, ν has finite second moment. If ς preserves μ, the distribution of $(F(b), \varsigma(b))$, $b \sim \mu$, is a coupling of ν and μ, so $\|\varsigma - F\|^2 \geq W_2^2(\nu,\mu)$, the least value of $\mathbb{E}\|A-B\|^2$ over pairs of random vectors with $A \sim \nu$ and $B \sim \mu$. A pair that attains it is an optimal coupling.

Since ν is absolutely continuous, the optimal coupling is unique and deterministic, $B = \nabla\psi^*(A)$ with ψ^* convex (Villani, 2003), Brenier's theorem (Brenier, 1991) in the form of McCann (1995), which needs neither a bounded domain nor moment conditions. We call $\nabla\psi^*$ the optimal transport map from ν to μ, which is the optimal transport map of Section 3. Equality therefore holds if and only if $\varsigma = \nabla\psi^* \circ F$ μ-almost everywhere (Brenier's polar factorization; Brenier, 1991; Villani, 2003), and this map preserves μ because $\nabla\psi^*$ maps ν to μ. □

Lemma D.18 gives parts (c) and (d), with F the map from the latent values to the difference coordinates of the cell in case (c), and the map from the decimator output to the Haar coordinates of the cell in case (d).

Proof of Proposition D.15(c). Let $D = (C, Z)$ be the decimator of a level. A latent reparameterization changes this block map only, to $(C, R(C, Z))$, and changes no field of hidden variables. With the cell input in Haar coordinates, $Hy = (\bar{y}, y_\Delta)$, let $F_c(\zeta)$ be the difference coordinates of $H\,D^{-1}(c,\zeta)$, so that $y_\Delta = F_c(\zeta)$.

Since H is orthogonal, $\|T^{\mathrm{dec}}(y) - y\| = \|H\,T^{\mathrm{dec}}(y) - Hy\|$, so the displacement of $(C, R(C, Z))$ splits into a coarse and a latent part,

$$\mathbb{E}\,|C(y) - \bar{y}|^2 + \mathbb{E}\,\|R(C(y), Z(y)) - y_\Delta\|^2.$$

A latent reparameterization leaves the first alone. Given $C(y) = c$ the latent value $\zeta = Z(y)$ has distribution γ_3 (Lemma D.1) and $y_\Delta = F_c(\zeta)$, so the second is $\mathbb{E}_c\,\|R(c,\cdot) - F_c\|^2_{L^2(\gamma_3)}$.

Assume that for almost every c the conditional distribution $\nu_c = (F_c)_{\#}\gamma_3$ of y_Δ given $C(y) = c$ is absolutely continuous. By Lemma D.18, for almost every c the second term is bounded below by $W_2^2(\nu_c, \gamma_3)$, with equality exactly for $R^*(c,\cdot) = \nabla\psi_c^* \circ F_c$, $\nabla\psi_c^*$ being the optimal transport map from ν_c to γ_3. The resulting latent values are $\nabla\psi_c^*(y_\Delta)$.

Two latent reparameterizations that attain the bound agree almost everywhere, hence everywhere, being continuous. So the displacement has exactly one minimizer over latent reparameterizations, R^*, and under the regularity condition, which makes R^* a latent reparameterization whose decimator block map admits a straight flow, R^* also minimizes $\mathcal{K}$ uniquely. □

Proof of Proposition D.15(d). If the cells of h_ℓ are mutually independent and the hierarchy is normalized, the coarse values are independent with distribution γ_1 and independent of the latent values, so the output of every cell is γ_4 and the cell outputs are independent. A decimator D' fits the level with normalized coarse values if and only if $D' = \Xi \circ D$ with Ξ preserving γ_4, and the change leaves the distribution of $x_{\ell+1}$ unchanged, hence also the displacement of all later layers and, as in the proof of Lemma D.17, the fit.

The displacement of $\Xi \circ D$ is $\|\Xi - F\|^2_{L^2(\gamma_4)}$ with $F = H \circ D^{-1}$, and by Lemma D.18 it is minimized exactly by $\Xi = \nabla\psi^* \circ F$, for which $\Xi \circ D = \nabla\psi^* \circ H$.

Since H is orthogonal, $\nabla(\psi^* \circ H)(y) = H^\top \nabla\psi^*(Hy)$, so the block map $H^\top \nabla\psi^*(H\,\cdot)$ is the gradient of the convex function $\psi^* \circ H$ and is therefore the optimal transport map from the distribution of the cell to γ_4 (Villani, 2003). It is the identity if the cell is already γ_4. The same holds at the coarsest level, whose whole output is latent. □

This proves Proposition D.15. The rest of this subsection shows that the selected maps are symmetric.

Optimal transport maps between symmetric distributions are themselves symmetric, so the maps selected in (c) and (d) are equivariant and odd without this being imposed.

Lemma D.19 (Optimal transport maps inherit symmetry)**.** *Let μ, ν have finite second moments, μ absolutely continuous, with optimal transport map $\nabla\psi$ from μ to ν, and let A be an orthogonal matrix. Then the optimal transport map from $A_\#\mu$ to $A_\#\nu$ is $A \circ \nabla\psi \circ A^{-1}$. In particular $\nabla\psi$ commutes with A μ-almost everywhere if A preserves μ and ν, hence everywhere if $\nabla\psi$ is continuous and μ has a positive density.*

Proof. The map $A \circ \nabla\psi \circ A^{-1}$ carries $A_\#\mu$ to $A_\#\nu$ at the cost of $\nabla\psi$, which is the optimal cost between the images because A is an isometry, and the optimal transport map is unique (McCann, 1995). □

The input distribution of a cell is invariant under every P_σ and the flip (Lemma D.2). In case (d) Lemma D.19 therefore shows that the selected block map commutes with every P_σ and is odd.

