WEIGHTING SCHEDULES GOVERN WHAT AND WHEN SCORE-BASED GENERATIVE MODELS LEARN FROM MULTIMODAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based generative models generate new samples by integrating a time-dependent drift that carries Gaussian noise onto the target distribution. In practice this drift is modeled by a neural network, trained on a loss integrated over time t with a weighting schedule $w(t)$. Along the backward dynamics, and for multi-modal distributions, trajectories commit to modes of the target within a narrow time window, the *speciation time*. In this work, focusing on high-dimensional data, we decompose the integrated loss into its single-time contributions and analyze each at fixed signal-to-noise ratio $\Lambda(t)$: we show that $\Lambda(t)$ sets the rate at which each feature of a multimodal target—the mode directions and their relative weights—is acquired during training. Crucially, at high $\Lambda(t)$ all mode directions are acquired together, on a single timescale insensitive to their amplitudes, while the relative weights are not learned at all. Only near the speciation time, where $\Lambda(t)$ becomes of order one, do all features become learnable, each on its own timescale: the weights are acquired jointly with the directions, and the directions at rates set by their relative amplitudes. For models trained on time-integrated objectives, the learning dynamics is then governed by how much of the weighting effectively sits near the speciation time, which provides insights on $w(t)$ design choices. These results follow from an exact high-dimensional analysis of the training dynamics of unbalanced and hierarchical Gaussian mixtures. Numerical experiments on image and human genome haplotype generation recover the predicted hierarchy of learning timescales in more complex settings.

1 INTRODUCTION

Over the last decade, the fast-paced development of transport-based generative models has led to tremendous progress in diverse and complex sampling tasks across both natural and artificial modalities. Vision modeling represents one such novelty, with latent diffusion models constituting state-of-the-art image (Dhariwal & Nichol, 2021; Rombach et al., 2022) and video generators (Liu et al., 2024; Yang et al., 2025). In the physical and chemical sciences, Boltzmann generators (Noé et al., 2019) offer an alternative to traditional sampling methods such as MCMC, either directly replacing them or augmenting them (Gabrié et al., 2022; Grenioux et al., 2025; Rehman et al., 2026). While this empirical progress has repeatedly outpaced our ability to explain it, significant leverage still lies in a precise theoretical understanding of training and the subsequent generative process.

The unified framework of Stochastic Interpolants (SI, Albergo & Vanden-Eijnden, 2023; Albergo et al., 2025) proposes a shared theoretical setting to discuss score-based generative models. Building such models amounts to optimizing a drift field driving Gaussian samples towards the target distribution, with generation quality directly impacted by practitioners' **training** and **sampling** choices (Karras et al., 2022). Within this paradigm, the statistical physics analysis of reverse time dynamics (Biroli et al., 2024) underlying the generation process has led to the identification of a narrow critical timescale dubbed speciation time: the brief window in which the generative trajectory commits to one mode of the target distribution, plainly identifying where compute effort should be spent during the **sampling phase.** While the practical impact of this result is undeniable, it assumes the optimal drift to be known. In practice that drift is learned, and which features of the target the converged score actually resolves is decided by the choices made during the **training phase**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Chief among those is the weighting schedule, which sets the noise levels the objective emphasizes and drives much of the reported progress in sample quality and training efficiency (Choi et al., 2022; Hang et al., 2023; Esser et al., 2024). These gains, however, rest largely on empirical tuning: how the weighting schedule governs the acquisition of data features during training still lacks a theoretical account. We bridge this gap by considering the training dynamics of two controlled families of structured data distributions: unbalanced and hierarchical Gaussian Mixture Models (GMM). We show that the design of the training objective quantitatively sets the rate at which each feature — mode direction, relative weight, mode amplitude — is acquired. In turn we empirically demonstrate that these results generalize to standard SI pipelines, with direct implications for the choice of weighting schedule.

Main contributions. Stochastic interpolant optimization is generically recast as minimizing

$$\mathcal{L}(\boldsymbol{\theta}) = d^{-1}\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\xi}}\left[\int_0^{t_{\max}} \mathrm{d}t w(t)\left\|\beta(t)\mathbf{s}_{\boldsymbol{\theta}}\left(\alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\xi},\, t\right) + \boldsymbol{\xi}\right\|^2\right] \tag{1}$$

with respect to the model parameters $\boldsymbol{\theta}$, where $\mathbf{x}_0$ is drawn from the target distribution P_0 and $\boldsymbol{\xi}$ from a standard Gaussian $\mathcal{N}(0, \boldsymbol{I}_d)$. With unlimited training time and exact minimization, the integrated loss is minimized by the exact score at every noise level, whatever the weighting schedule w. In practice, however, training is stopped after a finite time and is at finite precision. The noising schedule (α, β) and weighting schedule w then constitute important design choices whose influence on learning dynamics for high-dimensional data is the primary concern of this paper.

- **Training at fixed noise level -** For both unbalanced and structured GMM variants characterized by finitely many modes with distinct directions and weights, we analyze the training dynamics of a network $\mathbf{s}_{\boldsymbol{\theta}}$ specifically tailored to the GMM datasets at a single noising time t^*. The timescale on which each mode direction and mode weight are learned critically depends on t^*. In particular, for t^* small with respect to the speciation time t_s, only the mode directions are recovered, on a timescale insensitive to the mode amplitudes, while focusing training around t_s leads to weights and directions being learned jointly, at rates set by the modes' relative amplitude.
- **Influence of the weighting schedule -** Using the signal-to-noise ratio as the shared noising metric, we precisely show how the choice of weighting schedule w imposes effective weights for the relevant noise levels. As a result, the dominant noising regions control the learning dynamics of mode direction and weights. Importantly, we emphasize that Diffusion and Flow Matching (FM) correspond to different effective weights, resulting in distinct learning dynamics.
- **Empirical validation on real datasets -** We evaluate our results against i) handmade unbalanced and hierarchical MNIST dataset and ii) Haplotypes from the Human Genome Dataset (HGD). We sweep over a breadth of SI frameworks, completing the GMM Denoising Score Matching experiments with Flow Matching, as well as the more recent Masked Discrete Diffusion framework for the HGD, and recover quantitative agreement between theory and the empirical timescales governing hierarchical feature emergence.

RELATED WORKS

Where structure appears at generation time. Mode commitment during the generation dynamics, first identified in Raya & Ambrogioni (2023), is thoroughly analyzed in Biroli & Mézard (2023); Biroli et al. (2024): they introduce the speciation time, characterize it analytically for high-dimensional GMMs through overlap summary statistics, and generalize it to arbitrary distributions. The analysis is further extended to mixtures of strongly log-concave densities in Li & Chen (2024). Behjoo & Chertkov (2025); Sclocchi et al. (2025) probe the same windows empirically through forward-backward U-turn protocols on pretrained denoisers. All of these assume the score to be exact and locate a narrow commitment window along the *sampling* trajectory. The complementary question, at what *training* time each feature becomes available, is the one we address.

Where structure appears at training time. A series of works characterizes the optimal score reachable at finite sample complexity for a given neural network architecture, without studying

the training dynamics: Cui et al. (2024) consider a two-layer autoencoder at a fixed noising time for a symmetric bimodal GMM, completed by Aranguri & Insulla (2024), who study mode weight recovery for asymmetric GMMs, while George et al. (2026) derive precise learning curves for random feature networks. Another line of work focuses on the training dynamics itself. Outside the diffusion paradigm, Bachtis et al. (2024) follow the training dynamics of energy-based models on hierarchically structured data, and find the levels of the hierarchy to be learned successively. Within the diffusion framework, Cui et al. (2025) extend the two-layer autoencoder to all noising times, at the price of an encoding of time with low expressivity, Nicoletti et al. (2026) track how data structure and imbalance shape the dynamics in a random feature model, and Bardone et al. (2026) show that moments of a distribution are learned sequentially. Closest to our question, Wang & Pehlevan (2026) find the covariance eigendirections of a Gaussian target to display sequential alignment during training. Whereas all the structure lies in the covariance, we here focus on multimodal structure.

Noising and weighting schedules. Kingma & Gao (2023) show that most standard diffusion objectives are the ELBO under different weightings; Choi et al. (2022), Karras et al. (2022); Hang et al. (2023) and Esser et al. (2024) propose state-of-the-art SNR-based schedules and weighting functions. On the theory side, Aranguri et al. (2025) optimize noise schedules in high dimension, while Aranguri & Insulla (2024) analyze mode weight recovery using ad-hoc weighting schedules, and Gagneux et al. (2025) study the influence of the choice of weighting scheme on the performance of trained neural networks. Instead, we here focus on learning of mode weights and hierarchy in standard SI pipelines and analyze the consequences of the resulting weighting schemes.

2 SETTING

Stochastic interpolants. The Stochastic Interpolant (SI) framework (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2025) encompasses both Diffusion Models (DM) (Sohl-Dickstein et al., 2015; Song et al., 2021; Ho et al., 2020) and Flow Matching (FM) (Lipman et al., 2023). A SI defines a bridge between a target distribution P_0 and Gaussian white noise by introducing the interpolation

$$\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\xi}, \tag{2}$$

where $(\mathbf{x}_0, \boldsymbol{\xi}) \sim P_0 \times \mathcal{N}(0, \boldsymbol{I}_d)$, $t \in [0, t_{\max}]$[1], and the schedule functions (α, β) defining the SI obey the boundary conditions $\alpha(0) = 1,\ \beta(0) = 0,\ \alpha(t_{\max}) = 0,\ \beta(t_{\max}) = 1$. By construction, the marginal distribution $\mathbf{x}_t \sim P_t$ interpolates between P_0 and $\mathcal{N}(0, \boldsymbol{I}_d)$. Following the pioneering works of Anderson (1982) and Haussmann & Pardoux (1986), the interpolant $\mathbf{x}_t$ can also be written as the solution at time t of the deterministic flow

$$\dot{\mathbf{x}}(t) = \mathbf{b}(\mathbf{x}(t), t), \quad \mathbf{x}(t_{\max}) \sim \mathcal{N}(0, \boldsymbol{I}_d) \tag{3}$$

integrated backward in time, with velocity field

$$\mathbf{b}(x, t) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}] = \frac{\dot{\alpha}(t)}{\alpha(t)}\,\mathbf{x} - \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\beta(t)\,\nabla_{\mathbf{x}} \log P_t(\mathbf{x}), \tag{4}$$

so that integrating equation 3 generates new samples from P_0. The term $\mathbf{s}^*(\mathbf{x}, t) \equiv \nabla_{\mathbf{x}} \log P_t(\mathbf{x})$, dubbed the score function, is generally not available in closed form, but it is the minimizer of the Denoising Score Matching (DSM) loss (Hyvärinen, 2005; Vincent, 2011)

$$\mathcal{L}_{\mathrm{DSM}}(\mathbf{s}) = \tfrac{1}{d}\mathbb{E}_{(\mathbf{x}_0, \boldsymbol{\xi}) \sim P_0 \times \mathcal{N}(0, \boldsymbol{I}_d)} \left[\int_0^{t_{\max}} \mathrm{d}t w(t) \|\beta(t)\mathbf{s}\left(\mathbf{x}_t, t\right) + \boldsymbol{\xi}\|^2 \right], \tag{5}$$

where the weighting function $w(t)$ is arbitrary. In practice the score function is approximated by a neural network with parameters $\boldsymbol{\theta}$ and optimized with local methods such as stochastic gradient descent (Robbins & Monro, 1951).

Schedule and weighting functions (α, β, w) are crucial design choices: they set the relative weights of various noise levels and thus which features of the target distribution P_0 can possibly be recovered. The precise interplay between these choices and the training or sampling dynamics remains poorly understood, and the few results available assume an exact, optimally trained score (Biroli & Mézard, 2023; Biroli et al., 2024; Aranguri et al., 2025). They underlie our own analysis and are recalled below.

[1] For instance $\alpha(t) = e^{-t},\ \beta(t) = \sqrt{1 - e^{-2t}},\ t_{\max} = \infty$ for DM and $\alpha(t) = 1 - t,\ \beta(t) = t,\ t_{\max} = 1$ for FM.

Data distributions. We focus on two Gaussian mixtures with closed-form scores, one unbalanced and one hierarchical (see Fig. 1 and 2 for illustrative histograms):

- $(\mathcal{D}_1)$, an unbalanced bimodal distribution,

$$P_0(\mathbf{x}) = \omega\,\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + (1-\omega)\,\mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d), \quad ||\boldsymbol{\mu}||^2 = d, \quad \omega, \sigma^2 = O_d(1),$$
$$\nabla \log P_t(\mathbf{x}) = -\frac{\mathbf{x}}{\Gamma_t} + \frac{\alpha(t)\boldsymbol{\mu}}{\Gamma_t} \tanh\left(\frac{\alpha(t)\,\boldsymbol{\mu}^T\mathbf{x}}{\Gamma_t} + \tfrac{1}{2}\log\tfrac{\omega}{1-\omega}\right), \tag{6}$$

  with $\Gamma_t = \alpha^2(t)\sigma^2 + \beta^2(t)$.

- $(\mathcal{D}_2)$, a quadrimodal distribution inspired by Bachtis et al. (2024),

$$P_0(\mathbf{x}) = \frac{1}{4}\sum_{s_1=\pm1}\sum_{s_2=\pm1} \mathcal{N}(s_1\boldsymbol{\mu}_1 + s_2\boldsymbol{\mu}_2,\, \sigma^2\boldsymbol{I}_d),$$
$$\nabla \log P_t(\mathbf{x}) = -\frac{\mathbf{x}}{\Gamma_t} + \frac{\alpha(t)\boldsymbol{\mu}_1}{\Gamma_t}\tanh\left(\frac{\alpha(t)\,\boldsymbol{\mu}_1^T\mathbf{x}}{\Gamma_t}\right) + \frac{\alpha(t)\boldsymbol{\mu}_2}{\Gamma_t}\tanh\left(\frac{\alpha(t)\,\boldsymbol{\mu}_2^T\mathbf{x}}{\Gamma_t}\right), \tag{7}$$

  where $\boldsymbol{\mu}_1 \in \mathbb{R}^d$ is supported on the first κd coordinates and $\boldsymbol{\mu}_2$ on the remaining $(1-\kappa)d$ with $\|\boldsymbol{\mu}_1\|^2 = \kappa d,\ \|\boldsymbol{\mu}_2\|^2 = (1-\kappa)d,\ \kappa = O_d(1)$. We work in the high-dimensional limit $d \gg 1$, for two reasons: realistic data are high-dimensional, and, as we show below, this is precisely the regime in which the choice of $w(t)$ matters most.

Diffusion Regimes and Speciation Time. Consider a symmetric bimodal target $P_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{I}_d)$ and a diffusion model, i.e. $\alpha(t) = e^{-t}$, $\beta(t) = \sqrt{1-e^{-2t}}$, $t \in [0, t_{\max}]$, for which the exact score $\mathbf{s}^*(\mathbf{x}, t)$ is known and the reverse-time drift equation 4 can be analyzed in closed form. Biroli et al. (2024) show that the mode overlap[2] $q_t = \boldsymbol{\mu}^T\mathbf{x}_t/d$ evolves in an effective time-dependent potential $V_t(q)$, whose shape sharply shifts around the speciation time $t_s \equiv \frac{1}{2}\log(d)$. In Regime I, corresponding to times $t \gg t_s$, V_t is effectively unimodal, and no information is retained about the bimodal structure of P_0. Conversely, in Regime II, $t \ll t_s$, V_t consists of two well-separated harmonic wells centered on ± 1: the bimodality of P_0 is fully resolved and the generated sample has committed to one of the modes. The speciation time t_s is the narrow window in which V_t develops a cusp that decides the generation outcome, and specifies where sampling compute is best spent when integrating equation 4. The speciation phenomenon has been confirmed in other theoretical models (Biroli & Mézard, 2023; Ambrogioni, 2023; Li & Chen, 2024) and in real cases by numerical experiments (Behjoo & Chertkov, 2025; Sclocchi et al., 2025).

Note however that these results only concern sampling and they do not address how a trained score approaches $\mathbf{s}^*$ as optimization of equation 5 is carried out, nor whether t_s plays a significant role, or whether the picture holds once modes carry structure beyond the simple bimodal balanced GMM - unequal mode weights, or a hierarchy of sub-modes at different scales. The aim of this paper is to study training dynamics under such structure, for general stochastic interpolants.

3 Analytical results

In the following, we study the training of a parametrized score model $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ by gradient descent on the population loss equation 5[3], in the infinitesimal learning rate limit known as Gradient Flow (GF)

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{DSM}}(\mathbf{s}_{\boldsymbol{\theta}}), \tag{8}$$

where the training time τ should not be confused with the sampling time t. The aim is to elucidate the dynamical timescale dependencies on the dimension d and structure parameters ω, κ. To isolate

[2] §A.1 works with the rescaled overlap $\tilde{q}_t = \sqrt{d}\,q_t = \boldsymbol{\mu}^T\mathbf{x}_t/\sqrt{d}$, there simply written q, which keeps the noise contribution $O_d(1)$ rather than $O_d(1/\sqrt{d})$ and is the natural scale for the signal-vs-noise argument; the two differ only by this rescaling.

[3] We work throughout with the population loss i.e. the expectation value is exact. This departs from a line of work on the empirical loss, which asks how many samples are needed to learn a given feature (Bardone et al., 2026)

dependencies between the schedules (α, β) and w, it is insightful to consider first a single noise level, equivalent to a Dirac weighting function $w(t) = \delta(t - t^*)$. In this case, the synthetic distributions $\mathcal{D}_{1,2}$, which allow for analytical treatment, explicitly showcase the sequential learning of structured data features.

3.1 SINGLE FIXED NOISING TIME

Score models. At fixed noising time t, we parametrize the trainable score as a two-layer network with skip connection and $\tanh$ activation: for $(\mathcal{D}_1)$,

$$\mathbf{s}_{\boldsymbol{\theta}=(c,b,\mathbf{w})}(\mathbf{x}, t) = c\mathbf{x} + \mathbf{w}\tanh(\mathbf{w}^T\mathbf{x} + b), \qquad c, b \in \mathbb{R},\ \mathbf{w} \in \mathbb{R}^d, \tag{9}$$

with the learned weight read out of the bias, $\rho = e^{2b}/(1 + e^{2b})$; and for $(\mathcal{D}_2)$,

$$\mathbf{s}_{\boldsymbol{\theta}=(c,\mathbf{w}_1,\mathbf{w}_2)}(\mathbf{x}, t) = c\mathbf{x} + \mathbf{w}_1\tanh(\mathbf{w}_1^T\mathbf{x}) + \mathbf{w}_2\tanh(\mathbf{w}_2^T\mathbf{x}), \tag{10}$$

where $\mathbf{w}_1$ lives in the same block dimensional space as $\boldsymbol{\mu}_1$ (resp. for $(\mathbf{w}_2, \boldsymbol{\mu}_2)$). Under the GF dynamics equation 8, we show in appendix (§A.2.1) and (§A.2.2) that $\mathcal{L}_{\mathrm{DSM}}$ depends on $\mathbf{w}$ (resp. $\mathbf{w}_1, \mathbf{w}_2$) only through its projection onto the mode direction(s) and its orthogonal component, so training reduces to a closed system of ODEs on the summary statistics

$$m = \frac{\mathbf{w}^T\boldsymbol{\mu}}{\alpha(t)\, d}, \quad q = \frac{\|\mathbf{w}^\perp\|^2}{\alpha(t)^2\, d} \qquad \left(\text{resp. } m_i = \frac{\mathbf{w}_i^T\boldsymbol{\mu}_i}{\alpha(t)\,\kappa_i d}, \quad q_i = \frac{\|\mathbf{w}_i^\perp\|^2}{\alpha(t)^2\,\kappa_i d},\ i = 1, 2\right) \tag{11}$$

together with c (and b for $\mathcal{D}_1$), where $\mathbf{w}^\perp$ is the component of $\mathbf{w}$ orthogonal to $\boldsymbol{\mu}$. The parameterizations equation 9 and equation 10 are expressive enough to represent the exact score. Our aim is therefore to determine whether, and on what timescales, the summary statistics reach their optimal values during training.

