Karlsruher Institut für Technologie

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

DASHH

# Efficient Sampling from Bayesian Network Posteriors for Optimal Uncertainties


**Sebastian Bieringer,** Gregor Kasieczka, Jan Kieseler, Maximilian Steffen, Mathias Trabs

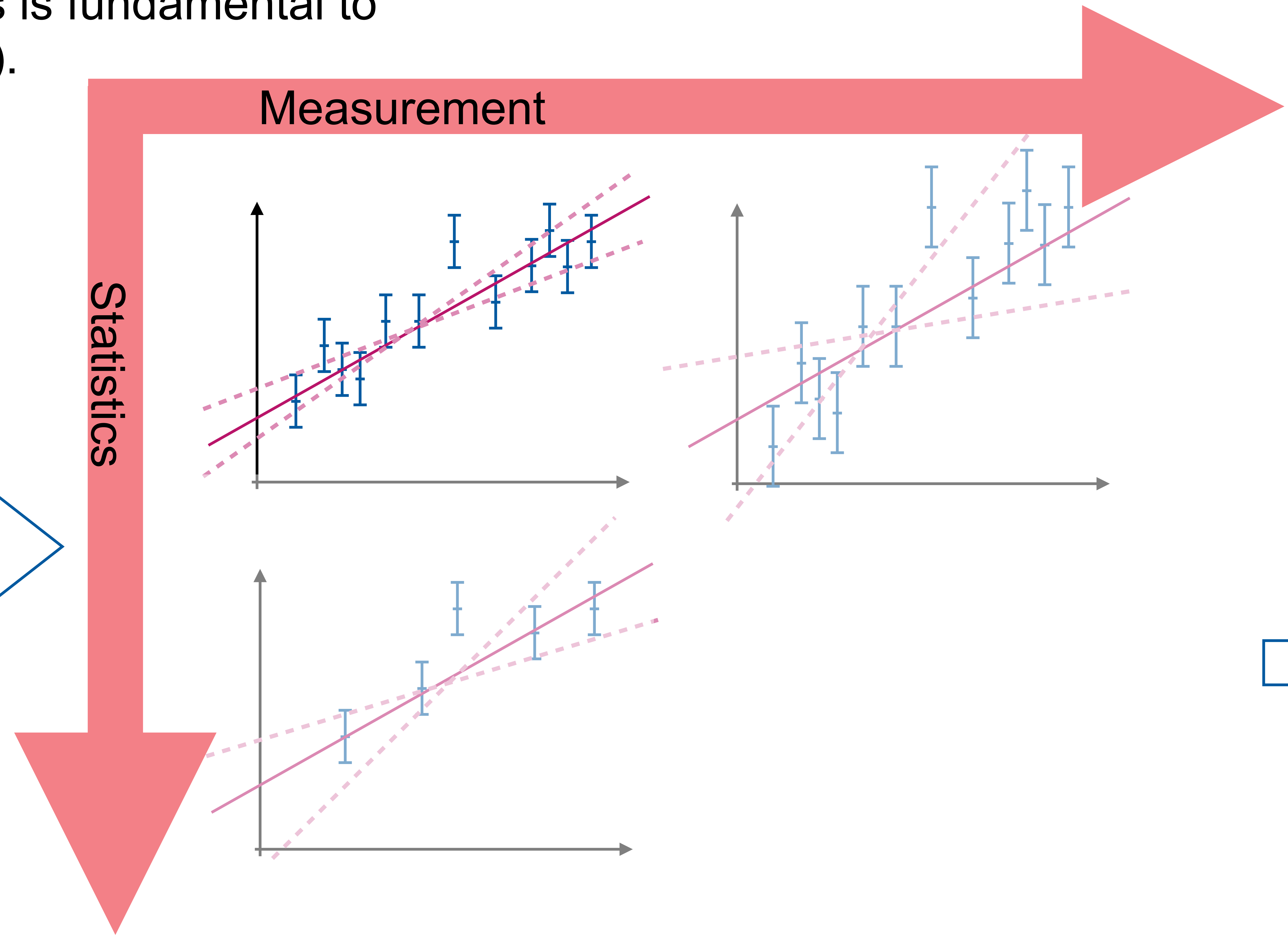
Institut für Experimentalphysik, Universität Hamburg, Germany
sebastian.guido.bieringer@uni-hamburg.de


01.12.2023 - Artificial Intelligence and the Uncertainty challenge in Fundamental Physics 2023

# Introduction

The estimation of uncertainties is fundamental to Science (and HEP specifically).
Both in **experiment** …

Measurements:
Values & Errors




Systematic Uncertainty

Aleatoric Uncertainty

„Data-distribution“

Statistical Uncertainty

Epistemic Uncertainty

Uncertainty from using a ML model

Results:
Values & Errors

… and in **evaluation**.

# „Bayesian Neural Networks“
## Mean Field Gaussian Variational Inference

**Description (1505.05424):**

- Estimate the posterior $p(\theta\,|\,\mathscr{D})$ with a simpler distribution $q(\theta)$
- Infer with gradient descent:

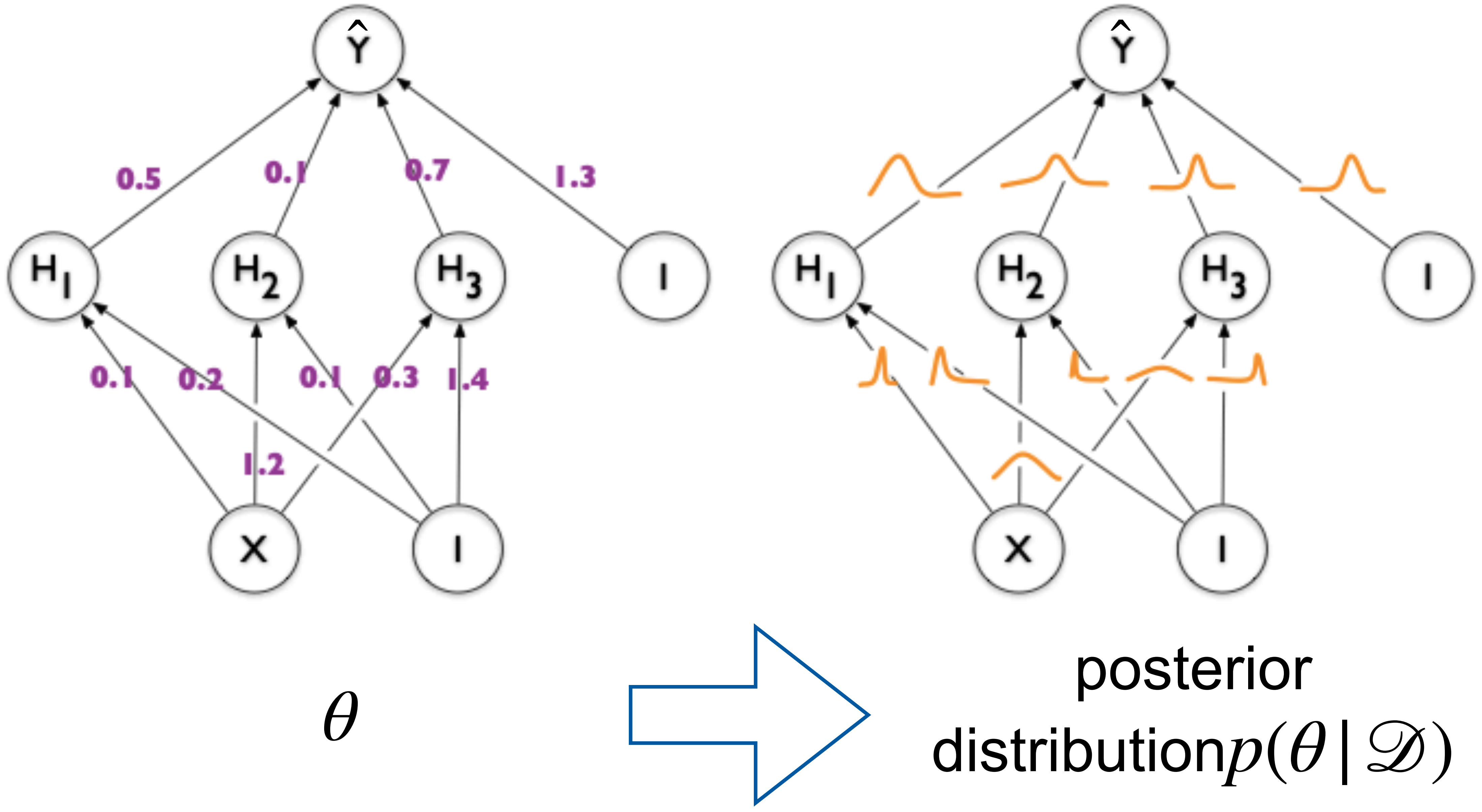
$$L_n(\hat{f}_\vartheta; \mathscr{D}_n) = \mathrm{KL}(p(\theta\,|\,\mathscr{D}_n)\,|\,q(\theta)) = -\int d\theta\, q(\theta)\, \log p(\mathscr{D}_n\,|\,\theta) + \mathrm{KL}(q(\theta)\,|\,p(\theta))$$

**Pros & Cons:**

+ Fast posterior sampling, active learning possible
- Additional loss term with high variance → influences performance
- Assumption: Posterior has uncorrelated Gaussian shape
- Doubles the number of parameters

**Adaptations:**

- Noise-Contrastive Priors (1807.09289), Flipout Layers (1803.04386)




taken from
Blundell, Charles, et al. "Weight uncertainty in neural network." *International conference on machine learning*. PMLR, 2015.


approximate posterior

sample from posterior

# Cyclic sgLD

***Description (1902.03932):***

- (Pretrain to optimal parameters $\theta^{(0)} = \theta^\star$)
- Construct a Markov-Chain with invariant distribution

$$p(\theta\,|\,\mathscr{D}) \propto \exp\left(-\lambda_{\text{LD}} L_{\text{NLL}}(\hat{f}_\theta; \mathscr{D})\right)$$

- Stochastic Gradient Langevin Dynamics (sgLD):

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k \nabla_\theta L_{\text{NLL},n}(\theta^{(k)}) + \sqrt{\frac{2\eta_k}{\lambda_{\text{LD}}}}\epsilon_k \text{ with } \epsilon_k \sim \mathcal{N}(0,1)$$

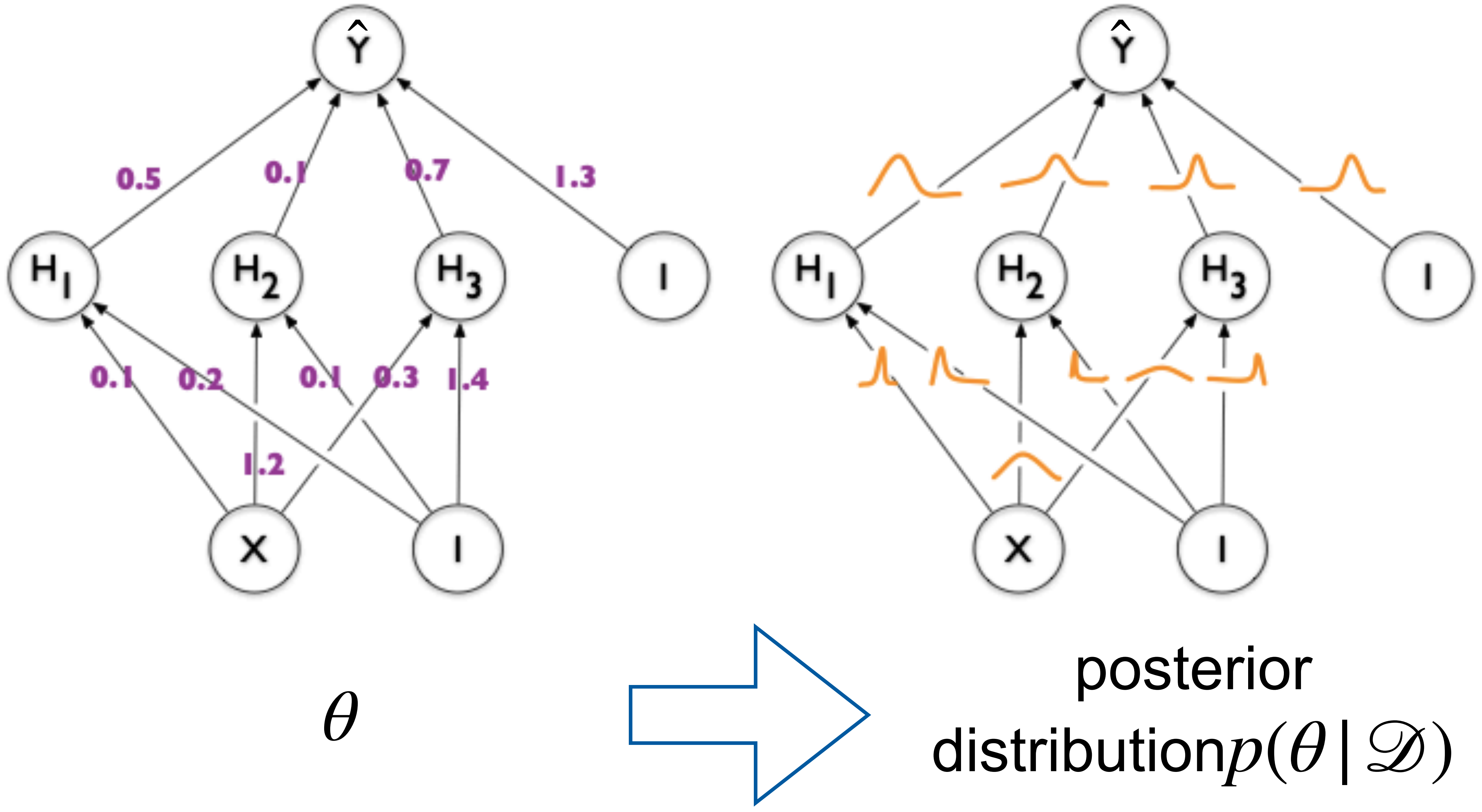
- Cyclic scheduling of stepsize $\eta_k$

***Pros & Cons:***

+ Exact sampling from the posterior
+ Good out-of-distribution detection
- Slow mixing rates
- Strongly dependent on the scheduling parameters

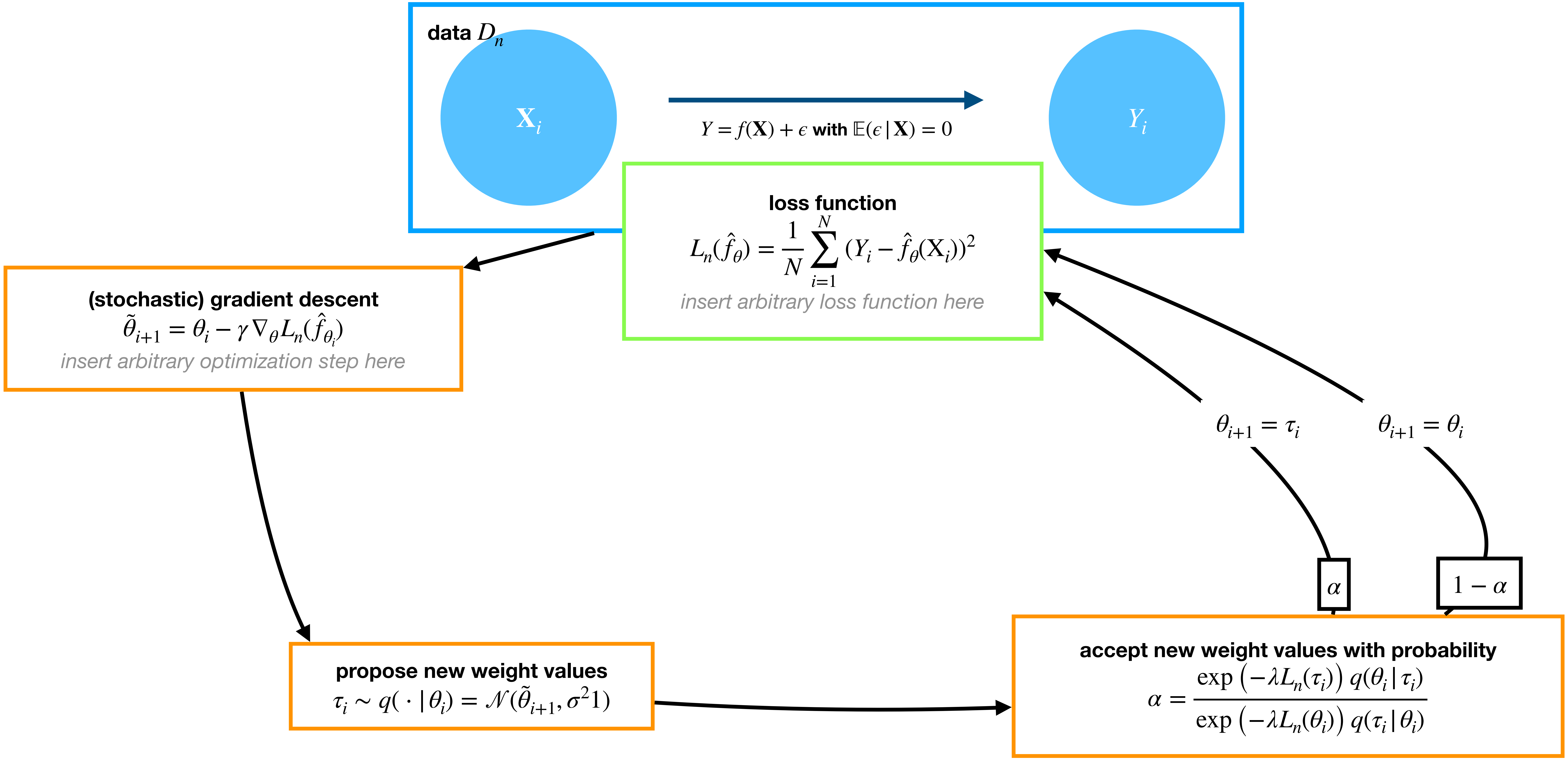
***Adaptations:***

- Hamiltonian Monte-Carlo (HMC) (1902.03932)

