Spherical Boltzmann machines: a solvable theory of learning and generation in energy-based models

Thomas Tulinski, Simona Cocco, Rémi Monasson
Laboratoire de Physique de l'École Normale Supérieure, PSL,
CNRS UMR8023, Sorbonne Université, 24 rue Lhomond, 75005 Paris, France

Jorge Fernandez-de-Cossio-Diaz
Institut de Physique Théorique, Université Paris-Saclay,
CNRS UMR3681, CEA, Gif-sur-Yvette, France
jorge.fdcd@ipht.fr

Abstract

Energy-based models (EBMs) are flexible generative architectures inspired by statistical physics, but their learning and generative properties remain poorly understood. Here, we analyze a solvable EBM in the high-dimensional limit: the spherical Boltzmann machine (SBM). Combining tools from random matrix theory and dynamical mean-field theory, we: derive and solve closed equations describing the training dynamics of the SBM in the high-dimensional limit; compute the asymptotic Bayesian evidence, which acts as a partition function in parameter space and encodes global properties of the trained model; and uncover cascades of phase transitions that occur both during training and as a function of hyperparameters, related to successive alignment and condensation of the top modes of the coupling matrix to the data. We connect these transitions to sampling-time generative phenomena in a teacher-student scenario, including: sampling temperature tuning, double descent as a function of regularization strength, tempered posterior effects, and out-of-equilibrium effects during training that induce biases in the trained model. We provide numerical evidence showing that all these phenomena appear in standard generative architectures, beyond the SBM.[1]

1 Introduction

Energy-based models (EBMs) define complex probability distributions over high-dimensional spaces, placing mass on regions where realistic data lie while suppressing it elsewhere. The main ingredients are an energy function $\mathcal{E}(\mathbf{x}; W)$, defined over data points $\mathbf{x}$ and depending on parameters W, which is used to define a probability distribution through a Boltzmann law $P(\mathbf{x}|W) = e^{-\beta\mathcal{E}(\mathbf{x};W)}/Z_\beta(W)$, where β is an inverse sampling temperature and $Z_\beta(W) = \int e^{-\beta\mathcal{E}(\mathbf{x};W)}\mathrm{d}\mathbf{x}$ a normalization constant known as the partition function in statistical physics. EBMs are competitive generative models in diverse applications, including images [Du and Mordatch, 2019], biological sequences [Fernandez-de-Cossio-Diaz et al., 2025, Russ et al., 2020], and others. But the computation of Z requires a typically intractable integration over the high-dimensional point $\mathbf{x}$. As a consequence, training and sampling of EBMs often relies on approximate Monte Carlo (MC) techniques. Beyond computational cost, the intractable Z also blocks theoretical progress on the analysis of the trained model.

[1] Code: https://github.com/cossio/spherical-bm-2026.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Contributions. We study *spherical Boltzmann machines* (SBMs) trained on data in the high-dimensional limit. SBMs are solvable EBMs defined on the N-sphere. Combining recent results on finite-rank spherical-integral asymptotics [Guionnet and Husson, 2022] which encode the alignment of W's eigenvectors with the data principal components, with a Coulomb-gas analysis of W's eigenvalues as repelling charges in an external potential set by the data and the prior, we obtain a closed-form large-N theory of this model. We derive the equilibrium phase diagram of the trained SBM, which has qualitatively different properties in different hyperparameter settings; solve the high-dimensional Langevin dynamics of training via dynamical mean-field theory (DMFT) [Agoritsas et al., 2018, Cugliandolo, 2024] uncovering a non-equilibrium phase transition in the training dynamics; and compute closed-form metrics of generative performance in a teacher-student setting.

Our solvable model predicts four effects, that we fully characterize: **sampling temperature tuning**, where tuning $\beta \neq 1$ post-training can improve sample fidelity; **double descent** of generative metrics as a function of regularization; **tempered posterior** effects, where the performance of an ensemble of models weighted by the Bayes posterior improves if the posterior is annealed at a posterior temperature $\eta \neq 1$; and **out-of-equilibrium** training dynamics leading to eventual deterioration of the model during training.

We provide numerical evidence that all these phenomena arise in standard generative architectures, in qualitative agreement with the predictions from our solvable SBM theory. These experiments show that the four behaviors recur beyond the SBM and, in several cases, exhibit spectral signatures compatible with the SBM picture.

Related works. Statistical-mechanics tools have been used to characterize the parameter space of discriminative networks [Gardner, 1988, Zambon et al., 2025, Zdeborová and Krzakala, 2016] and of generative models with random weights [Amit et al., 1985, Tubiana and Monasson, 2017, Genovese and Tantari, 2020, Decelle and Furtlehner, 2021], neglecting the impact of training on data. For trained EBMs, analytical progress has relied on early-training approximations [Bachtis et al., 2024, Fachechi et al., 2025, Thériault et al., 2025]. The theoretical understanding of the structure of the learned weights is challenged by the complexity of the partition function term. In a related line of works, DMFT has emerged as a standard tool to study the high-dimensional learning trajectories of supervised models [Montanari and Urbani, 2025, Bonnaire et al., 2024, Urbani, 2026], and has been applied to maximum-likelihood training of RBMs with finitely many hidden units, using an asymptotic reduction of the likelihood to analyze data with order-one covariance spikes and a sample size proportional to the data dimension [Xu et al., 2025]. This reduction makes the negative phase tractable, but leaves open the problem of describing the coupled dynamics of learning and persistent negative-phase sampling, as commonly used in practical EBM training.

Sampling temperature tuning and out-of-equilibrium training dynamics are familiar to EBM practitioners. For example, in generative models of biological sequences, setting $\beta > 1$ at sampling time improves the functionality of designed sequences in experiments [Russ et al., 2020, Fernandez-de-Cossio-Diaz et al., 2025], but this often comes at the expense of generation diversity [Calvanese et al., 2026] and the trade-off is not very well understood [Fields et al., 2025]. Persistent-sampling MC is frequently used to estimate the learning gradient during EBM training, but training dynamics often reach out-of-equilibrium regimes, inducing subtle and hard-to-control biases in the trained model [Decelle et al., 2023, Bachtis et al., 2024]; in other settings, models are deliberately trained out-of-equilibrium [Decelle et al., 2021, Nijkamp et al., 2019]. Analytical results for non-persistent short-run MCMC training have been obtained for Gaussian EBMs and mean-field Boltzmann machines [Agoritsas et al., 2023]. But overall, these strategies are largely empirical, and the regimes in which they operate remain poorly understood.

Double descent [Belkin et al., 2019, Mei and Montanari, 2022] and tempered posterior effects [Wenzel et al., 2020, Pitas and Arbel, 2024] have been extensively studied in the context of supervised learning, though the origins of these effects are still debated. In generative models, double descent has been reported in GANs Luzi et al. [2024], LLMs [Hernandez et al., 2022, Morris et al., 2025], diffusion models [Zhang et al., 2025, Fraij and Dauncey, 2025], autoencoders [Rahimi et al., 2025], RBMs [Cheema and Sugiyama, 2020], and others, in a variety of settings, including as a function of hyper-parameters, sample complexity, or training duration. Bayesian ensembles of generative models have also been proposed [Murray and Ghahramani, 2012,

Saatchi and Wilson, 2017], but to the best of our knowledge, tempered posterior effects have not been reported before in a generative context.

A solvable EBM accounting for all these effects is missing. The SBM closes this gap: equilibrium structure, training dynamics, and generative phenomenology are characterized in a unified asymptotic large-N framework from which the four effects emerge.

2 Spherical Boltzmann machines: Equilibrium and training dynamics

The SBM is defined on the N-sphere, $\mathcal{S}_N = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|^2 = N\}$, with the energy function

$$\mathcal{E}(\mathbf{x}; W) = -\frac{1}{2}\mathbf{x}^\top W \mathbf{x}, \qquad \mathbf{x} \in \mathcal{S}_N, \quad W \in \mathrm{Sym}_N, \tag{1}$$

The SBM is a classical model from the statistical physics of disordered systems [Berlin and Kac, 1952, Kosterlitz et al., 1976], arising as a continuous relaxation of Ising spin systems, in which the per-site constraint $x_i \in \{\pm 1\}$ is replaced by the global constraint $\|\mathbf{x}\|^2 = N$. Despite the quadratic energy, $P(\mathbf{x}|W)$ is not Gaussian: it is the Bingham distribution of directional statistics [Bingham, 1974, Kent, 1982]. The phenomenology described below depends critically on the spherical constraint. Whereas W is purely random in most prior works, a few exceptions consider learning, such as Decelle and Furtlehner [2020], who study Gaussian-spherical RBMs but do not derive the large-N spectral law of the trained weights. In the present work, the SBM is treated as an EBM trained on data, which imprints statistical structure in W that we characterize both at equilibrium and along the training dynamics.

Without loss of generality, one can assume that $\beta = 1$ during training, but $\beta \neq 1$ is often tuned post-training with the hope of improving the quality of generated samples. Given a training dataset $\mathcal{D} = \{\mathbf{x}^k\}_{k=1}^K \subseteq \mathcal{S}_N$, the likelihood assigned by the SBM depends on $\mathcal{D}$ only through the rescaled empirical covariance matrix $C = \frac{1}{N}\sum_k \mathbf{x}^k \mathbf{x}^{k\top}$, with the usual empirical covariance given by $(N/K)C$. The model is then trained by ascent of the log-posterior. In the continuous-time limit, the training dynamics can be described by a Langevin equation,

$$\partial_t W(t) = \frac{1}{2}\left(C - \frac{K}{N}\langle \mathbf{x}\mathbf{x}^\top \rangle_{\mathbf{x}\sim P_{W(t)}}\right) - \frac{\gamma}{2}W(t) + \frac{\Omega(t)}{\sqrt{\eta N}}, \tag{2}$$

where $\gamma > 0$ controls the strength of weight decay, $\Omega(t)$ is a symmetric matrix-valued white noise satisfying $\langle \Omega_{ij}(t)\Omega_{kl}(t')\rangle = (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\,\delta(t-t')$ arising *e.g.* from minibatch sampling in stochastic gradient ascent, and $\eta > 0$ an inverse *learning temperature* controlling the amplitude of this noise. It is important to distinguish the two inverse temperatures from each other: η controls the dispersion in the space of model parameters W, while β acts in the sample-space, $\mathbf{x}$.

The moment average of $\mathbf{x}\mathbf{x}^\top$, also called the negative phase in the EBM literature, must be computed by drawing equilibrated samples from $P(\mathbf{x}|W(t))$, a daunting task in high-dimensional settings. In practice, this term is approximated by MC sampling. A popular variant of the Boltzmann machine learning algorithm consists of running persistent Markov chains, which are frequently updated before each gradient update of the weights [Younes, 1989, Tieleman, 2008, Rosset et al., 2026]. In this case, the training dynamics can be modeled by the coupled Langevin equations [Dotsenko et al., 1994, Coolen et al., 1993]:

$$\partial_t W(t) = \frac{1}{2}\left(C - \frac{K}{N}\mathbf{x}(t)\mathbf{x}(t)^\top\right) - \frac{\gamma}{2}W(t) + \frac{\Omega(t)}{\sqrt{\eta N}}, \tag{3}$$

$$\partial_t \mathbf{x}(t) = \nu W(t)\mathbf{x}(t) - \kappa(t)\mathbf{x}(t) + \sqrt{2\nu}\,\boldsymbol{\xi}(t), \tag{4}$$

where $\mathbf{x}(t)$ is the persistent MC chain, with $\kappa(t)$ a Lagrange multiplier enforcing $\|\mathbf{x}(t)\|^2 = N$, the independent white noise $\boldsymbol{\xi}(t)$ satisfying $\langle \xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t-t')$, and a rate ν controlling how many sampling steps are performed on the persistent MC chain per gradient update of the weights. In practice, increasing ν requires spending more compute cycles updating the MC chains during training.

Equations (3)-(4) define the exact training dynamics of our model. In the following, we study the large-N asymptotic solution of this model in two Data Regimes:

DR1. *Undersampled regime.* The number of training samples $K = \mathcal{O}(1)$ stays finite as $N \to \infty$. The nonzero eigenvalues of C and the hyperparameters are $\mathcal{O}(1)$.

DR2. *Effective low-dimensional data.* $N, K \to \infty$. A finite number $K' = \mathcal{O}(1)$ of eigenvalues of C grow linearly with K, while the remaining nonzero eigenvalues are $o(K)$ (sublinear) as K grows. No restriction is imposed on the relative scaling of N and K; in particular, the proportional regime $K/N \to \alpha \in (0, \infty)$ is included. Intuitively, the K data points are effectively low-dimensional, with K' dominant directions. We consider two sub-cases: **DR2.1** the hyperparameters stay fixed as K grows, treated in App. C.2.1; and **DR2.2** the hyperparameters are rescaled with K, treated in App. C.2.2.

We focus on DR1 in the remainder of the main text for simplicity. The equilibrium and dynamical phase structure derived below carries over unchanged to DR2.2 in the rescaled variables, while DR2.1 is recovered as a boundary limit of DR2.2 (App. C.2.3).

2.1 Equilibrium phase diagram

In the limit $\nu \to \infty$, $\mathbf{x}(t)$ approaches an equilibrium sample from $P(\mathbf{x}|W)$ in a quasi-stationary W background. In this case, the trained $W(t \to \infty)$ converges to the stationary distribution:

$$P_\eta(W|\mathcal{D}) = \frac{e^{\mathcal{L}(W;C)}}{\mathcal{Y}(C)}, \quad \text{with} \quad \mathcal{L}(W;C) = \frac{N}{2}\eta \operatorname{tr} CW - \frac{N}{4}\gamma\eta \operatorname{tr} W^2 - K\eta \ln Z(W), \tag{5}$$

a tempered posterior at temperature $1/\eta$, that interpolates between the standard Bayes posterior at $\eta = 1$ and the maximum-a-posteriori (MAP) limit $\eta \to \infty$. The tempered evidence, $\mathcal{Y}(C) = \int_{W \in \mathrm{Sym}_N} e^{\mathcal{L}(W;C)}\, \mathrm{d}W$, plays the role of a partition function over the weights.

Changing variables to the eigendecomposition $W = \frac{1}{N} V \Lambda V^\top$, we can rewrite $\mathcal{L}(W;C)$ as a Coulomb gas over eigenvalues Λ: $\lambda_1 > \cdots > \lambda_N$ [Livan et al., 2018],

$$\mathcal{L}_\Lambda(\Lambda;C) = -\frac{N}{4}\gamma\eta \operatorname{tr} \Lambda^2 - K\eta \ln Z(\Lambda) + \ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda;C) + \sum_{i<j} \ln|\lambda_i - \lambda_j|, \tag{6}$$

where $\mathcal{J}_{\mathrm{HCIZ}}$ is the real analogue of the Harish–Chandra–Itzykson–Zuber (HCIZ) integral [Guionnet and Husson, 2022] over the eigenvectors $V = (\mathbf{v}_j)$, normalized to $\mathbf{v}_i \cdot \mathbf{v}_j = N\delta_{ij}$.

In the large-N limit, the eigenvalue positions are obtained by optimizing $\mathcal{L}_\Lambda$. A dominant $\mathcal{O}(N^2)$ contribution comes from the terms $-\frac{N}{4}\gamma\eta \operatorname{tr} \Lambda^2 + \sum_{i<j} \ln|\lambda_i - \lambda_j|$, which do not involve the training data. These terms describe a standard Coulomb gas, with half-width parameter $\sigma = 1/\sqrt{\gamma\eta}$. The bottom $N - K$ eigenvalues then form a Wigner semicircle bulk,

$$\rho(\lambda) = \frac{1}{N-K} \sum_{i=K+1}^{N} \delta(\lambda_i - \lambda) \underset{N\to\infty}{\to} \frac{\sqrt{4\sigma^2 - \lambda^2}}{2\pi\sigma^2} \mathbf{1}_{|\lambda| \le 2\sigma}. \tag{7}$$

The top K eigenvalues interact with the data and may detach from the bulk, so we must treat them separately (Fig. 1A). Let $C = \frac{1}{N}\sum_k c_k \mathbf{c}_k \mathbf{c}_k^\top$ be the eigendecomposition of C, with eigenvectors normalized to $\mathbf{c}_i \cdot \mathbf{c}_j = N\delta_{ij}$. As a consequence of the $\mathrm{O}(N)$-rotation invariance of (1), the following analysis depends on C only through its eigenvalues, $c_1 > \cdots > c_K > 0$.

We determine the positions of $\lambda_1, \ldots, \lambda_K$ by combining finite-rank large-N asymptotics of $\mathcal{J}_{\mathrm{HCIZ}}$ derived by Guionnet and Husson [2022] with a saddle-point calculation extremizing (6), reported in App. A.2. Intuitively, this yields a balance of $\mathcal{O}(N)$ forces acting on each of the top K eigenvalues λ_k. Besides the interaction with the bulk, the likelihood terms contribute two additional forces (arrows in Fig. 1A),

$$\text{HCIZ pull away from bulk: } \frac{N}{2}\eta c_k u_k^2 \qquad \text{partition push toward bulk: } -\frac{NK}{2}\eta h^2 \tag{8}$$

where $u_k^2 = \langle (\mathbf{c}_k \cdot \mathbf{v}_k/N)^2 \rangle$ is the overlap of the k-th eigenvector of W with the k-th eigenvector of C, $h_k^2 = \langle (\mathbf{x} \cdot \mathbf{v}_k/N)^2 \rangle$ the overlap of typical generated samples with the k-th eigenvector of W, and $h^2 = \sum_{k=1}^K h_k^2$. The order parameters u_k, h characterize the trained model parameters and typical data it generates. They concentrate around typical values, defining the following qualitatively distinct regimes:

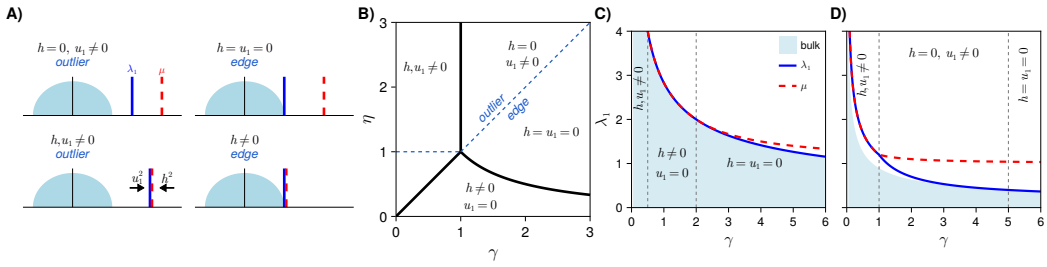


Figure 1: **Equilibrium phase diagram for $K = 1$. A)** Schematic eigenvalue configurations in each phase: bulk semicircle (light blue), top eigenvalue λ_1 and the Lagrange multiplier μ enforcing $\|\mathbf{x}\|^2 = N$. **B)** Phases in the (γ, η) plane; the blue dashed line separates outlier from edge regimes. **C)** λ_1 and Lagrange multiplier μ vs. γ in the small-η regime, where λ_1 sticks to the edge; phase transitions occur at the vertical dashed lines. **D)** Same in the larger-η regime, where λ_1 detaches before returning to the edge.

- $h = u_{1,\dots,K} = 0$ and $\lambda_{1,\dots,K} = 2\sigma$. If $\gamma\eta > 1$ and $\gamma > \eta c_1^2$, the eigenvalues stick to the bulk edge, the trained weights do not align macroscopically to the data, and generated samples are nearly isotropic on the sphere.
- $h = 0, u_{1,\dots,a} \neq 0$. If $\gamma > c_1$ and $\eta c_{a+1}^2 < \gamma < \eta c_a^2$; the top $a \in [K]$ modes detach from the bulk at $\lambda_k = (\eta c_k)^{-1} + c_k/\gamma$, aligning to the data. But this is insufficient to trigger condensation: generated samples are still nearly isotropic on the sphere.
- $h \neq 0, u_{1,\dots,d} \neq 0$: the top $d \in [K]$ eigenvalues coalesce at a common position λ_1 (see Eq. (41) in App. A.2) aligning to the data. Two sub-phases can be distinguished, according to whether λ_1 sticks to the semi-circle or detaches (conditions in Eqs. (52), (54)).
- $h \neq 0, u_{1,\dots,K} = 0$ and $\lambda_{1,\dots,K} = 2\sigma$. If $\eta^2 c_1^2 < \gamma\eta < 1$, generated samples condense along a random direction, unrelated to the data.

The phase diagram for $K = 1$ is shown in Fig. 1B. Fig. 1C,D show λ_1 as a function of γ for two selected values of η, illustrating the outlier detachment and condensation transitions. For $K > 1$ different modes can be in different phases, and top modes can coalesce and condense simultaneously (App. Fig. 7 for $K = 2$). In phases with $h = 0$, the projections $\mathbf{x} \cdot \mathbf{v}_j$ are all comparable and stay $\mathcal{O}(\sqrt{N})$, covering the sphere in a biased but nearly isotropic fashion. Phases with $h \neq 0$ are qualitatively different: samples *condense* along the top mode(s) of W, with significantly larger projections $\mathcal{O}(N)$ along $\mathbf{v}_{1,\dots,d}$, while projections along other directions remain $\mathcal{O}(\sqrt{N})$. See App. A.2 for derivations and App. A.2.3 for finite-N equilibrium comparisons. See also [Tulinski et al., 2026] and App. A.4 for an alternative approach based on the replica method.

2.2 Training dynamics

We next analyze how the Langevin dynamics (3)–(4) approach the equilibrium $P_\eta(W|\mathcal{D})$ described in Sec. 2.1. Equilibration is guaranteed only in the fast sampling limit $\nu \to \infty$; at finite ν training can settle into a non-equilibrium stationary state qualitatively distinct from the equilibrium phase diagram.

We initialize $W(0)$ at random from the prior and $\mathbf{x}(0)$ with a small seed overlap with the top data mode $\mathbf{c}_1$. Under these assumptions, Eq. (3) can be rewritten in the equivalent integrated form:

$$W(t) = W_{\mathrm{GOE}}(t) + \frac{1 - e^{-\gamma t/2}}{\gamma} C - \frac{K}{2} \int_0^t e^{-\gamma(t-u)/2} \frac{\mathbf{x}(u)\mathbf{x}(u)^\top}{N} \mathrm{d}u \tag{9}$$

where $W_{\mathrm{GOE}}(t)$ is a standard Gaussian orthogonal ensemble (GOE) matrix process,

$$\partial_t W_{\mathrm{GOE}}(t) = -\frac{\gamma}{2} W_{\mathrm{GOE}}(t) + \frac{\Omega(t)}{\sqrt{\eta N}}, \quad W_{\mathrm{GOE}}(0) = W(0), \tag{10}$$

The first term in (9) is a growing GOE noise, the second a deterministic spike in the data directions, and the third a negative-phase feedback from the generated samples. Their combination produces a

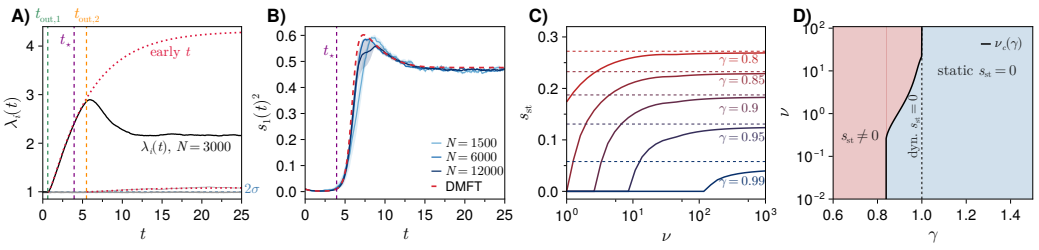


Figure 2: **Training dynamics. A)** Top eigenvalue trajectories $\lambda_i(t)$ from finite-N simulations (black) vs. the early-time predictions (14) (red dotted). Vertical dashed lines are the detachment (14) and condensation times t_*. Blue horizontal line is the bulk edge (2σ). **B)** $s_1(t)$ from DMFT (red) compared to finite-N simulations (blue). **C)** Stationary overlap $s_{\rm st}$ as a function of ν for selected values of γ (dashed lines indicate the equilibrium $\nu \to \infty$ asymptotes). **D)** MAP-limit dynamical phase diagram ($\eta \to \infty$), with $\nu_{\rm c}$ separating the $s_{\rm st} \neq 0$ from the $s_{\rm st} = 0$ phases. For $\gamma > 1$ (dashed), $s_{\rm st} = 0$ also at equilibrium.

cascade of eigenvalue detachments from the bulk. At every time, $W(t)$ consists of a Wigner bulk plus at most K upper outliers (proof in App. B.1).

Effective DMFT. We derive a closed set of equations describing the $N \to \infty$ limit of the process (9) through a DMFT saddle-point analysis. Although non-rigorous, the derivation retains all leading-order terms. As in many standard applications of DMFT, the resulting closure is therefore expected to be asymptotically exact. Briefly, representing the dynamics by a Martin–Siggia–Rose path integral [Martin et al., 1973], and integrating out the weights and noises yields, at the large-N saddle, the following closed equations for the overlaps $s_k(t) = \mathbf{c}_k \cdot \mathbf{x}(t)/N$, the correlator $Q(t,t') = \mathbf{x}(t) \cdot \mathbf{x}(t')/N$, and the causal response $R(t,t')$:

$$\partial_t s_k(t) = -\kappa(t) s_k(t) + \frac{\nu}{\gamma}(1 - e^{-\gamma t/2}) c_k s_k(t) + \int_0^t M(t,u) s_k(u)\, \mathrm{d}u, \tag{11}$$

$$\partial_t R(t,t') = -\kappa(t) R(t,t') + \int_{t'}^t M(t,u) R(u,t') \mathrm{d}u, \tag{12}$$

$$\begin{aligned} \partial_t Q(t,t') = {} & \frac{\nu}{\gamma}(1 - e^{-\gamma t/2}) \sum_{k=1}^K c_k s_k(t) s_k(t') - \kappa(t) Q(t,t') \\ & + \int_0^t M(t,u) Q(u,t') \mathrm{d}u + \int_0^{t'} D(t,u) R(t',u) \mathrm{d}u, \end{aligned} \tag{13}$$

subject to $Q(t,t') = Q(t',t)$, $R(t,t') = 0$ if $t < t'$, $R(t,t) = 1$, with $\kappa(t)$ at all times fixed by the spherical constraint $Q(t,t) = 1$. See App. B for derivations, including the definitions of the memory kernel $M(t,u)$ and the noise correlator $D(t,u)$, and further validation of the closure (11)–(13) against finite-N simulations.

Eigenvalue detachment cascade and condensation. Early in the dynamics the negative-phase contribution is weak ($s_k^2(t) \ll c_k/\gamma$) and Eq. (3) can be approximated by a deterministic spike growth, which places the k-th eigenvalue at the outlier position (App. B)

$$\lambda_k(t) = \begin{cases} \dfrac{1}{g_k(t)} + \dfrac{g_k(t)}{\gamma\eta} & t \geq t_{\mathrm{out},k}, \\ 2\sigma & t \leq t_{\mathrm{out},k}, \end{cases} \qquad \begin{aligned} &\text{with } g_k(t) = \frac{\gamma}{c_k(1 - e^{-\gamma t/2})}, \text{ and} \\ &\text{with } t_{\mathrm{out},k} = -\frac{2}{\gamma} \ln\left(1 - \frac{1}{c_k}\sqrt{\frac{\gamma}{\eta}}\right) \end{aligned} \tag{14}$$

where $g_k(t) = G(\lambda_k(t))$ is the Stieltjes transform (17) evaluated at the k-th outlier position, and $t_{\mathrm{out},k}$ is the detachment time, defined for $c_k > \sqrt{\gamma/\eta}$ (otherwise λ_k never detaches). Fig. 2A shows the top eigenvalue trajectories $\lambda_i(t)$ from finite-N Langevin simulations, in agreement with the early time prediction (14).

Fig. 2B shows $s_1(t)$ from DMFT in agreement with averaged finite-N simulations. If ν is large enough (see next paragraph), $s_1(t)$ first decays transiently before condensing along the top mode, after which $\lambda_1(t)$ deviates from the early time prediction (14). The time at which this happens can be computed as the t_* for which $s_1(t_*)$ reaches a prescribed small value, signaling onset of condensation of the training sample, after which $s_1(t)$ grows to a nonzero stationary value. We remark that here we initialize $\mathbf{x}(0)$ near the data, so $s_1(0) > 0$. Otherwise, if $\mathbf{x}(0)$ were initialized at random, t_* can be much larger (growing with N) as the training sample searches a distinguished spike direction in high-dimensional space [Bonnaire et al., 2024].

Stationary solution. For long times, the DMFT admits a stationary solution with $s_k(t) \to s_{\mathrm{st},k}$, $Q(t,t') \to Q_{\mathrm{st}}(t-t')$ and $R(t,t') \to R_{\mathrm{st}}(t-t')$. The trivial branch $s_{\mathrm{st},1} = 0$ is the only solution for low ν, but above a critical $\nu_c(\gamma,\eta)$, a condensed branch $s_{\mathrm{st},1} > 0$ appears (Fig. 2C and App. B). Fig. 2D shows the dynamical phase diagram in the MAP-limit ($\eta \to \infty$): for $\nu < \nu_c(\gamma)$ with $\gamma < 1$, training cannot reach the correct equilibrium. In this case, the negative phase is incorrectly estimated during training and the converged model is biased. EBM training dynamics have been studied in the literature [Bachtis et al., 2024, Decelle et al., 2021, Xu et al., 2025], but the DMFT above provides, to our knowledge, the first closed large-N description of the coupled dynamics of EBM weights and a persistent negative-phase sampler at finite sampling rate.

3 Generative metrics in a teacher–student scenario

We now evaluate the generative performance of an SBM student trained on K samples drawn from a teacher model $\mathbf{x}^k \sim P^*$ [Seung et al., 1992, Engel and Van den Broeck, 2001]. We consider two student models. A *typical* student commits to a single weight matrix $W \sim P_\eta(W|\mathcal{D})$ drawn from the posterior, and then generates data from: $\mathbf{x} \sim P_W(\mathbf{x}) = P(\mathbf{x}|W)$. The *posterior predictive* student instead marginalizes over the weights before generation, effectively drawing samples from an ensemble of trained models weighted by their posterior probabilities: $\mathbf{x} \sim P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D}) = \int P(\mathbf{x}|W) P_\eta(W|\mathcal{D})\, \mathrm{d}W$.

The quality of a student generated sample $\mathbf{x}$ is measured by its surprise under the teacher, $-\ln P^*(\mathbf{x})$. Averaging over $\mathbf{x} \sim P_W$ yields the cross-entropy $\mathcal{H}[P_W; P^*] = -\int P_W(\mathbf{x}) \ln P^*(\mathbf{x}) \mathrm{d}\mathbf{x}$. Minimizing this quantity alone, however, selects a singular student placing all its probability mass at the maximizer of P^*. Diversity collapse is undesirable for a generative model, so we are naturally led to the reverse Kullback-Leibler (KL) divergence, $D_{\mathrm{KL}}(P_W \| P^*) = \mathcal{H}[P_W; P^*] - \mathcal{H}[P_W]$, where the entropy $\mathcal{H}[P_W] = -\int P_W(\mathbf{x}) \ln P_W(\mathbf{x}) \mathrm{d}\mathbf{x}$ penalizes students with poor diversity. Reversing the arguments defines the forward-KL, $D_{\mathrm{KL}}(P^* \| P_W)$. Whereas the reverse-KL is mode-seeking, concentrating student mass where P^* is large, the forward-KL is mode-covering, penalizing regions of P^* the student misses [Bishop, 2006]. Both are widely used to train and evaluate generative models in practice. In the large N-limit we consider here, they are self-averaging: for any typical realization $W \sim P_\eta$, both KLs stay close to their average values, $\langle D_{\mathrm{KL}}(P_W \| P^*) \rangle_{W \sim P_\eta}$ and $\langle D_{\mathrm{KL}}(P^* \| P_W) \rangle_{W \sim P_\eta}$, respectively, so we may omit the average brackets to lighten the notation.

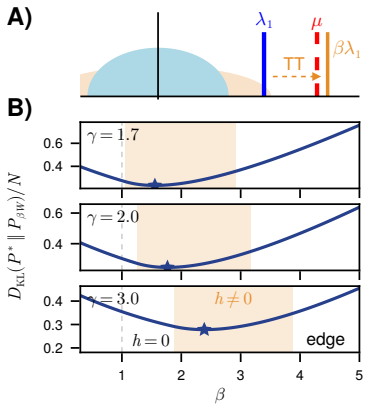


Figure 3: **Sampling temperature tuning** (TT). **A)** Mode rescue through temperature tuning, illustrating how the bulk and outlier position change. **B)** $D_{\mathrm{KL}}(P^* \| P_{\beta W})$ vs. β.

Similarly, the generative performance of the posterior predictive student can be measured by the reverse $D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*)$ or forward $D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}})$. The posterior predictive is popular in Bayesian prediction [Aitchison, 1975, Brown et al., 2008], but, with few exceptions [Saatchi and Wilson, 2017], has been less studied in the context of generative modeling, where the focus has been on typical or MAP weights. However, the predictive posterior is expected to outperform typical students: $D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) \leq \langle D_{\mathrm{KL}}(P_W \| P^*) \rangle_{W \sim P_\eta}$ and $D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) \leq \langle D_{\mathrm{KL}}(P^* \| P_W) \rangle_{W \sim P_\eta}$ by convexity (see App. D.8). These inequalities justify the interest in the posterior predictive also in a generative setting.

In what follows we consider a simple SBM teacher model with rank-one weights $W^* = \omega^* \mathbf{w}^* \mathbf{w}^{*\top} - \frac{1}{N}\omega^* I$, where I is the identity matrix, $\mathbf{w}^*$ is a unit vector and ω^* a parameter controlling the signal-to-noise ratio, with $\omega^* > 1$ placing the teacher in a condensed regime. The teacher samples $K = 2$ data points which are used to train a student SBM. This is the smallest sample size giving two nonzero covariance modes, one correlated with the teacher direction and the other asymptotically orthogonal to it (App. D.1), and is sufficient to demonstrate the phenomenology of interest. The forward and reverse KL divergences, for typical and posterior predictive students, are all computable in closed form in this setting. Each of the phenomena introduced above surfaces most naturally in a different KL, motivating our examination of all four KL variants. We emphasize that both students are constructed from the same tempered posterior (5), based on the regularized log-likelihood; the choice of KL only concerns their evaluation. Detailed calculations are deferred to App. D.

Sampling temperature tuning. After training, one can re-introduce a sampling temperature $1/\beta \neq 1$. What is the effect of tuning β at generation-time? The forward $D_{\mathrm{KL}}(P^* \| P_{\beta W})$ is a convex function of β. Its unique minimizer is > 1 or < 1 according to whether the entropy $\mathcal{H}[P_W]$ is larger or smaller than the cross-entropy $\mathcal{H}[P^*; P_W]$. Intuitively, $\beta > 1$ (< 1) compensates for an excess (lack) of entropy in the trained model.

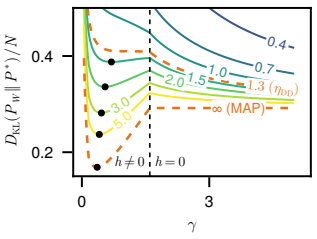


Figure 4: **Double descent.** η indicated in each curve. Black dots: minima. Orange dashed: thresholds $\eta = \eta_{\mathrm{DD}}$ (upper) and $\eta \to \infty$ (lower). Vertical dashed: boundary between $h \neq 0$ and $h = 0$ phases.

The model responds in qualitatively different ways to temperature tuning, depending on the starting phase. Furthermore, changing the value of β can transport the model from one phase to another. If the trained model starts in the phase $h \neq 0, u_1 = 0$ (condensed in a random direction), the optimal β is always < 1, and it is such that the temperature-tuned model lands in an uncondensed phase ($h \to 0$). This removes the random condensed direction ($u_1 = 0$), which results in a lower KL divergence from the teacher P^*.

If instead the model is initially in the $h = 0, u_1 \neq 0$ phase, the coupling W carries data-aligned modes, but the data signal imprinted in the weights is too weak and samples are swamped by noise. Tuning $\beta > 1$ is equivalent to rescaling $W \to \beta W$, which can push the aligned outlier modes beyond the condensation threshold, so that the learned modes become active in generated data, *i.e.*, the temperature-tuned model is transported to a $h, u_1 \neq 0$ phase. This is illustrated in Fig. 3. See App. D.9 for further discussion and derivations.

Double descent. Fig. 4 plots the reverse $D_{\mathrm{KL}}(P_W \| P^*)$ as a function of the regularization strength γ. If η is sufficiently large, it exhibits a double descent. First, for $\gamma \in [0, c_1]$, the reverse-KL reaches a minimum (black dots) inside the $h \neq 0$ phase, where the student model condenses along a teacher-aligned direction. Beyond the optimal γ, the student condensation (h) weakens and the reverse-KL increases, until it spikes at the condensation threshold $\gamma = c_1$ (vertical dashed line), where the student transitions into the $h = 0$ phase. The KL is decreasing throughout this last regime, according to $\frac{1}{N} D_{\mathrm{KL}}(P_W \| P^*) = \frac{\omega^* - 1 - \ln \omega^*}{2} + \frac{1}{4\gamma\eta}$ (see Eq. (148) for the general expression). The first contribution is an irreducible gap between the teacher and a student uniformly distributed on the sphere, $\frac{1}{N} D_{\mathrm{KL}}(P_{W=0} \| P^*)$, while the second term $1/(4\gamma\eta)$ accounts for the random anisotropy of the modes in the bulk of the student weight spectrum.

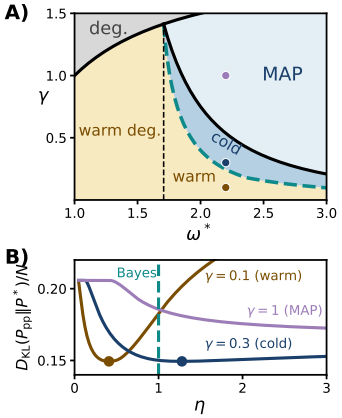


Figure 5: **Tempered-posterior**. **A)** Phases of the optimal $\eta_{\mathrm{pp}}^{\mathrm{rev}}(\gamma, \omega^*)$. **B)** Reverse KL vs. η for selected γ's (dots: minima).

The double descent behavior is present only in a window of posterior temperatures, $\eta_{\mathrm{DD}}(\omega^*) < \eta < \infty$. The value of $\eta_{\mathrm{DD}}(\omega^*)$ is determined in App. D.10 (upper orange dashed curve in Fig. 4). Below $\eta_{\mathrm{DD}}(\omega^*)$, the student bulk is over-spread and the reverse-KL is a monotonically decreasing function of γ, with no minima in the $h \neq 0$ phase. In the opposite MAP limit $\eta \to \infty$ (lower orange dashed in Fig. 4), the anisotropy contribution in the $h = 0$ phase is sent to zero, collapsing the second descent to a flat line, while preserving the first minimum.

