Learning Renormalization-Group-Like Latent Variables with a Hierarchical GNN-VAE for the 2D Ising Model (Extended Abstract)

Tingyu Meng
University of Wisconsin–Madison

Abstract

We study whether a generative, self-supervised neural network can learn *renormalization-group (RG)-like* variables from raw Monte Carlo configurations of the two-dimensional Ising model. Our model is a hierarchical graph neural network variational autoencoder (GNN-VAE) equipped with an *RG consistency* loss that encourages the latent representation to remain stable under repeated coarse-graining. We find that the learned latent space (i) correlates with magnetization and energy density, (ii) exhibits two basins consistent with ordered/disordered fixed points, and (iii) supports an empirical linear "RG map" whose dominant eigen-direction aligns with the leading principal component (PC1) of the latent space. These observations suggest that the network learns an interpretable, scale-robust coordinate beyond mere phase separation.

1 Motivation

Machine-learning approaches to the Ising model have shown that unsupervised methods can detect phase structure from raw configurations [1, 2], and recent work has also focused on making such solutions explainable [3]. 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?

2 Model and Training

We generate 2D Ising configurations on an $L \times L$ square lattice ($L = 16$) with periodic boundary conditions using Metropolis updates. Each configuration is encoded as a lattice graph with nearest-neighbor edges. A tokenizer augments each spin $s_i \in \{-1, +1\}$ with a local interaction feature $h_i = \sum_{j \in \text{n.n.}(i)} s_i s_j$ before message passing.

Our encoder is hierarchical: repeated GCN+pooling blocks perform coarse-graining $16 \to 8 \to 4$ and output latent parameters $(\mu^{(\ell)}, \log \sigma^{2(\ell)})$ at each level. A graph-level latent is sampled via the reparameterization trick $z^{(\ell)} = \mu^{(\ell)} + \sigma^{(\ell)}\epsilon$, and an MLP decoder reconstructs the spin configuration from the final latent.

The objective combines reconstruction, KL regularization, and an RG consistency term:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{recon}} + \beta\, D_{\text{KL}}(q(z|x) \,\|\, \mathcal{N}(0, I)) + \lambda_{\text{RG}} \mathcal{L}_{\text{RG}}, \\ \mathcal{L}_{\text{RG}} &= \frac{1}{L_s - 1} \sum_{\ell} \| r^{(\ell)} - r^{(\ell+1)} \|^2, \end{aligned} \tag{1}$$

where $r^{(\ell)}$ is a latent representation at level ℓ (in practice we use $\mu^{(\ell)}$ for stability) and L_s is the number of scales.

3 Key Results

Latent interpretability. Performing PCA on the learned latent means, PC1 varies systematically with the coupling K and correlates with standard observables (magnetization and energy density), indicating that the dominant variance direction is physically meaningful (Fig. 1).

RG-like flow and fixed points. Tracking the same configuration under repeated coarse-graining in latent space reveals trajectories that flow toward two distinct basins, consistent with ordered/disordered fixed points. *See Appendix Fig. 3 for a visualization of the learned flow.* **Empirical RG map.** We fit a linear map between latents at adjacent scales, $\mu_{\text{coarse}} \approx W \mu_{\text{fine}} + b$, and analyze the spectrum of W. The dominant eigen-direction aligns closely with PC1 (cosine similarity ≈ 1 in our runs), supporting the interpretation of PC1 as the learned "relevant" direction. Excluding samples near the critical region stabilizes the fitted spectrum, highlighting

