

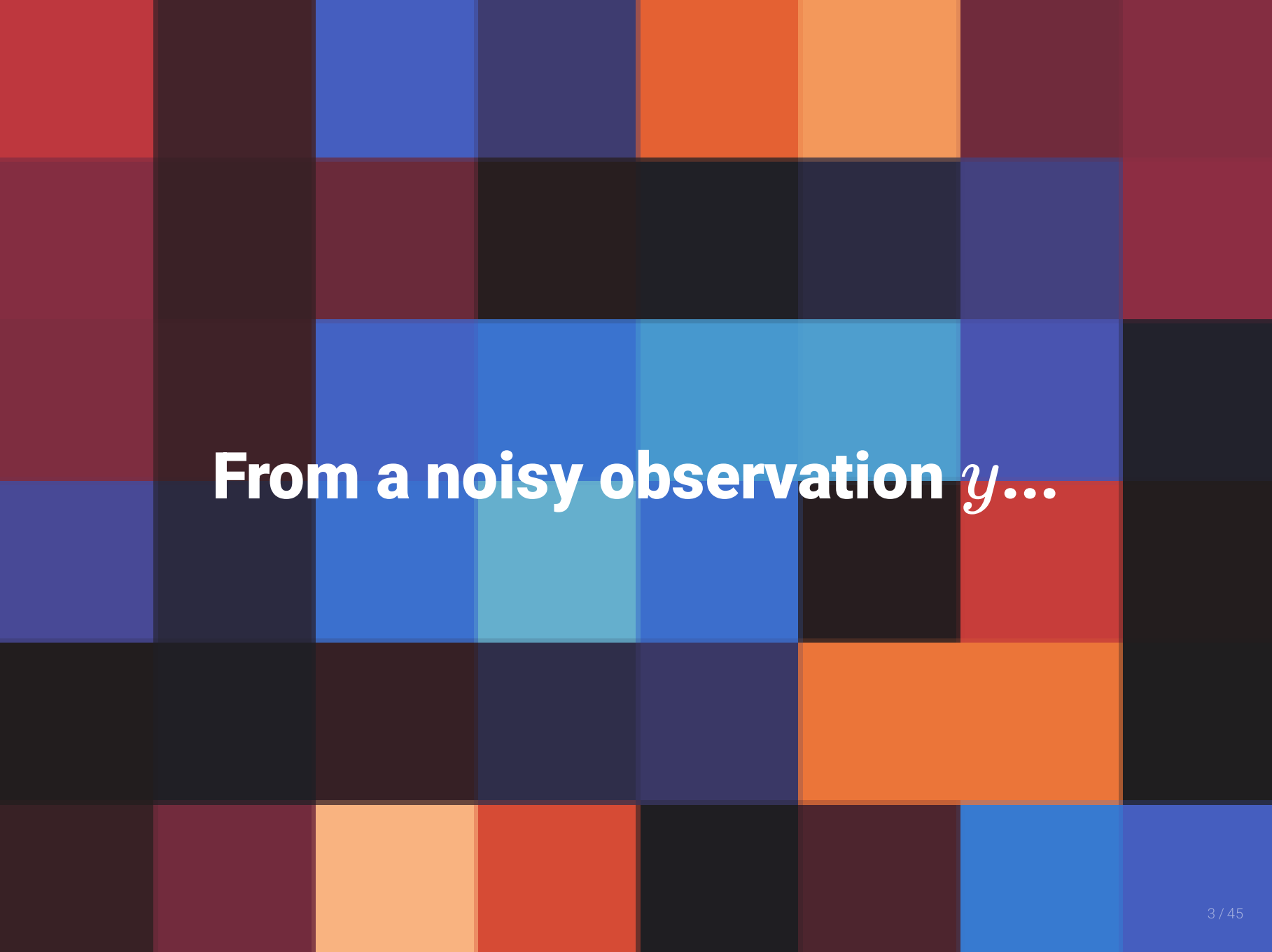
Inverting scientific images with generative models

Generative Modeling Summer School / Statlearn
April 3, 2025

Gilles Louppe
g.louppe@uliege.be



(G  rome, Fran  ois, Victor, Omer, Sacha, Matthias, Elise, Malavika, Arnaud, Joeri)



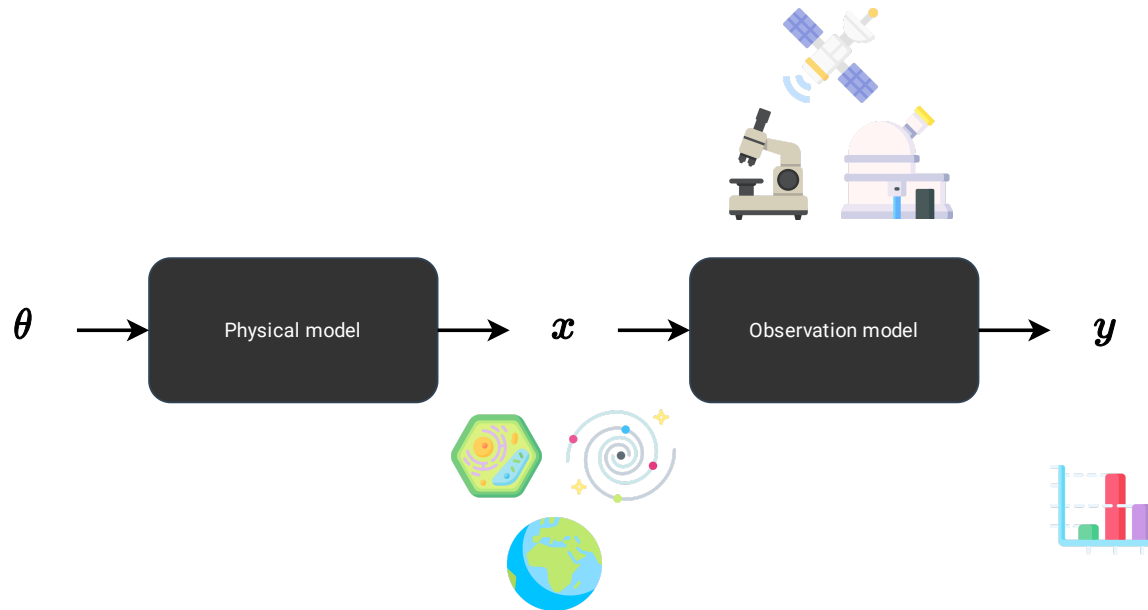
From a noisy observation y ...

**... can we recover
all plausible images x ?**

$$\dot{u} = -u \nabla u + \frac{1}{Re} \nabla^2 u - \frac{1}{\rho} \nabla p + f$$

$$0 = \nabla \cdot u$$

... or parameters $\theta = \{Re, \rho, f\}$?



Problem statement (Bayesian inference)

Given a noisy observation y , estimate either

- the posterior distribution $p(x|y) \propto p(x)p(y|x)$ of latent states x , or
- the posterior distribution $p(\theta|y)$ of parameters θ ,

assuming a prior $p(\theta)$ and forward models $p(x|\theta)$ and $p(y|x)$.



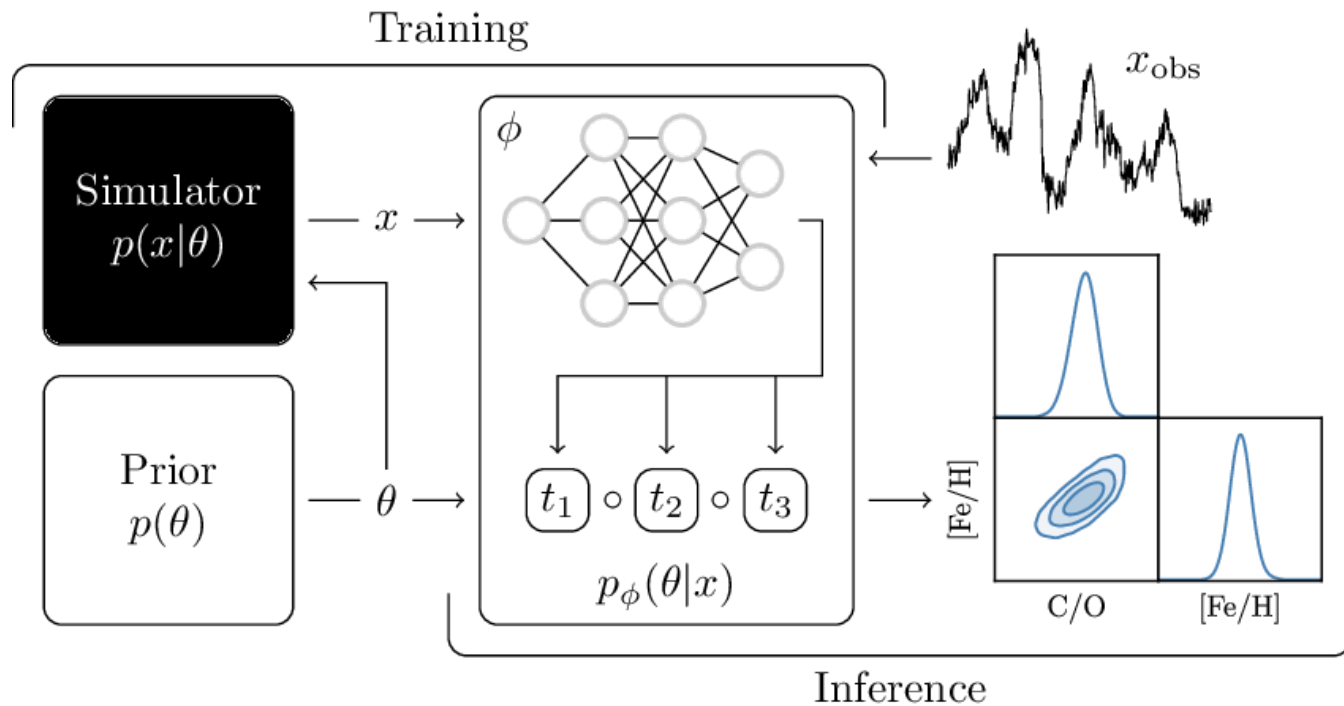
A word on notations...

- In the inverse problem literature, x is often used to denote the latent state, while y is used to denote the observation.
- In simulation-based inference, x is often used to denote the observation and the latent state is not denoted. Inference is done on the parameters θ .
- In the generative modeling literature, x often denotes an observation, while z is used to denote the latent state.

Sorry for the mess, not my fault!

Neural posterior estimation

Neural posterior estimation (NPE) is a **simulation-based inference** algorithm that learns to approximate the posterior distribution $p(\theta|x)$ of parameters θ given a noisy observation x .

