

Robust Generative Modeling of Extreme Events: Lipschitz-Regularized Divergences and CVaR- Penalized Wasserstein Flows

Markos A. Katsoulakis



Generative AI for extreme events

The challenge

- Predict the frequency and magnitude of extreme events from data
- Heavy tails: few samples but with dominant contribution
- Generative models struggle with heavy tails
 - Unstable training unless Lipschitz regularization is enforced [Chen et al., IMA I&I 2025]
 - Lipschitz maps of light-tailed sources remain light-tailed
 - Heavy-tail designs: require an accurate tail index to perform well

In this talk: Wasserstein gradient flows

- **Fine-tune** any pre-trained model from its samples alone
- **Physics-inspired:** Wasserstein gradient flow of a free energy
 - Lipschitz-regularized KL for the bulk + CVaR penalty on an observable
 - Tail-agnostic; target enters only via one observable, dimension-free estimation
 - Depth fitted to the target by free-energy dissipation

Some generative models for extreme events

	Model	Tail mechanism	Characteristics
Normalizing flows	TTF Hickling & Prangle 2025	tail-transform layer with learnable tail weights	tail-decay rate as input
	mTAF Laszkiewicz et al. 2022	heavy-tailed base on marginals	estimated tail indices as input
Diffusion	t-EDM Pandey et al. 2025	Student-t noise	tail parameter v as input
	DLPM Shariatian et al. 2025	α -stable (Lévy) noise	stability index α as input
	SHD Liu et al. 2024	heavy-tailed score-based diffusion	tail-decay rate as input
GAN	Tail-GAN Cont et al. 2026	VaR/CVaR of portfolios in the loss	fixed-depth generator
Gradient flow	Lip-KL-GPA Chen, Gu et al. 2025	Lipschitz-regularized KL flow	tail-agnostic, adaptive depth

Common limitation: tail index set a priori, or transport depth too short for the extreme region.

Why fine-tune with a gradient flow? Three capabilities:

$$\partial_t Q_t = \nabla \cdot \left(Q_t \nabla_x \frac{\delta \mathcal{F}}{\delta Q}(Q_t) \right), \quad Q_0 = P_{\text{pre}}$$

1. Start from any data

The initial condition can be any pre-trained model's samples; (GAN, diffusion model, flow,)

$$Q_0 = P_{\text{pre}}$$

“Model-agnostic” no need to start from Gaussian noise, ...

2. Design the free energy

Free to add the quantity of interest **QoI** you care about to the energy the flow minimizes.

$$\mathcal{F} = \text{KL}^L(Q \parallel P_{\text{tar}}) + \lambda \text{QoI}$$

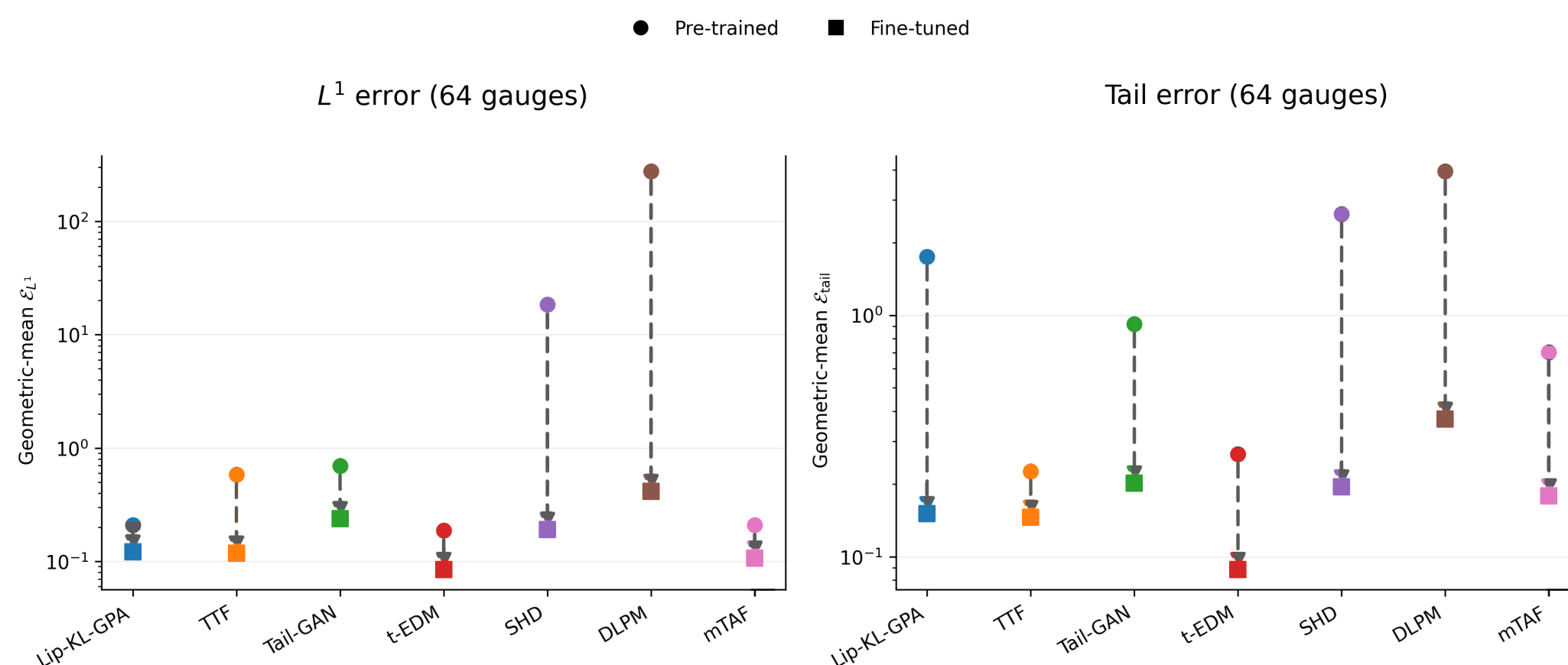
QoI = $(\Delta \text{CVaR})^2$: the flow learns heavy tails—“agnostic”

3. Go as deep as needed

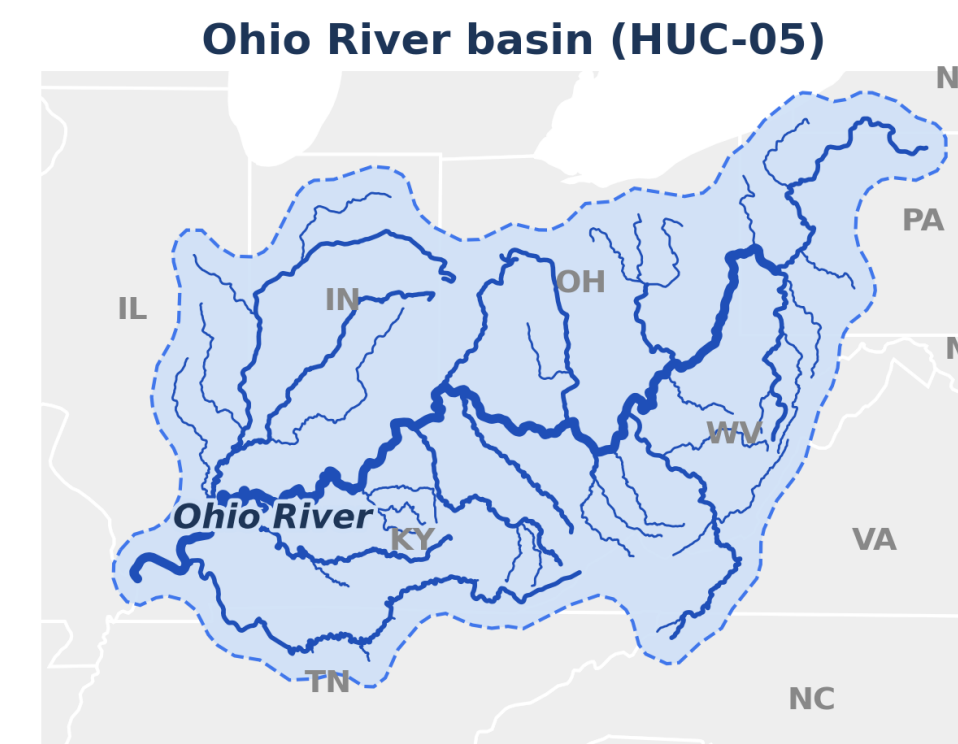
The gradient flow dissipates $\mathcal{F}$.

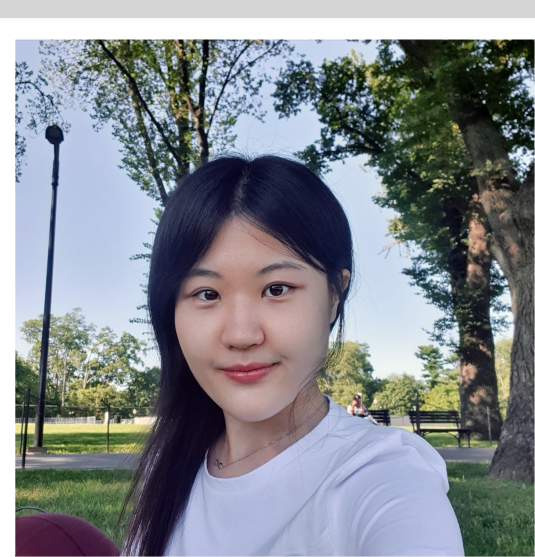
$$\frac{d}{dt} \mathcal{F}(Q_t) = - \int Q_t \left| \nabla_x \frac{\delta \mathcal{F}}{\delta Q} \right|^2 dx \leq 0$$

Architecture adapts to capturing the target

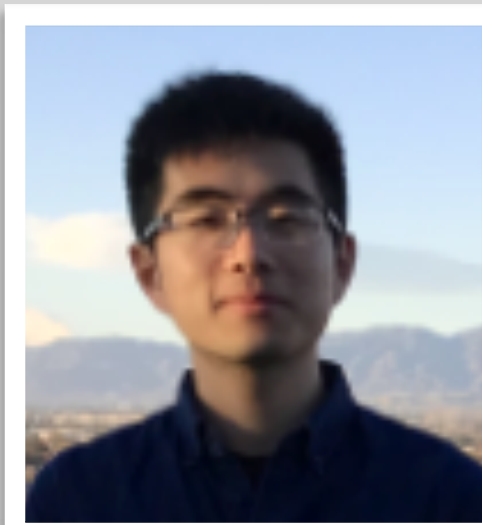


Data set:
Streamflow





Hyemin Gu,
UMass Amherst



Ziyu Chen,
UNC Chapel Hill



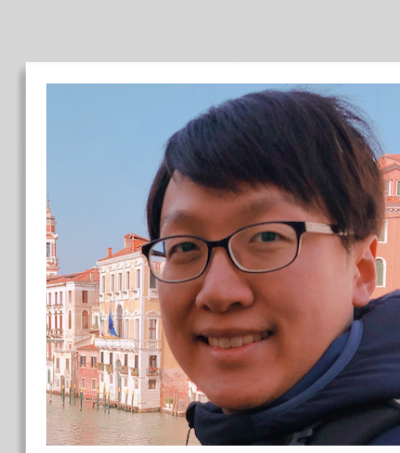
Thejani Gamage,
UMass Amherst



Zhizhen Zhang,
UMass Amherst
(see poster on 5.)



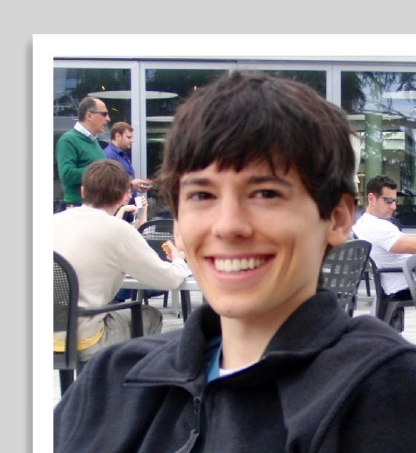
Luc Rey-Bellet
UMass



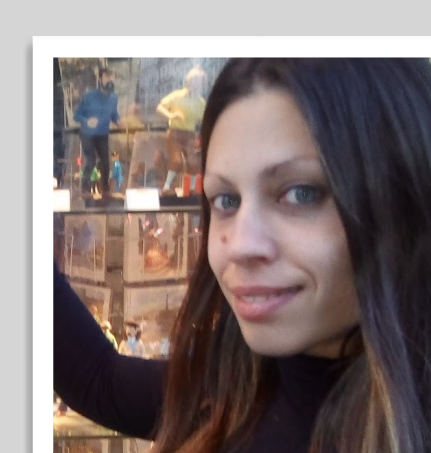
Yu-Chen Cheng,
UBC



Yannis Pantazis,
FORTH, Greece



Jeremiah Birrell,
Texas State

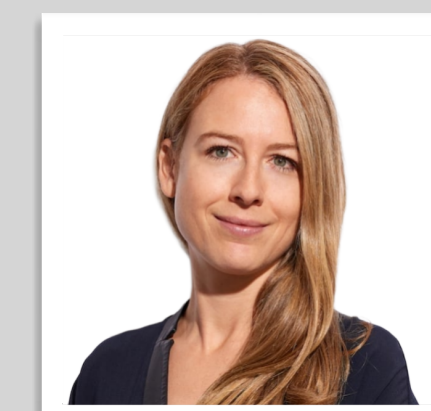


Panagiota Birmpa,
Heriot-Watt, UK

1. J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, L. Rey-Bellet, *(f, Γ) -Divergences: Interpolating between f -Divergences and Integral Probability Metrics*, **Journal of Machine Learning Research & NeurIPS**, (2022)
2. H. Gu, P. Birmpa, Y. Pantazis, M. A. Katsoulakis, and L. Rey-Bellet, *Lipschitz-regularized gradient flows & generative particle algorithms for high-dim. scarce data*, **SIAM Data Science**, (2024)
3. Z. Chen, H. Gu, M. A. Katsoulakis, L. Rey-Bellet, W. Zhu, *Generative Learning with Lipschitz-Regularized α -Divergences Allows Minimal Assumptions on Target Distributions*, **IMA Information & Inference** (2025)
4. Yu-Chen Cheng, H. Gu, M. A. Katsoulakis, F. Michor, et al. *PROFET Predicts Continuous Gene Expression Dynamics from scRNA-seq Data to Elucidate Heterogeneity of Cancer Treatment Responses*, **Cell Systems**, (2026)
5. T. Gamage, H. Gu, Z. Zhang, Z. Chen, M.K. & L. Rey-Bellet, *Fine-Tuning Generative Models for Extreme Events via CVaR-Penalized Wasserstein Gradient Flows*, **Arxiv** (2026)



W. Zhu, Gatech



Franziska Michor Harva
& Dana-Farber

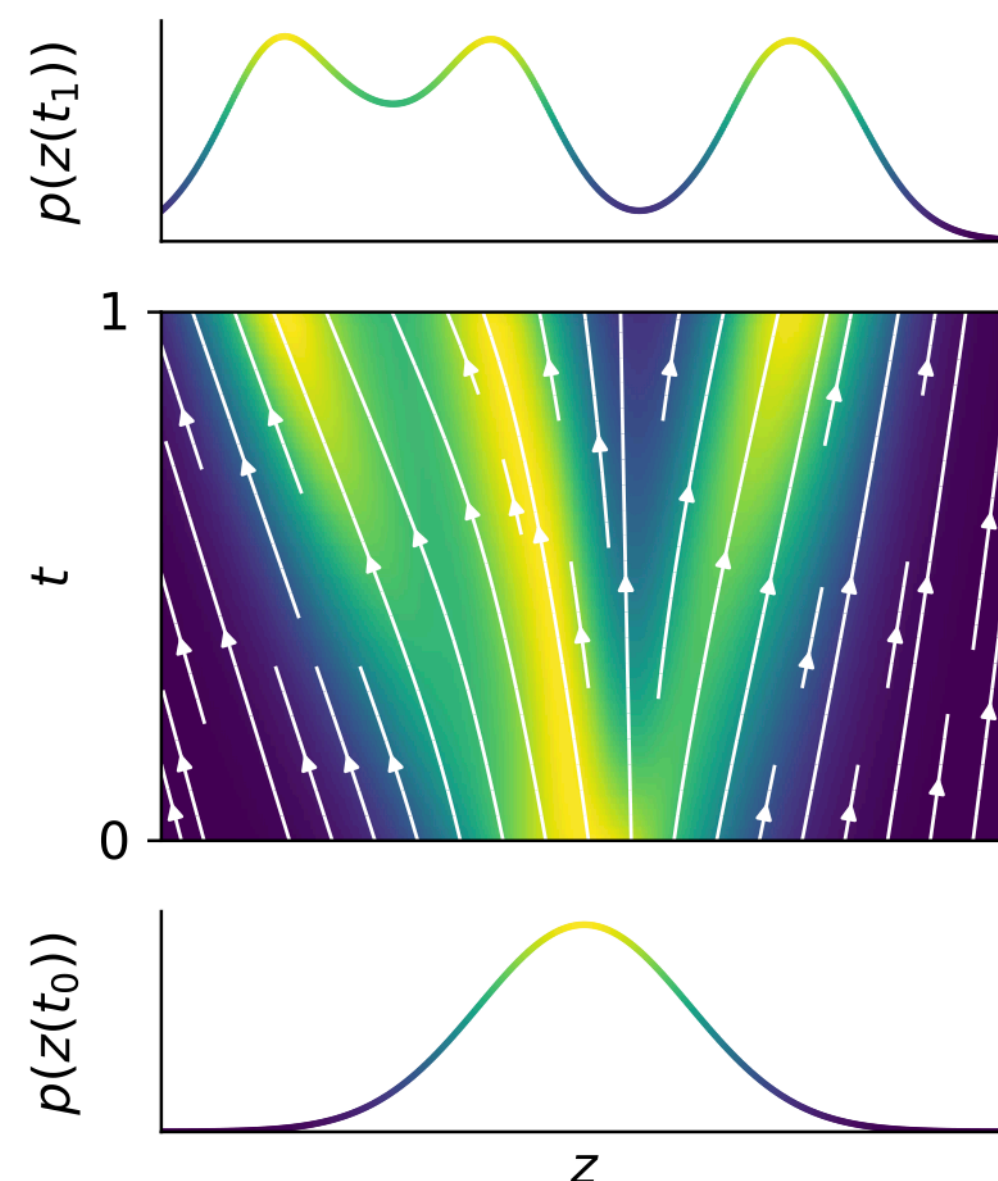


Paul Dupuis, Brown

State-of-the-art generative models are often **flows**

Goal: Given data $\{X_i\}_{i=1}^N \sim \pi$ from (unknown) **target distribution** π , produce new samples from π

