Probabilistic Models

Trainable parameters

$p_\theta(x)$ which enables

- Fitting to $p_{\text{data}}(x)$
- Likelihood evaluation $p_\theta(x_i)$
- Sampling $x \sim p_\theta(x)$

Maximum likelihood estimation

What would I like from a model

- Expressivity
- Efficiency

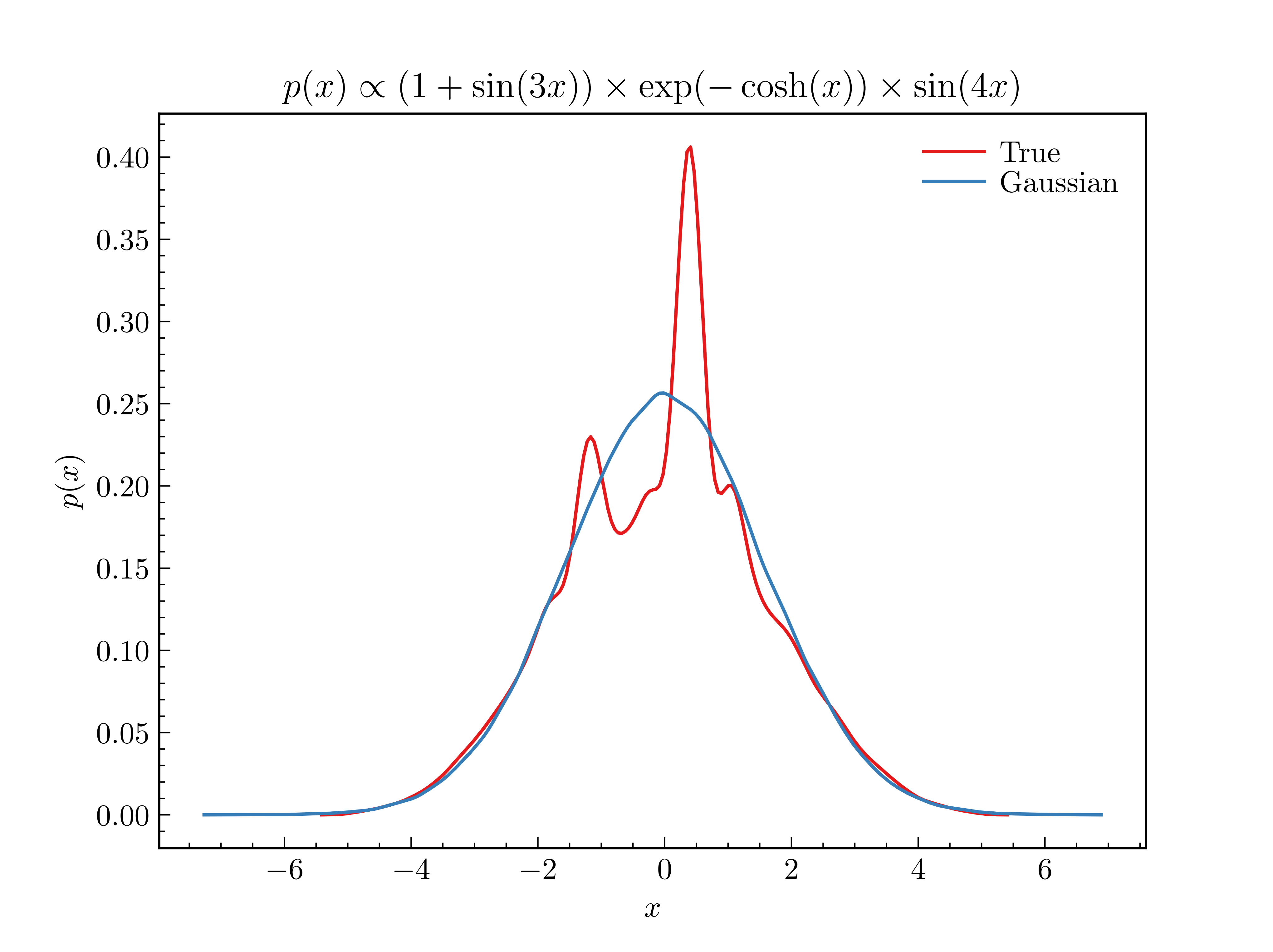
Can fit complicated $p_{\text{data}}(x)$

Fast sampling & likelihood evaluation

Gaussian distribution

$$p_\theta(x) = \mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Parameters: μ, σ



Not very expressive....

Latent Variable Models

More expressive models by combining simple ones

Latent variable $z \longrightarrow p(x, z)$

Dropping the θ notation

Interested only in marginal

$$p(x) = \int_{\mathcal{Z}} dz\, p(x, z) = \int_{\mathcal{Z}} dz\, p(z)\, p(x \mid z)$$