Tempered-posterior effects. The MAP limit ($\eta \to \infty$) in our framework corresponds to the standard practice of training a single generative model by minimizing a loss. Few works have attempted to target a Bayesian ($\eta = 1$) posterior predictive distribution [Saatchi and Wilson, 2017]. In our solvable setting, the optimal η may differ from both endpoints. The posterior-predictive KL divergences are not convex functions of η, and their behavior is consequently more complex than the sampling temperature-tuning scenario discussed above. We focus in this section on the reverse KL, and define an optimal value of η minimizing this quantity: $\eta_{\rm pp}^{\rm rev}(\gamma, \omega^*) = \operatorname{argmin}_{\eta>0} D_{\rm KL}\left(P_{\rm pp} \middle\| P^*\right)$. The optimal $\eta_{\rm pp}^{\rm rev}$ depends on the values of the parameters γ, ω^*. In our teacher-student scenario, small γ and/or small ω^* favor $\eta < 1$ (warm posterior effect), while larger γ and/or larger ω^* favors $\eta > 1$ (cold posterior effect).

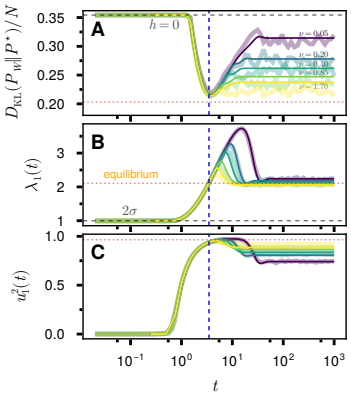


Figure 6: **Out-of-equilibrium training.** **A)** Reverse $D_{\rm KL}(P_{W(t)}\|P^*)$ versus training time t, **B)** $\lambda_1(t)$, **C)** $u_1(t)$. Color: ν. Blue dashed: optimal early-stopping time. Orange dashed: equilibrium values (at $\nu \to \infty$).

An optimal posterior temperature different from 1 (the Bayesian prescription) has been referred to as a *cold/warm-posterior effect* [Wenzel et al., 2020, Noci et al., 2021, Nabarro et al., 2022, Pitas and Arbel, 2024] in supervised learning. Here it arises in an exactly solvable EBM in the high-dimensional limit.

Sweeping (ω^*, γ) partitions the plane into the five regions shown in Fig. 5A, with closed-form thresholds derived in App. D.11.1. In the *warm* phase the optimum lies below the Bayesian value $\eta = 1$, because predictions benefit more from averaging across the posterior than from concentrating on its mode. The *cold* phase is the opposite: the posterior is already informative and concentrating on its mode at finite $\eta > 1$ outperforms averaging. The *MAP* phase is the extreme limit in which collapsing onto the posterior mode is optimal ($\eta_{\rm pp}^{\rm rev} \to \infty$). In the two *deg.* regions, either h or u_1 vanishes, and the loss becomes degenerate in η over a range of values, making the minimum $\eta_{\rm pp}^{\rm rev}$ non-unique. Within *warm deg.*, there is an interval of minimizers which are all < 1, whereas in *deg.* the reverse-KL becomes independent of η so the minimum is completely degenerate. Fig. 5B plots the reverse KL as a function of η for selected values of γ, having warm, cold, and MAP minima. The posterior predictive forward-KL has a similar behavior, with large ω^*, γ selecting cold/MAP optima ($\eta_{\rm pp}^{\rm fwd} > 1$), while small ω^*, γ select warm optima ($\eta_{\rm pp}^{\rm fwd} < 1$). See Fig. 13 in Appendix.

Out-of-equilibrium training dynamics. The dynamical regimes identified in Sec. 2.2 shape the generative quality of the trained model. Fig. 6 shows example training trajectories at varying sampling rates ν, where we also track the reverse $D_{\rm KL}(P_{W(t)}\|P^*)$ during training (panel A). For $\nu \gg \nu_{\rm c}$, the negative phase persistent MC chain equilibrates quickly. The reverse-KL, the top eigenvalue $\lambda_1(t)$ and the overlap $u_1(t)$ all converge to close to their equilibrium ($\nu \to \infty$, dashed orange) values.

As ν decreases, $\lambda_1(t)$ detaches from the bulk and initially follows the same high-ν trajectory, but then keeps growing past its equilibrium value, overshooting before slowly relaxing back. The overlap $u_1(t)$ stays close to 1 longer than at high ν, then drops to an asymptote below its equilibrium plateau. The KL loss dips almost reaching the equilibrium value, but then increases again and settles at a higher value. The early-time spike dynamics (14) can be used to estimate an optimal early-stopping time (dashed blue, Fig. 6; see Eq. (199)) that is independent of ν. In this setting, early stopping offers no benefit in the equilibrated limit $\nu \to \infty$.

4 Validation of phenomenology in standard generative architectures

We next test whether the four effects found above in the SBM also occur in standard architectures (Table 2). Although some have been reported previously, they remain poorly understood in generative models. In particular, posterior tempering is little explored in generative models to the best of our knowledge.

We first consider small BMs and RBMs with exactly computable partition functions and numerically estimated posterior averages, trained on spiked teachers as in Sec. 3. In the RBM, we verify beneficial warming and cooling of both sampling and posterior temperatures, together with nonmonotonic

typical reverse KL as regularization varies. Slower persistent sampling amplifies weight modes and causes alignment to decline after an initial rise, favoring early stopping over continued training; matched controls separate infrequent-refresh effects from an additional Gibbs contribution. These finite-size experiments do not determine SBM phase boundaries. Full results are reported in Apps. G.1 and G.2.

We next consider larger models. The relevant posterior averages and observables are difficult to evaluate jointly in high-dimensional models, and we could not probe all four effects in a single large model. We instead use complementary architectures, each suited to probing one or more aspects of the phenomenology. Although we do not claim quantitative agreement, and other factors may play a role in larger models, these experiments provide qualitative evidence that the four effects recur beyond the SBM and, where spectral observables are accessible, reveal signatures compatible with the SBM mechanism. Details are in App. G. Briefly,

Sampling temperature tuning. We trained Potts Boltzmann machines on protein sequences from Pfam [Weigt et al., 2009, Marks et al., 2011, Hopf et al., 2017], at various hyperparameter settings. In sufficiently regularized models, the learned coupling matrix contains modes aligned with the top covariance modes of the data that stay approximately silent in the generated data, reminiscent of the $h = 0, u_1 \neq 0$ phase in the SBM. Tuning $\beta > 1$ rescues the silent modes in generated data, in a manner compatible with the mechanism depicted in Fig. 3. See App. G.3.

Double descent. We confirmed double descent as a function of regularization in two generative architectures: a Gaussian–Bernoulli restricted Boltzmann machine (RBM) trained on a financial dataset [Fama and French, 1997, Laloux et al., 1999], and autoregressive belief networks trained on samples from the 2D Ising and Curie–Weiss models. See App. G.4.

Tempered-posterior effects. To the best of our knowledge, warm/cold posterior effects have not been studied in generative models. We trained Bayesian GANs Saatchi and Wilson [2017] on synthetic datasets, and demonstrate that the optimal posterior temperature can be < 1 or > 1, depending on the regularization, as in the SBM. See App. G.5.

Out-of-equilibrium training dynamics. We trained Potts Boltzmann machines on protein sequence alignments at varying sampling rates ν during training, observing overshooting of the top weight modes, with data overlaps that converge to sub-optimal values, reproducing the qualitative behavior of the SBM. See App. G.6.

5 Conclusion, limitations, and broader impact

Solvable models provide guidance to navigate complex phenomena in machine learning [Simon et al., 2026]. We introduced the SBM as a solvable EBM, derived its equilibrium phase diagram, solved its Langevin training dynamics via DMFT, and obtained teacher–student generative metrics in the high-dimensional limit. Standard generative architectures reproduce qualitatively the predicted temperature tuning, double-descent as a function of regularization strength, tempered-posterior effects, and out-of-equilibrium training behavior. These effects depend critically on the spherical constraint of the SBM, which drives the condensation transition, and would not occur in a multivariate Gaussian model (see App. E for further discussion).

Limitations. The SBM is a minimal solvable EBM rather than a realistic generative architecture. Our analysis assumes effectively low-dimensional data (App. C). Our derivations are heuristic in the statistical-physics sense (saddle-point evaluation of the Coulomb gas, MSR/DMFT closure of the Langevin dynamics) rather than rigorous proofs; results are expected to be asymptotically exact in the $N \to \infty$ limit, and we verify them against finite-N simulations throughout.

Broader impact. This is a theoretical paper on the statistics of generative models; it does not propose new models or release datasets, and we do not see a direct avenue for societal harm.

Acknowledgments and Disclosure of Funding

We acknowledge support from ANR grants MEMNET, ANR-24-CPJ1-0172-01 and ANR-25-CE45-5343 "MANUFACTURE". We used HPC resources from GENCI–IDRIS, Grant 2025-AD011015768R1.

References

Agoritsas, Elisabeth, Biroli, Giulio, Urbani, Pierfrancesco, and Zamponi, Francesco. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51:085002, 2018. doi: 10.1088/1751-8121/aaa68d. URL https://arxiv.org/abs/1710.04894.

Agoritsas, Elisabeth, Catania, Giovanni, Decelle, Aurélien, and Seoane, Beatriz. Explaining the effects of non-convergent MCMC in the training of energy-based models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 322–336. PMLR, 2023. URL https://proceedings.mlr.press/v202/agoritsas23a.html.

Aitchison, John. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975. doi: 10.1093/biomet/62.3.547.

Amit, Daniel J., Gutfreund, Hanoch, and Sompolinsky, Haim. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985. doi: 10.1103/PhysRevA.32.1007.

Bachtis, Dimitrios, Biroli, Giulio, Decelle, Aurélien, and Seoane, Beatriz. Cascade of phase transitions in the training of energy-based models, 2024.

Baik, Jinho, Arous, Gérard Ben, and Péché, Sandrine. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005. doi: 10.1214/009117905000000233. URL https://doi.org/10.1214/009117905000000233.

Baxter, Rodney J. *Exactly Solved Models in Statistical Mechanics*. Academic Press, London, reprint edition, 1990. ISBN 978-0120831821. Second paperback printing; originally published 1982.

Belkin, Mikhail, Hsu, Daniel, Ma, Siyuan, and Mandal, Soumik. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

Berlin, Theodore H and Kac, Mark. The spherical model of a ferromagnet. *Physical Review*, 86(6):821, 1952. doi: 10.1103/PhysRev.86.821.

Bingham, Christopher. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974. doi: 10.1214/aos/1176342874.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Bonnaire, Tony, Ghio, Davide, Krishnamurthy, Kamesh, Mignacco, Francesca, Yamamura, Atsushi, and Biroli, Giulio. High-dimensional non-convex landscapes and gradient descent dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104004, 2024.

Brown, Lawrence D., George, Edward I., and Xu, Xinyi. Admissible predictive density estimation. *The Annals of Statistics*, 36(3):1156–1170, 2008. doi: 10.1214/07-AOS506.

Calvanese, Francesco, Lombardi, Gianluca, Weigt, Martin, and Fernandez-de-Cossio-Diaz, Jorge. Steering sequence generation in protein language models through Iterative Lookback Monte Carlo Sampling. *Machine Learning: Science and Technology*, 7(5):055003, 2026. doi: 10.1088/2632-2153/ae923a. URL https://doi.org/10.1088/2632-2153/ae923a.

Cheema, Prasad and Sugiyama, Mahito. The volume of non-restricted Boltzmann machines and their double descent model complexity. In *NeurIPS 2020 Workshop on Deep Learning and Information Geometry*, 2020. URL https://openreview.net/forum?id=vU1QL3jGmV_.

Chen, Tianqi, Fox, Emily, and Guestrin, Carlos. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014. URL https://arxiv.org/abs/1402.4102.

Coolen, A.C.C., Penney, R., and Sherrington, D. Coupled dynamics of fast neurons and slow interactions. In Cowan, J., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/c0f168ce8900fa56e57789e2a2f2c9d0-Paper.pdf.

Cugliandolo, Leticia F. Recent applications of dynamical mean-field methods. *Annual Review of Condensed Matter Physics*, 15(Volume 15, 2024):177–213, 2024. ISSN 1947-5462. doi: https://doi.org/10.1146/annurev-conmatphys-040721-022848. URL https://www.annualreviews.org/content/journals/10.1146/annurev-conmatphys-040721-022848.

Decelle, Aurélien and Furtlehner, Cyril. Gaussian-spherical restricted boltzmann machines. *Journal of Physics A: Mathematical and Theoretical*, 53(18):184002, 2020. doi: 10.1088/1751-8121/ab79f3.

Decelle, Aurélien and Furtlehner, Cyril. Restricted boltzmann machine: Recent advances and mean-field theory. *Chinese Physics B*, 30(4):040202, 2021.

Decelle, Aurélien, Furtlehner, Cyril, and Seoane, Beatriz. Equilibrium and non-equilibrium regimes in the learning of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems*, volume 34, pages 5345–5359, 2021.

Decelle, Aurélien, Seoane, Beatriz, and Rosset, Lorenzo. Unsupervised hierarchical clustering using the learning dynamics of restricted Boltzmann machines. *Physical Review E*, 108(1):014110, 2023. doi: 10.1103/PhysRevE.108.014110.

Depeweg, Stefan, Hernández-Lobato, José Miguel, Doshi-Velez, Finale, and Udluft, Steffen. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning (ICML)*, 2018.

Dotsenko, Victor, Franz, Silvio, and Mézard, Marc. Partial annealing and overfrustration in disordered systems. *J. Phys. A: Math. Gen.*, 27:2351–2365, 1994. doi: 10.1088/0305-4470/27/7/016.

Du, Yilun and Mordatch, Igor. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Engel, Andreas and Van den Broeck, Chris. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. doi: 10.1017/CBO9781139164542.

Fachechi, Alberto, Agliari, Elena, Aquaro, Miriam, Coolen, Anthony, and Mulder, Menno. Fundamental operating regimes, hyper-parameter fine-tuning and glassiness: towards an interpretable replica-theory for trained restricted Boltzmann machines. *Journal of Physics A: Mathematical and Theoretical*, 58(6):065004, 2025. doi: 10.1088/1751-8121/ada986.

Fama, Eugene F. and French, Kenneth R. Industry costs of equity. *Journal of Financial Economics*, 43:153–193, 1997. doi: 10.1016/S0304-405X(96)00896-3.

Fernandez-de-Cossio-Diaz, Jorge, Hardouin, Pierre, Lyonnet du Moutier, Francois-Xavier, Di Gioacchino, Andrea, Marchand, Bertrand, Ponty, Yann, Sargueil, Bruno, Monasson, Rémi, and Cocco, Simona. Designing molecular RNA switches with restricted Boltzmann machines. *Nature Communications*, 16(1):11223, 2025.

Fields, Peter W., Ngampruetikorn, Vudtiwat, Schwab, David J., and Palmer, Stephanie E. Understanding temperature tuning in energy-based models, 2025. URL https://arxiv.org/abs/2512.09152.

Fraij, Ahmad and Dauncey, Sam. Double descent as a lens for sample efficiency in autoregressive vs. discrete diffusion models, 2025. URL https://arxiv.org/abs/2509.24974.

Frey, Brendan J. *Graphical models for machine learning and digital communication*. PhD thesis, MIT Press, Cambridge, MA, 1998.

Gardner, Elizabeth. The space of interactions in neural network models. *J. Phys. A: Math. Gen.*, 21(1):257–270, 1988. doi: 10.1088/0305-4470/21/1/030.

Genovese, Giuseppe and Tantari, Daniele. Legendre equivalences of spherical Boltzmann machines. *Journal of Physics A: Mathematical and Theoretical*, 53(9):094001, 2020. doi: 10.1088/1751-8121/ab6b92. URL https://arxiv.org/abs/1910.14559.

Guionnet, Alice and Husson, Jonathan. Asymptotics of k dimensional spherical integrals and applications. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 19(1):769–797, 2022.

Hernandez, Danny, Brown, Tom, Conerly, Tom, DasSarma, Nova, Drain, Dawn, El-Showk, Sheer, Elhage, Nelson, Hatfield-Dodds, Zac, Henighan, Tom, Hume, Tristan, Johnston, Scott, Mann, Ben, Olah, Chris, Olsson, Catherine, Amodei, Dario, Joseph, Nicholas, Kaplan, Jared, and McCandlish, Sam. Scaling laws and interpretability of learning from repeated data, 2022. URL https://arxiv.org/abs/2205.10487.

Hopf, Thomas A., Ingraham, John B., Poelwijk, Frank J., Schärfe, Charlotta P. I., Springer, Michael, Sander, Chris, and Marks, Debora S. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017. doi: 10.1038/nbt.3769.

Houlsby, Neil, Huszár, Ferenc, Ghahramani, Zoubin, and Lengyel, Máté. Bayesian active learning for classification and preference learning, 2011.

Kent, John T. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982. doi: 10.1111/j.2517-6161.1982.tb01189.x.

Kent, John T., Ganeiber, Asaad M., and Mardia, Kanti V. A new unified approach for the simulation of a wide class of directional distributions. *Journal of Computational and Graphical Statistics*, 27(2):291–301, 2018. doi: 10.1080/10618600.2017.1390468.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. *Proceedings of ICLR*, 2015.

Kosterlitz, John M, Thouless, David J, and Jones, Raymund C. Spherical model of a spin-glass. *Physical Review Letters*, 36(20):1217, 1976. doi: 10.1103/PhysRevLett.36.1217.

Lakshminarayanan, Balaji, Pritzel, Alexander, and Blundell, Charles. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Laloux, Laurent, Cizeau, Pierre, Bouchaud, Jean-Philippe, and Potters, Marc. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83:1467–1470, 1999. doi: 10.1103/PhysRevLett.83.1467.

Lindley, Dennis V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Livan, Giacomo, Novaes, Marcel, and Vivo, Pierpaolo. *Introduction to Random Matrices: Theory and Practice*. Springer, 2018. doi: 10.1007/978-3-319-70885-0.

Luzi, Lorenzo, Dar, Yehuda, and Baraniuk, Richard G. Double descent and other interpolation phenomena in GANs, 2024. URL `https://arxiv.org/abs/2106.04003`.

Maddox, Wesley J., Izmailov, Pavel, Garipov, Timur, Vetrov, Dmitry P., and Wilson, Andrew Gordon. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Marks, Debora S., Colwell, Lucy J., Sheridan, Robert, Hopf, Thomas A., Pagnani, Andrea, Zecchina, Riccardo, and Sander, Chris. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12): e28766, 2011. doi: 10.1371/journal.pone.0028766.

Martin, Paul C., Siggia, Eric D., and Rose, Harry A. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423–437, 1973. doi: 10.1103/PhysRevA.8.423.

Mei, Song and Montanari, Andrea. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.

Montanari, Andrea and Urbani, Pierfrancesco. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

Morris, John X., Sitawarin, Chawin, Guo, Chuan, Kokhlikyan, Narine, Suh, G. Edward, Rush, Alexander M., Chaudhuri, Kamalika, and Mahloujifar, Saeed. How much do language models memorize?, 2025. URL `https://arxiv.org/abs/2505.24832`.

Murray, Iain and Ghahramani, Zoubin. Bayesian learning in undirected graphical models: Approximate MCMC algorithms, 2012. URL `https://arxiv.org/abs/1207.4134`.

Murray, Iain, Ghahramani, Zoubin, and MacKay, David J. C. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2006. URL `https://homepages.inf.ed.ac.uk/imurray2/pub/06doubly_intractable/`.

Nabarro, Seth, Ganev, Stoil, Garriga-Alonso, Adrià, Fortuin, Vincent, van der Wilk, Mark, and Aitchison, Laurence. Data augmentation in Bayesian neural networks and the cold posterior effect. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *PMLR*, pages 1434–1444, 2022.

Nijkamp, Erik, Hill, Mitch, Zhu, Song-Chun, and Wu, Ying Nian. Learning non-convergent non-persistent short-run MCMC toward energy-based model. *Advances in Neural Information Processing Systems*, 32, 2019.

Noci, Lorenzo, Roth, Kevin, Bachmann, Gregor, Nowozin, Sebastian, and Hofmann, Thomas. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Pastore, Mauro, Di Gioacchino, Andrea, and Rotondo, Pietro. Large deviations of the free energy in the p-spin glass spherical model. *Phys. Rev. Res.*, 1(3):033116, 2019.

Pitas, Konstantinos and Arbel, Julyan. The fine print on tempered posteriors. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *PMLR*, pages 1087–1102, 2024.

Rahimi, Kobi, Refael, Yehonathan, Tirer, Tom, and Lindenbaum, Ofir. Unveiling multiple descents in unsupervised autoencoders, 2025. URL `https://arxiv.org/abs/2406.11703`.

Rosset, Lorenzo, Netti, Roberto, Muntoni, Anna Paola, Weigt, Martin, and Zamponi, Francesco. adabmDCA 2.0—A flexible but easy-to-use package for direct coupling analysis. In *Protein Evolution*, Methods in Molecular Biology, pages 83–104. Humana, New York, NY, 2026. doi: 10.1007/978-1-0716-4828-5_6.

Russ, William P., Figliuzzi, Matteo, Stocker, Christian, Barrat-Charlaix, Pierre, Socolich, Michael, Kast, Peter, Hilvert, Donald, Monasson, Remi, Cocco, Simona, Weigt, Martin, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Saatchi, Yunus and Wilson, Andrew Gordon. Bayesian GAN. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://arxiv.org/abs/1705.09558`.

Salakhutdinov, Ruslan and Murray, Iain. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th International Conference on Machine Learning*, pages 872–879, 2008. doi: 10.1145/1390156.1390266.

Seung, Hyunjune S., Sompolinsky, Haim, and Tishby, Naftali. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992. doi: 10.1103/PhysRevA.45.6056.

Simon, Jamie, Kunin, Daniel, Atanasov, Alexander, Boix-Adserà, Enric, Bordelon, Blake, Cohen, Jeremy, Ghosh, Nikhil, Guth, Florentin, Jacot, Arthur, Kamb, Mason, Karkada, Dhruva, Michaud, Eric J., Ottlik, Berkan, and Turnbull, Joseph. There will be a scientific theory of deep learning, 2026. URL `https://arxiv.org/abs/2604.21691`.

Steinberg, Julia, Adomaitytė, Urté, Fachechi, Alberto, Mergny, Pierre, Barbier, Damien, and Monasson, Rémi. Replica method for computational problems with randomness: principles and illustrations. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104002, 2024.

Thériault, Robin, Tosello, Francesco, and Tantari, Daniele. Modeling structured data learning with restricted boltzmann machines in the teacher–student setting. *Neural Networks*, 189:107542, 2025. doi: 10.1016/j.neunet.2025.107542.

Tieleman, Tijmen. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1064–1071, 2008. doi: 10.1145/1390156.1390290.

Tubiana, Jérôme and Monasson, Rémi. Emergence of compositional representations in restricted Boltzmann machines. *Physical Review Letters*, 118(13):138301, 2017. doi: 10.1103/PhysRevLett.118.138301.

Tulinski, Thomas, Fernandez-De-Cossio-Diaz, Jorge, Cocco, Simona, and Monasson, Rémi. Replica theory of spherical Boltzmann machine ensembles, 2026. URL `https://arxiv.org/abs/2604.17936`.

Urbani, Pierfrancesco. Theory of learning of high-dimensional controlled non-linear dynamical systems (I): models and methods, 2026. URL `https://arxiv.org/abs/2606.07247`.

Weigt, Martin, White, Robert A., Szurmant, Hendrik, Hoch, James A., and Hwa, Terence. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009. doi: 10.1073/pnas.0805923106.

Wenzel, Florian, Roth, Kevin, Veeling, Bastiaan S, Świątkowski, Jakub, Tran, Linh, Mandt, Stephan, Snoek, Jasper, Salimans, Tim, Jenatton, Rodolphe, and Nowozin, Sebastian. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pages 10248–10259. PMLR, 2020.

Xu, Yizhou, Krzakala, Florent, and Zdeborová, Lenka. Learning with restricted Boltzmann machines: Asymptotics of AMP and GD in high dimensions. In *Advances in Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2505.18046`.

Younes, Laurent. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989. doi: 10.1007/BF00341287.

Zambon, Alessandro, Malatesta, Enrico M., Tiana, Guido, and Zecchina, Riccardo. Sampling the space of solutions of an artificial neural network. *Phys. Rev. E*, 112:045303, 2025. doi: 10.1103/qs48-jzyq.

Zdeborová, Lenka and Krzakala, Florent. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393.

Zhang, Huijie, Huang, Zijian, Chen, Siyi, Zhou, Jinfan, Zhang, Zekai, Wang, Peng, and Qu, Qing. Understanding generalization in diffusion distillation via probability flow distance, 2025. URL `https://arxiv.org/abs/2505.20123`.

Appendix contents

A Equilibrium phase diagram **17**

A.1 Large N asymptotic expressions 17

A.2 Optimal outlier positions and general structure of the phase diagram 19

A.3 Computation of sample and weight statistics from the evidence integral 24

A.4 Replica approach to equilibrium properties of trained SBM ensembles 25

B Training dynamics: Martin–Siggia–Rose derivation **28**

B.1 Upper bound on the outlier count 31

B.2 MAP limit of the dynamical threshold 31

C Nearly finite-dimensional data **33**

C.1 Statics 33

C.2 Dynamics 33

D Teacher-student scenario **39**

D.1 Teacher-generated training data 39

D.2 Posterior predictive as a ratio of evidence integrals 39

D.3 MAP and flat limits of the posterior predictive 39

D.4 Posterior-predictive entropy 40

D.5 Reverse cross-entropies 41

D.6 Forward cross-entropies 41

D.7 KL divergences 41

D.8 Posterior-predictive gap identities 42

D.9 Sampling temperature tuning 42

D.10 Double descent 43

D.11 Tempered-posterior effects 44

D.12 Training dynamics of the KL divergences 46

D.13 Posterior-predictive forward KL 49

E Comparison of the SBM with a multivariate Gaussian model **50**

F First-order SBM reduction of a general spherical EBM **51**

G Computational experiments in generative models **53**

G.1 Experiments in small tractable BMs 54

G.2 Experiments in small tractable RBMs 59

G.3 Sampling temperature tuning in Potts BMs trained on protein sequences 64

G.4 Double descent 66

G.5 Tempered posterior effects 68

G.6 Out-of-equilibrium training dynamics in a Potts Boltzmann machine 70

A Equilibrium phase diagram

A.1 Large N asymptotic expressions

We consider an asymptotic expansion of the evidence at large N,

$$\mathcal{Y}(C) = \int \exp\left\{ -\tfrac{N}{4}\gamma\eta \,\mathrm{tr}\,\Lambda^2 - \eta\,\mathrm{tr}(C)\ln Z(\Lambda) + \ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) + \sum_{i<j} \ln|\lambda_i - \lambda_j| \right\} \mathrm{d}\Lambda. \quad (15)$$

Note that if $\mathrm{tr}(C) = K$ we recover the definition used in the main text. Here it will be convenient to consider a generalization where C is not required to satisfy this constraint.

The dominant terms are $\mathcal{O}(N^2)$ and can be written:

$$\frac{\ln \mathcal{Y}(C)}{N^2} \sim -\frac{\gamma\eta}{4}\int \lambda^2 \rho(\lambda)\mathrm{d}\lambda + \frac{1}{2}\int \ln|\lambda - \lambda'|\rho(\lambda)\rho(\lambda')\mathrm{d}\lambda\mathrm{d}\lambda' \sim -\frac{3}{8} + \frac{1}{2}\ln\sigma \quad (16)$$

up to terms of lower-order, where $\rho(\lambda)$ is given by Eq. (7), with $\sigma = 1/\sqrt{\gamma\eta}$. Thus the leading $\mathcal{O}(N^2)$ contribution comes from standard Coulomb-gas terms. Extremization of (16) with respect to ρ results in the Wigner semi-circle law (7), that describes the bulk of eigenvalues (substituting Eq. (7) gives the evaluated result on the right-hand side of Eq. (16)). It is convenient to define:

$$\begin{aligned} G(z) &= \int \frac{\rho(\lambda)\mathrm{d}\lambda}{z-\lambda} = \frac{1}{2\sigma}\left(\frac{z}{\sigma} - \sqrt{\frac{z^2}{\sigma^2} - 4}\right), \\ F(z) &= \int \rho(\lambda)\ln(z-\lambda)\mathrm{d}\lambda = \frac{\sigma^2}{2}G^2(z) - \ln G(z), \\ B(z) &= \int \frac{\lambda}{z-\lambda}\rho(\lambda)\mathrm{d}\lambda = \frac{z}{2\sigma}\left(\frac{z}{\sigma} - \sqrt{\frac{z^2}{\sigma^2} - 4}\right) - 1 = zG(z) - 1. \end{aligned} \quad (17)$$

We will use the notation $g_k = G(\lambda_k)$.

To capture the phenomenology of the outliers, it is important to consider the $\mathcal{O}(N)$ contribution. We write $\ln Y = N^2(\dots) + N\Phi(C) + \mathcal{O}(1)$. Once the K outliers are separated, the bulk integral runs over $M = N - K$ eigenvalues, and expanding its Coulomb-gas free energy in K/N produces an $\mathcal{O}(N)$ boundary correction $\frac{K}{2}\ln(\gamma\eta)$ (an accompanying additive $K/2$ is absorbed into the HCIZ normalization convention used below). We retain the Gaussian base-measure constant in $Z(\Lambda)$. Collecting all order-N contributions,

$$\Phi(C) = \frac{K}{2}\ln(\gamma\eta) + \sum_{i=1}^{K}\left(-\frac{\gamma\eta}{4}\lambda_i^2 + F(\lambda_i)\right) - \frac{\eta}{N}\,\mathrm{tr}(C)\ln Z(\Lambda) + \frac{1}{N}\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \quad (18)$$

In the following sections we evaluate $\ln Z(\Lambda)$ and $\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C)$ to leading order in N.

A.1.1 Evaluation of $Z(W)$ and condensation phase transition

Following standard steps (*e.g.* Ch. 5 in [Baxter, 1990]):

$$Z \propto \int_{-\infty}^{\infty} \mathrm{d}\mathbf{x} \int_{c-i\infty}^{c+i\infty} \mathrm{d}\mu \exp\left(\frac{1}{2}\sum_{i=1}^{N}(\lambda_i - \mu)x_i^2 + \frac{N}{2}\mu\right) \quad (19)$$

where μ traverses a line in the complex plane parallel to the imaginary axis with constant real part, $\Re\mu = c$ larger than all eigenvalues. In anticipation of the condensation along the top modes [Kosterlitz et al., 1976, Steinberg et al., 2024], we introduce order parameters $\mathbf{h} = (h_1, \dots, h_K)$

with $h_i = x_i/\sqrt{N}$, along the top K modes. From (19),

$$Z \propto \int \mathrm{d}h_1 \dots \mathrm{d}h_K \, \mathrm{d}x_{K+1} \dots \mathrm{d}x_N \, \mathrm{d}\mu \tag{20}$$

$$\exp\left(\frac{N}{2}\sum_{i=1}^{K}(\lambda_i - \mu)h_i^2 + \frac{1}{2}\sum_{i=K+1}^{N}(\lambda_i - \mu)x_i^2 + \frac{N\mu}{2}\right) \tag{21}$$

$$\propto \int \mathrm{d}h_1 \dots \mathrm{d}h_K \, \mathrm{d}\mu \, \exp\left\{\frac{N}{2}\left(\sum_{i=1}^{K}(\lambda_i - \mu)h_i^2 - F(\mu) + \mu\right)\right\}. \tag{22}$$

For large N, the integrations over $h_1, \dots, h_K$ and μ concentrate at a saddle-point, satisfying the stationarity conditions:

$$1 - h^2 = G(\mu), \quad (\lambda_i - \mu)h_i = 0 \tag{23}$$

where $h^2 = \sum_{i=1}^{K} h_i^2$. The solution of these equations can be written

$$h^2 = \begin{cases} 0, & \\ 1 - g_1, & \end{cases} \qquad \mu = \begin{cases} G^{-1}(1) & \text{if } g_1 \geq 1, \\ \lambda_1 & \text{if } g_1 \leq 1, \end{cases} \tag{24}$$

where $z = G^{-1}(a)$ for $a > 0$ denotes the unique root of the equation $G(z) = a$ satisfying $z \geq \lambda_{K+1}$.

Evaluating (20) at the saddle point yields

$$\frac{1}{N}\ln Z \sim \frac{\ln(2\pi) + \mu - F(\mu)}{2} = \frac{1}{2}\ln(2\pi) + \begin{cases} \frac{1}{2} + \frac{1}{4\gamma\eta} & g_1 \geq 1, \\ \frac{1}{2}\left(\frac{g_1}{\gamma\eta} + \frac{1}{g_1} - \frac{g_1^2}{2\gamma\eta} + \ln g_1\right) & g_1 \leq 1. \end{cases} \tag{25}$$

to leading order in N, where the additive $\frac{1}{2}\ln(2\pi)$ comes from the Gaussian base measure of (19) and is the constant carried explicitly by the teacher partition function (125).

To compute the **average energy** we use:

$$\langle\langle \mathcal{E}(\mathbf{x}; W)\rangle_{\mathbf{x}\sim P_W}\rangle_{W\sim P_\eta} = -\left\langle \frac{\partial \ln Z(\beta W)}{\partial \beta}\bigg|_{\beta=1}\right\rangle_{W\sim P_\eta} = \frac{N}{2}(1-\mu) \tag{26}$$

to leading order in N. The **entropy** of the SBM is given by,

$$\langle \mathcal{H}[P_W]\rangle_{W\sim P_\eta} = \langle\langle \mathcal{E}(\mathbf{x}; W)\rangle_{\mathbf{x}\sim P_W} + \ln Z(W)\rangle_{W\sim P_\eta} = \frac{N}{2}[\ln(2\pi e) - F(\mu)] \tag{27}$$

A.1.2 The real spherical HCIZ integral and its large-N asymptotics

The real (orthogonal-group) spherical HCIZ integral is defined as

$$\mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) = \int_{V\in \mathrm{O}(N)} \mathcal{D}V \exp\left(\tfrac{N\eta}{2}\,\mathrm{tr}\, CV\Lambda V^\top\right). \tag{28}$$

where $\mathrm{O}(N)$ is the group of $N \times N$ orthogonal matrices and $\mathcal{D}V$ is the flat Haar measure on $\mathrm{O}(N)$. For K fixed while $N \to \infty$, its leading-$\mathcal{O}(N)$ asymptotics [Guionnet and Husson, 2022] take the form

$$\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \underset{N\to\infty}{\sim} \frac{N}{2}\sum_{k=1}^{K}\left\{\eta c_k(\lambda_k u_k^2 + \chi_k - \chi_k u_k^2) - F(\chi_k) - \ln \eta c_k\right\}, \tag{29}$$

with saddle-point variables (u_k^2, χ_k) obtained by extremization of this expression. The stationarity conditions result in two branches:

$$\chi_k = \begin{cases} \lambda_k, & \\ G^{-1}(\eta c_k), & \end{cases} \qquad u_k^2 = \begin{cases} 1 - g_k/(\eta c_k) & \text{if } g_k \leq \eta c_k, \\ 0 & \text{if } g_k \geq \eta c_k. \end{cases} \tag{30}$$

Substituting the stationary values, (29) simplifies to

$$\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \underset{N\to\infty}{\sim} \frac{N}{2}\sum_{k=1}^{K}\left\{\eta c_k \chi_k - F(\chi_k) - \ln \eta c_k\right\}. \tag{31}$$

These asymptotics also cover coincident limiting outlier positions, by the continuity and uniform-approximation results of [Guionnet and Husson, 2022].

A.1.3 Evaluated form of Φ

Substituting the saddle values of $\ln Z(\Lambda)$ (25) and $\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C)$ (29) into (18) and extremizing over χ_k gives $\Phi(C)$ as a closed-form function of the data spectrum. For a source C with eigenvalues $c_1, \ldots, c_K$ ($K = \operatorname{rank} C$, $\operatorname{tr} C = \sum_k c_k$),

$$\Phi(C) = \frac{K}{2} \ln(\gamma\eta) - \frac{\eta}{2}[\ln(2\pi) + \mu - F(\mu)] \operatorname{tr} C + \sum_{k=1}^{K} \psi_k, \tag{32}$$

where the per-mode contribution ψ_k is

$$\psi_k = -\frac{\gamma\eta}{4}\lambda_k^2 + F(\lambda_k) + \frac{\eta c_k \chi_k - F(\chi_k) - \ln(\eta c_k)}{2} \tag{33}$$

$$= \begin{cases} \frac{\eta c_k^2}{4\gamma} - \frac{1}{2}\ln(\gamma\eta), & \text{uncondensed } (k > d), \\ -\frac{1}{2} - \frac{\gamma\eta}{4g_1^2} + \frac{c_k g_1}{2\gamma} + \frac{\eta c_k}{2g_1} - \frac{\ln(g_1\, \eta c_k)}{2}, & \text{condensed } (k \le d), \end{cases} \tag{34}$$

with λ_k and χ_k determined by the phase diagram. The first row applies to a mode in uncondensed position (λ_k either at the bulk edge or detached at $1/(\eta c_k) + c_k/\gamma$). The second row applies to aligned condensed modes with $\lambda_k = \chi_k = \lambda_1$ and $g_1 = G(\lambda_1)$; it spans the edge sub-phase ($g_1 = \sqrt{\gamma\eta}$, λ_1 pinned at the bulk edge) and outlier sub-phase ($g_1 < \sqrt{\gamma\eta}$, λ_1 detached at $1/g_1 + g_1/(\gamma\eta)$) of the $h, u_1 \neq 0$ phase continuously.

A.2 Optimal outlier positions and general structure of the phase diagram

We now optimize (18) with respect to the outlier eigenvalue positions. The optimal positions of $\lambda_1, \ldots, \lambda_K$ must be stable under small perturbations $\lambda_1 + \epsilon_1, \ldots, \lambda_K + \epsilon_K$.

