Supplemental Material:
Discovering and decoding latent mean-field structure with variational autoencoders

Marco Biroli,[1] Max Welling,[2,3] and Vincenzo Vitelli[1,4]

[1]*Department of Physics and the James Franck Institute, University of Chicago, Chicago, IL, USA*
[2]*CuspAI, Cambridge, England*
[3]*AMLab, University of Amsterdam, Amsterdam, Netherlands*
[4]*Leinweber Institute for Theoretical Physics*

(Dated: July 11, 2026)

This Supplemental Material collects material that complements the main text. References of the form Eq. (N), Fig. N or Sec. N without an "S" point to the main text; references of the form Eq. (SN), Fig. SN or Sec. SN point inside this document. Section S1 provides the geometric picture behind the capacity bound derived in the Methods and shows why the singular-prior loophole is unreachable by training. Section S2 collects the thermodynamic observables that the VAE reproduces for each solvable model, and Sec. S3 the diagnostics quantifying the recovery of the microscopic mean-field parameters; both complement the summary figures of the main text.

S1. CAPACITY OF THE LATENT CHANNEL: GEOMETRY AND THE INACCESSIBILITY OF SINGULAR PRIORS

The Methods derive the capacity criterion $d\log(1/\sigma) \gtrsim I_{\mathrm{bip}}(p)$ [Eq. (27)] by matching the message-passing bound $I(\mathbf{X};\mathbf{Z}) \geq I_{\mathrm{bip}}(p)$ [Eq. (24)] to the sphere-packing rate $R \simeq d\log(1/\sigma)$ [Eq. (55)] of a Gaussian-prior VAE. This section supplies the geometric picture behind that rate and shows why the trivial-but-singular representations that would formally violate the bound are never reached by training.

A. The fuzzy-ball picture

For a Gaussian prior $p_\psi(\mathbf{z}) = \mathcal{N}(0, \mathbb{I}_d)$ and a near-deterministic Gaussian encoder $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2\mathbb{I}_d)$ with $\sigma \ll 1$, the rate of Eq. (58) reduces to $R \simeq d\log(1/\sigma)$. This is the logarithm of the number $\sim (1/\sigma)^d$ of radius-σ balls that tile the unit-scale prior (Fig. S1). The encoder places each configuration $\mathbf{x}$ at a point $\mu(\mathbf{x})$ surrounded by a d-ball of radius σ; the aggregated posterior is the union of these balls, and the KL term of the ELBO forces it to fill the prior [1, 2]. The latent channel can therefore distinguish at most e^R equivalence classes of inputs, and any correlation structure that would require resolving more than $e^{I_{\mathrm{bip}}(p)}$ classes is lost into the encoder noise. A flow prior $p_\psi = g_\psi \# \mathcal{N}(0, \mathbb{I}_d)$ deforms each ball into an anisotropic ellipsoid but, being a fixed-capacity diffeomorphism, preserves the count up to an $\mathcal{O}(1)$ Jacobian term and leaves the asymptotic bound intact (Methods, Sec. X H).

B. Why singular priors do not rescue expressivity

Any joint distribution $p(\mathbf{x})$ admits a conditionally independent representation. The most pathological example is the following. Take $z \in [0, 1[$, write its binary digits as $z = 0.z_0 z_1 z_2 \cdots$, and denote by $(z)_k$ the number formed by the de-interleaved digits $(z)_k = 0.z_k\, z_{k+N}\, z_{k+2N} \cdots$. Identifying $x_i \equiv (z)_i$ and setting $p(z) = p\big(x_0 = (z)_0, \ldots, x_{N-1} = (z)_{N-1}\big)$ gives

$$p(\mathbf{x}) \;=\; \int \mathrm{d}z\, p(z) \prod_{i=0}^{N-1} \delta\big(x_i - (z)_i\big)\,, \tag{S1}$$

which is conditionally independent (indeed deterministic) given z and reproduces *any* target $p(\mathbf{x})$ with a single scalar latent. This degenerate solution formally satisfies the capacity criterion Eq. (27) by taking the limit $\sigma \to 0^+$, but the VAE training procedure never reaches it.