In case (c), P_σ fixes $\bar{y}$ and acts on y_Δ by a signed permutation τ_σ (a permutation matrix with some entries -1) that preserves ν_c, and the flip maps ν_c to ν_{-c}, so $\nabla\psi_c^*$ commutes with τ_σ and

$$\nabla\psi_{-c}^* = -\nabla\psi_c^*(-\,\cdot).$$

The selected maps are therefore in $\mathcal{C}_{\mathrm{sym}}$ whenever they are C^1 diffeomorphisms, jointly C^1 in (c, y_Δ) in case (c).

D.6 Uniqueness given the hidden distributions

Assume (D2). Let $\mathcal{H}$ be the set of hierarchies in $\mathcal{C}_{\mathrm{sym}}$ that are exactly fitting, are normalized, have block maps that admit straight flows and finite $\mathcal{K}$, satisfy the hypothesis of Lemma D.14 at every level, and for which the conditional distributions ν_c of Proposition D.15(c) are absolutely continuous.

Theorem D.20 (Theorem 3, precise form)**.** *Let $M, M' \in \mathcal{H}$ both minimize $\mathcal{K}$ over $\mathcal{H}$, let the disentangler outputs h_ℓ, $\ell \le n-2$, have the same distribution in M and in M', and let M and M' satisfy the displacement condition and the regularity condition of Appendix D.5. Then $M = M'$.*

The theorem of Section 3 follows: a real-analytic hierarchy of the class stated there lies in $\mathcal{H}$ by Lemma D.1(iii) and the regularity condition, and a minimizer over that class which lies in $\mathcal{H}$ minimizes over $\mathcal{H}$.

The proof is an induction over the layers, from the finest, and its three cases are the three mechanisms of Appendix D.5. Suppose the two hierarchies agree below level ℓ, so that the input of the disentangler of level ℓ is the same function of x in both. The two disentanglers map this common input distribution to the common distribution of h_ℓ, so they differ by a symmetry of the input (Lemma D.17), which the displacement condition prices out (Lemma D.16). The two decimators then act on the same h_ℓ. At a separating level their coarse values agree up to a monotone map (Theorem D.12), normalization reduces the map to a sign, the displacement condition excludes the sign, and both decimators are the unique minimizer of Proposition D.15(c). At a level with independent cells, and at the coarsest level, both are the unique minimizer of Proposition D.15(d). There most of the argument checks that the selected decimator keeps the hierarchy in $\mathcal{H}$. Throughout, ν_c is the distribution of a cell's difference coordinates y_Δ given its coarse value c, and F_c the map that expresses y_Δ through the latent values, as in the proof of Proposition D.15(c).

Proof of Theorem 3. We show by induction on ℓ that the levels below ℓ coincide, so that $x_\ell = x'_\ell$ as functions of x. Assume this for some $\ell \leq n-2$. Each of the three cases below gives one more common layer.

Disentangler. T and T' map the common distribution of x_ℓ to the common distribution of h_ℓ, so $T' = T \circ k$ for a symmetry k of the input, and replacing T' by T in M' gives an element of $\mathcal{H}$ (Lemma D.17; the replaced block map is in $\mathcal{C}_{\mathrm{sym}}$, and the hypothesis of Lemma D.14 and the absolute continuity of the ν_c depend only on the distributions of the fields of hidden variables and on the decimators, which are unchanged). By minimality of M' the displacement of T' is at most that of T, and conversely. By the displacement condition and Lemma D.16, $k \neq \mathrm{id}$ would give $T \circ k$ a strictly larger displacement than T. So $T' = T$ and $h'_\ell = h_\ell$.

Decimator, separating level. Both decimators fit the level on the same h_ℓ, so $D' = (\phi \circ C, R(C, Z))$ by Theorem D.12(c′), and ϕ preserves the one-site distribution γ_1 of the coarse variables, so $\phi = \pm\mathrm{id}$ (Lemma D.7(b)).

Replacing D' by D in M' changes $x_{\ell+1}$ at most by a sign. The distribution of $x_{\ell+1}$ is symmetric (Lemma D.2) and every later block map is odd, so a sign change of $x_{\ell+1}$ leaves the distributions of all later fields of hidden variables and the kinetic energies of all later layers unchanged. It changes the later part of the pyramid by a sign at most, which preserves the standard normal distribution. The argument of Lemma D.17 then places the modified hierarchy in $\mathcal{H}$: the ν_c at level ℓ are now those of M, and at later levels a sign change of the input replaces ν_c by $(-\mathrm{id})_\# \nu_{-c}$, so they stay absolutely continuous. As before, minimality of both hierarchies forces D and D' to have the same displacement, which we write δ^2.

If $\phi = -\mathrm{id}$, the two decimators have coarse values C and $-C$ and displacement δ^2 each, so $\|\bar{y}\| \leq \delta$ by the computation in the proof of Proposition D.15(b), which the displacement condition excludes. So $C' = C$ and D' is a latent reparameterization of D.

By the regularity condition and Lemma D.19 the map R^* of Proposition D.15(c) is a latent reparameterization with block map in $\mathcal{C}_{\mathrm{sym}}$, and the hierarchy it gives is in $\mathcal{H}$, since a latent reparameterization changes no field of hidden variables. Minimality of M forces the displacement of D to be at most that of the R^* version, so D attains the bound of Proposition D.15(c) and is the R^* version. The same holds for D'. The R^* version does not depend on the latent map one starts from: its coarse value is C and its latent values are $\nabla\psi_c^*(y_\Delta)$, with ν_c determined by C and the distribution of h_ℓ, both common to M and M'. So $D' = D$.

Decimator, other levels. If the level is not separating, its cells are independent (Theorem D.12(b′)) and $D' = \Xi \circ D$ with Ξ preserving γ_4 (proof of Proposition D.15(d)). Let $D_* = \nabla\psi^* \circ H = (C_*, Z_*)$ be the decimator selected there, in Haar coordinates. Its block map $H^\top \nabla\psi^*(H\,\cdot)$ is in $\mathcal{C}_{\mathrm{sym}}$ by the regularity condition and Lemma D.19, and admits a straight flow by Lemma D.3.