- **Flow-based generative modeling:** construct **ODE** or **SDE** to **transport probability measures**



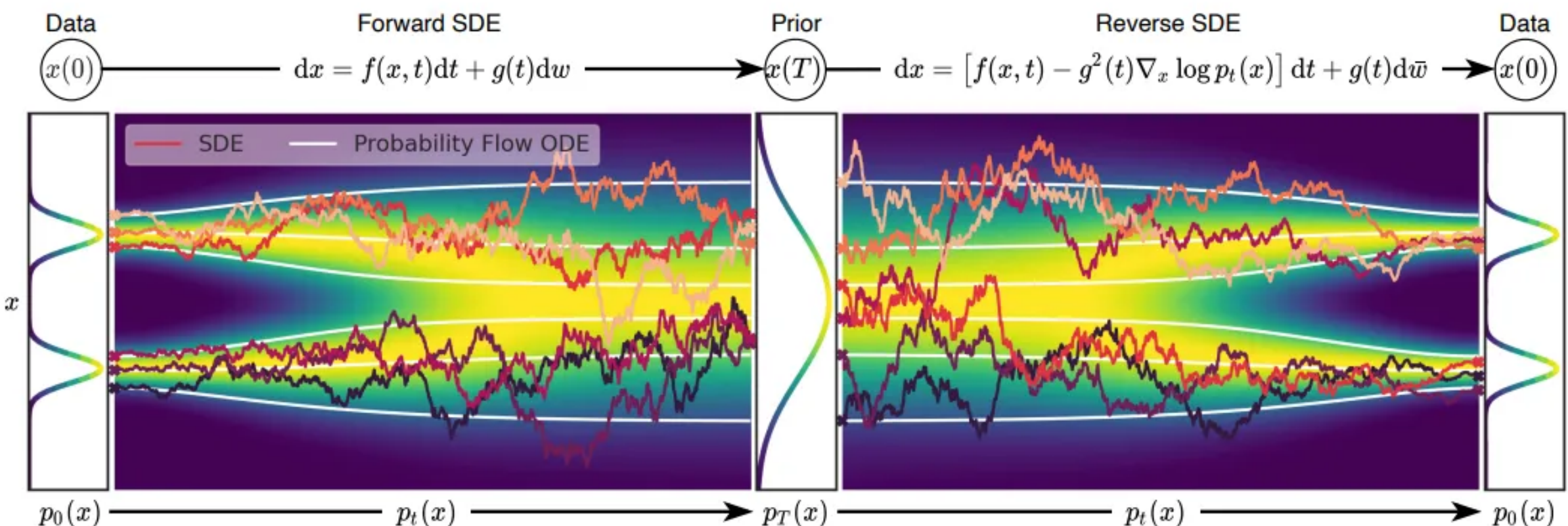
Continuous normalizing flows

Grathwohl et al. '18

Score-based generative models_x

Song et al. '20

Ho et al. '20



Stable Diffusion, DALL-E, Midjourney, ...

These algorithms are all secretly solving control problems and HJ PDEs

Related ideas: flow-matching, distillation, fine-tuning, discrete diffusion, ...

Sampling vs Generation: Complementary Perspectives

Sampling

- **Given** a **known** distribution $\pi(x)$, e.g. from Stat. Physics, Comput. Biology, Bayesian Modeling
- **“Known”** means we have an analytical formula $g(x)$ but we may not know the partition function Z :

$$\pi(x) = \frac{g(x)}{Z}, \quad Z = \int_{\mathbb{R}^d} g(x) dx$$

- **Task:** Generate samples from $\pi(x)$

$$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$$

Generation

- **Given** **only samples** from $\pi(x)$

$$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$$

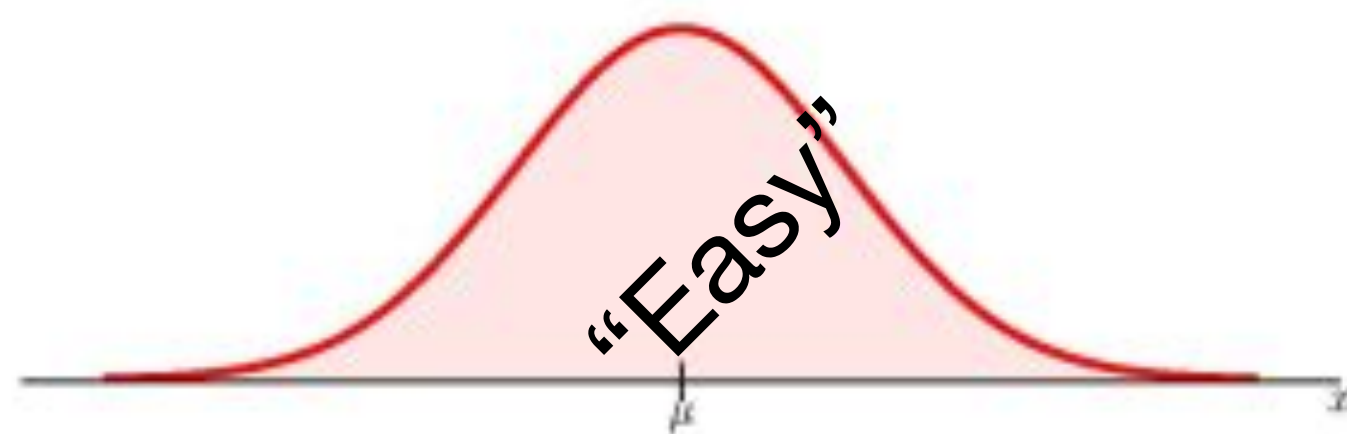
- However $\pi(x)$ is **unknown**, except for the data set $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$

- **Task:** Discover/Learn an approximation of $\pi(x)$ using a parametric family of “models” (to be constructed too) $p_\theta(x)$ for parameters $\theta \in \Theta$:

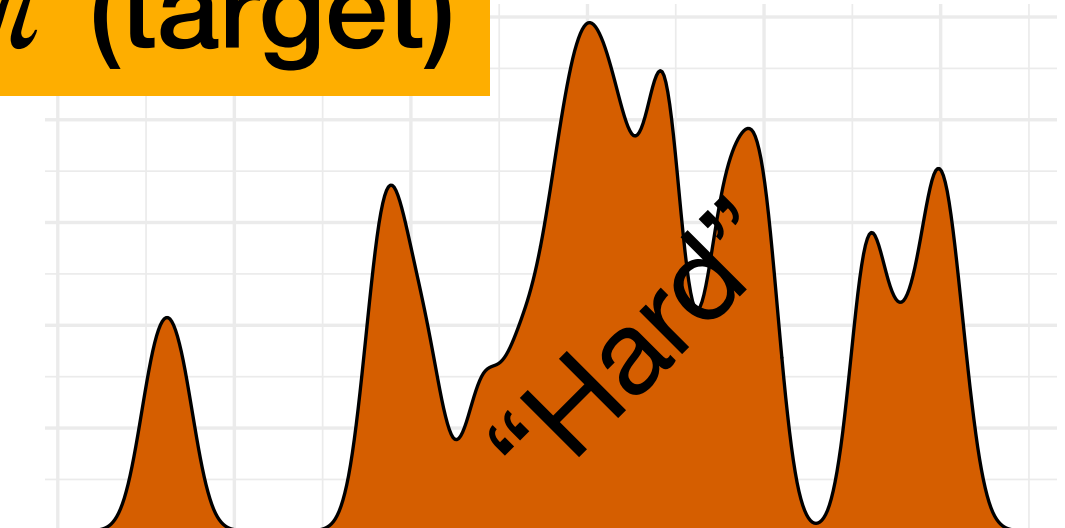
$$p_{\theta^*}(x) \approx \pi(x)$$

- Here “model” means probability distribution (known from Probability/Stat Physics or an **entirely new** one).

Multistep transport: $\mu_0 \mapsto \mu_1 \mapsto \dots \mapsto \mu_D \approx \pi$ (target)

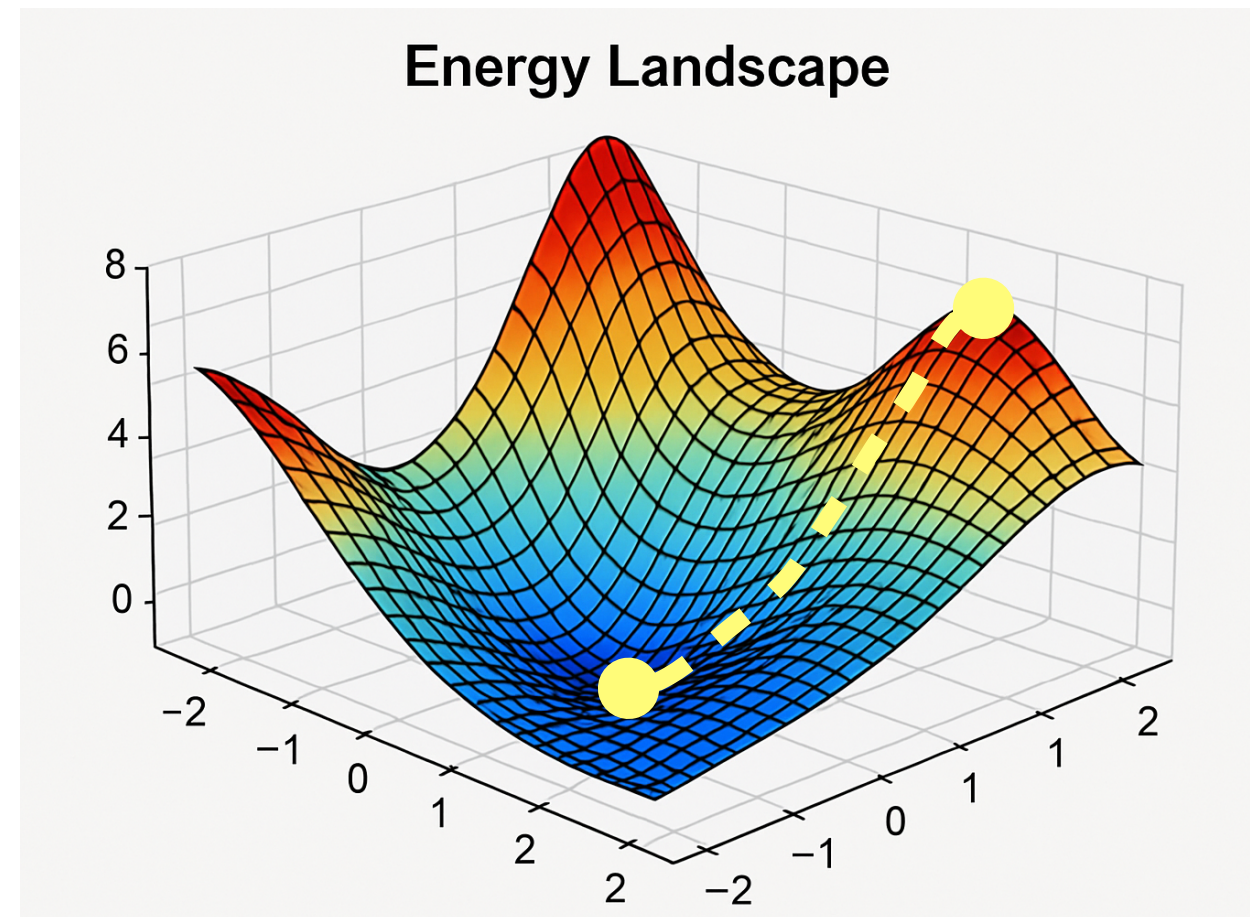


A Common Math Theme:
Measure Transportation



Two measure transport frameworks

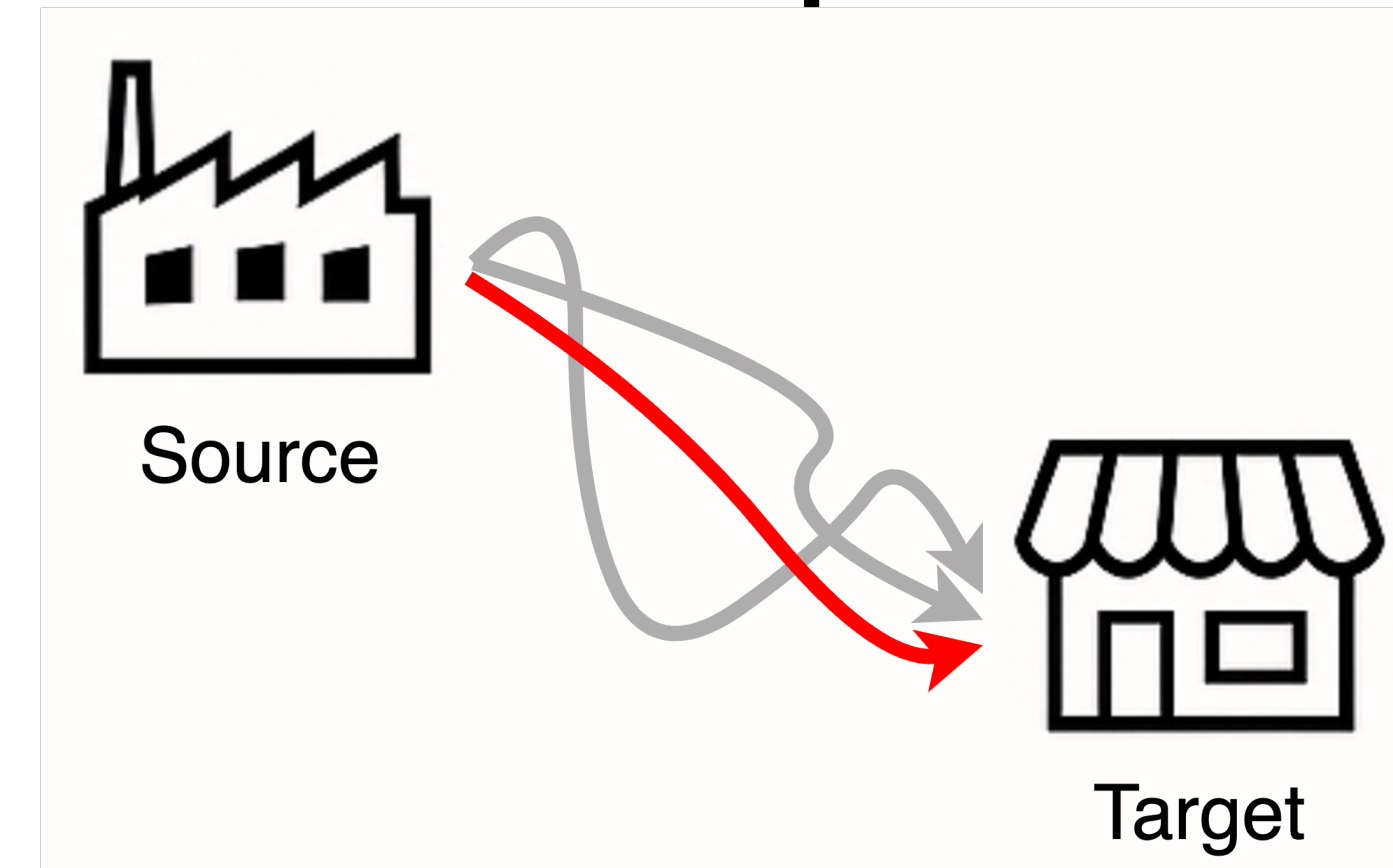
Gradient flows



- Particles evolve by descending the **Free Energy** landscape—flow determined by the system's geometry “natural laws”
- Goal is to converge to an equilibrium state

This talk

OT: Mean Field Optimal Control



- At each time and state, agents choose “controls” to minimize the future cumulative cost relative to the current state
- Dynamics are guided by agents decisions rather than being fixed
- Diffusion, flow-matching, normalizing flows

Gradient Flows in Probability Space & KL-Based Formulations: Computation and Limitations