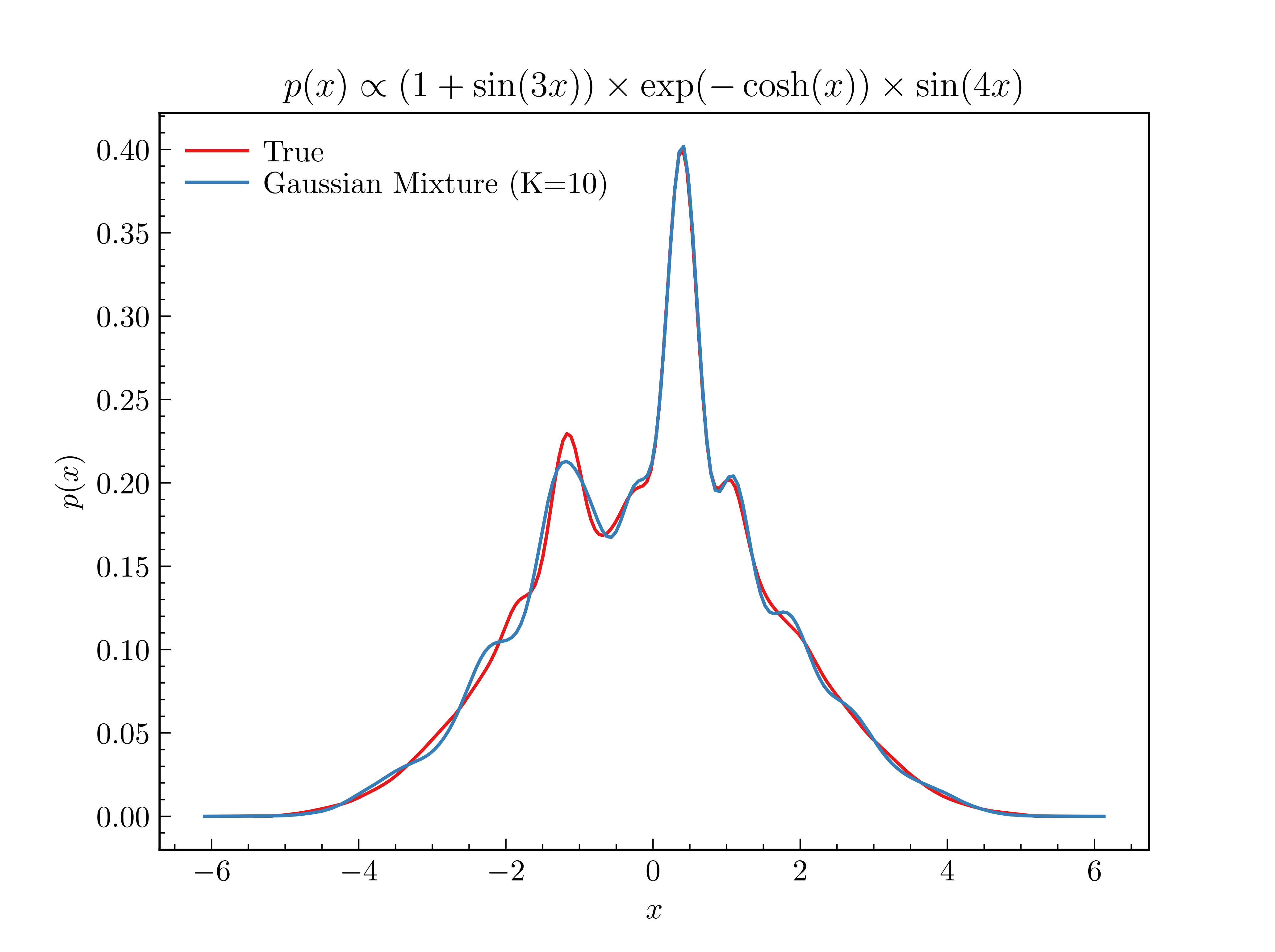
Sampling is still *simple*

$$z \sim p(z)$$
$$x \sim p(x \mid z)$$

Gaussian mixture model (discrete latent)

$$p(x) = \sum_{i=1}^{K} p_i \mathcal{N}(x \mid \mu_i, \sigma_i)$$

Parameters: p_i, μ_i, σ_i



Much better!

But doesn't scale well to higher dims…

Latent Variable Models

Marginal likelihood is usually *intractable*

$$p(x) = \int_{\mathcal{Z}} dz\, p(x,z) = \int_{\mathcal{Z}} dz\, p(z)\, p(x\,|\,z)$$

Too hard to compute efficiently

We would *really* like to have acces to the exact likelihood

- Training through maximum likelihood estimation
- Use for anomaly detection, likelihood-free inference, etc