Replacing D by D_* leaves the distribution of $x_{\ell+1}$, i.i.d. γ_1, and hence the distributions of all later fields of hidden variables unchanged, and keeps the hierarchy exactly fitting and normalized, as shown in the proof of Proposition D.15(d). The hypothesis of Lemma D.14 and the absolute continuity of the ν_c at later levels depend only on those distributions and on the unchanged later decimators.

It remains to check that D_* keeps the hierarchy in $\mathcal{H}$, whose definition requires ν_c to have a density at every level, including this one. At level ℓ itself, write $A = \nabla\psi^*$, a C^1 diffeomorphism of $\mathbb{R}^4$ whose Jacobian is symmetric and positive definite, being positive semidefinite by convexity and invertible. So is the Jacobian of A^{-1}, and hence its lower right 3×3 block. The output of D_* in Haar coordinates is

$$(c, \zeta) = A(Hy),$$

so given $C_*(y) = c$ the difference coordinates are

$$y_\Delta = F_c(\zeta), \qquad \zeta \sim \gamma_3,$$

with F_c as in the proof of Proposition D.15(c), here the last three coordinates of $A^{-1}(c, \zeta)$.

The Jacobian of F_c is that block, so F_c is a local C^1 diffeomorphism of $\mathbb{R}^3$, and $\nu_c = (F_c)_\# \gamma_3$ is absolutely continuous: $\mathbb{R}^3$ is a countable union of open sets on which F_c is a diffeomorphism, and a diffeomorphism maps Lebesgue-null sets to Lebesgue-null sets (Rudin, 1987). So the modified hierarchy is in $\mathcal{H}$.

Minimality of M forces the displacement of D to be at most that of D_*, so D attains the bound of Proposition D.15(d) and equals D_* almost everywhere, hence everywhere by continuity; likewise D'. At level $n-1$ the same argument applies, the whole output being latent.

In each case $x'_{\ell+1} = x_{\ell+1}$, which closes the induction, so M and M' have the same block map in every layer. □

Reflecting every block map of a hierarchy by a lattice symmetry gives its *mirror image*, which has the same $\mathcal{K}_j$ in every layer. Building the symmetry into the block maps removes these pairs by construction. Proposition D.21 shows that without it a unique minimizer of $\mathcal{K}$ is nevertheless equivariant.

Proposition D.21 (A unique minimizer is symmetric)**.** *Assume (D2). Let $\mathcal{H}_0$ be the set of normalized, exactly fitting hierarchies in $\mathcal{C}$ whose block maps admit straight flows, with no symmetry imposed on the block maps. If $\mathcal{K}$ has a unique minimizer on $\mathcal{H}_0$, that minimizer is in $\mathcal{C}_{\mathrm{sym}}$.*

Proof. Let g be a lattice symmetry or the flip, with position permutation σ, and for a hierarchy M let M^g have every block map T replaced by $P_\sigma T P_\sigma^{-1}$, or by $-T(-\cdot)$ for the flip. As in the proof of Lemma D.2,

$$g \circ \mathcal{L}_{\mathcal{P},T} = \mathcal{L}_{\mathcal{P},P_\sigma T P_\sigma^{-1}} \circ g, \qquad S_\ell \circ g = (g' \times \tau_\sigma) \circ S_\ell,$$

with τ_σ a signed permutation of the latent coordinates, because each P_σ fixes the first row of $H = H_1 \otimes H_1$ and permutes the other three up to signs.

Hence the pyramid of M^g at x is a signed permutation of the pyramid of M at $g^{-1}x$, and its fields of hidden variables are the images under g or g' of those of M at $g^{-1}x$.

Since p is g-invariant and the standard normal distribution is invariant under signed permutations, M^g is exactly fitting; its fields of hidden variables are those of M up to a permutation of the sites and a sign, so every site still has marginal $\mathcal{N}(0,1)$ and M^g is normalized; and it has the same layer kinetic energies as M, since conjugation by an orthogonal map preserves displacements and straight paths.

So $M \mapsto M^g$ maps $\mathcal{H}_0$ to itself and preserves $\mathcal{K}$, a unique minimizer satisfies $M^g = M$, and this for all g says that every block map commutes with every P_σ and is odd, that is, the minimizer is in $\mathcal{C}_{\mathrm{sym}}$. □

D.7 Uniqueness across hidden distributions

The first part of Corollary 4 removes part of the hypothesis of Theorem 3 that the fields h_ℓ have the same distributions in the two hierarchies: for almost every positive weighting w of the layers, all hierarchies in the class of Theorem 3 of least $\mathcal{K}_w = \sum_j w_j \mathcal{K}_j$ have the same kinetic energy in every layer. Weights matter because two hierarchies can have the same total kinetic energy but distribute it differently over the layers, and for almost every weighting this cannot happen among minimizers. The proof uses only that the least value, an infimum over the class, is a concave function of the weights, hence differentiable almost everywhere, and that at a point of differentiability every minimizer realizes the gradient. The second hypothesis of the corollary, which we call the *no-tie assumption*, is discussed after the proof.

Proof of Corollary 4. Let $\mathcal{S} \subset [0,\infty)^{2n-1}$ be the set of vectors $\boldsymbol{\mathcal{K}} = (\mathcal{K}_j(M))_j$ of the hierarchies M in the class of Theorem 3, and $f(w) = \inf_{\boldsymbol{\mathcal{K}} \in \mathcal{S}} w \cdot \boldsymbol{\mathcal{K}}$ for $w \in (0,\infty)^{2n-1}$. If the class is empty there is nothing to prove, so let $\boldsymbol{\mathcal{K}}_0 \in \mathcal{S}$. Then $0 \le f(w) \le w \cdot \boldsymbol{\mathcal{K}}_0 < \infty$, and f, an infimum of linear functions, is a finite concave function on an open convex set, hence differentiable at Lebesgue-almost every w (Rockafellar, 1970). The exceptional set depends on p through $\mathcal{S}$, and it is a cone, because $f(tw) = tf(w)$ for $t > 0$.

