Representation Learning of Collider Events

Jack Collins

SLAC NATIONAL ACCELERATOR LABORATORY

ML4Jets 2020

How Much Information is in a Jet?

Kaustuv Datta and Andrew Larkoski

Physics Department, Reed College, Portland, OR 97202, USA

How Much Information is in a Jet?

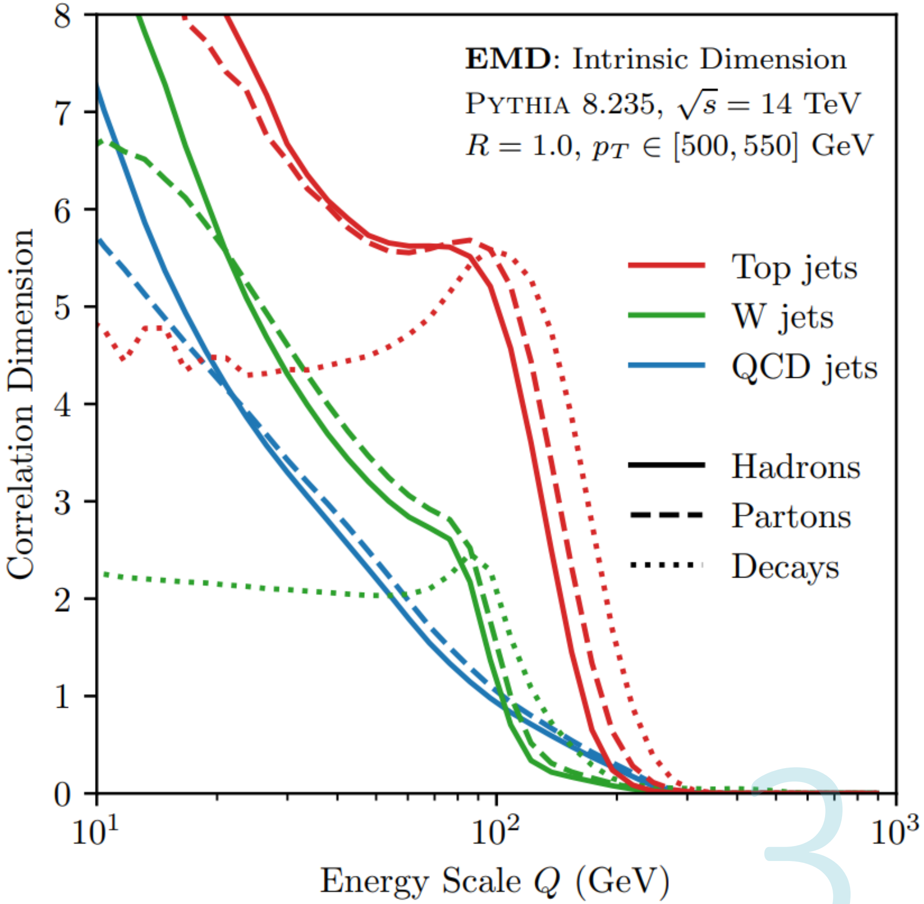
Kaustuv Datta and Andrew Larkoski

Physics Department, Reed College, Portland, OR 97202, USA

The Metric Space of Collider Events

Patrick T. Komiske,* Eric M. Metodiev,† and Jesse Thaler‡

Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and Department of Physics, Harvard University, Cambridge, MA 02138, USA

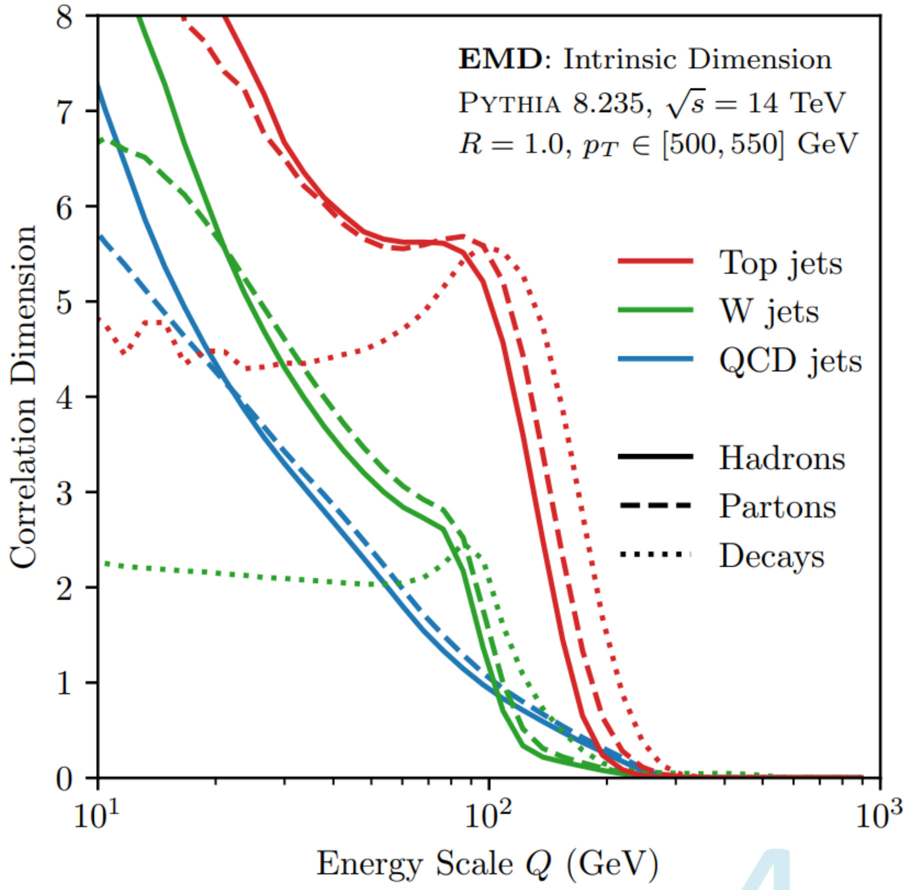


Conclusions

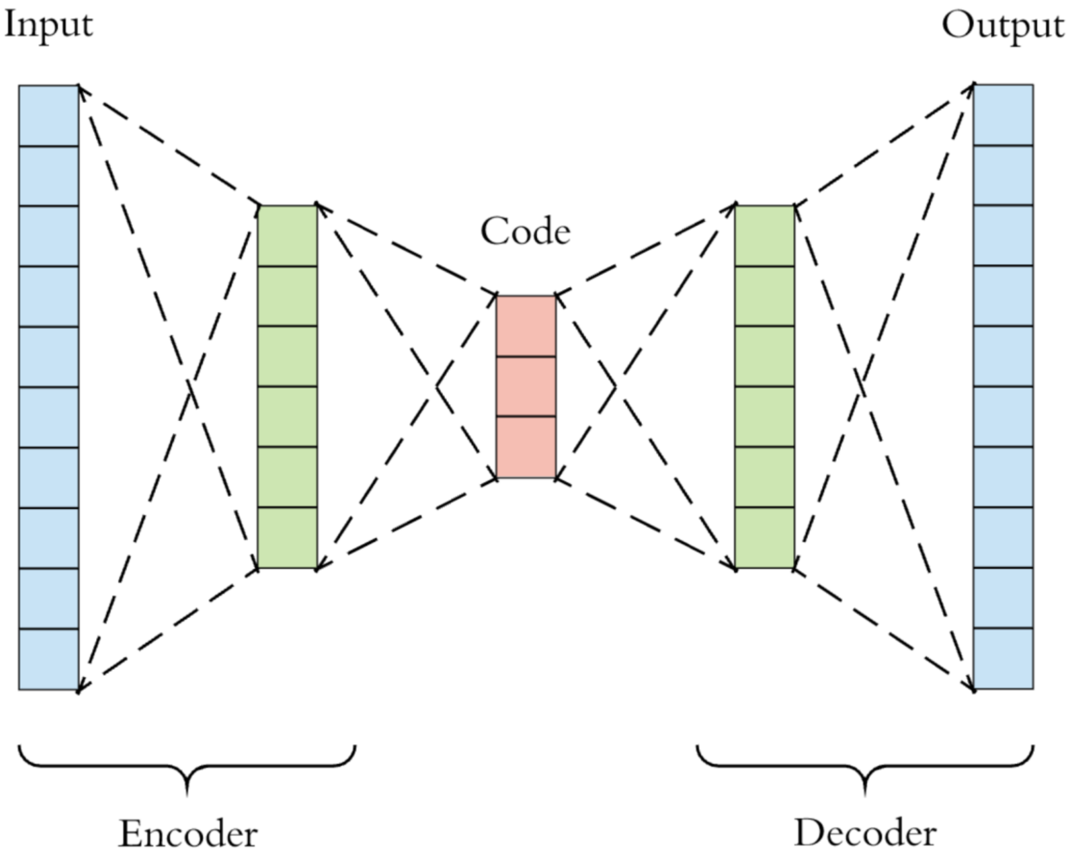
- I have been training Variational Autoencoders to reconstruct jets or collider events using Earth Movers Distance as the reconstruction metric.