⇨ Normalizing flows

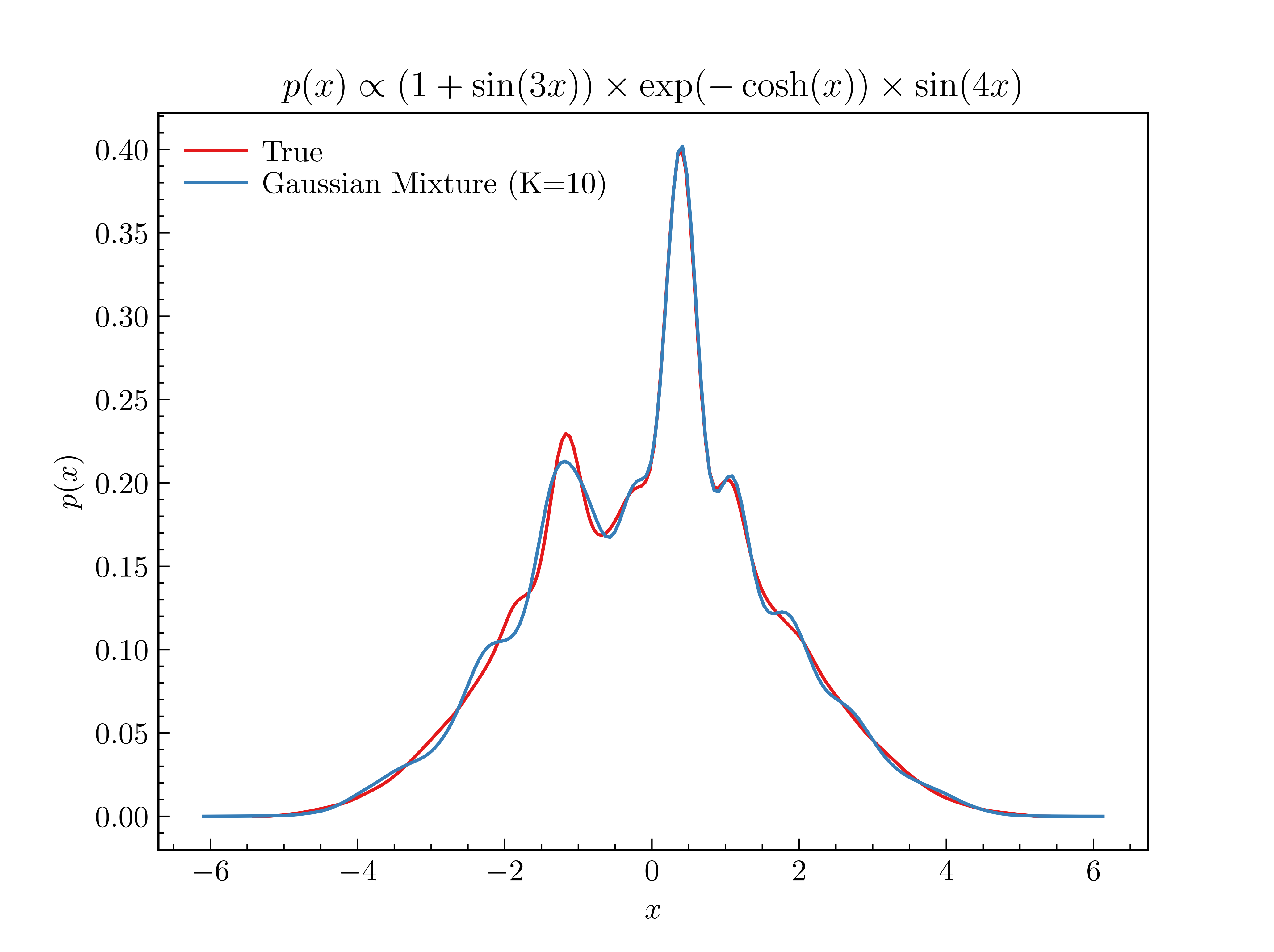
Other generative models actually also fit this latent variable description, and you can even combine then

RV: 2205.01697

Gaussian mixture model (discrete latent)

$$p(x) = \sum_{i=1}^{K} p_i \mathcal{N}(x\,|\,\mu_i, \sigma_i)$$

Parameters: p_i, μ_i, σ_i



Much better!

But doesn't scale well to higher dims…

Normalizing Flows

$$p(x) = \int_{\mathcal{Z}} dz\, p(x, z) = \int_{\mathcal{Z}} dz\, p(z)\, p(x \,|\, z)$$

