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 JORGE.FDCD@IPHT.FR
Institut de Physique Théorique, Université Paris-Saclay, CNRS UMR3681, CEA, Gif-sur-Yvette, France

Abstract

Energy-based models (EBMs) are flexible generative architectures inspired by statistical physics, but their learning and generative properties remain poorly understood. We analyze a solvable EBM in the high-dimensional limit, the spherical Boltzmann machine (SBM), combining random matrix theory and dynamical mean-field theory. We solve exact equations for its training dynamics; compute the Bayesian evidence, which acts as a partition function over parameters; and uncover cascades of phase transitions, both at equilibrium and along training, driven by the alignment and condensation of the top modes of the coupling matrix. In a teacher–student setting we connect these transitions to four generative phenomena: sampling temperature tuning, double descent of generative metrics versus regularization, tempered-posterior effects in posterior-predictive ensembles, and out-of-equilibrium training dynamics that bias the converged model. Numerical experiments on Potts BMs, normalizing flows, autoregressive networks, RBMs, and Bayesian GANs reproduce the same phenomenology beyond the SBM.[1]

1. Introduction

Energy-based models (EBMs) define complex high-dimensional distributions through an energy function $\mathcal{E}(\mathbf{x}; W)$ and a Boltzmann law $P(\mathbf{x}|W) = e^{-\beta\mathcal{E}(\mathbf{x};W)}/Z_\beta(W)$. They are competitive generative models for images [17], biological sequences [21, 50], and elsewhere. The normalization constant $Z_\beta(W) = \int e^{-\beta\mathcal{E}(\mathbf{x};W)}\mathrm{d}\mathbf{x}$, also known as the partition function, is an intractable high-dimensional integral. As a consequence, training and sampling often rely on Monte Carlo (MC) approximations, which also hinders theoretical analysis of the trained model.

Contributions. We study *spherical Boltzmann machines* (SBMs), a solvable EBM defined on the N-sphere. Combining recent finite-rank spherical-integral asymptotics [26] with a Coulomb-gas analysis of the coupling spectrum, we obtain a closed-form large-N theory: an equilibrium phase diagram of the trained SBM, a dynamical mean-field theory (DMFT) [12] of its Langevin training that reveals a non-equilibrium phase transition, and closed-form generative metrics in a teacher–student setting. Our solvable model predicts and characterizes four effects: **sampling temperature tuning** [8, 21, 22, 50], where tuning $\beta \neq 1$ post-training improves the fidelity of generated samples; **double descent** of generative metrics as a function of regularization, in line with reports in GANs [37], LLMs [27, 42], diffusion models [23, 66], autoencoders [47], and RBMs [9]; **tempered-**

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

© T. Tulinski, S. Cocco, R. Monasson & J. Fernandez-de-Cossio-Diaz.

posterior effects in posterior-predictive ensembles, the generative analogue of cold/warm-posterior effects in Bayesian deep learning [43, 44, 46, 58, 61]; and **out-of-equilibrium** training dynamics that result in persistent biases of the model after training [3, 14]. We also provide numerical evidence that all four effects arise in standard generative architectures.

Related work. Statistical-mechanics methods have been used to characterize discriminative networks [25, 64, 65] and random-weight generative models [2, 13, 57]. Analytical progress on trained EBMs has relied on early-training approximations [3, 19, 55]. DMFT has been used to study the high-dimensional learning trajectories of supervised models [7, 41], but has not been applied to maximum-likelihood EBM training, except in particular cases where the partition function reduces tractably [62]. A solvable EBM where we can characterize exactly the asymptotic equilibrium phase diagram and training dynamics in the high dimensional limit, while predicting various effects also found in realistic generative models is, to our knowledge, missing; the SBM closes this gap.

2. Spherical Boltzmann machines: equilibrium and training dynamics

The spherical Boltzmann machine (SBM) is an EBM defined on the N-sphere, $\mathcal{S}_N = \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|^2 = N\}$ with a pairwise energy function $\mathcal{E}(\mathbf{x}; W) = -\frac{1}{2}\mathbf{x}^\top W \mathbf{x}$, where $W \in \mathrm{Sym}_N$ is a parameter coupling matrix. It is a classical disordered-system model from statistical physics [5, 32], obtained as a continuous relaxation of Ising spins under the global constraint $\|\mathbf{x}\|^2 = N$. The probability distribution $P(\mathbf{x}|W) \propto e^{-\beta\mathcal{E}(\mathbf{x};W)}$ is known as the Bingham distribution in directional statistics [6, 30], and differs from a Gaussian distribution due to the spherical constraint. Most of the phenomenology below depends critically on this difference and would not arise in a purely Gaussian model.

Whereas in previous works on the SBM W follows a simple distribution (*e.g.* GOE) [5, 32], here it is trained by gradient ascent of the log-posterior. A set of training data $\mathcal{D} = \{\mathbf{x}^k\}_{k=1}^K \subseteq \mathcal{S}_N$ enters the likelihood only through the empirical covariance $C = \frac{1}{N}\sum_k \mathbf{x}^k \mathbf{x}^{k\top}$. Persistent-MC training then yields the coupled Langevin system [11, 16, 49, 56, 63]

$$\partial_t W(t) = \tfrac{1}{2}\big(C - \tfrac{K}{N}\mathbf{x}(t)\mathbf{x}(t)^\top\big) - \tfrac{\gamma}{2}W(t) + \Omega(t)/\sqrt{\eta N}, \tag{1}$$

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

where $\gamma > 0$ is the weight decay, $\Omega(t)$ a symmetric matrix-valued GOE white noise, $\eta > 0$ an inverse *learning* temperature controlling weight-space dispersion (distinct from the sample-space β), $\kappa(t)$ a Lagrange multiplier enforcing $\|\mathbf{x}(t)\|^2 = N$, and ν the sampling rate of the persistent MC chain per gradient step. In practice, increasing ν requires spending more compute cycles updating the MC chains during training. We focus on the undersampled regime K fixed, $N \to \infty$; the same theory extends to large but effectively low-dimensional data (App. C).

2.1. Equilibrium phase diagram

In the limit $\nu \to \infty$ the dynamics equilibrate to the tempered posterior $P_\eta(W|\mathcal{D}) \propto e^{\mathcal{L}(W;C)}$ with $\mathcal{L}(W;C) = \frac{N}{2}\eta \operatorname{tr} CW - \frac{N}{4}\gamma\eta \operatorname{tr} W^2 - K\eta \ln Z(W)$, normalised by the tempered evidence $\mathcal{Y}(C) = \int e^{\mathcal{L}(W;C)}\mathrm{d}W$. Diagonalising $W = \frac{1}{N}V\Lambda V^\top$ rewrites $\mathcal{L}$ as a Coulomb gas [36],

$$\mathcal{L}_\Lambda(\Lambda; C) = -\tfrac{N}{4}\gamma\eta \operatorname{tr}\Lambda^2 - K\eta \ln Z(\Lambda) + \ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) + \sum_{i<j} \ln|\lambda_i - \lambda_j|, \tag{3}$$

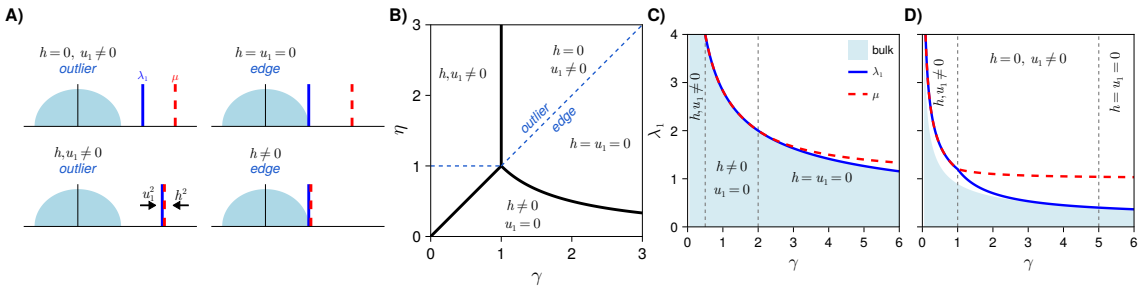

Figure 1: **Equilibrium phase diagram for K=1. A)** Schematic eigenvalue configurations: bulk semicircle (light blue), top eigenvalue λ_1, Lagrange multiplier μ. **B)** Phases in the (γ, η) plane; blue dashed line separates outlier from edge regimes. **C)** λ_1 and μ vs. γ at $\eta = 0.5$, where λ_1 sticks to the edge; vertical dashed lines mark phase transitions. **D)** Same at $\eta = 5$, where λ_1 detaches before returning to the edge.

where $\mathcal{J}_{\text{HCIZ}}$ is the real Harish–Chandra–Itzykson–Zuber integral [26] encoding the alignment of W's eigenvectors with the data principal components. The dominant $\mathcal{O}(N^2)$ contribution sets a Wigner semicircle bulk of width 2σ with $\sigma = 1/\sqrt{\gamma\eta}$ (Eq. (6), App. A.2).

The top K eigenvalues interact with the data and may detach (Fig. 1A). Stationarity of (3) reduces (App. A.2) to a balance of $\mathcal{O}(N)$ forces on each λ_k: an HCIZ pull away from the bulk, $\frac{N}{2}\eta c_k u_k^2$, and a partition push toward it, $-\frac{NK}{2}\eta h^2$. Here $u_k^2 = \langle(\mathbf{c}_k\cdot\mathbf{v}_k/N)^2\rangle$ is the overlap between the k-th eigenvector of W and the k-th eigenvector of C, and $h^2 = \sum_k h_k^2$ with $h_k^2 = \langle(\mathbf{x}\cdot\mathbf{v}_k/N)^2\rangle$ the projection of typical generated samples onto $\mathbf{v}_k$. Concentration of u_k, h around their typical values partitions the (γ, η, c_k) space into qualitatively distinct phases (paramagnetic edge, paramagnetic outlier, ferromagnetic edge/outlier, random-direction condensed; full conditions in App. A.2). Fig. 1B shows the K=1 diagram; Fig. 1C,D show the outlier-detachment and condensation transitions in γ at two η. In phases with h=0 the projections $\mathbf{x}\cdot\mathbf{v}_j$ are all $\mathcal{O}(\sqrt{N})$ and samples are biased but nearly isotropic; in phases with $h\neq 0$ they *condense* along the top mode(s) with $\mathcal{O}(N)$ projections. For $K > 1$, modes can sit in different phases or coalesce simultaneously (App. Fig. 3); see also [58] and App. A.4 for an alternative replica derivation.

2.2. Training dynamics: DMFT

Equilibration to $P_\eta(W|\mathcal{D})$ is guaranteed only as $\nu \to \infty$; at finite ν the dynamics can settle into a non-equilibrium stationary state. Initialising $W(0)$ from the prior and $\mathbf{x}(0)$ near the top data mode, the Martin–Siggia–Rose path integral [40] closes (1)–(2) in the large-N limit onto a DMFT for the data overlap $s_k(t) = \mathbf{c}_k\cdot\mathbf{x}(t)/N$, the correlator $Q(t,t') = \mathbf{x}(t)\cdot\mathbf{x}(t')/N$, and the causal response $R(t,t')$, with memory and noise kernels $M(t,u), D(t,u)$ and a Lagrange multiplier $\kappa(t)$ fixed by $Q(t,t) = 1$ (App. B). Early in training the negative-phase contribution is weak and (1) reduces to deterministic spike growth, placing the k-th outlier at $\lambda_k(t) = g_k(t)^{-1} + g_k(t)/(\gamma\eta)$ with $g_k(t) = \gamma/[c_k(1 - e^{-\gamma t/2})]$, above a detachment time $t_{\text{out},k} = -(2/\gamma)\ln(1 - c_k^{-1}\sqrt{\gamma/\eta})$ defined for $c_k > \sqrt{\gamma/\eta}$ (App. B). At every time $W(t)$ is a Wigner bulk plus at most K outliers (App. B.1). Fig. 2A,B confirms this against finite-N Langevin simulations.

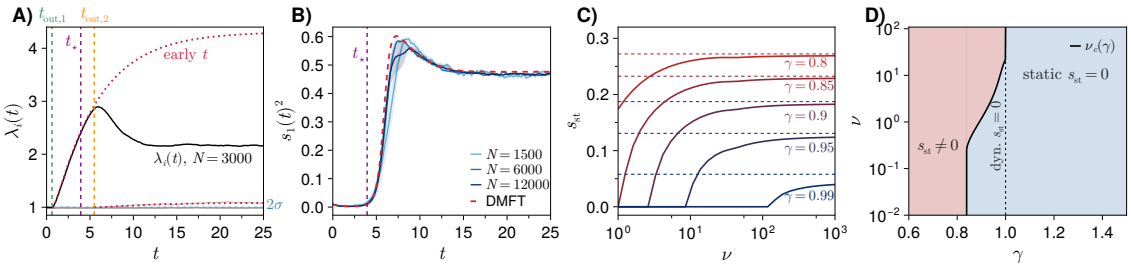

Figure 2: **Training dynamics.** **A)** Top eigenvalue trajectories $\lambda_i(t)$ from finite-N simulations (black) vs. early-time prediction (red dotted). Vertical dashed: detachment and condensation times; blue horizontal: bulk edge 2σ. **B)** $s_1(t)$ from DMFT (red) vs. averaged finite-N simulations (blue). **C)** Stationary overlap $s_{\rm st}$ vs. ν at selected γ; dashed lines: equilibrium $\nu \to \infty$ asymptotes. **D)** MAP-limit dynamical phase diagram ($\eta \to \infty$); $\nu_{\rm c}$ separates $s_{\rm st}{\neq}0$ from $s_{\rm st}{=}0$.

For long times the DMFT admits a stationary solution $s_k(t) \to s_{{\rm st},k}$, $Q(t,t') \to Q_{\rm st}(t-t')$, $R(t,t') \to R_{\rm st}(t-t')$. The trivial branch $s_{{\rm st},1} = 0$ is the only solution for low ν, but above a critical $\nu_{\rm c}(\gamma,\eta)$ a condensed branch $s_{{\rm st},1} > 0$ emerges (Fig. 2C and App. B). Fig. 2D shows the MAP-limit dynamical phase diagram: for $\nu < \nu_c(\gamma)$ with $\gamma < 1$, training never reaches the correct equilibrium and the converged model is biased. Out-of-equilibrium EBM training has been discussed previously [3, 14], but the DMFT above is, to our knowledge, the first closed-form large-N theory of EBM training dynamics.

3. Generative metrics in a teacher–student scenario

We evaluate a student trained on K samples from a teacher $\mathbf{x}^k \sim P^*$ [18, 53]. A *typical* student commits to one $W \sim P_\eta(W|\mathcal{D})$; a *posterior-predictive* (pp) student instead marginalises, $\mathbf{x} \sim P_{\rm pp}(\mathbf{x}|\mathcal{D}) = \int P(\mathbf{x}|W)P_\eta(W|\mathcal{D})\,{\rm d}W$. Sample quality is measured by the reverse KL $D_{\rm KL}(P_W\|P^*)$ (mode-seeking) and the forward KL $D_{\rm KL}(P^*\|P_W)$ (mode-covering), and analogously for $P_{\rm pp}$. By convexity, the pp student dominates the typical one in both directions (App. D.8). All four divergences are self-averaging at large N and computable in closed form for a rank-one teacher $W^* = \omega^*\mathbf{w}^*\mathbf{w}^{*\top} - N^{-1}\omega^* I$, whose $K = 2$ samples yield a C with eigenvalues $c_1 = 2 - 1/\omega^*$, $c_2 = 1/\omega^*$ (App. D). Each phenomenon below surfaces most naturally in one of the four divergences.

Sampling temperature tuning. After training, re-introducing a sampling temperature $1/\beta \neq 1$ rescales $W \to \beta W$ and can transport the model across phases. The forward $D_{\rm KL}(P^*\|P_{\beta W})$ is convex in β, with a unique optimum $\gtrless 1$ according to whether the entropy $\mathcal{H}[P_W]$ exceeds or falls short of the cross-entropy $\mathcal{H}[P^*;P_W]$, so $\beta > 1$ (< 1) compensates for an excess (lack) of entropy in the trained model. The response depends on the starting phase. From $h \neq 0, u_1 = 0$ (condensed along a random direction) the optimum is always $\beta < 1$, landing the model in an uncondensed phase and erasing the spurious direction ($u_1 \to 0$), which lowers the divergence from P^*. From $h = 0, u_1 \neq 0$ the coupling carries data-aligned modes but the imprinted signal is too weak, so samples are swamped by bulk noise; $\beta > 1$ pushes the aligned outliers past the condensation

threshold, transporting the model to a $h, u_1 \neq 0$ phase in which the learned modes become active in generated data (App. Fig. 8, App. D.9).

Double descent. The reverse $D_{\mathrm{KL}}(P_W \| P^*)$ exhibits a double descent in γ within a window $\eta_{\mathrm{DD}}(\omega^*) < \eta < \infty$ (App. Fig. 9). A first minimum lies inside the $h \neq 0$ phase where the student condenses along the teacher; beyond the optimal γ, condensation weakens and the curve spikes at the transition $\gamma = c_1$; in the $h = 0$ regime it descends as $\frac{1}{N} D_{\mathrm{KL}}(P_W \| P^*) = \frac{\omega^* - 1 - \ln \omega^*}{2} + \frac{1}{4\gamma\eta}$, the second descent driven by the random anisotropy of the bulk modes. Below η_{DD} the bulk is overspread and the curve is monotone; in the MAP limit the second descent collapses to a flat plateau and only the first minimum survives. The threshold $\eta_{\mathrm{DD}}(\omega^*)$ is derived in App. D.10.

Tempered-posterior effects. The MAP limit ($\eta \to \infty$) corresponds to standard single-model training; few works target the Bayesian ($\eta = 1$) posterior predictive in generation [51]. We define $\eta_{\mathrm{pp}}^{\mathrm{rev}}(\gamma, \omega^*) = \arg\min_{\eta > 0} D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*)$. The plane (γ, ω^*) partitions into five regions (App. Fig. 10A): a *warm* phase ($\eta_{\mathrm{pp}}^{\mathrm{rev}} < 1$) where averaging across the posterior beats sharpening on its mode; a *cold* phase ($\eta_{\mathrm{pp}}^{\mathrm{rev}} > 1$) where the posterior is informative and concentration helps; a *MAP* phase ($\eta_{\mathrm{pp}}^{\mathrm{rev}} \to \infty$); and two degenerate regions where h or u_1 vanish and the loss is η-independent over an interval. Closed-form thresholds are derived in App. D.11; the forward pp-KL has the same qualitative structure (App. Fig. 11). This is, to our knowledge, the first analytical instance of cold/warm-posterior effects in a generative EBM.

Out-of-equilibrium training dynamics. The dynamical regimes of Sec. 2.2 shape generative quality. At $\nu \gg \nu_{\mathrm{c}}$, $\lambda_1(t)$, $u_1(t)$ and the reverse-KL all converge to their equilibrium values. As ν decreases, $\lambda_1(t)$ overshoots its equilibrium value before slowly relaxing back; $u_1(t)$ stays high transiently and then drops below its equilibrium plateau; the reverse-KL dips and then settles at a higher value. The early-time spike formula yields an optimal early-stopping time independent of ν (App. D.12, Eq. (185); App. Fig. 12).

4. Validation in standard generative architectures

To show that the qualitative phenomenology predicted by the SBM extends beyond its theoretical setting, we run numerical experiments in standard generative architectures trained on real and synthetic data; in each case the SBM mechanism is reproduced at the qualitative level.

Sampling temperature tuning. Potts BMs trained on protein families from Pfam [28, 39, 59] carry data-aligned modes that stay silent in generation; tuning $\beta > 1$ rescues them, in agreement with the SBM $h{=}0, u_1{\neq}0$ to $h{\neq}0, u_1{\neq}0$ transport (App. G.1). **Double descent.** A normalising flow on $\mathcal{S}_N$ [48] targeting the rank-one SBM teacher, a Gaussian–Bernoulli RBM trained on the Kenneth French 49-industry portfolio [20, 34], and an autoregressive belief-network on 2D Ising and Curie–Weiss data all reproduce double descent in γ (App. G.2). **Tempered-posterior effects.** Bayesian GANs [51] on synthetic datasets show optimal posterior temperatures both < 1 and > 1 depending on regularisation, matching the SBM warm/cold structure (App. G.3). **Out-of-equilibrium training.** Potts BMs trained at varying ν exhibit the same overshoot of top weight modes and converged sub-optimal alignment (App. G.4).

