DISINTEGRATED STOCHASTIC INTERPOLANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based and flow-based models are often applied to structured data in scientific applications that obey conservation laws, symmetries, or equality constraints. This structural knowledge should, ideally, simplify the learning task, but, in practice, does the opposite due to data degeneracies. We provide the missing theoretical foundation to overcome such issues: we introduce the *disintegrated stochastic interpolants*, a generalisation of the stochastic interpolant framework that provides a unified approach to exploiting structure in both flow-based and diffusion-based models. Our construction begins with a structural map whose fibres partition the state space. Disintegrating probability measures along these fibres decomposes the transport between a source and a target measure *into a transport within and between fibres*. This decomposition formalises the exploitable structure: when the transport between fibres can be specified analytically, learning can focus on the remaining within-fibre transport. We establish conditions under which these components recover the desired ambient evolution and derive corresponding training objectives and sampling algorithms. Furthermore, we show that under suitable choices of structural map, this framework encompasses equality constraints, conservation laws, and symmetry. Across geometric and physics-based problems, our disintegrated models learn target distributions more accurately at matched training budgets and satisfy prescribed structure substantially better than alternatives.

1 INTRODUCTION

Generative modelling often involves learning distributions whose structure is partially known in advance, e.g., physical systems are restricted to obey conservation laws, equality constraints, and symmetries. Such restrictions offer an opportunity to simplify model learning, but we often lack the mathematical tools to realise the opportunity. For example, a density restricted to satisfying known constraints will *appear degenerate* to a naive learner, often resulting in theoretical inconsistencies and ill-conditioned optimisation problems (Liu et al., 2025; Zeng & Yan, 2025; see appendix D).

In this paper, we introduce *disintegrated stochastic interpolants*, a generalisation of *stochastic interpolants* (Albergo et al., 2025), that let us prescribe parts of the generative evolution. We show that under known structural information, the target density always decomposes into two parts: one is fully determined by the known structure and the other is free to be learned. Neatly, this decomposition is easily realised in the generating differential equation, giving simple and scalable algorithms.

To realise the above, we *first* develop a measure-theoretic framework for disintegrated stochastic interpolants and identify conditions under which prescribed structural dynamics and learned conditional dynamics recover the desired marginal evolution, including for singular targets. We explain how suitable choices of the structural map describe symmetry, conserved quantities, and equality constraints. *Second*, we derive training objectives for the remaining velocity and, under additional regularity assumptions, fibrewise score fields, yielding flow and diffusion constructions. *Third*, we develop the first numerical methods to implement this separation using explicit fibre coordinates or implicitly defined constraints. *Finally*, we evaluate distributional accuracy, structural fidelity, and computational cost on geometric and physics-based problems, with particular attention to singular targets and the regimes in which prescribing structure improves learning. *The practical consequence of our work is that prescribed structure need not be learned, paving the way for smaller models, higher accuracies, and easier training.*

Figure 1 illustrates two practical benefits of our work: enforcing hard constraints by construction and improving training efficiency by reducing what must be learned from data.

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.

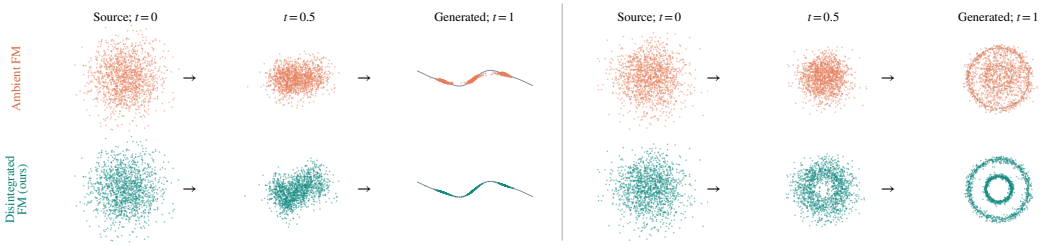


Figure 1: An illustration of how disintegrated flow matching exploits known structure for a singular target (left) and a symmetric target (right). Ambient (top) and our disintegrated (bottom) flow matching at $t \in 0, 0.5, 1$, with matched training budgets and 256 neural function evaluations for sampling. **Left:** For a sine-curve target, disintegrated flow matching analytically drives the vertical offset to zero while learning transport along the moving level sets, satisfying the constraint exactly where ambient flow matching retains off-curve error. **Right:** For a rotationally invariant target concentrated in two annuli, disintegrated flow matching preserves the uniform angular distribution and learns only radial transport, capturing the target's radial structure more accurately.

Background: stochastic interpolants. Our goal is to transport a tractable source measure ρ_0 with density p_0 to the data measure ρ_1 with density p_1, and the transport is built from the interpolant, $X_t = I(t, X_0, X_1) + \gamma_t Z$, where (X_0, X_1) is drawn from a coupling of ρ_0 and ρ_1, and $Z \sim \mathcal{N}(0, I)$ is independent. The map I is sufficiently regular and satisfies $I(0, x_0, x_1) = x_0$ and $I(1, x_0, x_1) = x_1$, while the noise schedule obeys $\gamma_0 = \gamma_1 = 0$, so that X_t interpolates between the two measures. Albergo et al. (2025) show that the velocity field $b(t, \mathbf{x}) = \mathbb{E}[\partial_t I(t, X_0, X_1) + \dot{\gamma}_t Z \mid X_t = \mathbf{x}]$ generates the marginals, i.e., that the solution of $\dot{\hat{X}}_t = b(t, \hat{X}_t)$ started at $\hat{X}_0 \sim \rho_0$ has the same distribution as X_t for every $t \in [0, 1]$. Samples from ρ_1 can therefore be obtained by drawing from ρ_0 and integrating this ordinary differential equation (ODE) to $t = 1$. The same construction yields the training objectives of score-based diffusion models. In this setting, everything is stated in terms of densities, and absolute continuity with respect to the Lebesgue measure is an explicit assumption. In appendix A we rederive the key results presented in Albergo et al. (2025) in measure-theoretic language, which is necessary to build our generalisations.

2 Theory: Disintegrated Stochastic Interpolants

In this section we extend stochastic interpolants to data whose structure is partly known in advance. Let ρ_0 and ρ_1 be probability measures on $\mathbb{R}^d$, and let $\mathbf{c} : \mathbb{R}^d \to \mathcal{Y}$ be a continuously differentiable map into a smooth Riemannian manifold $\mathcal{Y}$ of dimension $k < d$ (for simplicity, it is possible to think of $\mathcal{Y}$ as $\mathbb{R}^k$). We call $\mathbf{c}(x)$ the *structural value* of x. We can partition any distribution on $\mathbb{R}^d$ along the fibres

$$\mathcal{M}_y = \{x \in \mathbb{R}^d : \mathbf{c}(x) = y\}, \qquad y \in \mathcal{Y}, \tag{1}$$

together with the conditional distributions on the fibres. We write $\nu_0 = \mathbf{c}_{\#}\rho_0$ and $\nu_1 = \mathbf{c}_{\#}\rho_1$ for the base laws of the source and the target.

The function c encodes the data's known structure; specifically, the target base law ν_1 is known. To illustrate, consider two common kinds of structure: singularity and symmetry.

Singularity. The target satisfies equality constraints and lies on the zero set of some function $\mathbf{c} : \mathbb{R}^d \to \mathbb{R}^k$; then the structure map is $\mathbf{c}$ and its pushforward is

$$\mathcal{Y} = \mathbb{R}^k, \qquad \nu_1 = \delta_0.$$

For example, for a target on the unit circle, $\mathbf{c}(x) = \|x\| - 1$ and the fibres are circles centred at the origin.

Symmetry. The target is invariant under a compact Lie group G, and $\mathbf{c}$ is equivariant into a space $\mathcal{Y}$ on which G acts transitively,

$$\mathbf{c}(g \cdot x) = g \cdot \mathbf{c}(x), \qquad \nu_1 = \text{uniform on } \mathcal{Y},$$

since ν_1 is then invariant and the uniform law is the only invariant probability on $\mathcal{Y}$. E.g., for rotations of the plane, $\mathbf{c}(x)$ is the angle of x, and the fibres are rays from the origin.

In general, the fibres are curved submanifolds of $\mathbb{R}^d$, so a bit of geometry is needed to define our disintegrated stochastic interpolants and describe how they transport measures.

2.1 GEOMETRY OF FIBRES

To describe transport between and within fibres, we separate the part of an ambient velocity that changes the structural value from what leaves it unchanged. Since the fibres are curved, we need differential operators along them. We use the Euclidean metric on $\mathbb{R}^d$ and a Riemannian metric on $\mathcal{Y}$.

Assumption 2.1 (Regularity of the structure map)**.** The map $\mathbf{c} : \mathbb{R}^d \to \mathcal{Y}$ is C^3 and has rank k on an open set $U \subseteq \mathbb{R}^d$, and the paths of the interpolants considered below stay in U almost surely.

Fibres and vertical vectors. Under assumption 2.1, the preimage theorem makes each fibre, restricted to U, an embedded submanifold of dimension $d - k$. We call its tangent space at x,

$$\mathcal{V}_x := T_x \mathcal{M}_{\mathbf{c}(x)} = \ker D\mathbf{c}(x), \tag{2}$$

the *vertical space*. A vector in $\mathcal{V}_x$ is tangent to the fibre through x and leaves $\mathbf{c}(x)$ unchanged to first order. The vertical spaces therefore describe the directions along which the data-generating ODE can transport mass.

Horizontal space and the splitting. To turn a prescribed change in the structural value into an ambient velocity, we choose a complementary space $\mathbb{R}^d = \mathcal{H}_x \oplus \mathcal{V}_x$ that varies smoothly with x. This is an *Ehresmann connection*. The *horizontal lift* $\mathsf{H}_x : T_{\mathbf{c}(x)}\mathcal{Y} \to \mathcal{H}_x$ is the unique linear map with image $\mathcal{H}_x$ that satisfies $D\mathbf{c}(x)\mathsf{H}_x = I$, and the *vertical projector*: $\mathsf{P}_x = I_d - \mathsf{H}_x D\mathbf{c}(x)$, keeps the component tangent to the current fibre. Every ambient velocity v therefore admits the unique decomposition

$$v = \mathsf{H}_x D\mathbf{c}(x)v + \mathsf{P}_x v. \tag{3}$$

The first term is determined by the rate $D\mathbf{c}(x)v$ at which the structural value changes. The second is tangent to the fibre through x. Figure 2 illustrates the decomposition.

Unless stated otherwise, we take the orthogonal complement $\mathcal{H}_x = \mathcal{V}_x^{\perp}$, for which

$$\mathsf{H}_x = D\mathbf{c}(x)^{\top}\big(D\mathbf{c}(x)D\mathbf{c}(x)^{\top}\big)^{-1}, \qquad \mathsf{P}_x = I_d - \mathsf{H}_x D\mathbf{c}(x). \tag{4}$$

Intrinsic operators on a fibre. Each fibre inherits the Euclidean metric of $\mathbb{R}^d$ and its associated Riemannian volume σ_y. For a function f and a tangent vector field w on a fibre, the intrinsic gradient, divergence, and Laplacian are

$$\nabla_{\mathcal{M}} f = \mathsf{P}_x \nabla f, \qquad \operatorname{div}_{\mathcal{M}} w = \operatorname{tr}(\mathsf{P}_x Dw), \qquad \Delta_{\mathcal{M}} f = \operatorname{div}_{\mathcal{M}}(\nabla_{\mathcal{M}} f).$$

Together with the horizontal lift, they provide the geometric operations used below to describe transport and diffusion within moving fibres.

2.2 DISINTEGRATED STOCHASTIC INTERPOLANTS

We now define the disintegrated stochastic interpolants. A coupling $\pi \in \Pi(\rho_0, \rho_1)$ is a probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ρ_0 and ρ_1.

Definition 2.2 (Disintegrated stochastic interpolant)**.** Let $(X_0, X_1) \sim \pi$ for a coupling $\pi \in \Pi(\rho_0, \rho_1)$, and let Z be an auxiliary random variable independent of (X_0, X_1). A disintegrated stochastic interpolant between ρ_0 and ρ_1 consists of a process

$$X_t = I(t, X_0, X_1, Z), \qquad t \in [0, 1], \tag{5}$$

together with a base schedule $B_t : \mathcal{Y} \to \mathcal{Y}$, such that

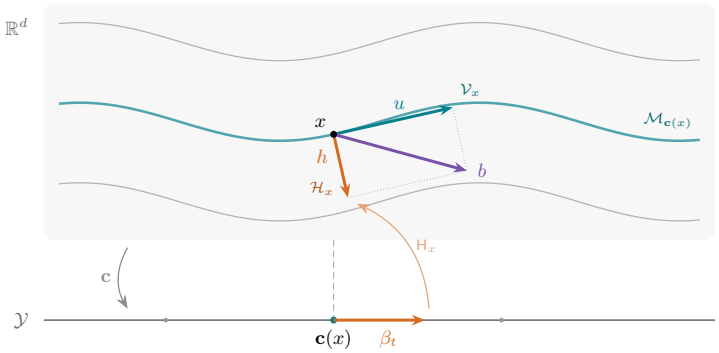


Figure 2: Velocity decomposition. The structure map $\mathbf{c}$ sends each fibre to a point of $\mathcal{Y}$. The known base velocity β_t is lifted to $h = \mathsf{H}_x \beta_t(\mathbf{c}(x)) \in \mathcal{H}_x$, and the learned component $u = \mathsf{P}_x b \in \mathcal{V}_x$ moves x along its fibre $\mathcal{M}_{\mathbf{c}(x)}$.

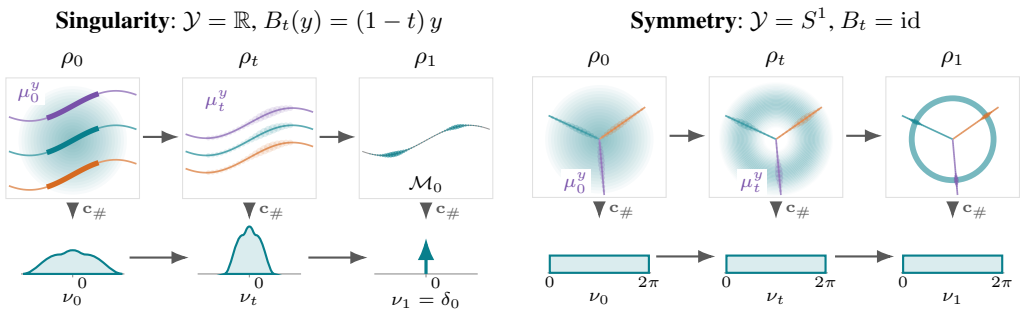


Figure 3: Disintegrated transport for the two structures. Each panel shows the laws ρ_0, ρ_t, ρ_1 above their base laws $\nu_t = \mathbf{c}_\#\rho_t$ on $\mathcal{Y}$. *Left, singularity:* the fibres are level sets of a constraint, and the schedule $B_t(y) = (1-t)y$ collapses ν_t to δ_0, so the target lies on $\mathcal{M}_0 = \mathbf{c}^{-1}(0)$. *Right, symmetry:* the fibres are rays and $\mathbf{c}$ is the angle, drawn on $[0, 2\pi)$, so $B_t = \mathrm{id}$ keeps ν_t uniform. The coloured curves are selected example fibres, and the marks on them show the conditional laws μ_t^y, which are learned. Density shapes are illustrative.

1. I is continuously differentiable in t, with $I(0, x_0, x_1, z) = x_0$ and $I(1, x_0, x_1, z) = x_1$. Writing $\dot{X}_t = \partial_t I(t, X_0, X_1, Z)$, we require $\mathbb{E}\int_0^1 \|\dot{X}_t\|^2 dt < \infty$.
2. $(t, y) \mapsto B_t(y)$ is continuously differentiable, with $B_0 = \mathrm{id}$ and $(B_1)_\#\nu_0 = \nu_1$. For $t < 1$, B_t is a diffeomorphism onto its image $\mathcal{Y}_t$, with base velocity $\beta_t(y) = \partial_t B_t\big(B_t^{-1}(y)\big)$.
3. The interpolant is *separable* with respect to $\mathbf{c}$: for almost every t, if

$$\int_{\mathbb{R}^d} \nabla\varphi(x) \cdot \mathsf{H}_x\big(D\mathbf{c}(x)\, b_t(x) - \beta_t(\mathbf{c}(x))\big)\, \rho_t(dx) = 0 \qquad \text{for every } \varphi \in C_c^1(\mathbb{R}^d), \quad (6)$$

where ρ_t is the distribution of X_t and $b_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x]$ is the average velocity of paths through x.

The endpoint conditions in (1) give $X_0 \sim \rho_0$ and $X_1 \sim \rho_1$. The rate at which the structural value of a path changes is $D\mathbf{c}(X_t)\dot{X}_t$, and the schedule prescribes the rate β_t at the current structural value. Condition (3) ensures that we can transport the structural value according to the base schedule and X_t according to the chosen interpolant, but still the difference between the two in the base direction does not change the distribution. The simplest satisfaction is to require $\mathbf{c}(X_t) = B_t(\mathbf{c}(X_0))$ for every t almost surely, since the difference then vanishes on every path. The weaker form is sufficient for symmetry. Let a compact Lie group G act on $\mathbb{R}^d$, with $\mathcal{H}_x$ along its orbits (section 2.1) and $B_t = \mathrm{id}$. If the endpoints and the noise are invariant under G, and the coupling and interpolant commute with G, then ρ_t stays invariant and the base part of the motion only moves mass along orbits, leaving ρ_t unchanged. For example, the independent coupling with $X_t = (1-t)X_0 + tX_1 + \gamma_t Z$ qualifies whenever G acts linearly and preserves the Gaussian source.

Measure-theoretically, the distribution ρ_t of X_t disintegrates with respect to $\mathbf{c}$ (Kallenberg, 2002),

$$\rho_t = \int_{\mathcal{Y}} \mu_t^y\, (\mathbf{c}_\#\rho_t)(dy), \qquad \mu_t^y(\mathcal{M}_y) = 1 \quad \text{for } \mathbf{c}_\#\rho_t\text{-almost every } y, \quad (7)$$

where $\mathbf{c}_\#\rho_t$ is its base part and μ_t^y is the conditional distribution of X_t given $\mathbf{c}(X_t) = y$. A disintegrated interpolant replaces the single transport of ρ_0 to ρ_1 by two transports acting on the output space and fibres of $\mathbf{c}$. On the base, the schedule transports ν_0 to ν_1 through $\nu_t = (B_t)_\#\nu_0$, which is known in advance. On the fibres, μ_t^y evolves from those of the source to those of the target: transport between fibres is the prescribed base velocity, lifted by H, while the transport within fibres is the vertical part of $\dot{X}_t$. Condition (3) ensures that the two transports are consistent. Figure 3 visualises this transport for singularity and symmetry.

Transport equations and velocity regression. Separability lets a known field replace the horizontal part of the path velocity. For $t < 1$ and $x \in \mathbb{R}^d$, define

$$h_t(x) = \mathsf{H}_x\beta_t(\mathbf{c}(x)), \qquad u_t(x) = \mathbb{E}\big[\mathsf{P}_{X_t}\dot{X}_t \mid X_t = x\big]. \quad (8)$$

The first moves samples between fibres along the schedule and is known in closed form. The second satisfies $D\mathbf{c}(x)u_t(x) = 0$ and is therefore tangent to the fibres. Identities involving conditional velocities hold up to $dt\,\rho_t(dx)$-null sets.

