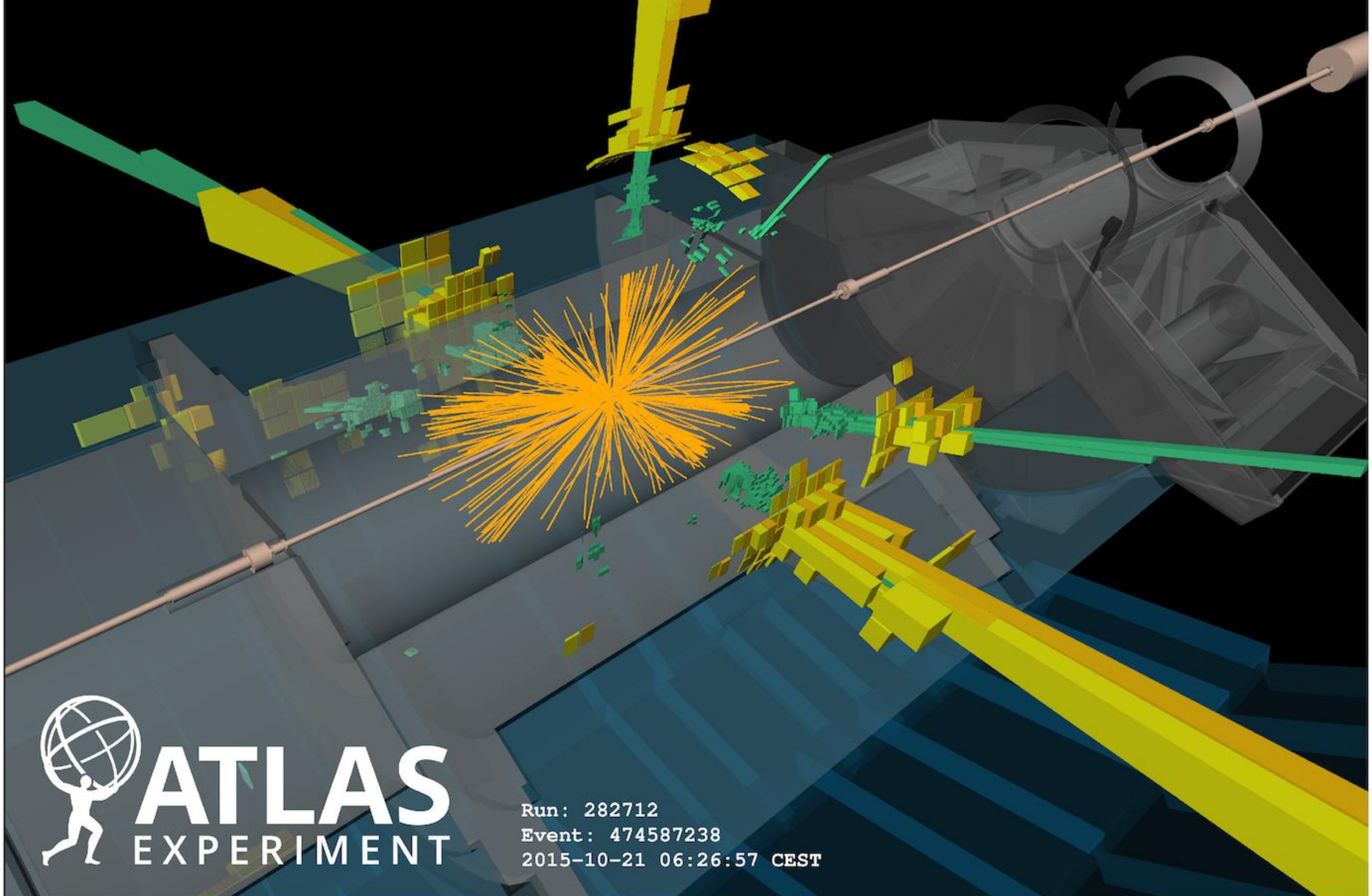
The Learnt Geometry of Collider Events

2109.10919 Jack H Collins

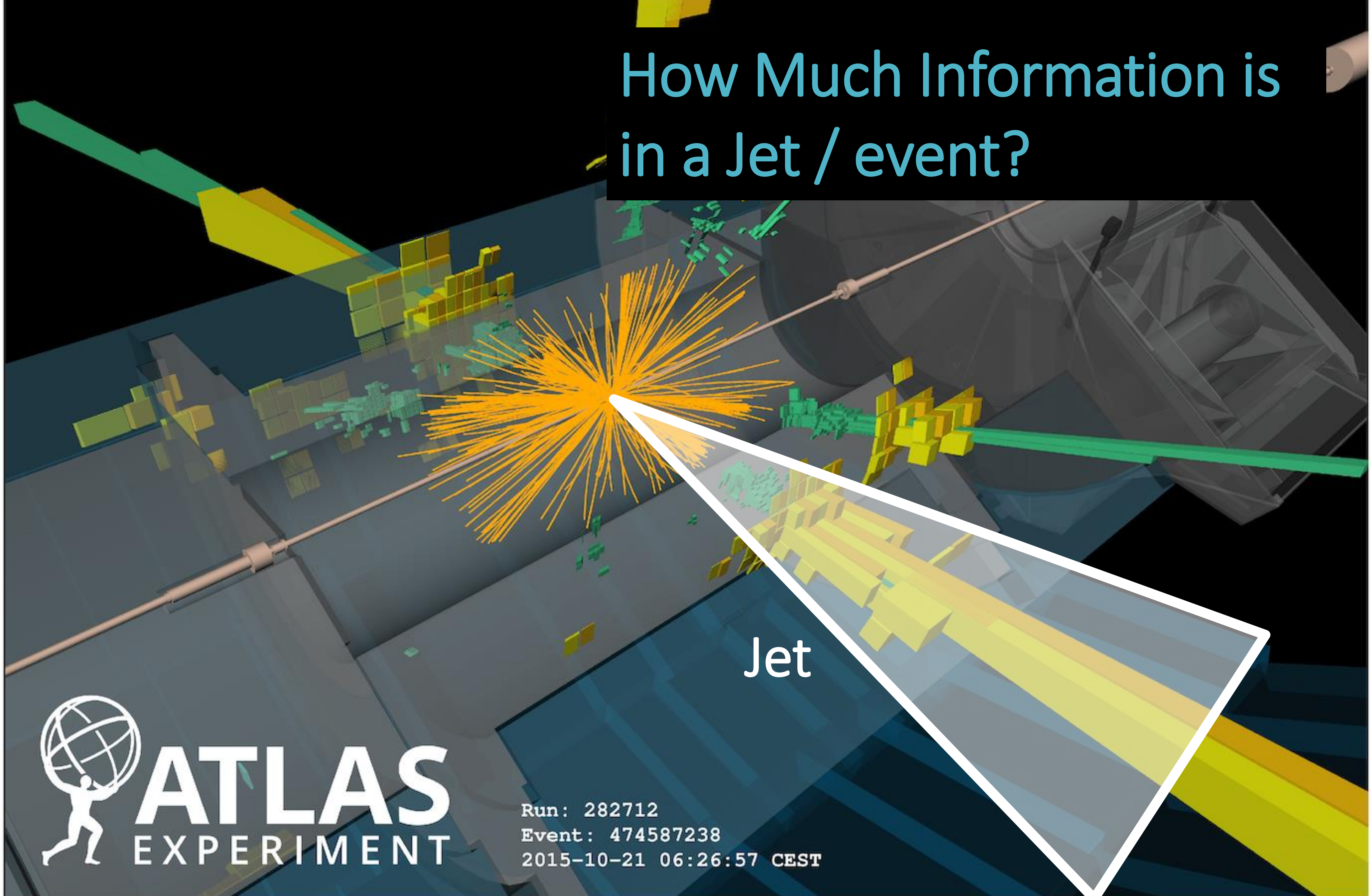
Jack Collins

SLAC NATIONAL ACCELERATOR LABORATORY

Heidelberg UNIVERSITY


ATLAS
EXPERIMENT
Run: 282712
Event: 474587238
2015-10-21 06:26:57 CEST

How Much Information is in a Jet / event?



Menu

(Absolutely no substitutions)

Aperetif
How much information is in a jet?