Fix intractability by removing stochastic component from $p(x \,|\, z)$

$$p(x \,|\, z) \rightarrow \delta(x - f(z))$$

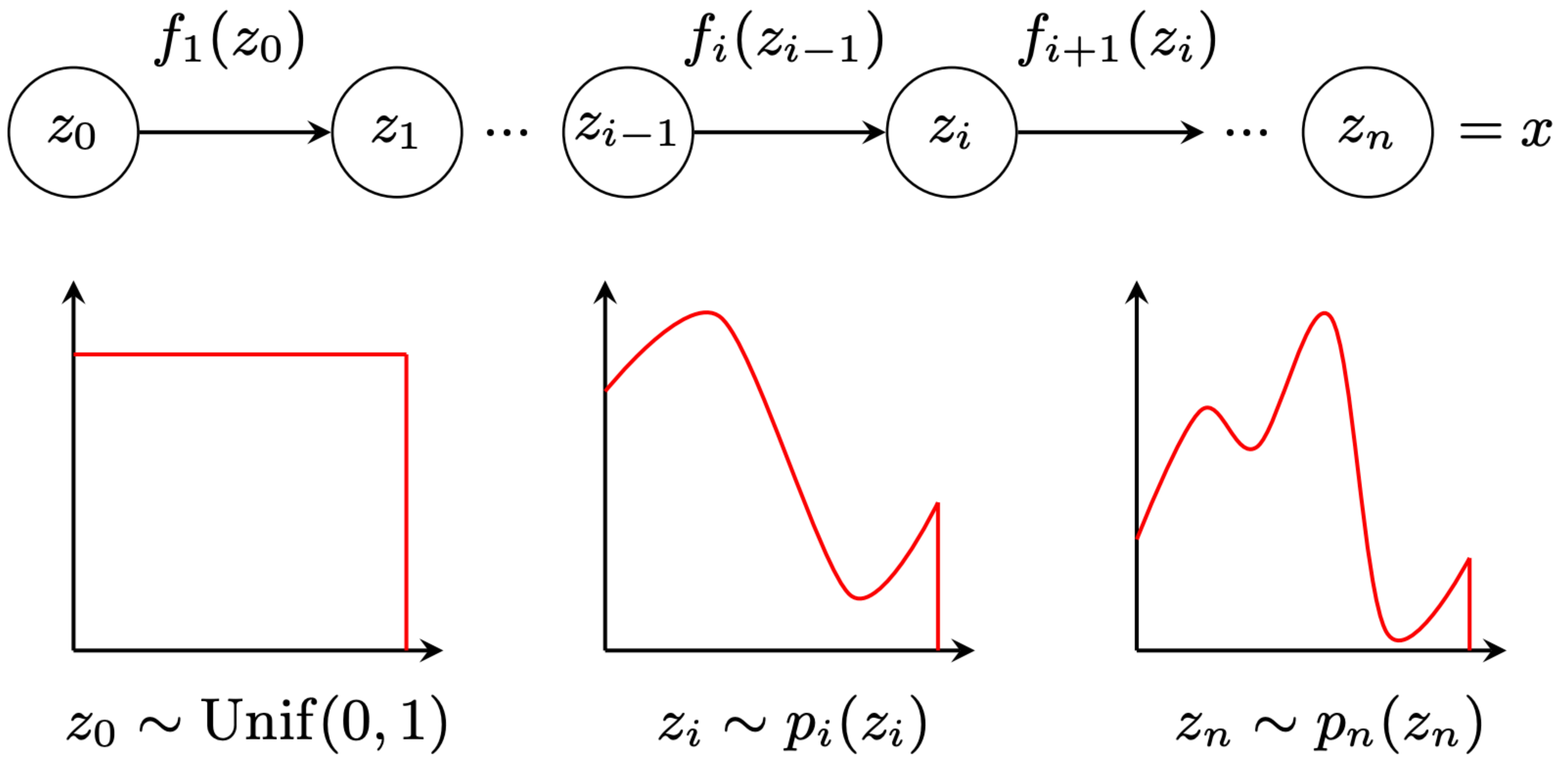
Parametric bijection

$$\Rightarrow \quad \log p(x) = \log \int_{\mathcal{Z}} dz\, p(z)\, \delta(x - f(z)) = \log p(z) + \log |\, J(x) \,|$$