5. Conclusion and limitations

We introduced the SBM as a solvable EBM, derived its equilibrium phase diagram, solved its Langevin training via DMFT, and obtained closed-form teacher–student generative metrics. Stan-

dard architectures qualitatively reproduce the predicted temperature tuning, double descent, tempered-posterior effects, and out-of-equilibrium training. The phenomenology hinges on the spherical constraint and is absent in an unconstrained Gaussian model (App. F). **Limitations.** The SBM is a minimal solvable EBM; we assume effectively low-dimensional data (App. C); our derivations rely on statistical-physics methods (saddle-point evaluation of the Coulomb gas, MSR/DMFT closure of the Langevin dynamics) rather than rigorous proofs, expected to be asymptotically exact at $N \to \infty$ and verified against finite-N simulations.

References

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

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

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

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

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

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

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

[8] Francesco Calvanese, Gianluca Lombardi, Martin Weigt, and Jorge FERNANDEZ-DE-COSSIO-DIAZ. Steering sequence generation in protein language models through iterative lookback monte carlo sampling. *bioRxiv*, 2026. doi: 10.64898/2026.05.01.722156. URL https://www.biorxiv.org/content/early/2026/05/07/2026.05.01.722156.

[9] Prasad Cheema and Mahito Sugiyama. 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_.

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

[11] A.C.C. Coolen, R. Penney, and D. Sherrington. Coupled dynamics of fast neurons and slow interactions. In J. Cowan, G. Tesauro, and J. Alspector, 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.

[12] Leticia F. Cugliandolo. 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.

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

[14] Aurélien Decelle, Cyril Furtlehner, and Beatriz Seoane. 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.

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

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

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

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

[19] Alberto Fachechi, Elena Agliari, Miriam Aquaro, Anthony Coolen, and Menno Mulder. 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.

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

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

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

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

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

[25] Elizabeth Gardner. 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.

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

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

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

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

[30] John T Kent. 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.

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

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

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

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

[35] Dennis V. Lindley. 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.

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

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

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

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

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

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

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

[43] Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. 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.

[44] Lorenzo Noci, Kevin Roth, Gregor Bachmann, Sebastian Nowozin, and Thomas Hofmann. 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.

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

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

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

[48] Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pages 8083–8092. PMLR, 2020.

[49] Lorenzo Rosset, Roberto Netti, Anna Paola Muntoni, Martin Weigt, and Francesco Zamponi. 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.

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

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

[52] Ruslan Salakhutdinov and Iain Murray. 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.

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

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

[55] Robin Thériault, Francesco Tosello, and Daniele Tantari. 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.

[56] Tijmen Tieleman. 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.

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

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

[59] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. 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.

[60] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

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

[62] Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. 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.

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

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

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

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

Appendix A. Equilibrium phase diagram

A.1. Large N asymptotic expressions

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

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

Note that if $\operatorname{tr}(C) = K$ 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:

$$\frac{\ln \mathcal{Y}(C)}{N^2} \sim -\frac{\gamma\eta}{4} \int \lambda^2 \rho(\lambda) \mathrm{d}\lambda + \frac{1}{2} \int \ln|\lambda - \lambda'| \rho(\lambda)\rho(\lambda') \mathrm{d}\lambda \mathrm{d}\lambda' \sim -\frac{3}{8} + \frac{1}{2} \ln \sigma \quad (5)$$

up to terms of lower-order, where $\rho(\lambda)$ is the Wigner semicircle

$$\rho(\lambda) = \frac{\sqrt{4\sigma^2 - \lambda^2}}{2\pi\sigma^2} \mathbf{1}_{|\lambda| \le 2\sigma}, \qquad \sigma = 1/\sqrt{\gamma\eta}. \quad (6)$$

Thus the leading $\mathcal{O}(N^2)$ contribution comes from standard Coulomb-gas terms. Extremization of (5) with respect to ρ recovers (6), and substituting it back gives the right-hand side of (5). It is convenient to define:

$$\begin{aligned} G(z) &= \int \frac{\rho(\lambda)\mathrm{d}\lambda}{z - \lambda} = \frac{1}{2\sigma} \left(\frac{z}{\sigma} - \sqrt{\frac{z^2}{\sigma^2} - 4} \right), \\ F(z) &= \int \rho(\lambda) \ln(z - \lambda) \mathrm{d}\lambda = \frac{\sigma^2}{2} G^2(z) - \ln G(z), \\ B(z) &= \int \frac{\lambda}{z - \lambda} \rho(\lambda) \mathrm{d}\lambda = \frac{z}{2\sigma} \left(\frac{z}{\sigma} - \sqrt{\frac{z^2}{\sigma^2} - 4} \right) - 1 = zG(z) - 1. \end{aligned} \quad (7)$$

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 same proportionality convention that drops N-independent constants in $Z(\Lambda)$ and $\mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C)$). Collecting all order-$N$ contributions,

$$\Phi(C) = \frac{K}{2} \ln(\gamma\eta) + \sum_{i=1}^{K} \left(-\frac{\gamma\eta}{4} \lambda_i^2 + F(\lambda_i) \right) - \frac{\eta}{N} \operatorname{tr}(C) \ln Z(\Lambda) + \frac{1}{N} \ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \quad (8)$$

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 [4]):

$$Z \propto \int_{-\infty}^{\infty} \mathrm{d}\mathbf{x} \int_{c-i\infty}^{c+i\infty} \mathrm{d}\mu \exp\left(\frac{1}{2}\sum_{i=1}^{N}(\lambda_i - \mu)x_i^2 + \frac{N}{2}\mu\right) \tag{9}$$

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 [32, 54], we introduce order parameters $\mathbf{h} = (h_1, \ldots, h_K)$ with $h_i = x_i/\sqrt{N}$, along the top K modes. From (9),

$$Z \propto \int \mathrm{d}h_1 \ldots \mathrm{d}h_K \, \mathrm{d}x_{K+1} \ldots \mathrm{d}x_N \, \mathrm{d}\mu \tag{10}$$

$$\exp\left(\frac{N}{2}\sum_{i=1}^{K}(\lambda_i - \mu)h_i^2 + \frac{1}{2}\sum_{i=K+1}^{N}(\lambda_i - \mu)x_i^2 + \frac{N\mu}{2}\right) \tag{11}$$

$$\propto \int \mathrm{d}h_1 \ldots \mathrm{d}h_K \, \mathrm{d}\mu \, \exp\left\{\frac{N}{2}\left(\sum_{i=1}^{K}(\lambda_i - \mu)h_i^2 - F(\mu) + \mu\right)\right\}. \tag{12}$$

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

$$1 - h^2 = G(\mu), \quad (\lambda_i - \mu)h_i = 0 \tag{13}$$

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

$$h^2 = \begin{cases} 0, \\ 1 - g_1, \end{cases} \qquad \mu = \begin{cases} G^{-1}(1) & \text{if } g_1 \geq 1, \\ \lambda_1 & \text{if } g_1 \leq 1, \end{cases} \tag{14}$$

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 (10) at the saddle point yields

$$\frac{1}{N}\ln Z \sim \frac{\ln(2\pi) + \mu - F(\mu)}{2} = \frac{1}{2}\ln(2\pi) + \begin{cases} \frac{1}{2} + \frac{1}{4\gamma\eta} & g_1 \geq 1, \\ \frac{1}{2}\left(\frac{g_1}{\gamma\eta} + \frac{1}{g_1} - \frac{g_1^2}{2\gamma\eta} + \ln g_1\right) & g_1 \leq 1. \end{cases} \tag{15}$$

to leading order in N, where the additive $\frac{1}{2}\ln(2\pi)$ comes from the Gaussian base measure of (9) and is the constant carried explicitly by the teacher partition function (111).

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

$$\langle\langle \mathcal{E}(\mathbf{x}; W)\rangle_{\mathbf{x}\sim P_W}\rangle_{W\sim P_\eta} = -\left\langle \frac{\partial \ln Z(\beta W)}{\partial \beta}\bigg|_{\beta=1}\right\rangle_{W\sim P_\eta} = \frac{N}{2}(1-\mu) \tag{16}$$

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

$$\langle \mathcal{H}[P_W]\rangle_{W\sim P_\eta} = \langle\langle \mathcal{E}(\mathbf{x}; W)\rangle_{\mathbf{x}\sim P_W} + \ln Z(W)\rangle_{W\sim P_\eta} = \frac{N}{2}[\ln(2\pi e) - F(\mu)] \tag{17}$$

A.1.2. THE REAL SPHERICAL HCIZ INTEGRAL AND ITS LARGE-N ASYMPTOTICS

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

$$\mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) = \int_{V\in \mathrm{O}(N)} \mathcal{D}V \exp\Big(\tfrac{N\eta}{2}\operatorname{tr} CV\Lambda V^{\top}\Big)\,. \tag{18}$$

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 [26] take the form

$$\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \underset{N\to\infty}{\sim} \frac{N}{2}\sum_{k=1}^{K}\left\{\eta c_k(\lambda_k u_k^2 + \chi_k - \chi_k u_k^2) - F(\chi_k) - \ln \eta c_k\right\}, \tag{19}$$

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

$$\chi_k = \begin{cases}\lambda_k, \\ G^{-1}(\eta c_k),\end{cases} \qquad u_k^2 = \begin{cases}1 - g_k/(\eta c_k) & \text{if } g_k \le \eta c_k, \\ 0 & \text{if } g_k \ge \eta c_k.\end{cases} \tag{20}$$

Substituting the stationary values, (19) simplifies to

$$\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C) \underset{N\to\infty}{\sim} \frac{N}{2}\sum_{k=1}^{K}\left\{\eta c_k \chi_k - F(\chi_k) - \ln \eta c_k\right\}. \tag{21}$$

A.1.3. EVALUATED FORM OF Φ

Substituting the saddle values of $\ln Z(\Lambda)$ (15) and $\ln \mathcal{J}_{\mathrm{HCIZ}}(\Lambda; C)$ (19) into (8) and extremizing over χ_k gives $\Phi(C)$ as a closed-form function of the data spectrum. For a source C with eigenvalues $c_1, \ldots, c_K$ ($K = \operatorname{rank} C$, $\operatorname{tr} C = \sum_k c_k$),

$$\Phi(C) = \frac{K}{2}\ln(\gamma\eta) - \frac{\eta}{2}(\mu - F(\mu))\operatorname{tr} C + \sum_{k=1}^{K}\psi_k, \tag{22}$$

where the per-mode contribution ψ_k is

$$\psi_k = -\frac{\gamma\eta}{4}\lambda_k^2 + F(\lambda_k) + \frac{\eta c_k \chi_k - F(\chi_k) - \ln(\eta c_k)}{2} \tag{23}$$

$$= \begin{cases}\frac{\eta c_k^2}{4\gamma} - \frac{1}{2}\ln(\gamma\eta), & \text{uncondensed } (k > d), \\ -\frac{1}{2} - \frac{\gamma\eta}{4g_1^2} + \frac{c_k g_1}{2\gamma} + \frac{\eta c_k}{2g_1} - \frac{\ln(g_1\,\eta c_k)}{2}, & \text{condensed } (k \le d),\end{cases} \tag{24}$$

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 (8) with respect to the outlier eigenvalue positions. The optimal positions of $\lambda_1, \dots, \lambda_K$ must be stable under small perturbations $\lambda_1 + \epsilon_1, \dots, \lambda_K + \epsilon_K$.

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:

$$-\frac{\eta}{2} h^2 \epsilon_* \operatorname{tr} C + \sum_k \left\{ -\frac{\gamma\eta}{2}\lambda_k + \frac{\eta}{2} c_k u_k^2 + g_k \right\} \epsilon_k, \tag{25}$$

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

$$-\frac{\eta \operatorname{tr} C}{2} h^2 \epsilon_* + \sum_{k \leq d} \left\{ -\frac{\gamma\eta}{2g_1} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_1}{2} \right\} \epsilon_k + \sum_{k > d} \left\{ -\frac{\gamma\eta}{2g_k} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_k}{2} \right\} \epsilon_k \,. \tag{26}$$

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 $\epsilon_k \geq 0$, so we only require $-\frac{\gamma\eta}{2g_k} + \frac{1}{2}\eta c_k u_k^2 + \frac{g_k}{2} \leq 0$. Solving these conditions gives the positions of uncondensed eigenvalues (*i.e.* $\lambda_k < \mu$ for $k > d$):

$$\begin{aligned} \lambda_k &= \frac{2}{\sqrt{\gamma\eta}}, & g_k &= \sqrt{\gamma\eta}, & u_k^2 &= 0, & &\text{if } \eta c_k \leq \sqrt{\gamma\eta}, \\ \lambda_k &= \frac{1}{\eta c_k} + \frac{c_k}{\gamma}, & g_k &= \frac{\gamma}{c_k}, & u_k^2 &= 1 - \frac{\gamma}{\eta c_k^2}, & &\text{if } \eta c_k \geq \sqrt{\gamma\eta}. \end{aligned} \tag{27}$$

Now we focus on the condensed part, $k \leq d$. 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 $\epsilon_1, \dots, \epsilon_d$ are free. Then

$$-\frac{\eta \operatorname{tr} C}{2} h^2 \epsilon_* + \sum_{k \leq d} \left\{ -\frac{\gamma\eta}{2g_1} + \frac{\eta}{2} c_k u_k^2 + \frac{g_1}{2} \right\} \epsilon_k \tag{28}$$

simplifies to

$$-\eta \epsilon_* (1 - g_1) \operatorname{tr} C - \frac{\gamma\eta}{g_1} \sum_{k \leq d} \epsilon_k + \eta \sum_{k \leq d} c_k \epsilon_k. \tag{29}$$

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

$$-(1 - g_1) \operatorname{tr} C - \frac{d\gamma}{g_1} + \sum_{k \leq d} c_k = 0 \tag{30}$$

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

$$g_1 = \frac{1}{2\operatorname{tr} C}\left\{\sum_{k=d+1}^{K} c_k + \sqrt{\left(\sum_{k=d+1}^{K} c_k\right)^2 + 4d\gamma \operatorname{tr} C}\right\}. \tag{31}$$

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

$$\epsilon_* = \max\{\lambda_1 + \epsilon_1, \ldots, \lambda_K + \epsilon_K\} - \lambda_1, \tag{32}$$

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

$$\max_k\{\lambda_k + \epsilon_k\} = \lambda_j + \epsilon_j \leq \lambda_1 + \epsilon_j \leq \lambda_1 + \max_k \epsilon_k, \tag{33}$$

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

A.2.1. PHASE DIAGRAM

- $h = u_{1,\ldots,K} = 0$. 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

$$\lambda_k = \frac{1}{\eta c_k} + \frac{c_k}{\gamma}, \qquad u_k^2 = 1 - \frac{\gamma}{\eta c_k^2} \qquad (1 \leq k \leq a), \tag{34}$$

  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

$$\eta c_{d+1} < \sqrt{\gamma\eta} < \min\{1, \eta c_d\} \quad \text{and} \quad \sqrt{\gamma\eta}\,(\eta \operatorname{tr} C - d) + \eta \sum_{k=1}^{d} c_k < \eta \operatorname{tr} C, \tag{35}$$

  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,

$$u_k^2 = 1 - \frac{1}{c_k}\sqrt{\frac{\gamma}{\eta}} \quad (k \leq d), \tag{36}$$

  and $h \neq 0$.

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

$$\frac{\gamma}{c_d} < g_1 < \min\left\{1,\, \eta c_d,\, \frac{\gamma}{c_{d+1}},\, \sqrt{\gamma\eta}\right\}, \tag{37}$$

for $d \in [K]$ (with $c_{d+1} = 0$ when $d = K$), with g_1 given by Eq. (31), 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 $k \leq d$; the remaining $\lambda_{d+1}, \dots, \lambda_K$ sit at $\lambda_k = 1/(\eta c_k) + c_k/\gamma$ with $u_k^2 = 1 - \gamma/(\eta c_k^2)$.

Figure 3 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.

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

In this Appendix we establish the following two formulas:

$$\langle W \rangle_{W \sim P_\eta} = \frac{1}{N\eta} \sum_{k=1}^{K} \left(\eta\chi_k - \frac{1}{c_k}\right) \left(\mathbf{c}_k \mathbf{c}_k^\top - I\right). \tag{38}$$

and

$$\frac{K}{N} \left\langle \left\langle \mathbf{x}\mathbf{x}^\top \right\rangle_{\mathbf{x} \sim P_W} \right\rangle_{W \sim P_\eta} = C - \frac{\gamma}{N\eta} \sum_{k=1}^{K} \left(\eta\chi_k - \frac{1}{c_k}\right) \left(\mathbf{c}_k \mathbf{c}_k^\top - I\right) \tag{39}$$

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

$$c'_k = c_k + \frac{z}{N} \mathbf{c}_k^\top U \mathbf{c}_k + \mathcal{O}(z^2) \tag{40}$$

for $k \leq K$ (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:

$$\begin{aligned} \frac{1}{N} \frac{\partial}{\partial z} \ln \mathcal{Y}(C + zU) \bigg|_{z=0} &= \frac{1}{N} \frac{\partial \ln \mathcal{J}_{\text{HCIZ}}(\Lambda; C + zU)}{\partial z} \bigg|_{z=0} \\ &= \frac{1}{2N} \sum_{k=1}^{K} \left(\eta\chi_k - \frac{1}{c_k}\right) \mathbf{c}_k^\top U \mathbf{c}_k \end{aligned} \tag{41}$$

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.

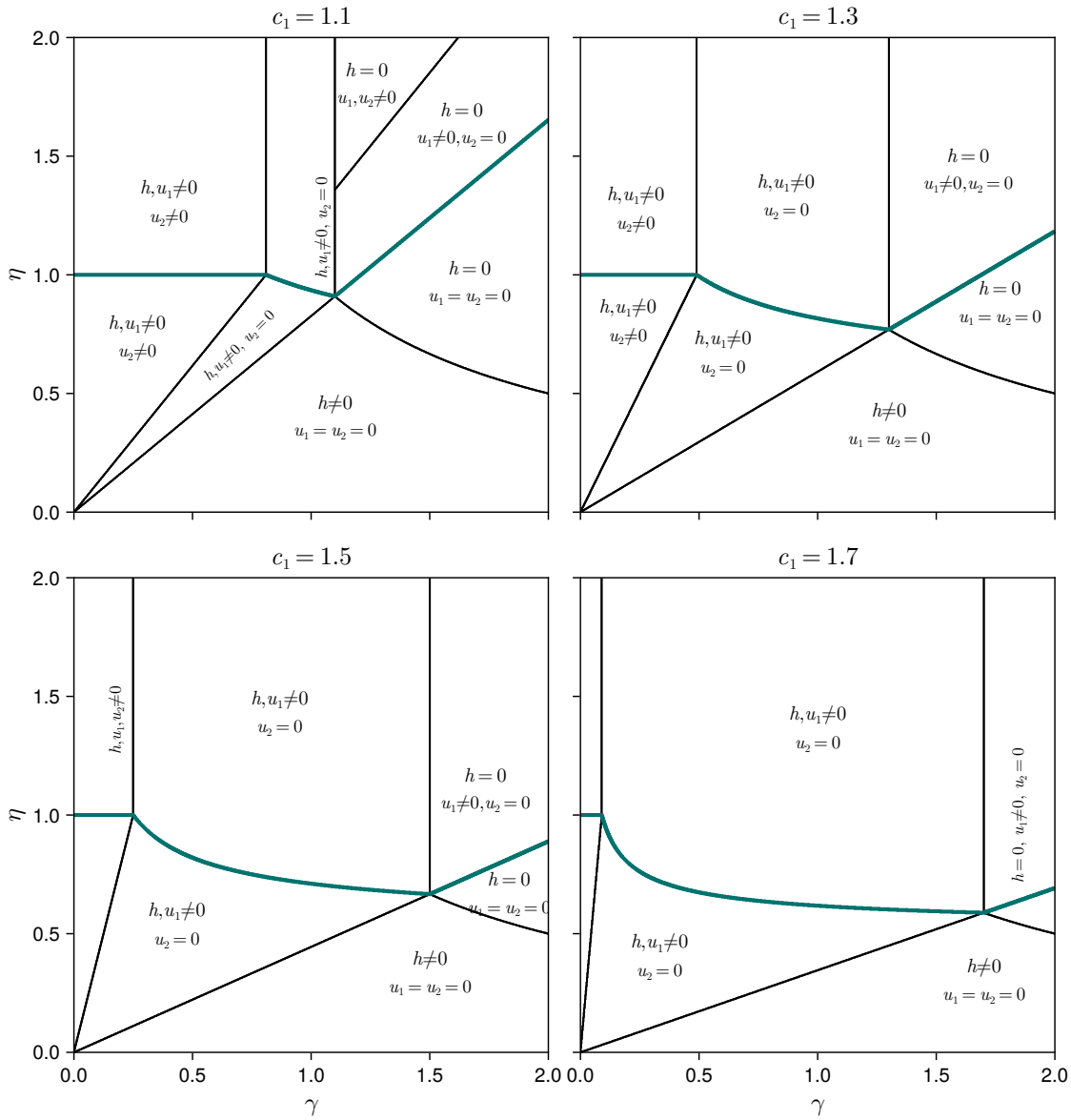

Figure 3: $K = 2$ **phase diagrams.** Phase diagram in the (γ, η) plane for $c_1 \in \{1.1, 1.3, 1.5, 1.7\}$ (with $c_2 = 2 - c_1$). Black lines are phase boundaries; the teal line separates outlier from edge phases.