Signal-to-noise ratio and speciation time for generic SI. As shown in appendix (§A.1), the speciation time analysis carried out in Biroli et al. (2024) holds for generic SI schedules (α, β). Mode selection during sampling is now governed by the signal-to-noise (SNR) ratio

$$\Lambda(t) = \frac{\alpha^2(t)\, d}{\alpha^2(t)\sigma^2 + \beta^2(t)}, \tag{12}$$

redefining the speciation time by the convention $\Lambda(t_s) \equiv 1$ (equivalently $\alpha/\beta \sim d^{-1/2}$). The underlying principle is clear: mode mixing occurs at speciation and this analysis carries over to the training paradigm. Indeed, training deep in Regime II, $\Lambda_t = O(d)$, exposes the network only to noise levels at which the data structure is fully resolved[4]; training around speciation exposes it to meaningfully collapsed data structures. By characterizing training at a given Λ_t in Regime II and around speciation, for each of $(\mathcal{D}_1), (\mathcal{D}_2)$, we show that its choice determines not only how fast (m, q) are learned, but what is learned at all. Our findings are summarized in the following results.

Result 1 (Training dynamics for $(\mathcal{D}_1)$ (§A.2.1))**.** *Let $\mathbf{w}(0) \in \mathbb{S}^{d-1}$ with $c(0), b(0) = O_d(1)$, and let $(c, \mathbf{w}, b)$ evolve under GF equation 8 for the score model equation 9 at fixed Λ_t. In the $d \to \infty$ limit the d-dimensional flow closes on the summary statistics (m, q, c, b), which obey a finite system of ODEs given in §A.2.1. Specifically:*

- *(a)* ***Regime II*** *($\Lambda_t = O_d(d)$). On $\tau = O_d(1)$ times $c \to -1/(\Gamma_t + \alpha_t^2)$, while (m, q, b) stay at initialization. On $\tau = O_d(d)$ times (m, q, b) relax to $(1/\Gamma_t, 0, b(0))$ while c relaxes instantaneously to $c^*(m, q)$.*
- *(b)* ***Speciation*** *($\Lambda_t = O_d(1)$). On $\tau = O_d(1)$ times $c \to -1/\Gamma_t$ while on $\tau = O_d(d)$ times $(m, q, b) \to (1/\Gamma_t, 0, \frac{1}{2}\log\frac{\omega}{1-\omega})$.*

On $\tau = O_d(1)$ times only the skip connection c evolves; the model learns the score associated to $\mathcal{N}(0, (\Gamma_t + \alpha_t^2)\boldsymbol{I}_d)$, the best unimodal approximation to the bimodal target. Multimodality is then acquired on $\tau = O_d(d)$ times and in a regime-dependent fashion: in Regime II the network recovers the exact score equation 6 up to its bias, so the direction is learned but the imbalance is not (Fig. 1, *left*); at speciation, direction and imbalance are recovered jointly (Fig. 1, *middle*).

[4]This presupposes that the structural parameters ω themselves remain $O_d(1)$, contrary to the results presented in Aranguri & Insulla (2024)

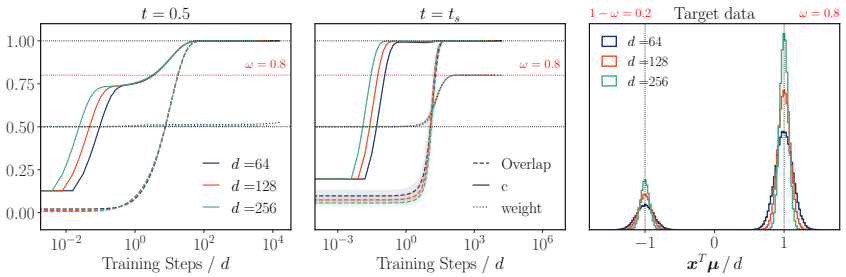

Figure 1: **Exact score parametrization for $(\mathcal{D}_1)$.** Overlap m (dashed), skip connection c (solid) and mode weight (dotted), against training steps rescaled by d, for the unbalanced GMM at $\omega = 0.8$, different values of d and two noising times, $t = 0.5$ *(left)* and $t = t_s$ *(middle)*. Typical bimodal histograms of projected samples are shown on the $(right)$ panel. In **Regime II** *(left)* c plateaus on $O(1)$ times and m converges on $O(d)$ times, while the weight stays frozen at its initial value $1/2$. At **speciation** *(right)*, c first reaches its final value and then weight and overlap rise together, on a shared $O(d)$ timescale.

Result 2 (Training dynamics for $(\mathcal{D}_2)$ (§A.2.2))**.** *Let $\kappa_1 = \kappa$, $\kappa_2 = 1 - \kappa$, $\mathbf{w}_1(0) \in \mathbb{S}^{\kappa_1 d-1}$, $\mathbf{w}_2(0) \in \mathbb{S}^{\kappa_2 d-1}$ and $c(0) = O_d(1)$. Let $(c, \mathbf{w}_1, \mathbf{w}_2)$ evolve under the GF equation 8 for the score model equation 10 at fixed Λ_t. In the $d \to \infty$ limit the d-dimensional flow closes on the summary statistics (m_1, q_1, m_2, q_2, c), each pair (m_i, q_i) obeying the system of Result 1 at $\omega = 1/2$, $b = 0$ and rescaled SNR $\gamma_i^2 = \kappa_i \Lambda_t$, the two systems being coupled only through c. Specifically:*

- *(a)* ***Regime II*** *($\Lambda_t = O_d(d)$). On $\tau = O_d(1)$ times $c \to -1/(\Gamma_t + \alpha_t^2)$, while (m_i, q_i) stay at initialization. On $\tau = O_d(d)$ times c relaxes instantaneously to $c^*(m_1, q_1, m_2, q_2)$ while $(m_i, q_i) \to (1/\Gamma_t, 0)$ on a κ-independent timescale.*
- *(b)* ***Speciation*** *($\Lambda_t = O_d(1)$). On $\tau = O_d(1)$ times $c \to -1/\Gamma_t$, the two systems decouple and, for $i = 1, 2$, $(m_i, q_i) \to (1/\Gamma_t, 0)$ on κ-dependent timescales τ_i. In the small κ limit, $\tau_1/\tau_2 \sim 1/\kappa$.*

As in the unbalanced case, on $\tau = O_d(1)$ times the model only adapts its skip connection and learns the best unimodal approximation of the target. The two directions are then acquired on a $O_d(d)$ timescale, in a regime-dependent fashion. In Regime II both overlaps evolve jointly, independently of the block asymmetry (Fig. 2, *left*); at speciation however each block evolves at its own SNR set by κ. Consequently, learning is hierarchical, with the dominant direction $\boldsymbol{\mu}_2$ acquired before $\boldsymbol{\mu}_1$ (Fig. 2, *middle*).

Since the dynamics of (m_i, q_i) are coupled, the rate of growth of m_i drifts as q_i relaxes, making its κ dependence intractable. To proceed further, we impose the extra constraint of fixed $\mathbf{w}_i$ norm.

Result 2 bis (Heuristic growth rate §A.2.3)**.** *Constrain each $\mathbf{w}_i$ to its target norm, $\|\mathbf{w}_i\| = \alpha(t)\|\boldsymbol{\mu}_i\|$, i.e. $m_i^2 + q_i = 1$. The small m_i linearized dynamics of Result 2(b) then reduces to the single ODE*

$$\dot{m}_i = d^{-1}\lambda(\gamma_i)\, m_i + O(m_i^3), \tag{13}$$

with $\lambda(\gamma)$ given in closed form in equation 103 of §A.2.3.

While this assumption is provably not exact, it is closed-form and recovers 2(b) exactly as $\kappa \to 0$. Consequently, we use $\tau_i \propto 1/\lambda(\gamma_i)$ as the reference timescale against which training time is rescaled in Fig. 2 and throughout.

3.2 Integrated loss

Standard SI pipelines optimize the integrated loss equation 5 rather than a single-time one, exposing the model to many noise levels at once. Since this loss is a weighted superposition of fixed-SNR denoising problems, Results 1 and 2 suggest that training depends on how much weight is placed on each SNR region, which we now make precise.

Rewriting a denoising objective in terms of SNR rather than time is now standard (Kingma & Gao, 2023), and underlies state-of-the-art schedule designs (Karras et al., 2022; Esser et al., 2024).

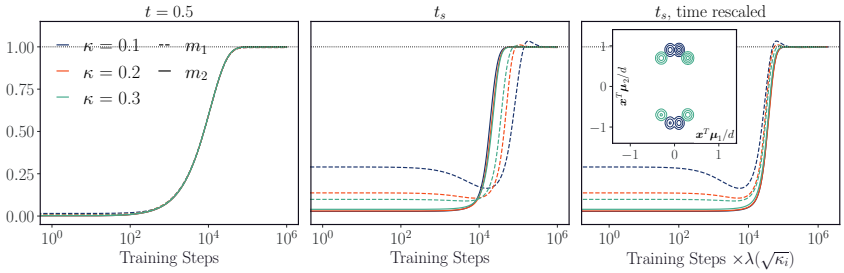

Figure 2: **Exact score parametrization for $(\mathcal{D}_2)$.** Block overlaps m_1 (dashed) and m_2 (solid), against training steps, for the quadrimodal GMM at $d = 1024$, structure parameter $\kappa \in \{0.1, 0.2, 0.3\}$ and two noising times, $t = 0.5$ *(left)* and $t = t_s$ *(middle and right)*. In **Regime II** *(left)* the two overlaps are learned on the same training timescale, independently of κ. At **speciation** *(middle)* they evolve on different timescales with $\boldsymbol{\mu}_1$ being learned after $\boldsymbol{\mu}_2$. Rescaling each block's training time by its own closed-form rate $\lambda(\gamma_i)$ of equation 13, with $\gamma_i^2 = \kappa_i \Lambda_t \Gamma_t$ and $\kappa_1 = \kappa,\ \kappa_2 = 1 - \kappa$, collapses the curves *(right)*. The inset shows $(\mathcal{D}_2)$ samples, projected on $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. The smaller κ, the closer the modes pair up along $\boldsymbol{\mu}_1$ and the more hierarchical the target.

Switching from $t \mapsto \Lambda$ in the DSM loss equation 5, we collect every schedule (α, β) and weighting w into a single effective SNR weight $w_{\text{eff}}(\Lambda) = w(t(\Lambda))|\mathrm{d}t/\mathrm{d}\Lambda|$, which fully characterizes a given SI framework (§A.3). In light of the fixed-SNR analysis, most of the structural information is passed to the model around speciation: the relevant signal peaks at $\Lambda = O_d(1)$, and comparing pipelines amounts to comparing the weight they place around that scale. Under this lens, standard Diffusion (DM) and Flow Matching (FM)[5] pipelines differ explicitly. For $\Lambda \ll d$,

$$w_{\text{eff}}^{\text{DM}}(\Lambda) \propto \Lambda^{-1} \qquad w_{\text{eff}}^{\text{FM}}(\Lambda) \propto \sqrt{d}\Lambda^{-3/2}. \tag{14}$$

Diffusion is scale-free and allocates the same denoising effort to every decade of SNR, exposing the network to a broad mixture of noise levels from Regime II down past speciation. Conversely, Flow Matching carries the extra factor $\sqrt{d/\Lambda}$ and tilts training towards low SNR. Note that this concentration is not a design choice but a consequence of the canonical FM schedule itself. We therefore expect the FM training dynamics to be dominated by speciation-specific effects, whereas Diffusion should display more mixed behavior. In turn, this predicts distinct structure learning signatures, which we test on $(\mathcal{D}_1)$ and $(\mathcal{D}_2)$ by replacing the analytical score of §3.1 with a generic ResNet denoiser trained under both FM and DM effective SNR weighting functions (§B.2). Neither pipeline is expected to reproduce the fixed-SNR analysis exactly — the network shares parameters across noise levels, so the fixed-SNR dynamics do not simply superpose — but we still find qualitative agreement.

On $(\mathcal{D}_1)$ (Fig. 3a), once two modes appear in the generated samples, uniform DSM first relaxes towards an almost even split, as predicted by Result 1(a), before converging to ω; its earlier rise in positive fraction is spurious, reporting a displaced mean while the histograms are still unimodal at (1). Under the FM weighting, weight and direction are instead acquired together, as at speciation. On $(\mathcal{D}_2)$ (Fig. 3b), the overlap with the subdominant direction $\boldsymbol{\mu}_1$ separates with κ under both weightings, more sharply under FM, for which rescaling training steps by $\lambda(\sqrt{\kappa})$ collapses the curves.

4 Structure Learning in Complex Data

While synthetic GMMs allow for a clear interpretation of the structural features composing the data, such identification in real world datasets is typically open to interpretation. Nevertheless we demonstrate that our theory captures learning dynamics across data modalities and SI design choices through three numerical experiments of increasing complexity. Precise details on all numerical experiments are provided in appendix (§B).

Weight learning in unbalanced MNIST. The paradigmatic multi-class MNIST dataset offers a natural route to test coordinated learning of class index and class proportion. To mimic a bimodal

[5]For Flow Matching the velocity regression must first be rewritten as a noise prediction, which contributes an extra Jacobian $1/\alpha^2(t)$ (§A.3).

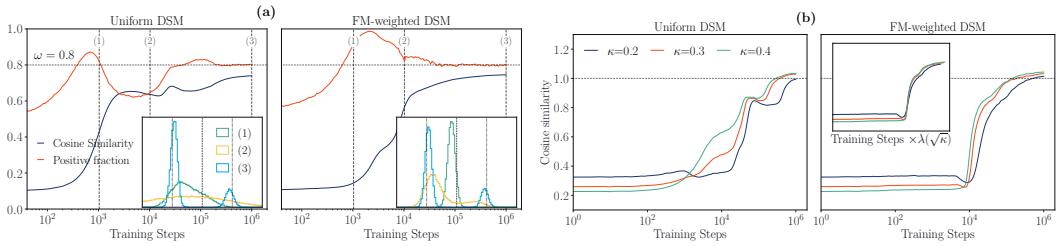

Figure 3: **ResNet for GMMs** at $d = 64$, under uniform DSM (*left*) and FM-weighted DSM (*right*) in each panel, against training steps. **(a)** ($\mathcal{D}_1$) at $\omega = 0.8$: cosine similarity with $\boldsymbol{\mu}$ (blue) and positive fraction of $p = \mathbf{x}^T\boldsymbol{\mu}/\|\boldsymbol{\mu}\|^2$ (red); insets show the histogram of p at the three marked steps. **(b)** ($\mathcal{D}_2$): cosine similarity with $\boldsymbol{\mu}_1$ for $\kappa \in \{0.2, 0.3, 0.4\}$; the inset replots the FM curves against training steps rescaled by $\lambda(\sqrt{\kappa})$.

distribution we prune MNIST to two classes $3/6$ with relative tunable proportion ω. We compare on Fig. 4a out-of-the-box DSM and FM implementations for a fixed imbalance $\omega = 0.8$ in favor of the 6s and use a pretrained digit classifier for evaluating empirical class proportion.

Under Flow Matching, class proportion and direction evolve jointly: as denoising progresses, the fraction of blurry images classified as 6s ramps up to ω concurrently with increasing image crispness, as expected of a pipeline overweighting low SNR regions. Unlike the controlled ResNet experiments discussed above, where only w changes, however, the comparison with DSM is inconclusive: proportion and direction appear to be learned on the same timescale under both objectives. We attribute this to the weaker mode separation of unbalanced MNIST, which shortens the relevant SNR range distinguishing the two weighting schemes (the two out-of-the-box pipelines also differ in more than w, §B.3.3).

Learning hierarchy in (Tinted MNIST/FM). To investigate hierarchical feature emergence during training we design a new dataset by adding a digit-orthogonal mode to the $3/6$ pruned MNIST. Specifically, we randomly colorize each image in either red or green with tunable intensity $\alpha \in (0, 1]$; see Fig. 6 for representative samples and appendix B.3.1 for details on the dataset construction. Importantly, the resulting dataset is reminiscent of $\mathcal{D}_2$, with orthogonal digit and color directions. To match the GMM analysis, the squared amplitude of the color direction scales as α^2 and can thus be made arbitrarily small.

Within the FM pipeline, training is dominated by the speciation region; consequently, digit learning occurs on a single α-independent timescale, whereas the color direction is acquired on an α dependent one, its onset spanning close to a decade across the sweep (Fig. 4b). We emphasize that even on this complex, image-based dataset, the GMM rescaling ansatz equation 13 still quantitatively captures the onset of color mode emergence, provided the experiment is matched to the single SNR analysis by identifying the small mode amplitude $|\boldsymbol{\mu}_1| \propto \alpha$.

Sequentially learning structured features in genomic data (HGD/MDLM). Departing from synthetic datasets, we finally consider haplotype data (binary sequences) from the Human Genome Dataset (HGD), following Bachtis et al. (2024). To test the scope of our theory, we consider yet another SI framework, Masked Discrete Diffusion, adapting the original implementation of Sahoo et al. (2024) to the binary data at hand (see App. B.4 for implementation details). Unlike in the GMM and tinted MNIST cases, feature structure in the HGD is not made obvious by construction. To assess sequential learning of data features we pre-compute the training data's PCA eigenvectors and track the statistics of inferred samples projected on these fixed directions, shown in Fig. 5.

