# Start Where the Data Are

## Statistical Priors for One-Step Generation on Manifolds

Yuyao Wang[1] Haoliang Sun[2†] Xiankai Lu[2] Xinxin Zhang[2] Yongshun Gong[2†]

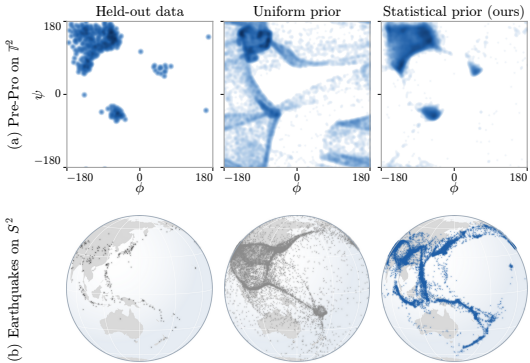
[1]Department of Mathematics and Statistics, Boston University
[2]School of Software, Shandong University [†]Corresponding authors

**arxiv.org/abs/2611.31856**

**github.com/olivia3395/statprior-onestep**

# One-step generators put samples where no data are

- Many scientific variables live on **manifolds**: protein torsion angles on tori, event locations on the sphere, rotations in SO(3).
- **One-step generators** (flow maps, Riemannian MeanFlows) sample with a **single network call**.
- But starting from **uniform** noise, they place mass in the **low-density gaps** between modes:
  - protein angles in sterically forbidden regions,
  - earthquakes in the open ocean.




Same network and loss; only the prior and pairing differ.

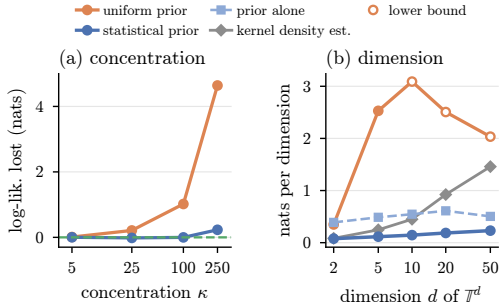
# Why: uniform noise asks for the most contraction

**Proposition 1: average log-volume change**

For a $C^1$ diffeomorphism $T$ of a compact manifold with $T_{\#}p_0 = p_1$,

$$\mathbb{E}_{p_0}[\log J_T] = H(p_1) - H(p_0),$$
$$-\mathbb{E}_{p_0}[\log J_T] \leq \mathrm{KL}(p_1 \,\|\, \mathscr{U}),$$

with equality when $p_0 = \mathscr{U}$.




(a) Sharper vMF modes ($\kappa$: 5 → 250): uniform prior loses 0.01 → 4.6 nats, ours stays below 0.23.

A **uniform** start demands the largest squeeze toward the data; one-step errors track it (rank corr. 0.90).

# Idea: start where the data are

**1. Fit a prior once**
mixture of tangent-space Gaussians on $\mathscr{M}$

→

**2. Pair by the posterior**
noise drawn from each point's own component

→

**3. Train as usual**
same network, loss and one-step sampler

The prior captures the **coarse modal structure**; the generator only learns the **remaining local transport**.

Works with any one-step generator whose training allows a new source distribution and a new noise–data coupling.

# Geometry-aware mixture priors

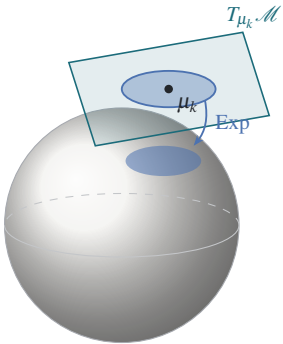
**Component:** a truncated tangent-space Gaussian mapped by $\mathrm{Exp}_\mu$,

$$q(x;\mu,\Sigma) = \frac{\mathcal{N}(v;0,\Sigma)}{Z(\Sigma)\, j(v)}, \quad v = B_\mu^\top \mathrm{Log}_\mu x.$$

- Volume Jacobian $j(v)$: $\frac{\sin r}{r}$ on $S^2$, $\frac{2(1-\cos r)}{r^2}$ on $\mathrm{SO}(3)$, 1 on $\mathbb{T}^d$.
- Valid density, exact sampling, tractable posterior $\hat{\gamma}(k \mid x)$.