$\dim(z) = \dim(x)$

Flow: Repeat a few times

$$\log p(x) = \log p(z_0) + \sum_{i=1} \log |\, J_i(z_i) \,| \,.$$

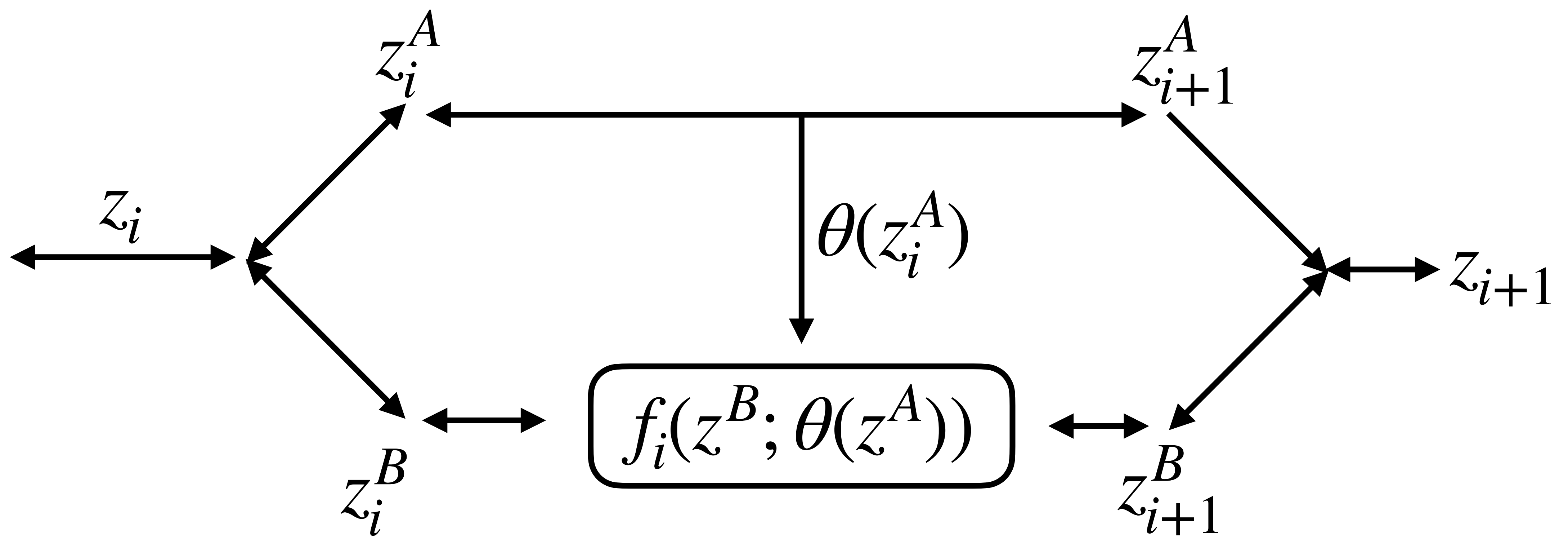


Two Normalizing Flow Architectures

Jacobian $|J_i(x)|$ must be easy to evaluate

Coupling layers $p(z) = p(z^A)p(z^B \mid z^A)$

Split into two pieces $z \to z^A, z^B$

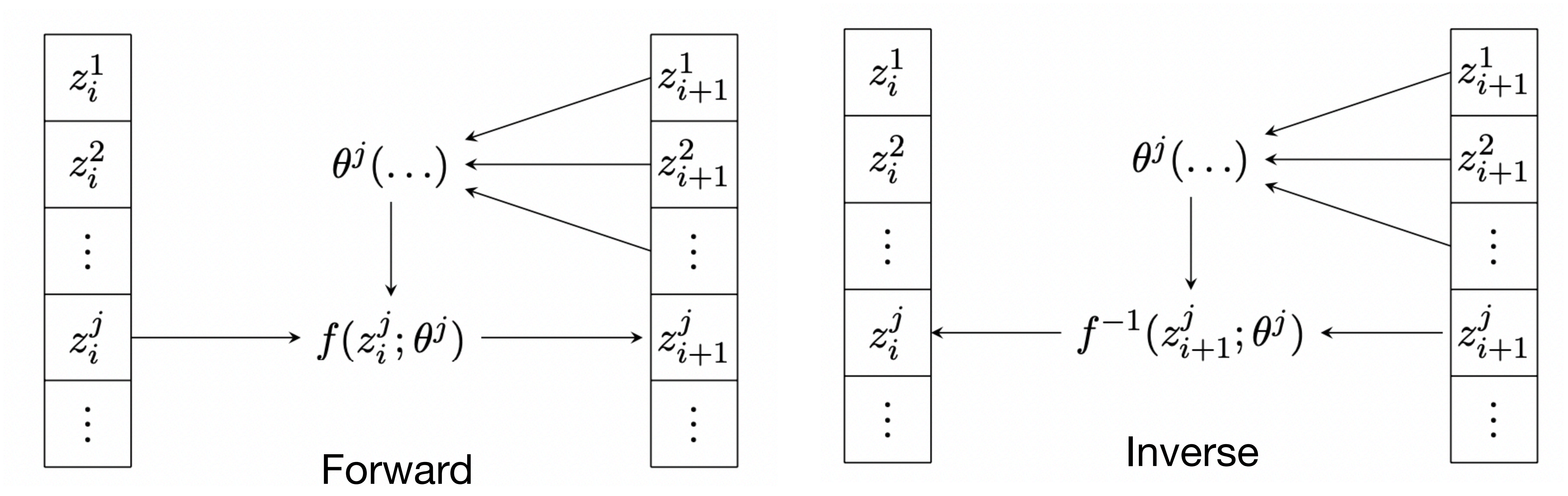


- Fast in both directions
- Simple Jacobian

$$|J| = \left|\frac{dz_{i+1}}{dz_i}\right| = \begin{vmatrix} \overrightarrow{1} & \overrightarrow{0} \\ \frac{df_i}{d\theta}\frac{d\theta}{dz_i^B} & \frac{df_i}{dz_i^B} \end{vmatrix} = \left|\frac{df_i}{dz_i^B}\right|$$

Autoregressive layers $p(z) = \prod_{j=1}^{D} p\left(z^j \mid z^{1:j-1}\right)$

Split into D 1-d transforms



- Fast in only one direction
- Lower-triangular Jacobian

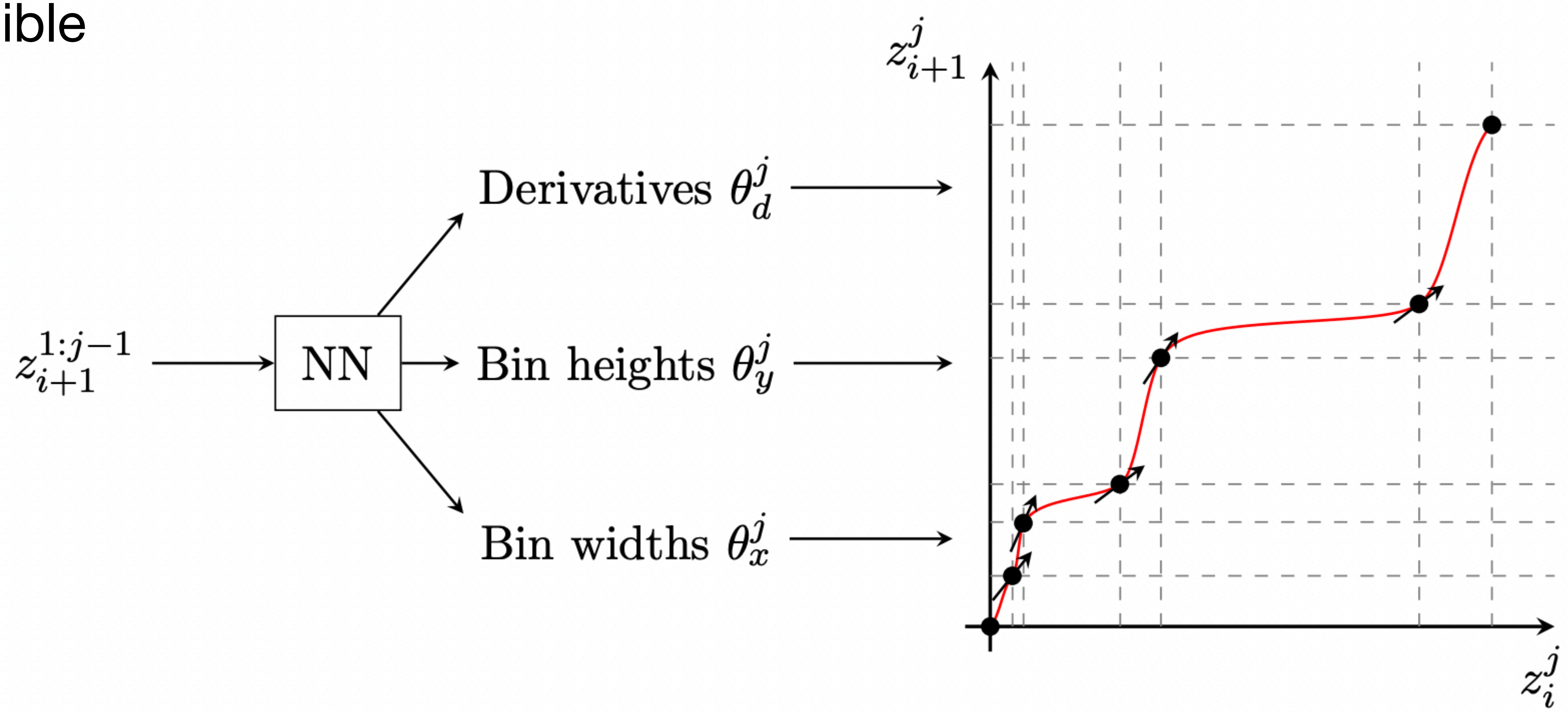
⇨ $\mathcal{O}(d)$ instead of $\mathcal{O}(d^3)$ determinant

Flow Transforms

Some requirements:

- Bijective functions $f_i(z) \leftrightarrow f_i^{-1}(x)$
- As expressive as possible

