MaDeL: Manifold-Decomposed Feature Losses for Generative Modeling

Anonymous authors
Paper under double-blind review

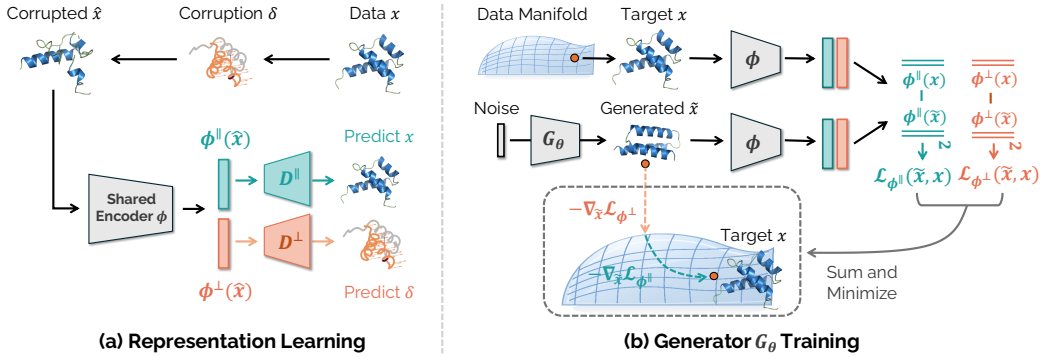


Figure 1: **Overview of MaDeL. (a)** A shared encoder splits a corrupted observation into manifold-parallel and manifold-orthogonal features, trained with $\boldsymbol{x}$ and $\boldsymbol{\delta}$-prediction. **(b)** These learned features define the decomposed generative loss. The resulting gradients produce complementary tangent and normal updates to the data manifold.

Abstract

Generative models are often trained with isotropic objectives such as mean-squared error. For data concentrated near a low-dimensional manifold, however, such losses conflate displacement along the manifold, which may represent valid variation, with displacement away from it, which produces invalid samples. This mismatch is especially problematic in sparse, highly constrained domains, where ambient-space regression can encourage off-manifold interpolation. *We ask whether a generative objective can distinguish manifold-parallel variation from manifold-orthogonal deviation directly from data, without explicitly estimating the manifold.* We introduce a manifold-decomposed feature loss (MaDeL) that learns complementary representations from corrupted observations: one is trained to recover the clean sample, while the other is trained to recover the corruption. We show that, under a feature bottleneck, their Jacobians align with the tangent and normal spaces, exactly for linear manifolds and locally for smooth manifolds. Together, these representations define an anisotropic objective that separately measures intrinsic variation and off-manifold deviation. Across synthetic, Earth and climate science, and torsion-angle benchmarks, MaDeL improves support recovery and average angular W_1 under single-step sampling; on protein backbones, it reduces steric clashes across one- and few-step sampling budgets.

1 Introduction

Generative models have achieved remarkable fidelity across images, video, molecules, and protein structures (Karras et al., 2022; Ho et al., 2022; Hoogeboom et al., 2022; Yim et al., 2023b). Despite substantial differences in architecture and formulation, many modern approaches share a common training principle: they learn a denoising, score, velocity, or reconstruction target using a largely isotropic regression objective. For example, diffusion and flow-based models commonly minimize an ℓ_2 discrepancy between predicted and target quantities (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2023). Such objectives measure *how far* a prediction is from its target, but are largely agnostic to *the direction in which the prediction is wrong*.

This isotropic treatment is poorly matched to the geometry of natural data. High-dimensional observations are hypothesized to concentrate near a lower-dimensional, structured subset of the ambient

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.

space (Tenenbaum et al., 2000; Narayanan & Mitter, 2010; Fefferman et al., 2016; Pope et al., 2021). Locally, perturbations of equal magnitude can have very different consequences depending on their orientation relative to the data manifold (Rifai et al., 2011a; Alain & Bengio, 2014). A displacement parallel to the manifold may correspond to meaningful variation, such as a change in object pose or a valid protein conformation, whereas an orthogonal displacement can produce perceptual artifacts or physically invalid structures. An isotropic loss cannot distinguish between them.

This distinction is especially important for generative modeling. Near the data manifold, a generator must both suppress deviations away from valid data and model variation along the manifold. The desired dynamics are therefore inherently anisotropic: *orthogonal directions should restore samples toward the manifold, whereas parallel directions should describe variation within it*. Recent analyses of diffusion models support this geometric perspective, showing that score functions recover directions associated with the normal bundle of the data manifold at low noise levels (Stanczuk et al., 2024). Ignoring this distinction encourages interpolation through off-manifold, low-density regions.

Ideally, a generative objective would therefore measure prediction errors according to the local geometry of the data distribution rather than a fixed isotropic metric. Explicitly estimating tangent and normal spaces for complex high-dimensional data, however, is challenging. Instead, we use learned feature spaces to induce data-dependent notions of similarity, building on their successful use as perceptual losses for image synthesis (Johnson et al., 2016; Zhang et al., 2018).

Concretely, given data $\boldsymbol{x}$, we construct a corrupted observation $\tilde{\boldsymbol{x}} = \boldsymbol{x}+\boldsymbol{\delta}$ and learn two complementary representations. *Manifold-parallel features* are learned by predicting $\boldsymbol{x}$ from $\tilde{\boldsymbol{x}}$ ($\boldsymbol{x}$*-prediction*), encouraging them to preserve intrinsic variation while suppressing corruption. *Manifold-orthogonal features* are learned by predicting $\boldsymbol{\delta}$ from the same observation ($\boldsymbol{\delta}$*-prediction*), encouraging them to capture deviations away from the manifold. Together, these representations provide a learned decomposition of errors into manifold-parallel and orthogonal components.

We use these representations to define a geometry-aware training objective for generative models: rather than penalizing all prediction errors uniformly, it measures them through feature spaces specialized for intrinsic variation and off-manifold deviation, and thus distinguishes predictions that are equally distant in ambient space but differ in their compatibility with the data distribution (Fig. 1).

Specifically, our contributions are:

- **Theoretical characterization of learned geometry (Secs. 4.1 and 4.2).** We show that, under suitable bottlenecks, optimal $\boldsymbol{x}$-prediction and $\boldsymbol{\delta}$-prediction encoders have Jacobian row spaces contained in the tangent and normal spaces of the data manifold, respectively, exactly for linear manifolds and to leading order locally for smooth manifolds. Their feature losses yield complementary manifold-parallel and manifold-orthogonal gradients.
- **Manifold-decomposed generative loss (Sec. 4.3).** We introduce MaDeL, a geometry-aware feature loss that separately measures manifold-parallel variation and manifold-orthogonal deviation using these representations, without access to the data manifold.
- **Validation across diverse manifolds (Sec. 5).** We evaluate MaDeL on a helix, Earth and climate science data, protein and RNA torsion angles, and protein backbone generation. MaDeL improves support recovery and average angular W_1 on the single-step benchmarks and reduces steric clashes in protein backbones across one- and few-step budgets.

2 Background

Isotropic regression objectives. Let $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ and $\boldsymbol{x}_{\mathrm{targ}}$ denote a model prediction and its regression target. Diffusion and Consistency Models (Song et al., 2023; Song & Dhariwal, 2024), for example, are commonly trained with isotropic ambient-space objectives of the form

$$\mathcal{L}_{\mathrm{iso}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}[\rho(\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}} - \boldsymbol{x}_{\mathrm{targ}}\|_2^2)], \tag{1}$$

where $\rho(u) = u$ yields mean-squared error (MSE) and $\rho(u) = \sqrt{u + c^2} - c$ yields the pseudo-Huber loss. Although these methods construct $\boldsymbol{x}_{\mathrm{targ}}$ differently, their losses depend only on the residual magnitude and therefore penalize equal-magnitude errors identically regardless of direction.

Geometry-aware generative modeling. Natural data often concentrate near low-dimensional subsets of the ambient space, making their local geometry anisotropic: perturbations of equal mag-

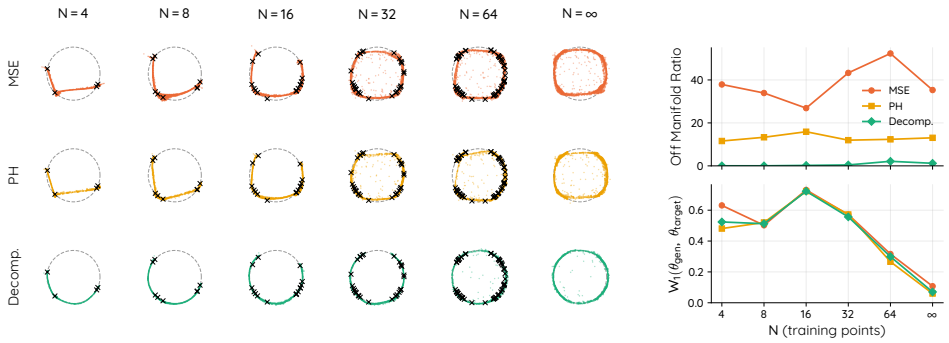


Figure 2: **A manifold-decomposed metric mitigates off-manifold interpolation.** With sparse training samples (crosses), Cartesian MSE and Pseudo-Huber interpolate through the circle, whereas the oracle decomposed metric preserves its support.

nitude can represent meaningful variation when tangent to the data manifold but invalid deviation when normal to it, a distinction ignored by isotropic regression.

Tab. 1 summarizes how prior work accounts for this geometry. Existing approaches primarily incorporate the geometry of a known state space: Riemannian diffusion models (Huang et al., 2022) and Riemannian Flow Matching (RFM) (Chen & Lipman, 2024) extend generative dynamics to non-Euclidean manifolds, while MeanFlow (MF) (Geng et al., 2025) and Riemannian MeanFlow (RMF) (Woo et al., 2026) enable one-step generation in Euclidean and manifold-valued spaces, respectively, and Tango-DM trains only the tangential score component of a known manifold (Liu et al., 2025). Other work targets the geometry of the embedded data distribution, but extracts it from a pretrained model: low-noise diffusion scores approximate its normal bundle (Stanczuk et al., 2024), SDDM decomposes scores into tangent and normal components (Sun et al., 2023), MPGD constructs manifold-preserving guidance from pretrained autoencoders (He et al., 2024), and the score Jacobian yields a tangent-aware metric (Saito & Matsubara, 2026). In contrast, we learn complementary tangent- and normal-sensitive features from data and place them directly in the training objective, with neither a known manifold nor a pretrained models.