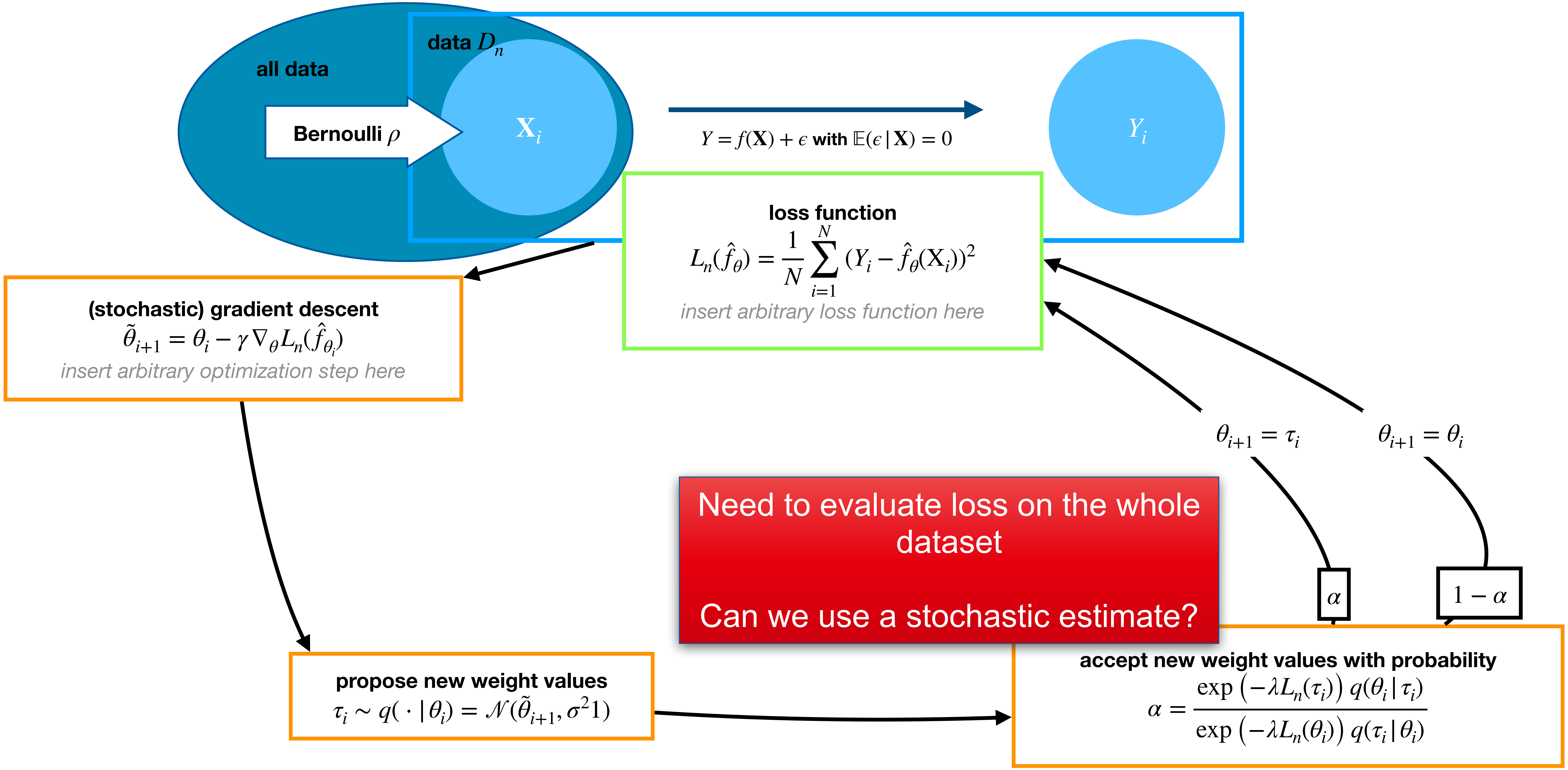



taken from
Blundell, Charles, et al. "Weight uncertainty in neural network." *International conference on machine learning*. PMLR, 2015.

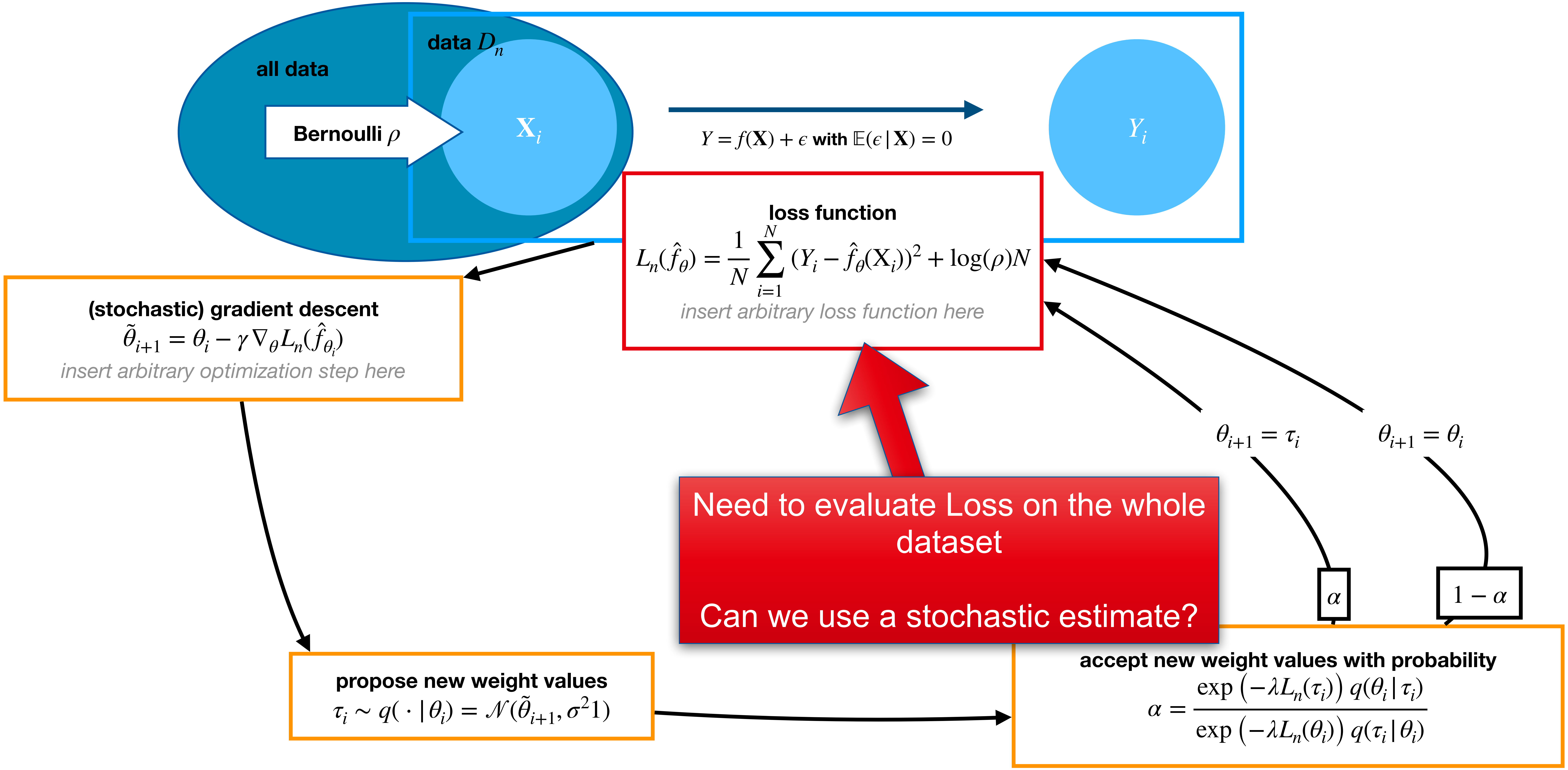
# Corrected Stochastic Metropolis Adjusted Langevin Algorithm



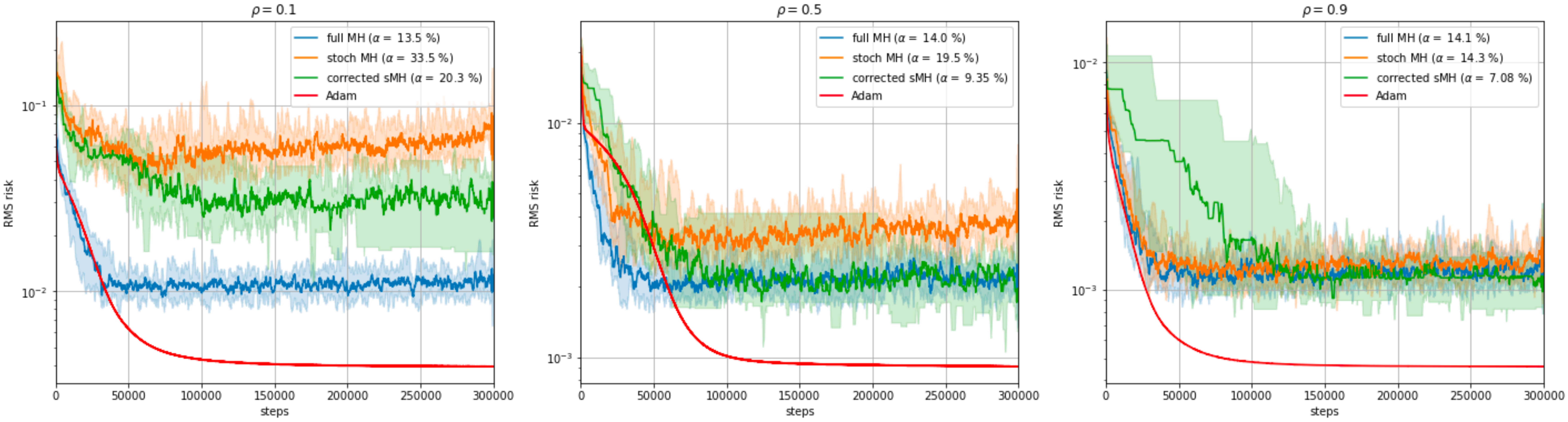
# Corrected Stochastic Metropolis Adjusted Langevin Algorithm