The complexity of the HGD is made apparent in the left panel of Fig. 5, where the first and second eigenvector projections display clear multi-modality. The right panel shows that the generated data acquires the top PCA eigendirections sequentially, the leading two being separated by close to half a decade of training while the subsequent ones only begin to rise at the end of the run. Following the single SNR analysis, we rescale training time by $\lambda(\sqrt{v_k})$, with v_k the variance explained along direction k (App. B.4). The collapse holds for the two leading directions, precisely those along which the projected data is visibly multimodal (left panel), and degrades beyond: the smaller eigendirections

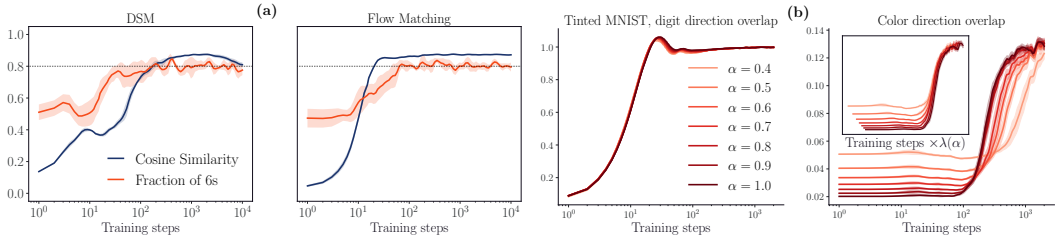

Figure 4: **Experiments on MNIST. (a)** Cosine similarity with the digit direction (blue) and fraction of 6s (orange), against training steps, for MNIST restricted to 3s and 6s, of which a fraction $\omega = 0.8$ are 6s (dotted line), under DSM *(left)* and FM *(right)*. Both objectives acquire direction and proportion within a single decade of training, MNIST offering too short an SNR range for the two weightings to separate. **(b)** Overlap with the digit direction *(left)* and with the color direction *(right)*, against training steps, for tinted MNIST trained with FM at tint intensities α, the dataset being built so that the squared amplitude of the color mode scales as α^2. The digit direction is learned at a rate nearly insensitive to α, while the color direction separates with it; rescaling training time by $\lambda(\alpha)$ collapses the color curves (inset).

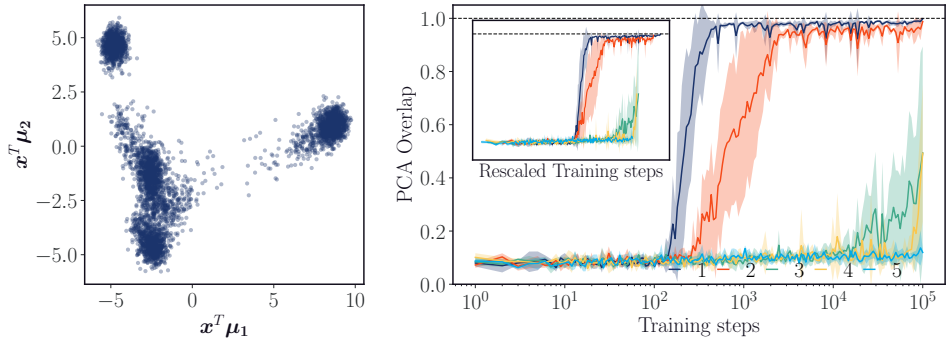

Figure 5: **Sequential feature learning on HGD.** *(Left)* HGD samples projected onto the first two PCA eigenvectors, showing a complex multimodal structure. *(Right)* Overlap of the generated samples with the leading PCA components $\boldsymbol{\mu}_i$, $i = 1, \ldots, 5$, against training steps, under Masked Discrete Diffusion: the directions are recovered one after the other, and rescaling training time by the closed-form rate $\lambda(\sqrt{v_i})$ of equation 13, with v_i the variance explained along the i-th reference PCA direction, collapses the two leading ones (inset).

carry no resolved modes, so the mode amplitude the ansatz rests on has no counterpart there. The MDLM loss is moreover scale free (App. A.5), which, as in the ResNet experiment of Fig. 3b, does not prevent sequential emergence.

5 Conclusion and Future Work

Reading a training objective through its effective SNR weighting turns an arbitrary design choice into a predictive one: which features a model resolves, and when, follows from where that weighting places its mass relative to the speciation scale. Gaussian mixtures make this concrete. Training deep in Regime II of Biroli et al. (2024) recovers every mode direction on a common timescale but leaves the relative weights untouched, however long one trains. Training around speciation instead acquires weights and directions jointly, and separates the directions themselves at rates set by their relative amplitudes. The same phenomenology holds for networks trained on unbalanced MNIST and, quantitatively, on tinted MNIST and genomic haplotype data.

The same statement, read backwards, is a design principle. Rather than inheriting a weighting schedule, one could tailor it to the structure of a given target, so as to acquire a chosen feature first or to reach a desired structure within a fixed compute budget. We leave this to future work.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15571.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://jmlr.org/papers/v26/23-1605.html.

Luca Ambrogioni. The statistical thermodynamics of generative diffusion models, 2023. URL https://arxiv.org/abs/2310.17467.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. URL https://doi.org/10.1016/0304-4149(82)90051-5.

Santiago Aranguri and Francesco Insulla. Phase-aware training schedule simplifies learning in flow-based generative models, 2024. URL https://arxiv.org/abs/2412.07972.

Santiago Aranguri, Giulio Biroli, Marc Mézard, and Eric Vanden-Eijnden. Optimizing noise schedules of generative models in high dimensions, 2025. URL https://arxiv.org/abs/2501.00988.

Dimitrios Bachtis, Giulio Biroli, Aurélien Decelle, and Beatriz Seoane. Cascade of phase transitions in the training of energy-based models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55591–55619, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/648a5a590ca6f2bb5de53f938e230160-Paper-Conference.pdf.

Lorenzo Bardone, Claudia Merger, and Sebastian Goldt. A theory of learning data statistics in diffusion models, from easy to hard, 2026. URL https://arxiv.org/abs/2603.12901.

Hamidreza Behjoo and Michael Chertkov. U-turn diffusion. *Entropy*, 27(4):343, March 2025. ISSN 1099-4300. doi: 10.3390/e27040343. URL http://dx.doi.org/10.3390/e27040343.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(9):093402, 2023. URL https://doi.org/10.1088/1742-5468/acf8ba.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. URL https://doi.org/10.1038/s41467-024-54281-3.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11462–11471, 2022. URL https://arxiv.org/abs/2204.00227.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. In *International Conference on Learning Representations*, pp. 51929–51955, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e45a448dfa778f6d62729a7bc8633c06-Paper-Conference.pdf.

Hugo Cui, Cengiz Pehlevan, and Yue Lu. A solvable model of learning generative diffusion: Theory and insights. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5253–5296, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/082d3d795520c43214da5123e56a3a34-Paper-Conference.pdf.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=OU98jZWS3x_.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.03206.

Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022. URL https://doi.org/10.1073/pnas.2109420119.

Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. The generation phases of flow matching: A denoising perspective, 2025. URL https://arxiv.org/abs/2510.24830.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves, 2026. URL https://arxiv.org/abs/2502.00336.

Louis Grenioux, Leonardo Galliano, Ludovic Berthier, Giulio Biroli, and Marylou Gabrié. Boltzmann generators for amorphous particle systems, 2025. URL https://arxiv.org/abs/2512.16607.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-SNR weighting strategy. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.09556.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4): 1188–1205, 1986. URL https://doi.org/10.1214/aop/1176992362.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL https://jmlr.org/papers/v6/hyvarinen05a.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.00848.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models, 2024. URL https://arxiv.org/abs/2403.01633.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. URL https://arxiv.org/abs/2402.17177.

Flavio Nicoletti, Chenxiao Ma, Enrico Ventura, Luca Saglietti, and Stefano Sarao Mannelli. The interplay of data structure and imbalance in the learning dynamics of diffusion models, 2026. URL https://arxiv.org/abs/2605.06367.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. URL https://doi.org/10.1126/science.aaw1147.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66377–66389, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d0da30e312b75a3fffd9e9191f8bc1b0-Paper-Conference.pdf.

Danyal Rehman, Charlie B. Tan, Yoshua Bengio, Avishek Joey Bose, and Alexander Tong. Autoregressive Boltzmann generators, 2026. URL https://arxiv.org/abs/2606.27361. ICML 2026 (spotlight).

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. URL https://doi.org/10.1214/aoms/1177729586.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.07524.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122 (1):e2408799121, 2025. URL https://doi.org/10.1073/pnas.2408799121.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. URL https://doi.org/10.1162/NECO_a_00142.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *Advances in Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=SDhOClkyqC.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

A Details on the theoretical results

A.1 Dynamical regimes and signal-to-noise ratio

In this section we adapt the computation of Biroli et al. (2024); Aranguri et al. (2025) to stochastic interpolants, in order to identify the speciation time i.e. the time at which the generative trajectory commits to one of the modes of the target. Let us consider as a target distribution a bimodal symmetric isotropic Gaussian distribution

$$P_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d), \qquad \|\boldsymbol{\mu}\|^2 = d. \tag{15}$$

Since the stochastic interpolant at time t is $\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\xi}$, the marginal P_t is also a bimodal symmetric isotropic mixture,

$$P_t = \tfrac{1}{2}\mathcal{N}(\alpha(t)\boldsymbol{\mu}, \Gamma_t \boldsymbol{I}_d) + \tfrac{1}{2}\mathcal{N}(-\alpha(t)\boldsymbol{\mu}, \Gamma_t \boldsymbol{I}_d), \qquad \Gamma_t = \alpha^2(t)\sigma^2 + \beta^2(t), \tag{16}$$

whose score reads

$$\nabla_{\mathbf{x}} \log P_t(\mathbf{x}) = -\frac{\mathbf{x}}{\Gamma_t} + \frac{\alpha(t)\boldsymbol{\mu}}{\Gamma_t} \tanh\left(\frac{\alpha(t)\, \boldsymbol{\mu}^T \mathbf{x}}{\Gamma_t}\right). \tag{17}$$

Signal-to-noise ratio. We study the evolution, during the generative dynamics, of the overlap

$$q_t = \frac{\boldsymbol{\mu}^T \mathbf{x}_t}{\sqrt{d}}. \tag{18}$$

Writing $\mathbf{x}_0 = s\boldsymbol{\mu} + \sigma\mathbf{z}$, with $s = \pm 1$ the label of the mode from which the sample is drawn and $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and projecting $\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\xi}$ on $\boldsymbol{\mu}/\sqrt{d}$ gives

$$q_t = \underbrace{s\,\alpha(t)\sqrt{d}}_{\text{signal}} + \underbrace{\alpha(t)\,\sigma\, g_1 + \beta(t)\, g_2}_{\text{noise}}, \qquad g_1, g_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{19}$$

where we used $\|\boldsymbol{\mu}\|^2 = d$ and the fact that $\boldsymbol{\mu}^T\mathbf{z}/\sqrt{d}$ and $\boldsymbol{\mu}^T\boldsymbol{\xi}/\sqrt{d}$ are standard Gaussians. The two noise contributions are independent, of total variance $\alpha^2(t)\sigma^2 + \beta^2(t) = \Gamma_t$. In turn, we define the time-dependent signal-to-noise ratio Λ_t as

$$\Lambda_t = \frac{\left(\mathbb{E}\left[q_t \mid s\right]\right)^2}{\mathrm{Var}\left(q_t \mid s\right)} = \frac{\alpha^2(t)\|\boldsymbol{\mu}\|^2}{\Gamma_t} = \frac{\alpha^2(t)\, d}{\alpha^2(t)\sigma^2 + \beta^2(t)}, \tag{20}$$

where Λ_t decreases from $\Lambda_t = d/\sigma^2$ at $t = 0$ to $\Lambda_t = 0$ at $t = t_{\max}$.

Effective potential. Project the generative ODE equation 4

$$\dot{\mathbf{x}}(t) = \mathbf{b}(\mathbf{x}(t), t), \qquad \mathbf{b}(\mathbf{x}, t) = \frac{\dot{\alpha}(t)}{\alpha(t)}\mathbf{x} - \left(\dot{\beta}(t) - \frac{\dot{\alpha}(t)}{\alpha(t)}\beta(t)\right)\beta(t)\nabla_{\mathbf{x}} \log P_t(\mathbf{x}) \tag{21}$$

onto $\boldsymbol{\mu}/\sqrt{d}$. Using equation 17 and $\boldsymbol{\mu}^T\mathbf{x} = \sqrt{d}\, q$ yields a closed equation for q,

$$-\frac{\mathrm{d}q_t}{\mathrm{d}t} = -\frac{\partial V_{\text{eff}}}{\partial q}(q_t), \tag{22}$$

where t is the forward time and with

$$V_{\text{eff}}(q) = \frac{A_t + B_t/\Gamma_t}{2}\, q^2 - B_t \log\cosh\left(\frac{\alpha(t)\sqrt{d}}{\Gamma_t}\, q\right) \tag{23}$$

$$= \frac{A_t + B_t/\Gamma_t}{2}\, q^2 - B_t \log\cosh\left(\sqrt{\Lambda_t/\Gamma_t}\, q\right) \tag{24}$$

where $A_t = \dot{\alpha}(t)/\alpha(t)$ and $B_t = \big(\dot{\beta}(t) - A_t\beta(t)\big)\beta(t)$. We realize that $\sqrt{\Lambda_t/\Gamma_t}$ is the scaling variable controlling the argument of the second term, and that Λ_t alone controls it wherever $\Gamma_t = O_d(1)$. Expanding the potential in low and high SNR limits, it reads, close to the origin $q = 0$,

$$V_{\text{eff}}(q) \simeq \begin{cases} \dfrac{1}{2}\left(A_t + \dfrac{B_t}{\Gamma_t}\,(1 - \Lambda_t)\right) q^2, & \Lambda_t \ll 1, \\[2ex] \dfrac{1}{2}\left(A_t + \dfrac{B_t}{\Gamma_t}\right) q^2 - B_t\, \dfrac{\alpha(t)\sqrt{d}}{\Gamma_t}\, |q|, & \Lambda_t \gg 1. \end{cases} \tag{25}$$

For $\Lambda_t \ll 1$ the potential is purely quadratic and the trajectory is not driven towards either mode. For $\Lambda_t \gg 1$ it acquires a linear cusp at the origin which makes $q = 0$ unstable: the trajectory commits to one of the two branches $q \gtrless 0$ and is carried towards the corresponding mode, reaching $q = \pm\sqrt{d}$ at the end of generation.

The transition between these two regimes occurs at some $\Lambda_t = O(1)$, which defines the speciation time t_s. We adopt the convention $\Lambda_{t_s} = 1$[6]. This yields

$$\frac{\alpha^2(t_s)\, d}{\alpha^2(t_s)\sigma^2 + \beta^2(t_s)} = 1 \qquad \Longleftrightarrow \qquad \frac{\alpha^2(t_s)}{\beta^2(t_s)} = \frac{1}{d - \sigma^2} \underset{d\gg 1}{\sim} \frac{1}{d}. \tag{26}$$

This generalizes the analysis of Biroli et al. (2024) to arbitrary stochastic interpolants.

Specific speciations times. Let us specialize this criterion for the two most widespread SI frameworks, namely Diffusion and Flow Matching.

- For Diffusion, $\alpha(t) = e^{-t}$ and $\beta(t) = \sqrt{1 - e^{-2t}}$, so that $\Gamma_t = 1 - e^{-2t}(1 - \sigma^2)$ and

$$\Lambda_t^{\mathrm{DM}} = \frac{e^{-2t} d}{1 - e^{-2t}(1 - \sigma^2)}, \qquad t_s = \frac{1}{2}\log\left(d + 1 - \sigma^2\right) \underset{d\gg 1}{\sim} \frac{1}{2}\log d, \tag{27}$$

  recovering the result of Biroli et al. (2024).

- For the linear interpolant of Flow Matching, $\alpha(t) = 1 - t$ and $\beta(t) = t$ with $t \in [0, 1]$, so that $\Gamma_t = (1 - t)^2\sigma^2 + t^2$ and

$$\Lambda_t^{\mathrm{FM}} = \frac{(1-t)^2 d}{(1-t)^2\sigma^2 + t^2}, \qquad \frac{t_s}{1 - t_s} = \sqrt{d - \sigma^2} \quad \Longleftrightarrow \quad 1 - t_s = \frac{1}{1 + \sqrt{d - \sigma^2}} \underset{d\gg 1}{\sim} \frac{1}{\sqrt{d}}. \tag{28}$$

A.2 Single time analysis

A.2.1 Unbalanced distribution: exact loss and gradient flow

This section is self-contained. We derive, for the score model actually used in our experiments, the DSM loss and gradient flow exactly as functions of a finite set of summary statistics, valid for any signal-to-noise ratio, and only then specialize to the two regimes of interest by scaling analysis.

Setup. The target is the unbalanced mixture

$$P_0 = \omega\,\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + (1 - \omega)\,\mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d), \qquad \|\boldsymbol{\mu}\|^2 = d, \qquad \omega, \sigma^2 = O_d(1), \tag{29}$$

and we model the score with a two-layer network with trainable skip connection,

$$\mathbf{s}_{\boldsymbol{\theta}=(c,\mathbf{w},b)}(\mathbf{x}, t) = c\,\mathbf{x} + \mathbf{w}\tanh\left(\mathbf{w}^T\mathbf{x} + b\right), \qquad c, b \in \mathbb{R},\ \mathbf{w} \in \mathbb{R}^d. \tag{30}$$

Note that equation 30 has no explicit dependence on the schedule (α, β): whatever amplitude and argument scale the exact score requires at the noising time t, the network must produce through training. We work at a fixed noising time t, writing $\alpha = \alpha(t)$, $\beta = \beta(t)$ and

$$\Gamma = \alpha^2\sigma^2 + \beta^2, \qquad \Lambda = \frac{\alpha^2 d}{\Gamma}, \qquad \gamma^2 := \alpha^2 d = \Lambda\Gamma, \tag{31}$$

where Λ is the signal-to-noise ratio and γ will turn out to be the natural scale controlling the $\tanh$. Training is gradient flow in training time τ on the single-time DSM loss,

$$\mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{d}\,\mathbb{E}_{\mathbf{x}_0\sim P_0,\ \boldsymbol{\xi}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\|\beta\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\xi}\|^2\right], \qquad \mathbf{x}_t = \alpha\mathbf{x}_0 + \beta\boldsymbol{\xi}, \qquad \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} = -\nabla_{\boldsymbol{\theta}}\mathcal{L}. \tag{32}$$

[6] Any other choice of $O(1)$ constant shifts t_s only at subleading order in d.

Summary statistics and initialization. Because equation 30 carries no α, the weight vector must itself grow to the scale set by the exact score. It is therefore convenient to measure $\mathbf{w}$ in units of α and define

$$m = \frac{\mathbf{w}^T \boldsymbol{\mu}}{\alpha\, d}, \qquad q = \frac{\|\mathbf{w}^\perp\|^2}{\alpha^2 d}, \qquad \text{so that} \qquad \|\mathbf{w}\|^2 = \alpha^2 d\left(m^2 + q\right), \tag{33}$$

with $\mathbf{w}^\perp$ the component of $\mathbf{w}$ orthogonal to $\boldsymbol{\mu}$. Equivalently, m and q are the usual overlap and orthogonal norm of the rescaled vector $\mathbf{u} = \mathbf{w}/\alpha$. Together with c and b, these are the only quantities the loss will depend on.

We initialize $c(0), b(0) = O_d(1)$ and draw $\mathbf{w}(0) = \mathbf{w}_0$ uniformly on the unit sphere $\mathbb{S}^{d-1}(1)$, as in our experiments. Then $\mathbf{w}_0^T \boldsymbol{\mu}$ has zero mean and variance $\|\boldsymbol{\mu}\|^2/d = 1$, while $\|\mathbf{w}_0^\perp\|^2 = 1 - O_d(1/d)$, so that

$$m_0 = O_d\!\left(\frac{1}{\alpha d}\right), \qquad q_0 = \frac{1}{\alpha^2 d}\left(1 + O_d\big(\tfrac{1}{d}\big)\right) = \frac{1}{\gamma^2} + O_d\big(\tfrac{1}{d}\big)\,. \tag{34}$$

The initial condition thus depends on the noising time through α alone. Everything below establishes the following statement, the long form of Result 1 of the main text.

