Event Generation and Density Estimation with Surjective Normalizing Flows

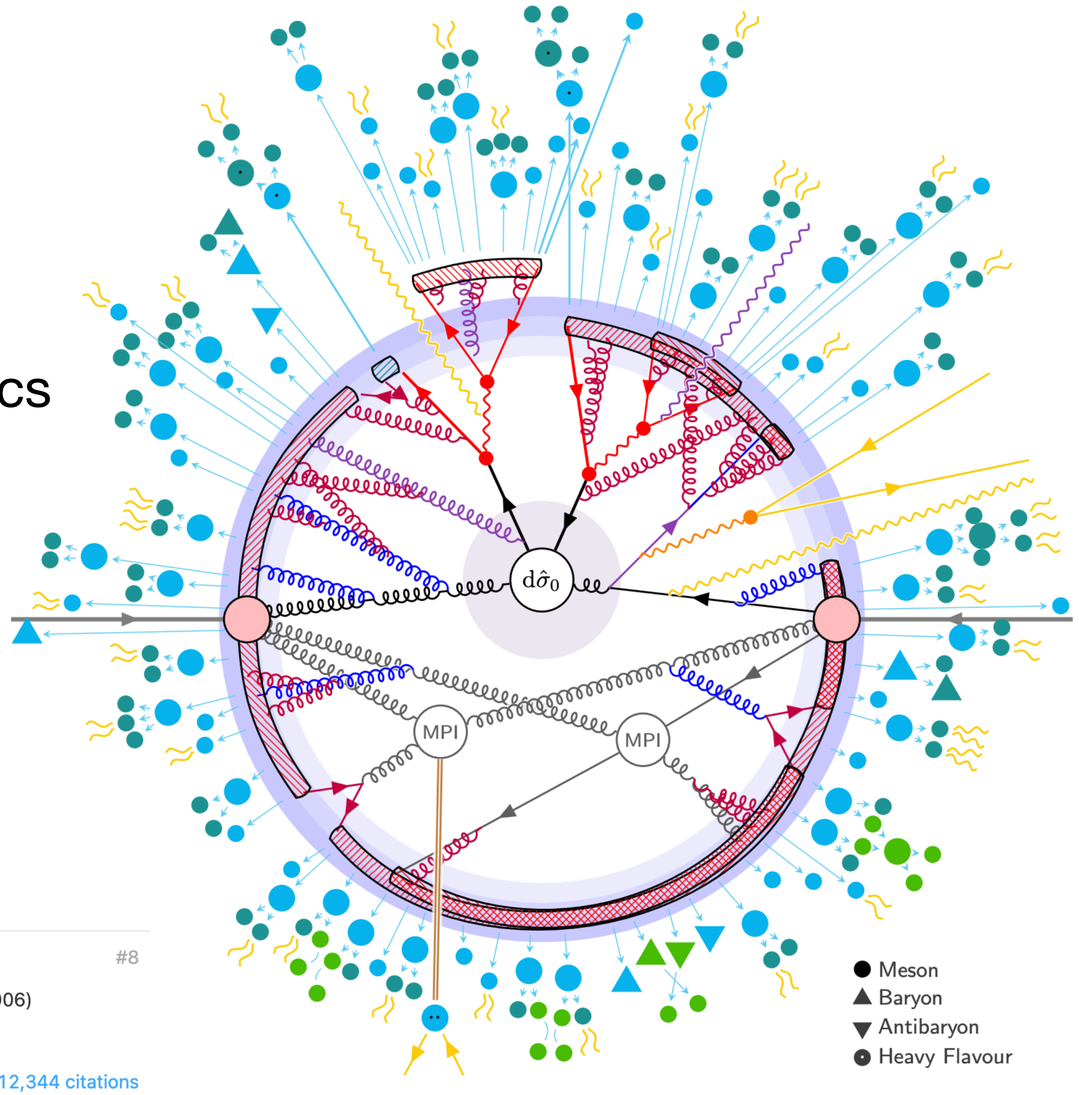
Rob Verheyen

2205.01697

European Research Council
Established by the European Commission

Event generators

Core component of modern particle physics



PYTHIA 6.4 Physics and Manual #8

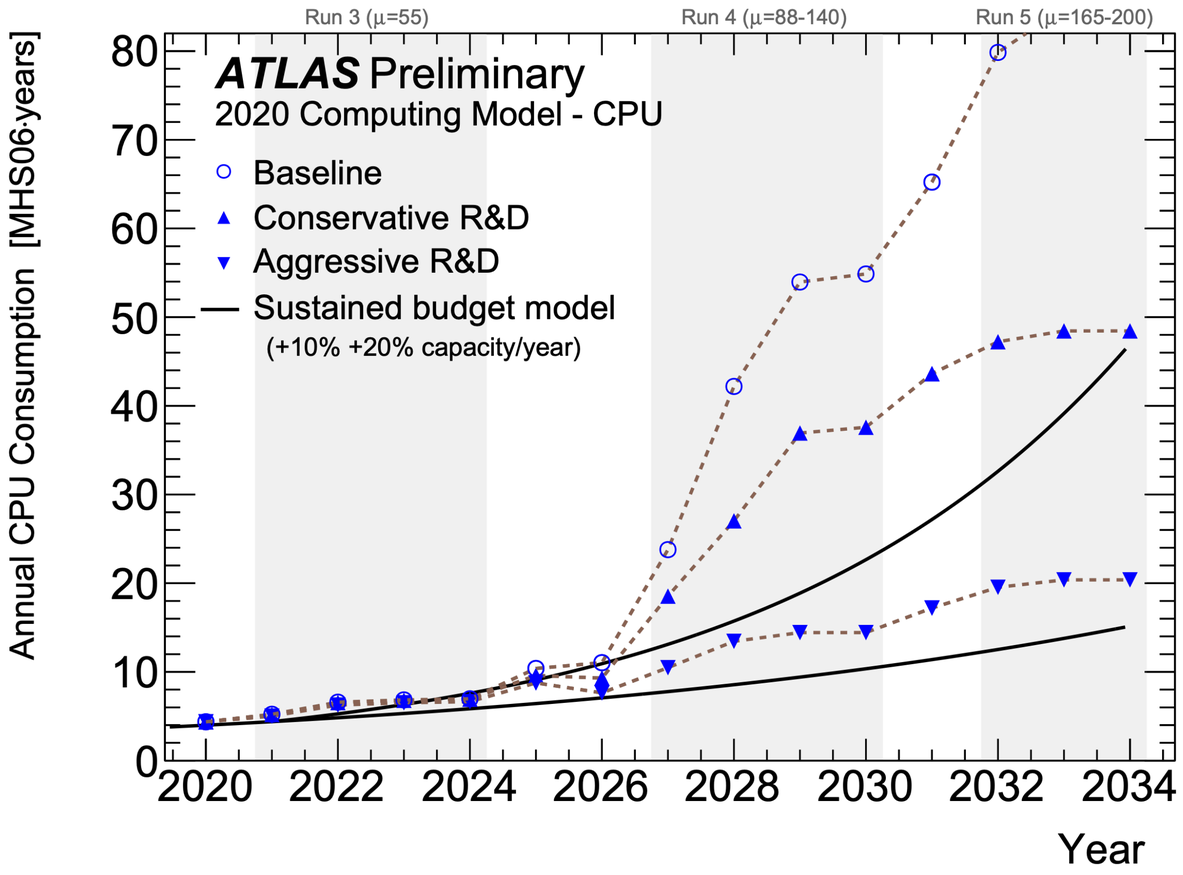
Torbjorn Sjostrand (Lund U., Dept. Theor. Phys.), Stephen Mrenna (Fermilab), Peter Z. Skands (Fermilab) (Mar, 2006)

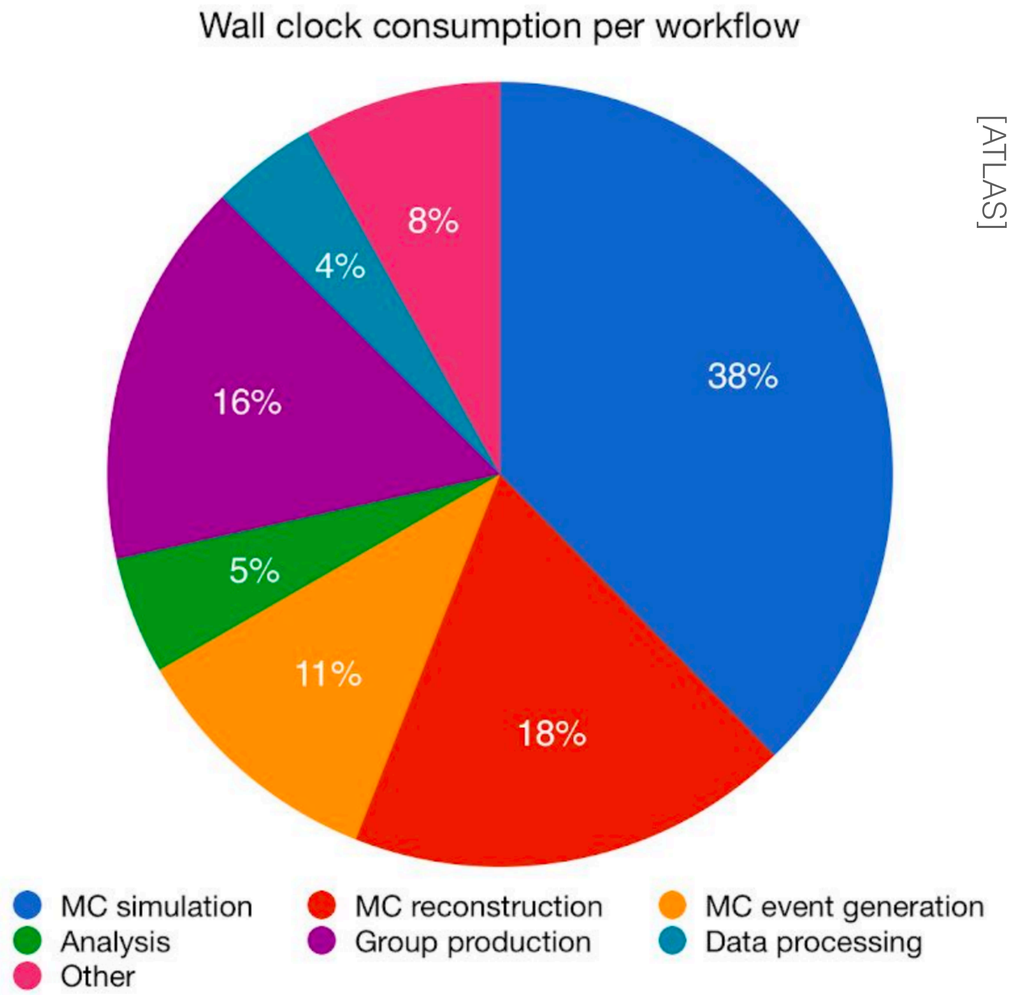
Published in: *JHEP* 05 (2006) 026 • e-Print: hep-ph/0603175 [hep-ph]