**Fitting:** EM responsibilities + geometric moment updates; $K$ by BIC.

**Sampling prior:** $p_0 = \sum_k \hat{w}_k\, q(\cdot; \hat{\mu}_k, s\hat{\Sigma}_k)$, $s = 2$.




one component on $\mathcal{M}$

# Posterior coupling: short pairs, prior preserved

**Problem.** Independent pairing $p_0 \otimes p_1$ links noise near one mode to data at another, even when $p_0 = p_1$.

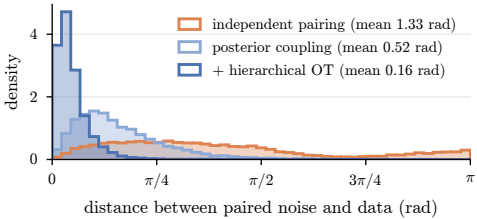
**Fix.** For each data point $x$: $k \sim \hat{\gamma}(\cdot \mid x)$, $\varepsilon \sim q(\cdot; \hat{\mu}_k, s\hat{\Sigma}_k)$.

**Proposition 2: training-noise marginal**

With $\bar{w}_k = \frac{1}{n}\sum_i \hat{\gamma}(k \mid x_i)$, training noise follows $\bar{p}_0 = \sum_k \bar{w}_k\, q(\cdot; \hat{\mu}_k, s\hat{\Sigma}_k)$, and

$$\mathrm{TV}(\bar{p}_0, p_0) \le \tfrac{1}{2}\sum_k |\bar{w}_k - \hat{w}_k|$$

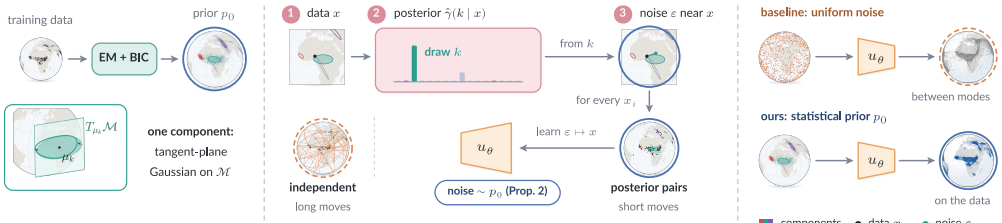
(at most **0.030** over all standard fits).




Distance between paired noise and data: independent 1.33 rad, posterior 0.52 rad. Optional within-component matching shortens it further.

No transport solver; training noise matches the sampling prior.

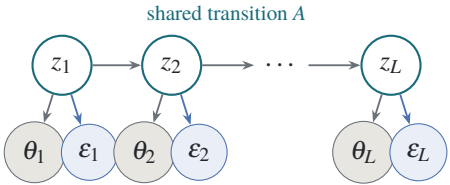
# The whole pipeline




(a) Fit a mixture of tangent-space Gaussians once. (b) Each data point draws a component from its posterior, then noise from that component: pairs are short and the noise still follows $p_0$. (c) Same one-step network: uniform noise lands between modes, the statistical prior lands on the data.

# Long chains: a hidden Markov prior

- A full-covariance mixture on $(\mathbb{T}^2)^L$ needs $O(L^2)$ covariance parameters per component.
- **HMM prior** (after TorusDBN): latent states $z_j$ form a Markov chain; each emits a tangent Gaussian on $\mathbb{T}^2$. Parameters are shared, so the count does not grow with $L$.
- **Posterior path coupling**: forward filtering, backward sampling of $z \mid \theta$, then noise $\varepsilon_j \sim q(\cdot; \hat{\mu}_{z_j}, s\hat{\Sigma}_{z_j})$. Cost $O(LS^2)$.




On 10-residue fragments: held-out log-likelihood $-7.8$ vs. $-17.5$ for a 32-component mixture with $10\times$ the parameters.

# A drop-in change

**Prior fitting** (once)
fit $(\hat{w}_k, \hat{\mu}_k, \hat{\Sigma}_k)$, choose $K$ by BIC

**Training** (per minibatch)
$k_i \sim \hat{\gamma}(\cdot \mid x_i), \quad \varepsilon_i \sim q(\cdot; \hat{\mu}_{k_i}, s\hat{\Sigma}_{k_i})$
update the generator with its **own loss** on $(\varepsilon_i, x_i)$

**Sampling**
$\varepsilon \sim p_0$, then **one network evaluation**

| | changed? |
|---|---|
| Source distribution | **yes** |
| Noise–data pairing | **yes** |
| Network architecture | no |
| Training objective | no |
| Sampler, network calls | no (1 call) |

Prior fit: 11 s to 8 min. Added cost per training step: at most 23%.