Method	Geometry	Tan. / Nor.	Train.
RFM (Chen & Lipman, 2024)	given	✓/✗	✓
RMF (Woo et al., 2026)	given	✓/✗	✓
Tango-DM (Liu et al., 2025)	given	✓/✗	✓
SDDM (Sun et al., 2023)	pretrained	✓/✓	✗
MPGD (He et al., 2024)	pretrained	✗/✓	✗
Saito & Matsubara (2026)	pretrained	✓/✓	✗
AYT (Kim et al., 2026)	learned	✗/✓	✓
MaDeL (Ours)	learned	✓/✓	✓

Table 1: Geometry-aware methods: source of geometry, tangent/normal modeling, and use in training.

Feature losses as learned directional metrics. Learned features introduce anisotropy without an explicit parameterization of the data manifold: perceptual losses measure errors in representations aligned with perceptual similarity (Johnson et al., 2016; Zhang et al., 2018), and regularized autoencoders capture local geometry through denoising, Jacobian contraction, or reconstruction fields (Vincent et al., 2008; Rifai et al., 2011b; Alain & Bengio, 2014). Features have also been incorporated directly into generative training through pretrained discriminator features (Sauer et al., 2021) and feature-based consistency objectives (Song et al., 2023), including manifold-aligned tangents (AYT) (Kim et al., 2026). These approaches, however, provide a single notion of similarity rather than complementary sensitivities to intrinsic variation and off-manifold deviation.

3 Motivation

Let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^n$ be a feature encoder and define $\mathcal{L}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}, \boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{\phi}(\hat{\boldsymbol{x}}) - \boldsymbol{\phi}(\boldsymbol{x})\|_2^2$. Its gradient is $\boldsymbol{J}_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}})^\top(\boldsymbol{\phi}(\hat{\boldsymbol{x}}) - \boldsymbol{\phi}(\boldsymbol{x}))$, so the encoder Jacobian determines the input-space directions to which the loss is sensitive. We learn complementary encoders $\boldsymbol{\phi}^\parallel$ and $\boldsymbol{\phi}^\perp$ for manifold-parallel variation and manifold-orthogonal deviation, respectively. Writing $a \in \parallel, \perp$ for the two branches, we define

$$\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \textstyle\sum_{a\in\parallel,\perp} \lambda^a \mathcal{L}_{\boldsymbol{\phi}^a}(\hat{\boldsymbol{x}}, \boldsymbol{x}), \tag{2}$$

$$\nabla_{\hat{\boldsymbol{x}}}\mathcal{L}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \textstyle\sum_{a\in\parallel,\perp} \lambda^a \boldsymbol{J}_{\boldsymbol{\phi}^a}(\hat{\boldsymbol{x}})^\top \left(\boldsymbol{\phi}^a(\hat{\boldsymbol{x}}) - \boldsymbol{\phi}^a(\boldsymbol{x})\right). \tag{3}$$

If the encoders $\phi^{\parallel}$ and $\phi^{\perp}$ specialize as intended, these terms provide complementary update directions along and away from the data manifold, inducing a learned directional geometry rather than a generic perceptual metric.

Sanity check on a toy example. We first isolate the effect of the training metric on a manifold that is known exactly, using an oracle decomposition that the learned features of Sec. 4 are designed to approximate. The target distribution is a mixture of four von Mises distributions on the unit circle in $\mathbb{R}^2$. For each dataset size N, identical generators are trained with the same OT-matched objective and differ only in the regression metric; the crosses in Fig. 2 denote the training samples, and $N = \infty$ denotes training with fresh samples from the underlying distribution.

As shown in Fig. 2 left, Cartesian MSE and Pseudo-Huber (PH) interpolate between sparse modes along Euclidean chords and produce off-manifold samples inside the circle. The oracle decomposed metric applies PH in polar coordinates with $(\lambda^{\parallel}, \lambda^{\perp}) = (1, 5)$, where angle and radius correspond to the manifold-parallel and manifold-orthogonal directions, and preserves the support.

Fig. 2 right quantifies this behavior: the decomposed metric reduces the fraction of off-manifold samples to nearly zero across dataset sizes while attaining a comparable angular W_1 distance to the target distribution. The improvement therefore arises from the manifold-orthogonal geometry without loss of variation along the manifold, which underscores the role of the training metric.

4 Main Contribution: MaDeL

Building on Sec. 3, we show that the two feature spaces arise from $\boldsymbol{x}$-prediction and $\boldsymbol{\delta}$-prediction on corrupted data, and turn them into a two-stage training procedure. We characterize the optimal encoders exactly for linear manifolds (Sec. 4.1), extend the characterization locally to smooth manifolds (Sec. 4.2), connect the encoder Jacobians to the update directions of the decomposed loss, and describe the shared encoder and generator training (Sec. 4.3).

4.1 Theory: Linear Manifolds and Autoencoders

Suppose the data $\boldsymbol{x}$ is supported on an r-dimensional linear subspace of $\mathbb{R}^d$. Concretely,

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad r := \operatorname{rank}(\boldsymbol{A}) < d, \tag{4}$$

such that data $\boldsymbol{x}$ is Gaussian distributed with covariance

$$\boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U} \operatorname{diag}(\lambda_1, \dots, \lambda_d)\boldsymbol{U}^\top, \tag{5}$$

with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_d$. We consider linear encoder-decoder pairs $(\boldsymbol{E}, \boldsymbol{D})$ with bottleneck dimensions $n_{\parallel}$ for $\boldsymbol{x}$-prediction and $n_{\perp}$ for $\boldsymbol{\delta}$-prediction, together with isotropic corruption $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Proposition 1 ($\boldsymbol{x}$-prediction selects high-variance directions)**.** *Consider the denoising objective*

$$\min_{\boldsymbol{E},\boldsymbol{D}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\delta}} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{E}(\boldsymbol{x} + \boldsymbol{\delta})\|_2^2, \quad n_{\parallel} \leq r. \tag{6}$$

Assume there are no ties between positive eigenvalues, i.e., $\lambda_{n_{\parallel}} > \lambda_{n_{\parallel}+1}$, and let $\boldsymbol{U}_{n_{\parallel}}$ contain the eigenvectors corresponding to the $n_{\parallel}$ largest eigenvalues of $\boldsymbol{\Sigma}$. Then the optimal composite map is

$$\boldsymbol{D}^\star \boldsymbol{E}^\star = \boldsymbol{U}_{n_{\parallel}} \boldsymbol{\Gamma}_{\parallel} \boldsymbol{U}_{n_{\parallel}}^\top, \quad \boldsymbol{\Gamma}_{\parallel} := \operatorname{diag}\left(\tfrac{\lambda_1}{\lambda_1+\sigma^2}, \dots, \tfrac{\lambda_{n_{\parallel}}}{\lambda_{n_{\parallel}}+\sigma^2}\right). \tag{7}$$

Moreover, every optimal rank-$n_{\parallel}$ factorization can be written as $\boldsymbol{E}^\star = \boldsymbol{R}\boldsymbol{U}_{n_{\parallel}}^\top$, $\boldsymbol{D}^\star = \boldsymbol{U}_{n_{\parallel}} \boldsymbol{\Gamma}_{\parallel} \boldsymbol{R}^{-1}$ for an arbitrary invertible matrix $\boldsymbol{R} \in \mathbb{R}^{n_{\parallel} \times n_{\parallel}}$.

Proposition 1 shows that $\boldsymbol{x}$-prediction allocates the limited representation capacity to directions of largest data variance. Importantly, the encoder is insensitive to every direction outside the selected principal subspace: $\boldsymbol{E}^\star \boldsymbol{v} = \boldsymbol{0}$ for all $\boldsymbol{v} \perp \operatorname{Span}(\boldsymbol{U}_{n_{\parallel}})$.

For the linear data manifold, $\ker(\boldsymbol{\Sigma})$ is exactly its normal space; hence, all manifold-orthogonal directions are annihilated by the $\boldsymbol{x}$-prediction encoder. The corruption in the input is essential for this result: unlike ordinary autoencoding of $\boldsymbol{x}$ from $\boldsymbol{x}$, denoising exposes the encoder to off-manifold directions while requiring the output to remain unchanged.

Proposition 2 ($\boldsymbol{\delta}$-prediction selects low-variance directions)**.** *Consider instead the objective*

$$\min_{\boldsymbol{E},\boldsymbol{D}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{\delta}} \|\boldsymbol{\delta} - \boldsymbol{D}\boldsymbol{E}(\boldsymbol{x}+\boldsymbol{\delta})\|_2^2. \tag{8}$$

Let $\boldsymbol{U}_{-n_\perp}$ contain eigenvectors corresponding to the $n_\perp$ smallest eigenvalues $\lambda_{d-n_\perp+1}, \ldots, \lambda_d$ of $\boldsymbol{\Sigma}$. Then the optimal composite map is

$$\boldsymbol{D}^\star\boldsymbol{E}^\star = \boldsymbol{U}_{-n_\perp}\boldsymbol{\Gamma}_\perp\boldsymbol{U}_{-n_\perp}^\top, \quad \boldsymbol{\Gamma}_\perp \coloneqq \operatorname{diag}\left(\tfrac{\sigma^2}{\lambda_{d-n_\perp+1}+\sigma^2}, \ldots, \tfrac{\sigma^2}{\lambda_d+\sigma^2}\right), \tag{9}$$

and every optimal rank-$n_\perp$ factorization can be written as $\boldsymbol{E}^\star = \boldsymbol{R}\boldsymbol{U}_{-n_\perp}^\top$, $\boldsymbol{D}^\star = \boldsymbol{U}_{-n_\perp}\boldsymbol{\Gamma}_\perp\boldsymbol{R}^{-1}$ for an arbitrary invertible matrix $\boldsymbol{R} \in \mathbb{R}^{n_\perp \times n_\perp}$.

In contrast to $\boldsymbol{x}$-prediction, $\boldsymbol{\delta}$-prediction prioritizes directions of smallest data variance. In particular, if $n_\perp \le d - r$, the selected directions can lie entirely within $\ker(\boldsymbol{\Sigma})$, which is exactly the normal space of the data manifold. When $n_\perp = d - r$, the encoder spans the complete normal space. Thus, under a suitable bottleneck, the two objectives induce complementary representations: $\boldsymbol{x}$-prediction selects manifold-parallel variation, whereas $\boldsymbol{\delta}$-prediction selects manifold-orthogonal variation.