Let $\mathcal{N}$ be a Riemannian manifold with gradient $\nabla_{\mathcal{N}}$ and divergence $\mathrm{div}_{\mathcal{N}}$. A curve of probability measures m_t on $\mathcal{N}$ solves $\partial_t m_t + \mathrm{div}_{\mathcal{N}}(v_t m_t) = 0$ in the *weak sense* on $(0,1)$ if

$$\frac{d}{dt}\int \varphi\, dm_t = \int \langle \nabla_{\mathcal{N}}\varphi, v_t\rangle\, dm_t \qquad \text{for every } \varphi \in C_c^1(\mathcal{N}), \tag{9}$$

with the derivative taken in the sense of distributions on $(0,1)$. We use this with $\mathcal{N} = \mathbb{R}^d$ and with $\mathcal{N} = \mathcal{Y}$. A curve on moving fibres is carried both by the fibre's motion and by transport within it, so its equation is tested with ambient functions. A curve m_t concentrated on $\mathcal{M}_{y_t}$ solves $\partial_t m_t + \nabla\cdot(h_t m_t) + \mathrm{div}_{\mathcal{M}_{y_t}}(u_t m_t) = 0$ in the weak sense if

$$\frac{d}{dt}\int \varphi\, dm_t = \int \left[\nabla\varphi\cdot h_t + \nabla_{\mathcal{M}_{y_t}}\varphi\cdot u_t\right] dm_t \qquad \text{for every } \varphi \in C_c^1(\mathbb{R}^d). \tag{10}$$

Theorem 2.3 (Disintegrated continuity equation)**.** *Let $(X_t)_{t\in[0,1]}$ be a disintegrated stochastic interpolant satisfying assumption 2.1, with h_t and u_t as in eq. 8. For $t < 1$, let μ_t^z be the conditional distribution of X_t given $\mathbf{c}(X_t) = z$. The following hold in the weak sense on $(0,1)$.*

1. ***Between fibres.*** *The base laws $\mathbf{c}_{\#}\rho_t$ coincide with $\nu_t = (B_t)_{\#}\nu_0$ and solve*
$$\partial_t \nu_t + \mathrm{div}_{\mathcal{Y}}\big(\beta_t \nu_t\big) = 0. \tag{11}$$
2. ***Within fibres.*** *For ν_0-almost every y, the curve $t \mapsto \mu_t^{B_t(y)}$ is concentrated on the moving fibre $\mathcal{M}_{B_t(y)}$ and solves*
$$\partial_t \mu_t^{B_t(y)} + \nabla\cdot\big(h_t \mu_t^{B_t(y)}\big) + \mathrm{div}_{\mathcal{M}_{B_t(y)}}\big(u_t \mu_t^{B_t(y)}\big) = 0. \tag{12}$$
3. ***Recombination.*** *Mixing the fibre curves over the base laws gives the interpolant, $\rho_t = \int_{\mathcal{Y}} \mu_t^z\, \nu_t(dz)$, and it solves*
$$\partial_t \rho_t + \nabla\cdot\big((h_t + u_t)\rho_t\big) = 0. \tag{13}$$

The horizontal field h_t is known in closed form, so only u_t has to be estimated from data, and combining u_t with h_t recovers a transport from ρ_0 to ρ_1. Separability is what makes this possible: it lets the known field h_t stand in for the horizontal motion, so the base equation is solved in closed form, and only the fibre equation involves the data.

Theorem 2.4 (Vertical velocity regression)**.** *Let $\mathcal{U}$ be the set of measurable fields $\hat{u}$ with $D\mathbf{c}(x)\hat{u}(t,x) = 0$ for all (t,x) and $\int_0^1 \mathbb{E}\|\hat{u}(t,X_t)\|^2 dt < \infty$. The objective*

$$\mathcal{L}_u[\hat{u}] = \int_0^1 \mathbb{E}\Big[\|\hat{u}(t,X_t)\|^2 - 2\,\mathsf{P}_{X_t}\dot{X}_t\cdot\hat{u}(t,X_t)\Big]dt \tag{14}$$

is uniquely minimised in $\mathcal{U}$, up to $dt\,\rho_t(dx)$-null sets, by

$$u_t(x) = \mathbb{E}\Big[\mathsf{P}_{X_t}\dot{X}_t \mid X_t = x\Big]. \tag{15}$$

Fibrewise scores. Diffusion also needs the score of ρ_t along the fibres. However, the within-fibre measures need not have densities, so we define a generalised score following Bogachev (1997, Chapter 6) (see appendix A), and on a fibre the natural candidate uses the manifold divergence to define the score. For $t < 1$, let μ_t^z be the conditional distribution of X_t given $\mathbf{c}(X_t) = z$, as in theorem 2.3. A field s_t tangent to the fibres is the *fibre score* of ρ_t if

$$\int \mathrm{div}_{\mathcal{M}_z}\, w\, d\mu_t^z = -\int s_t\cdot w\, d\mu_t^z \tag{16}$$

for $\mathbf{c}_{\#}\rho_t$-almost every z and every compactly supported C^1 vector field w tangent to $\mathcal{M}_z$.

The interpolant determines how the auxiliary randomness Z perturbs the sample. At a fixed time $t \in (0,1)$, a *noise target* is an integrable random vector ξ_t, tangent to the fibre through X_t and computed from (t, X_0, X_1, Z), such that

$$\mathbb{E}\big[\xi_t \cdot w(X_t) \mid X_0, X_1\big] = -\,\mathbb{E}\big[\operatorname{div}_{\mathcal{M}} w(X_t) \mid X_0, X_1\big] \tag{17}$$

for every compactly supported C^1 field w tangent to the fibres. For additive Gaussian noise $\gamma_t Z$ within a flat fibre, with Z standard Gaussian in its tangent space, Stein's lemma gives $\xi_t = -Z/\gamma_t$, recovering the ordinary stochastic-interpolant target (Albergo et al., 2025). Heat-kernel perturbations on a fibre satisfy the same identity with the spatial logarithmic gradient of the heat kernel as their target.

Theorem 2.5 (Vertical score regression)**.** *Let ξ_t be a noise target satisfying eq. 17 for almost every $t \in (0,1)$. Let λ be continuous and positive on $(0,1)$, with $\int_0^1 \lambda(t)\,\mathbb{E}\|\xi_t\|^2\,dt < \infty$. Among measurable tangent fields $\hat{s}$ with $\int_0^1 \lambda(t)\,\mathbb{E}\|\hat{s}(t,X_t)\|^2\,dt < \infty$, the objective*

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \lambda(t)\,\mathbb{E}\|\hat{s}(t,X_t) - \xi_t\|^2\,dt \tag{18}$$

is uniquely minimized, up to $dt\,\rho_t(dx)$-null sets, by

$$s_t(x) = \mathbb{E}[\xi_t \mid X_t = x], \tag{19}$$

which is the fibre score of ρ_t defined by eq. 16.

Sampling. Consider the Stratonovich sampler

$$d\widehat{X}_t = \big[h_t + \hat{u}_t + \varepsilon_t \hat{s}_t\big](\widehat{X}_t)\,dt + \sqrt{2\varepsilon_t}\,\mathsf{P}_{\widehat{X}_t} \circ dW_t, \qquad \widehat{X}_0 \sim \rho_0, \tag{20}$$

where $\hat{u}_t$ and $\hat{s}_t$ are tangent to the fibres, $\varepsilon_t \geq 0$ is a smooth deterministic schedule, and W_t is a standard Brownian motion independent of $\widehat{X}_0$. Taking $\varepsilon \equiv 0$ gives the ODE sampler. When $\varepsilon \not\equiv 0$, we use the orthogonal connection. Write $\Delta_{\mathcal{M}}\varphi(x)$ for the Laplacian of φ restricted to the fibre through x. The weak Fokker–Planck equation for a curve of probability measures m_t is

$$\frac{d}{dt}\int \varphi\,dm_t = \int \Big[(h_t + u_t + \varepsilon_t s_t)\cdot\nabla\varphi + \varepsilon_t \Delta_{\mathcal{M}}\varphi\Big]\,dm_t \qquad \text{for every } \varphi \in C_c^2(\mathbb{R}^d). \tag{21}$$

Theorem 2.6 (Sampling)**.** *Suppose eq. 20 has a continuous solution on $[0,1)$. Then, almost surely,*

$$\mathbf{c}(\widehat{X}_t) = B_t(\mathbf{c}(\widehat{X}_0)) \qquad \textit{for every } t < 1.$$

If $\hat{u} = u$ and $\hat{s} = s$, assume that the transport and score drifts are integrable against both the sampler and interpolant laws on every interval $[0,T]$, $T < 1$, and that the weak Fokker–Planck equation (21) has a unique narrowly continuous probability solution in this integrability class. Then $\widehat{X}_t \sim \rho_t$ for every $t < 1$, and its law converges narrowly to ρ_1 as $t \to 1$.

The prescribed base evolution therefore holds for any learned tangent fields. Exact fields additionally recover the interpolant marginals.

3 Methods

The central question is whether we can derive practical training algorithms from the theory presented above. We present practical variants of flow matching and diffusion based on theorems 2.4 and 2.5. Separability in definition 2.2 is a condition on averages and is hard to check for a given interpolant. We therefore impose the stricter pathwise form, $\mathbf{c}(X_t) = B_t(\mathbf{c}(X_0))$ for every t almost surely, which implies separability and can be enforced by construction. Algorithms 1 and 2 both require a coupling that satisfies $\mathbf{c}(X_1) = B_1(\mathbf{c}(X_0))$ almost surely.

Trivializations. The theoretical construction so far has been coordinate-free. When a global trivialization is known, the training objectives from section 2 simplify. A *trivialization* is a smooth invertible map $\Psi(y, \mathbf{w}) = \mathbf{x}$, with smooth inverse and $\mathbf{c}(\Psi(y,\mathbf{w})) = y$, so that $y \in \mathcal{Y}$ identifies a fibre and $\mathbf{w}$ identifies a position inside it. Differentiating this identity gives $D\mathbf{c}\,\partial_{\mathbf{w}}\Psi = 0$ and $D\mathbf{c}\,\partial_y\Psi = I$. The columns of $\partial_{\mathbf{w}}\Psi$ therefore span the vertical space. Taking $\mathcal{H}_x$ to be spanned

Algorithm 1 Training, explicit. FM: flow matching; diff.: diffusion; $\hat{s}_\theta$ is the coordinate score at $(t, y_t, \mathbf{w}_t)$.

repeat
  $(X_0, X_1) \sim \pi$ (compatible), $t \sim \mathcal{U}[0,1]$
  $\mathbf{z} \sim \mathcal{N}(0, I_{d-k})$
  $(y_i, \mathbf{w}_i) \leftarrow \Psi^{-1}(X_i),\ i = 0, 1$
  $y_t \leftarrow B_t(y_0)$
  $\mathbf{w}_t \leftarrow I(t, \mathbf{w}_0, \mathbf{w}_1) + \gamma_t \mathbf{z}$
  $\ell_v \leftarrow \left\| \hat{v}_\theta(t, y_t, \mathbf{w}_t) - \partial_t I - \dot{\gamma}_t \mathbf{z} \right\|^2$
  $\ell \leftarrow \begin{cases} \ell_v & \text{(FM)} \\ \ell_v + \lambda(t) \|\gamma_t \hat{s}_\theta + \mathbf{z}\|^2 & \text{(diff.)} \end{cases}$
  gradient step on ℓ
until converged

Algorithm 2 Training, implicit. FM: flow matching; diff.: approximate diffusion; $\hat{s}_\theta = \hat{s}_\theta(t, X_t)$.

repeat
  $(X_0, X_1) \sim \pi$ (compatible), $t \sim \mathcal{U}[0,1]$
  $\mathbf{z} \sim \mathcal{N}(0, I_d)$
  $y_t \leftarrow B_t(\mathbf{c}(X_0))$
  $C_t \leftarrow \mathsf{R}_{y_t}(I(t, X_0, X_1))$
  $X_t \leftarrow \mathsf{R}_{y_t}(C_t + \gamma_t \mathsf{P}_{C_t} \mathbf{z})$
  $\ell_u \leftarrow \left\| \mathsf{P}_{X_t} \hat{u}_\theta(t, X_t) - \mathsf{P}_{X_t} \dot{X}_t \right\|^2$
  $\ell \leftarrow \begin{cases} \ell_u & \text{(FM)} \\ \ell_u + \lambda(t) \|\mathsf{P}_{X_t}(\gamma_t \hat{s}_\theta + \mathbf{z})\|^2 & \text{(diff.)} \end{cases}$
  gradient step on ℓ
until converged

by the columns of $\partial_y \Psi$, the vertical projector keeps the fibre component of an ambient velocity, while the horizontal lift maps a base velocity to $\partial_y \Psi\, \beta_t$. The splitting (3) is then a splitting in these coordinates. Polar coordinates on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ give both running examples. For the unit-circle target, $\Psi(y, \theta) = (1+y)(\cos\theta, \sin\theta)$, with circles as fibres and $\theta \in S^1$ as the fibre coordinate. For rotational symmetry, $\Psi(\theta, r) = r(\cos\theta, \sin\theta)$, with rays as fibres and $r > 0$ as the fibre coordinate.

Explicit disintegration. With a trivialization, both endpoints are mapped to coordinates, $(y_i, \mathbf{w}_i) = \Psi^{-1}(X_i)$. The structural value follows the schedule, $y_t = B_t(y_0)$, and the fibre coordinate follows an ordinary stochastic interpolant $\mathbf{w}_t = I(t, \mathbf{w}_0, \mathbf{w}_1) + \gamma_t \mathbf{z}$, so separability holds pathwise by construction. This Gaussian construction assumes $\mathbf{w} \in \mathbb{R}^{d-k}$; periodic or restricted coordinates require adapted interpolants and noise laws. The coupling condition $y_1 = B_1(y_0)$ is automatic for a zero-set target with $B_1 \equiv 0$; for symmetry with $B_t = \mathrm{id}$, it pairs endpoints on the same fibre. Training is ordinary flow matching in $\mathbf{w}$, with y_t as an extra input, or diffusion with an additional coordinate score head (algorithm 1). This score is defined with respect to $d\mathbf{w}$. Sampling evaluates y_t in closed form, evolves $\mathbf{w}_t$ with the learned ODE or SDE, and returns $\Psi(y_1, \mathbf{w}_1)$. The decoded SDE generally differs from the intrinsic isotropic sampler in eq. 20.

Implicit disintegration. Without a trivialization, we work with the structure map $\mathbf{c}$ directly. Two operations are needed: projecting a velocity onto the vertical space, and moving a point onto a prescribed fibre. The first is the orthogonal projector:

$$\mathsf{P}_{\mathbf{x}} = I_d - D\mathbf{c}(\mathbf{x})^\top \left(D\mathbf{c}(\mathbf{x}) D\mathbf{c}(\mathbf{x})^\top \right)^{-1} D\mathbf{c}(\mathbf{x}), \tag{22}$$

and the second is the retraction onto the fibre $\mathcal{M}_y$, defined here by the nearest-point projection,

$$\mathsf{R}_y(\mathbf{x}) = \operatorname*{arg\,min}_{\mathbf{x}' \in \mathcal{M}_y} \|\mathbf{x}' - \mathbf{x}\|. \tag{23}$$

We assume this projection is single-valued and smooth where evaluated. A numerical implementation and differentiation are described in appendix C. Since an ordinary linear interpolant does not follow the prescribed schedule, we enforce it by retracting onto the scheduled fibre, adding tangent noise, and re-retracting:

$$C_t = \mathsf{R}_{y_t}(I_t), \qquad X_t = \mathsf{R}_{y_t}\left(C_t + \gamma_t \mathsf{P}_{C_t} \mathbf{z}\right), \tag{24}$$

with $I_t = I(t, X_0, X_1)$, $y_t = B_t(\mathbf{c}(X_0))$ and $\mathbf{z} \sim \mathcal{N}(0, I_d)$. This gives $\mathbf{c}(X_t) = y_t$ by construction. Flow matching regresses $\mathsf{P}_{X_t} \dot{X}_t$, with $\dot{X}_t$ obtained by differentiating through the full construction. For diffusion, an additional score head uses the projected-noise target in algorithm 2. We use this as a small-noise approximation; the retraction generally prevents the target from satisfying eq. 17 exactly. Sampling alternates a step of the learned vertical dynamics with the known base update and a retraction onto the new fibre. The structural value follows the schedule up to the retraction solver's tolerance.

4 Experiments

We evaluate our implicit methods for singular targets on PDE trajectory generation, QM9 molecular generation, and a synthetic Stiefel benchmark. We omit symmetry experiments here because the

		$\mathrm{SW}_2 \downarrow$	Constraint error $\downarrow$
Flow matching	Ambient	0.030 ± 0.003	2.0×10^{-1}
	Disintegrated	$\mathbf{0.017 \pm 0.001}$	$\mathbf{1.5 \times 10^{-7}}$
Diffusion	Ambient	0.041 ± 0.002	2.3×10^{-1}
	Disintegrated	$\mathbf{0.016 \pm 0.001}$	$\mathbf{1.4 \times 10^{-7}}$

Table 1: Generation on $\mathrm{St}(10, 4)$. Mean and standard deviation over three seeds. Lower is better.

corresponding diffusion construction exactly recovers quotient-space diffusion (Xu et al., 2026). We provide full experimental details in appendix F.

Summary of results. Disintegrated flow matching (DI-FM) is competitive with specialised methods across a diverse range of tasks with structured data. The clearest result is the comparison with conditional functional flow matching (CFFM; Kerrigan et al., 2024): DI-FM achieves lower mean-field error with approximately one quarter of the parameters, at the cost of somewhat higher standard-deviation error. On QM9, DI-FM achieves moment error comparable to the specialised baseline Stiefel flow matching (Stiefel-FM; Cheng et al., 2025) and greater diversity. In all experiments, we see significant improvement across the board over ordinary flow matching and diffusion.

PDE trajectory generation. We generate PDE trajectories conditioned on initial or boundary conditions, with constraints on the observations and conservation laws (details in section F. We replicate the experimental setting from Utkarsh et al. (2025). Each sample is a complete space-time solution field, and we assess both the statistics of the generated ensemble and constraint satisfaction. We compare the proposed disintegrated interpolants (DI) with unconditional functional flow matching (FFM) Kerrigan et al. (2024), PCFM Utkarsh et al. (2025) applied to the same unconditional models, and conditional FFM (CFFM) Kerrigan et al. (2024) at two network sizes.

QM9 molecular generation. We generate QM9 molecular geometries conditioned on atom identities and mass-weighted moments. We use a smaller, subsampled version of the QM9 reconstruction benchmark, with ten candidates per molecule on 256 test molecules. We compare DI-FM with ambient flow matching and Stiefel-FM Cheng et al. (2025), using the same network architecture.

Stiefel manifold. We evaluate implicit disintegration on $\mathrm{St}(10, 4)$ with a non-uniform target and an ambient Gaussian source. Disintegration improves distributional fit for both flow matching and diffusion and reduces constraint error from 10^{-1} to approximately 10^{-7}.

	Burgers			Reaction–diffusion		
Method	MMSE/10^{-3} $\downarrow$	SMSE/10^{-3} $\downarrow$	CE $\downarrow$	MMSE/10^{-3} $\downarrow$	SMSE/10^{-3} $\downarrow$	CE $\downarrow$
FFM	23.70 ± 1.79	20.35 ± 2.46	4.67e-1 ± 2.40e-1	56.07 ± 2.56	27.10 ± 4.43	2.80e-1 ± 4.14e-2
PCFM†	45.60 ± 33.83	81.89 ± 18.77	6.57e-6 ± 9.44e-7	63.87	1091.63	1.76e-5
CFFM (small)	5.178 ± 2.402	13.10 ± 4.88	4.62e-7 ± 1.01e-8	1.305 ± 0.649	1.086 ± 0.388	**9.66e-7** ± 1.78e-8
CFFM (large)	4.094 ± 0.961	**10.49** ± 2.27	**4.45e-7** ± 5.75e-8	0.618 ± 0.344	**0.528** ± 0.207	9.72e-7 ± 2.41e-8
DI (small)	**2.665** ± 0.633	12.79 ± 3.46	4.83e-7 ± 1.70e-8	**0.570** ± 0.307	0.593 ± 0.082	9.69e-7 ± 1.80e-8

Table 2: PDE generation with validation-selected checkpoints and 32 sampling steps. MMSE/SMSE measure errors in ensemble means/standard deviations; CE measures conservation error. Results are mean ± sample SD over three seeds, except PCFM†, which reports Burgers seeds 43/44 and reaction–diffusion seed 42, omitting higher-error runs. Full per-seed results in appendix F.1.1.

Method	CE $\downarrow$	Valid (%) $\uparrow$	Stable $\downarrow$	Diverse $\uparrow$	NFE
Ambient FM	9.53 ± 0.74	49.05 ± 1.86	-0.317 ± 0.024	1.4519 ± 0.0183	200
DI-FM	$(5.19 \pm 0.38) \times 10^{-5}$	56.82 ± 1.42	-0.521 ± 0.031	1.4882 ± 0.0147	200
Stiefel-FM	$(5.65 \pm 0.43) \times 10^{-5}$	$\mathbf{64.66} \pm 1.67$	-0.594 ± 0.027	1.4763 ± 0.0162	200

Table 3: QM9 molecular reconstruction over three training seeds. Metrics are defined in appendix F.2.