Result (1, long form: training dynamics, unbalanced distribution)**.** *Let $\mathbf{w}(0)$ be uniform on $\mathbb{S}^{d-1}(1)$, $c(0), b(0) = O_d(1)$, and let $(c, \mathbf{w}, b)$ evolve under GF equation 8 for the score model equation 9 at fixed t. As $d \to \infty$:*

(a) ***Regime II*** *($\Lambda = O_d(d)$). On $\tau = O_d(1)$ times, $c \to c_1^* = -1/(\Gamma + \alpha^2)$ while (m, q, b) remain at initialization; on $\tau = O_d(d)$ times,*

$$\dot{q} = -\frac{4\beta^2}{d}\, q, \qquad \dot{m} = -\frac{2\beta^2 \Gamma}{d(\Gamma + \alpha^2)}\Big(m - \frac{1}{\Gamma}\Big), \tag{35}$$

so $(m, q) \to (1/\Gamma, 0)$, c is slaved to the value of (m, q) and b stays frozen at its initial value: the network recovers the exact score equation 6 up to its bias, and the imbalance ω is ***not learned****.*

(b) ***Speciation time*** *($\Lambda = O_d(1)$). Here $\Gamma \to 1$ and $\gamma^2 = \Lambda\Gamma = O_d(1)$, with initialization $q(0) = 1/\gamma^2 = O_d(1)$, $m(0) = O_d(1/\sqrt{d})$. On $\tau = O_d(1)$ times $c \to -1/\Gamma$. On $\tau = O_d(d)$ times, i.e. in the rescaled training time $\check{\tau} = \beta^2\tau/d$, the orthogonal component relaxes on its own while, at small (m, b), the overlap and the bias leave the origin jointly,*

$$\frac{\mathrm{d}q}{\mathrm{d}\check{\tau}} = -4q\,\Psi(q, \gamma) + O(m^2, b^2), \qquad \frac{\mathrm{d}}{\mathrm{d}\check{\tau}}\begin{pmatrix} m \\ b \end{pmatrix} = \mathbf{M}(q, \gamma)\begin{pmatrix} m \\ b \end{pmatrix} + O(m^2, b^2), \tag{36}$$

with the scalar Ψ given in equation 62 and the explicit 2×2 matrix $\mathbf{M}$ in equation 64. The flow converges to $q \to 0$, $m \to 1/\Gamma$ and $b \to b^* = \frac{1}{2}\log\frac{\omega}{1-\omega}$: the imbalance is learned jointly with the mode direction, on the same $\tau = O_d(d)$ timescale on which Regime II recovers the direction alone.

Detailed dynamics.

1. Effective noises. Decompose the data as $\mathbf{x}_0 = s\boldsymbol{\mu} + \sigma\mathbf{z}$, with $s = \pm 1$ with probabilities $\omega, 1 - \omega$ and $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and introduce

$$q_{\mathbf{z}} = \frac{\mathbf{w}^T \mathbf{z}}{\alpha\sqrt{d(m^2 + q)}}, \qquad q_{\boldsymbol{\xi}} = \frac{\mathbf{w}^T \boldsymbol{\xi}}{\alpha\sqrt{d(m^2 + q)}}, \qquad q_{\mathbf{z}}, q_{\boldsymbol{\xi}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \tag{37}$$

using $\|\mathbf{w}\|^2 = \alpha^2 d(m^2 + q)$ from equation 33. The projections of data and noise onto $\mathbf{w}$ then read

$$\mathbf{w}^T \mathbf{x}_0 = \alpha\left(s\, m\, d + \sigma\sqrt{d(m^2 + q)}\, q_{\mathbf{z}}\right), \qquad \mathbf{w}^T \boldsymbol{\xi} = \alpha\sqrt{d(m^2 + q)}\, q_{\boldsymbol{\xi}}. \tag{38}$$

2. Effective Loss. Inserting $\mathbf{x}_t = \alpha\mathbf{x}_0 + \beta\boldsymbol{\xi}$ into equation 30 and collecting terms along $\mathbf{x}_0$, $\boldsymbol{\xi}$ and $\mathbf{w}$,

$$\beta\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\xi} = \alpha\beta c\,\mathbf{x}_0 + \left(1+\beta^2 c\right)\boldsymbol{\xi} + \beta\,T\,\mathbf{w}, \qquad T := \tanh\left(\mathbf{w}^T\mathbf{x}_t + b\right). \tag{39}$$

Squaring, dividing by d, and using the concentration of $\frac{1}{d}\|\mathbf{x}_0\|^2 \to 1+\sigma^2$, $\frac{1}{d}\|\boldsymbol{\xi}\|^2 \to 1$ and $\frac{1}{d}\mathbf{x}_0^T\boldsymbol{\xi} \to 0$, together with equation 38, gives the effective loss

$$\begin{aligned}\mathcal{L} = &\left(1+\beta^2 c\right)^2 + \alpha^2\beta^2\left[c^2\left(1+\sigma^2\right) + \left(m^2+q\right)\mathbb{E}\left[T^2\right] + 2c\,m\,\mathbb{E}[s\,T]\right] \\ &+ \frac{2\alpha\beta\sqrt{m^2+q}}{\sqrt{d}}\Big(\alpha\beta c\,\sigma\,\mathbb{E}[q_z T] + \left(1+\beta^2 c\right)\mathbb{E}[q_{\boldsymbol{\xi}} T]\Big).\end{aligned} \tag{40}$$

While the last line of equation 40 carries an explicit $1/\sqrt{d}$, it is not subleading in d. Indeed, T is correlated with $q_{\mathbf{z}}, q_{\boldsymbol{\xi}}$, and Gaussian integration by parts (Stein's lemma), using $\partial T/\partial q_{\boldsymbol{\xi}} = \alpha\beta\sqrt{d(m^2+q)}\,(1-T^2)$ and $\partial T/\partial q_{\mathbf{z}} = \alpha^2\sigma\sqrt{d(m^2+q)}\,(1-T^2)$ from equation 38, gives

$$\mathbb{E}[q_{\boldsymbol{\xi}} T] = \alpha\beta\sqrt{d(m^2+q)}\,\mathbb{E}\left[1-T^2\right], \qquad \mathbb{E}[q_{\mathbf{z}} T] = \alpha^2\sigma\sqrt{d(m^2+q)}\,\mathbb{E}\left[1-T^2\right]. \tag{41}$$

such that each expectation carries a compensating $\sqrt{d}$. Substituting into equation 40, the explicit $1/\sqrt{d}$ cancels exactly and, using $1+\beta^2 c+\alpha^2\sigma^2 c = 1+c\,\Gamma$, the whole line collapses to the $O_d(1)$ contribution

$$2\alpha^2\beta^2\left(m^2+q\right)(1+c\,\Gamma)\,\mathbb{E}\left[1-T^2\right]. \tag{42}$$

yielding

$$\begin{aligned}\mathcal{L} = &\left(1+\beta^2 c\right)^2 \\ &+ \alpha^2\beta^2\left[c^2\left(1+\sigma^2\right) + \left(m^2+q\right)\mathbb{E}\left[T^2\right] + 2c\,m\,\mathbb{E}[s\,T]\right] \\ &+ 2\alpha^2\beta^2\left(m^2+q\right)(1+c\,\Gamma)\,\mathbb{E}\left[1-T^2\right].\end{aligned} \tag{43}$$

3. $\tanh$ **analysis.** From equation 38, $\mathbf{w}^T\mathbf{x}_t = \alpha^2 smd + \alpha\sqrt{d(m^2+q)}\,(\alpha\sigma q_{\mathbf{z}} + \beta q_{\boldsymbol{\xi}})$. The two independent Gaussians combine into a single $g \sim \mathcal{N}(0,1)$ with total variance $\alpha^2\sigma^2+\beta^2 = \Gamma$, so that, using $\alpha^2 d = \gamma^2$,

$$\mathbf{w}^T\mathbf{x}_t + b = s\,\gamma^2 m + \gamma\sqrt{\Gamma\,(m^2+q)}\;g + b. \tag{44}$$

Averaging over s with probabilities $\omega, 1-\omega$, and flipping $g \to -g$ in the $s=-1$ branch, the two expectations left in equation 40 reduce to one-dimensional Gaussian integrals,

$$h := \mathbb{E}[s\,T] = \omega\,I_1(b) + (1-\omega)\,I_1(-b), \qquad k := \mathbb{E}\left[T^2\right] = \omega\,I_2(b) + (1-\omega)\,I_2(-b), \tag{45}$$

$$I_1(b) = \mathbb{E}_g\Big[\tanh\Big(\gamma^2 m + \gamma\sqrt{\Gamma(m^2+q)}\,g + b\Big)\Big], \qquad I_2(b) = \mathbb{E}_g\Big[\tanh^2\Big(\gamma^2 m + \gamma\sqrt{\Gamma(m^2+q)}\,g + b\Big)\Big].$$

4. Exact loss. Collecting equation 40, equation 42 and equation 45,

$$\mathcal{L}(c,m,q,b) = \left(1+\beta^2 c\right)^2 + \alpha^2\beta^2\Big[c^2\left(1+\sigma^2\right) + \left(m^2+q\right)\Big(k + 2\,(1+c\,\Gamma)\,(1-k)\Big) + 2c\,m\,h\Big], \tag{46}$$

exactly, for any Λ, ω and b, with h, k given by equation 45.

5. Exact gradient flow. The flow acts on $\mathbf{w} \in \mathbb{R}^d$, while equation 46 depends on it only through m and q. From equation 33,

$$\nabla_{\mathbf{w}} m = \frac{\boldsymbol{\mu}}{\alpha d}, \qquad \nabla_{\mathbf{w}} q = \frac{2\mathbf{w}^{\perp}}{\alpha^2 d}, \tag{47}$$

so that, projecting $\mathrm{d}\mathbf{w}/\mathrm{d}\tau = -\nabla_{\mathbf{w}}\mathcal{L}$ onto $\boldsymbol{\mu}$ and onto $\mathbf{w}^{\perp}$ and using $\boldsymbol{\mu}^T\mathbf{w}^{\perp} = 0$, $\|\boldsymbol{\mu}\|^2 = d$ and $\|\mathbf{w}^{\perp}\|^2 = \alpha^2 dq$,

$$\frac{\mathrm{d}c}{\mathrm{d}\tau} = -\frac{\partial\mathcal{L}}{\partial c}, \qquad \frac{\mathrm{d}m}{\mathrm{d}\tau} = -\frac{1}{\gamma^2}\frac{\partial\mathcal{L}}{\partial m}, \qquad \frac{\mathrm{d}q}{\mathrm{d}\tau} = -\frac{4q}{\gamma^2}\frac{\partial\mathcal{L}}{\partial q}, \qquad \frac{\mathrm{d}b}{\mathrm{d}\tau} = -\frac{\partial\mathcal{L}}{\partial b}. \tag{48}$$

The vector parameters m, q inherit a factor $1/\gamma^2 = 1/(\alpha^2 d)$ from the change of variables that the scalars c, b do not. This asymmetry, and the fact that γ^2 is $O_d(d)$ in one regime and $O_d(1)$ in the other, is what drives everything below. Equations equation 46–equation 48 are the exact reduction of the problem. The two regimes are now obtained by scaling analysis, i.e. by evaluating h, k from equation 45 in the large d limit.

Regime II: $\Lambda = O_d(d)$, i.e. $\alpha = O_d(1)$. Here $\gamma^2 = \alpha^2 d = O_d(d)$, so as soon as $m^2 + q = O_d(1)$ the Gaussian term in equation 44 has standard deviation $\gamma\sqrt{\Gamma(m^2+q)} = O_d(\sqrt{d})$: the $\tanh$ argument diverges for almost every g and $T \to \mathrm{sign}$. Consequently $k \to 1$, and the $O_d(1)$ bias b is swamped, so $h \to \mathbb{E}[\mathrm{sign}(\nu + g)] = \mathrm{erf}(\nu/\sqrt{2})$ with

$$\nu = \sqrt{\Lambda}\,\frac{m}{\sqrt{m^2+q}}. \tag{49}$$

Crucially $1 - k \to 0$ exponentially in γ (a Gaussian tail), so the term equation 42 is genuinely negligible here, and h is independent of b. Since the imbalance ω enters the exact score equation 6 only through the offset $\frac{1}{2}\log\frac{\omega}{1-\omega}$, and $\partial\mathcal{L}/\partial b \to 0$, the bias never moves in this regime. The loss reduces to

$$\mathcal{L}(c,m,q) = \left(1+\beta^2 c\right)^2 + \alpha^2\beta^2\left[c^2\left(1+\sigma^2\right) + m^2 + q + 2c\,m\,\mathrm{erf}\left(\frac{\nu}{\sqrt{2}}\right)\right], \tag{50}$$

and equation 48 closes on (c, m, q), with all the α^2 prefactors cancelling against $1/\gamma^2$ in the m, q equations:

$$\frac{\mathrm{d}c}{\mathrm{d}\tau} = -2\beta^2\left(1+\beta^2 c\right) - 2\alpha^2\beta^2\left[c\left(1+\sigma^2\right) + m\,\mathrm{erf}\left(\frac{\nu}{\sqrt{2}}\right)\right], \tag{51}$$

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = -\frac{2\beta^2}{d}\left[m + c\,\mathrm{erf}\left(\frac{\nu}{\sqrt{2}}\right) + \sqrt{\frac{2}{\pi}}\,\frac{c\,q\,m\sqrt{\Lambda}}{(m^2+q)^{3/2}}\,e^{-\nu^2/2}\right], \tag{52}$$

$$\frac{\mathrm{d}q}{\mathrm{d}\tau} = -\frac{4\beta^2}{d}\,q\left[1 - \sqrt{\frac{2}{\pi}}\,\frac{c\,m^2\sqrt{\Lambda}}{(m^2+q)^{3/2}}\,e^{-\nu^2/2}\right]. \tag{53}$$

The right-hand sides of the m and q equations are $O_d(1/d)$ while that of c is $O_d(1)$: the dynamics splits into two phases.

Phase 1, $\tau = O_d(1)$. Only c moves. By equation 34, $m_0, q_0 = O_d(1/d)$, so $m\,\mathrm{erf}(\nu/\sqrt{2})$ is negligible and

$$\frac{\mathrm{d}c}{\mathrm{d}\tau} = -2\beta^2\left(\Gamma+\alpha^2\right)(c - c_1^*), \qquad c_1^* = -\frac{1}{\Gamma+\alpha^2} = -\frac{d}{\mathbb{E}\|\mathbf{x}_t\|^2}, \tag{54}$$

so c relaxes exponentially to the value that best fits the isotropic Gaussian approximation of P_t, while (m, q, b) sit at initialization.

Phase 2, $\tau = O_d(d)$. Now m, q evolve while c, being d times faster, is slaved through $\partial_c\mathcal{L} = 0$, i.e. $c^* = -(1 + \alpha^2 m\,\mathrm{erf}(\nu/\sqrt{2}))/(\Gamma+\alpha^2)$. Note from equation 34 and equation 49 that $\nu = O_d(1)$ already at initialization in this regime: the drive $-c\,\mathrm{erf}(\nu/\sqrt{2}) > 0$ is present from the outset and m grows immediately, with no instability threshold to cross. Once $\nu \gg 1$ the error function saturates and, taking $m > 0$ without loss of generality, the flow linearizes,

$$\frac{\mathrm{d}q}{\mathrm{d}\tau} = -\frac{4\beta^2}{d}\,q, \qquad \frac{\mathrm{d}m}{\mathrm{d}\tau} = -\frac{2\beta^2\Gamma}{d\,(\Gamma+\alpha^2)}\left(m - \frac{1}{\Gamma}\right), \tag{55}$$

so $q \to 0$ and $m \to 1/\Gamma$ on $\tau = O_d(d)$. At the fixed point $\mathbf{w} = \alpha\boldsymbol{\mu}/\Gamma$ and $c = -1/\Gamma$, giving

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) \longrightarrow -\frac{\mathbf{x}}{\Gamma} + \frac{\alpha\boldsymbol{\mu}}{\Gamma}\tanh\left(\frac{\alpha\,\boldsymbol{\mu}^T\mathbf{x}}{\Gamma} + b_0\right), \tag{56}$$

which is the exact score equation 6 of the target except for the frozen bias b_0 in place of $b^* = \frac{1}{2}\log\frac{\omega}{1-\omega}$.

Speciation time: $\Lambda = O_d(1)$, i.e. $\alpha^2 = O_d(1/d)$. Now $\gamma^2 = \Lambda\Gamma = O_d(1)$ and the $\tanh$ argument equation 44 stays $O_d(1)$: no saturation, and the bias is no longer negligible. Two simplifications are available. First, $\alpha^2 = O_d(1/d) \to 0$ forces $t \to t_{\max}$, hence $\beta \to 1$ and

$$\Gamma = \alpha^2\sigma^2 + \beta^2 \longrightarrow 1, \tag{57}$$

so we set $\Gamma = 1$ and $\gamma^2 = \Lambda$ throughout. Second, by equation 34 the initial condition is now

$$q_0 = \frac{1}{\gamma^2} = O_d(1), \qquad m_0 = O_d\!\left(\frac{1}{\gamma\sqrt{d}}\right) = O_d\!\left(\frac{1}{\sqrt{d}}\right), \tag{58}$$

so m starts vanishingly small against an $O_d(1)$ value of q - the opposite of Regime II, where both were $O_d(1/d)$. Since $\alpha^2\beta^2 = O_d(1/d)$, the loss equation 46 is dominated by $(1+\beta^2 c)^2$ and c relaxes on $\tau = O_d(1)$ to $c^* = -1/\Gamma + O_d(1/d) = -1 + O_d(1/d)$, before (m, q, b) move. Substituting $c = c^*$ makes the $(1 + c\Gamma)$ term in equation 46 $O_d(1/d)$, leaving $\mathcal{L} = \text{const} + \alpha^2\beta^2\,\mathcal{F} + O_d(1/d^2)$ with

$$\mathcal{F}(m, q, b) = \mathbb{E}_{g,s}\Big[\,\big(m^2 + q\big)\,T^2 - 2\,s\,m\,T\Big], \qquad T = \tanh(A), \qquad A = s\,\gamma^2 m + \gamma\sqrt{m^2+q}\;g + b. \tag{59}$$

Measuring training time in units of $\check{\tau} = \beta^2\tau/d$, the exact flow equation 48 becomes

$$\frac{\mathrm{d}m}{\mathrm{d}\check{\tau}} = -\frac{\partial\mathcal{F}}{\partial m}, \qquad \frac{\mathrm{d}q}{\mathrm{d}\check{\tau}} = -4q\,\frac{\partial\mathcal{F}}{\partial q}, \qquad \frac{\mathrm{d}b}{\mathrm{d}\check{\tau}} = -\gamma^2\,\frac{\partial\mathcal{F}}{\partial b}, \tag{60}$$

where the α^2 of the loss has cancelled against the $1/\gamma^2 = 1/(\alpha^2 d)$ of equation 48 in the m, q equations, while the b equation retains a factor $\gamma^2 = O_d(1)$. All four parameters therefore move at comparable rates in $\check{\tau}$, i.e. $\tau = O_d(d)$.