A.3.2. WEIGHT STATISTICS

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

$$\frac{\partial}{\partial z} \ln \mathcal{Y}(C + zU)\Big|_{z=0} = \frac{N\eta}{2} \langle \operatorname{tr} UW \rangle_{W \sim P_\eta} . \tag{42}$$

Combined with (41),

$$\langle \operatorname{tr} UW \rangle_{W \sim P_\eta} = \frac{1}{N\eta} \sum_{k=1}^{K} \left(\eta \chi_k - \frac{1}{c_k} \right) \mathbf{c}_k^\top U \mathbf{c}_k \tag{43}$$

for any U with $\operatorname{tr} U = 0$. Since $\operatorname{tr} \langle W \rangle_{P_\eta} = 0$, this pins down the mean to Eq. (38).

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

$$\begin{aligned} \frac{\partial}{\partial z} \ln \int \exp \left\{ -K\eta \ln Z(W - zU) - \frac{N\gamma\eta}{4} \operatorname{tr} W^2 + \frac{N\eta}{2} \operatorname{tr} CW \right\} \mathrm{d}W \Big|_{z=0} \\ = \frac{K\eta}{2} \left\langle \left\langle \mathbf{x}^\top U \mathbf{x} \right\rangle_{P_W} \right\rangle_{W \sim P_\eta} . \end{aligned} \tag{44}$$

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 (41),

$$\begin{aligned} \frac{K\eta}{2} \left\langle \left\langle \mathbf{x}^\top U \mathbf{x} \right\rangle_{P_W} \right\rangle_{W \sim P_\eta} &= \frac{\partial}{\partial z} \ln \mathcal{Y}(C - z\gamma U)\Big|_{z=0} + \frac{N\eta}{2} \operatorname{tr} CU \\ &= -\frac{\gamma}{2} \sum_{k=1}^{K} \left(\eta \chi_k - \frac{1}{c_k} \right) \mathbf{c}_k^\top U \mathbf{c}_k + \frac{N\eta}{2} \operatorname{tr} CU. \end{aligned} \tag{45}$$

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. (39).

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 [58]. We summarize here how their approach relates to ours.

Effective theory. The derivation is based on the rewriting

$$\mathcal{Y}(C) = \lim_{n \to -K\eta} \langle Z(W)^n \rangle_{W \sim \exp(-\frac{N\gamma\eta}{4} \operatorname{tr} W^2 + \frac{N\eta}{2} \operatorname{tr}(CW))}, \tag{46}$$

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

$$\langle Z^n \rangle \propto \int \mathrm{d}M\, \mathrm{d}Q\, \exp\left(N\Phi_n^{\text{rep}}(M,Q)\right) \tag{47}$$

where

$$\Phi_{\text{rep}}(M,Q) = \frac{\eta \mathrm{tr}(C^2)}{4\gamma} + \frac{\mathrm{tr}(Q^2)}{4\gamma\eta} + \frac{1}{2\gamma}\mathrm{tr}(M\mathrm{diag}(c_k)M^\top) + \frac{1}{2}\mathrm{tr}\ln(Q - MM^\top) + \frac{n}{2}. \tag{48}$$

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

Saddle-point analysis. It was shown in [58] 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 $\mathrm{O}(n)$-invariants are related via

$$\mu = \frac{q_k}{\eta\gamma} + \frac{c_k}{\gamma} = \frac{q_k}{\eta\gamma} + \frac{1}{q_k - m_k^2} = \frac{\widetilde{q}}{\eta\gamma} + \frac{1}{\widetilde{q}}, \tag{49}$$

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,

$$\mathrm{tr}\, Q = n, \tag{50}$$

by $\mathrm{O}(n)$-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. (48)-(50) are valid provided $n > n_c$ where

$$n_c = -K\eta\, \frac{\sum_{k \leq d} \frac{c_k}{K}\left(1 - \frac{1}{c_k}\sqrt{\frac{\gamma}{\eta}}\right)}{1 - \sqrt{\eta\gamma}}. \tag{51}$$

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. (35). For $n < n_c$, the solution $\{m_k(n), \widetilde{q}(n)\}$ obtained for $n \geq n_c$ 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 [45, 58] for an interpretation of the freezing.

Eq. (50) is quadratic in $\widetilde{q}$. The solution, $\widetilde{q}(d)$, depends on d and equals the coalesced g_1, Eq. (31). The conditions determining d can be derived from the local stability of the d-coalesced saddle points [58], 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. (37), and $d = \#\{c_k : c_k > \sqrt{\gamma/\eta}\}$ for $n = -K\eta < n_c$, see Eqs. (35).

From Eqs. (49), 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

$$\frac{\eta c_k^2}{4\gamma} + \frac{q_k^2}{4\gamma\eta} + \frac{c_k m_k^2}{2\gamma} = \begin{cases} \frac{g_1^2}{4\gamma\eta} + \frac{\gamma\eta}{4g_1^2}, & k \le d, \\ \frac{\eta c_k^2}{4\gamma}, & k > d. \end{cases} \tag{52}$$

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

$$\Phi_{\text{rep}}(-K\eta) = \frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \frac{d\gamma\eta}{4g_1^2} - \frac{Kg_1^2}{4\gamma} + \frac{1}{2}\sum_{k=1}^{d} \ln\frac{\gamma}{c_k} - \frac{K\eta + d}{2}\ln g_1 - \frac{K\eta}{2}. \tag{53}$$

We turn to the total Φ predicted by Eq. (22). Collecting logarithms from the uncondensed and condensed modes,

$$\frac{K}{2}\ln(\gamma\eta) - \frac{K-d}{2}\ln(\gamma\eta) - \frac{K\eta}{2}\ln g_1 - \frac{1}{2}\sum_{k\le d}\ln(g_1\eta\, c_k) = \frac{1}{2}\sum_{k\le d}\ln\frac{\gamma}{c_k} - \frac{K\eta + d}{2}\ln g_1,$$

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 $\sum_{k\le d} c_k - K = -Kg_1 + d\gamma/g_1$. 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

$$\Phi'_{\text{rep}}(n_c) = \frac{1}{\sqrt{\gamma\eta}} + \frac{1}{4}\ln(\gamma\eta) - \frac{1}{4}. \tag{54}$$

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

$$\Phi_{\text{rep}} = \frac{\eta}{4\gamma}\sum_{k>d} c_k^2 + \sqrt{\frac{\eta}{\gamma}}\Big(\sum_{k\le d} c_k - K\Big) + \frac{K\eta}{4} - \frac{3d}{4} + \frac{1}{2}\sum_{k\le d}\ln\frac{\gamma}{c_k} - \frac{K\eta + d}{4}\ln(\gamma\eta), \tag{55}$$

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

We again turn to the Φ predicted by Eq. (22). 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 $-\frac{3d}{4} + \sqrt{\frac{\eta}{\gamma}}\sum_{k\le d} c_k - \frac{1}{2}\sum_{k\le d}\ln(\sqrt{\eta\gamma}\eta c_k)$, while $-K\eta(\ln Z)/N = -K\sqrt{\eta/\gamma} - \frac{K\eta}{4}\ln(\gamma\eta) + \frac{K\eta}{4}$. 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.

Appendix B. Training dynamics: Martin–Siggia–Rose derivation

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

$$Q(t,u) = \frac{\mathbf{x}(t)\cdot\mathbf{x}(u)}{N}, \qquad R(t,u) = \left.\frac{\delta\langle x_i(t)\rangle}{\delta j_i(u)}\right|_{j=0}, \tag{56}$$

with $j_i(t)$ a fictitious source added to the right-hand side of (2); the memory kernel and noise correlator are

$$\begin{aligned} M(t,u) &= e^{-\frac{\gamma}{2}(t-u)}\left[-\frac{K\nu}{2}Q(t,u) + \frac{\nu^2}{\eta\gamma}R(t,u)\right], \\ D(t,u) &= 2\nu\delta(t-u) + \frac{\nu^2}{\eta\gamma}e^{-\frac{\gamma}{2}|t-u|}Q(t,u). \end{aligned} \tag{57}$$

We sketch the derivation of the dynamical mean-field theory summarized in Sec. 2.2. The intermediate expressions (58)–(60) 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 (58) carries the K=1 data covariance at unit scalar eigenvalue. The general-K reduction to Eqs. (63)–(65) 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. 4 (including the $K = 2$ panels) runs with this multi-mode extension. Starting from the coupled Langevin equations (2)–(1), 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

$$W(t) = W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma}\frac{\mathbf{c}\mathbf{c}^\top}{N} - \frac{1}{2}\int_0^t e^{-\gamma(t-u)/2}\frac{\mathbf{x}(u)\mathbf{x}(u)^\top}{N}\,\mathrm{d}u. \tag{58}$$

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 (58) yields a closed generating functional over $(\mathbf{x}, \hat{\mathbf{x}})$ alone,

$$\mathcal{Z} \propto \int \mathcal{D}[\mathbf{x},\hat{\mathbf{x}}]\, e^{-N\mathcal{A}_{\mathrm{x}}[\mathbf{x},\hat{\mathbf{x}}]}, \tag{59}$$

with action

$$\begin{aligned} N\mathcal{A}_{\mathrm{x}} = \int \mathrm{d}t\, &\left\{\hat{\mathbf{x}}\cdot\dot{\mathbf{x}} + \kappa(t)\hat{\mathbf{x}}\cdot\mathbf{x} - \nu\|\hat{\mathbf{x}}\|^2 - \frac{\nu}{\gamma N}(1-e^{-\gamma t/2})\,\hat{\mathbf{x}}^\top\mathbf{c}\mathbf{c}^\top\mathbf{x}\right\} \\ &+ \frac{1}{N}\int \mathrm{d}t\,\mathrm{d}u\,\Theta(t-u)\,e^{-\gamma(t-u)/2}\left\{\tfrac{\nu}{2}(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u))\right. \\ &\left. - \tfrac{\nu^2}{\eta\gamma}(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u)) - \tfrac{\nu^2}{\eta\gamma}(\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u))(\mathbf{x}(t)\cdot\hat{\mathbf{x}}(u))\right\}, \end{aligned} \tag{60}$$

in which the Heaviside $\Theta(t-u)$ is inherited from the causal structure of (58). 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 (56) 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,

$$\mathcal{A}_{\mathrm{x}}[\mathbf{x}, \hat{\mathbf{x}}] = \frac{1}{N} \sum_i \mathfrak{a}[x_i, \hat{x}_i | c_i], \tag{61}$$

from which the effective single-site Gaussian process reads

$$\dot{x}(t) = -\kappa(t) x(t) + \frac{\nu}{\gamma}(1 - e^{-\gamma t/2}) s(t)\, c + \int_0^t M(t,u)\, x(u)\, \mathrm{d}u + \zeta(t), \tag{62}$$

with the kernels $M(t,u)$ and $D(t,u)$ of (57) and $\langle \zeta(t)\zeta(u)\rangle = D(t,u)$. Because (62) 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

$$\partial_t s_k(t) = -\kappa(t) s_k(t) + \tfrac{\nu}{\gamma}(1 - e^{-\gamma t/2}) c_k s_k(t) + \int_0^t M(t,u) s_k(u)\, \mathrm{d}u, \tag{63}$$

$$\partial_t R(t,t') = -\kappa(t) R(t,t') + \int_{t'}^t M(t,u) R(u,t')\, \mathrm{d}u, \tag{64}$$

$$\begin{aligned} \partial_t Q(t,t') = {} & \tfrac{\nu}{\gamma}(1 - e^{-\gamma t/2}) \sum_{k=1}^K c_k s_k(t) s_k(t') - \kappa(t) Q(t,t') \\ & + \int_0^t M(t,u) Q(u,t')\, \mathrm{d}u + \int_0^{t'} D(t,u) R(t',u)\, \mathrm{d}u, \end{aligned} \tag{65}$$

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 $\kappa(t)$.

Early-time spike growth. Early in training the negative-phase contribution in (58) is weak ($s_k^2(t) \ll c_k/\gamma$), and the spike along data direction $\mathbf{c}_k$ grows deterministically. The k-th outlier eigenvalue is then placed at

$$\lambda_k(t) = \frac{1}{g_k(t)} + \frac{g_k(t)}{\gamma\eta}, \qquad g_k(t) = \frac{\gamma}{c_k\,(1 - e^{-\gamma t/2})}, \qquad t_{\mathrm{out},k} = -\frac{2}{\gamma} \ln\!\left(1 - \frac{1}{c_k}\sqrt{\tfrac{\gamma}{\eta}}\right), \tag{66}$$

valid for $t \geq t_{\mathrm{out},k}$ and defined when $c_k > \sqrt{\gamma/\eta}$ (otherwise λ_k stays at the bulk edge 2σ). $g_k(t) = G(\lambda_k(t))$ is the Stieltjes transform (7) evaluated at the k-th outlier position.

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

$$M_{\mathrm{st}}(\tau) = e^{-\gamma\tau/2} \left\{ -\frac{K\nu}{2} Q_{\mathrm{st}}(\tau) + \frac{\nu^2}{\eta\gamma} R_{\mathrm{st}}(\tau) \right\}, \quad D_{\mathrm{st}}(\tau) = 2\nu\delta(\tau) + \frac{\nu^2}{\eta\gamma} e^{-\gamma|\tau|/2} Q_{\mathrm{st}}(\tau). \tag{67}$$

The stationary equations become,

$$0 = \left(-\kappa_{\rm st} + \frac{\nu c_a}{\gamma} + \int_0^\infty M_{\rm st}(\tau)\mathrm{d}\tau\right) s_{a,\rm st}, \qquad a = 1, \dots, K, \tag{68}$$

$$\partial_\tau R_{\rm st}(\tau) = -\kappa_{\rm st} R_{\rm st}(\tau) + \int_0^\tau M_{\rm st}(\tau-\sigma) R_{\rm st}(\sigma)\mathrm{d}\sigma, \tag{69}$$

$$\begin{aligned}\partial_\tau Q_{\rm st}(\tau) = \frac{\nu}{\gamma}\sum_{a=1}^K c_a s_{a,\rm st}^2 - \kappa_{\rm st} Q_{\rm st}(\tau) + \int_0^\tau M_{\rm st}(\tau-\sigma) Q_{\rm st}(\sigma)\mathrm{d}\sigma \\ + \int_0^\infty M_{\rm st}(\tau+\sigma) Q_{\rm st}(\sigma)\mathrm{d}\sigma + \int_0^\infty D_{\rm st}(\tau+\sigma) R_{\rm st}(\sigma)\mathrm{d}\sigma\end{aligned} \tag{70}$$

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

A condensed solution may also exist, satisfying $\kappa_{\rm st} = \nu c_a/\gamma + \int_0^\infty M_{\rm st}(\tau)\mathrm{d}\tau$. 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

$$z + \kappa_P - \check{M}_P(z) = \frac{\nu c_a}{\gamma}. \tag{71}$$

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

$$\check{R}_P(z) = \frac{1}{z + \kappa_P - \check{M}_P(z)}. \tag{72}$$

Substituting (72) into (71) yields the spectral equation

$$1 = \frac{\nu c_a}{\gamma}\int_0^\infty e^{-z\tau} R_P(\tau)\,\mathrm{d}\tau, \tag{73}$$

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 $\nu_c = \gamma/(c_1\chi_P)$.

Numerically, the two-time system (63)–(65) 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 $\kappa(t_n)$.

Numerical validation against finite-N SGD training. Figure 4 compares the DMFT solution of Eqs. (63)–(65) with finite-N SBM training simulated by Euler–Maruyama integration of the coupled Langevin system (2)–(1), in the mini-batch-size-one, persistent-MCMC regime, with the GOE noise $\Omega/\sqrt{\eta N}$ playing the role of the stochastic-gradient noise. This validates the large-N DMFT (63)–(65) 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

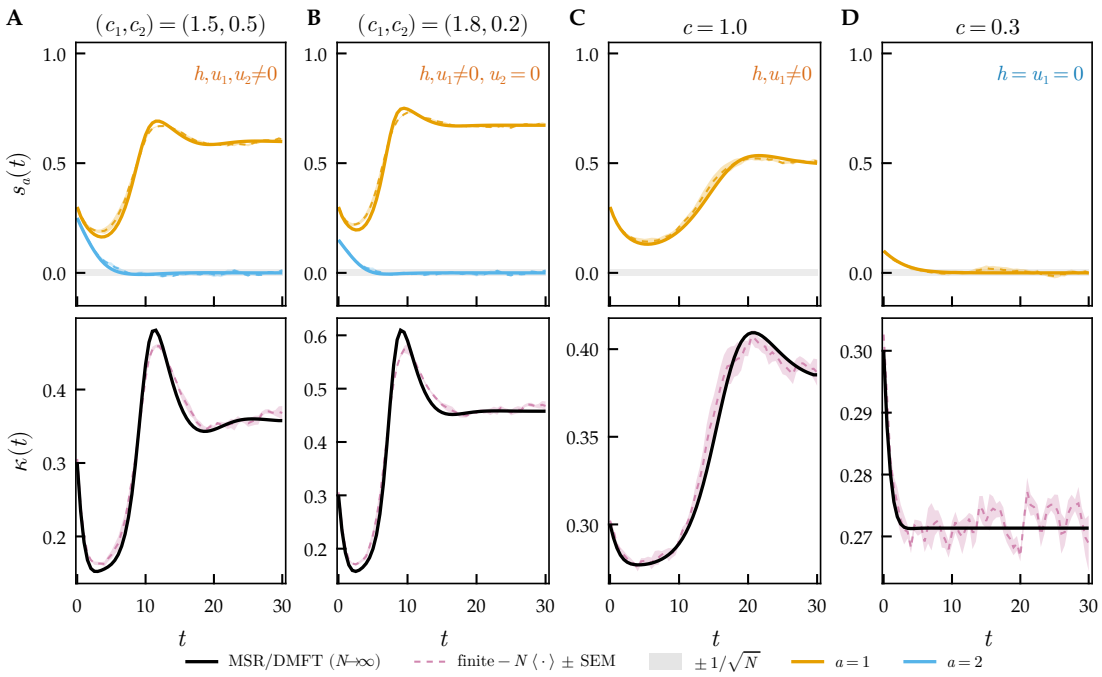

Figure 4: **Validation of the large-N DMFT against finite-N SBM training.** Each panel compares DMFT (solid) with finite-N SGD runs (persistent MCMC, mini-batch size $= 1$) simulated by Euler integration of the coupled Langevin system (2)–(1) at $N = 4000$, averaged over five seeds (dashed, with $\pm$SEM 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 (2)–(1).

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.

B.1. Upper bound on the outlier count

The decomposition (58) implies that $W(t)$ carries at most K upper outliers at every fixed time. The negative-phase integral

$$\int_0^t e^{-\gamma(t-u)/2} \frac{\mathbf{x}(u)\mathbf{x}(u)^\top}{N} \mathrm{d}u \ \succeq\ 0 \tag{74}$$

is positive semidefinite, so

$$W(t) \;\preceq\; W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma}\, C. \tag{75}$$

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

$$\lambda_{K+j}\left(W_{\mathrm{GOE}}(t) + \frac{1-e^{-\gamma t/2}}{\gamma} C\right) \le \lambda_j\left(W_{\mathrm{GOE}}(t)\right), \qquad (j \ge 1), \tag{76}$$

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,

$$\lambda_{K+1}(W(t)) \le \frac{2}{\sqrt{\gamma\eta}} + \varepsilon. \tag{77}$$

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 (57): $M(t,u) \to -\frac{\nu}{2} e^{-\gamma(t-u)/2} Q(t,u)$ 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

$$\tfrac{\gamma}{2}\left[\hat{Q}_P(\gamma/2) - \hat{Q}_P^2(\gamma/2)\right] \;=\; 1-\gamma. \tag{78}$$

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 5 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 $\nu_c \Delta\tau < 0.6$.