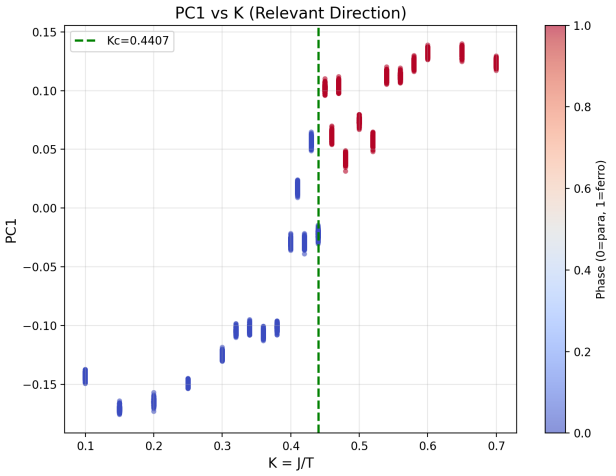

Figure 1: First principal component (PC1) of the latent mean μ versus coupling $K = J/T$. PC1 changes most rapidly near $K_c \simeq 0.4407$, consistent with a relevant thermal direction organizing the two phases.

the sensitivity of linearization near criticality. **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.

- In our baseline run, the eigenvalue magnitudes are 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.

4 Discussion and Deliverables

Our results provide multiple, complementary signatures of RG-like organization in a learned latent space: scale-consistent representations, physically interpretable latent axes, and agreement between PCA and a fitted linear RG map. Remaining work toward a publication-quality result includes controlled ablations (e.g., $\lambda_{\mathrm{RG}} = 0$) and likelihood choices appropriate for binary spins (Bernoulli/BCE instead of MSE).

Code and repository link. Please upload your GitHub URL here: `<YOUR_GITHUB_REPO_URL>`. The analysis figures are generated by `analyze_all.py` and saved in `analysis_plots/`. The implementation uses PyTorch and PyTorch Geometric [4, 5].

References

References

[1] Lei Wang. Discovering phase transitions with unsupervised learning. *Phys. Rev. B*, 94:195105, 2016.

[2] Sebastian J. Wetzel. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E*, 96:022140, 2017.

[3] R. C. Alamino. Explaining the machine learning solution of the ising model. *arXiv preprint arXiv:2402.11701*, 2024.

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[5] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Appendix: Mathematical details and additional results

Note: Appendix pages do not count toward the 2-page extended abstract limit.

A. Variational autoencoder (VAE) objective

Given an input configuration x (a spin configuration encoded as a graph), the encoder produces a diagonal Gaussian posterior

$$q_\phi(z|x) = \mathcal{N}\big(z; \mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x))\big)\,. \tag{2}$$

We sample latents via the reparameterization trick

$$z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

The KL term is

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(q_\phi(z|x)\,\|\,\mathcal{N}(0,I)) = -\frac{1}{2}\sum_{i=1}^{d}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right), \tag{4}$$

where d is the latent dimension.

B. Coarse-graining and RG consistency

We construct a 2×2 block-spin coarse-graining using a majority rule:

$$s'_I = \mathrm{sign}\left(\sum_{i \in \mathrm{block}(I)} s_i\right), \tag{5}$$

with ties assigned to $+1$ in our implementation.

At multiple scales ℓ (e.g., $16 \to 8 \to 4$), the hierarchical encoder outputs $(\mu^{(\ell)}, \log\sigma^{2(\ell)})$ and we enforce scale-consistency through

$$\mathcal{L}_{\mathrm{RG}} = \frac{1}{L_s - 1}\sum_{\ell=0}^{L_s-2}\left\|r^{(\ell)} - r^{(\ell+1)}\right\|^2, \tag{6}$$

where $r^{(\ell)}$ is a chosen representation (we use $\mu^{(\ell)}$ for stability) and L_s is the number of scales.

C. Linear RG map and eigenvalue spectrum

To probe the learned coarse-graining in latent space, we fit a linear map

$$\mu_{\mathrm{coarse}} \approx W\,\mu_{\mathrm{fine}} + b \tag{7}$$

by least-squares regression. The eigenvalues of W provide an empirical linearization: directions with $|\lambda| > 1$ behave as relevant perturbations (grow under coarse-graining), while $|\lambda| < 1$ correspond to irrelevant directions (shrink).