Linear stability of $m = b = 0$. Expanding equation 59 to second order in (m, b) at fixed q, and introducing the Gaussian averages

$$\Theta_1(q,\gamma) = \mathbb{E}_g\big[\tanh^2(\gamma\sqrt{q}\,g)\big], \qquad \Theta_2(q,\gamma) = \mathbb{E}_g\big[\big(1-\tanh^2\big)\,\big(1-3\tanh^2\big)(\gamma\sqrt{q}\,g)\big], \tag{61}$$

one finds $\partial_m\mathcal{F} = \partial_b\mathcal{F} = 0$ at $m = b = 0$, both integrands being odd. Writing

$$\begin{aligned}
&\Psi(q,\gamma) = \Theta_1 + \gamma^2 q\,\Theta_2, \\
&\mu(q,\gamma) = 4\gamma^2\,(1-\Theta_1) - 2\Theta_1 - 2\gamma^2 q\,\big(1+\gamma^2\big)\,\Theta_2, \\
&\Xi(q,\gamma) = (1-\Theta_1) - \gamma^2 q\,\Theta_2,
\end{aligned} \tag{62}$$

the zeroth order of the expansion gives the autonomous relaxation of q,

$$\frac{\mathrm{d}q}{\mathrm{d}\check{\tau}} = -4q\,\Psi(q,\gamma) + O(m^2, b^2), \tag{63}$$

which decays monotonically from $q_0 = 1/\gamma^2$, while the first order gives the coupled linear system obeyed by the overlap and the bias at fixed q,

$$\frac{\mathrm{d}}{\mathrm{d}\check{\tau}}\begin{pmatrix} m \\ b \end{pmatrix} = \mathbf{M}(q,\gamma)\begin{pmatrix} m \\ b \end{pmatrix}, \qquad \mathbf{M}(q,\gamma) = \begin{pmatrix} \mu & 2\,(2\omega-1)\,\Xi \\ 2\gamma^2\,(2\omega-1)\,\Xi & -2\gamma^2 q\,\Theta_2 \end{pmatrix}. \tag{64}$$

The off-diagonal terms are proportional to $2\omega - 1$: for balanced data the overlap and the bias decouple entirely, m growing at the rate $\mu(q,\gamma)$ while b decays to 0. For $\omega \neq 1/2$ the two are coupled and leave the origin together, the flow converging to $q = 0$, $m = 1/\Gamma$ and

$$b \longrightarrow b^* = \frac{1}{2}\log\frac{\omega}{1-\omega}, \qquad \text{i.e.} \qquad \rho \longrightarrow \omega. \tag{65}$$

Contrary to Regime II, where the bias stays frozen at b_0 and only the direction is recovered, training around the speciation time learns the mode direction and the imbalance, jointly and on the same $\tau = O_d(d)$ timescale.

A.2.2 GMM WITH FOUR MODES: EXACT LOSS AND GRADIENT FLOW

This section is self-contained, and mirrors §A.2.1 for the hierarchical target, using the score model of the main text, equation 10.

Setup. The target is the block-structured quadrimodal mixture

$$P_0(\mathbf{x}) = \frac{1}{4} \sum_{s_1=\pm 1} \sum_{s_2=\pm 1} \mathcal{N}\big(s_1 \boldsymbol{\mu}_1 + s_2 \boldsymbol{\mu}_2,\, \sigma^2 \boldsymbol{I}_d\big)\,, \tag{66}$$

where $\boldsymbol{\mu}_1$ is supported on the first κd coordinates and $\boldsymbol{\mu}_2$ on the remaining $(1-\kappa)d$. We write $\kappa_1 = \kappa$, $\kappa_2 = 1-\kappa$, both $O_d(1)$, and normalize $\|\boldsymbol{\mu}_i\|^2 = \kappa_i d$ so that every coordinate of the means is $O_d(1)$. The score model is

$$\mathbf{s}_{\boldsymbol{\theta}=(c,\mathbf{w}_1,\mathbf{w}_2)}(\mathbf{x}, t) = c\,\mathbf{x} + \mathbf{w}_1 \tanh\big(\mathbf{w}_1^T \mathbf{x}\big) + \mathbf{w}_2 \tanh\big(\mathbf{w}_2^T \mathbf{x}\big)\,, \tag{67}$$

with $\mathbf{w}_i$ supported on block i, so that $\mathbf{w}_1^T \mathbf{w}_2 = 0$ and $\mathbf{w}_i^T \boldsymbol{\mu}_j = 0$ for $i \neq j$. As in §A.2.1 we work at fixed noising time t, write $\alpha = \alpha(t), \beta = \beta(t)$, $\Gamma = \alpha^2\sigma^2 + \beta^2$, $\Lambda = \alpha^2 d/\Gamma$, and train by gradient flow on the single-time DSM loss $\mathcal{L} = \frac{1}{d}\mathbb{E}\|\beta \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\xi}\|^2$ with $\mathbf{x}_t = \alpha \mathbf{x}_0 + \beta \boldsymbol{\xi}$. Note that equation 67 carries no bias: each block of equation 66 is balanced, so there is no mode weight to learn, and the structural parameter is κ instead.

Summary statistics, block SNR, initialization. Per block we define, as in equation 11,

$$m_i = \frac{\mathbf{w}_i^T \boldsymbol{\mu}_i}{\alpha\, \kappa_i d}, \qquad q_i = \frac{\|\mathbf{w}_i^{\perp}\|^2}{\alpha^2 \kappa_i d}, \qquad \text{so that} \qquad \|\mathbf{w}_i\|^2 = \alpha^2 \kappa_i d \left(m_i^2 + q_i\right), \tag{68}$$

with $\mathbf{w}_i^{\perp}$ the component of $\mathbf{w}_i$ orthogonal to $\boldsymbol{\mu}_i$ inside block i. Because block i has $\kappa_i d$ coordinates and $\|\boldsymbol{\mu}_i\|^2 = \kappa_i d$, the natural signal-to-noise ratio and $\tanh$ scale of block i are those of §A.2.1 with $d \to \kappa_i d$:

$$\Lambda^{(i)} = \frac{\alpha^2 \|\boldsymbol{\mu}_i\|^2}{\Gamma} = \kappa_i \Lambda, \qquad \gamma_i^2 := \alpha^2 \kappa_i d = \kappa_i \Lambda \Gamma = \kappa_i \gamma^2. \tag{69}$$

The block asymmetry κ enters only through this rescaling of the SNR. Initializing each $\mathbf{w}_i(0)$ uniformly on the unit sphere of its own block gives, exactly as in equation 34,

$$m_i(0) = O_d\!\left(\frac{1}{\alpha \kappa_i d}\right), \qquad q_i(0) = \frac{1}{\gamma_i^2} + O_d\big(\tfrac{1}{d}\big)\,, \qquad \text{so that} \qquad \gamma_i \sqrt{q_i(0)} = 1 \tag{70}$$

for both blocks, whatever κ: the two blocks start at the same point in the rescaled variable that controls the $\tanh$, and differ only through γ_i. Together with (§A.2.3), everything below establishes the following statement, the long form of Result 2 of the main text.

Result (2, long form: training dynamics, hierarchical distribution)**.** *Let $\mathbf{w}_1(0), \mathbf{w}_2(0)$ be independent and uniform on the unit sphere of their respective κd (resp. $(1-\kappa)d$) dimensional blocks, $c(0) = O_d(1)$, and let $(c, \mathbf{w}_1, \mathbf{w}_2)$ evolve under GF equation 8 for the score model equation 10 at fixed t. Define $\kappa_1 = \kappa, \kappa_2 = 1-\kappa$, each pair (m_i, q_i) obeys the dynamics of Result 1 at $\omega = 1/2$, $b = 0$ and rescaled SNR $\Lambda \to \kappa_i \Lambda$, the two blocks being coupled only through the shared skip connection c. As $d \to \infty$:*

- *(a)* ***Regime II.*** *On $\tau = O_d(1)$ times only c moves, relaxing to $c_1^* = -1/(\Gamma + \alpha^2)$ while both (m_i, q_i) stay at initialization. On $\tau = O_d(d)$ times c is slaved to the overlaps, and (m_1, q_1), (m_2, q_2) both converge to $(1/\Gamma, 0)$, independently of κ_i: the two mode directions $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are recovered together, on the same timescale, regardless of the block asymmetry.*
- *(b)* ***Speciation time*** *($\Lambda = O_d(1)$). Here $\Gamma \to 1$, each block carries its own SNR $\gamma_i^2 = \kappa_i \Lambda \Gamma$ and starts from $q_i(0) = 1/\gamma_i^2$. On $\tau = O_d(1)$ times $c \to -1/\Gamma$ and stays there for the remainder of the dynamics; with c thus frozen the two blocks decouple entirely, and since $\omega = 1/2$ makes $\mathbf{M}(q_i, \gamma_i)$ diagonal, each pair obeys the closed autonomous system*

$$\frac{\mathrm{d}q_i}{\mathrm{d}\check{\tau}} = -4 q_i\, \Psi(q_i, \gamma_i)\,, \qquad \frac{\mathrm{d}m_i}{\mathrm{d}\check{\tau}} = \mu(q_i, \gamma_i)\, m_i + O(m_i^3), \tag{71}$$

with Ψ and the growth rate μ as given in equation 62. The orthogonal component q_i decreases monotonically on the same timescale as m_i, and μ is a strictly decreasing

function of q_i: it is the decay of q_i that drives the overlap unstable. In the $q_i \to 0$ limit, $\mu(q_i, \gamma_i) \sim 4\gamma_i^2 \equiv 4\kappa_i\Lambda$: the rate of block i is proportional to κ_i. The two directions are learned on different timescales - for κ small the dominant direction $\boldsymbol{\mu}_2$ is learned on $\tau = O_d(d)$, while the subdominant $\boldsymbol{\mu}_1$ lags by a factor $1/\kappa$, on $\tau = O_d(d/\kappa)$.

(c) ***Heuristic growth rate.*** *The system equation 71 is two-dimensional, so at finite γ_i the rate $\mu(q_i, \gamma_i)$ drifts as q_i relaxes. Assuming instead that $\|\mathbf{w}_i\|$ is initialized and kept at the norm $\alpha\|\boldsymbol{\mu}_i\|$ of the target direction, i.e. $m_i^2 + q_i = 1$, eliminates q_i from equation 71 and reduces the dynamics to a single ODE $\mathrm{d}m_i/\mathrm{d}\check{\tau} = \lambda(\gamma_i)\, m_i + O(m_i^3)$, with the closed-form rate equation 104 of §A.2.3,*

$$\lambda(\gamma) = \mu(1, \gamma) + 2\,\Psi(1, \gamma). \tag{72}$$

Note that this constraint is not exactly met by the initialization above, which gives $q_i(0) = 1/\gamma_i^2$ rather than $q_i(0) = 1$; for $\kappa_i = O_d(1)$, however, the two differ only by the $O_d(1)$ factor $\gamma_i^2 = \kappa_i\Lambda\Gamma$, so both start the dynamics at the same order in d. The reduction moreover recovers (b) exactly as $\kappa \to 0$, since $\lambda(\gamma) = 4\gamma^2 - 6\gamma^4 + O(\gamma^6)$. Away from that limit it is a heuristic, but a closed-form one, and we use $\tau_i \propto 1/\lambda(\gamma_i)$ as the reference timescale against which the measured dynamics are rescaled in Fig. 2.

Detailed dynamics.

1. Exact loss. Decompose $\mathbf{x}_0 = s_1\boldsymbol{\mu}_1 + s_2\boldsymbol{\mu}_2 + \sigma\mathbf{z}$ with $s_1, s_2 = \pm 1$ independent and uniform, $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, and set $q_{\mathbf{z}_i} = \mathbf{w}_i^T\mathbf{z}/(\alpha\sqrt{\kappa_i d(m_i^2 + q_i)})$ and $q_{\boldsymbol{\xi}_i} = \mathbf{w}_i^T\boldsymbol{\xi}/(\alpha\sqrt{\kappa_i d(m_i^2 + q_i)})$; the four are i.i.d. standard Gaussians because $\mathbf{w}_1^T\mathbf{w}_2 = 0$. Since $\mathbf{w}_i^T\boldsymbol{\mu}_j = 0$ for $i \neq j$, block i only ever sees s_i, and

$$\mathbf{w}_i^T\mathbf{x}_0 = \alpha\left(s_i\, m_i\, \kappa_i d + \sigma\sqrt{\kappa_i d(m_i^2 + q_i)}\, q_{\mathbf{z}_i}\right), \qquad \mathbf{w}_i^T\boldsymbol{\xi} = \alpha\sqrt{\kappa_i d(m_i^2 + q_i)}\, q_{\boldsymbol{\xi}_i}. \tag{73}$$

Inserting $\mathbf{x}_t = \alpha\mathbf{x}_0 + \beta\boldsymbol{\xi}$ into equation 67 and regrouping,

$$\beta\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\xi} = \alpha\beta c\,\mathbf{x}_0 + \left(1 + \beta^2 c\right)\boldsymbol{\xi} + \beta\sum_{i=1,2} T_i\,\mathbf{w}_i, \qquad T_i := \tanh\left(\mathbf{w}_i^T\mathbf{x}_t\right). \tag{74}$$

Squaring and dividing by d: the cross term between the two blocks vanishes by $\mathbf{w}_1^T\mathbf{w}_2 = 0$, $\frac{1}{d}\mathbb{E}\|\mathbf{x}_0\|^2 \to 1 + \sigma^2$ since $\|\boldsymbol{\mu}_1\|^2 + \|\boldsymbol{\mu}_2\|^2 = d$, and the terms in $q_{\mathbf{z}_i}, q_{\boldsymbol{\xi}_i}$ carry an explicit $1/\sqrt{d}$ which is again compensated by Stein's lemma,

$$\mathbb{E}[q_{\boldsymbol{\xi}_i} T_i] = \alpha\beta\sqrt{\kappa_i d(m_i^2 + q_i)}\,\mathbb{E}\left[1 - T_i^2\right], \quad \mathbb{E}[q_{\mathbf{z}_i} T_i] = \alpha^2\sigma\sqrt{\kappa_i d(m_i^2 + q_i)}\,\mathbb{E}\left[1 - T_i^2\right], \tag{75}$$

the two recombining through $1 + \beta^2 c + \alpha^2\sigma^2 c = 1 + c\,\Gamma$. Finally, from equation 73 and $\alpha\sqrt{\kappa_i d} = \gamma_i$, the argument of the i-th $\tanh$ is

$$\mathbf{w}_i^T\mathbf{x}_t = s_i\,\gamma_i^2 m_i + \gamma_i\sqrt{\Gamma\,(m_i^2 + q_i)}\; g_i, \qquad g_i \sim \mathcal{N}(0, 1), \tag{76}$$

and, s_i being uniform, averaging over it with $g_i \to s_i g_i$ removes s_i altogether. Collecting,

$$\mathcal{L}(c, m_1, q_1, m_2, q_2) = \left(1 + \beta^2 c\right)^2 + \alpha^2\beta^2\left[c^2\left(1 + \sigma^2\right) + \sum_{i=1,2}\kappa_i\,\mathcal{G}_i\right], \tag{77}$$

$$\mathcal{G}_i = \left(m_i^2 + q_i\right)\Big(k_i + 2\,(1 + c\,\Gamma)\,(1 - k_i)\Big) + 2c\,m_i\,h_i, \tag{78}$$

with the one-dimensional Gaussian averages

$$h_i = \mathbb{E}_g[\tanh A_i], \quad k_i = \mathbb{E}_g\left[\tanh^2 A_i\right], \quad A_i = \gamma_i^2 m_i + \gamma_i\sqrt{\Gamma\,(m_i^2 + q_i)}\; g. \tag{79}$$

Comparing with equation 46: each $\mathcal{G}_i$ is exactly the bracket of the unbalanced problem at $\omega = 1/2$ and $b = 0$, with $\gamma \to \gamma_i$. The hierarchical target is therefore two independent copies of the single-mode problem at rescaled SNR $\kappa_i\Lambda$, weighted by κ_i and coupled only through the shared skip connection c.

2. Exact gradient flow. With $\nabla_{\mathbf{w}_i} m_i = \boldsymbol{\mu}_i/(\alpha\kappa_i d)$ and $\nabla_{\mathbf{w}_i} q_i = 2\mathbf{w}_i^{\perp}/(\alpha^2\kappa_i d)$, projecting $\mathrm{d}\mathbf{w}_i/\mathrm{d}\tau = -\nabla_{\mathbf{w}_i}\mathcal{L}$ as in equation 47 gives

$$\frac{\mathrm{d}c}{\mathrm{d}\tau} = -\frac{\partial\mathcal{L}}{\partial c}, \qquad \frac{\mathrm{d}m_i}{\mathrm{d}\tau} = -\frac{1}{\gamma_i^2}\frac{\partial\mathcal{L}}{\partial m_i} = -\frac{\alpha^2\beta^2\kappa_i}{\gamma_i^2}\frac{\partial\mathcal{G}_i}{\partial m_i}, \qquad \frac{\mathrm{d}q_i}{\mathrm{d}\tau} = -\frac{4q_i}{\gamma_i^2}\frac{\partial\mathcal{L}}{\partial q_i} = -\frac{4q_i\,\alpha^2\beta^2\kappa_i}{\gamma_i^2}\frac{\partial\mathcal{G}_i}{\partial q_i}. \tag{80}$$

Using $\gamma_i^2 = \alpha^2\kappa_i d$ and the rescaled training time $\check{\tau} = \beta^2\tau/d$ yields

$$\frac{\mathrm{d}m_i}{\mathrm{d}\check{\tau}} = -\frac{\partial\mathcal{G}_i}{\partial m_i}, \qquad \frac{\mathrm{d}q_i}{\mathrm{d}\check{\tau}} = -4q_i\,\frac{\partial\mathcal{G}_i}{\partial q_i}. \tag{81}$$

The block asymmetry acts on the dynamics exclusively through $\gamma_i^2 = \kappa_i\Lambda\Gamma$ inside $\mathcal{G}_i$, i.e. purely as a rescaling of the block SNR.