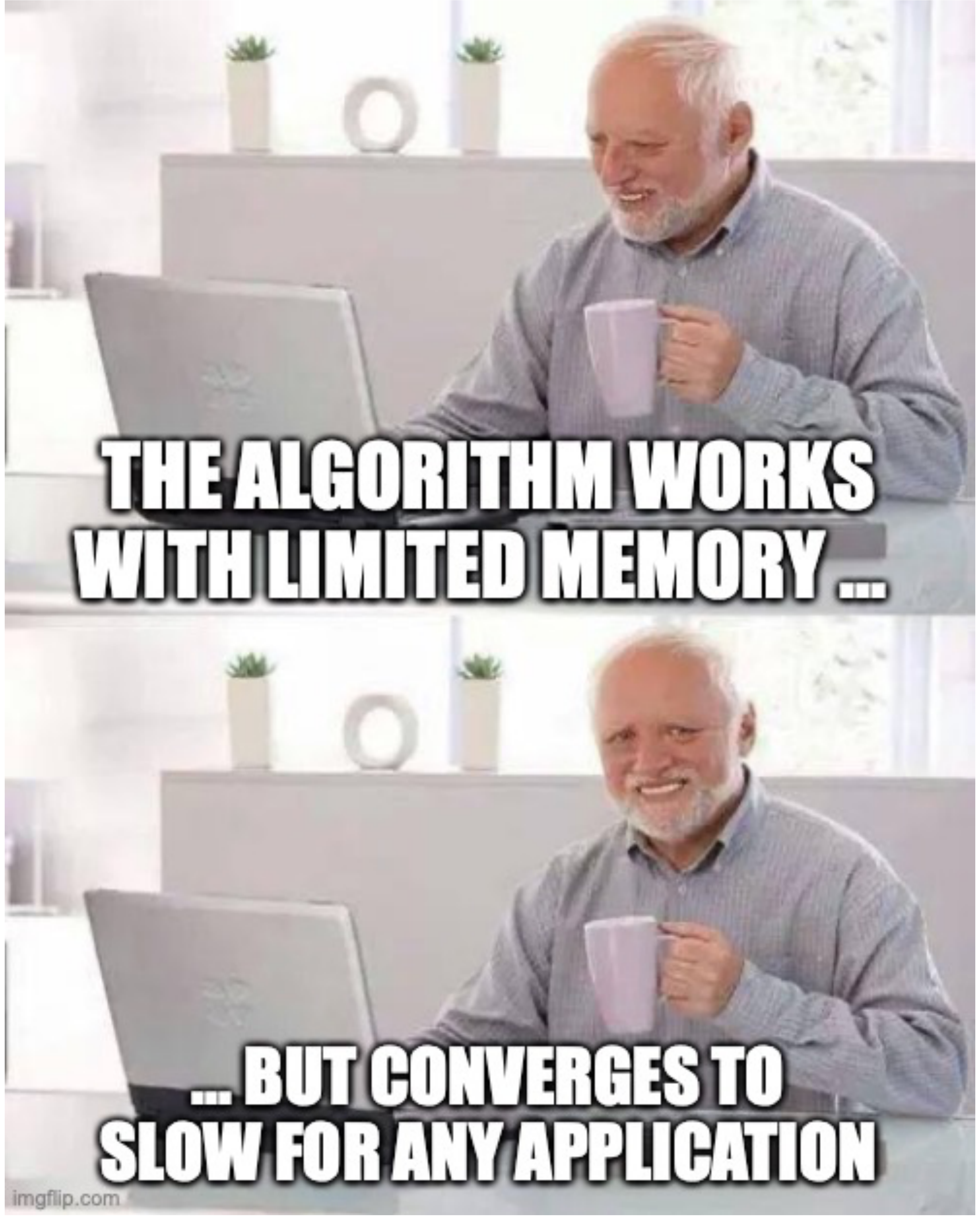


# Corrected Stochastic Metropolis Adjusted Langevin Algorithm

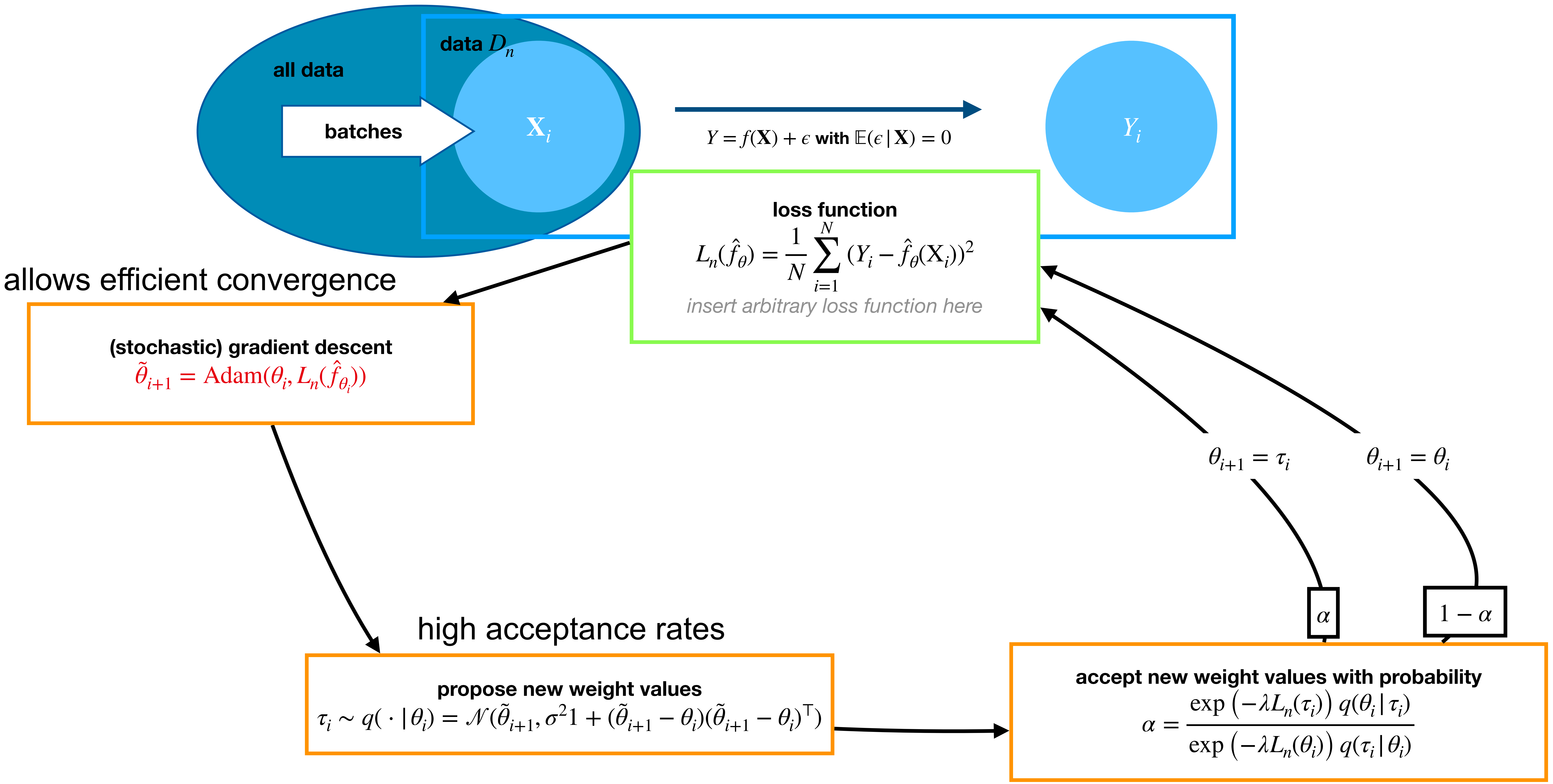


# Fits with Stochastic Gradient MALA




THE ALGORITHM WORKS WITH LIMITED MEMORY ...
... BUT CONVERGES TO SLOW FOR ANY APPLICATION
imgflip.com

# Adam-MCMC



all data
data $D_n$
batches
$\mathbf{X}_i$
$Y = f(\mathbf{X}) + \epsilon$ with $\mathbb{E}(\epsilon \,|\, \mathbf{X}) = 0$
$Y_i$
loss function
$L_n(\hat{f}_\theta) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_\theta(\mathrm{X}_i))^2$
insert arbitrary loss function here
allows efficient convergence
(stochastic) gradient descent
$\tilde{\theta}_{i+1} = \mathrm{Adam}(\theta_i, L_n(\hat{f}_{\theta_i}))$
high acceptance rates
propose new weight values
$\tau_i \sim q(\,\cdot\,|\,\theta_i) = \mathcal{N}(\tilde{\theta}_{i+1}, \sigma^2 \mathbb{1} + (\tilde{\theta}_{i+1} - \theta_i)(\tilde{\theta}_{i+1} - \theta_i)^\top)$
accept new weight values with probability
$\alpha = \frac{\exp\left(-\lambda L_n(\tau_i)\right) q(\theta_i \,|\, \tau_i)}{\exp\left(-\lambda L_n(\theta_i)\right) q(\tau_i \,|\, \theta_i)}$
$\alpha$
$1 - \alpha$
$\theta_{i+1} = \tau_i$
$\theta_{i+1} = \theta_i$

# Adam-MCMC

## - Invariant distribution -

**Theorem 1.** *For $\rho_l^2 = (1-\beta_l^2)s^2, l = 1,2$, and arbitrary proposal distributions $q_{1,k}(\tau^{(k)}|\vartheta^{(k)}, m^{(k)})$ the Markov chain $(\vartheta^{(k)}, m^{(k)})_{k\geq 1}$ admits the invariant distribution $f(\vartheta, m) = p_\lambda(\vartheta|\mathcal{D}_n)\varphi_{g(\vartheta),s^2}(m_1)\varphi_{g(\vartheta)^2,s^2}(m_2)$. In particular, the marginal distribution of $f(\vartheta, m)$ in $\vartheta$ is the Gibbs posterior distribution $p_\lambda(\cdot|\mathcal{D}_n)$.*

## - Convergence -

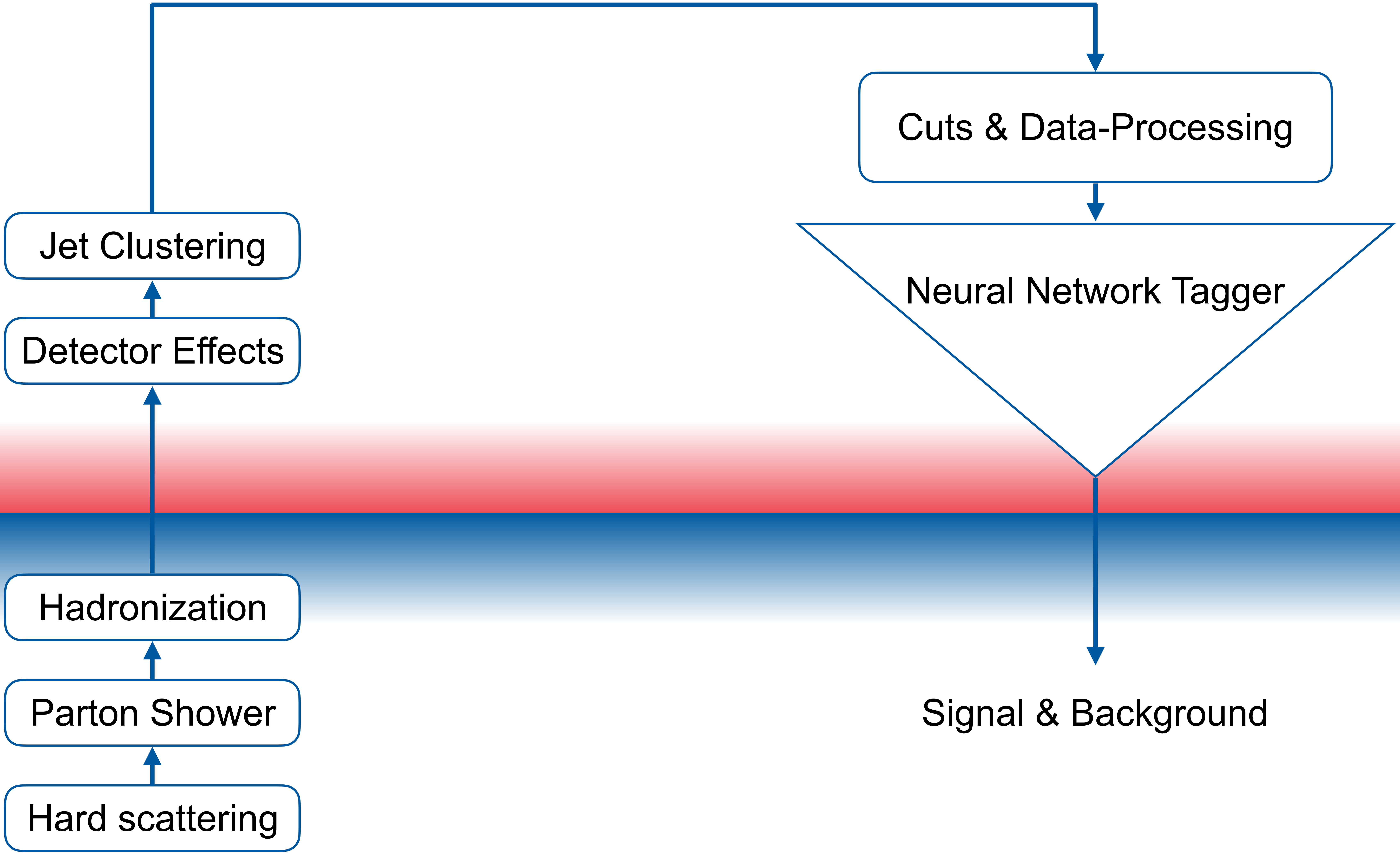
**Theorem 2.** *Let $\rho_l^2 = (1-\beta_l^2)s^2, l = 1,2$, take $q_{1,k}$ from (4) and $m^{(0)} \sim \mathcal{N}((g(\vartheta^{(0)}), g(\vartheta^{(0)})^2), s^2 I_{2\#\vartheta})$ where $\vartheta^{(0)}$ is an arbitrary random initialization of the chain. Then the distribution of $\vartheta^{(k)}$ converges in total variation distance to the Gibbs posterior $p_\lambda(\cdot|\mathcal{D}_n)$:*

$$\mathrm{TV}(\mathbb{P}^{\vartheta^{(k)}}, \Pi_\lambda(\cdot|\mathcal{D}_n)) \lesssim (1-a)^k \xrightarrow{k\to\infty} 0 \qquad \textit{for some } a \in (0,1).$$

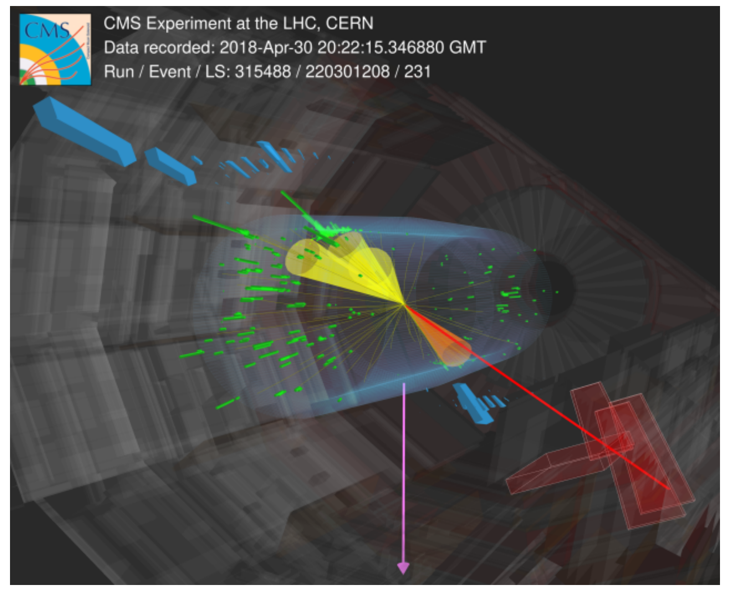
# Physics Use-case

# Classification Surrogates

DASHH

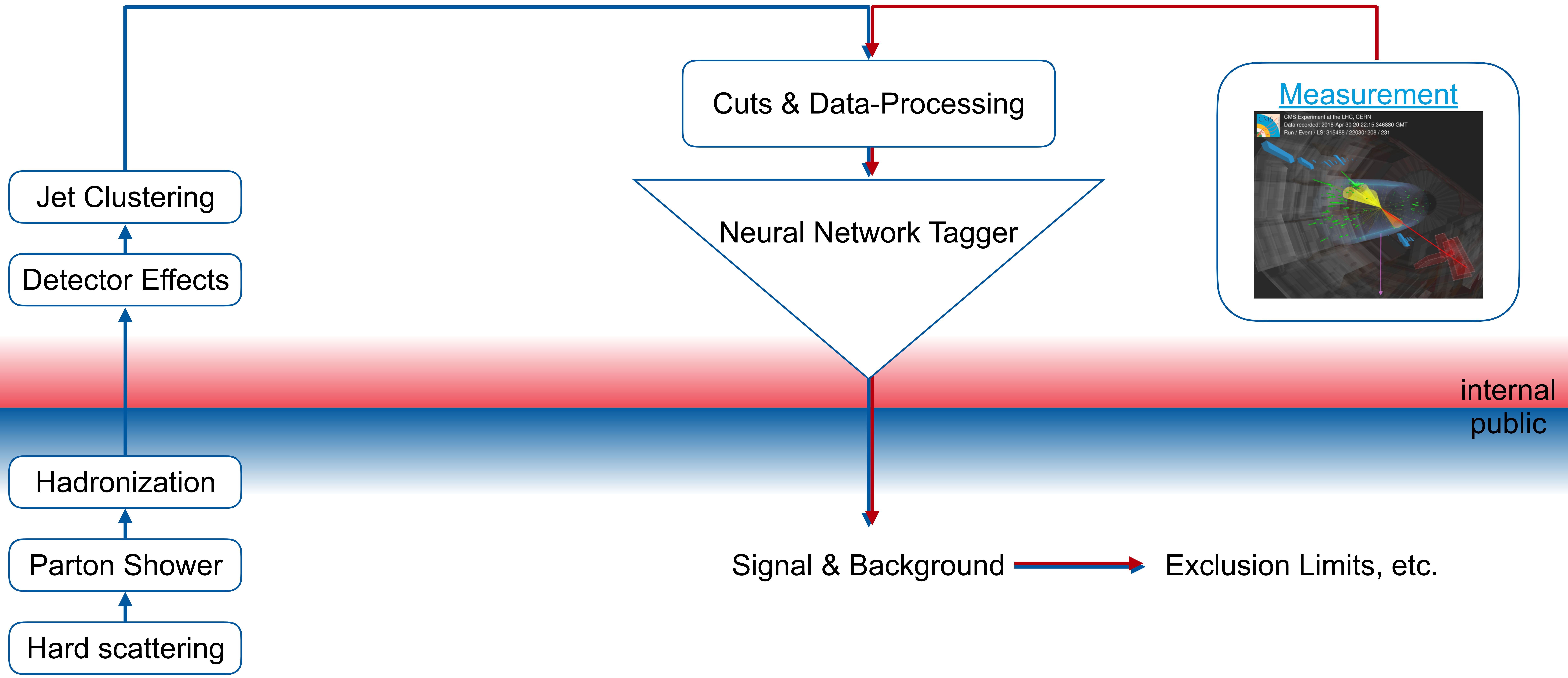



Measurement




internal
public

# Classification Surrogates



# Classification Surrogates

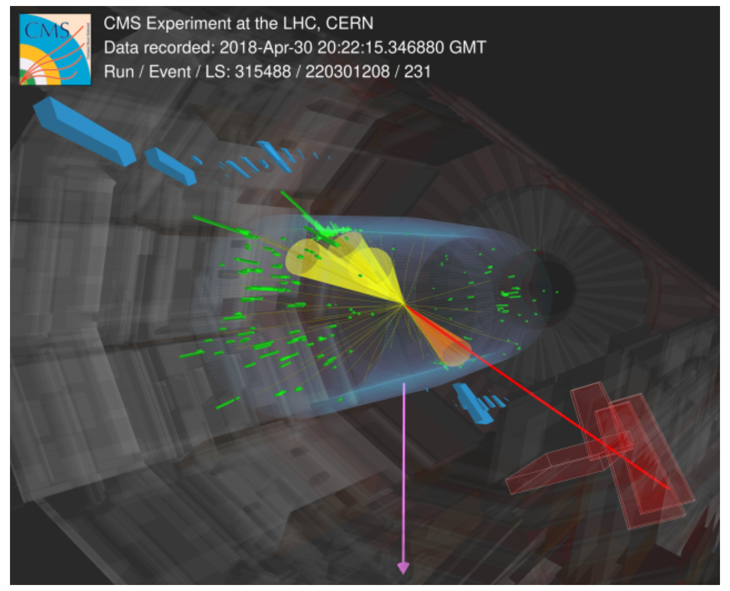
Is this evaluation also sensitive to
$X \rightarrow Y + Z$?