The distinction can also be seen directly from their direction-wise gains. For an eigen-direction with data variance λ, $\boldsymbol{x}$-prediction and $\boldsymbol{\delta}$-prediction assign $g_\parallel(\lambda) = \frac{\lambda}{\lambda+\sigma^2}$ and $g_\perp(\lambda) = \frac{\sigma^2}{\lambda+\sigma^2}$ which satisfy $g_\parallel(\lambda) + g_\perp(\lambda) = 1$. $\boldsymbol{x}$-prediction therefore increasingly favors high-variance directions, while $\boldsymbol{\delta}$-prediction increasingly favors low-variance directions.

4.2 Theory: General Smooth Manifolds and Autoencoders

We now extend the linear analysis to a general smooth data manifold with smooth autoencoder. Let $\mathcal{M} \subset \mathbb{R}^d$ be an r-dimensional smooth embedded manifold and consider a point $\boldsymbol{x}_0 \in \mathcal{M}$. We parameterize nearby clean samples on the manifold as $\boldsymbol{x}_\epsilon = \iota(\exp_{\boldsymbol{x}_0}(\epsilon\boldsymbol{\xi}))$, where the random tangent vector $\boldsymbol{\xi} \in T_{\boldsymbol{x}_0}\mathcal{M}$ possesses the first and second moments

$$\mathbb{E}[\boldsymbol{\xi}] = \boldsymbol{0}, \quad \mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = \boldsymbol{\Sigma}_T, \tag{10}$$

and $\iota : \mathcal{M} \to \mathbb{R}^d$ denotes the embedding. Let $\boldsymbol{J}_\iota \coloneqq d\iota_{\boldsymbol{x}_0}$ and

$$\boldsymbol{C}_T \coloneqq \boldsymbol{J}_\iota\boldsymbol{\Sigma}_T\boldsymbol{J}_\iota^\top = \boldsymbol{U}\operatorname{diag}(\lambda_1, \ldots, \lambda_d)\boldsymbol{U}^\top, \quad \lambda_1 \ge \cdots \ge \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_d, \tag{11}$$

so that $\operatorname{Im}(\boldsymbol{C}_T) = T_{\boldsymbol{x}_0}\mathcal{M}$ and $\ker(\boldsymbol{C}_T) = N_{\boldsymbol{x}_0}\mathcal{M}$ whenever $\boldsymbol{\Sigma}_T$ has full rank on $T_{\boldsymbol{x}_0}\mathcal{M}$. As before, independent isotropic corruption $\boldsymbol{\delta}_\sigma \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is added to $\boldsymbol{x}_\epsilon$.

For each objective, let the encoder $\boldsymbol{E}$ and decoder $\boldsymbol{D}$ be twice continuously differentiable, with bottleneck dimension $n_\parallel$ for $\boldsymbol{x}$-prediction and $n_\perp$ for $\boldsymbol{\delta}$-prediction. Denote their local Jacobians by

$$\boldsymbol{E}_1 \coloneqq d\boldsymbol{E}_{\boldsymbol{x}_0}, \quad \boldsymbol{D}_1 \coloneqq d\boldsymbol{D}_{\boldsymbol{E}(\boldsymbol{x}_0)}. \tag{12}$$

The key observation is that, after first-order expansion around $\boldsymbol{x}_0$, both objectives reduce exactly to the corresponding linear problems addressed in Sec. 4.1, with the local tangent covariance $\boldsymbol{C}_T$ replacing the global covariance $\boldsymbol{\Sigma}$.

Proposition 3 (Local behavior of $\boldsymbol{x}$-prediction)**.** *Assume $\boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_0)) = \boldsymbol{x}_0$. As $\epsilon, \sigma \to 0$,*

$$\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma}\|\boldsymbol{x}_\epsilon - \boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma))\|_2^2 = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma}\|\epsilon\boldsymbol{J}_\iota\boldsymbol{\xi} - \boldsymbol{D}_1\boldsymbol{E}_1(\epsilon\boldsymbol{J}_\iota\boldsymbol{\xi} + \boldsymbol{\delta}_\sigma)\|_2^2 + o(\epsilon^2 + \sigma^2). \tag{13}$$

For $n_\parallel \le r$, the leading-order objective is minimized by

$$\boldsymbol{D}_1^\star\boldsymbol{E}_1^\star = \boldsymbol{U}_{n_\parallel}\boldsymbol{\Gamma}_\parallel^{\mathrm{loc}}\boldsymbol{U}_{n_\parallel}^\top, \quad \boldsymbol{\Gamma}_\parallel^{\mathrm{loc}} \coloneqq \operatorname{diag}\left(\tfrac{\epsilon^2\lambda_1}{\epsilon^2\lambda_1+\sigma^2}, \ldots, \tfrac{\epsilon^2\lambda_{n_\parallel}}{\epsilon^2\lambda_{n_\parallel}+\sigma^2}\right), \tag{14}$$

with $\boldsymbol{E}_1^\star = \boldsymbol{R}\boldsymbol{U}_{n_\parallel}^\top$, $\boldsymbol{D}_1^\star = \boldsymbol{U}_{n_\parallel}\boldsymbol{\Gamma}_\parallel^{\mathrm{loc}}\boldsymbol{R}^{-1}$ for an arbitrary invertible matrix $\boldsymbol{R} \in \mathbb{R}^{n_\parallel \times n_\parallel}$.

Thus, to leading order, the Jacobian of the $\boldsymbol{x}$-prediction encoder lies in the local tangent space and is insensitive to normal perturbations: $\boldsymbol{E}_1^\star\boldsymbol{v} = 0$ for all $\boldsymbol{v} \in N_{\boldsymbol{x}_0}\mathcal{M}$.

Proposition 4 (Local behavior of $\boldsymbol{\delta}$-prediction)**.** *Assume $\boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_0)) = \boldsymbol{0}$. As $\epsilon, \sigma \to 0$,*

$$\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma}\|\boldsymbol{\delta}_\sigma - \boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma))\|_2^2 = \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma}\|\boldsymbol{\delta}_\sigma - \boldsymbol{D}_1\boldsymbol{E}_1(\epsilon\boldsymbol{J}_\iota\boldsymbol{\xi} + \boldsymbol{\delta}_\sigma)\|_2^2 + o(\epsilon^2 + \sigma^2). \tag{15}$$

The leading-order objective is minimized by

$$\boldsymbol{D}_1^\star\boldsymbol{E}_1^\star = \boldsymbol{U}_{-n_\perp}\boldsymbol{\Gamma}_\perp^{\mathrm{loc}}\boldsymbol{U}_{-n_\perp}^\top, \quad \boldsymbol{\Gamma}_\perp^{\mathrm{loc}} \coloneqq \operatorname{diag}\left(\tfrac{\sigma^2}{\epsilon^2\lambda_{d-n_\perp+1}+\sigma^2}, \ldots, \tfrac{\sigma^2}{\epsilon^2\lambda_d+\sigma^2}\right), \tag{16}$$

with $\boldsymbol{E}_1^\star = \boldsymbol{R}\boldsymbol{U}_{-n_\perp}^\top$, $\boldsymbol{D}_1^\star = \boldsymbol{U}_{-n_\perp}\boldsymbol{\Gamma}_\perp^{\mathrm{loc}}\boldsymbol{R}^{-1}$ for an arbitrary invertible matrix $\boldsymbol{R} \in \mathbb{R}^{n_\perp \times n_\perp}$.

Algorithm 1 MaDeL: Learning Manifold-Decomposed Features and the Generative Model

Require: Encoder ϕ, projections $\boldsymbol{P}_{\|}, \boldsymbol{P}_{\perp}$ onto its first $n_{\|}$ and last $n_{\perp}$ features, decoders $\boldsymbol{D}^{\|}, \boldsymbol{D}^{\perp}$, generator $\boldsymbol{G}_{\theta}$; steps $N_{\text{rep}}, N_{\text{gen}}$; weights $(\lambda_{\boldsymbol{x}}, \lambda_{\boldsymbol{\delta}}, \lambda_{\text{iso}}, \lambda^{\|}, \lambda^{\perp})$

Stage 1: Representation learning

for step $n = 1, \ldots, N_{\text{rep}}$ **do**
  Sample a clean sample $\boldsymbol{x}$, corruption $\boldsymbol{\delta}$, and let $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x} + \boldsymbol{\delta}$
  Compute representations $\boldsymbol{h}^{\|} \leftarrow (\boldsymbol{P}_{\|} \circ \boldsymbol{\phi})(\tilde{\boldsymbol{x}})$ and $\boldsymbol{h}^{\perp} \leftarrow (\boldsymbol{P}_{\perp} \circ \boldsymbol{\phi})(\tilde{\boldsymbol{x}})$
  Minimize $\lambda_{\boldsymbol{x}} \cdot \|\boldsymbol{x} - \boldsymbol{D}^{\|}(\boldsymbol{h}^{\|})\|_2^2 + \lambda_{\boldsymbol{\delta}} \cdot \|\boldsymbol{\delta} - \boldsymbol{D}^{\perp}(\boldsymbol{h}^{\perp})\|_2^2$ w.r.t. $(\boldsymbol{\phi}, \boldsymbol{D}^{\|}, \boldsymbol{D}^{\perp})$
end for
Freeze $\boldsymbol{\phi}^{\|} := \boldsymbol{P}_{\|} \circ \boldsymbol{\phi}$ and $\boldsymbol{\phi}^{\perp} := \boldsymbol{P}_{\perp} \circ \boldsymbol{\phi}$

Stage 2: Generative model learning

for step $n = 1, \ldots, N_{\text{gen}}$ **do**
  Sample prior $\boldsymbol{\varepsilon}$ and generative model output $\hat{\boldsymbol{x}}_{\theta} \leftarrow \boldsymbol{G}_{\theta}(\boldsymbol{\varepsilon})$
  Assign a target $\boldsymbol{x}_{\text{targ}}$ to $\hat{\boldsymbol{x}}_{\theta}$ by the underlying generative method
  Minimize $\lambda_{\text{iso}}\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}} - \boldsymbol{x}_{\text{targ}}\|_2^2 + \lambda^{\|}\|\boldsymbol{\phi}^{\|}(\hat{\boldsymbol{x}}_{\theta}) - \boldsymbol{\phi}^{\|}(\boldsymbol{x}_{\text{targ}})\|_2^2 + \lambda^{\perp}\|\boldsymbol{\phi}^{\perp}(\hat{\boldsymbol{x}}_{\theta}) - \boldsymbol{\phi}^{\perp}(\boldsymbol{x}_{\text{targ}})\|_2^2$ w.r.t. θ
end for

For $n_{\perp} \leq d - r$, the vanishing eigenvalues of $\boldsymbol{C}_T$ correspond to normal directions; hence, the local Jacobian of the $\boldsymbol{\delta}$-prediction encoder lies in the normal space and spans it when $n_{\perp} = d - r$. Propositions 3 and 4 thus provide complementary local sensitivities without access to the tangent or normal spaces. Fig. 3 verifies this on a two-dimensional spiral, where the $\boldsymbol{x}$- and $\boldsymbol{\delta}$-prediction encoders have Jacobians parallel and orthogonal to the manifold.