5 Related work

Structured generative modelling. The closest related work is quotient-space diffusion (QSD; Xu et al. (2026)). Our theoretical framework is more general and, under the conditions detailed in appendix E, recovers exactly the QSD training objectives. Other existing methods are instances of the explicit disintegrated interpolants where a trivialization is known. Torsional Diffusion takes bond lengths and angles and holds them fixed during generation, learning only the distribution of rotations around bonds conditional on this geometry (Jing et al., 2022). AssembleFlow makes a similar separation for molecules: internal shapes remain fixed, while the model learns relative positions and orientations (Guo et al., 2025).

Singular targets. Recent theoretical studies of learning low-dimensional target distributions establish adaptation rates governed primarily by intrinsic dimension for diffusion models (Tang & Yang, 2024) and flow matching (Kumar et al., 2026). Flows starting from an ambient density collapse onto the target only in the endpoint limit, through a loss of regularity of the velocity field. Learning near this limit is increasingly ill-conditioned, as documented for diffusion (Liu et al., 2025) and flow-based models (Zeng & Yan, 2025). We address this difficulty by prescribing the transport towards the constraint set analytically, leaving the network to learn only motion within the fibres.

Conditional generation. Pretrained unconditional models can be conditioned during sampling: DDRM treats linear inverse problems (Kawar et al., 2022), while diffusion posterior sampling provides approximate conditioning for noisy nonlinear observations (Chung et al., 2023). Several methods apply constraint corrections during sampling from a pretrained flow model (Utkarsh et al., 2025; Christopher et al., 2024). Conditional flow matching and diffusion can also be trained on paired data and observations, with the condition supplied to the network (Zheng et al., 2023; Shysheya et al., 2024). Our approach goes further by exploiting the relation between observation and data to decompose the transport. We prescribe the evolution of the observation and learn only motion within its level sets, thus simplifying the learning problem.

Intrinsic methods enforce geometric properties. Riemannian continuous normalising flows, flow matching and diffusion models all place the source and target on the same manifold and evolve samples therein (Mathieu & Nickel, 2020; Chen & Lipman, 2024; Huang et al., 2022). They need a source distribution that can be sampled on that manifold, alongside geodesics or exponential maps. Most benchmarks are restricted to spheres and rotation groups where these operations are analytically tractable and computationally efficient. Our construction needs neither and is more widely applicable.

Stochastic Interpolants construct transport dynamics from paths between endpoint samples, with velocity and score fields obtained by regression (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2025). Data-dependent couplings extend the interpolant framework to conditional generation (Albergo et al., 2024b). Further extensions have been proposed (Albergo et al., 2024a; Yu et al., 2026).

6 Conclusion

We introduced disintegrated stochastic interpolants, a generalisation of stochastic interpolants that exploits known structure in data to simplify the learning problem. The framework accommodates singular targets and recovers the training objectives of quotient-space diffusion (Xu et al., 2026) under the stated symmetry assumptions. Several existing methods also fit within its formulation through explicit coordinates (Guo et al., 2025; Jing et al., 2022) that separate prescribed structure from learned motion. We also propose novel generative modelling algorithms that emerge from this theory. In our experience, the proposed methods simplify the learning problem compared to not incorporating structure. This desired property is often not found in structural machine learning methods.

The main limitations are that the *implicit* method introduces computational overhead through projections and retractions, which can offset the benefits of a simpler learning problem. Our proposed algorithms are naive implementations of the developed theory; we expect more elaborate methods will close the performance gap. For the future, our disintegration lets us study how known structure affects generalisation and memorisation, and how it can guide representation learning. The algorithms explored here cover only a small part of the possibilities permitted by the theory. Future work will investigate further constructions, consider structures beyond singularity and symmetry, and apply the framework to larger-scale experiments.

AI USE STATEMENT

In this work, we used generative AI tools primarily for language editing and as a coding assistant for implementing baselines and converting PyTorch code to JAX. We have not used generative AI tools for other tasks with required disclosure such as developing the theory, implelemnting the core method, and the rest of the required disclosure tasks do not apply to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Michael Albergo, Nicholas Boffi, Michael Lindsey, and Eric Vanden-Eijnden. Multimarginal generative modeling with stochastic interpolants. In *International Conference on Learning Representations*, volume 2024, pp. 55884–55901, 2024a.

Michael 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.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Michael Samuel Albergo, Mark Goldstein, Nicholas Matthew Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 921–937. PMLR, 2024b. URL https://proceedings.mlr.press/v235/albergo24a.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2nd edition, 2008. ISBN 978-3-7643-8722-8.

VI Bogachev. Differentiable measures and the malliavin calculus. *Journal of Mathematical Sciences*, 87(4):3577–3731, 1997.

Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Soc., 2015.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

Austin H Cheng, Alston Lo, Kin Long Kelvin Lee, Santiago Miret, and Alán Aspuru-Guzik. Stiefel flow matching for moment-constrained structure elucidation. In *International Conference on Learning Representations*, volume 2025, pp. 21025–21055, 2025.

Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

David Chin-Lung Fong and Michael Saunders. Lsmr: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

Hongyu Guo, Yoshua Bengio, and Shengchao Liu. Assembleflow: Rigid flow matching with inertial frames for molecular assembly. In *International Conference on Learning Representations*, volume 2025, pp. 755–787, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

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

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in neural information processing systems*, 35:24240–24253, 2022.

Olav Kallenberg. *Foundations of Modern Probability*. Springer-Verlag, New York, second edition, 2002. ISBN 978-0387953137.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11793.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3934–3942. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kerrigan24a.html.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures. *arXiv preprint arXiv:2602.22486*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Zichen Liu, Wei Zhang, and Tiejun Li. Improving the euclidean diffusion generation of manifold data by mitigating score function singularity. *Advances in Neural Information Processing Systems*, 38:110885–110917, 2025.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in neural information processing systems*, 33:2503–2515, 2020.

Hrittik Roy, Søren Hauberg, and Nicholas Krämer. Matrix-free least squares solvers: Values, gradients, and what to do with them. *arXiv preprint arXiv:2510.19634*, 2025.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José M Hernández-Lobato, Richard E Turner, and Emile Mathieu. On conditional diffusion models for pde simulations. *Advances in Neural Information Processing Systems*, 37:23246–23300, 2024.

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://openreview.net/forum?id=PxTIG12RRHS.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International conference on artificial intelligence and statistics*, pp. 1648–1656. PMLR, 2024.

Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gómez-Bombarelli, and Christopher Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *Advances in Neural Information Processing Systems*, 2025.

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

Yixian Xu, Yusong Wang, Shengjie Luo, Kaiyuan Gao, Tianyu He, Di He, and Chang Liu. Quotient-space diffusion models. In *International Conference on Learning Representations*, 2026.

James Boran Yu, RuiKang OuYang, Julien Horwood, and José Miguel Hernández-Lobato. Stochastic interpolants in hilbert spaces. *arXiv preprint arXiv:2602.01988*, 2026.

Weili Zeng and Yichao Yan. Flow matching in the low-noise regime: Pathologies and a contrastive remedy. *arXiv preprint arXiv:2509.20952*, 2025.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

A Stochastic Interpolants in Measure-Theoretic Language

A.1 Outline of Albergo et al. (2025)

We begin by outlining the argument of Albergo et al. (2025). Their stated goal is to provide a versatile way to build generative models that unifies and extends many existing algorithms. Suppose we have a source ρ_0 and a target ρ_1 on $\mathbb{R}^d$ with densities p_0 and p_1 with respect to Lebesgue measure, a coupling π of the two, and the interpolant $X_t = I(t, X_0, X_1) + \gamma_t Z$. Four results follow from this construction, and together they account for the flow-matching and diffusion algorithms in common use.

Transport. Suppose ρ_t is the distribution of X_t and p_t is its density. The velocity that moves p_t is a conditional average of the path velocity.

$$b_t(x) = \mathbb{E}\big[\partial_t I(t, X_0, X_1) + \dot{\gamma}_t Z \,\big|\, X_t = x\big]. \tag{25}$$

This field averages the velocities of all interpolant paths passing through x at time t. Albergo et al. (2025) show that p_t solves the transport equation

$$\partial_t p_t + \nabla\cdot(b_t p_t) = 0, \qquad p_{t=0} = p_0, \quad p_{t=1} = p_1. \tag{26}$$

The crucial step for us here is to redefine what the transport equation means without reference to densities. We define it in a weak sense and say the transport equation holds if, for any test function, it holds when integrated with respect to ρ_t the equation holds.

Velocity regression. The conditional expectation in eq. 25 typically has no closed form. Albergo et al. (2025) show that it is the unique minimiser of a quadratic objective that requires only samples of the interpolant,

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\Big[\|\hat{b}(t, X_t)\|^2 - 2\big(\partial_t I(t, X_0, X_1) + \dot{\gamma}_t Z\big)\cdot \hat{b}(t, X_t)\Big]\, dt. \tag{27}$$

Minimising this objective over a neural network by stochastic gradient descent gives a training algorithm. When $\gamma \equiv 0$ and $I(t, x_0, x_1) = (1-t)x_0 + t x_1$, the regression target is $X_1 - X_0$, and generating samples of ρ_1 by integrating $\dot{\hat{X}}_t = b_t(\hat{X}_t)$ from $\hat{X}_0 \sim \rho_0$ recovers the conditional flow matching objective (Lipman et al., 2023). The proof of this is pretty direct; it requires manipulating integrals and using the tower law of conditional expectations, and most of the argument can be easily generalised to the measure-theoretic setting.

Score. Diffusion models also require the score $s_t = \nabla \log p_t$. For $\gamma_t > 0$ the Gaussian term makes p_t positive and smooth, and Gaussian integration by parts expresses the score through the noise,

$$s_t(x) = -\gamma_t^{-1}\, \mathbb{E}[Z \mid X_t = x]. \tag{28}$$

The score is therefore learned by regression as well. The denoiser $\eta_t(x) = \mathbb{E}[Z \mid X_t = x]$ is the unique minimizer of

$$\mathcal{L}_\eta[\hat{\eta}] = \int_0^1 \mathbb{E}\|\hat{\eta}(t, X_t) - Z\|^2\, dt, \tag{29}$$

and the score is recovered as $s_t = -\eta_t/\gamma_t$. This is proved using Stein's identity and applying the tower law of conditional expectations. If we can define the score without reference to density, the remaining machinery can be reused.

Sampling. The score turns the transport equation into a family of Fokker–Planck equations. Since $\nabla p_t = s_t p_t$, adding and subtracting $\varepsilon_t \Delta p_t$ gives, for every $\varepsilon_t \geq 0$,

$$\partial_t p_t + \nabla\cdot\big((b_t + \varepsilon_t s_t)p_t\big) = \varepsilon_t \Delta p_t. \tag{30}$$

Albergo et al. (2025) conclude that the probability flow ODE $\dot{\hat{X}}_t = b_t(\hat{X}_t)$ and the SDE

$$d\hat{X}_t = (b_t + \varepsilon_t s_t)(\hat{X}_t)\, dt + \sqrt{2\varepsilon_t}\, dW_t, \tag{31}$$

both started from $\hat{X}_0 \sim \rho_0$, have marginals ρ_t for every t.

Generative models. The results above unify different generative models depending on the choice of interpolant. Flow matching (Lipman et al., 2023) takes $\gamma = 0$ and $I(t, x_0, x_1) = (1-t)x_0 + tx_1$. The regression target in eq. 27 is then $X_1 - X_0$. A single network learns b_t, and samples are generated by integrating the ODE.

With $\gamma > 0$, learning the denoiser as well as b_t gives the score, and samples can then be drawn from the SDE for any choice of noising schedule ε_t. The ODE and the SDE share the marginals ρ_t. They differ in the noise injected along the way, and the term $\varepsilon_t s_t$ in the drift compensates for it, so $\varepsilon_t = 0$ recovers the ODE. For a general interpolant, b_t and s_t can both be learned with different networks and samples can be generated with either the SDE or ODE.

Score-based diffusion avoids the second network using a specific choice for the interpolant. Let $Z \sim \mathcal{N}(0, I)$ be independent of X_1, and consider the one-sided interpolant

$$X_t = \alpha_t Z + \beta_t X_1, \qquad \alpha_0 = \beta_1 = 1, \quad \alpha_1 = \beta_0 = 0, \tag{32}$$

so that $X_0 = Z$ and $\rho_0 = \mathcal{N}(0, I)$. Applying the score matching result gives $s_t(x) = -\alpha_t^{-1}\,\mathbb{E}[Z \mid X_t = x]$. The denoiser minimises

$$\int_0^1 \mathbb{E}\big\|\hat{\eta}(t, \alpha_t Z + \beta_t X_1) - Z\big\|^2 dt, \tag{33}$$

which is the noise-prediction objective of denoising diffusion (Ho et al., 2020). The velocity is $b_t(x) = \dot{\alpha}_t\,\mathbb{E}[Z \mid X_t = x] + \dot{\beta}_t\,\mathbb{E}[X_1 \mid X_t = x]$. Since $x = \alpha_t \mathbb{E}[Z \mid X_t = x] + \beta_t \mathbb{E}[X_1 \mid X_t = x]$, the velocity is a known function of x and the score. For variance-preserving schedules, $\alpha_t^2 + \beta_t^2 = 1$, this reduces to

$$b_t(x) = \frac{\dot{\beta}_t}{\beta_t}\big(x + s_t(x)\big), \qquad 0 < t < 1. \tag{34}$$

One network therefore determines both samplers. With $\varepsilon_t = \dot{\beta}_t/\beta_t$ the SDE becomes

$$d\hat{X}_t = \frac{\dot{\beta}_t}{\beta_t}\big(\hat{X}_t + 2s_t(\hat{X}_t)\big)dt + \sqrt{\frac{2\dot{\beta}_t}{\beta_t}}\,dW_t, \tag{35}$$

The reverse-time SDE of variance-preserving diffusion (Song et al., 2021) is written in generative time, and $\varepsilon \equiv 0$ gives its probability flow ODE.

We rewrite the crucial results in measure-theoretic language since density-based formulations assume an ambient Lebesgue measure. Since we define transports on fibres which may be non-Euclidean, we need to rewrite these results without relying on densities. Most of the results carry over with just a slight change in notation from the paper. The important changes are that we change the assumptions in the definition of stochastic interpolant to be somewhat weaker, and we define a score without reference to the density function.

A.2 Stochastic interpolants between measures

Let ρ_0 and ρ_1 be probability measures on $\mathbb{R}^d$.

Definition A.1 (Stochastic interpolant)**.** A stochastic interpolant between ρ_0 and ρ_1 is a stochastic process $(X_t)_{t\in[0,1]}$ in $\mathbb{R}^d$ with $X_0 \sim \rho_0$ and $X_1 \sim \rho_1$, whose paths are almost surely absolutely continuous with derivative $\dot{X}_t$, and which satisfies

$$\mathbb{E}\int_0^1 \|\dot{X}_t\|\,dt < \infty. \tag{36}$$

The law of X_t is written ρ_t, and $b_t(x) = \mathbb{E}[\dot{X}_t \mid X_t = x]$ denotes a version of this conditional expectation that is jointly measurable in (t, x), which exists because $\dot{X}_t$ is integrable for almost every t by eq. 36. When the conditional expectation in eq. 25 can be computed from densities, it agrees with b_t for ρ_t-almost every x.

Theorem A.2 (Transport)**.** *For every $\varphi \in C_c^1([0,1] \times \mathbb{R}^d)$,*

$$\int \varphi(1,\cdot)\,d\rho_1 - \int \varphi(0,\cdot)\,d\rho_0 = \int_0^1 \!\!\int \big[\partial_t\varphi + b_t \cdot \nabla\varphi\big]\,d\rho_t\,dt. \tag{37}$$

In particular, $\partial_t \rho_t + \nabla\cdot(b_t\rho_t) = 0$ in the sense of distributions on $(0,1) \times \mathbb{R}^d$.

Proof. Fix $\varphi \in C_c^1([0,1] \times \mathbb{R}^d)$. Its derivative is continuous and vanishes outside a compact set, so it is bounded, and the mean value theorem then shows that φ is Lipschitz. Composing a Lipschitz function with an absolutely continuous path gives an absolutely continuous function, so $t \mapsto \varphi(t, X_t)$ is absolutely continuous for almost every path. At every t where the path is differentiable, and hence for almost every t, the chain rule gives its derivative as $\partial_t \varphi(t, X_t) + \nabla \varphi(t, X_t) \cdot \dot{X}_t$. Integrating over $[0,1]$,

$$\varphi(1, X_1) - \varphi(0, X_0) = \int_0^1 \big[\partial_t \varphi(t, X_t) + \nabla \varphi(t, X_t) \cdot \dot{X}_t\big] dt \quad \text{almost surely.} \tag{38}$$

We now take expectations on both sides of eq. 38. The integrand is bounded by $C(1 + \|\dot{X}_t\|)$, with C the larger of $\sup |\partial_t \varphi|$ and $\sup \|\nabla \varphi\|$, and this bound is integrable by eq. 36, so the expectation may be exchanged with the time integral. This gives

$$\mathbb{E}[\varphi(1, X_1)] - \mathbb{E}[\varphi(0, X_0)] = \int_0^1 \mathbb{E}[\partial_t \varphi(t, X_t)]\, dt + \int_0^1 \mathbb{E}[\nabla \varphi(t, X_t) \cdot \dot{X}_t]\, dt. \tag{39}$$

Since $X_0 \sim \rho_0$ and $X_1 \sim \rho_1$, the left-hand side of eq. 39 equals $\int \varphi(1, \cdot)\, d\rho_1 - \int \varphi(0, \cdot)\, d\rho_0$. Since $X_t \sim \rho_t$, the first term on the right equals $\int_0^1 \int \partial_t \varphi(t, \cdot)\, d\rho_t\, dt$. For the second term, fix t such that $\dot{X}_t$ is integrable, which holds for almost every t by eq. 36. The vector $\nabla \varphi(t, X_t)$ is bounded and is a function of X_t, so the tower property gives

$$\mathbb{E}[\nabla \varphi(t, X_t) \cdot \dot{X}_t] = \mathbb{E}\big[\nabla \varphi(t, X_t) \cdot \mathbb{E}[\dot{X}_t \mid X_t]\big] = \mathbb{E}[\nabla \varphi(t, X_t) \cdot b_t(X_t)] = \int \nabla \varphi(t, \cdot) \cdot b_t\, d\rho_t.$$

Substituting these expressions into eq. 39 gives eq. 37.

Finally, if φ is supported in $(0,1) \times \mathbb{R}^d$, then $\varphi(0, \cdot) = \varphi(1, \cdot) = 0$ and the left-hand side of eq. 37 is zero. The identity eq. 37 then states that $\int_0^1 \int [\partial_t \varphi + b_t \cdot \nabla \varphi]\, d\rho_t\, dt = 0$ for every such φ, which is the definition of $\partial_t \rho_t + \nabla \cdot (b_t \rho_t) = 0$ in the sense of distributions. □

A.3 Training objectives

Theorem A.3 (Velocity regression)**.** *Let X be a stochastic interpolant with $\mathbb{E} \int_0^1 \|\dot{X}_t\|^2 dt < \infty$. For every measurable $\hat{b}$ with $\mathbb{E} \int_0^1 \|\hat{b}(t, X_t)\|^2 dt < \infty$, the objective*

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\Big[\|\hat{b}(t, X_t)\|^2 - 2\, \dot{X}_t \cdot \hat{b}(t, X_t)\Big] dt \tag{40}$$

satisfies

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \int \|\hat{b}(t, \cdot) - b_t\|^2\, d\rho_t\, dt - \int_0^1 \int \|b_t\|^2\, d\rho_t\, dt. \tag{41}$$

Its unique minimizer, up to $dt\, \rho_t(dx)$-null sets, is b.