Rational Quadratic Spline



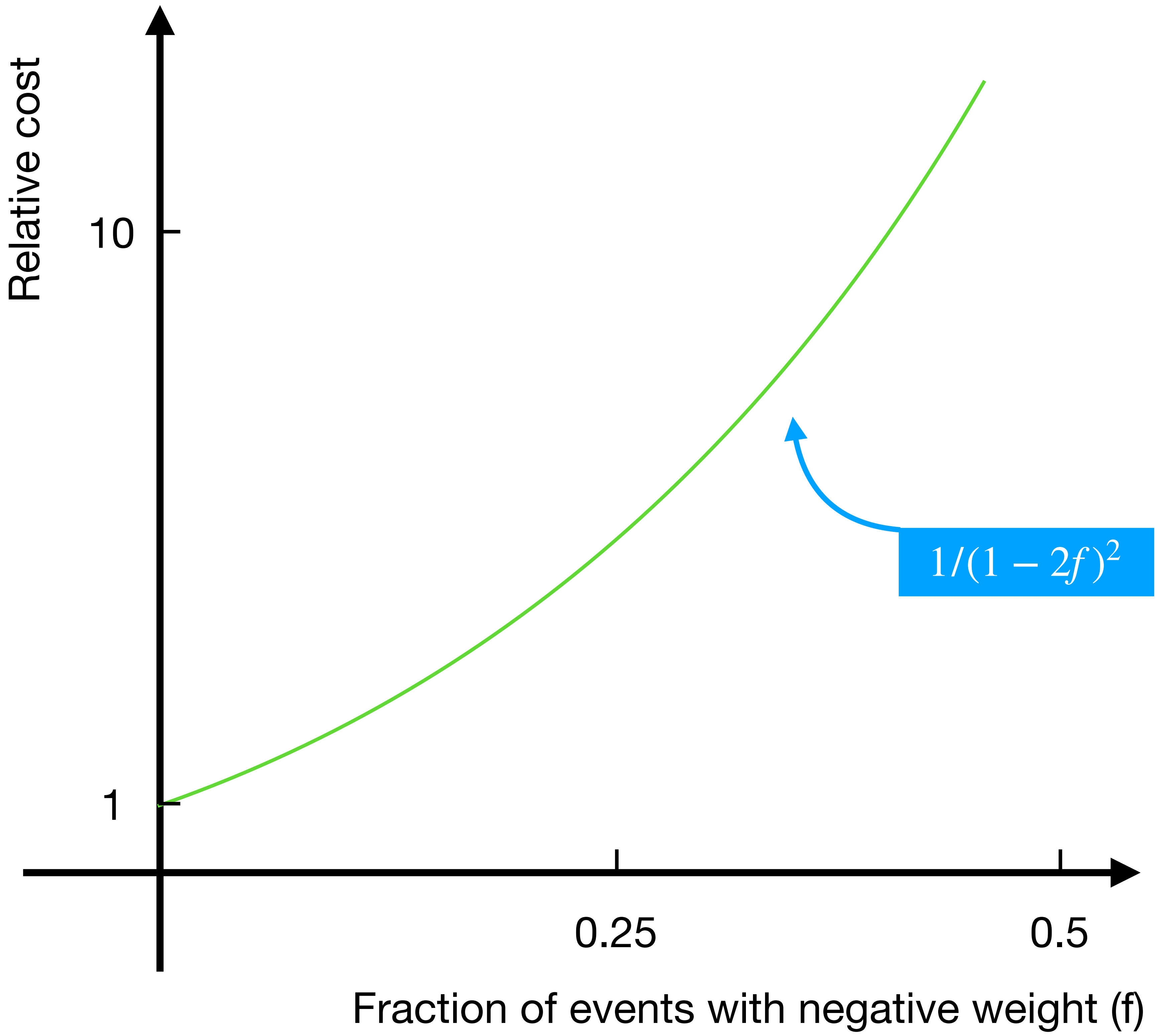
Example Application: Negative Weights

Stienen, RV: 2011.13445

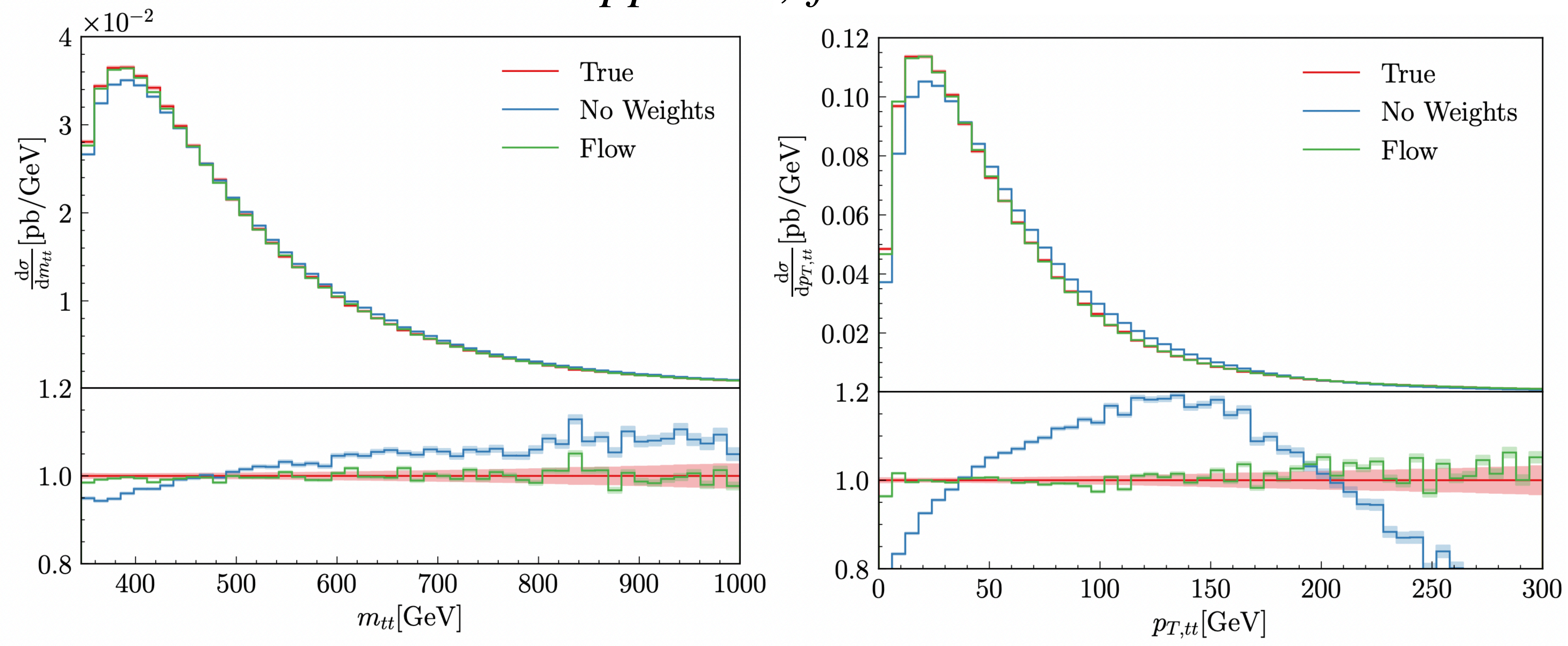
Common in ME/PS matching (MC@NLO)