D

Cuts & Data-Processing

Neural Network Tagger

Measurement




internal

public

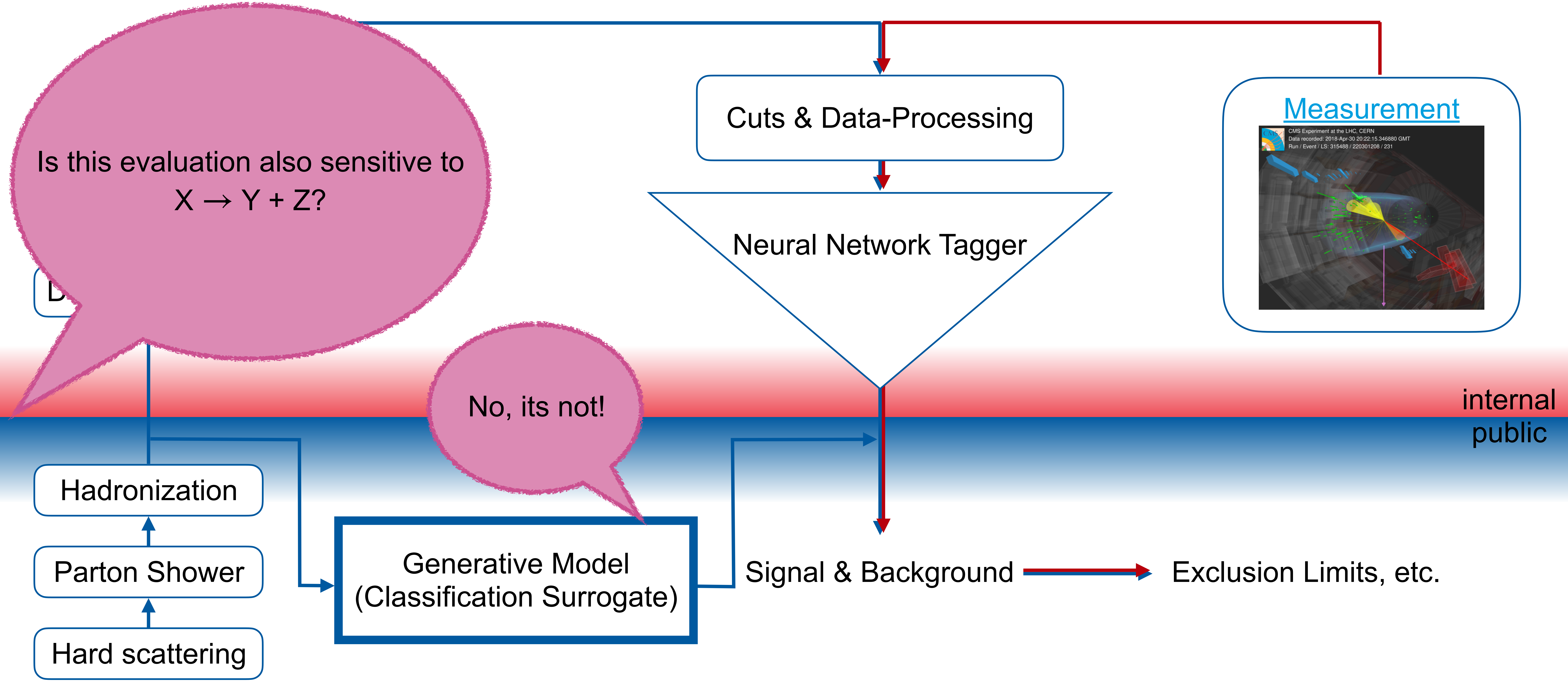
Hadronization

Parton Shower

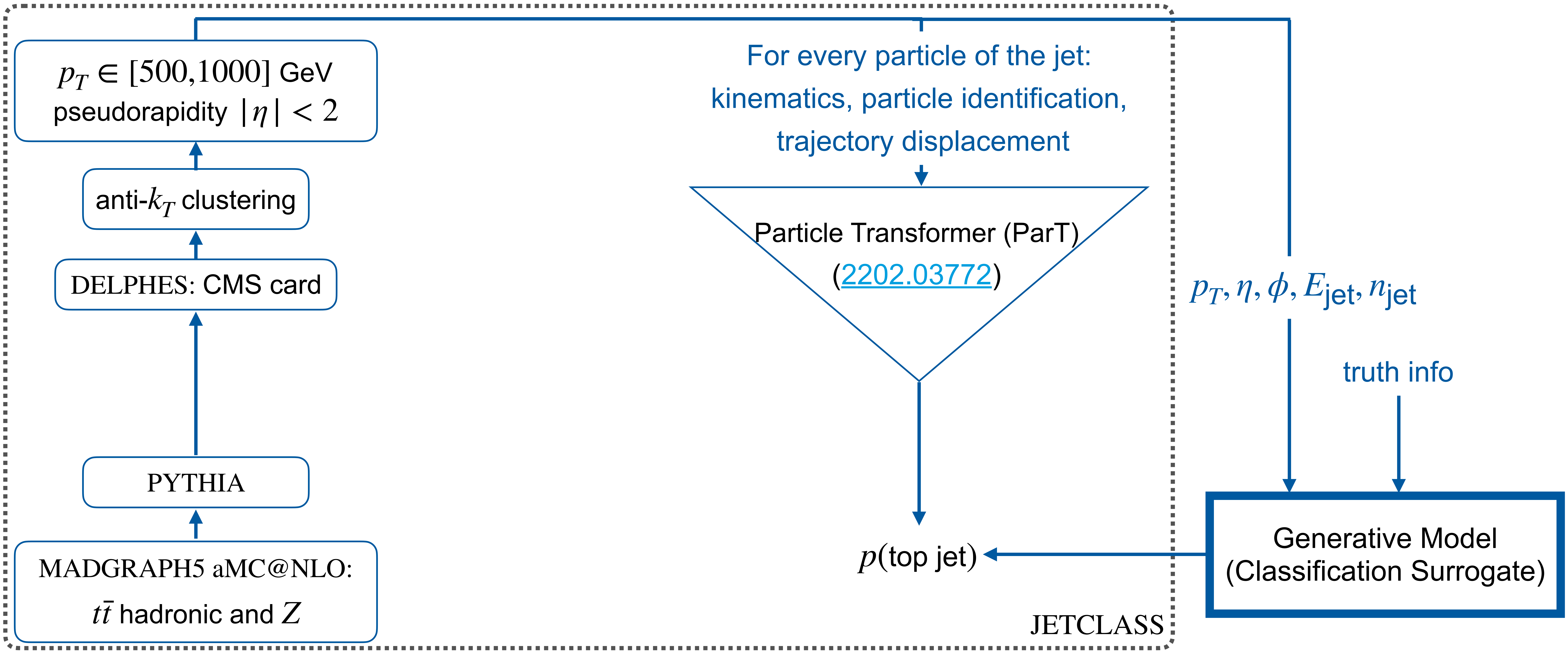
Hard scattering

Signal & Background → Exclusion Limits, etc.

# Classification Surrogates

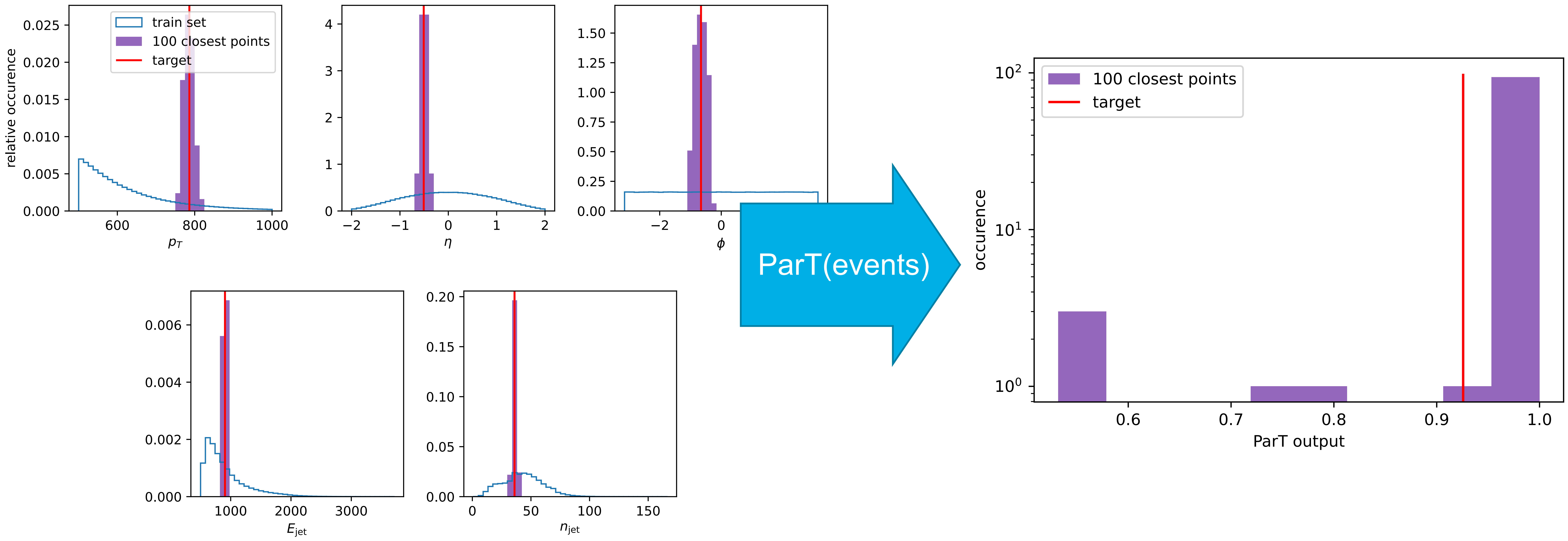


# The Toy Setup



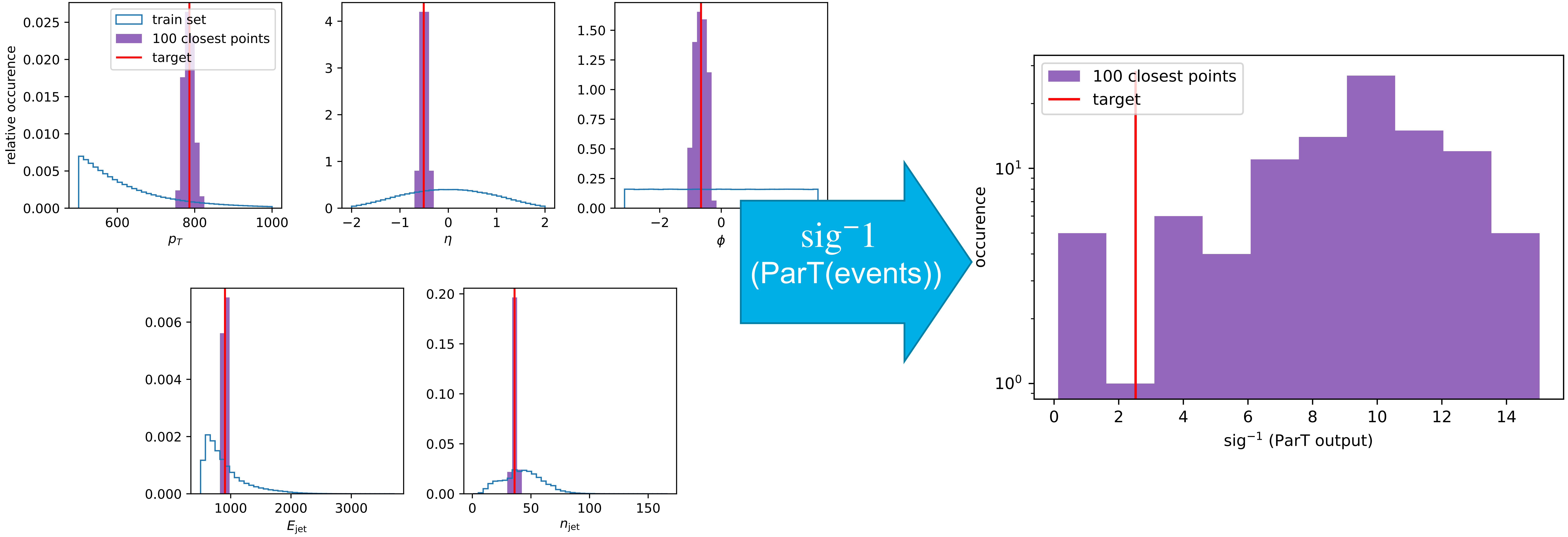
# Detector Smearing Distribution

- pick a jet event
- select the 100 events with $p_T, \eta, \phi, E_{\text{jet}}, n_{\text{jet}}$ closest

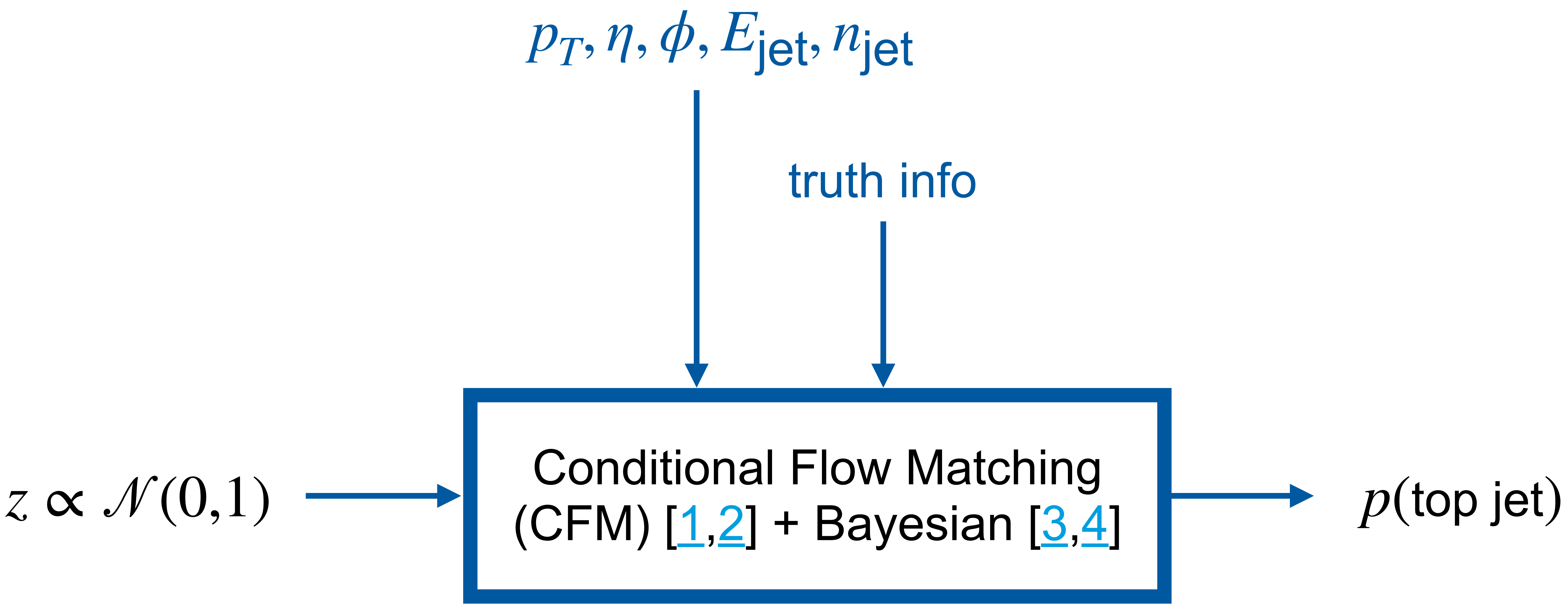


# Detector Smearing Distribution

- pick a jet event
- select the 100 events with $p_T, \eta, \phi, E_{\text{jet}}, n_{\text{jet}}$ closest



# The Generative Model



# Variational Inference Bayesian CFM

DASHH

**Continuous Normalizing Flow:**

- Flow $\phi : [0,1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined via

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = v_t(\phi_t(x)) = \tilde{v}_t(x, \theta)$$

- solve the ODE to train and sample
- linear trajectory
- transforms probability distributions

$$p_t(x) = p_0\left(\phi_t^{-1}(x)\right) \det\left[\frac{\partial \phi_t^{-1}}{\partial x}(x)\right]$$

**Conditional Flow Matching:**

- loss that does not ODE solving

$$\mathscr{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \left\| v_t(x) - \tilde{v}_t(x,\theta)) \right\|^2$$

- by choice of $p_t$ and $v_t$

$$\mathscr{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,p_t(x),\epsilon}\left[\tilde{v}_t\left((1-t)x_0 + t\epsilon, \theta\right) - \left(\epsilon - x_0\right)\right]^2$$

- not a $\log$-Likelihood loss

**Variational Inference Bayesian Conditional Flow Matching:**

- Bayesian loss $\mathscr{L}_{\mathrm{BNN}} = \mathrm{KL}\left[q(\theta), p\left(\theta \mid x\right)\right] = -\int \mathrm{d}\theta\, q(\theta) \log p\left(x \mid \theta\right) + \mathrm{KL}[q(\theta), p(\theta)] +$ const.
- connect both $\mathscr{L}_{\mathrm{B-CFM}} = \left\langle \mathscr{L}_{\mathrm{CFM}} \right\rangle_{\theta \sim q(\theta)} + c\mathrm{KL}[q(\theta), p(\theta)]$, with $q(\theta)$ uncorrelated Gaussian shape

# Adam-MCMC Bayesian CFM

**Continuous Normalizing Flow:**

- Flow $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ defined via

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = v_t(\phi_t(x)) = \tilde{v}_t(x, \theta)$$

- solve the ODE to train and sample
- linear trajectory
- transforms probability distributions

$$p_t(x) = p_0\left(\phi_t^{-1}(x)\right) \det\left[\frac{\partial \phi_t^{-1}}{\partial x}(x)\right]$$

**Conditional Flow Matching:**

- loss that does not ODE solving

$$\mathscr{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \left\| v_t(x) - \tilde{v}_t(x, \theta)) \right\|^2$$

- by choice of $p_t$ and $v_t$

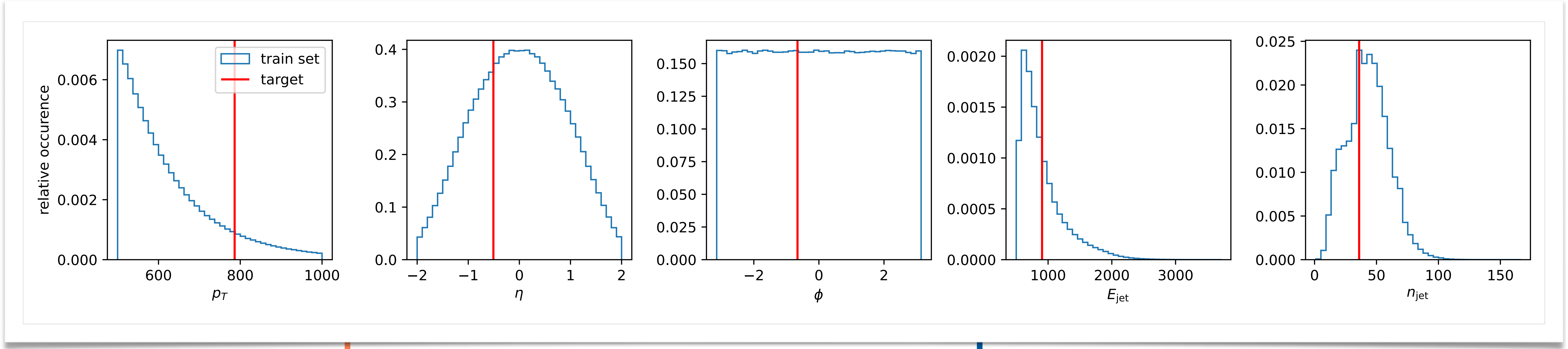
$$\mathscr{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,p_t(x),\epsilon}\left[\tilde{v}_t\left((1-t)x_0 + t\epsilon, \theta\right) - \left(\epsilon - x_0\right)\right]^2$$