Proof. Fix t with $\mathbb{E}\|\dot{X}_t\|^2 < \infty$. This holds for almost every t by assumption. The variable $b_t(X_t)$ is square integrable because it is a conditional expectation of a square-integrable vector, and $\hat{b}(t, X_t)$ is a square-integrable function of X_t, so the tower property gives

$$\mathbb{E}[\dot{X}_t \cdot \hat{b}(t, X_t)] = \mathbb{E}[b_t(X_t) \cdot \hat{b}(t, X_t)].$$

Substituting this into the integrand of eq. 40 and completing the square,

$$\mathbb{E}\Big[\|\hat{b}(t, X_t)\|^2 - 2\, \dot{X}_t \cdot \hat{b}(t, X_t)\Big] = \mathbb{E}\|\hat{b}(t, X_t) - b_t(X_t)\|^2 - \mathbb{E}\|b_t(X_t)\|^2.$$

Integrating over t and using $X_t \sim \rho_t$ gives eq. 41, in which the second integral is finite and independent of $\hat{b}$ while the first is nonnegative and vanishes exactly when $\hat{b}(t, \cdot) = b_t$ for $dt\, \rho_t(dx)$-almost every (t, x). □

Write $C_c^1(\mathbb{R}^d;\mathbb{R}^d)$ for the continuously differentiable vector fields with compact support and $\nabla\cdot w = \sum_i \partial_i w_i$ for the divergence. Definitions of this kind characterise the logarithmic gradient of a measure through an integration-by-parts identity. They come from the theory of differentiable measures, which is developed in detail by Bogachev (1997, Chapters 6).

Definition A.4 (Score)**.** Let μ be a probability measure on $\mathbb{R}^d$. A measurable map $s:\mathbb{R}^d\to\mathbb{R}^d$ with $\int \|s\|\,d\mu<\infty$ is the score of μ if

$$\int \nabla\cdot w\,d\mu = -\int s\cdot w\,d\mu \qquad \text{for every } w\in C_c^1(\mathbb{R}^d;\mathbb{R}^d). \tag{42}$$

Testing eq. 42 against $w=\psi e_i$, with $\psi\in C_c^1(\mathbb{R}^d)$ arbitrary, gives $\int \psi\,(s_i-s_i')\,d\mu=0$ for two scores s and s', so they agree μ-almost everywhere. If μ has a positive continuously differentiable density p with $\int\|\nabla\log p\|\,d\mu<\infty$, integration by parts shows that $\nabla\log p$ is its score.

Theorem A.5 (Denoising)**.** *Let $X_t=I(t,X_0,X_1)+\gamma_t Z$ be a stochastic interpolant with $Z\sim\mathcal{N}(0,I_d)$ independent of (X_0,X_1), and let $\eta_t(x)=\mathbb{E}[Z\mid X_t=x]$ be a version of this conditional expectation that is jointly measurable in (t,x). The objective*

$$\mathcal{L}_\eta[\hat\eta]=\int_0^1 \mathbb{E}\|\hat\eta(t,X_t)-Z\|^2 dt \tag{43}$$

has η as its unique minimizer, up to $dt\,\rho_t(dx)$-null sets. For every t with $\gamma_t>0$, the map $s_t=-\eta_t/\gamma_t$ is the score of ρ_t.

Proof. The objective eq. 43 equals $\int_0^1\mathbb{E}[\|\hat\eta(t,X_t)\|^2-2Z\cdot\hat\eta(t,X_t)]\,dt+d$, and the proof of theorem A.3 with Z in place of $\dot X_t$ applies to the first term.

For the second claim, fix t with $\gamma_t>0$ and $w\in C_c^1(\mathbb{R}^d;\mathbb{R}^d)$. For a bounded continuously differentiable function f with bounded derivative, Gaussian integration by parts gives $\mathbb{E}[Z_i f(Z)]=\mathbb{E}[\partial_i f(Z)]$. Since Z is independent of (X_0,X_1), we may apply it conditionally on (X_0,X_1) with $f(z)=w_i(I(t,X_0,X_1)+\gamma_t z)$. Its partial derivative in z_i is $\gamma_t\,\partial_i w_i(I(t,X_0,X_1)+\gamma_t z)$. Summing over i and taking expectations gives

$$\mathbb{E}[Z\cdot w(X_t)]=\gamma_t\,\mathbb{E}[\nabla\cdot w(X_t)]=\gamma_t\int\nabla\cdot w\,d\rho_t. \tag{44}$$

The field $w(X_t)$ is bounded and is a function of X_t, so the tower property gives

$$\mathbb{E}[Z\cdot w(X_t)]=\mathbb{E}[\eta_t(X_t)\cdot w(X_t)]=\int\eta_t\cdot w\,d\rho_t.$$

Equating the two expressions for $\mathbb{E}[Z\cdot w(X_t)]$ gives $\int\nabla\cdot w\,d\rho_t=-\int s_t\cdot w\,d\rho_t$ with $s_t=-\eta_t/\gamma_t$. The integrability required in definition A.4 holds because $\int\|s_t\|\,d\rho_t\le\gamma_t^{-1}\mathbb{E}\|Z\|$. □

A.4 Fokker–Planck equations and sampling

Write $C_c^2([0,1)\times\mathbb{R}^d)$ for the twice continuously differentiable functions with compact support in $[0,1)\times\mathbb{R}^d$, and Δ for the Laplacian in x.

Theorem A.6 (Fokker–Planck)**.** *Let X be a stochastic interpolant. Let $s:(0,1)\times\mathbb{R}^d\to\mathbb{R}^d$ be a measurable map such that s_t is the score of ρ_t for almost every t, and let $\varepsilon:[0,1)\to[0,\infty)$ be measurable and locally bounded with*

$$\int_0^T \varepsilon_t\int\|s_t\|\,d\rho_t\,dt<\infty \qquad \textit{for every } T<1. \tag{45}$$

Then for every $\varphi\in C_c^2([0,1)\times\mathbb{R}^d)$,

$$\int\varphi(0,\cdot)\,d\rho_0+\int_0^1\int\left[\partial_t\varphi+(b_t+\varepsilon_t s_t)\cdot\nabla\varphi+\varepsilon_t\Delta\varphi\right]d\rho_t\,dt=0. \tag{46}$$

Proof. Fix φ and $T < 1$ such that φ vanishes on $[T, 1) \times \mathbb{R}^d$. Extending φ by zero at $t = 1$ gives a function in $C_c^1([0, 1] \times \mathbb{R}^d)$ with $\varphi(1, \cdot) = 0$, so theorem A.2 gives

$$\int \varphi(0, \cdot)\, d\rho_0 + \int_0^1 \int \big[\partial_t \varphi + b_t \cdot \nabla \varphi\big]\, d\rho_t\, dt = 0.$$

It remains to show that the terms involving ε_t integrate to zero. For almost every t, s_t is the score of ρ_t, and $w = \nabla\varphi(t, \cdot)$ belongs to $C_c^1(\mathbb{R}^d; \mathbb{R}^d)$ because φ is twice continuously differentiable with compact support. Its divergence is $\Delta\varphi(t, \cdot)$, so eq. 42 gives

$$\int \big[s_t \cdot \nabla\varphi(t, \cdot) + \Delta\varphi(t, \cdot)\big]\, d\rho_t = 0 \qquad \text{for almost every } t.$$

With C the larger of $\sup \|\nabla\varphi\|$ and $\sup |\Delta\varphi|$, the integrand in t multiplied by ε_t is bounded by $C\varepsilon_t\big(1 + \int \|s_t\|\, d\rho_t\big)$, which is integrable on $[0, T]$ by eq. 45 and the local boundedness of ε. Its integral over $[0, T]$ is therefore zero, and adding it to the transport identity gives eq. 46. □

With $\varepsilon \equiv 0$, eq. 46 reduces to the transport identity. A curve (μ_t) of probability measures is narrowly continuous if $t \mapsto \int f\, d\mu_t$ is continuous for every bounded continuous f. The maps b and s are determined only up to null sets under $dt\, \rho_t(dx)$, while a sampler may visit such sets, so we fix versions of b and s for the rest of this section. Let $\mathcal{C}$ be the set of narrowly continuous curves $(\mu_t)_{t\in[0,1)}$ of probability measures with $\int_0^T \int \|b_t + \varepsilon_t s_t\|\, d\mu_t\, dt < \infty$ for every $T < 1$.

Theorem A.7 (Sampling). *Let s and ε be as in theorem A.6, and assume that eq. 46 has at most one solution in $\mathcal{C}$. Let W be a standard Brownian motion in $\mathbb{R}^d$. Let $(\hat{X}_t)_{t\in[0,1)}$ be a continuous process with $\hat{X}_0 \sim \rho_0$ independent of W that solves*

$$d\hat{X}_t = (b_t + \varepsilon_t s_t)(\hat{X}_t)\, dt + \sqrt{2\varepsilon_t}\, dW_t, \tag{47}$$

in the sense that $\hat{X}_t = \hat{X}_0 + \int_0^t (b_r + \varepsilon_r s_r)(\hat{X}_r)\, dr + \int_0^t \sqrt{2\varepsilon_r}\, dW_r$ for all $t < 1$ almost surely, and with $\mathbb{E}\int_0^T \|(b_t + \varepsilon_t s_t)(\hat{X}_t)\|\, dt < \infty$ for every $T < 1$. Then $\hat{X}_t \sim \rho_t$ for every $t \in [0, 1)$. Moreover, $\rho_t \to \rho_1$ narrowly as $t \to 1$, and if $\hat{X}_t$ converges almost surely as $t \to 1$, its limit has law ρ_1.

Proof. Let μ_t be the law of $\hat{X}_t$. We first show that μ and ρ both belong to $\mathcal{C}$. The paths of $\hat{X}$ are continuous, so for bounded continuous f and $t_n \to t$, dominated convergence gives $\int f\, d\mu_{t_n} \to \int f\, d\mu_t$. The integrability required in $\mathcal{C}$ is the drift assumption written in terms of μ_t. The paths of X are also continuous, and the same argument shows that ρ is narrowly continuous on $[0, 1]$, including at $t = 1$. For almost every t, $\|b_t(X_t)\| \le \mathbb{E}[\|\dot{X}_t\| \mid X_t]$, since the norm of a conditional expectation is at most the conditional expectation of the norm. Integrating gives $\int_0^1 \int \|b_t\|\, d\rho_t\, dt \le \mathbb{E}\int_0^1 \|\dot{X}_t\|\, dt$, which is finite by eq. 36, and together with eq. 45 this places ρ in $\mathcal{C}$.

It remains to show that μ solves eq. 46. Fix φ and T as in the previous proof. Itô's formula applied to $\varphi(t, \hat{X}_t)$ on $[0, T]$, with $\varphi(T, \cdot) = 0$, gives

$$-\varphi(0, \hat{X}_0) = \int_0^T \big[\partial_t \varphi + (b_t + \varepsilon_t s_t) \cdot \nabla\varphi + \varepsilon_t \Delta\varphi\big](t, \hat{X}_t)\, dt + \int_0^T \sqrt{2\varepsilon_t}\, \nabla\varphi(t, \hat{X}_t) \cdot dW_t.$$

The stochastic integrand is bounded on $[0, T]$, so the stochastic integral has expectation zero. The time integrand is bounded by a constant times $1 + \|(b_t + \varepsilon_t s_t)(\hat{X}_t)\|$, which is integrable by assumption, so the expectation may be exchanged with the time integral. Taking expectations and using $\mu_0 = \rho_0$ shows that μ satisfies eq. 46. Both μ and ρ are solutions in $\mathcal{C}$, so $\mu_t = \rho_t$ for every $t < 1$ by uniqueness.

Narrow continuity of ρ at $t = 1$ gives $\rho_t \to \rho_1$. If $\hat{X}_t$ converges almost surely to a limit $\hat{X}_1$, dominated convergence gives $\int f\, d\mu_t \to \mathbb{E} f(\hat{X}_1)$ for bounded continuous f. Since $\mu_t = \rho_t$ converges narrowly to ρ_1, the limit $\hat{X}_1$ has law ρ_1. □

Uniqueness holds, for instance, when $b_t + \varepsilon_t s_t$ is bounded and Lipschitz in x, uniformly on $[0, T] \times \mathbb{R}^d$ for each $T < 1$ (Ambrosio et al., 2008; Bogachev et al., 2015). The exact fields need not satisfy this near $t = 1$, which is why the statement is posed on $[0, 1)$.

B PROOFS

B.1 PRELIMINARIES

Throughout this appendix $\mathbf{c} : \mathbb{R}^d \to \mathcal{Y}$ is a C^3 submersion, and $(X_t)_{t\in[0,1]}$ is a disintegrated stochastic interpolant in the sense of definition 2.2. We use the separability condition eq. 6. The paths are almost surely absolutely continuous, with $\int_0^1 \mathbb{E}\|\dot{X}_t\|^2 dt < \infty$. The schedule B_t is continuously differentiable, is a diffeomorphism onto its image for $t < 1$, and satisfies $B_0 = \mathrm{id}$ and $(B_1)_\# \nu_0 = \nu_1$.

For the transport argument we assume

$$\int_0^T \mathbb{E}\|h_t(X_t)\|\, dt < \infty \qquad \text{for every } T < 1. \tag{48}$$

We also require $\int_0^1 \mathbb{E}\|\mathsf{P}_{X_t}\dot{X}_t\|^2 dt < \infty$; for the orthogonal connection, this follows from the bound on $\dot{X}_t$. Conditional expectations below are taken in jointly measurable versions, defined up to $dt\,\rho_t(dx)$-null sets. Tangent versions are obtained by setting the fields to zero on any exceptional set. All evolution equations are understood in the sense of distributions in time, before $t = 1$.

The derivatives of the curves of measures are obtained as follows. For $\varphi \in C_c^1(\mathbb{R}^d)$ absolute continuity of the paths gives $\varphi(X_t) - \varphi(X_0) = \int_0^t \nabla\varphi(X_s)\cdot\dot{X}_s\, ds$ almost surely. The integrand is dominated by $\|\nabla\varphi\|_\infty\|\dot{X}_s\|$, whose integral over $[0,1]\times\Omega$ is finite. Fubini therefore gives

$$\frac{d}{dt}\mathbb{E}[\varphi(X_t)] = \mathbb{E}[\nabla\varphi(X_t)\cdot\dot{X}_t] = \int \nabla\varphi\cdot b_t\, d\rho_t \qquad \text{for almost every } t. \tag{49}$$

The curve ρ_t is narrowly continuous by path continuity and bounded convergence.

Each fibre carries the metric induced by $\mathbb{R}^d$ and its Riemannian volume σ_y. For a tangent field w, the distributional operators on measures are defined by

$$\langle \mathrm{div}_{\mathcal{M}}(w\mu), \varphi\rangle = -\int \langle \nabla_{\mathcal{M}}\varphi, w\rangle\, d\mu, \qquad \langle \Delta_{\mathcal{M}}\mu, \varphi\rangle = \int \Delta_{\mathcal{M}}\varphi\, d\mu. \tag{50}$$

Here $\nabla_{\mathcal{M}}\varphi$ is the intrinsic gradient of the restriction of φ to the fibre. For every tangent w, $\langle\nabla_{\mathcal{M}}\varphi, w\rangle = \langle\nabla\varphi, w\rangle$. The divergence of a tangent vector field is taken with respect to σ_y. These definitions do not require μ to have a density.

Lemma B.1. *The lift and projection satisfy $D\mathbf{c}(x)\mathsf{H}_x = \mathrm{id}_{T_{\mathbf{c}(x)}\mathcal{Y}}$, $D\mathbf{c}(x)\mathsf{P}_x = 0$, and $\mathsf{P}_x^2 = \mathsf{P}_x$. Consequently $u_t(x) = \mathsf{P}_x b_t(x)$ is tangent to the fibre and*

$$\int_0^1 \mathbb{E}\|u_t(X_t)\|^2 dt \le \int_0^1 \mathbb{E}\|\mathsf{P}_{X_t}\dot{X}_t\|^2 dt.$$

For the orthogonal connection, P_x is the orthogonal projector onto $\mathcal{V}_x$, and $\|\mathsf{P}_x v\| \le \|v\|$.

Proof. The first identity is the defining property of the lift. Substituting $\mathsf{P}_x = I_d - \mathsf{H}_x D\mathbf{c}(x)$ gives the other two identities. Conditioning on $X_t = x$ fixes P_x, so it passes out of the expectation and gives $u_t = \mathsf{P}_x b_t$. The integral bound follows from conditional Jensen. For the orthogonal choice in eq. 4, $\mathsf{H}_x D\mathbf{c}(x)$ is symmetric and idempotent. Its complement is therefore the orthogonal projector onto $\ker D\mathbf{c}(x)$ and has operator norm at most one. □

B.2 PROOF OF THEOREM 2.3

Theorem (theorem 2.3, restated)**.** *Under the assumptions above, $\mathbf{c}_\#\rho_t = (B_t)_\#\nu_0 = \nu_t$ for every $t < 1$, and ν_t solves eq. 11. The disintegrations can be chosen so that, for ν_0-almost every y, the curve $\mu_t^{B_t(y)}$ is narrowly continuous, is concentrated on $\mathcal{M}_{B_t(y)}$, and solves eq. 12. Their mixture is ρ_t, which solves eq. 13.*

Proof. We first establish the ambient equation. By lemma B.1,

$$b_t - h_t - u_t = \mathsf{H}_x\big(D\mathbf{c}(x)b_t(x) - \beta_t(\mathbf{c}(x))\big).$$

Separability makes the pairing of this difference with $\nabla\varphi$ vanish after integration against ρ_t. Thus eq. 49 becomes

$$\frac{d}{dt}\int \varphi\, d\rho_t = \int \nabla\varphi \cdot (h_t + u_t)\, d\rho_t,$$

which is eq. 13.

Fix $T < 1$. By eq. 48 and lemma B.1, $h+u$ is integrable against $dt\,\rho_t(dx)$ on $[0,T]$. The superposition principle for the continuity equation therefore gives a process $\widetilde{X}_t$ with absolutely continuous paths such that

$$\mathrm{Law}(\widetilde{X}_t) = \rho_t, \qquad \dot{\widetilde{X}}_t = (h_t + u_t)(\widetilde{X}_t) \quad \text{for almost every } t,$$

almost surely; see Ambrosio et al. (2008). The chain rule and tangency of u_t give

$$\frac{d}{dt}\mathbf{c}(\widetilde{X}_t) = D\mathbf{c}(\widetilde{X}_t)(h_t + u_t)(\widetilde{X}_t) = \beta_t(\mathbf{c}(\widetilde{X}_t)). \tag{51}$$

Differentiating $B_t^{-1}(\mathbf{c}(\widetilde{X}_t))$ shows that it is constant. Consequently $\mathbf{c}(\widetilde{X}_t) = B_t(\mathbf{c}(\widetilde{X}_0))$ for every $t \le T$, almost surely. This identity is used for the process $\widetilde{X}_t$ constructed above.

Taking laws gives $\mathbf{c}_{\#}\rho_t = (B_t)_{\#}\nu_0$. For $\psi \in C_c^1(\mathcal{Y})$, the chain rule along $B_t(y)$ gives

$$\frac{d}{dt}\int \psi\, d\nu_t = \int \langle \nabla_{\mathcal{Y}}\psi, \beta_t\rangle\, d\nu_t,$$

which is eq. 11. The calculation is local in time and on the support of ψ. At the endpoint, continuity of B_t and bounded convergence give $\nu_t \to (B_1)_{\#}\nu_0 = \nu_1$ narrowly.

We next turn to the fibre curves. Disintegrate the law of the paths $\widetilde{X}$ with respect to $\mathbf{c}(\widetilde{X}_0) = y$. Under this conditional law the paths lie in $\mathcal{M}_{B_t(y)}$ at time t. Since B_t is injective, their time marginals give a version of $\mu_t^{B_t(y)}$, the conditional law at the current structural value. Path continuity makes these conditional curves narrowly continuous. Fubini and the integrability of $h + u$ imply that their expected total variation is finite for ν_0-almost every y. Applying the pathwise chain rule under each such conditional law yields

$$\frac{d}{dt}\int \varphi\, d\mu_t^{B_t(y)} = \int \langle \nabla\varphi, h_t\rangle\, d\mu_t^{B_t(y)} + \int \langle \nabla_{\mathcal{M}}\varphi, u_t\rangle\, d\mu_t^{B_t(y)}. \tag{52}$$

A countable dense family of compactly supported test functions provides a single exceptional set in y. This is eq. 12.

Repeating the construction along a sequence $T \uparrow 1$ gives versions on $[0,1)$. On overlapping intervals these versions agree for every rational time outside a common ν_0-null set, by uniqueness of disintegration. Narrow continuity then gives agreement at every time. Thus the conditional curves can be chosen consistently on $[0,1)$.