Two aspects of the pipeline forbid singular priors. First, the learned prior $p_\psi(\mathbf{z})$ is the pushforward of a Gaussian through a conditional spline flow of fixed depth and width, which is by construction a diffeomorphism; its log-density is Lipschitz and cannot concentrate on isolated point masses. Second, the encoder outputs a Gaussian

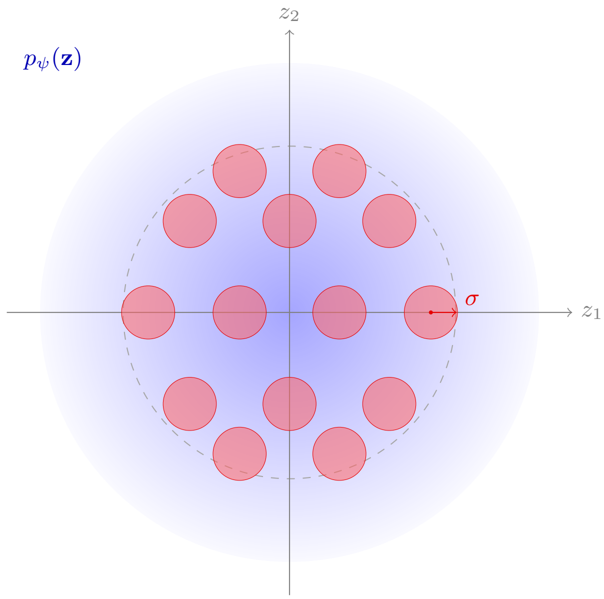

FIG. S1. **Fuzzy-ball picture of the latent channel** in $d = 2$. The prior $p_\psi(\mathbf{z}) = \mathcal{N}(0, \mathbb{I}_d)$ is the blue Gaussian (dashed line at one standard deviation). The encoder places each configuration $\mathbf{x}$ at $\mu(\mathbf{x})$ surrounded by a d-ball of radius σ (red); the aggregated posterior is the union of these balls, and the ELBO forces it to reproduce the prior. The number of distinguishable codes is the sphere-packing number $\sim (1/\sigma)^d$, and the channel rate is $R \simeq d \log(1/\sigma)$ nats [Eq. (55)].

$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mu, \sigma^2\mathbb{I})$ with strictly positive variance, and the KL divergence of a Gaussian against any prior with a singular component is $+\infty$,

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu, \sigma^2\mathbb{I}) \,\|\, p_\psi\big) = +\infty \quad \text{whenever } p_\psi \text{ has a singular component.} \tag{S2}$$

The reparameterization trick therefore confines optimization to the submanifold of smooth priors, on which the capacity bound is an honest ceiling. It is thus the training procedure *in conjunction with* the conditional-independence ansatz that enforces the mean-field expressivity bound: a purely combinatorial representation with exponentially many tightly concentrated latent classes is mathematically legal but practically unreachable.

S2. THERMODYNAMIC OBSERVABLES REPRODUCED ACROSS THE MODEL HIERARCHY

The main text quantifies the fidelity of the VAE reconstruction through the conditional total correlation $\mathrm{TC}_{|\mathbf{z}}$ (Fig. 3) and, for the Curie-Weiss model, through the standard thermodynamic observables (Fig. 1). Here we report the same observables: mean order parameter, susceptibility, and Binder cumulant as functions of temperature for the remaining three systems. In every panel open circles are Monte-Carlo ground truth, open squares are VAE samples, error bars are bootstrap standard deviations, and the solid line is the mean-field (thermodynamic-limit) prediction. Notice that since these observables rely only on one-point functions the VAE is able to recover them precisely for the 2D Ising model even without being able to model the spatial correlations (as was shown by $\mathrm{TC}_{|\mathbf{z}}$ in Fig. 3). The failure of 2D Ising will only show up in observables requiring 2-point functions (like $\langle x_i x_j \rangle$) and above.