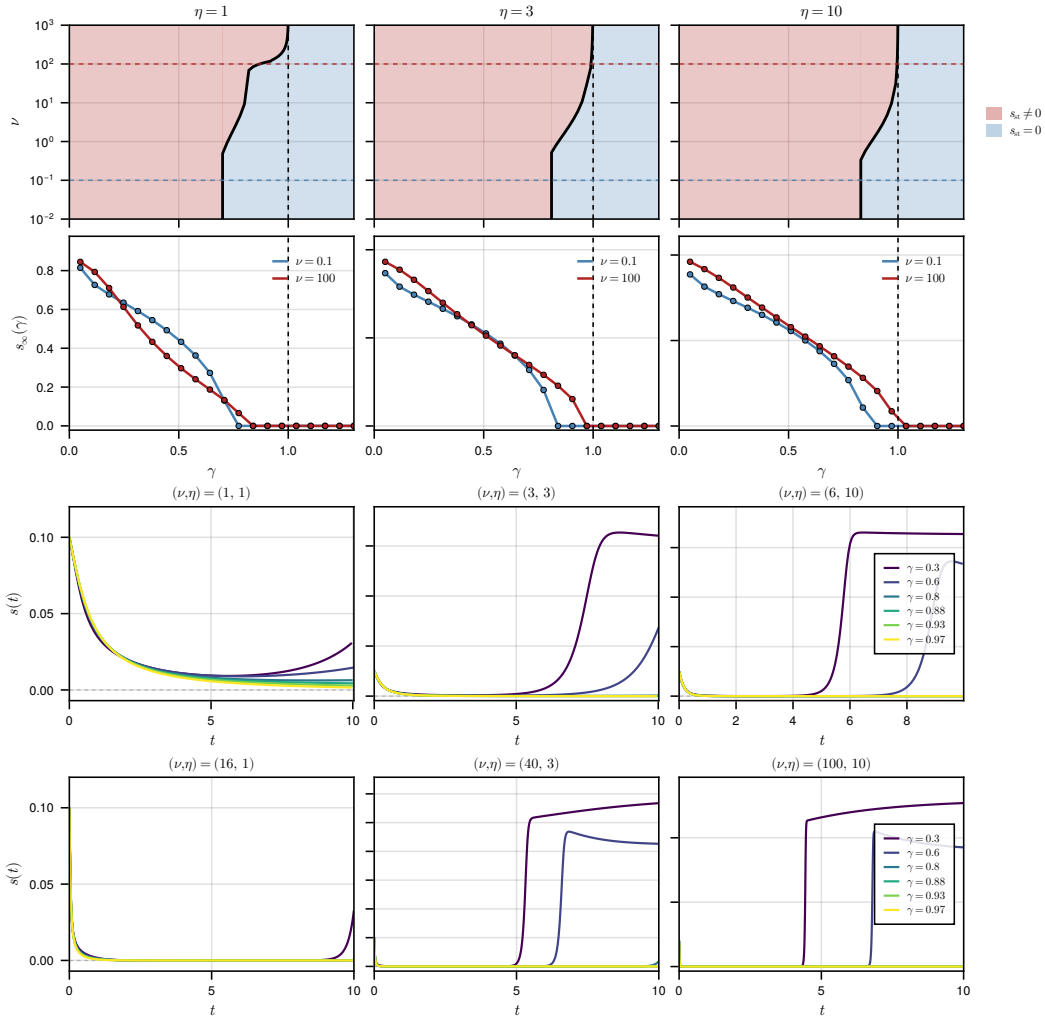

Figure 5: **Dynamical phase structure of the $K{=}1$ DMFT at finite η.** Three values $\eta \in \{1, 3, 10\}$ are shown in columns. **Row 1**: dynamical phase diagram $\nu_c(\gamma)$ (solid black), obtained by warm-started path continuation on a fine τ-grid ($\Delta\tau \simeq 4 \times 10^{-3}$) with 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.

Appendix 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 $o(K)$. The relative scaling of N and K, it turns out, is irrelevant. 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 [58]. 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

$$\gamma > \gamma_{\text{bulk}} \quad \text{where} \quad \gamma_{\text{bulk}} = c_{K'+1}\left(\delta_K + \frac{K'}{K} c_{K'+1}\right), \quad \text{with} \quad \delta_K = 1 - \frac{1}{K}\sum_{\ell=1}^{K'} c_\ell \,. \tag{79}$$

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 γ/δ_K^2 for thick slabs, $\delta_K \gg K^{-1/2}$.

C.2. Dynamics

C.2.1. BARE LARGE-K REGIME

We consider the regime

$$t, \gamma, \eta, \nu = \mathcal{O}(1). \tag{80}$$

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

$$\widetilde{M}(t,u) \equiv \frac{M(t,u)}{K} = -\frac{\nu}{2} e^{-\frac{\gamma}{2}(t-u)} Q(t,u) + \mathcal{O}(K^{-1}), \qquad t > u, \tag{81}$$

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

$$\widetilde{\kappa}(t) \equiv \frac{\kappa(t)}{K\nu} = \frac{1 - e^{-\gamma t/2}}{\gamma} \sum_{\ell=1}^{K'} \widetilde{c}_\ell s_\ell(t)^2 - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u)^2 + \mathcal{O}(K^{-1}). \tag{82}$$

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

$$\begin{cases} 0 = -\widetilde{\kappa}(t) s_\ell(t) + (1 - e^{-\gamma t/2})\frac{\widetilde{c}_\ell}{\gamma} s_\ell(t) - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) s_\ell(u) + o(1), \\ 0 = -\widetilde{\kappa}(t) s_k(t) - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) s_k(u) + o(1). \end{cases} \tag{83}$$

The correlation function also behaves quasistatically on finite timescales,

$$0 = -\widetilde{\kappa}(t)Q(t,t') + (1-e^{-\gamma t/2})\sum_{\ell}\frac{\widetilde{c}_\ell}{\gamma}s_\ell(t)s_\ell(t') - \frac{1}{2}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}Q(t,u)Q(u,t'), \quad (84)$$

while the response function obeys

$$\partial_t R(t,t') = -K\nu\widetilde{\kappa}(t)R(t,t') - \frac{K\nu}{2}\int_{t'}^t \mathrm{d}u\, e^{-\gamma(t-u)/2}Q(t,u)R(u,t') + O(1), \quad t > t', \quad (85)$$

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 $Q(t,t'), R(t,t') \to Q_{\mathrm{st}}(\tau), R_{\mathrm{st}}(\tau)$, $s_\ell(t), s_k(t), \kappa(t) \to s_{\ell,\mathrm{st}}, s_{k,\mathrm{st}}, \kappa_{\mathrm{st}}$.

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

$$0 = \left[\widetilde{\kappa}_{\mathrm{st}} + \frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2}Q_{\mathrm{st}}(\tau) - \frac{\widetilde{c}_\ell}{\gamma}\right]s_{\ell,\mathrm{st}}, \qquad \ell = 1,\ldots,K'. \quad (86)$$

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

The stationary response function satisfies

$$\partial_\tau R_{\mathrm{st}}(\tau) = -K\nu\left[\widetilde{\kappa}_{\mathrm{st}}R_{\mathrm{st}}(\tau) + \frac{1}{2}\int_0^\tau \mathrm{d}\sigma\, e^{-\gamma\sigma/2}Q_{\mathrm{st}}(\sigma)R_{\mathrm{st}}(\tau-\sigma)\right] \quad (87)$$

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

$$\check{R}_{\mathrm{st}}(p) = \frac{1}{p + K\nu(\widetilde{\kappa}_{\mathrm{st}} + \frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-(p+\gamma/2)\tau}Q_{\mathrm{st}}(\tau))}. \quad (88)$$

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

The equation for the stationary correlation function is

$$0 = -\widetilde{\kappa}_{\mathrm{st}}Q_{\mathrm{st}}(\tau) + \frac{1}{\gamma}\sum_{\ell}\widetilde{c}_\ell s_{\ell,\mathrm{st}}^2 - \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2}Q_{\mathrm{st}}(\sigma)Q_{\mathrm{st}}(\tau-\sigma). \quad (89)$$

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

$$\frac{1}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2}C(\tau)^2 = \frac{1}{\gamma}(1-q)(q-\widetilde{c}) + \frac{1-2q}{2}\int_0^\infty \mathrm{d}\tau\, e^{-\gamma\tau/2}C(\tau). \quad (90)$$

From Eq. (62), fluctuations obey the linear equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta\mathbf{x} = -K\nu\left[\widetilde{\kappa}_{\mathrm{st}}\delta\mathbf{x}(t) + \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2}Q_{\mathrm{st}}(\sigma)\delta\mathbf{x}(t-\sigma)\right] + \boldsymbol{\zeta}(t). \quad (91)$$

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 $\frac{\mathrm{d}}{\mathrm{d}t}\delta\mathbf{x} \approx -K\nu[\widetilde{\kappa}_{\text{st}} + \frac{1}{2}\int_0^\infty \mathrm{d}\sigma\, e^{-\gamma\sigma/2} Q_{\text{st}}(\sigma)]\delta\mathbf{x}(t)$. 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. (90) gives $q \geq \widetilde{c}_{\max}$ since $q \leq 1$ and LHS > 0, while Eq. (86) gives $\gamma\widetilde{\kappa}_{\text{st}} = \widetilde{c}_{\max} - q$, whence $q \leq \widetilde{c}_{\max}$. As a result, $q = \widetilde{c}_{\max}$ and $\Delta = \widetilde{\kappa}_{\text{st}} = 0$. We conclude that $Q_{\text{st}}(\tau) = q$ with $\tau > 0$ for $K \to \infty$. Contrast this with the finite-K case, where we expect $Q_{\text{st}}(\tau)$ to plateau only when $\tau \to \infty$.

The non-equilibrium stationary state is thus characterized by

$$q = \sum_\ell \delta_{c_\ell, c_{\max}} s_\ell^2 = \min\{1, \widetilde{c}_{\max}\}, \qquad \widetilde{\kappa}_{\text{st}} = \frac{(\widetilde{c}_{\max} - 1)_+}{\gamma}, \tag{92}$$

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. (58). As a result, at long times, the top mode of W is not the sampled direction, but the largest extensive unsampled direction.

Validity of 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. Inspection of Eq. (60) shows that the relative scaling of K and N is irrelevant, as long $N, K \to \infty$. After power counting, $N\mathcal{A}_{\mathbf{x}} = NK\mathcal{A}_{\text{eff}} + o(NK)$ with

$$\begin{aligned}\mathcal{A}_{\text{eff}} = \int \mathrm{d}t\, \widetilde{\kappa}(t) \frac{\hat{\mathbf{x}}(t)\cdot\mathbf{x}(t)}{N} - \int \mathrm{d}t\, (1 - e^{-\gamma t/2}) \sum_{\ell \leq K'} \frac{\widetilde{c}_\ell}{\gamma} \frac{\mathbf{x}(t)\cdot\mathbf{c}_\ell}{N} \frac{\hat{\mathbf{x}}(t)\cdot\mathbf{c}_\ell}{N} \\ + \frac{1}{2}\int_{t>u} \mathrm{d}u\, e^{-\gamma(t-u)/2} Q(t,u) \frac{\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u)}{N}.\end{aligned} \tag{93}$$

The correction consists of the kinetic term, the spin noise, the weight noise, and the bulk modes

$$o(NK) = \begin{cases} \int \mathrm{d}t\, \hat{\mathbf{x}}(t)\cdot\mathbf{x}(t), \\ -\nu \int \mathrm{d}t\, \kappa(t) \hat{\mathbf{x}}(t)\cdot\mathbf{x}(t), \\ \frac{K\nu}{2N} \int_{t>u} \mathrm{d}t\, \mathrm{d}u\, e^{-\gamma(t-u)/2} \left[(\mathbf{x}(t)\cdot\mathbf{x}(u))(\hat{\mathbf{x}}(t)\cdot\hat{\mathbf{x}}(u)) + (\hat{\mathbf{x}}(t)\cdot\mathbf{x}(u))(\mathbf{x}(t)\cdot\hat{\mathbf{x}}(u))\right], \\ -\nu \int \mathrm{d}t\, (1 - e^{-\gamma t/2}) \sum_{k>D} \frac{c_k}{\gamma} (\mathbf{x}(t)\cdot\mathbf{c}_k)(\hat{\mathbf{x}}(t)\cdot\mathbf{c}_k)). \end{cases}$$

The leading action $\mathcal{A}_{\text{eff}}$ reproduces the large-K equations already analyzed.

C.2.2. INVARIANT LARGE-K REGIME

We now show that the DMFT (63)-(57) is left invariant under the rescaling

$$\bar{t} = \frac{K}{K'} t, \quad \gamma = \frac{K}{K'}\overline{\gamma}, \quad \eta = \frac{K'}{K}\overline{\eta}, \quad \nu = \frac{K}{K'}\overline{\nu}, \qquad \text{with} \qquad \bar{t}, \overline{\gamma}, \overline{\eta}, \overline{\nu} = \mathcal{O}(1), \tag{94}$$

and describes the large-K Langevin training dynamics Eq. (1)-(2) on nearly finite-K' training data.

We introduce

$$\overline{c}_{1,\ldots,K'} = K'\widetilde{c}_{1,\ldots,K'} \Big(= \frac{K'}{K} c_{1,\ldots,K'}\Big), \qquad \overline{C} = \frac{K'}{K} C, \tag{95}$$

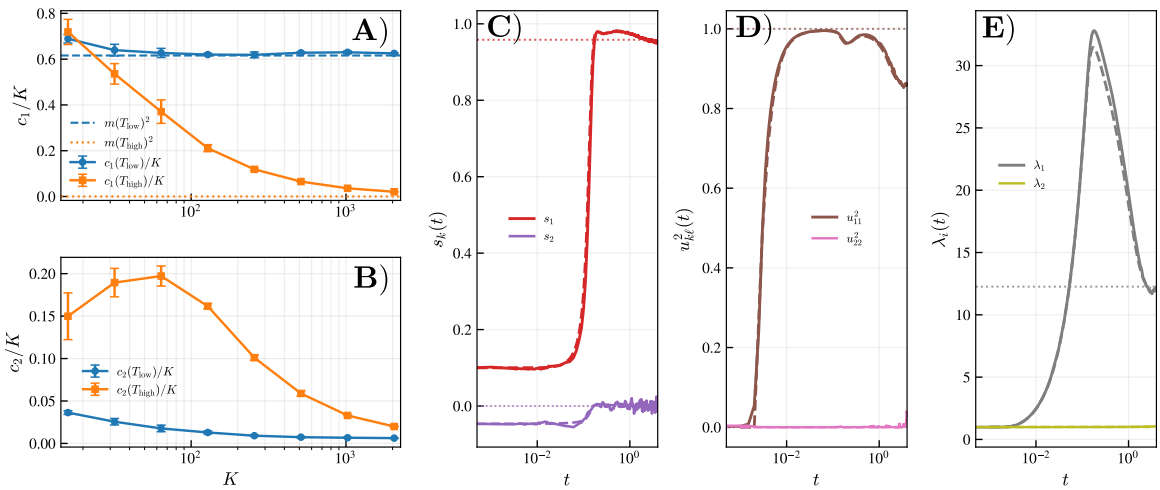

Figure 6: **Validation of the large-K DMFT equations for nearly finite-dimensional Ising data. A, B)** Spectrum of the empirical covariance matrix C for K configurations of the two-dimensional periodic Ising model at fixed linear size $L = 32$, $N = L \times L (= 1024)$. Below the critical temperature, $T_{\rm low} = 2.2 < T_c \simeq 2.269$, the leading eigenvalue is extensive, $c_1 \simeq K m(T_{\rm low})^2$, while c_2/K decreases with K, consistent with effectively $K' = 1$ data. Above the transition, $T_{\rm high} = 3.2 > T_c$, 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.

together with the scaled-time dynamical variables

$$\overline{W}(\bar{t}) = W(t), \qquad \overline{\mathbf{x}}(\bar{t}) = \mathbf{x}(t), \qquad \overline{\kappa}(\bar{t}) = \frac{K'}{K}\kappa(t). \tag{96}$$

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

$$\frac{\mathrm{d}\overline{W}}{\mathrm{d}\bar{t}} = \frac{1}{2}\left(\overline{C} - \frac{K'}{N}\overline{\mathbf{x}}(\bar{t})\overline{\mathbf{x}}(\bar{t})^\top\right) - \frac{\overline{\gamma}}{2}\overline{W}(\bar{t}) + \frac{\overline{\Omega}(\bar{t})}{\sqrt{N\overline{\eta}}}, \tag{97}$$

$$\frac{\mathrm{d}\overline{\mathbf{x}}}{\mathrm{d}\bar{t}} = \overline{\nu}\overline{W}(\bar{t})\overline{\mathbf{x}}(\bar{t}) - \overline{\kappa}(t)\overline{\mathbf{x}}(\bar{t}) + \sqrt{2\overline{\nu}}\,\overline{\boldsymbol{\xi}}(\bar{t}), \tag{98}$$

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,

$$\frac{\mathrm{d}\bar{s}_k}{\mathrm{d}\bar{t}} = -\overline{\mu}(\bar{t})\,\bar{s}_k(\bar{t}) + \frac{\overline{\nu}}{\overline{\gamma}}\left(1 - e^{-\overline{\gamma}\,\bar{t}/2}\right)\bar{c}_k\,\bar{s}_k(\bar{t}) + \int_0^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\bar{s}_k(\overline{u})\,\mathrm{d}\overline{u}, \tag{99}$$

$$\partial_{\bar{t}}\overline{R}(\bar{t},\bar{t}') = -\overline{\mu}(\bar{t})\,\overline{R}(\bar{t},\bar{t}') + \int_{\bar{t}'}^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\overline{R}(\overline{u},\bar{t}')\,\mathrm{d}\overline{u}, \tag{100}$$

$$\partial_{\bar{t}}\overline{Q}(\bar{t},\bar{t}') = \frac{\overline{\nu}}{\overline{\gamma}}\left(1 - e^{-\overline{\gamma}\,\bar{t}/2}\right)\sum_{k=1}^{K}\bar{c}_k\,\bar{s}_k(\bar{t})\,\bar{s}_k(\bar{t}') - \overline{\mu}(\bar{t})\,\overline{Q}(\bar{t},\bar{t}') \tag{101}$$

$$+ \int_0^{\bar{t}} \overline{M}(\bar{t},\overline{u})\,\overline{Q}(\overline{u},\bar{t}')\,\mathrm{d}\overline{u} + \int_0^{\bar{t}'} \overline{D}(\bar{t},\overline{u})\,\overline{R}(\bar{t}',\overline{u})\,\mathrm{d}\overline{u}. \tag{102}$$

where

$$\overline{M}(\bar{t},\overline{u}) = \frac{K'^2}{K^2}M(t,t') = e^{-\overline{\gamma}(\bar{t}-\overline{u})/2}\left\{\frac{\overline{\nu}^2}{\overline{\eta}\,\overline{\gamma}}\,\overline{R}(\bar{t},\overline{u}) - \frac{K'\overline{\nu}}{2}\,\overline{Q}(\bar{t},\overline{u})\right\}, \tag{103}$$

$$\overline{D}(\bar{t},\overline{u}) = \frac{K'^2}{K^2}D(t,t') = 2\overline{\nu}\,\delta(\bar{t}-\overline{u}) + \frac{\overline{\nu}^2}{\overline{\eta}\,\overline{\gamma}}\,e^{-\overline{\gamma}|\bar{t}-\overline{u}|/2}\overline{Q}(\bar{t},\overline{u}), \tag{104}$$

and we introduced the scaled-time dynamical order parameters

$$\bar{s}_k(\bar{t}) = s_k(t), \qquad \overline{Q}(\bar{t},\bar{t}') = Q(t,t'), \qquad \overline{R}(\bar{t},\bar{t}') = R(t,t'). \tag{105}$$

To complete the reduction, we show that

$$\sum_{k=1}^{K}\bar{c}_k\,\bar{s}_k(\bar{t})\,\bar{s}_k(\bar{t}') \approx \sum_{k=1}^{K'}\bar{c}_k\,\bar{s}_k(\bar{t})\,\bar{s}_k(\bar{t}'). \tag{106}$$

Let

$$\Delta(\bar{t},\bar{t}') = \frac{1}{N}\overline{\mathbf{x}}(\bar{t})^\top\overline{B}\overline{\mathbf{x}}(\bar{t}'), \qquad \text{with} \qquad \overline{B} = \frac{1}{N}\sum_{a=K'+1}^{K}\bar{c}_a\,\mathbf{c}_a\mathbf{c}_a^\top, \tag{107}$$

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

$$|\Delta(\bar{t},\bar{t}')| \le \|\overline{B}\|_{\mathrm{op}}\sqrt{\frac{\|\overline{\mathbf{x}}(\bar{t})\|^2}{N}\frac{\|\overline{\mathbf{x}}(\bar{t}')\|^2}{N}} = \|\overline{B}\|_{\mathrm{op}} = \bar{c}_{K'+1} = o(K). \tag{108}$$

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. Fig. 7 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 ($K' = 1$).