We also compute the alignment between the dominant eigenvector of W and the first principal component (PC1) direction of the latent space via cosine similarity.

D. Figures and additional results

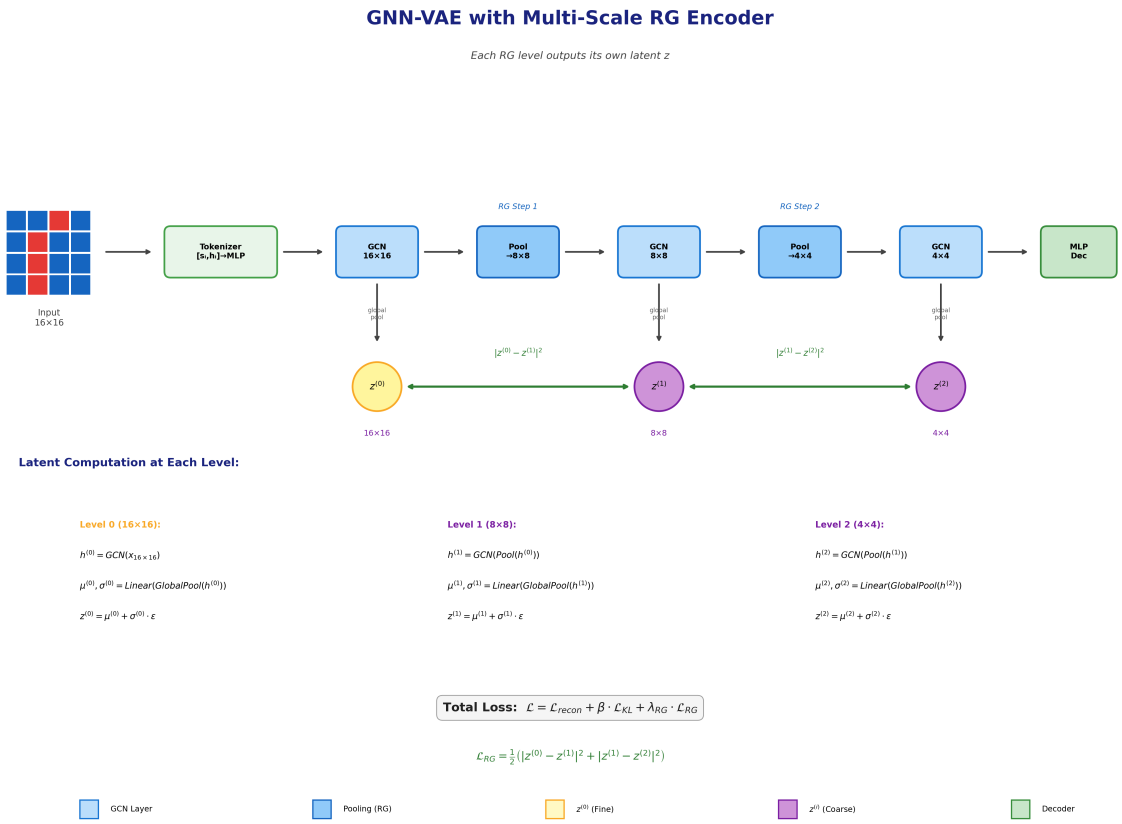

Figure 2: Architecture of the hierarchical GNN-VAE with multi-scale latents and RG consistency loss.