Suppose a stable configuration has $\lambda_1 = \cdots = \lambda_d$ of degeneracy d, while $\lambda_d > \lambda_{d+1}$. To first-order in the perturbations ϵ_k, the variation of the objective is:

$$-\frac{\eta}{2} h^2 \epsilon_* \operatorname{tr} C + \sum_k \left\{ -\frac{\gamma\eta}{2}\lambda_k + \frac{\eta}{2} c_k u_k^2 + g_k \right\} \epsilon_k, \tag{35}$$

where $\epsilon_* = \max\{\lambda_1 + \epsilon_1, \ldots, \lambda_d + \epsilon_d\} - \lambda_1 \le \max\{\epsilon_1, \ldots, \epsilon_d\}$ (the inequality follows from $\lambda_1 \ge \lambda_k$ for all k and it is proved below). After some algebra,

$$-\frac{\eta \operatorname{tr} C}{2} h^2 \epsilon_* + \sum_{k \le d} \left\{ -\frac{\gamma\eta}{2g_1} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_1}{2} \right\} \epsilon_k + \sum_{k > d} \left\{ -\frac{\gamma\eta}{2g_k} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_k}{2} \right\} \epsilon_k \,. \tag{36}$$

The ϵ_k for $k > d$ can be varied independently. They are free if $\lambda_k > \frac{2}{\sqrt{\gamma\eta}}$, in which case we must have $-\frac{\gamma\eta}{2g_k} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_k}{2} = 0$. On the other hand if $\lambda_k = \frac{2}{\sqrt{\gamma\eta}}$, we can only perturb $\epsilon_k \ge 0$, so we only require $-\frac{\gamma\eta}{2g_k} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_k}{2} \le 0$. Solving these conditions gives the positions of uncondensed eigenvalues (*i.e.* $\lambda_k < \mu$ for $k > d$):

$$\begin{aligned} &\lambda_k = \frac{2}{\sqrt{\gamma\eta}}, && g_k = \sqrt{\gamma\eta}, \quad u_k^2 = 0, && \text{if } \eta c_k \le \sqrt{\gamma\eta}, \\ &\lambda_k = \frac{1}{\eta c_k} + \frac{c_k}{\gamma}, && g_k = \frac{\gamma}{c_k}, \quad u_k^2 = 1 - \frac{\gamma}{\eta c_k^2}, && \text{if } \eta c_k \ge \sqrt{\gamma\eta}. \end{aligned} \tag{37}$$

Now we focus on the condensed part, $k \le d$. If $h^2 = 0$ we have the same conditions as before and these eigenvalues are also in their uncondensed positions. So let's assume we are in the $h \neq 0$ phase with $h^2 > 0$. We then have $h^2 = 1 - g_1$ and $u_k^2 = 1 - \frac{g_1}{\eta c_k}$, and in this case λ_1 is an outlier so $\epsilon_1, \ldots, \epsilon_d$ are free. Then

$$-\frac{\eta \operatorname{tr} C}{2} h^2 \epsilon_* + \sum_{k \le d} \left\{ -\frac{\gamma\eta}{2g_1} + \frac{\eta}{2} c_k u_k^2 + \frac{g_1}{2} \right\} \epsilon_k \tag{38}$$

simplifies to

$$-\eta \epsilon_* (1 - g_1) \operatorname{tr} C - \frac{\gamma\eta}{g_1} \sum_{k \le d} \epsilon_k + \eta \sum_{k \le d} c_k \epsilon_k. \tag{39}$$

Note that $\eta c_k \geq \frac{\gamma\eta}{g_1}$ for $k \leq d$, because these eigenvalues are outliers that have been pushed away from their uncondensed positions by μ towards the bulk. Therefore

$$-\frac{\gamma\eta}{g_1}\sum_{k\leq d}\epsilon_k + \eta\sum_{k\leq d} c_k\epsilon_k \geq 0,$$

so the best perturbation we can do (in the sense of trying to maximize the objective) is to set all $\epsilon_k = \epsilon_*$ equal to the maximum. Then, setting $\epsilon_1 = \cdots = \epsilon_d$ all equal, gives the condition

$$-(1-g_1)\operatorname{tr} C - \frac{d\gamma}{g_1} + \sum_{k\leq d} c_k = 0 \tag{40}$$

that solves for the optimal position of the coalesced eigenvalues (given in terms of their Stieltjes transform g_1):

$$g_1 = \frac{1}{2\operatorname{tr} C}\left\{\sum_{k=d+1}^{K} c_k + \sqrt{\left(\sum_{k=d+1}^{K} c_k\right)^2 + 4d\gamma\operatorname{tr} C}\right\}. \tag{41}$$

An auxiliary inequality. The first-order perturbation analysis above invokes the bound $\epsilon_* \leq \max_k \epsilon_k$ on the quantity

$$\epsilon_* = \max\{\lambda_1+\epsilon_1, \ldots, \lambda_K+\epsilon_K\} - \lambda_1, \tag{42}$$

where $\lambda_1 = \max\{\lambda_1, \ldots, \lambda_K\}$ because the eigenvalues are sorted in decreasing order. Setting $j = \operatorname{argmax}_k\{\lambda_k + \epsilon_k\}$,

$$\max_k\{\lambda_k+\epsilon_k\} = \lambda_j + \epsilon_j \leq \lambda_1 + \epsilon_j \leq \lambda_1 + \max_k \epsilon_k, \tag{43}$$

and subtracting λ_1 from both sides yields $\epsilon_* \leq \max_k \epsilon_k$ as claimed.

A.2.1 Concavity and uniqueness of the equilibrium saddle

The previous argument establishes a saddle-point of the action that determines the positions of the outlier eigenvalues. However, this does not rule out the existence of competing saddle-points. In this section, we show by a concavity argument that the saddle-point we described above is in fact the unique global maximum.

After fixing the semicircular bulk, the remaining order-N variational problem can be written, up to terms independent of the saddle variables, as

$$\Phi(C) = \max_{\mu\geq\lambda_1\geq\cdots\geq\lambda_K\geq 2\sigma} \mathcal{J}(\mu,\lambda_1,\ldots,\lambda_K) + \text{const.}, \tag{44}$$

where

$$\mathcal{J}(\mu,\lambda_1,\ldots,\lambda_K) = -\frac{\eta}{2}\operatorname{tr} C\big[\mu - F(\mu)\big] + \sum_{k=1}^{K}\psi_k(\lambda_k). \tag{45}$$

Here $\psi_k(\lambda_k)$ is defined in Eq. (33), with χ_k set to its saddle value in Eq. (30), before optimizing over λ_k.

We now show that $\mathcal{J}$ is strictly concave. From Eq. (17),

$$F'(z) = G(z), \qquad G'(z) = -\int \frac{\rho(x)}{(z-x)^2}\,dx < 0, \qquad z > 2\sigma. \tag{46}$$

The dependence on μ therefore satisfies

$$\frac{\partial\mathcal{J}}{\partial\mu} = -\frac{\eta}{2}\operatorname{tr} C\,\big[1-G(\mu)\big], \qquad \frac{\partial^2\mathcal{J}}{\partial\mu^2} = \frac{\eta}{2}\operatorname{tr} C\,G'(\mu) < 0. \tag{47}$$

The constrained maximization over $\mu \geq \lambda_1$ consequently gives $G(\mu) = 1$ when the unconstrained maximum is feasible, and $\mu = \lambda_1$ otherwise, reproducing Eq. (24).

For the outlier contribution, using $g_k = G(\lambda_k)$ and the two HCIZ branches in Eq. (30), one obtains

$$\psi_k'(\lambda_k) = \begin{cases} -\dfrac{\gamma\eta}{2}\lambda_k + g_k, & g_k \geq \eta c_k, \\ -\dfrac{\gamma\eta}{2}\lambda_k + \dfrac{g_k + \eta c_k}{2}, & g_k \leq \eta c_k. \end{cases} \tag{48}$$

Equivalently, the derivative of the optimized HCIZ contribution is zero on the $u_k = 0$ branch and $(\eta c_k - g_k)/2$ on the $u_k \neq 0$ branch. Away from the switching point,

$$\psi_k''(\lambda_k) = \begin{cases} -\dfrac{\gamma\eta}{2} + G'(\lambda_k), & g_k > \eta c_k, \\ -\dfrac{\gamma\eta}{2} + \dfrac{1}{2}G'(\lambda_k), & g_k < \eta c_k, \end{cases} \quad < 0. \tag{49}$$

At $g_k = \eta c_k$, the two expressions in Eq. (48) coincide. Thus $\psi_k'(\lambda_k)$ is continuous and strictly decreasing throughout $\lambda_k \geq 2\sigma$, and ψ_k is globally strictly concave, including across the alignment transition.

It follows that $\mathcal{J}$ is strictly concave on the convex domain

$$\mu \geq \lambda_1 \geq \cdots \geq \lambda_K \geq 2\sigma. \tag{50}$$

Moreover, the objective is coercive and its maximum is attained. Strict concavity then implies that this maximizer is unique. In particular, every feasible solution of the corresponding stationarity and complementarity conditions is necessarily the global maximizer.

A.2.2 Phase diagram

The different phases are given by:

- $h = u_{1,\ldots,K} = 0$. If $\gamma\eta > 1$ and $\gamma > \eta c_1^2$. The eigenvalues stick to the edge, $\lambda_1 = \cdots = \lambda_K = 2\sigma$.
- $h = 0$, $u_{1,\ldots,a} \neq 0$. If $\gamma > c_1$, $\gamma\eta > 1$, and $\eta c_{a+1}^2 < \gamma < \eta c_a^2$ for some $a \in [K]$. The top a eigenvalues detach from the bulk at

  $$\lambda_k = \frac{1}{\eta c_k} + \frac{c_k}{\gamma}, \qquad u_k^2 = 1 - \frac{\gamma}{\eta c_k^2} \qquad (1 \leq k \leq a), \tag{51}$$

  while $\lambda_{a+1} = \cdots = \lambda_K = 2\sigma$ stick to the edge with $u_{a+1} = \cdots = u_K = 0$.
- $h \neq 0$, $u_{1,\ldots,K} = 0$. If $\eta^2 c_1^2 < \gamma\eta < 1$, then $h^2 = 1 - g_1$. All eigenvalues stick to the edge, $\lambda_1 = \cdots = \lambda_K = 2\sigma$.
- $h \neq 0$, $u_{1,\ldots,d} \neq 0$ (edge). If

  $$\eta c_{d+1} < \sqrt{\gamma\eta} < \min\{1, \eta c_d\} \quad \text{and} \quad \sqrt{\gamma\eta}\,(\eta \operatorname{tr} C - d) + \eta \sum_{k=1}^{d} c_k < \eta \operatorname{tr} C, \tag{52}$$

  for some $d \in [K]$ (taking $c_{d+1} = 0$ if $d = K$), then all eigenvalues stick to the edge, $\lambda_1 = \cdots = \lambda_K = 2\sigma$, the top d of them align with the data,

  $$u_k^2 = 1 - \frac{1}{c_k}\sqrt{\frac{\gamma}{\eta}} \quad (k \leq d), \tag{53}$$

  and $h \neq 0$.
- $h \neq 0$, $u_{1,\ldots,d} \neq 0$ (outlier). If

  $$\frac{\gamma}{c_d} < g_1 < \min\left\{1,\, \eta c_d,\, \frac{\gamma}{c_{d+1}},\, \sqrt{\gamma\eta}\right\}, \tag{54}$$

  for $d \in [K]$ (with $c_{d+1} = 0$ when $d = K$), with g_1 given by Eq. (41), the top d eigenvalues coalesce at a common outlier position obtained by inverting $g_1 = G(\lambda_1)$. Then $h^2 = 1 - g_1$ and $u_k^2 = 1 - g_1/(\eta c_k)$ for $k \leq d$. For the remaining modes $k > d$, if $\eta c_k^2 \geq \gamma$, then $\lambda_k = 1/(\eta c_k) + c_k/\gamma$ and $u_k^2 = 1 - \gamma/(\eta c_k^2)$; otherwise, $\lambda_k = 2/\sqrt{\gamma\eta}$ and $u_k^2 = 0$.

Figure 7 shows how the $K = 2$ phase diagrams deform as c_1 varies across the allowed range $1 < c_1 < 2$. Near the symmetric limit $c_1 = c_2 = 1$ the $h \neq 0$ regions with $d = 2$ (both edge and outlier) are largest; as c_1 increases toward 2 these regions shrink and the outlier–edge boundary (teal) flattens.

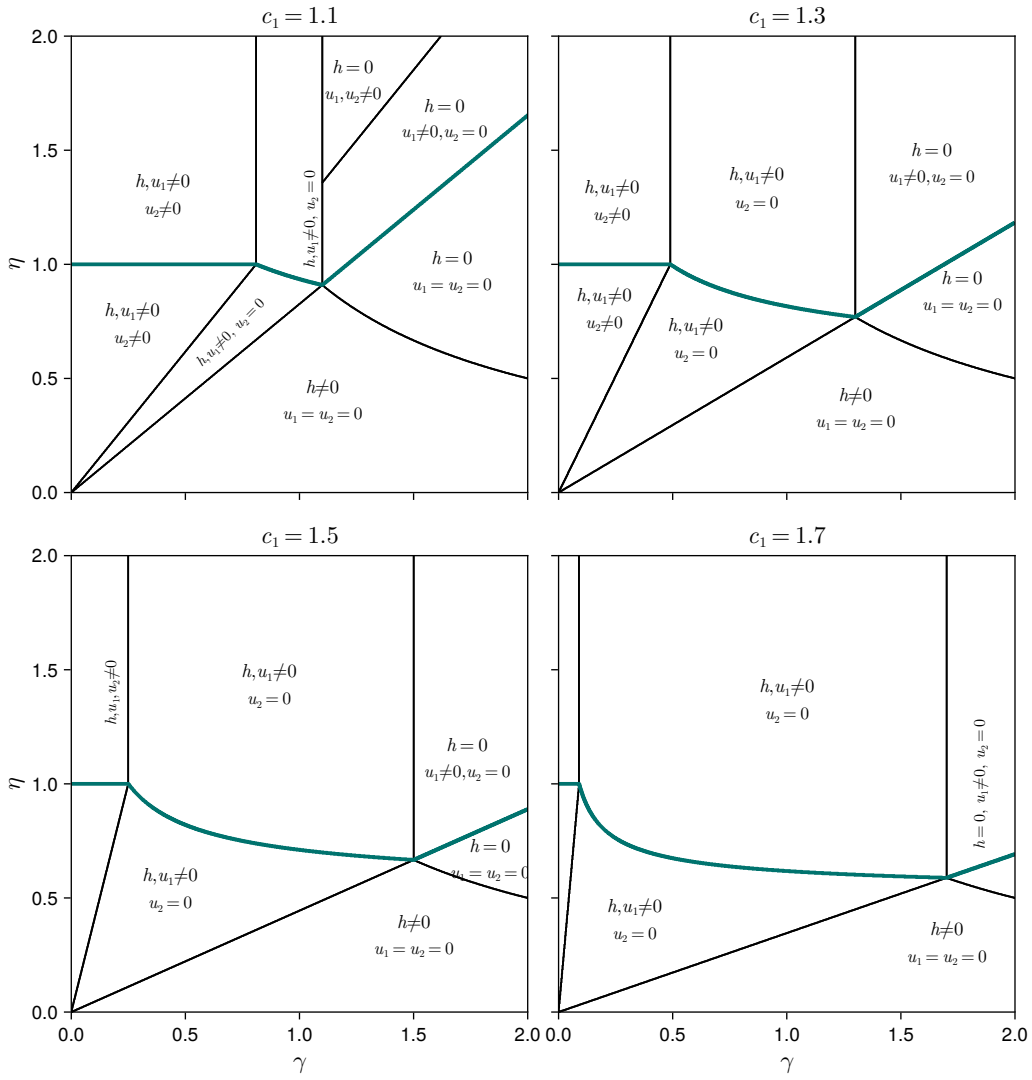


Figure 7: $K = 2$ **phase diagrams.** Phase diagram in the (γ, η) plane for $c_1 \in \{1.1, 1.3, 1.5, 1.7\}$ (with $c_2 = 2 - c_1$). Black lines are phase boundaries; the teal line separates outlier from edge phases.

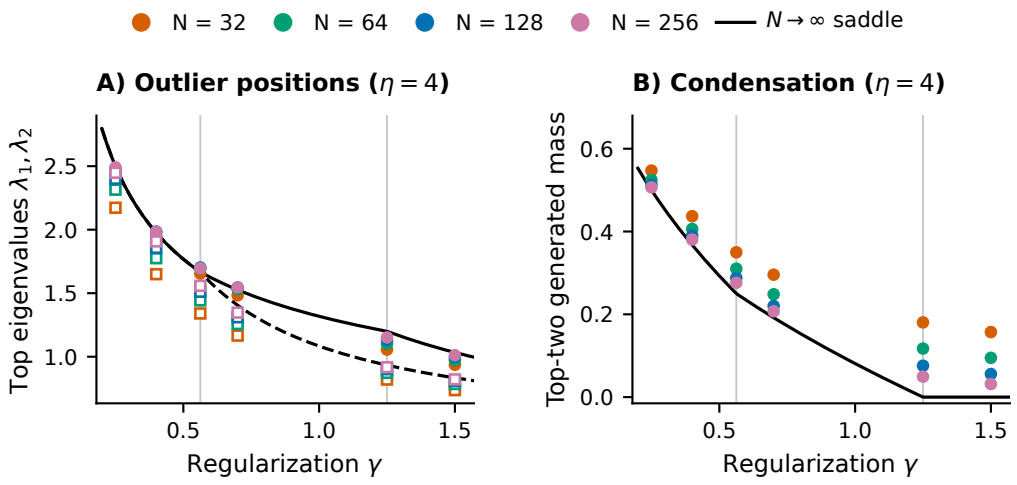


Figure 8: **Finite-size equilibrium validation.** We sample the finite-N posterior with $K = 2$ and $(c_1, c_2) = (1.25, 0.75)$. Colors denote $N = 32, 64, 128, 256$ and bars show 95% Monte Carlo intervals. **A)** Top eigenvalues at $\eta = 4$ (filled circles: λ_1; open squares: λ_2); black solid and dashed curves give the saddle values of λ_1 and λ_2. **B)** Generated mass $\sum_{j=1}^{2} \langle (\mathbf{x} \cdot \mathbf{v}_j/N)^2 \rangle$ compared with the limiting h^2. The uncondensed modes contribute a nonzero finite-size baseline. Gray vertical lines mark the coalescence and condensation boundaries.

A.2.3 Finite-size equilibrium validation

We sample the finite-N posterior (5) with the exchange algorithm [Murray et al., 2006]. We write $W = A + tI$ with $\operatorname{tr} A = 0$. The spherical constraint $\|\mathbf{x}\|^2 = N$ gives $Z(A + tI) = e^{Nt/2} Z(A)$, so the normalized spherical law is independent of t. The posterior therefore factorizes into the law of the traceless matrix A and an independent $t \sim \mathcal{N}(0, 2/(N^2 \gamma \eta))$. We update A using proposals reversible with respect to the Gaussian factor of its posterior. Each proposal A' draws $b = K\eta$ independent samples $\mathbf{y}^r \sim P(\mathbf{x} \mid A')$ by exact Bingham rejection sampling [Kent et al., 2018] and is accepted with probability

$$\min \left\{ 1, \exp \left[\frac{1}{2} \sum_{r=1}^{b} (\mathbf{y}^r)^\top (A - A') \mathbf{y}^r \right] \right\}. \tag{55}$$

Here $K\eta$ is integer. The auxiliary densities cancel the partition functions, retaining the finite-N matrix posterior without a saddle approximation. We restore the independent trace when measuring weight observables. Controls against exact $N = 2$ quadrature and the GOE spectrum at $K = \eta = 1$ agree within 1.6 Monte Carlo standard errors.

We compare $K = 2$, $(c_1, c_2) = (1.25, 0.75)$ and $N \in \{32, 64, 128, 256\}$ at $\eta = 4$, with $\gamma \in \{0.25, 0.4, 0.5625, 0.7, 1.25, 1.5\}$ spanning the coalescence and condensation boundaries. Four chains start from different spectra and alignments. For $N = 32, 64, 128$, we use 20,000 warmup transitions followed by 150,000, 300,000 and 600,000 production transitions, respectively. For $N = 256$, we use 100,000 warmup transitions followed by 1.2×10^6 production transitions (2.4×10^6 at $\gamma = 0.25$). We freeze proposal adaptation after warmup, retain every twentieth state, and use sixteen fresh spherical samples per retained state to measure the generated mass. Across the plotted settings, the maximum rank-normalized split or folded $\widehat{R}$ is 1.0094 and the minimum bulk or tail effective sample size is 1112, meeting thresholds 1.01 and 400. A separate four-chain replication at $(N, \eta, \gamma) = (128, 4, 0.25)$ uses 50,000 warmup transitions and the same production length; it meets the same gates, with a largest difference of 0.98 combined Monte Carlo standard errors across eleven prespecified observables (Bonferroni threshold 2.84).

Figure 8 separates Monte Carlo uncertainty from finite-size discrepancies. At $N = 256$, the maximum discrepancies from the saddle are 6.5% (relative) for the eigenvalues and 0.049 (absolute) for the

generated mass. These measurements test selected equilibrium observables, with finite-size rounding near phase boundaries; they neither estimate the absolute evidence nor replace the uniqueness argument in App. A.2.1.

A.3 Computation of sample and weight statistics from the evidence integral

In this Appendix we establish the following two formulas:

$$\langle W\rangle_{W\sim P_\eta} = \frac{1}{N\eta}\sum_{k=1}^{K}\left(\eta\chi_k - \frac{1}{c_k}\right)\left(\mathbf{c}_k\mathbf{c}_k^\top - I\right). \tag{56}$$

and

$$\frac{K}{N}\left\langle\left\langle \mathbf{x}\mathbf{x}^\top\right\rangle_{\mathbf{x}\sim P_W}\right\rangle_{W\sim P_\eta} = C - \frac{\gamma}{N\eta}\sum_{k=1}^{K}\left(\eta\chi_k - \frac{1}{c_k}\right)\left(\mathbf{c}_k\mathbf{c}_k^\top - I\right) \tag{57}$$

A.3.1 Perturbation theory of the evidence

Let U be a symmetric matrix of finite rank, with $\operatorname{tr} U = 0$. Calling c'_k the eigenvalues of $C' = C + zU$, first-order perturbation theory gives

$$c'_k = c_k + \frac{z}{N}\mathbf{c}_k^\top U\mathbf{c}_k + \mathcal{O}(z^2) \tag{58}$$

for $k \le K$ (the non-zero eigenvalues of C are assumed non-degenerate, with eigenvectors $\mathbf{c}_k$ normalized to $\mathbf{c}_k^\top\mathbf{c}_k = N$). Let $U' = \mathbb{P}_{\ker(C)}U\mathbb{P}_{\ker(C)}$, where $\mathbb{P}_{\ker(C)} = I - \frac{1}{N}\sum_{k=1}^{K}\mathbf{c}_k\mathbf{c}_k^\top$. Then C' gains $R = \operatorname{rank} U'$ additional non-zero eigenvalues,

$$c'_j = z\alpha_j + \mathcal{O}(z^2), \quad (j > K)$$

where α_j are the eigenvalues of U'.

The only explicit dependence on the eigenvalues of C comes from the HCIZ integral:

$$\begin{aligned}\frac{1}{N}\frac{\partial}{\partial z}\ln\mathcal{Y}(C+zU)\Big|_{z=0} &= \frac{1}{N}\left.\frac{\partial\ln\mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C+zU)}{\partial z}\right|_{z=0}\\ &= \frac{1}{2N}\sum_{k=1}^{K}\left(\eta\chi_k - \frac{1}{c_k}\right)\mathbf{c}_k^\top U\mathbf{c}_k\end{aligned} \tag{59}$$

where we used $\frac{1}{\eta c'_k}G(\chi_k) = 1 - u_k^2$ and $(\lambda_k - \chi_k)u_k^2 = 0$. For $j > K$, $c'_j \to 0$ gives $\chi_j = G^{-1}(\eta c'_j) \sim (\eta c'_j)^{-1}$, so $(\eta\chi_j - 1/c'_j)(\partial c'_j/\partial z) = \mathcal{O}(c'_j) \to 0$; the newborn modes $j > K$ drop out.

A.3.2 Weight statistics

Directly differentiating $\ln\mathcal{Y}(C + zU)$ in its integral representation gives

$$\left.\frac{\partial}{\partial z}\ln\mathcal{Y}(C+zU)\right|_{z=0} = \frac{N\eta}{2}\left\langle\operatorname{tr} UW\right\rangle_{W\sim P_\eta}. \tag{60}$$

Combined with (59),

$$\left\langle\operatorname{tr} UW\right\rangle_{W\sim P_\eta} = \frac{1}{N\eta}\sum_{k=1}^{K}\left(\eta\chi_k - \frac{1}{c_k}\right)\mathbf{c}_k^\top U\mathbf{c}_k \tag{61}$$

for any U with $\operatorname{tr} U = 0$. Since $\operatorname{tr}\langle W\rangle_{P_\eta} = 0$, this pins down the mean to Eq. (56).

A.3.3 Sample statistics

The same generating trick on the other side of the $\ln Z(W)$ factor yields sample moments under the posterior predictive. Shifting $W \to W - zU$ inside the posterior-weighted integral gives

$$\frac{\partial}{\partial z} \ln \int \exp\left\{-K\eta \ln Z(W - zU) - \frac{N\gamma\eta}{4} \operatorname{tr} W^2 + \frac{N\eta}{2} \operatorname{tr} CW\right\} \mathrm{d}W \Big|_{z=0} = \frac{K\eta}{2} \left\langle \left\langle \mathbf{x}^\top U \mathbf{x} \right\rangle_{P_W} \right\rangle_{W \sim P_\eta}. \quad (62)$$

Completing the square, the log of the integral equals $\ln \mathcal{Y}(C - z\gamma U) - \frac{N}{4} z^2 \gamma\eta \operatorname{tr} U^2 + \frac{N\eta}{2} z \operatorname{tr} CU$. Differentiating and using (59),

$$\begin{aligned} \frac{K\eta}{2} \left\langle \left\langle \mathbf{x}^\top U \mathbf{x} \right\rangle_{P_W} \right\rangle_{W \sim P_\eta} &= \frac{\partial}{\partial z} \ln \mathcal{Y}(C - z\gamma U) \Big|_{z=0} + \frac{N\eta}{2} \operatorname{tr} CU \\ &= -\frac{\gamma}{2} \sum_{k=1}^{K} \left(\eta\chi_k - \frac{1}{c_k}\right) \mathbf{c}_k^\top U \mathbf{c}_k + \frac{N\eta}{2} \operatorname{tr} CU. \end{aligned} \quad (63)$$

Since $\operatorname{tr} \langle \langle \mathbf{x}\mathbf{x}^\top \rangle_{P_W} \rangle_{P_\eta} = N$, the arbitrariness of U then pins down the full covariance to Eq. (57).

A.4 Replica approach to equilibrium properties of trained SBM ensembles

The equilibrium properties of SBM ensembles learned from data were recently derived in an independent manner using the replica method from the statistical physics of disordered systems [Tulinski et al., 2026]. We summarize here how their approach relates to ours.

Effective theory. The derivation is based on the rewriting

$$\mathcal{Y}(C) = \lim_{n \to -K\eta} \langle Z(W)^n \rangle_{W \sim \exp(-\frac{N\gamma\eta}{4} \operatorname{tr} W^2 + \frac{N\eta}{2} \operatorname{tr}(CW))}, \quad (64)$$

where n is the replica number. Pretending that n is a positive integer, $Z(W)^n$ is the partition function of a system made of n "replicas" of the SBM. The samples $\|\mathbf{x}^a\|^2 = N$ generated by replicas $a = 1, \ldots, n$ are mutually independent conditionally on W. Integrating over W creates effective interactions of replicas with each other, $Q_{ab} = \mathbf{x}_a \cdot \mathbf{x}_b / N$, and with the data, $M_{ak} = \mathbf{x}_a \cdot \mathbf{c}_k / N$, both of which are introduced as auxiliary integration variables, leading to the expression

$$\langle Z^n \rangle \propto \int \mathrm{d}M \, \mathrm{d}Q \, \exp\left(N \Phi_n^{\text{rep}}(M, Q)\right) \quad (65)$$

where

$$\begin{aligned} \Phi_{\text{rep}}(M, Q) = \frac{\eta \operatorname{tr}(C^2)}{4\gamma} + \frac{\operatorname{tr}(Q^2)}{4\gamma\eta} + \frac{1}{2\gamma} \operatorname{tr}(M \operatorname{diag}(c_k) M^\top) \\ + \frac{1}{2} \operatorname{tr} \ln(Q - MM^\top) + \frac{n}{2} \ln(2\pi e). \end{aligned} \quad (66)$$

The spherical constraint $\operatorname{tr} Q = n$ is enforced by a Lagrange multiplier $\Phi_{\text{rep}} \mapsto \Phi_{\text{rep}} - \frac{\mu}{2}(\operatorname{tr}(Q) - n)$.

Saddle-point analysis. It was shown in [Tulinski et al., 2026] that, to leading order in N, the integral $\langle Z^n \rangle$ is invariant not only under permutations of replicas $a = 1, \ldots, n$, but also under orthogonal transformations in n-dimensional replica space. As a result, the effective action Φ_{rep} depends on M and Q only through the squared norm of the $d \leq K$ nonzero columns $\mathbf{m}_1, \ldots, \mathbf{m}_d$ of the former and eigenvalues of the latter, with d determined *a posteriori* from the optimizing of Φ_{rep}.

Integrating over Q and M by saddle point reveals that

(i) columns of the optimal M are eigenvectors of the optimal Q, with eigenvalues $q_1, \ldots, q_d$;

(ii) the optimal Q is isotropic in subspace orthogonal to their span, with eigenvalue $\widetilde{q}$;

(iii) these O(n)-invariants are related via

$$\mu = \frac{q_k}{\eta\gamma} + \frac{c_k}{\gamma} = \frac{q_k}{\eta\gamma} + \frac{1}{q_k - m_k^2} = \frac{\widetilde{q}}{\eta\gamma} + \frac{1}{\widetilde{q}}, \tag{67}$$

where $m_k^2 = \|\mathbf{m}_k\|^2/n$;

(iv) it is sufficient for the spherical constraints $\|\mathbf{x}^a\|^2 = N$ to hold on average,

$$\operatorname{tr} Q = n, \tag{68}$$

by O(n)-invariance to leading order in N.

The replica order parameter m_k tracks the overlap $\mathbf{x} \cdot \mathbf{c}_k/N$ of generated samples with the data eigenvector $\mathbf{c}_k$, whereas the condensation parameter h_k of §2.1 tracks the overlap $\mathbf{x} \cdot \mathbf{v}_k/N$ with the k-th eigenvector $\mathbf{v}_k$ of W. At saddle the two are related by $m_k = h_k u_k$ with $u_k = \mathbf{v}_k \cdot \mathbf{c}_k/N$ the W-to-data alignment, so $m_k = 0$ whenever $h_k = 0$ or $u_k = 0$. The phase labels below are written in the main-text (h, u) convention of §2.1 for direct cross-referencing.

Eqs. (66)-(68) are valid provided $n > n_c$ where

$$n_c = -K\eta \, \frac{\sum_{k \le d} \frac{c_k}{K}\left(1 - \frac{1}{c_k}\sqrt{\frac{\gamma}{\eta}}\right)}{1 - \sqrt{\eta\gamma}}. \tag{69}$$

The boundary of the $h, u_1 \neq 0$ edge phase solves the equation $n(\eta) = n_c(\gamma, \eta, c_1, \ldots, c_d)$, which is equivalent to Eq. (52). For $n < n_c$, the solution $\{m_k(n), \widetilde{q}(n)\}$ obtained for $n \ge n_c$ is frozen, and the analytical continuation of $\Phi_{\text{rep}}(n)$ viewed as a function of n is $\Phi_{\text{rep}}(n) = \Phi_{\text{rep}}(n_c) + (n - n_c)\Phi'_{\text{rep}}(n_c)$, see [Pastore et al., 2019, Tulinski et al., 2026] for an interpretation of the freezing.

Eq. (68) is quadratic in $\widetilde{q}$. The solution, $\widetilde{q}(d)$, depends on d and equals the coalesced g_1, Eq. (41). The conditions determining d can be derived from the local stability of the d-coalesced saddle points [Tulinski et al., 2026], and are $\gamma/c_d < \widetilde{q}(d) < \min\{1, \sqrt{\eta\gamma}, \gamma/c_{d+1}\}$ for $n = -K\eta > n_c$, consistent with Eq. (54), and $d = \#\{c_k : c_k > \sqrt{\gamma/\eta}\}$ for $n = -K\eta < n_c$, see Eqs. (52).

From Eqs. (67), we have that $(Q - MM^\top)^{-1}$ has d eigenvalues c_k/γ along with an $(n-d)$-fold degenerate eigenvalue g_1^{-1}, and $q_k = g_1 + \eta(\gamma/g_1 - c_k)$, $m_k^2 = q_k - \gamma/c_k$. It follows that

$$\frac{\eta c_k^2}{4\gamma} + \frac{q_k^2}{4\gamma\eta} + \frac{c_k m_k^2}{2\gamma} = \begin{cases} \frac{g_1^2}{4\gamma\eta} + \frac{\gamma\eta}{4g_1^2}, & k \le d, \\ \frac{\eta c_k^2}{4\gamma}, & k > d. \end{cases} \tag{70}$$

Substituting into Φ_{rep} evaluated at $n = -K\eta > n_c$ yields

$$\begin{aligned} \Phi_{\text{rep}}(-K\eta) = {} & \frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \frac{d\gamma\eta}{4g_1^2} - \frac{Kg_1^2}{4\gamma} \\ & + \frac{1}{2}\sum_{k=1}^{d} \ln\frac{\gamma}{c_k} - \frac{K\eta + d}{2}\ln g_1 - \frac{K\eta}{2}\ln(2\pi e). \end{aligned} \tag{71}$$

We turn to the total Φ predicted by Eq. (32). Collecting logarithms from the uncondensed and condensed modes and the base measure,

$$\begin{aligned} & \frac{K}{2}\ln(\gamma\eta) - \frac{K-d}{2}\ln(\gamma\eta) - \frac{K\eta}{2}\ln g_1 - \frac{1}{2}\sum_{k\le d}\ln(g_1\eta\, c_k) - \frac{K\eta}{2}\ln(2\pi) \\ & \quad = \frac{1}{2}\sum_{k\le d}\ln\frac{\gamma}{c_k} - \frac{K\eta + d}{2}\ln g_1 - \frac{K\eta}{2}\ln(2\pi), \end{aligned}$$

which is the logarithmic part of Φ_{rep}. The non-logarithmic terms of Φ give

$$\frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \Big(\frac{\eta}{2g_1} + \frac{g_1}{2\gamma}\Big)\Big(\sum_{k\le d} c_k - K\Big) + \frac{Kg_1^2}{4\gamma} - \frac{d}{2} - \frac{d\gamma\eta}{4g_1^2} = \frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \frac{d\gamma\eta}{4g_1^2} - \frac{Kg_1^2}{4\gamma} - \frac{K\eta}{2},$$

where we used $\sum_{k\leq d} c_k - K = -Kg_1 + d\gamma/g_1$. Summing both logarithmic and non-logarithmic parts shows that $\Phi = \Phi_{\text{rep}}(n)$ for $n = -K\eta > n_c$, which describes the $h = u_1 = 0$, $h = 0, u_1 \neq 0$ and $h, u_1 \neq 0$ outlier phases.

We now turn to $n = -K\eta < n_c$. In that case, the replica solution is $\widetilde{q} = \sqrt{\gamma\eta}$, $\mu = 2/\sqrt{\gamma\eta}$, for which the transverse modes of the Hessian $\partial^2\Phi_{\text{rep}}/\partial Q_{ab}\partial M_{ak}$ are flat. The slope of Φ_{rep} is given by

$$\Phi'_{\text{rep}}(n_c) = \frac{1}{\sqrt{\gamma\eta}} + \frac{1}{4}\ln(\gamma\eta) - \frac{1}{4} + \frac{1}{2}\ln(2\pi). \tag{72}$$

Using $(1 - 1/\sqrt{\gamma\eta})n_c = \sqrt{\eta/\gamma}\sum_{k\leq d} c_k - d$, we obtain the expression

$$\begin{aligned}\Phi_{\text{rep}} = \frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \sqrt{\frac{\eta}{\gamma}}\Big(\sum_{k\leq d} c_k - K\Big) + \frac{K\eta}{4} - \frac{3d}{4} \\ + \frac{1}{2}\sum_{k\leq d}\ln\frac{\gamma}{c_k} - \frac{K\eta + d}{4}\ln(\gamma\eta) - \frac{K\eta}{2}\ln(2\pi),\end{aligned} \tag{73}$$

valid for $n = -K\eta < n_c$.

We again turn to the Φ predicted by Eq. (32). The contributions from the uncondensed and condensed modes are, respectively, $\frac{\eta}{4\gamma}\sum_{k>d} c_k^2 - \frac{K-d}{2}\ln(\gamma\eta)$ and $-\frac{3d}{4} + \sqrt{\frac{\eta}{\gamma}}\sum_{k\leq d} c_k - \frac{1}{2}\sum_{k\leq d}\ln(\sqrt{\eta\gamma}\eta c_k)$, while $-K\eta(\ln Z)/N = -K\sqrt{\eta/\gamma} - \frac{K\eta}{4}\ln(\gamma\eta) + \frac{K\eta}{4} - \frac{K\eta}{2}\ln(2\pi)$. Summing these contributions together with $\frac{K}{2}\ln(\gamma\eta)$, we see that $\Phi = \Phi_{\text{rep}}(n)$ for $n = -K\eta < n_c$ too, thus capturing the $h \neq 0, u_1 = 0$ and $h, u_1 \neq 0$ edge phases.

B Training dynamics: Martin–Siggia–Rose derivation

Equations (3)–(4) model persistent contrastive divergence in continuous time. Here t denotes parameter-training time, and the chain accumulates ν units of Langevin sampling time per unit t. The weight update uses the data covariance C and the current chain state $\mathbf{x}(t)$ for the positive and negative phases, respectively, while the same chain continues under the evolving weights without requiring equilibration at each update.