Finally, disintegration and $\nu_t = (B_t)_{\#}\nu_0$ give

$$\rho_t = \int_{\mathcal{Y}} \mu_t^z\, \nu_t(dz) = \int_{\mathcal{Y}} \mu_t^{B_t(y)}\, \nu_0(dy).$$

Integrating eq. 52 against ν_0, and using tangency of u_t, returns

$$\frac{d}{dt}\int \varphi\, d\rho_t = \int \langle \nabla\varphi, h_t + u_t\rangle\, d\rho_t.$$

This proves the recombination statement. □

B.3 Proof of theorem 2.4

Theorem (theorem 2.4, restated)**.** *The objective eq. 14 has a unique minimiser among square-integrable measurable tangent fields, up to $dt\,\rho_t(dx)$-null sets. It is the vertical drift $u_t(x) = \mathbb{E}[\mathsf{P}_{X_t}\dot{X}_t \mid X_t = x]$.*

Proof. We first check that u_t is admissible. By lemma B.1 it is tangent, and conditional Jensen gives

$$\int_0^1 \mathbb{E}\|u_t(X_t)\|^2 dt \le \int_0^1 \mathbb{E}\|\mathsf{P}_{X_t}\dot{X}_t\|^2 dt < \infty.$$

Let now $\hat{u}$ be an admissible field. The cross term of eq. 14 is integrable by Cauchy–Schwarz, and the tower property replaces $\mathsf{P}_{X_t}\dot{X}_t$ by $u_t(X_t)$ under the expectation. Completing the square,

$$\begin{aligned}\mathcal{L}_u[\hat{u}] &= \int_0^1 \mathbb{E}\big[\|\hat{u}(t,X_t)\|^2 - 2\langle u_t(X_t), \hat{u}(t,X_t)\rangle\big] dt \\ &= \int_0^1 \mathbb{E}\|\hat{u}(t,X_t) - u_t(X_t)\|^2 dt - \int_0^1 \mathbb{E}\|u_t(X_t)\|^2 dt.\end{aligned}$$

The second integral does not depend on $\hat{u}$, and the first is nonnegative and vanishes precisely when $\hat{u}(t,x) = u_t(x)$ for $dt\,\rho_t(dx)$-almost every (t,x). This is the stated conclusion. □

B.4 Proof of theorem 2.5

Theorem (theorem 2.5, restated)**.** *Let the tangent noise target ξ_t satisfy eq. 17, and let $\lambda(t) > 0$ for $t \in (0,1)$, with $\int_0^1 \lambda(t)\mathbb{E}\|\xi_t\|^2 dt < \infty$. The objective eq. 18 has a unique minimiser among measurable tangent fields with finite weighted second moment, up to $dt\,\rho_t(dx)$-null sets. It is the weak fibre score, given by $s_t(x) = \mathbb{E}[\xi_t \mid X_t = x]$.*

Proof. Write $\bar{\xi}_t(x) = \mathbb{E}[\xi_t \mid X_t = x]$. Conditioning fixes the tangent space, so $\bar{\xi}_t(x)$ is tangent. Conditional Jensen gives

$$\int_0^1 \lambda(t)\mathbb{E}\|\bar{\xi}_t(X_t)\|^2 dt \le \int_0^1 \lambda(t)\mathbb{E}\|\xi_t\|^2 dt < \infty.$$

We first identify $\bar{\xi}_t$ with the weak fibre score. Fix a time at which the noise identity and the second-moment bound hold. Averaging eq. 17 over the endpoints and applying the tower property gives, for every compactly supported C^1 tangent field w,

$$\int \langle \bar{\xi}_t, w\rangle \, d\rho_t = -\int \operatorname{div}_{\mathcal{M}} w \, d\rho_t.$$

For $\psi \in C_c^1(\mathcal{Y})$ the field $(\psi \circ \mathbf{c})w$ is also tangent. Since $\psi \circ \mathbf{c}$ is constant on each fibre,

$$\operatorname{div}_{\mathcal{M}}\big((\psi \circ \mathbf{c})w\big) = (\psi \circ \mathbf{c}) \operatorname{div}_{\mathcal{M}} w.$$

Testing with this field and disintegrating ρ_t gives

$$\int_{\mathcal{Y}} \psi(y) \left[\int_{\mathcal{M}_y} \big(\langle \bar{\xi}_t, w\rangle + \operatorname{div}_{\mathcal{M}} w\big)\, d\mu_t^y\right] \nu_t(dy) = 0.$$

As ψ is arbitrary, the bracket vanishes for ν_t-almost every y. Using a countable dense family of tangent fields in submersion charts gives one exceptional set for all tests. Compactly supported tangent fields on a fibre extend locally to tangent fields on neighbouring fibres; a partition of unity then gives the identity for every such field. Thus

$$\int_{\mathcal{M}_y} \langle \bar{\xi}_t, w\rangle \, d\mu_t^y = -\int_{\mathcal{M}_y} \operatorname{div}_{\mathcal{M}} w \, d\mu_t^y \qquad \text{for } \nu_t\text{-almost every } y. \tag{53}$$

This is the defining identity for the weak fibre score. Such a tangent field is unique up to μ_t^y-null sets, since compactly supported tangent test fields determine its pairing with the measure. Hence $\bar{\xi}_t = s_t$ for ρ_t-almost every x.

It remains to minimise the objective. For an admissible $\hat{s}$, the tower property makes the cross term between $\hat{s}(t,X_t) - \bar{\xi}_t(X_t)$ and $\xi_t - \bar{\xi}_t(X_t)$ vanish. Expanding the square therefore gives

$$\begin{aligned}\mathcal{L}_s[\hat{s}] = &\int_0^1 \lambda(t)\mathbb{E}\|\hat{s}(t,X_t) - \bar{\xi}_t(X_t)\|^2 dt \\ &+ \int_0^1 \lambda(t)\mathbb{E}\|\xi_t - \bar{\xi}_t(X_t)\|^2 dt.\end{aligned}$$

Both terms are finite. The second does not depend on $\hat{s}$, and the first vanishes precisely when $\hat{s} = \bar{\xi} = s$ almost everywhere. Since $\lambda(t) > 0$ on $(0, 1)$, uniqueness holds up to $dt\,\rho_t(dx)$-null sets. This proves the score representation and uniqueness. □

B.5 Proof of theorem 2.6

Theorem (theorem 2.6, restated)**.** *For the orthogonal connection, any continuous solution of eq. 20 on $[0, 1)$ satisfies $\mathbf{c}(\widehat{X}_t) = B_t(\mathbf{c}(\widehat{X}_0))$ almost surely for every $t < 1$. With exact fields and the stated integrability and Fokker–Planck uniqueness assumptions, $\widehat{X}_t \sim \rho_t$ for every $t < 1$, and these laws converge narrowly to ρ_1 as $t \to 1$.*

Proof. For the orthogonal connection, lemma B.1 identifies P_x as the orthogonal projector onto the tangent space of the fibre through x. All coefficients in the sampler are taken in the fixed versions specified in the theorem, with the learned fields tangent at every point. The diffusion schedule ε is nonnegative and continuous on $[0, 1)$.

The Stratonovich chain rule, together with $D\mathbf{c}\,\hat{u} = D\mathbf{c}\,\hat{s} = 0$ and $D\mathbf{c}\,\mathsf{P} = 0$, gives

$$d\mathbf{c}(\widehat{X}_t) = D\mathbf{c}(\widehat{X}_t)h_t(\widehat{X}_t)\,dt = \beta_t(\mathbf{c}(\widehat{X}_t))\,dt.$$

As in eq. 51, differentiating $B_t^{-1}(\mathbf{c}(\widehat{X}_t))$ gives $\mathbf{c}(\widehat{X}_t) = B_t(\mathbf{c}(\widehat{X}_0))$ for all $t < 1$ on one event of probability one. Taking laws gives the base law ν_t.

We next identify the diffusion operator. Let e_j, $1 \le j \le d$, be the standard basis of $\mathbb{R}^d$ and set $E_j(x) = \mathsf{P}_x e_j$. For $\varphi \in C_c^2(\mathbb{R}^d)$, write $E_j\varphi = E_j \cdot \nabla\varphi$. Orthogonal projection gives $\nabla_{\mathcal{M}}\varphi = \sum_j (E_j\varphi)E_j$ and $\sum_j (\mathrm{div}_{\mathcal{M}} E_j)E_j = 0$. The latter follows by differentiating $\mathsf{P}^2 = \mathsf{P}$ along tangent directions: the derivative of P sends tangent vectors into the normal space. Taking the divergence of the gradient therefore gives

$$\sum_{j=1}^{d} E_j(E_j\varphi) = \Delta_{\mathcal{M}}\varphi.$$

Consequently the Stratonovich noise in eq. 20 contributes $\varepsilon_t\Delta_{\mathcal{M}}$ to the generator.

Now take $\hat{u} = u$ and $\hat{s} = s$. The uniqueness class consists of narrowly continuous probability curves m_t with $m_0 = \rho_0$ and

$$\int_0^T \int \big(\|h_t + u_t\| + \varepsilon_t\|s_t\|\big)\,dm_t\,dt < \infty \qquad \text{for every } T < 1.$$

By the assumptions in the theorem, both ρ_t and the laws of $\widehat{X}_t$ belong to this class. Itô's formula and expectation give the weak Fokker–Planck equation for the sampler laws,

$$\frac{d}{dt}\int \varphi\,dm_t = \int \big[(h_t + u_t + \varepsilon_t s_t)\cdot\nabla\varphi + \varepsilon_t\Delta_{\mathcal{M}}\varphi\big]dm_t. \tag{54}$$

The martingale has zero expectation: P is bounded, $\nabla\varphi$ is bounded, and ε is integrable on each $[0, T]$. The remaining terms are integrable by the assumptions above and the local regularity of P.

To check the same equation for ρ_t, apply the weak score identity to $w = \nabla_{\mathcal{M}}\varphi$ on each fibre and integrate over ν_t. Since s_t is tangent, this gives

$$\int s_t \cdot \nabla\varphi\,d\rho_t = -\int \Delta_{\mathcal{M}}\varphi\,d\rho_t.$$

The score and diffusion terms therefore cancel. The transport equation (13) is exactly eq. 54 for ρ_t. Uniqueness in the stated class yields $\mathrm{Law}(\widehat{X}_t) = \rho_t$ for every $t < 1$. When $\varepsilon \equiv 0$, the same argument uses only the transport equation and requires no score.

Finally, the interpolant satisfies $X_t \to X_1$ almost surely. For every bounded continuous f, bounded convergence gives $\int f\,d\rho_t \to \int f\,d\rho_1$, proving the narrow convergence. If $\widehat{X}_t$ has an almost sure limit as $t \to 1$, the same argument shows that its law is ρ_1. □

Remark B.2. The transport and velocity regression arguments apply to any smooth choice of horizontal complement under the stated integrability assumptions. The score uses the induced fibre metric and its Riemannian volume. Changing that metric requires the intrinsic operators and the diffusion to be changed consistently. The projected Stratonovich realisation used above is for the metric induced by the ambient Euclidean space.

Remark B.3. Taking $k = 0$ gives $\mathsf{P}_x = I_d$, a single fibre $\mathcal{M} = \mathbb{R}^d$, Lebesgue reference measure, and $h_t = 0$. Then $u_t = b_t$, and theorem 2.3 reduces to $\partial_t \rho_t + \nabla \cdot (b_t \rho_t) = 0$. Theorem 2.4 gives the velocity regression objective of Albergo et al. (2025). For additive Gaussian noise $\gamma_t Z$, with Z independent of the endpoints and $\gamma_t > 0$, Gaussian integration by parts gives $\xi_t = -Z/\gamma_t$. Thus theorem 2.5 returns the ambient weak score $s_t(x) = -\gamma_t^{-1}\mathbb{E}[Z \mid X_t = x]$ whenever the stated weighted integrability condition holds. The sampling result then gives the usual Fokker–Planck family.

B.6 Score-only diffusion and implicit score matching

As discussed in section 3, general stochastic-interpolant diffusion uses both a velocity and a score. In (20), these are the learned tangent fields $u_t = \mathsf{P} b_t$ and s_t, together with the prescribed field h_t. The two learned fields may share a network.

When one score model suffices. For the ordinary one-sided Gaussian interpolant $X_t = a_t X_1 + \sigma_t Z$, with $Z \sim \mathcal{N}(0, I_d)$ independent of X_1, take $(a_0, \sigma_0) = (0, 1)$ and $(a_1, \sigma_1) = (1, 0)$, with $a_t, \sigma_t > 0$ in the interior. For $0 < t < 1$, Gaussian integration by parts and conditioning give

$$\mathbb{E}[Z \mid X_t = x] = -\sigma_t s_t(x), \qquad a_t \mathbb{E}[X_1 \mid X_t = x] = x + \sigma_t^2 s_t(x), \tag{55}$$

$$b_t(x) = \frac{\dot a_t}{a_t} x + \left(\sigma_t^2 \frac{\dot a_t}{a_t} - \sigma_t \dot\sigma_t \right) s_t(x). \tag{56}$$

Thus one score model determines the sampling drift $b_t + \varepsilon_t s_t$ (Albergo et al., 2025). This identity uses the affine Gaussian construction and does not extend to arbitrary interpolants. The nonlinear retraction in algorithm 2 removes this affine relation. It also gives

$$D_Z X_t = \gamma_t D\mathsf{R}_{y_t}\big(C_t + \gamma_t \mathsf{P}_{C_t} Z\big)\mathsf{P}_{C_t}, \tag{57}$$

which generally differs from $\gamma_t \mathsf{P}_{X_t}$. Consequently the projected Gaussian target need not satisfy the fibre score identity, and there is no general algebraic recovery of u_t from s_t.

Implicit and denoising score matching. Assume smooth positive densities, vanishing integration-by-parts boundary terms, and finite weighted second moments. Expectations below include $t \sim \mathcal{U}(0, 1)$, and $\lambda(t) > 0$. For an ordinary interpolant with density p_t and $s_t = \nabla \log p_t$, integration by parts yields $\mathbb{E}[s_\theta \cdot s_t] = -\mathbb{E}[\nabla \cdot s_\theta]$ at each time. Hence implicit score matching (ISM; Hyvärinen (2005)) minimises

$$\mathcal{L}_{\mathrm{ISM}} = \mathbb{E}\big[\lambda(t)\,\big(\tfrac{1}{2}\|s_\theta(t, X_t)\|^2 + \nabla \cdot s_\theta(t, X_t)\big)\big] \tag{58}$$

$$= \tfrac{1}{2}\mathbb{E}\big[\lambda(t)\|s_\theta(t, X_t) - s_t(X_t)\|^2\big] + C, \tag{59}$$

where C is independent of θ. For $X_t = I_t + \gamma_t Z$ with independent Gaussian Z, denoising score matching (DSM) instead minimises $\frac{1}{2}\mathbb{E}[\lambda(t)\|s_\theta(t, X_t) + Z/\gamma_t\|^2]$, with the same population minimiser (Vincent, 2011). DSM uses a known conditional score; ISM requires only samples and spatial derivatives of the model.

Fibrewise objective and algorithm. Write $\mu_t^y = q_t^y \sigma_y$ and $s_t = \nabla_{\mathcal{M}_y} \log q_t^y$ on each fibre. Applying the same integration by parts fibrewise and averaging over ν_t gives, for $s_\theta(t, x) = \mathsf{P}_x \hat s_\theta(t, x)$,

$$\mathcal{L}_{\mathrm{ISM}}^{\mathcal{M}} = \mathbb{E}\big[\lambda(t)\,\big(\tfrac{1}{2}\|s_\theta(t, X_t)\|^2 + \mathrm{div}_{\mathcal{M}}\, s_\theta(t, X_t)\big)\big] \tag{60}$$

$$= \tfrac{1}{2}\mathbb{E}\big[\lambda(t)\|s_\theta(t, X_t) - s_t(X_t)\|^2\big] + C. \tag{61}$$

An unbiased divergence estimate uses $\eta \sim \mathcal{N}(0, I_d)$ and $\zeta = \mathsf{P}_x \eta$:

$$\mathrm{div}_{\mathcal{M}}\, s_\theta(t, x) = \mathbb{E}_\eta\big[\zeta^\top D_x s_\theta(t, x)\zeta\big]. \tag{62}$$

Sample (t, X_t) using algorithm 2, draw η, and replace its denoising loss by the resulting estimate of (61). The Jacobian–vector product differentiates the full projected field, including P_x, with ζ held

fixed as the direction. Retain the velocity loss and sampler. This trains the score of the actual retracted law without a projected-noise approximation; it does not remove the velocity model.

A score-only alternative. To remove the velocity model, one can instead specify a known forward diffusion on each fixed fibre (De Bortoli et al., 2022). Starting from data $Y_0 \sim \mu_1^y$, choose a known tangent drift f_τ and scalar $\kappa_\tau > 0$:

$$dY_\tau = f_\tau(Y_\tau)\, d\tau + \sqrt{2\kappa_\tau}\mathsf{P}_{Y_\tau} \circ dW_\tau. \tag{63}$$

Train its fibre score $s_\theta(\tau, \cdot)$ by (61) using simulated forward samples. Under the usual time-reversal regularity assumptions, generation from the terminal law at time T uses

$$d\widehat{X}_t = [-f_{T-t} + 2\kappa_{T-t} s_\theta(T-t, \cdot)]\,(\widehat{X}_t)\, dt + \sqrt{2\kappa_{T-t}}\mathsf{P}_{\widehat{X}_t} \circ dW_t. \tag{64}$$

Only the score is learned. Sampling $y \sim \nu_1$ and then reversing the corresponding fibre diffusion recovers the target mixture with exact scores and terminal initialisation. This replaces the retracted-interpolant construction and requires a samplable terminal law on each fibre; ISM alone does not provide that law.

C Numerical Methods

We will first describe the linear algebra problems we need to solve to implement our method and then describe various solutions to it.

C.1 Linear Algebra Required by the Method

Let $\mathbf{c} : \mathbb{R}^d \to \mathbb{R}^k$ be the structure map and $J_x = D_x\mathbf{c}(x)$. We use the Euclidean orthogonal connection and assume J_x has full row rank wherever the following operations are applied. The numerical method requires minimum-norm linear solves for projection, lifting and nonlinear correction, together with differentiation of the constructed training paths.

Projection and lifting. The basic problem is to find the smallest displacement producing a prescribed change b in the structure:

$$\delta_x(b) = \operatorname*{arg\,min}_{\delta \in \mathbb{R}^d} \frac{1}{2}\|\delta\|^2 \quad \text{subject to} \quad J_x\delta = b. \tag{65}$$

The same solve defines both the horizontal lift and the tangent projection:

$$\mathsf{H}_x b = \delta_x(b), \qquad \mathsf{P}_x v = v - \delta_x(J_x v). \tag{66}$$

During training, we project the network output and path velocity to form the loss

$$\frac{1}{2}\left\|\mathsf{P}_{X_t}\left(v_\theta(t, X_t) - \dot{X}_t\right)\right\|^2. \tag{67}$$

Since the geometry is fixed with respect to the network parameters, the backward pass uses the same self-adjoint projector. During sampling, projection and lifting combine into one minimum-norm solve:

$$h_t(x) + \mathsf{P}_x v_\theta(t, x) = v_\theta(t, x) + \delta_x(\beta_t(\mathbf{c}(x)) - J_x v_\theta(t, x))\,. \tag{68}$$

Retractions. The map $\mathsf{R}_y(z)$ corrects a candidate z onto the fibre $\mathcal{M}_y = \{x : \mathbf{c}(x) = y\}$. It is used to construct training paths and to restore the prescribed structural value after a sampling step. Starting from $x^{(0)} = z$, a general correction iterates

$$x^{(\ell+1)} = x^{(\ell)} + \alpha_\ell \mathsf{H}_{x^{(\ell)}}\left(y - \mathbf{c}(x^{(\ell)})\right), \tag{69}$$

where $\alpha_\ell \in (0, 1]$ is a step length. Each lift is the minimum-norm solution of

$$J_{x^{(\ell)}}\Delta x = y - \mathbf{c}(x^{(\ell)}). \tag{70}$$

On convergence, the corrected point defines $\mathsf{R}_y(z)$ to numerical tolerance. This enforces the constraint but need not return the nearest point to z. Below, we describe more direct corrections for the constraints in each experiment.