Let w be a point of differentiability and let $\boldsymbol{\mathcal{K}} \in \mathcal{S}$ attain the infimum. Since f is an infimum over $\mathcal{S}$, $f(w') \le w' \cdot \boldsymbol{\mathcal{K}}$ for every w', with equality at w, hence $f(w') \le f(w) + \boldsymbol{\mathcal{K}} \cdot (w' - w)$ for all w'. A concave function that is differentiable at w admits only $\nabla f(w)$ in this inequality (Rockafellar, 1970), so $\boldsymbol{\mathcal{K}} = \nabla f(w)$, the same vector for every minimizer.

Each exchange in the proof of Theorem 3 replaces the block map of a single layer and leaves the kinetic energy of every other layer unchanged, so comparing two weighted totals reduces to comparing that one layer and uses only that its weight is positive. Theorem 3 therefore holds with $\mathcal{K}_w$ in place of $\mathcal{K}$.

Consequently, let w be outside the exceptional set and let M, M' be real-analytic hierarchies of least $\mathcal{K}_w$ over the class of Theorem 3 that satisfy the displacement and regularity conditions. They have the same kinetic energy in every layer, hence, under the second hypothesis of the corollary, the same distributions of $h_0, \dots, h_{n-2}$. They lie in $\mathcal{H}$ (Appendix D.6), and since $\mathcal{H}$ is contained in the class they minimize $\mathcal{K}_w$ over $\mathcal{H}$, so they are equal by Theorem 3 with $\mathcal{K}_w$ in place of $\mathcal{K}$. □

The no-tie assumption rules out two normalized, exactly fitting minimizers of $\mathcal{K}_w$ that have the same kinetic energy in every layer but give some field h_ℓ different distributions. We do not prove it, but the agreement between independently trained models in Section 4 is consistent with it.

The exceptional set of layer weightings depends on p and is a cone, so rescaling the weights does not help, and the corollary does not guarantee that the uniform weighting used in training lies outside it.

D.8 Several disentanglers per level

When a level applies several disentanglers, alternately on the offset cells and on the aligned cells, the results above extend with the changes recorded in Proposition D.22. The proofs go through because consecutive layers still act on different partitions.

Let level $\ell \le n-2$ apply d_ℓ disentanglers in turn, d_ℓ odd, the first on the offset cells, the second on the aligned cells, the third on the offset cells, and so on, so that the last acts on the offset cells. The decimator on the aligned cells and the Haar split follow, and the coarsest level keeps its single decimator. Write $h_{\ell,0} = x_\ell$ and $h_{\ell,r}$ for the output of the r-th disentangler, so that h_{ℓ,d_ℓ} is the input of the decimator. The fields of hidden variables are the $h_{\ell,r}$, $1 \le r \le d_\ell$, and the x_ℓ, and *normalized* means that every site of every one of them has marginal $\mathcal{N}(0,1)$.

The number of layers is $J = n + \sum_{\ell \le n-2} d_\ell$, and $\mathcal{K}$ is the average of the $\mathcal{K}_j$ over these J layers. A symmetry of the input of a disentangler is defined as in Appendix D.5, with the cell layer taken on the disentangler's own partition and the distribution preserved that of its own input field.

Proposition D.22 (Several disentanglers)**.** *In this layout:* (i) *Theorem 1 holds with a sitewise map, one monotone map of $\mathbb{R}$ per position class, at the output of every disentangler;* (ii) *Theorem 2 and Proposition D.15 hold with h_ℓ replaced by h_{ℓ,d_ℓ}, and the displacement condition is required of every disentangler on its own input;* (iii) *Theorem 3 holds when every $h_{\ell,r}$, $1 \le r \le d_\ell$, has the same distribution in M and in M';* (iv) *Corollary 4 holds with J layer weights and the no-tie assumption for every $h_{\ell,r}$ (Remark D.23), and Proposition D.21 holds unchanged.*

Proof. Consecutive layers of a level act on different partitions, and two cells of different partitions that meet share one site when $L_\ell \ge 4$. Every added layer commutes with translations by two sites and with the lattice symmetries by (D.1), and with the flip because the block maps are odd, so Lemma D.2 holds for every $h_{\ell,r}$. Every $h_{\ell,r}$ is a bijective C^1 function of x_ℓ, and the pyramid is still a bijective function of x with $(z_0, \dots, z_{\ell-1})$ independent of x_ℓ at exact fit, so Lemma D.1 holds for every field of hidden variables with the same proof.

(i). At the output of a disentangler the next layer is a cell layer on the other partition, so Lemmas D.5 and D.6 apply, with the roles of offset cells and aligned cells exchanged when the next layer acts on the offset cells. At a decimator output the proof of Theorem 1 is unchanged: in Lemma D.8(a) the added disentanglers of level $\ell+1$ are set to the identity, and the first layer of level $\ell+1$ is still a cell layer on the offset cells of $\Lambda_{\ell+1}$, as Lemma D.8(c) requires.

(ii). Appendix D.4 uses only that the decimator fits its level and the density of its input field, and Appendix D.5 only the input distribution of each block map and the exchange of Lemma D.17, whose proof applies to any disentangler of the stack: replacing $T'_{\ell,r}$ by $T_{\ell,r}$ leaves the distribution of $h_{\ell,r}$, hence of every later field, unchanged, and $(z_0, \dots, z_{\ell-1})$ is independent of x_ℓ, of which the rest of the pyramid is a function.