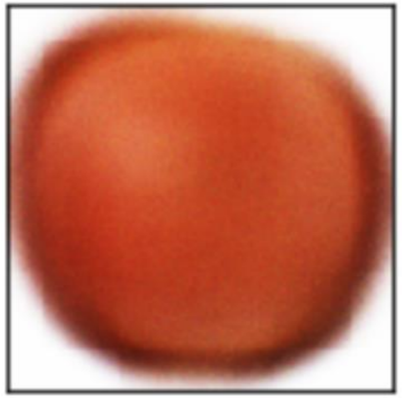
- The learnt representation:
  1. *Is scale dependent*
  2. *Is orthogonalized*
  3. *Is hierarchically organized by scale*
  4. *Has fractal dimension which relates to that of the data manifold*

- This is because:
  1. *The VAE is trained to be parsimonious with information*
  2. *The metric space is physically meaningful and structured*



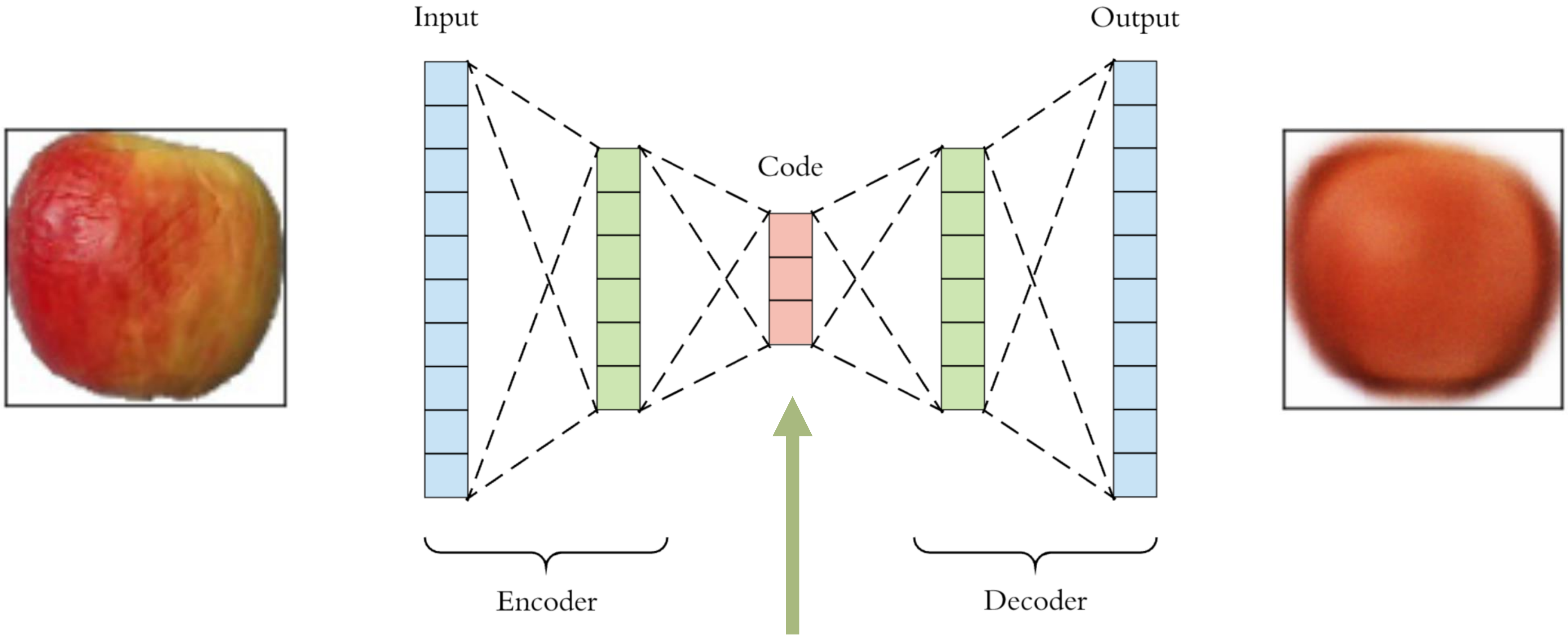
The Plain Autoencoder





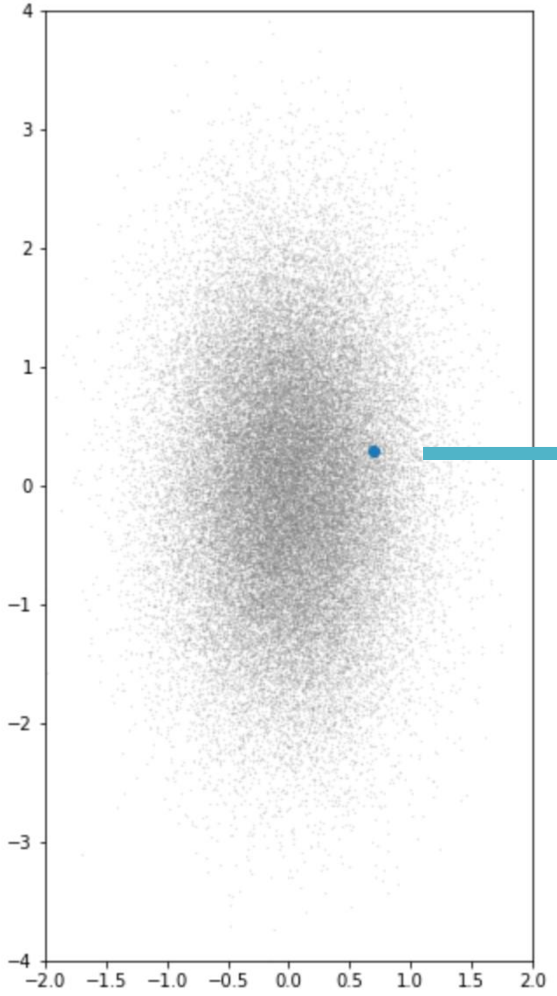
Loss = |Output – Input| *(what is this for jets?)*

The Plain Autoencoder



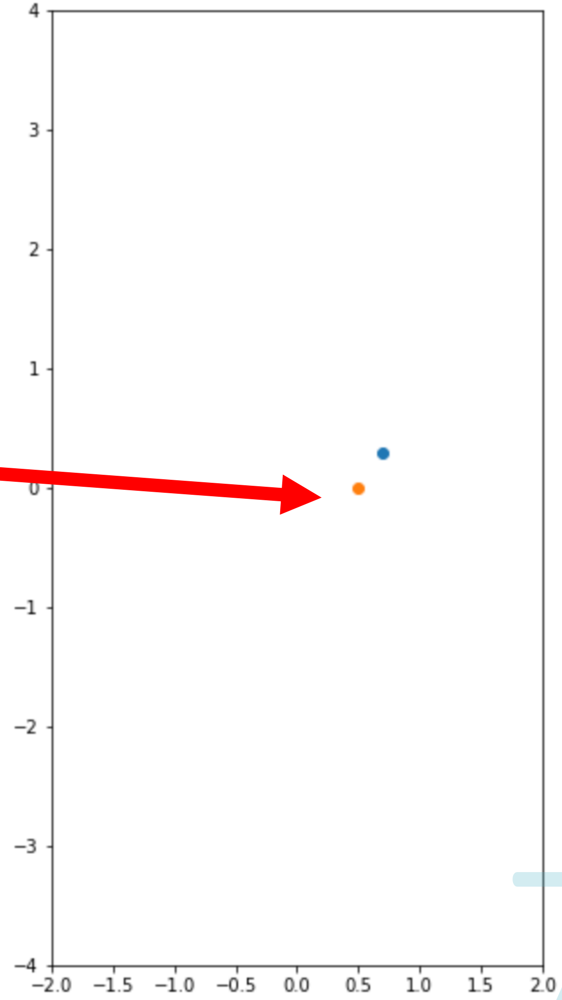
Latent space =?= Learnt representation

The Plain Autoencoder: *a toy example*



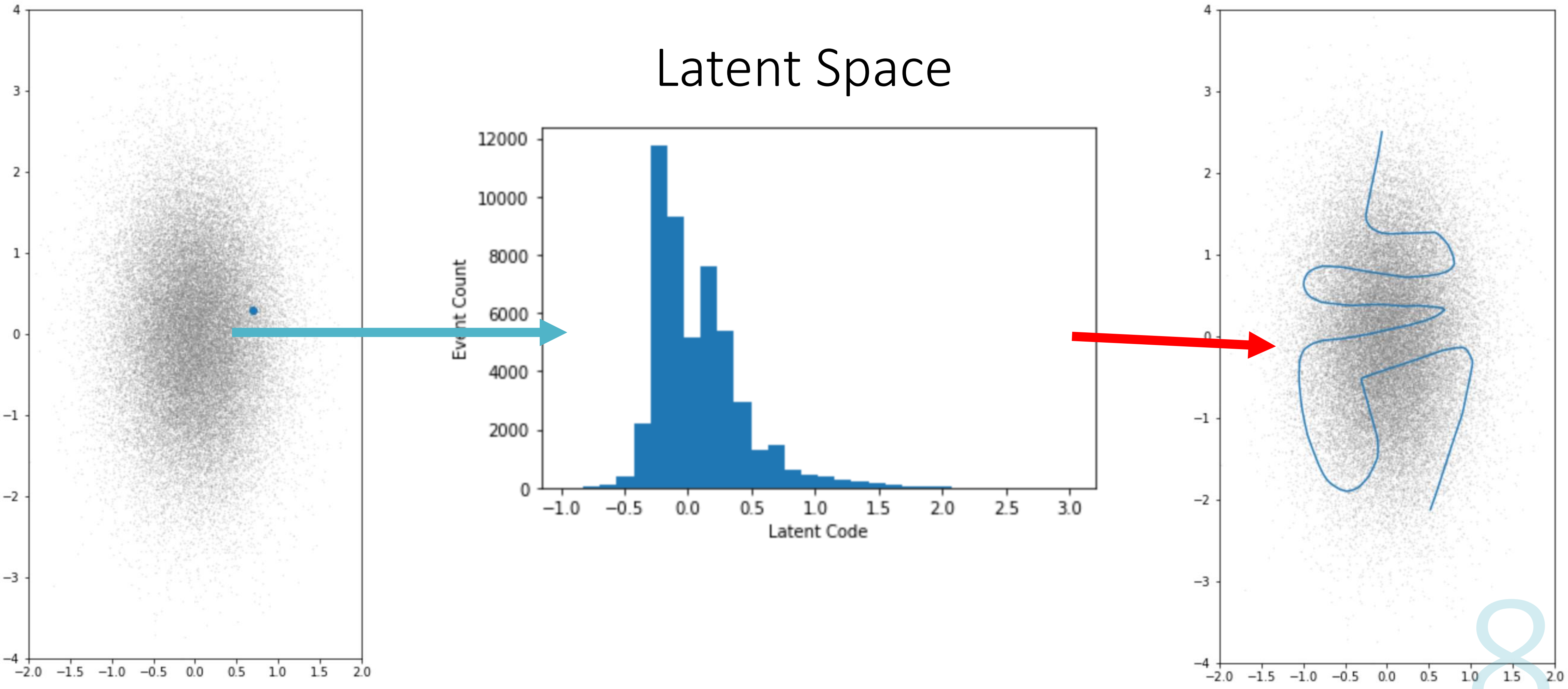
AE

(1D latent space)