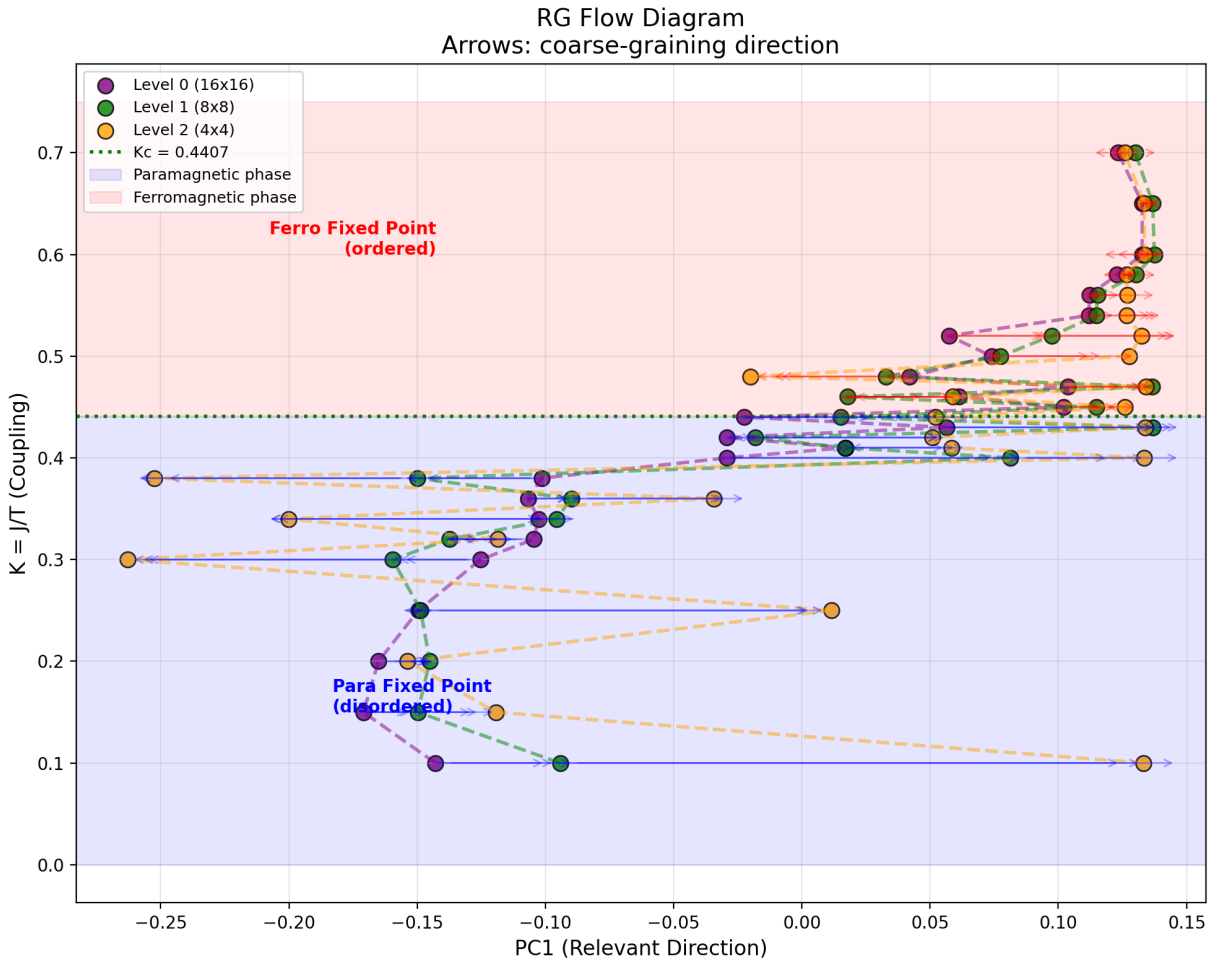

Figure 3: RG flow diagram in $(K, \mathrm{PC1})$ space. Horizontal arrows connect the same configuration across coarse-graining levels (fine→coarse).