- not a $\log$-Likelihood loss

**Adam-MCMC Bayesian Conditional Flow Matching:**

- train the network with CFM
- start Markov Chain from this point (independent of starting point):
  - 1D problem: Solve ODE to get $\log$-Likelihood of batch for update steps and acceptance rates

# Learned Detector Smearing Distribution

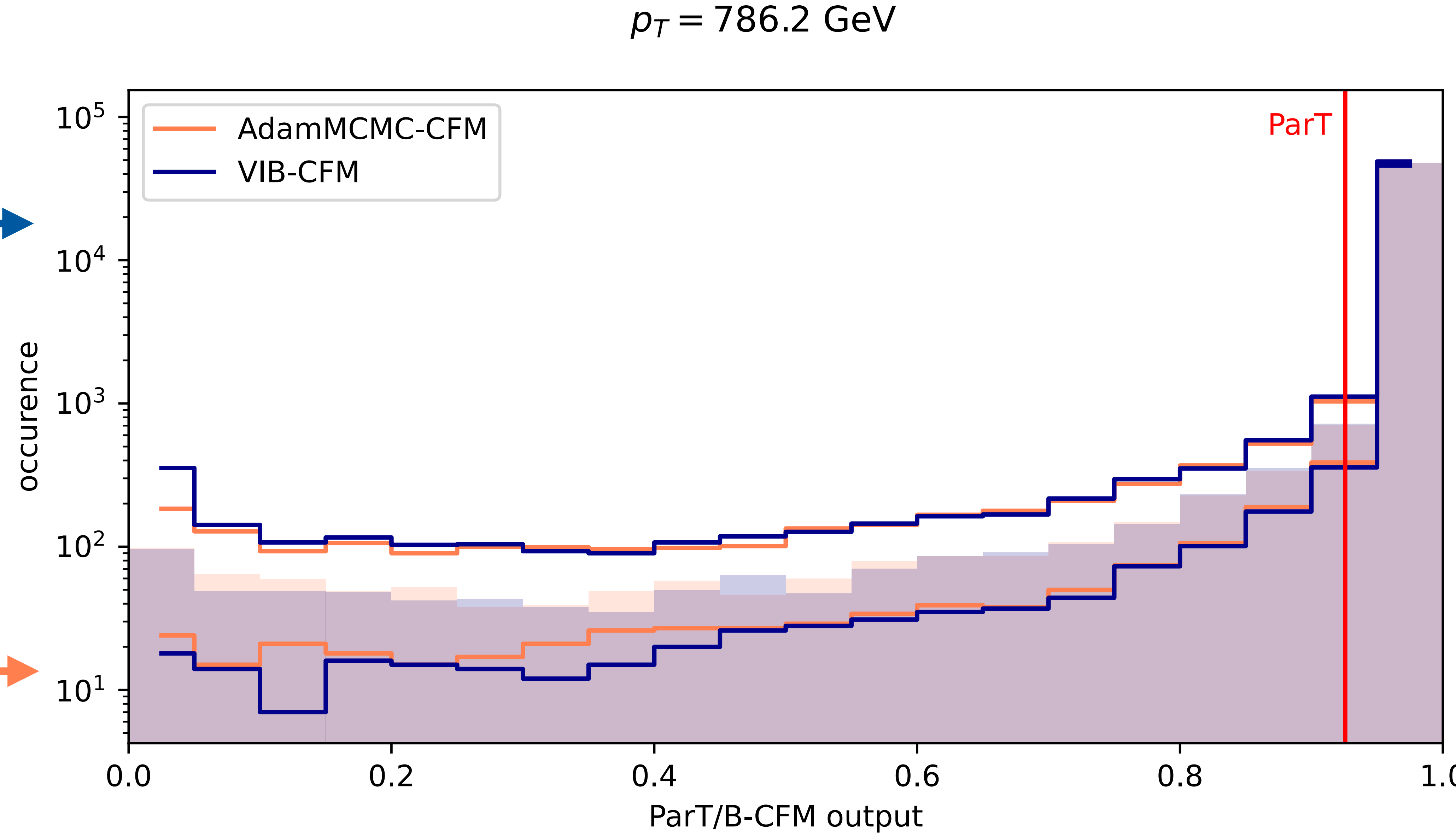



1

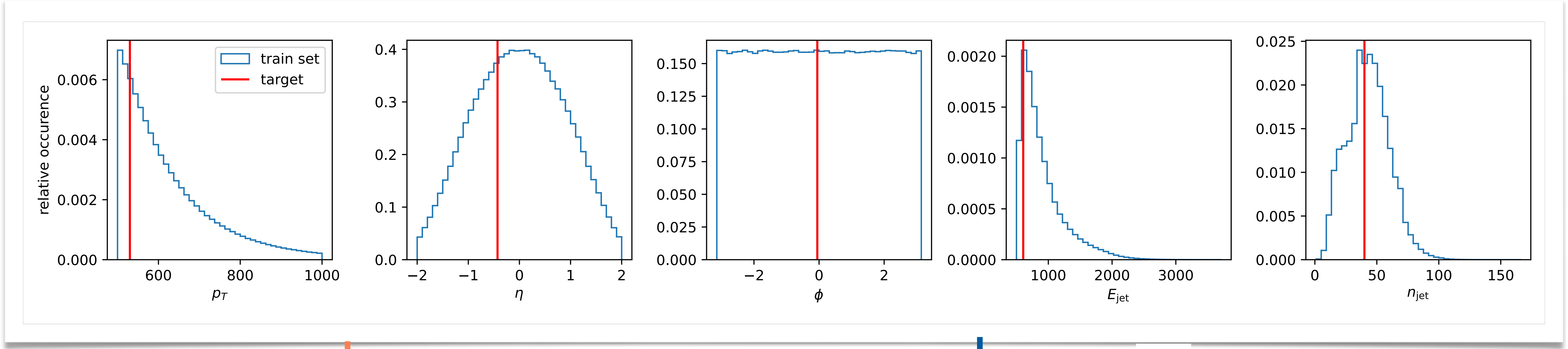
$z \propto \mathcal{N}(0,1)$ → Conditional Flow Matching (CFM) + Variational Inference Bayes

1

$z \propto \mathcal{N}(0,1)$ → Conditional Flow Matching (CFM) + AdamMCMC



# Learned Detector Smearing Distribution




1

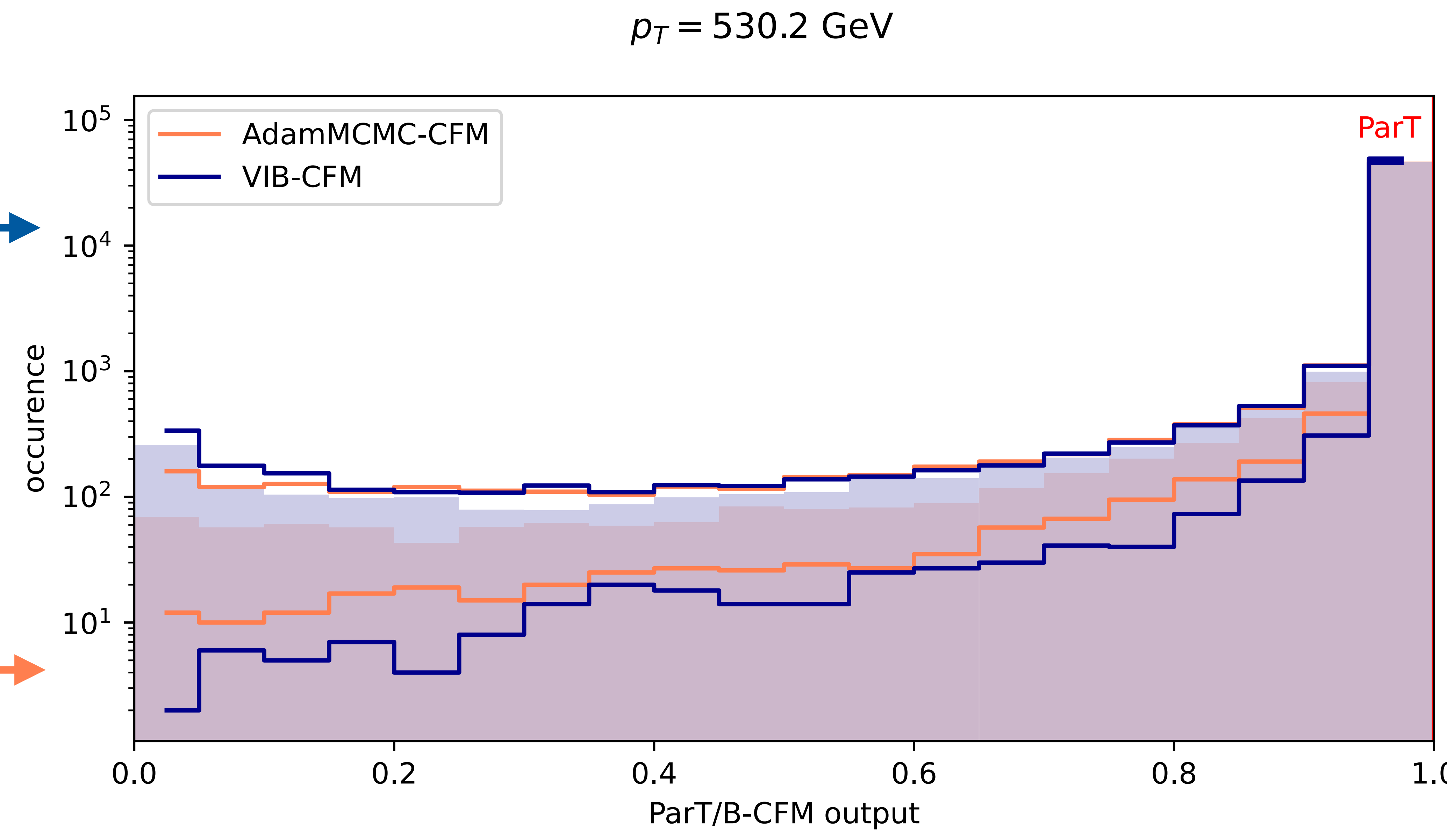
$z \propto \mathcal{N}(0,1)$

Conditional Flow Matching (CFM)
+ Variational Inference Bayes

1

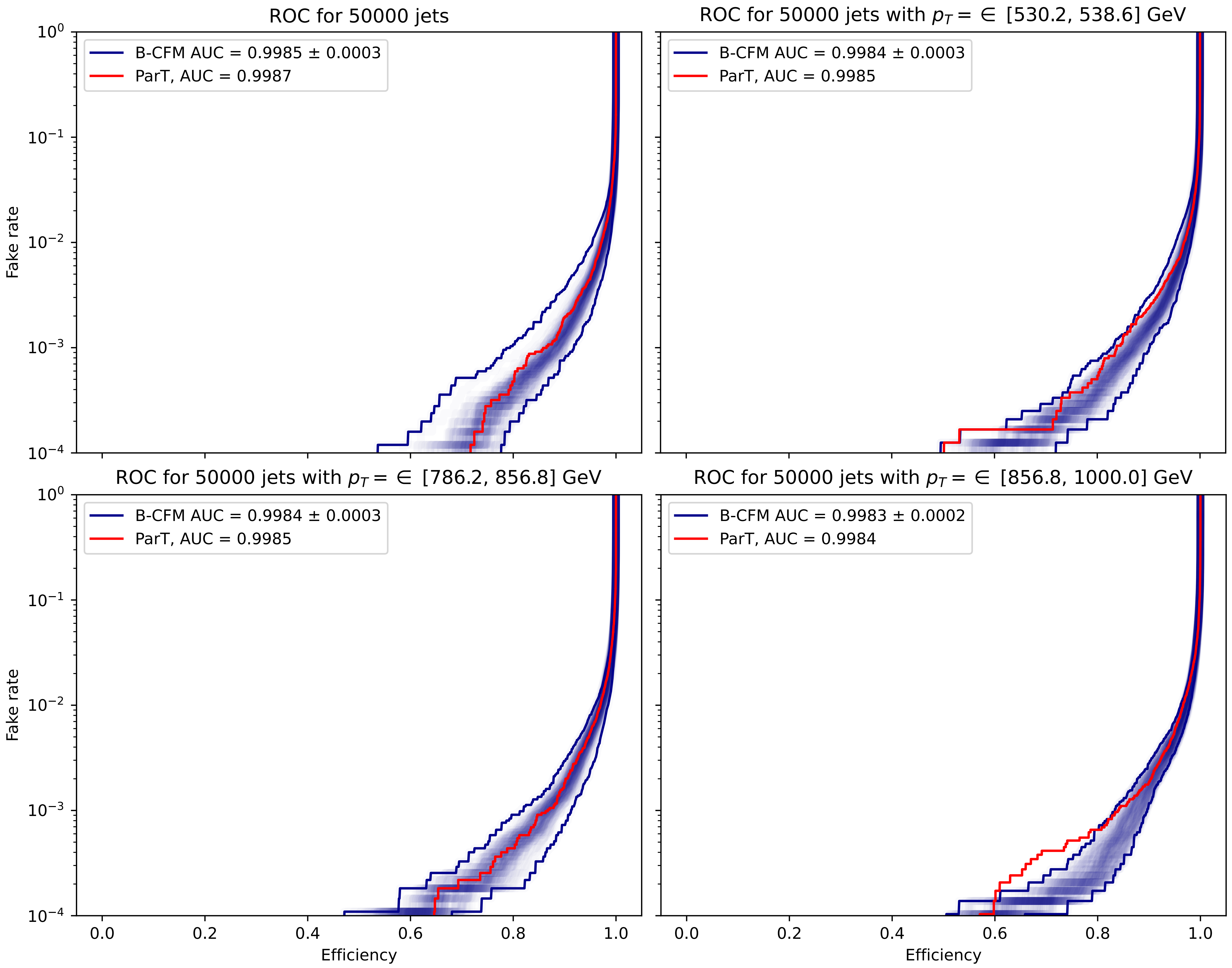
$z \propto \mathcal{N}(0,1)$

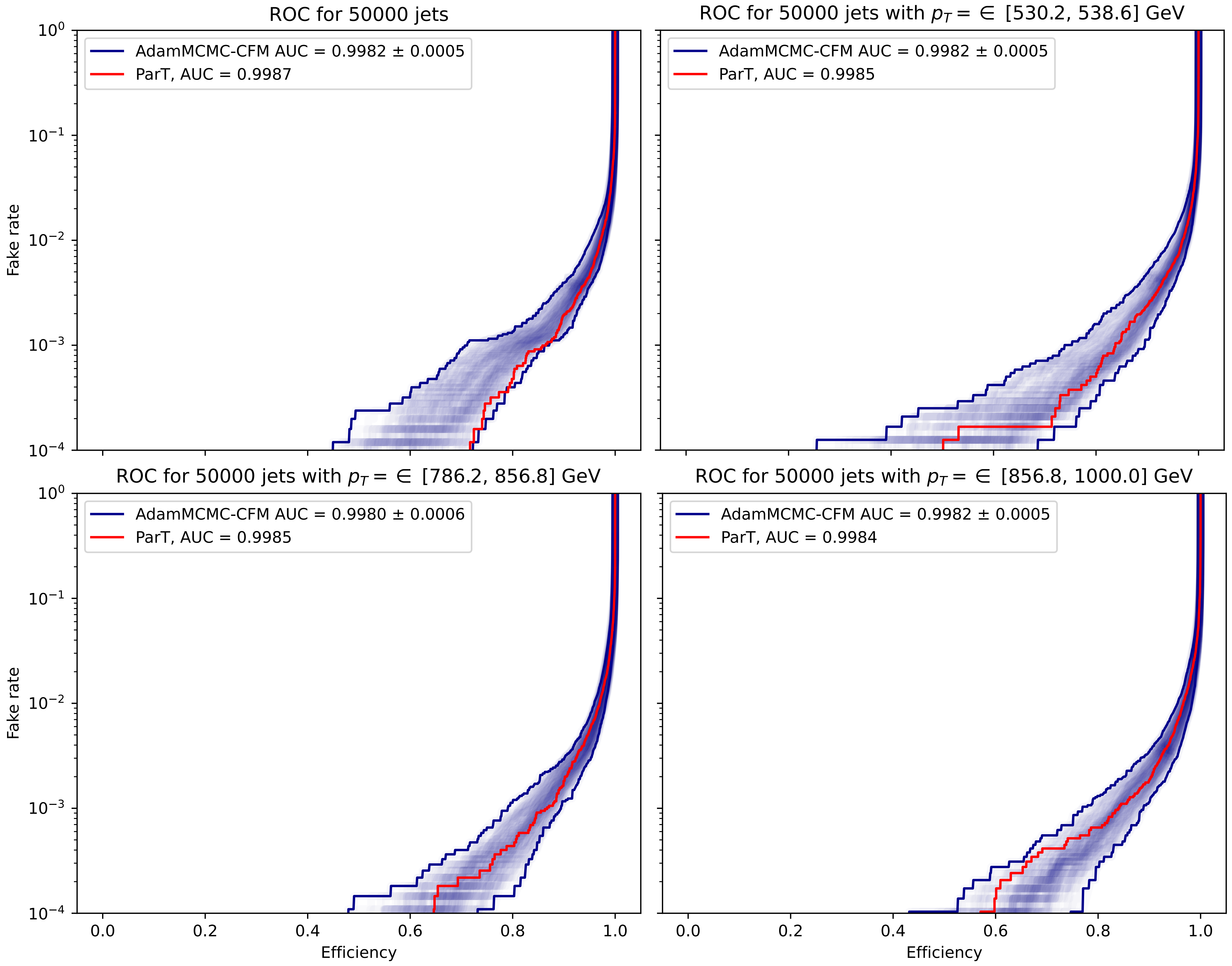
Conditional Flow Matching (CFM)
+ AdamMCMC