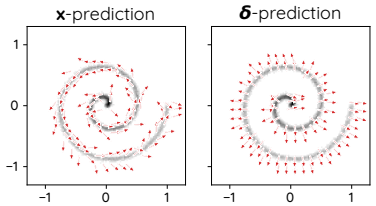


Figure 3: **Encoder Jacobians on a 2D spiral.**

The linear setting is therefore not merely illustrative: after locally linearizing the manifold and encoder-decoder, the general problem reduces to the rank-constrained regression of Sec. 4.1. This motivates the construction below, where $\boldsymbol{x}$- and $\boldsymbol{\delta}$-prediction encoders define the manifold-parallel and manifold-orthogonal losses.

4.3 From Autoencoders to Generative Learning

We now connect these local Jacobians to the generative objective in Eq. (2). Recall from Eq. (3) that the gradient of a feature loss is obtained by mapping the feature residual back to the ambient space through the transpose of the feature Jacobian; the row space of the encoder Jacobian therefore determines the directions along which the loss can update a generated sample.

Proposition 5 (Directional decomposition of the feature loss)**.** *Let $\boldsymbol{x}_0 \in \mathcal{M}$, $n_{\perp} \leq d - r$, and $\boldsymbol{\phi}^{\|}$ and $\boldsymbol{\phi}^{\perp}$ denote encoders satisfying the locally optimal characterizations of Prop. 3 and 4. Then*

$$\operatorname{Im}((\boldsymbol{J}_{\boldsymbol{\phi}^{\|}}(\boldsymbol{x}_0))^{\top}) \subseteq T_{\boldsymbol{x}_0}\mathcal{M}, \quad \operatorname{Im}((\boldsymbol{J}_{\boldsymbol{\phi}^{\perp}}(\boldsymbol{x}_0))^{\top}) \subseteq N_{\boldsymbol{x}_0}\mathcal{M}. \tag{17}$$

Thus, the gradient of the decomposed feature loss in Eq. (2), evaluated locally at $\boldsymbol{x}_0$, decomposes

$$\nabla_{\hat{\boldsymbol{x}}}\mathcal{L} = \lambda^{\|}\boldsymbol{g}^{\|} + \lambda^{\perp}\boldsymbol{g}^{\perp}, \quad \boldsymbol{g}^{\|} \in T_{\boldsymbol{x}_0}\mathcal{M}, \quad \boldsymbol{g}^{\perp} \in N_{\boldsymbol{x}_0}\mathcal{M}. \tag{18}$$

Proposition 5 connects the representation-learning objectives to the generative loss: the $\boldsymbol{x}$-prediction encoder yields feature gradients confined to manifold-parallel directions and the $\boldsymbol{\delta}$-prediction encoder to manifold-orthogonal directions, so that their combination in Eq. (2) updates a generated sample along complementary directions.

Learning the encoders. We use a shared-encoder dual-decoder architecture. Given $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$, the encoder $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^n$, with $n = n_{\|} + n_{\perp}$, is split by coordinate projections $\boldsymbol{P}_{\|}$ and $\boldsymbol{P}_{\perp}$ into

$$\boldsymbol{\phi}^{\|} := \boldsymbol{P}_{\|} \circ \boldsymbol{\phi}, \qquad \boldsymbol{\phi}^{\perp} := \boldsymbol{P}_{\perp} \circ \boldsymbol{\phi}. \tag{19}$$

These branch encoders correspond to Propositions 3 and 4. Under the assumptions of Sec. 4.2, $n_{\|} \leq r$ restricts $\boldsymbol{x}$-prediction features to dominant tangent directions, while $n_{\perp} \leq d - r$ restricts

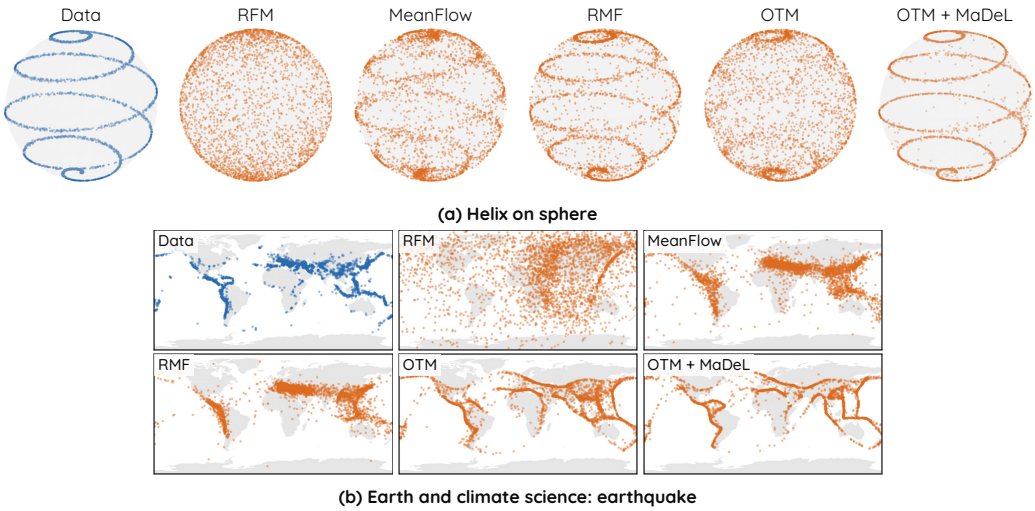


Figure 4: **Single-step generation on manifold-valued benchmarks.**

$\boldsymbol{\delta}$-prediction features to the normal space, which they span when $n_\perp = d - r$. Because these dimensions are unknown, $n_\parallel$ and $n_\perp$ are treated as branch-capacity hyperparameters.

Each feature subset is provided exclusively to its corresponding decoder: the parallel decoder $\boldsymbol{D}^\parallel$ is trained for $\boldsymbol{x}$-prediction, while the orthogonal decoder $\boldsymbol{D}^\perp$ is trained for $\boldsymbol{\delta}$-prediction:

$$\mathcal{L}_{\mathrm{AE}} := \lambda_{\boldsymbol{x}} \cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{\delta}}[\|\boldsymbol{x} - \boldsymbol{D}^\parallel(\boldsymbol{\phi}^\parallel(\boldsymbol{x} + \boldsymbol{\delta}))\|_2^2] + \lambda_{\boldsymbol{\delta}} \cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{\delta}}[\|\boldsymbol{\delta} - \boldsymbol{D}^\perp(\boldsymbol{\phi}^\perp(\boldsymbol{x} + \boldsymbol{\delta}))\|_2^2] \tag{20}$$

The partition prevents the two decoders from sharing task-specific features while retaining a common feature extractor. By the analysis of Sec. 4.1, $\boldsymbol{x}$-prediction renders $\boldsymbol{\phi}^\parallel$ insensitive to off-manifold corruption, whereas $\boldsymbol{\delta}$-prediction drives $\boldsymbol{\phi}^\perp$ to capture deviations from the data manifold. These feature subsets enter the generative loss in Eq. (2) directly.

Training the generator. Our method is two-stage: after training, the encoders are frozen and used to train the generator. Given $\boldsymbol{\varepsilon}$ from the prior, the generator produces $\hat{\boldsymbol{x}}_\theta := \boldsymbol{G}_\theta(\boldsymbol{\varepsilon})$; the underlying method constructs a regression target $\boldsymbol{x}_{\mathrm{targ}}$, and we modify only the loss comparing these quantities.

For OT matching, $\boldsymbol{x}_{\mathrm{targ}}$ is assigned by minibatch OT between current outputs and real samples, following explicit Wasserstein minimization (Fatras et al., 2020; Chen et al., 2019). Given batches $\{\boldsymbol{\varepsilon}_i\}$ and $\{\boldsymbol{x}_j\}$, with $\hat{\boldsymbol{x}}_{\theta,i} = \boldsymbol{G}_\theta(\boldsymbol{\varepsilon}_i)$, we compute

$$\pi^\star := \arg\min_{\pi \in \Pi(B)} \textstyle\sum_{i,j} \pi_{ij} \|\hat{\boldsymbol{x}}_{\theta,i} - \boldsymbol{x}_j\|_2^2. \tag{21}$$

We then set $\boldsymbol{x}_{\mathrm{targ}} = \boldsymbol{x}_j$ for the data sample assigned to $\hat{\boldsymbol{x}}_{\theta,i}$, treating the assignment as fixed during the generator update. For a flow-map semigroup objective such as Riemannian MeanFlow (Woo et al., 2026), let $\boldsymbol{G}_\theta^{s\to t}(\boldsymbol{x}_s)$ denote the model prediction obtained by transporting $\boldsymbol{x}_s$ directly from time s to t. The regression target is constructed by composing predictions over two shorter intervals:

$$\hat{\boldsymbol{x}}_\theta = \boldsymbol{G}_\theta^{s\to t}(\boldsymbol{x}_s), \quad \boldsymbol{x}_{\mathrm{targ}} = \mathrm{sg}\left[\boldsymbol{G}_\theta^{r\to t}\left(\boldsymbol{G}_\theta^{s\to r}(\boldsymbol{x}_s)\right)\right]. \tag{22}$$

Here, $s \le r \le t$, and sg denotes stop-gradient. Additional terms required by the underlying method, such as the MeanFlow boundary loss, are retained. Given $\boldsymbol{x}_{\mathrm{targ}}$, we train the generator using

$$\mathcal{L}_{\mathrm{gen}}(\theta) := \mathbb{E}[\lambda_{\mathrm{iso}} \|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}} - \boldsymbol{x}_{\mathrm{targ}}\|_2^2 + \lambda^\parallel \mathcal{L}_{\boldsymbol{\phi}^\parallel}(\hat{\boldsymbol{x}}_\theta, \boldsymbol{x}_{\mathrm{targ}}) + \lambda^\perp \mathcal{L}_{\boldsymbol{\phi}^\perp}(\hat{\boldsymbol{x}}_\theta, \boldsymbol{x}_{\mathrm{targ}})]. \tag{23}$$

Thus, the proposed metric is independent of how the regression target is constructed. The squared discrepancy in either ambient or feature space may be replaced by any differentiable isotropic loss; the gradient-decomposition analysis extends directly, up to a loss-dependent scalar weighting of each residual. Algorithm 1 summarizes the complete two-stage procedure.