DMFT notation recap. For reference, we restate here the main-text DMFT quantities. The two-time order parameters are

$$Q(t,u) = \frac{\mathbf{x}(t)\cdot\mathbf{x}(u)}{N}, \qquad R(t,u) = \left.\frac{\delta\langle x_i(t)\rangle}{\delta j_i(u)}\right|_{j=0}, \tag{74}$$

with $j_i(t)$ a fictitious source added to the right-hand side of (4); the memory kernel and noise correlator are

$$\begin{aligned} M(t,u) &= e^{-\frac{\gamma}{2}(t-u)}\left[-\frac{K\nu}{2}Q(t,u) + \frac{\nu^2}{\eta\gamma}R(t,u)\right], \\ D(t,u) &= 2\nu\delta(t-u) + \frac{\nu^2}{\eta\gamma}e^{-\frac{\gamma}{2}|t-u|}Q(t,u). \end{aligned} \tag{75}$$

We sketch the derivation of the dynamical mean-field theory summarized in Sec. 2.2. The intermediate expressions (76)–(78) are displayed at $K = 1$ with a single data direction $\mathbf{c}$ and scalar eigenvalue c. Throughout, data eigenvectors are normalized to $\|\mathbf{c}_k\|^2 = \sum_i c_{k,i}^2 = N$, so that overlaps $\mathbf{c}_k\cdot\mathbf{v}_k/N$ are $\mathcal{O}(1)$ and the rank-one matrix $\mathbf{c}\mathbf{c}^\top/N$ in (76) carries the K=1 data covariance at unit scalar eigenvalue. The general-K reduction to Eqs. (11)–(13) follows by linearity, summing the positive-phase source over the K data modes (restoring the eigenvalue factors c_k) and restoring the factor K in the negative-phase self-interaction. The finite-N SGD comparison in Fig. 9 (including the $K = 2$ panels) runs with this multi-mode extension. Starting from the coupled Langevin equations (4)–(3), the weight equation can be integrated explicitly against the Ornstein–Uhlenbeck process $W_{\mathrm{GOE}}(t)$ that shares the same noise Ω but drops the source terms. With the stationary GOE initial condition $W(0) \sim e^{-\frac{N\gamma\eta}{4}\operatorname{tr} W^2}$, the noise correlator satisfies $\langle W_{\mathrm{GOE},ij}(t)W_{\mathrm{GOE},kl}(t')\rangle = (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})e^{-\gamma|t-u|/2}/(N\gamma\eta)$, and

$$W(t) = W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma}\frac{\mathbf{c}\mathbf{c}^\top}{N} - \frac{1}{2}\int_0^t e^{-\gamma(t-u)/2}\frac{\mathbf{x}(u)\mathbf{x}(u)^\top}{N}\,\mathrm{d}u. \tag{76}$$

Representing the Langevin equations as delta functionals with response fields $\hat{\mathbf{x}}(t)$ and $\hat{W}(t)$, averaging over the noises, and integrating out $W, \hat{W}$ using (76) yields a closed generating functional over $(\mathbf{x}, \hat{\mathbf{x}})$ alone,

$$\mathcal{Z} \propto \int \mathcal{D}[\mathbf{x},\hat{\mathbf{x}}]\, e^{-N\mathcal{A}_{\mathrm{x}}[\mathbf{x},\hat{\mathbf{x}}]}, \tag{77}$$

with action

$$\begin{aligned} N\mathcal{A}_{\mathrm{x}} = \int \mathrm{d}t\, \Big\{ &\hat{\mathbf{x}}\cdot\dot{\mathbf{x}} + \kappa(t)\hat{\mathbf{x}}\cdot\mathbf{x} - \nu\|\hat{\mathbf{x}}\|^2 - \frac{\nu}{\gamma N}(1-e^{-\gamma t/2})\,\hat{\mathbf{x}}^\top\mathbf{c}\mathbf{c}^\top\mathbf{x}\Big\} \\ &+ \frac{1}{N}\int \mathrm{d}t\,\mathrm{d}u\,\Theta(t-u)\,e^{-\gamma(t-u)/2}\Big\{\tfrac{\nu}{2}(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u)) \\ &\qquad - \tfrac{\nu^2}{\eta\gamma}(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u)) - \tfrac{\nu^2}{\eta\gamma}(\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u))(\mathbf{x}(t)\cdot\hat{\mathbf{x}}(u))\Big\}, \end{aligned} \tag{78}$$

in which the Heaviside $\Theta(t-u)$ is inherited from the causal structure of (76). The bilocal terms depend on $\mathbf{x}, \hat{\mathbf{x}}$ only through the rotation-invariant combinations $\mathbf{x}(t)\cdot\mathbf{x}(u)$, $\mathbf{x}(t)\cdot\hat{\mathbf{x}}(u)$, and $\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u)$. Introducing the order parameters Q, R of (74) and the signal s through Hubbard–Stratonovich identities, and by $\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u) = \mathcal{O}(1)$ (response fields generate no equal-time variance), the saddle reduces to a sum of decoupled one-site contributions at fixed c,

$$\mathcal{A}_{\mathrm{x}}[\mathbf{x},\hat{\mathbf{x}}] = \frac{1}{N}\sum_i \mathfrak{a}[x_i,\hat{x}_i|c_i], \tag{79}$$

from which the effective single-site Gaussian process reads

$$\dot{x}(t) = -\kappa(t)x(t) + \frac{\nu}{\gamma}(1 - e^{-\gamma t/2})s(t)\, c + \int_0^t M(t,u)\, x(u)\, \mathrm{d}u + \zeta(t), \tag{80}$$

with the kernels $M(t,u)$ and $D(t,u)$ of (75) and $\langle \zeta(t)\zeta(u)\rangle = D(t,u)$. Because (80) is linear, $\langle x(t)|c\rangle = c\, h(t)$ for a scalar $h(t)$, and the self-consistency $s(t) = \frac{1}{N}\sum_i c_i \langle x(t)|c_i\rangle = h(t)$ (using $\sum_i c_i^2 = N$) closes the signal equation. The two-time order parameters then satisfy the ODEs (11)–(13) of Sec. 2.2, supplemented by $R(t,u) = 0$ for $t < u$, $\lim_{t\to u^+} R(t,u) = 1$, $Q(t,u) = Q(u,t)$, and the sphericity constraint $Q(t,t) = 1$ which also fixes $\kappa(t)$.

Stationary regime. For long times, the system reaches a stationary, time-translation invariant, regime, for which: $Q(t,u) \to Q_{\mathrm{st}}(t-u)$, $R(t,u) \to R_{\mathrm{st}}(t-u)$, $s(t) \to s_{\mathrm{st}}$, $\kappa(t) \to \kappa_{\mathrm{st}}$, with $Q_{\mathrm{st}}(-\tau) = Q_{\mathrm{st}}(\tau)$. The stationary kernels collapse to

$$M_{\mathrm{st}}(\tau) = e^{-\gamma\tau/2}\left\{-\frac{K\nu}{2}Q_{\mathrm{st}}(\tau) + \frac{\nu^2}{\eta\gamma}R_{\mathrm{st}}(\tau)\right\}, \quad D_{\mathrm{st}}(\tau) = 2\nu\delta(\tau) + \frac{\nu^2}{\eta\gamma}e^{-\gamma|\tau|/2}Q_{\mathrm{st}}(\tau). \tag{81}$$

The stationary equations become,

$$0 = \left(-\kappa_{\mathrm{st}} + \frac{\nu c_a}{\gamma} + \int_0^\infty M_{\mathrm{st}}(\tau)\mathrm{d}\tau\right) s_{a,\mathrm{st}}, \qquad a = 1, \dots, K, \tag{82}$$

$$\partial_\tau R_{\mathrm{st}}(\tau) = -\kappa_{\mathrm{st}} R_{\mathrm{st}}(\tau) + \int_0^\tau M_{\mathrm{st}}(\tau-\sigma)R_{\mathrm{st}}(\sigma)\mathrm{d}\sigma, \tag{83}$$

$$\begin{aligned}\partial_\tau Q_{\mathrm{st}}(\tau) = \frac{\nu}{\gamma}\sum_{a=1}^K c_a s_{a,\mathrm{st}}^2 - \kappa_{\mathrm{st}}Q_{\mathrm{st}}(\tau) + \int_0^\tau M_{\mathrm{st}}(\tau-\sigma)Q_{\mathrm{st}}(\sigma)\mathrm{d}\sigma \\ + \int_0^\infty M_{\mathrm{st}}(\tau+\sigma)Q_{\mathrm{st}}(\sigma)\mathrm{d}\sigma + \int_0^\infty D_{\mathrm{st}}(\tau+\sigma)R_{\mathrm{st}}(\sigma)\mathrm{d}\sigma\end{aligned} \tag{84}$$

These equations always admit a solution with $s_{\mathrm{st},k} = 0$. The corresponding solution of $R_{\mathrm{st}}(\tau), Q_{\mathrm{st}}(\tau), \kappa_{\mathrm{st}}$ with $s_{\mathrm{st}} = 0$ will be denoted by $R_P(\tau), Q_P(\tau), \kappa_P$.

A condensed solution may also exist, satisfying $\kappa_{\mathrm{st}} = \nu c_a/\gamma + \int_0^\infty M_{\mathrm{st}}(\tau)\mathrm{d}\tau$. To locate its onset, we linearize the stationary signal equation around the uncondensed bath (R_P, Q_P, κ_P) and set $s(t) = s_0 e^{zt}$. The memory integral evaluates as $\int_0^t M_P(t-u)e^{zu}\mathrm{d}u \to e^{zt}\check{M}_P(z)$ at large t, with $\check{M}_P(z) \equiv \int_0^\infty e^{-z\tau}M_P(\tau)\,\mathrm{d}\tau$, so the signal equation reduces to

$$z + \kappa_P - \check{M}_P(z) = \frac{\nu c_a}{\gamma}. \tag{85}$$

The Laplace transform of the uncondensed-bath response equation $\partial_\tau R_P(\tau) = -\kappa_P R_P(\tau) + \int_0^\tau M_P(\tau-\sigma)R_P(\sigma)\,\mathrm{d}\sigma$ with $R_P(0) = 1$ gives

$$\check{R}_P(z) = \frac{1}{z + \kappa_P - \check{M}_P(z)}. \tag{86}$$

Substituting (86) into (85) yields the spectral equation

$$1 = \frac{\nu c_a}{\gamma}\int_0^\infty e^{-z\tau}R_P(\tau)\,\mathrm{d}\tau, \tag{87}$$

whose marginal solution $z = 0$ isolates the static condensation threshold of mode a, $\nu_{c,a} = \gamma/(c_a\chi_P)$, with $\chi_P = \int_0^\infty R_P(\tau)\,\mathrm{d}\tau$. Since $c_1 > \cdots > c_K$, mode 1 destabilizes first, fixing the dynamical phase boundary at $\nu_c = \gamma/(c_1\chi_P)$. This boundary marks the linear instability of the zero-overlap solution.

Note that if $N \to \infty$ is taken before the long-time limit with random initialization, Eq. (11) preserves $s_1(t) = 0$ even for $\nu > \nu_c$; reaching a condensed branch requires a nonzero seed or finite-N fluctuations. For this reason, we initialize $s_1(0)$ at a small nonzero value.

Numerically, the two-time system (11)–(13) is solved by causal row-by-row time marching with an implicit predictor–corrector at each new time, enforcing $Q(t_n, t_n) = 1$ through $\kappa(t_n)$.

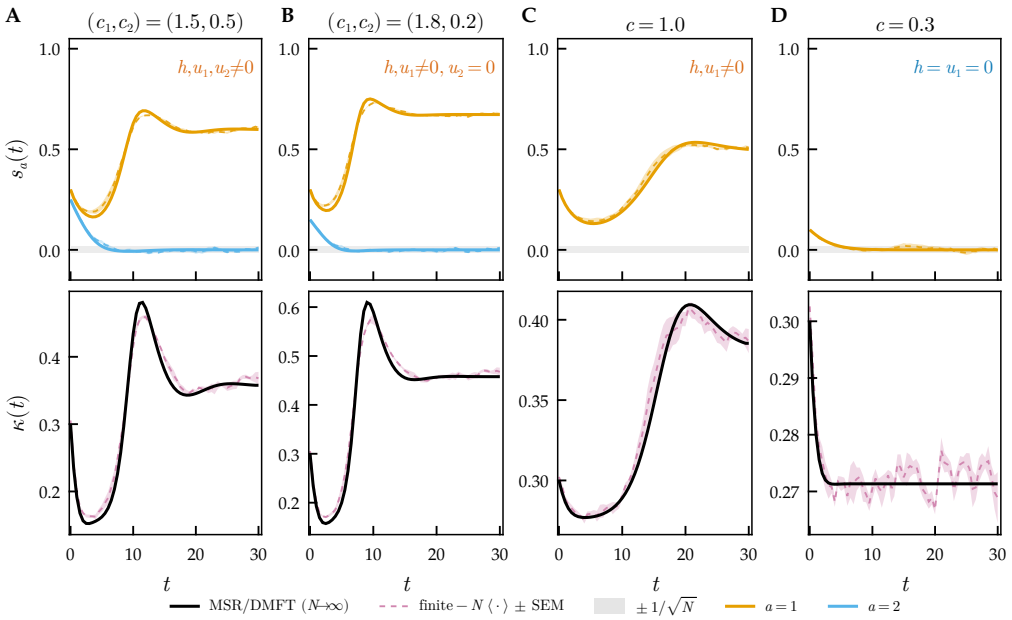


Figure 9: **Validation of the large-N DMFT against finite-N SBM training.** Each panel compares DMFT (solid) with finite-N SGD runs (full positive phase, one persistent negative-phase particle) simulated by Euler integration of the coupled Langevin system (4)–(3) at $N = 4000$, averaged over five seeds (dashed, with $\pm$SEM bands). Top row: signal overlaps $s_k(t)$; bottom row: Lagrange multiplier $\kappa(t)$. The gray strip marks the $\pm 1/\sqrt{N}$ finite-N noise floor (≈ 0.016). **A)** $K = 2$, $\mathbf{c} = (1.5, 0.5)$: one channel condenses ($s_1 \to 0.60$), the other stays at $s = 0$. **B)** $K = 2$, $\mathbf{c} = (1.8, 0.2)$: stronger signal imbalance, $s_1 \to 0.67$. **C)** $K = 1$, $c = 1$: pure $h \neq 0$ phase. **D)** $K = 1$, $c = 0.3$: pure $h = 0$ phase; both s and κ fluctuate near zero. Common parameters: $\gamma = 0.5$, $\eta = 3$, $\nu = 0.3$, $T_{\max} = 30$. Finite-N seeds are sign-aligned to the DMFT solution to absorb the $\mathbf{x} \to -\mathbf{x}$ symmetry of (4)–(3).

Numerical validation against finite-N SGD training. Figure 9 compares the DMFT solution of Eqs. (11)–(13) with finite-N SBM training simulated by Euler–Maruyama integration of the coupled Langevin system (4)–(3), using the full positive phase and one persistent negative-phase particle, with the GOE noise $\Omega/\sqrt{\eta N}$ playing the role of the stochastic-gradient noise. This validates the large-N DMFT (11)–(13) as a description of finite-N SGD training of the SBM. Four representative cases are shown, spanning $K \in \{1, 2\}$ and covering both phases; in each case we plot the signal overlaps $s_k(t)$ and the spherical-constraint multiplier $\kappa(t)$. Parameters are chosen so that the transient dip of $s_1(t)$, caused by the weight matrix W needing some time to build the signal eigenvector before $\mathbf{x}$ can condense onto it, remains larger than the $1/\sqrt{N}$ noise floor; below that floor, finite-N fluctuations dominate the trajectory and the comparison is obscured by thermal noise. Paramagnetic channels and the pure-PM case stay at the $1/\sqrt{N}$ floor in finite-N realisations, consistent with DMFT setting them exactly to zero. The agreement confirms that the large-N DMFT reduction reproduces the macroscopic observables of finite-N SGD training. We also tested discrete SGD with one randomly sampled training example per update and no added weight noise, finding good agreement with the $\eta \to \infty$ DMFT at small learning steps (not shown).

B.1 Upper bound on the outlier count

The decomposition (9) implies that $W(t)$ carries at most K upper outliers at every fixed time. The negative-phase integral

$$\int_0^t e^{-\gamma(t-u)/2} \frac{\mathbf{x}(u)\mathbf{x}(u)^\top}{N} \mathrm{d}u \;\succeq\; 0 \tag{88}$$

is positive semidefinite, so

$$W(t) \;\preceq\; W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma}\, C. \tag{89}$$

Monotonicity of ordered eigenvalues under the Loewner order then gives $\lambda_j(W(t)) \le \lambda_j(W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma} C)$ for every $j \ge 1$, where $\lambda_j(\,\cdot\,)$ labels eigenvalues in decreasing order. Since C has rank K and is positive semidefinite, the min-max principle yields

$$\lambda_{K+j}\left(W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma} C\right) \le \lambda_j\left(W_{\mathrm{GOE}}(t)\right), \qquad (j \ge 1), \tag{90}$$

and in particular $\lambda_{K+1}(W(t)) \le \lambda_1(W_{\mathrm{GOE}}(t))$. With the stationary GOE initialization $W(0) \sim e^{-N\gamma\eta\,\mathrm{tr}\,W^2/4}$, the process $W_{\mathrm{GOE}}(t)$ remains GOE at every time, with semicircle edge $2/\sqrt{\gamma\eta}$; hence for every $\varepsilon > 0$, with high probability at large N,

$$\lambda_{K+1}(W(t)) \le \frac{2}{\sqrt{\gamma\eta}} + \varepsilon. \tag{91}$$

Therefore, no more than K eigenvalues of $W(t)$ are detached above the bulk edge at any given time.

B.2 MAP limit of the dynamical threshold

In the MAP regime $\eta \to \infty$, the learning-noise feedback drops out of both kernels in (75): $M(t,u) \to -\frac{\nu}{2} e^{-\gamma(t-u)/2} Q(t,u)$ and $D(t,u) \to 2\nu\,\delta(t-u)$. On the stationary uncondensed bath, the one-sided Laplace transform $\hat{f}(s) \equiv \int_0^\infty e^{-s\tau} f(\tau)\,\mathrm{d}\tau$ yields $\chi_P = 1/[\kappa_P + (\nu/2)\hat{Q}_P(\gamma/2)]$, while the spherical constraint gives $\kappa_P = \nu - (\nu/2)\hat{Q}_P^2(\gamma/2)$. We work at $K = 1$ with the unit-eigenvalue normalization $c_1 = 1$ inherited from $\|\mathbf{c}\|^2 = N$, so the leading-mode condensation condition $\nu c_1 \chi_P = \gamma$ reduces to $\nu\chi_P = \gamma$ and rearranges into

$$\tfrac{\gamma}{2}\left[\hat{Q}_P(\gamma/2) - \hat{Q}_P^2(\gamma/2)\right] \;=\; 1 - \gamma. \tag{92}$$

Because the LHS is the integral of $e^{-\gamma\tau/2} Q_P(\tau)[1 - Q_P(\tau)] \ge 0$, it is non-negative; hence $\gamma > 1$ admits no solution and the system is uncondensed for *every* ν. For $\gamma < 1$ the LHS depends on ν through the bath auto-correlation $Q_P(\tau;\nu,\gamma)$, and a genuine ν-dependent boundary survives. The resulting MAP phase diagram is shown in Fig. 2D of the main text: the boundary $\nu_c(\gamma)$ occupies only a narrow sliver below $\gamma = 1$, diverging vertically as $\gamma \to 1^-$ and disappearing below $\gamma_{\min} \approx 0.84$ where the uncondensed bath turns condensed for all ν.

Finite-η dynamical phase diagram. The MAP-limit structure survives at finite η: a grid-convergence study at fixed $\gamma = 2$ finds the apparent crossing $\nu\chi_P = \gamma$ scaling as $\nu_c \propto 1/\Delta\tau$ for every η tested, confirming that no physical boundary exists for $\gamma > 1$. Figure 10 shows the grid-converged sliver at $\eta \in \{1, 3, 10\}$: $\gamma_{\min}(\eta)$ approaches the MAP value ≈ 0.84 and the boundary steepens toward $\gamma = 1$ as η grows. The stationary $s_{\mathrm{st}}(\gamma)$ sweeps (row 2) and causal MSR trajectories (rows 3–4) confirm this structure over two decades of ν; the ceiling $\nu_c \lesssim 10^2$ visible in row 1 reflects the convergence guard $\nu_c \Delta\tau < 0.6$.

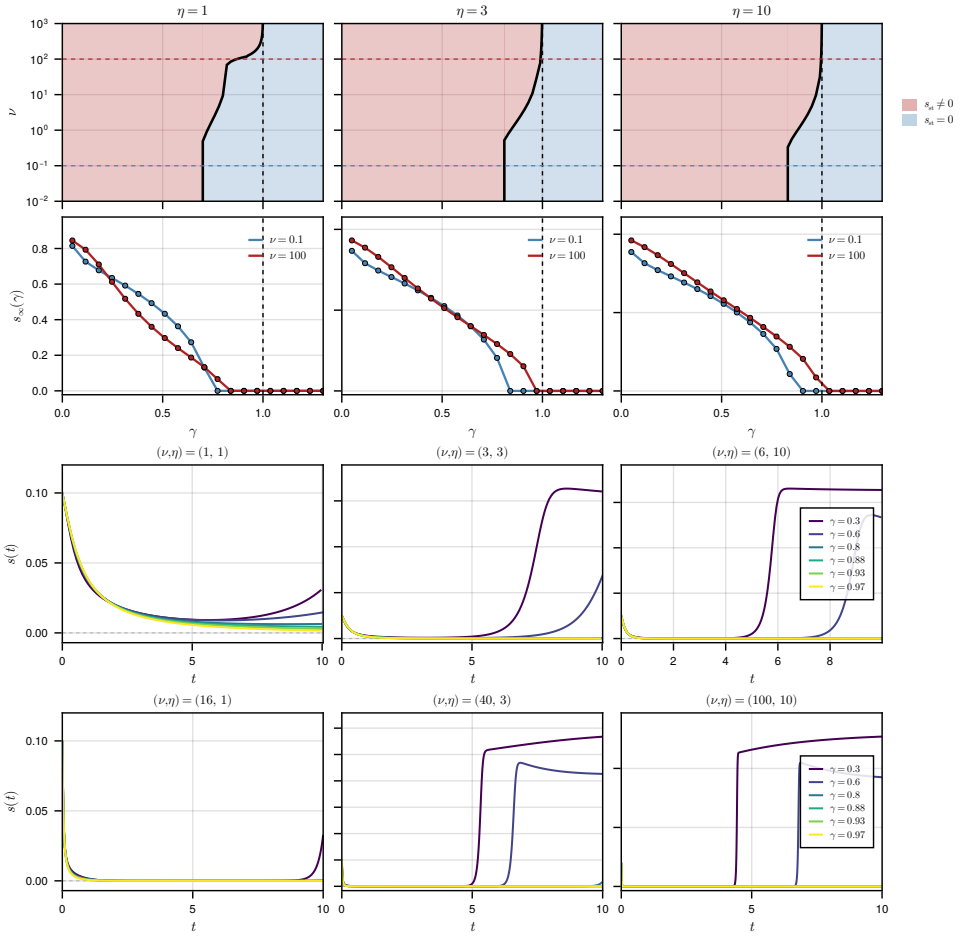


Figure 10: **Dynamical phase structure of the $K{=}1$ DMFT at finite η.** Three values $\eta \in \{1, 3, 10\}$ are shown in columns. **Row 1**: dynamical phase diagram $\nu_c(\gamma)$ (solid black), obtained by warm-started path continuation on a fine τ-grid ($\Delta\tau \simeq 4 \times 10^{-3}$) with the guard $\nu_c \Delta\tau < 0.6$; red shading is the condensed branch ($s_{\rm st} \neq 0$) and blue shading the uncondensed branch ($s_{\rm st} \to 0$). The dashed vertical line marks $\gamma = 1$, the vertical asymptote of the boundary at every finite η. Horizontal dashed lines at $\nu = 0.1$ (blue) and $\nu = 100$ (red) mark the two fixed sampling rates swept in row 2. **Row 2**: stationary $s_{\rm st}(\gamma)$ from the time-translation invariant DMFT branch-follower at $\nu = 0.1$ (blue) and $\nu = 100$ (red), evaluated on a 20-point γ-grid (circles) and connected by straight segments. The blue curve vanishes at a strictly smaller γ than the red curve, quantifying the widening of the condensed window as ν increases; both collapse to zero at $\gamma = 1$. **Rows 3–4**: families of causal MSR trajectories $s(t)$ starting from $s(0) = 0.1$, integrated to $t = 10$, at six log-spaced integer sampling rates $\nu \in \{1, 3, 6, 16, 40, 100\}$, paired per η column as $(\nu_{\rm lo}, \nu_{\rm hi}) = (1, 16), (3, 40), (6, 100)$ (titled in each panel) so each pair brackets the numerical ν_c range of its column, with a sweep of six γ values bracketing the transition (low-γ curves in dark purple, high-γ in yellow, viridis gradient). Trajectories with γ below the boundary condense to a finite plateau; those with γ above decay to zero. The t axis is identical across columns within each row.

C Nearly finite-dimensional data

This Appendix shows that the static and dynamical solutions obtained here are valid descriptions of SBM training on $K \to \infty$ data living in a K'-dimensional slab embedded in N-dimensional space, $N \to \infty$. The condition is that C has $K' = \mathcal{O}(1)$ eigenvalues which are $\mathcal{O}(K)$ while the remaining $\min\{N, K\} - K'$ stay $o(K)$. Depending on the scaling of t and γ, η, ν with K, different regimes arise, some of which we characterize below.

C.1 Statics

C.1.1 Bare large-K regime

We consider the regime $\gamma, \eta = \mathcal{O}(1)$. This regime was first characterized through a replica analysis in [Tulinski et al., 2026]. We summarize their main findings in the following paragraph.

A typical $W \sim P_\eta(\cdot|\mathcal{D})$ after training is found to exhibit the following properties. Its eigenvectors accurately recover the manifold directions, $\mathbf{v}_{1,\dots,K'} = \mathbf{c}_{1,\dots,K'}$, provided extensive modes are non-degenerate, and $h^2_{1,\dots,K'} = \widetilde{c}_{1,\dots,K'}$ where $\widetilde{c}_{1,\dots,K'} = c_{1,\dots,K'}/K$. Generated data $\mathbf{x} \sim P_W$ have vanishing projections along the other directions as long as

$$\gamma > \gamma_{\text{bulk}} \quad \text{where} \quad \gamma_{\text{bulk}} = c_{K'+1}\left(\delta_K + \frac{K'}{K} c_{K'+1}\right), \quad \text{with} \quad \delta_K = 1 - \frac{1}{K}\sum_{\ell=1}^{K'} c_\ell \,. \tag{93}$$

Here δ_K is the fraction of variance unexplained by the K' projections of the generated data along $\mathbf{v}_{1,\dots,K'}$. When $\gamma > \gamma_{\text{bulk}}$, the $\eta_c(\gamma, c_{1,\dots,K})$ delimiting the boundary of the edge phases with the others, scales like K'/K for thin slabs, $\delta_K \ll K^{-1/2}$, and like δ_K^2/γ for thick slabs, $\delta_K \gg K^{-1/2}$.

C.2 Dynamics

C.2.1 Bare large-K regime

We consider the regime

$$t, \gamma, \eta, \nu = \mathcal{O}(1). \tag{94}$$

Large-K limit of the DMFT We take $K \to \infty$ in the DMFT derived sending $N \to \infty$ first while holding K fixed. For large K, the noise correlator stays $\mathcal{O}(1)$. The memory kernel behaves as

$$\widetilde{M}(t,u) \equiv \frac{M(t,u)}{K} = -\frac{\nu}{2} e^{-\frac{\gamma}{2}(t-u)} Q(t,u) + \mathcal{O}(K^{-1}), \qquad t > u, \tag{95}$$

dominated by the negative phase.

Using that $\kappa(t)$ satisfies $\partial_t Q(t,t) = 0$, we get that $\kappa(t)$ also scales linearly with K, as

$$\widetilde{\kappa}(t) \equiv \frac{\kappa(t)}{K\nu} = \frac{1 - e^{-\gamma t/2}}{\gamma} \sum_{\ell=1}^{K'} \widetilde{c}_\ell s_\ell(t)^2 - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u)^2 + \mathcal{O}(K^{-1}). \tag{96}$$

On $\mathcal{O}(1)$ timescales, the signals $s_\ell(t), s_k(t)$ ($\ell \le K' < k$) are quasistatic, with

$$\begin{cases} 0 = -\widetilde{\kappa}(t) s_\ell(t) + (1 - e^{-\gamma t/2})\frac{\widetilde{c}_\ell}{\gamma} s_\ell(t) - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) s_\ell(u) + o(1), \\ 0 = -\widetilde{\kappa}(t) s_k(t) - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) s_k(u) + o(1). \end{cases} \tag{97}$$

The correlation function also behaves quasistatically on finite timescales,

$$0 = -\widetilde{\kappa}(t) Q(t,t') + (1 - e^{-\gamma t/2}) \sum_\ell \frac{\widetilde{c}_\ell}{\gamma} s_\ell(t) s_\ell(t') - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) Q(u,t'), \tag{98}$$

while the response function obeys

$$\partial_t R(t,t') = -K\nu \widetilde{\kappa}(t) R(t,t') - \frac{K\nu}{2}\int_{t'}^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) R(u,t') + O(1), \quad t > t', \tag{99}$$

together with $R(t,t) = 1$. Since the restoring force is linear in $K\nu$, we expect small perturbations at a time $t' < t$ to be suppressed on fast time scales $t - t' \sim 1/K\nu$. Inspecting $R(t,t')$ close to the diagonal, $t = t' + \theta/K\nu$, we see that the integrand of the memory term is $\mathcal{O}(1)$, so the integral is $\sim 1/K\nu$, and $R(t,t') \sim e^{-K\nu\widetilde{\kappa}(t')(t-t')} \downarrow 0$ ($t > t'$).

Large K non-equilibrium stationary state Now take $t, t' \to \infty$, with $\tau = t - t' = \mathcal{O}(1)$, and write $Q(t,t'), R(t,t') \to Q_{\rm st}(\tau), R_{\rm st}(\tau)$, $s_\ell(t), s_k(t), \kappa(t) \to s_{\ell,\rm st}, s_{k,\rm st}, \kappa_{\rm st}$.

The $\mathcal{O}(1)$ bulk modes stay uncondensed, $s_k = 0$ ($k > D$), meanwhile extensive modes satisfy

$$0 = \left[\widetilde{\kappa}_{\rm st} + \frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2} Q_{\rm st}(\tau) - \frac{\widetilde{c}_\ell}{\gamma}\right] s_{\ell,\rm st}, \qquad \ell = 1, \ldots, K'. \tag{100}$$

Hence, only modes with the same maximum rescaled eigenvalue $\widetilde{c}_{\max}$ can condense.

The stationary response function satisfies

$$\partial_\tau R_{\rm st}(\tau) = -K\nu\left[\widetilde{\kappa}_{\rm st} R_{\rm st}(\tau) + \frac{1}{2}\int_0^\tau \mathrm{d}\sigma\, e^{-\gamma\sigma/2} Q_{\rm st}(\sigma) R_{\rm st}(\tau - \sigma)\right] \tag{101}$$

together with $R_{\rm st}(0^+) = 1$. Taking a one-sided Laplace transform $\check{R}(p) = \int_0^\infty \mathrm{d}\tau\, e^{-\tau p} R(\tau)$ yields

$$\check{R}_{\rm st}(p) = \frac{1}{p + K\nu(\widetilde{\kappa}_{\rm st} + \frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-(p+\gamma/2)\tau} Q_{\rm st}(\tau))}. \tag{102}$$

In particular, $\int_0^\infty \mathrm{d}\tau\, R_{\rm st}(\tau) = \check{R}(0) = \frac{\gamma}{K\nu\widetilde{c}_{\max}} \downarrow 0$.

The equation for the stationary correlation function is

$$0 = -\widetilde{\kappa}_{\rm st} Q_{\rm st}(\tau) + \frac{1}{\gamma}\sum_\ell \widetilde{c}_\ell s_{\ell,\rm st}^2 - \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2} Q_{\rm st}(\sigma) Q_{\rm st}(\tau - \sigma). \tag{103}$$

Writing $\mathbf{x}(t) = \langle \mathbf{x}(t)\rangle + \delta\mathbf{x}(t)$ with $\langle \delta\mathbf{x}(t)\rangle = 0$, one has $Q_{\rm st}(\tau) = q + C(\tau)$, $q = \sum_{\ell \le K'} \delta_{c_\ell, c_{\max}} s_\ell^2$, $C(\tau) = \langle \delta\mathbf{x}(t)\cdot\delta\mathbf{x}(t-\tau)\rangle/N$. Combining the multiplier and signal equations,

$$\frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2} C(\tau)^2 = \frac{1}{\gamma}(1-q)(q - \widetilde{c}) + \frac{1-2q}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2} C(\tau). \tag{104}$$

From Eq. (80), fluctuations obey the linear equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta\mathbf{x} = -K\nu\left[\widetilde{\kappa}_{\rm st}\delta\mathbf{x}(t) + \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2} Q_{\rm st}(\sigma)\delta\mathbf{x}(t-\sigma)\right] + \boldsymbol{\zeta}(t). \tag{105}$$

When $\delta\mathbf{x}(t)$ varies slowly, $\delta\mathbf{x}(t-\sigma) \approx \delta\mathbf{x}(t)$ over the memory window $\sigma \sim 2\gamma^{-1}$ and then $\frac{\mathrm{d}}{\mathrm{d}t}\delta\mathbf{x} \approx -K\nu[\widetilde{\kappa}_{\rm st} + \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2} Q_{\rm st}(\sigma)]\delta\mathbf{x}(t)$. Stationarity implies that the bracket equals $\widetilde{c}_{\max}/\gamma > 0$, hence slow fluctuations are suppressed, and the overlap of $C(\tau)$ with kernels of a fixed $\mathcal{O}(1)$ width vanishes. In particular, $\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2} C(\tau) = 0$. Using $\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2} C(\tau) = 0$, Eq. (104) gives $q \ge \widetilde{c}_{\max}$ since $q \le 1$ and LHS > 0, while Eq. (100) gives $\gamma\widetilde{\kappa}_{\rm st} = \widetilde{c}_{\max} - q$, whence $q \le \widetilde{c}_{\max}$. As a result, $q = \widetilde{c}_{\max}$ and $\Delta = \widetilde{\kappa}_{\rm st} = 0$. We conclude that $Q_{\rm st}(\tau) = q$ with $\tau > 0$ for $K \to \infty$. Contrast this with the finite-K case, where we expect $Q_{\rm st}(\tau)$ to plateau only when $\tau \to \infty$.

The non-equilibrium stationary state is thus characterized by

$$q = \sum_\ell \delta_{c_\ell, c_{\max}} s_\ell^2 = \min\{1, \widetilde{c}_{\max}\}, \qquad \widetilde{\kappa}_{\rm st} = \frac{(\widetilde{c}_{\max} - 1)_+}{\gamma}, \tag{106}$$

independently of η and ν. To leading order in K, then, the negative phase cancels the contribution to W of the top extensive mode in the positive phase, see Eq. (9). As a result, at long times, the top mode of W is not the sampled direction, but the largest extensive unsampled direction.

Action in the bare large-K regime The solution above was derived by sending $K \to \infty$ after having first taken $N \to \infty$ holding K fixed. Using $\kappa = K\nu\widetilde{\kappa}$ and $c_\ell = K\widetilde{c}_\ell$ for the extensive modes, the general-K form of Eq. (78) decomposes as $N\mathcal{A}_{\mathbf{x}} = NK\nu\mathcal{A}_{\rm eff} + \mathcal{A}_{\rm rem}$, where

$$\begin{aligned}\mathcal{A}_{\rm eff} = &\int \mathrm{d}t\, \widetilde{\kappa}(t)\frac{\hat{\mathbf{x}}(t)\cdot\mathbf{x}(t)}{N} \\ &- \int \mathrm{d}t\, (1 - e^{-\gamma t/2}) \sum_{\ell \le K'} \frac{\widetilde{c}_\ell}{\gamma}\frac{\mathbf{x}(t)\cdot\mathbf{c}_\ell}{N}\frac{\hat{\mathbf{x}}(t)\cdot\mathbf{c}_\ell}{N} \\ &+ \frac{1}{2}\int_{t>u} \mathrm{d}t\, \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u)\frac{\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u)}{N}.\end{aligned} \tag{107}$$

The remainder contains the kinetic term, the spin noise, the weight noise, and the subextensive data modes:

$$\begin{aligned}\mathcal{A}_{\text{rem}} = &\int \mathrm{d}t\ \left[\hat{\mathbf{x}}(t)\cdot\dot{\mathbf{x}}(t) - \nu\|\hat{\mathbf{x}}(t)\|^2\right] \\ &- \frac{\nu^2}{\eta\gamma N}\int_{t>u} \mathrm{d}t\,\mathrm{d}u\, e^{-\gamma(t-u)/2}(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u)) \\ &- \frac{\nu^2}{\eta\gamma N}\int_{t>u} \mathrm{d}t\,\mathrm{d}u\, e^{-\gamma(t-u)/2}(\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u))(\mathbf{x}(t)\cdot\hat{\mathbf{x}}(u)) \\ &- \frac{\nu}{\gamma N}\int \mathrm{d}t\,(1-e^{-\gamma t/2})\sum_{k=K'+1}^{\operatorname{rank} C} c_k(\mathbf{x}(t)\cdot\mathbf{c}_k)(\hat{\mathbf{x}}(t)\cdot\mathbf{c}_k).\end{aligned}$$

The formal leading action $\mathcal{A}_{\text{eff}}$ gives the quasistatic signal and correlation equations above.

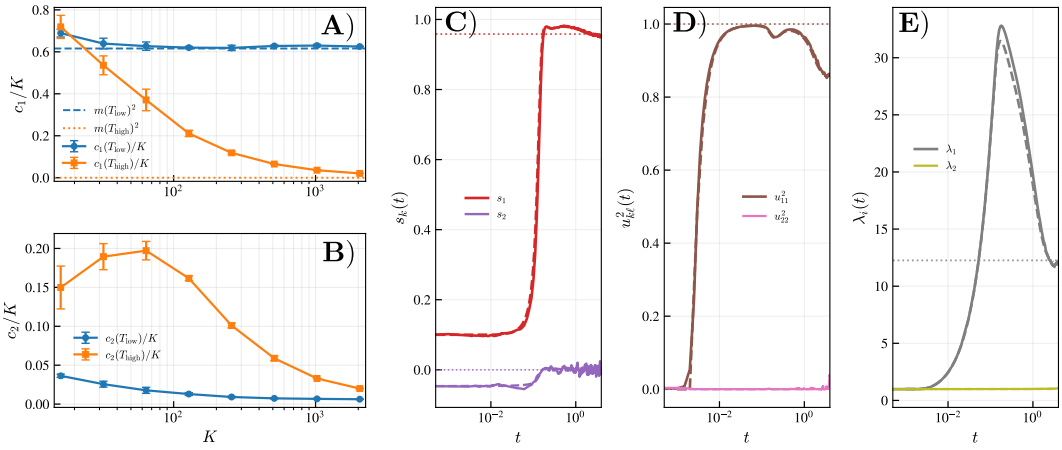


Figure 11: **Validation of the large-K DMFT equations for nearly finite-dimensional Ising data. A, B)** Spectrum of the empirical covariance matrix C for K configurations of the two-dimensional periodic Ising model at fixed linear size $L = 32$, $N = L \times L (= 1024)$. Below the critical temperature, $T_{\text{low}} = 2.2 < T_c \simeq 2.269$, the leading eigenvalue is extensive, $c_1 \simeq Km(T_{\text{low}})^2$, while c_2/K decreases with K, consistent with effectively $K' = 1$ data. Above the transition, $T_{\text{high}} = 3.2 > T_c$, no extensive mode is present and both c_1/K and c_2/K vanish with K. Points and error bars show the mean and standard deviation over three independently sampled datasets. **C–E)** Comparison between finite-N coupled Langevin simulations and the DMFT prediction for a representative below-critical dataset with $K = 512$. Solid curves show finite-N simulations, dashed curves show the DMFT solution using the empirical spectrum of C, and dotted horizontal lines show the large-K stationary predictions. **C)** Signal overlaps $s_k(t) = c_k \cdot x(t)/N$. The dynamics condenses onto the leading Ising mode, while the second mode remains negligible. **D)** Squared eigenvector overlaps $u_k^2(t)$, showing alignment of the top learned eigenvector with the leading data direction. **E)** Top eigenvalues of $W(t)$, showing the detachment of λ_1 from the Wigner bulk edge while λ_2 remains near the edge.