Regime II: $\Lambda = O_d(d)$. Then $\gamma_i^2 = \kappa_i\Lambda\Gamma = O_d(d)$ for both blocks, the argument equation 76 diverges, $T_i \to \mathrm{sign}$, $k_i \to 1$ with $1 - k_i$ exponentially small, and $h_i \to \mathrm{erf}(\nu_i/\sqrt{2})$ with $\nu_i = \sqrt{\kappa_i\Lambda}\,m_i/\sqrt{m_i^2 + q_i}$. The loss becomes

$$\mathcal{L} = \left(1 + \beta^2 c\right)^2 + \alpha^2\beta^2\left[c^2\left(1+\sigma^2\right) + \sum_{i=1,2}\kappa_i\left(m_i^2 + q_i + 2c\,m_i\,\mathrm{erf}\left(\frac{\nu_i}{\sqrt{2}}\right)\right)\right], \tag{82}$$

and equation 80 gives

$$\frac{\mathrm{d}c}{\mathrm{d}\tau} = -2\beta^2\left(1+\beta^2 c\right) - 2\alpha^2\beta^2\left[c\left(1+\sigma^2\right) + \sum_i \kappa_i\,m_i\,\mathrm{erf}\left(\frac{\nu_i}{\sqrt{2}}\right)\right], \tag{83}$$

$$\frac{\mathrm{d}m_i}{\mathrm{d}\tau} = -\frac{2\beta^2}{d}\left[m_i + c\,\mathrm{erf}\left(\frac{\nu_i}{\sqrt{2}}\right) + \sqrt{\frac{2}{\pi}}\,\frac{c\,q_i\,m_i\sqrt{\kappa_i\Lambda}}{\left(m_i^2+q_i\right)^{3/2}}e^{-\nu_i^2/2}\right], \tag{84}$$

$$\frac{\mathrm{d}q_i}{\mathrm{d}\tau} = -\frac{4\beta^2}{d}\,q_i\left[1 - \sqrt{\frac{2}{\pi}}\,\frac{c\,m_i^2\sqrt{\kappa_i\Lambda}}{\left(m_i^2+q_i\right)^{3/2}}e^{-\nu_i^2/2}\right]. \tag{85}$$

Again the c equation is $O_d(1)$ and the others $O_d(1/d)$, giving two phases. On $\tau = O_d(1)$, with $m_i, q_i = O_d(1/d)$ at initialization, c relaxes to

$$c_1^* = -\frac{1}{\Gamma + \alpha^2} \tag{86}$$

while both (m_i, q_i) stay put. On $\tau = O_d(d)$, c is slaved through $\partial_c\mathcal{L} = 0$, i.e. $c^* = -\left(1 + \alpha^2\sum_j \kappa_j m_j\,\mathrm{erf}(\nu_j/\sqrt{2})\right)/(\Gamma + \alpha^2)$ and linearizing

$$\frac{\mathrm{d}q_i}{\mathrm{d}\tau} = -\frac{4\beta^2}{d}\,q_i, \qquad \frac{\mathrm{d}m_i}{\mathrm{d}\tau} = -\frac{2\beta^2}{d}\left[m_i - \frac{1 + \alpha^2\sum_j \kappa_j m_j}{\Gamma + \alpha^2}\right]. \tag{87}$$

Both blocks obey the same dynamics and converge together. Using $\sum_j \kappa_j = 1$, the fixed point is $m_1 = m_2 = 1/\Gamma$, $q_i = 0$, $c = -1/\Gamma$, whence $\mathbf{w}_i = \alpha\boldsymbol{\mu}_i/\Gamma$ and

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) \longrightarrow -\frac{\mathbf{x}}{\Gamma} + \sum_{i=1,2}\frac{\alpha\boldsymbol{\mu}_i}{\Gamma}\tanh\left(\frac{\alpha\,\boldsymbol{\mu}_i^T\mathbf{x}}{\Gamma}\right), \tag{88}$$

exactly the target score equation 7. In Regime II both mode directions are recovered on the same $\tau = O_d(d)$ timescale, independently of κ: training at high SNR is blind to the block asymmetry.

Speciation time: $\Lambda = O_d(1)$. Now $\gamma_i^2 = \kappa_i\Lambda\Gamma = O_d(1)$, and as in §A.2.1 $\alpha^2 = O_d(1/d)$ forces $\Gamma \to 1$, so $\gamma_i^2 \to \kappa_i\Lambda$. The skip connection relaxes on $\tau = O_d(1)$ to $c^* = -1 + O_d(1/d)$; substituting it kills the $(1 + c\Gamma)$ term in equation 78 and leaves $\mathcal{G}_i = \mathcal{F}_i + O_d(1/d)$ with

$$\mathcal{F}_i(m_i, q_i) = \mathbb{E}_g\left[\left(m_i^2 + q_i\right)\tanh^2 A_i - 2\,m_i\tanh A_i\right], \qquad A_i = \gamma_i^2 m_i + \gamma_i\sqrt{m_i^2 + q_i}\;g. \tag{89}$$

With c frozen, the two blocks are now completely decoupled: by equation 81 each pair (m_i, q_i) follows its own autonomous flow, on the timescale

$$\check{\tau} = O_d(1) \qquad \Longleftrightarrow \qquad \tau = O_d(d). \tag{90}$$

Expanding equation 89 in m_i at fixed q_i (the $b = 0$, $\omega = 1/2$), and writing $\Theta_{1,2}$ for the Gaussian averages equation 61 evaluated at (q_i, γ_i),

$$\frac{\mathrm{d}q_i}{\mathrm{d}\check{\tau}} = -4q_i\,\Psi(q_i,\gamma_i) + O(m_i^2), \qquad \frac{\mathrm{d}m_i}{\mathrm{d}\check{\tau}} = \mu(q_i,\gamma_i)\,m_i + O(m_i^3), \tag{91}$$

with μ as in equation 62. Since, $\Theta_1 \to 0$ and $\Theta_2 \to 1$ as $q_i \to 0$, μ increases monotonically along the flow towards

$$\mu(q_i \to 0, \gamma_i) = 4\gamma_i^2 = 4\,\kappa_i\Lambda\Gamma \;\xrightarrow[\Gamma\to 1]{}\; 4\,\kappa_i\Lambda : \tag{92}$$

the asymptotic growth rate of block i is linear in κ_i. We expect the escape time of m_i to therefore scale as $1/\kappa_i$ in $\check{\tau}$, i.e.

$$\tau_i = O_d\!\left(\frac{d}{\kappa_i}\right). \tag{93}$$

Taking κ small, so that $\kappa_1 = \kappa \ll 1$ and $\kappa_2 = 1 - \kappa \simeq 1$, the dominant direction $\boldsymbol{\mu}_2$ is learned on $\tau = O_d(d)$ while the subdominant direction $\boldsymbol{\mu}_1$ lags by a factor $1/\kappa$, on $\tau = O_d(d/\kappa)$. Once escaped, each block converges to $q_i = 0$, $m_i = 1/\Gamma$, recovering equation 88.

A.2.3 Fixed-norm ansatz: a one-dimensional reduction at the speciation time

The analyses of §A.2.1 and §A.2.2 involve two coupled order parameters, the overlap m and the orthogonal norm q, whose joint relaxation makes the growth rate of m depend on the instantaneous value of q. We show here that constraining the weight vector to a fixed norm removes q entirely, reduces the dynamics to a single scalar ODE, and yields a closed-form growth rate at small overlap m.

Fixed norm constraint. We keep the notation of §A.2.1: score model equation 30, order parameters equation 33, and $\gamma^2 = \alpha^2 d = \Lambda\Gamma$. We now impose that $\mathbf{w}$ retain throughout training the norm of the target direction,

$$\|\mathbf{w}\| = \alpha\sqrt{d} \qquad \Longleftrightarrow \qquad m^2 + q = 1, \tag{94}$$

using $\|\mathbf{w}\|^2 = \alpha^2 d(m^2 + q)$. This is the natural scale: the exact score equation 6 is reproduced at $\mathbf{w}^* = \alpha\boldsymbol{\mu}/\Gamma$, of norm $\alpha\sqrt{d}/\Gamma$. We place ourselves at the speciation time, $\Lambda = O_d(1)$, and set $\Gamma = 1$ throughout, so that $\gamma^2 = \Lambda$.

The loss loses its dependence on q. The whole (m, q) dependence of the loss enters through the law equation 44 of the $\tanh$ argument, whose mean is $\propto m$ and whose standard deviation is $\gamma\sqrt{\Gamma(m^2 + q)}$. Under equation 94 the latter is constant,

$$\mathbf{w}^T\mathbf{x}_t + b = s\,\gamma^2 m + \gamma\,g + b, \qquad g \sim \mathcal{N}(0,1), \tag{95}$$

so the constraint freezes the width of the Gaussian and lets only its mean move with m. Writing $\mathcal{F}$ for the reduced loss equation 59 and substituting $q = 1 - m^2$, the prefactor $(m^2 + q)$ becomes 1 and

$$\Phi(m, b) := \mathcal{F}\big(m, 1 - m^2, b\big) = k(m, b) - 2\,m\,h(m, b), \tag{96}$$

with h, k the Gaussian averages equation 45 evaluated on equation 95. For the balanced case ($\omega = 1/2$, $b = 0$) relevant to each block of $(\mathcal{D}_2)$ these reduce to

$$h(m) = \mathbb{E}_g\big[\tanh\big(\gamma^2 m + \gamma g\big)\big], \qquad k(m) = \mathbb{E}_g\big[\tanh^2\big(\gamma^2 m + \gamma g\big)\big]. \tag{97}$$

The problem is now one-dimensional: a single scalar m, and a single parameter γ.

Constrained gradient flow. Restricting the flow to the sphere $\|\mathbf{w}\| = \alpha\sqrt{d}$ means projecting out the radial component, $\mathrm{d}\mathbf{w}/\mathrm{d}\tau = -\left(\boldsymbol{I} - \hat{\mathbf{w}}\hat{\mathbf{w}}^T\right)\nabla_{\mathbf{w}}\mathcal{L}$ with $\hat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$. Using $\nabla_{\mathbf{w}} m = \boldsymbol{\mu}/(\alpha d)$ and $\nabla_{\mathbf{w}} q = 2\mathbf{w}^{\perp}/(\alpha^2 d)$ from equation 47, together with $\mathbf{w}^T\nabla_{\mathbf{w}}\mathcal{L} = m\,\partial_m\mathcal{L} + 2q\,\partial_q\mathcal{L}$, one finds

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = -\frac{1}{\gamma^2}\Big[\left(1-m^2\right)\partial_m\mathcal{L} - 2mq\,\partial_q\mathcal{L}\Big] = -\frac{1-m^2}{\gamma^2}\Big[\partial_m\mathcal{L} - 2m\,\partial_q\mathcal{L}\Big], \tag{98}$$

the second equality using $q = 1 - m^2$. The bracket is precisely the total derivative of the loss along the constraint surface, $\Phi'(m) = \partial_m\mathcal{F} - 2m\,\partial_q\mathcal{F}$. Since $\mathcal{L} = \mathrm{const} + \alpha^2\beta^2\mathcal{F} + O_d(1/d^2)$ and $\alpha^2\beta^2/\gamma^2 = \beta^2/d$, the flow closes on m alone in the rescaled time $\check{\tau} = \beta^2\tau/d$ of §A.2.1,

$$\frac{\mathrm{d}m}{\mathrm{d}\check{\tau}} = -\left(1-m^2\right)\Phi'(m), \tag{99}$$

so that the dynamics again unfolds on $\tau = O_d(d)$.

Linearized dynamics. Everything is controlled by the same two Gaussian integrals equation 61 as in the unconstrained case, now evaluated at $q = 1$ since the width in equation 95 is fixed:

$$\theta_1(\gamma) := \Theta_1(1,\gamma) = \mathbb{E}_g\left[\tanh^2(\gamma g)\right], \tag{100}$$

$$\theta_2(\gamma) := \Theta_2(1,\gamma) = \mathbb{E}_g\left[\left(1-\tanh^2(\gamma g)\right)\left(1-3\tanh^2(\gamma g)\right)\right]. \tag{101}$$

Expanding equation 97 in m around 0 and using that $\tanh$ is odd (so $\mathbb{E}_g[\tanh(\gamma g)] = \mathbb{E}_g[\tanh\tanh'(\gamma g)] = 0$),

$$h(m) = \gamma^2\left(1-\theta_1\right)m + O(m^3), \qquad k(m) = \theta_1 + \gamma^4\theta_2\,m^2 + O(m^4), \tag{102}$$

so that Φ is even in m and

$$\Phi(m) = \theta_1 - \frac{\lambda(\gamma)}{2}\,m^2 + O(m^4), \qquad \lambda(\gamma) = 4\gamma^2\left(1-\theta_1(\gamma)\right) - 2\gamma^4\,\theta_2(\gamma). \tag{103}$$

Inserting into equation 99, the overlap grows exponentially from its $O_d(1/\sqrt{d})$ initial value,

$$\frac{\mathrm{d}m}{\mathrm{d}\check{\tau}} = \lambda(\gamma)\,m + O(m^3). \tag{104}$$

One can check that $\lambda(\gamma) > 0$ for all $\gamma > 0$ and, unlike the unconstrained rate $\mu(q,\gamma)$ of equation 62, which drifts as q relaxes, λ is a fixed number once γ is chosen.

Small γ and consistency with §A.2.2. For $\gamma \ll 1$, $\tanh(\gamma g) = \gamma g - \frac{1}{3}\gamma^3 g^3 + O(\gamma^5)$ gives $\theta_1 = \gamma^2 - 2\gamma^4 + O(\gamma^6)$ and $\theta_2 = 1 - 4\gamma^2 + O(\gamma^4)$, hence

$$\lambda(\gamma) = 4\gamma^2\left(1-\gamma^2\right) - 2\gamma^4 + O(\gamma^6) = 4\gamma^2 - 6\gamma^4 + O(\gamma^6). \tag{105}$$

Applied blockwise to $(\mathcal{D}_2)$, where $\gamma_i^2 = \kappa_i\Lambda\Gamma \to \kappa_i\Lambda$ by equation 69, this gives

$$\lambda(\gamma_i) \simeq 4\gamma_i^2 = 4\,\kappa_i\Lambda, \tag{106}$$

identical to the limiting rate equation 92 obtained there without the constraint.

Note that the two descriptions do not coincide away from this limit. For instance, a direct comparison at $q = 1$ gives the exact identity

$$\lambda(\gamma) - \mu(1,\gamma) = 2\left(\theta_1 + \gamma^2\theta_2\right) = 2\,\Psi(1,\gamma) > 0. \tag{107}$$

A.3 Integrated loss: SNR reparametrization and effective weightings

The ResNet experiments of Fig. 3 do not compare two distinct pipelines. Both panels train the same denoiser, on the same denoising score matching objective equation 5 and the same noising schedule; the two curves differ only through the weighting function $w(t)$, taken either uniform or chosen so as to reproduce the allocation of noise levels that Flow Matching performs implicitly. This appendix makes that construction precise. We first show that, once expressed in the signal-to-noise ratio, any stochastic interpolant pipeline is characterized by a single effective SNR weighting $w_{\mathrm{eff}}(\Lambda)$, so that two pipelines sharing it train identically whatever their time parametrization. We then compute w_{eff} in closed form for Diffusion and Flow Matching; the weighting $w(t)$ that installs the latter inside a DSM run is derived in §A.4, and is the one used in Fig. 3.

Setup. In practice a single network represents the denoiser at every noise level at once, and is trained on the loss integrated over the noising time with a weighting function $w(t)$,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{d}\int_0^{t_{\max}} \mathrm{d}t\; w(t)\; \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\xi}} \left\| \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\left(\alpha(t)\mathbf{x}_0 + \beta(t)\boldsymbol{\xi},\, t\right)\right\|^2 . \tag{108}$$

Here $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}} := -\beta(t)\,\mathbf{s}_{\boldsymbol{\theta}}$ is the noise prediction associated with the score model, so that equation 108 is identically the loss $\|\beta\mathbf{s}_{\boldsymbol{\theta}} + \boldsymbol{\xi}\|^2$ of equation 5 used in the single time analysis: the two are the same residual written in two parametrizations, for every t, with no approximation on β. The single SNR results of §A.2.1–§A.2.2 may therefore be inserted into equation 108 as they stand.

SNR reparametrization. Two pipelines with different (α, β, w) may present the network with the same collection of denoising problems, since what a training example teaches depends on its noise level only through the SNR

$$\Lambda_t = \frac{\alpha^2(t)\, d}{\alpha^2(t)\sigma^2 + \beta^2(t)}, \tag{109}$$

which, assuming β/α monotonically increases, decreases monotonically from $\Lambda = d/\sigma^2$ at $t = 0$ to $\Lambda = 0$ at $t = t_{\max}$ (§A.1). Being monotone, $t \mapsto \Lambda_t$ is invertible and may be used as the integration variable in equation 108,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{d}\int_0^{d/\sigma^2} \mathrm{d}\Lambda\; w_{\mathrm{eff}}(\Lambda)\; \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\xi}} \left\| \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}\right\|^2 , \qquad w_{\mathrm{eff}}(\Lambda) = w\big(t(\Lambda)\big)\left|\frac{\mathrm{d}t}{\mathrm{d}\Lambda}\right| . \tag{110}$$

We call w_{eff} the effective SNR weighting of the scheme. Two pipelines with the same w_{eff} train identically whatever their time parametrizations; conversely, the same w produces very different w_{eff} for different interpolants. All the comparisons below are therefore comparisons of w_{eff}.

Flow Matching as a reweighted denoising loss. Diffusion regresses the noise directly, so equation 108 applies verbatim. Flow Matching instead regresses the velocity field along the linear interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\xi}$, whose target is $\dot{\mathbf{x}}_t = \boldsymbol{\xi} - \mathbf{x}_0$,

$$\mathcal{L}_{\mathrm{FM}} = \int_0^1 \mathrm{d}t\, w(t)\, \mathbb{E}\left\|\hat{\boldsymbol{v}} - (\boldsymbol{\xi} - \mathbf{x}_0)\right\|^2 . \tag{111}$$

Eliminating $\mathbf{x}_0 = (\mathbf{x}_t - t\boldsymbol{\xi})/(1-t)$ gives $\boldsymbol{\xi} - \mathbf{x}_0 = (\boldsymbol{\xi} - \mathbf{x}_t)/(1-t)$, so that with the change of variables $\hat{\boldsymbol{\xi}} = (1-t)\hat{\boldsymbol{v}} + \mathbf{x}_t$,

$$\mathcal{L}_{\mathrm{FM}} = \int_0^1 \mathrm{d}t\; \frac{w(t)}{(1-t)^2}\; \mathbb{E}\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|^2 . \tag{112}$$

Flow Matching is thus a denoising loss in disguise, but one that silently reweights the noise levels by $1/(1-t)^2 = 1/\alpha^2(t)$. This Jacobian is not a cosmetic detail: since $\alpha \to 0$ at the noisy end, it is precisely what will concentrate FM on the low-SNR region below.

Effective weightings of DM and FM. We take $w \equiv 1$; a non-uniform w simply multiplies the results by $w(t(\Lambda))$.