Training-path derivatives. Training requires the velocity $\dot{X}_t$ of the corrected interpolation path. For example, suppose we correct the straight path $z_t = (1-t)x_0 + tx_1$ along fixed directions L:

$$X_t = z_t + La_t, \qquad \mathbf{c}(X_t) = y_t. \tag{71}$$

The coefficients a_t are chosen to satisfy the constraint. Differentiating this equation gives a linear system for how those coefficients change with time:

$$(J_{X_t}L)\dot{a}_t = \dot{y}_t - J_{X_t}(x_1 - x_0). \tag{72}$$

When $J_{X_t}L$ is invertible, we solve for $\dot{a}_t$ and obtain

$$\dot{X}_t = (x_1 - x_0) + L\dot{a}_t. \tag{73}$$

The training target is then $\mathsf{P}_{X_t}\dot{X}_t$. Other path constructions require differentiating their own defining equations, as described below.

C.2 Matrix-Free Solution with LSMR

LSMR (Fong & Saunders, 2011) solves least-squares problems using only products with a linear operator and its transpose. This avoids storing the constraint Jacobian or forming its Gram matrix, reducing memory requirements and avoiding the numerical loss associated with explicitly forming normal equations. For $J_x = D\mathbf{c}(x)$, these products are supplied by automatic differentiation as JVPs and VJPs. We write $\mathrm{LstSq}(A, b, 0)$ for the minimum-norm least-squares solution, computed by LSMR initialised at zero. For a full-rank operator A, define

$$\mathrm{LstSq}(A, b, \lambda) = \begin{cases} \arg\min_z \|Az - b\|^2 + \lambda^2\|z\|^2, & \lambda > 0 \text{ or } A \text{ is tall}, \\ \arg\min_{Az=b} \|z\|^2, & \lambda = 0 \text{ and } A \text{ is wide}. \end{cases} \tag{74}$$

Thus, the unregularised tall problem minimises the residual, whereas the wide problem returns the minimum-norm exact solution. For square invertible A, both definitions coincide at $\lambda = 0$.

Projection and lifting. A horizontal lift requires one solve,

$$\mathsf{H}_x b = \mathrm{LstSq}(J_x, b, 0). \tag{75}$$

To project v, use $b = J_x v$ and subtract the resulting correction:

$$\mathsf{P}_x v = v - \mathrm{LstSq}(J_x, J_x v, 0). \tag{76}$$

The same projection is used for the training target, network output and backward cotangent.

Retractions. Retraction requires repeated LSMR solves. Starting from $x^{(0)} = z$, each iteration computes

$$\delta^{(\ell)} = \mathrm{LstSq}\Big(J_{x^{(\ell)}}, y - \mathbf{c}(x^{(\ell)}), 0\Big), \tag{77}$$

$$x^{(\ell+1)} = x^{(\ell)} + \alpha_\ell \delta^{(\ell)}. \tag{78}$$

We recompute the Jacobian products at each iterate and stop when the constraint residual meets the prescribed tolerance.

Path derivatives. Implicit differentiation computes the path velocity through an additional linear solve, without differentiating through the retraction iterations. Its cost is comparable to a forward linear solve when convergence is comparable, following the differentiable-solver approach of Roy et al. (2025).

We next exploit the structure of each experiment to replace these general iterative solves with specialised computations.

C.3 Exploiting Structure in the Experiments

PDE trajectories. The PDE constraints combine affine initial or boundary conditions with integrated conservation laws. Write

$$\mathbf{c}(x) = \begin{bmatrix} Ax - a \\ r(x) \end{bmatrix}, \tag{79}$$

and let P_A project onto $\ker A$. The affine conditions can be handled analytically by fixing observed entries or equating boundary entries. We eliminate them before solving for the conservation correction.

Although the conservation residuals are cumulative in physical time, their consecutive differences depend only on adjacent time slices. Applying the first-difference operator E therefore gives a reduced Jacobian

$$K_x = E\, Dr(x)\mathsf{P}_A \tag{80}$$

whose Gram matrix $K_x K_x^\top$ is tridiagonal: rows associated with nonadjacent time intervals have disjoint support. To project a vector v, we solve

$$(K_x K_x^\top)q = K_x \mathsf{P}_A v, \qquad \mathsf{P}_x v = \mathsf{P}_A v - K_x^\top q. \tag{81}$$

We use a tridiagonal $LDL^\top$ factorisation, followed by forward and backward substitution. Lifting uses the same factorisation. For a prescribed derivative $b = (b_A, b_r)$, first compute the affine lift $h_A = A^\top (AA^\top)^{-1} b_A$, then solve

$$(K_x K_x^\top)q = E\big(b_r - Dr(x)h_A\big), \qquad \mathsf{H}_x b = h_A + K_x^\top q. \tag{82}$$

This preserves the orthogonal projection and lift while replacing the general iterative solver with a direct solve.

For training-path correction, we choose spatial shifts that change mass while preserving the affine conditions. In Burgers, shifting only the interior entries leaves the boundary fluxes unchanged, so the conservation correction is obtained directly from the mass discrepancy. In reaction–diffusion, a constant shift of each later time slice preserves the derivative-based boundary fluxes but changes the quadratic reaction term. After temporal differencing, the corrections are scalar quadratic equations solved sequentially in physical time. Differentiating these correction equations gives the path velocity through a direct solve for Burgers and a triangular solve for reaction–diffusion. A damped nonlinear correction is used if the structured construction fails its residual check. During sampling, affine conditions are restored after each step, while nonlinear conservation drift is removed by the terminal correction.

Molecular moments. After mass and moment normalisation, the molecular constraints have the same Gram structure, together with the centre-of-mass condition $a^\top Z = 0$, where $\|a\| = 1$. We enforce centring analytically:

$$Z_c = Z - a(a^\top Z), \qquad V_c = V - a(a^\top V). \tag{83}$$

We then apply the preceding projection, lifting and polar-correction formulas to the centred matrices with $r = 3$. Polar correction preserves centring because it only multiplies the coordinates on the right. Consequently, the molecular geometry requires atomwise reductions and 3×3 matrix solves, rather than a Jacobian over all atomic coordinates. These operations are orthogonal in the normalised coordinates. Training uses the differentiated polar path; sampling combines the learned tangent velocity with the analytic Gram lift and corrects each step to the next scheduled Gram matrix.

C.4 Computational Complexity

Let d be the ambient dimension, k the number of independent constraints, and q the number of LSMR iterations. Denote the costs of a Jacobian–vector product and a transpose-Jacobian product by C_J and $C_{J^\top}$. One projection or lift costs

$$O(q(C_J + C_{J^\top} + d + k))\,. \tag{84}$$

LSMR requires $O(d + k)$ working storage, excluding automatic-differentiation storage. Its cost depends on both the operator evaluations and convergence; matrix-free computation alone does not guarantee linear-time complexity. A retraction taking p nonlinear iterations requires p linear solves, together with constraint and line-search evaluations.

The specialised methods remove the dependence on iterative linear-solver convergence. For a PDE grid with N_x spatial points and N_s physical-time slices, constructing and applying the reduced Jacobian costs $O(N_x N_s)$, while factoring and solving its tridiagonal Gram system costs $O(N_s)$. The direct Burgers correction and sequential reaction–diffusion correction also cost $O(N_x N_s)$, excluding nonlinear fallbacks.

Method	Computational Complexity	Memory Complexity
Dense Jacobian and Gram	$O(dk^2 + k^3)$	$O(dk + k^2)$
Matrix-free LSMR	$O(q(C_J + C_{J^\top} + d + k))$	$O(d + k)$
PDE temporal structure	$O(N_x N_s)$	$O(N_x N_s)$
Gram structure	$O(nr^2 + r^3)$	$O(nr + r^2)$

Table 4: Operation costs per projection or Lift, including factorisation where applicable. Dense costs exclude Jacobian construction; LSMR storage excludes automatic-differentiation storage. Network costs are excluded throughout.

Method	GPU sampling (s)
Ambient FM	82.36
DI-FM	93.92
Stiefel-FM	118.44

Table 5: Recorded QM9 average sampling times for 2,560 candidates at 200 function evaluations, using a matched network architecture and one training seed. Chemistry scoring is excluded;

For an $n \times r$ matrix with a Gram constraint, forming the Gram matrix and applying its corrections costs $O(nr^2)$; the small eigendecomposition costs $O(r^3)$. Projection, lifting, polar correction and their directional derivatives therefore cost $O(nr^2 + r^3)$. With $r = 2$ for synthetic Stiefel matrices and $r = 3$ for molecules, these operations are linear in the number of rows. These bounds exclude path rank checks and exceptional waypoint construction.

All geometric costs are additional to network evaluation and differentiation. Factors can be reused across right-hand sides at the same point. During sampling, geometric operations performed at every step accumulate with the number of steps, so equal numbers of network evaluations need not imply equal runtime.

Molecule sampling times. To give an idea about the empirical aspects of computational cost: On QM9, generating 2,560 candidates with the same network architecture and 200 function evaluations takes 82.36 s for ambient FM, 83.76 s, and 93.92 s for DI-FM, and 118.44 s for Stiefel-FM. DI-FM therefore adds approximately 14% to ambient sampling time, or 12% relative to ambient FM. These measurements include sampling geometry but exclude chemistry scoring. This is a particularly favourable case because the linear algebra is on very small systems.

PDE sampling times. For the PDE experiments, the width-16 models at 32 sampling steps require 5.98 ± 2.66 ms per trajectory for CFFM and 11.22 ± 6.07 ms for DI on Burgers, and 7.19 ± 0.18 ms versus 9.09 ± 0.85 ms on reaction–diffusion. In summary, the exact computational cost of our method depends on the exact method being used.

D But why care about Singularity?

We examine how transport towards a singular target affects the learning problem. We will see how a Gaussian-to-Gaussian transport example shows how approaching a singular target can make flow matching harder to optimise. We then examine how target dimension affects the number of training updates needed to reach a fixed error.

D.1 A Gaussian Conditioning Example

Interpolant and velocity. Let U, Z be independent standard Gaussian variables. Consider the coupled endpoints $X_0 = (U, Z)$ and $X_1 = (U, 0)$, with linear interpolant

$$X_t = (U, (1-t)Z). \tag{85}$$

The target is a one-dimensional Gaussian supported on $\{(u, y) : y = 0\}$. For $t < 1$, writing $\varepsilon = 1 - t$, the optimal velocity satisfies

$$v_t^\star(u, y) = \left(0, -\frac{y}{\varepsilon}\right), \tag{86}$$

PDE	CFFM (small)	DI (small)
Burgers	5.98	11.22
Reaction–diffusion	7.19	9.09

Table 6: Average Sampling time in milliseconds per trajectory at 32 sampling steps.

Also the normal spatial derivative diverges as $t \to 1$. Along the interpolant, however, the normal velocity equals $-Z$ and retains unit variance.

Conditioning of the regression problem. At a fixed time, approximate the normal velocity by $v^{\perp}_{a,b}(u, y) = au + by$. Its population flow-matching loss is

$$\begin{aligned}\mathcal{L}_\varepsilon(a, b) &= \frac{1}{2}\mathbb{E}\big[(aU + b\varepsilon Z + Z)^2\big] \\ &= \frac{1}{2}\left[a^2 + (\varepsilon b + 1)^2\right]. \end{aligned} \tag{87}$$

The coefficient a describes dependence of the normal prediction on the tangential input. The minimiser and Hessian are

$$(a^\star, b^\star) = (0, -\varepsilon^{-1}), \qquad \nabla^2\mathcal{L}_\varepsilon = \mathrm{diag}(1, \varepsilon^2). \tag{88}$$

Consequently,

$$\kappa\big(\nabla^2\mathcal{L}_\varepsilon\big) = \varepsilon^{-2}. \tag{89}$$

Contraction of the normal input therefore produces a flatter loss in the coefficient b, even as the required coefficient grows.

Starting from $(a_0, b_0) = (1, 0)$, population gradient descent with learning rate η gives

$$\mathcal{L}_\varepsilon(a_n, b_n) = \frac{1}{2}\left[(1-\eta)^{2n} + (1-\eta\varepsilon^2)^{2n}\right]. \tag{90}$$

For fixed $0 < \eta < 1$, reducing the loss to a fixed tolerance below $1/2$ requires a number of updates proportional to ε^{-2} as $\varepsilon \to 0$.

Numerical illustration. Figure 4 compares two independently optimised problems at $t = 0$ and $t = 0.9$. Both use learning rate $\eta = 0.5$ and the same initial parameters, giving initial loss one. Their Hessian condition numbers are 1 and 100, respectively. Reaching loss 10^{-2} requires four updates at $t = 0$ and 391 updates at $t = 0.9$. The optimisation curves use exact population moments.

D.2 Dependence on Target Dimension

Next we explore how this phenomenon evolves with dimensions of the target distribution. We fix the ambient dimension at eight and vary the target dimension $r \in \{1, 2, 4, 6, 7, 8\}$, using

$$\rho_0 = \mathcal{N}(0, I_8), \qquad \rho_1^{(r)} = \mathcal{N}(0, I_r) \otimes \delta_0^{\otimes(8-r)}. \tag{91}$$

Flow matching uses independent endpoints and linear paths; diffusion uses cosine variance-preserving noising with noise prediction. Both use two 32-unit tanh hidden layers and Adam, with batch size 512 and 20,000 training updates. Training times are sampled uniformly from $[0.01, 0.99]$.

We measure mean squared error per coordinate against the analytic optimal predictor, excluding irreducible regression noise. Evaluation uses 8,192 held-out examples every 250 updates. Training cost is the first of three consecutive evaluations with error at most 0.02.

As target dimension increases from one to eight, the mean training cost in fig. 5 decreases from 5,083 to 2,000 updates for flow matching and from 3,583 to 667 for diffusion. Thus, lower-dimensional targets generally require more training in this experiment. The trends are compared within each method, since velocity and noise prediction have different target scales.

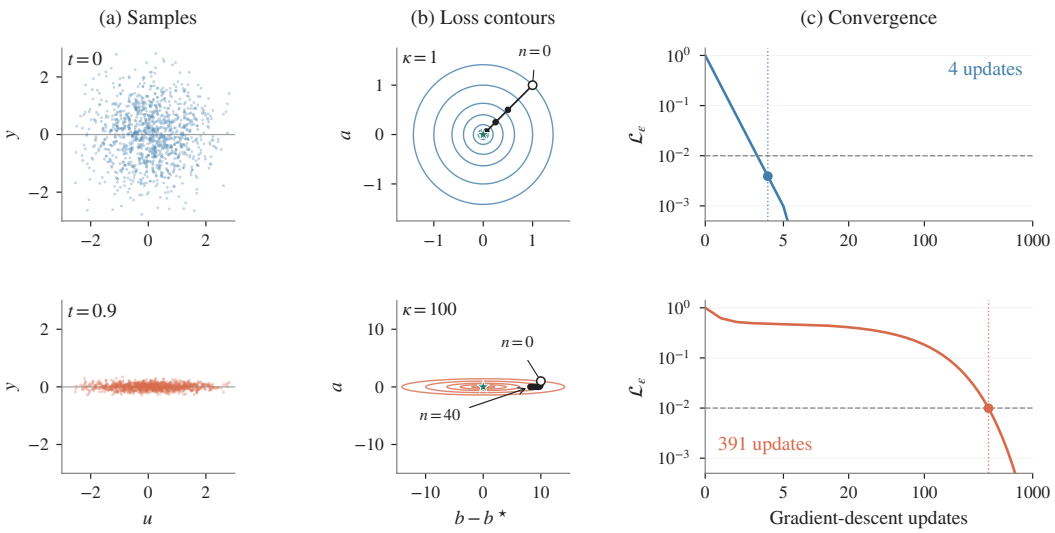


Figure 4: Contracting normal inputs worsen the conditioning of linear flow-matching regression. Rows correspond to $t = 0$ and $t = 0.9$. Loss contours are shown in coordinates $(b - b^\star, a)$. Gradient descent uses a common initialisation and learning rate; the dashed loss threshold is 10^{-2}.

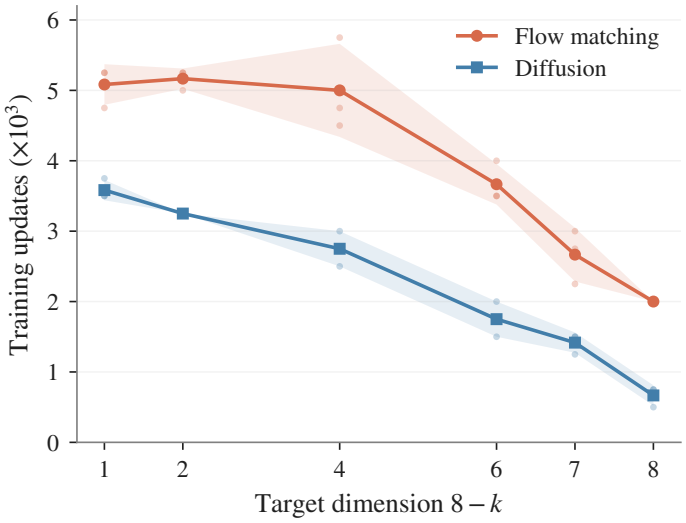


Figure 5: Training updates to reach excess prediction error 0.02 versus target dimension in $\mathbb{R}^8$. Means and sample standard deviations over three seeds.

D.3 Learning Efficiency on the Heat Equation

We compare DI small with CFFM small and CFFM large using 2,048 heat-equation trajectories on a 100×100 grid. Each method is trained for 3,000 updates over three seeds. Small models have width 16; the large model has width 32. Validation uses two conditions, with 64 generated and reference trajectories per condition and 32 sampling steps.

Figure 6 shows the learning efficiency of DI. DI first reaches MMSE and SMSE at most 0.01 at 1,000 updates in every seed. This requires 40% fewer updates than CFFM small and 25% fewer than CFFM large on average. At the end of training, DI small performs similarly to CFFM small, but it gets there faster. After 3,000 updates, large CFFM has the lowest errors (MMSE 1.02×10^{-3}, SMSE 1.62×10^{-3}), while DI small (1.61×10^{-3}, 1.69×10^{-3}) and CFFM small (1.45×10^{-3}, 1.70×10^{-3}) are comparable within seed variation.

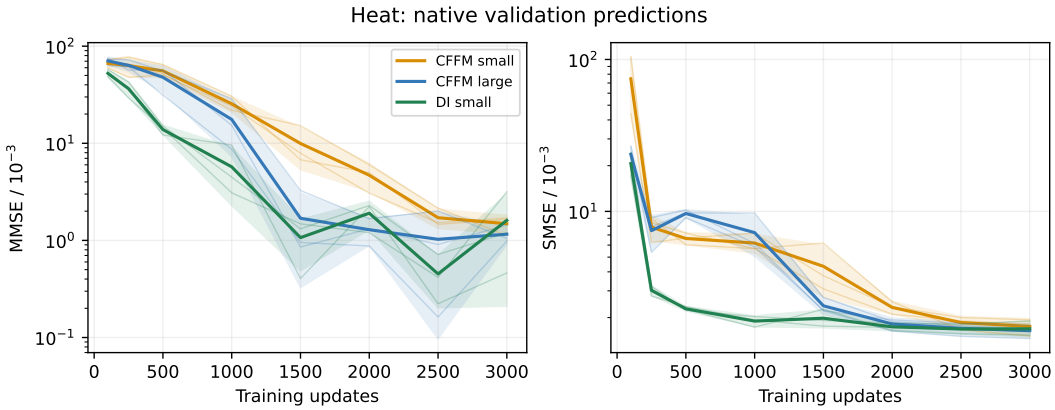


Figure 6: Heat-equation validation MMSE and SMSE versus training updates. Lines and bands show means and sample standard deviations over three seeds.

E Relation to Quotient Space Diffusion (Xu et al., 2026)

E.1 Relation to quotient-space diffusion

QSD construction. Consider the Euclidean setting of Xu et al. (2026), with a compact group G acting orthogonally, a G-invariant target ρ_1, and an independent Gaussian source. We work on the regular part of the action, where the quotient is a smooth manifold, and assume that the sampling processes remain there. Translations, when present, are removed by centring before applying this construction. QSD projects onto the orthogonal complement of the group-orbit tangent space and includes a geometric correction in its diffusion sampler.