C.2.2 Invariant large-K regime

We now show that the DMFT (11)-(75) is left invariant under the rescaling

$$\overline{t} = \frac{K}{K'}t, \quad \gamma = \frac{K}{K'}\overline{\gamma}, \quad \eta = \frac{K'}{K}\overline{\eta}, \quad \nu = \frac{K}{K'}\overline{\nu}, \qquad \text{with} \qquad \overline{t},\overline{\gamma},\overline{\eta},\overline{\nu} = \mathcal{O}(1), \tag{108}$$

and describes the large-K Langevin training dynamics Eq. (3)-(4) on nearly finite-K' training data.

We introduce

$$\overline{c}_{1,\ldots,K'} = K'\widetilde{c}_{1,\ldots,K'}\left(= \frac{K'}{K}c_{1,\ldots,K'}\right), \qquad \overline{C} = \frac{K'}{K}C, \tag{109}$$

together with the scaled-time dynamical variables

$$\overline{W}(\bar{t}) = W(t), \qquad \overline{\mathbf{x}}(\bar{t}) = \mathbf{x}(t), \qquad \overline{\kappa}(\bar{t}) = \frac{K'}{K}\kappa(t). \tag{110}$$

The scaled covariance $\overline{C}$ has eigenvalues $\bar{c}_{1,\dots,K} = \frac{K'}{K} c_{1,\dots,K}$, for which $\bar{c}_{1,\dots,K'} = \mathcal{O}(1)$ while $\bar{c}_{K'+1,\dots,K} = o(1)$. Under these scalings,

$$\frac{\mathrm{d}\overline{W}}{\mathrm{d}\bar{t}} = \frac{1}{2}\left(\overline{C} - \frac{K'}{N}\overline{\mathbf{x}}(\bar{t})\overline{\mathbf{x}}(\bar{t})^\top\right) - \frac{\overline{\gamma}}{2}\overline{W}(\bar{t}) + \frac{\overline{\Omega}(\bar{t})}{\sqrt{N\overline{\eta}}}, \tag{111}$$

$$\frac{\mathrm{d}\overline{\mathbf{x}}}{\mathrm{d}\bar{t}} = \overline{\nu}\overline{W}(\bar{t})\overline{\mathbf{x}}(\bar{t}) - \overline{\kappa}(\bar{t})\overline{\mathbf{x}}(\bar{t}) + \sqrt{2\overline{\nu}}\,\overline{\boldsymbol{\xi}}(\bar{t}), \tag{112}$$

where $\overline{\Omega}(\bar{t}), \overline{\boldsymbol{\xi}}(\bar{t})$, denote white noises in the scaled time $\bar{t}$. Following the same steps as in App. B,

$$\frac{\mathrm{d}\overline{s}_k}{\mathrm{d}\bar{t}} = -\overline{\kappa}(\bar{t})\,\overline{s}_k(\bar{t}) + \frac{\overline{\nu}}{\overline{\gamma}}\left(1 - e^{-\overline{\gamma}\,\bar{t}/2}\right)\bar{c}_k\,\overline{s}_k(\bar{t}) + \int_0^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\overline{s}_k(\overline{u})\,\mathrm{d}\overline{u}, \tag{113}$$

$$\partial_{\bar{t}}\overline{R}(\bar{t},\bar{t}') = -\overline{\kappa}(\bar{t})\,\overline{R}(\bar{t},\bar{t}') + \int_{\bar{t}'}^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\overline{R}(\overline{u},\bar{t}')\,\mathrm{d}\overline{u}, \tag{114}$$

$$\partial_{\bar{t}}\overline{Q}(\bar{t},\bar{t}') = \frac{\overline{\nu}}{\overline{\gamma}}\left(1 - e^{-\overline{\gamma}\,\bar{t}/2}\right)\sum_{k=1}^{K}\bar{c}_k\,\overline{s}_k(\bar{t})\,\overline{s}_k(\bar{t}') - \overline{\kappa}(\bar{t})\,\overline{Q}(\bar{t},\bar{t}') \tag{115}$$

$$+ \int_0^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\overline{Q}(\overline{u},\bar{t}')\,\mathrm{d}\overline{u} + \int_0^{\bar{t}'} \overline{D}(\bar{t},\overline{u})\,\overline{R}(\bar{t}',\overline{u})\,\mathrm{d}\overline{u}. \tag{116}$$

where

$$\overline{M}(\bar{t},\overline{u}) = \frac{K'^2}{K^2}M(t,u) = e^{-\overline{\gamma}(\bar{t}-\overline{u})/2}\left\{\frac{\overline{\nu}^2}{\overline{\eta}\,\overline{\gamma}}\,\overline{R}(\bar{t},\overline{u}) - \frac{K'\overline{\nu}}{2}\,\overline{Q}(\bar{t},\overline{u})\right\}, \tag{117}$$

$$\overline{D}(\bar{t},\overline{u}) = \frac{K'^2}{K^2}D(t,u) = 2\overline{\nu}\,\delta(\bar{t}-\overline{u}) + \frac{\overline{\nu}^2}{\overline{\eta}\,\overline{\gamma}}\,e^{-\overline{\gamma}|\bar{t}-\overline{u}|/2}\overline{Q}(\bar{t},\overline{u}), \tag{118}$$

and we introduced the scaled-time dynamical order parameters

$$\overline{s}_k(\bar{t}) = s_k(t), \qquad \overline{Q}(\bar{t},\bar{t}') = Q(t,t'), \qquad \overline{R}(\bar{t},\bar{t}') = R(t,t'). \tag{119}$$

To complete the reduction, we show that

$$\sum_{k=1}^{K}\bar{c}_k\,\overline{s}_k(\bar{t})\,\overline{s}_k(\bar{t}') \approx \sum_{k=1}^{K'}\bar{c}_k\,\overline{s}_k(\bar{t})\,\overline{s}_k(\bar{t}'). \tag{120}$$

Let

$$\Delta(\bar{t},\bar{t}') = \frac{1}{N}\overline{\mathbf{x}}(\bar{t})^\top\overline{B}\overline{\mathbf{x}}(\bar{t}'), \qquad \text{with} \qquad \overline{B} = \frac{1}{N}\sum_{a=K'+1}^{K}\bar{c}_a\,\mathbf{c}_a\mathbf{c}_a^\top, \tag{121}$$

be the remainder of the above approximation. By the spherical constraint:

$$|\Delta(\bar{t},\bar{t}')| \le \|\overline{B}\|_{\mathrm{op}}\sqrt{\frac{\|\overline{\mathbf{x}}(\bar{t})\|^2}{N}\frac{\|\overline{\mathbf{x}}(\bar{t}')\|^2}{N}} = \|\overline{B}\|_{\mathrm{op}} = \bar{c}_{K'+1} = o(1). \tag{122}$$

Thus $\Delta(\bar{t},\bar{t}') \to 0$ as K grows, uniformly in $\bar{t},\bar{t}'$ by the spherical constraint, and the claim follows. This reduction requires no restriction on the relative scaling of N and K beyond the stated spectral assumptions. Fig. 12 validates this scaled DMFT against finite-N Langevin simulations on nearly finite-dimensional training data, sampled from the two-dimensional Ising model below the critical temperature ($K' = 1$).

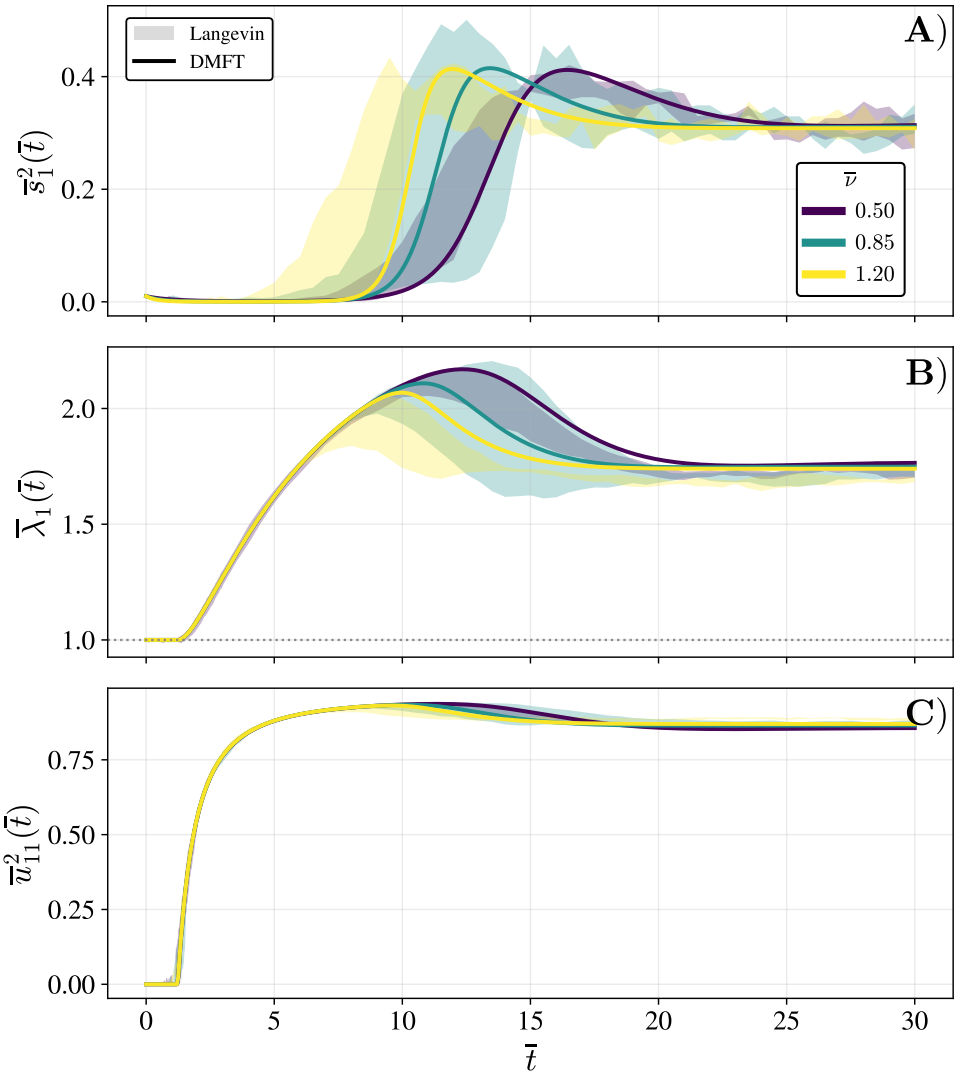


Figure 12: **Large-K training dynamics for nearly finite-dimensional Ising data.** Comparison between the effective $K' = 1$ large-K DMFT prediction and finite-N coupled Langevin simulations for below-critical two-dimensional Ising data at $T = 1.8$, with $K = 512$ and $N = 2025$. Time is shown in the rescaled variable $\overline{t} = Kt/K'$, and colors indicate the rescaled persistent-chain update rate $\overline{\nu}$. Solid curves show the effective $K' = 1$ DMFT solution, while shaded regions show finite-N Langevin simulations over independent noise seeds, plotted as mean $\pm$ one standard deviation; the simulation mean is not separately drawn. **A)** Squared signal overlap $\overline{s}_1^2(\overline{t})$, showing condensation of the persistent chain onto the leading Ising covariance mode. **B)** Top eigenvalue $\overline{\lambda}_1(\overline{t})$ of the learned weight matrix, with the dotted line marking the bulk edge. **C)** Squared eigenvector overlap $\overline{u}_{11}^2(\overline{t})$, showing alignment of the top eigenvector with the leading data mode. Increasing $\overline{\nu}$ accelerates the onset of the negative phase, reducing the eigenvalue overshoot while leaving the learned direction largely unchanged.

C.2.3 Recovering the bare from the invariant large-K regime

We now check that the bare large-K regime of Sec. C.2.1 is recovered as a boundary limit of the invariant scaling of Sec. C.2.2.

Let

$$\varepsilon = \frac{K'}{K} \to 0.$$

If the bare variables t, γ, ν, η are kept $O(1)$, then in the invariant variables one has

$$\overline{\gamma} = \varepsilon\gamma \to 0, \qquad \overline{\nu} = \varepsilon\nu \to 0, \qquad \overline{\eta} = \frac{\eta}{\varepsilon} \to \infty, \qquad \overline{t} = \frac{t}{\varepsilon} \to \infty.$$

This is the boundary of the invariant scaling corresponding to the bare regime.

Since the scaled covariance is $\overline{C} = \varepsilon C$, its leading eigenvalues translate across the two regimes as

$$\overline{c}_\ell = \varepsilon c_\ell = K'\widetilde{c}_\ell, \qquad \ell = 1, \dots, K'.$$

We use

$$\overline{s}_\ell(\overline{t}) = s_\ell(t), \qquad \overline{Q}(\overline{t}, \overline{t}') = Q(t, t'), \qquad \overline{R}(\overline{t}, \overline{t}') = R(t, t'), \qquad \overline{\kappa}(\overline{t}) = \varepsilon\kappa(t).$$

Injecting

$$\overline{\gamma}(\overline{t} - \overline{u}) = \gamma(t - u), \qquad \frac{\overline{\nu}^2}{\overline{\eta}\,\overline{\gamma}} = \frac{\varepsilon^2\nu^2}{\eta\gamma} = O(\varepsilon^2), \qquad K'\overline{\nu} = K'\varepsilon\nu,$$

in the scaled memory kernel (117) gives

$$\overline{M}(\overline{t}, \overline{u}) = -\frac{K'\varepsilon\nu}{2} e^{-\gamma(t-u)/2} Q(t, u) + O(\varepsilon^2).$$

Since $\mathrm{d}\overline{u} = \mathrm{d}u/\varepsilon$, the memory integral has a finite limit

$$\int_0^{\overline{t}} \mathrm{d}\overline{u}\, \overline{M}(\overline{t}, \overline{u}) \overline{s}_\ell(\overline{u}) = -\frac{K'\nu}{2} \int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t, u) s_\ell(u) + o(1).$$

Using $\partial_{\overline{t}} = \varepsilon\partial_t$ and $\varepsilon \to 0$ in Eq. (113) gives the bare large-K quasistatic signal equation, Eq. (97).

The same reduction applies to the correlation equation. The smooth part of $\overline{D}$ is $O(\varepsilon^2)$, and the white-noise term does not contribute to the off-diagonal equation $t > t'$. From there, using $\partial_{\overline{t}} = \varepsilon\partial_t$ and $\varepsilon \to 0$ once again gives back the bare large-K quasistatic correlation equation, Eq. (98).

Applying similar manipulations turns Eq. (114) into Eq. (99), thereby concluding the reduction.

D Teacher-student scenario

This appendix collects the four teacher–student KL divergences used in the main text for the rank-one teacher of Sec. 3.

Table 1: KL divergences. Columns (students): typical $W \sim P_\eta$ vs. posterior predictive $P_{\rm pp}$. Rows (metrics): forward $(P^*\|\cdot)$ vs. reverse $(\cdot\|P^*)$ KL.

	Typical $W \sim P_\eta$	Post. pred. $P_{\rm pp}$
Forward KL	$\langle D_{\rm KL}(P^*\|P_W)\rangle$	$D_{\rm KL}(P^*\|P_{\rm pp})$
Reverse KL	$\langle D_{\rm KL}(P_W\|P^*)\rangle$	$D_{\rm KL}(P_{\rm pp}\|P^*)$

D.1 Teacher-generated training data

The teacher is the rank-one SBM $W^* = \omega^* \mathbf{w}^* \mathbf{w}^{*\top} - (\omega^*/N)I$ of Sec. 3, against which all four KL divergences are evaluated. For $\omega^* > 1$, the teacher generates data that condenses along $\mathbf{w}^*$. For $K = 2$, we have almost surely that C has the two eigenvalues:

$$c_1 = 2 - \frac{1}{\omega^*}, \qquad c_2 = \frac{1}{\omega^*}. \tag{123}$$

The corresponding eigenvectors $\mathbf{c}_{1,2}$ satisfy:

$$\frac{(\mathbf{c}_1 \cdot \mathbf{w}^*)^2}{N} = 1 - \frac{1}{2\omega^* - 1}, \quad \frac{(\mathbf{c}_2 \cdot \mathbf{w}^*)^2}{N} = 0 \tag{124}$$

to leading order in N. These deterministic limits also imply self-averaging of the four KL divergences per spin (that we compute below) over the training set. The teacher partition function and entropy read:

$$\frac{1}{N} \ln Z(W^*) = \frac{1}{2} \ln(2\pi e) + \frac{\omega^* - 1 - \ln \omega^*}{2}, \quad \frac{1}{N} \mathcal{H}[P^*] = \frac{1}{2} \ln(2\pi e) - \frac{1}{2} \ln \omega^* \tag{125}$$

where we used $\ln Z(0) = \frac{N}{2} \ln(2\pi e)$.

D.2 Posterior predictive as a ratio of evidence integrals

We compute:

$$\begin{aligned} P_{\rm pp}(\mathbf{x}|C) &= \langle P(\mathbf{x}|W) \rangle_{W \sim P_\eta} = \int P(\mathbf{x}|W) P_\eta(W|C) \mathrm{d}W \\ &= \frac{1}{\mathcal{Y}(C)} \int \exp\left\{ \frac{N}{2} \operatorname{tr}\left[\left(\eta C + \frac{\mathbf{x}\mathbf{x}^\top}{N} \right) W \right] - (K\eta + 1) \ln Z(W) - \frac{N}{4} \gamma \eta \operatorname{tr} W^2 \right\} \mathrm{d}W \end{aligned} \tag{126}$$

Defining $C_{\mathbf{x}} = C + \frac{1}{\eta} \frac{\mathbf{x}\mathbf{x}^\top}{N}$, so that $\eta \operatorname{tr} C_{\mathbf{x}} = K\eta + 1$, the integrand is $e^{\mathcal{L}(W; C_{\mathbf{x}})}$. Then

$$P_{\rm pp}(\mathbf{x}|C) = \frac{\mathcal{Y}(C_{\mathbf{x}})}{\mathcal{Y}(C)}, \qquad C_{\mathbf{x}} = C + \frac{1}{\eta} \frac{\mathbf{x}\mathbf{x}^\top}{N}. \tag{127}$$

This identity, noticed in [Tulinski et al., 2026], reduces the posterior predictive to evaluating the evidence at a low-rank perturbation of C, so we can reuse the evidence formulas derived above rather than computing a separate W integral.

D.3 MAP and flat limits of the posterior predictive

$P_{{\rm pp},\eta}(\mathbf{x}|C)$ interpolates between the flat measure on $\mathcal{S}_N$ for $\eta \downarrow 0$ and the MAP likelihood for $\eta \uparrow \infty$.

Flat posterior predictive In the $\eta \to 0^+$ limit, the posterior predictive is uniform on the sphere,

$$\lim_{\eta\to 0^+} \frac{1}{N} \log P_{\mathrm{pp},\eta}(\mathbf{x}|C) = -\frac{1}{2}\log(2\pi e) + o_N(1). \tag{128}$$

$P_{\mathrm{pp},\eta}(\mathbf{x}|C)$ depends on η through $P_\eta(W|C) \propto e^{-\frac{N\gamma\eta}{4}\operatorname{tr} W^2 + \frac{N\eta}{2}\operatorname{tr} CW - K\eta \ln Z(W)}$.

We argue that, under the rescaling $W \mapsto \eta^{-1/2}\widetilde{W}$, and assuming that $\eta^{1/2}\operatorname{tr} C \to 0^+$, the dominant contribution to the exponent is $-\frac{N\gamma}{4}\operatorname{tr}\widetilde{W}^2 + o(1)$. To start, since $|\operatorname{tr} C\widetilde{W}| \le \operatorname{tr} C\, \|\widetilde{W}\|_{\mathrm{op}}$, we have

$$\eta \operatorname{tr} CW = \eta^{1/2} \operatorname{tr} \widetilde{W} C = o(1) \tag{129}$$

for a typical $\widetilde{W}$ with $\|\widetilde{W}\|_{\mathrm{op}} = \mathcal{O}(1)$. To bound $\eta \ln Z(\eta^{-1/2}\widetilde{W})$, use

$$N\lambda_{\min}(W) \le \mathbf{x}^\top W \mathbf{x} \le N\lambda_{\max}(W), \tag{130}$$

which holds pointwise on $\mathcal{S}_N$. Exponentiating and integrating over $\mathcal{S}_N$ gives

$$0 < Z(0) e^{\frac{N}{2}\lambda_{\min}} \le Z(W) \le Z(0) e^{\frac{N}{2}\lambda_{\max}}. \tag{131}$$

Taking logarithms and using $|\lambda_{\min}|, |\lambda_{\max}| \le \|W\|_{\mathrm{op}}$, we obtain

$$|\ln Z(W) - \ln Z(0)| \le \frac{N}{2}\|W\|_{\mathrm{op}}. \tag{132}$$

Injecting $W = \eta^{-1/2}\widetilde{W}$, multiplying by η and using $\eta \ln Z(0) = \mathcal{O}(\eta)$,

$$\frac{\eta}{N} \log Z(\eta^{-1/2}\widetilde{W}) = \mathcal{O}(\eta^{1/2}). \tag{133}$$

Then $P_{0^+}(\widetilde{W}|C) \propto e^{-\frac{N\gamma}{4}\operatorname{tr}\widetilde{W}^2 + o(1)}$. The Jacobian from $W \mapsto \widetilde{W}$ is independent of C and $\mathbf{x}$, and absorbed into the normalization of the posterior over $\widetilde{W}$. Using $P_{0^+}(O^\top \widetilde{W} O|C) = P_{0^+}(\widetilde{W}|C)$ and $P_W(O\mathbf{x}) = P_{O^\top W O}(\mathbf{x})$ for any orthogonal matrix O, we have $P_{\mathrm{pp},0^+}(O\mathbf{x}) = P_{\mathrm{pp},0^+}(\mathbf{x})$. Any normalized rotationally invariant density on $\mathcal{S}_N$ is constant, whence the claimed

$$P_{\mathrm{pp},0^+}(\mathbf{x}|C) = \frac{1}{Z(0)}. \tag{134}$$

MAP posterior predictive In the $\eta \to +\infty$ limit, $P_\eta(W|C) \Longrightarrow \delta(W - W_{\mathrm{MAP}})$, where W_{MAP} is the unique maximizer of $P_\eta(W|C)$. Hence

$$\lim_{\eta\to+\infty} P_{\mathrm{pp},\eta}(\mathbf{x}|C) = P_{W_{\mathrm{MAP}}}(\mathbf{x}). \tag{135}$$

Stationarity of the log-posterior in W shows that W_{MAP} solves

$$\gamma W_{\mathrm{MAP}} = C - \frac{K}{N}\langle \mathbf{x}\mathbf{x}^\top \rangle_{P_{W_{\mathrm{MAP}}}}. \tag{136}$$

D.4 Posterior-predictive entropy

Applying (127) to the posterior-predictive entropy gives

$$\mathcal{H}[P_{\mathrm{pp}}] = \ln \mathcal{Y}(C) - \langle \ln \mathcal{Y}(C_{\mathbf{x}}) \rangle_{\mathbf{x}\sim P_{\mathrm{pp}}}. \tag{137}$$

Under $\mathbf{x} \sim P_{\mathrm{pp}}$, we have the overlaps (by (57))

$$m_k^2 = \frac{(\mathbf{c}_k \cdot \mathbf{x})^2}{N^2} = \frac{c_k}{2} - \frac{\gamma}{2}\left(\chi_k - \frac{1}{\eta c_k}\right), \quad k = 1, 2, \tag{138}$$

The three nonzero eigenvalues $\tilde{c}_{1,2,3}$ of $C_{\mathbf{x}}$ (127) are the roots of

$$m_1^2 \tilde{c}(\tilde{c} - c_2) + m_2^2 \tilde{c}(\tilde{c} - c_1) + (1 - m_1^2 - m_2^2 - \eta\tilde{c})(\tilde{c} - c_1)(\tilde{c} - c_2) = 0, \tag{139}$$

with $\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3 = 2 + 1/\eta$. Evaluating $\ln \mathcal{Y}$ in (137) at leading order (App. A.1.3),

$$\frac{1}{N}\mathcal{H}[P_{\mathrm{pp}}] = \Phi(c_1, c_2) - \Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3). \tag{140}$$

D.5 Reverse cross-entropies

Since the cross-entropy is linear in its first argument, the reverse cross-entropies of a typical P_W and the $P_{\rm pp}$ coincide:

$$\langle \mathcal{H}[P_W; P^*] \rangle_{W\sim P_\eta} = \mathcal{H}[\langle P_W \rangle_{W\sim P_\eta}; P^*] = \mathcal{H}[P_{\rm pp}; P^*], \tag{141}$$

so the posterior-average and posterior-predictive reverse cross-entropies coincide. Expanding,

$$\mathcal{H}[P_{\rm pp}; P^*] = \langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} + \ln Z(W^*). \tag{142}$$

For $K = 2$ with the rank-one teacher W^*, the sample statistics Eq. (57) combined with the leading-order identities $\frac{1}{N}\mathbf{c}_2^\top W^* \mathbf{c}_2 = 0$ and $\frac{1}{N}\mathbf{c}_1^\top W^* \mathbf{c}_1 = 2\omega^*(\omega^* - 1)/(2\omega^* - 1)$ yield

$$\frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} = \left(\frac{\gamma}{\eta} \frac{\omega^{*2}}{(2\omega^* - 1)^2} (\eta c_1 \chi_1 - 1) - 1 \right) \frac{\omega^* - 1}{2}, \tag{143}$$

where χ_1 is the saddle-point value from the spherical integral (29).

D.6 Forward cross-entropies

Posterior-average forward cross-entropy. The forward cross-entropy of a typical sample W is

$$\langle \mathcal{H}[P^*; P_W] \rangle_{W\sim P_\eta} = \langle \langle \mathcal{E}(\mathbf{x}; W) \rangle_{\mathbf{x}\sim P^*} + \ln Z(W) \rangle_{W\sim P_\eta}. \tag{144}$$

For $K = 2$ with the rank-one teacher, the weight statistics Eq. (56) combined with (124) give

$$\frac{1}{N}\langle \langle \mathcal{E}(\mathbf{x}; W) \rangle_{\mathbf{x}\sim P^*} \rangle_{W\sim P_\eta} = -\left(c_1\chi_1 - \frac{1}{\eta} \right) \frac{(\omega^* - 1)^2}{(2\omega^* - 1)^2}. \tag{145}$$

Posterior-predictive forward cross-entropy. From (127),

$$\mathcal{H}[P^*; P_{\rm pp}] = \ln \mathcal{Y}(C) - \langle \ln \mathcal{Y}(C_{\mathbf{x}}) \rangle_{\mathbf{x}\sim P^*}. \tag{146}$$

The eigenvalues of $C_{\mathbf{x}}$ when $\mathbf{x} \sim P^*$ concentrate (for $K = 2$) to three deterministic values $(\lambda_+, \lambda_-, c_2)$ where

$$\lambda_\pm = \frac{1}{2}\left\{ c_1 + \frac{1}{\eta} \pm \sqrt{\left(c_1 - \frac{1}{\eta} \right)^2 + \frac{2}{\eta}(c_1 - c_2)^2} \right\}, \tag{147}$$

with $\lambda_+ + \lambda_- + c_2 = 2 + 1/\eta$.

D.7 KL divergences

We now assemble the four KL divergences from the cross-entropies above and the entropies of P^* and P_W in (125) and (27). Differences of $\ln \mathcal{Y}$ are taken at leading $\mathcal{O}(N)$ from App. A.1.3.

Posterior-average reverse KL. Combining (142) with (27),

$$\frac{1}{N}\langle D_{\rm KL}(P_W \| P^*) \rangle_{W\sim P_\eta} = \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} + \frac{1}{2}\left(\omega^* - 1 - \ln \omega^* + F(\mu) \right). \tag{148}$$

with $\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}}$ from (143).

Posterior-predictive reverse KL. Combining (142) with (140),

$$\begin{aligned} \frac{1}{N} D_{\rm KL}(P_{\rm pp} \| P^*) = \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} + \frac{\ln(2\pi) + \omega^* - \ln \omega^*}{2} \\ + \Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3) - \Phi(c_1, c_2), \end{aligned} \tag{149}$$

with $\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}}$ from (143).

Posterior-average forward KL. Combining (145) with the teacher entropy $\mathcal{H}[P^*]/N = \frac{1}{2}\ln(2\pi e) - \frac{1}{2}\ln \omega^*$ from (125) and the posterior-average $\langle \ln Z(W) \rangle_{W\sim P_\eta}/N$ from (25) (the $\frac{1}{2}\ln(2\pi)$ base-measure constant cancels between the two),

$$\frac{1}{N}\langle D_{\rm KL}(P^* \| P_W) \rangle_{W\sim P_\eta} = -\frac{\eta c_1 \chi_1 - 1}{\eta} \frac{(\omega^* - 1)^2}{(2\omega^* - 1)^2} + \frac{\mu - F(\mu)}{2} - \frac{1 - \ln \omega^*}{2}. \tag{150}$$

Posterior-predictive forward KL. Subtracting the teacher entropy from (146) and evaluating $\ln \mathcal{Y}$ at the eigenvalues (147),

$$\frac{1}{N} D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = -\frac{\ln(2\pi e) - \ln \omega^*}{2} - \Phi(\lambda_+, \lambda_-, c_2) + \Phi(c_1, c_2), \tag{151}$$

with Φ the $\mathcal{O}(N)$ coefficient of the evidence (App. A.1.3). The optimum $\eta^{\mathrm{fwd}}_{\mathrm{pp}} = \arg\min_{\eta>0} D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}})$ admits closed-form thresholds in the (ω^*, γ) plane derived in App. D.11.2.

D.8 Posterior-predictive gap identities

We show that:

$$\langle D_{\mathrm{KL}}(P_W \| P^*) \rangle_{W \sim P_\eta} - D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \mathrm{MI}(W; \mathbf{x}|\mathcal{D}) \geq 0, \tag{152}$$

$$\langle D_{\mathrm{KL}}(P^* \| P_W) \rangle_{W \sim P_\eta} - D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = \left\langle D_{\mathrm{KL}}(P_\eta(W|\mathcal{D}) \| P_\eta(W|\mathcal{D}, \mathbf{x})) \right\rangle_{\mathbf{x} \sim P^*} \geq 0, \tag{153}$$

proving the inequalities stated in Sec. 3 of the main text. These gaps vanish in the MAP limit.

Reverse KL gap. Using (141):

$$\langle D_{\mathrm{KL}}(P_W \| P^*) \rangle - D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \mathcal{H}[P_{\mathrm{pp}}] - \langle \mathcal{H}[P_W] \rangle_{W \sim P_\eta} \tag{154}$$

The right-hand side equals the mutual information $\mathrm{MI}(W; \mathbf{x}|\mathcal{D})$ between the posterior weights W and student-generated data $\mathbf{x}$:

$$\begin{aligned} \mathrm{MI}(W; \mathbf{x}|\mathcal{D}) &= \int P_\eta(W|\mathcal{D}) P(\mathbf{x}|W) \ln \frac{P(\mathbf{x}|W)}{P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D})} \mathrm{d}W \mathrm{d}\mathbf{x} \\ &= \mathcal{H}[P_{\mathrm{pp}}] - \langle \mathcal{H}[P_W] \rangle_{W \sim P_\eta} \geq 0 \end{aligned} \tag{155}$$

establishing (152).

Forward KL gap. We compute:

$$\langle D_{\mathrm{KL}}(P^* \| P_W) \rangle_{W \sim P_\eta} - D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = \int \mathrm{d}W \, P_\eta(W|\mathcal{D}) \int \mathrm{d}\mathbf{x} \, P^*(\mathbf{x}) \ln \frac{P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D})}{P(\mathbf{x}|W)} \tag{156}$$

Using the Bayes update $P_\eta(W|\mathcal{D}, \mathbf{x}) = P(\mathbf{x}|W) P_\eta(W|\mathcal{D}) / P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D})$, the inner integral over W becomes $D_{\mathrm{KL}}(P_\eta(W|\mathcal{D}) \,\|\, P_\eta(W|\mathcal{D}, \mathbf{x}))$, yielding (153).

Discussion. Both identities are classical Bayesian facts from two largely disjoint literatures. The reverse gap is the mutual information $\mathrm{MI}(W; \mathbf{x})$, equivalent to Lindley's expected-information gain [Lindley, 1956] and to the epistemic/aleatoric decomposition of predictive entropy used in Bayesian deep learning [Houlsby et al., 2011, Depeweg et al., 2018, Lakshminarayanan et al., 2017]. The forward gap goes back to [Aitchison, 1975], who showed that P_{pp} dominates typical P_W under KL loss. The distinction between posterior predictive and a typical posterior sample is mainstream in feed-forward Bayesian models but, to our knowledge, has not been emphasized for generative models.

D.9 Sampling temperature tuning

We focus on the effect of sampling temperature tuning on the forward KL. We have:

$$D_{\mathrm{KL}}(P^* \| P_{\beta W}) = \beta \mathcal{H}[P^*; P_W] - \mathcal{H}[P^*] + \ln Z(\beta W) - \beta \ln Z(W) \tag{157}$$

This is a convex function of β, which therefore has a single minimum:

$$\beta_{\mathrm{TT}} = \operatorname*{argmin}_\beta D_{\mathrm{KL}}(P^* \| P_{\beta W}) \tag{158}$$

To determine whether the optimal β_{TT} is > 1 or < 1, it suffices to compute the sign of:

$$\left. \frac{\partial D_{\mathrm{KL}}(P^* \| P_{\beta W})}{\partial \beta} \right|_{\beta=1} = \mathcal{H}[P^*; P_W] - \mathcal{H}[P_W] \tag{159}$$

$\beta_{\rm TT}$ is < 1 or > 1 according to whether (159) is positive or negative, respectively. Note that equations (157) and (159) are general identities, valid for any temperature-tuned energy-based model.

Now we specialize to the teacher-student scenario introduced in the main text, with $K = 2$ and the rank-one teacher. Averaging over the posterior $W \sim P_\eta$ and substituting the forward average energy (145) and the model average energy (26) gives, to leading order in N,

$$\frac{1}{N}\left.\frac{\partial D_{\rm KL}(P^*\|P_{\beta W})}{\partial\beta}\right|_{\beta=1} = \frac{\mu-1}{2} - \left(\frac{\omega^*-1}{2\omega^*-1}\right)^2\left(c_1\chi_1 - \frac{1}{\eta}\right) \tag{160}$$

$$= \begin{cases} \frac{1}{2\gamma\eta} - \frac{1}{\gamma}\left(\frac{\omega^*-1}{\omega^*}\right)^2 & h = 0 \\ \frac{1}{\sqrt{\gamma\eta}} - \frac{1}{2} - \frac{1}{\gamma}\left(\frac{\omega^*-1}{\omega^*}\right)^2 & h \neq 0, u_1 = 0 \\ \frac{\lambda_1-1}{2} - \left(\frac{\omega^*-1}{2\omega^*-1}\right)^2\left(c_1\lambda_1 - \frac{1}{\eta}\right) & h, u_1 \neq 0 \end{cases} \tag{161}$$

after substituting the saddle-point values for μ and χ_1 in each phase.

After temperature tuning, the model can be transported from one phase to another.

- If $h = u_1 = 0$, then $\beta_{\rm TT}$ is < 1 or > 1 according to whether $2\eta[(\omega^*-1)/\omega^*]^2$ is < 1 or > 1. In this case, the temperature-tuned model remains in the initial phase (with $h = u_1 = 0$).
- If $h \neq 0$ and $u_1 = 0$, then $\beta_{\rm TT} < 1$ is such that the temperature-tuned model lands in a phase with $h = u_1 = 0$, killing the condensed random direction.
- If $h = 0$ and $u_1 \neq 0$, $\beta_{\rm TT} > 1$ or < 1 according to whether $\frac{\omega^{*2}}{2(\omega^*-1)^2}$ is $< \eta$ or $> \eta$. If $\frac{(\omega^*-1)^2(2\omega^*-1)}{\omega^*} < \frac{\gamma}{2\eta}$ the model stays in the same phase ($h = 0, u_1 \neq 0$), but otherwise it activates the dormant aligned mode, landing in $h, u_1 \neq 0$, which is beneficial because this mode was already aligned to the teacher.
- If $h, u_1 \neq 0$, both $\beta_{\rm TT} > 1$ or < 1 are possible, and the model can stay in the same phase or move to $h = 0, u_1 \neq 0$.

Generally, since β only rescales the spectrum of W without deforming it, temperature tuning cannot create new data-aligned outliers or merge two distinct outliers. Temperature tuning can only activate or deactivate an already present aligned mode.

D.10 Double descent

We derive the threshold $\eta_{\rm DD}(\omega^*)$ above which the typical reverse KL of Fig. 4 acquires a local minimum inside the data-aligned condensed phase $h \neq 0, u_1 \neq 0$, creating the first descent of the double-descent profile. The post-condensation $h = 0$ branch is governed by

$$\tfrac{1}{N}\left\langle D_{\rm KL}(P_W\|P^*)\right\rangle_{W\sim P_\eta} = \frac{\omega^* - 1 - \ln\omega^*}{2} + \frac{1}{4\gamma\eta}, \tag{162}$$

with γ-derivative $-1/(4\eta\gamma^2) < 0$ strictly negative for any finite η; the second descent therefore exists trivially, and the analysis reduces to the condensed branch. We work in the $K = 2$ rank-one teacher of App. D.1 with eigenvalues $c_1 = 2 - 1/\omega^*$, $c_2 = 1/\omega^*$ from (123).