5 Experiments

We evaluate MaDeL on four benchmark families spanning synthetic and structured real-world distributions. Full experimental details are provided in Sec. C.

Method	Earth and Climate Science Datasets ($W_1 \downarrow$)					Protein and RNA Torsion-Angle Datasets ($W_1 \downarrow$)					
	Earthquake	Fire	Flood	Volcano	Avg.	General	Glycine	Proline	Pre-Pro	RNA	Avg.
RFM	27.19°	33.04°	23.68°	28.40°	28.08°	57.79°	64.80°	60.74°	32.68°	63.55°	55.912°
MF	6.43°	9.22°	6.67°	14.95°	9.32°	17.27°	11.57°	16.25°	13.77°	35.62°	18.896°
RMF	9.91°	10.21°	8.39°	16.82°	11.33°	34.33°	24.63°	31.84°	15.21°	41.84°	29.570°
OTM	6.22°	6.25°	**4.88°**	6.22°	5.89°	4.81°	8.27°	4.97°	**5.97°**	23.68°	9.540°
+ MaDeL	**5.81°**	**5.93°**	5.16°	**5.46°**	**5.59°**	**4.51°**	**6.69°**	**4.39°**	6.03°	**22.52°**	**8.828°**

Table 2: **Single-step generation error on Earth and torsion-angle datasets.** Angular W_1 distance (degrees) between generated and ground-truth distributions; lower is better. RFM: Riemannian Flow Matching; MF, RMF: (Riemannian) MeanFlow; OTM: coordinate-space OT matching; + MaDeL: OTM with our loss. All methods use single-step sampling.

Benchmarks. *Helix on sphere*: a helix in $\mathbb{S}^2$, embedded isometrically into $\mathbb{S}^{512} \subset \mathbb{R}^{513}$ by a fixed orthonormal map and perturbed by tangential Gaussian noise. *Earth and climate science*: geographic coordinates of volcanic eruptions, earthquakes, floods, and wildfires represented on $\mathbb{S}^2$, following prior benchmarks for manifold-valued generation (Bortoli et al., 2022; Chen & Lipman, 2024). *Protein and RNA torsion angles*: the preprocessed benchmarks of Riemannian Flow Matching (Chen & Lipman, 2024), following Huang et al. (2022), comprising four Ramachandran-angle distributions (General, Glycine, Proline, and Pre-Pro) on $\mathbb{T}^2$ and a seven-dimensional RNA torsion distribution on $\mathbb{T}^7$. *Protein backbones*: SCOPe protein structures (Chandonia et al., 2022), following the preprocessing and evaluation protocol of FrameFlow (Yim et al., 2023a).

Baselines. We compare with Riemannian Flow Matching (RFM) (Chen & Lipman, 2024), MeanFlow (MF) (Geng et al., 2025), and Riemannian MeanFlow (RMF) (Woo et al., 2026), each trained with its native objective. We also compare with coordinate-space OT matching (OTM), which computes a minibatch OT coupling between generated samples and real samples and regresses each generated output toward its assigned real-data target. OTM and OTM + MaDeL use the same coupling, generator architecture, and sampling procedure, differing only in the regression loss.

5.1 Main Results

Helix on sphere. In Fig. 4 top, RFM, MF, RMF, and OTM distribute samples broadly over the sphere, failing to concentrate on the helical support. The Riemannian methods preserve the geometry of the spherical state space but do not recover the lower-dimensional data manifold embedded within it. In contrast, OTM + MaDeL reconstructs the global helical structure and places more probability mass near the data support using the same single network evaluation.

Earth and climate science. As shown in Fig. 4 bottom, RFM spreads samples broadly across the surface, while MF and RMF recover the major high-density regions but blur their finer structure. OTM captures much of the global support but retains noticeable dispersion away from the narrow data structures. Using the same OT assignments, augmenting the coordinate-space regression loss with MaDeL produces more sharply localized samples while preserving broad geographic coverage. In Tab. 2 left, OTM + MaDeL achieves the lowest angular distance on three of the four datasets and improves the Earth average from $5.89°$ to $5.59°$; OTM remains strongest on Flood.

Protein and RNA torsion angles. In Tab. 2 right, OTM + MaDeL achieves the lowest angular W_1 distance on four of the five datasets and improves the mean distance from $9.540°$ to $8.828°$ over OTM. It remains essentially tied on Pre-Pro ($6.03°$ versus $5.97°$), with the largest improvement on Glycine, the broadest Ramachandran distribution.

Protein backbone generation. We train identical RMF models, with MaDeL adding a frozen decomposed-feature term to the standard semigroup objective. Although MaDeL does not improve designability, it improves local steric geometry. We measure the maximum van der Waals overlap between backbone N, C_α, and C atoms of residues separated by at least two positions; lower values indicate fewer steric clashes, and real SCOPe backbones score 0.28–0.30 Å. As shown in Fig. 5, across eight paired prior batches from the one-million-step checkpoints, MaDeL reduces the median maximum overlap at all ten tested lengths and all three sampling budgets. At $L = 128$, it reduces overlap from 1.375 to 1.011 Å at $\mathrm{NFE} = 10$ (lower in $8/8$ batches), from 1.742 to 1.503 Å at $\mathrm{NFE} = 5$ ($7/8$), and from 2.732 to 2.681 Å at $\mathrm{NFE} = 1$ ($6/8$). This is not explained by chain expansion: at $L = 60$, MaDeL produces backbones that are both more compact and less clashing in

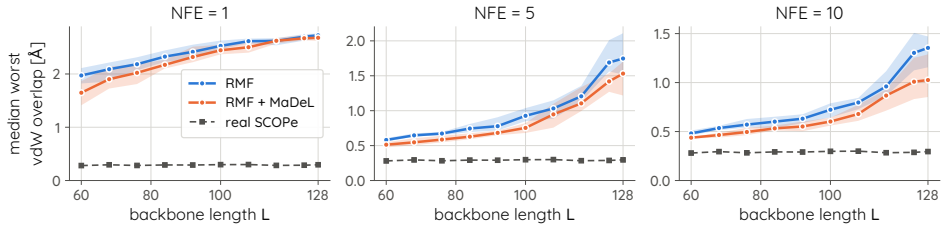


Figure 5: **Steric violation versus backbone length.** Deepest backbone van der Waals overlap per structure (N/Cα/C, residues $\geq$ 2 apart), median over 8 independent batches of 128 paired samples from each model's 1M-step checkpoint; bands are the full batch range, dashed is real SCOPe. MaDeL is lower at every length and budget.

6–7 of eight batches. Designability remains essentially unchanged ($10/640$ versus $12/640$), as does median self-consistency RMSD (8.63 versus 8.66 Å at $L = 128$ and NFE $= 10$).

5.2 Ablation Study and Analysis

Ablation. We ablate the two feature branches under the same OT pairing (Tab. 3). While Flood is an exception, with all feature-based variants underperforming coordinate-space OTM, the average results show a clear trend: the overparameterized autoencoder yields only a modest improvement from $5.89°$ to $5.84°$, either manifold-decomposed branch alone performs worse ($6.13°$ for $\|$ and $6.08°$ for $\perp$), and combining both branches achieves the best average of $5.59°$. MaDeL also attains the lowest error on Earthquake, Fire, and Volcano, indicating that the complementary branches provide the strongest overall performance despite variation across datasets.

Loss	Earthquake	Fire	Flood	Volcano	Avg.
OTM	6.22°	6.25°	**4.88°**	6.22°	5.89°
+ Overparam. AE	5.89°	6.16°	5.08°	6.24°	5.84°
+ $\|$ only ($\lambda^{\perp} = 0$)	6.12°	6.94°	5.41°	6.04°	6.13°
+ $\perp$ only ($\lambda^{\|} = 0$)	6.59°	6.42°	5.17°	6.15°	6.08°
+ MaDeL	**5.81°**	**5.93°**	5.16°	**5.46°**	**5.59°**

Table 3: **Ablation of the decomposed loss.** Angular W_1 under the same OT pairing.

Analysis. To test whether the encoders realize the decomposition predicted by Sec. 4.2, we use the tangent and normal spaces of the helix and measure the ratio of manifold-parallel to manifold-orthogonal Jacobian energy (Tab. 4). The ratio is evaluated at samples perturbed by noise with standard deviation σ. Across noise levels, $\phi^{\|}$ produces large ratios (15.88–45.04), whereas $\phi^{\perp}$ produces ratios near zero (0.02–0.04), confirming that energies concentrate in complementary directions on and near the manifold.

	$\sigma = 0.0$	0.01	0.05	0.1
$\phi^{\|}$	15.88	16.75	45.04	33.21
$\phi^{\perp}$	0.02	0.02	0.04	0.03

Table 4: **Jacobian tangent-to-normal energy ratio.**

6 Conclusion

We introduced MaDeL, a manifold-decomposed feature loss for generative modeling that distinguishes variation along the data manifold from deviation away from it. We showed that, under a feature bottleneck, $\boldsymbol{x}$-prediction and $\boldsymbol{\delta}$-prediction learn encoders whose Jacobians are confined to the tangent and normal spaces of the manifold, exactly for linear manifolds and locally for smooth manifolds, so that the resulting feature loss yields complementary tangent- and normal-aligned generative updates without explicit access to the manifold. Across controlled manifolds and structured real-world domains, MaDeL better preserves data support and improves distributional or local geometric accuracy relative to isotropic regression, which suggests that explicitly shaping the geometry of the training signal is a practical alternative to parameterizing the data manifold itself, particularly in sparse and highly constrained generative problems.