pdf links DOI cite 12,344 citations

Computational Cost

Higher accuracy ⟶ higher cost

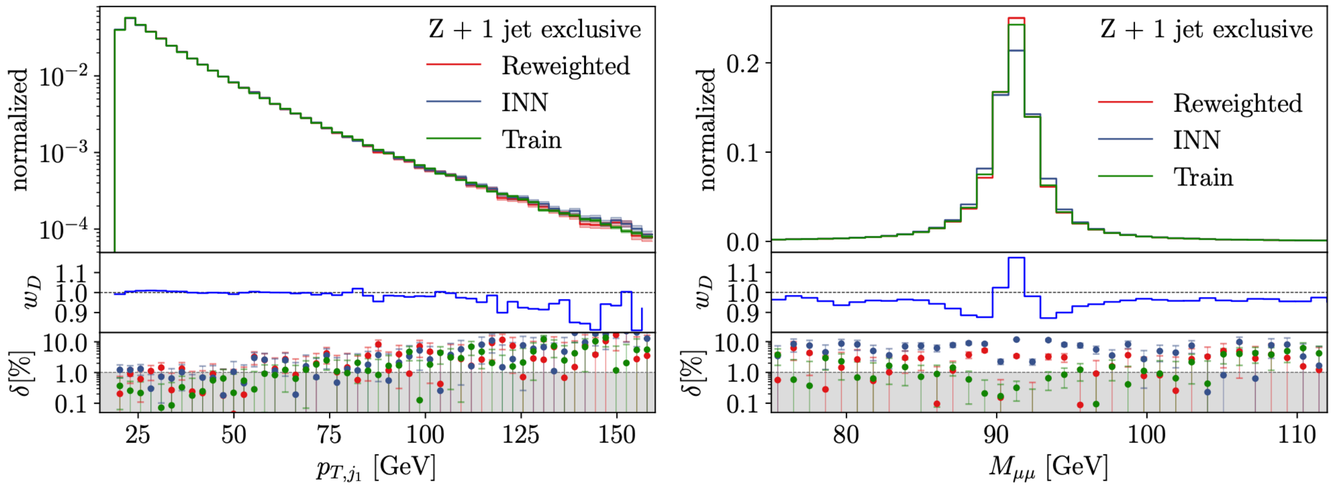




Generative Models: A way out?

Machine learning models that learn probability distributions

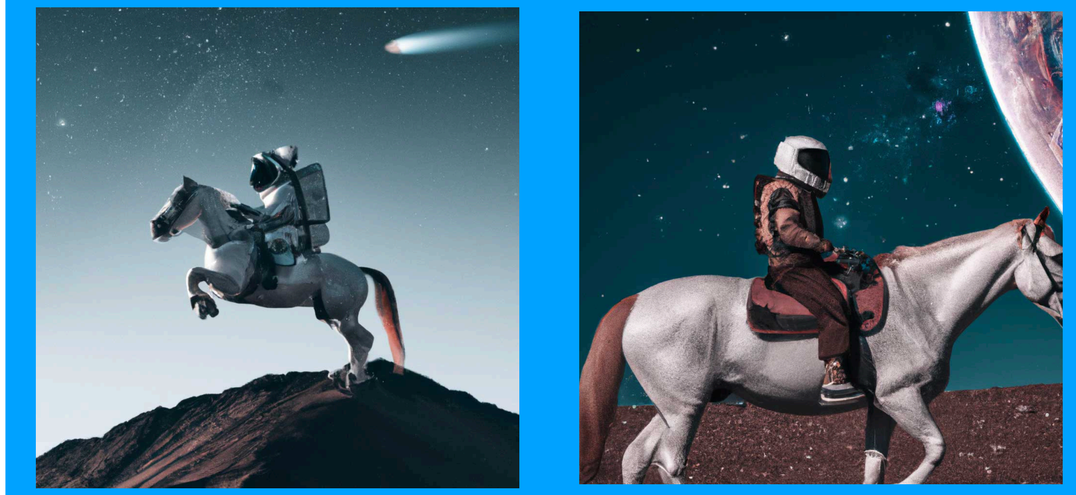
- Generative Adversarial Networks (GAN)
- Variational Autoencoders (VAE)
- Normalizing Flows (NF)
- Diffusion Models?



Butter, Heimel, Hummerich, Krebs, Plehn, Rousselot, Vent: 2110.13632

DALLE 2: conditional image generation

A bowl of soup that is a portal to another dimension as digital art

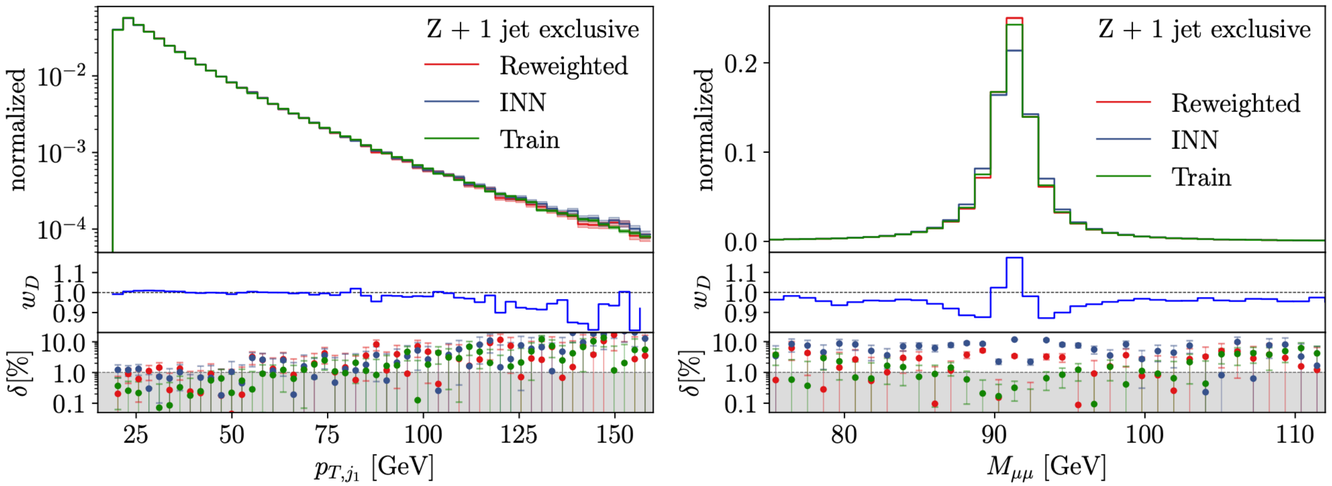


An astronaut riding a horse in a photorealistic style

Generative Models: A way out?

Machine learning models that learn probability distributions

- Generative Adversarial Networks (GAN)
- Variational Autoencoders (VAE)
- **Normalizing Flows (NF)**
- Diffusion Models?



Butter, Heimel, Hummerich, Krebs, Plehn, Rousselot, Vent: 2110.13632

DALLE 2: conditional image generation

A bowl of soup that is a portal to another dimension as digital art

An astronaut riding a horse in a photorealistic style

Normalizing Flows

Generative models with exact likelihood evaluation

Also useful in other areas:

- Anomaly detection

  Caron, Hendriks, RV: 2106.10164

  Hallin, Isaacson, Kasieczka, Krause, Nachman,
  Quadfasel, Schlaffer, Shih, Sommerhalder: 2109.00546.

  Nachman, Shih: 2001.04990

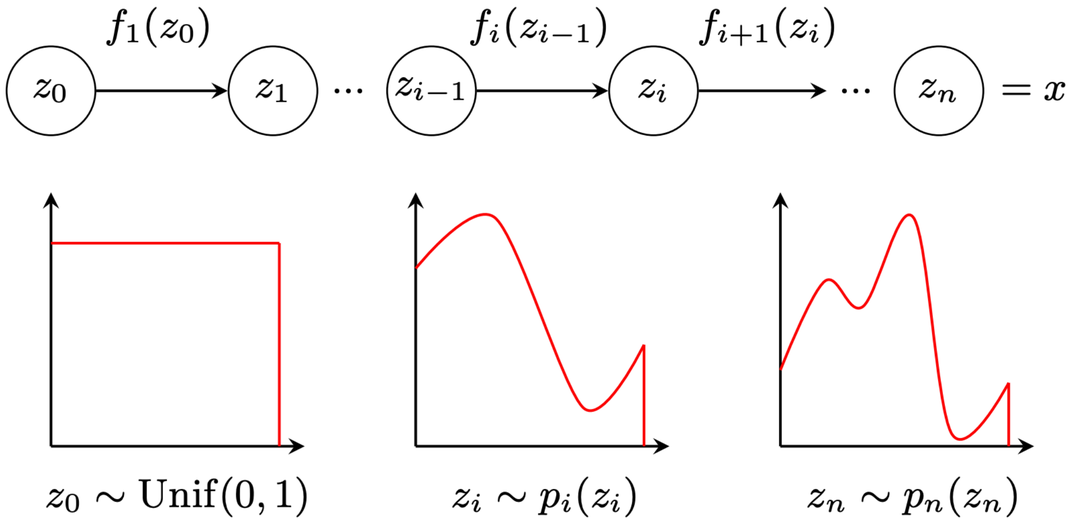
  Buss, Dillon, Finke, Krämer, Morandini, Mück, Oleksiyuk, Plehn: 2202.00686

- Likelihood-free inference

  Bieringer, Butter, Heimel, Höche, Köthe, Plehn, Radev: 2012.09873

  Vandegar, Kagan, Wehenkel, Louppe: 2011.05836

  Shirobokov, Belavin, Kagan, Ustyuzhanin, Baydin: 2002.04632.



This talk: Improve NF flexibility → better match with particle physics events

Outline

- Latent variable models
  1. Normalizing flows
  2. Variational autoencoders
  3. Surjective normalizing flows
- Applications in particle physics
  1. Permutation invariance
  2. Varying dimensionality
  3. Discrete features
- Anomaly detection

Probabilistic Models

What do I mean by a model?

Trainable parameters

$p_\theta(x)$ which enables

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

What would I like from a model

- Expressivity
- Efficiency

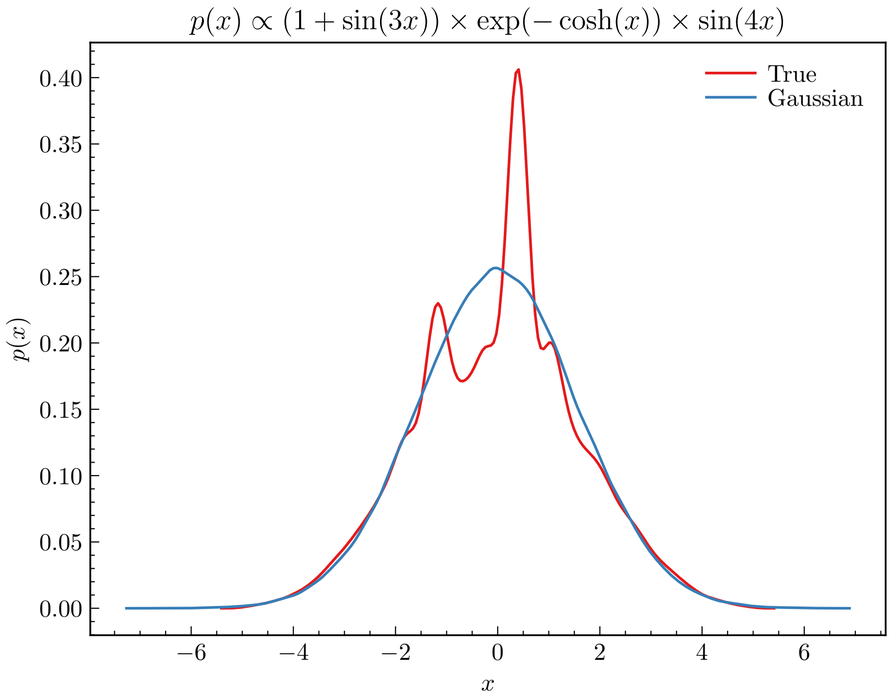
Can fit complicated $p_{\text{data}}(x)$