On the condensed branch the top-outlier Stieltjes transform $g_1 = G(\lambda_1)$ provides a smooth parametrization,

$$\gamma(g_1) = 2g_1^2 - \frac{g_1}{\omega^*}, \qquad \lambda_1 = \frac{1}{g_1} + \frac{g_1}{\gamma\eta}, \qquad \frac{1}{\omega^*} < g_1 < 1, \tag{163}$$

with the outlier-detachment condition $g_1 < \sqrt{\gamma\eta}$ equivalent to $\eta > \eta_F(g_1) := \omega^* g_1/(2\omega^* g_1 - 1)$, and the branch terminating at the condensation threshold $g_1 = 1$, $\gamma = c_1$. Substituting (163) into (148) and using $G(\mu) = 1 - h^2 = g_1$ on this branch yields the intensive reverse KL as a function of g_1,

$$\tfrac{1}{N}\left\langle D_{\rm KL}(P_W\|P^*)\right\rangle_{W\sim P_\eta} = \\ \frac{\omega^* - 1 - \ln(\omega^* g_1)}{2} - \frac{\omega^*(\omega^*-1)(1-g_1)}{2\omega^*-1}\left[1 - \frac{\omega^* g_1}{\eta(2\omega^*-1)}\right] + \frac{\omega^* g_1}{4\eta(2\omega^* g_1 - 1)}, \tag{164}$$

which matches (162) continuously at $g_1 = 1$.

Since $\mathrm{d}\gamma/\mathrm{d}g_1 = 4g_1 - 1/\omega^* > 0$ on this branch, extrema in γ and in g_1 coincide. With

$$g_{\mathrm{MAP}} := \frac{2\omega^* - 1}{2\omega^*(\omega^* - 1)}, \tag{165}$$

the g_1-derivative writes,

$$\frac{\mathrm{d}}{\mathrm{d}g_1}\Big[\tfrac{1}{N}\big\langle D_{\mathrm{KL}}(P_W\|P^*)\big\rangle_{W\sim P_\eta}\Big] = \frac{\omega^*(\omega^*-1)(g_1 - g_{\mathrm{MAP}})}{g_1(2\omega^*-1)} + \frac{1}{\eta}\left[\frac{\omega^{*2}(\omega^*-1)(1-2g_1)}{(2\omega^*-1)^2} - \frac{\omega^*}{4(2\omega^* g_1 - 1)^2}\right]. \tag{166}$$

In the MAP limit only the first term survives, and is positive on $(g_{\mathrm{MAP}}, 1)$ when $g_{\mathrm{MAP}} < 1$, equivalently $2\omega^{*2} - 4\omega^* + 1 > 0$, equivalently

$$\omega^* > \omega_{\mathrm{DD}} := 1 + \frac{1}{\sqrt{2}} \tag{167}$$

Below ω_{DD} the condensed branch is monotone for every η and no first minimum exists.

When (167) holds, an interior local minimum appears at $g_1 \in (g_{\mathrm{MAP}}, 1)$ as soon as the slope (166) is positive and the outlier-detachment condition $\eta > \eta_F(g_1)$ holds. The sharp threshold is the one-dimensional variational formula

$$\eta_{\mathrm{DD}}(\omega^*) = \inf_{g_{\mathrm{MAP}} < g_1 < 1} \frac{\omega^* g_1}{2\omega^* g_1 - 1} \times \max\left\{1,\ \frac{2\omega^* g_1 - 1}{(2\omega^*-1)(g_1 - g_{\mathrm{MAP}})}\left[(2g_1 - 1) + \frac{(2\omega^*-1)^2}{4\omega^*(\omega^*-1)(2\omega^* g_1 - 1)^2}\right]\right\}, \tag{168}$$

with $\eta_{\mathrm{DD}}(\omega^*) = \infty$ for $\omega^* \le \omega_{\mathrm{DD}}$. For $\eta_{\mathrm{DD}} < \eta < \infty$ the reverse KL has the double-descent profile of Fig. 4; the MAP limit collapses the post-transition branch to a flat line, so the second descent ceases to be strict while the first minimum survives.

D.11 Tempered-posterior effects

The optimal posterior temperature η_{pp} can be defined by minimising either the reverse KL $D_{\mathrm{KL}}(P_{\mathrm{pp}}\|P^*)$ or the forward KL $D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}})$ over $\eta > 0$. In both cases the (ω^*, γ) plane splits into warm ($\eta_{\mathrm{pp}} < 1$), cold ($\eta_{\mathrm{pp}} > 1$), and MAP ($\eta_{\mathrm{pp}} \to \infty$) regions, and the two definitions agree at the qualitative level: increasing ω^* or γ shifts the optimum from warm to cold under either KL. The phase boundaries and the structure of the optimum nevertheless differ quantitatively, as we work out in the following two subsections.

D.11.1 Optimal tempered-posterior according to the reverse KL

We minimise (149) over $\eta > 0$ at fixed (ω^*, γ) and obtain the five regions of Fig. 5A. On a smooth condensed branch, evaluating $\Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3)$ along the $m^2_{1,2}(\eta)$-parametrisation determined by (139) reduces (149) to

$$\tfrac{1}{N} D_{\mathrm{KL}}(P_{\mathrm{pp}}\|P^*) = \frac{\omega^* - 1 - \ln\omega^*}{2} - \frac{m_1^2}{2g_{\mathrm{MAP}}} - \frac{1}{2}\ln(1 - m_1^2 - m_2^2), \tag{169}$$

with g_{MAP} as in (165); all (γ, η)-dependence of the right-hand side is carried by m_1^2, m_2^2, while ω^* and g_{MAP} are teacher-only constants. The term $m_1^2/(2g_{\mathrm{MAP}})$ is the energetic gain along the teacher-aligned mode; m_2^2 enters only through the entropy because the second empirical mode is teacher-orthogonal. Differentiating in η gives the stationarity condition

$$\partial_\eta m_1^2 = \frac{g_{\mathrm{MAP}}\big(\partial_\eta m_1^2 + \partial_\eta m_2^2\big)}{1 - m_1^2 - m_2^2}. \tag{170}$$

Zero-overlap regions. The bound $-\ln(1-s) \ge s$ on $s = m_1^2 + m_2^2$ applied to (169) gives $D_{\mathrm{KL}}(P_{\mathrm{pp}}\|P^*)/N - (\omega^* - 1 - \ln\omega^*)/2 \ge [(1 - 1/g_{\mathrm{MAP}})m_1^2 + m_2^2]/2$, so for $g_{\mathrm{MAP}} \ge 1$, equivalently

$\omega^* \le \omega_0 := 1+1/\sqrt{2}$, the leading reverse KL is minimised at $m_1^2 = m_2^2 = 0$ and equals the uniform-student baseline $D_{\mathrm{KL}}(P_{W=0}\|P^*)/N = (\omega^* - 1 - \ln\omega^*)/2$ throughout the zero-overlap sector. For $\gamma < c_1$ this sector is the warm interval $\eta \in (0, \gamma/c_1^2]$ and the optimum is degenerate and warm (*warm deg.* in Fig. 5A). For $\gamma \ge c_1$, at any teacher strength, both overlaps vanish for all $\eta > 0$ and the leading reverse KL equals the uniform-student baseline throughout (*deg.*).

Single-mode branch. For $\omega^* > \omega_0$ and $\gamma < c_1$ a positive-overlap optimum exists. On the $d = 1$ outlier branch, $m_2^2 \equiv 0$, and (170) reduces to the scalar target

$$(m_1^2)^* = 1 - g_{\mathrm{MAP}}. \tag{171}$$

Tracking $m_1^2(\eta)$ along the smooth branch with the g_1-parametrisation of (176), the value $\eta_{\mathrm{pp}}^{\mathrm{rev}} = 1$ is hit on

$$\gamma_{\mathrm{wc}}^{(1)}(\omega^*) = 2g_{\mathrm{wc}}^2 - c_2\, g_{\mathrm{wc}}, \qquad g_{\mathrm{wc}} = \tfrac{1}{4}\Big[(c_2+3c_1) - \sqrt{(c_2+3c_1)^2 - 16c_1 g_{\mathrm{MAP}}}\Big], \tag{172}$$

and $\eta_{\mathrm{pp}}^{\mathrm{rev}} \to \infty$ on

$$\gamma_{\infty}^{\mathrm{rev}}(\omega^*) = 2g_{\mathrm{MAP}}^2 - c_2\, g_{\mathrm{MAP}} = \frac{2\omega^* - 1}{2\omega^*(\omega^* - 1)^2}. \tag{173}$$

Two-mode correction. The $d = 2$ co-condensed branch exists only for $\gamma < c_2^2$ and pre-empts the $d = 1$ optimum when its turn-on tangent is downhill, $g_{\mathrm{MAP}} < 1 - (c_1 - c_2)^2/(2c_1)$, equivalently $\omega^* > \omega_2 := (3+\sqrt{3})/2$. Substituting the edge-branch overlaps $m_i^2 = (c_i/2)(1 - \sqrt{\gamma/\eta}/c_i)^2$ into (170) fixes the warm/cold boundary at $\gamma_{\mathrm{wc}}^{(2)} = \alpha^2$, with $\alpha \in (0, c_2)$ the unique root of

$$\alpha\,(c_1 - \alpha)(2c_1c_2 - \alpha) = 2c_1\, g_{\mathrm{MAP}}\,(c_1c_2 - \alpha). \tag{174}$$

The cold/MAP boundary (173) is unchanged.

Phase classification. For $\omega^* \le \omega_0$ the optimum is the zero-overlap plateau (*warm deg.* for $\gamma < c_1$, *deg.* for $\gamma \ge c_1$). For $\omega^* > \omega_0$, set $\gamma_{\mathrm{wc}} = \gamma_{\mathrm{wc}}^{(1)}$ when $\omega^* \le \omega_2$ and $\gamma_{\mathrm{wc}} = \gamma_{\mathrm{wc}}^{(2)}$ when $\omega^* > \omega_2$; the regions are then *warm* for $\gamma < \gamma_{\mathrm{wc}}$, *cold* for $\gamma_{\mathrm{wc}} < \gamma < \gamma_{\infty}^{\mathrm{rev}}$, *MAP* for $\gamma_{\infty}^{\mathrm{rev}} \le \gamma < c_1$, and *deg.* for $\gamma \ge c_1$. At the MAP optimum, the predictive reverse KL equals the $\eta \to \infty$ limit of the general typical-posterior expression (148).

D.11.2 Optimal tempered-posterior according to the forward KL

We minimise (151) over $\eta > 0$ at fixed (ω^*, γ):

$$\eta_{\mathrm{pp}}^{\mathrm{fwd}} = \arg\min_{\eta>0} D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}}) \tag{175}$$

and obtain three closed-form thresholds, $\gamma_{\mathrm{wc}} < \gamma_{\mathrm{flat}} < \gamma_{\infty}$. Together with $\gamma = c_1$, they partition the (ω^*, γ) plane into the five regions of Fig. 13A.

Smooth branch. On the condensed branch outside the warm flat regime defined below, the perturbed and original sources carry the same number of outliers and the η-dependence of (151) does not cancel. The interior stationary condition $\partial_\eta D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}}) = 0$ reduces to matching the evidence saddle variable g_1 between $\mathcal{Y}(C_{\mathbf{x}})$ and $\mathcal{Y}(C)$, namely $g_1^{(C_{\mathbf{x}})} = g_1^{(C)}$, which using (147) simplifies to $\lambda_- = g_1/\eta$. Combined with the elementary identities $\lambda_+\lambda_- = \big[c_1 - \frac{1}{2}(c_1 - c_2)^2\big]/\eta$ and $\lambda_+ + \lambda_- = c_1 + 1/\eta$, this gives the smooth-branch stationary point $\eta_0(\gamma, \omega^*)$ of the forward KL,

$$\eta_0(\gamma, \omega^*) = \frac{g_1\,(1 - g_1)}{c_1(1 - g_1) - \frac{1}{2}(c_1 - c_2)^2}, \qquad g_1(\gamma, \omega^*) = \frac{c_2 + \sqrt{c_2^2 + 8\gamma}}{4}. \tag{176}$$

Since g_1 is monotone in γ, η_0 diverges at $g_1 = 1 - (c_1 - c_2)^2/(2c_1)$. Substituting back into $\gamma = 2g_1^2 - c_2 g_1$ yields the cold/MAP boundary

$$\gamma_{\infty}(\omega^*) = 2\Big[1 - \tfrac{(c_1-c_2)^2}{2c_1}\Big]^2 - c_2\Big[1 - \tfrac{(c_1-c_2)^2}{2c_1}\Big] = \frac{(3\omega^* - 2)(4\omega^* - 3)}{\omega^{*2}(2\omega^* - 1)^2}. \tag{177}$$

The smooth branch crosses $\eta_0 = 1$ at $g_1 = c_2$, equivalently

$$\gamma_{\mathrm{wc}}(\omega^*) = c_2^2 = 1/\omega^{*2}. \tag{178}$$

Warm flat interval. For $\eta < 1$ one has $\lambda_-(\eta) > c_2$, and when $\lambda_-(\eta) \geq \sqrt{\gamma/\eta}$ the perturbed source $(\lambda_+, \lambda_-, c_2)$ acquires one outlier above the bulk edge in addition to those of (c_1, c_2). In this regime the η-dependent pieces of $-\Phi(\lambda_+, \lambda_-, c_2) + \Phi(c_1, c_2)$ cancel exactly, producing a flat plateau in $D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}})$. Setting $\lambda_-(\eta) = \sqrt{\gamma/\eta}$ yields a quadratic in $\sqrt{\gamma/\eta}$,

$$\frac{\gamma}{\eta} - \left[\gamma + c_1 - \tfrac{1}{2}(c_1 - c_2)^2\right]\sqrt{\gamma/\eta} + \gamma\, c_1 = 0, \tag{179}$$

whose two positive roots delimit the warm interval

$$I_{\mathrm{warm}}(\gamma, \omega^*) = \left[\eta_-,\ \min(1, \eta_+)\right], \tag{180}$$

with endpoints

$$\eta_\pm = \frac{4\gamma}{\left[\gamma + c_1 - \tfrac{1}{2}(c_1 - c_2)^2 \mp \sqrt{\left[\gamma + c_1 - \tfrac{1}{2}(c_1 - c_2)^2\right]^2 - 4\gamma c_1}\right]^2}. \tag{181}$$

Every $\eta \in I_{\mathrm{warm}}$ is a global minimiser. The discriminant of (179) vanishes at

$$\gamma_{\mathrm{flat}}(\omega^*) = \left(\sqrt{c_1} - (c_1 - c_2)/\sqrt{2}\right)^2, \tag{182}$$

above which the warm flat regime no longer exists. For $\omega^* > 1$ the three thresholds satisfy $\gamma_{\mathrm{wc}} < \gamma_{\mathrm{flat}} < \gamma_\infty$, with equality only in the degenerate limit $\omega^* \downarrow 1$.

Uncondensed plateau. For $\gamma > c_1$, both the original and perturbed sources are uncondensed for $\eta \geq \eta_{\mathrm{PM}}$. Setting $\lambda_+(\eta) = \gamma$ in (147) gives the threshold

$$\eta_{\mathrm{PM}} = \frac{\gamma - c_1 + \tfrac{1}{2}(c_1 - c_2)^2}{\gamma(\gamma - c_1)}.$$

Substituting the uncondensed branch of (32) into (151) gives, throughout this interval,

$$\frac{1}{N} D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = \frac{\ln \omega^*}{2} - \frac{(c_1 - c_2)^2}{4\gamma}.$$

The leading forward KL therefore attains its MAP value at every $\eta \in [\eta_{\mathrm{PM}}, \infty)$, giving a degenerate optimum (*deg.* in Fig. 13A). Unlike the reverse-KL plateau, this interval does not include all positive η.

Phase classification. The minimisation has five outcomes (Fig. 13A). For $\gamma < \gamma_{\mathrm{wc}}$ the warm flat interval $I_{\mathrm{warm}} \subset (0, 1]$ is the optimum-set, with smooth-branch representative $\eta_0 < 1$ inside; at $\gamma = \gamma_{\mathrm{wc}}$, $\eta_0 = 1$. In the sliver $\gamma_{\mathrm{wc}} < \gamma \leq \gamma_{\mathrm{flat}}$ a warm flat sub-interval $I_{\mathrm{warm}} \subset (0, 1)$ and the cold isolated point $\eta_0 > 1$ are simultaneously global minimisers; $D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}})/N$ traces a non-convex W shape with the two troughs at the same value to leading order in N. At $\gamma = \gamma_{\mathrm{flat}}$ the warm interval collapses to a single point, still distinct from the cold minimiser. For $\gamma_{\mathrm{flat}} < \gamma < \gamma_\infty$ the unique global minimiser is the cold smooth-branch point $\eta_{\mathrm{pp}}^{\mathrm{fwd}} = \eta_0 > 1$. For $\gamma_\infty \leq \gamma \leq c_1$ the infimum is reached only as $\eta \to \infty$ (MAP), whereas for $\gamma > c_1$ every $\eta \geq \eta_{\mathrm{PM}}$ is a global minimiser.

D.12 Training dynamics of the KL divergences

We derive asymptotically exact expressions tracking the large-N limit of the intensive instantaneous KL divergences over the course of training. We focus on the typical KL divergences

$$\frac{1}{N} D_{\mathrm{KL}}(P^* \| P_t) = \frac{1}{N}\mathcal{H}[P^*; P_t] - \frac{1}{N}\mathcal{H}[P^*], \tag{183}$$

$$\frac{1}{N} D_{\mathrm{KL}}(P_t \| P^*) = \frac{1}{N}\mathcal{H}[P_t; P^*] - \frac{1}{N}\mathcal{H}[P_t], \tag{184}$$

where we used the shorthand $P_t = P_{W(t)}$ and $P^* = P_{W^*}$, with W^* the rank-one teacher of App. D.1.

D.12.1 Student average energy and entropy

Let $\lambda_1(t)$ be the top eigenvalue of $W(t)$ and $\mu(t) = \max\{G^{-1}(1), \lambda_1(t)\}$. Then, similar to Eqs. (25), (26) and (27),

$$\begin{aligned} \frac{1}{N}\ln Z(W(t)) &= \frac{\ln(2\pi) + \mu(t) - F(\mu(t))}{2} \\ \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W(t))\rangle_{\mathbf{x}\sim P_t} &= \frac{1 - \mu(t)}{2}, \\ \frac{1}{N}\mathcal{H}[P_{W(t)}] &= \frac{\ln(2\pi e) - F(\mu(t))}{2} \end{aligned} \tag{185}$$

D.12.2 Forward cross entropy

We first compute

$$\frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W(t))\rangle_{\mathbf{x}\sim P^*} = -\frac{1}{2}\left(1-\frac{1}{\omega^*}\right)\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^* \tag{186}$$

where

$$\Theta(t) = \frac{1-e^{-\gamma t/2}}{\gamma}C - \frac{1}{N}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}\mathbf{x}(u)\mathbf{x}(u)^\top \tag{187}$$

and where we used $\frac{\langle \mathbf{x}^\top W(t)\mathbf{x}\rangle_{P^*}}{N} = \frac{\langle(\mathbf{w}^*\cdot\mathbf{x})^2\rangle_{P^*}}{N}\mathbf{w}^{*\top}W(t)\mathbf{w}^*$ and $\frac{\langle(\mathbf{w}^*\cdot\mathbf{x})^2\rangle_{P^*}}{N} = 1-1/\omega^*$, to leading order, since $\mathbf{w}^{*\top}W_{\mathrm{GOE}}(t)\mathbf{w}^* = o(1)$. Here $\Theta(t)$ has at most K positive eigenvalues $\theta_k(t)$, with eigenvectors $\mathbf{t}_k$. Decomposing $\mathbf{w}^*$ onto $\mathbf{c}_1,\dots,\mathbf{c}_K$, one gets

$$\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^* = \left(1-\frac{1}{2\omega^*-1}\right)\left(\frac{1-e^{-\gamma t/2}}{\gamma}c_1 - \int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}s_1(u)^2\right), \tag{188}$$

where we used Eq. (124).

We then find:

$$\frac{1}{N}\mathcal{H}[P^*; P_t] = \frac{\ln 2\pi + \mu(t) - F(\mu(t))}{2} - \frac{1}{2}\left(1-\frac{1}{\omega^*}\right)\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^*. \tag{189}$$

Determining $\mu(t)$ or $g_t := G(\mu(t))$ requires knowledge of $\lambda_1(t)$. To make progress, we use the approximation that $W_{\mathrm{GOE}}(t)$ is asymptotically independent of $\Theta(t)$. Note that, for $u \le t$, $\Theta(t)$ depends on $W_{\mathrm{GOE}}(u)$ through $\mathbf{x}(u)$. Therefore, this approximation is valid for t small enough that the negative phase is small relative to the positive phase.

Properties of finite-rank deformed Wigner matrices then imply

$$g_t = \begin{cases} \min\{1, \sqrt{\gamma\eta}\} & \theta_1(t) \le \theta_c, \\ \theta_1(t)^{-1}, & \theta_1(t) > \theta_c, \end{cases} \qquad \text{with} \qquad \theta_c = \max\{1, 1/\sqrt{\gamma\eta}\}, \tag{190}$$

$h_t^2 = (1-g_t)_+$ and $\mu(t) = g_t/(\gamma\eta) + 1/g_t$.

D.12.3 Reverse cross entropy

Injecting W^* gives $\frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*)\rangle_{\mathbf{x}\sim P_t} = -\frac{\omega^* h_t^2}{2}\frac{\langle(\mathbf{v}_1\cdot\mathbf{w}^*)^2\rangle}{N} + o(1)$ upon using $\frac{\omega^*}{2}\frac{\|\mathbf{x}\|_2^2}{N} = \mathcal{O}(1)$ and $\frac{\langle(\mathbf{w}^*\cdot\mathbf{x})^2\rangle}{N} = h_t^2\frac{\langle(\mathbf{v}_1\cdot\mathbf{w}^*)^2\rangle}{N} + o(1)$. Therefore,

$$\frac{1}{N}\mathcal{H}[P_t; P^*] = \frac{1}{2}\ln 2\pi + \frac{1}{2}(\omega^* - \ln\omega^*) - \frac{1}{2}\omega^* h_t^2\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} + o(1). \tag{191}$$

We again make the early-training approximation that $\Theta(t)$ is asymptotically independent of $W_{\mathrm{GOE}}(t)$. Once $\lambda_1(t)$ has detached one has, from spectral properties of finite-rank deformed Wigner matrices,

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{t}_1(t))^2}{N^2} = \left(1-\frac{1}{\gamma\,\eta\,\theta_1(t)^2}\right)_+ + o(1). \tag{192}$$

The component of $\mathbf{v}_1(t)$ orthogonal to $\mathbf{t}_1(t)$ is asymptotically isotropic in the $(N-1)$-dimensional subspace orthogonal to $\mathbf{t}_1(t)$ and is therefore asymptotically orthogonal to $\mathbf{w}^*$. As a result,

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} = \left(1-\frac{1}{\gamma\,\eta\,\theta_1(t)^2}\right)_+\frac{(\mathbf{t}_1(t)\cdot\mathbf{w}^*)^2}{N} + o(1). \tag{193}$$

To evaluate $\mathbf{t}_1(t)\cdot\mathbf{w}^*$, expand $\mathbf{t}_1(t)$ in the eigenbasis of the empirical covariance of the training data and use $(\mathbf{c}_1\cdot\mathbf{w}^*)^2/N = 1-1/(2\omega^*-1)$, $(\mathbf{c}_k\cdot\mathbf{w}^*)^2/N = o(1)$ $(k > d)$. In this way, we find

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} = \left(1-\frac{1}{\gamma\eta\,\theta_1(t)^2}\right)\left(1-\frac{1}{2\omega^*-1}\right)\frac{(\mathbf{t}_1(t)\cdot\mathbf{c}_1)^2}{N^2}. \tag{194}$$

D.12.4 Forward KL

Combining Eqs. (125) and (189) yields

$$\frac{1}{N}D_{\mathrm{KL}}(P^*\,\|\,P_t) = \frac{1}{2}\ln\omega^* - \frac{1}{2} + \frac{1}{2}\left[\mu(t) + \ln g_t - \frac{g_t^2}{2\gamma\eta}\right] \\ - \frac{1}{2}\left(1-\frac{1}{\omega^*}\right)\left(1-\frac{1}{2\omega^*-1}\right)\left(\frac{1-e^{-\gamma t/2}}{\gamma}c_1 - \int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}s_1(u)^2\right). \tag{195}$$

D.12.5 Reverse KL

Using Eqs. (191), (194),

$$\frac{1}{N}D_{\mathrm{KL}}(P_t\,\|\,P^*) = \frac{1}{2}(\omega^* - 1 - \ln\omega^*) - \frac{1}{2}\ln g_t + \frac{g_t^2}{4\gamma\eta} \\ - \frac{1}{2}\omega^* h_t^2\left(1-\frac{1}{\gamma\eta\,\theta_1(t)^2}\right)\left(1-\frac{1}{2\omega^*-1}\right)\frac{\left(\mathbf{t}_1(t)\cdot\mathbf{c}_1\right)^2}{N^2}. \tag{196}$$

At very early times, $t < t_p = \inf\{t : \theta_1(t) > \max\{1, 1/\sqrt{\gamma\eta}\}\}$, until the student condenses along an outlier direction, the reverse KL stays flat, with plateau height

$$\frac{1}{N}D_{\mathrm{KL}}(P_t\,\|\,P^*) = \frac{1}{2}(\omega^* - 1 - \ln\omega^*) - \frac{1}{2}\ln g_0 + \frac{g_0^2}{4\gamma\eta}, \qquad g_0 = \begin{cases}1, & \gamma\eta \geq 1\\ \sqrt{\gamma\eta}, & \gamma\eta < 1.\end{cases} \tag{197}$$

Optimal stopping time. For $t > t_p$, the saddle is controlled by the data-aligned spike, $g_t = \theta_1(t)^{-1}$, and the behavior of the reverse KL results from the trade-off between the (i) marginal gain of increasing the student's alignment with the teacher and (ii) the marginal cost of decreasing the student's diversity. The optimal stopping time t^* is reached when (i) and (ii) are balanced,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\omega^* h_t^2\left(1-\tfrac{1}{\gamma\eta\,\theta_1(t)^2}\right)\left(1-\tfrac{1}{2\omega^*-1}\right)\tfrac{(\mathbf{t}_1(t)\cdot\mathbf{c}_1)^2}{N^2}\right)\Big|_{t=t^*} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\ln\theta_1(t) + \tfrac{1}{2\gamma\eta\theta_1(t)^2}\right)\Big|_{t=t^*}.$$

When $u_1^2(t^*) \simeq 1$, setting $\partial_\theta D_{\mathrm{KL}}(P_t\,\|\,P^*)|_{\theta=\theta^*} = 0$ gives a quadratic in θ^*, whose positive root is

$$\theta^\star = \frac{1}{2}\left[\omega^*\left(1-\frac{1}{2\omega^*-1}\right) + \sqrt{\omega^{*2}\left(1-\frac{1}{2\omega^*-1}\right)^2 + \frac{4}{\gamma\eta}}\right]. \tag{198}$$

The early-training approximation, $\theta_1(t) \simeq \frac{c_1}{\gamma}(1-e^{-\gamma t/2})$, leads to the ν-independent estimate

$$t^* \simeq -\frac{2}{\gamma}\ln\left(1-\frac{\gamma\,\theta^*}{c_1}\right). \tag{199}$$

Injecting the parameter values of Fig. 6 yields $t^* \simeq 5\ln 2 \simeq 3.47$, in excellent agreement with the numerical value.

D.13 Posterior-predictive forward KL

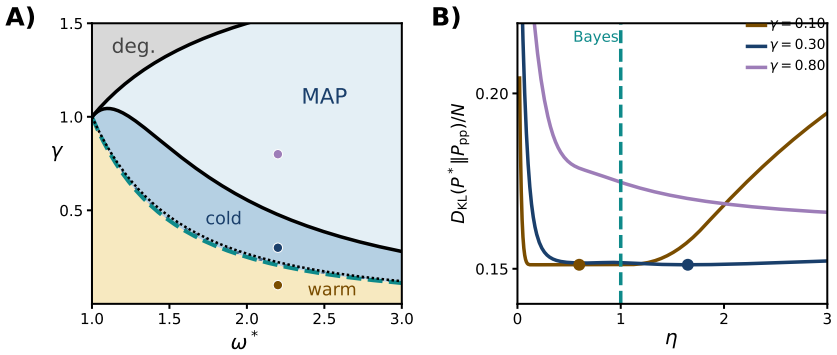


Figure 13: **Optimum** $\eta^{\mathrm{fwd}}_{\mathrm{pp}} = \arg\min_{\eta>0} D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}})$ **for the rank-one** $K{=}2$ **teacher. A)** Phases in the (ω^*, γ) plane: warm flat (yellow), unique cold (medium blue), MAP (light blue), and degenerate (gray), where all $\eta \geq \eta_{\mathrm{PM}}$ attain the MAP value (App. D.11.2); the hatched sliver is the mixed warm/cold tie. Boundaries: dashed teal γ_{wc} (Bayes-crossing, $\eta^{\mathrm{fwd}}_{\mathrm{pp}} = 1$), dotted black γ_{flat} (upper edge of the warm flat interval), solid black γ_∞ (cold/MAP) and $\gamma = c_1$ (MAP/degenerate). **B)** $D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}})/N$ vs. η at $\omega^* = 2.2$ for $\gamma \in \{0.10, 0.30, 0.80\}$ marked in panel A; teal dashed at $\eta = 1$.

E Comparison of the SBM with a multivariate Gaussian model

The SBM and a multivariate Gaussian both have quadratic energies, but the spherical constraint permits sample condensation, in which a finite number of modes carries a nonzero fraction of the total squared norm. Spectral outlier transitions also occur in standard spiked Gaussian covariance models [Baik et al., 2005], where fixed spike strengths do not produce sample condensation. Spectral separation and sample condensation are therefore distinct phenomena.

Figure 14 illustrates this distinction using the $K = 1$ SBM phase diagram. The uncondensed sectors ($h = 0$), with or without a spectral outlier, have Gaussian analogues. The condensed sectors ($h \neq 0$) exhibit the additional condensation mechanism induced by the spherical constraint.

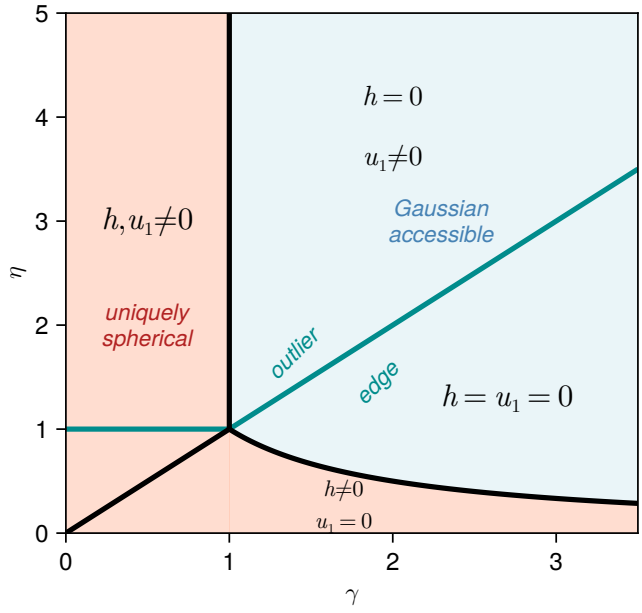


Figure 14: **Spectral outlier transitions and spherical condensation.** Phase diagram of the $K = 1$ SBM, with phases labeled by (h, u_1). Blue shading identifies uncondensed sectors whose spectral behavior has analogues in spiked Gaussian models. Salmon shading identifies sectors with sample condensation, absent from the Gaussian comparison with fixed spike strengths.

F First-order SBM reduction of a general spherical EBM

We describe a simple sense in which the SBM arises as a leading quadratic approximation of a more general EBM on the sphere. Let μ_0 denote the uniform probability measure on

$$\mathcal{S}_N = \left\{ \mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|^2 = N \right\}, \tag{200}$$

and consider an arbitrary spherical EBM

$$\frac{\mathrm{d}P_\epsilon}{\mathrm{d}\mu_0}(\mathbf{x}) = \frac{\exp[-\epsilon \mathcal{E}(\mathbf{x})]}{\langle \exp[-\epsilon \mathcal{E}] \rangle_0}, \tag{201}$$

where $\langle \cdot \rangle_0$ denotes averaging under μ_0 and $\mathcal{E}(\mathbf{x})$ is an arbitrary function, for example parameterized by a neural network. The parameter ϵ is a bookkeeping parameter controlling the departure from the uniform distribution. Here $\mathbf{x} \in \mathcal{S}_N$ may represent either original data on the sphere, or a normalized feature representation of a more general EBM.

Writing

$$\widetilde{\mathcal{E}}(\mathbf{x}) = \mathcal{E}(\mathbf{x}) - \langle \mathcal{E} \rangle_0, \tag{202}$$

the normalized density has the expansion

$$\frac{\mathrm{d}P_\epsilon}{\mathrm{d}\mu_0}(\mathbf{x}) = 1 - \epsilon \widetilde{\mathcal{E}}(\mathbf{x}) + \mathcal{O}(\epsilon^2). \tag{203}$$

Equivalently,

$$\log \frac{\mathrm{d}P_\epsilon}{\mathrm{d}\mu_0}(\mathbf{x}) = -\epsilon \widetilde{\mathcal{E}}(\mathbf{x}) + \mathcal{O}(\epsilon^2). \tag{204}$$

At order ϵ, the response of any centered quadratic observable depends only on the degree-two spherical-harmonic component of this log-density.

The degree-two spherical harmonics are precisely the traceless quadratic forms

$$\mathbf{x} \mapsto \frac{1}{2} \mathbf{x}^\top W \mathbf{x}, \qquad W \in \mathrm{Sym}_N, \qquad \mathrm{tr}\, W = 0.$$

The degree-two component is uniquely represented as $\frac{1}{2}\mathbf{x}^\top W_{\text{eff}} \mathbf{x}$, where $\mathrm{tr}\, W_{\text{eff}} = 0$. Equivalently,

$$W_{\text{eff}} = \underset{\substack{W \in \mathrm{Sym}_N \\ \mathrm{tr}\, W = 0}}{\arg\min} \left\langle \left[-\epsilon \widetilde{\mathcal{E}}(\mathbf{x}) - \frac{1}{2} \mathbf{x}^\top W \mathbf{x} \right]^2 \right\rangle_0. \tag{205}$$

This projection can be evaluated explicitly. The uniform spherical moments are

$$\langle x_i x_j \rangle_0 = \delta_{ij}, \qquad \langle x_i x_j x_k x_l \rangle_0 = \frac{N}{N+2} \left(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} \right). \tag{206}$$

Consequently, for any traceless $U \in \mathrm{Sym}_N$,

$$\left\langle \left(\mathbf{x}\mathbf{x}^\top - I \right) \frac{1}{2} \mathbf{x}^\top U \mathbf{x} \right\rangle_0 = \frac{N}{N+2} U. \tag{207}$$

The normal equations associated with Eq. (205) therefore give

$$W_{\text{eff}} = -\epsilon \frac{N+2}{N} \left\langle \left(\mathbf{x}\mathbf{x}^\top - I \right) \mathcal{E}(\mathbf{x}) \right\rangle_0. \tag{208}$$

Notice that $\mathrm{tr}\, W_{\text{eff}} = 0$ automatically, since $\|\mathbf{x}\|^2 = N$. Equivalently, for every traceless $U \in \mathrm{Sym}_N$,

$$\left\langle \left[-\epsilon \widetilde{\mathcal{E}}(\mathbf{x}) - \frac{1}{2} \mathbf{x}^\top W_{\text{eff}} \mathbf{x} \right] \mathbf{x}^\top U \mathbf{x} \right\rangle_0 = 0. \tag{209}$$

Thus $\frac{1}{2}\mathbf{x}^\top W_{\text{eff}} \mathbf{x}$ is exactly the degree-two component of the first-order log-density. If the EBM is invariant under $\mathbf{x} \mapsto -\mathbf{x}$, as in the zero-field setting studied here, its degree-one spherical harmonic vanishes. Otherwise, the degree-one component remains orthogonal to all the quadratic observables considered below.

The distribution obtained by retaining only this quadratic component is

$$P_\epsilon^{(2)}(\mathbf{x}) = \frac{1}{Z(W_{\text{eff}})} \exp\left(\frac{1}{2}\mathbf{x}^\top W_{\text{eff}}\mathbf{x}\right), \qquad \mathbf{x} \in \mathcal{S}_N \tag{210}$$

which is precisely an SBM. The general EBM and its projected SBM have identical quadratic observables to first order. In particular, defining the sampling covariance

$$\Sigma_\epsilon = \left\langle \mathbf{x}\mathbf{x}^\top \right\rangle_{P_\epsilon},$$

Eqs. (203), (208), and (207) give

$$\Sigma_\epsilon = I + \frac{N}{N+2} W_{\text{eff}} + \mathcal{O}(\epsilon^2) = \left\langle \mathbf{x}\mathbf{x}^\top \right\rangle_{P_\epsilon^{(2)}} + \mathcal{O}(\epsilon^2). \tag{211}$$

Therefore, to first order around the isotropic distribution, W_{eff} and $\Sigma_\epsilon - I$ have the same eigenvectors and proportional eigenvalues. Spectral alignment and the corresponding quadratic sample-overlap observables of a general spherical EBM thus have direct counterparts in the SBM.

This reduction is of course a first-order approximation only. It does not imply that the complete distribution is an SBM: higher spherical harmonics can affect higher-order observables already at order ϵ. A distribution-level approximation additionally requires the non-quadratic components of the log-density to be of order ϵ^2, and the perturbative argument by itself does not establish the non-perturbative condensation transition occurring at finite coupling.

We acknowledge an anonymous reviewer for suggesting this derivation.

G Computational experiments in generative models

We test the four SBM phenomena in tractable small models and larger generative architectures. Table 2 summarizes the observables and their correspondence with the SBM predictions.

Table 2: Summary of the experiments and their correspondence with the SBM predictions.

System / data	**Observables**	**Phenomenology**	**Comparison with SBM**
Small BM on Curie–Weiss data (App. G.1)	All state observables (exact enumeration)	Sampling-temperature tuning; double descent; warm and cold posteriors; out-of-equilibrium training	Qualitative agreement for all four effects
Small RBM on two-pattern teacher data (App. G.2)	All state observables (exact enumeration)	Sampling-temperature tuning; double descent; warm and cold posteriors; out-of-equilibrium training	Qualitative agreement for all four effects
Potts BM on Pfam and lattice-protein data (Apps. G.3, G.6)	Coupling and covariance matrices	Sampling-temperature tuning; out-of-equilibrium training	Qualitative spectral signatures
FVSBN on Curie–Weiss and Ising data (App. G.4.1)	Reverse KL; sample moments	Double descent	Qualitative agreement for double descent
RBM on financial returns (App. G.4.2)	Estimated NLL; weight matrix	Double descent	Qualitative spectral correspondence
BayesGAN on a 2D Gaussian target (App. G.5)	Estimated KL divergences	Warm and cold posteriors	Qualitative agreement with posterior-temperature dependence

G.1 Experiments in small tractable BMs

We numerically study fully connected zero-field binary Boltzmann machines in a controlled teacher–student setting. Configurations are vertices of the N-cube, $s \in \{-1,+1\}^N$, and the teacher is a rank-one Curie–Weiss model,

$$P^\star(s) = \frac{1}{Z^\star} \exp\left[\frac{\omega^\star}{2N}\left(\sum_{i=1}^N s_i\right)^2\right]. \tag{212}$$

The parameter $\omega^\star$ controls the strength of the teacher interaction. Equivalently, up to an additive constant (independent of s), the teacher coupling matrix has entries $J^\star_{ij} = \frac{\omega^\star}{N}$ $(i \neq j)$ and $J^\star_{ii} = 0$, and $Z^\star (= Z(J^\star))$. Its finite-N leading eigenvalue is therefore $\omega^\star(1 - 1/N)$.