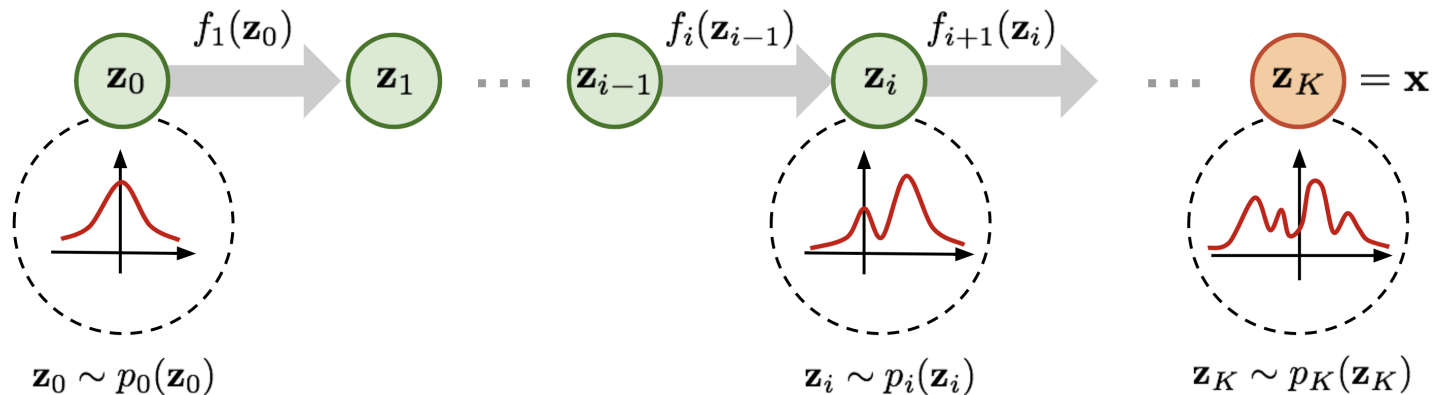


NPE amortizes **variational inference** by training a neural density estimator q_ϕ that minimizes

$$\mathbb{E}_{p(x)} [\text{KL}(p(\theta|x) || q_\phi(\theta|x))] .$$

Normalizing flows

A normalizing flow is a sequence of invertible transformations f_k that map a simple distribution p_0 to a more complex distribution p_K :



By the change of variables formula, the log-likelihood of a sample $\mathbf{x}$ is given by

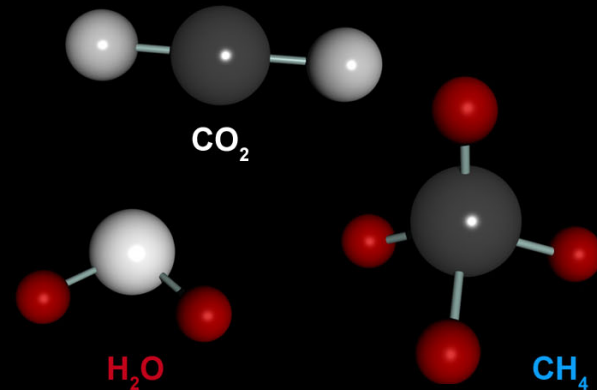
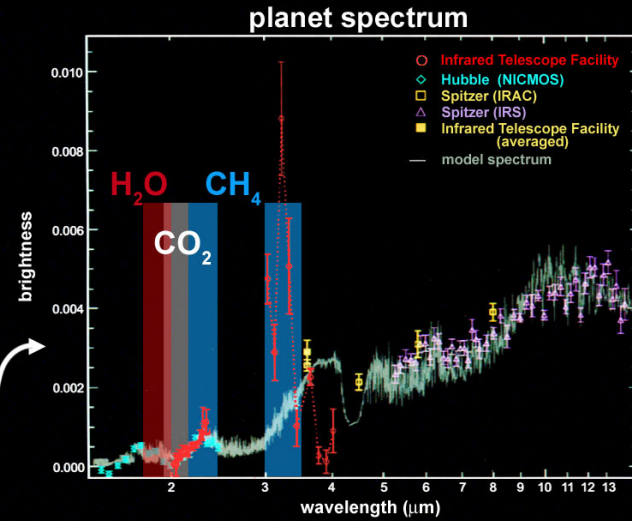
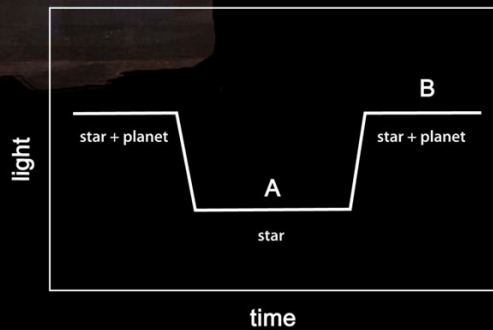
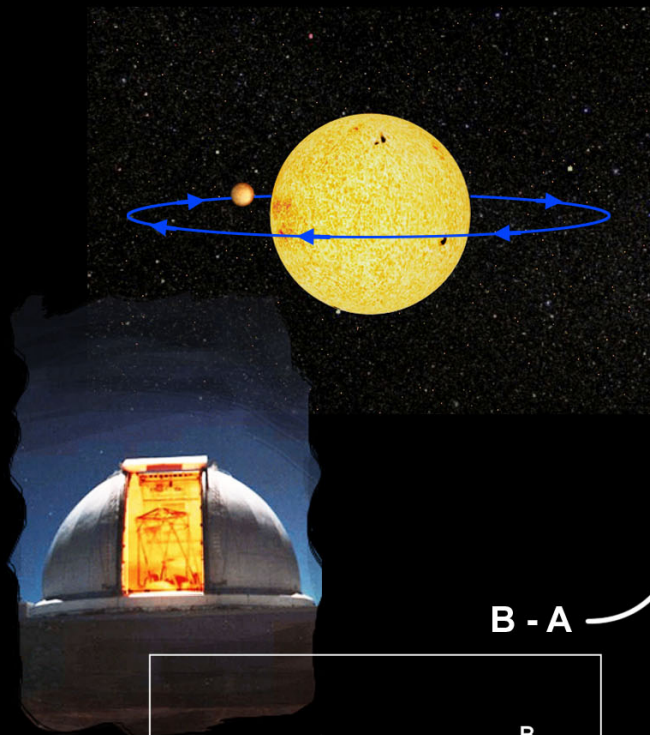
$$\log p(\mathbf{x}) = \log p(\mathbf{z}_0) - \sum_{k=1}^K \log |\det J_{f_k}(\mathbf{z}_{k-1})|.$$

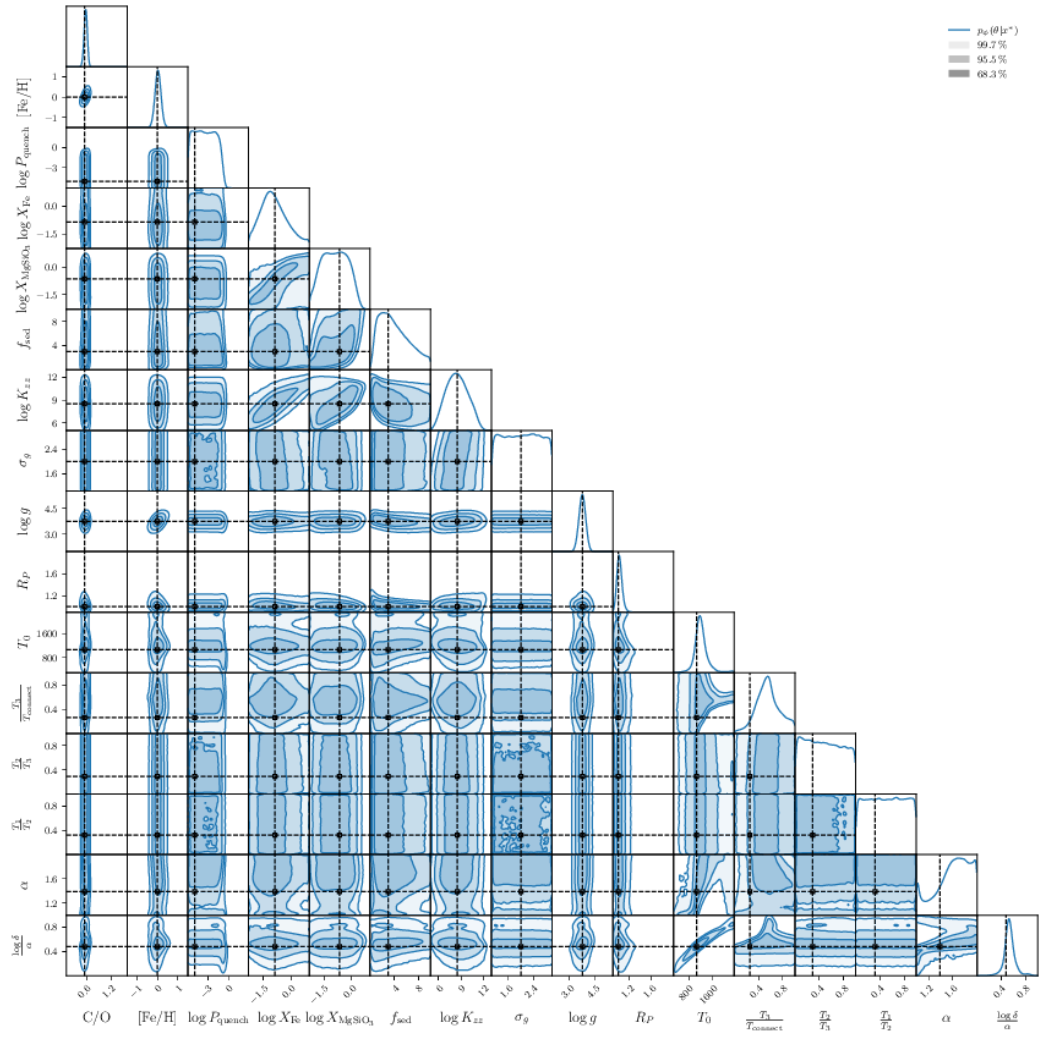
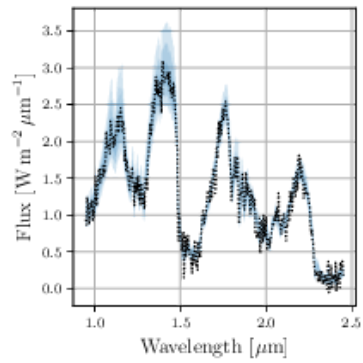
NPE often relies on normalizing flows to model the posterior distribution $p(\theta|x)$ because

- they are fast and easy to train,
- they are easy to sample from,
- they are universal approximators,
- they provide exact likelihood estimates.