Fast sampling & likelihood evaluation

Gaussian distribution

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

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

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

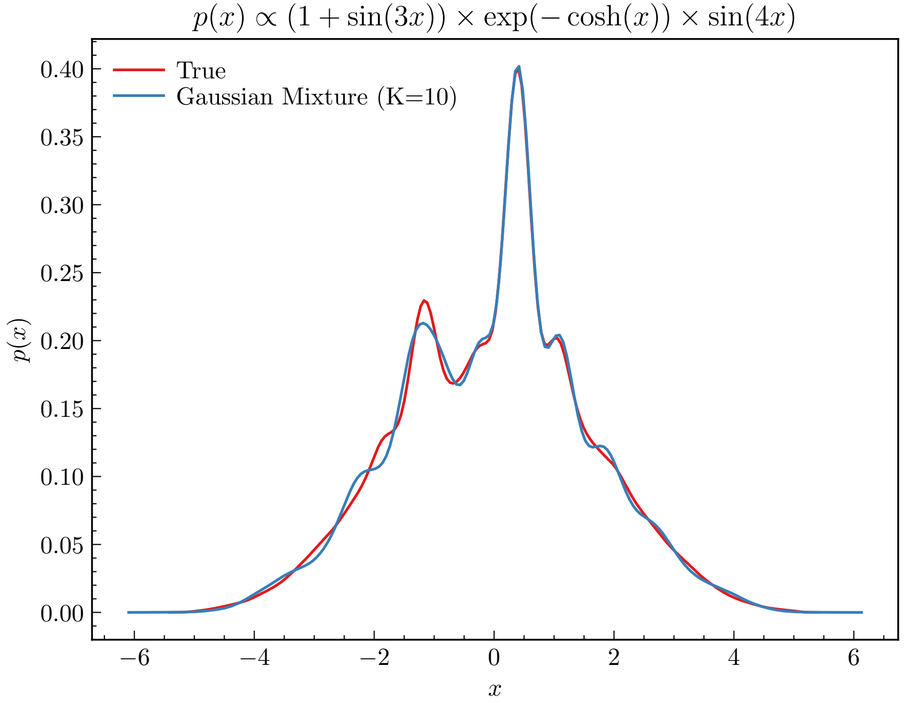
Sampling is still *simple*

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

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*

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

Too hard to compute efficiently

We *need* the likelihood

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

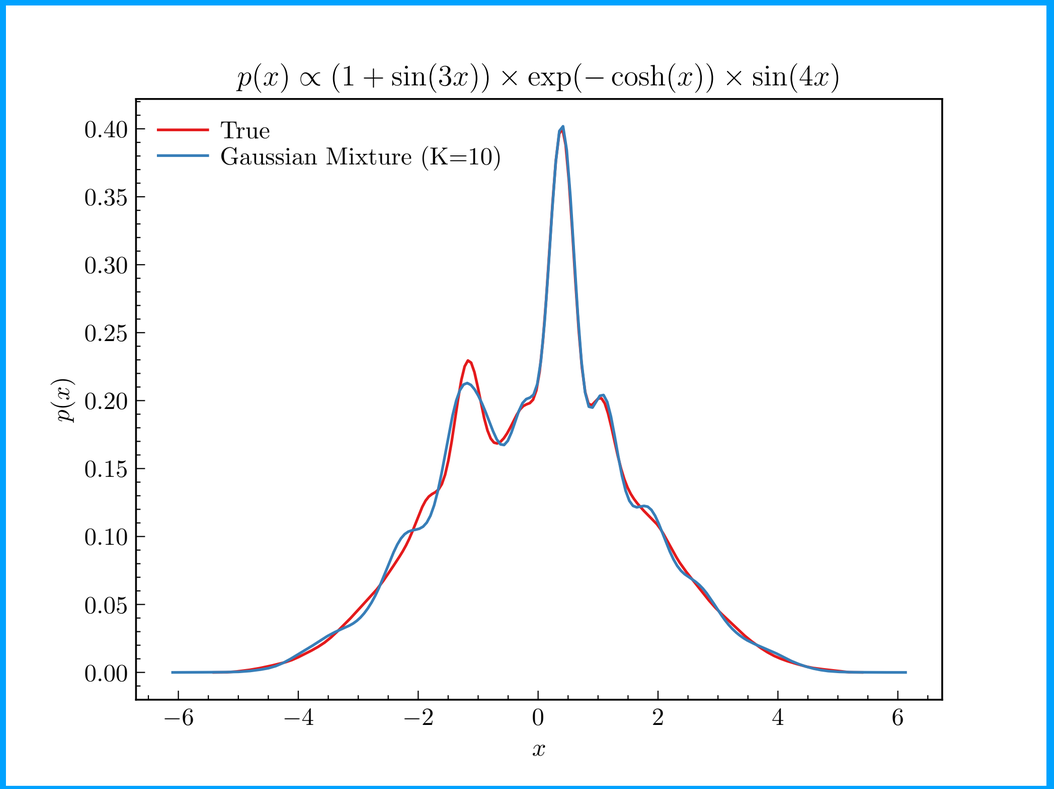
⇨ Look at two solutions:

- Normalizing flows
- Variational autoencoders

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…

Solution 1: Normalizing Flows

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

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

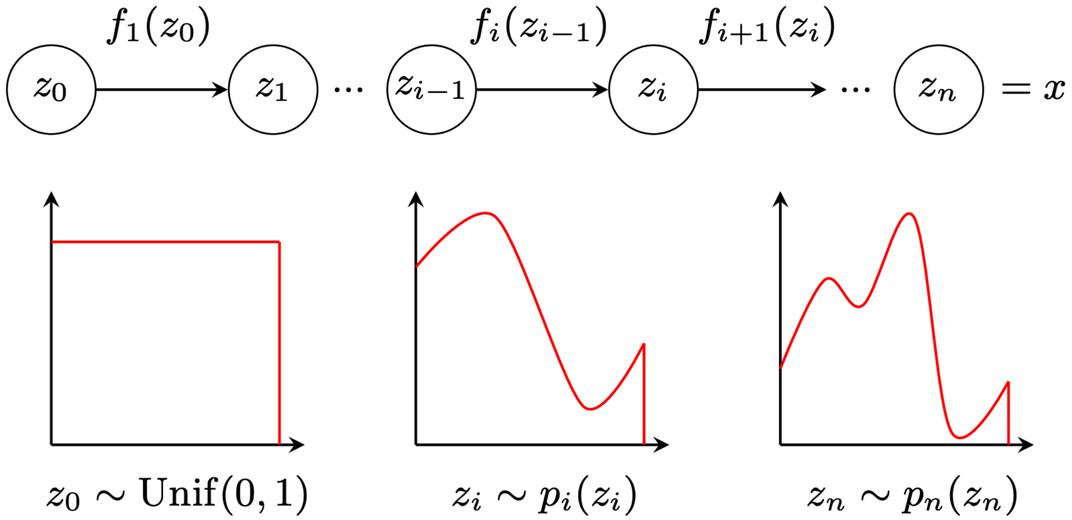
$$p(x \mid z) \to \delta(x - f(z)) \quad \Rightarrow \quad \log p(x) = \log \int_{\mathcal{Z}} dz\, p(z)\, \delta(x - f(z)) = \log p(z) + \log |J(x)|$$

Parametric bijection

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

Flow: Repeat a few times

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



Solution 2: Variational Inference

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

Introduce a second model $q(z \mid x)$

$$\log p(x) = \log p(x)$$

Solution 2: Variational Inference

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

Introduce a second model $q(z \mid x)$

$$\log p(x) = \log p(x)$$

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

$= 1$

Solution 2: Variational Inference

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

Introduce a second model $q(z \mid x)$

$$\begin{aligned}
\log p(x) &= \log p(x) \\
&= \int_{\mathcal{Z}} dz\, q(z \mid x)\, \log p(x) \\
&= \int_{\mathcal{Z}} dz\, q(z \mid x)\, \log \frac{p(x \mid z) p(z)}{p(z \mid x)} \frac{q(z \mid x)}{q(z \mid x)}
\end{aligned}$$

$= 1$

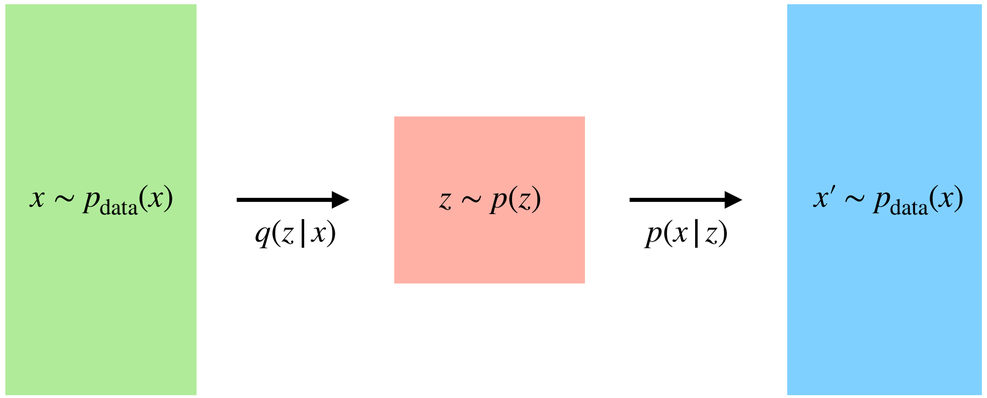
Bayes rule

Solution 2: Variational Inference

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

Introduce a second model $q(z \mid x)$

Variational Autoencoder



$$\log p(x) = \log p(x)$$

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

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

$$= \int_{\mathcal{Z}} dz\, q(z \mid x) \left[\log p(x \mid z) - \log \frac{q(z \mid x)}{p(z)} + \log \frac{q(z \mid x)}{p(z \mid x)} \right]$$

$$= \underbrace{\mathbb{E}_{q(z|x)}\left[\log p(x \mid z)\right] - \mathbb{D}_{\text{KL}}\left[q(z \mid x), p(z)\right]}_{\text{ELBO}} + \underbrace{\mathbb{D}_{\text{KL}}\left[q(z \mid x), p(z \mid x)\right]}_{\text{Bound looseness}}$$

$\Rightarrow$ Maximize the ELBO

$$\mathscr{E}(x, z)$$

Solutions

	Normalizing flow	Variational autoencoder
Pros	Exact likelihood evaluation Straightforward training Guaranteed to cover full feature space	Maximally flexible
Cons	Not very flexible Limitations due to efficiency requirement	No access to exact likelihood Loss with multiple components

The best of both worlds

Nielsen, Jaini, Hoogeboom, Winther, Welling: 2007.02731

Take the likelihood expression from the VAE and rearrange

$$p(x) = \int_{\mathcal{Z}} dz\, q(z\,|\,x) \left[\log p(z) + \log \frac{p(x\,|\,z)}{q(z\,|\,x)} + \log \frac{q(z\,|\,x)}{p(z\,|\,x)} \right]$$

If we consider a normalizing flow

$$p(x\,|\,z) = \delta(x - f(z)) \qquad q(z\,|\,x) = \delta(z - f^{-1}(x))$$

We find

$$p(x) = \int_{\mathcal{Z}} dz\, \delta(z - f^{-1}(x)) \left[\log p(z) + \log \frac{\delta(x - f(z))}{\delta(z - f^{-1}(x))} + \log \frac{\delta(z - f^{-1}(x))}{\delta(z - f^{-1}(x))} \right]$$