Limitations. Our analysis is local and holds to leading order in the corruption level, while the branch dimensions and loss weights must be chosen without knowledge of the intrinsic dimension. Our experiments cover a limited set of synthetic and structured scientific benchmarks, and do not yet establish scalability across models, datasets, or modalities. Future work will develop MaDeL for scalable manifold generative modeling; Sec. B discusses these limitations.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to support language editing, such as correcting grammar, refining phrasing, and enhancing the clarity and readability of text originally written by the authors. We also used AI-based tools to assist with finding related works, code inspection and debugging, mathematical proof verification, as well as to explore initial designs for figures. All AI-assisted outputs and modifications were carefully reviewed by the authors, who assume full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available scientific datasets and involves no human subjects, personally identifiable information, or private data. Although generative protein models may eventually support biological design and thus present potential dual-use concerns, our experiments are limited to computational generation and geometric evaluation: the models are not conditioned on biological function or pathogenicity, and no generated structures are synthesized or experimentally validated. Downstream biological applications should follow appropriate biosafety, biosecurity, data-use, and institutional review procedures.

REPRODUCIBILITY STATEMENT

The assumptions and derivations supporting our theoretical results are presented in Secs. 4.1 and 4.2, with complete proofs in Appendix A. Algorithm 1 specifies the two-stage representation and generator training procedure, while Sec. C documents dataset construction and preprocessing, architectures, corruption processes, optimization hyperparameters, data splits and random seeds, pairing procedures, sampling budgets, and evaluation metrics. Comparisons use matched architectures and sampling procedures and, where applicable, identical OT assignments or paired prior samples. Together with the cited public data sources, these details support independent reproduction of the reported results.

REFERENCES

Guillaume Alain and Yoshua Bengio. What Regularized Auto-Encoders Learn from the Data-Generating Distribution. *JMLR*, 2014.

Abraham Bondi. Van der Waals Volumes and Radii. *The Journal of Physical Chemistry*, 68(3): 441–451, 1964.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian Score-Based Generative Modelling. In *NeurIPS*, 2022.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K. Fox, and Steven E. Brenner. SCOPe: Improvements to the Structural Classification of Proteins—Extended Database to Facilitate Variant Interpretation and Machine Learning. *Nucleic Acids Research*, 50(D1):D553–D559, 2022.

Ricky T. Q. Chen and Yaron Lipman. Flow Matching on General Geometries. In *ICLR*, 2024.

Yucheng Chen, Matus Telgarsky, Chao Zhang, Bolton Bailey, Daniel Hsu, and Jian Peng. A Gradual, Semi-Discrete Approach to Generative Network Training via Explicit Wasserstein Minimization. In *ICML*, 2019.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, et al. Robust Deep Learning-Based Protein Sequence Design Using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with Minibatch Wasserstein: Asymptotic and Gradient Properties. In *AISTATS*, 2020.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the Manifold Hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean Flows for One-Step Generative Modeling. In *NeurIPS*, 2025.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold Preserving Guided Diffusion. In *ICLR*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. In *NeurIPS*, 2022.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *ICML*, 2022.

Chin-Wei Huang, Milad Aghajohari, Avishek Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian Diffusion Models. In *NeurIPS*, 2022.

Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*, 2016.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*, 2022.

Beomsu Kim, ByungHee Cha, and Jong Chul Ye. Align Your Trajectory Tangent: Training Better Consistency Models via Manifold-Aligned Tangents. In *ICML*, 2026.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science*, 379(6637):1123–1130, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

Zichen Liu, Wei Zhang, and Tiejun Li. Improving the Euclidean Diffusion Generation of Manifold Data by Mitigating Score Function Singularity. In *NeurIPS*, 2025.

Hariharan Narayanan and Sanjoy Mitter. Sample Complexity of Testing the Manifold Hypothesis. In *NeurIPS*, 2010.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The Intrinsic Dimension of Images and Its Impact on Learning. In *ICLR*, 2021.

Salah Rifai, Yann N. Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The Manifold Tangent Classifier. In *NeurIPS*, 2011a.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive Auto-Encoders: Explicit Invariance During Feature Extraction. In *ICML*, 2011b.

Shinnosuke Saito and Takashi Matsubara. Be Tangential to Manifold: Discovering Riemannian Metric for Diffusion Models. In *ECCV*, 2026.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs Converge Faster. In *NeurIPS*, 2021.

Yang Song and Prafulla Dhariwal. Improved Techniques for Training Consistency Models. In *ICLR*, 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 *ICLR*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *ICML*, 2023.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In *ICML*, 2024.

Shikun Sun, Longhui Wei, Junliang Xing, Jia Jia, and Qi Tian. SDDM: Score-Decomposed Diffusion Models on Manifolds for Unpaired Image-to-Image Translation. In *ICML*, 2023.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 2000.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *ICML*, 2008.

Dongyeop Woo, Marta Skreta, Seonghyun Park, Kirill Neklyudov, and Sungsoo Ahn. Riemannian MeanFlow. In *ICML*, 2026.

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast Protein Backbone Generation with SE(3) Flow Matching. *arXiv preprint arXiv:2310.05297*, 2023a.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) Diffusion Model with Application to Protein Backbone Generation. In *ICML*, 2023b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018.

A Proofs

A.1 Linear manifolds

Proof of Proposition 1. Let $\boldsymbol{y} \coloneqq \boldsymbol{x} + \boldsymbol{\delta}$ and $\boldsymbol{C} \coloneqq \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = \boldsymbol{\Sigma} + \sigma^2 \boldsymbol{I}$. Since $\sigma > 0$, $\boldsymbol{C}$ is positive definite. The objective depends on the encoder and decoder only through $\boldsymbol{F} \coloneqq \boldsymbol{D}\boldsymbol{E}$, where $\operatorname{rank}(\boldsymbol{F}) \le n_\parallel$. Using $\mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] = \boldsymbol{\Sigma}$ gives

$$\mathbb{E}\|\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}\|_2^2 = \operatorname{tr}(\boldsymbol{\Sigma}) - 2\operatorname{tr}(\boldsymbol{F}\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{F}\boldsymbol{C}\boldsymbol{F}^\top) \tag{24}$$

$$= \|\boldsymbol{F}\boldsymbol{C}^{1/2} - \boldsymbol{\Sigma}\boldsymbol{C}^{-1/2}\|_F^2 + \operatorname{tr}(\boldsymbol{\Sigma}) - \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{C}^{-1}\boldsymbol{\Sigma}). \tag{25}$$

Hence, after defining $\boldsymbol{G} \coloneqq \boldsymbol{F}\boldsymbol{C}^{1/2}$, minimizing the objective is equivalent to

$$\min_{\operatorname{rank}(\boldsymbol{G}) \le n_\parallel} \|\boldsymbol{G} - \boldsymbol{\Sigma}\boldsymbol{C}^{-1/2}\|_F^2. \tag{26}$$

Because $\boldsymbol{C} = \boldsymbol{U}\operatorname{diag}(\lambda_i + \sigma^2)\boldsymbol{U}^\top$ commutes with $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma}\boldsymbol{C}^{-1/2} = \boldsymbol{U}\operatorname{diag}\left(\frac{\lambda_1}{\sqrt{\lambda_1 + \sigma^2}}, \dots, \frac{\lambda_d}{\sqrt{\lambda_d + \sigma^2}}\right)\boldsymbol{U}^\top. \tag{27}$$

The scalar map $\lambda \mapsto \lambda/\sqrt{\lambda + \sigma^2}$ is strictly increasing on $[0, \infty)$. The Eckart–Young–Mirsky theorem and the assumed eigengap therefore show that the unique optimal rank-$n_\parallel$ approximation retains the first $n_\parallel$ eigendirections:

$$\boldsymbol{G}^\star = \boldsymbol{U}_{n_\parallel}\operatorname{diag}\left(\frac{\lambda_1}{\sqrt{\lambda_1 + \sigma^2}}, \dots, \frac{\lambda_{n_\parallel}}{\sqrt{\lambda_{n_\parallel} + \sigma^2}}\right)\boldsymbol{U}_{n_\parallel}^\top. \tag{28}$$

Multiplying by $\boldsymbol{C}^{-1/2}$ on the right yields

$$\boldsymbol{F}^\star = \boldsymbol{U}_{n_\parallel} \boldsymbol{\Gamma}_\parallel \boldsymbol{U}_{n_\parallel}^\top. \tag{29}$$

Since $n_\parallel \leq r$, $\boldsymbol{\Gamma}_\parallel$ is invertible and $\operatorname{rank}(\boldsymbol{F}^\star) = n_\parallel$. Thus, any factorization $\boldsymbol{F}^\star = \boldsymbol{D}^\star \boldsymbol{E}^\star$ through an $n_\parallel$-dimensional bottleneck has $\operatorname{rank}(\boldsymbol{D}^\star) = \operatorname{rank}(\boldsymbol{E}^\star) = n_\parallel$ and $\operatorname{Im}(\boldsymbol{D}^\star) = \operatorname{Span}(\boldsymbol{U}_{n_\parallel})$. Write $\boldsymbol{D}^\star = \boldsymbol{U}_{n_\parallel} \boldsymbol{A}$ for an invertible $\boldsymbol{A}$. Then $\boldsymbol{E}^\star = \boldsymbol{A}^{-1} \boldsymbol{\Gamma}_\parallel \boldsymbol{U}_{n_\parallel}^\top$. Setting $\boldsymbol{R} \coloneqq \boldsymbol{A}^{-1} \boldsymbol{\Gamma}_\parallel$ gives

$$\boldsymbol{E}^\star = \boldsymbol{R} \boldsymbol{U}_{n_\parallel}^\top, \qquad \boldsymbol{D}^\star = \boldsymbol{U}_{n_\parallel} \boldsymbol{\Gamma}_\parallel \boldsymbol{R}^{-1}, \tag{30}$$

which characterizes every optimal factorization. □

Proof of Proposition 2. Again let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\delta}$, $\boldsymbol{C} = \boldsymbol{\Sigma} + \sigma^2 \boldsymbol{I}$, and $\boldsymbol{F} = \boldsymbol{D}\boldsymbol{E}$. Since $\mathbb{E}[\boldsymbol{\delta} \boldsymbol{y}^\top] = \sigma^2 \boldsymbol{I}$,

$$\mathbb{E}\|\boldsymbol{\delta} - \boldsymbol{F}\boldsymbol{y}\|_2^2 = d\sigma^2 - 2\sigma^2 \operatorname{tr}(\boldsymbol{F}) + \operatorname{tr}(\boldsymbol{F}\boldsymbol{C}\boldsymbol{F}^\top) \tag{31}$$

$$= \|\boldsymbol{F}\boldsymbol{C}^{1/2} - \sigma^2 \boldsymbol{C}^{-1/2}\|_F^2 + d\sigma^2 - \sigma^4 \operatorname{tr}(\boldsymbol{C}^{-1}). \tag{32}$$