Appetizer
The Metric Space of Collider Events

Fish Course
The Variational Autoencoder:
a pedagogical introduction

Main Course
Application to W jets

Dessert
Unsupervised Classification

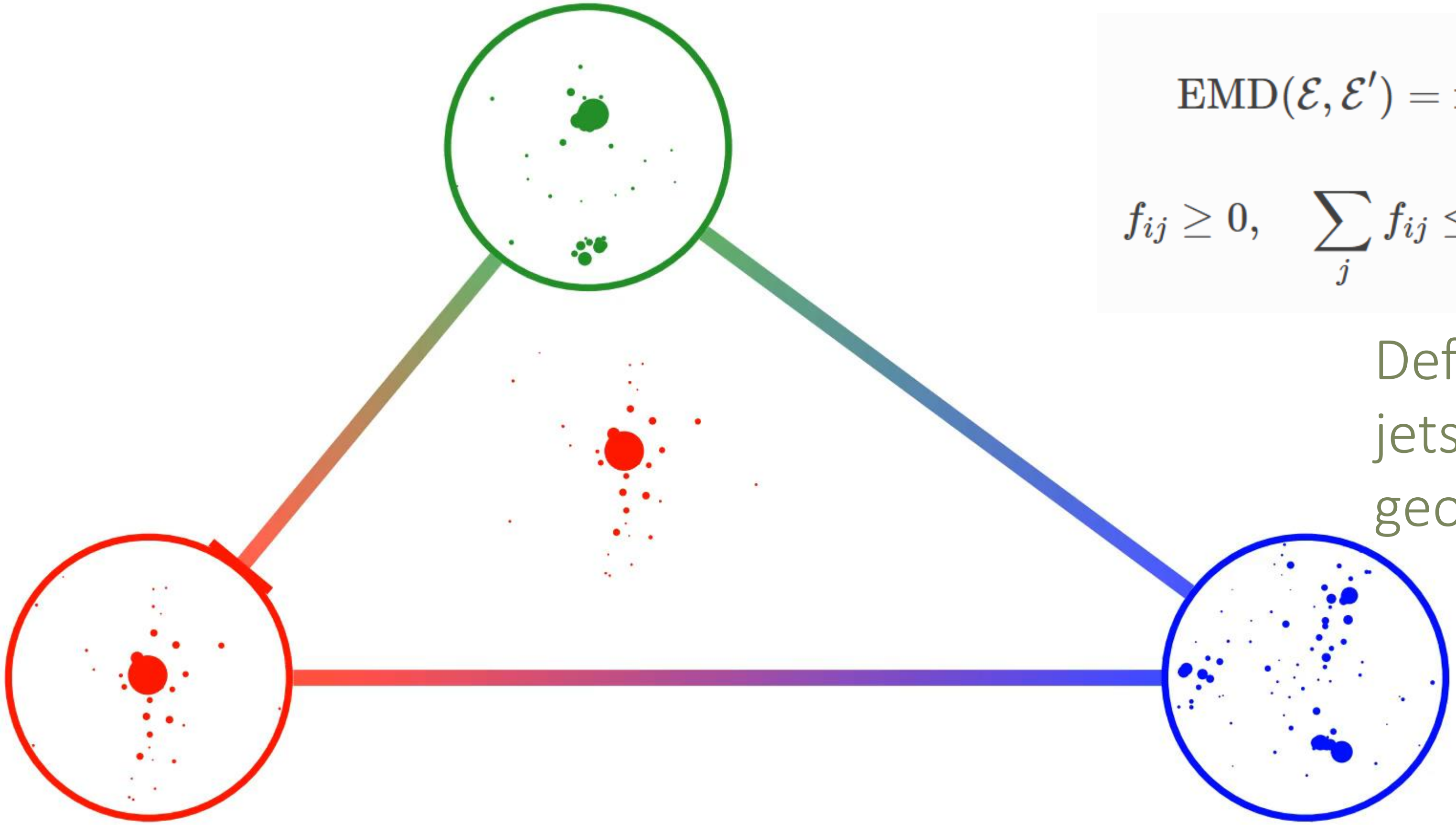
Digestif
Conclusions

Appetizer

The Metric Space of Collider Events

Earth Movers Distance

*Cost to transform one jet into another = Energy * distance*



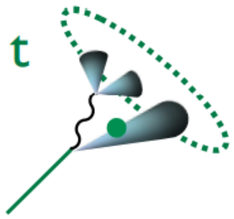