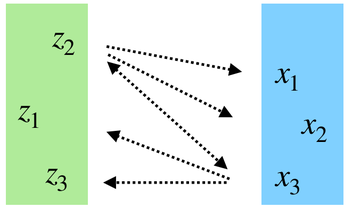
$$= \log p(z) + \log |\, J(x)\,|$$

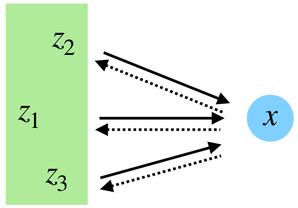
The best of both worlds

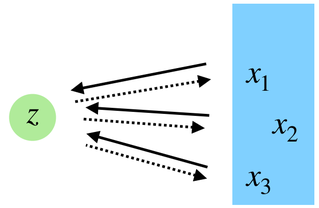
Nielsen, Jaini, Hoogeboom, Winther, Welling: 2007.02731

Describes both models (and more)!

$$p(x) = \int_{\mathcal{Z}} dz\, q(z\,|\,x) \left[\log p(z) + \log \frac{p(x\,|\,z)}{q(z\,|\,x)} + \underbrace{\log \frac{q(z\,|\,x)}{p(z\,|\,x)}}_{\mathcal{E}(x,y)} \right]$$







	NF	VAE	Surjective NF (forward)	Surjective NF (backward)
$p(x\,\|\,z)$	Deterministic	Stochastic	Deterministic	Stochastic
$q(z\,\|\,x)$	Deterministic	Stochastic	Stochastic	Deterministic
$\mathcal{E}(x,z)$	$= 0$	> 0	> 0	$= 0$

Application in Particle Physics

A Benchmark Process

Model the distribution of matrix element-level

$$gg \to \tilde{g}\tilde{g}\tilde{g}\tilde{g}$$

$\sqrt{s} = 3$ TeV

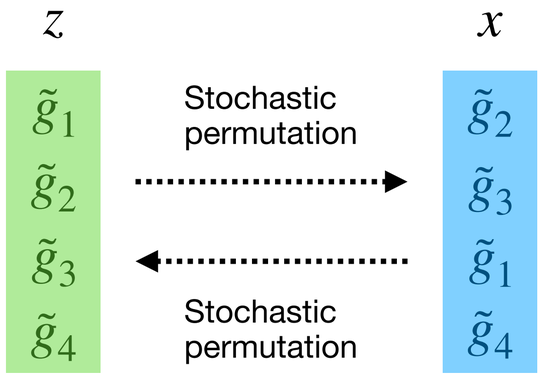
$m_{\tilde{g}} \approx 600$ GeV

- Four-fold permutation symmetry
  - 8-dim phase space neatly divides into 4x2
- Rich discrete structure
  - 6 partons in adjoint rep ⟶ many colour configurations
  - masses ⟶ mass-suppressed helicity configurations

Permutation Invariance

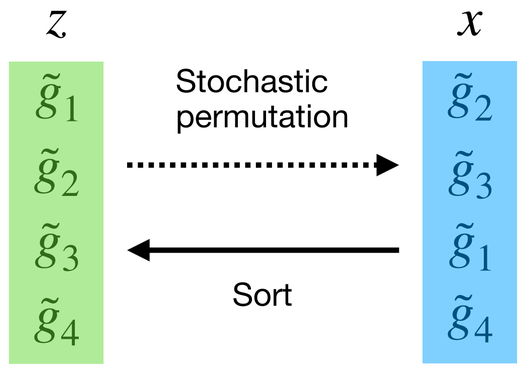
Two ways of encoding permutation invariance into flow

