Learning Renormalization-Group-Like Latent Variables with a Hierarchical Graph Neural Network Variational Autoencoder for the 2D Ising Model

Tingyu Meng[1]
[1] *University of Wisconsin–Madison*
(Dated: December 18, 2025)

We present a self-supervised Graph Neural Network variational autoencoder (GNN-VAE) designed to learn renormalization-group (RG) structure from Monte Carlo configurations of the two-dimensional Ising model. The key inductive bias is an RG consistency loss that encourages a learned latent variable to be approximately invariant under repeated coarse graining. We show that the resulting latent representation correlates strongly with standard physical observables (magnetization and energy density), exhibits two stable fixed points corresponding to the ordered and disordered phases, and supports an empirical "RG map" whose dominant eigen-direction aligns with the leading principal component of the latent space. These findings provide evidence that the model learns RG-like variables beyond mere phase classification.

I. INTRODUCTION

The renormalization group (RG) is one of the most powerful conceptual frameworks in theoretical physics, providing a systematic way to understand how physical systems behave across different length scales [1, 2]. For the 2D Ising model, RG theory predicts:

- Two stable fixed points: the paramagnetic (high-temperature) and ferromagnetic (low-temperature) phases
- One unstable fixed point at the critical temperature $K_c \approx 0.4407$ [3]
- A relevant direction (temperature) that determines which fixed point the system flows to

The central question we address is: **Can a neural network learn these RG structures without explicit supervision?**

Machine-learning approaches to the Ising model have shown that unsupervised methods can detect phase structure from raw configurations [4, 5], and recent work has also focused on making such solutions explainable [6]. Here we focus on a generative, self-supervised setting and add an explicit scale-consistency inductive bias to encourage learning RG-like variables.

In the Wilsonian picture of RG, one integrates out short-wavelength (high-momentum) modes to obtain an effective free energy

$$F'[\phi] = F[\phi^-; a', b', c', \dots]$$

with a new set of couplings $(a', b', c', \dots)$ that depend on the coarse-graining scale. The evolution

$$(a, b, c, \dots) \longrightarrow (a', b', c', \dots)$$

defines an RG flow in coupling space, with relevant directions growing under coarse-graining and irrelevant directions shrinking. In this project we ask a concrete version of the question posed in *"Can a neural network act as a renormalization group?"*: can a GNN-VAE, equipped with a physics-motivated inductive bias, learn both the effective couplings and their RG flow directly from Monte Carlo configurations of the 2D Ising model?

II. RG BACKGROUND AND NEURAL NETWORK ANALOGY

A. Wilsonian RG in a Nutshell

In field theory language, the partition function of a scalar field ϕ with a local action

$$F[\phi] = \int d^d x \left[\frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}\mu^2\phi^2 + g\phi^4 + \dots\right] \tag{1}$$

is

$$Z = \int \mathcal{D}\phi\, e^{-F[\phi]}. \tag{2}$$

Introducing a momentum cutoff $\Lambda \sim 1/a$ and separating modes into "slow" ($|\mathbf{k}| < \Lambda'$) and "fast" ($\Lambda' < |\mathbf{k}| < \Lambda$), the Wilsonian RG proceeds by integrating out the fast modes:

$$Z = \int \mathcal{D}\phi^- e^{-F_0[\phi^-]} \int \mathcal{D}\phi^+ e^{-F_0[\phi^+]} e^{-F_I[\phi^-,\phi^+]} = \int \mathcal{D}\phi^- e^{-F'[\phi^-;a',} \tag{3}$$

thereby generating a new effective action with renormalized couplings $(a', b', c', \dots)$. Repeating this procedure defines a flow in the space of couplings, with fixed points corresponding to scale-invariant theories.

The same idea can be illustrated with a simple zero-dimensional example

$$Z(a, b) = \int dx \int dy\, e^{-a(x^2+y^2)} e^{-b(x+y)^4}, \tag{4}$$

where integrating out y produces an effective single-variable theory

$$\int dy\, e^{-a(x^2+y^2)} e^{-b(x+y)^4} = e^{-a'x^2 - b'x^4 - c'x^6 - \dots}. \tag{5}$$

The transformation $(a, b) \to (a', b', c', \dots)$ is again an RG step in coupling space; relevant couplings grow, irrelevant ones shrink, and their flow organizes the phase structure.