(iii). The induction of the proof of Theorem 3 runs over the $d_\ell + 1$ layers of a level in order. For the r-th disentangler the induction hypothesis gives $h_{\ell,r-1} = h'_{\ell,r-1}$ as functions of x, and the hypothesis of (iii) gives $h_{\ell,r}$ and $h'_{\ell,r}$ the same distribution, so the disentangler case of that proof applies, with the exchange of (ii), and gives $T'_{\ell,r} = T_{\ell,r}$. The decimator cases are unchanged, since they use only the common input $h_{\ell,d_\ell} = h'_{\ell,d_\ell}$.

(iv). The proof of Corollary 4 uses only that the layer energies form a vector in $[0,\infty)^J$, and that of Proposition D.21 only that each layer is a cell layer on the offset cells or on the aligned cells. □

Remark D.23 (Layers on the same partition). With $d_\ell \geq 3$, two disentanglers of a level act on the same partition. If the layers between them are the identity, a block map may be placed at either of them: both placements realize the same map of the level and, when both are normalized, are exactly fitting together, with the same layer energies in a different order. Such pairs tie under $\mathcal{K}_w$ exactly when the two layers have the same weight, as in the uniform weighting. Whether they occur among the minimizers is not known. The caution about the uniform weighting in Appendix D.7 therefore applies with more force here, and unequal weights for layers on the same partition avoid this particular tie.

D.9 Fields with several components

This subsection extends the results to a field with $m \geq 1$ components per site and the orthogonal group $O(m)$ as internal symmetry. The scalar case is $m = 1$ with $O(1) = \{\pm 1\}$, the flip being -1. Proposition D.25 states the results for m components, with one exception. Part (b) of Theorem 2, the identifiability of the coarse value, needs a hypothesis on the data, separation in dimension, which analyticity supplies for $m = 1$ but not for $m \geq 2$, and Theorem 3 and Corollary 4 inherit it. The XY model of Section 4 is the case $m = 2$. For $m = 2$ the symmetry assumption of Proposition D.25 is the invariance of the data under translations, the lattice symmetries and global $O(2)$ transformations of the spins, rotations and reflections, which the model has. The block maps of Appendix A.5 are built to commute with the actions of D_4 and $O(2)$ defined below, and that appendix uses three disentanglers per level, which Proposition D.25 covers through Proposition D.22. For the XY model, then, Proposition D.25 applies with $m = 2$, and the one hypothesis on the data that the theory does not derive is that every level be separating in dimension or have mutually independent cells.

Layout and symmetry. A field is $x \in (\mathbb{R}^m)^{\Lambda_0}$. A cell of either partition carries a value in $(\mathbb{R}^m)^4 = \mathbb{R}^{4m}$, and a block map is a map of $\mathbb{R}^{4m}$. The Haar transform acts on each component, as $H \otimes I_m$, and the split forwards the normalized cell average, an m-vector, to $x_{\ell+1}$ and the $3m$ difference coordinates to z_ℓ, whose prior is γ_{3m} per aligned cell. The group D_4 acts on $\mathbb{R}^{4m}$ by $P_\sigma \otimes I_m$ and $O(m)$ by $I_4 \otimes Q$. The two actions commute with each other. The action $I_4 \otimes Q$ commutes with $H \otimes I_m$, and $(H \otimes I_m)(P_\sigma \otimes I_m)(H \otimes I_m)^\top = \mathrm{diag}(1, \tau_\sigma) \otimes I_m$ with τ_σ the signed permutation of Appendix D.5. Let $\mathcal{C}^{(m)}$ be the set of hierarchies whose block maps are C^1 diffeomorphisms of $\mathbb{R}^{4m}$, and $\mathcal{C}^{(m)}_{\mathrm{sym}}$ the subset whose block maps commute with both actions. For $m = 1$ these are $\mathcal{C}$ and $\mathcal{C}_{\mathrm{sym}}$, since commuting with $-I$ is oddness. In (D2) the flip is replaced by the action of $O(m)$ on the values. A hierarchy is normalized if every site of every field of hidden variables has distribution γ_m. A symmetry of the input of a disentangler is a C^1 diffeomorphism of $\mathbb{R}^{4m}$ that commutes with both actions and whose cell layer preserves the distribution of x_ℓ. The coarse value $C(y) \in \mathbb{R}^m$ of a cell, its latent values in $\mathbb{R}^{3m}$, equivalence of cell values and a decimator that fits the level are defined as in Appendices D.1 and D.4, with γ_3 replaced by γ_{3m}.

Separation in dimension. Every class of mutually equivalent cell values is a union of level sets of C, which are submanifolds of dimension $3m$ (Lemma D.13, whose proof is dimension-free). We call the level *separating in dimension* at a cell if every class of mutually equivalent values of that cell has Hausdorff dimension at most $3m$, and separating in dimension if this holds at some cell. Under (D2) it then holds at every cell. If a decimator $D = (C, Z)$ fits the level, a class is $C^{-1}(S) = D^{-1}(S \times \mathbb{R}^{3m})$ for a set $S \subset \mathbb{R}^m$ of coarse values (Lemma D.13), and since D^{-1} is locally bi-Lipschitz and the Hausdorff dimension of $S \times \mathbb{R}^{3m}$ is that of S plus $3m$ (Falconer, 2003), the condition says that S has Hausdorff dimension zero: the surroundings distinguish coarse values up to a set of dimension zero.

For $m = 1$ this implies the separating property of Appendix D.4, since a class of dimension at most 3 contains no open set, and analyticity supplies separation in dimension at every level whose

cells are not independent, as follows. Under analyticity of p and the block maps, p_h is real-analytic (Lemma D.1(iii)), and a class other than all of $\mathbb{R}^{4m}$ lies in the zero set of a nonzero real-analytic function $y' \mapsto p_h(y,s)p_h(y',s') - p_h(y,s')p_h(y',s)$, which has Hausdorff dimension at most $4m-1$ (Krantz & Parks, 2002). For $m = 1$ this bound is $3m$, so a level whose cells are not independent is separating in dimension. For $m \geq 2$ it exceeds $3m$. Analyticity still gives the dichotomy of Lemma D.14 between levels that are separating in the sense of Appendix D.4 and independent cells, but separation in dimension is a hypothesis on the data. A density under which the surroundings depend on a cell only through fewer than m functions of class C^1 of its values has some classes of dimension larger than $3m$, and determines only those functions.