The student is a fully connected binary Boltzmann machine,

$$P_J(s) = \frac{1}{Z(J)} \exp\left[\sum_{i<j} J_{ij} s_i s_j\right]. \tag{213}$$

The coupling matrix J is symmetric, with $J_{ii} = 0$, and its $N(N-1)/2$ off-diagonal entries are the free parameters. There are no external fields, and the couplings are subject to quadratic regularisation of intensity γ, as in Eq. (5).

Training datasets consist of K independent configurations sampled from the teacher.

Because the systems are small, all 2^N configurations are enumerated. For any fixed coupling matrix J, partition functions, equilibrium negative phases, state probabilities, forward and reverse Kullback–Leibler divergences, covariance matrices, and teacher-aligned observables are then evaluated exactly in configuration space. Integration over couplings remains numerical. For posterior tempering, we use four independent Metropolis-adjusted Hamiltonian Monte Carlo chains per dataset and parameter pair, retaining 4000 draws per chain after 1000 warmup steps, extended to 16000 draws at $\gamma = 0.1$ and $\eta \in \{0.05, 0.075, 0.1, 0.15\}$. Across the tested grid, the maximum rank-normalised split-$\widehat{R}$ is below 1.005 and the minimum bulk effective sample size exceeds 2300. For the $K = 1$ regularisation sweep, we average exactly over teacher-generated datasets and spin-sign gauges, using independent Gaussian-prior gauge-orbit draws at $\eta = 1$ and importance sampling at the other temperatures, supplemented by adjusted Hamiltonian Monte Carlo and Laplace importance proposals in the cold regime. State enumeration is exact to numerical precision; posterior averages retain Monte Carlo uncertainty.

The common performance measures are

$$D_{\mathrm{KL}}(P^\star \| Q) \qquad \text{and} \qquad D_{\mathrm{KL}}(Q \| P^\star), \tag{214}$$

where Q is either an individual trained model or a posterior-predictive distribution. For posterior prediction, we average P_J over the coupling posterior before evaluating the KL. The typical reverse KL in the regularisation sweep instead averages $D_{\mathrm{KL}}(P_J \| P^\star)$ over posterior draws and teacher-generated datasets.

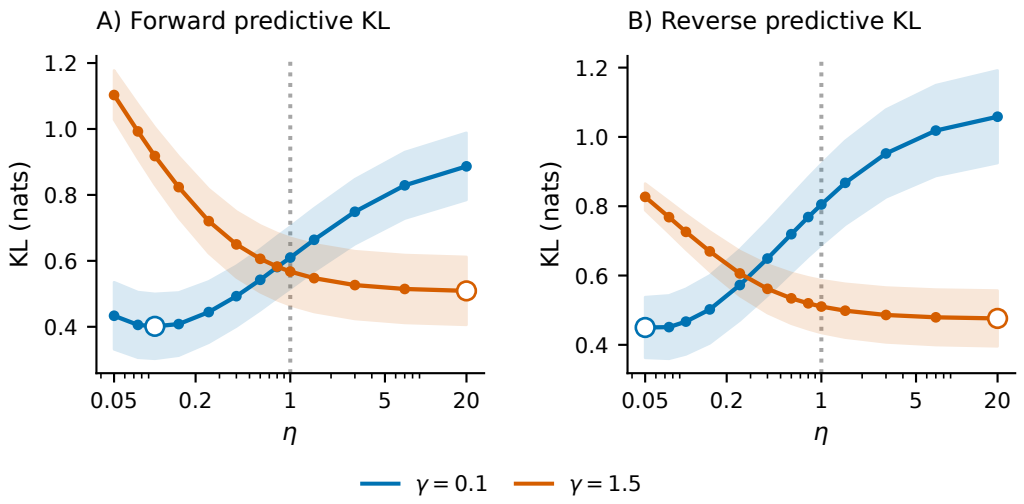


Figure 15: **Posterior tempering in a zero-field binary BM student with a Curie–Weiss teacher.** The forward **A)** and reverse **B)** KL losses of the posterior predictive are shown versus posterior inverse temperature η, for weak ($\gamma = 0.1$, blue) and strong ($\gamma = 1.5$, orange) regularisation. Curves are means over 100 independently sampled datasets; shading gives pointwise 99% confidence intervals over datasets. The vertical dotted line marks the ordinary Bayes posterior, $\eta = 1$, and open circles mark minima on the tested grid. Under weak regularisation, the forward and reverse minima occur at $\eta = 0.1$ and 0.05, respectively; under strong regularisation, both occur at the largest tested value, $\eta = 20$. Shared parameters: $N = 8$, $K = 4$, $\omega^\star = 1.5$; posterior convention as in Eq. (5).

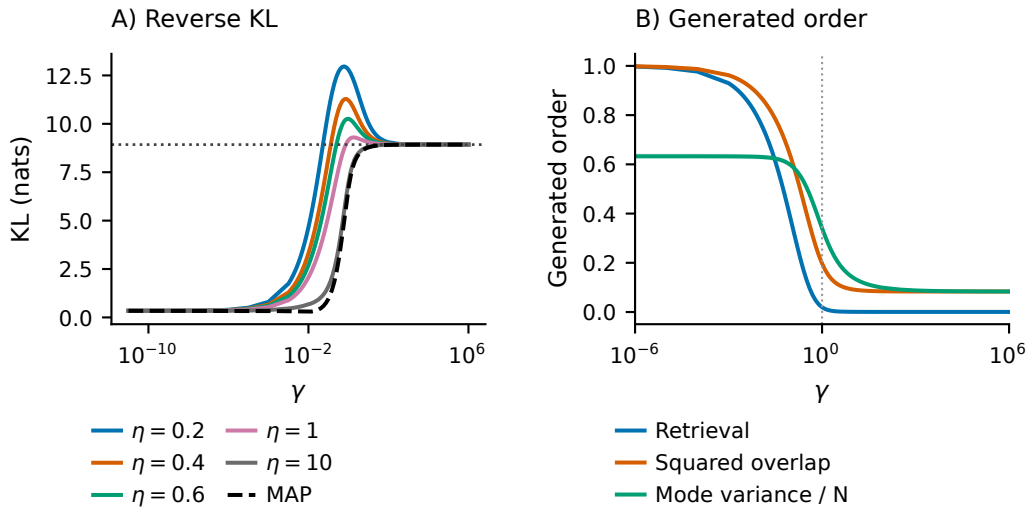


Figure 16: **Double descent with regularization strength in a zero-field binary BM student with a Curie–Weiss teacher. A)** Posterior-average reverse KL versus regularisation γ for various inverse temperatures η. For $\eta = 0.2, 0.4, 0.6, 1$, the loss rises from its weak regularisation minimum, overshoots the uniform student limit 8.9265 nats (horizontal dotted line), and then decreases towards that limit. Cooling reduces the overshoot, which is absent at $\eta = 10$. The dashed black curve gives the exact MAP loss, obtained from the strictly convex objective. It has a single minimum, 0.2846 nats at $\gamma \simeq 0.006604$, and then increases towards the uniform limit from below, without a second descent. **B)** Retrieval probability $P_J(\pm\xi)$, squared overlap with the observed sample $\langle (s\cdot\xi/N)^2 \rangle$, and variance along the leading coupling eigenvector divided by N, at $\eta = 1$. Their decline accompanies the loss of the data-aligned mode; the vertical dotted line marks $\gamma = 1$ as a reference in panel **B)**. Averages over all possible one-example datasets are exact. Shading gives approximate 95% Monte Carlo intervals for the coupling integrals. Common parameters: $N = 12$, $K = 1$, $\omega^\star = 3$, with Eq. (5) defining γ and η.

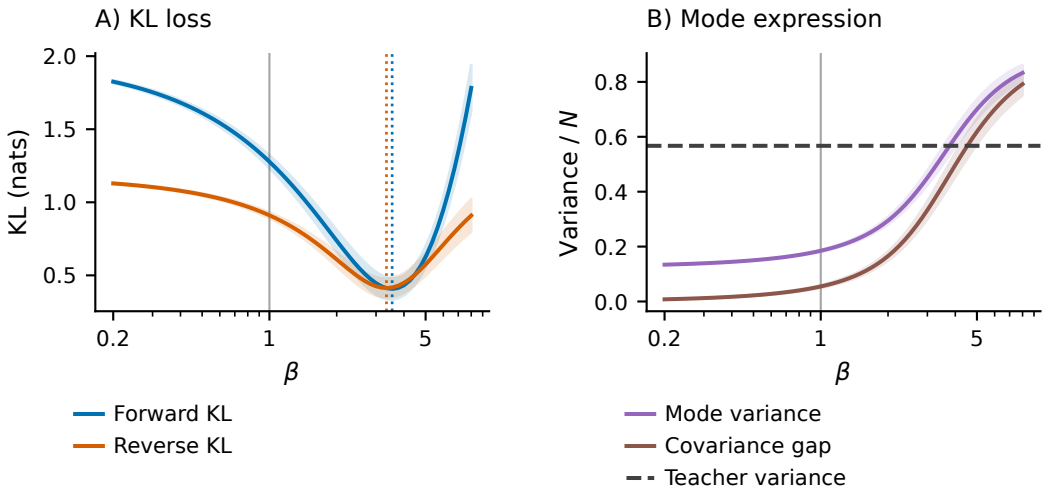


Figure 17: **Generative temperature tuning in a zero-field binary BM student with a Curie–Weiss teacher.** We hold the exact MAP couplings fixed while varying the generation inverse temperature β. Curves average 200 independently sampled datasets; shading gives pointwise 99% confidence intervals. **A)** Forward (blue) and reverse (orange) KL losses have cold optima at $\beta_{\text{opt}}^{\text{fwd}} \simeq 3.54$ and $\beta_{\text{opt}}^{\text{rev}} \simeq 3.34$, marked by the corresponding dotted lines. **B)** Generated variance along $v_1(\widehat{J})$ (purple), generated covariance eigenvalue gap (brown), and teacher variance along $v_1(\widehat{J})$ (horizontal dashed), all divided by N. The learned direction is underexpressed at $\beta = 1$; cooling restores its variance towards the teacher value and improves generation. Grey vertical lines mark $\beta = 1$. Common parameters: $N = 8$, $K = 4$, $\omega^\star = 1.5$, $\gamma = 6$ in Eq. (5).

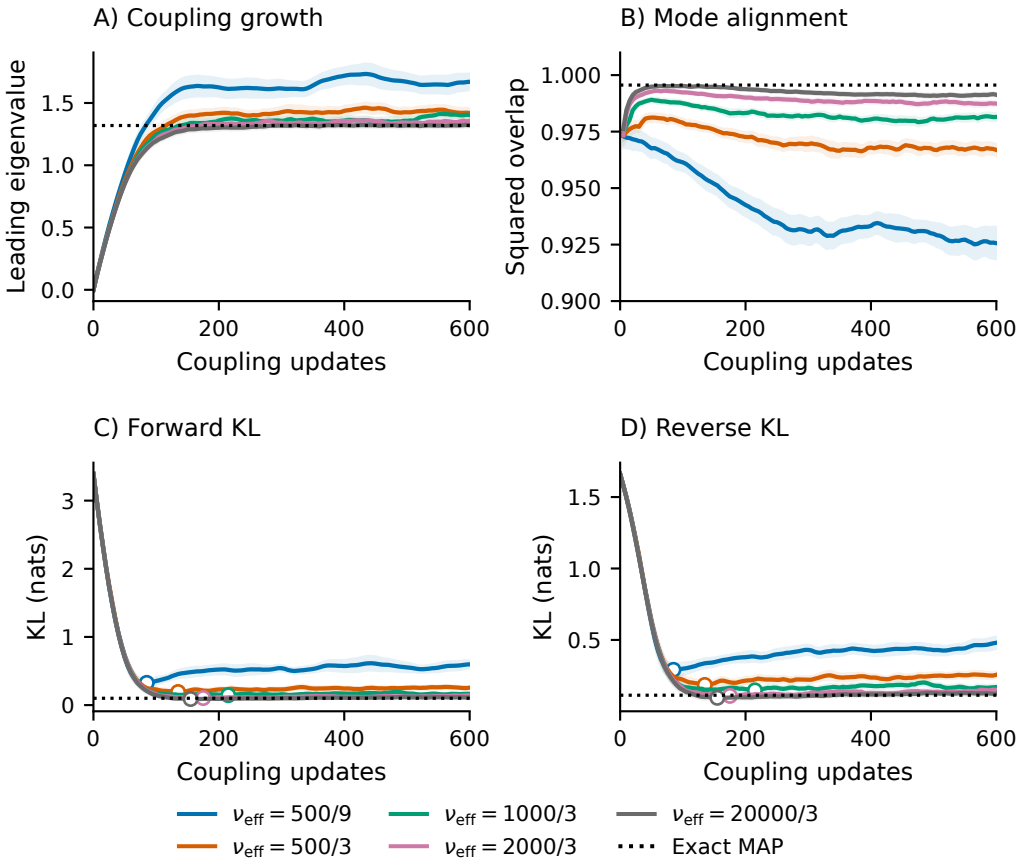


Figure 18: **Out-of-equilibrium learning in a zero-field binary BM student with a Curie–Weiss teacher.** Persistent contrastive-divergence training uses effective sampling rates $\nu_{\text{eff}} \in \{500/9, 500/3, 1000/3, 2000/3, 20000/3\}$, with one persistent particle per replica. We define $\nu_{\text{eff}} = m/(N\Delta t)$ for $m = 1, 3, 6, 12, 120$ random single-spin Gibbs updates per coupling update. In the time convention of Eq. (3), the learning rate $\epsilon = 0.003$ gives $\Delta t = 2N\epsilon/K = 0.0015$. Thus ν_{eff} is the Gibbs refresh rate per spin per unit learning time. Couplings start at zero and particles start uniformly; random-scan heat-bath updates precede each coupling update. Curves average 50 independent datasets after averaging three training replicas per dataset; shading gives pointwise 95% dataset confidence intervals. Horizontal dotted lines denote exact MAP values. **A)** Slow sampling produces an overshoot of the leading coupling eigenvalue, reaching 1.670 at step 600, compared with 1.322 for the fastest sampler and 1.319 at MAP. **B)** Squared overlap of the leading learned eigenvector with the principal empirical direction, showing reduced late-time alignment under slower sampling. **C,D)** Exact forward and reverse KL losses; open circles mark minima on the five-update checkpoint grid. Faster sampling reduces late-time degradation, while the precise optimum times need not be ordered by sampling rate. Common parameters: $N = 12$, $K = 48$, $\omega^\star = 1.5$, and $\gamma = 3.2$ in Eq. (5). The learning rate 0.003 multiplies the gradient of the negative log-posterior per training example.

G.2 Experiments in small tractable RBMs

We use a zero-field binary RBM with visible spins $\mathbf{v} \in \{\pm 1\}^N$, hidden spins $\mathbf{h} \in \{\pm 1\}^M$, and joint distribution

$$P_\beta(\mathbf{v}, \mathbf{h} \mid W) = \frac{\exp(\beta\, \mathbf{v}^\top W \mathbf{h})}{Z_\beta(W)}, \qquad P_\beta(\mathbf{v} \mid W) = \sum_{\mathbf{h}} P_\beta(\mathbf{v}, \mathbf{h} \mid W), \tag{215}$$

where $W \in \mathbb{R}^{N \times M}$, $Z_\beta(W)$ is the partition function, and β is the generation inverse temperature. The student has $N = 16$ visible and $M = 3$ hidden spins. Each training set contains $K = 16$ independent visible configurations sampled from a teacher with two hidden spins and weights $W^\star = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2/2)$, where the $\boldsymbol{\xi}_r$ are orthogonal ± 1 patterns. Both teacher sampling and training use $\beta = 1$. For training data $\mathcal{D}$, the posterior is

$$P_{\eta,\gamma}(W \mid \mathcal{D}) \propto \exp\left\{ \eta \left[\sum_{\mathbf{v} \in \mathcal{D}} \log P_1(\mathbf{v} \mid W) - \frac{\gamma}{2} \|W\|_F^2 \right] \right\}. \tag{216}$$

Here γ sets the weight penalty and η is the posterior inverse temperature. We sum over spin configurations exactly, use exact-gradient optimization for MAP fits, and use replica-exchange HMC for averages over weights.

We observe warm and cold posterior effects: warming improves prediction under weak regularization, while cooling helps under stronger regularization (Fig. 19). Increasing regularization produces double descent, with reverse KL rising to an intermediate maximum before falling again (Fig. 20). Near the maximum, the generated samples retain strong correlations while expressing the teacher pattern less strongly. Generation-temperature tuning also improves prediction by warming or cooling the distribution at fixed fitted weights (Fig. 21). Some posterior-predictive estimates retain unresolved Monte Carlo uncertainty beyond the dataset error bars.

To study out-of-equilibrium learning, we estimate the model term in the training gradient using one persistent visible configuration. We vary how often a Gibbs sweep refreshes it; its contribution is evaluated at the current weights at every update. Slow refresh amplifies the leading mode of $WW^\top$. Its eigenvector initially aligns with the principal data direction, then loses alignment as generation deteriorates; early stopping therefore helps (Fig. 22). We compare with exact gradients and independent equilibrium sampling. Infrequent exact-equilibrium redraws reproduce much of the degradation, consistent with the persistent configuration lagging behind the changing model.

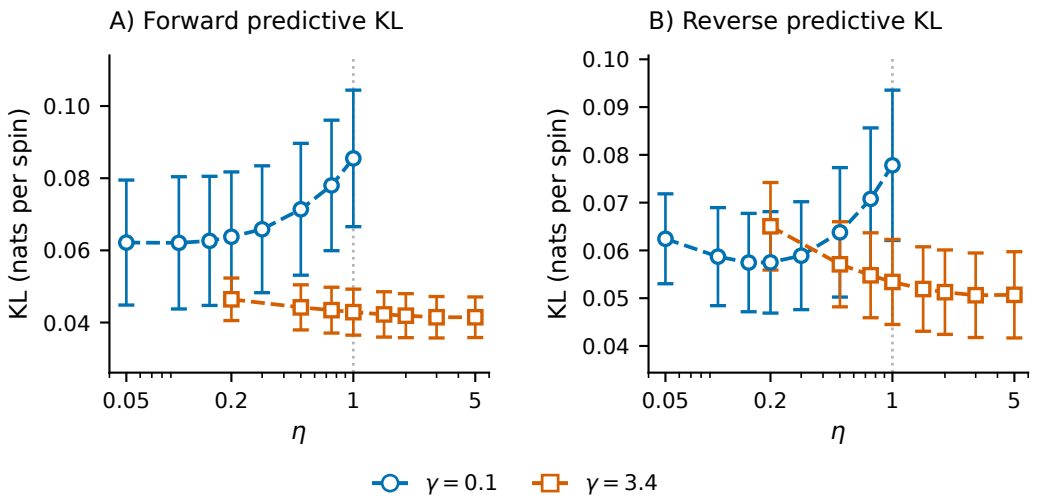


Figure 19: **Posterior tempering in the small RBM. A,B)** Forward and reverse KL of the posterior-predictive mixture, in nats per visible spin, at weak ($\gamma = .1$, blue) and stronger ($\gamma = 3.4$, orange) regularization. Each curve contains eight tested temperatures, spanning $.05 \leq \eta \leq 1$ and $.2 \leq \eta \leq 5$, respectively. Points average over eight independent training sets from the teacher defined in the text; bars give 95% dataset confidence intervals, and dashed segments guide the eye. The vertical dotted line marks the ordinary posterior, $\eta = 1$. Warming ($\eta < 1$) improves prediction under weak regularization, while cooling ($\eta > 1$) improves it under stronger regularization. Some posterior-predictive estimates retain unresolved Monte Carlo precision limitations; this uncertainty is not included in the dataset intervals. Common parameters: $N = 16$, $M = 3$, $K = 16$.

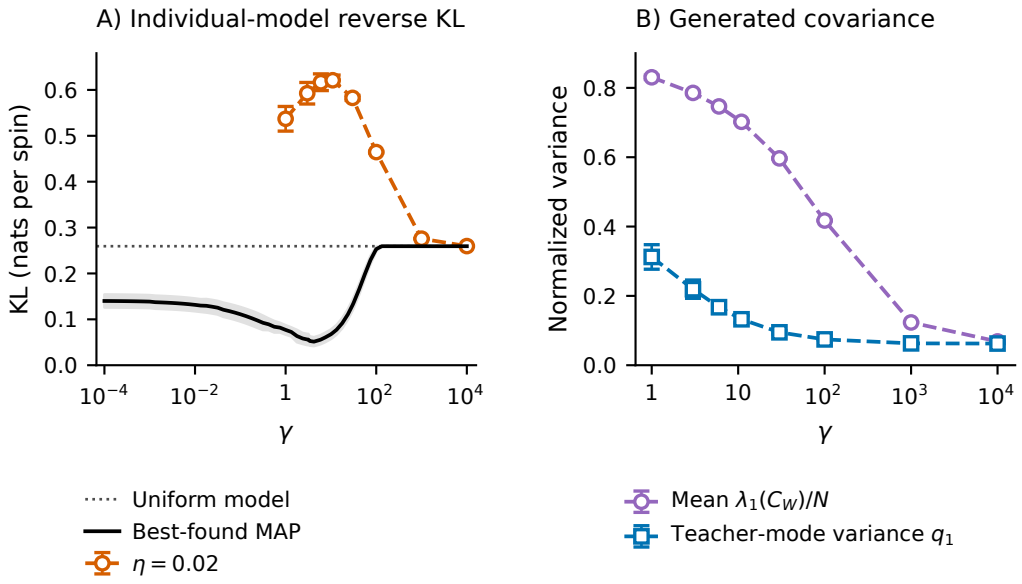


Figure 20: **Double descent and generated covariance in the small RBM.** Posterior averages use eight regularization strengths from 1 to 10^4 at $\eta = .02$; the MAP comparison uses 70 values from 10^{-4} to 10^4. Both average over the same eight independently drawn training sets, each containing 16 visible configurations from the teacher defined at the start of App. G.2. **A)** Mean individual-model reverse KL per spin. Orange markers show posterior averages; the solid black curve shows best-found MAP fits selected by the training objective from multiple exact-gradient initializations and bidirectional continuation. Their sampled mean minimum is 0.05134 nats per spin at $\gamma \simeq 4.217$; the subsequent high-regularization peak and second descent are absent. The horizontal dotted line is the uniform-model limit. **B)** At $\eta = .02$, mean leading eigenvalue of the individual-model covariance $C_W = \mathbb{E}[\mathbf{v}\mathbf{v}^\top \mid W]$, divided by N (purple), and mean variance along the dominant teacher pattern, $\mathbb{E}_W[\boldsymbol{\xi}_1^\top C_W \boldsymbol{\xi}_1]/N^2$ (blue). Error bars and grey shading give pointwise 95% dataset intervals; dashed segments guide the eye between posterior estimates. The posterior reverse-KL rise accompanies weakened teacher expression while individual models remain anisotropic. Common parameters: $N = 16$, $M = 3$, $K = 16$.

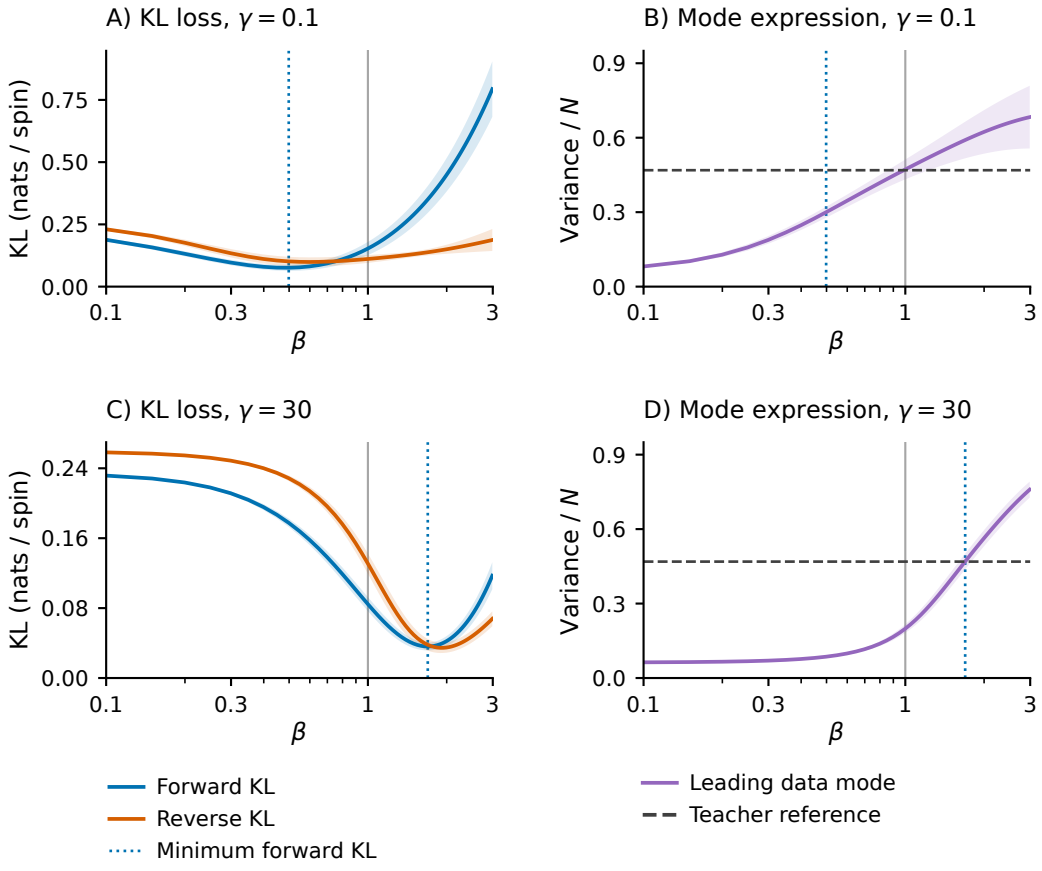


Figure 21: **Generative temperature tuning in the small RBM.** We vary the joint-energy inverse temperature β at fixed fitted weights, with $\gamma = .1$ (top) and $\gamma = 30$ (bottom). **A,C)** Forward (blue) and reverse (orange) KL per spin; dotted lines mark the minima of the mean forward-KL curves. **B,D)** Generated variance along the leading direction of the training-data second-moment matrix, divided by N (purple); horizontal dashed lines give the teacher variance along that direction. Grey vertical lines mark $\beta = 1$. Curves average over eight independent training sets; shading gives pointwise 95% dataset intervals. Warming suppresses mode expression and cooling amplifies it; KL improvement need not reduce the mode-variance mismatch. The minima are descriptive selections on these datasets. Common parameters: $N = 16$, $M = 3$, $K = 16$.

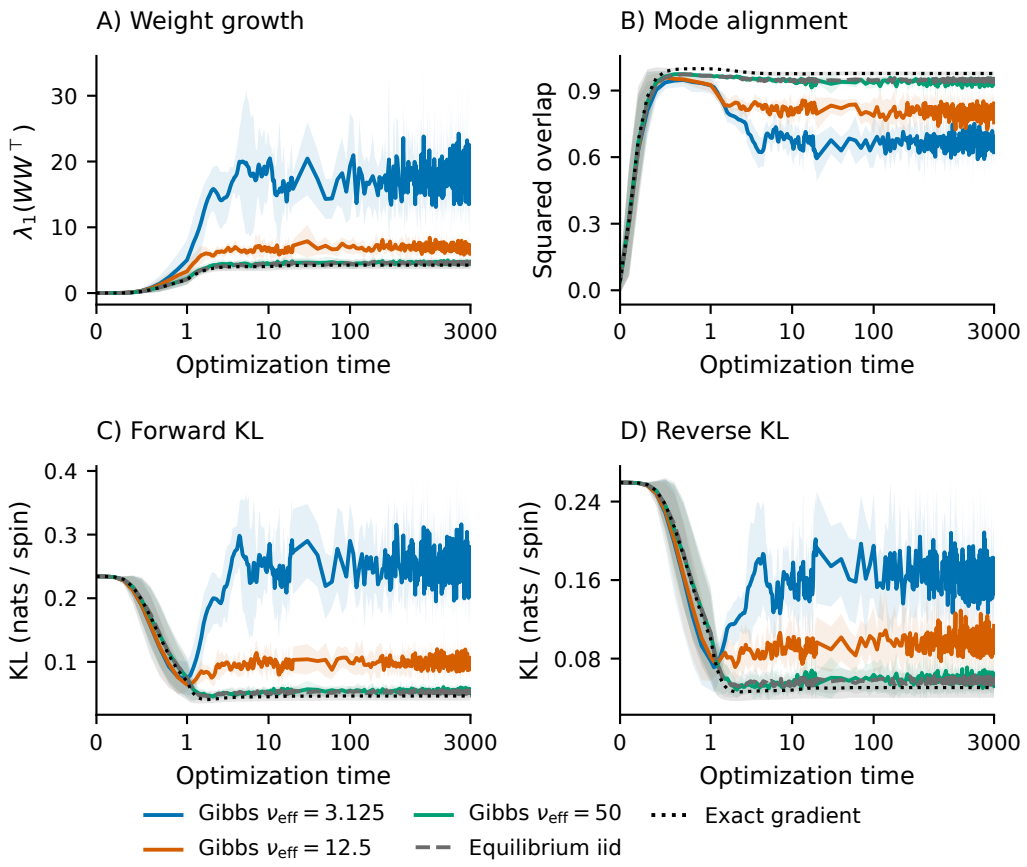


Figure 22: **Out-of-equilibrium learning in the small RBM.** One persistent particle receives a full blocked Gibbs sweep with probability p per weight update. We measure training time as $t = n\Delta t$ after n updates, with learning rate $\Delta t = .02$. The effective sampling rate $\nu_{\text{eff}} = p/\Delta t$ counts expected sweeps per unit training time. The dashed grey curve uses an independent exact-equilibrium sample at every weight update; the dotted black curve uses exact gradients. **A)** Leading eigenvalue of $WW^\top$. **B)** Squared overlap of its leading eigenvector with the principal direction of the training-data second-moment matrix. Under slower sampling, alignment first rises and then declines. **C,D)** Exact forward and reverse KL per spin. Curves average over eight additional independent training sets after averaging three stochastic repeats within each set; shading gives pointwise 95% dataset intervals. The time axis is linear up to $t = 1$ and logarithmic thereafter. Slower sampling amplifies weight modes and worsens late-time generation, favoring early stopping. Common parameters: $N = 16$, $M = 3$, $K = 16$, $\gamma = 3.4$.

G.3 Sampling temperature tuning in Potts BMs trained on protein sequences

A prediction of the SBM phase diagram is that L_2 regularization shrinks the coupling eigenvalues and pushes the trained model toward the $h = 0$ phase, where generated samples are near-isotropic noise. Raising the sampling inverse temperature $\beta > 1$ amplifies signal modes and can trigger an $h = 0$ to $h \neq 0$ condensation that rescues these modes, giving a principled mechanism for the widely-used practice of sampling energy-based models at temperatures below the training value [Fields et al., 2025]. We now show that this picture quantitatively describes pairwise Potts Boltzmann machines trained on real protein sequence data.

Expanding the sampling covariance of a model trained at $\beta = 1$ with regularization γ around the training point and choosing β to optimize the fit of the model covariance to the empirical data covariance, to leading order in γ, yields the closed-form optimum

$$\beta_{\mathrm{TT}}^{\mathrm{cov}} = 1 + \gamma \, \frac{\sum_k \lambda_k^2 \, c_k^2}{\sum_k \lambda_k (\lambda_k - \mu') \, c_k^4}, \tag{217}$$

where λ_k and c_k are, respectively, the coupling eigenvalues of the trained model and the data covariance eigenvalues in a common eigenbasis, and μ' encodes the on-site restoring force set by the normalization constraint: $\mu'_{\mathrm{SBM}} = \sum_k \lambda_k c_k^2 / \sum_k c_k^2$ for the spherical constraint, and $\mu' = 0$ for the per-site softmax of the Potts model. We use covariance matching rather than minimizing a KL divergence because any KL between the model and the data distribution requires the Potts partition function $Z(\beta J)$ and is intractable in this regime, while the model and data covariances are directly accessible from samples.

Derivation. Work in the joint eigenbasis of the trained coupling J and the data covariance C, with eigenvalues λ_k and c_k, and let $\sigma_k^2(\beta)$ denote the sampling-covariance eigenvalue at inverse temperature β. The L_2-regularized stationarity condition $\langle \mathbf{x}\mathbf{x}^\top \rangle_J = C - \gamma J$ at $\beta = 1$ reads in this basis

$$\sigma_k^2(1) = c_k - \gamma \lambda_k + \mathcal{O}(\gamma^2). \tag{218}$$

Sampling at β rescales $J \mapsto \beta J$. For the spherical model $P_\beta(\mathbf{x}) \propto e^{(\beta/2)\mathbf{x}^\top J \mathbf{x}} \, \delta(\|\mathbf{x}\|^2 - N)$ the Lagrange multiplier μ_β gives $\sigma_k^2(\beta) = (\mu_\beta - \beta\lambda_k)^{-1}$, hence

$$\partial_\beta \sigma_k^2 \big|_{\beta=1} = (\lambda_k - \mu') \, \sigma_k^2(1)^2 = (\lambda_k - \mu') \, c_k^2 + \mathcal{O}(\gamma), \tag{219}$$

with $\mu' \equiv \partial_\beta \mu_\beta |_{\beta=1}$. Differentiating the trace constraint $\sum_k \sigma_k^2(\beta) = N$ at $\beta = 1$ fixes

$$\mu'_{\mathrm{SBM}} = \frac{\sum_k \lambda_k c_k^2}{\sum_k c_k^2} + \mathcal{O}(\gamma); \tag{220}$$

the Potts softmax enforces normalization site-wise, leaving $\mu' = 0$. Expanding $\sigma_k^2(\beta)$ to first order around $\beta = 1$ via (218)–(219) gives the residual

$$\sigma_k^2(\beta) - c_k = -\gamma \lambda_k + (\beta - 1)(\lambda_k - \mu') \, c_k^2 + \mathcal{O}(\gamma^2). \tag{221}$$

A scalar β cannot zero this mode-by-mode; we project onto the coupling direction weighted by the data covariance,

$$\sum_k \lambda_k \, c_k^2 \big(\sigma_k^2(\beta_{\mathrm{TT}}^{\mathrm{cov}}) - c_k \big) = 0. \tag{222}$$

Substituting (221) into (222) yields

$$-\gamma \sum_k \lambda_k^2 c_k^2 + (\beta_{\mathrm{TT}}^{\mathrm{cov}} - 1) \sum_k \lambda_k (\lambda_k - \mu') c_k^4 = 0, \tag{223}$$

which solves to Eq. (217).

G.3.1 Response-regulator PF00072

We trained pairwise Potts models on the Pfam response-regulator family PF00072 ($L = 111$, $q = 21$) via adabmDCA [Rosset et al., 2026] at sixteen regularizations $\gamma \in [0.01, 1.0]$. For each trained model we sampled sequences at thirteen inverse temperatures using parallel tempering with 10^4 chains and recorded (i) the Pearson correlation between model and data connected correlations, whose

maximum defines the empirical optimum $\beta_{\rm opt}(\gamma)$, and (ii) the covariance eigenvalue spectrum $\sigma_k^2(\beta)$ of generated samples. Figure 23 summarizes the result. Panel A shows the Pearson correlation as a function of β for each regularization; the red dots mark the maximum of each curve and define $\beta_{\rm opt}(\gamma)$. Panel B tracks the top six sampling covariance eigenvalues at $\gamma = 1$: as β increases, the dominant modes detach from the near-degenerate bulk and grow sharply near $\beta \simeq \beta_{\rm opt} = 1.5$, the direct finite-$N$ signature of the SBM condensation. Past $\beta_{\rm opt}$, however, the top eigenvalue itself declines: as $\beta \to \infty$ the Gibbs measure concentrates on the ground state, so the covariance of generated samples collapses in *every* direction and all σ_k^2 vanish. This is made explicit by the green dashed line in panel B (right axis), which plots the bulk sum $\sum_{k>6} \sigma_k^2(\beta)$ and decays monotonically toward zero with β. The total trace $\sum_k \sigma_k^2$ is therefore *not* conserved along β, unlike the idealized spherical model whose hard constraint $|\mathbf{x}|^2 = N$ fixes it exactly. As a result, $\beta_{\rm opt}$ emerges from the competition between the condensation amplification of signal modes (which grows σ_1^2) and the freezing-driven overall shrinkage (which eventually overwhelms it), so that sampling at β well past $\beta_{\rm opt}$ degrades the generative model. Panel C shows $\beta_{\rm opt}(\gamma)$ rising monotonically from 1 as γ increases and tracking the SBM prediction (217) across the full regularization range; the naive $1/\lambda_1$ criterion (not shown) predicts values in the range 0.15–0.52, wrong by factors of 3–7. Residual deviations at large γ reflect higher-order corrections in γ and a per-site softmax rather than global spherical constraint in the Potts model, but do not affect the qualitative agreement. The same picture holds on a second family, the SH3 domain PF00018, with a more pronounced temperature-tuning effect owing to its smaller effective sample size; we show the corresponding data in App. G.3.2 below.