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 (103) 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. (99) gives the bare large-K quasistatic signal equation, Eq. (83).

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. (84).

Applying similar manipulations turns Eq. (100) into Eq. (85), thereby concluding the reduction.

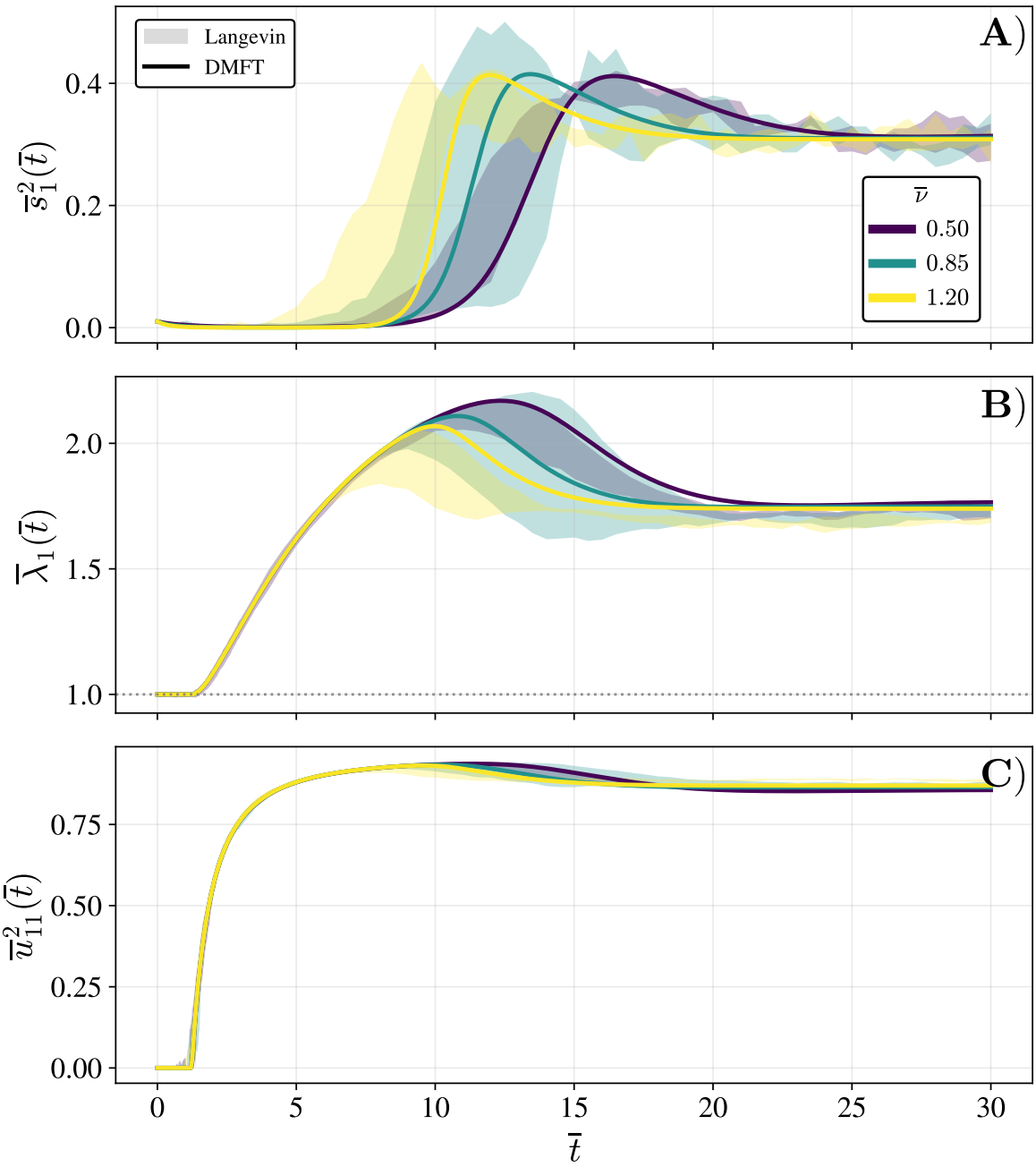

Figure 7: **Large-K training dynamics for nearly finite-dimensional Ising data.** Comparison between the effective $K' = 1$ large-K DMFT prediction and finite-N coupled Langevin simulations for below-critical two-dimensional Ising data at $T = 1.8$, with $K = 512$ and $N = 2025$. Time is shown in the rescaled variable $\bar{t} = Kt/K'$, and colors indicate the rescaled persistent-chain update rate $\bar{\nu}$. 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 $\bar{s}_1^2(\bar{t})$, showing condensation of the persistent chain onto the leading Ising covariance mode. **B)** Top eigenvalue $\bar{\lambda}_1(\bar{t})$ of the learned weight matrix, with the dotted line marking the bulk edge. **C)** Squared eigenvector overlap $\bar{u}_{11}^2(\bar{t})$, showing alignment of the top eigenvector with the leading data mode. Increasing $\bar{\nu}$ accelerates the onset of the negative phase, reducing the eigenvalue overshoot while leaving the learned direction largely unchanged.

Appendix 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:

$$c_1 = 2 - \frac{1}{\omega^*}, \qquad c_2 = \frac{1}{\omega^*}. \tag{109}$$

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

$$\frac{(\mathbf{c}_1 \cdot \mathbf{w}^*)^2}{N} = 1 - \frac{1}{2\omega^* - 1}, \qquad \frac{(\mathbf{c}_2^\top \mathbf{w}^*)^2}{N} = 0 \tag{110}$$

to leading order in N. The teacher partition function and entropy read:

$$\frac{1}{N} \ln Z(W^*) = \frac{1}{2}\ln(2\pi e) + \frac{\omega^* - 1 - \ln \omega^*}{2}, \qquad \frac{1}{N}\mathcal{H}[P^*] = \frac{1}{2}\ln(2\pi e) - \frac{1}{2}\ln \omega^* \tag{111}$$

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:

$$\begin{aligned} P_{\rm pp}(\mathbf{x}|C) &= \langle P(\mathbf{x}|W)\rangle_{W\sim P_\eta} = \int P(\mathbf{x}|W) P_\eta(W|C) \mathrm{d}W \\ &= \frac{1}{\mathcal{Y}(C)} \int \exp\left\{ \frac{N}{2} \operatorname{tr}\left[\left(\eta C + \frac{\mathbf{x}\mathbf{x}^\top}{N}\right) W\right] - (K\eta + 1)\ln Z(W) - \frac{N}{4}\gamma\eta \operatorname{tr} W^2 \right\} \mathrm{d}W \end{aligned} \tag{112}$$

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

$$P_{\rm pp}(\mathbf{x}|C) = \frac{\mathcal{Y}(C_{\mathbf{x}})}{\mathcal{Y}(C)}, \qquad C_{\mathbf{x}} = C + \frac{1}{\eta}\frac{\mathbf{x}\mathbf{x}^\top}{N}. \tag{113}$$

This identity, noticed in [58], 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

$P_{\mathrm{pp},\eta}(\mathbf{x}|C)$ 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,

$$\lim_{\eta\to 0^+} \frac{1}{N} \log P_{\mathrm{pp},\eta}(\mathbf{x}|C) = -\frac{1}{2}\log(2\pi e) + o_N(1). \tag{114}$$

$P_{\mathrm{pp},\eta}(\mathbf{x}|C)$ 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

$$\eta \operatorname{tr} CW = \eta^{1/2}\operatorname{tr}\widetilde{W}C = o(1) \tag{115}$$

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

$$N\lambda_{\min}(W) \le \mathbf{x}^\top W\mathbf{x} \le N\lambda_{\max}(W), \tag{116}$$

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

$$0 < Z(0)e^{\frac{N}{2}\lambda_{\min}} \le Z(W) \le Z(0)e^{\frac{N}{2}\lambda_{\max}}. \tag{117}$$

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

$$|\ln Z(W) - \ln Z(0)| \le \frac{N}{2}\|W\|_{\mathrm{op}}. \tag{118}$$

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

$$\frac{\eta}{N}\log Z(\eta^{-1/2}\widetilde{W}) = \mathcal{O}(\eta^{1/2}). \tag{119}$$

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 $P_W(O\mathbf{x}) = P_{O^\top WO}(\mathbf{x})$ 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

$$P_{\mathrm{pp},0^+}(\mathbf{x}|C) = \frac{1}{Z(0)}. \tag{120}$$

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

$$\lim_{\eta\to+\infty} P_{\mathrm{pp},\eta}(\mathbf{x}|C) = P_{W_{\mathrm{MAP}}}(\mathbf{x}). \tag{121}$$

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

$$\gamma W_{\mathrm{MAP}} = C - \frac{K}{N}\langle \mathbf{x}\mathbf{x}^\top\rangle_{P_{W_{\mathrm{MAP}}}}. \tag{122}$$

D.4. Posterior-predictive entropy

Applying (113) to the posterior-predictive entropy gives

$$\mathcal{H}[P_{\rm pp}] = \ln \mathcal{Y}(C) - \langle \ln \mathcal{Y}(C_{\mathbf{x}}) \rangle_{\mathbf{x}\sim P_{\rm pp}} . \tag{123}$$

Under $\mathbf{x} \sim P_{\rm pp}$, we have the overlaps (by (39))

$$m_k^2 = \frac{(\mathbf{c}_k \cdot \mathbf{x})^2}{N^2} = \frac{c_k}{2} - \frac{\gamma}{2}\left(\chi_k - \frac{1}{\eta c_k}\right), \quad k = 1, 2, \tag{124}$$

The three nonzero eigenvalues $\tilde{c}_{1,2,3}$ of $C_{\mathbf{x}}$ (113) are the roots of

$$m_1^2 \tilde{c}(\tilde{c} - c_2) + m_2^2 \tilde{c}(\tilde{c} - c_1) + (1 - m_1^2 - m_2^2 - \eta\tilde{c})(\tilde{c} - c_1)(\tilde{c} - c_2) = 0, \tag{125}$$

with $\tilde{c}_1 + \tilde{c}_2 + \tilde{c}_3 = 2 + 1/\eta$. Evaluating $\ln \mathcal{Y}$ in (123) at leading order (App. A.1.3),

$$\frac{1}{N}\mathcal{H}[P_{\rm pp}] = \Phi(c_1, c_2) - \Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3). \tag{126}$$

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:

$$\langle \mathcal{H}[P_W; P^*] \rangle_{W\sim P_\eta} = \mathcal{H}[\langle P_W \rangle_{W\sim P_\eta}; P^*] = \mathcal{H}[P_{\rm pp}; P^*], \tag{127}$$

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

$$\mathcal{H}[P_{\rm pp}; P^*] = \langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} + \ln Z(W^*). \tag{128}$$

For $K = 2$ with the rank-one teacher W^*, the sample statistics Eq. (39) combined with $\mathbf{c}_2^\top W^* \mathbf{c}_2 = 0$ and $\mathbf{c}_1^\top W^* \mathbf{c}_1 = 2\omega^*(\omega^* - 1)/(2\omega^* - 1)$ yield

$$\frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*) \rangle_{P_{\rm pp}} = \left(\frac{\gamma}{\eta}\frac{\omega^{*2}}{(2\omega^* - 1)^2}(\eta c_1 \chi_1 - 1) - 1\right)\frac{\omega^* - 1}{2}, \tag{129}$$

where χ_1 is the saddle-point value from the spherical integral (19).

D.6. Forward cross-entropies

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

$$\langle \mathcal{H}[P^*; P_W] \rangle_{W\sim P_\eta} = \langle \langle \mathcal{E}(\mathbf{x}; W) \rangle_{\mathbf{x}\sim P^*} + \ln Z(W) \rangle_{W\sim P_\eta} . \tag{130}$$

For $K = 2$ with the rank-one teacher, the weight statistics Eq. (38) combined with (110) give

$$\frac{1}{N}\langle \langle \mathcal{E}(\mathbf{x}; W) \rangle_{\mathbf{x}\sim P^*} \rangle_{W\sim P_\eta} = -\left(c_1\chi_1 - \frac{1}{\eta}\right)\frac{(\omega^* - 1)^2}{(2\omega^* - 1)^2}. \tag{131}$$

Posterior-predictive forward cross-entropy. From (113),

$$\mathcal{H}[P^*; P_{\rm pp}] = \ln \mathcal{Y}(C) - \langle \ln \mathcal{Y}(C_{\mathbf{x}}) \rangle_{\mathbf{x}\sim P^*} . \tag{132}$$

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

$$\lambda_\pm = \frac{1}{2}\left\{c_1 + \frac{1}{\eta} \pm \sqrt{\left(c_1 - \frac{1}{\eta}\right)^2 + \frac{2}{\eta}(c_1 - c_2)^2}\right\}, \tag{133}$$

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 (111) and (17). Differences of $\ln \mathcal{Y}$ are taken at leading $\mathcal{O}(N)$ from App. A.1.3.

Posterior-average reverse KL. Combining (128) with (17),

$$\frac{1}{N}\langle D_{\mathrm{KL}}(P_W \| P^*)\rangle_{W\sim P_\eta} = \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*)\rangle_{P_{\mathrm{pp}}} + \frac{1}{2}\left(\omega^* - 1 - \ln\omega^* + F(\mu)\right). \tag{134}$$

with $\langle \mathcal{E}(\mathbf{x}; W^*)\rangle_{P_{\mathrm{pp}}}$ from (129).

Posterior-predictive reverse KL. Combining (128) with (126),

$$\frac{1}{N} D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W^*)\rangle_{P_{\mathrm{pp}}} + \frac{\omega^* - \ln\omega^*}{2} + \Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3) - \Phi(c_1, c_2), \tag{135}$$

with $\langle \mathcal{E}(\mathbf{x}; W^*)\rangle_{P_{\mathrm{pp}}}$ from (129).

Posterior-average forward KL. Combining (131) with the teacher entropy $\mathcal{H}[P^*]/N = \frac{1}{2}\ln(2\pi e) - \frac{1}{2}\ln\omega^*$ from (111) and the posterior-average $\langle \ln Z(W)\rangle_{W\sim P_\eta}/N$ from (15) (the $\frac{1}{2}\ln(2\pi)$ base-measure constant cancels between the two),

$$\frac{1}{N}\langle D_{\mathrm{KL}}(P^* \| P_W)\rangle_{W\sim P_\eta} = -\frac{\eta c_1 \chi_1 - 1}{\eta}\frac{(\omega^* - 1)^2}{(2\omega^* - 1)^2} + \frac{\mu - F(\mu)}{2} - \frac{1 - \ln\omega^*}{2}. \tag{136}$$

Posterior-predictive forward KL. Subtracting the teacher entropy from (132) and evaluating $\ln \mathcal{Y}$ at the eigenvalues (133),

$$\frac{1}{N} D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = -\frac{1 - \ln\omega^*}{2} - \Phi(\lambda_+, \lambda_-, c_2) + \Phi(c_1, c_2), \tag{137}$$

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

D.8. Posterior-predictive gap identities

We show that:

$$\langle D_{\mathrm{KL}}(P_W \| P^*)\rangle_{W\sim P_\eta} - D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \mathrm{MI}(W; \mathbf{x}|\mathcal{D}) \geq 0, \tag{138}$$

$$\langle D_{\mathrm{KL}}(P^* \| P_W)\rangle_{W\sim P_\eta} - D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = \left\langle D_{\mathrm{KL}}(P_\eta(W|\mathcal{D}) \| P_\eta(W|\mathcal{D}, \mathbf{x}))\right\rangle_{\mathbf{x}\sim P^*} \geq 0, \tag{139}$$

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

Reverse KL gap. Using (127):

$$\langle D_{\mathrm{KL}}(P_W \| P^*) \rangle - D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \mathcal{H}[P_{\mathrm{pp}}] - \langle \mathcal{H}[P_W] \rangle_{W \sim P_\eta} \tag{140}$$

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

$$\begin{aligned} \mathrm{MI}(W; \mathbf{x}|\mathcal{D}) &= \int P_\eta(W|\mathcal{D}) P(\mathbf{x}|W) \ln \frac{P(\mathbf{x}|W)}{P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D})} \mathrm{d}W \mathrm{d}\mathbf{x} \\ &= \mathcal{H}[P_{\mathrm{pp}}] - \langle \mathcal{H}[P_W] \rangle_{W \sim P_\eta} \geq 0 \end{aligned} \tag{141}$$

establishing (138).

Forward KL gap. We compute:

$$\langle D_{\mathrm{KL}}(P^* \| P_W) \rangle_{W \sim P_\eta} - D_{\mathrm{KL}}(P^* \| P_{\mathrm{pp}}) = \int \mathrm{d}W \, P_\eta(W|\mathcal{D}) \int \mathrm{d}\mathbf{x} \, P^*(\mathbf{x}) \ln \frac{P_{\mathrm{pp}}(\mathbf{x}|\mathcal{D})}{P(\mathbf{x}|W)} \tag{142}$$

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 (139).

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 [35] and to the epistemic/aleatoric decomposition of predictive entropy used in Bayesian deep learning [15, 29, 33]. The forward gap goes back to [1], 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:

$$D_{\mathrm{KL}}(P^* \| P_{\beta W}) = \beta \mathcal{H}[P^*; P_W] - \mathcal{H}[P^*] + \ln Z(\beta W) - \beta \ln Z(W) \tag{143}$$

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

$$\beta_{\mathrm{TT}} = \operatorname*{argmin}_{\beta} D_{\mathrm{KL}}(P^* \| P_{\beta W}) \tag{144}$$

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

$$\left. \frac{\partial D_{\mathrm{KL}}(P^* \| P_{\beta W})}{\partial \beta} \right|_{\beta=1} = \mathcal{H}[P^*; P_W] - \mathcal{H}[P_W] \tag{145}$$

β_{TT} is > 1 or < 1 according to whether (145) is positive or negative, respectively. Note that equations (143) and (145) 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 (131) and the model average energy (16) gives, to leading order in N,

$$\frac{1}{N}\frac{\partial D_{\mathrm{KL}}(P^*\|P_{\beta W})}{\partial\beta}\bigg|_{\beta=1} = \frac{\mu-1}{2} - \left(\frac{\omega^*-1}{2\omega^*-1}\right)^2\left(c_1\chi_1 - \frac{1}{\eta}\right) \tag{146}$$

$$= \begin{cases} \frac{1}{2\gamma\eta} - \frac{1}{\gamma}\left(\frac{\omega^*-1}{\omega^*}\right)^2 & h=0 \\ \frac{1}{\sqrt{\gamma\eta}} - \frac{1}{2} - \frac{1}{\gamma}\left(\frac{\omega^*-1}{\omega^*}\right)^2 & h\neq 0, u_1=0 \\ \frac{\lambda_1-1}{2} - \left(\frac{\omega^*-1}{2\omega^*-1}\right)^2\left(c_1\lambda_1 - \frac{1}{\eta}\right) & h, u_1\neq 0 \end{cases} \tag{147}$$

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 β_{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 $\beta_{\mathrm{TT}} < 1$ 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 $u_1 \neq 0$, $\beta_{\mathrm{TT}} > 1$ 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 $\beta_{\mathrm{TT}} > 1$ 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 $\eta_{\mathrm{DD}}(\omega^*)$ above which the typical reverse KL of Fig. 9 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

$$\tfrac{1}{N}\left\langle D_{\mathrm{KL}}(P_W\|P^*)\right\rangle_{W\sim P_\eta} = \frac{\omega^*-1-\ln\omega^*}{2} + \frac{1}{4\gamma\eta}, \tag{148}$$

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 (109).