B. Neural Network Analogy

The above RG picture has a natural analogy in deep learning. A feed-forward network can be viewed as a sequence of coarse-grainings:

$$x^{(0)} \to x^{(1)} = f_1(x^{(0)}; \theta_1) \to x^{(2)} = f_2(x^{(1)}; \theta_2) \to \cdots, \tag{6}$$

where each layer f_ℓ discards some information while preserving features relevant for the final task. In convolutional or graph-based architectures with pooling, early layers focus on local patterns, while deeper layers encode increasingly coarse, global summaries. This mirrors the RG idea of integrating out microscopic degrees of freedom to obtain effective long-distance variables.

In our GNN-VAE:

- The *tokenizer* enriches spins with local interaction information, analogous to constructing local energy densities.
- Each GCN + pooling block plays the role of a block-spin transformation, reducing the number of sites while propagating information to a coarser lattice.
- The latent variables $z^{(l)}$ at different levels are neural analogues of effective couplings at different RG scales.

The physics consistency loss

$$\mathcal{L}_{\text{RG}} = \frac{1}{L-1} \sum_{l=0}^{L-2} \|z^{(l)} - z^{(l+1)}\|^2 \tag{7}$$

then plays the role of an RG prior: it encodes the belief that true RG variables should be approximately invariant under coarse-graining. Rather than hand-coding the RG map, we let the network learn a representation in which

$$z_{\text{fine}} \approx z_{\text{coarse}} \approx z_{\text{coarse}^2}, \tag{8}$$

and subsequently verify that these learned variables behave as expected from RG theory: they organize the phase diagram, identify fixed points, and align with the relevant eigen-directions of an explicitly fitted linear RG map.

III. METHOD

A. Model Architecture

We employ a hierarchical GNN-VAE with the following structure:

1. **Tokenizer**: Maps each spin s_i and its local field $h_i = \sum_{j \in \text{neighbors}} s_i s_j$ to a d-dimensional token via an MLP
2. **Hierarchical RG Encoder**: Multiple RG blocks, each consisting of:
   - GCN message passing (local information aggregation)
   - Graph coarsening via Graclus pooling (coarse-graining)
3. **Latent Space**: Each level outputs $\mu^{(l)}, \sigma^{(l)}$ via linear projections, then samples $z^{(l)} = \mu^{(l)} + \sigma^{(l)} \cdot \epsilon$
4. **MLP Decoder**: Reconstructs spin configuration from the final latent z

B. Loss Function

The total loss combines three terms:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \beta \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{RG}} \cdot \mathcal{L}_{\text{RG}} \tag{9}$$

where:

- $\mathcal{L}_{\text{recon}}$: reconstruction loss (baseline: mean-squared error on spins)
- $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q(z|x) \| \mathcal{N}(0, I))$: KL divergence
- $\mathcal{L}_{\text{RG}} = \frac{1}{L-1} \sum_{l=0}^{L-2} \|z^{(l)} - z^{(l+1)}\|^2$: RG consistency loss

The RG consistency loss is the key innovation: it enforces that the latent representation should be **scale-invariant**, i.e., the same physical state should map to the same z regardless of the observation scale.

C. Dataset and Graph Construction

We generate Monte Carlo configurations of the 2D nearest-neighbor ferromagnetic Ising model on an $L \times L$ square lattice with periodic boundary conditions, using a Metropolis update. The coupling is $K = J/T$ (we set $k_B = 1$). Unless otherwise stated, we use $L = 16$ and sample 26 couplings spanning both phases and the critical region near $K_c \simeq 0.4407$.

Each configuration is represented as an undirected graph with $N = L^2$ nodes. Node features are spins $s_i \in \{-1, +1\}$, and edges connect nearest neighbors on the lattice. For analysis we also compute physical observables directly from configurations:

$$|m| = \left| \frac{1}{N} \sum_i s_i \right|, \qquad e = -\frac{1}{N} \sum_{\langle ij \rangle} s_i s_j, \tag{10}$$

where $\langle ij \rangle$ denotes nearest neighbors (counted once).