$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\}} \sum_{ij} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_i E_i - \sum_j E'_j\right|,$$

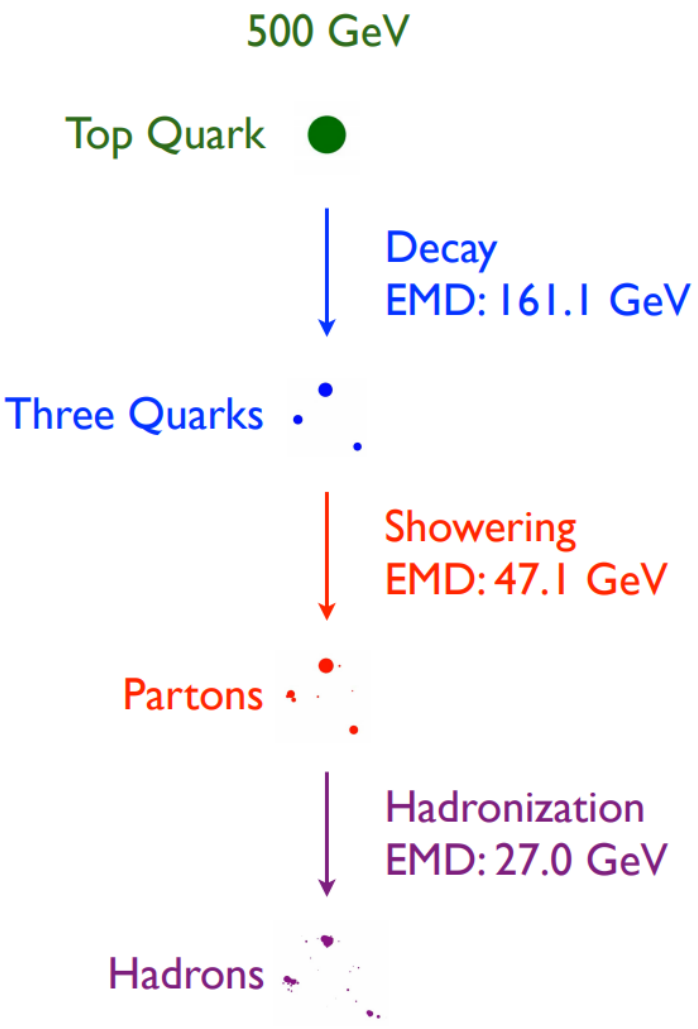
$$f_{ij} \geq 0, \quad \sum_j f_{ij} \leq E_i, \quad \sum_i f_{ij} \leq E'_j, \quad \sum_{ij} f_{ij} = E_{\min},$$

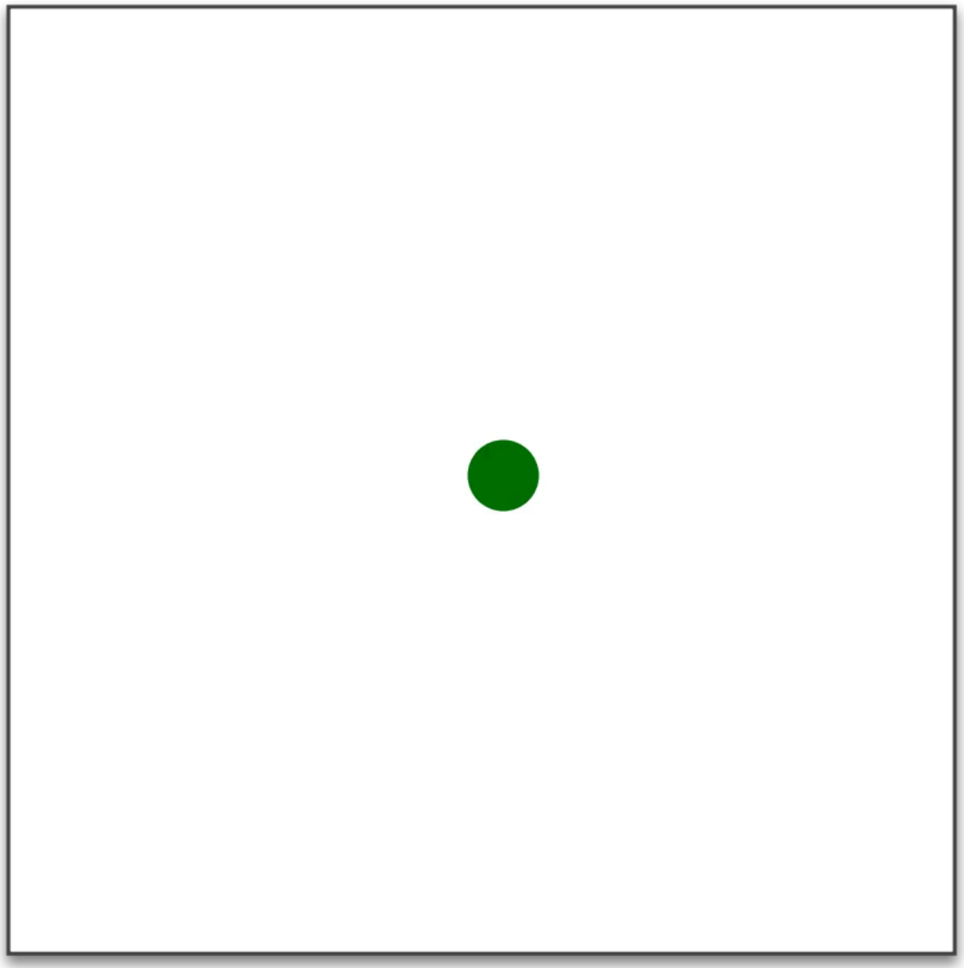
Defines a metric space in which jets or collider events form a geometric manifold.