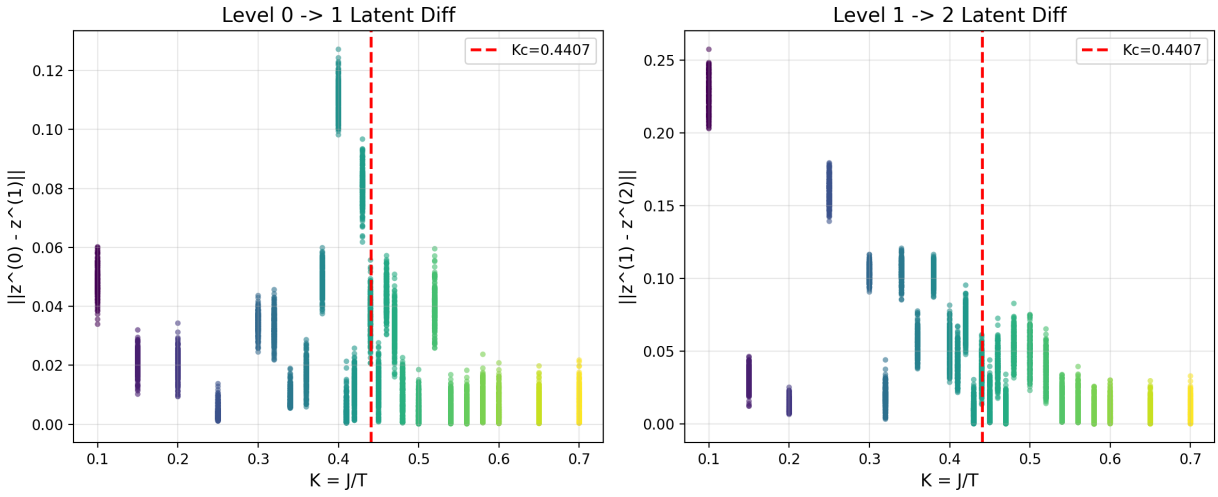

Figure 4: Multi-scale latent consistency $\|r^{(\ell)} - r^{(\ell+1)}\|$ versus coupling K.

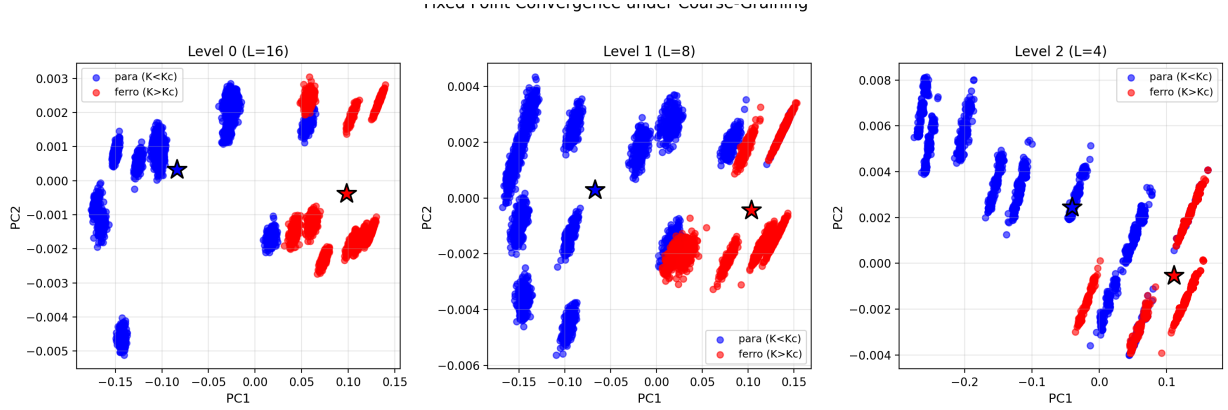

Figure 5: Fixed point convergence in latent PCA space across scales (phase-colored).