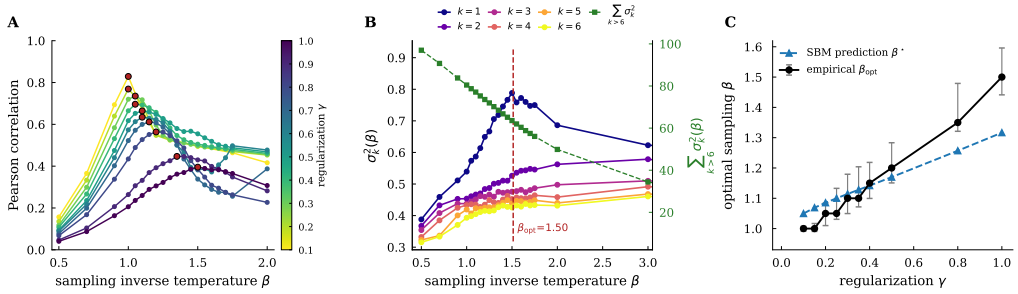


Figure 23: **Temperature tuning on Potts BM trained on PF00072.** Pairwise Potts models on the Pfam response-regulator family PF00072 ($L = 111$, $q = 21$), trained via adabmDCA [Rosset et al., 2026]. **A)** Pearson correlation between generated and data connected correlations as a function of sampling inverse temperature β, for sixteen L_2 regularizations γ (color bar). Red dots mark the peak of each curve, defining $\beta_{\rm opt}(\gamma)$. **B)** Top six covariance eigenvalues $\sigma_k^2(\beta)$ of sequences generated by the model trained at $\gamma = 1$ (left axis, solid lines). The dominant modes amplify as β increases, with the empirical optimum $\beta_{\rm opt} = 1.50$ marked by the red dashed line. The green dashed line with square markers (right axis) shows the sum of the remaining eigenvalues, $\sum_{k>6} \sigma_k^2(\beta)$, which decreases monotonically with β and approaches zero as $\beta \to \infty$. **C)** Empirical $\beta_{\rm opt}(\gamma)$ (black circles, from panel A) compared against the SBM prediction $\beta_{\rm TT}^{\rm cov}$ of (217) (blue triangles, dashed). For each γ, $\beta_{\rm opt}$ is extracted from the Pearson curve on the combined original + refined β grid; the error bar shows the range of β over which the Pearson correlation stays within one standard error of its peak (standard error estimated from an independent replicate scan with a different random seed, yielding $\sigma_P \simeq 5 \times 10^{-3}$).

G.3.2 SH3 domain PF00018

The SH3 domain PF00018 ($L = 48$, $q = 21$) sits deeper in the undersampled regime than PF00072; protocols are identical to App. G.3. Fig. 24 shows the temperature-tuning effect is correspondingly stronger, with $\beta_{\rm opt} \simeq 1.75$ already at $\gamma = 0.8$; Eq. (217) again tracks the empirical optimum across γ, with the same large-γ bias attributable to finite-order expansion and the softmax constraint.

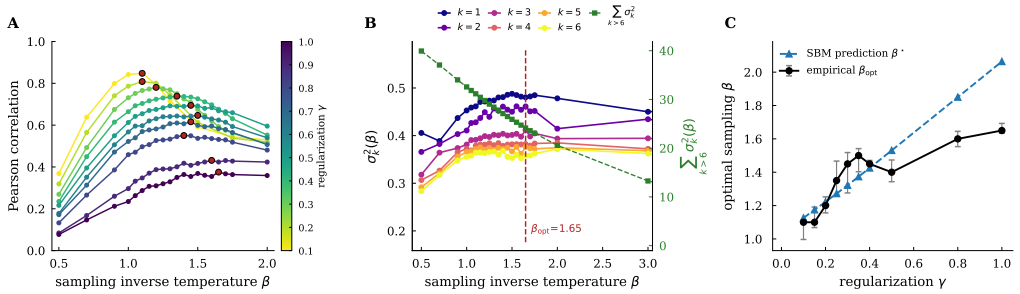


Figure 24: **Temperature tuning on Potts BM trained on PF00018.** Pairwise Potts models on the Pfam SH3 domain PF00018 ($L = 48$, $q = 21$), trained via adabmDCA. **A)** Pearson correlation between generated and data connected correlations as a function of sampling inverse temperature β, for sixteen L_2 regularizations (color bar); red dots mark β_{opt}. **B)** Top six covariance eigenvalues $\sigma_k^2(\beta)$ at $\gamma = 1$ (left axis, solid lines); green dashed line with square markers (right axis) shows the bulk sum $\sum_{k>6}\sigma_k^2(\beta)$. **C)** Empirical $\beta_{\text{opt}}(\gamma)$ compared against the SBM prediction $\beta_{\text{TT}}^{\text{cov}}$ of (217); error bars show the Pearson-plateau half-width (see Fig. 23 caption). Conventions as in Fig. 23.

G.4 Double descent

G.4.1 Double descent in a binary FVSBN, large-K behavior and the rank-1-data condition

We test double descent in a discrete student trained on two teachers whose covariances are dominated by a single mode. We vary the number of training samples from $K \ll N$ to $K = 2N$. This tests the regime-(b) extension in Sec. 2.1: the equilibrium theory remains valid as $K \to \infty$ provided the empirical covariance C retains $\mathcal{O}(1)$ outliers above its bulk.

The student is a binary fully-visible sigmoid belief network (FVSBN) [Frey, 1998] on $\{\pm 1\}^N$, with $N = 16$ and autoregressive log-likelihood

$$\ln q_\theta(s) = \sum_i \ln \sigma\big((2s_i)\big(b_i + \sum_{j<i} W_{ij} s_j\big)\big), \tag{224}$$

where σ is the logistic function. Its parameters are the strictly lower-triangular matrix $W \in \mathbb{R}^{N\times N}$ and biases $b \in \mathbb{R}^N$. We first maximize the likelihood with 1500 Adam steps. We then collect SWAG snapshots every 15 steps during 1500 SGD-Langevin updates at constant learning rate and $T_{\text{lang}} = 0.3$. We fit these snapshots with a diagonal Gaussian, draw 30 students, and average their reverse KL against the teacher. Setting the optimizer's weight decay to $\text{wd} = N\gamma/(2K)$ makes the loss averaged over training samples match the SBM prior $\propto \exp(-N\gamma\,\text{tr}\,W^2/4)$, so γ remains comparable across sample sizes.

The two teachers have different interaction structures. The first is a rank-1 Curie–Weiss / Hopfield model with $\mathcal{E}(s) = -\frac{1}{2N}(\xi \cdot s)^2$, $\beta = 2.0$, and a random $\xi \in \{\pm 1\}^N$ (Fig. 25A). The second is a 2D Ising model on the $L = 4$ periodic square lattice ($N = L^2 = 16$) at $\beta = 0.5$, deep in the ferromagnetic phase (Fig. 25B). We compute $\ln Z$ exactly by exhaustive enumeration and a column transfer matrix, respectively; for the first teacher, we also evaluate the reverse KL exactly. We use $K \in \{4, 8, 16, 32\}$ with 10 seeds each and record the spectrum of $C = \frac{1}{N}\sum_k x^k x^{k\top}$. As predicted, $\lambda_1(C)$ grows linearly with K, while $\lambda_2(C)$ grows much more slowly. From $K = 4$ to 32, the ratio λ_1/λ_2 rises from 9.7 to 27.3 for CW and from 9.5 to 18.8 for Ising: the empirical covariance remains dominated by its leading mode.

Figure 25 shows the SWAG-averaged reverse KL per spin. For both teachers, the double-descent bump persists across the full K range, rising above the $\gamma = 0$ baseline by 0.14–0.18 nats/N (CW) and 0.22–0.27 nats/N (Ising). Its height depends little on K, but its position γ_{peak} moves from $\gamma \approx 0.30$ at $K \le 8$ to $\gamma \approx 1$ at $K \ge 16$, tracking the retarded-learning threshold $\gamma \sim \eta c_1^2$ with $c_1 = \lambda_1(C) \propto K$. Thus the same qualitative signature appears for the rank-1 Hopfield teacher without spatial structure and the Ising teacher with nearest-neighbor lattice interactions.

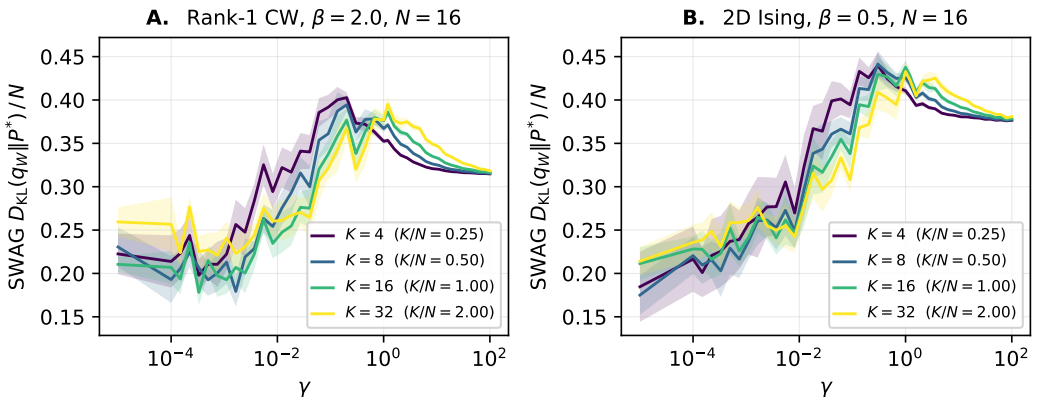


Figure 25: **Double descent in a binary FVSBN persists across K on rank-1-dominated teachers.** SWAG posterior-averaged reverse KL per spin, $D_{\mathrm{KL}}(q_W \| P^*)/N$, vs. the paper-convention regularization γ, for the binary FVSBN trained on **A)** a rank-1 Curie–Weiss / Hopfield teacher at $\beta = 2.0$, $N = 16$, and **B)** a 2D Ising teacher on the $L = 4$ periodic square lattice ($N = 16$) at $\beta = 0.5$, deep in the ferromagnetic phase. Each panel overlays four sample sizes $K \in \{4, 8, 16, 32\}$, i.e. $K/N \in \{0.25, 0.5, 1, 2\}$, 10 seeds each; shaded bands are ± 1 SEM. The optimizer weight decay is $\mathrm{wd} = N\gamma/(2K)$, matching the SBM prior $\propto \exp(-N\gamma \operatorname{tr} W^2/4)$ at every K. Across both teachers the SWAG-DD bump above the $\gamma = 0$ baseline persists at every K ($+0.14$ to $+0.18$ nats/N on CW, $+0.22$ to $+0.27$ on Ising), while the empirical-covariance gap $\lambda_1(C)/\lambda_2(C)$ grows monotonically with K (from 9.7 to 27.3 on CW and from 9.5 to 18.8 on Ising). The peak position γ_{peak} shifts with K in the direction predicted by the retarded-learning threshold $\gamma \sim \eta c_1^2$ with $c_1 \propto K$. The replacement of the geometry-free rank-1 Hopfield teacher by the 2D Ising NN lattice leaves the signature qualitatively unchanged, confirming that the relevant condition is a low-effective-rank empirical covariance and not the spatial structure of the teacher.

G.4.2 Double descent in an RBM trained on financial data

The double-descent phenomenon predicted by the theory is not limited to the SBM. A qualitatively similar behavior as a function of regularization strength also appears in a restricted Boltzmann machine (RBM) trained on financial data, where the peak in test loss coincides with a spectral transition in the learned weight matrix, qualitatively consistent with the spectral picture underlying double descent in the SBM.

Data. We use the Kenneth R. French 49 industry portfolio daily value-weighted returns [Fama and French, 1997, Laloux et al., 1999], spanning July 1926 to February 2026 (14 286 trading days after removing dates with missing entries). Each industry's returns are standardized to zero mean and unit variance on the training set, and winsorized at $\pm 5\sigma$ to limit the effect of extreme events. The empirical covariance of the standardized training returns has a dominant eigenvalue $\lambda_1 \approx 25$, well separated from $\lambda_2 \approx 1.6$ ($\lambda_1/\lambda_2 \approx 15$), reflecting the single "market mode" that drives correlated daily fluctuations across sectors. The data is split chronologically into 80% training (11 428 days) and 20% test (2 858 days).

Model and training. We train a Gaussian-visible, Bernoulli-hidden RBM with $N = 49$ visible units (one per industry) and $M = 20$ hidden units, using persistent contrastive divergence (PCD) [Tieleman, 2008] with L_2 weight decay of strength γ applied to the weight matrix $W \in \mathbb{R}^{N \times M}$. The visible biases and precisions are unregularized. Training consists of 6 000 Adam [Kingma and Ba, 2015] burn-in iterations followed by 2 000 SGD iterations at constant learning rate (the latter phase used for SWAG snapshot collection [Maddox et al., 2019], though posterior averaging turns out to have a negligible effect here). The log-partition function is estimated by annealed importance sampling [Salakhutdinov and Murray, 2008] with 10 000 inverse-temperature steps and 15 chains. We sweep γ from 10^{-6} to 10 and repeat each run for 8 independent training seeds at fixed data subset; error bars show ± 1 standard deviation across seeds.

Results. Figure 26 displays the eigenvalue spectrum of the empirical training-set covariance C (panel **A**), the test negative log-likelihood (NLL) per day (panel **B**), and the two largest singular values σ_1, σ_2 of the trained weight matrix W (panel **C**), for three training-set sizes $M_{\text{train}} \in \{1000, 2000, 11428\}$.

At small γ, the test NLL reaches a low plateau ($\approx$ 63 nats/day) where the RBM captures both the market-mode correlation and sub-leading sector structure: both σ_1 and σ_2 are of comparable magnitude ($\sigma_1 \approx 3$, $\sigma_2 \approx 2.5$). As γ increases past $\sim 10^{-2}$, σ_2 drops sharply toward zero while σ_1 remains stable: L_2 regularization kills the secondary modes before the dominant one. At this transition, the test NLL rises to a local maximum (the "bump"), with amplitude $\Delta\text{NLL} \approx 0.3$ nats/day, reproducible across runs (5–6 σ). Beyond the transition ($\gamma \approx 0.3$), the NLL partially recovers as the model commits to a clean rank-one representation of the market mode (only σ_1 survives), before finally collapsing to the uncorrelated-Gaussian floor ($\approx$ 73 nats/day) when $\gamma \gtrsim 1$ kills all weights.

The non-monotonic NLL profile (plateau, bump, partial recovery, then collapse) mirrors the double-descent structure predicted by the SBM theory (Fig. 4): the left minimum corresponds to the $h \neq 0$ phase (signal captured), the bump to the spectral transition where secondary modes merge into the effective noise floor, and the right-side recovery to the regime where the model operates with a single clean outlier. As M_{train} decreases, the bump shifts to larger γ (from $\gamma \approx 3 \times 10^{-2}$ at $M_{\text{train}} = 11428$ to $\gamma \approx 10^{-1}$ at $M_{\text{train}} = 1000$), consistent with a noisier empirical covariance requiring stronger regularization to trigger the spectral transition.

The bump is visible already in the MAP (point-estimate) test NLL, without requiring Bayesian posterior averaging. The reported MAP is the endpoint of the 2 000-step constant-learning-rate SGD phase rather than a true posterior argmax, and so retains residual minibatch noise; the SWAG posterior-predictive NLL, averaging over snapshots from this same phase, follows the single-endpoint curve within error bars. The spectral mechanism underlying the double descent, L_2-driven merging of a data eigenvalue into the noise bulk, is therefore robust enough to affect even the point estimate, provided the data has a clean rank-one structure.

G.5 Tempered posterior effects

G.5.1 BayesGAN

The closed-form analysis of Fig. 5 predicts that the posterior-predictive reverse KL of an SBM student has a γ-dependent optimum $\eta^{\text{rev}}_{\text{pp}}(\gamma)$: warm ($\eta^{\text{rev}}_{\text{pp}} < 1$) at small γ and cold ($\eta^{\text{rev}}_{\text{pp}} > 1$) at large γ. We test whether the same warm-to-cold migration is visible in a deep generative model, taking the Bayesian GAN of [Saatchi and Wilson, 2017] as a convenient example. Tempering is straightforward, even though the posterior is genuinely multimodal in weight space, as the original sampler (SGHMC [Chen et al., 2014]) admits the tempered target without modification.

Setting. We replace the joint posterior of [Saatchi and Wilson, 2017] by its tempered counterpart $p_\eta(\theta) \propto [\mathcal{L}(\theta)\, p(\theta)]^\eta$, with η=1 the Saatchi–Wilson Bayesian setting and $\eta \to \infty$ the classical maximum-likelihood GAN. The SGHMC update of [Chen et al., 2014] with gradient term scaled by η targets p^η at stationarity. The generator $G(z; \theta_g)$ and discriminator $D(x; \theta_d)$ are two-layer ReLU MLPs (hidden 32×32, $z \in \mathbb{R}^4$). Algorithm 1 of [Saatchi and Wilson, 2017] is run with J_g=4 generator chains and J_d=1 discriminator chain, M=2 inner steps, friction α=0.1, step size $\varepsilon = 5 \times 10^{-5}$, an isotropic Gaussian prior $\mathcal{N}(0, \sigma_p^2 I)$ on every weight, and a 500-iteration Adam warm-up before SGHMC sampling, as recommended in [Saatchi and Wilson, 2017]. The predictive density P_{pp} is estimated by isotropic Gaussian KDE (σ_{KDE}=0.1) on samples pooled across all generator chains and post-burn-in snapshots, which marginalizes weights and noise jointly.

Target and sweep. The target is a single 2-D isotropic Gaussian at the origin with σ_{target}=0.5. Figure 27 sweeps $(\sigma_p, \eta) \in \{0.3, 1, 3, 10\} \times \{0.1, 0.3, 1, 3, 10\}$ at up to twelve seeds per cell. The Gaussian prior $\mathcal{N}(0, \sigma_p^2 I)$ on the BayesGAN weights identifies σ_p with the SBM prior standard deviation $\sqrt{2/(N\gamma)}$, so smaller σ_p corresponds to larger γ. The closed-form prediction of Fig. 5 (warm at small γ, cold at large γ) therefore maps onto warm at large σ_p, cold at small σ_p. In $D_{\text{KL}}(P_{\text{pp}} \| P^*)$ (panel A), the argmin shifts as predicted: from $\eta^{\text{rev}}_{\text{pp}}$=0.3 at weak priors ($\sigma_p \geq 1$) to $\eta^{\text{rev}}_{\text{pp}}$=3 at the strongest prior (σ_p=0.3). The forward $D_{\text{KL}}(P^* \| P_{\text{pp}})$ (panel B) has a cold optimum at σ_p=1. Both forward and reverse KLs place the optimum cold for small σ_p and warm for large σ_p. At

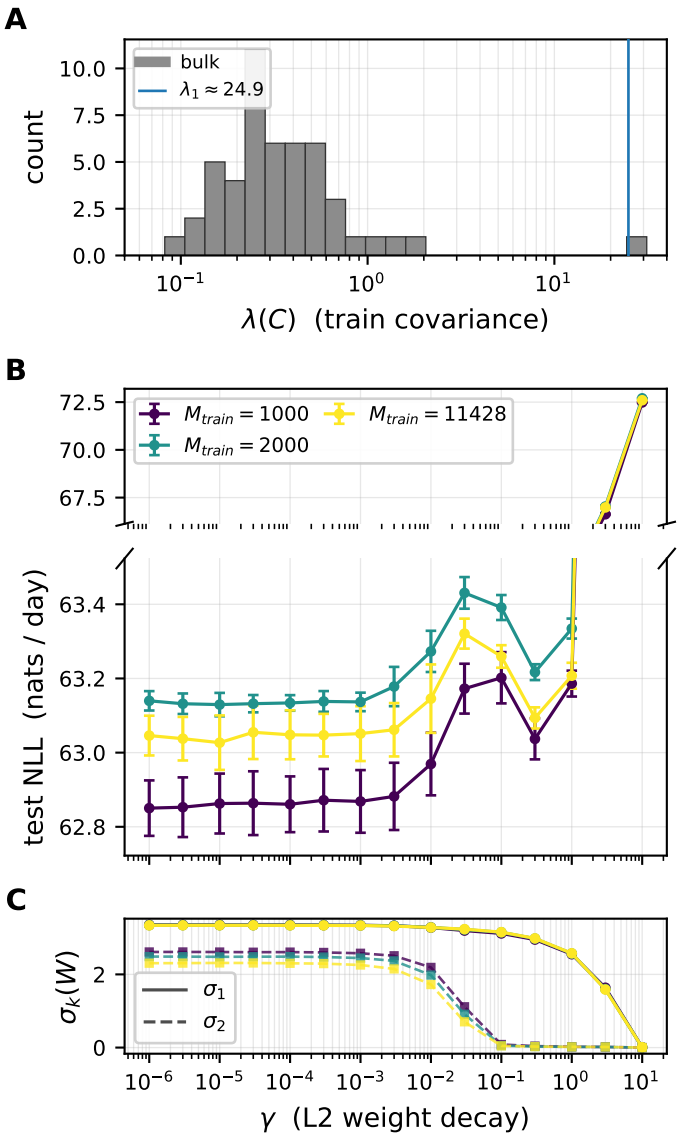


Figure 26: **Double descent in a Gaussian-visible RBM trained on financial return data. A)** Eigenvalue spectrum of the empirical training-set covariance C on a log-λ axis. The bulk (gray histogram) spans $\lambda \sim 10^{-1}$ to $\lambda \sim 2$, dominated by sectoral correlations; a single rank-one outlier sits at $\lambda_1 \approx 25$ (blue line), more than an order of magnitude above the bulk maximum $\lambda_2 \approx 1.6$, the "market mode" that drives the retarded-learning transition seen in panel **B**. **B)** Test negative log-likelihood (NLL) per day as a function of L_2 weight decay γ, for training-set sizes $M_{\text{train}} = 1000$, 2000, and 11428 (full training set). A broken y-axis separates the detailed view of the double-descent bump (lower part, 62.8–63.5 nats/day) from the high-γ collapse to the uncorrelated-Gaussian floor (upper part, $\approx$ 67–73 nats/day). **C)** Top two singular values σ_1 (solid) and σ_2 (dashed) of the trained weight matrix W. At $\gamma \lesssim 10^{-2}$, both are comparable ($\sigma_1 \approx 3$, $\sigma_2 \approx 2.5$); at $\gamma \approx 3 \times 10^{-2}$–$10^{-1}$, σ_2 drops to zero while σ_1 persists, marking the spectral transition that coincides with the NLL bump. Data: Kenneth French 49 industry daily value-weighted returns (1926–2026), standardized and winsorized at $\pm 5\sigma$. Model: Gaussian–Bernoulli RBM (N=49, M=20), trained by PCD with Adam burn-in; log-partition function estimated by AIS. Error bars: $\pm 1\,\sigma$ over 8 training seeds at fixed data subset.

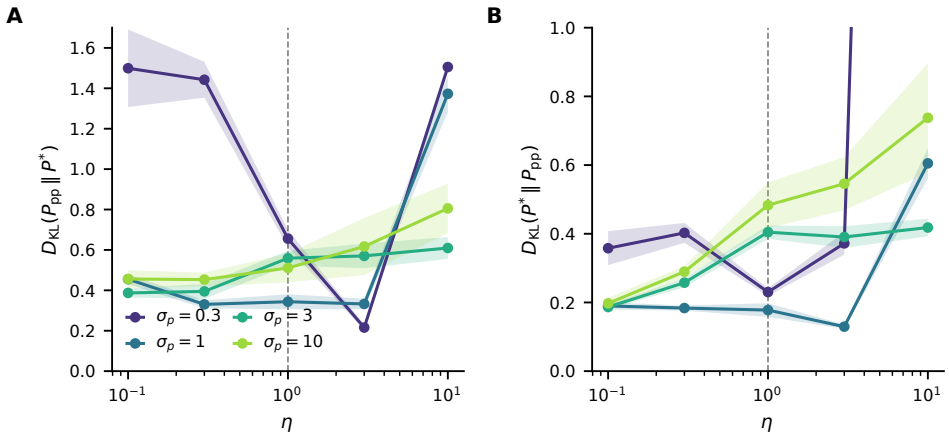


Figure 27: **Tempered Bayesian GAN on a unimodal target: warm-to-cold migration with prior strength. A)** Posterior-predictive reverse KL, $D_{\mathrm{KL}}(P_{\mathrm{pp}}\|P^*)$, versus η, for four prior widths $\sigma_p \in \{0.3, 1, 3, 10\}$. **B)** Forward $D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}})$. Target: single isotropic Gaussian at the origin with σ_{target}=0.5. Up to twelve seeds per (σ_p, η) cell. Bands are ± 1 SEM. Vertical dashed line: $\eta = 1$ (Bayes).

σ_p=0.3, η=10 the tight prior freezes the chain near $\theta \approx 0$, off the target; the forward KL in panel B then diverges because the KDE of P_{pp} leaves the support of P^* uncovered, while the reverse KL in panel A stays bounded since P^* is positive everywhere.

G.6 Out-of-equilibrium training dynamics in a Potts Boltzmann machine

The training dynamics of Sec. 2.2 predicts that the stationary sample–data overlap s_{st} undergoes a phase transition at the critical sampling rate $\nu_c = \gamma/(c_1\chi_P)$ [Fig. 2C]: above ν_c the persistent chain aligns with the data spike, while below ν_c the chain remains uncondensed ($s = 0$) even though the weight matrix has already acquired the data signal. This out-of-equilibrium (OOE) phenomenon mirrors the empirical findings of [Decelle et al., 2021], who demonstrated that RBMs trained by persistent contrastive divergence operate in qualitatively distinct equilibrium or OOE regimes depending on whether the per-step MCMC budget exceeds or falls below the model's mixing time.

We test whether this OOE mechanism is observable in a discrete Potts Boltzmann machine trained on real sequence data. Specifically, we train a pairwise Potts model on sequences sampled from a lattice-protein teacher at $\beta_{\mathrm{sel}} = 1000$ ($L = 27$ sites, $q = 20$ amino acids), using PCD in the MAP limit ($\eta \to \infty$, i.e. no weight noise), with L_2 regularization $\gamma = 0.01$ and learning rate $\eta_{\mathrm{lr}} = 0.01$. We control the sampling rate via k, the number of single-site Gibbs update attempts per gradient step (one full sweep corresponds to $k = L = 27$).

Order parameters. We monitor the chain variances $\sigma_a^2 = \langle[(\mathrm{oh}(\mathbf{x}) - f_1)^\top \mathbf{c}_a]^2\rangle$ along the data-spike directions, the Potts analog of the SBM signal-overlap variance $\langle s_a^2\rangle$, and the top coupling eigenvalue $\max_k |\lambda_k(J)|$, the analog of $\lambda_1(W)$. Here $f_1 \equiv \langle \mathrm{oh}(\mathbf{x})\rangle_{\mathrm{data}}$ are the empirical one-site frequencies and $\mathbf{c}_a$ are the eigenvectors of the centered one-hot data covariance. The top coupling eigenvalue integrates the OOE bias over the full training trajectory: when the negative phase is biased ($\nu < \nu_c$), the data term in (3) is not canceled, driving the coupling eigenvalues beyond their equilibrium values. The weight *direction*, measured by the subspace overlap $\sum_a \|\mathrm{Proj}_J \mathbf{c}_a\|^2$ between the top eigenvectors of J and the data eigenvectors, is predicted to be sampling-rate independent, since eigenvalue detachment occurs at the early-time scale (14) regardless of ν.

Results. Figure 28A shows the final $\max_k |\lambda_k(J)|$ after $t_{\mathrm{age}} = 5000$ gradient steps, as a function of k/L (the number of sweeps per step), with a single persistent chain ($N_{\mathrm{chains}} = 1$). At $k = 1$ (one site update per step), the top eigenvalue reaches $\lambda \approx 19.5$, overshooting the equilibrium value $\lambda_{\mathrm{eq}} \approx 12.4$ (attained at $k \geq 81$, dashed line) by $\sim 57\%$. The overshoot decreases monotonically with k and plateaus above $k/L \approx 3$. This behavior is robust across five independent training seeds (error bars in Fig. 28A).

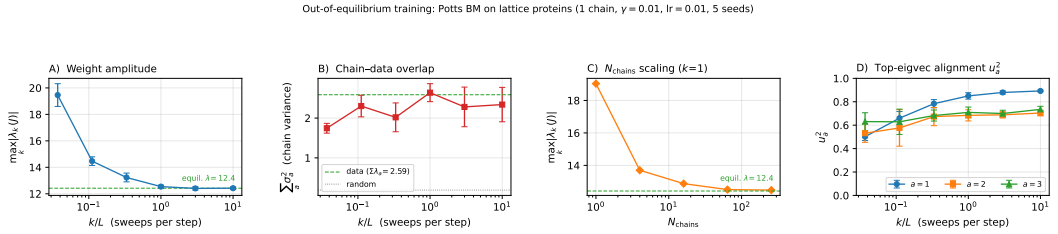


Figure 28: **Out-of-equilibrium training in a Potts BM.** We train a Potts BM (L=27, q=20) on lattice-protein sequences generated at β_{sel}=1000, using PCD–MAP with γ=0.01 and η_{lr}=0.01. Training lasts t_{age}=5000 steps with N_{chains}=1, except in panel C. Error bars in A, B and D show $\pm 1\sigma$ over five seeds.
A) Top coupling eigenvalue $\max_k |\lambda_k(J)|$ versus k/L, the number of Gibbs sweeps per gradient step. At small k, the single-chain negative phase fails to balance the data gradient, and the weights overshoot equilibrium (dashed line). **B)** Chain variance in the data-spike directions, $\sum_a \sigma_a^2$, averaged over the final 20% of training. It rises from ≈ 1.7 at k=1 to ≈ 2.5 at $k \geq L$, matching the data value $\sum \lambda_a = 2.59$ (dashed line). **C)** At k=1, the weight overshoot disappears as N_{chains} increases and averaging reduces single-chain noise. This agrees with the absence of visible OOE in standard DCA training ($N_{\text{chains}} \sim 500$) on these families. **D)** Alignment $u_a^2 = \max_k(\hat{v}_k(J) \cdot \hat{\mathbf{c}}_a)^2$ for each data mode. The leading-mode alignment rises from 0.50 at k=1 to 0.89 at k=10L. Other modes lose comparable alignment at small k, extending the $u_1^2 < 1$ signature to several data directions.

Panel B shows the chain-side observable: the variance $\sum_a \sigma_a^2$ of the persistent chain in the data-spike directions. At $k = 1$ the chain explores only $\sum \sigma^2 \approx 1.7$, roughly 30% below the data eigenvalue sum $\sum \lambda_a = 2.59$ (dashed), while at $k \geq L$ the chain variance rises to match the data value. This is the Potts analog of the $s_{\text{st}}(\nu)$ bifurcation of Fig. 2C, though noisier because a single chain provides only one sample per measurement.

Panel C shows the complementary effect: at fixed $k = 1$, increasing N_{chains} from 1 to 256 suppresses the weight overshoot, converging to $\lambda_{\text{eq}} \approx 12.5$ by $N_{\text{chains}} \approx 64$. On the protein families considered here, standard DCA runs with $N_{\text{chains}} \sim 500$ and $k \sim 10$ show no visible OOE signature, consistent with this observation.

All three panels confirm that the weight *direction* (eigenvector alignment with data) is k-independent ($\sum_a \|\text{Proj}_J \mathbf{c}_a\|^2 \approx 2.5$ throughout, not shown), consistent with the theory's prediction that representation acquisition precedes, and is independent of, sample condensation: the "representation without generation" signature.

Extension to real protein data. The same protocol on the SH3 domain PF00018 (alignment of App. G.3.2; $t_{\text{age}} = 10^4$, three seeds) reproduces the three OOE signatures (Fig. 29): at k=1 the weight amplitude overshoots $\lambda_{\text{eq}} \approx 13.3$ by a factor 3.6, the normalized chain variance $\sigma_a^2/\lambda_a^{\text{data}} \approx 0.7$ is suppressed in all three top data directions, and the J-to-data subspace overlap stays within 1% of its N_{chains}=256 value. All three signatures collapse onto equilibrium by $k \approx L/5$–$L/3$, confirming that the OOE mechanism persists when the lattice-protein teacher is replaced by real MSA correlations.

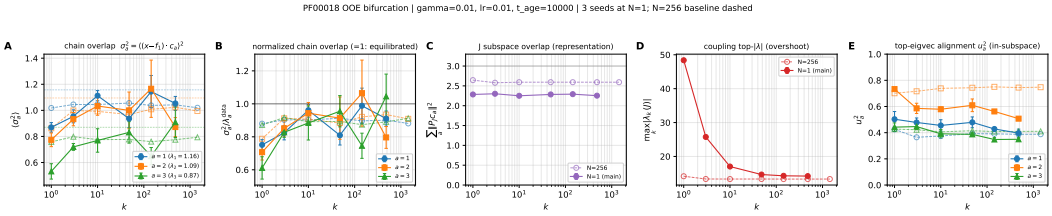


Figure 29: **Out-of-equilibrium training dynamics on real protein data (SH3 domain PF00018).** Following Fig. 28, we train a pairwise Potts BM on PF00018 (L=48, q=21, M=1.76 $\times$ 10^5). Training uses PCD–MAP with γ=0.01, $\eta_{\rm lr}$=0.01, and $t_{\rm age}$=10^4. Solid markers with error bars show results for $N_{\rm chains}$=1, averaged over 3 seeds. Open markers and dashed lines show the $N_{\rm chains}$=256 equilibrium baseline. The data directions $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ are the leading eigenvectors of the reweighted centered one-hot MSA covariance. Their eigenvalues are $\lambda_1^{\rm data}$=1.16, $\lambda_2^{\rm data}$=1.09, $\lambda_3^{\rm data}$=0.87 (dotted lines in A). Time averages use the second half of training.
A) Chain variance σ_a^2 along each data direction. **B)** Normalized chain variance $\sigma_a^2/\lambda_a^{\rm data}$. Below $k \approx L/5$, the ratio remains below 1: the chain has not equilibrated in the data subspace. At larger k, it fluctuates about 1 for all three modes. **C)** Subspace overlap $\sum_a \|{\rm Proj}_J \mathbf{c}_a\|^2$ (maximum 3). It is independent of k with both 1 and 256 chains. This confirms the "representation without generation" prediction: J acquires the data directions regardless of sampler equilibration. **D)** Top coupling eigenvalue $\max_k |\lambda_k(J)|$. With one chain at k=1, it overshoots the equilibrium value $\lambda_{\rm eq} \approx 13.3$ by a factor ≈ 3.6. This reproduces the overshoot seen in Fig. 28. **E)** Per-mode alignment $u_a^2 = \max_k(\hat{v}_k(J)\cdot\hat{\mathbf{c}}_a)^2$. Mode a=2 stays ~ 0.20 below the baseline with 256 chains throughout the k range; modes a=1, 3 agree within the variability across seeds. This alignment loss depends on the training regime and is weaker than for the lattice-protein teacher because the three leading data eigenvalues are nearly degenerate.

H Parameter values used in main-text figures

This appendix collects the numerical parameter values used to produce each main-text figure.

Fig. 1 (Equilibrium phase diagram for K=1). Panel A is a schematic; panel B is the analytic phase diagram in the (γ, η) plane. Panel C uses $\eta = 0.5$ (small-η regime, λ_1 sticks to the bulk edge); panel D uses $\eta = 5$ (larger-η regime, λ_1 detaches before returning to the edge).

Fig. 2 (Training dynamics). Panels A and B use $\gamma = 0.4$, $\eta = 10$, $\nu = 0.7$, $c_1 = 1.7$, $c_2 = 0.3$, with seed overlaps $s_1(0) = s_2(0) = 0.1$. Panel C uses $\gamma = 0.9$, $\eta = 3$. Panel D shows the MAP limit $\eta \to \infty$.

Fig. 3 (Sampling temperature tuning). Rank-one teacher with $\omega^* = 2.5$ ($c_1 = 1.6$, $c_2 = 0.4$); student at $\eta = 5$. Panel B shows curves at $\gamma \in \{1.7, 2.0, 3.0\}$.

Fig. 4 (Double descent). Rank-one teacher with $\omega^* = 2.5$ ($c_1 = 1.6$, $c_2 = 0.4$). Solid curves cover $\eta \in \{0.1, 0.2, 0.4, 0.7, 1.0, 1.5, 2.0, 3.0, 5.0\}$. The two orange dashed curves are the critical analytic anchors: the saddle-node threshold $\eta = \eta_{\mathrm{DD}}(\omega^*{=}2.5) \approx 1.342$ (upper), below which the first descent disappears, and the MAP limit $\eta \to \infty$ (lower), in which the second descent collapses to a flat PMo plateau.

Fig. 5 (Tempered-posterior). Panel A sweeps (ω^*, γ). Panel B fixes $\omega^* = 2.2$ and shows the posterior-predictive reverse KL versus η at $\gamma \in \{0.1, 0.3, 1.0\}$, representative of the warm, cold, and MAP optima.

Fig. 6 (Out-of-equilibrium training trajectories). Finite-N Langevin simulations with $N = 1500$ and rank-one teacher $\omega^* = 2.5$, at $\gamma = 0.4$, $\eta = 10$. Color encodes the sampling rate ν, swept over the 15 values $\nu \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.55, 0.70, 0.85, 1.00, 1.30, 1.70, 2.20, 3.00\}$ (roughly log-spaced, straddling ν_c).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions listed in Sec. 1 match the results derived in Secs. 2.1–3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A *Limitations* paragraph in Sec. 5 states the scope boundaries and the heuristic status of the replica/DMFT analysis.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Derivations are given in Apps. A.2, A.1.3, D.2, D.8, B, B.2, and A.4; spherical-integral asymptotics cite Guionnet and Husson [2022].

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Datasets, architectures, optimizers, and hyperparameters are documented in Apps. G.2, G.3, G.4.2, G.4.1, and G.6. Theoretical predictions are closed-form.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/cossio/spherical-bm-2026`. Datasets are public (Pfam, Kenneth French) and `adabmDCA` is open-source.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. G.2 specifies the small-RBM model and comparisons. Its posterior sampler uses eight independent replica-exchange HMC ladders per dataset, with $3000/12000$ warmup/production updates for the regularization comparison and $4000/20000$ for posterior-temperature comparisons ($8000/40000$ for one warm ensemble after a mixing check failed); losses use 512 and 4000 draws per ladder, respectively. Apps. G.3, G.4, and G.6 describe the other training settings.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Empirical curves report seed-averaged means with $\pm$SEM or $\pm 1\sigma$ bands (Figs. 9, 26, 25, 28, 29). App. G.2 reports a paired 95% interval across datasets for the slow-sampling comparison and identifies the posterior Monte Carlo precision limitation. Theoretical curves are asymptotic large-N predictions.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The small-RBM experiments used CPU workers on Jean Zay: 21.77 recorded CPU-hours including validation runs, maximum job wall time 22.6 minutes, and maximum recorded batch memory 4.87 GiB. The dense posterior-predictive evaluations used up to 16 local CPU workers and 29.7 minutes of total evaluation wall time; other local analysis is excluded. Other empirical experiments ran on standard CPU/GPU workstations under 48h wallclock; the DMFT pipeline converges in minutes on CPU.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Theoretical analysis and empirical validation on public datasets; no human subjects, scraped data, or deployed models.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The "Broader impact" paragraph in Sec. 5 addresses this.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No models or datasets are released.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Pfam (CC0), Kenneth French (academic-use), and `adabmDCA` Rosset et al. [2026] are credited and used within license terms.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pretrained models are released.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects or crowdsourcing.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A]

Justification: LLMs were not used in the core methods; any use was limited to writing assistance, which per NeurIPS policy does not require declaration.