With $\boldsymbol{G} \coloneqq \boldsymbol{F}\boldsymbol{C}^{1/2}$, it remains to find the best rank-$n_\perp$ approximation of

$$\sigma^2 \boldsymbol{C}^{-1/2} = \boldsymbol{U} \operatorname{diag}\left(\frac{\sigma^2}{\sqrt{\lambda_1 + \sigma^2}}, \ldots, \frac{\sigma^2}{\sqrt{\lambda_d + \sigma^2}} \right) \boldsymbol{U}^\top. \tag{33}$$

These singular values decrease as λ_i increases. Thus, the Eckart–Young–Mirsky theorem selects the $n_\perp$ smallest-variance eigendirections, giving

$$\boldsymbol{G}^\star = \boldsymbol{U}_{-n_\perp} \operatorname{diag}\left(\frac{\sigma^2}{\sqrt{\lambda_{d-n_\perp+1} + \sigma^2}}, \ldots, \frac{\sigma^2}{\sqrt{\lambda_d + \sigma^2}} \right) \boldsymbol{U}_{-n_\perp}^\top. \tag{34}$$

Therefore,

$$\boldsymbol{F}^\star = \boldsymbol{G}^\star \boldsymbol{C}^{-1/2} = \boldsymbol{U}_{-n_\perp} \boldsymbol{\Gamma}_\perp \boldsymbol{U}_{-n_\perp}^\top. \tag{35}$$

The diagonal entries of $\boldsymbol{\Gamma}_\perp$ are strictly positive, so $\operatorname{rank}(\boldsymbol{F}^\star) = n_\perp$. The same factorization argument as in Proposition 1 then yields

$$\boldsymbol{E}^\star = \boldsymbol{R} \boldsymbol{U}_{-n_\perp}^\top, \qquad \boldsymbol{D}^\star = \boldsymbol{U}_{-n_\perp} \boldsymbol{\Gamma}_\perp \boldsymbol{R}^{-1} \tag{36}$$

for an arbitrary invertible $\boldsymbol{R}$. If eigenvalues are tied at the truncation boundary, $\boldsymbol{U}_{-n_\perp}$ may be any orthonormal basis of an optimal subspace within the tied eigenspace. □

A.2 General smooth manifolds

Proof of Proposition 3. Define the local autoencoder map $\boldsymbol{A}_{\mathrm{AE}} \coloneqq \boldsymbol{D} \circ \boldsymbol{E}$. The manifold parameterization has the first-order expansion

$$\boldsymbol{x}_\epsilon = \boldsymbol{x}_0 + \epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{r}_\epsilon, \qquad \boldsymbol{r}_\epsilon = o(\epsilon) \tag{37}$$

in mean square. Since $\boldsymbol{A}_{\mathrm{AE}}$ is twice continuously differentiable and $\boldsymbol{A}_{\mathrm{AE}}(\boldsymbol{x}_0) = \boldsymbol{x}_0$, the chain rule gives

$$\boldsymbol{A}_{\mathrm{AE}}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma) = \boldsymbol{x}_0 + \boldsymbol{D}_1 \boldsymbol{E}_1 (\epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{\delta}_\sigma) + \boldsymbol{r}_{\mathrm{AE}}, \tag{38}$$

where the combined Taylor remainder satisfies $\mathbb{E}\|\boldsymbol{r}_\epsilon - \boldsymbol{r}_{\mathrm{AE}}\|_2^2 = o(\epsilon^2 + \sigma^2)$ and its cross term with the first-order residual is also $o(\epsilon^2 + \sigma^2)$. Consequently,

$$\mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\delta}_\sigma} \|\boldsymbol{x}_\epsilon - \boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma))\|_2^2 \tag{39}$$

$$= \mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\delta}_\sigma} \|\epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} - \boldsymbol{D}_1 \boldsymbol{E}_1 (\epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{\delta}_\sigma)\|_2^2 + o(\epsilon^2 + \sigma^2). \tag{40}$$

The leading-order problem is precisely the reduced-rank denoising problem of Proposition 1, with clean covariance $\epsilon^2 \boldsymbol{C}_T$ and corruption covariance $\sigma^2 \boldsymbol{I}$. Its eigendirections are those of $\boldsymbol{C}_T$, and the corresponding gains are

$$\frac{\epsilon^2 \lambda_i}{\epsilon^2 \lambda_i + \sigma^2}. \tag{41}$$

Applying Proposition 1 therefore gives the stated $\boldsymbol{D}_1^\star \boldsymbol{E}_1^\star$, $\boldsymbol{E}_1^\star$, and $\boldsymbol{D}_1^\star$. □

Proof of Proposition 4. Use the same expansion $\boldsymbol{x}_\epsilon = \boldsymbol{x}_0 + \epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{r}_\epsilon$. Now $\boldsymbol{A}_{\mathrm{AE}}(\boldsymbol{x}_0) = \boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_0)) = \mathbf{0}$, so Taylor expansion and the chain rule yield

$$\boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma)) = \boldsymbol{D}_1\boldsymbol{E}_1(\epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{\delta}_\sigma) + \boldsymbol{r}_{\mathrm{AE}}. \tag{42}$$

As in the preceding proof, the squared remainder and its cross term with the first-order residual are $o(\epsilon^2 + \sigma^2)$. Hence,

$$\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma} \|\boldsymbol{\delta}_\sigma - \boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_\epsilon + \boldsymbol{\delta}_\sigma))\|_2^2 \tag{43}$$

$$= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\delta}_\sigma} \|\boldsymbol{\delta}_\sigma - \boldsymbol{D}_1\boldsymbol{E}_1(\epsilon \boldsymbol{J}_\iota \boldsymbol{\xi} + \boldsymbol{\delta}_\sigma)\|_2^2 + o(\epsilon^2 + \sigma^2). \tag{44}$$

This is the linear $\boldsymbol{\delta}$-prediction problem of Proposition 2 with data covariance $\epsilon^2 \boldsymbol{C}_T$. The optimal rank-$n_\perp$ map therefore retains the $n_\perp$ smallest eigenvalues of $\boldsymbol{C}_T$ and assigns gain

$$\frac{\sigma^2}{\epsilon^2 \lambda_i + \sigma^2} \tag{45}$$

to the eigendirection with eigenvalue λ_i. Proposition 2 then gives the stated composite map and all of its optimal factorizations. □

A.3 Directional decomposition of the generative loss

Proof of Proposition 5. By Proposition 3, the Jacobian of a locally optimal manifold-parallel encoder has the form

$$\boldsymbol{J}_{\boldsymbol{\phi}^\parallel}(\boldsymbol{x}_0) = \boldsymbol{R}_\parallel \boldsymbol{U}_{n_\parallel}^\top. \tag{46}$$

The columns of $\boldsymbol{U}_{n_\parallel}$ are positive-eigenvalue eigenvectors of $\boldsymbol{C}_T$ and therefore lie in $\mathrm{Im}(\boldsymbol{C}_T) = T_{\boldsymbol{x}_0}\mathcal{M}$. Since $\boldsymbol{R}_\parallel$ is invertible,

$$\mathrm{Im}((\boldsymbol{J}_{\boldsymbol{\phi}^\parallel}(\boldsymbol{x}_0))^\top) = \mathrm{Span}(\boldsymbol{U}_{n_\parallel}) \subseteq T_{\boldsymbol{x}_0}\mathcal{M}. \tag{47}$$

Similarly, Proposition 4 gives

$$\boldsymbol{J}_{\boldsymbol{\phi}^\perp}(\boldsymbol{x}_0) = \boldsymbol{R}_\perp \boldsymbol{U}_{-n_\perp}^\top. \tag{48}$$

Because $n_\perp \leq d - r$, the columns of $\boldsymbol{U}_{-n_\perp}$ can be chosen entirely from the zero-eigenspace of $\boldsymbol{C}_T$, which equals $N_{\boldsymbol{x}_0}\mathcal{M}$. Thus,

$$\mathrm{Im}((\boldsymbol{J}_{\boldsymbol{\phi}^\perp}(\boldsymbol{x}_0))^\top) = \mathrm{Span}(\boldsymbol{U}_{-n_\perp}) \subseteq N_{\boldsymbol{x}_0}\mathcal{M}. \tag{49}$$

For a target $\boldsymbol{x}$, define the two local feature-gradient components at $\hat{\boldsymbol{x}} = \boldsymbol{x}_0$ by

$$\boldsymbol{g}^\parallel := (\boldsymbol{J}_{\boldsymbol{\phi}^\parallel}(\boldsymbol{x}_0))^\top \big(\boldsymbol{\phi}^\parallel(\boldsymbol{x}_0) - \boldsymbol{\phi}^\parallel(\boldsymbol{x})\big), \tag{50}$$

$$\boldsymbol{g}^\perp := (\boldsymbol{J}_{\boldsymbol{\phi}^\perp}(\boldsymbol{x}_0))^\top \big(\boldsymbol{\phi}^\perp(\boldsymbol{x}_0) - \boldsymbol{\phi}^\perp(\boldsymbol{x})\big). \tag{51}$$

The two image-space inclusions above imply $\boldsymbol{g}^\parallel \in T_{\boldsymbol{x}_0}\mathcal{M}$ and $\boldsymbol{g}^\perp \in N_{\boldsymbol{x}_0}\mathcal{M}$. Substituting these definitions into the gradient identity in Eq. (3) yields

$$\nabla_{\hat{\boldsymbol{x}}}\mathcal{L} = \lambda^\parallel \boldsymbol{g}^\parallel + \lambda^\perp \boldsymbol{g}^\perp, \tag{52}$$

which proves the claimed directional decomposition. □

B Scope of the Analysis and Limitations

Our characterization is local and holds to leading order in the corruption level; finite corruption, strong curvature, manifold boundaries, and imperfect optimization can therefore mix the two components. The corruption distribution, the branch dimensions $n_\parallel$ and $n_\perp$, and the loss weights $\lambda^\parallel$ and $\lambda^\perp$ must be selected without knowledge of the intrinsic dimension, and the feature geometry requires a separate representation-training stage for each domain. Our experiments concern sparse structured distributions and single-step generators; broader validation across architectures, modalities, and scale remains to be established.