⇨ Require more events for same statistics

- Train normalizing flow on weighted events
- Generate unweighted events



$pp \to t\bar{t},\, f = 0.239$

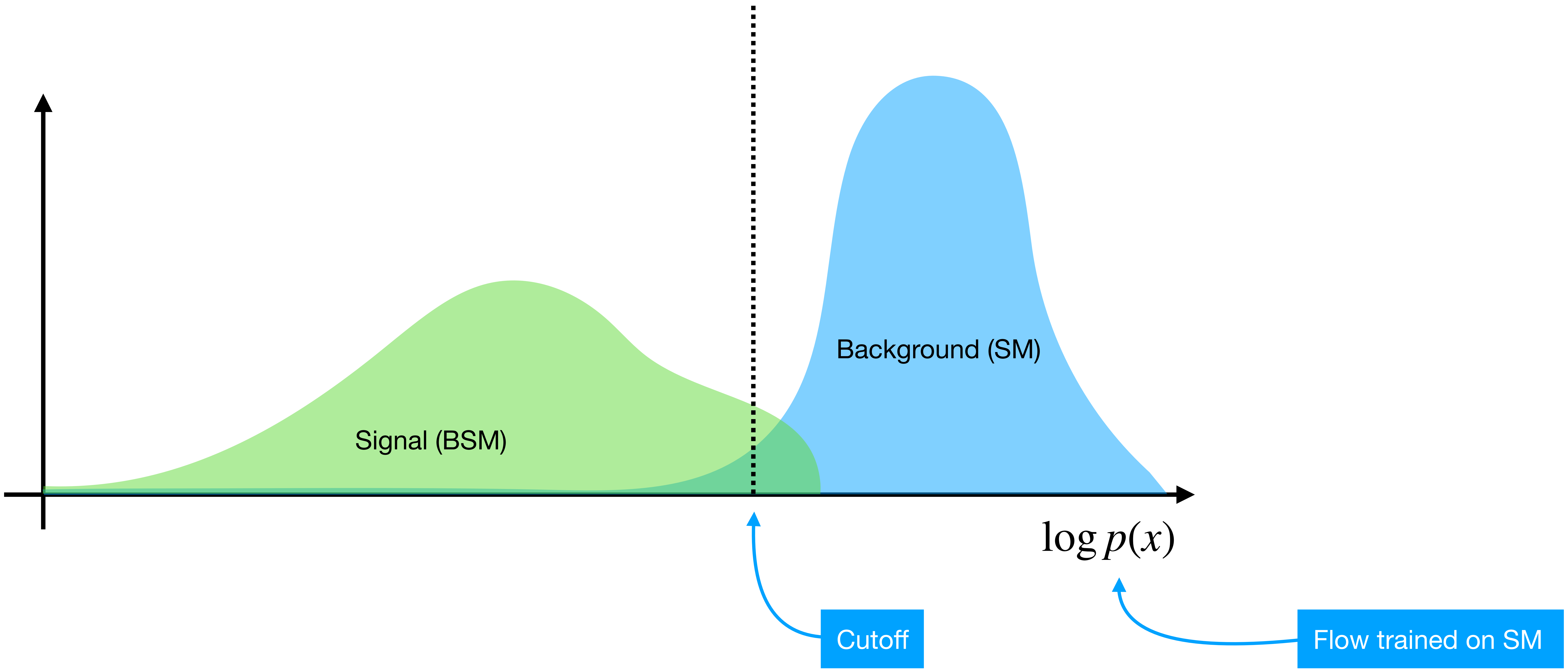


Example Application: Anomaly Detection

Caron, Hendriks, RV: 2011.13445

Search for out-of-distribution events

⇨ Identify regions of phase space for further study



Dark Machines Anomaly Detection Challenge

1. > 1B SM events:

Four channels

- Channel 1: Hadronic activity with lots of missing energy (214k events)
- Channel 2a: At least three identified leptons (20k events)
- Channel 2b: At least two identified leptons (340k events)
- Channel 3: Inclusive with moderate missing energy (8.5M events)

2. Validation set:

Events from various BSM models (Z', SUSY, etc.)