We write the Gaussian interpolant as

$$X_t = \alpha_t Z + \eta_t X_1, \qquad Z \sim \mathcal{N}(0, I_d), \qquad Z \perp X_1, \tag{92}$$

$$(\alpha_0, \eta_0) = (1, 0), \qquad (\alpha_1, \eta_1) = (0, 1). \tag{93}$$

Here $\alpha_t, \eta_t > 0$ for $0 < t < 1$. Write $\mathsf{P}^{\mathrm{Q}}_x$ for the orthogonal projection onto $T_x(G \cdot x)^{\perp}$. A denoiser $D_\theta(t, x)$ predicts the clean endpoint through the projected loss

$$\mathcal{L}_{\mathrm{Q}}(\theta) = \mathbb{E}\left[w(t) \left\| \mathsf{P}^{\mathrm{Q}}_{X_t} \left(D_\theta(t, X_t) - X_1 \right) \right\|^2 \right], \tag{94}$$

where t is uniform on $(0, 1)$ and w is a positive weight. Training draws t, Z, X_1, constructs the ambient interpolant, and projects the prediction error. No endpoint alignment or retraction is required for this objective.

Define

$$A_t = \frac{\dot{\alpha}_t}{\alpha_t}, \qquad C_t = \dot{\eta}_t - A_t \eta_t. \tag{95}$$

The corresponding velocity and ambient-score estimates are

$$\hat{b}_\theta(t, x) = A_t x + C_t D_\theta(t, x), \tag{96}$$

$$\hat{s}^{\mathrm{amb}}_\theta(t, x) = \frac{\eta_t D_\theta(t, x) - x}{\alpha_t^2}. \tag{97}$$

Let $V(x)$ denote the Riemannian volume of the orbit through x, and set

$$k(x) = \mathsf{P}^{\mathrm{Q}}_x \nabla \log V(x). \tag{98}$$

With the trace convention for mean curvature, k is the negative of QSD's lifted orbit mean-curvature vector. Its Euclidean sampler is therefore the Itô SDE

$$d\widehat{X}_t = \left[\mathsf{P}^{\mathrm{Q}}_{\widehat{X}_t} \left(\hat{b}_\theta + \varepsilon_t \hat{s}^{\mathrm{amb}}_\theta \right)(t, \widehat{X}_t) + \varepsilon_t k(\widehat{X}_t) \right] dt + \sqrt{2\varepsilon_t}\, \mathsf{P}^{\mathrm{Q}}_{\widehat{X}_t}\, dW_t, \qquad \widehat{X}_0 \sim \mathcal{N}(0, I_d). \tag{99}$$

The choice $\varepsilon_t = 0$ gives the projected ODE. The term k accounts for variation in orbit volume; it is required for the diffusion construction.

DSI specialization. Suppose, additionally, that an equivariant structure map $\mathbf{c}$ selects the symmetry coordinate and has fibres orthogonal to the group orbits:

$$\ker D\mathbf{c}(x) = T_x(G \cdot x)^{\perp}. \tag{100}$$

Assume these fibres are locally affine on the regular set under consideration. This condition holds for radial rays under planar rotations. The DSI connection is then orthogonal, and its fibre projector satisfies $\mathsf{P}_x = \mathsf{P}_x^{\mathrm{Q}}$.

Choose the *same ambient interpolant* and prescribe

$$B_t = \mathrm{id}, \qquad \nu_t = \mathrm{Unif}(\mathcal{Y}), \qquad h_t = 0, \tag{101}$$

where $\mathcal{Y}$ is the homogeneous space of symmetry coordinates. Invariance of X_t gives the stated base law. Individual interpolant paths may change their symmetry coordinate.

To check separability, let $D_*(t,x) = \mathbb{E}[X_1 \mid X_t = x]$. Gaussian integration by parts gives

$$s_t^{\mathrm{amb}}(x) = \frac{\eta_t D_*(t,x) - x}{\alpha_t^2}, \qquad b_t(x) = A_t x + C_t D_*(t,x). \tag{102}$$

The ambient score is orthogonal to the orbits because the law is invariant. The vector x is also orthogonal to the orbits because the action is orthogonal. Consequently D_* and b_t lie in the retained space:

$$\mathsf{P}_x b_t(x) = b_t(x), \qquad D\mathbf{c}(x) b_t(x) = 0. \tag{103}$$

Thus the flux-separability condition holds with the identity base schedule, although pathwise separability need not hold.

Equality of the training objectives. Parameterize the learned fibre velocity by

$$\hat{u}_\theta(t,x) = \mathsf{P}_x \hat{b}_\theta(t,x). \tag{104}$$

The interpolant satisfies the samplewise identity

$$\dot{X}_t = A_t X_t + C_t X_1. \tag{105}$$

Therefore, wherever $C_t \neq 0$, the DSI velocity loss with weight $w(t)/C_t^2$ is exactly

$$\begin{aligned} \mathcal{L}_u(\theta) &= \mathbb{E}\left[\frac{w(t)}{C_t^2}\left\|\hat{u}_\theta(t,X_t) - \mathsf{P}_{X_t}\dot{X}_t\right\|^2\right] \\ &= \mathbb{E}\Big[w(t)\left\|\mathsf{P}_{X_t}\big(D_\theta(t,X_t) - X_1\big)\right\|^2\Big] = \mathcal{L}_{\mathrm{Q}}(\theta). \end{aligned} \tag{106}$$

The equality holds for every network parameter, so it also gives identical stochastic gradients when the same training samples are used.

Equality of the diffusion samplers. The DSI score is defined relative to Riemannian volume on each fibre. Disintegrating ambient volume over the symmetry coordinate introduces the orbit-volume factor. Equivalently, a tangent field w satisfies

$$\mathrm{div}\, w = \mathrm{div}_{\mathcal{M}} w + k \cdot w. \tag{107}$$

Applying the ambient score identity to this equation gives

$$s_t = \mathsf{P} s_t^{\mathrm{amb}} + k. \tag{108}$$

Thus the geometric correction appears automatically when the ambient score is converted into the fibre score. Use the corresponding estimate

$$\hat{s}_\theta = \mathsf{P}\hat{s}_\theta^{\mathrm{amb}} + k. \tag{109}$$

Substitution into the DSI sampler yields

$$\begin{aligned} d\widehat{X}_t = &\left[\mathsf{P}\big(\hat{b}_\theta + \varepsilon_t \hat{s}_\theta^{\mathrm{amb}}\big) + \varepsilon_t k\right](t,\widehat{X}_t)\, dt \\ &+ \sqrt{2\varepsilon_t}\, \mathsf{P}_{\widehat{X}_t} \circ dW_t. \end{aligned} \tag{110}$$

Since the fibres are locally affine, the Stratonovich-to-Itô correction for this projected noise vanishes. This is exactly the QSD sampler in (99).

Exact equivalence under an orthogonal fibre decomposition

With the same Gaussian interpolant, denoiser parameterization and time weighting, DSI with an identity base schedule has the same training loss as QSD. Its fibre score includes the orbit-volume correction, giving the same ODE and SDE. The equality holds for every learned denoiser; using the same numerical scheme, initial samples and Brownian increments therefore gives the same numerical updates.

For planar rotations, this correspondence is explicit:

$$\mathbf{c}(x) = \frac{x}{\|x\|}, \qquad \mathcal{M}_y = \{ry : r > 0\}, \tag{111}$$

$$\mathsf{P}_x = \frac{xx^\top}{\|x\|^2}, \qquad V(x) = 2\pi\|x\|, \qquad k(x) = \frac{x}{\|x\|^2}. \tag{112}$$

The prescribed base is the uniform direction, and the learned dynamics evolve the radius. The additional fibre-score term is $1/r$ in the radial direction, exactly matching QSD's orbit-volume correction.

The orthogonal fibre decomposition is an additional restriction. General QSD horizontal spaces need not be integrable, so they need not be level-set tangent spaces of any structure map. The equivalence established here therefore applies to symmetry actions admitting the stated decomposition; it does not identify an arbitrary angle-based DSI construction with QSD.

E.2 Joint symmetry and constraints: a planar chain

This experiment tests whether one structure map can exploit rotational symmetry while enforcing geometric constraints. We generate flexible planar chains with fixed bond lengths and a non-uniform distribution of shapes. The task separates two requirements: generated chains must satisfy the prescribed geometry, and their bend angles must follow the target distribution. All Cartesian models use the same rotation-equivariant architecture, allowing us to examine the effect of prescribing structural dynamics when equivariance is already supplied.

Target distribution. Each configuration consists of eight points $x = (p_1, \dots, p_8) \in \mathbb{R}^{16}$. The seven bonds have unit length, and the centroid is zero. We construct the shape from six bend angles. A mixture component is drawn with probabilities $(0.5, 0.3, 0.2)$, after which the angles are sampled independently from von Mises distributions with concentration κ and component means

$$\mu_{\mathrm{C}} = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5), \tag{113}$$

$$\mu_{\mathrm{S}} = (0.6, 0.6, 0.6, -0.6, -0.6, -0.6), \tag{114}$$

$$\mu_{\mathrm{Z}} = (0.9, -0.9, 0.9, -0.9, 0.9, -0.9). \tag{115}$$

These components favour curved, S-shaped and zigzag chains. An independent uniform first-bond direction determines the global orientation; the chain is then assembled and centred. We use $\kappa = 20$ for the full comparison and $\kappa = 100$ for a more concentrated target. Each concentration has 200,000 training examples, two independent reference sets of 20,000 examples, and a separate validation set of 4,096 examples.

Disintegrated construction. Write $b_i = p_{i+1} - p_i$, let $\bar{p}$ denote the centroid, and let $\theta(x)$ denote the first-bond direction, interpreted modulo 2π. Full DI uses

$$\mathbf{c}(x) = \big(\|b_1\|^2 - 1, \dots, \|b_7\|^2 - 1, \bar{p}, \theta(x)\big), \tag{116}$$

$$B_t(r, z, \theta) = \big((1-t)r, (1-t)z, \theta\big). \tag{117}$$

Here $r \in \mathbb{R}^7$ contains the bond residuals and $z \in \mathbb{R}^2$ is the centroid. The target pushforward is the product of a point mass at zero in the nine constraint coordinates and the uniform orientation law. Starting from an ambient Gaussian, the schedule drives the bonds and centroid to their target values while keeping the initial orientation fixed. The learned field acts within the resulting six-dimensional fibres.

For each source–target pair, we rotate the target to match the source's first-bond direction. The training path is constructed by 16 continuation stages: each stage moves towards the target and retracts onto

the corresponding scheduled fibre. Retraction uses minimum-norm Gauss–Newton updates with a trust radius of two and residual backtracking, with five updates per stage and 30 final updates. We differentiate through the complete construction and regress the projected path derivative. The implementation uses the structure map and its Jacobian without requiring bend-angle coordinates for the learned dynamics. Geometry is evaluated in float64.

Sampling uses 200 projected Euler steps. After each tangent update, eight Gauss–Newton iterations retract the sample onto the next scheduled fibre; a final 30-iteration retraction completes the sampler. The reported constraint errors therefore describe the output of this numerical procedure.

Baselines. Ambient flow matching (FM) learns the full velocity along straight Cartesian interpolants from an ambient Gaussian. FM + PCFM (Utkarsh et al., 2025) uses the same trained model and imposes the bond and centroid constraints during sampling. Its endpoint projection uses five Newton iterations, followed by ten relaxed correction steps of size 0.01, and a terminal projection. The penalty is selected separately for each seed and checkpoint from $\{0, 0.001, 0.01, 0.1\}$ using validation MMD.

The baseline labelled **QSD** is our flow-matching ODE implementation of the quotient-space projection principle of Xu et al. (2026). It starts from a centred Gaussian, aligns target configurations to their sources by planar Kabsch rotation, and projects both the training residual and sampling velocity orthogonally to translation and rotation directions. It does not impose bond constraints. This is a task-specific ODE adaptation of QSD rather than a reproduction of its molecular diffusion experiments.

DI (bonds) prescribes the bond and centroid schedules but omits orientation from the structure map. It uses Kabsch alignment, whereas DI (full) uses first-bond alignment. Their comparison therefore includes the effects of both the structure map and the endpoint coupling. Finally, torus FM learns directly in the six bend angles, using a uniform source on the torus and shortest wrapped linear interpolants. Samples are converted into centred unit-bond chains with an independent uniform orientation. This baseline measures what can be achieved when an explicit parameterization is available.

All methods use minibatch optimal-transport permutations with batches of 64. The matching cost is Cartesian distance for ambient FM, distance after optimal rotation for QSD and DI (bonds), distance in the first-bond frame for DI (full), and squared wrapped angular distance for torus FM. PCFM inherits the ambient model's coupling.

Training and evaluation. Cartesian models share an MLP with four hidden layers of width 256, SiLU activations and 32 sinusoidal time features, giving 214,032 parameters. The equivariant wrapper rotates the input into its first-bond frame and rotates the predicted velocity back. Torus FM uses periodic angle features and width 258, giving 213,630 parameters.

Each seed uses an independently generated bank of 262,144 training tuples containing interpolation times, states and regression targets. For DI, caching these tuples avoids repeatedly differentiating through the retraction solver. Training uses float32 network parameters, batches of 512 and Adam for 20,000 updates. The learning rate warms up for 200 updates to 3×10^{-4}, then follows cosine decay to 3×10^{-5}. We clip the global gradient norm at one and evaluate exponential-moving-average parameters with decay 0.999. A mixture of uniform and target-norm-based sampling is corrected by reciprocal probability weights, preserving the empirical squared-error objective in expectation.

We evaluate checkpoints at 2,000, 5,000 and 20,000 updates over three seeds. Each evaluation generates 5,000 samples using 200 network calls. Shape discrepancy is measured by unbiased squared MMD with an RBF kernel on the periodic bend-angle features $(\cos\phi, \sin\phi)$. The bandwidth is fixed from the reference data. Comparing independent reference samples gives MMD estimates of 7.05×10^{-6} and 4.17×10^{-5} for $\kappa = 20$ and 100, respectively; these indicate the scale of sampling fluctuations, rather than a strict lower bound.

We also measure mode-weight total variation by assigning each sample to its nearest angular prototype. For geometric accuracy, each sample receives the maximum absolute bond-length error over its seven bonds; we report the median and 99th percentile across samples. Orientation diagnostics comprise the Kolmogorov–Smirnov statistic against a uniform angle and the largest empirical cross-moment between centred shape features and the complex orientation $e^{i\theta}$. Metrics use sampler outputs without further repair. Timing includes geometric operations and excludes compilation.

κ	Method	2,000 updates ↓	5,000 updates ↓	20,000 updates ↓
20	Ambient FM	1322 ± 60	12.21 ± 1.20	4.08 ± 1.12
	FM + PCFM	714 ± 140	9.77 ± 1.60	15.66 ± 3.24
	QSD	997 ± 30	8.06 ± 1.40	3.13 ± 0.73
	DI (bonds)	523 ± 120	4.49 ± 2.70	1.25 ± 1.49
	DI (full)	$\mathbf{458 \pm 91}$	1.84 ± 0.48	-0.018 ± 0.197
	Torus FM	824 ± 45	$\mathbf{1.08 \pm 0.44}$	$\mathbf{-0.138 \pm 0.595}$
100	Ambient FM	876 ± 76	10.94 ± 1.90	3.09 ± 2.88
	FM + PCFM	327 ± 68	2.03 ± 0.89	0.68 ± 1.36
	DI (full)	$\mathbf{194 \pm 83}$	$\mathbf{0.71 \pm 1.50}$	$\mathbf{0.20 \pm 0.81}$

Table 7: Shape MMD^2 multiplied by 10^4, reported as mean $\pm$ sample standard deviation over three seeds. Bold marks the lowest mean within each concentration and budget. Differences between estimates near zero should be interpreted relative to their uncertainty.

κ	Method	Median bond error ↓	99th-percentile bond error ↓	Seconds / 1,000 samples ↓
20	Ambient FM	$(5.85 \pm 0.13) \times 10^{-2}$	0.262 ± 0.004	0.0367 ± 0.0016
	FM + PCFM	$(2.22 \pm 0) \times 10^{-16}$	$(3.33 \pm 0) \times 10^{-16}$	0.1973 ± 0.0154
	QSD	$(4.34 \pm 0.14) \times 10^{-2}$	0.215 ± 0.016	0.0292 ± 0.0148
	DI (bonds)	$(1.11 \pm 0) \times 10^{-16}$	$(2.22 \pm 0) \times 10^{-16}$	0.2765 ± 0.0248
	DI (full)	$(1.11 \pm 0) \times 10^{-16}$	$(2.22 \pm 0) \times 10^{-16}$	0.2779 ± 0.0002
	Torus FM	$(2.22 \pm 0) \times 10^{-16}$	$(4.44 \pm 0) \times 10^{-16}$	0.0228 ± 0.00004
100	Ambient FM	$(3.63 \pm 0.06) \times 10^{-2}$	0.168 ± 0.034	0.0189 ± 0.0001
	FM + PCFM	$(2.22 \pm 0) \times 10^{-16}$	$(3.33 \pm 0) \times 10^{-16}$	0.1786 ± 0.0007
	DI (full)	$(1.11 \pm 0) \times 10^{-16}$	$(2.22 \pm 0) \times 10^{-16}$	0.2774 ± 0.0007

Table 8: Results after 20,000 training updates, using 200 sampling steps. Entries are mean $\pm$ sample standard deviation over three seeds. Bond-error differences at machine precision have no practical significance.

Shape accuracy across training budgets. Full DI has lower mean MMD than ambient FM and QSD at every reported budget for $\kappa = 20$ (table 7). At 5,000 updates, its mean discrepancy is approximately 6.6 times smaller than ambient FM and 4.4 times smaller than QSD. By 20,000 updates, full DI and torus FM both reach the scale of the reference-sample fluctuations. The negative estimates are permitted by the unbiased MMD estimator and do not indicate negative population discrepancy.

DI (bonds) also improves on ambient FM and QSD, showing that prescribing the geometric constraints accounts for a substantial part of the improvement. Full DI has lower mean MMD than DI (bonds), although the experiment does not isolate the orientation schedule from the change in coupling. At $\kappa = 100$, full DI again has the lowest mean MMD among the evaluated methods. Its final discrepancy overlaps that of PCFM within the reported standard deviations.

Constraint accuracy and computational cost. Both DI variants, PCFM and torus FM produce bond errors at float64 precision (table 8). QSD improves the bond errors over ambient FM, but removing rigid-body motion does not enforce bond lengths. Its centred source and projected velocity do preserve the centroid to numerical precision.

PCFM illustrates why constraint accuracy and distributional accuracy must be assessed separately. It satisfies the bonds in both settings, yet at $\kappa = 20$ its final shape MMD exceeds that of the ambient model it corrects. At $\kappa = 100$, it substantially improves the ambient model's mean MMD. The measurements therefore show that sampling-time correction can affect shape fit differently across targets.

The additional orientation diagnostics do not show a consistent advantage for full DI, as expected from the shared equivariant architecture. At $\kappa = 20$, final mean mode-weight errors range from 0.0071 to 0.0127, and angle KS statistics from 0.0084 to 0.0173. Full DI's angle statistic is unchanged across training budgets because its sampler preserves the source orientation.

Full DI approaches the shape accuracy of the explicit torus model while working with implicit constraints, but its retractions make sampling approximately twelve times slower at the final $\kappa = 20$ checkpoint. Its improvement at matched training updates establishes an advantage in shape learning under this protocol; it does not establish a reduction in total training time.

F Experimental Details

F.1 PDE Trajectory Generation

Data and tasks. We closely follow the PCFM experimental setup (Utkarsh et al., 2025) for the PDE definitions, conditioning tasks, Fourier neural operator architecture and Gaussian source prior. We implement the experiments in JAX using finite datasets generated from the released simulation procedures. We consider Burgers trajectories conditioned on boundary data and reaction–diffusion trajectories conditioned on their initial field. The models generate complete space–time fields; physical time is distinct from the generative interpolation time.

For Burgers, we solve the inviscid equation $\partial_s u + \partial_x(u^2/2) = 0$ on $[0,1] \times [0,1]$, using 101 spatial points and 101 physical-time points. The initial field is $u(x,0) = [1 + \exp((x-\ell)/0.02)]^{-1}$, where ℓ is uniform on $[0.2, 0.8]$. The observed condition is the left boundary value b, drawn uniformly from $[0,1]$ and imposed after the initial time; the right boundary has zero gradient. We use the released first-order Godunov update. Conditional generation averages over the unobserved initial location ℓ.