Exoplanet atmosphere characterization





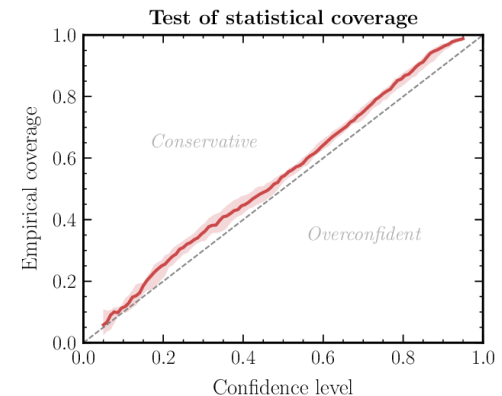
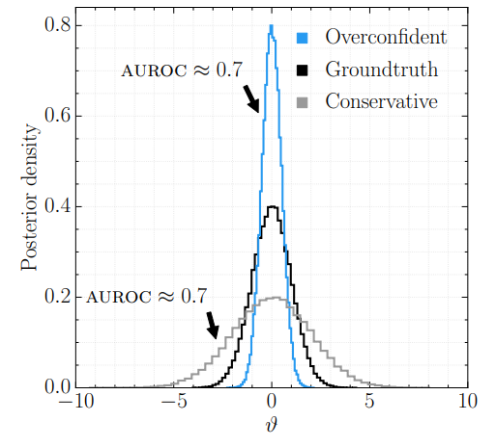


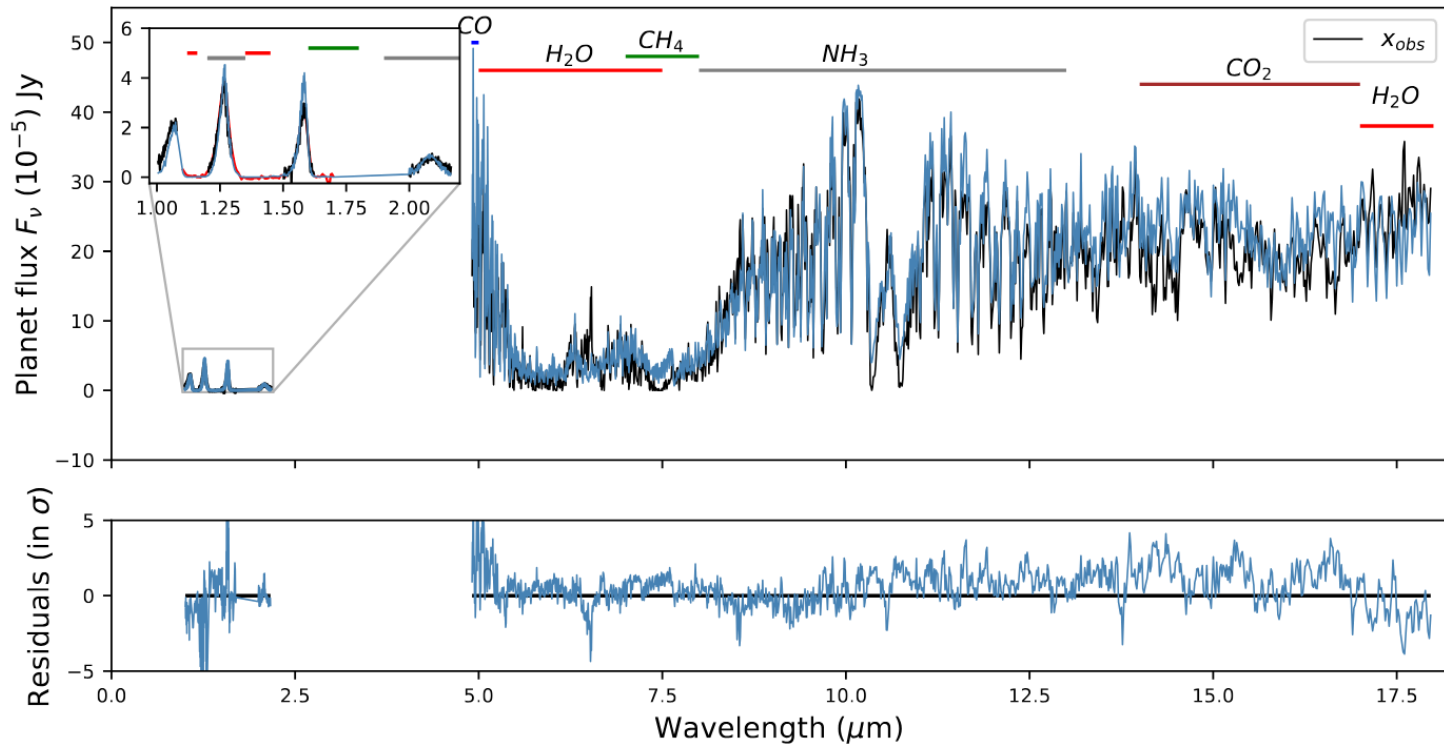
Expected coverage

$$\text{EC}(\hat{p}, \alpha) = \mathbb{E}_{p(\theta, \mathbf{x})}[\theta \in \Theta_{\hat{p}(\theta|\mathbf{x})}(\alpha)]$$

If the expected coverage is close to the nominal coverage probability α , then the approximate posterior $\hat{p}$ is calibrated.

- If $\text{EC} < \alpha$, then the posterior is underdispersed and overconfident.
- If $\text{EC} > \alpha$, then the posterior is overdispersed and conservative.





Panchromatic characterization of WISE 1738 using JWST/MIRI
(Vasist et al, submitted).



What if the posterior distribution is high-dimensional? ($O(10^3+)$)

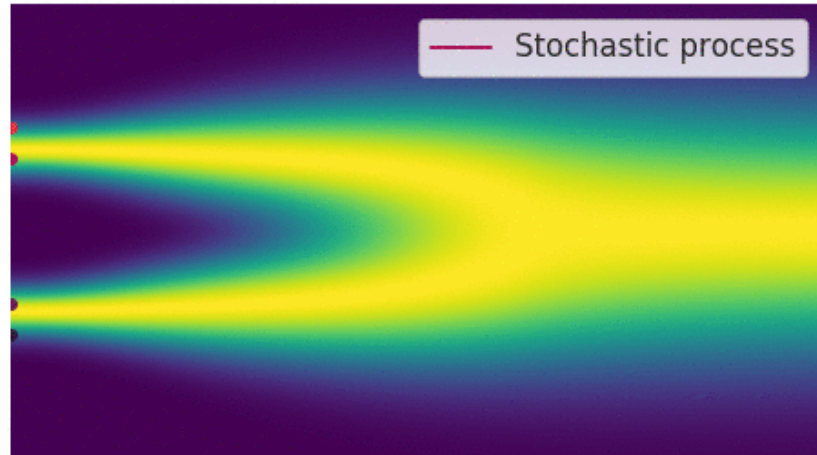
Diffusion models for inverse problems

Diffusion models 101

Samples $x \sim p(x)$ are progressively perturbed through a diffusion process described by the forward SDE

$$dx_t = f_t x_t dt + g_t dw_t,$$

where x_t is the perturbed sample at time t .

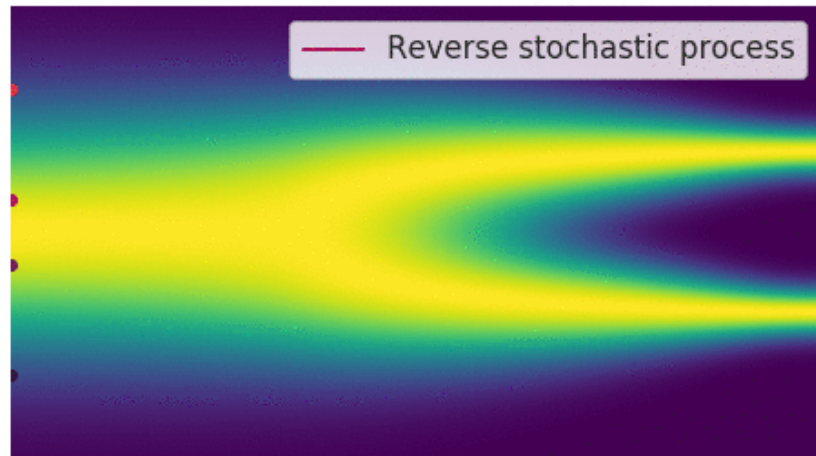


Forward diffusion process.

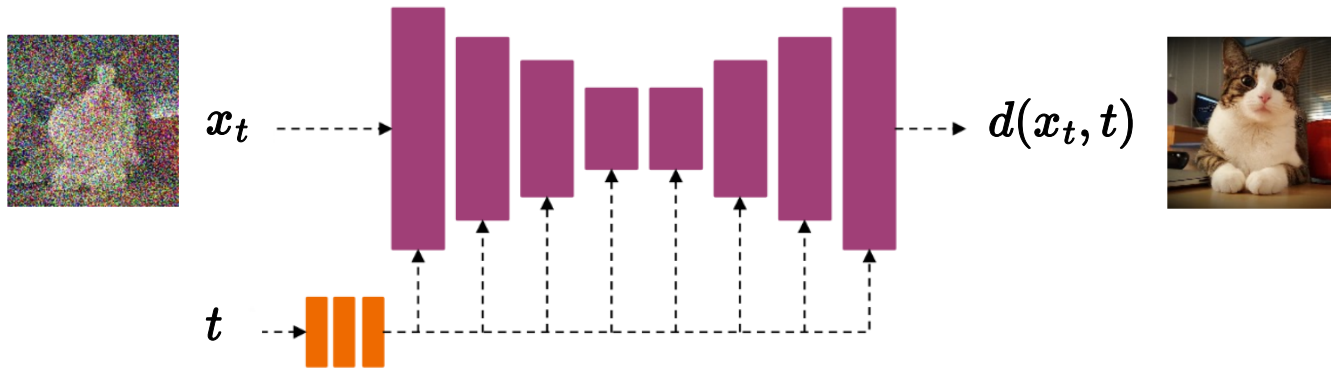
The reverse process satisfies a reverse-time SDE that can be derived analytically from the forward SDE as

$$dx_t = [f_t x_t - g_t^2 \nabla_{x_t} \log p(x_t)] dt + g_t dw_t.$$

Therefore, to generate data samples $x_0 \sim p(x_0) \approx p(x)$, we can draw noise samples $x_1 \sim p(x_1) \approx \mathcal{N}(0, \Sigma_1)$ and gradually remove the noise therein by simulating the reverse SDE from $t = 1$ to 0 .



Reverse denoising process.



The score function $\nabla_{x_t} \log p(x_t)$ is unknown, but can be approximated by a neural network $d_\theta(x_t, t)$ by minimizing the denoising score matching objective

$$\mathbb{E}_{p(x)p(t)p(x_t|x)} \left[\|d_\theta(x_t, t) - x\|_2^2 \right].$$

The optimal denoiser d_θ is the mean $\mathbb{E}[x|x_t]$ which, via Tweedie's formula, allows to use

$$s_\theta(x_t, t) = \Sigma_t^{-1} (d_\theta(x_t, t) - x_t)$$

as a score estimate of $\nabla_{x_t} \log p(x_t)$ in the reverse SDE.

Inverting single observations

To turn a diffusion model $p_\theta(\mathbf{x}_{0:T})$ into a conditional model, we can add conditioning information y at each step of the reverse process, as

$$p_\theta(\mathbf{x}_{0:T}|y) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, y).$$

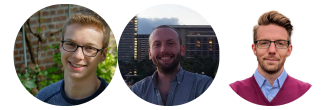
This can be done by adding y as an additional input to the denoiser $d_\theta(x_t, t, y)$.



Using the Bayes' rule, the posterior score $\nabla_{x_t} \log p(x_t|y)$ to inject in the reverse SDE can be decomposed as

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y|x_t) - \cancel{\nabla_{x_t} \log p(y)}.$$

This enables **zero-shot posterior sampling** from a diffusion prior $p(x_0)$ without having to pre-wire the neural denoiser to the observation model $p(y|x)$.