C Experimental Detail

C.1 Shared Setup for the Helix, Earth, and Torsion Experiments

The helix, Earth, and torsion-angle experiments use a common training and evaluation framework. RFM, MF, and RMF use the same network architecture, consisting of four hidden layers of width 256 and a Fourier time embedding with four frequency octaves. Each method is independently initialized with seed 0, and evaluation uses exponential-moving-average weights with decay 0.999. The flow-based methods are evaluated at $\mathrm{NFE} \in \{1, 2, 5, 10\}$, whereas OTM and OTM + MaDeL directly generate samples with one network evaluation.

For OTM and OTM + MaDeL, we compute an exact Hungarian assignment between generated and real samples using ambient squared distance. Both methods therefore use identical generated-to-data pairings and differ only in their regression losses. For Earth and torsion angles, OTM + MaDeL uses

$$\mathcal{L}_{\mathrm{OTM+MaDeL}} = \|\hat{\boldsymbol{x}} - \boldsymbol{x}_\pi\|_2^2 + 0.1\, h_c\Big(\boldsymbol{\phi}^{\parallel}(\hat{\boldsymbol{x}}) - \boldsymbol{\phi}^{\parallel}(\boldsymbol{x}_\pi)\Big) + 0.1\, h_c\big(\boldsymbol{\phi}^{\perp}(\hat{\boldsymbol{x}}) - \boldsymbol{\phi}^{\perp}(\boldsymbol{x}_\pi)\big)\,, \quad (53)$$

where $\boldsymbol{x}_\pi$ is the assigned real sample and $h_c(\boldsymbol{u}) = \sqrt{\|\boldsymbol{u}\|_2^2 + c^2} - c$ with $c = 0.01$. The helix experiment omits the ambient term and uses feature weights $\lambda^{\parallel} = 1$ and $\lambda^{\perp} = 5$.

The feature encoder is trained separately and frozen during generative training. Its parallel decoder predicts the clean sample, while its orthogonal decoder predicts the normalized corruption $\boldsymbol{\delta}/\operatorname{rms}(\boldsymbol{\delta})$. Normalizing the corruption prevents the orthogonal representation from collapsing when the corruption magnitude is small.

C.2 Helix on Sphere

We construct a spherical helix on $\mathbb{S}^2$ with winding parameter 5.0, latitude oscillation 0.5, and parameter range $[-\pi, \pi]$. Clean samples are obtained by applying tangent Gaussian perturbations with standard deviation $\sigma_{\mathrm{signal}} = 0.01$ and are then embedded into $\mathbb{S}^{512} \subset \mathbb{R}^{513}$ through a fixed random orthonormal map from $\mathbb{R}^3$ to $\mathbb{R}^{513}$. Encoder inputs are formed by applying an additional Gaussian perturbation with per-coordinate standard deviation $\sigma_{\mathrm{perturb}} = 0.10$ in the ambient tangent space of $\mathbb{S}^{512}$. The clean distribution therefore remains in the three-dimensional image of the embedding, while the corruption can leave this subspace.

We use 50,000 data samples. All generative models are trained for 50,000 iterations using RAdam with learning rate 10^{-3}, batch size 256, gradient clipping at 2.0, and EMA decay 0.999. The encoder uses $n_{\parallel} = 3$ parallel and $n_{\perp} = 509$ orthogonal features and is trained with clean- and corruption-prediction weights 1 and 5, respectively. OTM and OTM + MaDeL use a three-dimensional Gaussian prior and exact Hungarian assignment. For MF, the prior standard deviation is $1/\sqrt{513}$ so that its expected norm matches the unit-norm data.

Evaluation uses 2,048 generated samples. The qualitative comparison in Fig. 4 uses $\mathrm{NFE} = 1$ and projects samples back to the underlying $\mathbb{S}^2$ coordinates. We additionally monitor sliced W_2^2, energy distance, distance to the underlying helix, displacement outside the embedding subspace, and coverage along the helix.

C.3 Earth and Climate Science

The Earth benchmark contains event locations on $\mathbb{S}^2$ for volcanoes, earthquakes, floods, and wildfires, with 827, 6,120, 4,875, and 12,809 records, respectively. We use deterministic $80/10/10$ train, validation, and test splits with seed 42. Duplicate coordinates are retained during training because repeated events at one location encode the empirical event density.

Clean locations are corrupted in the tangent space of $\mathbb{S}^2$ using Gaussian noise with standard deviation 0.15. The encoder uses $n_{\parallel} = n_{\perp} = 2$ and equal clean- and corruption-prediction weights. It is trained for 10,000 iterations with batch size 2,048. Generative models are trained for 20,000 iterations using Adam with learning rate 10^{-3} and EMA decay 0.999. Flow-based methods use batch size 2,048, while the exact-Hungarian OTM variants use batch size 256.

We report geodesic W_1 distance between generated and real event distributions. The cost is the great-circle distance on $\mathbb{S}^2$, and the empirical transport problem is solved by exact Hungarian assignment. Datasets with more than 2,048 records use a fixed seeded reference subsample, shared across all methods and checkpoints. Distances are converted from radians to degrees for Tab. 2. The qualitative Earthquake comparison in Fig. 4 uses 4,096 generated samples at $\mathrm{NFE} = 1$.

C.4 Protein and RNA Torsion Angles

We use the torsion-angle benchmarks of Riemannian Flow Matching (Chen & Lipman, 2024), following the preprocessing of Riemannian diffusion models (Huang et al., 2022; Bortoli et al., 2022). The benchmark contains four two-dimensional Ramachandran distributions and one seven-dimensional RNA torsion distribution:

Dataset	Manifold	Number of records
General	$\mathbb{T}^2$	137,718
Glycine	$\mathbb{T}^2$	13,183
Proline	$\mathbb{T}^2$	7,618
Pre-Pro	$\mathbb{T}^2$	6,900
RNA	$\mathbb{T}^7$	9,457

Each angle is represented by its cosine and sine, embedding $\mathbb{T}^d$ isometrically into $\mathbb{R}^{2d}$ without an angular discontinuity. Corruption is applied independently to each angle using wrapped Gaussian noise with standard deviation 0.15 radians.

All generative models use batch size 256 and are trained for 20,000 iterations using Adam with learning rate 10^{-3} and EMA decay 0.999. The encoder is trained separately for 10,000 iterations with batch size 2,048. Its parallel and orthogonal feature dimensions are: $n_{\|} = n_{\perp} = 2$ for $\mathbb{T}^2$ and $n_{\|} = n_{\perp} = 7$ for $\mathbb{T}^7$.

We evaluate geodesic W_1 using exact Hungarian assignment under the flat-torus distance, defined as the Euclidean norm of the wrapped per-angle differences. The generated and test distributions are restricted to at most 2,048 samples using a fixed seeded subsample, and the resulting distances are reported in degrees in Tab. 2. All values in the main table use $\mathrm{NFE} = 1$.

C.5 SCOPe Protein Backbone Generation

Dataset and generator. We use 3,673 SCOPe backbones containing between 60 and 128 residues (Chandonia et al., 2022), matching the dataset used by RMF (Woo et al., 2026). Each backbone is represented by residue-local rigid frames in $\mathrm{SE}(3)^L$. The generator follows the FrameFlow architecture (Yim et al., 2023a), using node and edge dimensions 256 and 128, six invariant-point-attention blocks, eight attention heads, and self-conditioning.

Both the baseline and MaDeL models use the RMF SplitMeanFlow objective, comprising a flow-matching anchor and a semigroup composition term. Rotational and translational losses are weighted by 1 and 2, respectively, with translations scaled by 0.1. The time floor is 0.01, and both objective terms are clipped at 10 after 10,000 iterations. Models are trained with AdamW using learning rate 10^{-4}, weight decay 0.01, gradient clipping at 1.0, and EMA decay 0.9999. Batches contain at most 54 proteins under a length-dependent memory budget. Although training runs for up to two million iterations, the reported steric evaluation uses the one-million-step checkpoints.

Frozen protein encoder. The protein feature encoder is a split-latent U-Net over backbone frames with 64 feature planes, encoder depths $[4, 4, 4, 4]$, and decoder depths $[2, 2, 2]$. It is trained for 250,000 iterations using batch size 2,048, learning rate 10^{-4}, and equal clean- and corruption-prediction weights. Structured corruptions include rigid segment and loop displacements, chain breaks, and atom-level bond, angle, and torsion perturbations. Each of the 13 corruption operations is independently activated according to a base probability of 0.15, producing approximately two corruptions per sample. The EMA encoder checkpoint is frozen during RMF training.

The baseline sets the feature-semigroup weight to zero. The MaDeL model adds the decomposed feature semigroup term with weight 0.0029 while retaining the standard RMF objective. All other architecture, optimization, conditioning, and sampling settings are identical between the two runs.

Designability and self-consistency. For each checkpoint, backbone length, and NFE, we generate 64 backbones from identical priors for both models. Following standard protein-backbone evaluation, ProteinMPNN (Dauparas et al., 2022) generates eight sequences per backbone at temperature 0.1, which are refolded using ESMFold (Lin et al., 2023). Self-consistency RMSD is the minimum RMSD over the eight refolded sequences, and a backbone is considered designable when this value is below 2.0 Å.

Steric validity. We evaluate the one-million-step checkpoints at lengths

$$\{60, 68, 76, 84, 92, 100, 108, 116, 124, 128\}$$

and $\mathrm{NFE} \in \{1, 5, 10\}$. For each length and NFE, we draw eight independent batches of 128 samples, using identical priors for the baseline and MaDeL within each batch.

For every generated structure, we compute the maximum van der Waals overlap

$$\max_{(a,b)} \left(r_a + r_b - \|\boldsymbol{x}_a - \boldsymbol{x}_b\|_2\right) \tag{54}$$

over backbone N, C_α, and C atoms belonging to residues separated by at least two sequence positions. We use Bondi radii of 1.55 Å for nitrogen and 1.70 Å for carbon (Bondi, 1964). Backbone oxygen atoms are excluded because their placement is determined by the PDB-writing convention rather than directly emitted by the model. The reported statistic is the median maximum overlap within each batch. Real SCOPe backbones within four residues of each target length provide the reference distribution and have median overlaps of 0.28–0.30 Å.