Stochastic permutation

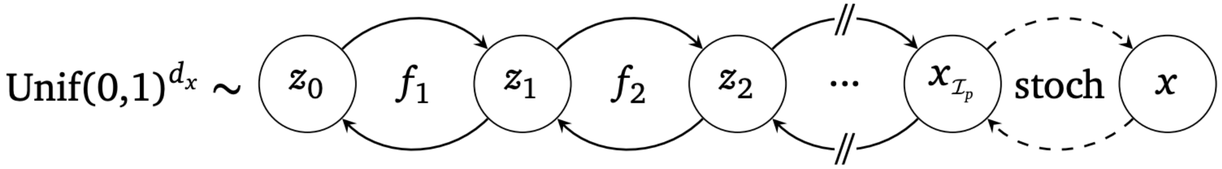


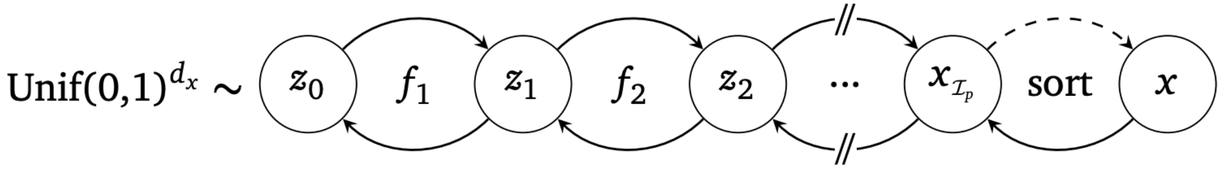
$$\mathscr{E}(x, z) > 0$$

Sort surjection



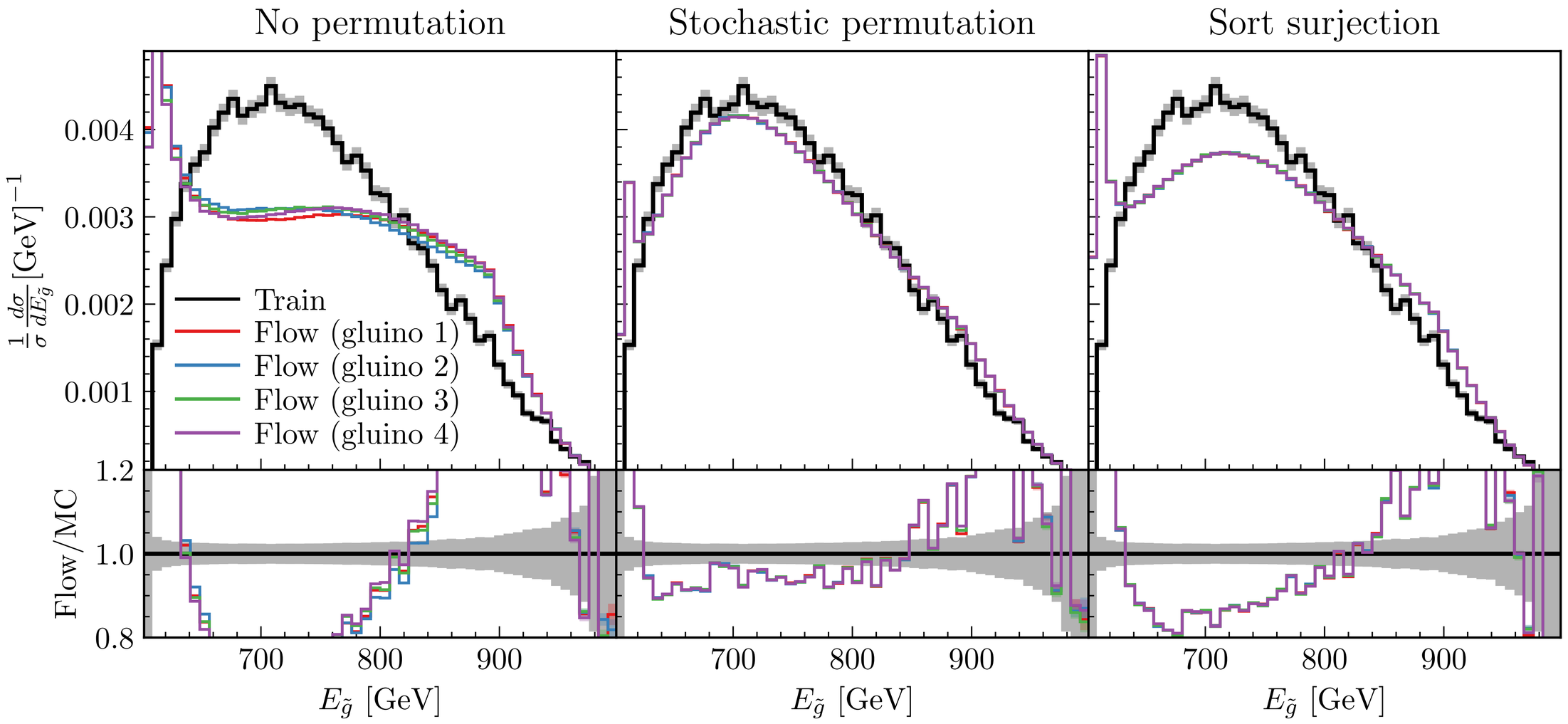
$$\mathscr{E}(x, z) = 0$$



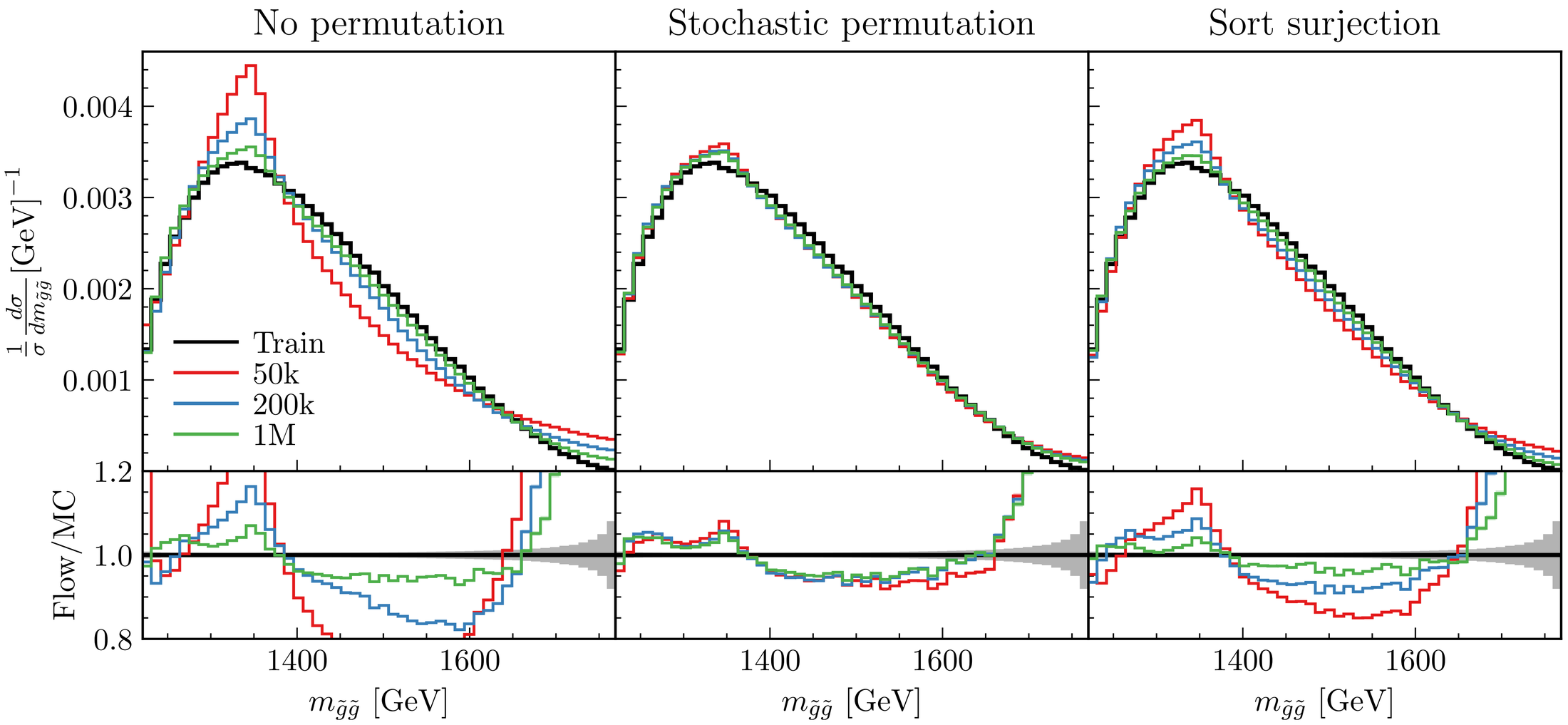


Permutation Invariance: Experiments with Small Stats

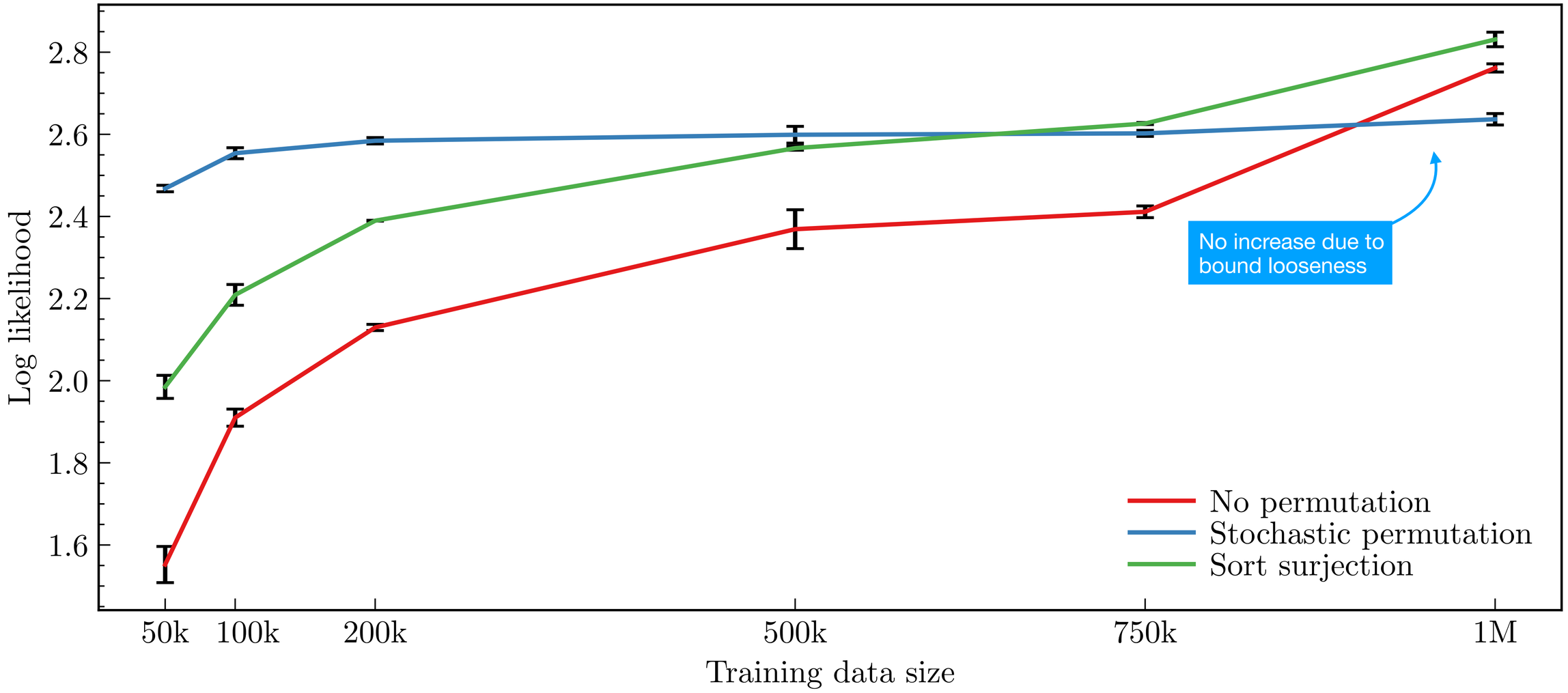
50k training events



Permutation Invariance: Experiments with Varying Stats



Permutation Invariance: Test Log Likelihood

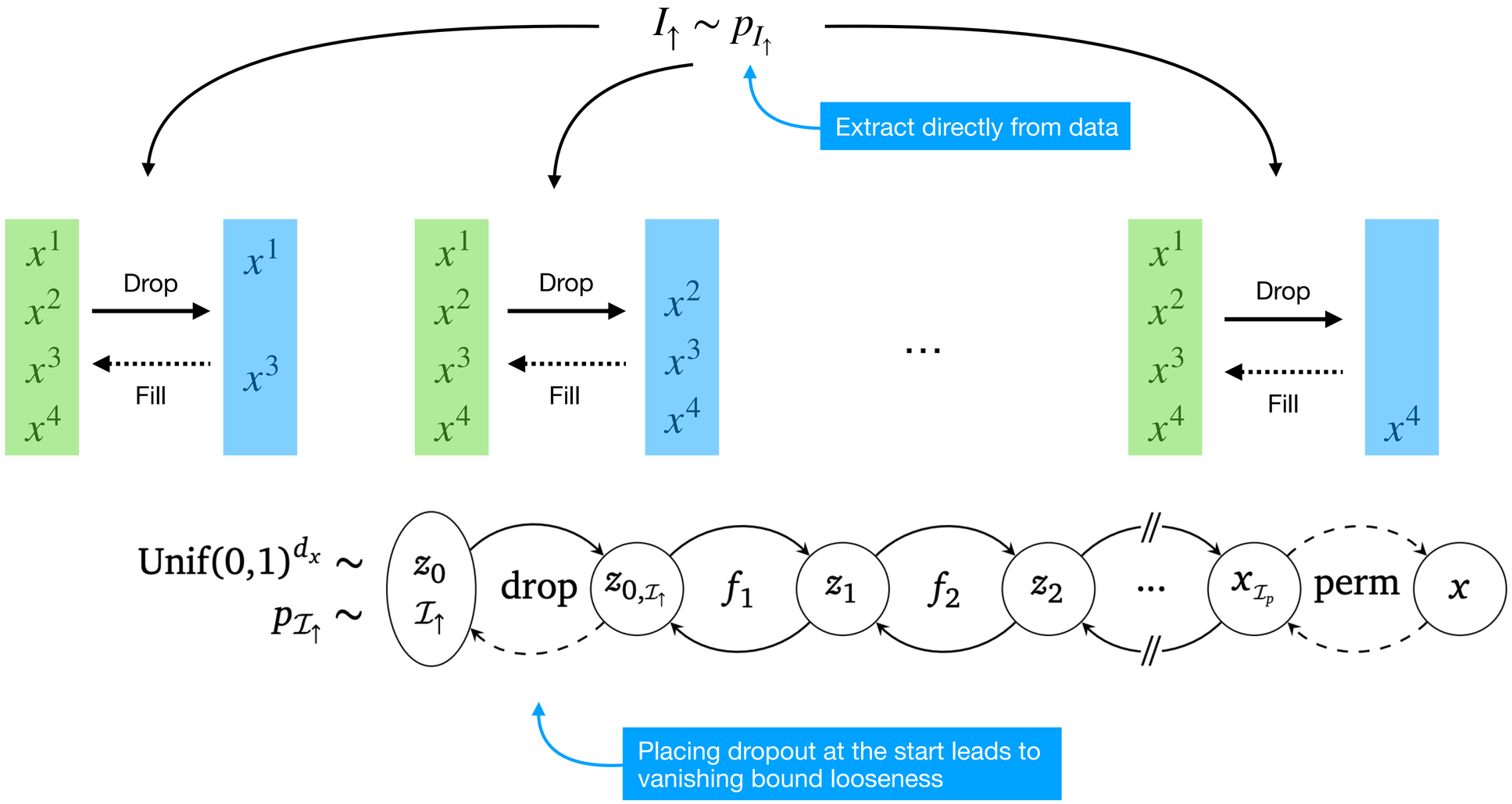


Varying Dimensionality

One solution: multiple conditional flows

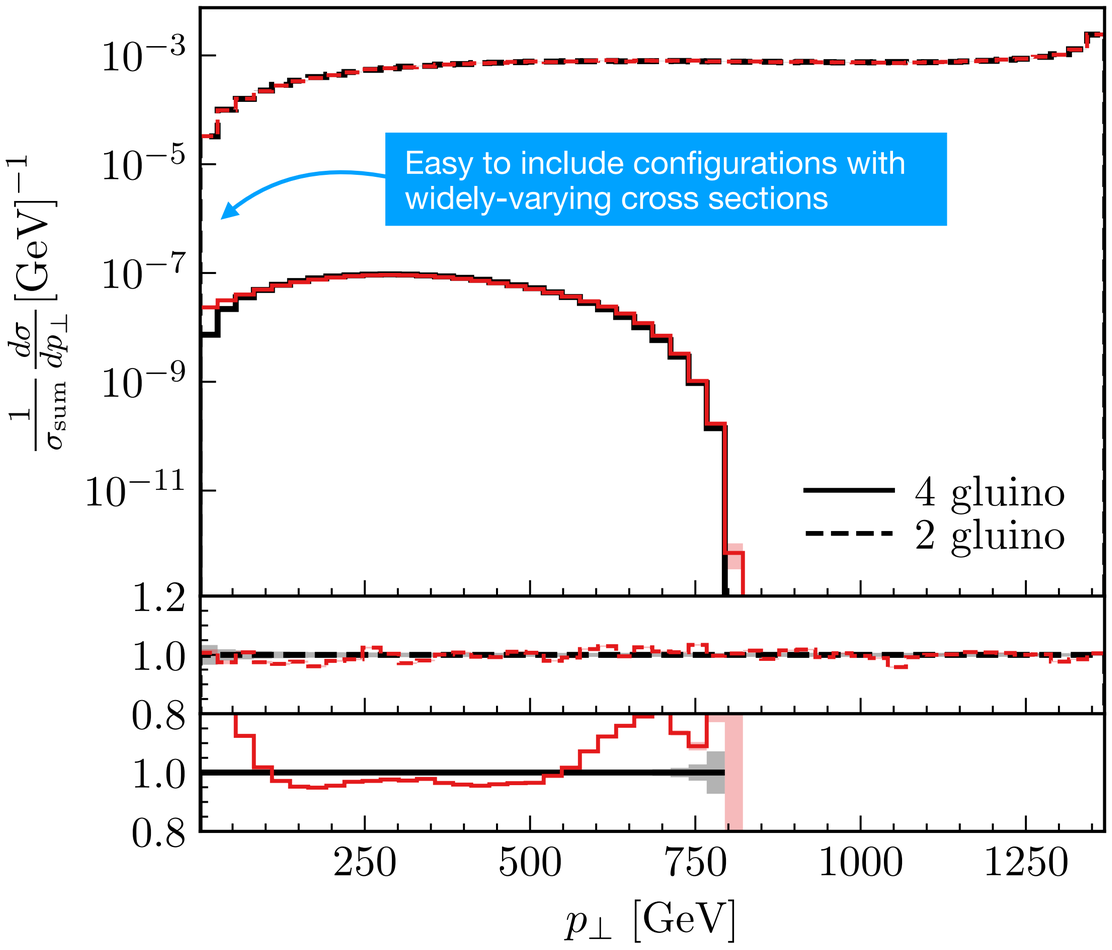
$Z + 1/2/3\,j$ Butter, Heimel, Hummerich, Krebs, Plehn, Rousselot, Vent: 2110.13632