Approximating $\nabla_{x_t} \log p(y|x_t)$

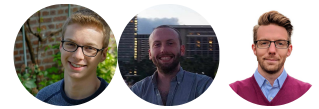
We want to estimate the score $\nabla_{x_t} \log p(y|x_t)$ of the noise-perturbed likelihood

$$p(y|x_t) = \int p(y|x)p(x|x_t)dx.$$

Our approach:

- Assume a linear Gaussian observation model $p(y|x) = \mathcal{N}(y|Ax, \Sigma_y)$.
- Assume the approximation $p(x|x_t) \approx \mathcal{N}(x|\mathbb{E}[x|x_t], \mathbb{V}[x|x_t])$, where $\mathbb{E}[x|x_t]$ is estimated by the denoiser and $\mathbb{V}[x|x_t]$ is estimated using Tweedie's covariance formula.
- Then $p(y|x_t) \approx \mathcal{N}(y|A\mathbb{E}[x|x_t], \Sigma_y + A\mathbb{V}[x|x_t]A^T)$.
- The score $\nabla_{x_t} \log p(y|x_t)$ then approximates to

$$\nabla_{x_t} \mathbb{E}[x|x_t]^T A^T (\Sigma_y + A\mathbb{V}[x|x_t]A^T)^{-1} (y - A\mathbb{E}[x|x_t]).$$



Approximation

We want to
likelihood

Our approach

- Assume
- Assume $\mathbb{E}[x|x_t]$
Tweedie
- Then $p(\dots)$
- The score

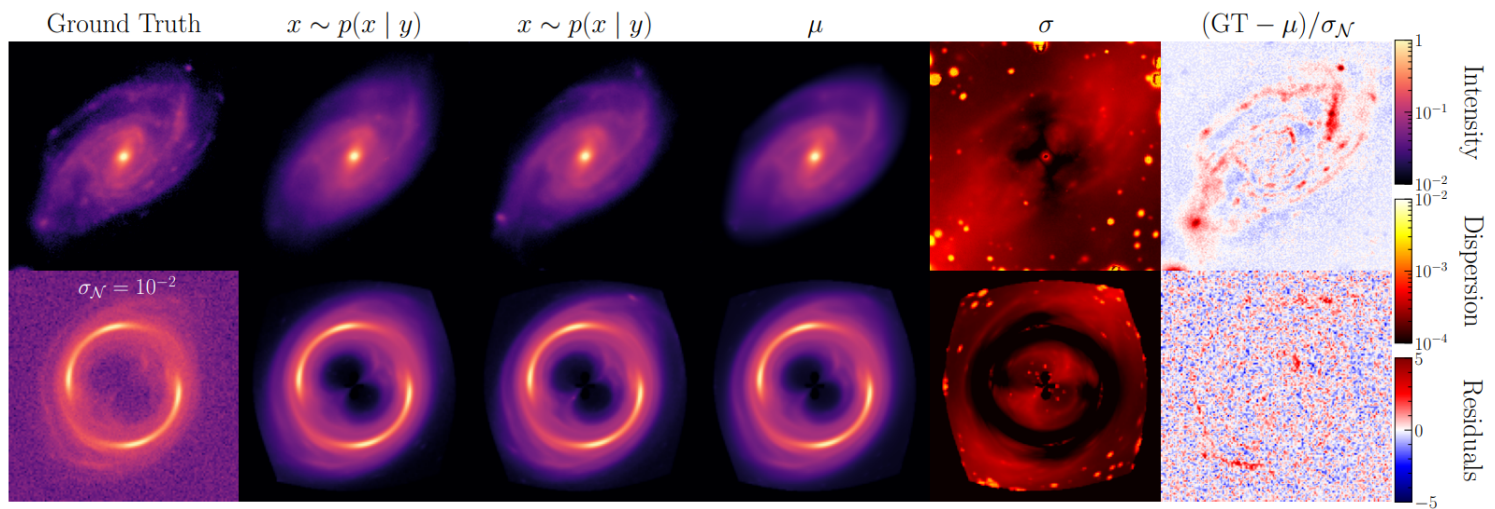
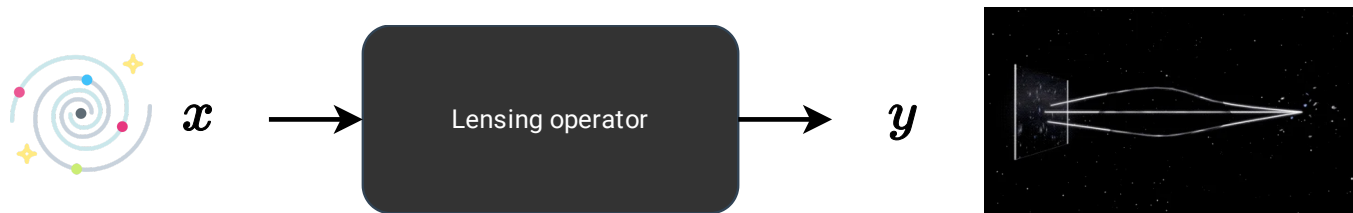


ed

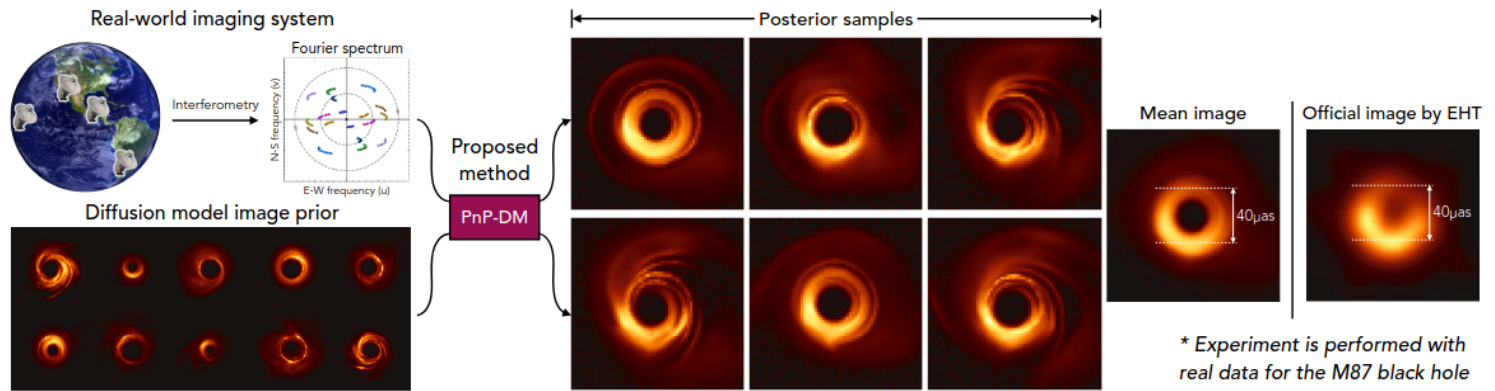
x, Σ_y).

where
using

$t]$).



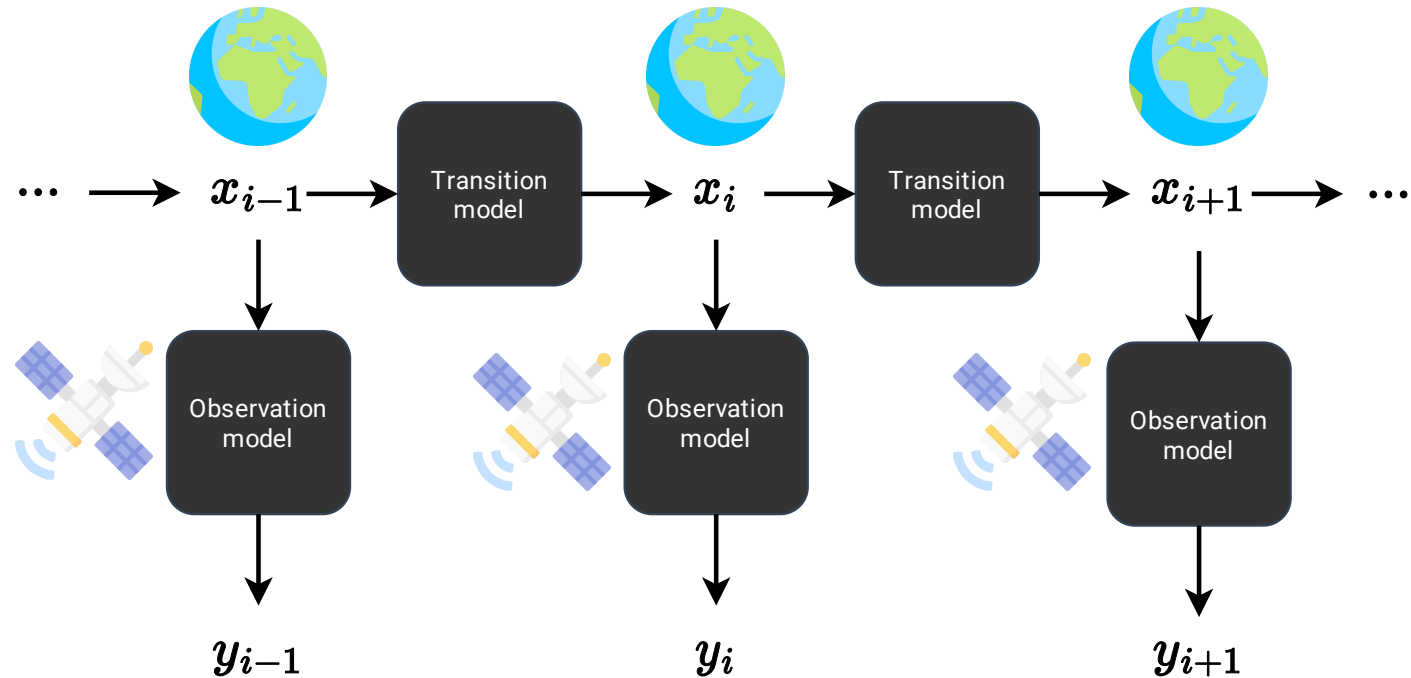
Posterior source galaxies x can be recovered from gravitational lenses y by zero-shot posterior sampling from a diffusion prior $p(x)$ of galaxy images.



Posterior M87 black hole images x using a diffusion prior $p(x)$ based on GRMHD simulations.