To encourage the model to "see" local energetics, we introduce a tokenizer that maps each spin and a local interaction feature

$$h_i = \sum_{j \in \text{n.n.}(i)} s_i s_j \tag{11}$$

to a learned token embedding via a small MLP.

D. Training Protocol and Hyperparameters

We train using Adam with learning rate 10^{-3} for 60 epochs and batch size 32. The encoder uses two RG steps (coarsening $16 \to 8 \to 4$), hidden dimension 64, token dimension 16, and latent dimension 4 (unless otherwise noted). The loss weights in the baseline run are $\beta = 0.1$ and $\lambda_{\text{RG}} = 0.5$.

We emphasize that for binary spin data, a Bernoulli likelihood (binary cross entropy on logits) is often better than MSE.).

IV. RESULTS

A. The latent space captures physical information

Figure 2 demonstrates that PC1 is not an abstract mathematical quantity but has clear physical meaning:

- PC1 increases monotonically with K (inverse temperature)
- PC1 correlates strongly with magnetization $|m|$
- The sharpest change occurs near the critical point K_c

Physical interpretation: PC1 corresponds to the **relevant direction** in RG theory—it encodes the "distance from criticality" or equivalently the degree of order in the system, as confirmed by its strong correlation with both magnetization and energy density.

B. Scale invariance of the latent representation

Figure 4 reveals a striking pattern:

- **Ordered phase** ($K > K_c$): $\|z^{(l)} - z^{(l+1)}\| \to 0$ — the latent representation is nearly identical across scales
- **Disordered phase** ($K < K_c$): larger differences persist

Physical interpretation: The ordered phase has already "flowed" to the ferromagnetic fixed point, where the system is scale-invariant. The disordered phase is still flowing toward the paramagnetic fixed point.

C. Two fixed points in latent space

Figure 5 shows that:

- The two phases remain clearly separated at all scales
- Each phase clusters around its own "fixed point" (marked by stars)
- This structure persists from $L = 16$ to $L = 4$

Physical interpretation: The model has learned the two **stable RG fixed points**:

- Blue star: Paramagnetic fixed point ($K^* = 0$, completely disordered)
- Red star: Ferromagnetic fixed point ($K^* = \infty$, completely ordered)

D. RG flow trajectories

Figure 6 visualizes the RG flow:

- **Red trajectories** ($K > K_c$): Short arrows, already near the ferromagnetic fixed point
- **Blue trajectories** ($K < K_c$): Longer arrows, still flowing toward the paramagnetic fixed point
- The flow direction is always from fine to coarse scale (the essence of RG)

E. RG flow diagram in $(K, \text{PC1})$ space

In Figure 7, each arrow shows how a single configuration moves in PC1 as it is coarse-grained, at fixed K. For $K < K_c$, arrows flow toward the paramagnetic fixed point cluster at negative PC1; for $K > K_c$, they flow toward the ferromagnetic fixed point cluster at positive PC1. The region near K_c shows the largest "velocity" in PC1, consistent with an unstable critical fixed point separating the two basins of attraction.

F. Linear RG map and eigenvalue spectrum

We fit a linear RG transformation $\mu' = W\mu + b$ and analyze its properties:

- **Left panel**: Fine (circles) and Coarse (crosses) latents overlap well, confirming $z_{\text{fine}} \approx z_{\text{coarse}}$
- **Middle panel**: Points lie along $y = x$, indicating the linear map accurately predicts coarse latents

