TDA-Guided Temporal Sampling for Multifidelity Flow Matching

Cristian Villatoro

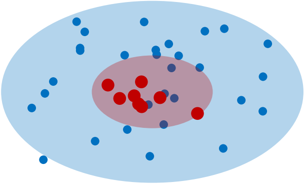
December 10, 2025

Outline

The Data Scarcity Challenge

Scientific Computing Reality:

- High-fidelity (HF) simulations are **expensive**
- Days of compute time per run
- Limited number of samples



Low-fidelity (LF) simulations are cheap:

- Minutes per run
- Can afford many samples
- Captures trends, but less accurate

Goal: Learn generative model $p_{\mathrm{HF}}(z|x)$ with limited HF data

Why Generative Models?

Beyond point prediction: We need to generate *samples* from $p_{\mathrm{HF}}(z|x)$

Applications:

- Uncertainty quantification
- Scenario analysis
- Data augmentation
- Rare event sampling
- Ensemble forecasting

Challenge:

- GANs, VAEs, Diffusion models need **lots of data**
- With N_{HF} small, standard methods fail
- Need to leverage LF information

Key Insight: LF and HF are *correlated*—can we transfer knowledge?

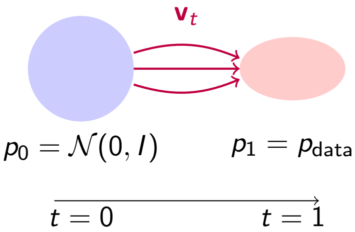
Flow Matching: Core Idea

Goal: Transform noise $z_0 \sim \mathcal{N}(0, I)$ into data $z_1 \sim p_{\text{data}}$

Learn a **vector field** $\mathbf{v}_t(z, t)$ defining particle flow:

$$\frac{dz_t}{dt} = \mathbf{v}_t(z_t, t), \quad z_0 \sim \mathcal{N}(0, I)$$

Integrating from $t = 0$ to $t = 1$ gives sample $z_1 \sim p_{\text{data}}$



Training: Regress onto ground truth vector field

$$\mathcal{L} = \mathbb{E}_{t,z_1,z_t}\left[\|\mathbf{v}_t(z_t) - \mathbf{u}_t(z_t|z_1)\|^2\right]$$

Conditional Flow Matching

Linear interpolation path:

$$z_t = (1 - t) \cdot z_0 + t \cdot z_1, \quad z_0 \sim \mathcal{N}(0, I), \quad z_1 = z^{\text{data}}$$

Ground truth vector field:

$$\mathbf{u}_t(z_t | z_1) = \frac{z_1 - z_t}{1 - t}$$

Training objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1]} \mathbb{E}_{z_0, z_1} \left[\| \mathbf{v}_t(z_t; \theta) - \mathbf{u}_t(z_t | z_1) \|^2 \right]$$

Key: Time t sampled **uniformly**—but is this optimal?

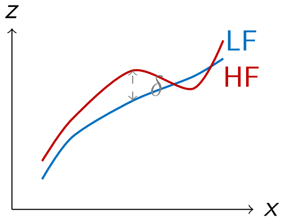
The Multifidelity Paradigm

Observation: LF and HF outputs are *correlated*

Classical decomposition:

$$f_{\mathrm{HF}}(x) = \rho \cdot f_{\mathrm{LF}}(x) + \delta(x)$$

- ρ: scaling factor
- $\delta(x)$: residual discrepancy



Strategy:

1. Learn f_{LF} from abundant LF data
2. Learn ρ, δ from limited HF data
3. Use LF as strong prior for HF

Multifidelity for Generative Models

Our extension: Apply multifidelity to *vector fields*

Multifidelity Vector Field Structure

$$\mathbf{v}_t^{\mathrm{HF}}(z \mid x) = \underbrace{A \cdot \mathbf{v}_t^{\mathrm{LF}}(z \mid x)}_{\text{scaled LF dynamics}} + \underbrace{B \cdot x + C}_{\text{affine correction}} + \underbrace{\Delta(z, x, \mathbf{v}_t^{\mathrm{LF}})}_{\text{residual network}}$$

Components:

- $\mathbf{v}_t^{\mathrm{LF}}$: Pre-trained LF vector field (**frozen**)
- A, B, C: Learnable affine parameters
- Δ: Neural network for complex residual dynamics

Training: Only learn (A, B, C, Δ) from limited HF data

The Key Question

Where should we focus learning?