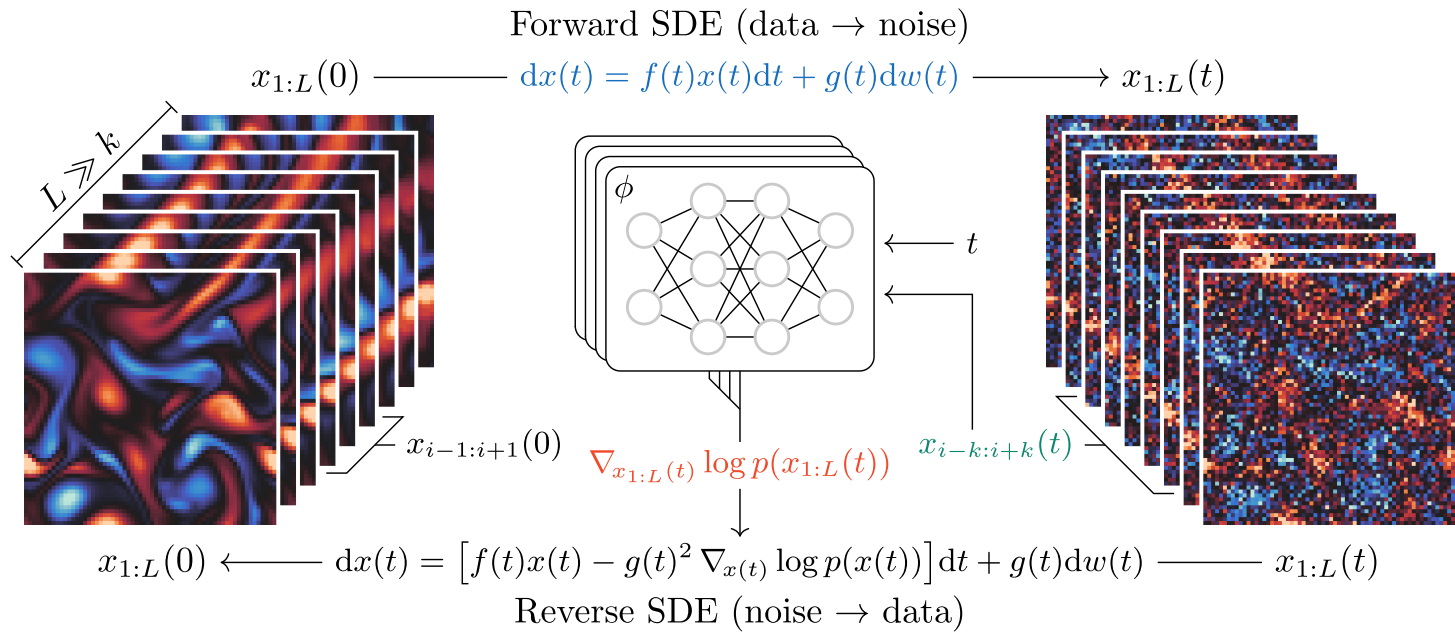


Score-based data assimilation (SDA) in dynamical systems



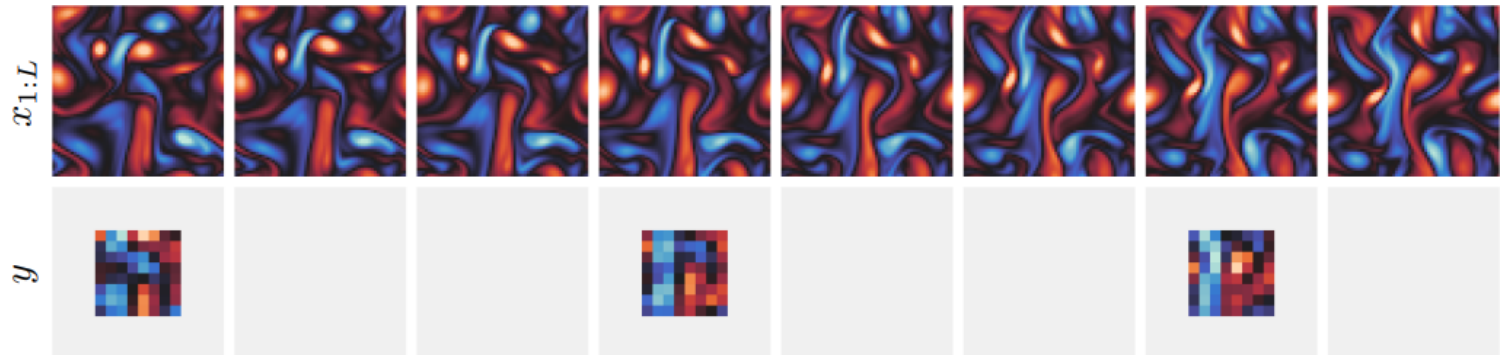
The goal of **data assimilation** is to estimate plausible trajectories $x_{1:L}$ given one or more noisy observations y (or $y_{1:L}$) as the posterior

$$p(x_{1:L}|y) = \frac{p(y|x_{1:L})}{p(y)} p(x_0) \prod_{i=1}^{L-1} p(x_{i+1}|x_i).$$

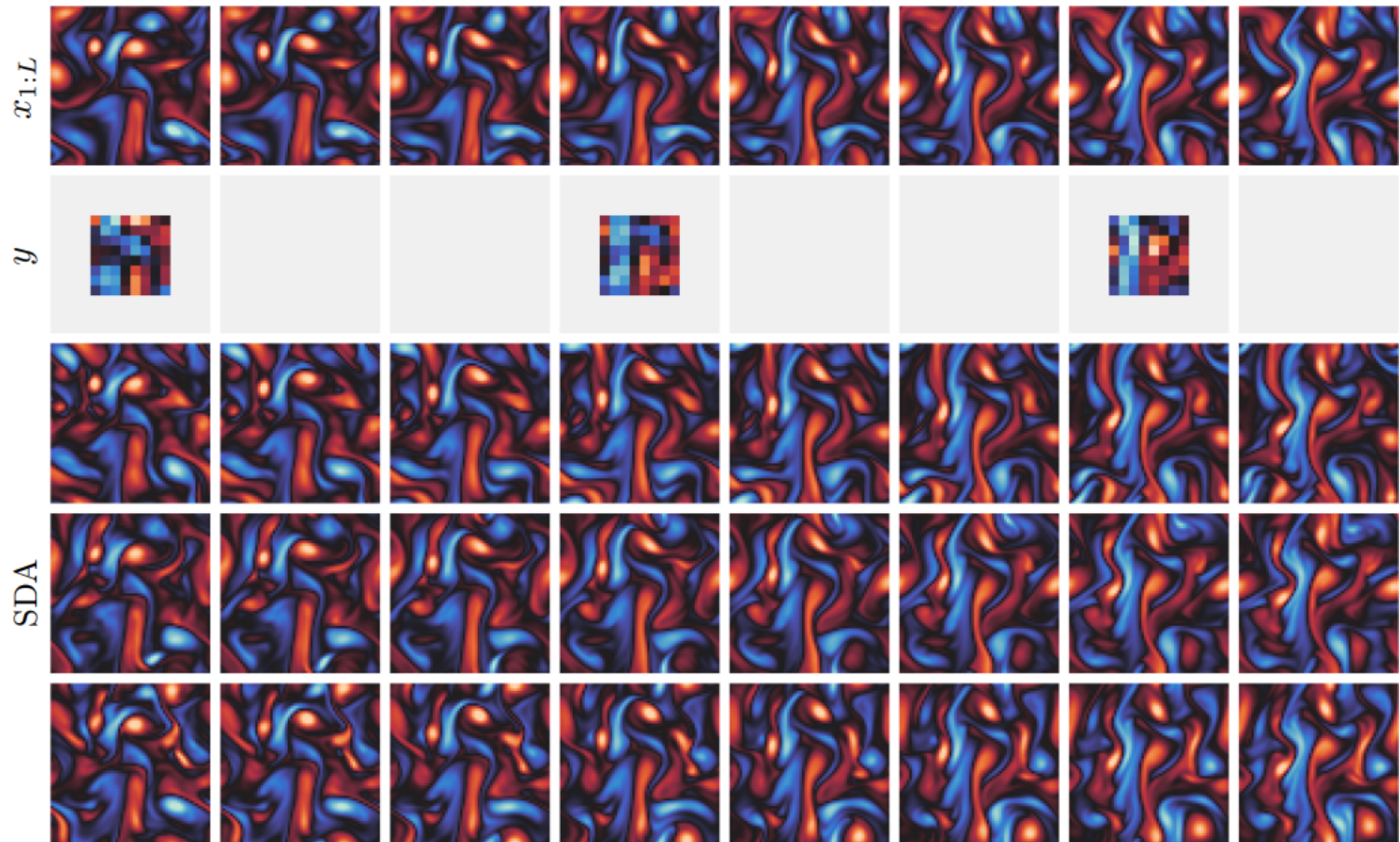


Our approach:

- Build a score-based generative model $p(x_{1:L})$ of arbitrary-length trajectories*.
- Use zero-shot posterior sampling to generate plausible trajectories from noisy observations y .



Sampling trajectories $x_{1:L}$ from
noisy, incomplete and coarse-grained observations y .



Sampling trajectories $x_{1:L}$ from
noisy, incomplete and coarse-grained observations y .

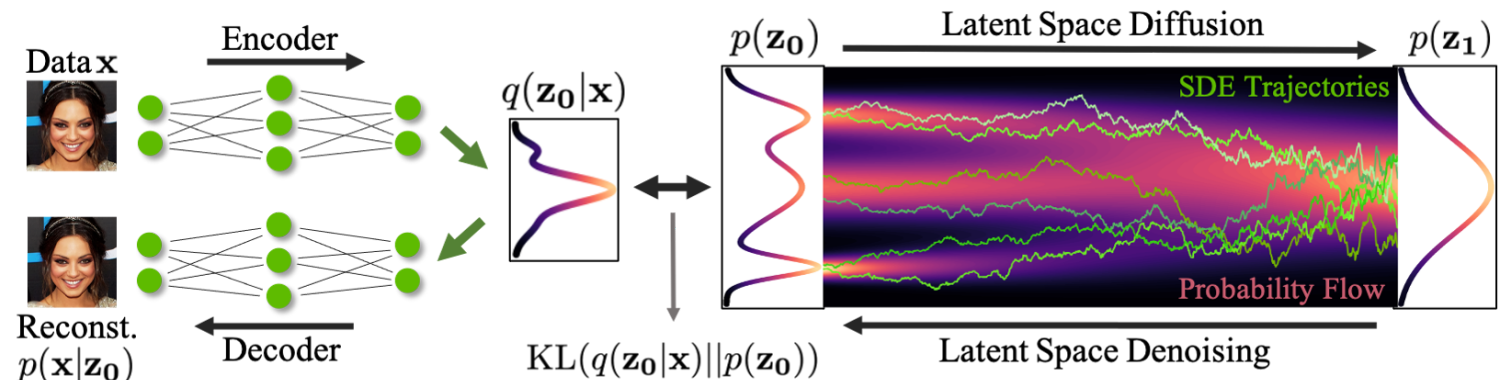


What if the posterior distribution is even larger? ($O(10^6 +)$)

Latent diffusion models (LDMs)

Directly modeling the data distribution becomes difficult for high-dimensional data, such as images or videos.

A more effective approach (à la Stable Diffusion) is to model a latent distribution $p_\theta(z)$ with a diffusion prior and then use a decoder $D(z)$ to map z to the data space x .