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

$$\gamma(g_1) = 2g_1^2 - \frac{g_1}{\omega^*}, \qquad \lambda_1 = \frac{1}{g_1} + \frac{g_1}{\gamma\eta}, \qquad \frac{1}{\omega^*} < g_1 < 1, \tag{149}$$

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 (149) into (134) and using $G(\mu) = 1 - h^2 = g_1$ on this branch yields the intensive reverse KL as a function of g_1,

$$\frac{1}{N}\langle D_{\mathrm{KL}}(P_W \| P^*)\rangle_{W\sim P_\eta} = \frac{\omega^* - 1 - \ln(\omega^* g_1)}{2} - \frac{\omega^*(\omega^*-1)(1-g_1)}{2\omega^*-1}\left[1 - \frac{\omega^* g_1}{\eta(2\omega^*-1)}\right] + \frac{\omega^* g_1}{4\eta(2\omega^* g_1 - 1)}, \quad (150)$$

which matches (148) 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

$$g_{\mathrm{MAP}} := \frac{2\omega^* - 1}{2\omega^*(\omega^*-1)}, \quad (151)$$

the g_1-derivative writes,

$$\frac{\mathrm{d}}{\mathrm{d}g_1}\left[\frac{1}{N}\langle D_{\mathrm{KL}}(P_W \| P^*)\rangle_{W\sim P_\eta}\right] = \frac{\omega^*(\omega^*-1)(g_1 - g_{\mathrm{MAP}})}{g_1(2\omega^*-1)} + \frac{1}{\eta}\left[\frac{\omega^{*2}(\omega^*-1)(1-2g_1)}{(2\omega^*-1)^2} - \frac{\omega^*}{4(2\omega^* g_1 - 1)^2}\right]. \quad (152)$$

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

$$\omega^* > \omega_{\mathrm{DD}} := 1 + \frac{1}{\sqrt{2}} \quad (153)$$

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

When (153) holds, an interior local minimum appears at $g_1 \in (g_{\mathrm{MAP}}, 1)$ as soon as the slope (152) is positive and the outlier-detachment condition $\eta > \eta_F(g_1)$ holds. The sharp threshold is the one-dimensional variational formula

$$\eta_{\mathrm{DD}}(\omega^*) = \inf_{g_{\mathrm{MAP}} < g_1 < 1} \frac{\omega^* g_1}{2\omega^* g_1 - 1} \times \max\left\{1, \frac{2\omega^* g_1 - 1}{(2\omega^*-1)(g_1 - g_{\mathrm{MAP}})}\left[(2g_1 - 1) + \frac{(2\omega^*-1)^2}{4\omega^*(\omega^*-1)(2\omega^* g_1 - 1)^2}\right]\right\}, \quad (154)$$

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. 9; 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 (135) over $\eta > 0$ at fixed (ω^*, γ) and obtain the five regions of Fig. 10A. On a smooth condensed branch, evaluating $\Phi(\tilde{c}_1, \tilde{c}_2, \tilde{c}_3)$ along the $m_{1,2}^2(\eta)$-parametrisation determined by (125) reduces (135) to

$$\tfrac{1}{N} D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*) = \frac{\omega^* - 1 - \ln \omega^*}{2} - \frac{m_1^2}{2 g_{\mathrm{MAP}}} - \frac{1}{2} \ln(1 - m_1^2 - m_2^2), \tag{155}$$

with g_{MAP} as in (151); 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

$$\partial_\eta m_1^2 = \frac{g_{\mathrm{MAP}} \left(\partial_\eta m_1^2 + \partial_\eta m_2^2 \right)}{1 - m_1^2 - m_2^2}. \tag{156}$$

Zero-overlap regions. The bound $-\ln(1-s) \geq s$ on $s = m_1^2 + m_2^2$ applied to (155) gives $D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*)/N - (\omega^* - 1 - \ln \omega^*)/2 \geq [(1 - 1/g_{\mathrm{MAP}}) m_1^2 + m_2^2]/2$, so for $g_{\mathrm{MAP}} \geq 1$, equivalently $\omega^* \leq \omega_0 := 1 + 1/\sqrt{2}$, 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. 10A); for $\gamma \geq c_1$ it covers all $\eta > 0$ and the leading KL is flat throughout (*deg.*).

Single-mode branch. For $\omega^* > \omega_0$ a positive-overlap optimum exists. On the $d = 1$ outlier branch, $m_2^2 \equiv 0$, and (156) reduces to the scalar target

$$(m_1^2)^* = 1 - g_{\mathrm{MAP}}. \tag{157}$$

Tracking $m_1^2(\eta)$ along the smooth branch with the g_1-parametrisation of (162), the value $\eta_{\mathrm{pp}}^{\mathrm{rev}} = 1$ is hit on

$$\gamma_{\mathrm{wc}}^{(1)}(\omega^*) = 2 g_{\mathrm{wc}}^2 - c_2\, g_{\mathrm{wc}}, \qquad g_{\mathrm{wc}} = \tfrac{1}{4} \left[(c_2 + 3c_1) - \sqrt{(c_2 + 3c_1)^2 - 16 c_1 g_{\mathrm{MAP}}} \right], \tag{158}$$

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

$$\gamma_{\infty}^{\mathrm{rev}}(\omega^*) = 2 g_{\mathrm{MAP}}^2 - c_2\, g_{\mathrm{MAP}} = \frac{2\omega^* - 1}{2\omega^*(\omega^* - 1)^2}. \tag{159}$$

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 (156) fixes the warm/cold boundary at $\gamma_{\mathrm{wc}}^{(2)} = \alpha^2$, with $\alpha \in (0, c_2)$ the unique root of

$$\alpha\, (c_1 - \alpha)(2c_1 c_2 - \alpha) = 2c_1\, g_{\mathrm{MAP}}\, (c_1 c_2 - \alpha). \tag{160}$$

The cold/MAP boundary (159) is unchanged.

Phase classification. For $\omega^* \leq \omega_0$ the optimum is the zero-overlap plateau (*warm deg.* for $\gamma < c_1$, *deg.* for $\gamma \geq c_1$). For $\omega^* > \omega_0$, set $\gamma_{\mathrm{wc}} = \gamma_{\mathrm{wc}}^{(1)}$ when $\omega^* \leq \omega_2$ and $\gamma_{\mathrm{wc}} = \gamma_{\mathrm{wc}}^{(2)}$ when $\omega^* > \omega_2$; the regions are then *warm* for $\gamma < \gamma_{\mathrm{wc}}$, *cold* for $\gamma_{\mathrm{wc}} < \gamma < \gamma_{\infty}^{\mathrm{rev}}$, and *MAP* for $\gamma \geq \gamma_{\infty}^{\mathrm{rev}}$. In the MAP region the typical-posterior reverse KL of (148) is recovered.

D.11.2. OPTIMAL TEMPERED-POSTERIOR ACCORDING TO THE FORWARD KL

We minimise (137) over $\eta > 0$ at fixed (ω^*, γ):

$$\eta_{\rm pp}^{\rm fwd} = \arg\min_{\eta>0} D_{\rm KL}(P^* \| P_{\rm pp}) \tag{161}$$

and obtain three closed-form thresholds, $\gamma_{\rm wc} < \gamma_{\rm flat} < \gamma_\infty$, that partition the (ω^*, γ) plane into the four regions of Fig. 11A.

Smooth branch. Outside the warm flat regime defined below, the perturbed and original sources carry the same number of outliers and the η-dependence of (137) does not cancel. The interior stationary condition $\partial_\eta D_{\rm KL}(P^* \| P_{\rm pp}) = 0$ 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 (133) simplifies to $\lambda_- = g_1/\eta$. Combined with the elementary identities $\lambda_+\lambda_- = \left[c_1 - \frac{1}{2}(c_1 - c_2)^2\right]/\eta$ and $\lambda_+ + \lambda_- = c_1 + 1/\eta$, this gives the smooth-branch stationary point $\eta_0(\gamma, \omega^*)$ of the forward KL,

$$\eta_0(\gamma, \omega^*) = \frac{g_1\,(1 - g_1)}{c_1(1 - g_1) - \frac{1}{2}(c_1 - c_2)^2}, \qquad g_1(\gamma, \omega^*) = \frac{c_2 + \sqrt{c_2^2 + 8\gamma}}{4}. \tag{162}$$

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

$$\gamma_\infty(\omega^*) = 2\Big[1 - \tfrac{(c_1-c_2)^2}{2c_1}\Big]^2 - c_2\Big[1 - \tfrac{(c_1-c_2)^2}{2c_1}\Big] = \frac{(3\omega^* - 2)(4\omega^* - 3)}{\omega^{*2}(2\omega^* - 1)^2}. \tag{163}$$

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

$$\gamma_{\rm wc}(\omega^*) = c_2^2 = 1/\omega^{*2}. \tag{164}$$

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 $D_{\rm KL}(P^* \| P_{\rm pp})$. Setting $\lambda_-(\eta) = \sqrt{\gamma/\eta}$ yields a quadratic in $\sqrt{\gamma/\eta}$,

$$\frac{\gamma}{\eta} - \left[\gamma + c_1 - \tfrac{1}{2}(c_1 - c_2)^2\right]\sqrt{\gamma/\eta} + \gamma\, c_1 = 0, \tag{165}$$

whose two positive roots delimit the warm interval

$$I_{\rm warm}(\gamma, \omega^*) = \left[\eta_-,\ \min(1, \eta_+)\right], \tag{166}$$

with endpoints

$$\eta_\pm = \frac{4\gamma}{\left[\gamma + c_1 - \frac{1}{2}(c_1 - c_2)^2 \mp \sqrt{\left[\gamma + c_1 - \frac{1}{2}(c_1 - c_2)^2\right]^2 - 4\gamma c_1}\right]^2}. \tag{167}$$

Every $\eta \in I_{\rm warm}$ is a global minimiser. The discriminant of (165) vanishes at

$$\gamma_{\rm flat}(\omega^*) = \left(\sqrt{c_1} - (c_1 - c_2)/\sqrt{2}\right)^2, \tag{168}$$

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

Phase classification. The minimisation has four exclusive outcomes (Fig. 11A). For $\gamma < \gamma_{\rm wc}$ the warm flat interval $I_{\rm warm} \subset (0, 1]$ is the optimum-set, with smooth-branch representative $\eta_0 < 1$ inside. In the sliver $\gamma_{\rm wc} \le \gamma \le \gamma_{\rm flat}$ a warm flat sub-interval $I_{\rm warm} \subset (0, 1)$ and the cold isolated point $\eta_0 > 1$ are simultaneously global minimisers; $D_{\rm KL}(P^* \| P_{\rm pp})/N$ traces a non-convex W shape with the two troughs at the same value to leading order in N. For $\gamma_{\rm flat} < \gamma < \gamma_\infty$ the unique global minimiser is the cold smooth-branch point $\eta_{\rm pp}^{\rm fwd} = \eta_0 > 1$. Above γ_∞ the infimum is reached only as $\eta \to \infty$ (MAP).

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

$$\frac{1}{N} D_{\rm KL}(P^* \| P_t) = \frac{1}{N}\mathcal{H}[P^*; P_t] - \frac{1}{N}\mathcal{H}[P^*], \tag{169}$$

$$\frac{1}{N} D_{\rm KL}(P_t \| P^*) = \frac{1}{N}\mathcal{H}[P_t; P^*] - \frac{1}{N}\mathcal{H}[P_t], \tag{170}$$

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. (15), (16) and (17),

$$\begin{aligned} \frac{1}{N}\ln Z(W(t)) &= \frac{\ln(2\pi) + \mu(t) - F(\mu(t))}{2} \\ \frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W(t))\rangle_{\mathbf{x}\sim P_t} &= \frac{1-\mu(t)}{2}, \\ \frac{1}{N}\mathcal{H}[P_{W(t)}] &= \frac{\ln(2\pi e) - F(\mu(t))}{2} \end{aligned} \tag{171}$$

D.12.2. FORWARD CROSS ENTROPY

We first compute

$$\frac{1}{N}\langle \mathcal{E}(\mathbf{x}; W(t))\rangle_{\mathbf{x}\sim P^*} = -\frac{1}{2}\left(1 - \frac{1}{\omega^*}\right)\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^* \tag{172}$$

where

$$\Theta(t) = \frac{1 - e^{-\gamma t/2}}{\gamma} C - \frac{1}{N}\int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}\mathbf{x}(u)\mathbf{x}(u)^\top \tag{173}$$

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 $\mathbf{w}^{*\top}W_{\rm GOE}(t)\mathbf{w}^* = o(1)$. 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

$$\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^* = \left(1 - \frac{1}{2\omega^* - 1}\right)\left(\frac{1 - e^{-\gamma t/2}}{\gamma}c_1 - \int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}s_1(u)^2\right), \tag{174}$$

where we used Eq. (110).

We then find:

$$\frac{1}{N}\mathcal{H}[P^*;P_t] = \frac{\ln 2\pi + \mu(t) - F(\mu(t))}{2} - \frac{1}{2}\left(1-\frac{1}{\omega^*}\right)\mathbf{w}^{*\top}\Theta(t)\mathbf{w}^*. \tag{175}$$

Determining $\mu(t)$ or $g_t := G(\mu(t))$ requires knowledge of $\lambda_1(t)$. To make progress, we use the approximation that $W_{\rm GOE}(t)$ is asymptotically independent of $\Theta(t)$. Note that, for $u \le t$, $\Theta(t)$ depends on $W_{\rm GOE}(u)$ 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

$$g_t = \begin{cases} \min\{1,\sqrt{\gamma\eta}\} & \theta_1(t) \le \theta_c, \\ \theta_1(t)^{-1}, & \theta_1(t) > \theta_c, \end{cases} \qquad \text{with} \qquad \theta_c = \max\{1, 1/\sqrt{\gamma\eta}\}, \tag{176}$$

$h_t^2 = (1-g_t)_+$ 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,

$$\frac{1}{N}\mathcal{H}[P_t;P^*] = \frac{1}{2}\ln 2\pi + \frac{1}{2}(\omega^* - \ln\omega^*) - \frac{1}{2}\omega^* h_t^2\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} + o(1). \tag{177}$$

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

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{t}_1(t))^2}{N^2} = \left(1 - \frac{1}{\gamma\,\eta\,\theta_1(t)^2}\right)_+ + o(1). \tag{178}$$

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,

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} = \left(1 - \frac{1}{\gamma\,\eta\,\theta_1(t)^2}\right)_+ \frac{(\mathbf{t}_1(t)\cdot\mathbf{w}^*)^2}{N} + o(1). \tag{179}$$

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

$$\frac{(\mathbf{v}_1(t)\cdot\mathbf{w}^*)^2}{N} = \left(1 - \frac{1}{\gamma\eta\,\theta_1(t)^2}\right)\left(1 - \frac{1}{2\omega^*-1}\right)\frac{(\mathbf{t}_1(t)\cdot\mathbf{c}_1)^2}{N^2}. \tag{180}$$

D.12.4. FORWARD KL

Combining Eqs. (111) and (175) yields

$$\begin{aligned}\frac{1}{N}D_{\rm KL}(P^*\,\|\,P_t) &= \frac{1}{2}\ln\omega^* - \frac{1}{2} + \frac{1}{2}\left[\mu(t) + \ln g_t - \frac{g_t^2}{2\gamma\eta}\right] \\ &- \frac{1}{2}\left(1-\frac{1}{\omega^*}\right)\left(1-\frac{1}{2\omega^*-1}\right)\left(\frac{1-e^{-\gamma t/2}}{\gamma}c_1 - \int_0^t \mathrm{d}u\, e^{-\gamma(t-u)/2}s_1(u)^2\right).\end{aligned} \tag{181}$$

D.12.5. REVERSE KL

Using Eqs. (177), (180),

$$\frac{1}{N}D_{\mathrm{KL}}(P_t \,\|\, P^*) = \frac{1}{2}(\omega^* - 1 - \ln\omega^*) - \frac{1}{2}\ln g_t + \frac{g_t^2}{4\gamma\eta} - \frac{1}{2}\omega^* h_t^2 \left(1 - \frac{1}{\gamma\eta\,\theta_1(t)^2}\right)\left(1 - \frac{1}{2\omega^* - 1}\right)\frac{(\mathbf{t}_1(t)\cdot\mathbf{c}_1)^2}{N^2}. \tag{182}$$

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

$$\frac{1}{N}D_{\mathrm{KL}}(P_t \,\|\, P^*) = \frac{1}{2}(\omega^* - 1 - \ln\omega^*) - \frac{1}{2}\ln g_0 + \frac{g_0^2}{4\gamma\eta}, \qquad g_0 = \begin{cases} 1, & \gamma\eta \geq 1 \\ \sqrt{\gamma\eta}, & \gamma\eta < 1. \end{cases} \tag{183}$$

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

$$\theta^\star = \frac{1}{2}\left[\omega^*\left(1 - \frac{1}{2\omega^* - 1}\right) + \sqrt{\omega^{*2}\left(1 - \frac{1}{2\omega^* - 1}\right)^2 + \frac{4}{\gamma\eta}}\right]. \tag{184}$$

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

$$t^* \simeq -\frac{2}{\gamma}\ln\left(1 - \frac{\gamma\,\theta^*}{c_1}\right). \tag{185}$$

Injecting the parameter values of Fig. 12 yields $t^* \simeq 5\ln 2 \simeq 3.47$, in excellent agreement with the numerical value.

Appendix E. Generative-metrics figures (teacher–student scenario)

The figures referenced from Sec. 3 are collected here; parameters are listed in App. H, derivations in App. D.

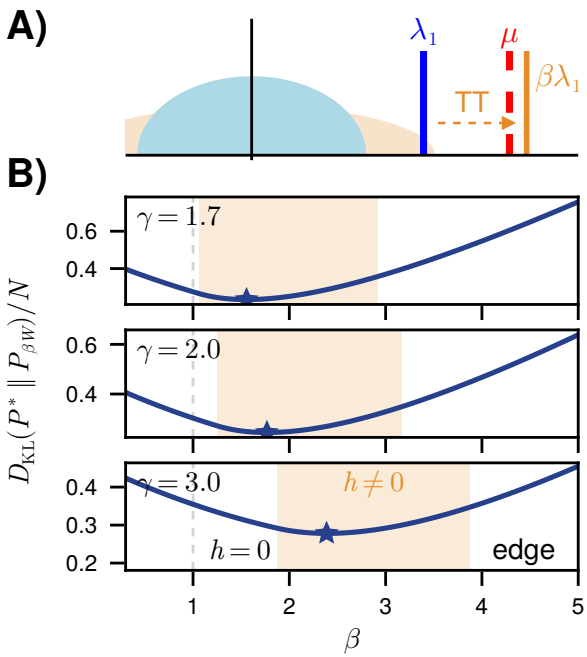

Figure 8: **Sampling temperature tuning** (TT). **A)** Mode rescue through temperature tuning, illustrating how the bulk and outlier position change. **B)** $D_{\mathrm{KL}}(P^* \| P_{\beta W})$ vs. β.

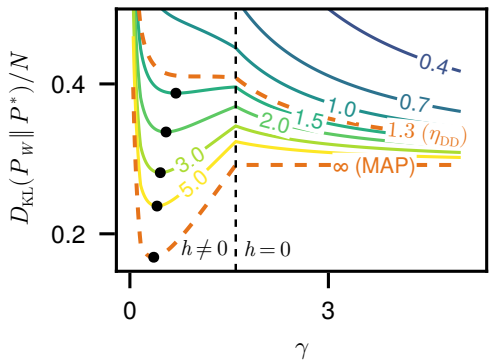

Figure 9: **Double descent.** $D_{\mathrm{KL}}(P_W \| P^*)$ vs. γ for the rank-one teacher with $\omega^* = 2.5$; η indicated on each curve. Black dots: minima. Orange dashed: thresholds $\eta = \eta_{\mathrm{DD}} \approx 1.342$ (upper) and $\eta \to \infty$ (lower). Vertical dashed: boundary between $h{\neq}0$ and $h{=}0$ phases.

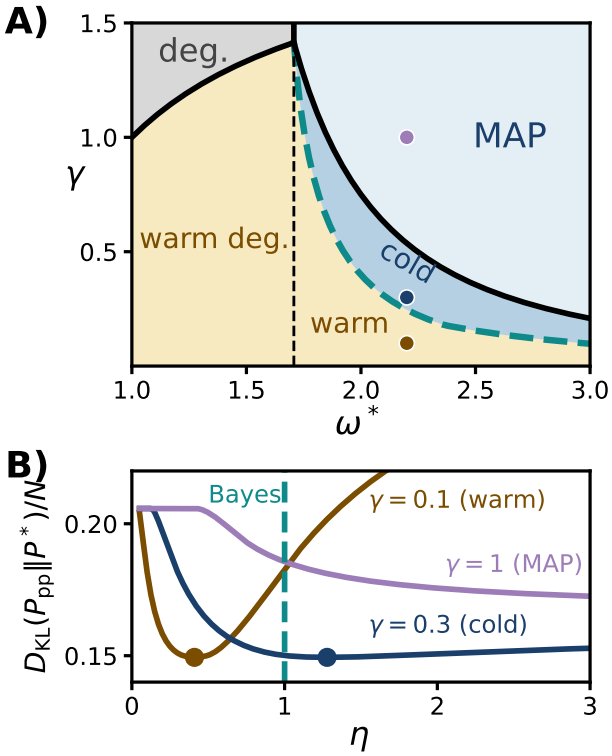

Figure 10: **Tempered-posterior, reverse KL. A)** Phases of the optimal $\eta_{\mathrm{pp}}^{\mathrm{rev}}(\gamma, \omega^*)$. **B)** Reverse KL $D_{\mathrm{KL}}(P_{\mathrm{pp}} \| P^*)$ vs. η for selected γ (dots: minima).