Standard flow matching samples $t \sim \mathcal{U}[0, 1]$ uniformly

But different time points have different **complexity**:

- Some t: LF dynamics $\approx$ HF dynamics (simple correction)
- Some t: LF dynamics $\neq$ HF dynamics (complex correction needed)

Our Insight

Use **Topological Data Analysis (TDA)** to identify *when* the multifidelity correction Δ is most complex, then sample those times more frequently.

Temporal Complexity Analysis with TDA

For each time $t \in [0, 1]$:

1. Compute vector field residual:

$$\mathbf{r}_t = \underbrace{\mathbf{u}_t^{\text{true}}}_{\text{ground truth HF}} - \underbrace{\mathbf{v}_t^{\text{HF,pred}}}_{\text{MF prediction}}$$

2. Build point cloud of residuals across all samples
3. Compute **persistent homology** of residual structure:
   - H_0: Multiple clusters $\Rightarrow$ missing dynamical modes
   - H_1: Loops $\Rightarrow$ rotational dynamics not captured
4. Extract complexity score: $\mathcal{C}(t) = \sum$ persistence

Result: Temporal complexity map $\mathcal{C} : [0, 1] \to \mathbb{R}_+$

TDA-Guided Time Sampling

Standard (Uniform):

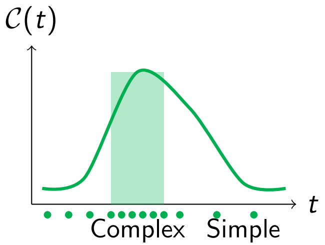
$$t \sim \mathcal{U}[0, 1]$$

Equal probability for all times

TDA-Guided:

$$t \sim p_{\mathcal{C}}(t) \propto \mathcal{C}(t) + \epsilon$$

More samples at complex times



More samples where $\mathcal{C}(t)$ is high

Physical interpretation: High $\mathcal{C}(t)$ often corresponds to:

- Phase transitions
- Bifurcations
- Regime changes

Flexible Sampling Strategies

Strategy	**TDA Recomputation**	**Sampling**	**Use Case**
Uniform	Never	Always $t \sim \mathcal{U}[0,1]$	Baseline
TDA	Every epoch	Always $t \sim p_{\mathcal{C}}(t)$	Max efficiency
Mixed	Every k epochs	$t \sim p_{\mathcal{C}}(t)$ until next	Balance

Key insight: Recompute $\mathcal{C}(t)$ periodically as model improves!

Example: Mixed with $k = 50$

- Epoch 0: Compute $\mathcal{C}(t)$, sample $t \sim p_{\mathcal{C}}$
- Epochs 1–49: Use same $\mathcal{C}(t)$, sample $t \sim p_{\mathcal{C}}$
- Epoch 50: **Recompute** $\mathcal{C}(t)$ with improved model
- Epochs 51–99: Use updated $\mathcal{C}(t)$
- ...