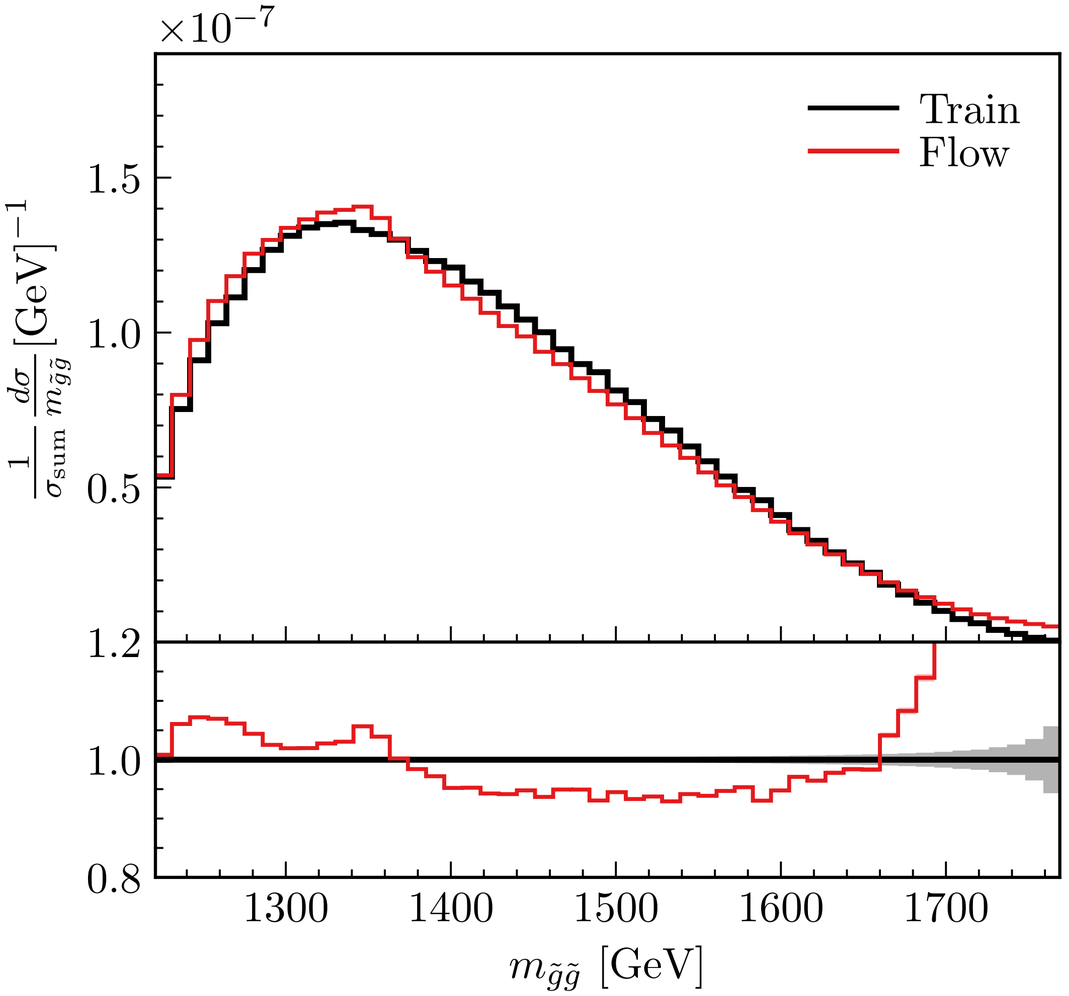
Dropout surjection



Varying Dimensionality: Experiments

Mix $gg \to \tilde{g}\tilde{g}\tilde{g}\tilde{g}$ and $gg \to \tilde{g}\tilde{g}$ events



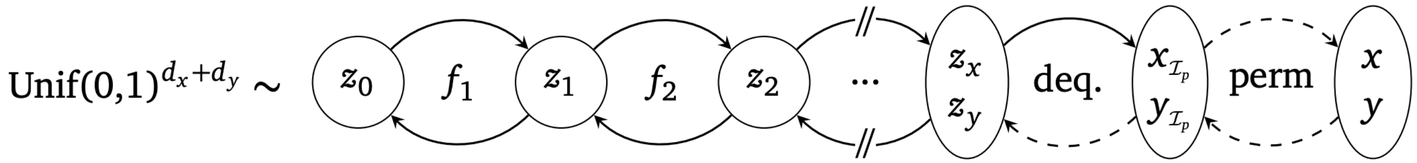


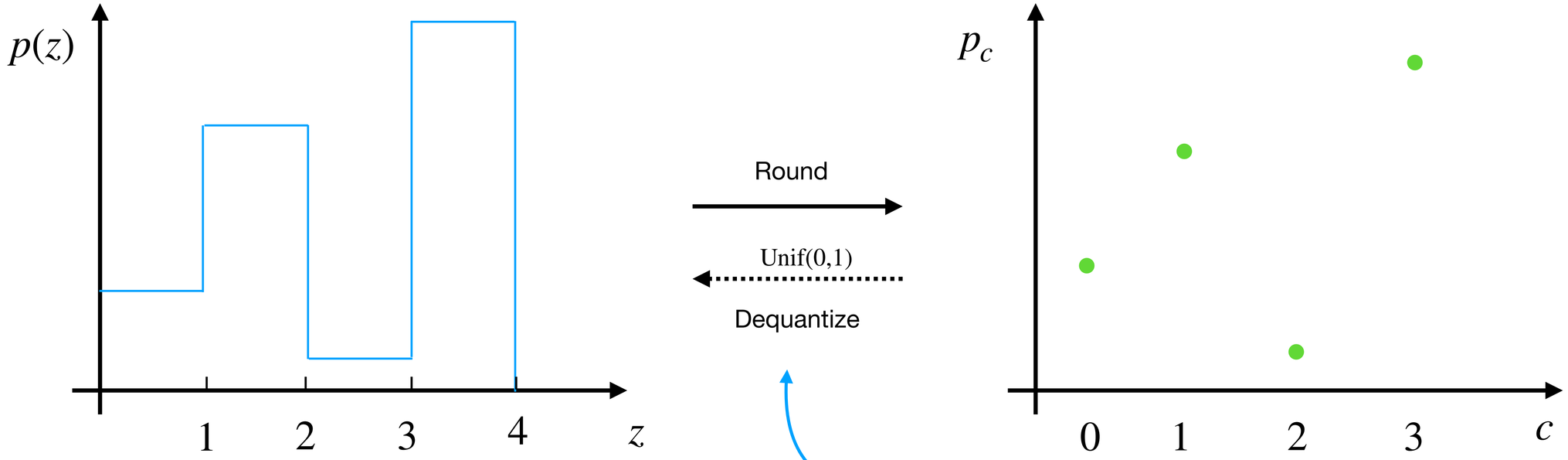
Discrete Features: Uniform Dequantization

Unique situation in particle physics

⇨ Events have mixed continuous-discrete features

Phase space &
- particle id
- helicity
- colour
- …



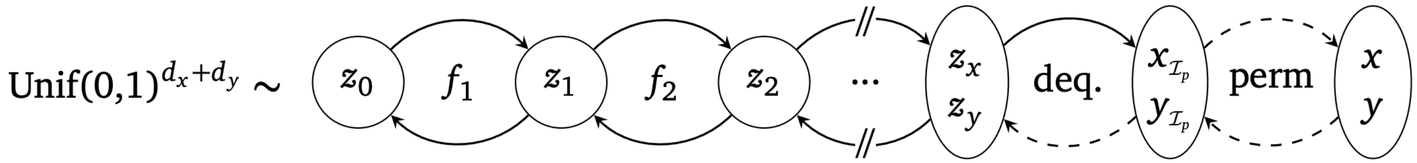


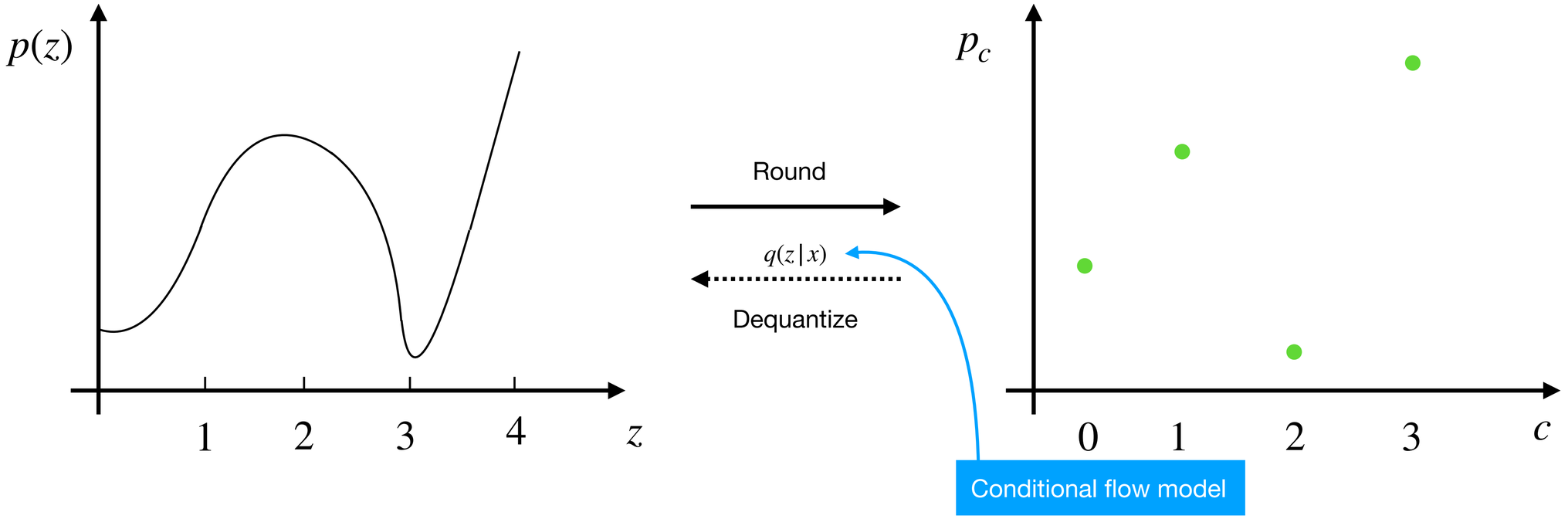
Discrete Features: Variational Dequantization