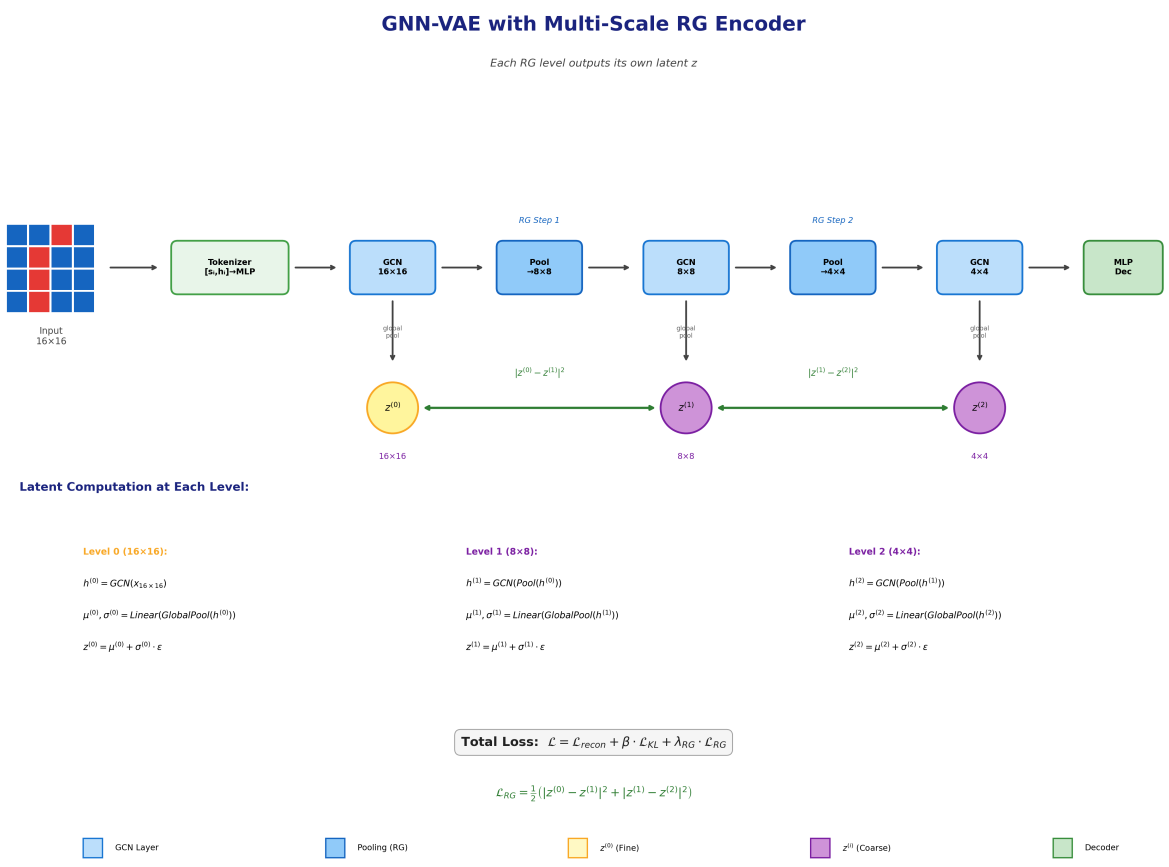
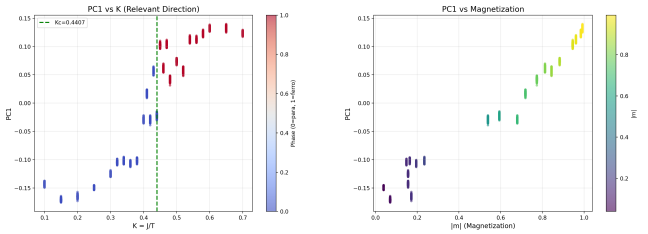
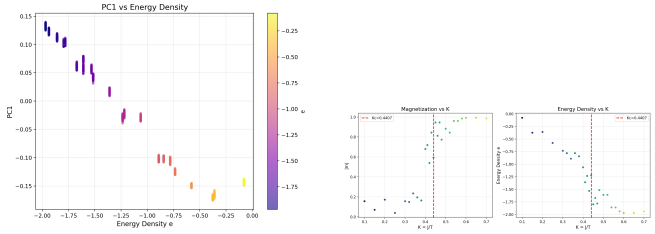


FIG. 1: Architecture of the GNN-VAE with hierarchical RG encoder and multi-scale latent outputs.



(a) PC1 vs coupling K (b) PC1 vs magnetization $|m|$

FIG. 2: The first principal component (PC1) of the latent space shows strong correlation with physical quantities: (a) the coupling constant $K = J/kT$, and (b) the magnetization $|m|$. The vertical dashed line marks the critical point $K_c \approx 0.44$.



(a) PC1 vs energy density e (b) Magnetization and energy vs K

FIG. 3: Relation between the latent PC1 and traditional observables. PC1 varies smoothly with energy density and mirrors the behavior of magnetization and energy across the phase transition.