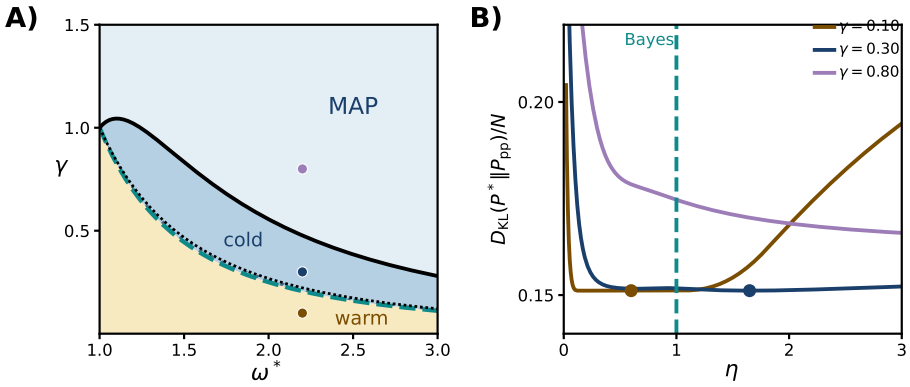

Figure 11: **Tempered-posterior, forward KL.** Optimum $\eta_{\rm pp}^{\rm fwd} = \arg\min_{\eta>0} D_{\rm KL}(P^*\|P_{\rm pp})$ for the rank-one $K{=}2$ teacher. **A)** Phases in the (ω^*, γ) plane: warm flat (yellow), unique cold (medium blue), MAP (light blue); the hatched sliver is the mixed warm/cold tie. Boundaries: dashed teal $\gamma_{\rm wc}$ (Bayes-crossing, $\eta_{\rm pp}^{\rm fwd} = 1$), dotted black $\gamma_{\rm flat}$ (upper edge of the warm flat interval), solid black γ_∞ (cold/MAP). **B)** $D_{\rm KL}(P^*\|P_{\rm pp})/N$ vs. η at $\omega^* = 2.2$ for $\gamma \in \{0.10, 0.30, 0.80\}$ marked in panel A; teal dashed at $\eta = 1$.

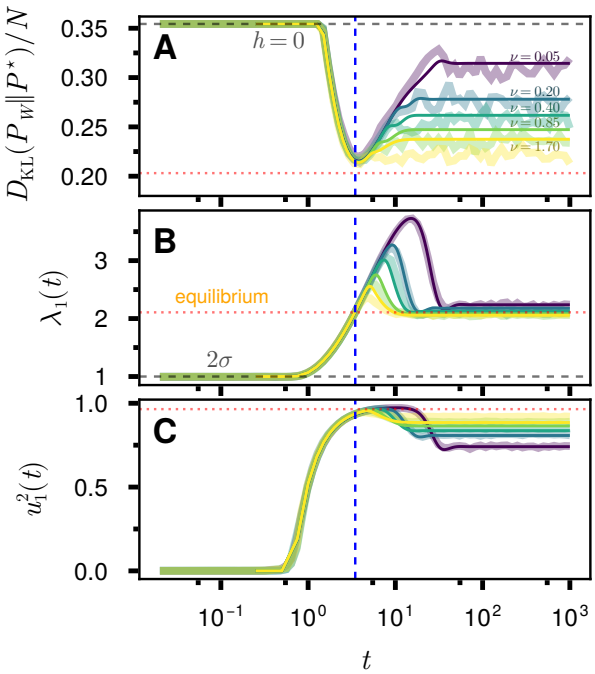

Figure 12: **Out-of-equilibrium training.** **A)** Reverse $D_{\mathrm{KL}}(P_{W(t)} \| P^*)$ during training, **B)** $\lambda_1(t)$, **C)** $u_1(t)$. Color: ν. Blue dashed: optimal early-stopping time. Orange dashed: equilibrium values (at $\nu \to \infty$).

Appendix F. Comparison of the SBM with a multivariate Gaussian model

The unconstrained multivariate Gaussian baseline $P_{\mathrm{G}}(\mathbf{x}|W) \propto \exp(\frac{1}{2}\mathbf{x}^\top W\mathbf{x})$ on $\mathbb{R}^N$ (normalizable only for $W \prec 0$) has the additive log-partition function $\ln Z_{\mathrm{G}}(\Lambda) = \frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_i \ln(-\lambda_i)$, which contributes only an $\mathcal{O}(1)$ force $K\eta/(2\lambda_k)$ on each top eigenvalue. The SBM partition function instead has a $\mathcal{O}(N)$ condensation force $-NK\eta h^2/2$ that couples all top modes through a shared order parameter. This structural difference is responsible for the phenomenology of the SBM.

In the $h = 0$ regime the collective force vanishes and the two models agree to leading order: same Wigner bulk $\sigma = 1/\sqrt{\gamma\eta}$, same retarded-learning threshold $\gamma < \eta c_k^2$, same outlier positions $\lambda_k = (\eta c_k)^{-1} + c_k/\gamma$ (Fig. 13). The condensed phases ($h \neq 0$) have no Gaussian analog, and the double-descent spike at the $h \neq 0$ to $h = 0$ boundary (Fig. 9) together with the warm-posterior effect $\eta_{\mathrm{opt}} \neq 1$ (Fig. 10B) trace to the non-conjugate, phase-dependent $\ln Z$: under the Gaussian with its conjugate prior, Bayesian inference is exact at $\eta = 1$, no warm-posterior effect arises, and the KL divergences are smooth in γ.

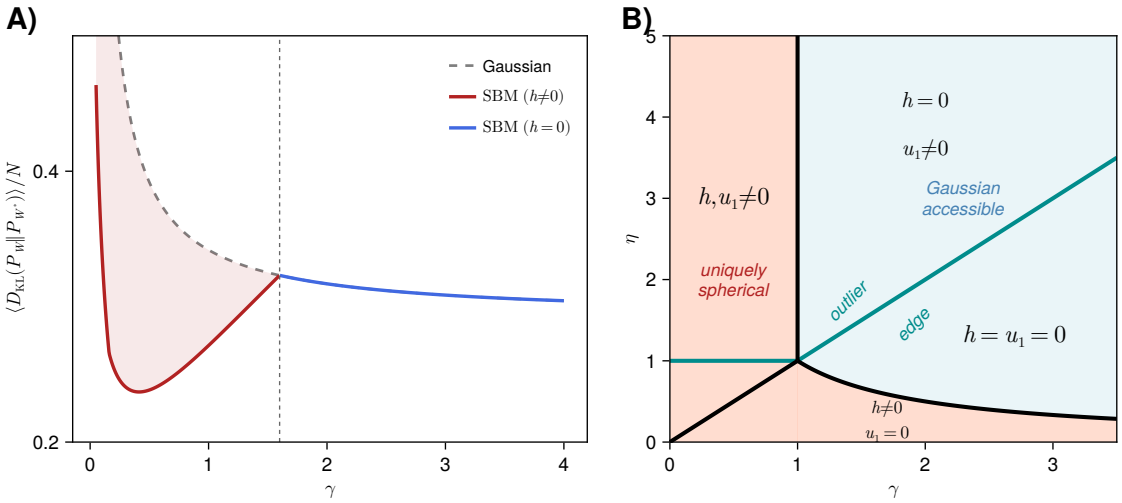

Figure 13: **Comparison of the SBM with the unconstrained Gaussian model. A)** Typical reverse KL divergence $\langle D_{\mathrm{KL}}(P_W\|P^*)\rangle/N$ as a function of the prior strength γ, at $\eta = 5$ and $\omega^* = 2.5$ ($K = 2$). The SBM curve (solid) is colored by phase: condensed $h \neq 0$ (red) and uncondensed $h = 0$ (blue). The Gaussian equivalent (dashed gray) is the $h = 0$ formula $(\omega^* - \ln\omega^* - 1 + 1/(2\gamma\eta))/2$, which coincides with the SBM throughout the $h = 0$ region. In the condensed region ($\gamma \lesssim 1.6$), condensation lowers the SBM's KL divergence below the Gaussian baseline; the shaded area quantifies this benefit. The vertical dashed line marks the $h = 0$ to $h \neq 0$ phase boundary. **B)** Phase diagram of the K=1 SBM, with phases labeled by the (h, u_1) order parameters as in Fig. 1. The uncondensed region ($h = 0$, blue shading) is the only regime where the Gaussian model and SBM agree to leading order. The condensed phases ($h \neq 0$, salmon shading) are uniquely attributable to the spherical constraint and have no Gaussian analog. Black lines: $h = 0$ to $h \neq 0$ boundaries. Teal lines: outlier–edge transitions.

Appendix G. Computational experiments in generative models

G.1. 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 [22]. 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

$$\beta_{\mathrm{TT}}^{\mathrm{cov}} = 1 + \gamma \, \frac{\sum_k \lambda_k^2 \, c_k^2}{\sum_k \lambda_k (\lambda_k - \mu') \, c_k^4}, \tag{186}$$

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

$$\sigma_k^2(1) = c_k - \gamma \lambda_k + \mathcal{O}(\gamma^2). \tag{187}$$

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

$$\partial_\beta \sigma_k^2 \big|_{\beta=1} = (\lambda_k - \mu') \, \sigma_k^2(1)^2 = (\lambda_k - \mu') \, c_k^2 + \mathcal{O}(\gamma), \tag{188}$$

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

$$\mu'_{\mathrm{SBM}} = \frac{\sum_k \lambda_k c_k^2}{\sum_k c_k^2} + \mathcal{O}(\gamma); \tag{189}$$

the Potts softmax enforces normalization site-wise, leaving $\mu' = 0$. Expanding $\sigma_k^2(\beta)$ to first order around $\beta = 1$ via (187)–(188) gives the residual

$$\sigma_k^2(\beta) - c_k = -\gamma \lambda_k + (\beta - 1)(\lambda_k - \mu') \, c_k^2 + \mathcal{O}(\gamma^2). \tag{190}$$

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

$$\sum_k \lambda_k \, c_k^2 \big(\sigma_k^2(\beta_{\mathrm{TT}}^{\mathrm{cov}}) - c_k \big) = 0. \tag{191}$$

Substituting (190) into (191) yields

$$-\gamma \sum_k \lambda_k^2 c_k^2 + (\beta_{\mathrm{TT}}^{\mathrm{cov}} - 1) \sum_k \lambda_k (\lambda_k - \mu') c_k^4 = 0, \tag{192}$$

which solves to Eq. (186).

G.1.1. RESPONSE-REGULATOR PF00072

We trained pairwise Potts models on the Pfam response-regulator family PF00072 ($L = 111$, $q = 21$, effective sample size $M_{\mathrm{eff}} \approx 3.6 \times 10^4$) via adabmDCA [49] 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 $\beta_{\mathrm{opt}}(\gamma)$, and (ii) the covariance eigenvalue spectrum $\sigma_k^2(\beta)$ of generated samples. Figure 14 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 $\beta_{\mathrm{opt}}(\gamma)$. 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 $\beta \simeq \beta_{\mathrm{opt}} = 1.5$, the direct finite-$N$ signature of the SBM condensation. Past β_{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, β_{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 β_{opt} degrades the generative model. Panel C shows $\beta_{\mathrm{opt}}(\gamma)$ rising monotonically from 1 as γ increases and tracking the SBM prediction (186) 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.1.2 below.

G.1.2. SH3 DOMAIN PF00018

The SH3 domain PF00018 ($L = 48$, $q = 21$, $M_{\mathrm{eff}} \approx 1.33 \times 10^4$, reweighting ratio $M_{\mathrm{eff}}/M \approx 0.075$) sits deeper in the undersampled regime than PF00072; protocols are identical to App. G.1. Fig. 15 shows the temperature-tuning effect is correspondingly stronger, with $\beta_{\mathrm{opt}} \simeq 1.75$ already at $\gamma = 0.8$; Eq. (186) again tracks the empirical optimum across γ, with the same large-γ bias attributable to finite-order expansion and the softmax constraint.

G.2. Double descent

G.2.1. DOUBLE DESCENT IN A TRACTABLE NORMALIZING FLOW ON $\mathcal{S}_N$

To check that this pattern is not a peculiarity of the spherical Bingham parameterization, we reproduce it in a different tractable student: a normalizing flow on $\mathcal{S}_N$ built from Householder reflections

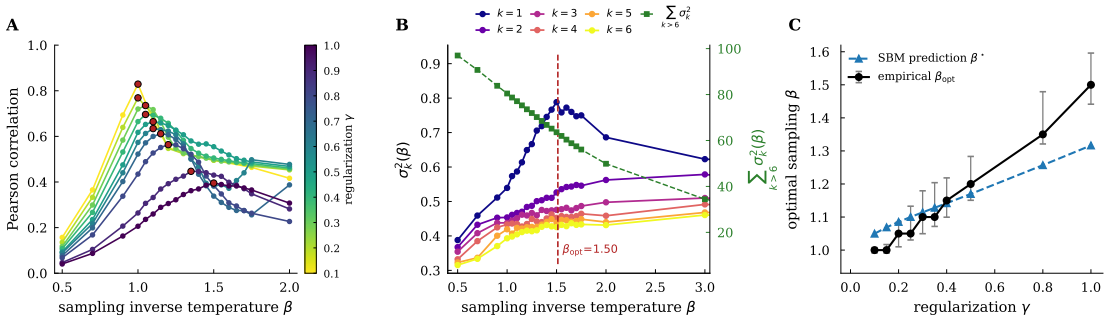

Figure 14: **Temperature tuning on Potts BM trained on PF00072.** Pairwise Potts models on the Pfam response-regulator family PF00072 ($L = 111$, $q = 21$), trained via adabmDCA [49]. **A)** Pearson correlation between generated and data connected correlations as a function of sampling inverse temperature 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 (186) (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 $\sigma_P \simeq 5 \times 10^{-3}$).

and altitude transforms [48], trained by reverse-KL variational inference against the same teacher ($\omega^* = 2.5$, $K = 2$, $N = 30$), with an L_2 penalty γ on all flow parameters. The $\gamma \to \infty$ limit drives these parameters to zero, collapsing the flow to the identity and P_θ to the uniform measure on $\mathcal{S}_N$, matching the SBM limit. The parameter posterior is collected by stochastic gradient Langevin dynamics [60] at $T_{\rm L} = 10^{-2}$ and fit locally with SWAG [38]. Figure 16 shows the posterior-averaged reverse KL and the $h \neq 0$ *fraction* (fraction of training seeds with $\langle(\mathbf{x} \cdot \mathbf{w}^*)^2\rangle/N$ above threshold) versus γ, for several Boltzmann reweighting temperatures T on the collected samples ($w_i \propto e^{-L(\theta_i)/T}$, self-normalized). Tracking the $h \neq 0$ fraction lets us locate the flow's $h \neq 0$ to $h = 0$ transition and coincides with the KL peak. As in Fig. 9, the KL has a minimum in the $h \neq 0$ phase, a peak at the transition, and a descent in the $h = 0$ phase toward the uniform value $(\omega^* - 1 - \ln \omega^*)/2 \approx 0.29$. Warming T deepens both features and shifts $\gamma_{\rm min}$ (stars) from ≈ 0.02 near MAP to ≈ 0.5 at $T = 5$: a broader posterior needs a stronger prior to suppress its high-loss tails. The MAP curve is flat in the $h \neq 0$ phase with no second descent, confirming that the post-spike descent is a Bayesian-averaging effect: it comes from the $\mathcal{O}(1/\eta)$ contribution absent from the $\eta \to \infty$ limit of (148).

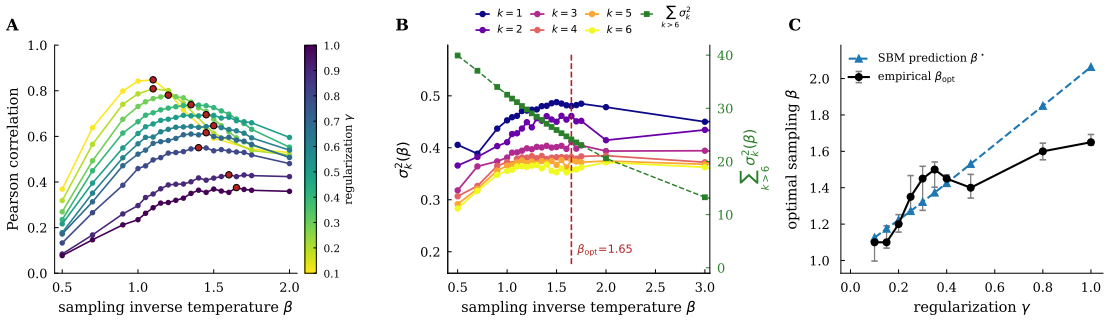

Figure 15: **Temperature tuning on Potts BM trained on PF00018.** Pairwise Potts models on the Pfam SH3 domain PF00018 ($L = 48$, $q = 21$), trained via adabmDCA. **A)** Pearson correlation between generated and data connected correlations as a function of sampling inverse temperature β, for sixteen 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 (186); error bars show the Pearson-plateau half-width (see Fig. 14 caption). Conventions as in Fig. 14.

G.2.2. DOUBLE DESCENT IN A BINARY FVSBN, LARGE-K BEHAVIOR AND THE RANK-1-DATA CONDITION

We complement the flow experiment of App. G.2.1 with a discrete student trained on two different rank-1-dominated teachers, varying the number of training samples K from $K \ll N$ to $K = 2N$ to test the regime-(b) extension discussed in Sec. 2.1: the equilibrium theory remains valid for $K \to \infty$ provided the empirical covariance C keeps $\mathcal{O}(1)$ outliers above its bulk.

We train a binary fully-visible sigmoid belief network (FVSBN) [24] on $\{\pm 1\}^N$ with $N = 16$, autoregressive log-likelihood $\ln q_\theta(s) = \sum_i \ln \sigma\big((2s_i)\big(b_i + \sum_{j<i} W_{ij} s_j\big)\big)$, and only the strictly lower-triangular $W \in \mathbb{R}^{N\times N}$ and biases $b \in \mathbb{R}^N$ as parameters. The student is trained by Adam maximum-likelihood for 1500 steps, then SWAG snapshots are collected every 15 steps over 1500 constant-LR SGD-Langevin updates at $T_{\text{lang}} = 0.3$, fit by a diagonal Gaussian, and used to draw 30 posterior-predictive students whose reverse KL against the teacher we average. The optimizer's weight decay is set to $\text{wd} = N\gamma/(2K)$, so the per-sample-averaged training loss matches the SBM prior $\propto \exp(-N\gamma \operatorname{tr} W^2/4)$ at every K and the γ-axis is directly comparable across K.

Two teachers with the same rank-1-dominated population covariance but otherwise unrelated structure are used: a rank-1 Curie–Weiss / Hopfield teacher $\mathcal{E}(s) = -\frac{1}{2N}(\xi \cdot s)^2$ at $\beta = 2.0$ with a random $\xi \in \{\pm 1\}^N$ (Fig. 17A), and 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. 17B). Both teachers admit exact $\ln Z$ (exhaustive enumeration and column transfer matrix, respectively); the former also lets us evaluate the reverse KL exactly. We sweep $K \in \{4, 8, 16, 32\}$ at 10 seeds each, logging the spectrum of $C = \frac{1}{N}\sum_k x^k x^{k\top}$ per seed. As predicted, $\lambda_1(C)$ grows linearly with K while $\lambda_2(C)$ grows much more slowly; the gap λ_1/λ_2 rises from 9.7 to 27.3 on CW and from 9.5 to 18.8 on Ising as $K\colon 4 \to 32$, so the empirical covariance stays low-effective-rank across the entire sweep.