Taken from https://energyflow.network/docs/emd/, Eric Metediov, Patrick Komiske III, Jesse Thaler

Visualizing Top Quark Evolution

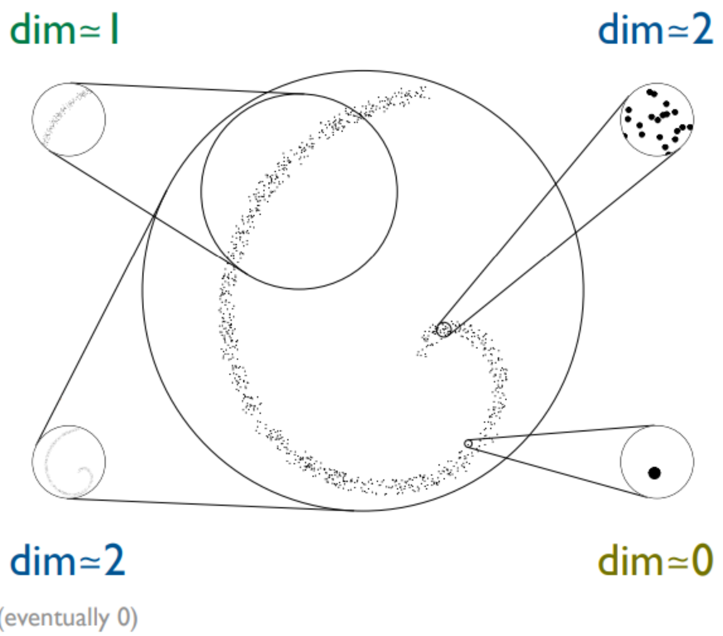






Quantifying Dimensionality

Correlation Dimension: $\mathrm{dim}(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta\big(\mathrm{EMD}(\mathcal{E}_i, \mathcal{E}_j) < Q\big)$



$$N_{\text{neighbors}}(r) \sim r^{\text{dim}}$$

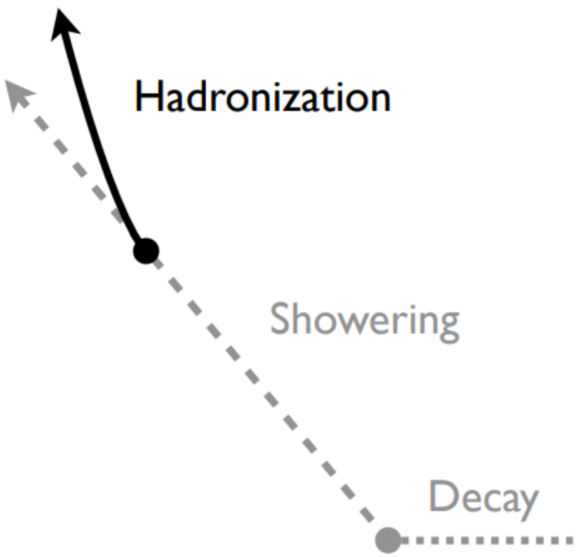
$$\Downarrow$$

$$\mathrm{dim}(r) \sim r \frac{\partial}{\partial r} \ln N_{\text{neighbors}}(r)$$