For DM, $\alpha = e^{-t}$ and $\beta = \sqrt{1 - e^{-2t}}$, so $\Gamma_t = 1 - e^{-2t}(1-\sigma^2)$ and equation 109 inverts as $e^{-2t} = \Lambda/(d + (1-\sigma^2)\Lambda)$. Differentiating $t = -\frac{1}{2}\log e^{-2t}$,

$$w_{\mathrm{eff}}^{\mathrm{DM}}(\Lambda) = \frac{d}{2\Lambda\,(d + (1-\sigma^2)\Lambda)} \;\underset{\Lambda \ll d}{\simeq}\; \frac{1}{2\Lambda}. \tag{113}$$

For FM, $\alpha = 1 - t$ and $\beta = t$, so $\Gamma_t = (1-t)^2\sigma^2 + t^2$ and, writing $r = (1-t)/t$, equation 109 reads $\Lambda = r^2 d/(1 + r^2\sigma^2)$, i.e. $r = \sqrt{\Lambda/(d - \sigma^2\Lambda)}$. Including the Jacobian of equation 112,

$$w_{\mathrm{eff}}^{\mathrm{FM}}(\Lambda) = \frac{d}{2\Lambda^{3/2}\sqrt{d - \sigma^2\Lambda}} \;\underset{\Lambda \ll d}{\simeq}\; \frac{\sqrt{d}}{2\Lambda^{3/2}}. \tag{114}$$

The two differ by a factor $\sqrt{\Lambda}$: $w_{\text{eff}}^{\text{FM}}/w_{\text{eff}}^{\text{DM}} \simeq \sqrt{d/\Lambda}$ for $\Lambda \ll d$, so FM up-weights low SNR ever more strongly as Λ decreases. Since the SNR spans several decades, it is enlightening to consider the alternative log-SNR coordinate $u = \log \Lambda$, $w_{\text{eff}}\, \mathrm{d}\Lambda = \tilde{w}(u)\, \mathrm{d}u$ with

$$\tilde{w}(u) = \Lambda\, w_{\text{eff}}(\Lambda), \qquad \tilde{w}_{\text{DM}} \simeq \frac{1}{2}, \qquad \tilde{w}_{\text{FM}} \simeq \frac{1}{2}\sqrt{\frac{d}{\Lambda}}, \tag{115}$$

for $\Lambda \ll d$. Under this change of variable, it clearly appears that DM is scale-free: it deposits the same weight on every decade of SNR, while FM is not: its weight per decade grows as $\sqrt{d/\Lambda}$ towards low SNR. Evaluated at speciation scale yields

$$\left.\frac{\tilde{w}_{\text{FM}}}{\tilde{w}_{\text{DM}}}\right|_{\Lambda = O_d(1)} = O_d\Big(\sqrt{d}\Big)\,. \tag{116}$$

A.4 Matching Diffusion and Flow Matching at equal SNR

The comparison of §A.3 is between weightings, not between implementations. To isolate that effect experimentally we keep everything else fixed – same $\boldsymbol{\xi}$ prediction network, same VP forward process $\mathbf{x}_t = e^{-t}\mathbf{x}_0 + \sqrt{1-e^{-2t}}\,\boldsymbol{\xi}$, same sampler – and change only the weighting $w(t)$ in equation 108: the baseline uses $w \equiv 1$, i.e. plain DSM, and the "FM-weighted" run uses the $w(t)$ that reproduces the effective SNR weighting of Flow Matching. We derive that weight here.

The matched time. Write $u(t) = \sqrt{e^{2t}-1}$, so that the VP interpolant has $\alpha_{\text{DM}}/\beta_{\text{DM}} = e^{-t}/\sqrt{1-e^{-2t}} = 1/u$. The linear interpolant $\mathbf{x}_r = r\mathbf{x}_0 + (1-r)\boldsymbol{\xi}$ has $\alpha_{\text{FM}}/\beta_{\text{FM}} = r/(1-r)$. Since Λ is a monotone function of α/β alone, the two processes present the same denoising problem when

$$\frac{r}{1-r} = \frac{1}{u} \qquad \Longleftrightarrow \qquad r(t) = \frac{1}{1+u(t)}. \tag{117}$$

Differentiating equation 117 with $\mathrm{d}u/\mathrm{d}t = e^{2t}/u$ gives $|\mathrm{d}r/\mathrm{d}t| = e^{2t}/\big(u(1+u)^2\big)$. The two combine into

$$w_{\text{FM}\to\text{DM}}(t) = (1+u)^2 \times \frac{e^{2t}}{u\,(1+u)^2} = \frac{e^{2t}}{\sqrt{e^{2t}-1}}. \tag{118}$$

One checks directly that uniform t sampling with this weight induces exactly $w_{\text{eff}}^{\text{FM}}$ of equation 114. The two limits are the ones anticipated in §A.3: at large t, $u \simeq e^t$ and $w_{\text{FM}\to\text{DM}} \simeq e^t = \sqrt{d/\Lambda}$, recovering the tilt of equation 115; at small t it diverges as $1/\sqrt{2t}$, which is regularized in practice by the cutoff $t_{\min} > 0$ used in training.

A.5 Masked discrete diffusion

The SNR construction of §A.3 is not tied to Gaussian interpolants. We treat here the masked discrete diffusion (MDLM) pipeline of Sahoo et al. (2024) used on the genomic data of §B.4.

Noise injection and SNR. In this case, each of the d tokens of $\mathbf{x}_0$ is independently replaced by `MASK` with probability $1 - e^{-\sigma(t)}$, so that a token survives with probability $e^{-\sigma(t)}$, and σ increases from $\sigma_{\min}$ at $t = 0$ to $\sigma_{\max}$ at $t = 1$.

In the discrete masking paradigm, a masked token carries no information at all, while an unmasked one carries it intact: masking destroys information by removing tokens rather than by attenuating them. The natural signal-to-noise ratio is therefore simply the expected number of tokens that survive,

$$\Lambda_t = e^{-\sigma(t)}\, d, \tag{119}$$

which plays exactly the role that $\Lambda_t = \alpha^2 d/\Gamma_t$ plays in the Gaussian case, both counting effective informative units (we absorb the $O_d(1)$ information per token into the normalization).

Effective weighting. As implemented, the ELBO equation 130 averages the cross-entropy over the masked positions and reweights it by $\dot{\sigma}/(1-e^{-\sigma})$, so that the per-token loss carries the time weighting $w(t) = \dot{\sigma}/(1-e^{-\sigma})$. With $\Lambda = e^{-\sigma}d$ we have $|\mathrm{d}t/\mathrm{d}\Lambda| = 1/(\dot{\sigma}e^{-\sigma}d)$, and the $\dot{\sigma}$ cancels:

$$w_{\text{eff}}^{\text{MDLM}}(\Lambda) = \frac{\dot{\sigma}}{1-e^{-\sigma}} \cdot \frac{1}{\dot{\sigma}\, e^{-\sigma} d} = \frac{d}{\Lambda\,(d-\Lambda)} \underset{\Lambda \ll d}{\simeq} \frac{1}{\Lambda}. \tag{120}$$

Two consequences follow. First, equation 120 does not involve $\sigma(t)$: the effective weighting is independent of the noise schedule, a discrete counterpart of the known schedule invariance of the masked diffusion ELBO. Only the endpoints $\sigma_{\min}, \sigma_{\max}$ matter, through the range of Λ they expose. Second, comparing with equation 113, MDLM has the same $1/\Lambda$ low SNR behavior as Diffusion: measured in SNR, masked discrete diffusion is scale-free and rather trains like a standard DSM than a FM.

B Numerical details

B.1 Fig. 1 and Fig. 2

In these experiments, we consider synthetic Gaussian-mixture data in $\mathbb{R}^d$ as described in the main text, focusing on two datasets:

- **Unbalanced** (two-mode GMM): a single signal direction $\boldsymbol{\mu}$ with unequal mode probability $\omega = 0.8$, swept over the dimension $d \in \{64, 128, 256\}$ at fixed ω.
- **Kappa dependence** (Quadrimodal GMM): two orthogonal signal directions $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_2$ of unequal strength (κ vs. $1-\kappa$), each with its own mode probability $\omega_1 = \omega_2 = 0.7$ (asymmetric $\pm$ sign per mode). $d = 1024$ fixed, swept over $\kappa \in \{0.1, 0.2, 0.3\}$.

In these experiments the score's functional form is known exactly from the Gaussian-mixture log-density at all noise levels, and only a handful of scalar/vector coefficients in that closed form are learned, by denoising score matching (DSM) of a VP-SDE at a single, fixed noise level. In both settings we compare two fixed DSM noise times $t \in \{0.5,\, t_s\}$ side by side, where $t_s = \frac{1}{2}\log d$. There is no sampling or sample generation anywhere in either experiment: because the score's closed form is known exactly, model performance is evaluated directly from the learned coefficients themselves, rather than from samples generated by the model.

B.1.1 Data

During training, data samples $\mathbf{x}$ are generated as follows:

Unbalanced GMM (dimension sweep). $\mathbf{x} = s\boldsymbol{\mu} + \mathbf{z}$, $\boldsymbol{\mu} = \mathbf{1}_d$ (all-ones, so $\|\boldsymbol{\mu}\|^2 = d$), $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $s = +1$ with probability $\omega = 0.8$ (else -1). Swept over $d \in \{64, 128, 256\}$ at fixed $\omega = 0.8$.

Quadrimodal GMM (kappa-dependence sweep). $\mathbf{x} = s_1\boldsymbol{\mu}_1 + s_2\boldsymbol{\mu}_2 + \mathbf{z}$, $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $d = 1024$. $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are orthogonal block-indicator vectors ($\|\boldsymbol{\mu}_1\|^2 = k_1 = \lfloor \kappa d \rfloor$, $\|\boldsymbol{\mu}_2\|^2 = k_2 = \lfloor (1-\kappa) d \rfloor$), with unbalanced signs: $s_1 = +1$ with probability $\omega_1 = 0.7$ (else -1), independently $s_2 = +1$ with probability $\omega_2 = 0.7$. Swept over $\kappa \in \{0.1, 0.2, 0.3\}$.

B.1.2 Model architecture

As discussed in depth in the main text, both experiments exploit closed-form knowledge of the true score, parameterizing it with a small number of scalar/vector coefficients rather than a generic function approximator; metrics are read directly off these learned coefficients.

Unbalanced. The exact score of the two-mode mixture factorizes as

$$\nabla_{\mathbf{x}} \log P_t(\mathbf{x}) = -c\,\mathbf{x} + \mathbf{w}\tanh(\mathbf{x}^T\mathbf{w} + b), \tag{121}$$

with $\mathbf{w} \in \mathbb{R}^d$ playing the role of the learned $\boldsymbol{\mu}$-direction, c the pseudo variance, and the log weight $b \to \frac{1}{2}\log\frac{\omega}{1-\omega}$ at convergence. The model thus has $(d+2)$ learnable parameters in total.

Quadrimodal. Because the four modes have equal weight (given s_1, s_2) and $\boldsymbol{\mu}_1 \perp \boldsymbol{\mu}_2$ by construction, the exact score of the VP-SDE-noised mixture at a fixed time t similarly factorizes:

$$\nabla_{\mathbf{x}} \log P_t(\mathbf{x}) = -c\,\mathbf{x} + \mathbf{w}_1 \tanh(\mathbf{x}^T \mathbf{w}_1 + b_1) + \mathbf{w}_2 \tanh(\mathbf{x}^T \mathbf{w}_2 + b_2),$$

with $\mathbf{w}_1 \in \mathbb{R}^{k_1}$, $\mathbf{w}_2 \in \mathbb{R}^{k_2}$ each compactly supported on their own mode's subspace, plus scalars c, b_1, b_2: $d{+}3$ learnable parameters total.

Both models use a small initialization: $\mathbf{w}_i \sim d^{-\frac{1}{2}}\mathcal{N}(0, \boldsymbol{I}_d)$ (respectively $\mathbf{w}_i \sim k_i^{-\frac{1}{2}}\mathcal{N}(0, \boldsymbol{I}_{k_i})$), such that initially $\|\mathbf{w}_i\|^2 = O(1)$

B.1.3 Training procedure

For each fixed DSM time t, we let $\alpha_t = e^{-t}$, $\sigma_t = \sqrt{1 - e^{-2t}}$, and optimize the loss

$$\mathcal{L} = \mathbb{E}\|\sigma_t \,\mathrm{score}(\mathbf{x}_t) + \boldsymbol{\xi}\|^2 \tag{122}$$

where $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\xi}$ and the expectation is taken over training data $\mathbf{x}_0$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

B.1.4 Unbalanced, dimension dependence, Fig. 1

The two panels are the two fixed noising times: $t = 0.5$ *(left)*, which sits deep in **Regime II**, and $t = t_s = \frac{1}{2}\log d$ *(right)*, the **speciation** time. colors index $d \in \{64, 128, 256\}$, and the three curves of the legend are, for each d:

- **Overlap** - the direction overlap $\mathbb{E}_+[\mathbf{w}^T \boldsymbol{\mu}]/(e^{-t}d)$, normalized to 1 at convergence.
- **c** - the learned scalar c, which plays the role of an inverse-variance-like precision term in the exact-score formula.
- **weight** - the learned mode weight $\omega = 0.5 + |\mathrm{sigmoid}(2b) - 0.5|$, read off the bias. The dotted horizontal line marks its target value $\omega = 0.8$.

The x axis is the number of SGD steps divided by d.

B.1.5 Kappa dependence, Fig. 2

The figure has three panels: $t = 0.5$ *(left)*, deep in **Regime II**; $t = t_s$ *(middle)*, at **speciation**; and the same speciation data replotted against a rescaled x axis *(right)*. colors index $\kappa \in \{0.1, 0.2, 0.3\}$, and for each κ the two curves of the legend are the overlaps along the two orthogonal directions,

$$m_1 = \frac{\mathbb{E}_+[\mathbf{w}_1^T \boldsymbol{\mu}_1]}{e^{-t}\kappa d}, \qquad m_2 = \frac{\mathbb{E}_+[\mathbf{w}_2^T \boldsymbol{\mu}_2]}{e^{-t}(1-\kappa) d}, \tag{123}$$

each normalized to 1 at convergence (m_1 dashed, m_2 solid). The learned biases b_i are not displayed here.

The x axis of the first two panels is the raw number of SGD steps. In the third it is rescaled by the growth rate of the corresponding block, written $\lambda(\sqrt{\kappa_i})$ in the figure and equal to the closed-form rate $\lambda(\gamma_i)$ of equation 13, evaluated at $\gamma_i^2 = \kappa_i \Lambda_t \Gamma_t$ with $\kappa_1 = \kappa$ and $\kappa_2 = 1 - \kappa$.

Relation to $(\mathcal{D}_2)$. The hierarchical target equation 7 of the main text is balanced, all four modes carrying weight $1/4$, whereas the runs here draw each sign with probability $\omega_1 = \omega_2 = 0.7$. The exact score model keeps one bias per block, b_1, b_2, initialized at zero, so this imbalance is representable and is learned alongside the overlaps. Strictly, each block then follows the coupled (m_i, b_i) dynamics of Result 1 rather than the balanced reduction of Result 2, the off-diagonal coupling being proportional to $2\omega_i - 1$. This does not affect what the figure tests: the growth rate of block i remains a function of (q_i, γ_i) alone, hence of $\kappa_i \Lambda$, so the collapse variable is unchanged and the balanced closed form $\lambda(\gamma_i)$ is used as the rescaling ansatz. The ResNet experiments below instead use the balanced target exactly.

Training hyperparameters are reported below:

	Unbalanced and Kappa-dependence
Optimizer	plain SGD (no momentum, no weight decay, constant LR)
Learning rate	10^{-1}
Batch size	512
Training steps	10^6 (SGD steps, no early stopping)
t_{noise}	0.5 or $t_s = \frac{1}{2}\log d$
Seeds per config	20
Sweep values	$d \in \{64, 128, 256\}$ at $\omega = 0.8$ $\kappa \in \{0.1, 0.2, 0.3\}$ at $d = 1024$

Table 1: Hyperparameters shared by all 20 runs averaged to produce each figure. Metrics are logged at 200 log-spaced steps.

B.2 Fig. 3a and Fig. 3b

In these experiments, we return to the same two Gaussian-mixture datasets introduced above — unbalanced and quadrimodal GMMs, with data generated as in §B.1.1 except that the quadrimodal signs are now drawn with equal probability, $\omega_1 = \omega_2 = 0.5$, so that the target is exactly the balanced ($\mathcal{D}_2$) of equation 7; the unbalanced target keeps $\omega = 0.8$ — but replace the analytically exact score parameterization with a generic fully-connected ResNet epsilon-predictor, trained by denoising score matching (DSM) over a continuous range of noise levels rather than at a single fixed t. Correspondingly, since the score is no longer known in closed form, model performance can no longer be read directly off a handful of coefficients: instead, we evaluate it by drawing samples from the model via reverse-time SDE integration and comparing statistics of those samples to the training distribution. We contrast two training objectives, standard DSM and FM-weighted DSM.

B.2.1 Model architecture

Both experiments use the same fully-connected ResNet epsilon-predictor $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{x}, t) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$: a linear input projection to width 128, 4 residual blocks (LayerNorm, additive sinusoidal-time-embedding conditioning via a linear projection, then a 2-layer SiLU MLP, residual add), and a zero-initialized linear output projection back to $\mathbb{R}^d$. Sinusoidal time embedding dimension 128. The score is reparameterized with a skip connection. The input/output projections are the only d-dependent layers.

B.2.2 Training procedure

In both experiments we follow the standard denoising score matching procedure, where noised samples are obtained via the forward continuous-time variance-preserving Ornstein-Uhlenbeck SDE: $\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{b}_t$, with closed-form marginal $\mathbf{x}_t \mid \mathbf{x}_0 \sim \mathcal{N}(e^{-t}\mathbf{x}_0, \sqrt{(1 - e^{-2t})}\boldsymbol{I})$. The network $\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}$ predicts the injected noise $\boldsymbol{\xi}$, and the score used in the backward inference dynamics is recovered via Tweedie's formula, $\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = -\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{x}, t)/\sigma(t)$. Sampling integrates the reverse SDE with Euler–Maruyama from $t_{\max}$ down to $t_{\min} > 0$.

To evaluate the impact of weighting schemes, we use a weighted denoising score matching loss

$$\mathcal{L} = \mathbb{E}_{t\sim\mathcal{U}(t_{\min}, t_{\max})}\left[w(t)\,\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|^2\right] \tag{124}$$

with two choices of $w(t)$:

- **Uniform DSM**: $w(t) = 1$.
- **FM-weighted DSM**: $w(t) = e^{2t}/\sqrt{e^{2t} - 1}$, the matched weight equation 118 derived in §A.4.

For each training we use plain SGD with a constant learning rate and neither momentum nor weight decay. The curves report the mean over the generated samples at each logged step and the shaded band the corresponding 95% CI, so the band measures sampling error at fixed model rather than run-to-run variability. Inference and training hyperparameter values are reported below:

	Unbalanced GMM	Quadrimodal GMM
d	64	64
Optimizer	plain SGD (no momentum, no weight decay, constant LR)	
Learning rate	10^{-2}	10^{-2}
Batch size	2048	2048
Training steps (SGD)	10^6	10^6
$t_{\min}, t_{\max}$	$5 \times 10^{-3},\ 6.2$	$10^{-3},\ 6.2$
Reverse SDE steps (eval)	30	30
Eval samples / logged step	5×10^4	5×10^4
Sweep values	$\omega = 0.8$	$\kappa \in \{0.2, 0.3, 0.4\}$

Table 2: Hyperparameters associated to each figure. Both objectives (uniform / FM-weighted DSM) use identical hyperparameters otherwise.