# Learned Detector Smearing Distribution

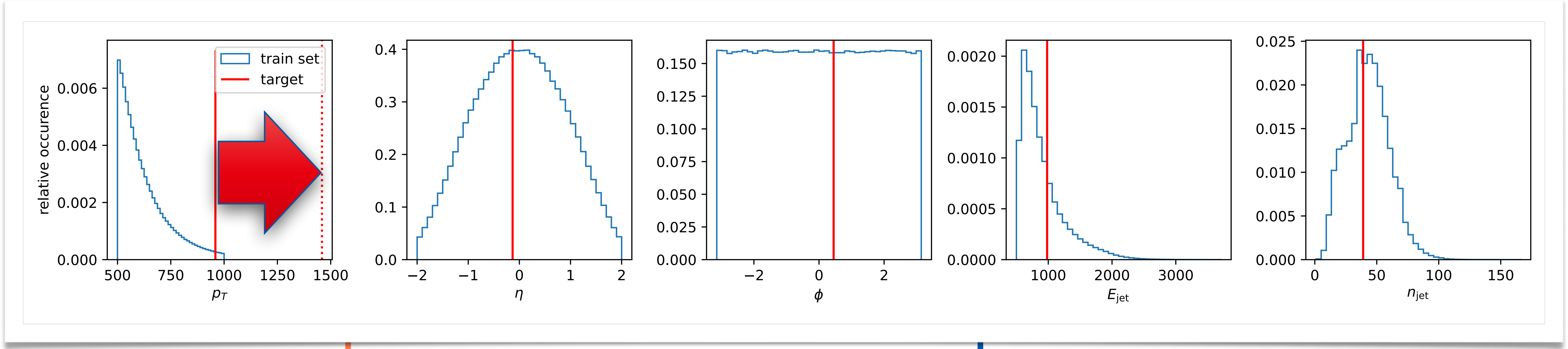
# Predicted ROC



# Unphysical Inputs

DASHH




1

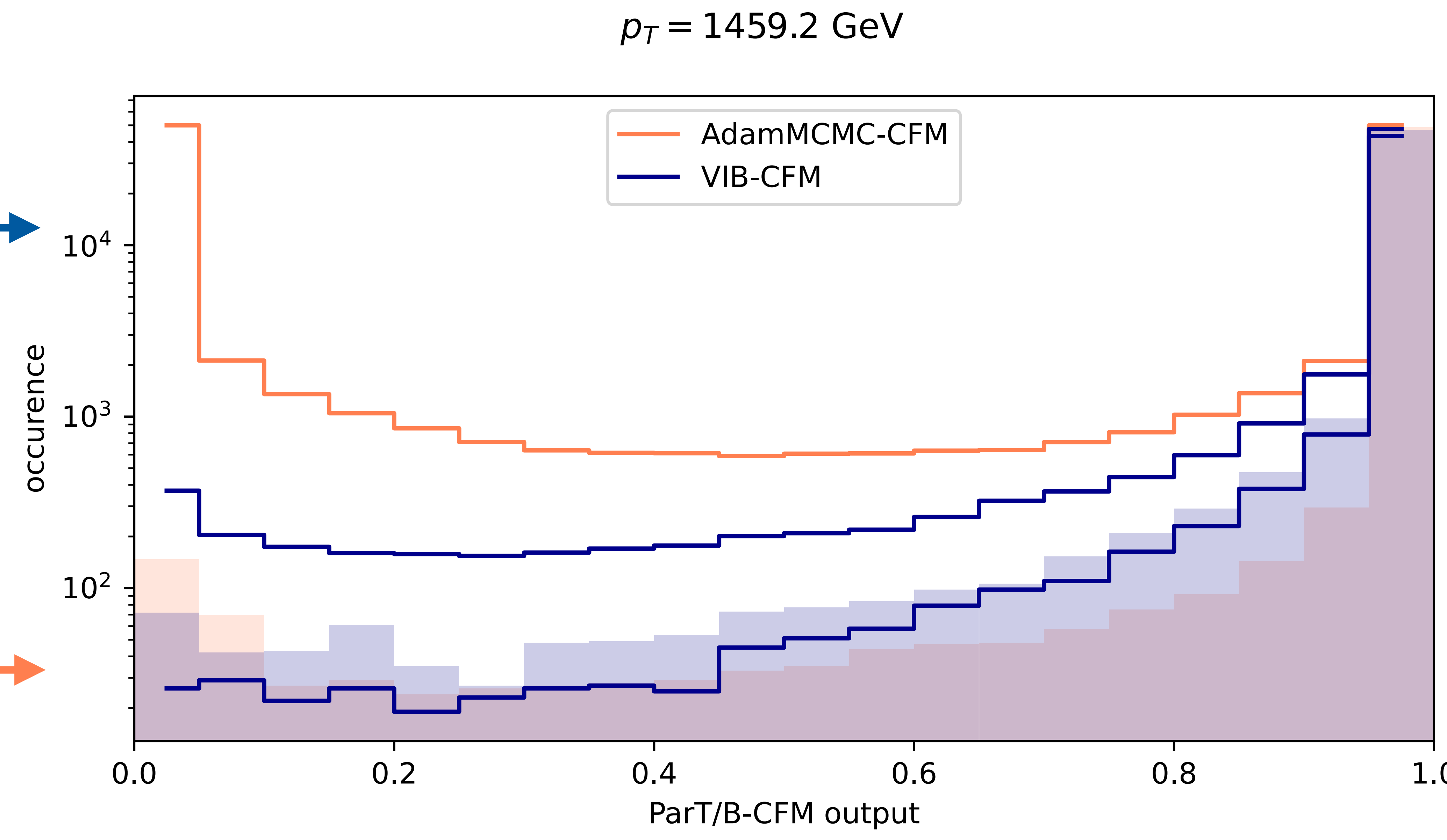
$z \propto \mathcal{N}(0,1)$

Conditional Flow Matching (CFM)
+ Variational Inference Bayes

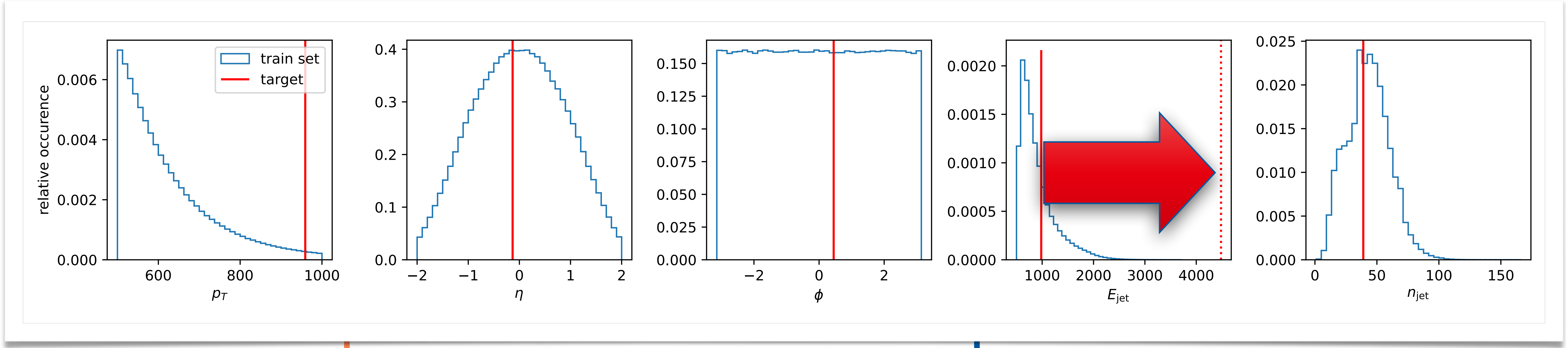
1

$z \propto \mathcal{N}(0,1)$

Conditional Flow Matching (CFM)
+ AdamMCMC



# Unphysical Inputs




1

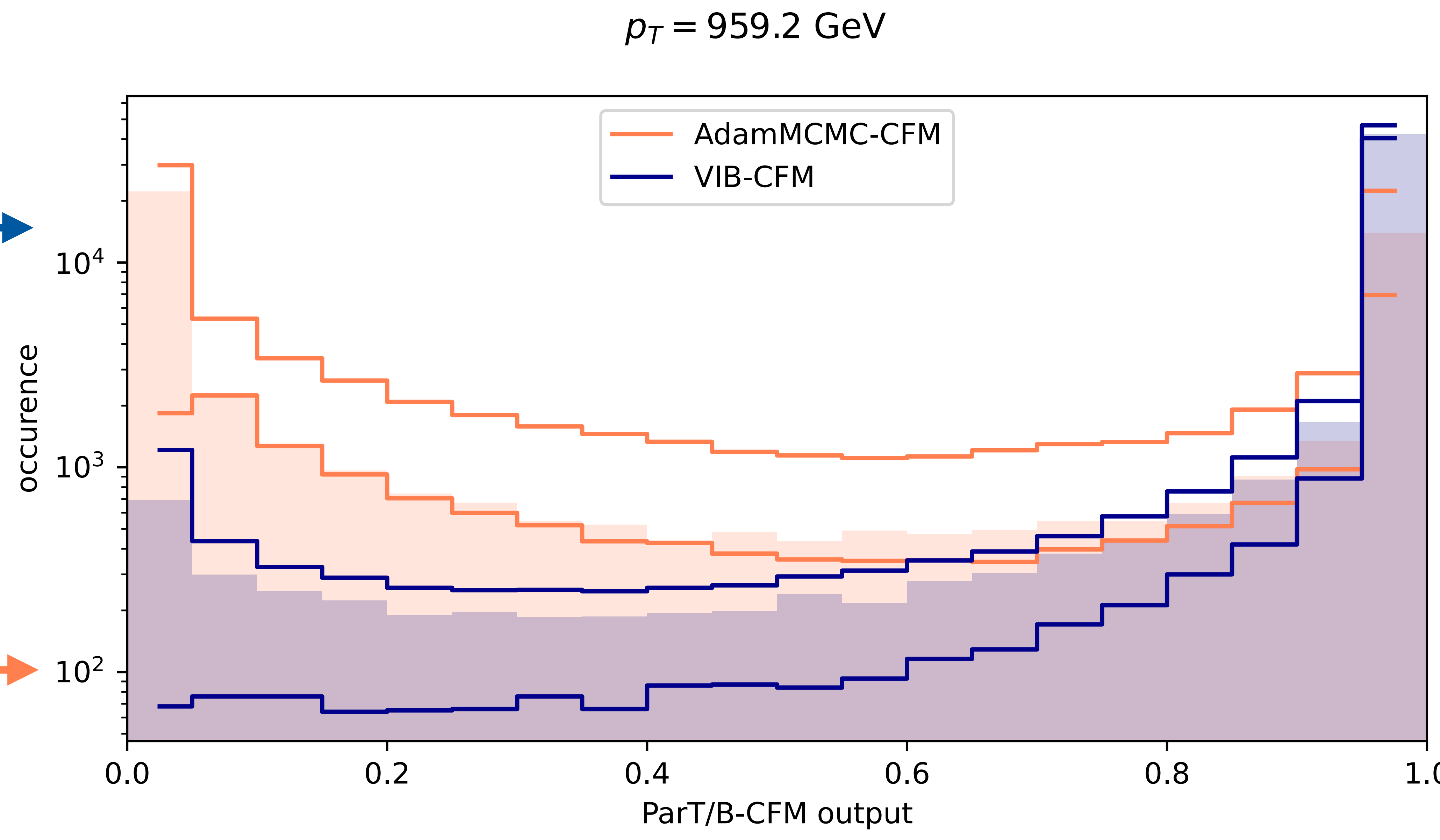
$z \propto \mathcal{N}(0,1)$

Conditional Flow Matching (CFM)
+ Variational Inference Bayes

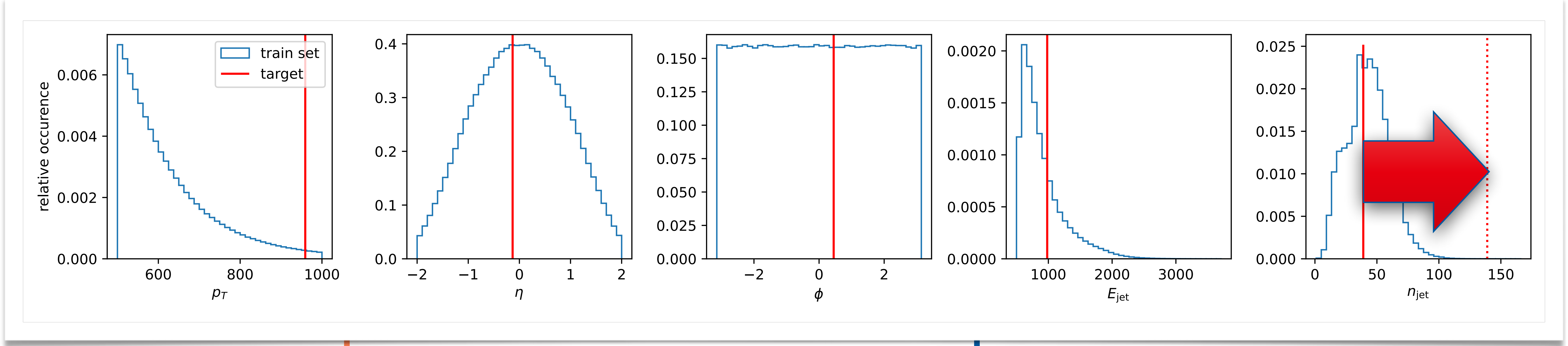
1

$z \propto \mathcal{N}(0,1)$

Conditional Flow Matching (CFM)
+ AdamMCMC



# Unphysical Inputs

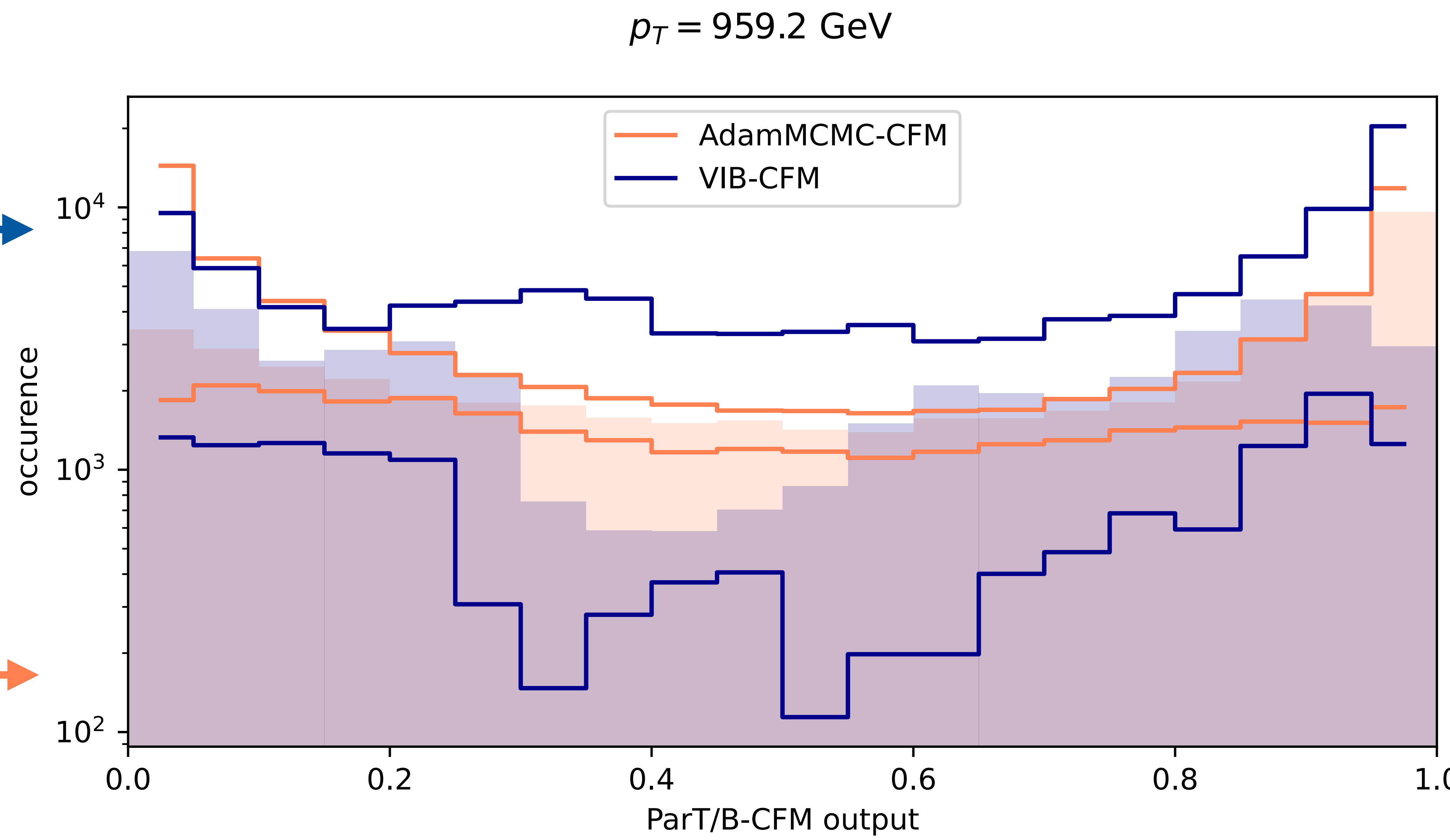



1

$z \propto \mathcal{N}(0,1)$ → Conditional Flow Matching (CFM) + Variational Inference Bayes →