Step-like distribution might be hard to learn

⇨ Jointly train dequantization model

Larochelle, Murray, Uria: 1306.0186



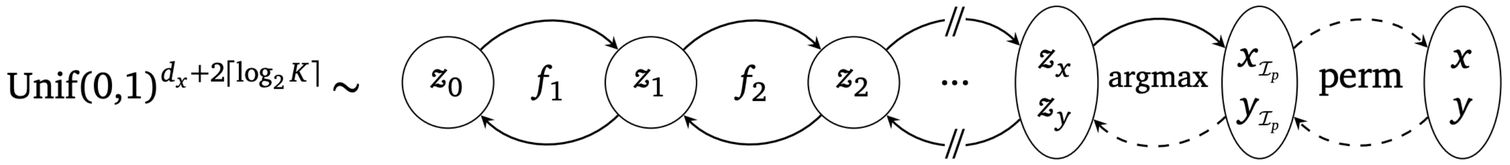


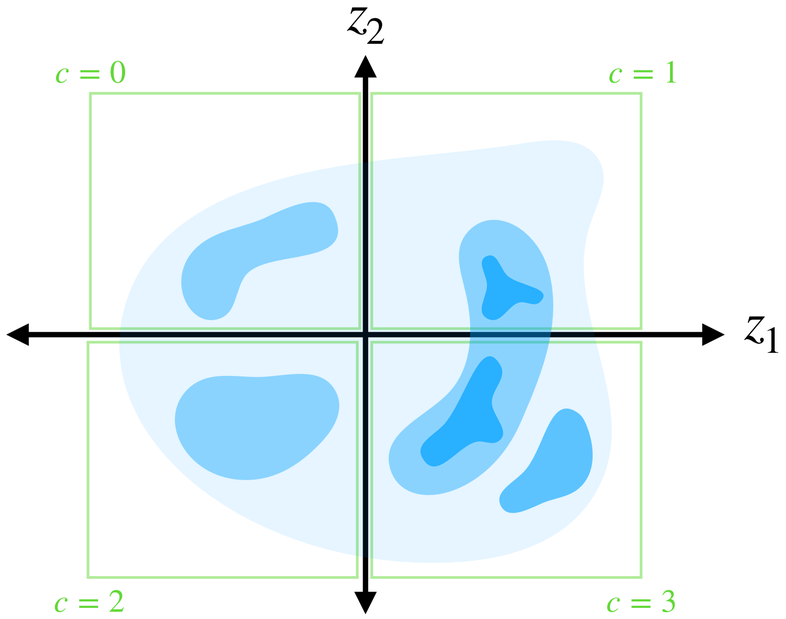
Discrete Features: Argmax Surjection

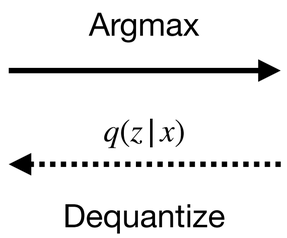
Hoogeboom, Nielsen, Jaini, Forré , Welling: 2102.05379

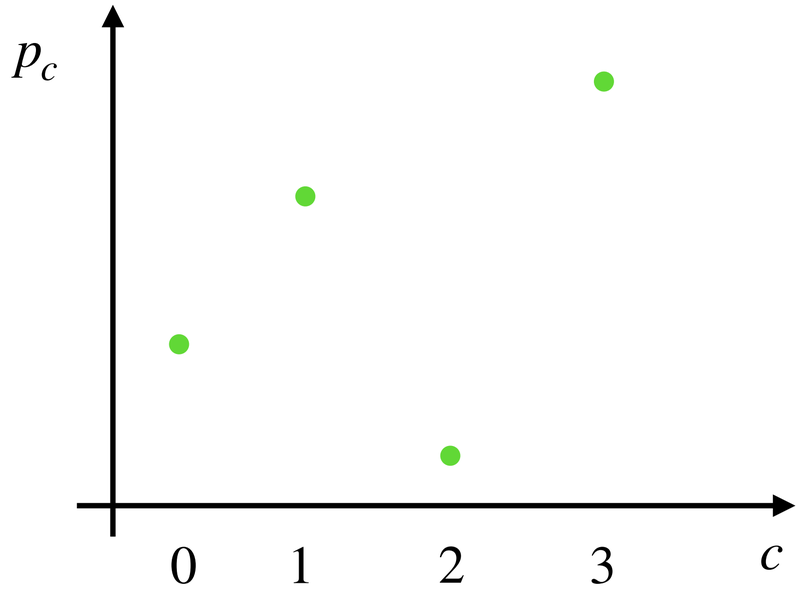
Dequantization makes sense for ordinal data

⇨ Discrete features are fundamentally categorical









Discrete Features: Factorized Models

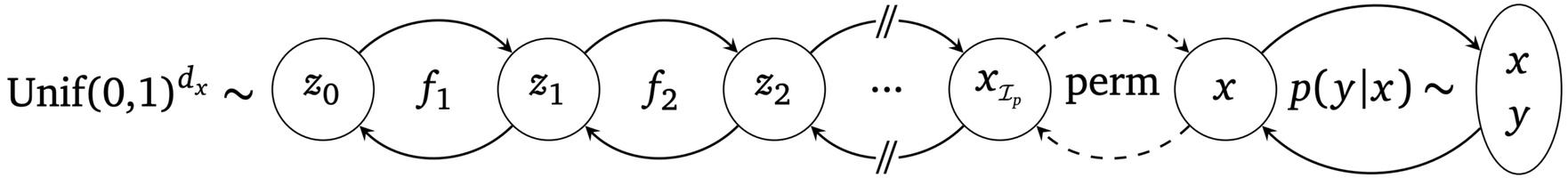
Exploit the mixed continuous-discrete features