Lemma D.24 (Equivariant maps that preserve the standard normal)**.** *Let $m \geq 1$ and let ϕ be a C^1 diffeomorphism of $\mathbb{R}^m$ that commutes with every $Q \in O(m)$ and preserves γ_m. Then $\phi = \pm\mathrm{id}$.*

Proof. For $m = 1$, commuting with -1 means that ϕ is odd, and Lemma D.7(b) gives $\phi = \pm\mathrm{id}$. Let $m \geq 2$.

The orbits of $O(m)$ are the spheres $S_r = \{|c| = r\}$, and equivariance maps orbits onto orbits. So $\phi(S_r) = S_{f(r)}$ for a bijection f of $[0,\infty)$ with $f(0) = 0$, which is continuous because ϕ is, hence increasing. Since ϕ preserves γ_m, the radius $|\phi(c)| = f(|c|)$ has the same distribution as $|c|$ when $c \sim \gamma_m$, and that distribution has a positive density on $(0,\infty)$. An increasing map that preserves it is the identity, so $f(r) = r$: ϕ preserves every sphere.

Fix $c \neq 0$. The stabilizer of c in $O(m)$ consists of the orthogonal maps that fix c and act on its orthogonal complement, and its fixed points are the multiples of c. For Q in the stabilizer, $Q\phi(c) = \phi(Qc) = \phi(c)$, so $\phi(c)$ is a multiple of c, and $|\phi(c)| = |c|$ gives $\phi(c) = \epsilon(c)\,c$ with $\epsilon(c) = \pm 1$. The sign ϵ is continuous on $\mathbb{R}^m \setminus \{0\}$, which is connected for $m \geq 2$, hence constant. Also $\phi(0) = 0$. So $\phi = \pm\mathrm{id}$. □

Proposition D.25 (Fields with several components)**.** *Let the field have m components in the layout above.*

(i) Reparameterization freedom. *The reparameterizations that keep the architecture and leave the likelihood of every hierarchy in $\mathcal{C}^{(m)}$ unchanged are exactly the compositions of a sitewise reparameterization, which applies C^1 diffeomorphisms of $\mathbb{R}^m$ site by site, one per position class of h_ℓ and a single one for x_ℓ, and a latent reparameterization, the same map for every cell of the $3m$ latent values of the cell, which may depend on the cell's coarse m-vector and preserves γ_{3m} for each value of it, together with, at the output of the coarsest decimator, which goes to the prior entirely, any map of $\mathbb{R}^{4m}$ that preserves γ_{4m}.*

(ii) Sufficiency of the coarse value. *In an exactly fitting hierarchy, a cell and its surroundings are independent given the coarse vector.*

(iii) Identifiability of the coarse value. *At a level that is separating in dimension, any two decimators $D = (C, Z)$ and $D' = (C', Z')$ in $\mathcal{C}^{(m)}$ that fit the level satisfy $C' = \phi \circ C$ for a C^1 diffeomorphism ϕ of $\mathbb{R}^m$, and $D' \circ D^{-1}(c,\zeta) = (\phi(c), R(c,\zeta))$ with $R(c,\cdot)$ preserving γ_{3m} for every c. If D and D' are in $\mathcal{C}^{(m)}_{\mathrm{sym}}$, then ϕ commutes with $O(m)$. If moreover both coarse values have distribution γ_m, then $\phi = \pm\mathrm{id}$.*

(iv) Selection and uniqueness. *Assume (D1) and (D2) with $O(m)$ in place of the flip, and read the displacement and regularity conditions for the m-vector coarse value and the $3m$ difference coordinates. For normalized, exactly fitting hierarchies in $\mathcal{C}^{(m)}_{\mathrm{sym}}$ of finite $\mathcal{K}$ whose block maps admit straight flows, Proposition D.15 holds, with "equivariant" in its part (a) read as keeping the block maps in $\mathcal{C}^{(m)}_{\mathrm{sym}}$, and apart from its sentence about the canonical normalization. Let $\mathcal{H}^{(m)}$ be the set of such hierarchies every level of which is separating in dimension or has mutually independent cells, and whose conditional distributions ν_c are absolutely continuous. If two hierarchies in $\mathcal{H}^{(m)}$ minimize $\mathcal{K}$ over $\mathcal{H}^{(m)}$, satisfy the displacement and regularity conditions, and give every h_ℓ the same distribution, they have the same block map in every layer. Proposition D.21 holds with $\mathcal{H}_0$ the set of normalized, exactly fitting hierarchies in $\mathcal{C}^{(m)}$ whose block maps admit straight flows, and with $\mathcal{C}^{(m)}_{\mathrm{sym}}$ in place of $\mathcal{C}_{\mathrm{sym}}$. Corollary 4 holds with the class of Theorem 3 replaced by the hierarchies considered in this part and "real-analytic hierarchy" replaced by "hierarchy in $\mathcal{H}^{(m)}$".*

Proposition D.22 extends parts (i) to (iv) to several disentanglers per level as it extends the scalar results, with h_{ℓ,d_ℓ} in place of h_ℓ in the definition of $\mathcal{H}^{(m)}$.

Proof. The proofs of the scalar results apply with γ_3, γ_4 and $\mathbb{R}^4$ read as γ_{3m}, γ_{4m} and $\mathbb{R}^{4m}$, with "separating" read as separating in dimension, and with the hypothesis of Lemma D.14 in the definition of $\mathcal{H}$ replaced by the requirement on the levels in the definition of $\mathcal{H}^{(m)}$. The rest of this proof treats what changes. Several lemmas are dimension-free under this reading: Lemmas D.1 and D.3, which use only the bijectivity of the layers, the standard normal pyramid and the Cauchy–Schwarz inequality, Lemma D.5, Lemma D.7(a), Lemma D.13, Lemma D.16, Lemma D.18, and Lemma D.17 with oddness read as commuting with $I_4 \otimes Q$. Lemma D.2 holds with the flip replaced by $O(m)$: a map $I_4 \otimes Q$ commutes with every cell layer of $\mathcal{C}_{\mathrm{sym}}^{(m)}$ and with $H \otimes I_m$, so the invariance of p under $O(m)$ passes to every field of hidden variables, and the one-site distributions are $O(m)$-invariant.