Setting up gradient flows in probability space

- **Key goal:** **sample/generate** from π — fixed **target** measure
- Define suitable **“free energy”** functional $\mathcal{F} : \mathcal{P} \rightarrow \mathbb{R} \cup \{+\infty\}$, e.g. a **divergence** $D(\rho_t \parallel \pi)$.
- We want to construct an approximation of target π by **minimizing over** $\rho \in \mathcal{P}$ or computational approximation $\mathcal{P}_{approx}$

$$\min_{\rho} \mathcal{F}(\rho) = \min_{\rho \in \mathcal{P}_{approx}} D(\rho \parallel \pi)$$

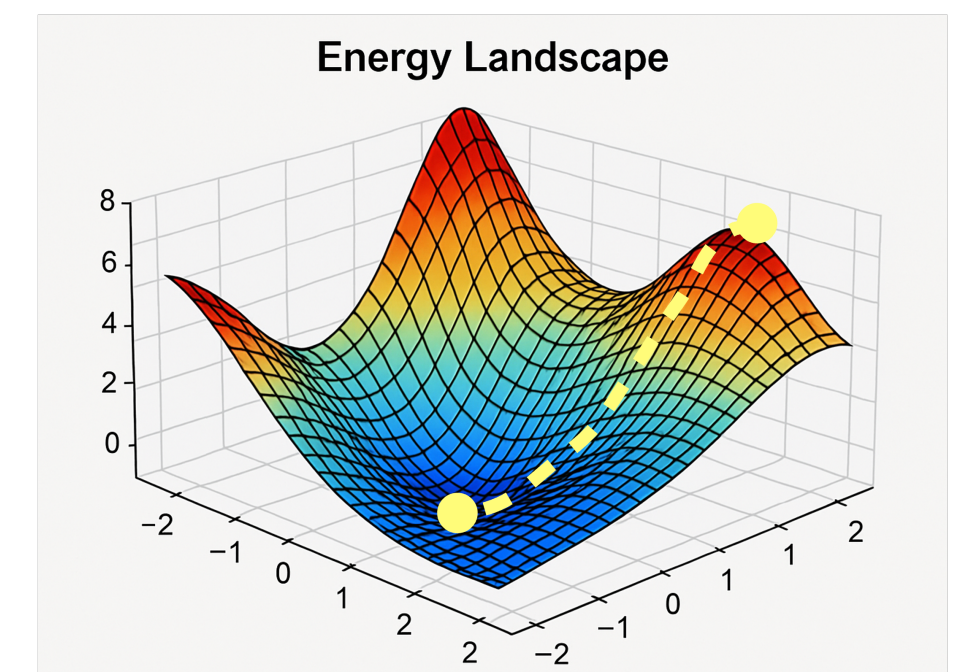
- **Gradient flow** for $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ given (simple) $\rho = \rho_0$: we want to construct/approximate ρ_t (e.g. by particles)

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t) \right), \quad \frac{d}{dt} \mathcal{F}(\rho_t) \leq 0 \quad (\text{dissipation})$$

- **Example: KL divergence** $\mathcal{F}(\rho) = \text{KL}(\rho \parallel \pi) = \int \log\left(\frac{\rho}{\pi}\right) \rho \, dx, \quad \frac{\delta \mathcal{F}}{\delta \rho}(x) = \log\left(\frac{\rho(x)}{\pi(x)}\right) + 1$

- Then:
$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho_t) \right) = \nabla \cdot \left(\rho_t \nabla \log\left(\frac{\rho_t}{\pi}\right) \right)$$

- $\partial_t \rho_t = \Delta \rho_t - \nabla \cdot (\rho_t \nabla \log \pi) = \Delta \rho_t + \nabla \cdot (\rho_t \nabla V)$ if $\pi(x) \propto e^{-V(x)}$ **(Fokker-Planck Equation), sampling...**



Dual formulations: Variational derivatives + Lagrangian dynamics

1. **Dual (variational) formulation:** $\text{KL}(\rho \parallel \pi) = \sup_{\varphi \in \mathcal{M}_b} \left\{ \int \varphi(x) d\rho(x) - \log \int e^{\varphi(x)} d\pi(x) \right\}$
2. $\varphi^*(x) = \arg \max \left\{ \int \varphi(x) d\rho(x) - \log \int e^{\varphi(x)} d\pi(x) \right\} = \log \left(\frac{\rho(x)}{\pi(x)} \right)$ (up to additive const.)
3. **Variational derivative** representation (see again later):

$$\frac{\delta \text{KL}}{\delta \rho}(\rho \parallel \pi) = \varphi^* + 1$$

4. A **Lagrangian** formulation of KL Gradient Flow [Degond et al, Russo, Maoutsa et al]: Rewrite as continuity equation:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \quad v_t(x) = - \nabla \left(\frac{\delta \text{KL}}{\delta \rho}(\rho_t)(x) \right) = - \nabla \varphi^*(x) = \nabla \log \pi(x) - \nabla \log \rho(x) \quad (\text{relative score})$$

5. **Lagrangian dynamics:**

$$\dot{X}_t = - \nabla \phi_t^*(X_t)$$

If $X_0 \sim \rho_0$, then the empirical law of $\{X_t\}_{t \geq 0}$ evolves according to the KL gradient flow. No noise needed!

“Probability Flows”

Limitations of KL and KL-based gradient flows

- **Absolute Continuity** (AC) required: if ρ is **singular** with respect to π , then $\text{KL}(\rho \parallel \pi) = \infty$, e.g.
 - Measures supported on low-dimensional **manifolds**
 - Fractals and strange **attractors**
 - **Empirical** distributions (finite particle systems)
 - **Heavy tails**
- The **variational derivative** also requires AC and involves the (notoriously unstable) **likelihood ratio**:

$$\frac{\delta}{\delta \rho} \text{KL}(\rho \parallel \pi) = \log\left(\frac{\rho}{\pi}\right) + 1$$

**So we need something else...we seek divergences whose gradient flows remain
meaningful in all these challenging regimes**

Learning objectives: probability metrics and information divergences

Mass redistribution

f -divergences

- $f: (0, \infty) \rightarrow \mathbb{R}$ convex, $f(1) = 0$, lower-semicontinuous, super-linear

$$D_f(P \parallel Q) = \begin{cases} \mathbb{E}_Q \left[f \left(\frac{dP}{dQ} \right) \right], & P \ll Q \\ \infty, & \text{otherwise} \end{cases}$$

- Differentiable (Gateux derivative w.r.t. P exist)
- Convex

- Cannot compare non-absolutely continuous measures

Mass transportation

Wasserstein-1 distance

- $$W_1(P, Q) = \inf_{\gamma \in \Pi(P, Q)} \int ||x - y|| d\gamma(x, y)$$

- Do not require absolute continuity

- Non-differentiable (no unique derivative)

Robust Gradient Flows in Probability Space

Lipschitz-regularized f-divergences

(Dupuis, Mao **ESAIM** (2022); Birrell, Dupuis, K., Pantazis, Rey-Bellet, **JMLR** 2022)

The f -divergence is defined as

$$D_f(P\|Q) = \int f\left(\frac{dP}{dQ}\right) dQ, \text{ or equivalently}$$

Used for estimating divergences: Nguyen, Wainright, Jordan, IEEE Trans Info. Th., 2010

$$D_f(P\|Q) = \sup_{\gamma \in \mathcal{M}_b(\mathbb{R}^d)} \{E_P[\gamma] - E_Q[f^*(\gamma)]\},$$

where f^* denotes the Legendre transform of f , e.g.

$$f(x) = f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha - 1)}, \alpha > 1. \alpha \rightarrow 1 \text{ gives the KL-divergence.}$$

The **Lipschitz-regularized** f-divergence is defined as

$$D_f^L(P\|Q) = \sup_{\gamma \in \text{Lip}_L(\mathbb{R}^d)} \{E_P[\gamma] - E_Q[f^*(\gamma)]\}.$$

Lipschitz regularization

Lipschitz-regularized f-divergences are Wasserstein-1 Proximals

Lipschitz+dual
formulation $\implies$ PyTorch
implementations spectral
normalization, one-sided
GP, Birell et al JMLR2022

$$D_f^L(P\|Q) = \sup_{\gamma \in \text{Lip}_L(\mathbb{R}^d)} \{E_P[\gamma] - E_Q[f^*(\gamma)]\}.$$

Source

Target

Lipschitz $\implies$ sample
complexity guarantees,
structure
Z. Chen, W. Zhu et al,
ICML2023, SIAM JUQ

Proximal Structure: The Lipschitz-regularized f-divergences have an equivalent
primal formulation:

Interpolation between D_f and Wasserstein-1:
mass transportation/redistribution

$$D_f^L(P\|Q) = \inf_{\eta \in \mathcal{P}(\mathbb{R}^d)} \{D_f(\eta\|Q) + L \cdot \mathcal{W}_1(P, \eta)\} \leq \begin{cases} D_f(P\|Q) \\ L \cdot \mathcal{W}_1(P, Q) \end{cases}$$

Proximal Structure $\implies$ No need for absolute continuity: inherited from Wasserstein 1

Dual formulation $\implies$ Differentiability $\frac{\delta D_f^L(P\|Q)}{\delta P}$: inherited from f-divergence

Data-agnostic variational derivatives: robustness without assumptions on target

Theorem (abbreviated) [Z. Chen, H. Gu, et al, *Inf. Inference* 2025]

Suppose $\alpha \geq 1$, and $P \in \mathcal{P}_1(\mathbb{R}^d)$. Then $D^{Lip}(P||\pi) < \infty$. Let ρ be a signed measure of total mass 0 and let $\rho = \rho_+ - \rho_-$, where both $\rho_{\pm} \in \mathcal{P}_1(\mathbb{R}^d)$. If $P + \epsilon\rho \in \mathcal{P}_1(\mathbb{R}^d)$ for sufficiently small $\epsilon > 0$, then

First variation formula

$$\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left(D^{Lip}(P + \epsilon\rho||\pi) - D^{Lip}(P||\pi) \right) = \int \phi^{\star} d\rho, \quad \phi^{\star}(x) = \frac{\delta D_{\alpha}^L(P||\pi)}{\delta P}$$

where $\phi^{\star} := \arg \max_{\gamma \in \text{Lip}_L(\mathbb{R}^d)} \left\{ \mathbb{E}_P[\gamma] - \mathbb{E}_{\pi}[f_{\alpha}^*(\gamma)] \right\}$

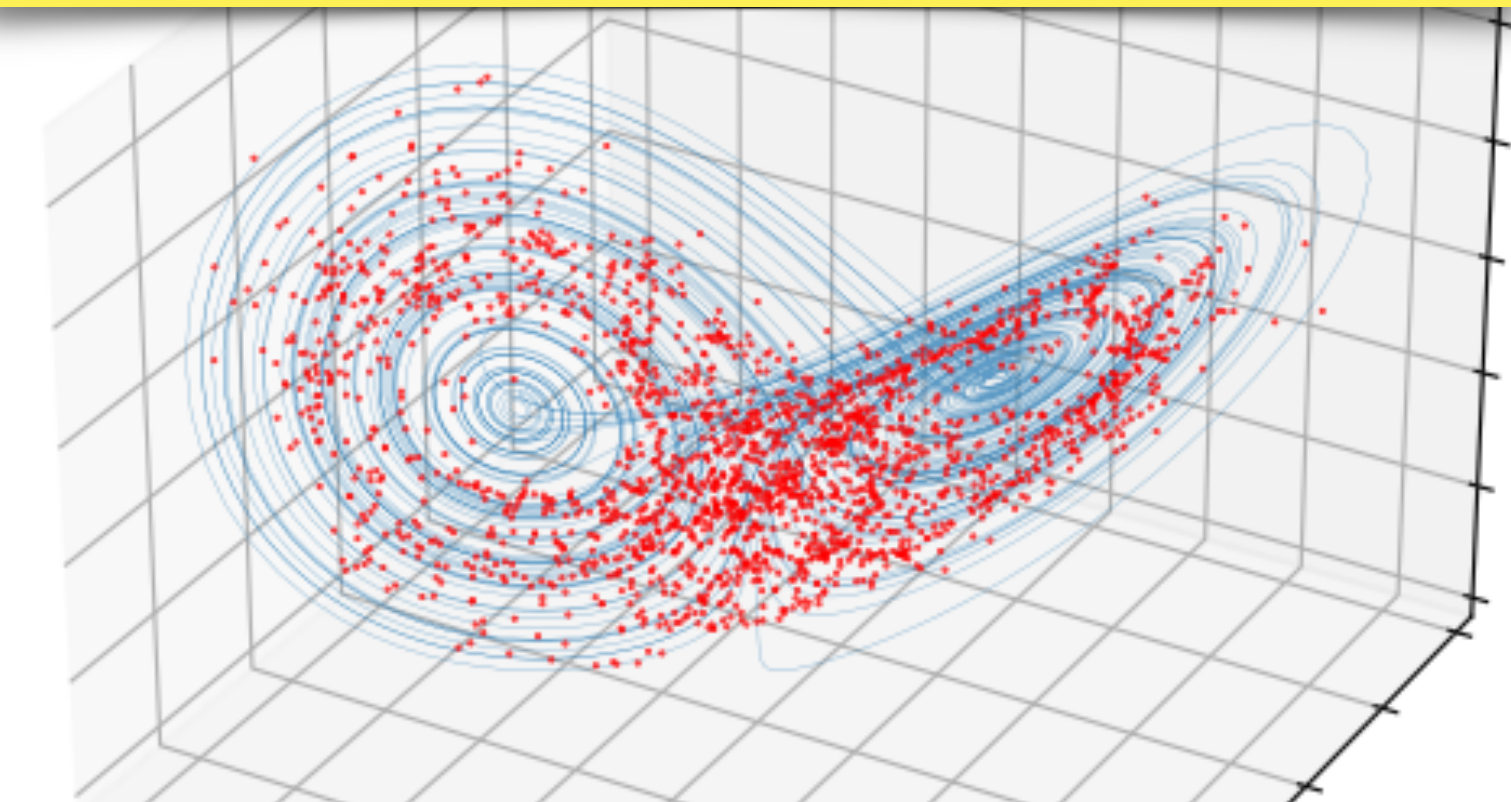
This includes cases when target π

- Has **heavy** tails
- Supported on **low-dimensional** manifolds
- Supported on attractors, **fractals**

Lipschitz-Regularized Divergences Allow **Minimal** Assumptions on Target Distributions

- Let P and π be arbitrary probability measures on $\mathbb{R}^d$, and we only assume that P has a finite first moment; i.e., $P \in \mathcal{P}_1(\mathbb{R}^d)$.

No assumptions on target π !



Generative particle algorithm (GPA) via Wasserstein gradient flow

Gu, Birmpa, Pantazis, Katsoulakis, Rey-Bellet, SIAM J. Math. Data Sci. 2024

Transport equation: $\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta D_f^L}{\delta \rho}(\rho_t || \pi) \right), \quad t \in [0, \infty)$

$$\frac{\delta D_f^L}{\delta \rho}(\rho_t || \pi) = \phi_t^* = \operatorname{argmax}_{\phi \in \text{Lip}_L(\mathbb{R}^d)} \left\{ \mathbb{E}_{\rho_t}[\phi] - \mathbb{E}_{\pi}[f_{\alpha}^*(\phi)] \right\}$$

Relation to Fokker-Planck equation when $f = KL$: $\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \log(\rho_t || \pi) \right) = - \nabla \cdot \left(\rho_t \nabla \log \pi \right) + \Delta \rho_t$

First variation of KL for probability flows score & noise for SDEs

Lagrangian dynamics / Probability flow

$$\frac{dX_t}{dt} = v_t(X_t) = - \nabla \phi_t^*(X_t), \quad X_0 \sim \rho_0$$

$$X_i^{k+1} = X_i^k - \Delta t \nabla \phi_k^*(X_i^k), \quad i = 1, \dots, N$$

Stability from bounded velocity

$$\|\nabla \phi_k^*\|_{\infty} \leq L \implies \max \text{ particle speed} \leq L$$

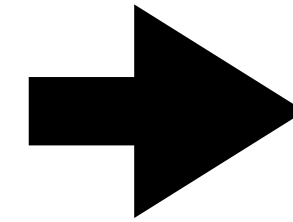
The Lipschitz bound is a maximum speed, so the Forward Euler scheme meets a **CFL condition** in coordinate with time step size

Dissipativity from gradient flow structure

As $t \rightarrow \infty$, $D_f^L(\rho_t || \pi) \rightarrow 0$ and the kinetic energy $\mathcal{K}(t) = \frac{1}{2} \int |v_t(x)|^2 d\rho_t(x) \rightarrow 0 \Rightarrow$ stationary distribution

Due to the data-agnostic properties of $D_{\text{KL}}^{\text{Lip}_L}$:

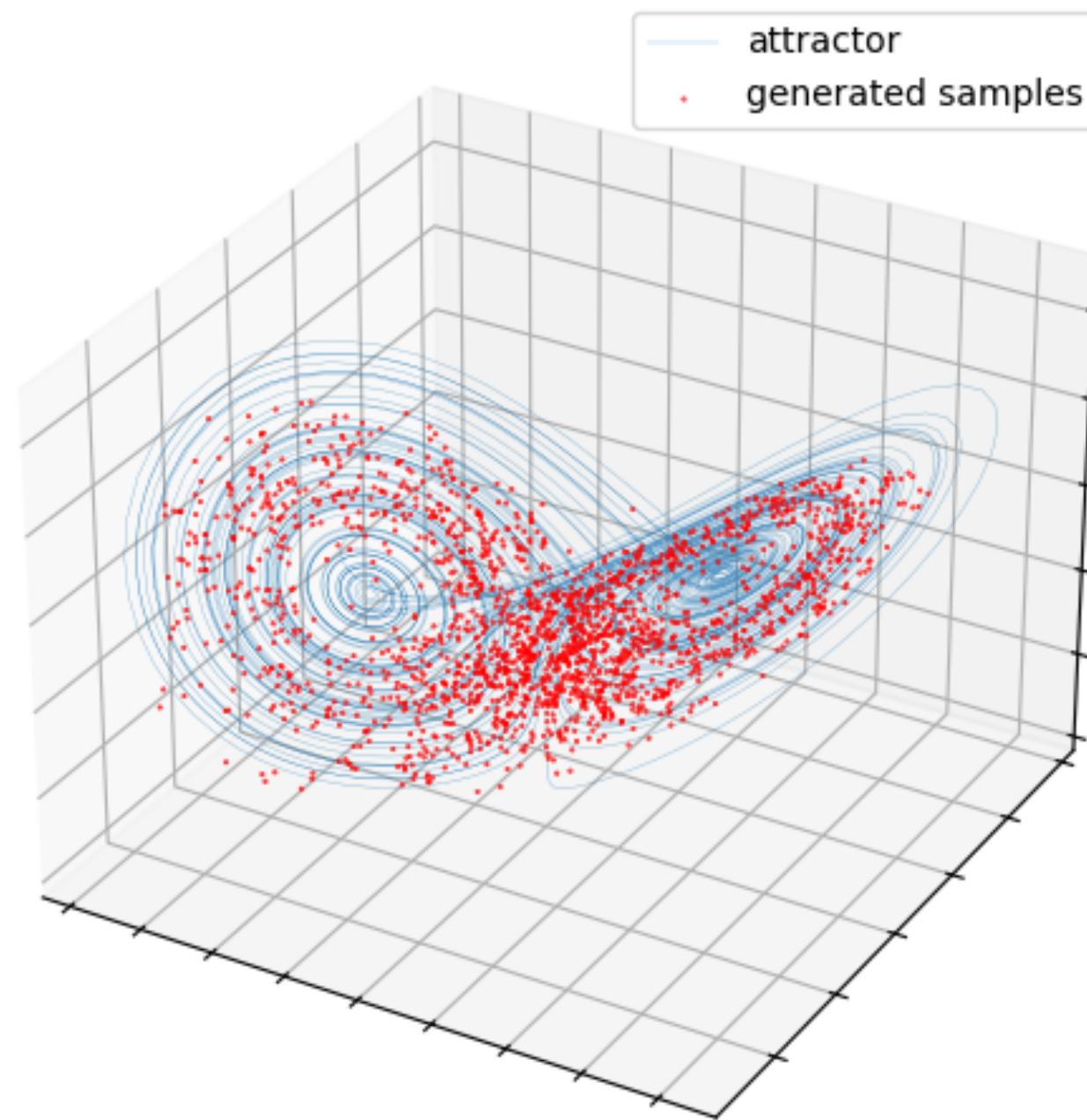
Overcoming distributional regularity assumptions in **OT**



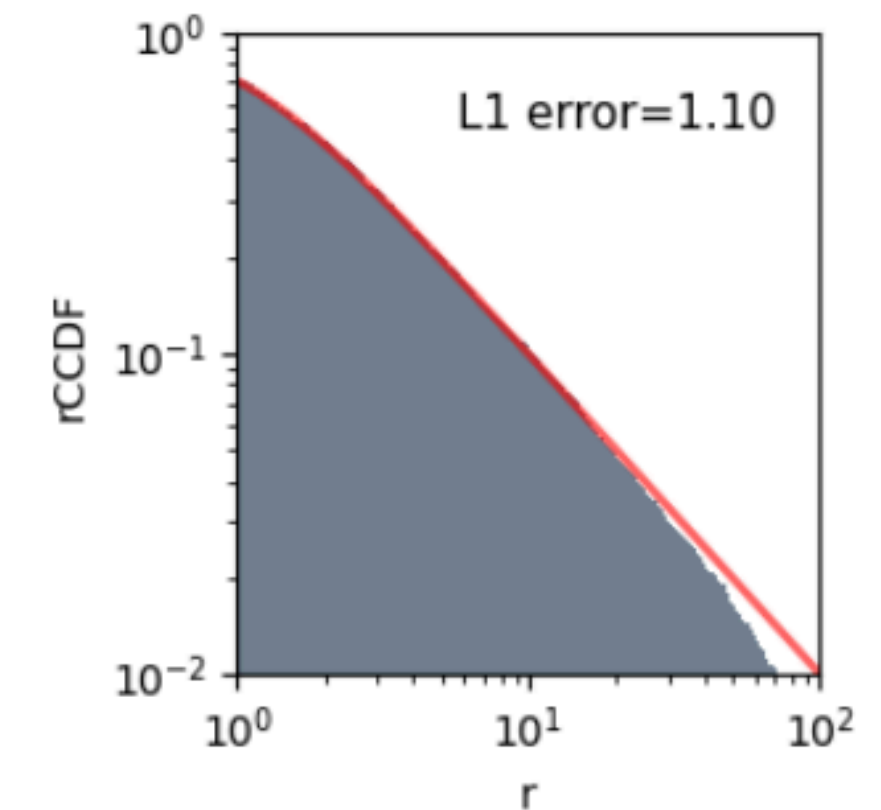
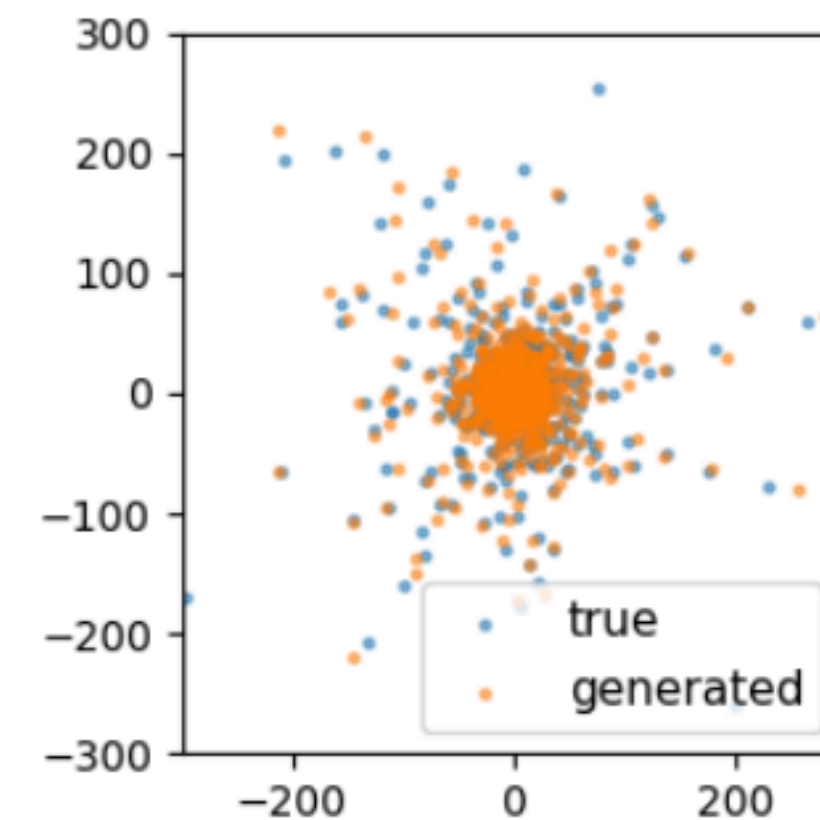
Gradient flow with $D_{\text{KL}}^{\text{Lip}_L}$ enable transport towards general target distributions:

Fractal structure

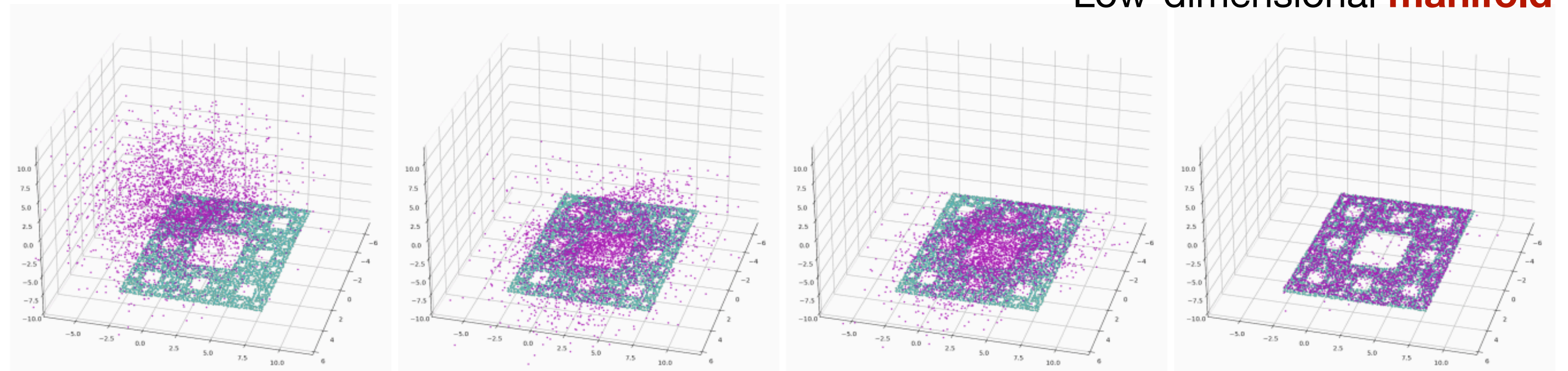
KL-Lipschitz=1 GPA (1000 samples)



Heavy-tailed data



Low-dimensional **manifold**



Pushing the limits of these methods: Generative Particle Flows for Reconstructing Nonlinear **Cell Dynamics**

Hyemin Gu, M. K. (UMass Amherst),
Yu-Chen Cheng, Franziska Michor (Harvard, Dana-Farber), et al.
(Cell Systems, August 2026)



- Learn from temporally sparse, very high dimensional data sets: **d=50K features**
- “moving target” (empirical) distribution

Problem Setting & Overview: chasing a “moving target”

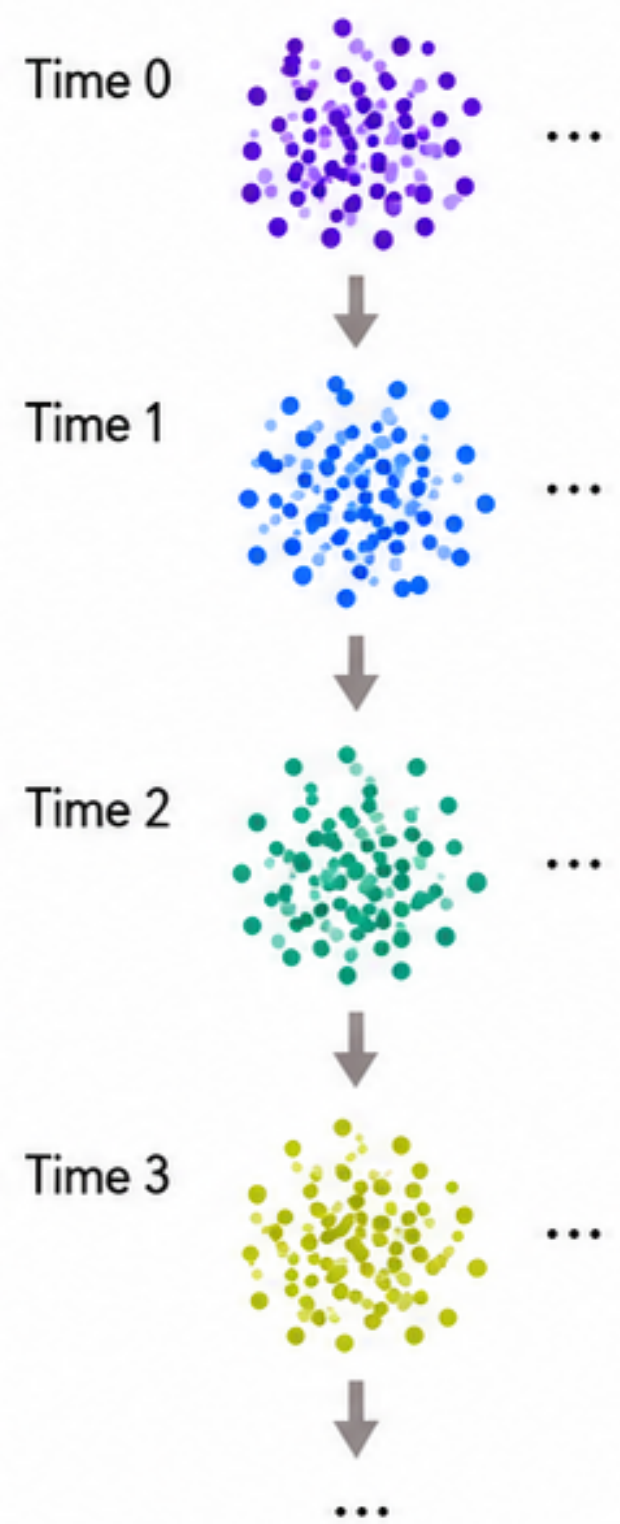
Single-cell RNA datasets

- **50K dimensional** normalized/log-transformed gene expression level
- **Sparse time snapshots** of cell population distributions

Goal:
Model and reconstruct cell population distribution dynamics

1. scRNA-seq snapshots (static observations)