LDMs for inverse problems

$$x \sim p_\theta(x|y) \Leftrightarrow z \sim p_\theta(z|y), x = D(z)$$

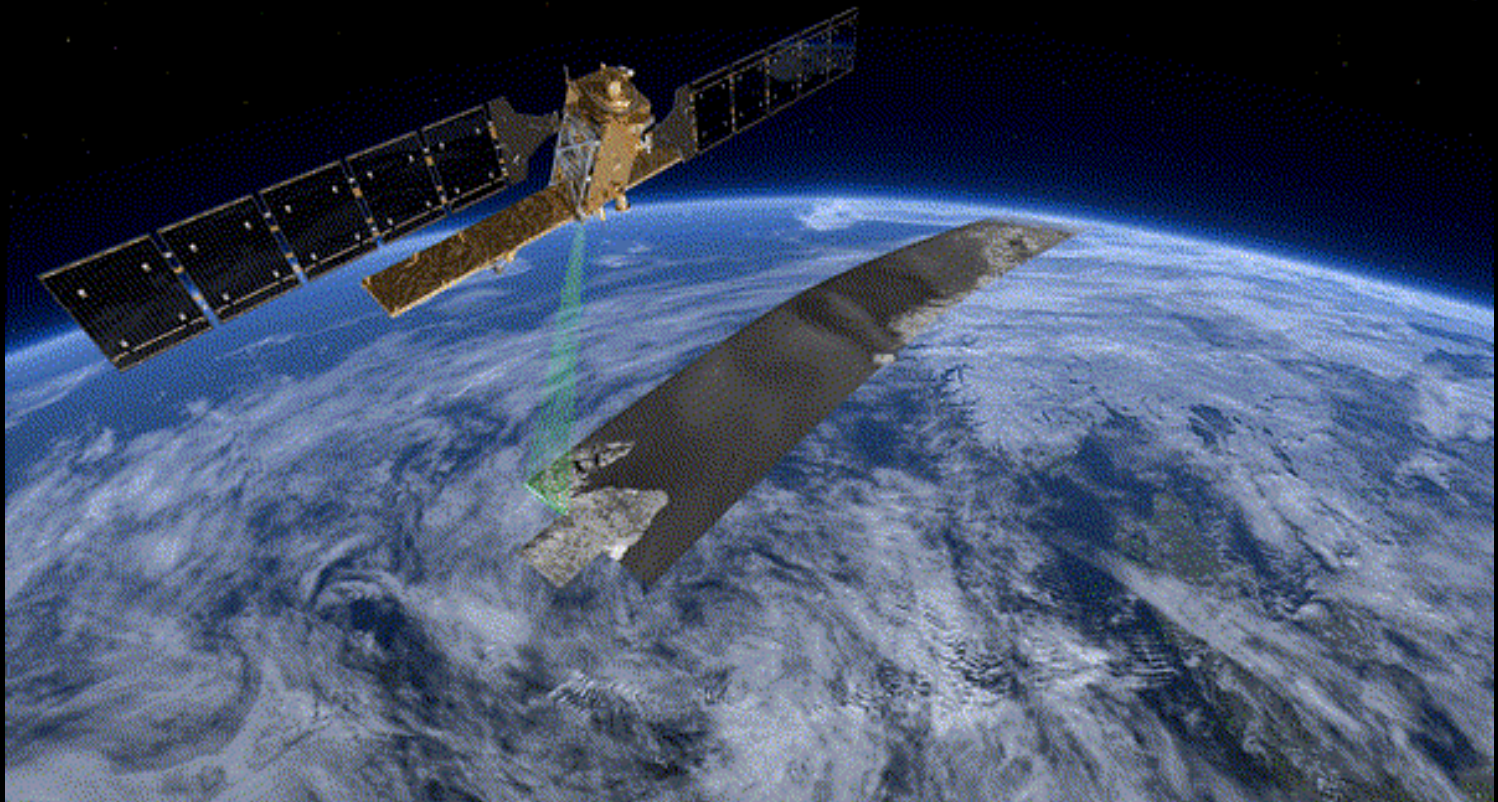
But how to sample from $p_\theta(z|y)$?

- If the observation model in data space is $p(y|x) = \mathcal{N}(y|M(x), \Sigma_y)$, then an observation model in the latent space can be defined as $p(y|z) = \mathcal{N}(y|M(D(z)), \Sigma_y)$.
- Regular posterior sampling algorithms can then be used to sample from the posterior distribution $p_\theta(z|y)$.

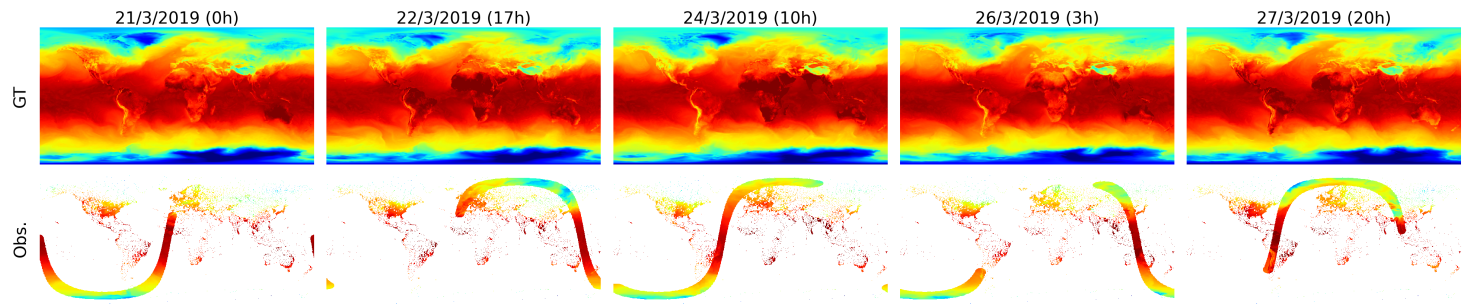


Can we scale SDA (Rozet and Louppe, 2023) to a whole Earth model?

At 0.25° resolution, for 6 atmospheric variables, 13 pressure levels, hourly time steps, and 14 days of simulation, a trajectory $x_{1:L}$ contains $721 \times 1440 \times 6 \times 13 \times 24 \times 14 = 27 \times 10^9$ variables.



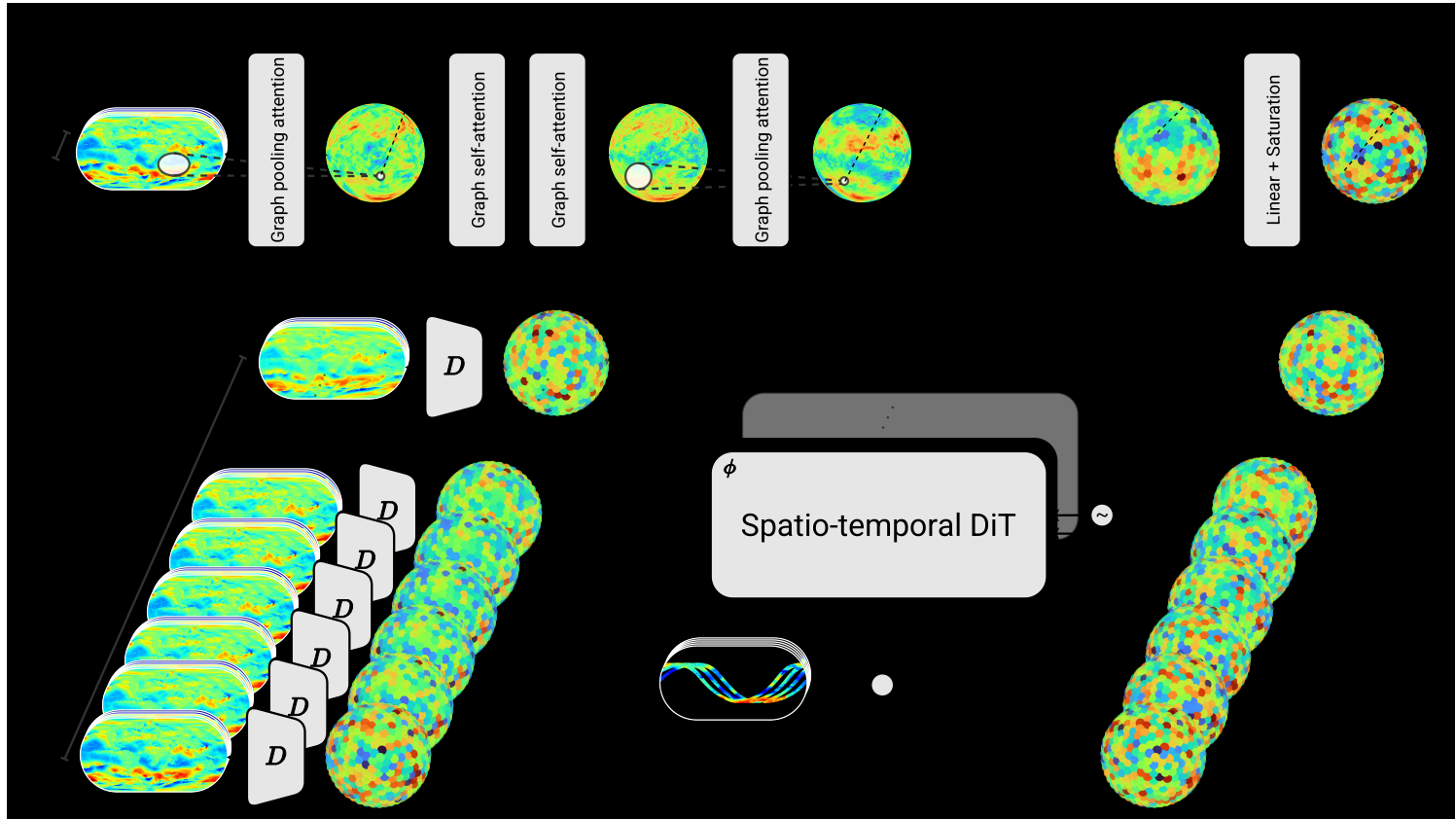
... and even more challenging, can we sample from the trajectory posterior $p(\mathbf{x}_{1:L}|\mathbf{y})$ given noisy and partial observations $\mathbf{y}$?



(2m temperature, Spring)



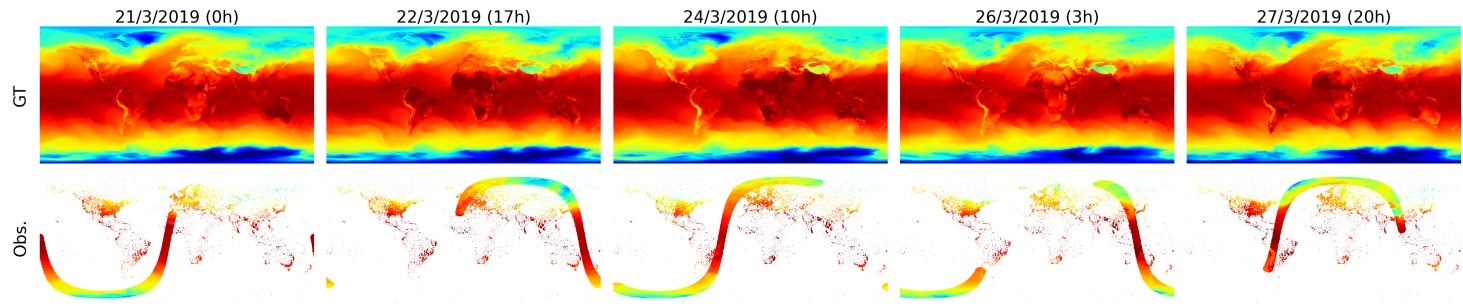
Appa: Bending weather dynamics with LDMs