- **Right panel**: The fitted map exhibits two near-degenerate eigenvalues with $|\lambda| \gtrsim 1$ and a set of eigenvalues with $|\lambda| < 1$, consistent with the existence of a low-dimensional "relevant" subspace and multiple irrelevant directions.

Physical interpretation:

- Eigen-directions with $|\lambda| > 1$ correspond to **relevant** perturbations of the coarse-grained description; $|\lambda| < 1$ are **irrelevant** directions that decay under coarse-graining.
- Directions with $|\lambda| < 1$ are **irrelevant**—they decay under RG transformation
- In our baseline run, the eigenvalue magnitudes are

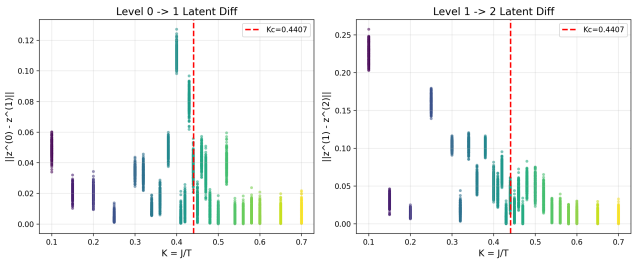


FIG. 4: Latent difference $\|z^{(l)} - z^{(l+1)}\|$ between adjacent coarse-graining levels as a function of K. In the ordered phase ($K > K_c$), the difference approaches zero, indicating convergence to a fixed point.

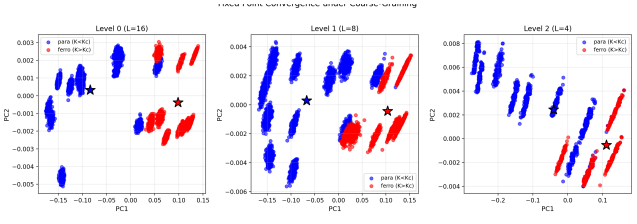


FIG. 5: Latent space at different coarse-graining levels. Blue points: paramagnetic samples ($K < K_c$); Red points: ferromagnetic samples ($K > K_c$). Stars mark the centroid of each phase (approximate fixed points).

approximately $|\lambda| = \{1.17, 0.76, 0.74, 0.6\}$ when excluding data near criticality ($|K - K_c| > 0.1$). We interpret only one$|\lambda| > 1$ as a empirical feature of learned relevant direction for RG process.

G. Model diagnostics: reconstruction and posterior utilization

Because our goal is interpretability of the learned RG-like representation, it is important to also report standard VAE diagnostics that measure whether the model makes effective use of the latent variables.

To keep the paper scientifically precise, we therefore distinguish between two questions: (i) whether the learned latent *organizes RG structure* (supported by Figs. 2–8), and (ii) whether the model is a high-fidelity *generative model* of Ising configurations (which the baseline does not yet fully achieve). Addressing (ii) is nonetheless important and motivates improved likelihoods and training schedules .

V. RELATION TO PREVIOUS WORK

Our approach is closely related to, but distinct from, several earlier attempts to connect RG and machine learning.

Koch-Janusz and Ringel [7] proposed to learn RG

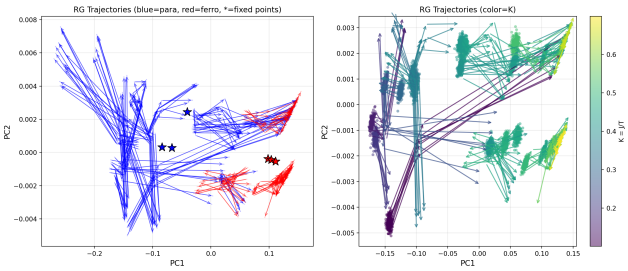


FIG. 6: RG flow trajectories in latent space. Arrows indicate the direction of coarse-graining (fine → coarse). Left: colored by phase; Right: colored by K value.