Cells are sacrificed at discrete time points

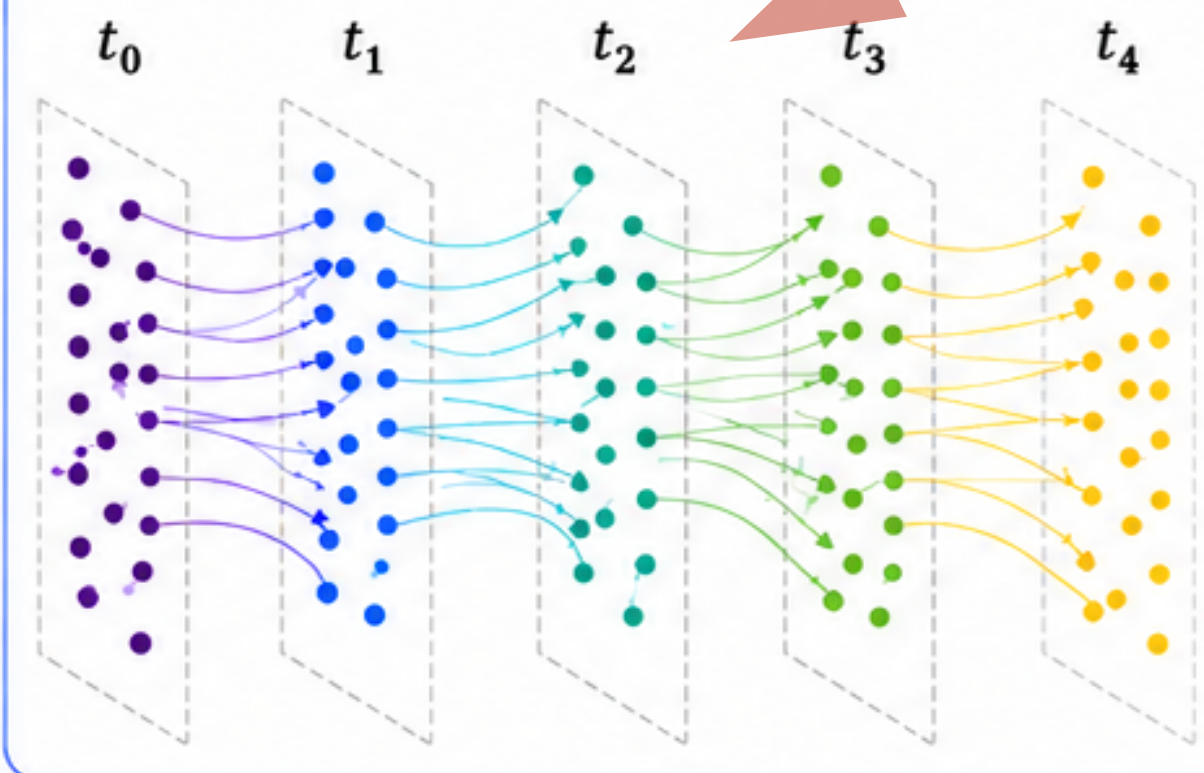


✗ Individual cell continuous trajectory cannot be measured experimentally

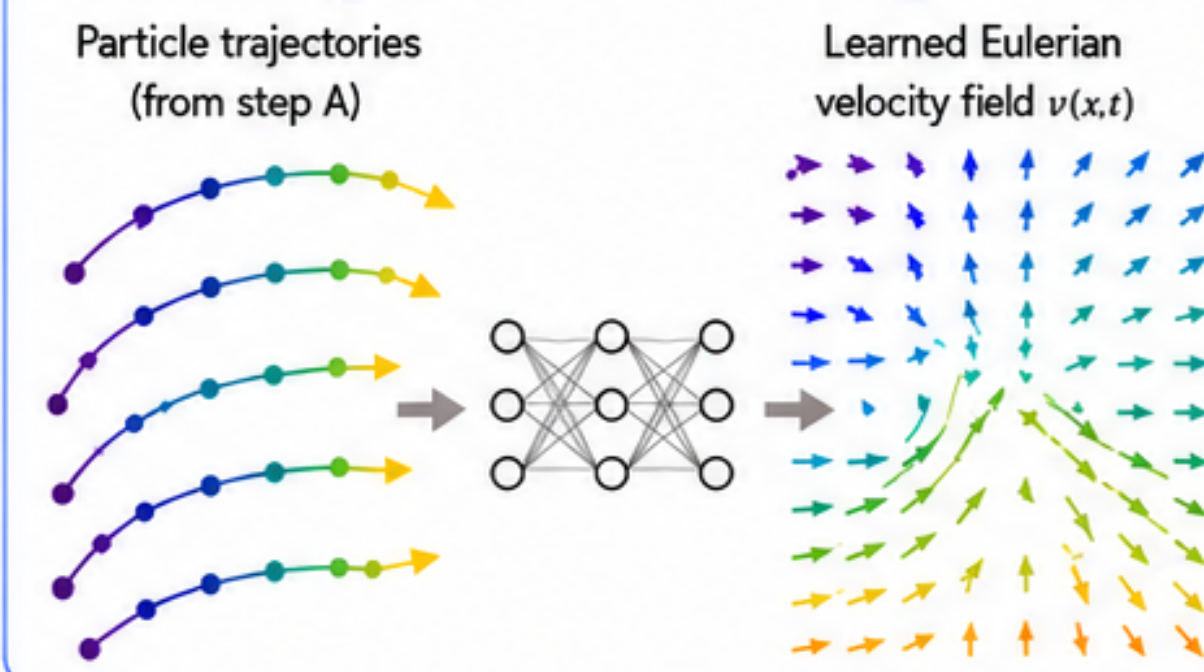
2. PROFET framework

PROFET

A. Particle flow with GPA (Lipschitz-regularized gradient flow)

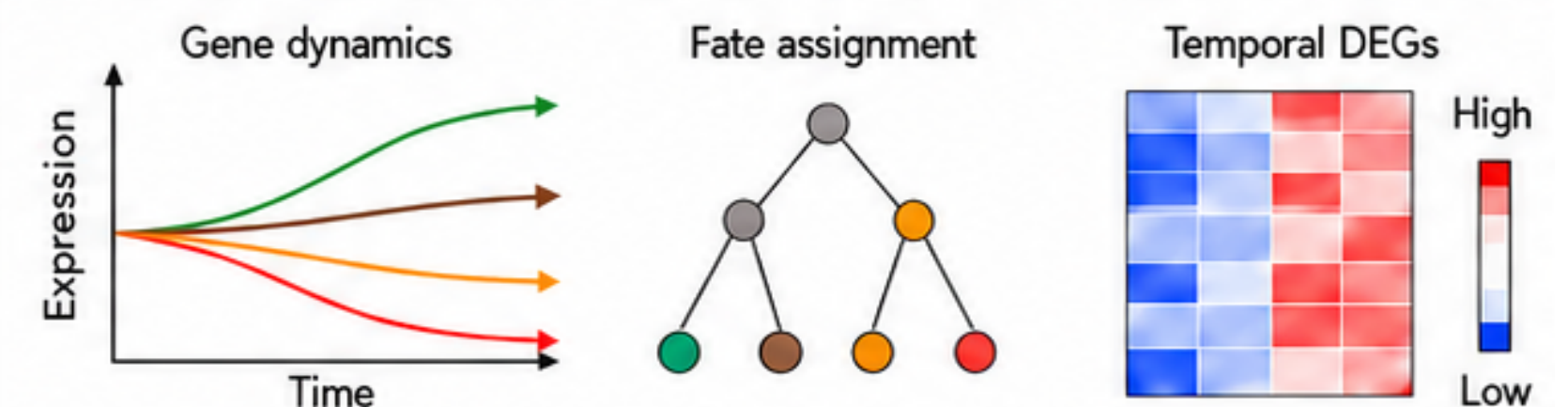
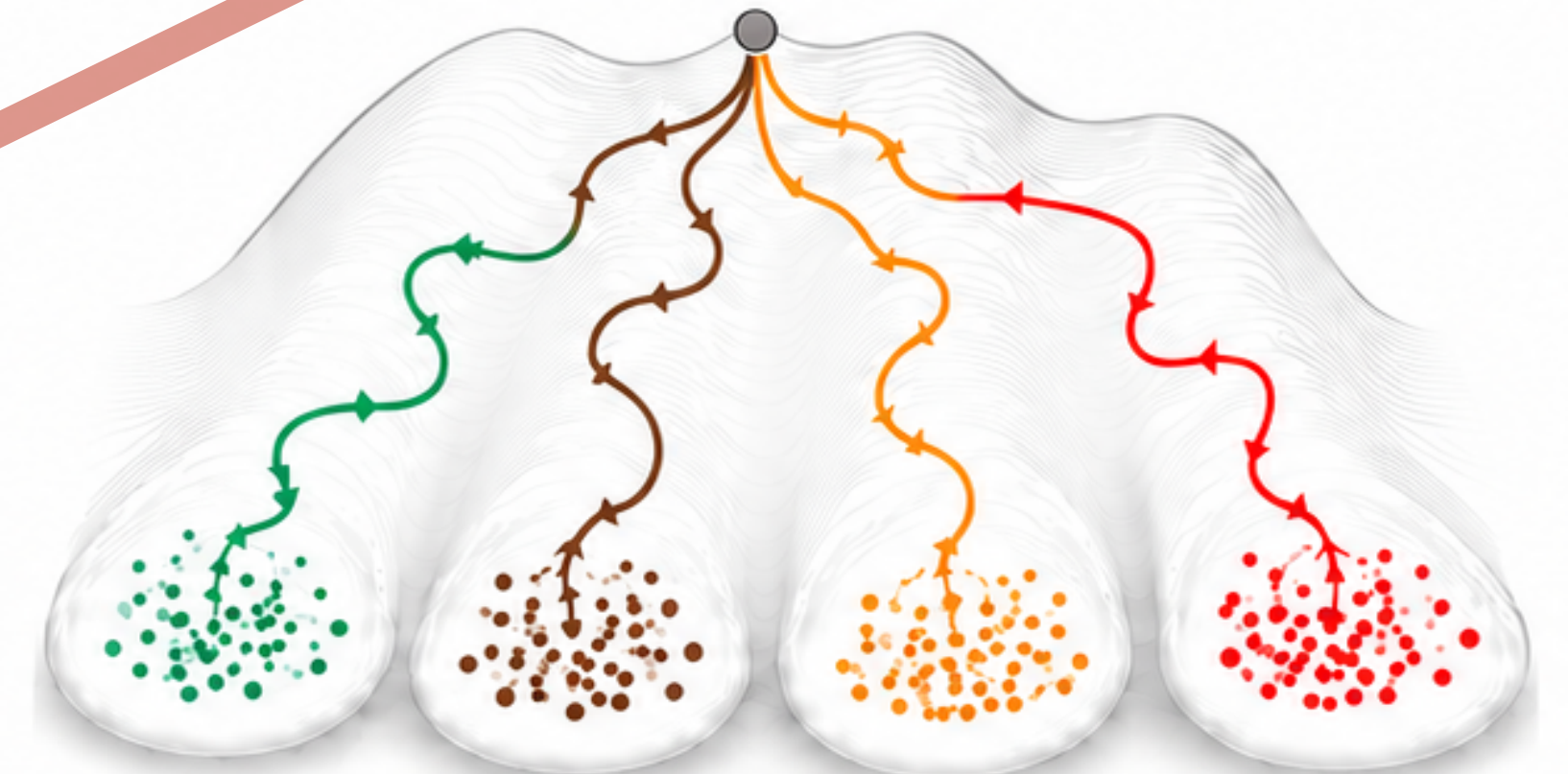


B. Neural force matching (global vector field learning)

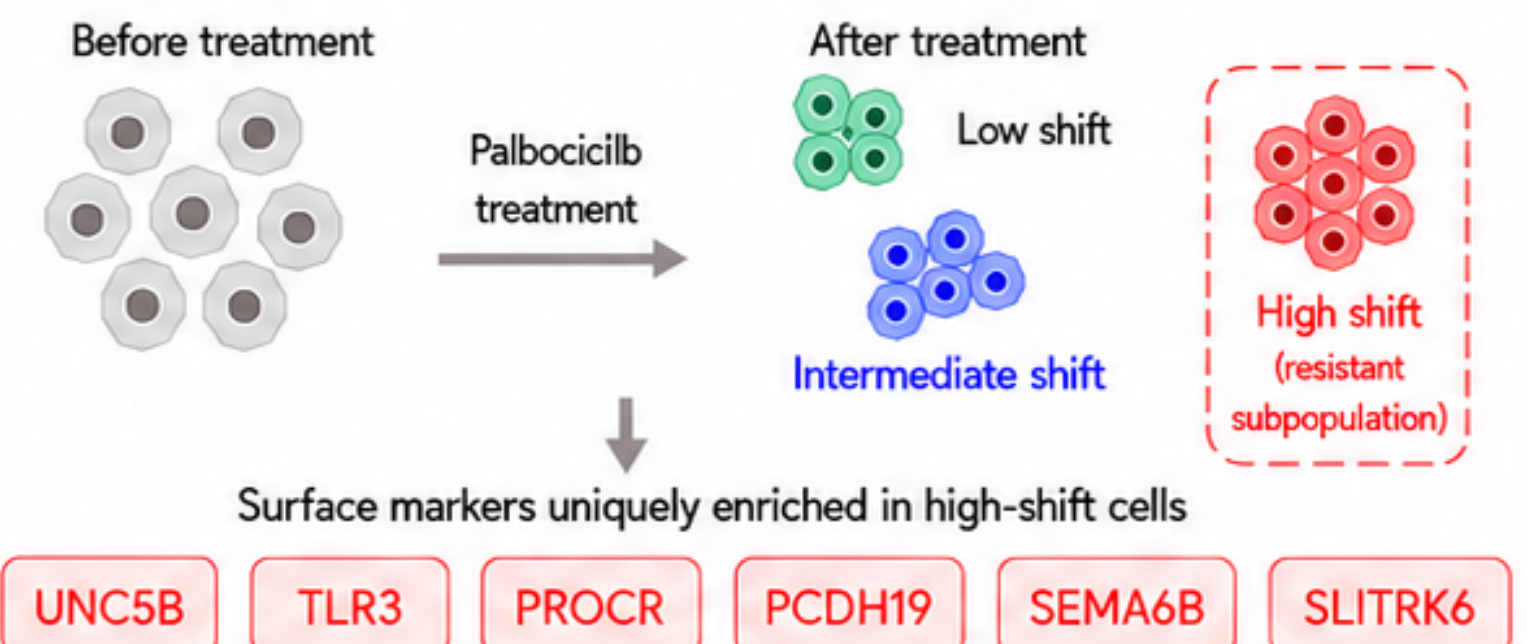


3. Reconstructed single-cell dynamics

Continuous, nonlinear trajectories in gene expression space



4. Application to palbociclib-treated HR+ breast cancer

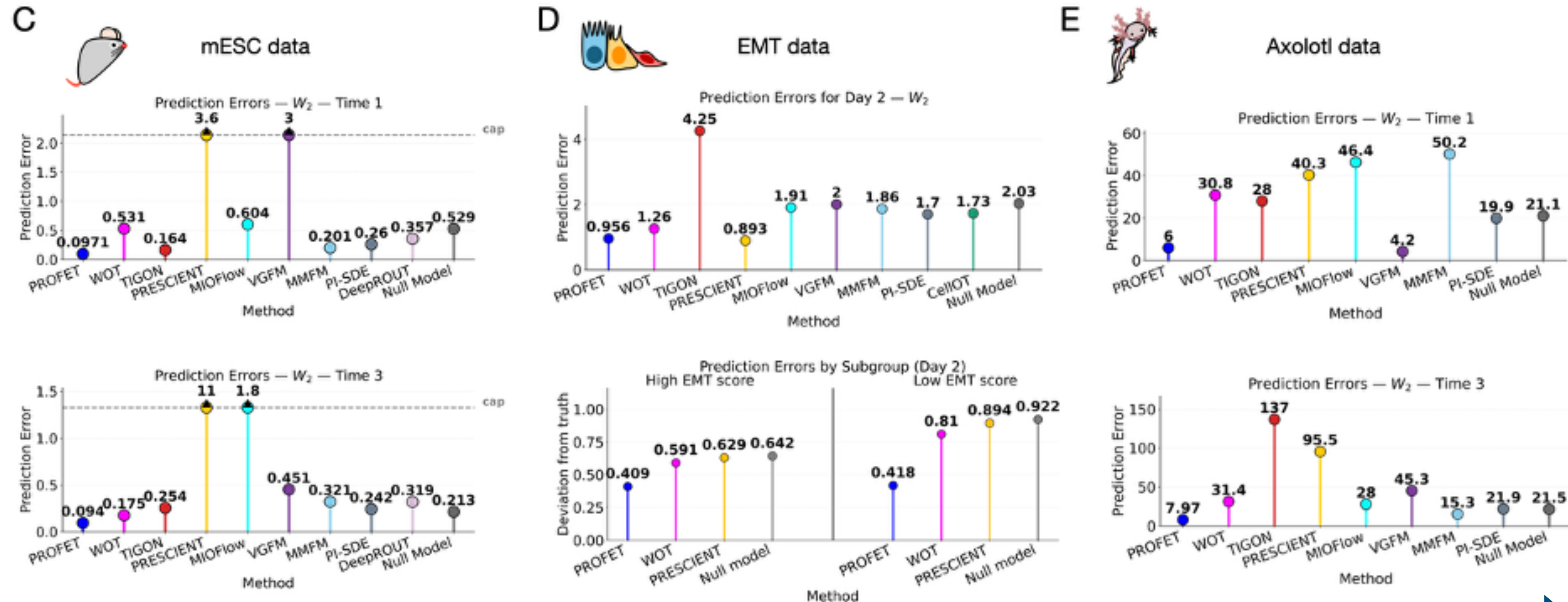


Comparisons of PROFET to SOTA

Cell Systems, August 2026



PROFET (blue/left) attains **2.6-12.5x** lower prediction error (W2), across data sets, scenarios



As dataset complexity increases, our performance gains over SOTA become larger

Risk-Aware Objectives for Extreme Event Generation & Fine-tuning

Arxiv: T. Gamage, H. Gu, Z. Zhang, Z. Chen, M.K. & L. Rey-Bellet

Why fine-tuning using a gradient flow?

$$\partial_t Q_t = \nabla \cdot \left(Q_t \nabla_x \frac{\delta \mathcal{F}}{\delta Q}(Q_t) \right), \quad Q_0 = P_{\text{pre}}$$

1. Start from any data

The initial condition can be any pre-trained model's samples; (GAN, diffusion model, flow,)

$$Q_0 = P_{\text{pre}}$$

“Model-agnostic”, could be just a data distribution or...