1

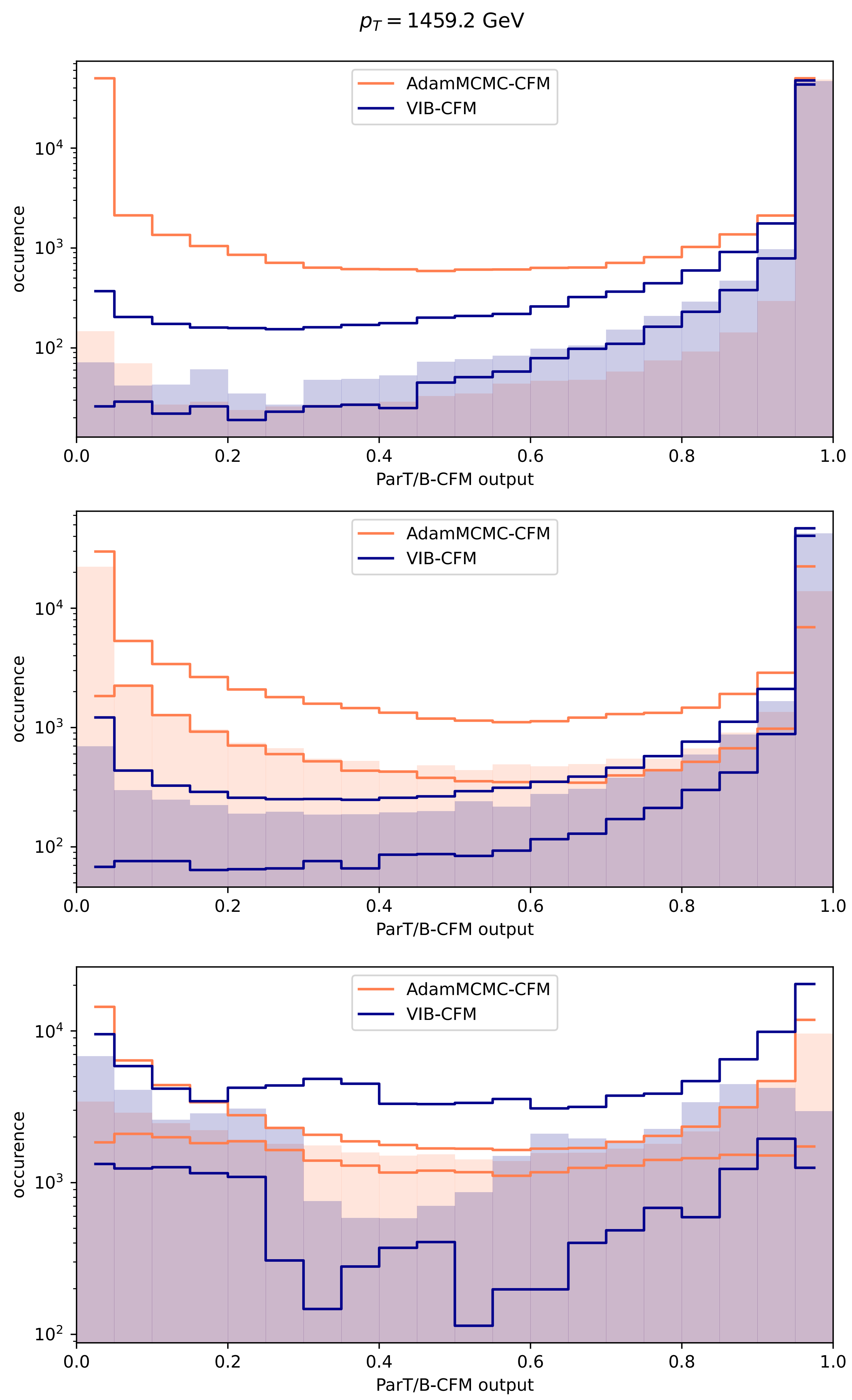
$z \propto \mathcal{N}(0,1)$ → Conditional Flow Matching (CFM) + AdamMCMC →



# Conclusion

- Adam-MCMC can provide improved error estimates over more common Bayesian architectures

- CFM model can can predict the in-distribution behavior of a large classifier well
  - Independent of detector-level data
  - Can be shared with analysis

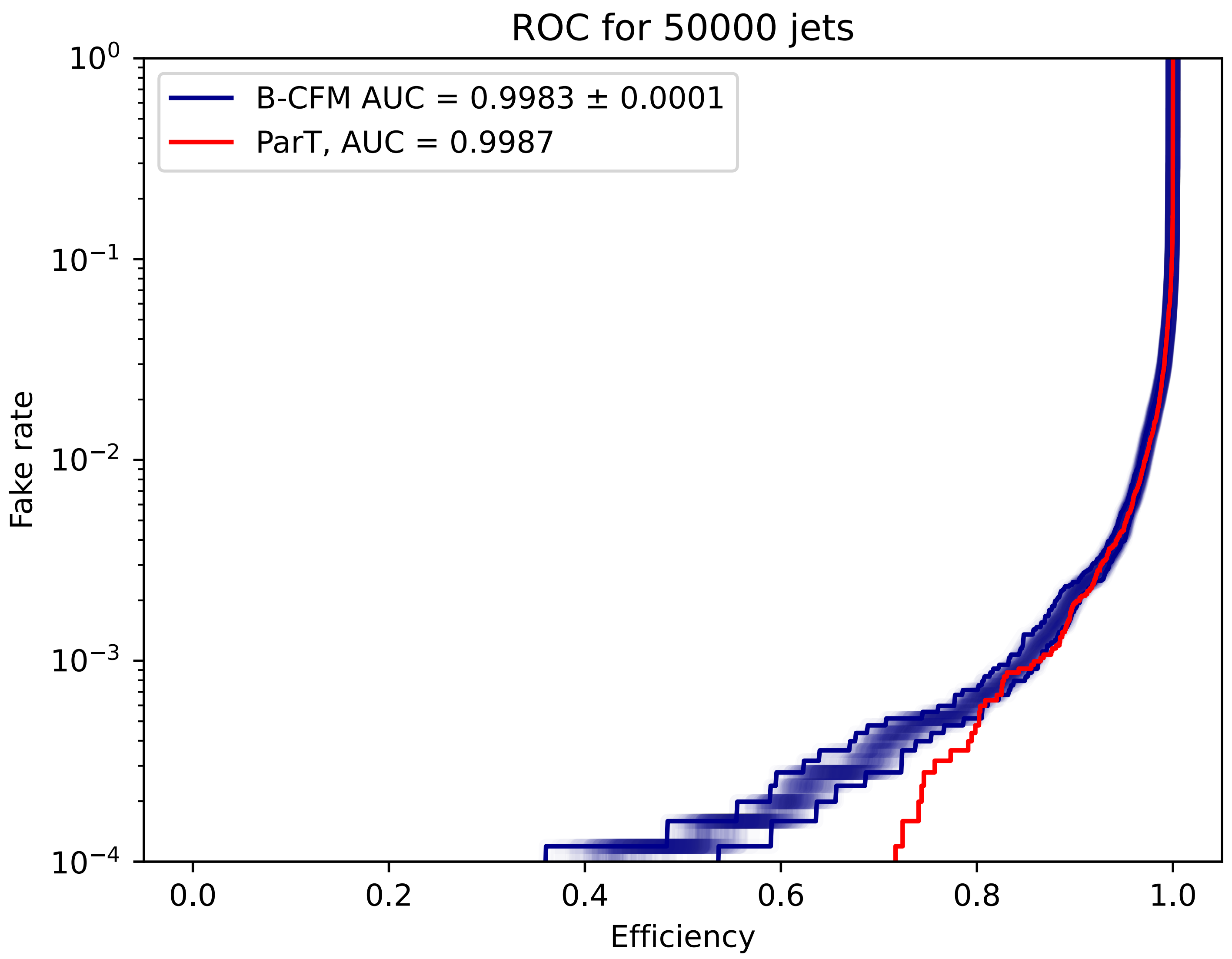


# Effects of the Prior Parameter $c$

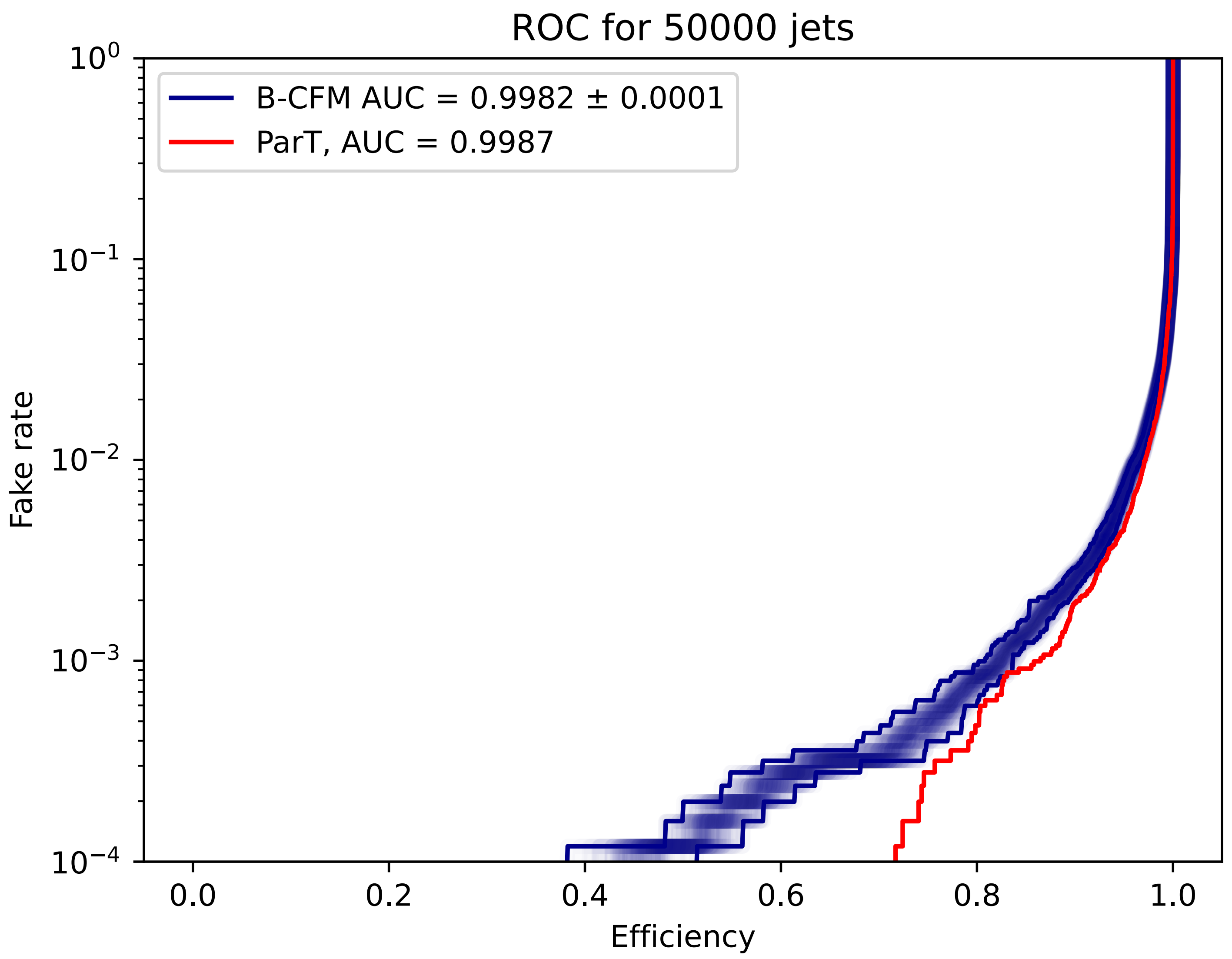
**Bayesian Conditional Flow Matching:**

- Bayesian loss $\mathscr{L}_{\text{BNN}} = \text{KL}\left[q(\theta), p\left(\theta \mid x\right)\right] = -\int \mathrm{d}\theta\, q(\theta)\, \log p\left(x \mid \theta\right) + \text{KL}[q(\theta), p(\theta)] +$ const.
- connect both $\mathscr{L}_{\text{B-CFM}} = \left\langle \mathscr{L}_{\text{CFM}} \right\rangle_{\theta \sim q(\theta)} + c\text{KL}[q(\theta), p(\theta)]$, with $q(\theta)$ uncorrelated Gaussian shape

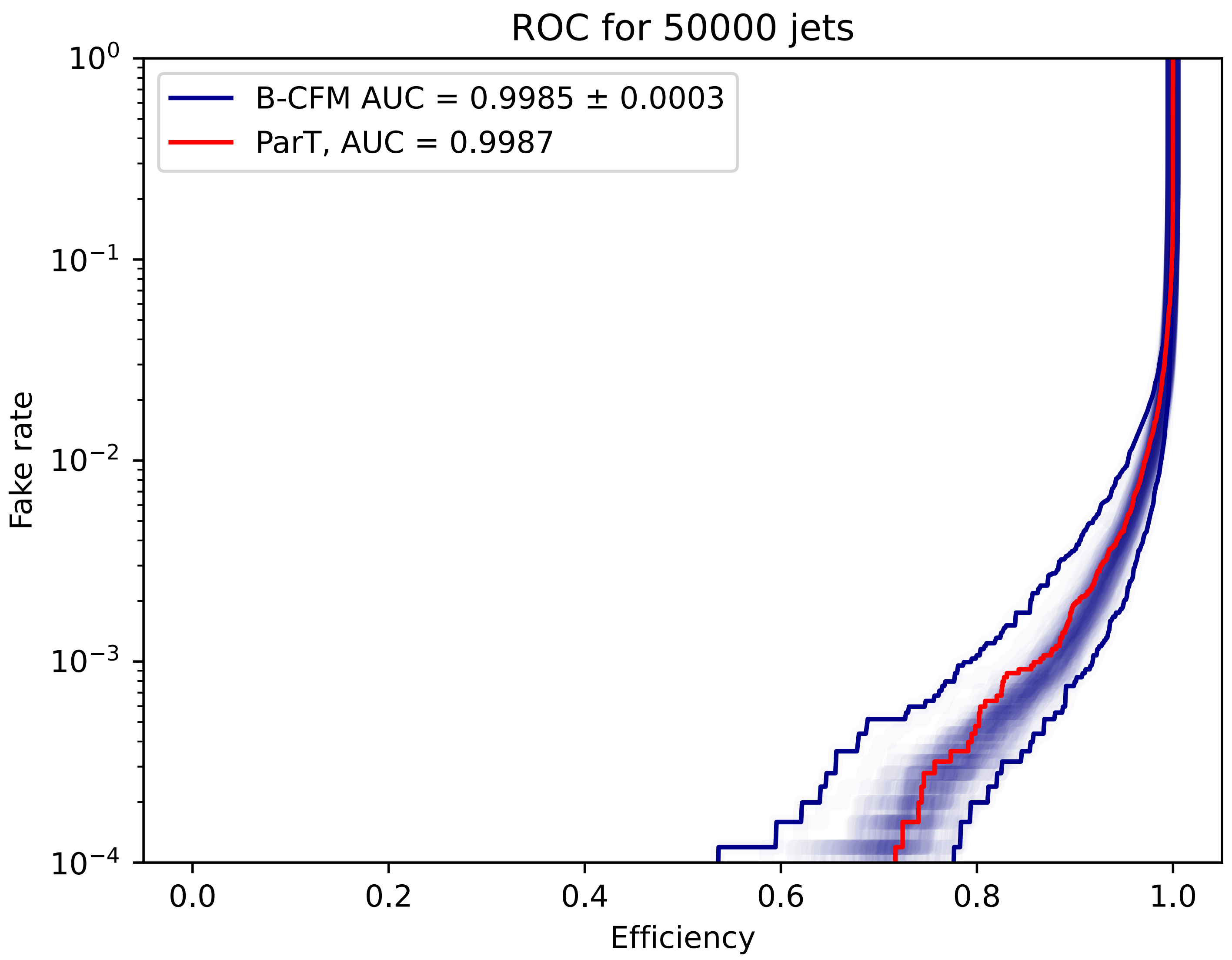
$c = 0.01$




$c = 1$




$c = 100$



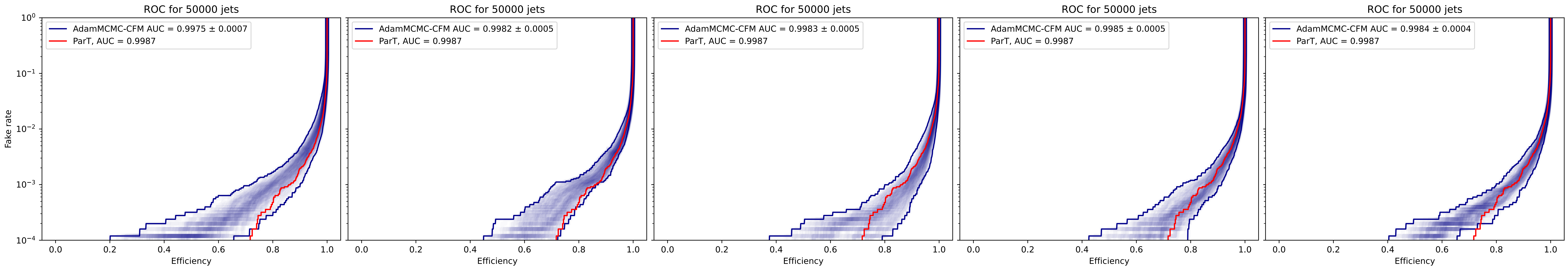
# Effects of high inverse temperature $\lambda$