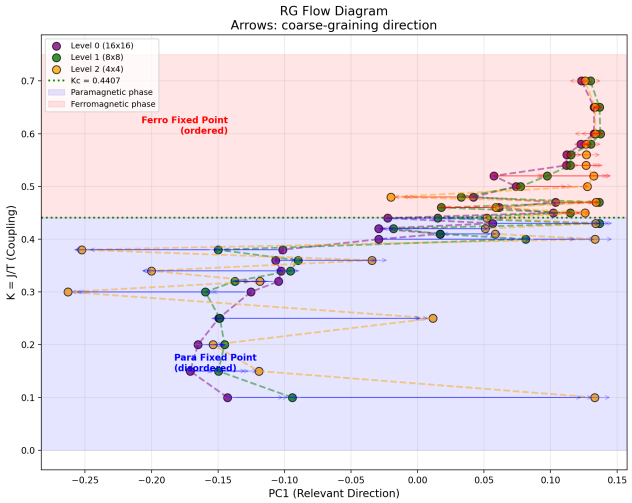


FIG. 7: RG flow diagram in the plane of coupling K and latent PC1. Each horizontal arrow connects the same configuration at different coarse-graining levels ($16\times16 \to 8\times8 \to 4\times4$). The blue (red) shaded regions indicate the paramagnetic (ferromagnetic) phase, and the dashed line marks the critical coupling $K_c \approx 0.4407$.

transformations by maximizing mutual information between inner and outer regions of a spin block, using neural networks as variational ansätze for the coarse-graining rule. Their work demonstrated that information-theoretic optimality naturally leads to block-spin rules that resemble real-space RG. In contrast, we do not explicitly optimize mutual information; instead, we enforce an RG-style *consistency loss* in the latent space and show that the resulting representation exhibits the expected fixed points and relevant directions.

Wang [4] showed that unsupervised methods such as PCA and clustering can detect phase transitions directly from raw spin configurations, even without explicit RG structure. Our results can be viewed as a representation-learning analogue of this idea: we first learn a latent embedding with a physics-informed GNN-VAE, and then perform PCA in the latent space. The fact that the leading principal component aligns with the relevant eigen-

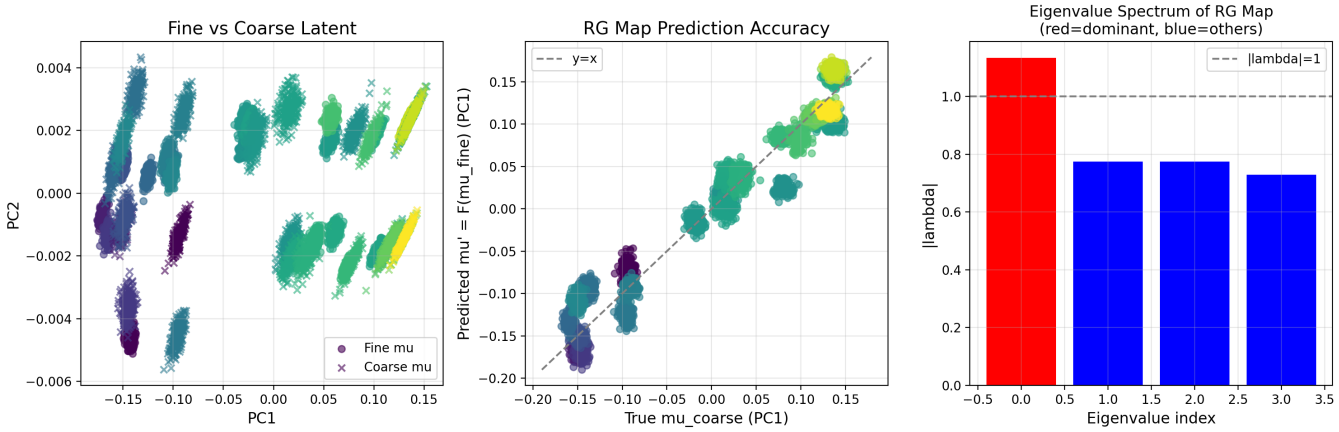


FIG. 8: Analysis of the learned RG map $\mu_{\text{coarse}} \approx W\mu_{\text{fine}} + b$. Left: Fine vs Coarse latent in PCA space. Middle: Prediction accuracy of the linear map. Right: Eigenvalue spectrum of W(exclude phase transition temperature).

direction of a learned RG map suggests that the network has discovered a non-trivial coarse-grained description rather than a purely geometric separation in configuration space.