2. Design the free energy

Free to add the quantity of interest **QoI** you care about to the energy the flow minimizes.

$$\mathcal{F} = \text{KL}^L(Q \parallel P_{\text{tar}}) + \lambda \text{QoI}$$

QoI = $(\Delta CVaR)^2$: the flow should learn heavy tails

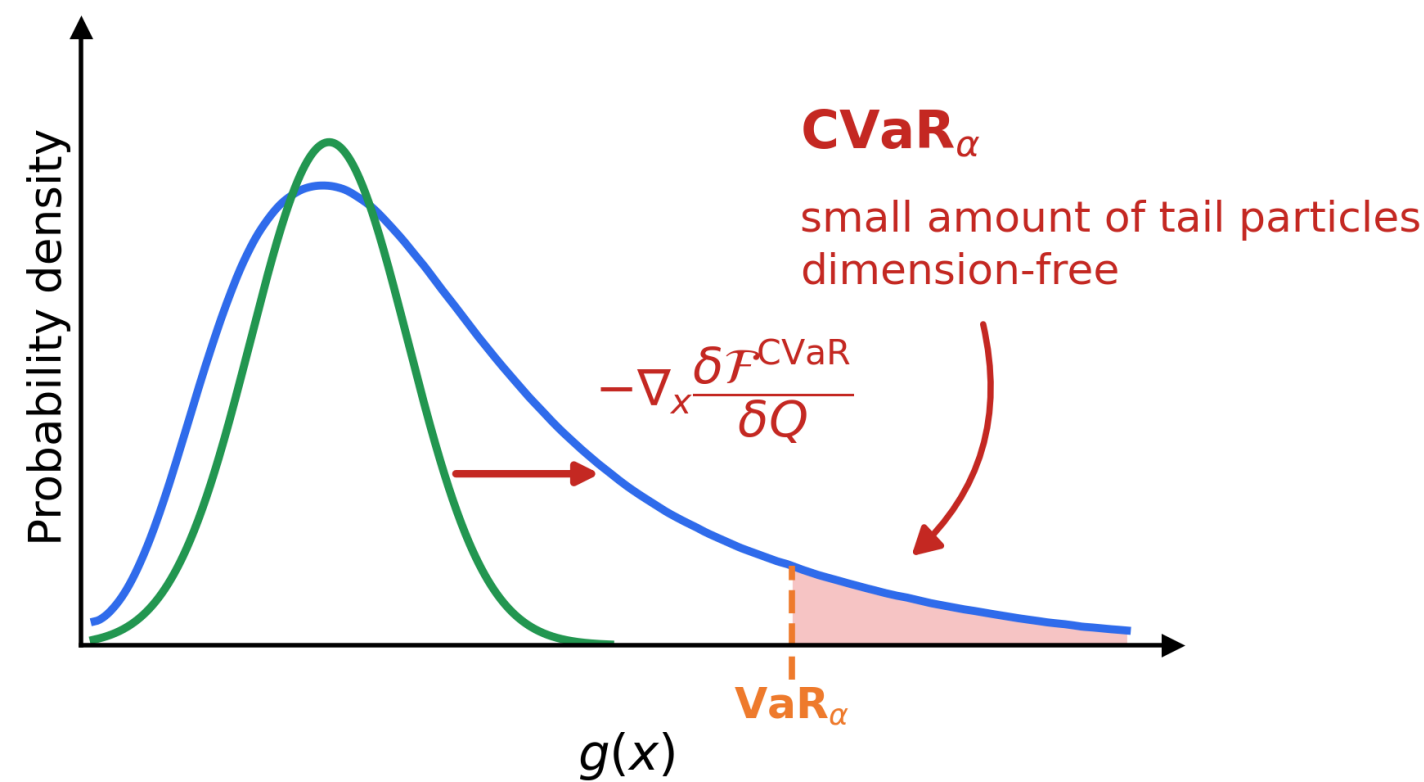
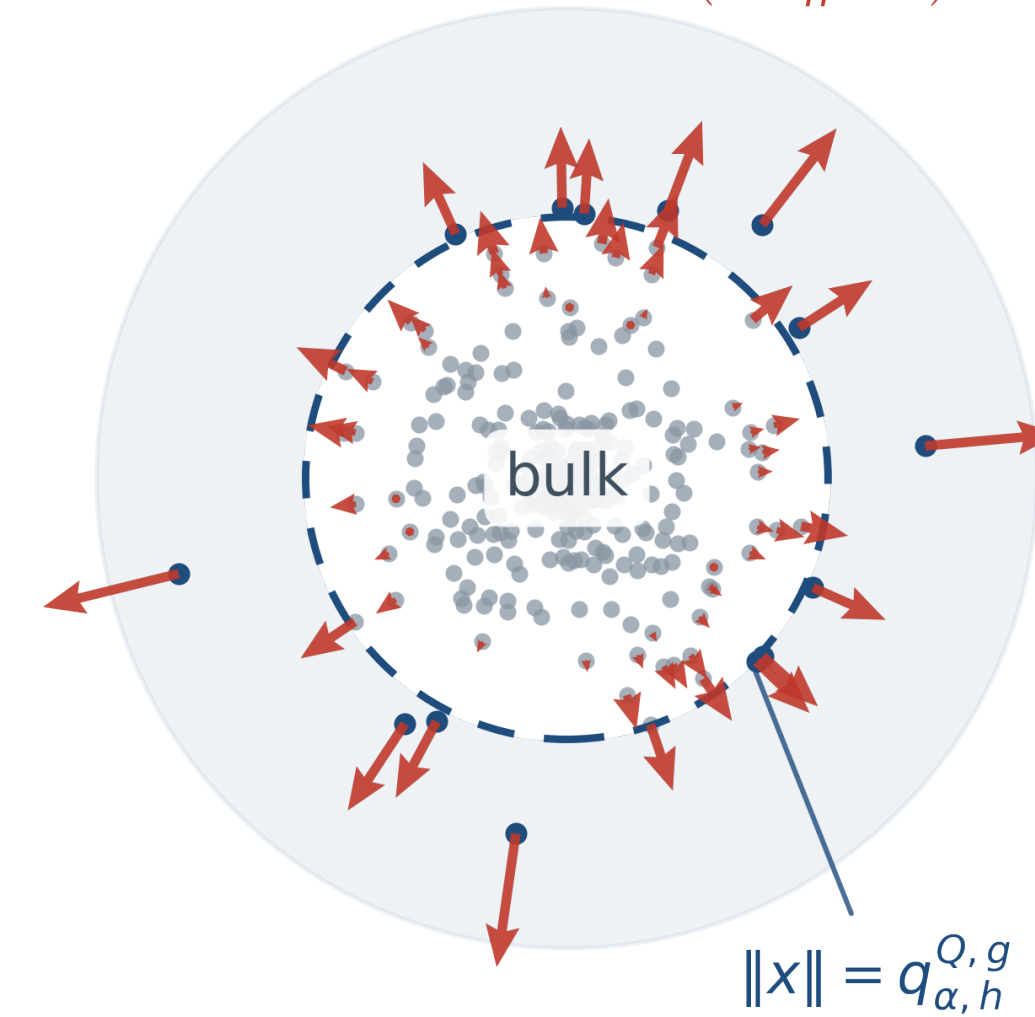
Fine-tuned Generative Models for Extreme Events via CVaR-penalized Wasserstein gradient flows

Change “free energy” to encourage a flow that learns heavy tails

$$\mathcal{F}^{CVaR}(Q; P_{tar}) = KL_{lip}(Q \| P_{tar}) + \lambda \left(\Delta C(Q; P_{tar}) \right)^2$$

$$\Delta C(Q; P_{tar}) := CVaR_{\alpha}^{P_{tar}, g} - CVaR_{\alpha}^{Q, g}$$

CVaR=captures the (expected) severity of **tail risk**



Wasserstein gradient flows are adaptable to **fine-tuning** tasks:

$$\partial_t Q_t - \nabla \cdot \left(Q_t \nabla_x \left(\frac{\delta \mathcal{F}^{CVaR}(Q; P^{tar})}{\delta Q}(x) \right) \Big|_{Q=Q_t} \right) = 0, \quad Q_0 = P_{pre}$$

Risk-aware extension of data-agnostic $D_{KL}^{Lip_L}$ enables effective fine-tuning for **multivariate, anisotropic** heavy tails

Value-at-Risk and Conditional Value-at-Risk

Definition

CDF F_X continuous, strictly increasing; $\mathbb{E}|X| < \infty$; $\alpha \in (0,1)$

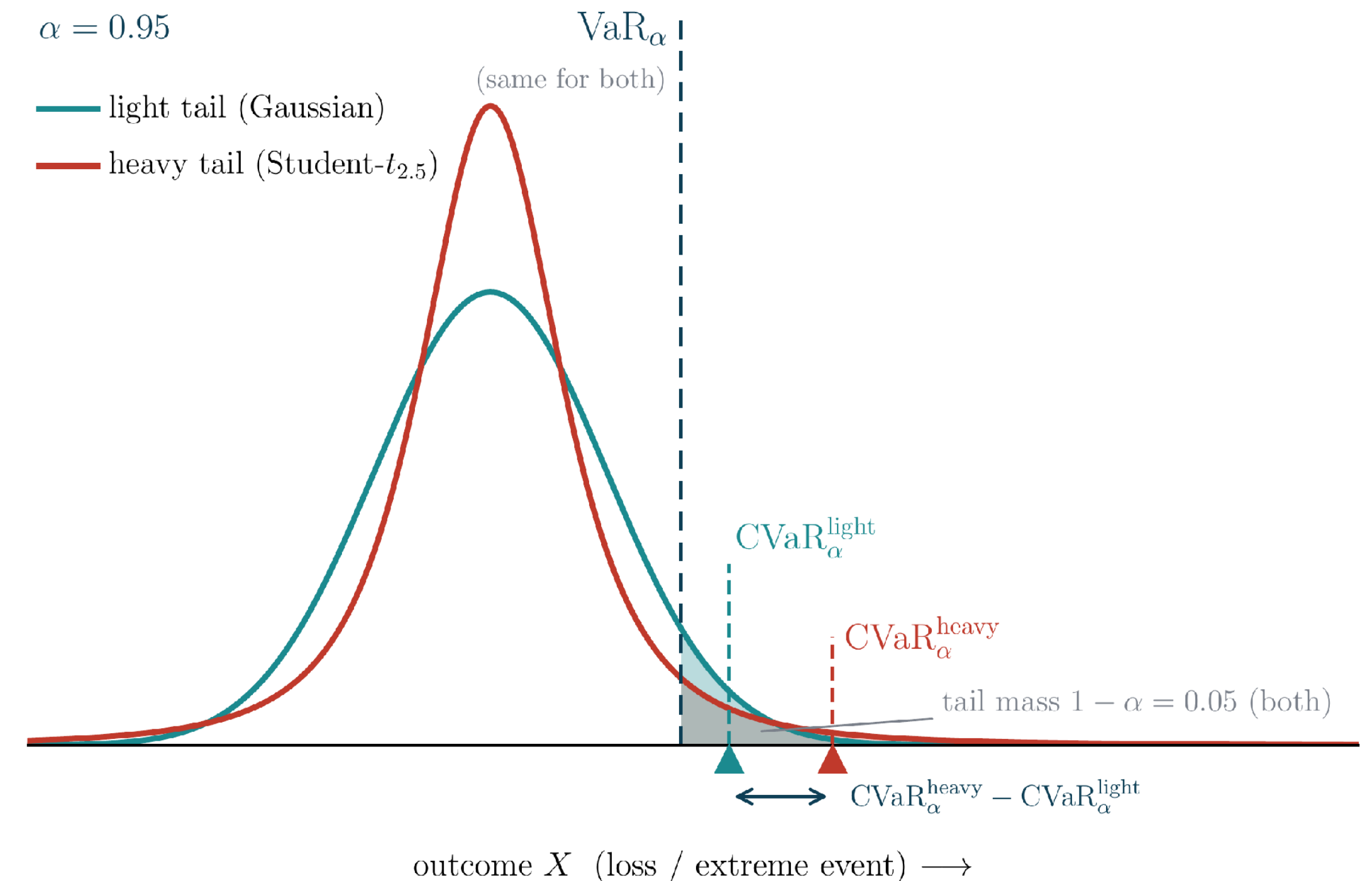
$$\text{VaR}_\alpha(X) = F_X^{-1}(\alpha)$$

$$\begin{aligned} \text{CVaR}_\alpha(X) &= \mathbb{E}[X \mid X \geq \text{VaR}_\alpha(X)] \\ &= \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-\alpha} \mathbb{E}[(X - t)_+] \right\} \quad (\text{RU}) \end{aligned}$$

RU = Rockafellar–Uryasev formula; minimizer $t^ = \text{VaR}_\alpha$*

Interpretation: tail mean

- **VaR** is the α -quantile: the level exceeded with probability $1-\alpha$. It only locates the tail threshold.
- **CVaR** is the mean of the worst $(1-\alpha)$ fraction of outcomes.
- CVaR also depends on the **magnitude** of the tail beyond that threshold; it separates light from heavy tails that share a quantile.



Gaussian vs. rescaled Student- $t_{2.5}$, same $\text{VaR}_{0.95} \approx 1.645$

Same VaR, same tail mass 0.05; different CVaR:
2.06 vs 2.96

The velocity field for the radial risk function $g(x) = \|x\|$

$$v_{Q,h}(x) = \underbrace{-\nabla_x \phi^*(x)}_{\text{from } D_{\text{KL}}^L(Q \| P^{\text{tar}})} + \underbrace{\lambda \frac{2}{1-\alpha} \Delta C_h(Q; P^{\text{tar}}) \sigma\left(\frac{g(x) - q_{\alpha,h}^{Q,g}}{h}\right) \nabla_x g(x)}_{\hat{v}_{Q,h}(x): \text{ from CVaR-penalization}}$$

$$g(x) = \|x\| \implies \nabla_x g(x) = \frac{x}{\|x\|}, \quad \sigma(t) = \frac{1}{1 + e^{-t}}$$

Radial

$g(x) = \|x\|$ captures extremes in every direction through one observable, so the push is along each particle's own outward ray.

$$\nabla_x g(x) = \frac{x}{\|x\|}$$

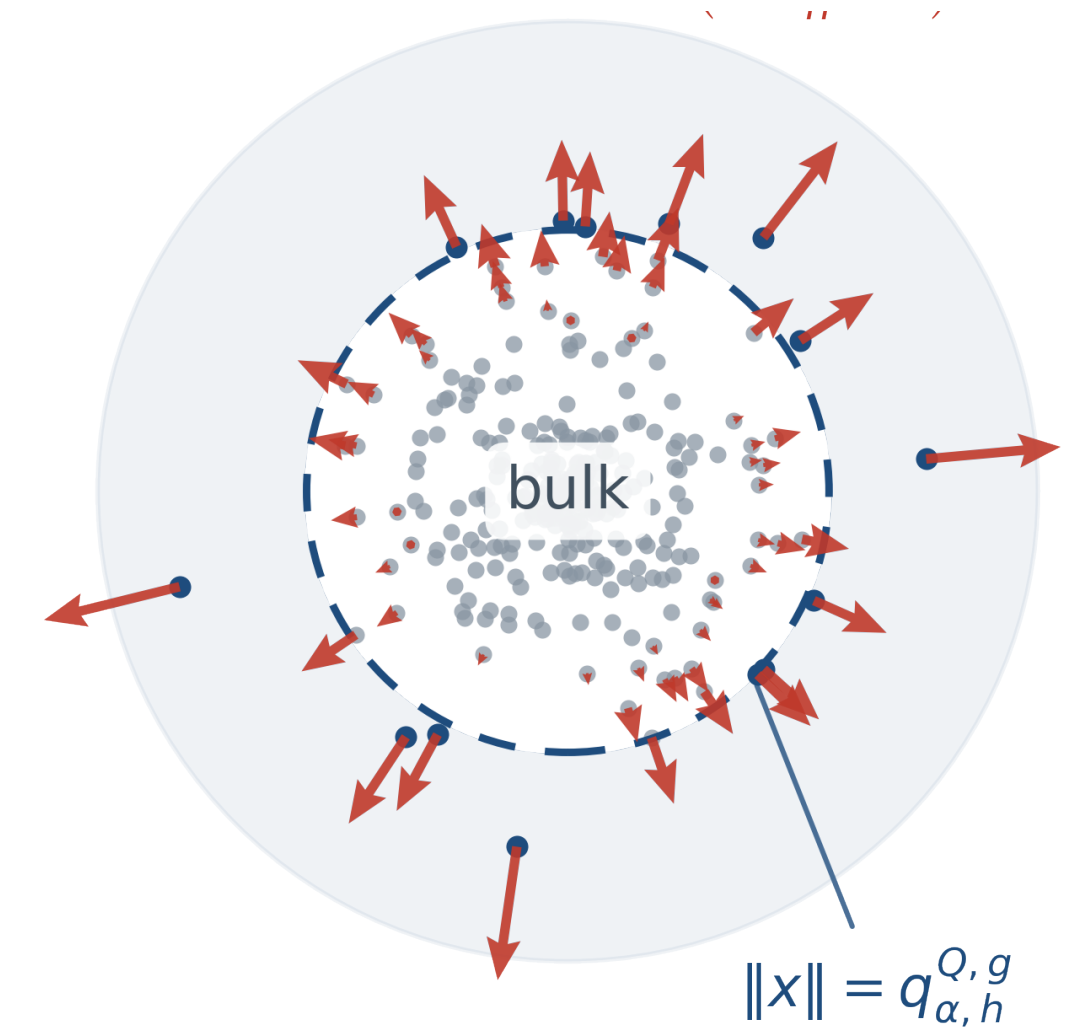
Tail-localized

The logistic factor is a smooth surrogate for the exceedance indicator.

$$\sigma\left(\frac{\|x\| - q_{\alpha,h}^{Q,g}}{h}\right) \approx \mathbf{1}_{\{\|x\| > q_{\alpha,h}^{Q,g}\}}$$

Dimension-free

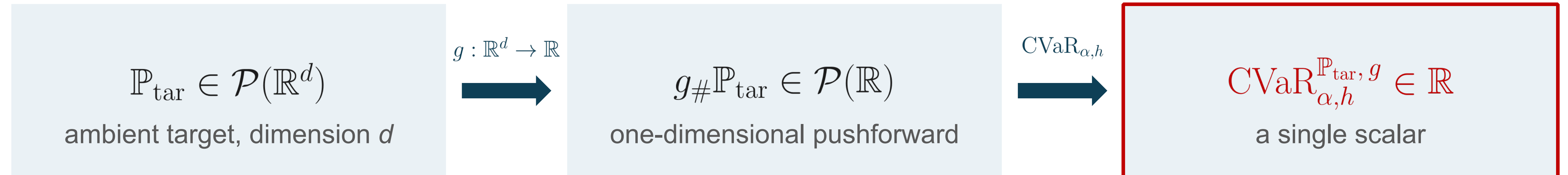
The target enters only through the scalar CVaR of its pushforward under g , whose estimation error is of $O\left(N^{\min\{\frac{1}{2}, \frac{q-1}{q}\}}\right)$
Independent of both d and of h .



Active in the tail: since $\sigma > 0$, the CVaR component is nonzero whenever $\Delta C_h \neq 0$ and $x \neq 0$.
stays active in the **under-sampled tail** however these are important particles (heavy!)