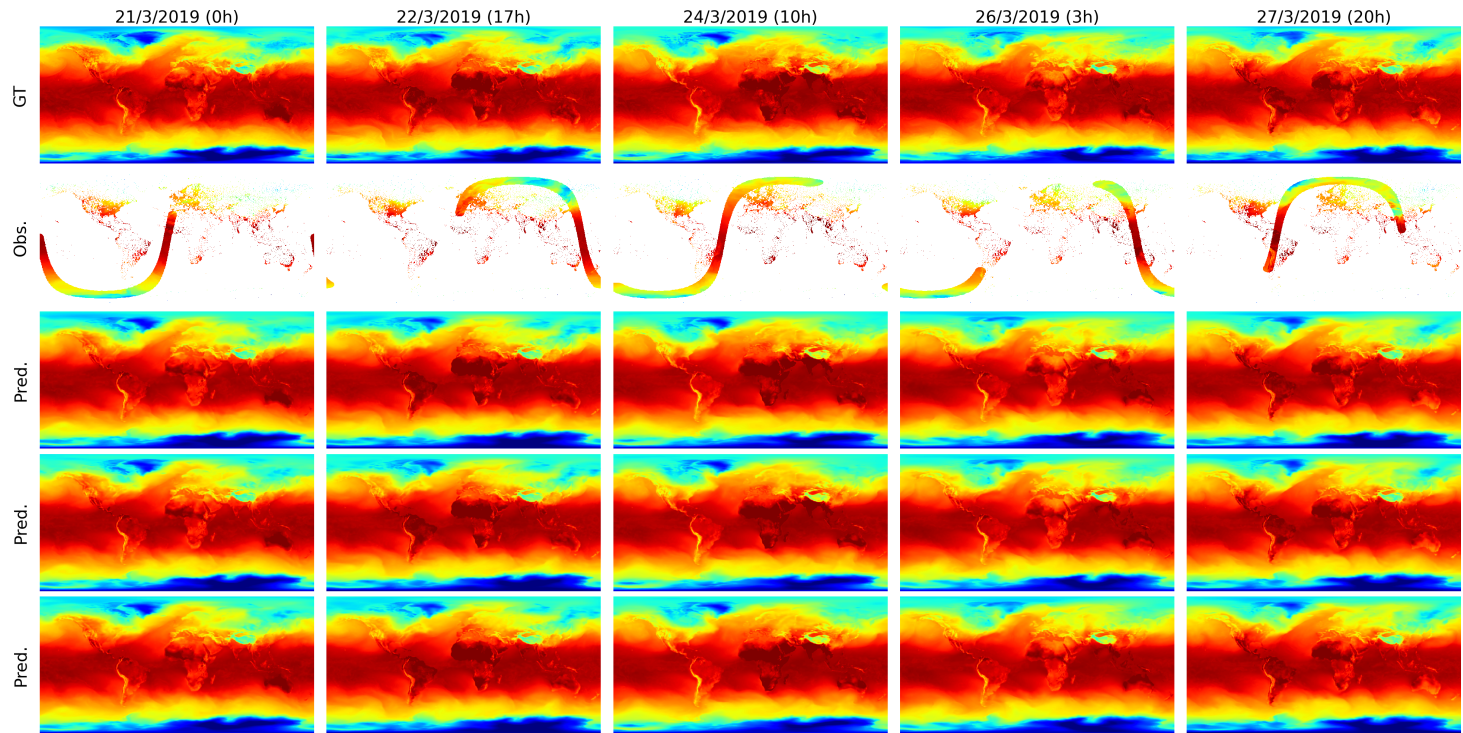


Appa is made of three components:

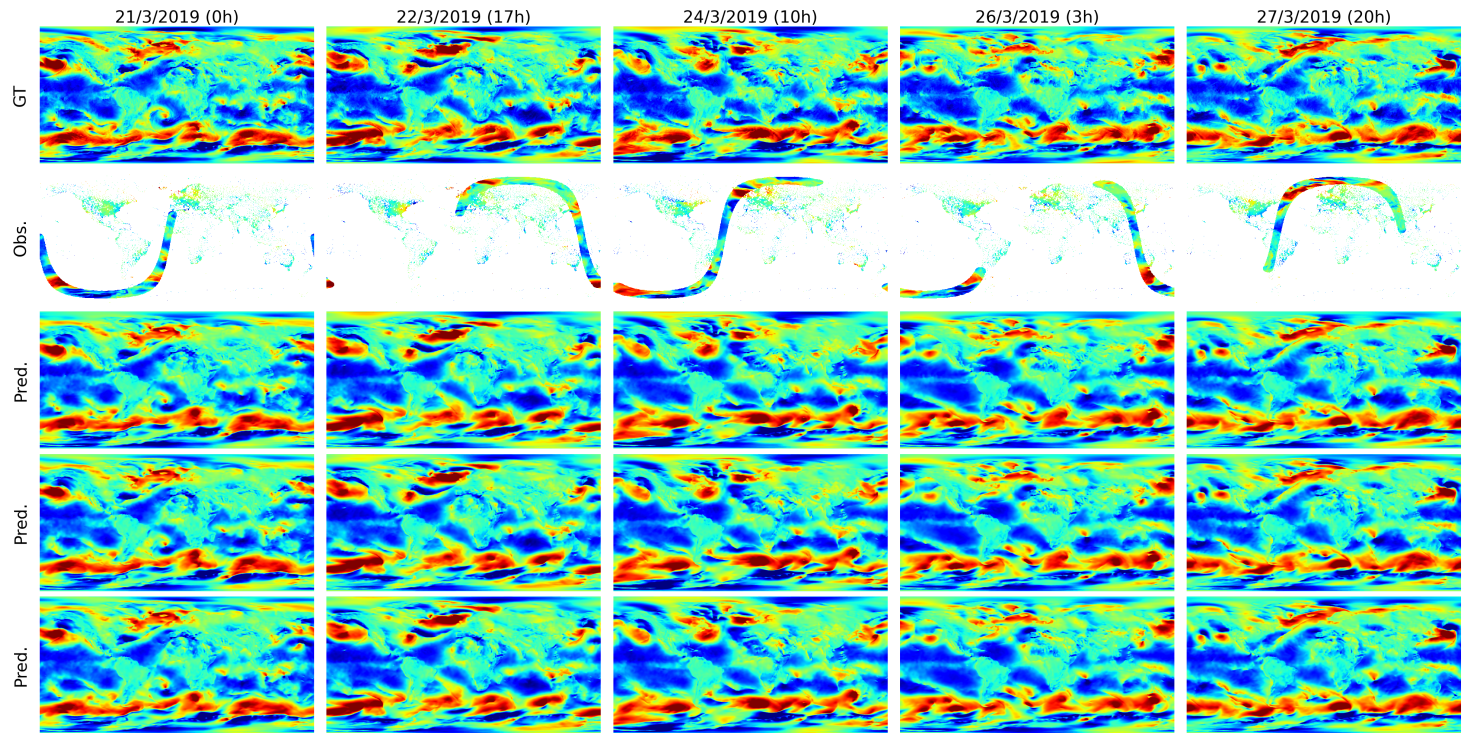
- a 500M-parameter **autoencoder** that compresses the data space x into a latent space z with a 450x compression factor;
- a 1B-parameter **diffusion model** that generates latent trajectories $z_{1:L}$;
- a **posterior sampling algorithm** adapted from MMPS (Rozet et al, 2024) that samples from the posterior distribution $p(z_{1:L} | y)$.



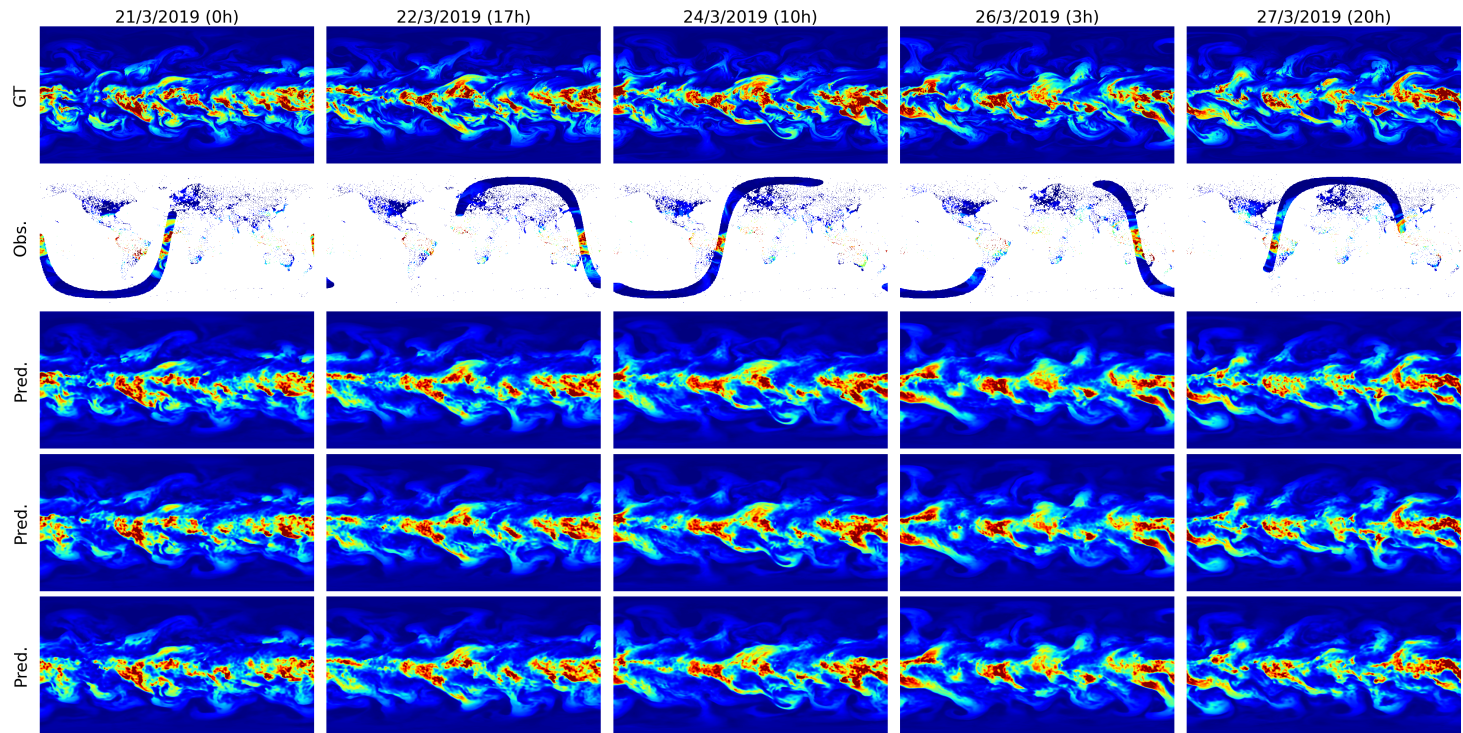
(2m temperature, Spring)



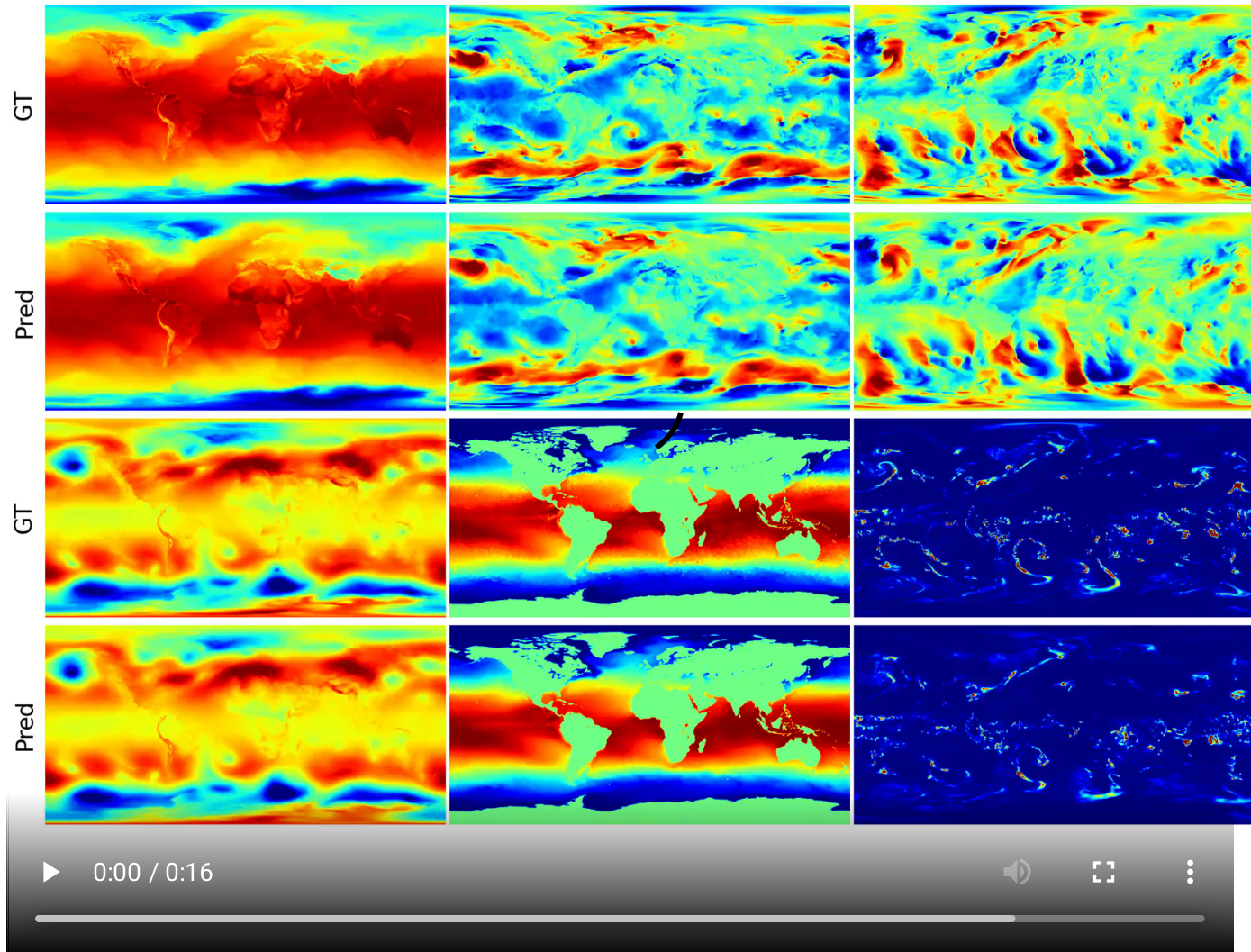
(2m temperature, Spring)