(i). In Lemma D.6 the shared-site argument shows that $g(v)_i$, now a vector in $\mathbb{R}^m$, is a function of v_i alone, and the product structure makes it a C^1 diffeomorphism of $\mathbb{R}^m$ per position class.

Lemma D.8 needs a different family of densities. Part (a) now uses the products $\prod_i \rho_1(v_i)$ over sites of a density $\rho_1 = \rho^{(1)} \otimes \cdots \otimes \rho^{(m)}$ that is itself a product over components of positive smooth densities on $\mathbb{R}$. The site product $\prod_i \rho_1(v_i)$ is realized by the block map $g^{\times 4}$ with $g = (\Phi^{-1} \circ F^{(1)}, \dots, \Phi^{-1} \circ F^{(m)})$ acting componentwise. It also uses the products over aligned cells of centred normal densities on $\mathbb{R}^{4m}$, realized by a linear decimator. Perturbing one component, $\rho^{(k)} \mapsto \rho^{(k)}(1 + \vartheta\eta_\sigma)/c_\vartheta$, the argument of parts (b) and (c) shows that a map Ψ preserving all these densities has a block map e that preserves, for every k, the multiset of k-th components of the four site vectors. On the open dense set where these four numbers are distinct, the k-th components of $e(b)$ are those of b permuted by a permutation π_k, locally constant by continuity and constant by the wall argument applied to the k-th components. So e permutes positions componentwise. The cell-wise normal density with covariance $(I + \vartheta\Sigma_0) \otimes I_m$, whose block between two sites $i \neq j$ of a cell is $\vartheta(\Sigma_0)_{ij} I_m$, is then preserved only if every π_k is the identity.

Part (b) becomes the following. If the ratios of all realizable densities at v and v' agree, then each component field of v' is a permutation of the same component field of v, so v' is one of finitely many candidates. This is all that Lemma D.9 uses, since it only needs the level set $\Gamma_{\hat{c}}$ to lie in a finite union of disjoint closed sets $\{c'\} \times \mathbb{R}^{3m|\Lambda_{\ell+1}|}$. Lemma D.9 and the proof of Theorem 1 then apply with $c \in V_{\ell+1} = (\mathbb{R}^m)^{\Lambda_{\ell+1}}$ and $\zeta \in \mathbb{R}^{3m|\Lambda_{\ell+1}|}$. Every vector of $\mathbb{R}^m$ occurs as an entry c_κ, so $R_1(c, \cdot)$ preserves γ_{3m} for every c.

(ii). The derivation of (a′) from Lemma D.13 uses only that $Z(y)$ is γ_{3m} and independent of $(C(y), s)$.

(iii). Let $\Gamma = \{C = c_0\}$, the image of $\{c_0\} \times \mathbb{R}^{3m}$ under D^{-1}: a connected C^1 submanifold of dimension $3m$, all of whose points are equivalent (Lemma D.13 for D). Every y with $C'(y) \in C'(\Gamma)$ is equivalent to a point of Γ (Lemma D.13 for D'), so $C'^{-1}(C'(\Gamma))$ lies in one class.

Suppose $S = C'(\Gamma)$ has more than one point, and fix $b_0 \in S$. Since S is connected, the Lipschitz map $(b, \zeta) \mapsto (|b - b_0|, \zeta)$ carries $S \times \mathbb{R}^{3m}$ onto $I \times \mathbb{R}^{3m}$ with I an interval of positive length, so $S \times \mathbb{R}^{3m}$ has Hausdorff dimension at least $3m + 1$, Lipschitz maps not increasing dimension. Now $C'^{-1}(S) = D'^{-1}(S \times \mathbb{R}^{3m})$, and D'^{-1} is locally Lipschitz, so $C'^{-1}(S)$ has Hausdorff dimension at least $3m + 1$ as well, Hausdorff dimension being stable under countable unions (Falconer, 2003). This contradicts separation in dimension. Hence $C'(\Gamma)$ is a point, C' is constant on the level sets of C, and $C' = \phi \circ C$ with $\phi(c) = C'(D^{-1}(c, 0))$, a C^1 map.

From $\nabla C' = \nabla\phi(C)\, \nabla C$, and the rank of ∇C and $\nabla C'$ being m, $\nabla\phi$ is invertible everywhere. The map ϕ is onto because C' is. It is also injective, as follows. The level set $\{C' = c'\}$ is connected, being the image of $\{c'\} \times \mathbb{R}^{3m}$ under D'^{-1}, and equals $C^{-1}(\phi^{-1}(c'))$. The set $\phi^{-1}(c') \subset \mathbb{R}^m$ is closed, and discrete because ϕ is a local diffeomorphism, so for $c_1 \in \phi^{-1}(c')$ the set $E = \phi^{-1}(c') \setminus \{c_1\}$ is closed. Then $\{C' = c'\} = C^{-1}(c_1) \cup C^{-1}(E)$ is a union of two disjoint closed sets, both nonempty if E is, since C is onto. Connectedness forces $E = \emptyset$. So ϕ is a bijection with invertible differential, hence a C^1 diffeomorphism, and $D' \circ D^{-1}$ has the stated form with $R(c, \cdot)$ preserving γ_{3m} for every c, by the argument of the proof of (c′).