Why a scalar risk function?

The CVaR penalty reduces to a one-dimensional problem



Lemma: finite-sample estimation error

[Fournier–Guillin 2015]

$\hat{\mathbb{P}}_{\text{tar}}$ = empirical measure of N i.i.d. samples of $\mathbb{P}_{\text{tar}}$; $g_{\#}\mathbb{P}_{\text{tar}}$ has a finite q -th moment, $q > 1$

$$\mathbb{E}_{(\mathbb{P}_{\text{tar}})^{\otimes N}} \left| \text{CVaR}_{\alpha,h}^{\hat{\mathbb{P}}_{\text{tar}},g} - \text{CVaR}_{\alpha,h}^{\mathbb{P}_{\text{tar}},g} \right| = O\left(N^{-\min\{\frac{1}{2}, \frac{q-1}{q}\}}\right)$$

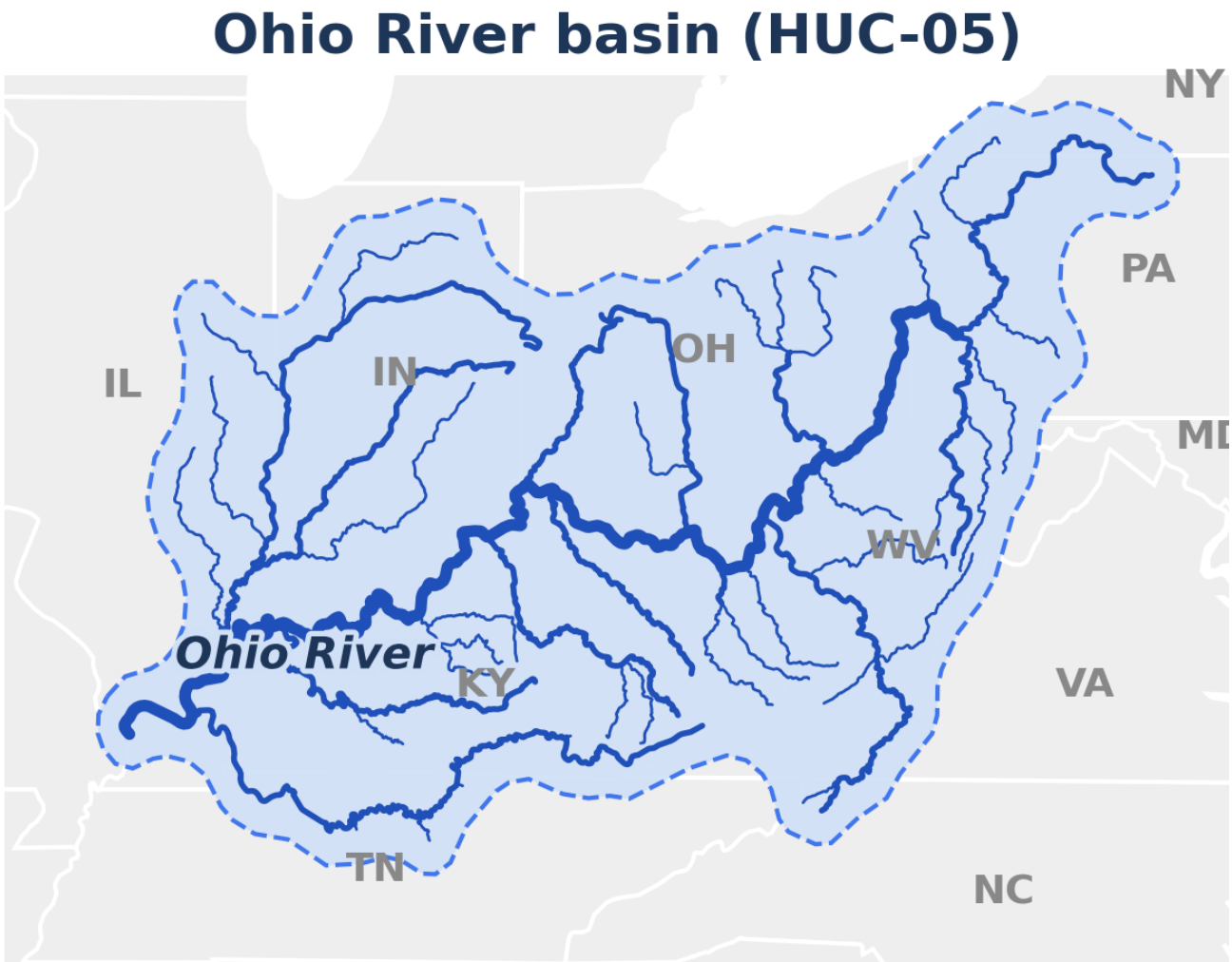
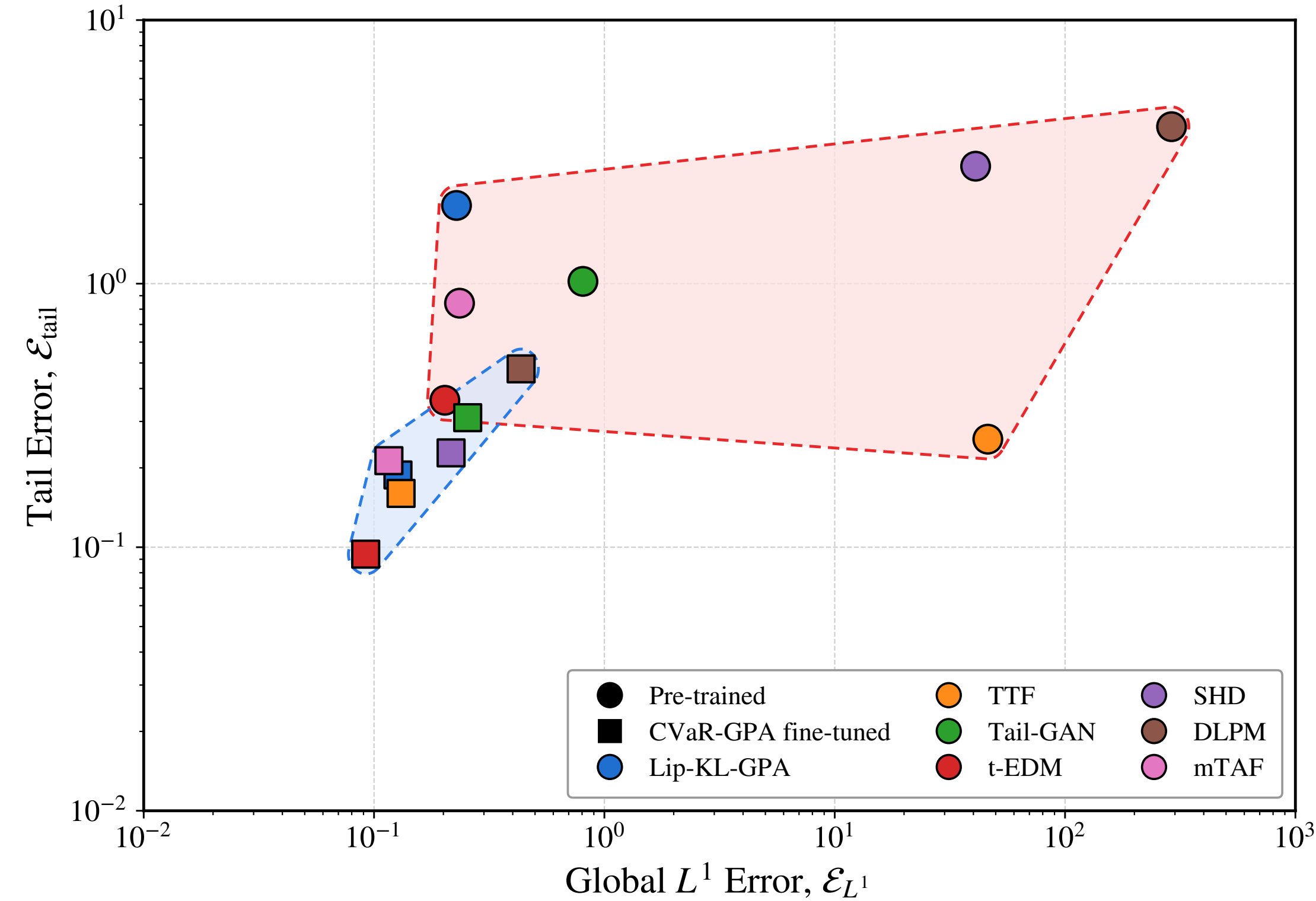
(for $q = 2$, with q replaced by any $q' \in (1, 2)$)

Implicit constant **independent of the dimension d and the smoothing parameter $h > 0$**

Our risk-aware fine-tuned model gets better accuracy learning the distribution globally and in the tail region

Why streamflow? Floods are heavy-tailed, multi-site extremes.
Data: 64 USGS gauges, Ohio River basin, daily specific discharge (2005–2024). **Hill index 1.05–3.34; infinite variance at 51/64 gauges; anisotropic, spatially dependent heavy tails.**

Pre-trained vs. CVaR-GPA Fine-Tuned Models

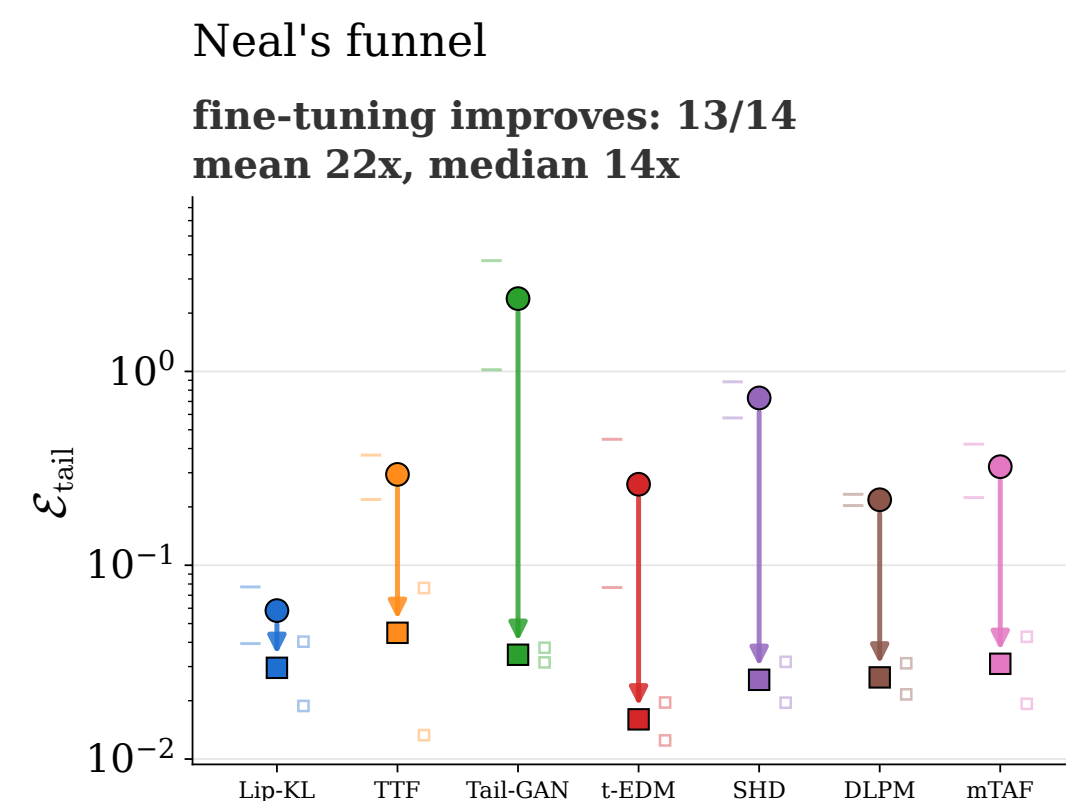
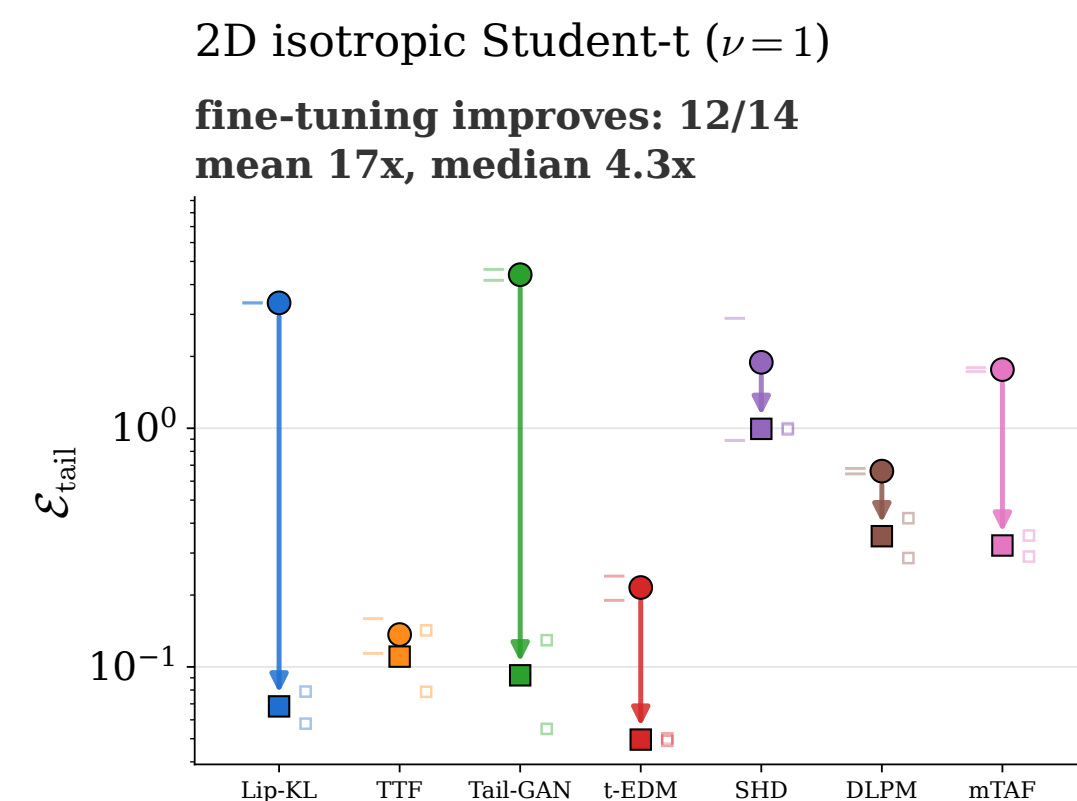


Robust to pretraining quality:
From a weak start to strong results.

Seven pre-trained models across: **GANs** (Tail-GAN), **diffusion** (t-EDM, DLPM, SHD) and **flows** (mTAF, TTF, Lip-KL).

One fine-tuning with **fixed** hyper parameters:
Global error $\downarrow 9.5\times$, tail $\downarrow 5.4\times$ (geometric mean).