1. Classifier

$$p(x, y) = p(x)\, p(y \mid x)$$

Classifier

Regular flow model

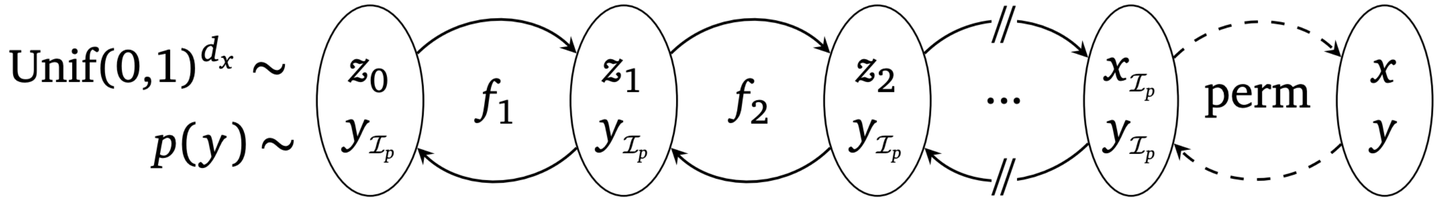


2. Mixture

$$p(x, y) = p(y)\, p(x \mid y)$$

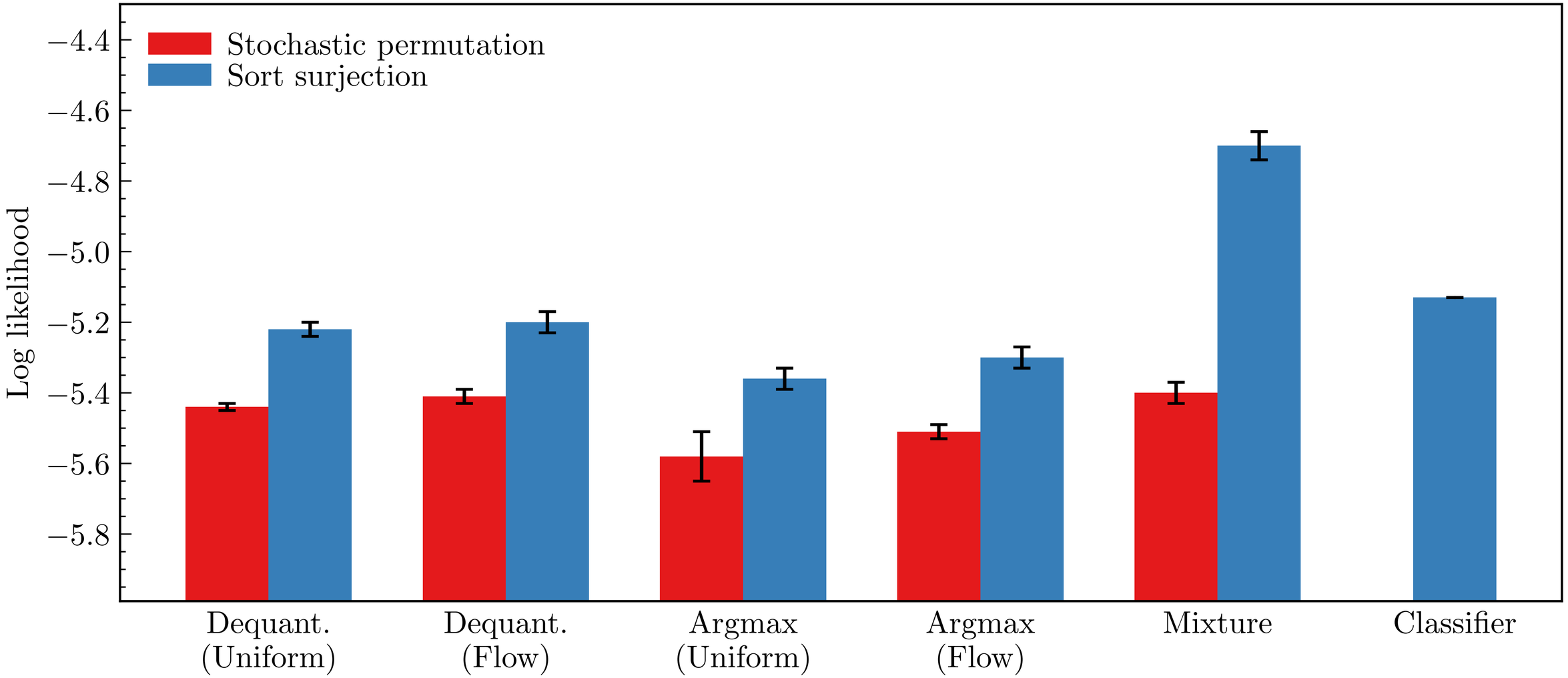
Conditional flow model

Discrete model



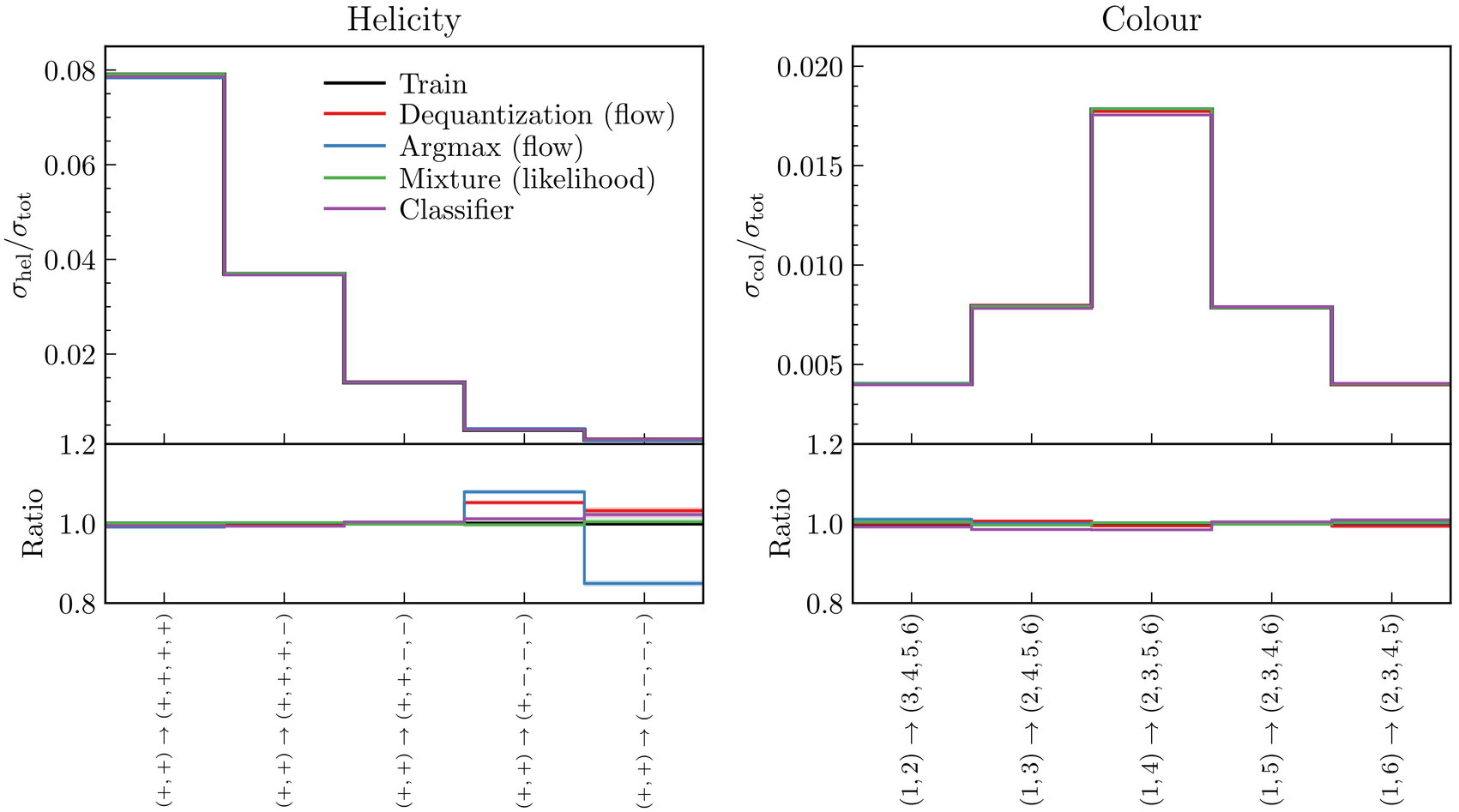
Discrete Features: Experiments

120 colour flows x 64 helicity configurations = 7680 categories



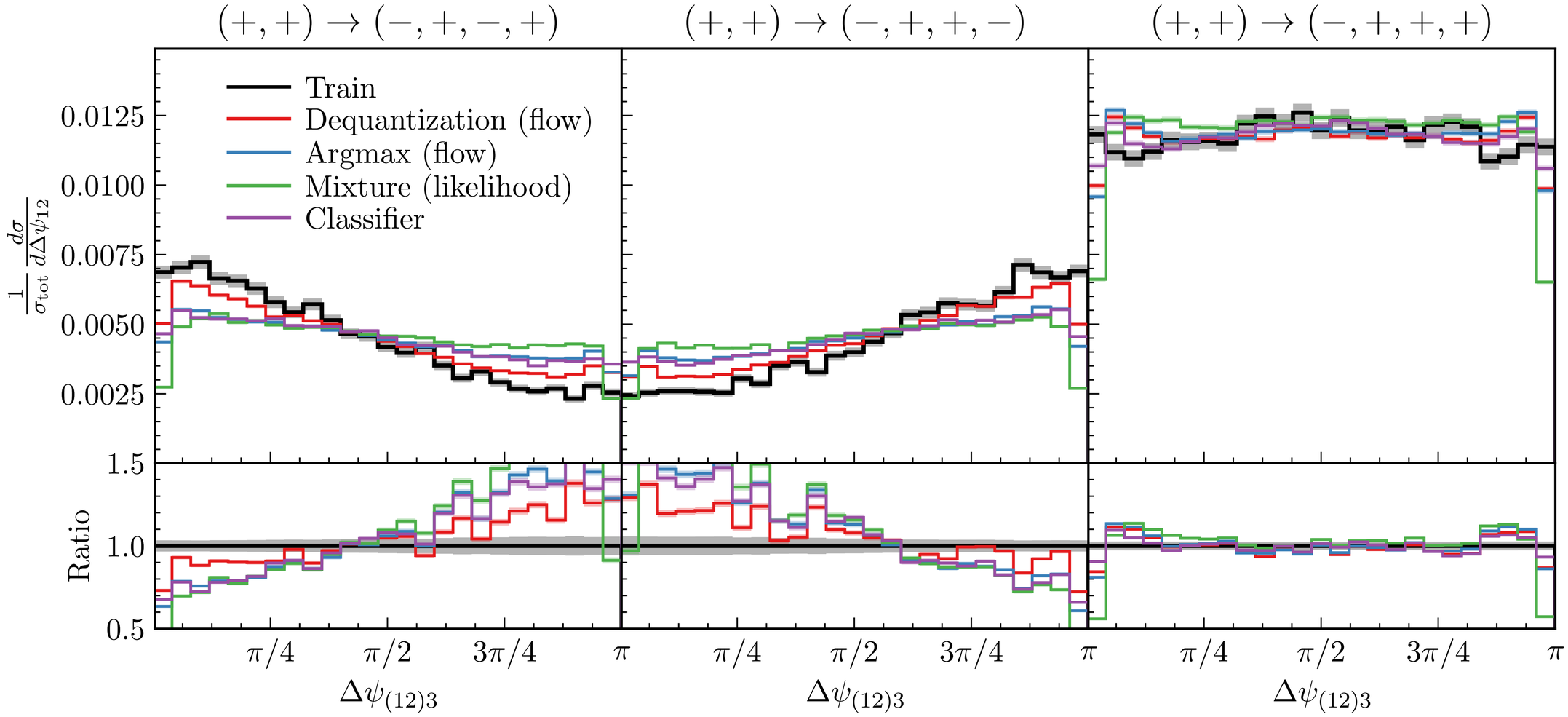
Discrete Features: Experiments

Discrete marginalised distributions



Anomaly Detection

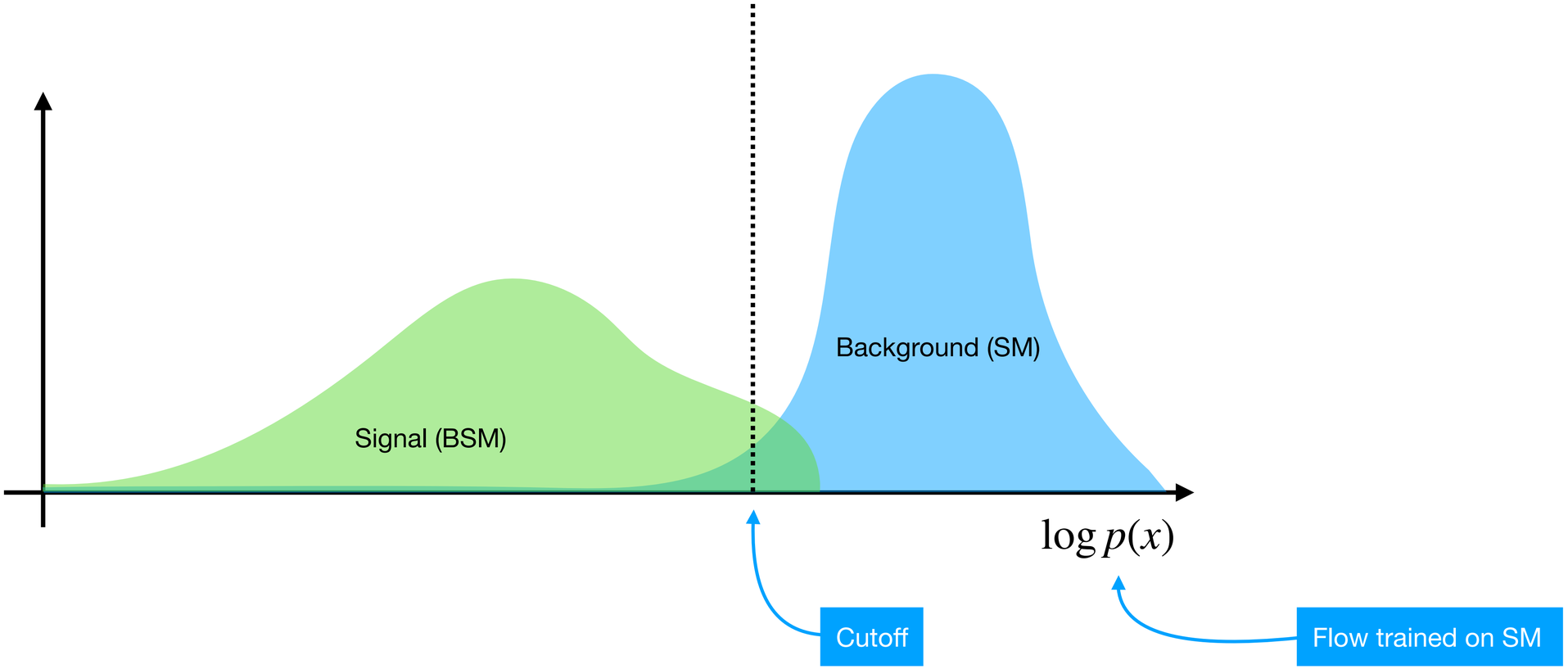
Conditional continuous distributions



Anomaly Detection

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
   - Channel 2a: At least three identified leptons
   - Channel 2b: At least two identified leptons
   - Inclusive with moderate missing energy

2. Validation set:

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

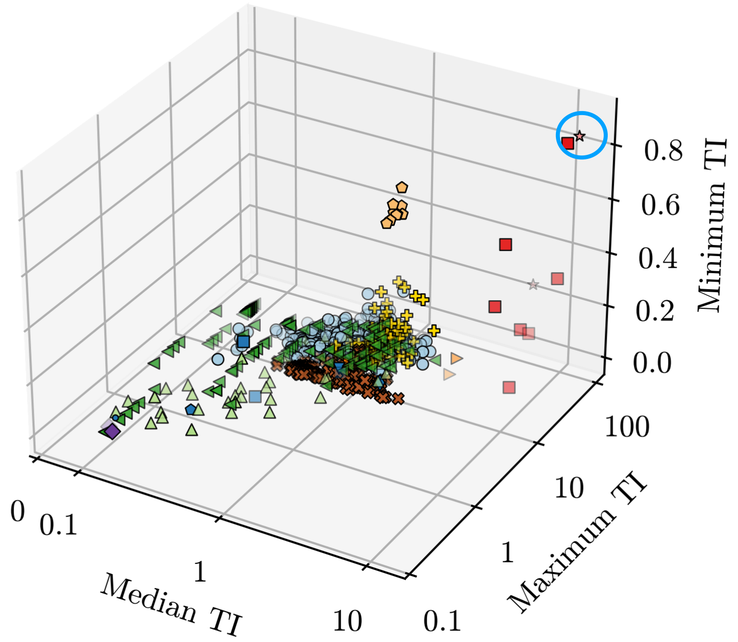
3. Test set:

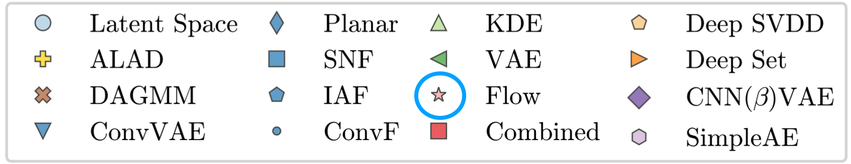
   Secret dataset with labels not known to model authors

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[c] 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[a,a] Z. Zhang[d]

Anomaly Detection