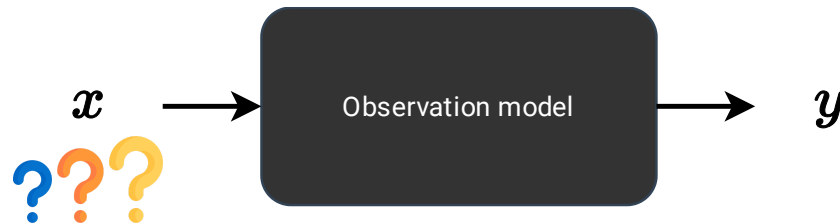
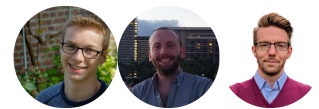
(10m u-component of wind, Spring)



(Specific humidity, Spring)



Bayesian inference without priors



Learning priors from noisy observations

Assume only observations $y \sim p(y)$ and a known observation model $p(y|x)$.

The objective of **Empirical Bayes** is find a prior model $q_\theta(x)$ such that

$$q_\theta(y) = \int p(y|x)q_\theta(x)dx$$

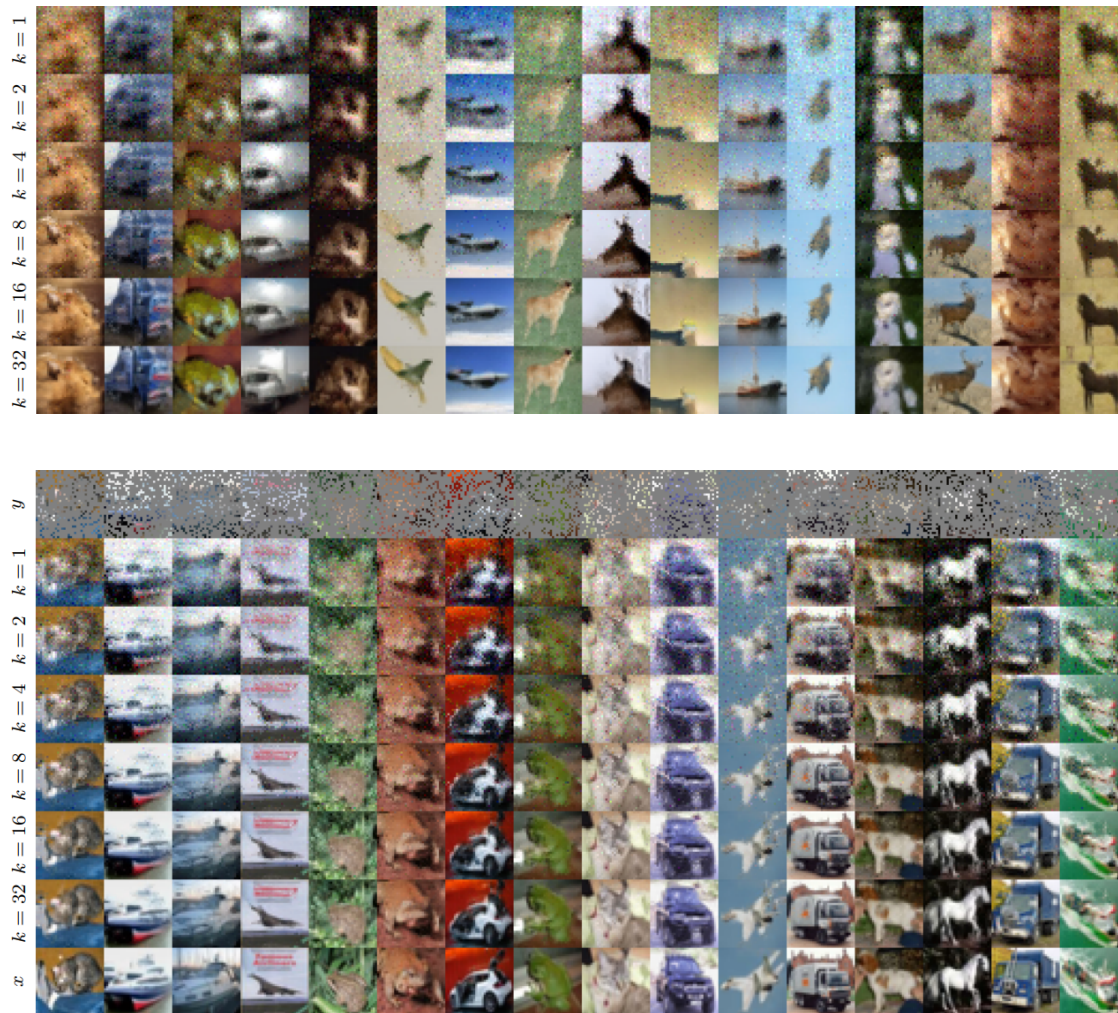
is closest to $p(y)$.

Our approach:

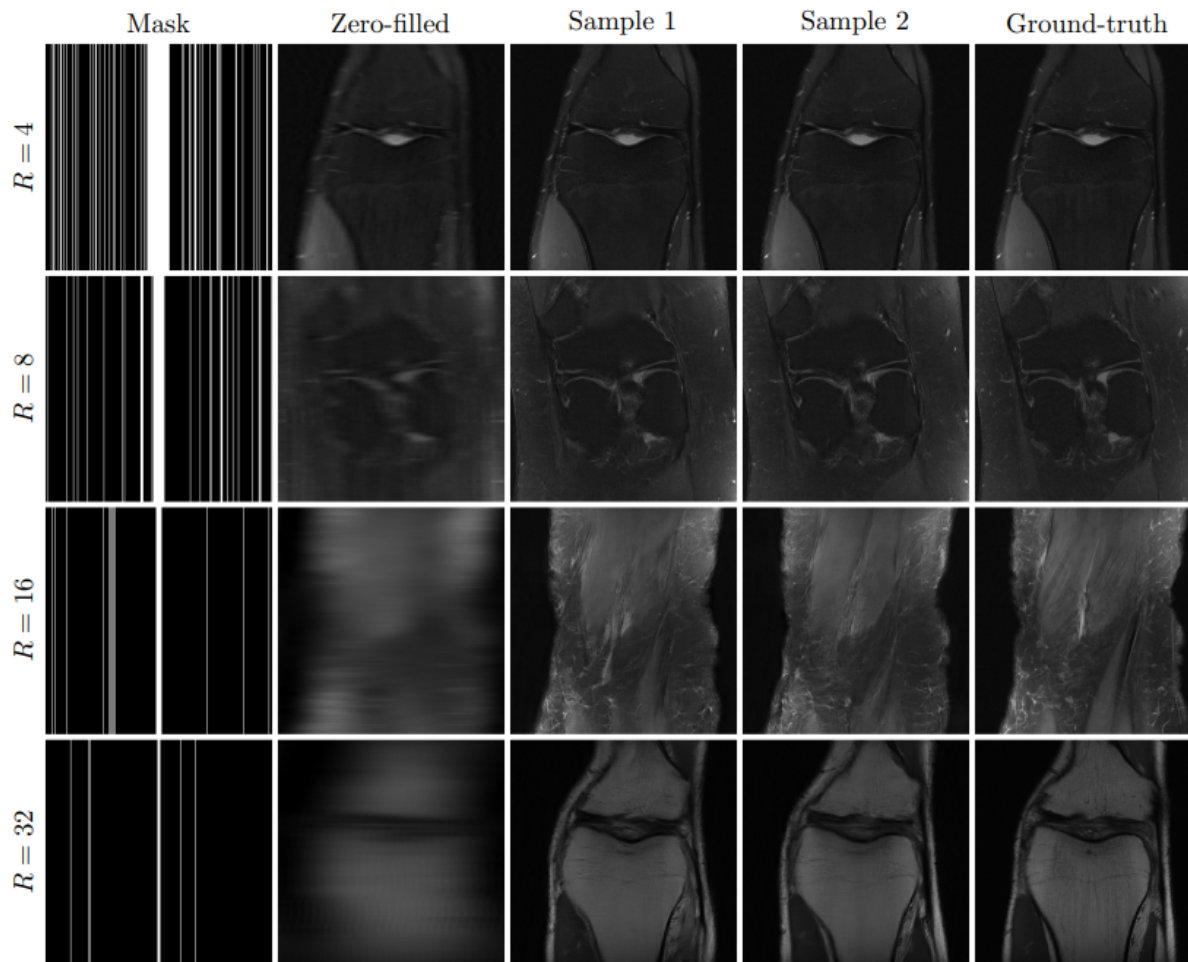
- If we parameterize the latent state x with a diffusion prior $q_\theta(x)$, then Expectation-Maximization can be used to maximize $q_\theta(y)$.
- It can be shown that the EM update

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{p(y)} \mathbb{E}_{q_{\theta_k}(x|y)} [\log q_\theta(x)] ,$$

where $q_{\theta_k}(x|y)$ is obtained by posterior sampling from $q_{\theta_k}(x)$, leads to a sequence of parameters θ_k such that $\mathbb{E}_{p(y)} [\log q_{\theta_k}(y)]$ is monotonically increasing and converges to a local optimum.



Samples from the prior $q_{\theta_k}(x)$ (top) and the posterior $q_{\theta_k}(x|y)$ (bottom) along the EM iterations when training from corrupted CIFAR-10 images.



Posterior samples for accelerated MRI using a diffusion prior trained only from observations with subsampled frequencies.



Conclusions

Score-based generative models...

- can be used for high-dimensional inverse problems;
- enable zero-shot posterior sampling, without pre-wiring the network to observations;
- do not require paired data.

Next challenges:

- Rigorous diagnostics for the quality of the approximation.