B.2.3 UNBALANCED GMM, FIG. 3A

The two panels are the two weighting schemes of equation 124, labeled **Uniform DSM** ($w(t) = 1$) and **FM-weighted DSM**, at the single imbalance $\omega = 0.8$ marked by the dotted horizontal line. For each we monitor the two curves of the legend:

- **Cosine Similarity** - the cosine between an inferred sample and the mode direction, $\cos(\mathbf{x}, \boldsymbol{\mu}) = \mathbf{x}^T\boldsymbol{\mu}/(|\mathbf{x}||\boldsymbol{\mu}|)$. Since the target is symmetric, we average the mean cosine over the two signs of $\mathbf{x}^T\boldsymbol{\mu}$ and report $\frac{1}{2}\big(|\mathbb{E}_+[\cos(\mathbf{x}, \boldsymbol{\mu})]| + |\mathbb{E}_-[\cos(\mathbf{x}, \boldsymbol{\mu})]|\big)$. This tracks direction learning: a transition to $O(1)$ values indicates that the inferred samples are spread along the preferential training direction $\boldsymbol{\mu}$. Note that this does not mean the inferred distribution is necessarily bimodal, only that its covariance matrix is no longer isotropic.
- **Positive fraction** - the fraction of generated samples falling on the dominant side of the mode direction, $\frac{1}{N}\#\{\mathbf{x} : \mathbf{x}^T\boldsymbol{\mu} > 0\}$. It converges to ω once the imbalance has been acquired.

Each panel carries an inset showing the histogram of the projection $p = \mathbf{x}^T\boldsymbol{\mu}/\|\boldsymbol{\mu}\|^2$ of the generated samples, at the three training times marked by the vertical dashed lines; the dotted verticals mark the target mode locations ± 1.

B.2.4 QUADRIMODAL GMM, FIG. 3B

Again the two panels are the two weighting schemes, **Uniform DSM** and **FM-weighted DSM**, and colors index $\kappa \in \{0.2, 0.3, 0.4\}$. The monitored quantity is the cosine similarity between the inferred samples and the subdominant direction $\boldsymbol{\mu}_1$, defined as in §B.2.3 with $\boldsymbol{\mu}$ replaced by $\boldsymbol{\mu}_1$; since $\|\boldsymbol{\mu}_1\|^2 = \kappa d$, smaller κ means a weaker subdominant mode.

Only the **FM-weighted DSM** panel carries an inset, which replots that panel's curves against the rescaled x axis, training steps $\times\, \lambda(\sqrt{\kappa})$, with λ the closed-form rate equation 13 evaluated at $\gamma_i^2 = \kappa_i \Lambda$.

B.3 FIG. 4A AND FIG. 4B

We extend the ResNet/GMM experiments with two companion experiments on MNIST digits $\{3, 6\}$, sharing the same generative architecture and varying a different axis of the data-generating process:

- **Weight learning** (class-imbalanced grayscale MNIST): only digit identity varies, but the two classes appear in an imbalanced ratio $\omega = P(\text{class} = 6)$ in the training set.
- **Structure learning** (tinted MNIST): both digits are shown in a random, class-independent color (red or green) at a fixed contrast α. The data has two orthogonal directions of variation — digit identity ("3 vs. 6") and color ("red vs. green") — and α controls the strength of the (weaker) color signal.

B.3.1 DATA

Both experiments restrict MNIST (train split) to digits 3 and 6 ($n_3 = 6{,}131$, $n_6 = 5{,}918$ images, 28×28).

Class-imbalanced MNIST. Grayscale (1 channel), normalized to $[-1, 1]$. All n_6 sixes are kept and threes are subsampled to hit a target mixture weight $\omega \in \{0.5, 0.8\}$: $n_{\text{tot}} = \lfloor n_6/\omega \rfloor$, $n_3 = n_{\text{tot}} - n_6$ (so $\omega = 0.5$: balanced, $n_3 = n_6 = 5{,}918$; $\omega = 0.8$: $n_3 = 1{,}479$, $n_6 = 5{,}918$, total 7,397).

Tinted MNIST. All 12,049 available 3s and 6s are used, no subsampling. Each grayscale image $\mathbf{x} \in [0, 1]^{28\times 28}$ is expanded to 3 channels and randomly tinted red or green (uniformly, independently of digit label) at contrast $\alpha \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$:

$$\mathbf{x}_{\text{tinted}} = (1 - \alpha)\, \bar{\mathbf{x}} + \alpha\, \bar{\mathbf{x}} \odot c, \qquad \bar{\mathbf{x}} = (\mathbf{x}, \mathbf{x}, \mathbf{x}), \quad c \in \{(1, 0, 0), (0, 1, 0)\}, \tag{125}$$

then normalized to $[-1, 1]$ per channel. $\alpha = 0$ would give plain grayscale (no hue signal); $\alpha = 1$ gives a fully saturated tint; see Fig. 6 for a sample.



Figure 6: Image sample at tint $\alpha = 0.7$.

B.3.2 MODEL ARCHITECTURE

Both experiments use the same U-Net backbone (`torchcfm.models.unet.UNetModel`): model channels 32, 2 residual blocks per resolution for the imbalance experiment and 1 for the tinted one, channel multiplier $(1, 2, 2)$ (the package's default for 28×28 inputs), self-attention at the coarsest resolution (single head), no class conditioning, dropout 0, GroupNorm+SiLU residual blocks with a sinusoidal timestep embedding. Input/output channels are 3 for the tinted experiment and 1 for the grayscale one, both $\approx 1.08 \times 10^6$ parameters.

That backbone is used in two different roles. Under Flow Matching it is the velocity field $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ regressed in equation 128. Under DSM it parametrizes the score through a trainable skip connection,

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}, t) = -\mathbf{x} + \text{UNet}_{\boldsymbol{\theta}}(\mathbf{x}, t), \tag{126}$$

rather than the usual $-\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}}(\mathbf{x}, t)/\sigma(t)$: at initialization the network output is small, so the score defaults to $-\mathbf{x}$, the score of a standard normal, instead of an untrained noise prediction divided by a possibly tiny $\sigma(t)$. This keeps reverse-SDE sampling stable before the network has learned anything.

B.3.3 FLOW-MATCHING TRAINING PROCEDURE

In the colorized experiment, we use the standard conditional Flow Matching (CFM): for a clean sample $\mathbf{x}_1$ and $\mathbf{x}_0 \sim \mathcal{N}(0, \boldsymbol{I})$,

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad t \sim \mathcal{U}(0, 1), \tag{127}$$

and the network is trained to regress the constant conditional velocity $\dot{\mathbf{x}}_t = \mathbf{x}_1 - \mathbf{x}_0$ with an MSE loss,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2 \tag{128}$$

	Structure sweep (tinted)	Weight sweep (imbalanced)
Optimizer	AdamW, weight decay 10^{-2} (default), constant LR	
Learning rate	10^{-3}	
Batch size	128	
Training steps	2,000	10,000
Sweep values	$\alpha \in \{0.4, \dots, 1.0\}$ (step 0.1)	$\omega = 0.8$
Seeds per value	10	5
Eval samples / logged step	10,000 (1,000 $\times$ 10 repeats)	5,000 (1,000 $\times$ 5 repeats)
Classifier epochs	6	20

Table 3: Hyperparameters for the two experiments.

Sampling integrates $\mathbf{v}_{\boldsymbol{\theta}}$ from $t = 0$ (Gaussian noise) to $t = 1$ with an adaptive-step Dormand–Prince (`dopri5`) ODE solver (`torchdyn`, $\text{atol} = \text{rtol} = 10^{-4}$). The training and inference parameters are reported below:

The imbalance experiment is run twice, once with the Flow Matching objective above and once with the score-based DSM objective of §B.2.2, which is the pair compared in Fig. 4a. The DSM side uses the same U-Net with the score skip connection equation 126, a unit-rate Ornstein-Uhlenbeck forward process $\mathrm{d}\mathbf{x} = -\mathbf{x}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{b}_t$ run over $t \in [5 \times 10^{-3}, 3.3]$, and generates by integrating the reverse SDE over 50 steps with cutoff $\epsilon_t = 5 \times 10^{-3}$. All optimization hyperparameters are those of Tab. 3.

B.3.4 Weight learning, Fig. 4a

Both panels show the same unbalanced dataset at fixed $\omega = 0.8$ (dotted horizontal line); what differs is the training objective, **DSM** on the left and **Flow Matching** on the right, so the two panels isolate the effect of SI pipeline at fixed data. Curves are the mean over the 5 seeds, the band a 95% CI across seeds, Gaussian-smoothed along the step axis.

At each logged step we draw samples from the current model and pass them through two fixed evaluation probes, neither of which is ever used to train the generative model:

- **An auxiliary digit classifier**, a 1-channel, 2-layer CNN trained once per run on real data with a matching noise-augmentation scheme (classify $(1-t)\mathbf{x}_0 + t\mathbf{x}_1$ against the true label, $t \sim \mathcal{U}(0,1)$, so it stays calibrated at the high-noise end of the path). It is always fit on a balanced ($\omega = 0.5$) split, so it judges generated class frequency without inheriting the training-set imbalance.
- **Class-mean projections.** $\boldsymbol{\mu}_3, \boldsymbol{\mu}_6$, the pixel-space mean image of each digit class in the training set, precomputed per run; every generated sample $\mathbf{x}$ is projected onto both, $\mathbf{x}^T\boldsymbol{\mu}_3$ and $\mathbf{x}^T\boldsymbol{\mu}_6$.

The two curves of the legend are then:

- **Cosine Similarity** - $\mathbb{E}\big[\max\big(\cos(\mathbf{x}, \boldsymbol{\mu}_3), \cos(\mathbf{x}, \boldsymbol{\mu}_6)\big)\big]$, the cosine between a generated sample and the nearer of the two class-mean images.
- **Fraction of 6s** - $\mathbb{E}[\mathbf{1}_{C(\mathbf{x})=6}]$, the empirical fraction of generated samples the classifier assigns to digit 6.

B.3.5 Structure learning, Fig. 4b

Two panels, the mean over the 10 seeds per α, the band a 95% CI across seeds, Gaussian-smoothed along the step axis. The digit direction is read off the same class-mean projections as above; the color direction uses a hue projection, the mean red minus mean green pixel intensity of a generated image, $r - g$ ("color spread").

- Left panel - the digit direction overlap, defined as $\mathbb{E}\big[\max(\frac{|\mathbf{x}^T\boldsymbol{\mu}_3|}{|\boldsymbol{\mu}_3|^2}, \frac{|\mathbf{x}^T\boldsymbol{\mu}_6|}{|\boldsymbol{\mu}_6|^2})\big]$.

- Right panel - the color spread. Since $r-g$ is bimodal by construction we keep only its positive part and monitor $\mathbb{E}_+[\alpha\cdot(r-g)]/\alpha^2$. The α^{-2} normalization reflects the squared norm of the training color spread scaling as α^2, mirroring the $|\boldsymbol{\mu}_{3,6}|^2$ normalization of the digit direction.

To rescale the time axis of the subdominant direction we identify the relative direction amplitude $|\boldsymbol{\mu}_1|^2/|\boldsymbol{\mu}_2|^2$ as proportional to α^2; rescaling training time by equation 13 then collapses the curves.

B.4 FIG. 5

B.4.1 DATASET

We use the 805-SNP subset of phased haplotypes from the 1000 Genomes Project (805_SNP_1000G_real.hapt). Each of the $2{,}504$ sequenced individuals contributes two phased haplotypes, giving $N = 5{,}008$ binary sequences of length $L = 805$, $\mathbf{x} \in \{0,1\}^L$. The data is split into a $90/10$ training and validation set. A reference PCA basis $\{\boldsymbol{\mu}_k\}_{k=1}^K$, $K = 50$, together with the reference eigenvalues $\{\lambda_k\}_{k=1}^K$, is fit once on the training split and kept fixed throughout training for evaluation purposes.

B.4.2 MASKED DISCRETE DIFFUSION PROCESS

We use continuous-time absorbing-state (masked) discrete diffusion (MDLM) (Sahoo et al., 2024), adapted to a binary alphabet. Each token from the original sequence $\mathbf{x}_0$ is independently replaced in the noised sequence $\mathbf{x}_t$ by **MASK** with probability $\alpha(t) = 1 - e^{-\sigma(t)}$, under a log-linear noise schedule

$$\sigma(t) = \sigma_{\min}^{1-t}\,\sigma_{\max}^{t}, \qquad t \in [0,1], \tag{129}$$

with $\sigma_{\min} = 10^{-4}$, $\sigma_{\max} = 20$. The model is trained with the continuous-time ELBO (masked cross-entropy at masked positions only, reweighted by $\dot\sigma(t)/(1 - e^{-\sigma(t)})$), with $t \sim \mathcal{U}(\boldsymbol{\xi}, 1)$, $\epsilon = 10^{-4}$, resampled independently for every sequence in a minibatch:

$$\mathcal{L}(\boldsymbol{\theta}) = -\int_\epsilon^1 \frac{\dot\sigma(t)}{(1 - e^{-\sigma(t)})}\,\mathbb{E}_{\mathbf{x}_t,\mathbf{x}_0}\left[\log(p_{\boldsymbol{\theta}}(\mathbf{x}_0|\mathbf{x}_t))\right]. \tag{130}$$

Sampling uses the ancestral reverse process as originally proposed in the MDLM paper, starting from a fully masked sequence and unmasking tokens over $n_{\text{steps}} = 30$ discretized reverse steps. At each logged step, 10×500 samples are drawn from the current model to evaluate PCA overlap statistics.

B.4.3 MODEL ARCHITECTURE

The denoiser $p_{\boldsymbol{\theta}}$ is a small Transformer encoder, operating on a 3-token vocabulary $\{0, 1, \text{MASK}\}$:

- **Input embedding.** A learned token embedding ($3 \times d_{\text{model}}$) plus a learned absolute positional embedding ($L \times d_{\text{model}}$), summed.
- **Time conditioning.** The noise level enters through $\log\sigma(t)$, mapped by a sinusoidal embedding of dimension d_{cond} followed by a 2-layer MLP ($d_{\text{cond}} \to 4d_{\text{cond}} \to d_{\text{model}}$, GELU nonlinearity), and added to every token position (broadcast over the sequence).
- **Backbone.** n_{layers} pre-norm blocks (d_{model}, n_{heads} heads, feed-forward width $4d_{\text{model}}$, GELU, dropout p), full self-attention.
- **Output head.** LayerNorm $\to$ linear projection to 3 logits, followed by the SUBS parameterization of MDLM: the MASK logit is set to $-\infty$; at unmasked positions the output distribution is forced to a point mass on the observed token; at masked positions the model outputs a free $\mathrm{softmax}$ over $\{0,1\}$. The output projection is zero-initialized.

For the run analyzed here: $d_{\text{model}} = 64$, $n_{\text{heads}} = 4$, $n_{\text{layers}} = 2$, $d_{\text{cond}} = 64$, dropout $p = 0.1$, giving $\approx 1.85 \times 10^5$ trainable parameters.

Training parameters are reported below

Optimizer	AdamW, weight decay 10^{-2}
Peak learning rate	3×10^{-2}
LR schedule	linear warm-up (50 steps) $\rightarrow$ constant $\rightarrow$ linear decay to $10^{-3}\times$ peak over the final 20% of training (decay starts at step $80{,}000$)
Batch size	256
Training steps	$100{,}000$
Gradient clipping	global norm 1.0
Seeds	4 independent runs

Table 4: Optimization hyperparameters for the run shown in Fig. 5

B.4.4 FIGURE 5: PCA-LEARNING COLLAPSE

At every logged training step s, and for each of the $5{,}000$ freshly generated samples, we:

1. Fit a fresh K-component PCA, $\{\boldsymbol{\mu}_k^{\mathrm{inf}}(s)\}$, on the generated batch ("inferred" PCA at step s).
2. Compute the overlap matrix between the inferred and the fixed reference eigenvectors,
$$O_{ij}(s) = \frac{1}{|\boldsymbol{\mu}_j^{\mathrm{ref}}|^2}\, \boldsymbol{\mu}_i^{\mathrm{inf}}(s) \cdot \boldsymbol{\mu}_j^{\mathrm{ref}}, \qquad i,j = 1,\dots,K, \tag{131}$$
3. Pair inferred and reference directions one to one, through the permutation π_s of $\{1,\dots,K\}$ maximizing the total matched overlap $\sum_j |O_{\pi_s(j)j}(s)|$ (Hungarian algorithm). The matched value $|\boldsymbol{\mu}_{\mathrm{inf}}\cdot\boldsymbol{\mu}_{\mathrm{ref}}|_k(s) := |O_{\pi_s(k)k}(s)|$ tracks how well the k-th reference principal direction has emerged in the generated samples by training step s.

Ranking the entries of O and reading the k-th largest as the k-th direction would be ambiguous on two counts: the curves would then be non-increasing in k by construction, whatever the model has learned, and a single dominant inferred component leaking onto several reference eigenvectors could occupy several of the top slots. Both matter here, since the inset of Fig. 5 rescales curve k by the timescale of reference direction k specifically. The two constructions agree on the two leading directions, each carried by a single inferred component, and differ on the subleading ones. The reported curves are the mean and 95% CI of $|\boldsymbol{\mu}_{\mathrm{inf}}\cdot\boldsymbol{\mu}_{\mathrm{ref}}|_k(s)$ across the 4 training seeds, lightly smoothed along the step axis (Gaussian filter, $\sigma = 0.1$ index units), for the top 5 modes.

Left panel. A plain scatter of $\mathbf{x}^T\boldsymbol{\mu}_0$ vs. $\mathbf{x}^T\boldsymbol{\mu}_1$ (the projections of each haplotype onto the top-two reference PCA directions), using every one of the $N_{\mathrm{train}} = 4{,}508$ real training haplotypes – no subsampling. Both the scatter and the reference directions $\{\boldsymbol{\mu}_k\}$ come from the same reference PCA fit.

Right panel. $|\boldsymbol{\mu}_{\mathrm{inf}}\cdot\boldsymbol{\mu}_{\mathrm{ref}}|_k$, labeled **PCA Overlap** on the figure, plotted against the raw training step s; the inset shows the same curves against the rescaled step $\lambda_k \times s$, where they collapse. To identify the rescaling timescales we extract the explained variance v_k along each of the reference PCA directions. The GMM growth rate equation 13 is a function of the block SNR γ_i, which at the speciation scale reduces to $\gamma_i = \sqrt{\kappa_i}$, with $\kappa_i = \|\boldsymbol{\mu}_i\|^2/d$ the squared amplitude of direction i. Identifying that squared amplitude with the variance explained along direction k, we rescale the time axis by
$$\lambda_k \equiv \lambda\left(\sqrt{v_k}\right). \tag{132}$$
The PCA projected Human Genome dataset is of course not a GMM, and the rescaling should only be expected to hold where the projected data actually resolves into modes. This is the case for the two leading directions, which collapse onto one another; the subsequent ones, along which no mode separation is visible, do not, and the ansatz carries less meaning there.

C LLM USAGE

We acknowledge the use of LLMs to assist in drafting the manuscript and refining its clarity. The tool was also used to support the derivation of mathematical proofs and the writing of the code. All AI-generated content was rigorously reviewed, verified, and edited by the authors. The authors bear full responsibility for the originality and scientific integrity of this work.