Rec. Loss = $|\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^{(2)}$

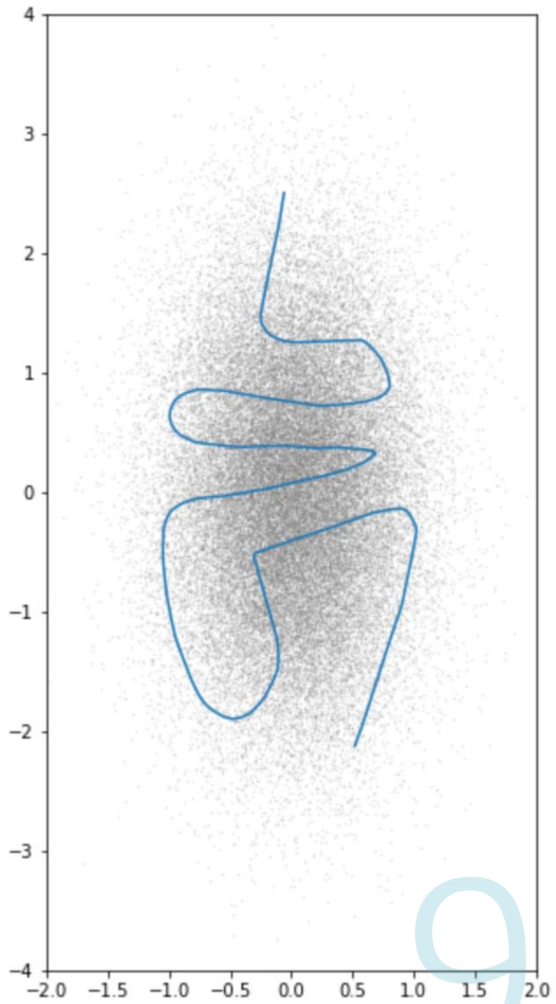
The Plain Autoencoder: *a toy example*



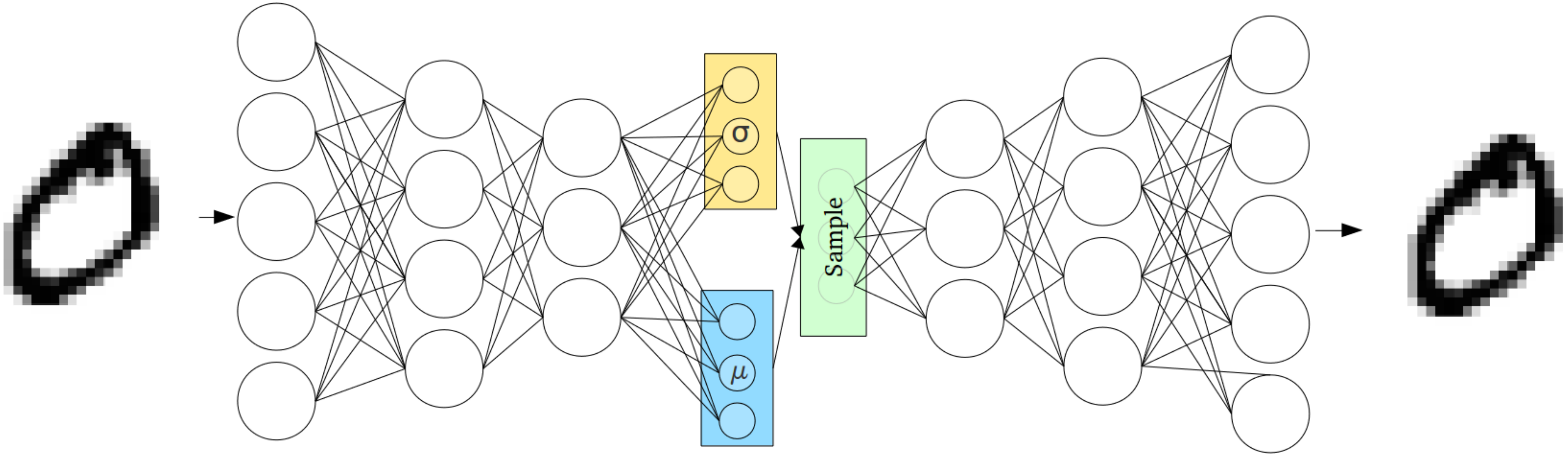
The Plain Autoencoder: *a toy example*

1. The AE learns some dense packing of the data space

2. The latent representation is highly coupled with the expressiveness of the network architecture of the encoder and decoder



The Variational Autoencoder



$$\text{Loss} = \underbrace{|\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2}_{\textit{Reconstruction error}} - \underbrace{\textstyle\sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)}_{KL(q(z|x)||p(z)) \sim \text{“Information cost”}}$$

The Variational Autoencoder:

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder:

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

1) $\boldsymbol{\beta}$ is dimensionful!

The same dimension as the distance metric,
e.g. GeV.

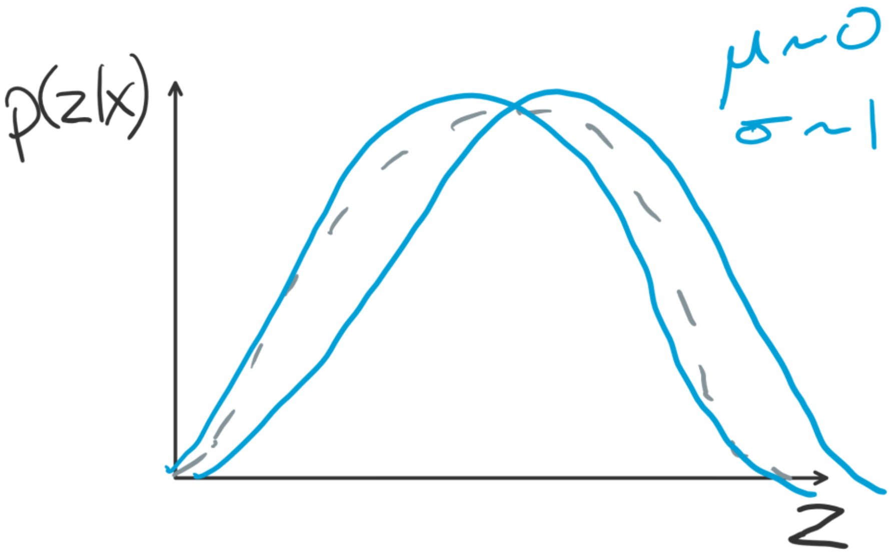
$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder:

Information and the loss function

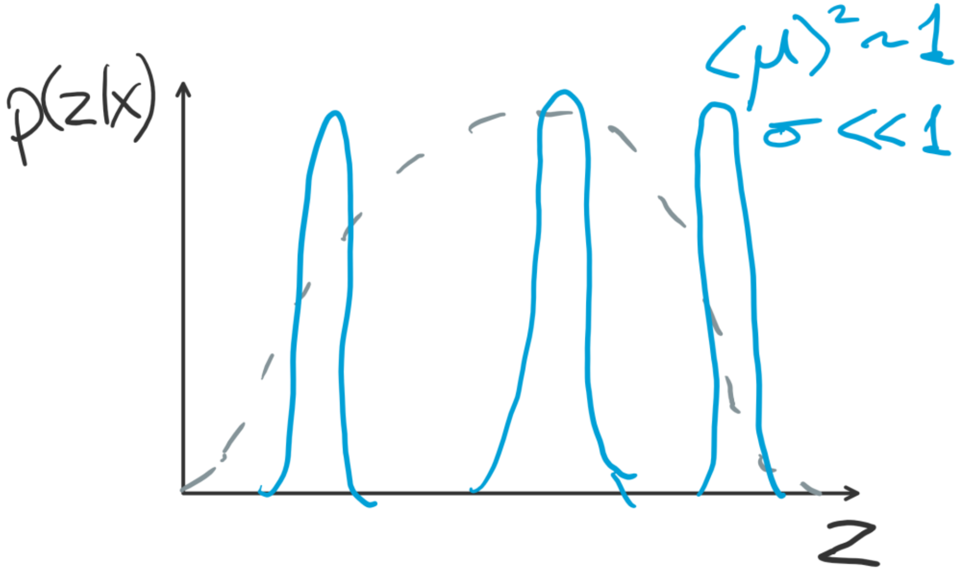
$\boldsymbol{\beta} \rightarrow \infty$

No info encoded in latent space



$\boldsymbol{\beta} \ll$ ***Lengthscale***

Info encoded in latent space



$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder:

Information and the loss function

$\boldsymbol{\beta} \rightarrow \infty$

No info encoded in latent space

$\boldsymbol{\beta} \ll$ ***Lengthscale***

Info encoded in latent space

2) $\boldsymbol{\beta}$ is the cost for encoding information

The encoder will only encode information about the input to the extent that its usefulness for reconstruction is sufficient to justify the cost.