**Adam-MCMC Bayesian Conditional Flow Matching:**

- Metropolis-Hastings correction: Accept new weight values with probability $\alpha = \dfrac{\exp\left(-\lambda L_n(\tau_i)\right)\, q(\theta_i \,|\, \tau_i)}{\exp\left(-\lambda L_n(\theta_i)\right)\, q(\tau_i \,|\, \theta_i)}$

- $\lambda$ gives the inverse temperature of a tempered Gibbs-Posterior $p_\lambda(\vartheta \,|\, D_n) \propto \exp(-\lambda L_n(\vartheta))\; p(\vartheta)$

$\lambda = 1$ $\lambda = 50$ $\lambda = 100$ $\lambda = 250$ $\lambda = 500$

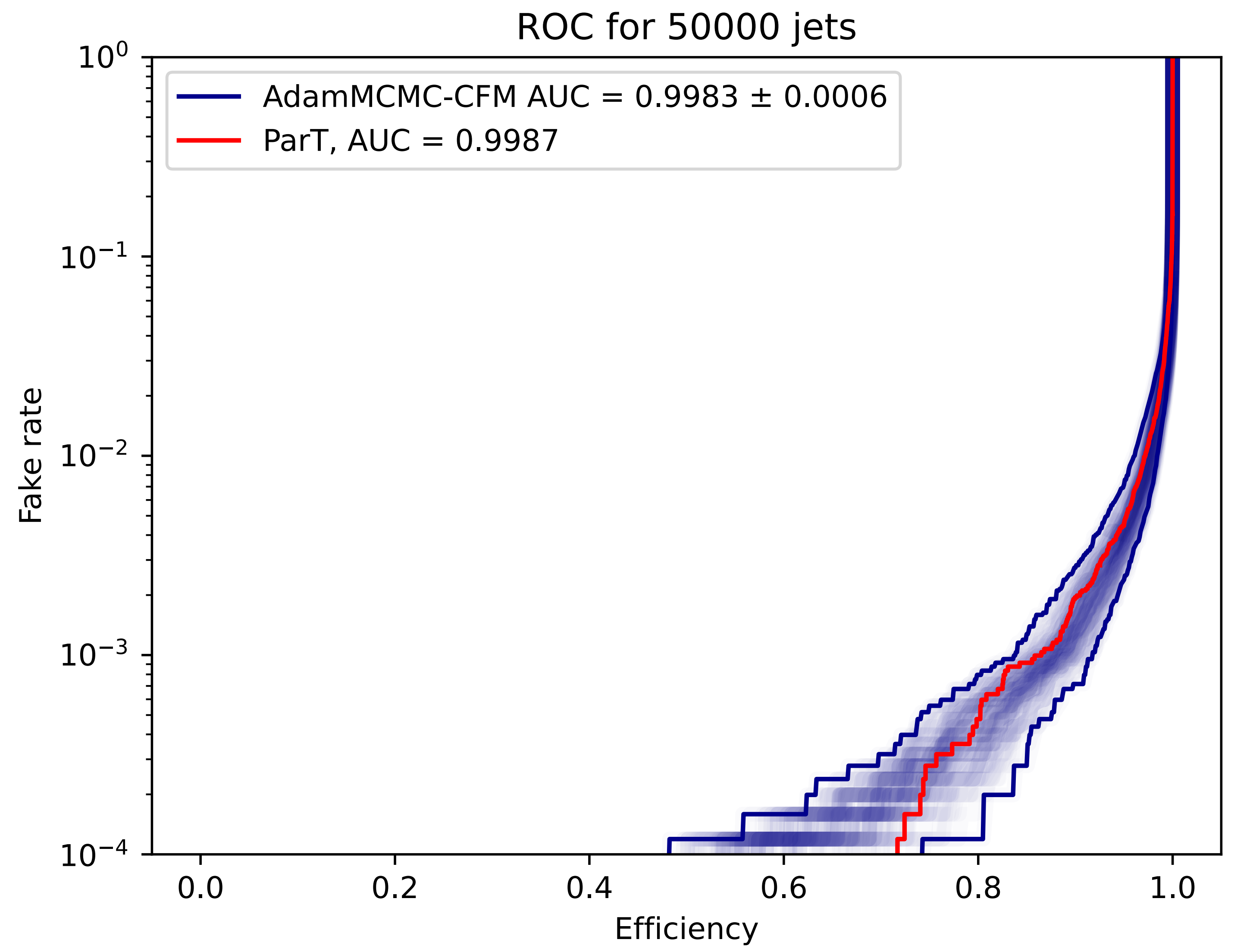


# Effects of low inverse temperature $\lambda$

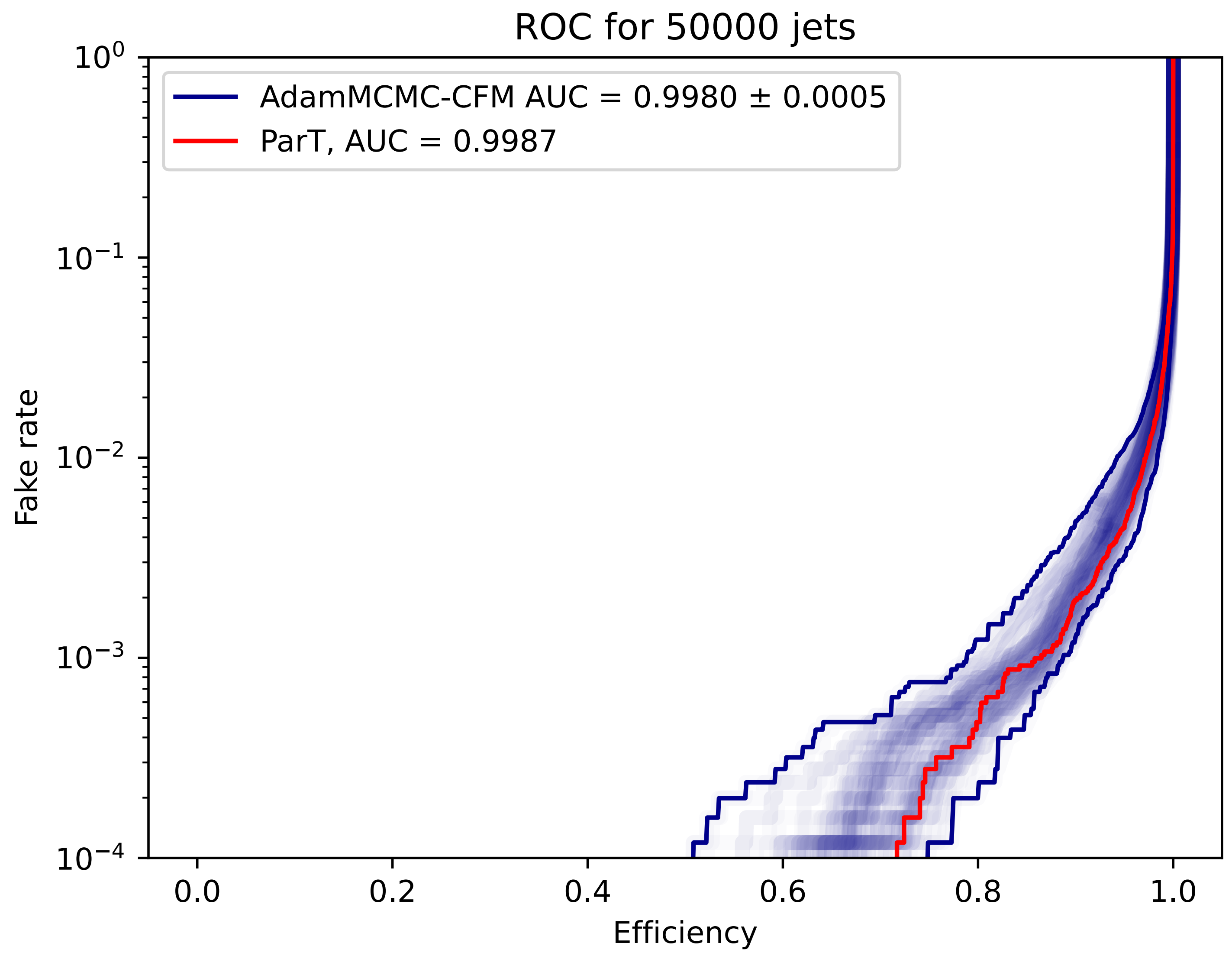
**Adam-MCMC Bayesian Conditional Flow Matching:**

- Metropolis-Hastings correction: Accept new weight values with probability $\alpha = \dfrac{\exp\left(-\lambda L_n(\tau_i)\right) q(\theta_i \,|\, \tau_i)}{\exp\left(-\lambda L_n(\theta_i)\right) q(\tau_i \,|\, \theta_i)}$
- $\lambda$ gives the inverse temperature of a tempered Gibbs-Posterior $p_\lambda(\vartheta \,|\, D_n) \propto \exp(-\lambda L_n(\vartheta))\, p(\vartheta)$

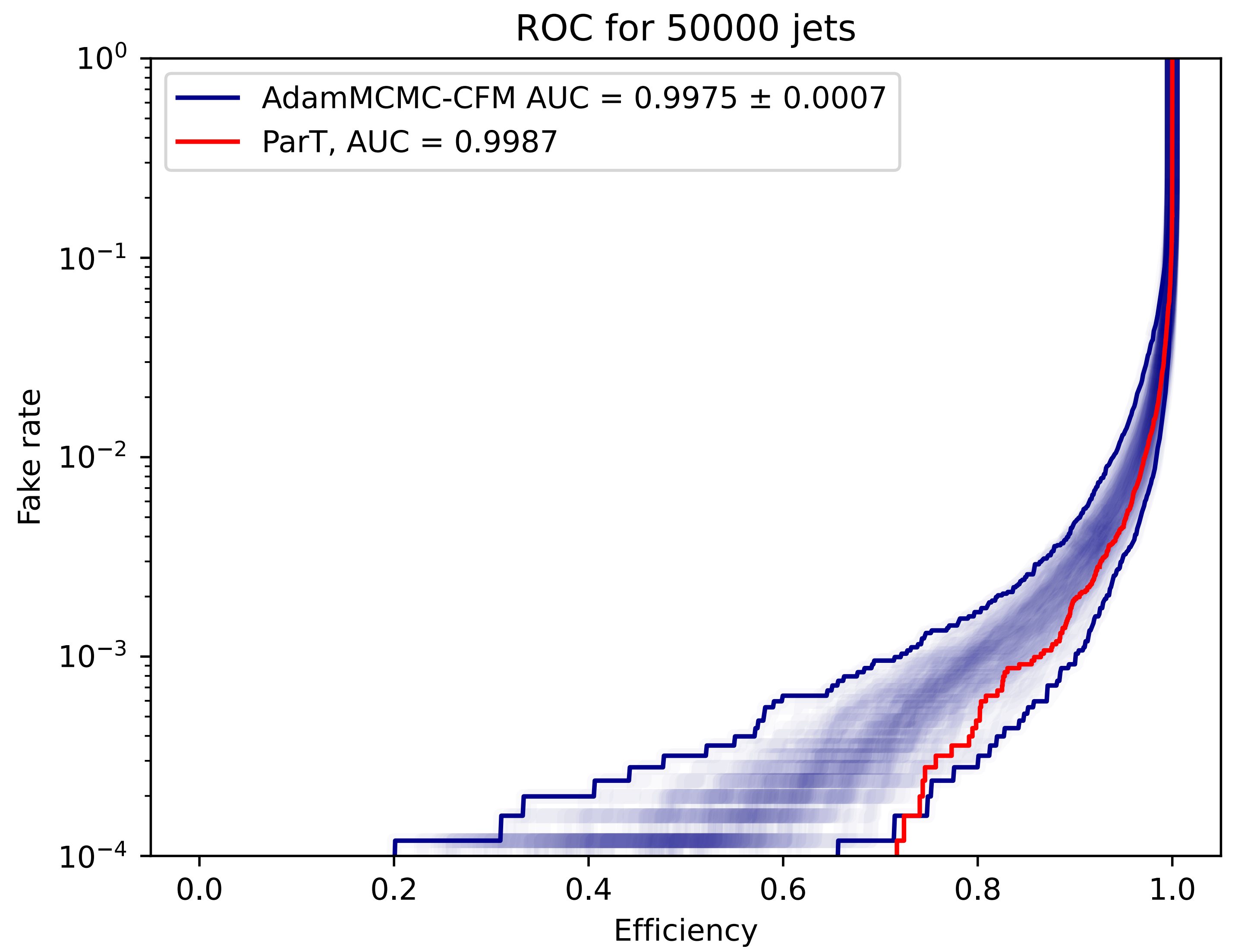
$\lambda = 0.01$




$\lambda = 0.1$

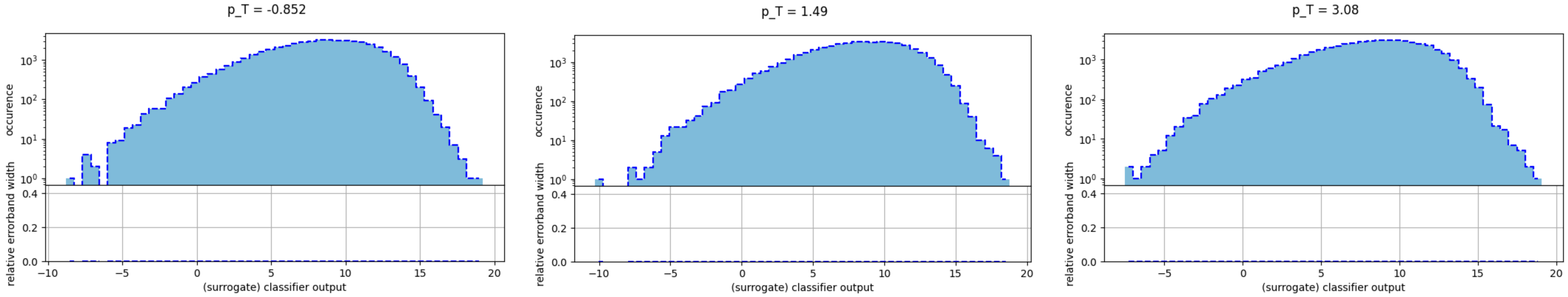



$\lambda = 1$

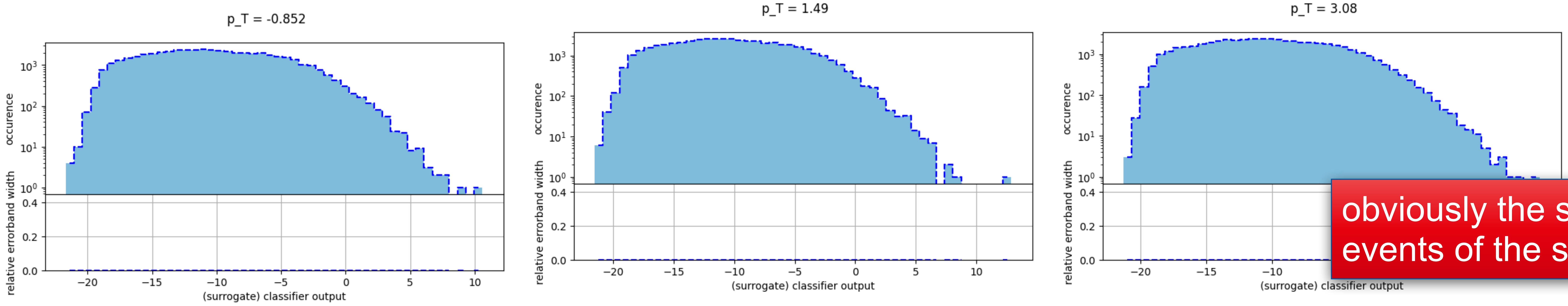


# What if only trained on truth?

## top jets




## not top jets




obviously the same for events of the same class