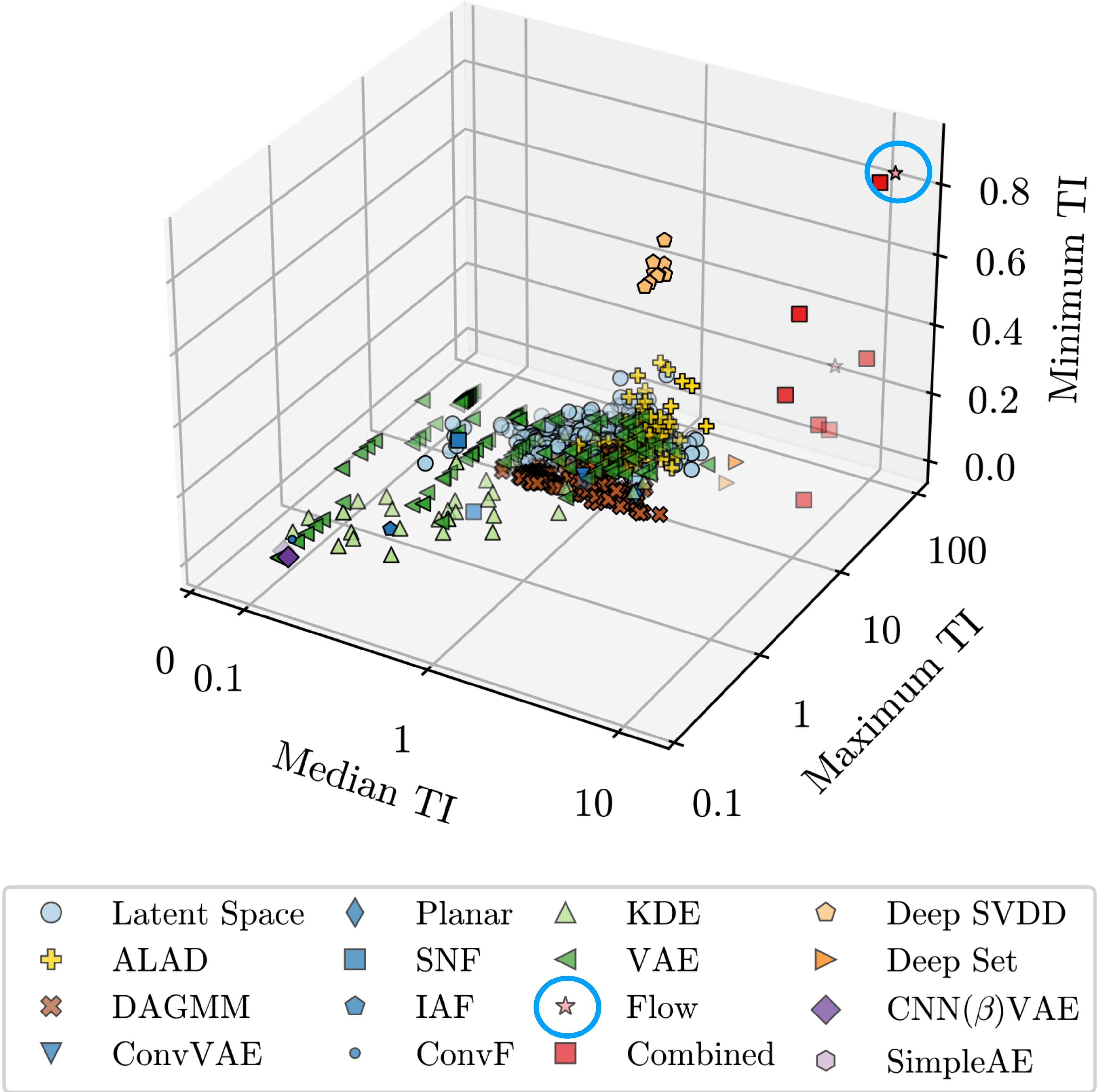
3. Test set:

Secret dataset with labels not known to model authors

Figure of merit:

$$\text{Max SI} = \max_{\epsilon_B} \epsilon_S(\epsilon_B)/\sqrt{\epsilon_B}$$

where $\epsilon_B \in \{10^{-2}, 10^{-3}, 10^{-4}\}$



The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider

T. Aarrestad[a] M. van Beekveld[b] M. Bona[c] A. Boveia[e] S. Caron[d] J. Davies[c] A. De Simone[f,g] C. Doglioni[h] J. M. Duarte[i] A. Farbin[j] H. Gupta[k] L. Hendriks[d] L. Heinrich[a] J. Howarth[l] P. Jawahar[m,a] A. Jueid[n] J. Lastow[h] A. Leinweber[o] J. Mamuzic[p] E. Merényi[q] A. Morandini[r] P. Moskvitina[d] C. Nellist[d] J. Ngadiuba[s,t] B. Ostdiek[u,v] M. Pierini[a] B. Ravina[l] R. Ruiz de Austri[p] S. Sekmen[w] M. Touranakou[x,a] M. Vaškevičiūte[l] R. Vilalta[y] J.-R. Vlimant[t] R. Verheyen[z] M. White[o] E. Wulff[h] E. Wallin[h] K.A. Wozniak[α,a] Z. Zhang[d]

Discussion

- Applications in physics
  - Event generation/numerical integration
  - Anomaly detection
  - Likelihood-free inference
  - ???
- Differences with other generative models
  - Easy to train
  - Not as flexible
- How to obtain the best performance?
  - Architecture/loss function
  - Discriminator-assisted training
- Not nearly as flexible as a MC event generator
  - Cuts/sharp features in phase space
  - Physical parameters