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2 - \beta^2 \sum_i \frac{1}{2}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

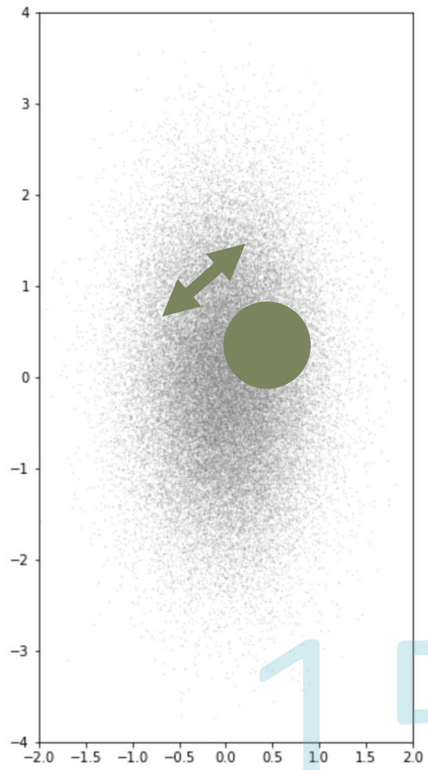
The Variational Autoencoder:

Information and the loss function

$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/\beta^2 - \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

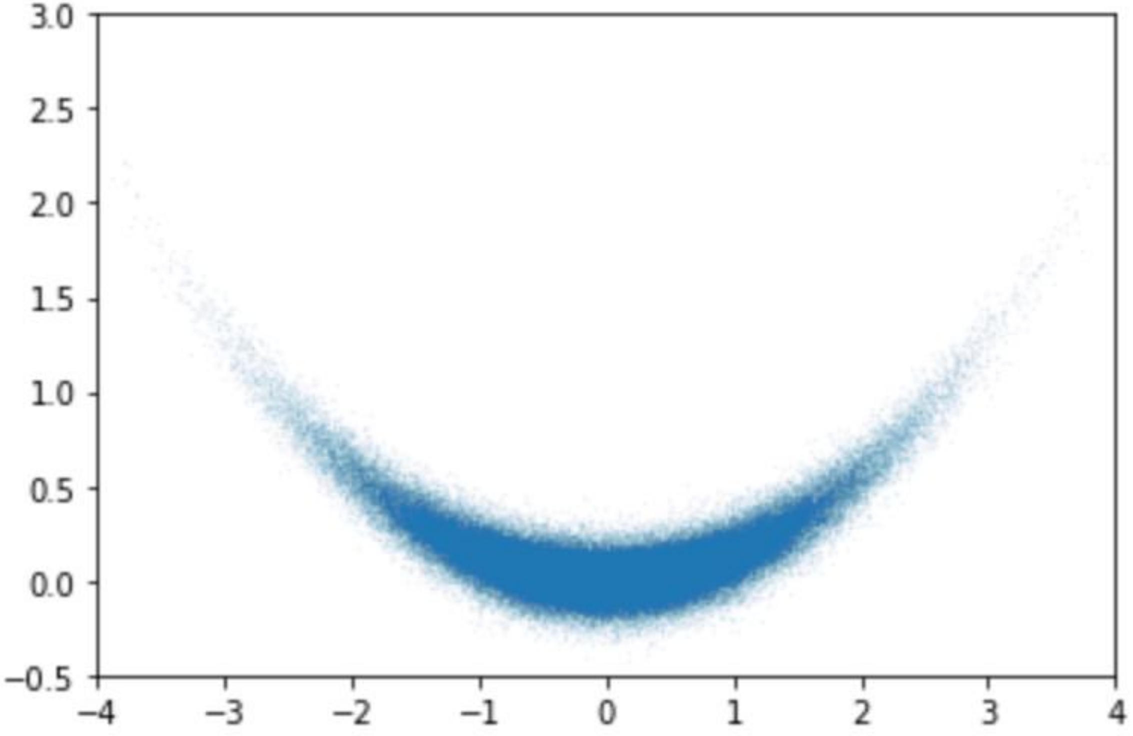
3) $\boldsymbol{\beta}$ is the distance resolution in reconstruction space

The stochasticity of the latent sampling will smear the reconstruction at scale $\sim \beta$



The Variational Autoencoder:

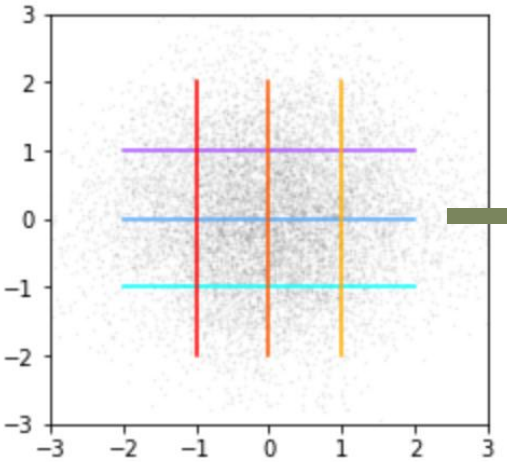
Bananas



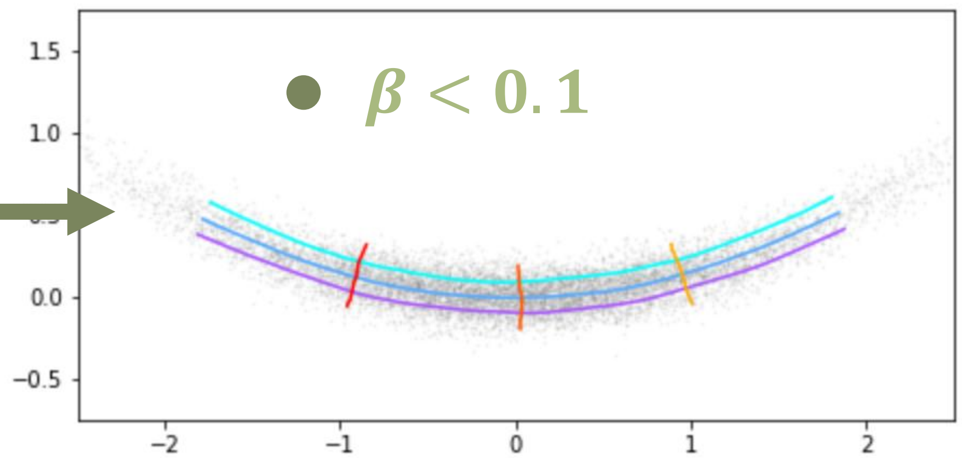
The Variational Autoencoder:

Bananas

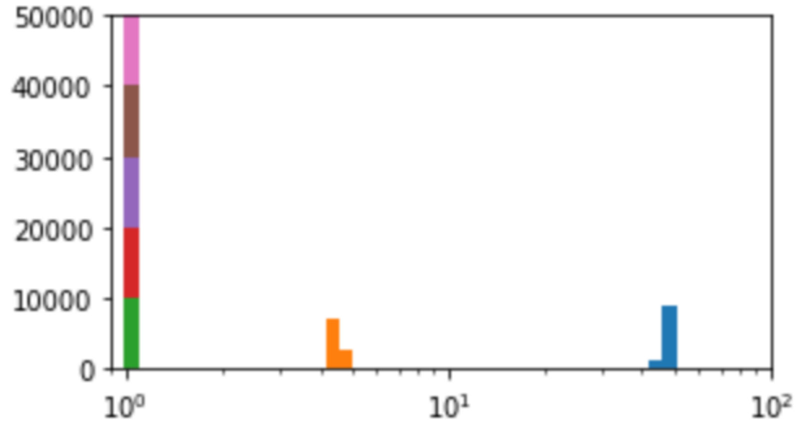
Latent Space



Reconstruction Space



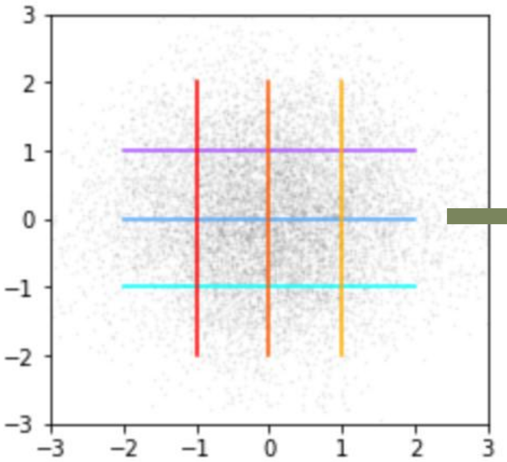
- $\beta < 0.1$



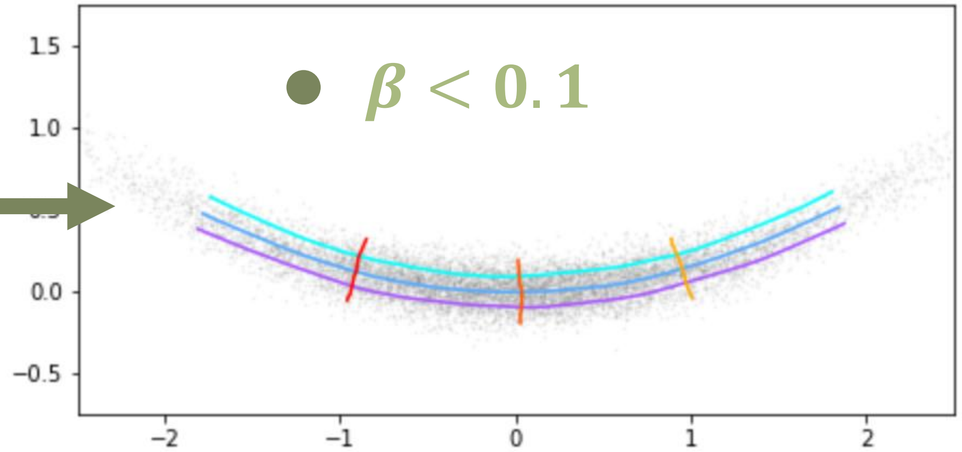
The Variational Autoencoder:

Bananas

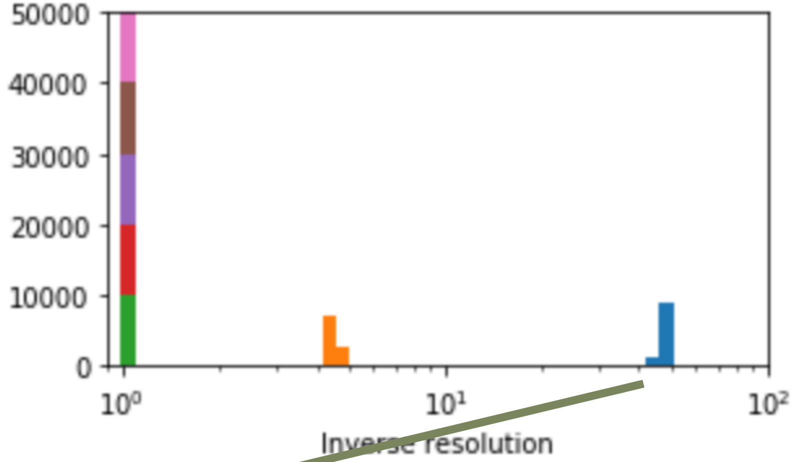
Latent Space



Reconstruction Space



- $\beta < 0.1$



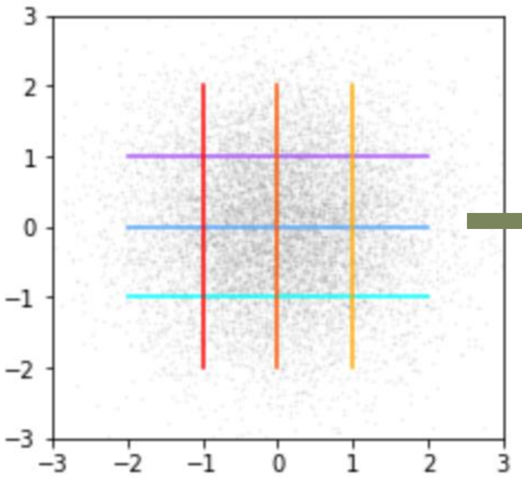
$$\text{Size} = {}^{\beta}/_{\sigma}$$

The VAE is doing non-linear PCA

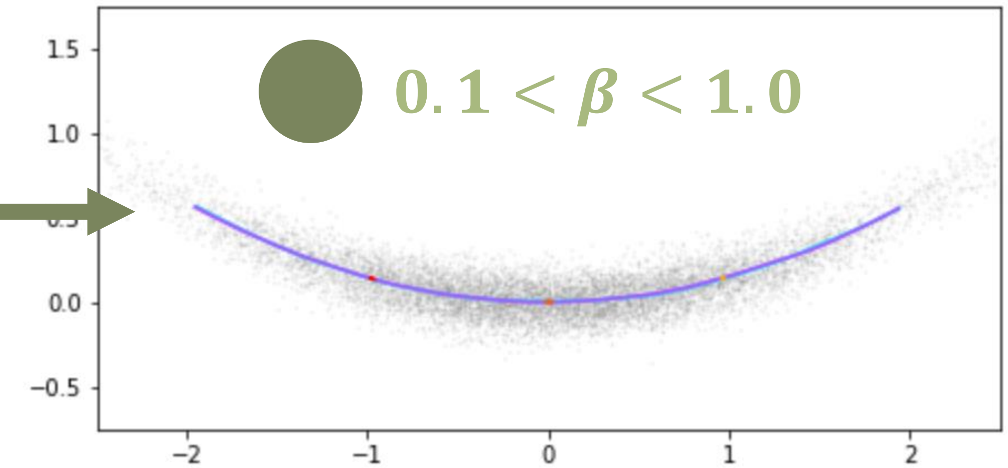
The Variational Autoencoder:

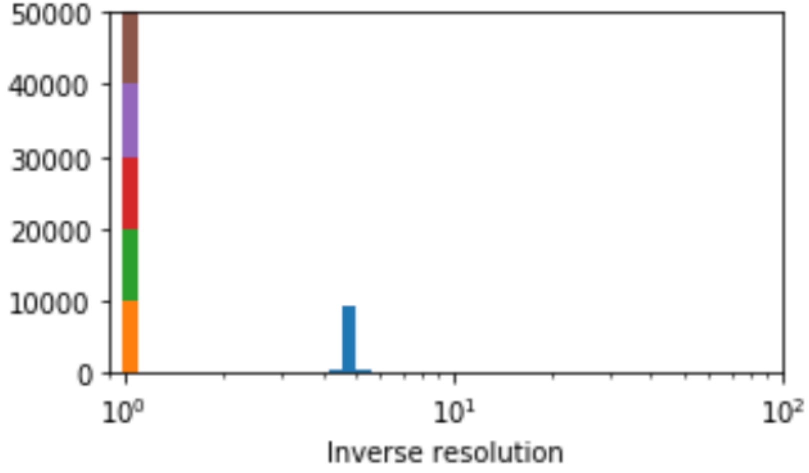
Bananas

Latent Space



Reconstruction Space

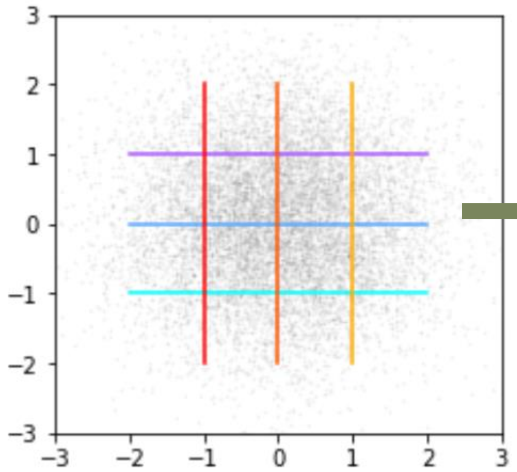




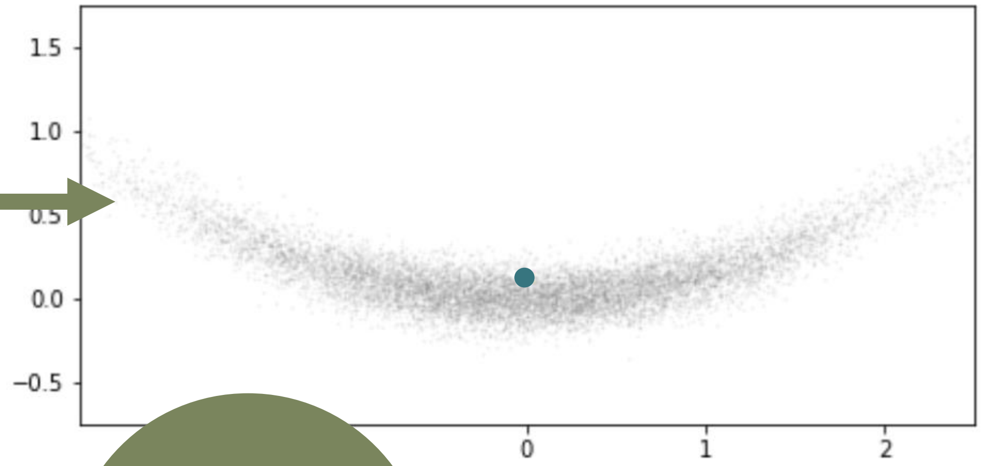
The Variational Autoencoder:

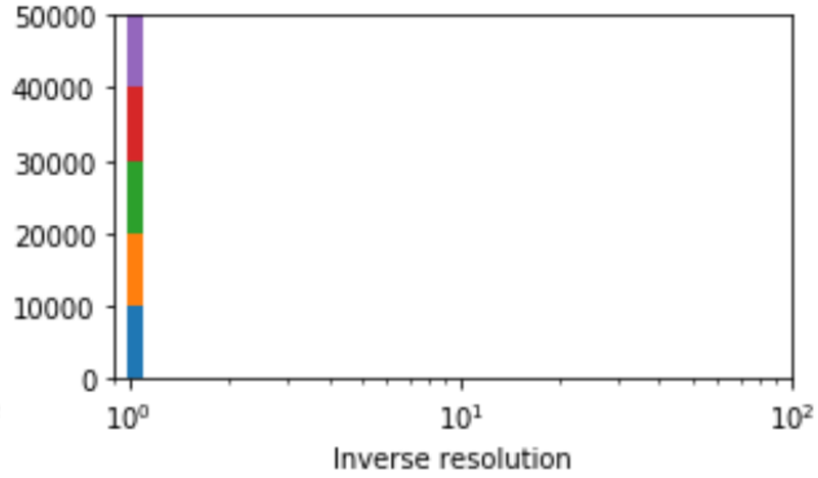
Bananas

Latent Space



Reconstruction Space





$\beta \gg 1$

The Variational Autoencoder:

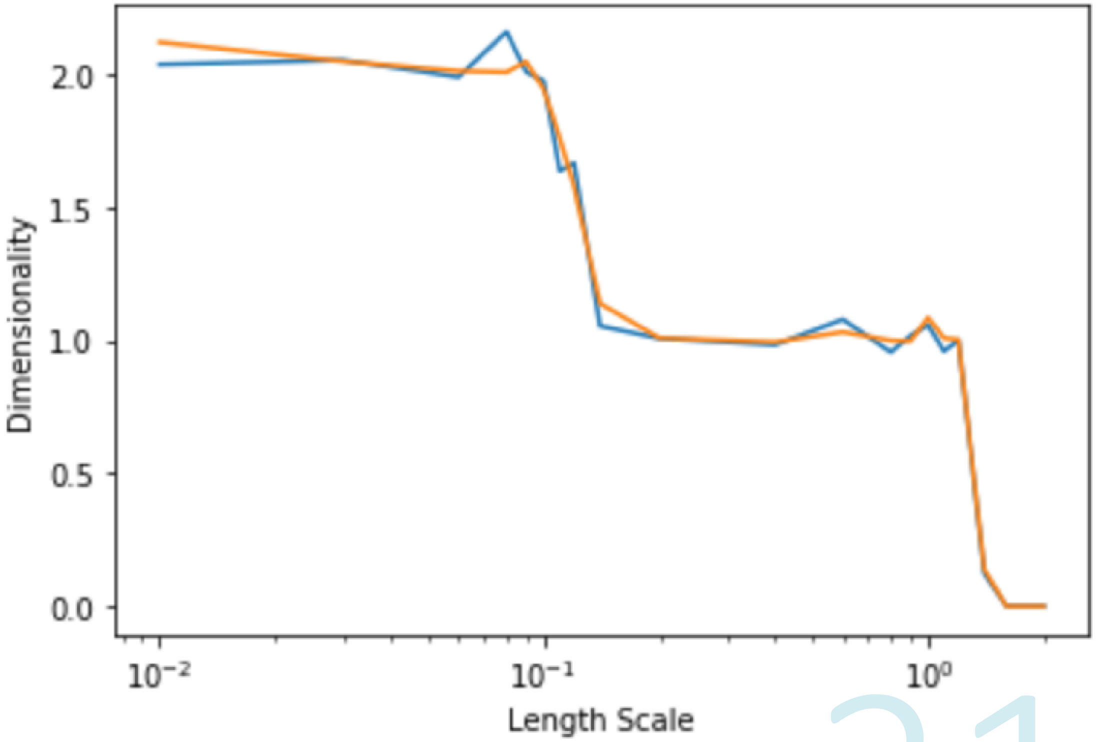
Dimensionality

$$\langle|\Delta \boldsymbol{x}|^2\rangle = \sum\langle|\Delta x_i|^2\rangle = D\rho^2 + \sum_{i>D} S_i^2$$

$$D = \frac{d\langle|\Delta \boldsymbol{x}|^2\rangle}{d\rho^2}$$

Setting $\frac{dL}{d\sigma} = 0$ implies:

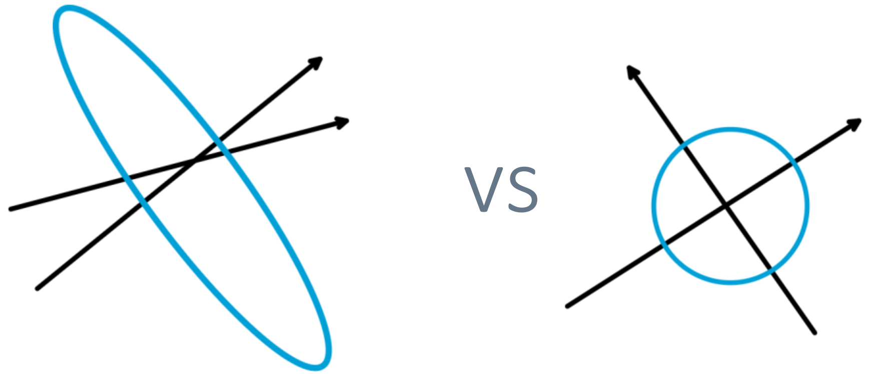
1. $\rho = \beta$
2. $D = \frac{d\,KL}{d\log\beta}$



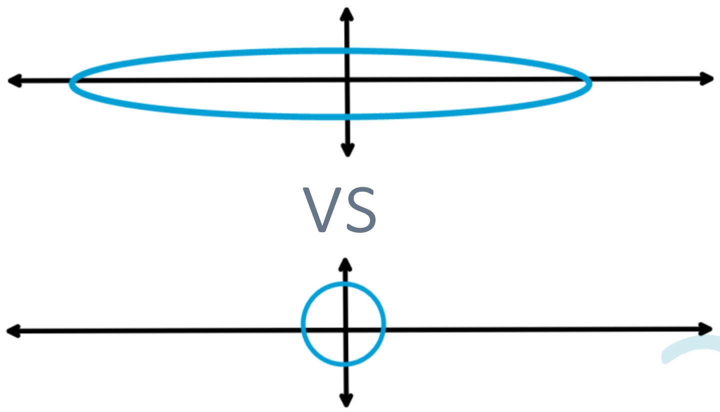
The Variational Autoencoder:

Orthogonalization and Organization is Information-Efficient

Orthogonalization:

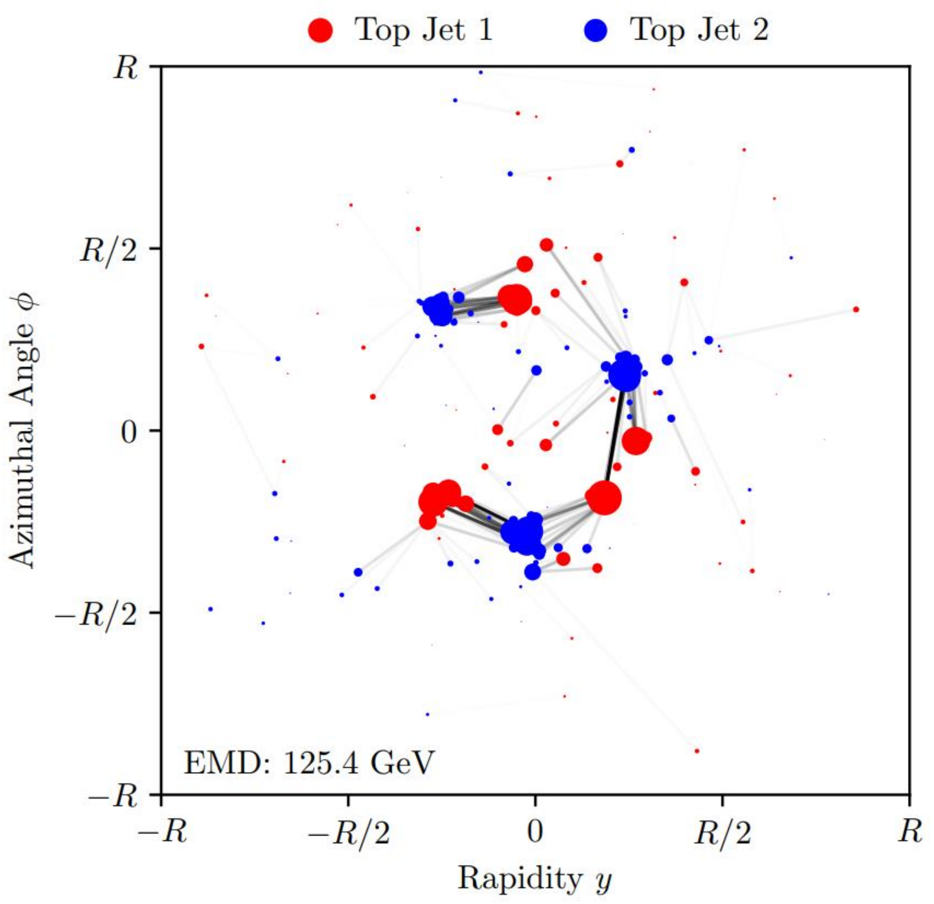


Organization:



Distance between Jets:

Optimal Transport



arXiv:1902.02346

EMD ≈ Sinkhorn Distance

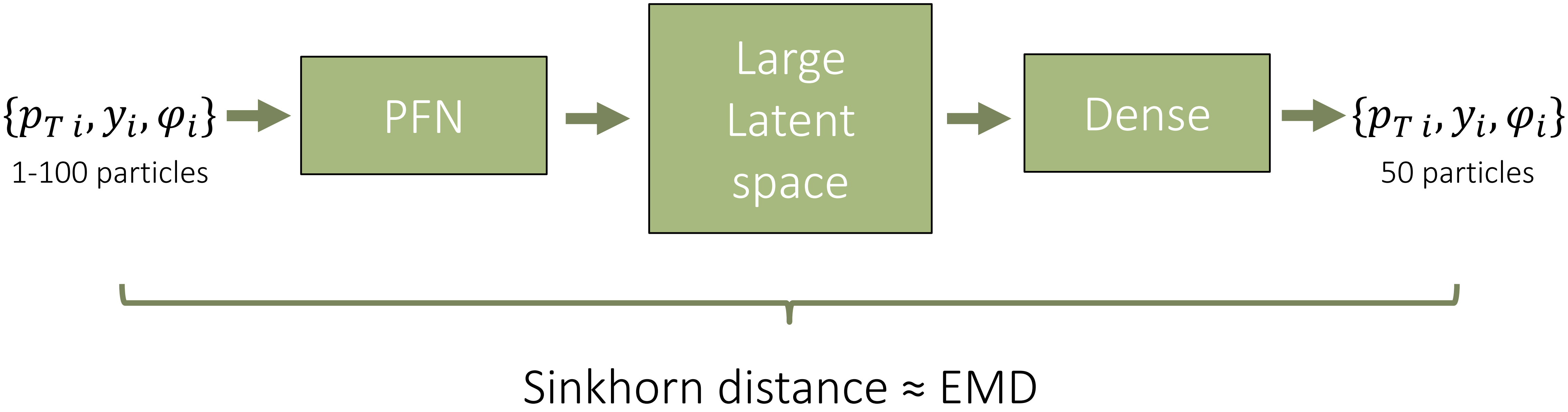
I wish I had an extra 15 minutes to talk about this. Critical papers (for me):

arXiv:1306.0895 [stat.ML] M. Cuturi
Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances

arXiv:1706.00292 [stat.ML] A. Genevay, G. Peyré, M. Cuturi
Learning Generative Models with Sinkhorn Divergences

arXiv:1805.11897 [stat.ML] G. Luise, A. Rudi, M. Pontil, C. Ciliberto
Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