S3. ROBUSTNESS OF MICROSCOPIC-PARAMETER RECOVERY

The main text reports the recovery of the microscopic mean-field parameters from the trained decoder. Namely, the Hopfield pattern matrix (Fig. 4a). Fig. S5 shows the quality of the recovery at every temperature and every memory load $\alpha = P/N$. We observe a near perfect recovery almost everywhere with only a slight degradation at low-temperature well-above the Amit-Gutfreund-Sompolinsky bound, i.e. deep in the glassy phase.

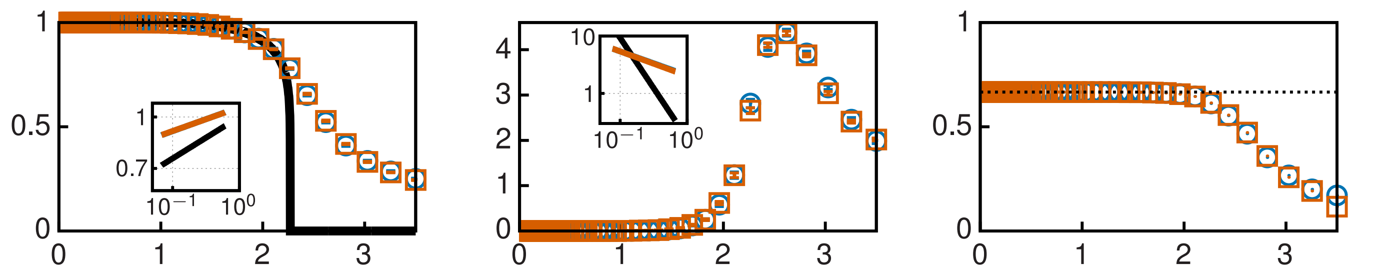

FIG. S2. **Two-dimensional Ising single-site observables.** Mean absolute magnetization $\langle|m|\rangle$ (left), susceptibility χ (middle) and Binder cumulant U_4 (right) versus T/J. The solid line is the infinite-dimensional mean-field prediction, shown for reference; the true two-dimensional curve is the Monte-Carlo data (circles). The VAE (squares) reproduces every one-point observable across the transition, even though it fails to capture the two-point correlations—the residual that $\mathrm{TC}_{|\mathbf{z}}$ detects (Fig. 3, top row).

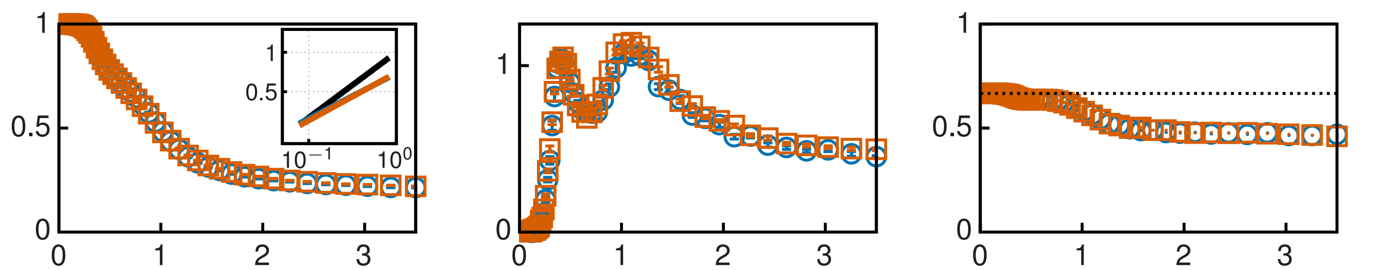

FIG. S3. **Hopfield model observables** ($N = 64$, $P = 4$). Maximum pattern overlap $\langle|m|_{\max}\rangle$ (left), susceptibility (middle) and Binder cumulant (right) versus T/J, with the mean-field self-consistency $m = \tanh(\beta J m)$ as the solid line. The VAE (squares) reproduces the retrieval transition at $T_c/J = 1$, including the condensed low-temperature plateau and the $1/\sqrt{N}$ isotropic baseline above T_c, substantiating the statement in Sec. V B of the main text.

[1] M. D. Hoffman and M. J. Johnson, Workshop in advances in approximate Bayesian inference, NIPS **1** (2016).
[2] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, Proceedings of the 35th International Conference on Machine Learning (ICML) , 159 (2018).