# Setup

## Data

**Torus:** Top500 backbone $(\phi, \psi)$ in 4 classes; side chains of 14 residue types; windows up to $\mathbb{T}^{10}$; fragments $\mathbb{T}^{20}, \mathbb{T}^{40}$; RNA on $\mathbb{T}^{7}$.

**Sphere:** volcanoes, earthquakes, floods, fires.

**SO(3)** and synthetic sets with known densities (tori up to $d = 50$).

## Generators

Five one-step generators, each with its original architecture, loss and schedule:

- RMF (Zhong et al.)
- RMF-E / -L / -S (Woo et al.)
- G-LSD generalized flow maps (Davis et al.)

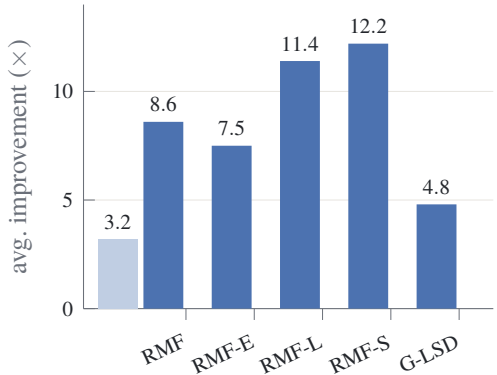
## Metric

Unbiased $\mathrm{MMD}^2$ to held-out data, divided by its **noise floor** (spread between disjoint sets of real data).

$\leq 1$: indistinguishable from real data at that scale.

Plus protein geometry: Ramachandran and rotamer outliers, C$\alpha$ clashes.

# It improves every one-step generator




Geometric mean over 9 datasets of uniform-prior error / our error. “Prior alone”: samples from the fitted prior, no network, vs. RMF with the uniform prior.

**45 / 45**

generator–dataset pairs improved; 40 by more than 3 noise-floor standard deviations

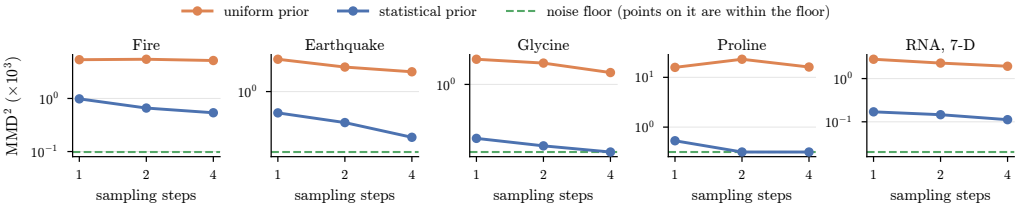
**3.4–125×**

lower error on torsion angles (1.3–6× on Earth data)

**$\leq 1$**

noise floor reached under the RMF kernel on all protein and RNA datasets

# One step from our prior beats more steps from uniform




One step from the statistical prior: **3–30**× lower error than **four** steps from the uniform prior, on every dataset.

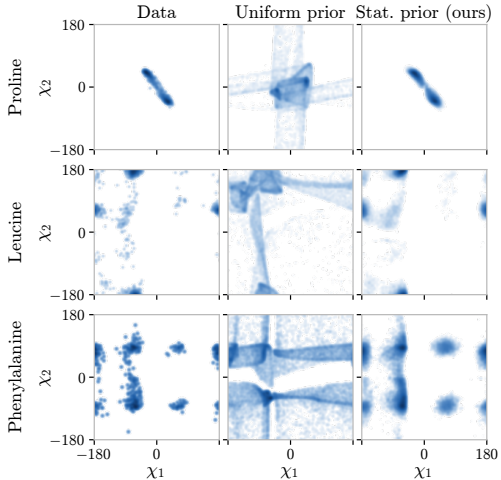
Proline: 5× longer training + 4 steps bring the uniform prior to 6.3; ours reaches **3.4** in one step at the original budget.