Benefit: Complexity map adapts as model learns

Complete Algorithm

Algorithm 1 TDA-Guided Multifidelity Flow Matching

Input: LF data $\mathcal{D}_{\mathrm{LF}}$, HF data $\mathcal{D}_{\mathrm{HF}}$, PH computation frequency k

Train LF flow: $\mathbf{v}_t^{\mathrm{LF}} \leftarrow \mathrm{FlowMatching}(\mathcal{D}_{\mathrm{LF}})$
Initialize: A, B, C, Δ

for epoch $= 1$ to N **do**
  if epoch $\bmod k = 0$ **then** ▷ Periodic recomputation
    $\mathcal{C}(t) \leftarrow \mathrm{TDA\_Analysis}(\text{paired data}, \mathbf{v}_t^{\mathrm{LF}}, A, B, C, \Delta)$ ▷ Rips filtration
    Build sampler $p_{\mathcal{C}}(t) \propto \mathcal{C}(t) + \epsilon$
  end if

  Sample time: $t \sim p_{\mathcal{C}}(t)$ ▷ Use current complexity map
  Compute: $\mathbf{v}_t^{\mathrm{HF}} = A \cdot \mathbf{v}_t^{\mathrm{LF}} + B \cdot x + C + \Delta$
  Update (A, B, C, Δ) via $\mathcal{L} = \|\mathbf{v}_t^{\mathrm{HF}} - \mathbf{u}_t^{\mathrm{true}}\|^2$
end for

Experimental Setup

Synthetic Benchmark:

- LF: $z^{\text{LF}} = \sin(2x) + \epsilon$
- HF: $z^{\text{HF}} = \sin(2x) + 2 \cdot \mathbb{I}_{|x|<1} + \epsilon$
- Step discontinuity in $|x| < 1$ region that LF misses

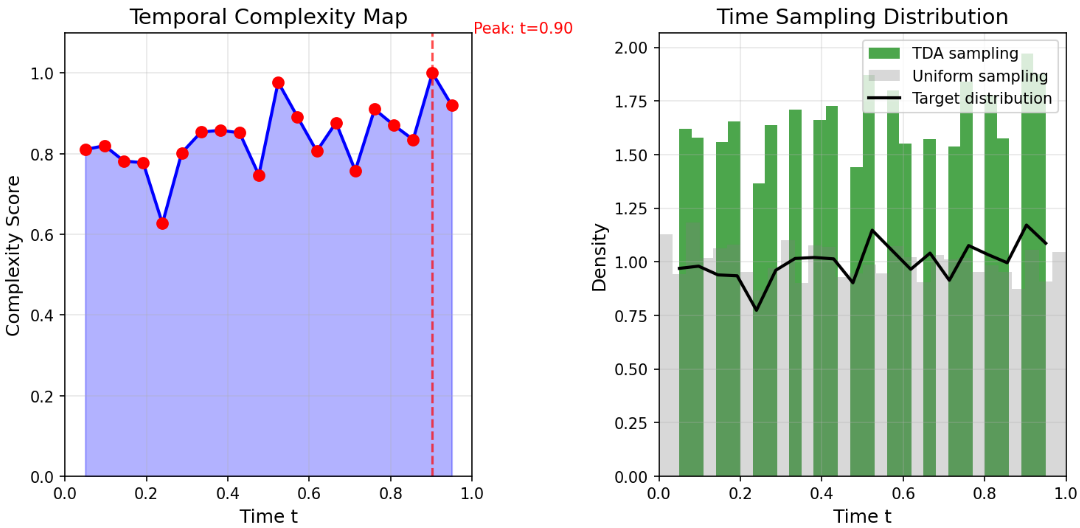
Data:

- $N_{\text{LF}} = 1000$ samples, $N_{\text{HF}} = 100$ samples
- TDA recomputation every 100 epochs for 5000 epochs

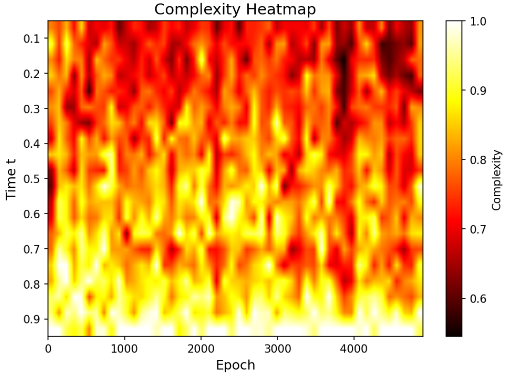
Comparisons:

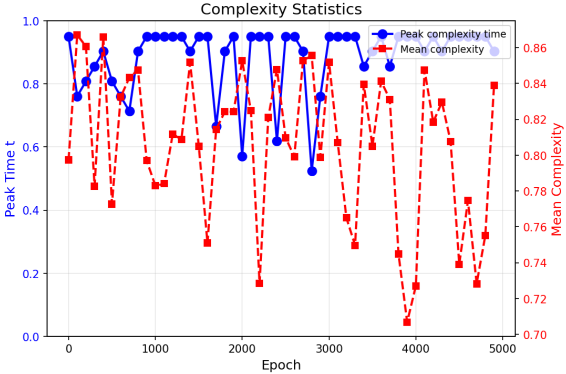
1. Uniform time sampling (baseline)
2. TDA-guided (recompute & sample adaptively)
3. Mixed (recompute periodically, always sample adaptively)

Temporal Complexity Map



Complexity Evolution During Training

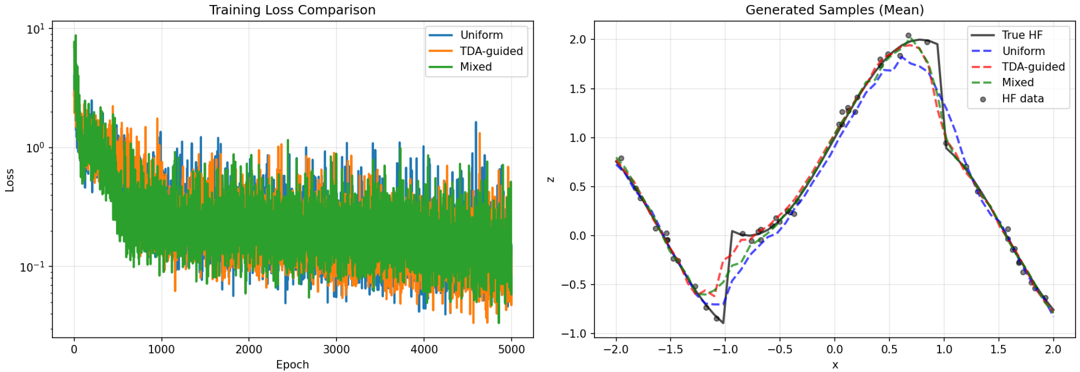




Left: Complexity heatmap over time and epochs **Right:** Complexity statistics

Key: Complexity map adapts as model learns and periodic recomputation matters!

Training Loss and Generated Samples



Left: Training loss comparison (log scale), TDA methods converge differently
Right: Generated samples capture HF step function; TDA-guided is most accurate

Results Summary

Strategy	TDA Recomputations	TDA Sampling	Benefit	MSE
Uniform	0	0%	baseline	0.027355
TDA-guided	every epoch	100%	faster convergence	0.021788
Mixed	every 100 epochs	1%	balanced	0.016657

Key findings:

- TDA identifies critical time points where LF→HF correction is complex
- Adaptive sampling focuses learning on informative regions
- Computational overhead: Persistent Homology calculations are worth considering based on dimensionality of data and number of samples

Summary

Contributions:

1. **TDA Temporal Analysis:** Identify when corrections are complex
2. **Adaptive Time Sampling:** Focus learning on informative times
3. **Flexible Strategies:** Uniform, TDA, or mixed sampling

Key Result: TDA-guided sampling improves learning efficiency by focusing on critical flow times

Future Work

- **Spatial-temporal TDA:** Joint analysis in space and time
- **Multi-level hierarchies:** LowF $\rightarrow$ MedF $\rightarrow$ HighF chains
- **Uncertainty-aware sampling:** Combine topology with epistemic uncertainty
- **Real-world applications:**
  - Computational fluid dynamics (coarse $\rightarrow$ fine resolution)
- **Theoretical analysis:** Convergence guarantees for Homology-guided sampling

Thank You

Questions? Comments?