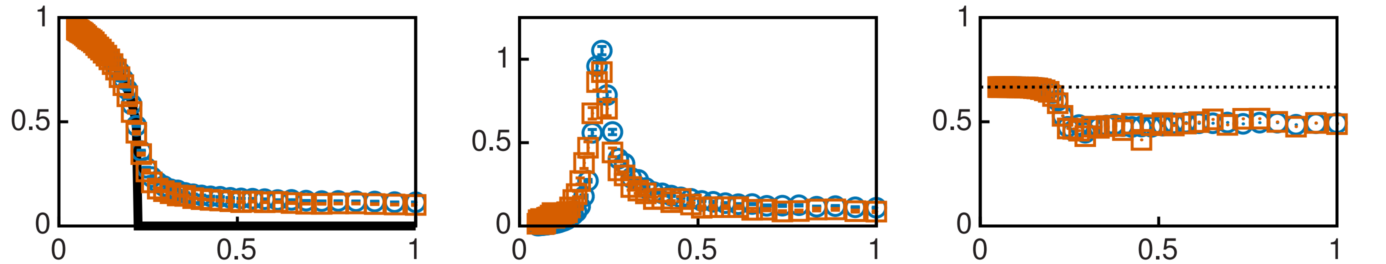

FIG. S4. **Maier–Saupe nematic observables** ($N = 64$). Scalar nematic order parameter S (left), susceptibility (middle) and Binder cumulant (right) versus T/J. The VAE (squares) reproduces the location of the first-order nematic–isotropic transition at $T_{NI}/J \approx 0.22$ and the Monte-Carlo observables (circles) on both sides of it, substantiating the statement in Sec. V C of the main text.

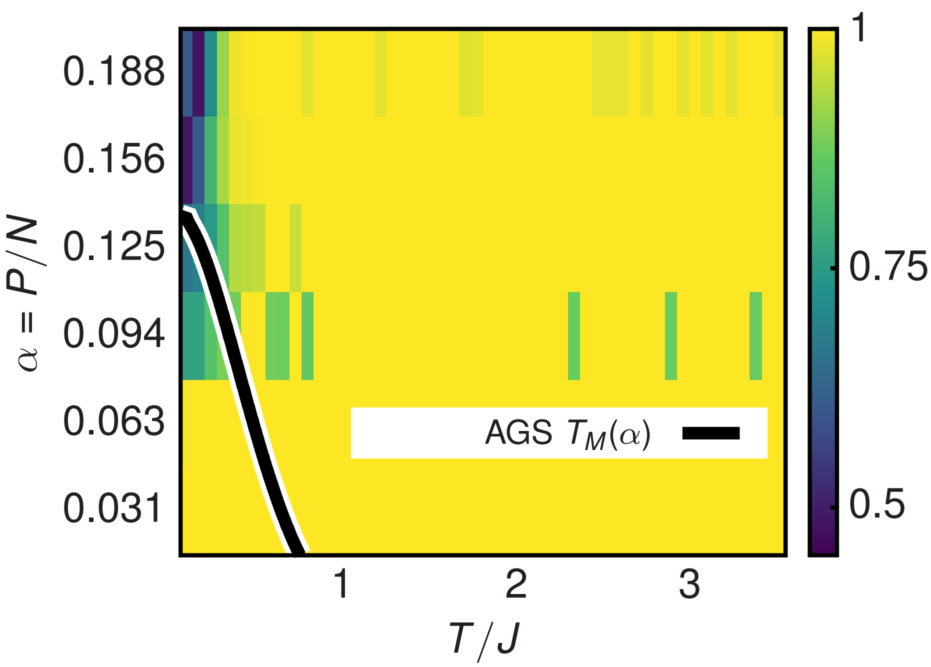

FIG. S5. **Hopfield pattern recovery.** Average pattern recovery fidelity as a function of the storage ratio $\alpha = P/N$ (at $N = 64$) and normalized temperature T/J. The black line marks the Amit-Gutfreund-Sompolinsky (AGS) bound. The fidelity stays close to unity well beyond the AGS retrieval bound, with only mild degradation appearing at low temperature with $\alpha > \alpha_c \approx 0.138$.