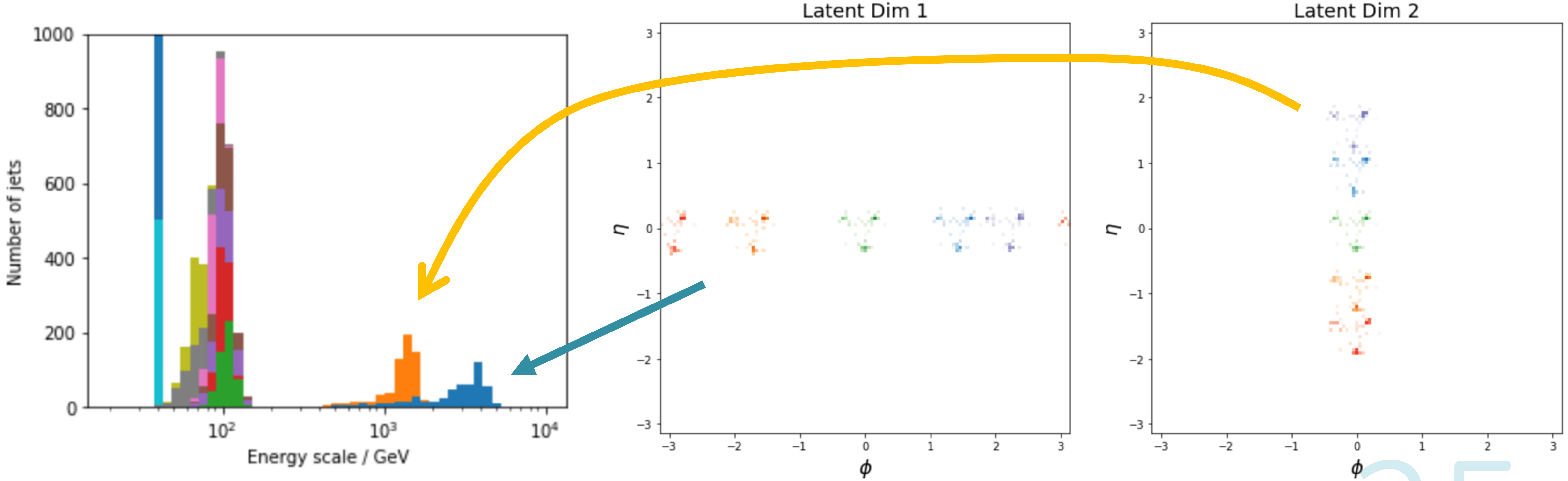
Jet VAE


$\{p_{T\,i}, y_i, \varphi_i\}$
1-100 particles
PFN
Large Latent space
Dense
$\{p_{T\,i}, y_i, \varphi_i\}$
50 particles
Sinkhorn distance ≈ EMD

Exploring the Learnt Representation:

Top Jets

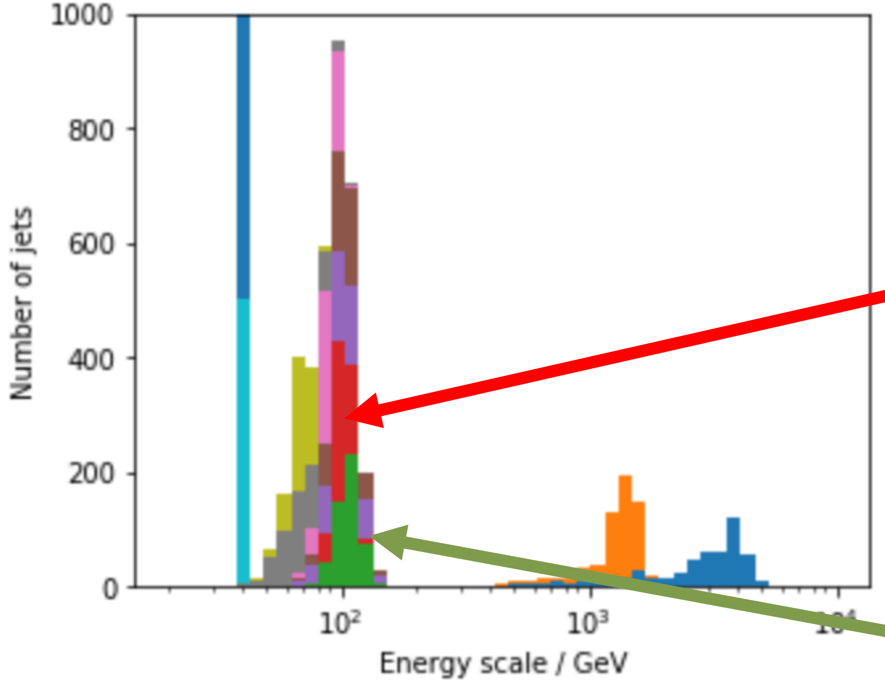
$\beta = 40\ GeV$

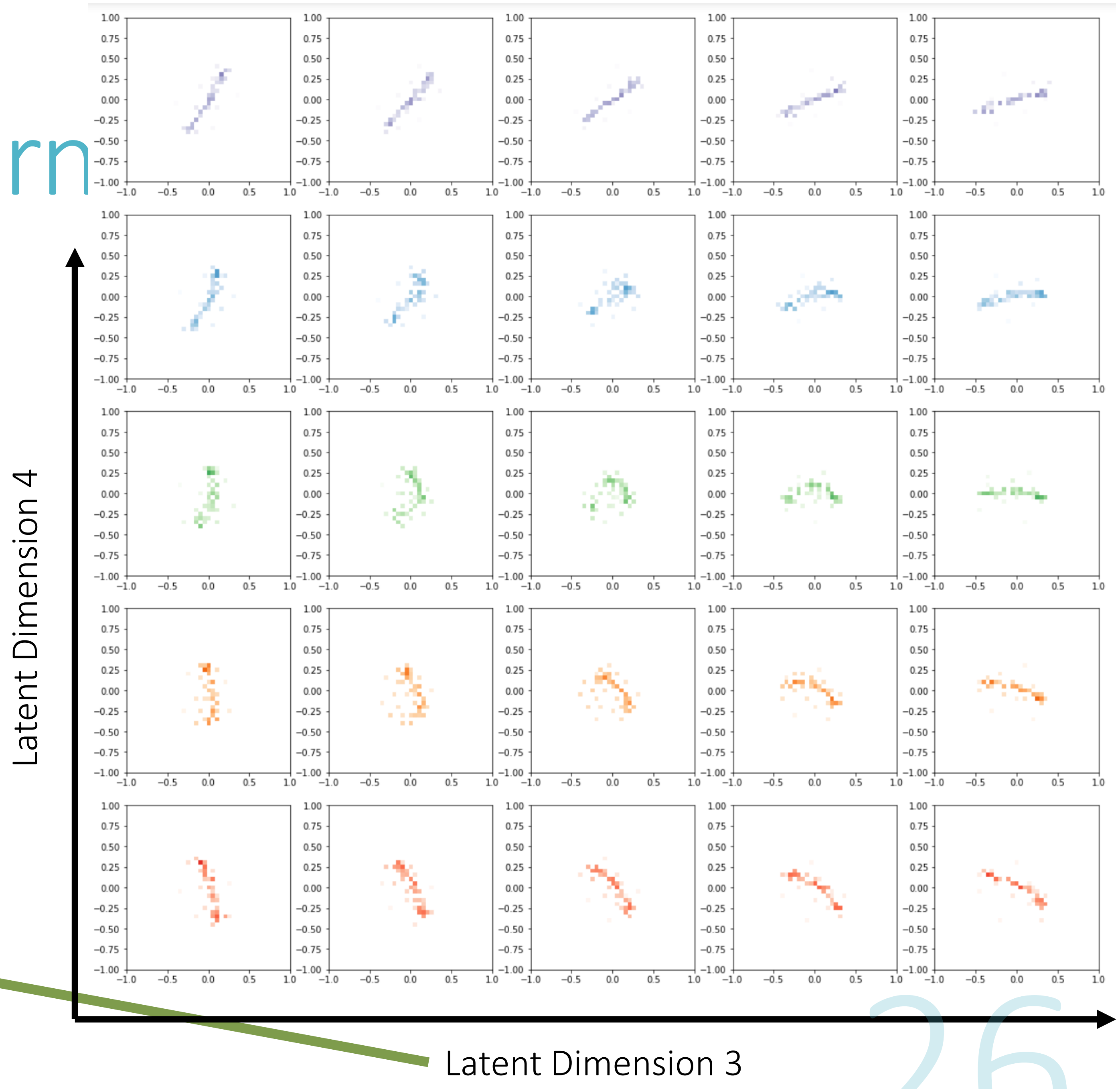


Exploring the Learn

Top Jets

$\beta = 40\ GeV$

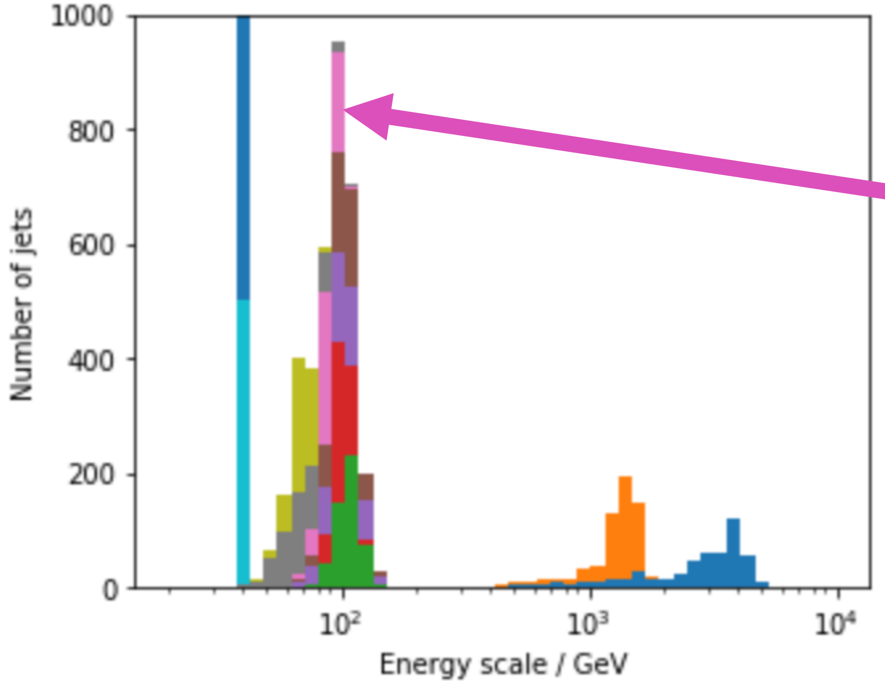


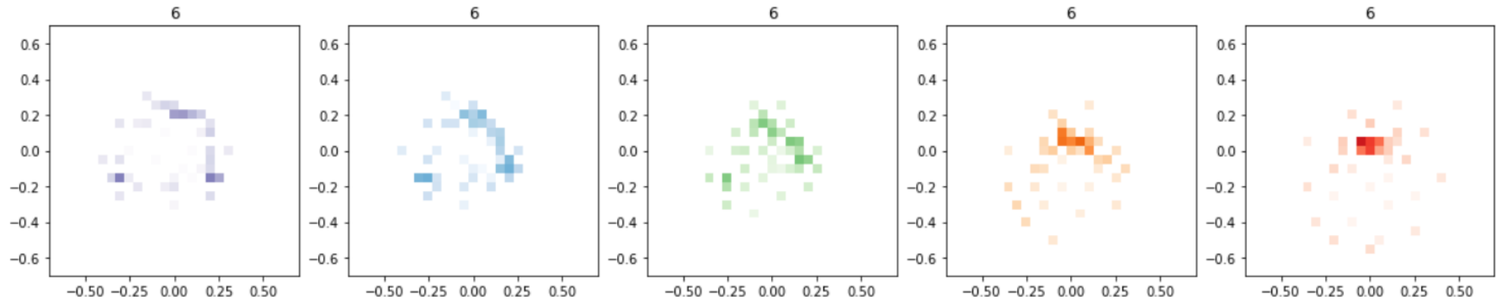


Exploring the Learnt Representation:

Top Jets

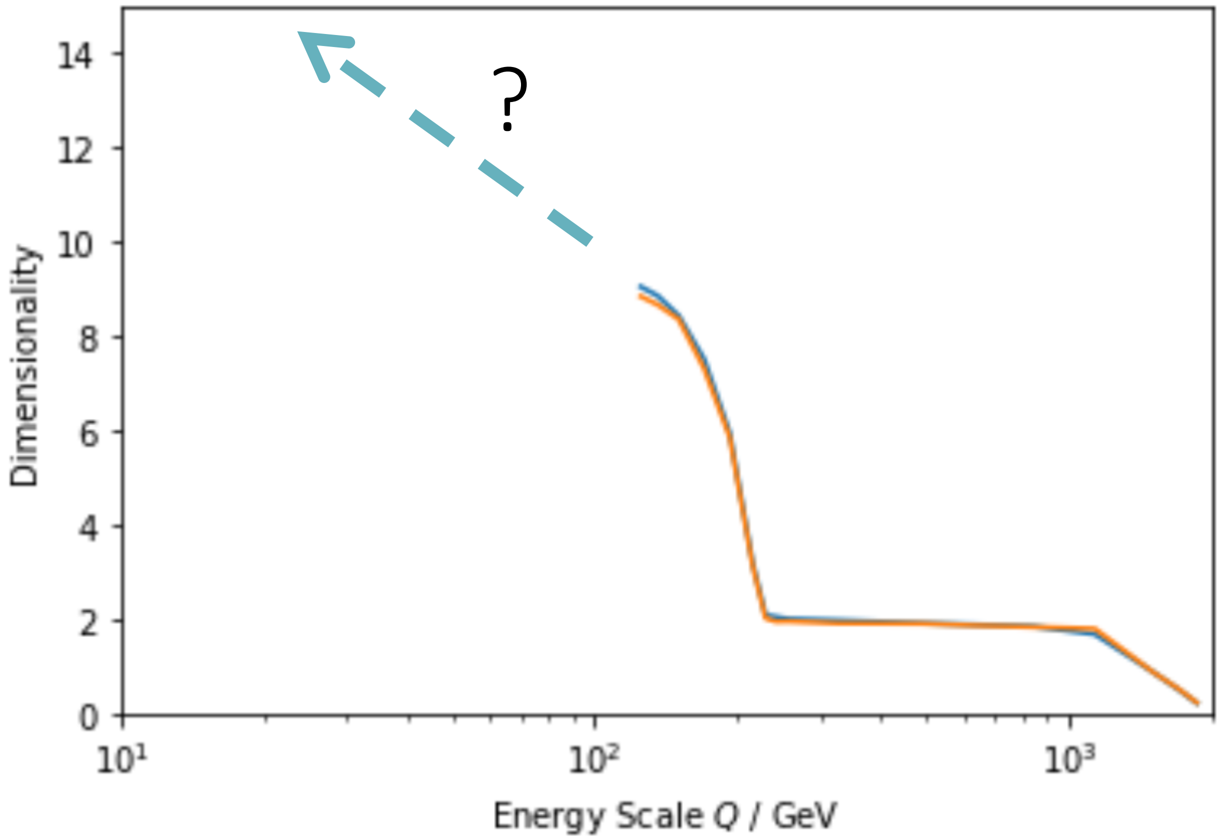
$\beta = 40\ GeV$

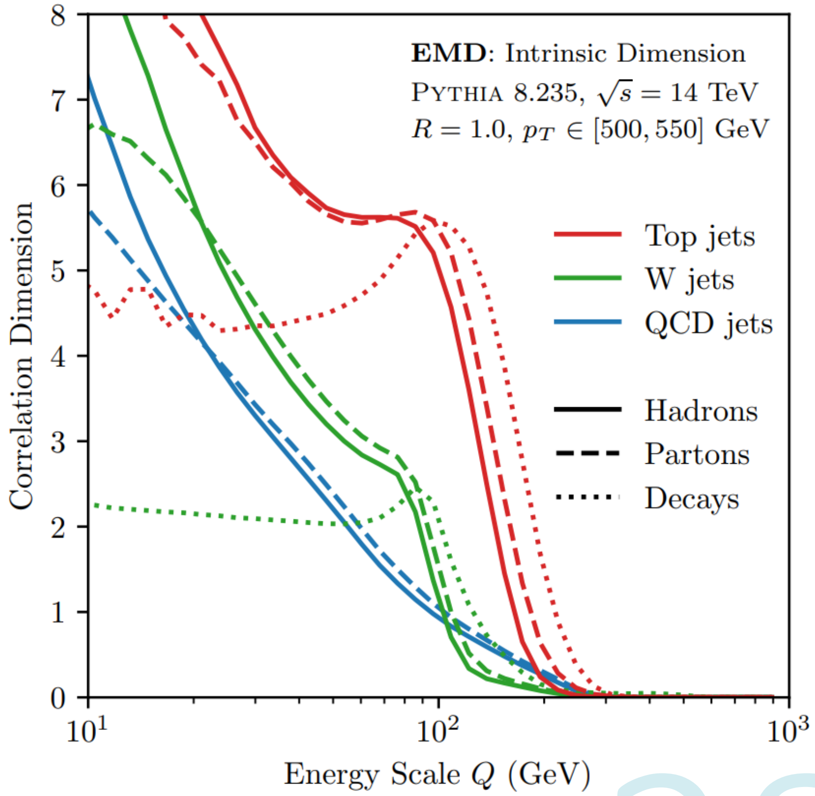




Exploring the Learnt Representation:

Dimensionality





Future Directions

1. What is the point?

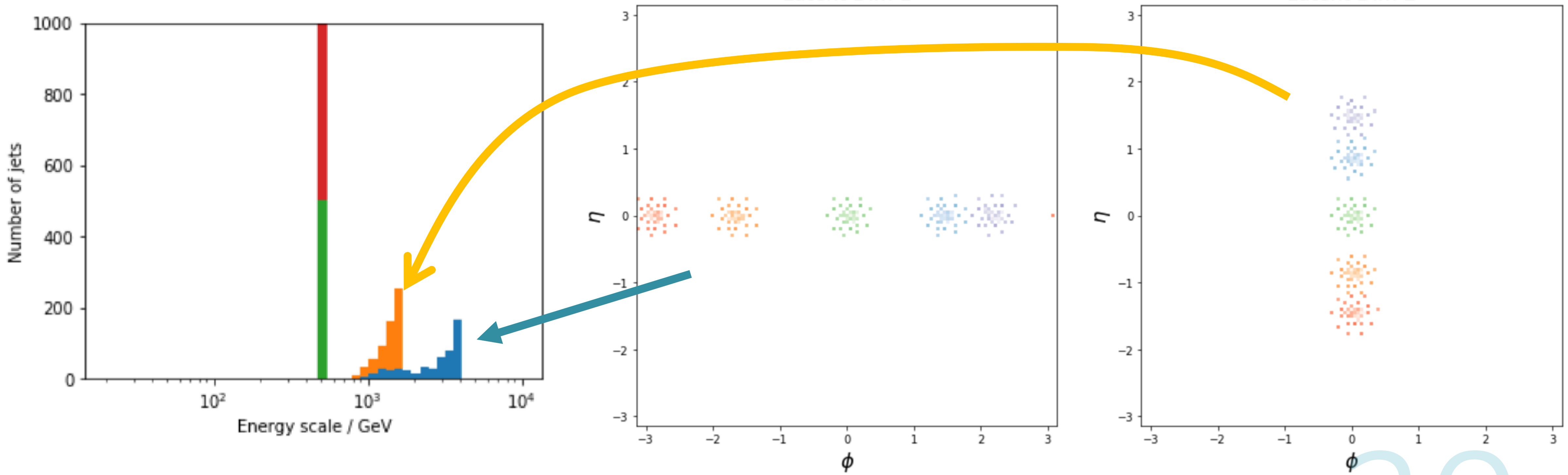
2. Alternative latent priors?

3. Alternative metrics?

Exploring the Learnt Representation:

Top Jets

$\beta = 400\ GeV$



The Variational Autoencoder

ML Engineer:

“A VAE is a fancy AE with regulated stochastic latent space sampling”

Bayesian statistician:

*“A VAE is a probability model trained to extremize the **E**vidence **L**ower **BO**und on the posterior distribution p(x)”*