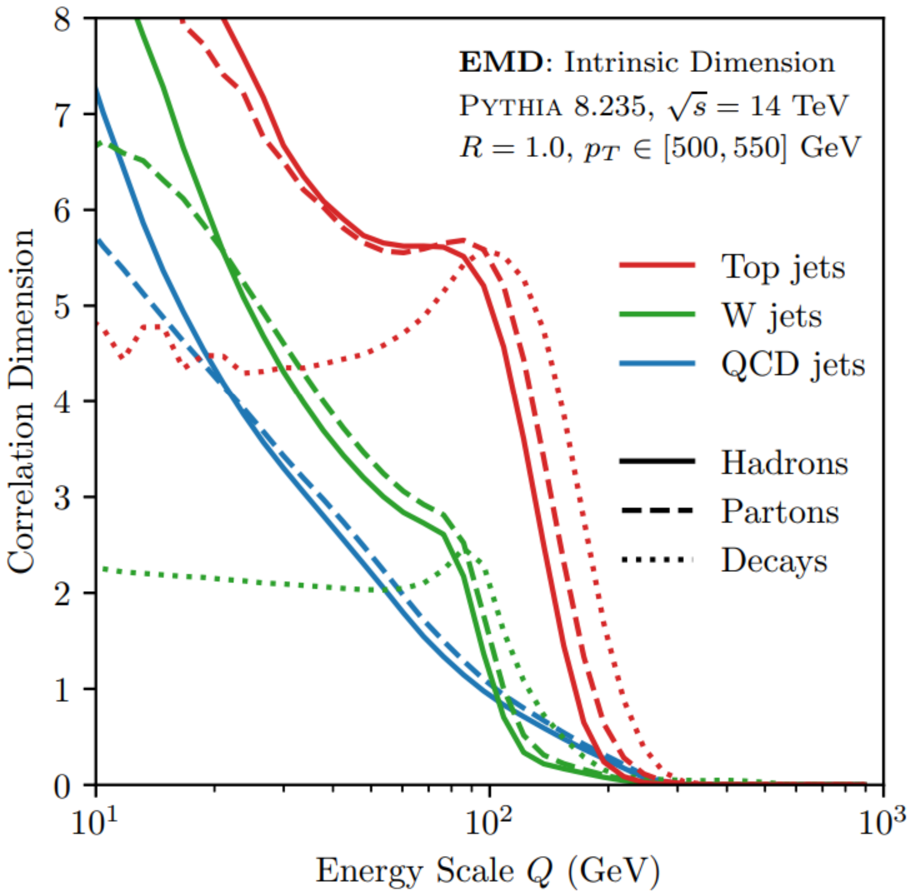
[Grassberger, Procaccia, PRL 1983; Kégl, NIPS 2002]

Hadron-Level Dimension

$$\dim(Q) = Q \frac{\partial}{\partial Q} \ln \sum_i \sum_j \Theta(\text{EMD}(\mathcal{E}_i, \mathcal{E}_j) < Q)$$



Increasing complexity: multi-body phase space
perturbative emissions
non-perturbative dynamics

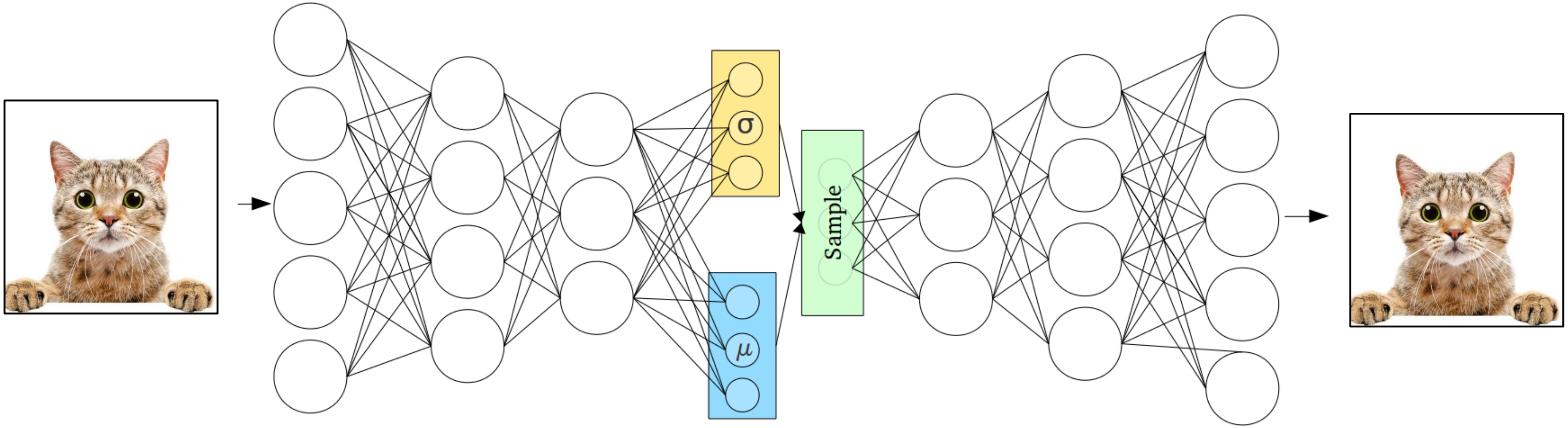


[Komiske, Metodiev, JDT, 1902.02346]