For reaction–diffusion, we solve $\partial_s u = 0.005\,\partial_{xx} u + 0.01\,u(1-u)$ on 128 spatial cell centres in $[0,1]$, recording 100 times $s_j = j/100$. Initial fields follow the released random sine-series construction, with random amplitudes and phases, optional absolute-value and window transformations, and rescaling to $[0,1]$. The observed condition is the entire initial field. Left and right diffusive boundary fluxes are drawn from $[0, 0.05]$ and $[-0.05, 0]$, respectively, rounded to 0.001, and remain unobserved. Trajectories are generated in float64 using the released logistic reaction update and a two-stage diffusion update.

Each task contains 2,048 training trajectories from 64 conditions with 32 realizations each, 128 validation trajectories from two conditions, and 512 test trajectories from eight conditions. Validation and test conditions each have 64 reference trajectories. Conditions are disjoint across splits, and fields remain in physical units without fitted normalisation.

Methods and architectures. We compare disintegrated interpolants (DI) with unconditional functional flow matching (FFM), conditional FFM (CFFM) (Kerrigan et al., 2024), and PCFM applied to the unconditional FFM checkpoints. CFFM and DI receive two additional channels containing observed values and a binary observation mask. Hidden initial locations and boundary fluxes are never supplied to the networks.

The structure map combines the observed initial or boundary conditions with integrated conservation residuals. For Burgers, these impose the left boundary, the zero-gradient right boundary, and the mass balance determined by the boundary flux $u^2/2$. For reaction–diffusion, they impose the initial field and the mass balance including reaction and boundary diffusion fluxes. Spatial mass uses a rectangular grid sum, and physical-time integrals use cumulative trapezoidal quadrature. Reaction–diffusion boundary fluxes are estimated from the field using fourth-order one-sided derivative stencils. These constraints do not impose the local PDE at every grid point.

FFM and CFFM use independent source–target pairs and straight-line interpolants. DI prescribes $\mathbf{c}(X_t) = (1-t)\mathbf{c}(X_0)$, correcting the ambient chord along fixed directions to obtain the training path. Its velocity is obtained by differentiating the defining constraint equation, and training regresses the orthogonal tangent component. Burgers corrections use interior mass shifts that preserve the boundary conditions; reaction–diffusion corrections use spatially constant shifts with a sequential quadratic solve. Normal lifts use a structured tridiagonal solve after eliminating affine constraints. CFFM and DI share constraint-corrected training targets, whereas unconditional FFM uses raw trajectories pooled over conditions. All methods are evaluated against raw simulator trajectories.

Every network is a two-dimensional Fourier neural operator with four spectral blocks, 32×32 retained modes, a flow-time embedding of width 32, and lifting and projection hidden widths of 256. We use GELU activations. Small CFFM and DI have channel width 16 and 1,133,649 parameters

each. Large CFFM has width 32 and 4,487,329 parameters; unconditional FFM has width 32 and 4,486,817 parameters. PCFM introduces no additional learned parameters. The source is a zero-mean, unit-variance Matérn-$1/2$ Gaussian field with length scale 0.001 on normalised space–time coordinates, sampled through a positive block-circulant FFT embedding.

Training and hyperparameters. Models are trained for 20,000 updates with batch size eight and seeds 42, 43 and 44. We use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical stabiliser 10^{-8}, no weight decay, and global gradient clipping at norm 100. The learning rate starts at 3×10^{-4}, without warmup, and is halved after more than ten validation checks without improvement, down to 10^{-4}. Generative times are sampled uniformly, and regression losses are averaged over all field entries. Model computation uses float32 without exponential moving averages.

Production validation is performed every 1,000 updates using 32 sampling steps. Checkpoints minimise validation MMSE plus SMSE, evaluated after correction for CFFM and DI and on native outputs for FFM. PCFM inherits the selected FFM checkpoint. Runs continued from earlier checkpoints retain their optimiser state and count those updates towards the 20,000-update budget.

Sampling and evaluation. Main results use 32 sampling steps. FFM, CFFM and DI use forward Euler; DI additionally enforces the prescribed residual derivative during integration.

For both PDEs, PCFM uses one damped Newton endpoint correction per sampling step, with ridge 10^{-6}, followed by interpolation from the original source; additional interpolation guidance is disabled. Its terminal correction uses an anchored Newton iteration with at most 32 iterations and tolerance 10^{-5}. The reported float64 variant retries failed terminal corrections on the retained samples in double precision. Generation is not repeated, and samples are neither replaced nor filtered.

For each test condition, we compare 64 generated trajectories with 64 reference trajectories. MMSE is the mean squared difference between their pointwise ensemble means, and SMSE is the mean squared difference between their pointwise ensemble standard deviations, using sample standard deviations within each ensemble. Both errors are averaged over the space–time grid. Conservation error (CE) is the ensemble mean of the Euclidean norm of the integrated conservation residual vector, without normalisation by the number of constraints. Metrics are averaged equally over the eight test conditions within each seed, then reported as the mean and sample standard deviation across the three training seeds. MMSE and SMSE, including their reported standard deviations, are multiplied by 10^3; CE remains unscaled.

Conditioning and constraints. We use the conditioning tasks and discrete conservation residuals from the PCFM setup (Utkarsh et al., 2025). Write $U_{i,j}$ for a trajectory evaluated at spatial point x_i and physical time s_j. CFFM and DI receive the observations through two additional input channels: a field containing the observed values and zeros elsewhere, and a binary mask identifying their locations.

For **Burgers**, the observation is the constant left boundary value b. The initial step location remains unobserved, so the task is to generate different trajectories compatible with the same boundary condition. The observation channels contain b at the left boundary for physical-time indices $j \geq 1$. We impose

$$U_{0,j} - b = 0, \qquad U_{N_x-1,j} - U_{N_x-2,j} = 0, \qquad j = 1, \ldots, N_s - 1. \tag{118}$$

The second equality imposes the prescribed zero-gradient right boundary. Neither boundary condition is imposed at the initial time, where the simulator retains the sampled initial field. We additionally enforce the integrated mass balance

$$M_j - M_0 + \mathcal{T}_j\left[\frac{U^2_{N_x-1,\cdot} - U^2_{0,\cdot}}{2}\right] = 0, \qquad j = 1, \ldots, N_s - 1, \tag{119}$$

where $M_j = \Delta x \sum_{i=0}^{N_x-1} U_{i,j}$ is the discrete mass, including both spatial endpoints, and $\mathcal{T}_j$ denotes cumulative trapezoidal integration up to physical time s_j:

$$\mathcal{T}_j[a] = \sum_{q=1}^{j} \frac{s_q - s_{q-1}}{2}(a_{q-1} + a_q). \tag{120}$$

The conservation residual uses the boundary values of the candidate trajectory, including when that trajectory does not yet satisfy the boundary conditions.

For **reaction–diffusion**, the observation is the full initial field y, and the observation channels are nonzero only at the initial physical-time slice. The boundary diffusion fluxes remain unobserved and vary between trajectories sharing the same initial field. We impose

$$U_{i,0} - y_i = 0, \qquad i = 0, \ldots, N_x - 1, \tag{121}$$

together with the integrated reaction–diffusion mass balance

$$M_j - M_0 - \mathcal{T}_j\left[0.01\,\Delta x \sum_{i=0}^{N_x-1} U_{i,\cdot}(1 - U_{i,\cdot})\right] - \mathcal{T}_j[F_L(U) - F_R(U)] = 0, \qquad j = 1, \ldots, N_s - 1. \tag{122}$$

Here M_j is the same rectangular spatial sum, now evaluated on the reaction–diffusion grid. The fluxes $F_L(U)$ and $F_R(U)$ approximate $-0.005\,\partial_x u$ at the left and right edges using fourth-order one-sided five-point derivative stencils. They are computed from the candidate trajectory; the hidden flux values used to generate the reference data are not supplied to the model or constraint calculation.

For each task, the structure map $\mathbf{c}$ concatenates the initial or boundary residuals with the conservation residuals above, with the observed condition held fixed. DI prescribes $\mathbf{c}(X_t) = (1-t)\mathbf{c}(X_0)$, where t is generative time, and learns the tangent component of the transport. CFFM receives the same observation channels and learns the full velocity. PCFM receives the observations through its sampling constraints, using an unconditional FFM network. No method in this comparison explicitly enforces the local PDE residual at every grid point. The reported CE measures only the integrated conservation residuals.

F.1.1 Baselines

We compare DI with unconditional functional flow matching (FFM) (Kerrigan et al., 2024), physics-constrained flow matching (PCFM) (Utkarsh et al., 2025), and conditional functional flow matching (CFFM). These comparisons address different questions: FFM measures performance without conditioning, PCFM imposes conditions during sampling, and CFFM learns the conditional distribution during training.

Unconditional FFM. FFM is trained on trajectories pooled across conditions, without receiving the initial or boundary observations. It uses straight-line interpolants and learns the full ambient velocity. Its samples therefore represent the marginal trajectory distribution. We include it to establish the effect of conditioning and to provide the pretrained model used by PCFM. The network has 4,486,817 parameters.

PCFM. PCFM applies physics-based constraints while sampling from a pretrained unconditional FFM model, without additional network training (Utkarsh et al., 2025). This permits the same model to be reused, in principle, with different observation values and constraint types, including initial or boundary conditions. We apply PCFM to our locally trained unconditional FFM models using the observation and integrated conservation constraints defined above.

Our implementation follows the published constraint formulation and sampling procedure, with a reduced training dataset and model size. For Burgers and reaction–diffusion, the FFM network has four Fourier layers, 32 modes per dimension, width 32 and 4,486,817 parameters. Each model is trained on 2,048 trajectories for 20,000 updates. Checkpoints are selected by the unconditional FFM validation MMSE plus SMSE and shared by FFM and PCFM; no separate checkpoint selection is performed for PCFM.

The main comparison uses 32 sampling steps. Each PCFM step performs one Newton correction with ridge 10^{-6}, with interpolation guidance disabled. The terminal correction uses at most 32 anchored Newton iterations and a maximum-residual tolerance of 10^{-5}. Failed terminal corrections are retried in float64 on the retained samples, without regenerating or replacing them. The appendix distinguishes the original evaluations from those requiring terminal repair.

PCFM exhibits large between-seed variation under this protocol. Table 2 reports a selected subset comprising seeds 43 and 44 for Burgers and seed 42 for reaction–diffusion, all of which completed without terminal repair. The omitted runs include substantially larger distributional errors and one terminal-correction failure subsequently recovered in float64. This subset therefore does not represent performance across all three seeds. Full per-seed results and repair outcomes are reported in section F.1.1.

These experiments evaluate PCFM under our data and model rather than reproduce its published benchmark. PCFM addresses a different setting from DI and CFFM: its advantage is the ability to condition an existing model without retraining. Its results contextualise the trade-off between post-hoc conditioning and training for a specified conditional task. CFFM remains the direct comparison with DI because both methods train for the specified conditioning task. Full per-seed results and terminal-repair outcomes are reported in table 9.

PDE	Seed	Selected update	MMSE/10^{-3}	SMSE/10^{-3}	CE (CL)
Burgers	42	2,000	22,635.51	24,036.96	5.26×10^{-6}
	43	15,000	69.52	68.62	7.24×10^{-6}
	44	5,000	21.68	95.17	5.91×10^{-6}
Reaction–diffusion	42	2,000	63.87	1,091.63	1.76×10^{-5}
	43†	13,000	1,407.32	111,475.53	2.51×10^{-5}
	44	1,000	373.26	9,322.69	1.08×10^{-5}

Table 9: Full per-seed PCFM results for validation-selected checkpoints at 32 sampling steps. Checkpoints are selected using the paired unconditional FFM validation MMSE+SMSE. CE (CL) denotes conservation error. † Reaction–diffusion seed 43 initially failed its terminal correction; the displayed metrics follow float64 terminal repair of the retained samples, without regeneration. All other runs completed without repair. The main table reports only Burgers seeds 43/44 and reaction–diffusion seed 42.

Conditional FFM. CFFM receives the observed values and their mask as network inputs. Each training trajectory is paired with its own observations, so the regression objective learns the corresponding conditional distribution. The network predicts the full ambient velocity, including the components that bring samples into agreement with the observations. We use a small model with 1,133,649 parameters and a large model with 4,487,329 parameters. The small model has the same architecture and parameter count as DI; the large model has approximately four times as many parameters.

CFFM is the direct comparison for assessing the benefit of the disintegrated construction. CFFM and DI share the conditioning inputs, constraint-corrected training targets. DI additionally prescribes the evolution of the constraint residual and learns only the tangent component of transport. Comparing against small CFFM assesses this change at fixed network capacity, while large CFFM tests whether additional capacity yields comparable accuracy. Both conditional methods accept unseen observation values for the task on which they were trained; changing the conditioning task requires retraining in our implementation.

F.2 Molecular Reconstruction

Data and task. We follow the moment-constrained structure elucidation task of Stiefel-FM (Cheng et al., 2025), generating all-atom molecular geometries from atom identities and mass-weighted moments. The molecular bond graph is not supplied. We use a reduced QM9 experiment with 25,000 training molecules, 256 validation molecules and 256 test molecules, sampled from the processed upstream splits. All methods use the same subsets and three training seeds, 24. Generation-based checkpoint selection uses 64 of the validation molecules. This experiment uses a smaller dataset and network than the full Stiefel-FM benchmark.

Conditioning and constraints. Coordinates are expressed in Å, centred at the centre of mass and represented in their principal-axis frame. For a molecule with N atoms, let $X \in \mathbb{R}^{N\times 3}$ contain its coordinates, m_i its atomic masses, and $\lambda_1, \lambda_2, \lambda_3$ the prescribed principal moments of the mass-weighted second-moment matrix. These are the eigenvalues of the planar dyadic matrix, rather than the inertia tensor. The constraints are

$$\sum_{i=1}^{N} m_i X_{i,\alpha} = 0, \qquad \alpha = 1, 2, 3, \tag{123}$$

$$\sum_{i=1}^{N} m_i X_{i,\alpha} X_{i,\beta} = \lambda_\alpha \delta_{\alpha\beta}, \qquad 1 \le \alpha \le \beta \le 3. \tag{124}$$

Thus the conditions specify the molecular composition and spatial mass distribution while leaving the geometry to be generated. Atom identities and masses enter through node features, and the three moments are supplied explicitly to every model.

For the geometric methods, we use the normalised coordinates

$$Z_{i,\alpha} = \sqrt{\frac{m_i}{\lambda_\alpha}} X_{i,\alpha}, \qquad a_i = \sqrt{\frac{m_i}{\sum_j m_j}}, \tag{125}$$

so the constraints become $a^\top Z = 0$ and $Z^\top Z = I_3$. Centre-of-mass constraints are enforced analytically. DI uses the Gram residual $Z^\top Z - I_3$ as its structure map and prescribes

$$Z_t^\top Z_t = (1-t) Z_0^\top Z_0 + t I_3. \tag{126}$$

The implementation requires at least five real atoms and strictly positive target moments.

F.2.1 Baselines

All baselines receive the same atom identities and moment conditions.

Ambient FM. Ambient FM learns the full Cartesian velocity using straight-line paths from a centre-of-mass-projected Gaussian source. Its velocity is projected to preserve the centre of mass, while agreement with the supplied moments is learned from paired examples. This is the conditional ambient baseline for assessing the benefit of the geometric construction.

Stiefel-FM. Stiefel-FM (Cheng et al., 2025) is the specialised intrinsic baseline. Its source already satisfies the moment constraints, and its training paths remain on the corresponding Stiefel manifold. We use the upstream three-column logarithm construction and canonical-metric regression objective, without optimal-transport matching, stochastic path noise or inference annealing. The baseline is trained locally on the same subset and with the same network capacity as DI.

F.2.2 Implementation and Evaluation

Disintegrated construction. DI-FM starts from a random mass-centred orthonormal frame Q_0, extended away from the constraint set by

$$Z_0 = Q_0 R, \qquad R = \exp\big(0.1(A + A^\top)\big), \qquad A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{127}$$

The training path is obtained by polar-normalising the source–target chord and applying the square root of the prescribed Gram matrix. Training regresses the tangent component of its velocity. A rank check uses 65 points along the chord and a spectral-norm bound between them; when the bound is below 10^{-4}, a waypoint is introduced while retaining both endpoints. Geometry is evaluated through atomwise reductions and 3×3 Gram and Sylvester solves, without constructing a dense atom-space constraint Jacobian.

Architecture and training. Every method uses the same sign-equivariant network with 6,688,719 trainable parameters. It has six blocks, hidden width 256, conditioning width 128, eight attention heads and feed-forward expansion factor four. Atom embeddings have width 32, edge features have width 64, and activations are SiLU. The model receives sinusoidal embeddings of generative time and the moments, together with the raw moment values.

Models are trained for 120,000 updates with batch size 64 using AdamW. The learning rate increases linearly to 10^{-4} over 2,000 updates and remains constant thereafter. Weight decay is 0.01, excluding embeddings, layer normalisations and biases. Adaptive gradient clipping uses a queue of 50 gradient norms, with threshold equal to 1.5 times their mean plus twice their population standard deviation. Exponential moving averages use decay 0.9995. The backbone runs in float32, with float64 used for selected geometric computations. Runs continue from their 40,000-update checkpoints while retaining optimiser and moving-average states.

Sampling and evaluation. All flow methods use 200 network evaluations. Ambient FM uses forward Euler. Stiefel-FM uses the upstream QR and matrix-exponential integrator. DI combines the learned tangent velocity with the prescribed Gram lift and retracts each step to the next scheduled Gram matrix.

We compare moving-average checkpoints at 40,000, 60,000, 80,000, 100,000 and 120,000 updates, generating ten candidates for each of 64 validation molecules. Selection first minimises failed candidates, then maximises the fraction of molecules with a candidate below 0.25 Å RMSD, and finally minimises mean best-of-ten RMSD, with ties resolved by the earlier checkpoint. Ambient FM selects 80,000 updates; DI-FM and Stiefel-FM select 120,000. Test evaluation generates ten candidates for each of the same 256 molecules.

Constraint error is the upstream moment RMSE: the Frobenius norm of the upper-triangular moment-matrix residual divided by three, corresponding to an elementwise mean over the full 3×3 array before taking the square root. Validity follows the upstream RDKit convention, which counts a candidate as valid when its inferred structure is connected and sanitizable, or when the reference itself is invalid. Diversity is the average pairwise RMSD among candidates for the same molecule. Stability is the mean base-10 logarithm of the energy-gradient norm reported by a GFN2-xTB single-point calculation; lower values indicate smaller gradients.

RMSD evaluation accounts for principal-axis sign reflections and permutations of atoms of the same element. Candidate-level metrics are averaged within each molecule and then across molecules. The study uses three training seeds.

F.3 STIEFEL EXPERIMENT.

Targets are obtained by sign-corrected QR decomposition of $A = M + E$, where E has independent standard Gaussian entries and $M = [I_4; 0] \in \mathbb{R}^{10\times4}$. Sources have independent $\mathcal{N}(0, 1)$ entries. The constraint is the upper-triangular part of $X^\top X - I_4$.

Model and training. All models use an MLP with three width-96 SiLU hidden layers, taking flattened coordinates and five time features. The diffusion model outputs both velocity and noise predictions. Training uses Adam with learning rate 8×10^{-4} and gradient clipping at norm 1.

DI-FM. Source-target chords are corrected towards the prescribed level $\mathbf{c}(X_t) = (1 - t)\mathbf{c}(X_0)$. Path velocities are estimated by finite differences with increment 2×10^{-3}; targets and predictions are tangent-projected. Projection and iterative correction use LSMR.

DI diffusion. We add tangent-projected Gaussian noise with amplitude $\gamma(t) = 0.35 \sin(\pi t)$ and correct back to the scheduled fibre. The model learns projected path velocity and noise using squared-error losses weighted 1 and $1/2$. Sampling uses projected drift and noise followed by fibre correction.

The default run uses three seeds, 150 updates per model, batch size 8, and 128 evaluation samples with 24 sampling steps. Distributional quality is measured by sliced Wasserstein distance over 64 fixed directions.