Finally, from the perspective of field-theoretic RG expositions such as Shankar [8], our linear RG map analysis provides an empirical counterpart to the linearization of RG flows near fixed points. The eigenvalues of the fitted map W play the role of scaling factors b^{y_i} for different operators, with $|\lambda_i| > 1$ corresponding to relevant perturbations and $|\lambda_i| < 1$ to irrelevant ones. This offers a concrete example where such eigen-directions are reconstructed numerically from data by a neural network.

VI. DISCUSSION: BEYOND CLASSIFICATION

A natural question arises: *Is the model just learning to classify phases?*

We argue that the GNN-VAE learns genuine RG structure, not mere classification:

First, the **physics consistency loss** $\mathcal{L}_{\text{RG}}$ acts as an explicit inductive bias, encoding the expectation that true RG variables should be invariant under coarse-graining. This bias is inspired by the Wilsonian view that integrating out short-distance degrees of freedom should leave the effective long-distance couplings unchanged, up to flow along relevant directions. In our implementation, enforcing $z^{(l)} \approx z^{(l+1)}$ across scales encourages the latent representation to discard microscopic noise and retain only scale-robust information.

Second, even with this bias, it is *not* guaranteed that the emergent latent coordinate that best separates phases must align with the RG relevant direction. The fact that the PCA direction with largest variance (PC1) also coincides with the dominant eigen-direction of the learned RG map (Section 8) is a non-trivial emergent property of training on Ising data with this loss.

The key evidence that we learn RG (not just classification):

1. **Scale invariance**: z is approximately the same at 16×16, 8×8, and 4×4 scales
2. **Physical interpretability**: PC1 correlates with $|m|$, K, and energy density
3. **Correct fixed point structure**: Two stable fixed points connected by the relevant direction
4. **Unsupervised learning**: The model never sees phase labels during training

VII. CONCLUSION

We have demonstrated that a GNN-VAE with RG consistency loss can learn the Renormalization Group structure of the 2D Ising model:

- The latent space PC1 corresponds to the **relevant direction** (temperature/order)
- The model discovers **two stable fixed points** (paramagnetic and ferromagnetic)
- The latent representation is **scale-invariant**, especially in the ordered phase
- A fitted linear RG map reveals a low-dimensional relevant subspace whose dominant direction aligns with PC1

This work shows that neural networks can discover deep physical principles—in this case, the RG structure—without explicit supervision, opening possibilities for applying similar techniques to systems where the RG structure is unknown.

VIII. LIMITATIONS AND FUTURE WORK

Our baseline model demonstrates clear RG-like organization in latent space but also exhibits imperfect reconstruction and partial posterior under-utilization. We view these as opportunities for systematic improvement. Promising directions include:

- **Bernoulli likelihood for spins**: replace MSE with BCE on logits to better match binary data.
- **KL scheduling and regularization**: KL annealing and/or free-bits to encourage robust latent usage.
- **Decoder inductive bias**: a graph-structured decoder may improve reconstruction without sacrificing interpretability.
- **Oversmoothing and deeper GNNs**: deeper GCN stacks can oversmooth node features; spectral filter designs (e.g., Bernstein polynomial filters) and residual connections are plausible remedies.
- **More faithful RG linearization**: estimate W in restricted neighborhoods (e.g., near fixed points) and quantify uncertainties to connect more directly to scaling exponents.

[1] K. G. Wilson, Phys. Rev. B **4**, 3174 (1971).
[2] L. P. Kadanoff, Physics **2**, 263 (1966).
[3] L. Onsager, Phys. Rev. **65**, 117 (1944).
[4] L. Wang, Phys. Rev. B **94**, 195105 (2016).
[5] S. J. Wetzel, Phys. Rev. E **96**, 022140 (2017).
[6] R. C. Alamino, arXiv preprint arXiv:2402.11701 (2024).
[7] M. Koch-Janusz and Z. Ringel, Nature Physics **14**, 578 (2018).
[8] R. Shankar, *Quantum Field Theory and Condensed Matter: An Introduction* (Cambridge University Press, 2017).