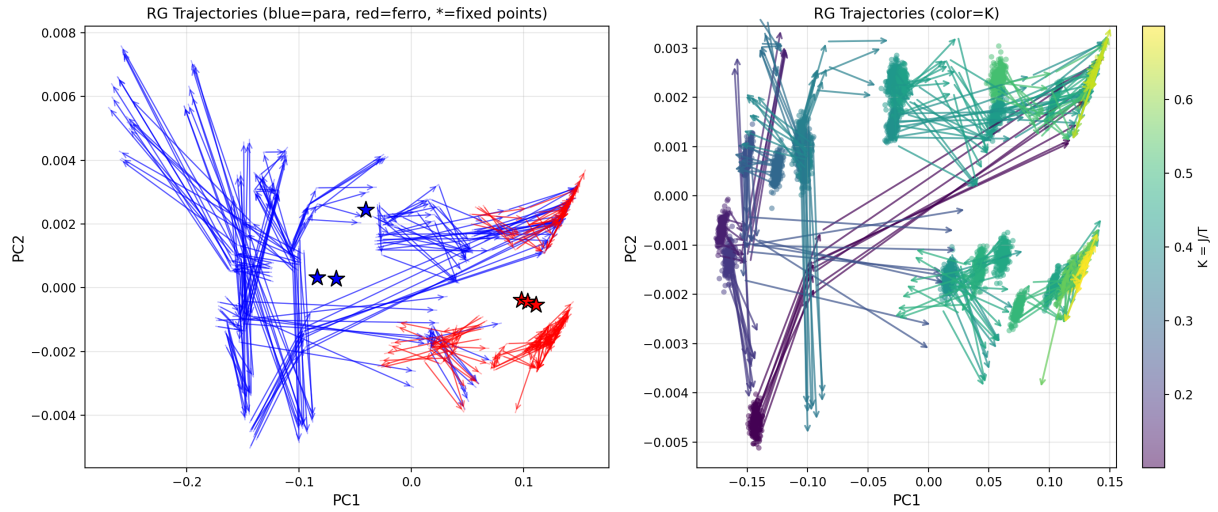

Figure 6: Latent RG trajectories under repeated coarse-graining (fine→coarse), shown in PCA space.

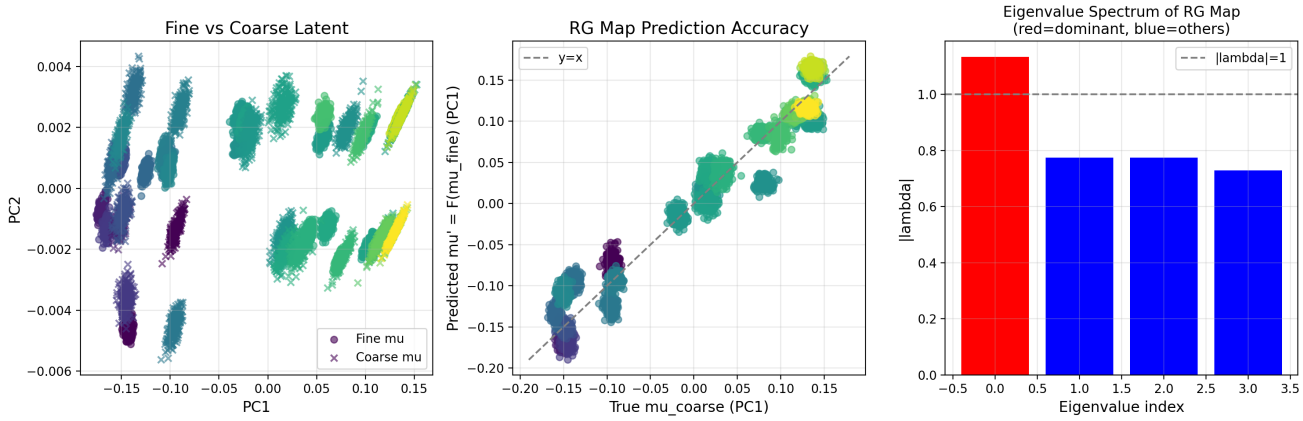

Figure 7: RG map analysis: fine vs coarse latent overlap, linear map accuracy, and eigenvalue spectrum.