# Proteins: samples obey structural chemistry

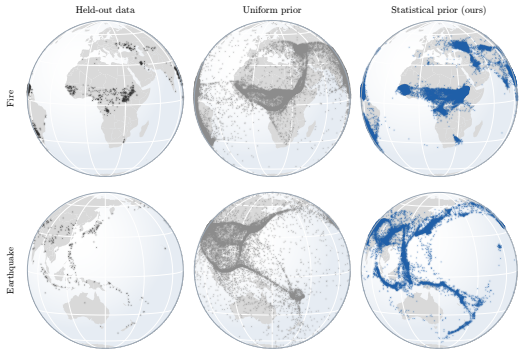
| | uniform | ours |
|---|---|---|
| Usable fragments | 34% → | **93%** |
| Ramachandran outliers | 12.1% → | **0.55%** |
| Rotamer outliers | 43% → | **3.9%** |

Real data: 95%, 0.24%, 1.9%. Four uniform-prior steps reach only 69% usable fragments.

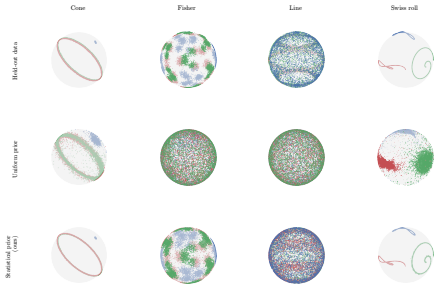
> Uniform-prior samples form bands through forbidden side-chain regions; ours stay on the rotamer modes.



# Earth and rotations



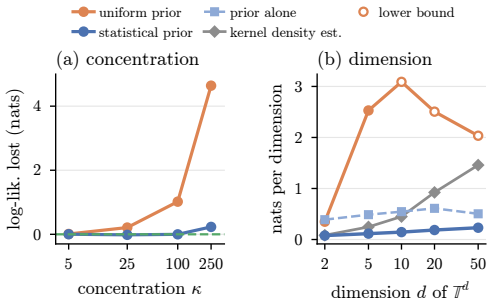

Fire and earthquakes: our samples follow the fire regions and seismic belts; uniform-prior samples spread over the ocean.




| MMD ↓ | Cone | Fisher | Line | Swiss |
|---|---|---|---|---|
| best reported | 0.044 | 0.039 | 0.035 | 0.032 |
| ours | **0.012** | **0.025** | **0.025** | **0.017** |

SO(3): lowest error on all four benchmarks, 1.4–3.7× below the best result reported by Zhong et al.

# Higher dimensions and longer chains




(b) Synthetic tori, $d = 2$ to 50, misspecified mixture: ours loses 0.08–0.23 nats per dimension, lowest at every $d$.

**Protein windows.** $\mathbb{T}^4, \mathbb{T}^6$: within the noise floor. $\mathbb{T}^{10}$: 9.2 vs. 305–540 for uniform-prior baselines.

**10-residue fragments rebuilt in 3D**

| | clash ↓ | helix | sheet |
|---|---|---|---|
| held-out fragments | 0.012 | 0.496 | 0.415 |
| RMF + uniform prior | 0.164 | 0.417 | 0.225 |
| RMF + mixture prior | 0.055 | 0.486 | 0.341 |
| RMF + hidden-Markov prior | **0.043** | **0.501** | **0.407** |

# What does each part do?

| finest-scale error | uniform | fitted prior | + coupling |
|---|---|---|---|
| Cone (SO(3)) | 158 | $\leq 1$ | $\leq 1$ |
| Pre-Pro | 58 | 52 | **6.5** |
| Fire | 110 | 123 | **20** |

Noise-floor multiples. Under the coarser RMF kernel, the fitted prior alone already fixes Pre-Pro ($83 \rightarrow 2.8$).

**The prior places the mass; the coupling recovers its shape.**

- **The generator still matters.** Fire: uniform 110, prior alone 35, after one step **20**. It cuts the prior-only error by 2.1–12$\times$ on torsion angles.
- **No memorization.** Memorization ratio 1.0–1.2 on torsion angles (ideal: 1), vs. 1.8–6.4 from the uniform prior.
- **Robust to the prior.** Fitted on 500 points or with EM stopped after 3 iterations: still 2.4–9.3$\times$ better. Misspecified (thin rings, long arcs): 3.7–19$\times$.

SUMMARY

# Takeaways

## Why

For a fixed target on a compact manifold, a uniform start requires the largest average volume contraction. Errors grow with it.

## How

Fit a geometry-aware mixture (or HMM) prior once; pair each data point with noise from its posterior component.

## Result

Drop-in for five generators: up to **125**× lower error, Ramachandran outliers **12.1%** → **0.55%**, still one network call.

# Thank you



github.com/olivia3395/statprior-onestep

arxiv.org/abs/2611.31856

yuyaow@bu.edu haolsun@sdu.edu.cn ysgong@sdu.edu.cn