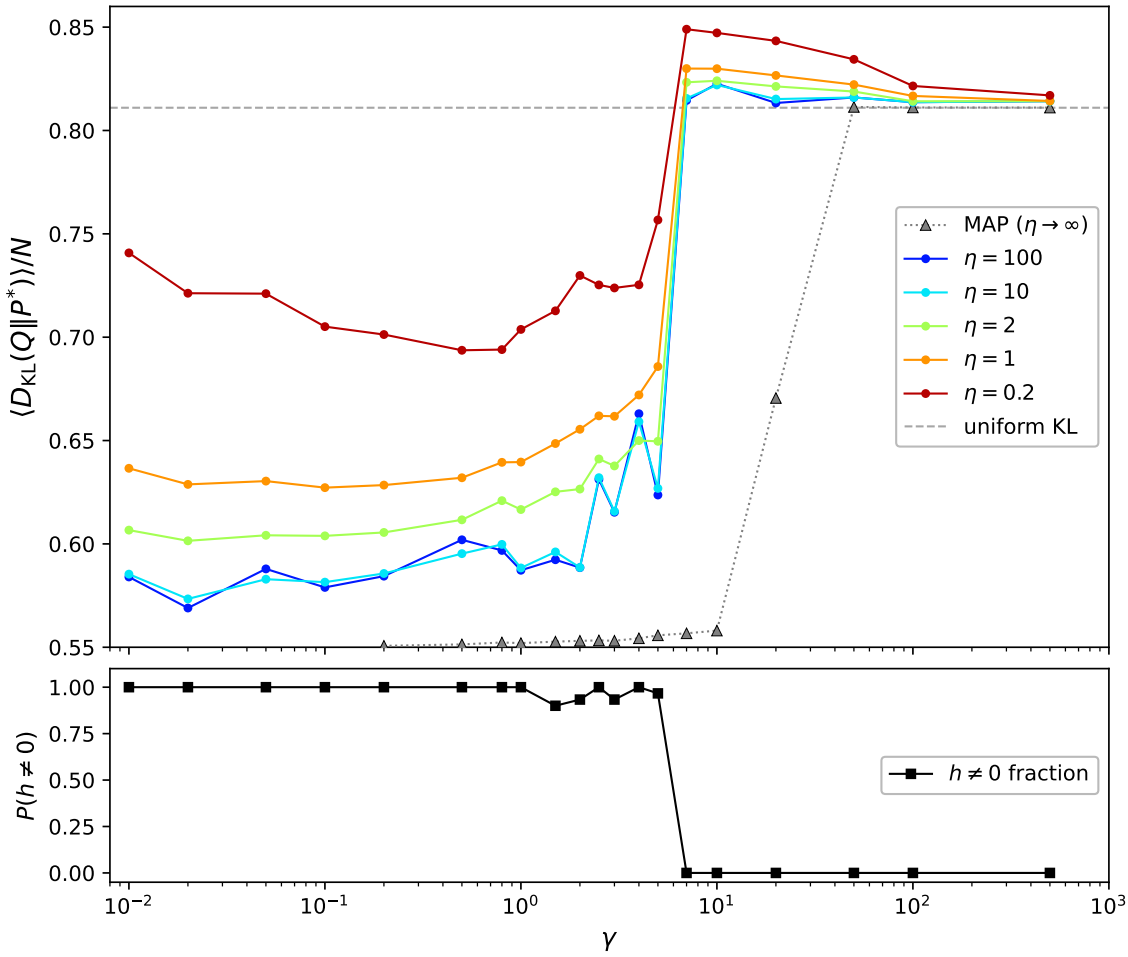

Figure 16: **Double descent in a tractable flow generative model.** Posterior-averaged reverse KL divergence per dimension (*top*) and $h \neq 0$ fraction across training seeds (*bottom*) as functions of the L_2 prior strength γ, for a Householder-and-altitude normalizing flow on $\mathcal{S}_N$ [48] trained by reverse-KL variational inference against the same rank-one Bingham teacher as in Fig. 9 ($\omega^* = 2.5$, $K = 2$, $N = 30$, $L = 8$ flow layers, 30 training seeds). The parameter posterior is collected by stochastic gradient Langevin sampling [60] and fit locally with SWAG [38]; the five coloured curves use different Boltzmann reweighting temperatures T on the collected samples ($w_i \propto e^{-L(\theta_i)/T}$, self-normalized), interpolating between the MAP limit ($T \to 0$, black triangles) and a broader posterior average ($T = 5$, red). Stars ($\star$) mark the γ minimizing each curve. As T grows, $\gamma_{\min}$ shifts to larger values and a pronounced bump at $\gamma \approx 5$–10 emerges, coinciding with the $h \neq 0$ to $h = 0$ transition of the trained flow (sharp drop of the $h \neq 0$ fraction in the lower panel). Past the transition the curves descend toward the uniform value $(\omega^* - 1 - \ln \omega^*)/2 \approx 0.29$ (dashed line), reproducing the double-descent pattern predicted analytically for the SBM in Fig. 9. The MAP curve is flat across the $h \neq 0$ phase and shows no second descent, consistent with the $\eta \to \infty$ analytical limit of (148) in which the $\mathcal{O}(1/\eta)$ residual-anisotropy term vanishes.

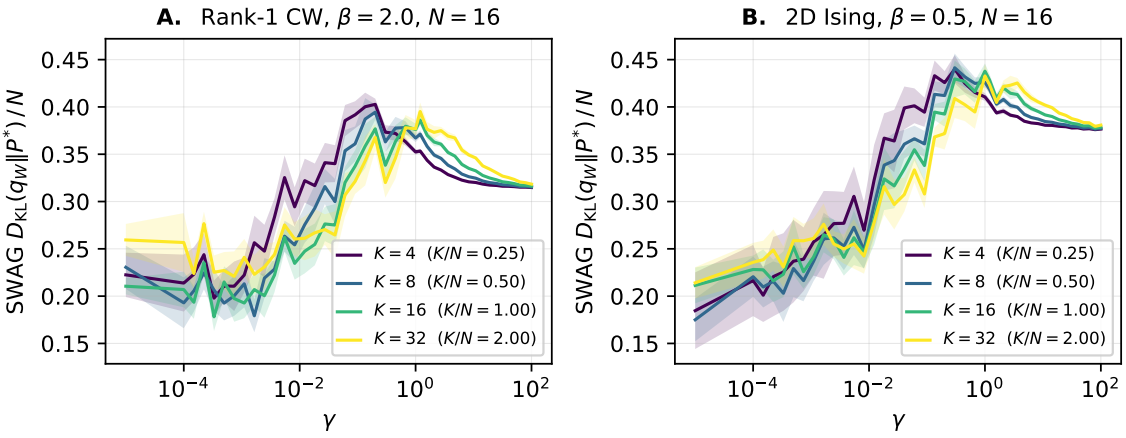

Figure 17: **Double descent in a binary FVSBN persists across K on rank-1-dominated teachers.** SWAG posterior-averaged reverse KL per spin, $D_{\mathrm{KL}}(q_W \| P^*)/N$, vs. the paper-convention regularization γ, for the binary FVSBN trained on **A)** a rank-1 Curie–Weiss / Hopfield teacher at $\beta = 2.0$, $N = 16$, 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.

The SWAG-averaged reverse KL per spin is shown in Fig. 17. On both teachers the double-descent bump above the $\gamma = 0$ baseline persists across the entire K range, in the band $+0.14$ to $+0.18$ nats/N (CW) and $+0.22$ to $+0.27$ nats/N (Ising). The peak position γ_{peak} shifts from $\gamma \approx 0.30$ at $K \leq 8$ to $\gamma \approx 1$ at $K \geq 16$, tracking the retarded-learning threshold $\gamma \sim \eta c_1^2$ with $c_1 = \lambda_1(C) \propto K$. The peak amplitude is essentially K-independent. Spatial geometry is not load-bearing: the geometry-free rank-1 Hopfield teacher of panel A and the 2D Ising NN lattice of panel B yield the same SWAG-DD signature, with no qualitative change.

G.2.3. 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, the same mechanism that drives the SBM double descent at the $h \neq 0$ to $h = 0$ boundary.

Data. We use the Kenneth R. French 49 industry portfolio daily value-weighted returns [20, 34], 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) [56] 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 [31] burn-in iterations followed by 2 000 SGD iterations at constant learning rate (the latter phase used for SWAG snapshot collection [38], though posterior averaging turns out to have a negligible effect here). The log-partition function is estimated by annealed importance sampling [52] 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 18 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 (≈ 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 (≈ 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. 9): 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.

Unlike the flow experiment of Fig. 16, the bump here is visible already in the MAP (point-estimate) test NLL, without requiring Bayesian posterior averaging. The reported MAP is the end-point 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 (Jensen gap $\lesssim 0.1$ nats), bounding this noise well below the bump amplitude $\Delta\text{NLL} \approx 0.3$ nats/day and confirming that

the RBM posterior is narrow around the MAP on this dataset. 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.3. Tempered posterior effects

G.3.1. BAYESGAN

The closed-form analysis of Fig. 10 predicts that the posterior-predictive reverse KL of an SBM student has a γ-dependent optimum $\eta_\star(\gamma)$: warm ($\eta_* < 1$) at small γ and cold ($\eta_* > 1$) at large γ. We test whether the same warm-to-cold migration is visible in a deep generative model, taking the Bayesian GAN of [51] as a convenient example. Tempering is straightforward, even though the posterior is genuinely multimodal in weight space, as the original sampler (SGHMC [10]) admits the tempered target without modification.

Setting. We replace the joint posterior of [51] 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 [10] 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 [51] is run with J_g=4 generator chains and J_d=1 discriminator chain, M=2 inner steps, friction α=0.1, step size ε=5×10^{-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 [51]. The predictive density $P_{\rm pp}$ is estimated by isotropic Gaussian KDE ($\sigma_{\rm KDE}$=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 $\sigma_{\rm target}$=0.5. Figure 19 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. 10 (warm at small γ, cold at large γ) therefore maps onto warm at large σ_p, cold at small σ_p. In $D_{\rm KL}(P_{\rm pp}\|P^*)$ (panel A), the argmin shifts as predicted: from $\eta_\star$=0.3 at weak priors ($\sigma_p \geq 1$) to $\eta_\star$=3 at the strongest prior (σ_p=0.3). The forward $D_{\rm KL}(P^*\|P_{\rm pp})$ (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 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 $P_{\rm pp}$ leaves the support of P^* uncovered, while the reverse KL in panel A stays bounded since P^* is positive everywhere.

G.4. Out-of-equilibrium training dynamics in a Potts Boltzmann machine

The training dynamics of Sec. 2.2 predicts that the stationary sample–data overlap $s_{\rm st}$ 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 [14], 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

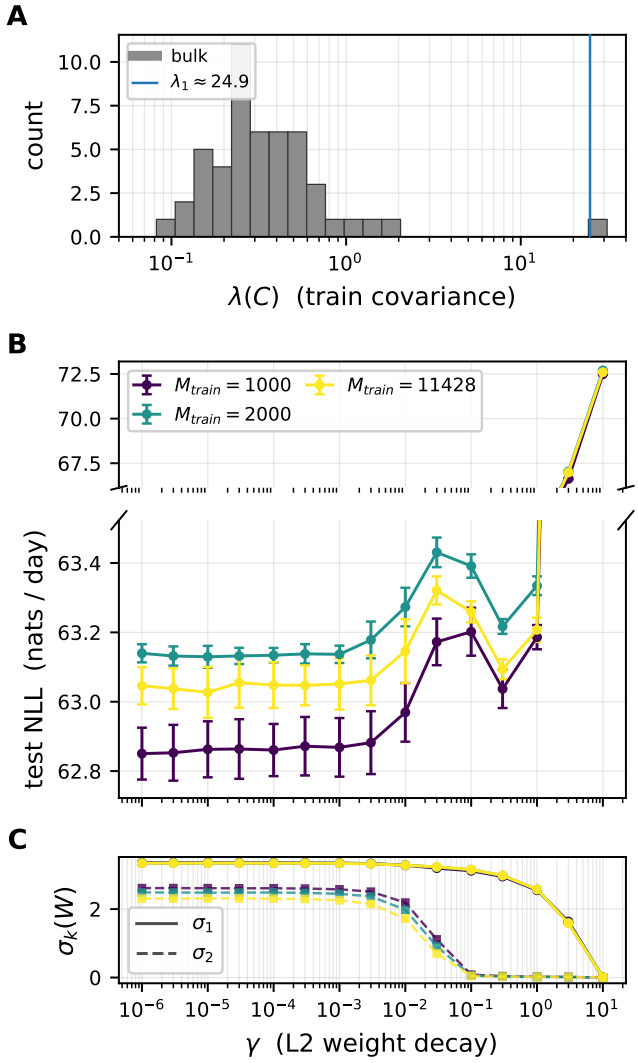

Figure 18: **Double descent in a Gaussian-visible RBM trained on financial return data. A)** Eigenvalue spectrum of the empirical training-set covariance C on a log-λ axis. The bulk (gray histogram) spans $\lambda \sim 10^{-1}$ to $\lambda \sim 2$, dominated by sectoral correlations; a single rank-one 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.

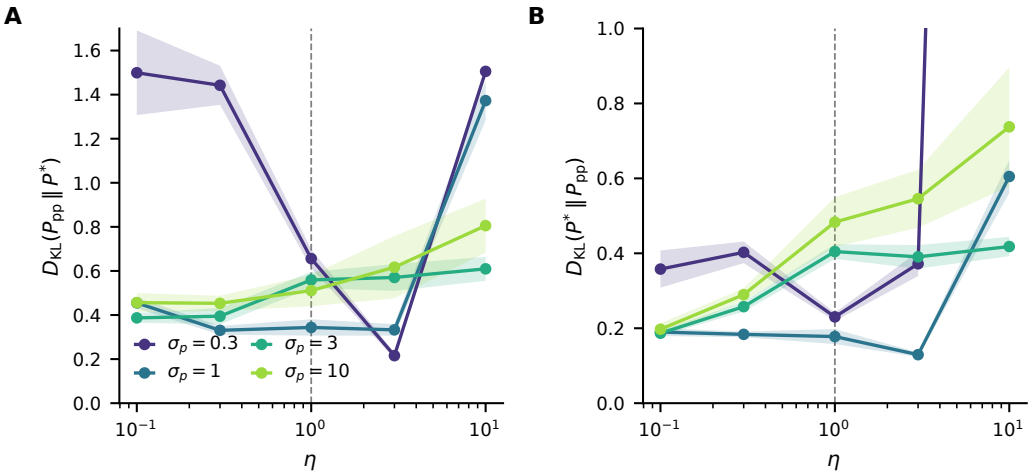

Figure 19: **Tempered Bayesian GAN on a unimodal target: warm-to-cold migration with prior strength. A)** Posterior-predictive reverse KL, $D_{\mathrm{KL}}(P_{\mathrm{pp}}\|P^*)$, versus η, for four prior widths $\sigma_p \in \{0.3, 1, 3, 10\}$. **B)** Forward $D_{\mathrm{KL}}(P^*\|P_{\mathrm{pp}})$. 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).

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 $\sigma_a^2 = \langle(\mathrm{oh}(\mathbf{x}) - f_1)^\top \mathbf{c}_a\rangle^2$ 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 (1) 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 (66) regardless of ν.

Results. Figure 20A 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. 20A).

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

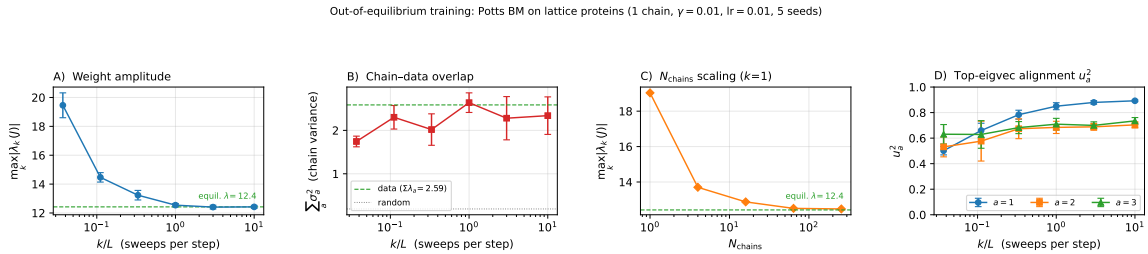

Figure 20: **Out-of-equilibrium training in a Potts BM.** A Potts Boltzmann machine (L=27, q=20) is trained by PCD-MAP on lattice-protein sequences (β_{sel}=1000), with γ=0.01, η_{lr}=0.01, t_{age}=5000, N_{chains}=1. **A)** Top coupling eigenvalue $\max_k |\lambda_k(J)|$ vs. sampling rate k/L (site updates per gradient step, normalized by chain length). At small k the weights overshoot the equilibrium value (dashed line) because the single chain negative phase cannot cancel the data gradient. **B)** Chain variance in the data-spike directions, $\sum_a \sigma_a^2$, averaged over the last 20% of training. At k=1 the chain explores less of the data-aligned subspace ($\sum \sigma^2 \approx 1.7$) than at $k \geq L$ ($\sum \sigma^2 \approx 2.5$, matching the data eigenvalue sum $\sum \lambda_a = 2.59$, dashed line). **C)** Weight overshoot vs. N_{chains} at k=1: the bias vanishes as more chains average out the single-chain noise, consistent with the absence of visible OOE in standard DCA training ($N_{\text{chains}} \sim 500$) on the families studied here. **D)** Per-mode best alignment $u_a^2 = \max_k(\hat{v}_k(J) \cdot \hat{\mathbf{c}}_a)^2$: the wrap-figure signature $u_1^2 < 1$ generalizes within the multi-mode subspace, with u_1^2 rising from 0.50 at k=1 to 0.89 at $k = 10L$ and the subleading modes degrading by comparable amounts at small k. Error bars in A,B,D: $\pm 1\sigma$ over 5 seeds.

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 $N_{\text{chains}} \approx 64$. With many chains the gradient estimate becomes unbiased regardless of individual-chain equilibration, washing out the OOE signal. 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 mechanism.

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.1.2; $t_{\text{age}} = 10^4$, three seeds) reproduces the three OOE signatures (Fig. 21): 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.

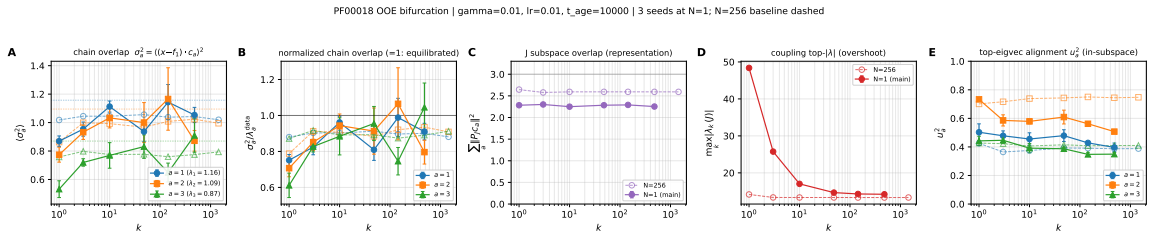

Figure 21: **Out-of-equilibrium training dynamics on real protein data (SH3 domain PF00018).** Same protocol as Fig. 20, applied to a pairwise Potts BM trained on the Pfam family PF00018 (L=48, q=21, M=1.76×10^5, $M_{\rm eff}\approx 1.33\times10^4$) with PCD–MAP, γ=0.01, $\eta_{\rm lr}$=0.01, $t_{\rm age}$=10^4. Solid markers with error bars: $N_{\rm chains}$=1 main sweep, averaged over 3 seeds; open markers (dashed): $N_{\rm chains}$=256 equilibrium baseline. Data directions $\mathbf{c}_1,\mathbf{c}_2,\mathbf{c}_3$ are the top three eigenvectors of the reweighted centered one-hot MSA covariance, with eigenvalues $\lambda_1^{\rm data}$=1.16, $\lambda_2^{\rm data}$=1.09, $\lambda_3^{\rm data}$=0.87 (dotted horizontal lines in panel A). Time averages are taken over the second half of training. **A)** Absolute chain variance σ_a^2 in each data-spike direction. **B)** Normalized chain variance $\sigma_a^2/\lambda_a^{\rm data}$. Below $k\approx L/5$ the ratio sits systematically below 1 (chain has not equilibrated in the data subspace); above, it fluctuates about 1 for all three modes. **C)** Subspace overlap $\sum_a\|{\rm Proj}_J\mathbf{c}_a\|^2$ (maximum 3): k-independent at both $N_{\rm chains}$=1 and 256, confirming the "representation without generation" prediction that J acquires the data directions regardless of the sampler's equilibration state. **D)** Top coupling eigenvalue $\max_k|\lambda_k(J)|$. At k=1 the $N_{\rm chains}$=1 weights overshoot the equilibrium value $\lambda_{\rm eq}\approx 13.3$ (reached at $N_{\rm chains}$=256, dashed) by a factor ≈ 3.6, matching the lattice-protein signature of Fig. 20. **E)** Per-mode alignment $u_a^2=\max_k(\hat v_k(J)\cdot\hat{\mathbf{c}}_a)^2$. The a=2 direction sits below the $N_{\rm chains}$=256 baseline by ~ 0.20 throughout the swept k-range, while a=1, 3 are within the seed scatter; the per-mode alignment loss is regime-dependent and weaker than on the lattice-protein teacher because the three top data eigenvalues are nearly degenerate ($\lambda_{1,2,3}^{\rm data}=1.16, 1.09, 0.87$).

Appendix 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. 8 (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. 9 (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. 10 (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. 12 (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).