If D and D' commute with $I_4 \otimes Q$, so do their coarse values, because $H \otimes I_m$ does: $C(Qy) = QC(y)$ and $C'(Qy) = QC'(y)$, so $\phi(Qc) = Q\phi(c)$. If moreover $C(y)$ and $C'(y) = \phi(C(y))$ both have distribution γ_m, then ϕ preserves γ_m, and Lemma D.24 gives $\phi = \pm\mathrm{id}$.

(iv). Selection. In Proposition D.15(a), a sitewise reparameterization that keeps the block maps in $\mathcal{C}_{\mathrm{sym}}^{(m)}$ applies one map g of $\mathbb{R}^m$ at every site, commuting with $O(m)$, and normalization makes g preserve γ_m, so $g = \pm\mathrm{id}$ by Lemma D.24. In part (b), the decimator half compares C and $-C$ through $\|C(y) - \bar{y}\|$ and $\|-C(y) - \bar{y}\|$ with $\bar{y} \in \mathbb{R}^m$ the cell average, as before. In parts (c) and (d), Lemma D.18 is applied in $\mathbb{R}^{3m}$ and $\mathbb{R}^{4m}$, using the orthogonality of $H \otimes I_m$, and the symmetry of the selected maps comes from Lemma D.19 with A ranging over the orthogonal maps $P_\sigma \otimes I_m$ and $I_4 \otimes Q$, which preserve the input distribution of a cell by the extended Lemma D.2. In case (c), $P_\sigma \otimes I_m$ fixes $\bar{y}$ and acts on y_Δ by $\tau_\sigma \otimes I_m$, which preserves ν_c, while $I_4 \otimes Q$ maps $\bar{y}$ to $Q\bar{y}$, acts on y_Δ by $I_3 \otimes Q$, and maps ν_c to ν_{Qc}. Hence $\nabla\psi^*_{Qc} = (I_3 \otimes Q)\,\nabla\psi^*_c\big((I_3 \otimes Q)^{-1}\,\cdot\big)$, and the selected maps commute with both actions.

Uniqueness. In the proof of Theorem 3, every level is separating in dimension or has mutually independent cells, by the definition of $\mathcal{H}^{(m)}$, so every level falls under one of the two decimator cases. The disentangler case uses Lemmas D.17 and D.16 only. At a level that is separating in dimension, part (iii) gives $D' = (\phi \circ C, R(C, Z))$ with $\phi = \pm\mathrm{id}$, the coarse fields being normalized. The sign is excluded by the displacement condition as before, since that computation compares C and $-C$ only, and the latent part is fixed by Proposition D.15(c). At a level whose cells are independent, and at the coarsest level, the argument is that of the scalar case with γ_{4m}. The absolute continuity of ν_c for the selected decimator follows from the invertibility of the lower right $3m \times 3m$ block of the Jacobian of A^{-1}.

Symmetric minimizers, weights and several disentanglers. Proposition D.21 uses conjugation by the orthogonal maps $P_\sigma \otimes I_m$ and $I_4 \otimes Q$. These intertwine the split with orthogonal maps of the coarse field and of the latent coordinates, $S_\ell \circ g = (g' \times (\tau_\sigma \otimes I_m)) \circ S_\ell$ and $S_\ell \circ Q = (Q \times (I_3 \otimes Q)) \circ S_\ell$, and preserve displacements and straight paths. The proof of Corollary 4 concerns the layer energies only. Proposition D.22 concerns the partition on which each layer acts and is unaffected by the number of components. □

E Summary of the theory and its limitations

Table E.1 lists how two exactly fitting hierarchies can differ, what removes each difference, and under which condition. For the experiments the selection means that, for hierarchies that fit exactly, minimize the kinetic energy and satisfy the two conditions, a decimator's block map in canonical normalization is determined by the distribution of the field h_ℓ it acts on, so two such models whose fields h_ℓ agree in distribution have the same fingerprint at that level, and that p itself fixes these distributions is the content of Corollary 4. The remaining paragraphs state what the results assume and what they do not cover.

Analyticity and the regularity condition are assumed, not derived, as is separation in dimension for fields with several components (Appendix D.9), and a trained hierarchy need not satisfy the factorization hypothesis of Lemma D.14 at every level. The displacement condition can be verified on a trained model only in part (Appendix D.5.2).

What is assumed about the distributions of the hidden variables is collected in Appendix D.7, and the corollary covers almost every layer weighting, not necessarily the uniform one used in training. Theorems 2 and 3 concern exactly fitting hierarchies, and we have no result that converts a small divergence of a trained model into closeness of its coarse level sets, which would require a quantitative form of the separating property. Finally, the results on kinetic energy concern block maps that admit straight flows, for which the least kinetic energy is the displacement. The kinetic energy of a trained flow bounds this value from above. The selection results compare normalized hierarchies, whereas training enforces normalization only through the penalty $\mathcal{M}$, and the canonical normalization applied afterwards changes the displacements.

Table E.1: How two exactly fitting hierarchies can differ, what removes each difference, and under which condition.

Difference	Removed by	Condition
Units of a hidden variable (monotone map)	$\mathcal{M}$ or canonical normalization (Prop. D.15(a))	equivariant block maps; up to a sign
Sign, or a symmetry of a disentangler's input	kinetic energy (Prop. D.15(b))	displacement condition
Latent values	kinetic energy (Prop. D.15(c))	regularity condition
What the decimator keeps, separating level	likelihood (Thm. D.12(c′))	exact fit; up to a monotone map of $\mathbb{R}$, which normalization (first row) reduces to a sign and the kinetic energy (second row) fixes
What the decimator keeps, independent cells and coarsest level	kinetic energy (Prop. D.15(d))	regularity condition; among decimators that fit the level with normalized coarse values
Mirror image under a lattice symmetry	equivariant block maps, or uniqueness (Prop. D.21)	(D2); by construction, or if $\mathcal{K}$ has a unique minimizer
Distributions of the h_ℓ	layer weights and the no-tie assumption (Cor. 4)	almost every positive weighting, not necessarily the uniform one