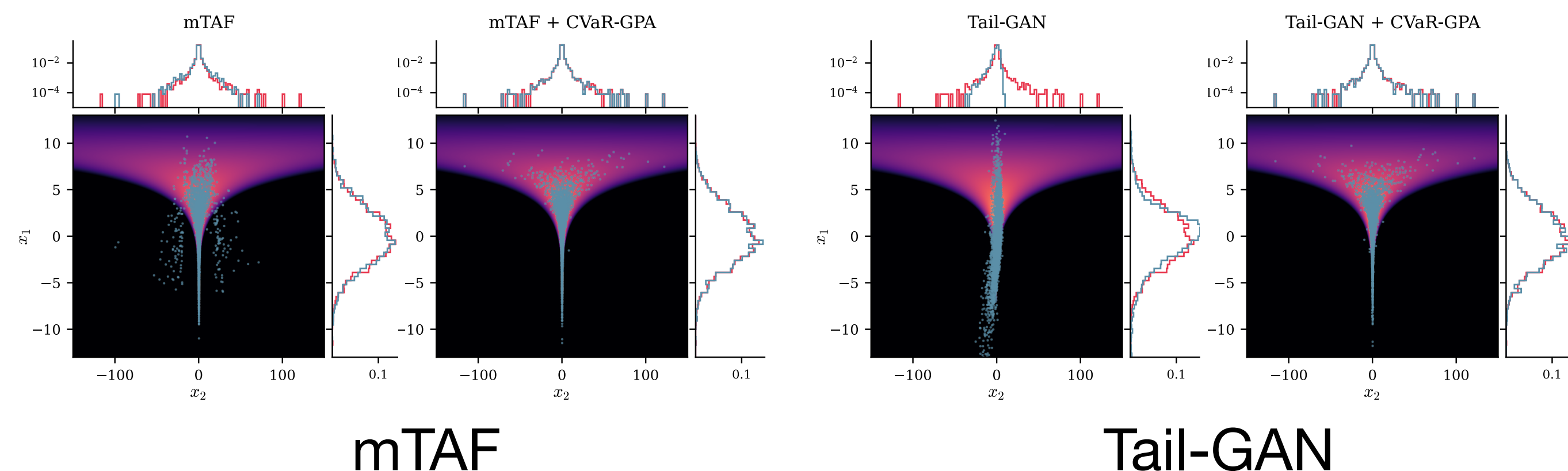
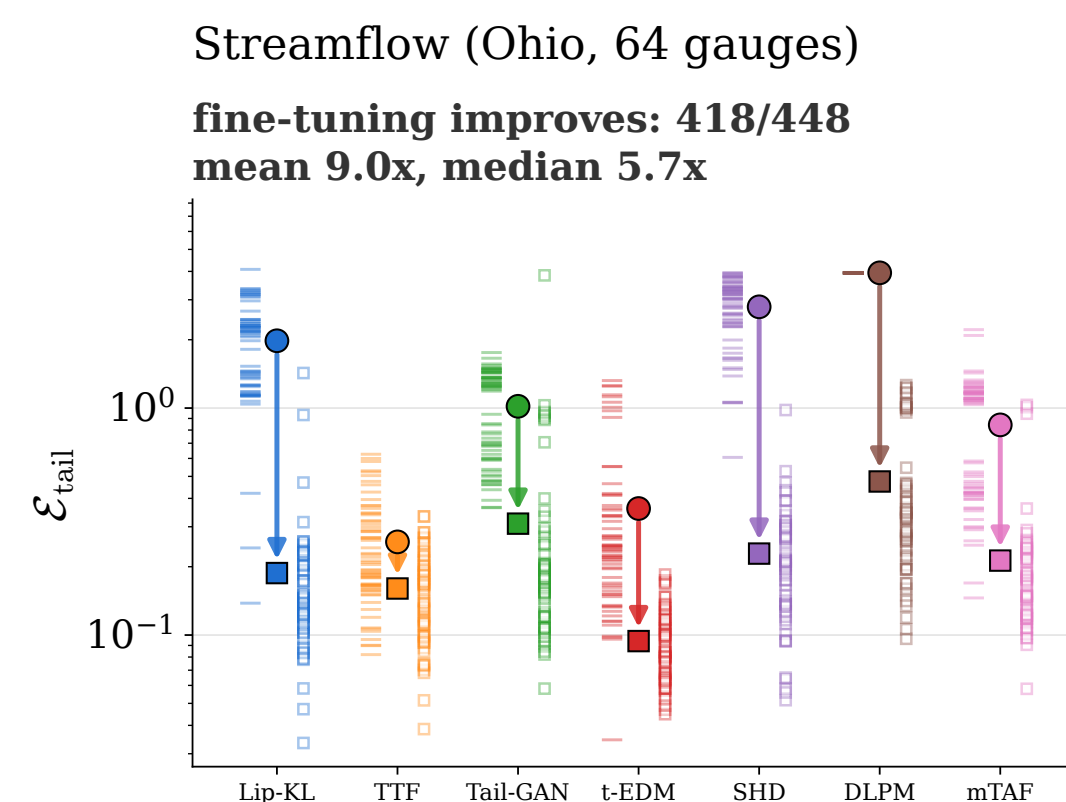
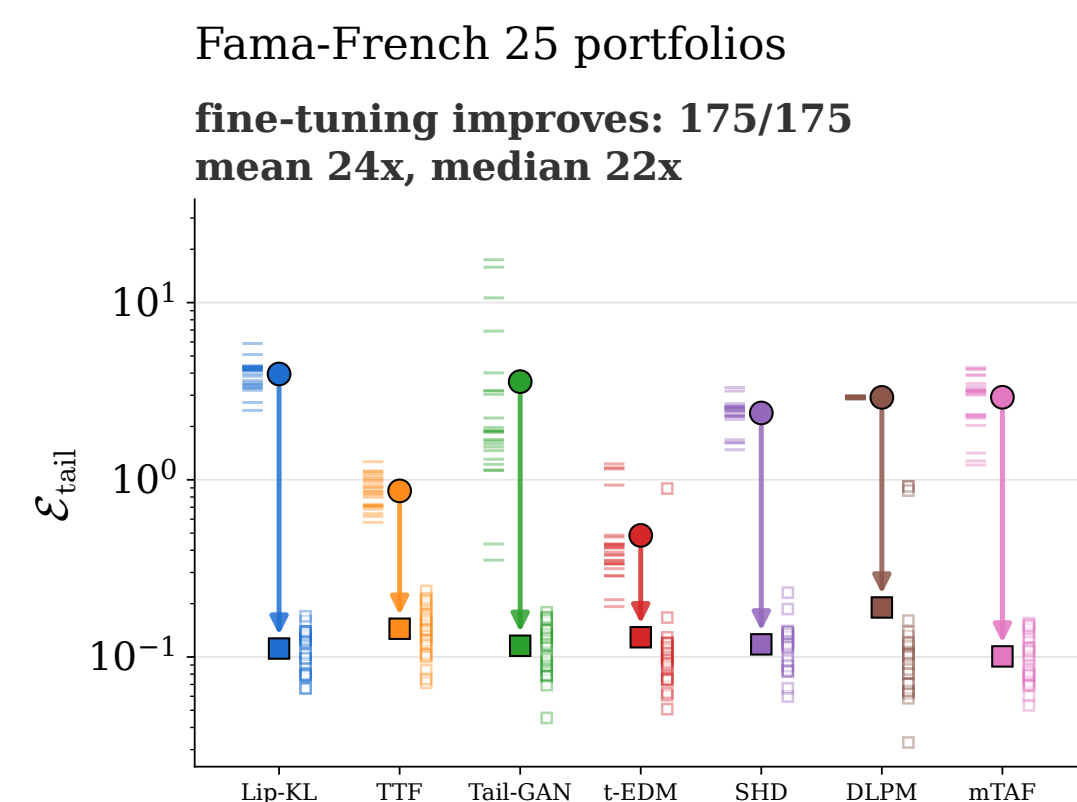
Tail accuracy vs seven generative baselines, four target datasets



Four targets — 2-d Cauchy, Neal's funnel, Fama–French 25, Ohio streamflow.

Seven pre-trained models, same hyperparameters — nothing retuned.

618 of 651 dimensions improve in tail error (**641 of 651** globally).



- pre-trained (mean over dimensions)
- CVaR-GPA fine-tuned (mean)
- per-dimension, pre-trained
- per-dimension, fine-tuned

CVaR-GPA repairs the tails while preserving the structure.

Takeaway CVaR-GPA is a fine-tuning layer: it improves whatever model you already trained.

Why fine-tuning works with this gradient flow?

$$\partial_t Q_t = \nabla \cdot \left(Q_t \nabla_x \frac{\delta \mathcal{F}}{\delta Q}(Q_t) \right), \quad Q_0 = P_{\text{pre}}$$

1. Start from any data

The initial condition can be any pre-trained model's samples; (GAN, diffusion model, flow,)

$$Q_0 = P_{\text{pre}}$$

“Model-agnostic”, could be just a data distribution

2. Design the free energy

Free to add the quantity of interest **QoI** you care about to the energy the flow minimizes.

$$\mathcal{F} = \text{KL}^L(Q \parallel P_{\text{tar}}) + \lambda \text{QoI}$$

QoI = $(\Delta \text{CVaR})^2$: the flow learns heavy tails—“agnostic”

3. Go as deep as needed

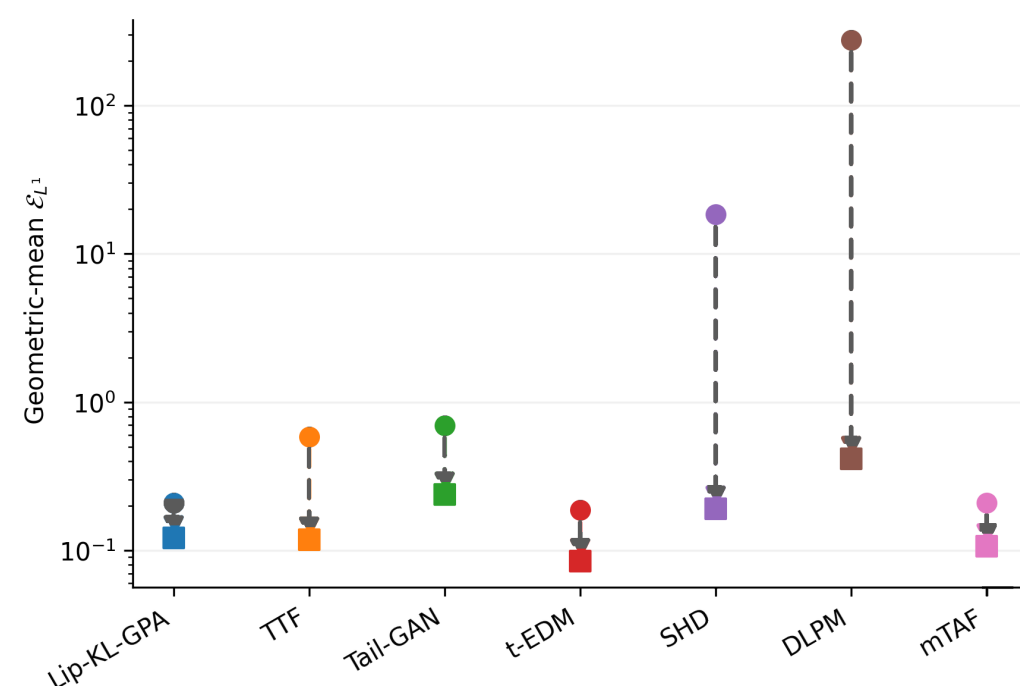
The gradient flow dissipates $\mathcal{F}$.

$$\frac{d}{dt} \mathcal{F}(Q_t) = - \int Q_t \left| \nabla_x \frac{\delta \mathcal{F}}{\delta Q} \right|^2 dx \leq 0$$

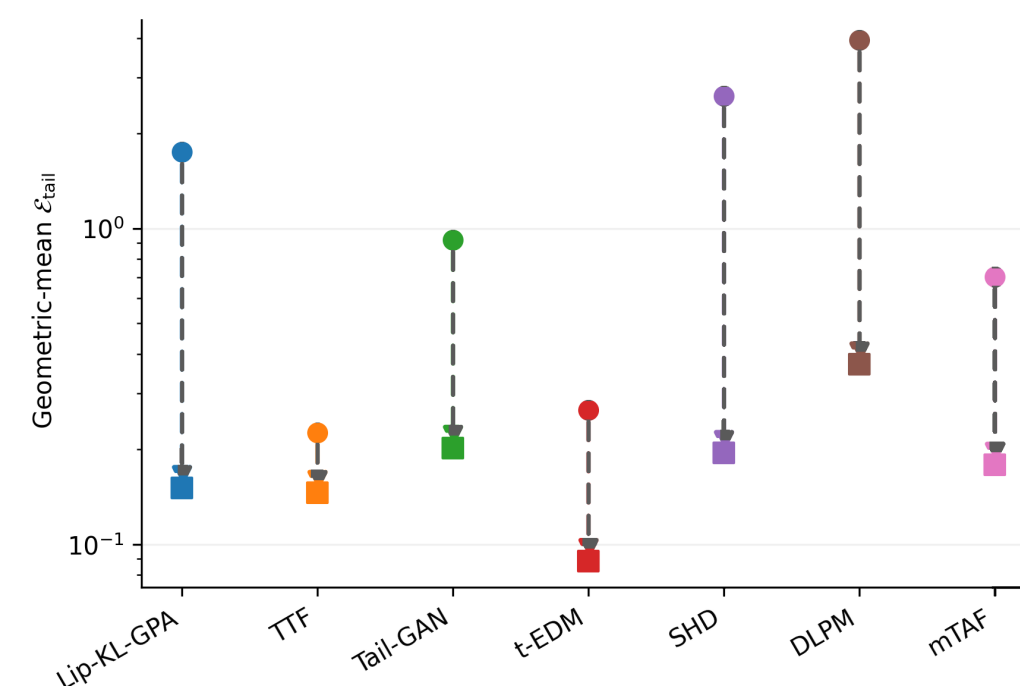
Architecture adapts to capturing the target

● Pre-trained ■ Fine-tuned

L^1 error (64 gauges)



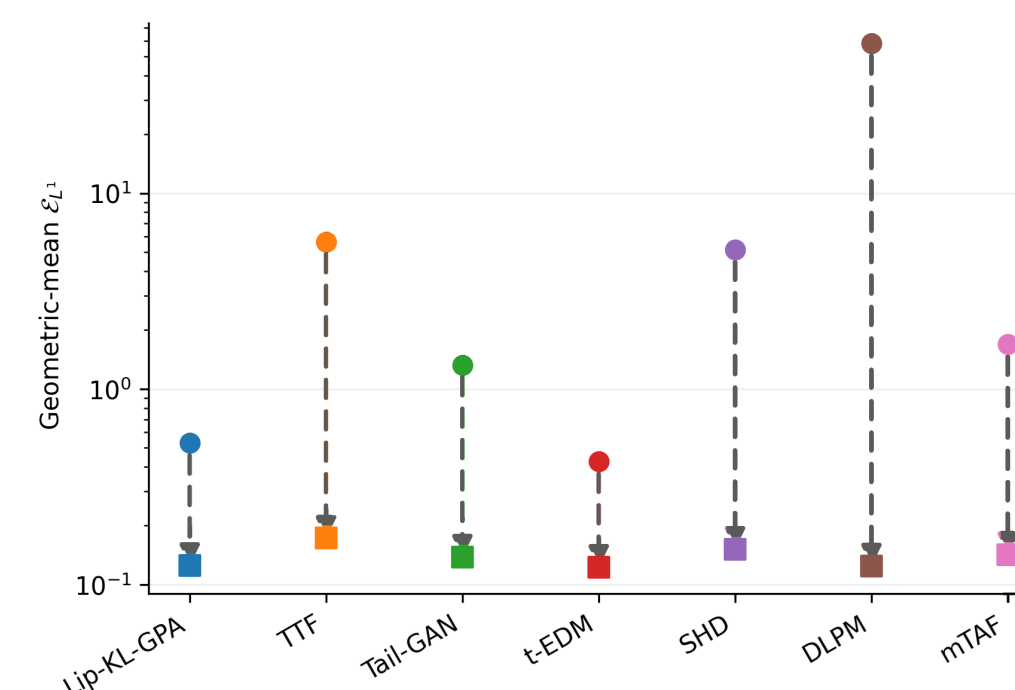
Tail error (64 gauges)



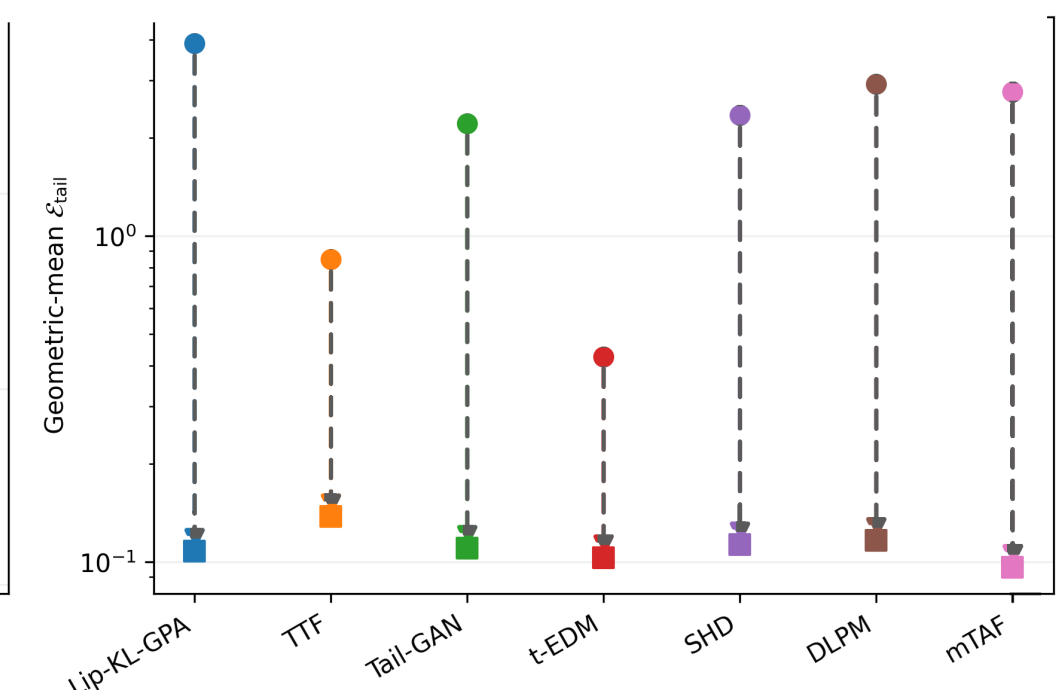
Data sets:
Streamflow
French25

● Pre-trained ■ Fine-tuned

L^1 error (25 portfolios)



Tail error (25 portfolios)



Conclusions

CVaR-GPA

- Fine-tune any pre-trained model from its samples alone
- Physics-inspired: Wasserstein gradient flow of a free energy
 - Lipschitz-regularized KL for the bulk + CVaR penalty on an observable
 - Tail-agnostic; target enters only via one observable, dimension-free estimation
 - Depth fitted to the target by free-energy dissipation

Outlook

- Connections with sampling methods for extreme events?
 - Generative modeling and sampling: two sides of the same coin
- New and upcoming generative AI methods for extreme events
 - e.g. extreme-event-aware (η -) learning [Chang & Sapsis, Nat. Commun. 2026]
 - CVaR-GPA fine-tuning can apply to their output samples
- Richer free energies: other risk measures, observables, multiple observables