Fish Course

The Variational Autoencoder

The Variational Autoencoder



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

The Variational Autoencoder

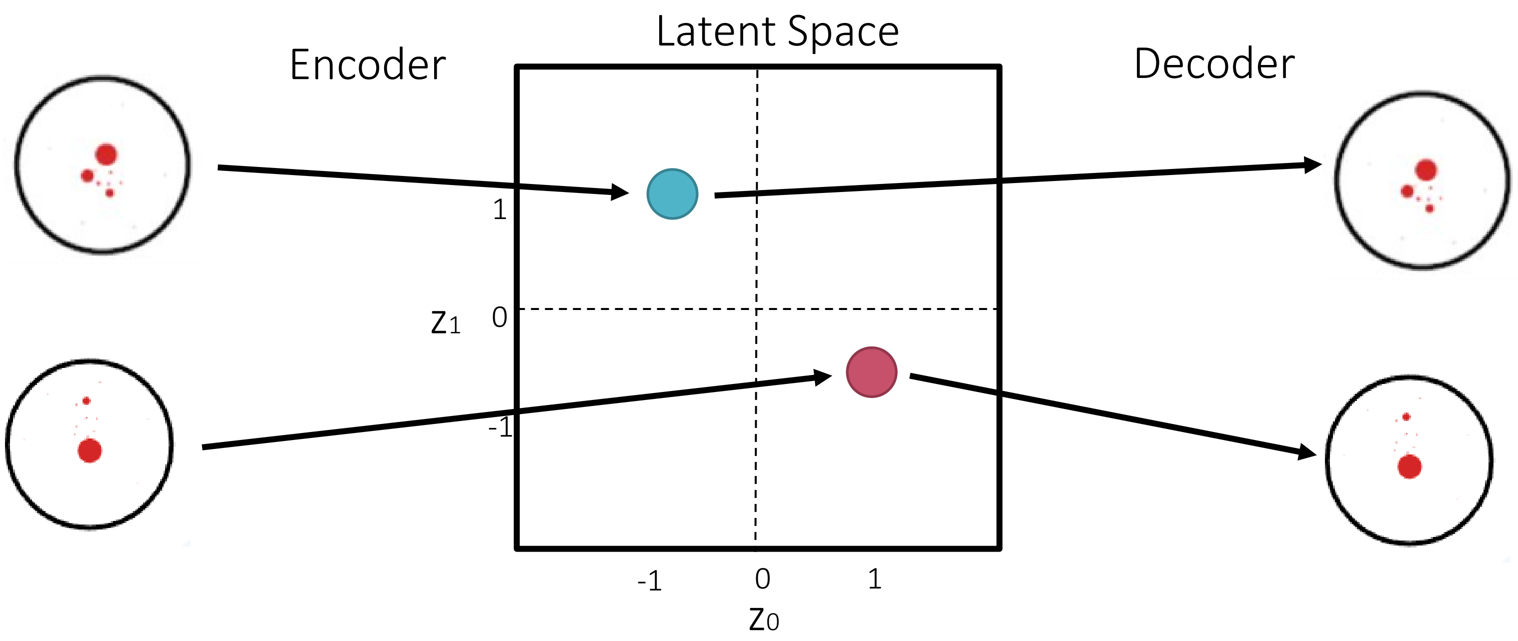
Loss $= -\langle \log(p(x|z)) \rangle + D_{KL}(q(z|x)||P(z))$

Loss $= -\langle \log(\exp(-d(x, \rho(z))^2/2\beta^2)) \rangle + D_{KL}(q(z|x)||P(z))$

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

The Variational Autoencoder:

Information and the loss function

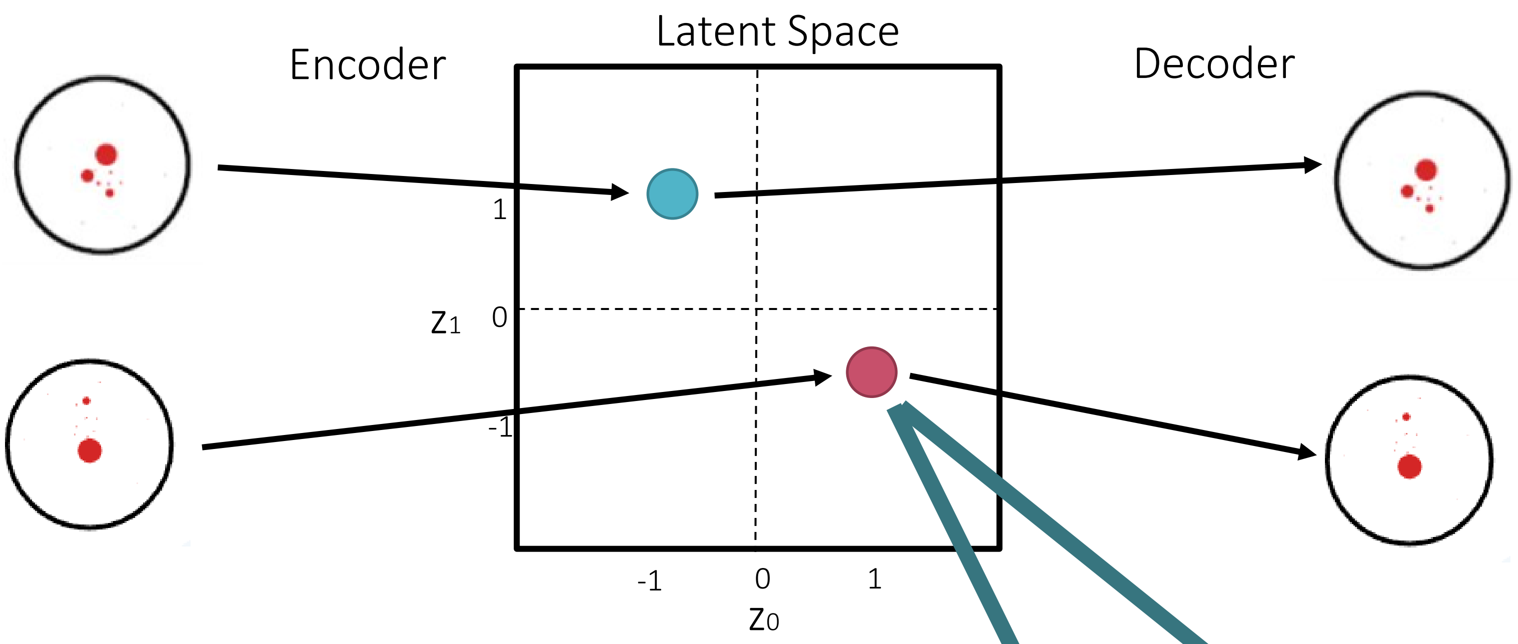


Precise encoding in latent space is penalized by KL term but favoured for reconstruction

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

The Variational Autoencoder:

Information and the loss function

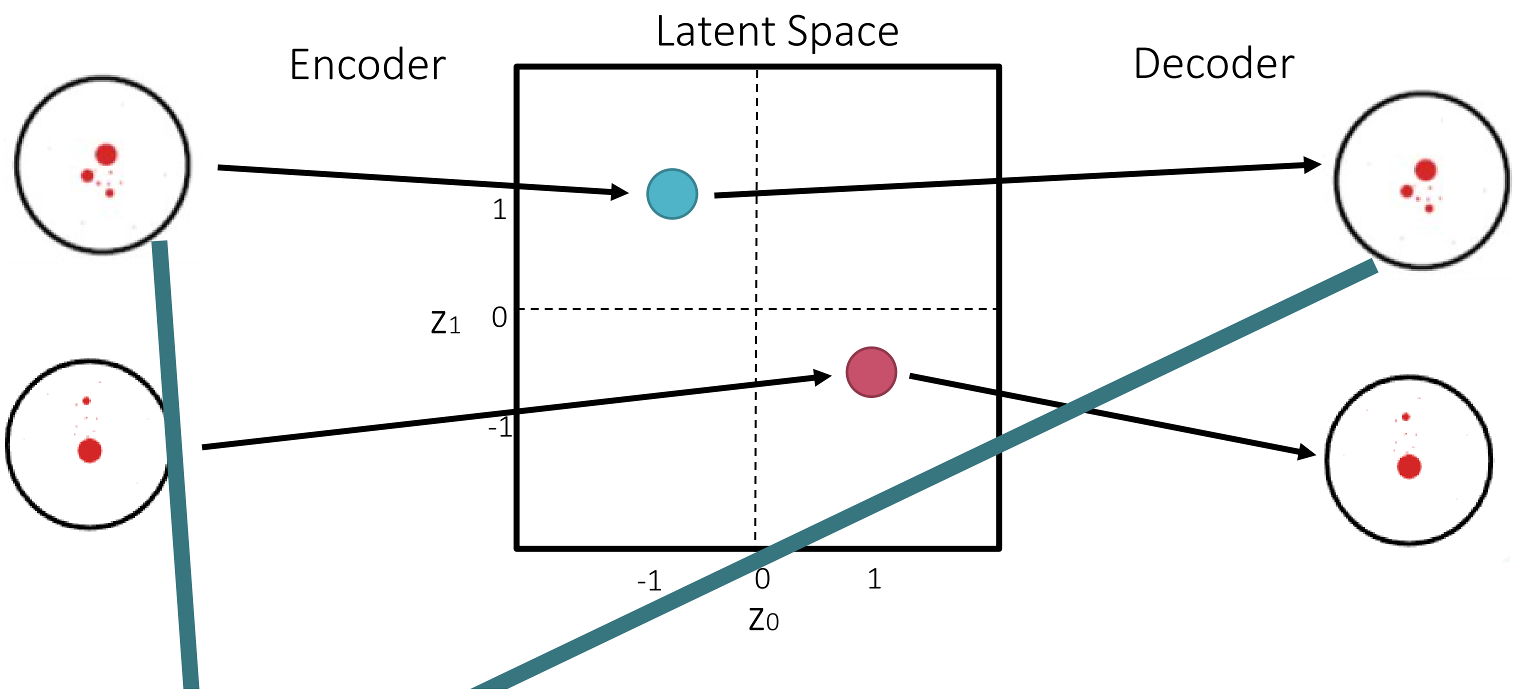


Precise encoding in latent space is penalized by KL term but favoured for reconstruction

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

The Variational Autoencoder:

Information and the loss function

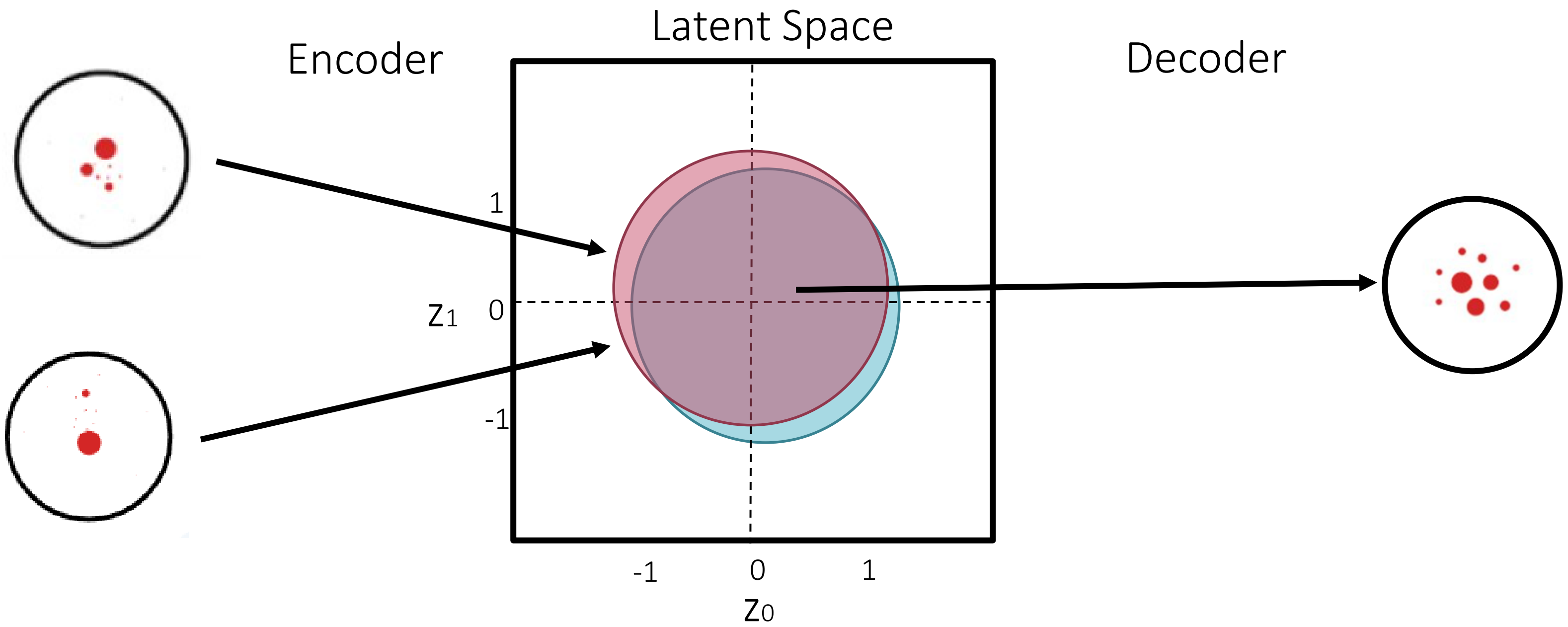


Precise encoding in latent space is penalized by KL term but favoured for reconstruction

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

The Variational Autoencoder:

Information and the loss function



Imprecise encoding in latent space is favoured by KL term but penalized by reconstruction

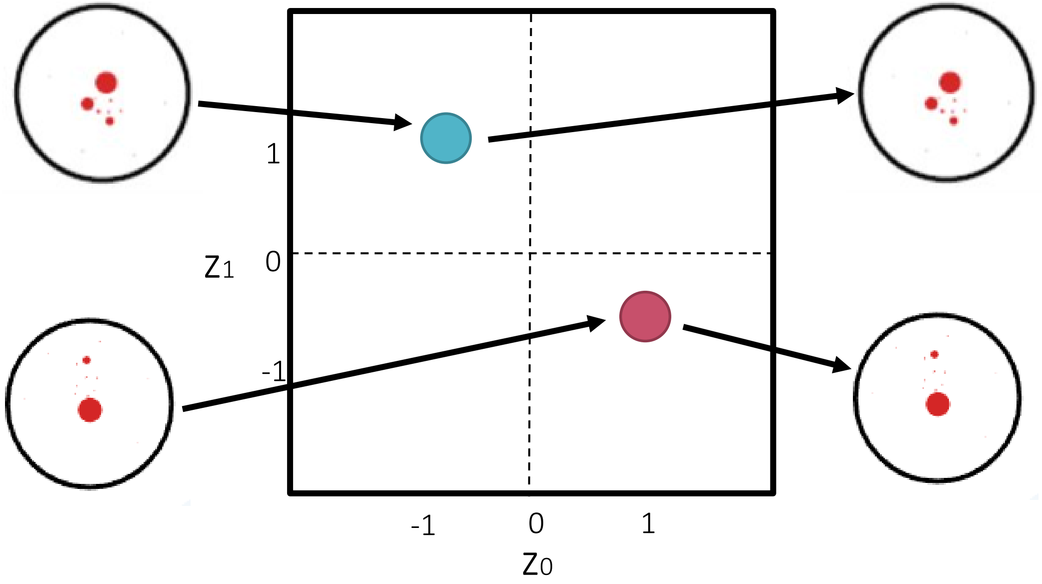
$$\text{Loss} = |\boldsymbol{x}_{out} - \boldsymbol{x}_{in}|^2/2\beta^2 - \sum_i \frac{1}{2}\left(1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)$$

The Variational Autoencoder:

Information and the loss function

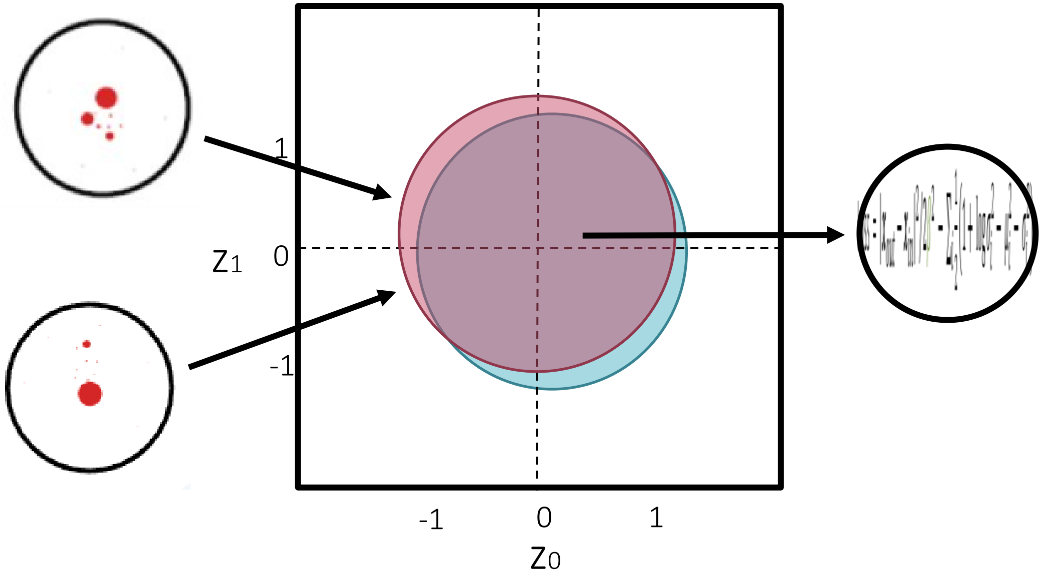
$\boldsymbol{\beta} \to \mathbf{0}$

Info precisely encoded in latent space



$\boldsymbol{\beta} \to \infty$

No info encoded in latent space



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

The Variational Autoencoder:

Information and the loss function

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

1) $\boldsymbol{\beta}$ is 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.

2) $\boldsymbol{\beta}$ is dimensionful

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

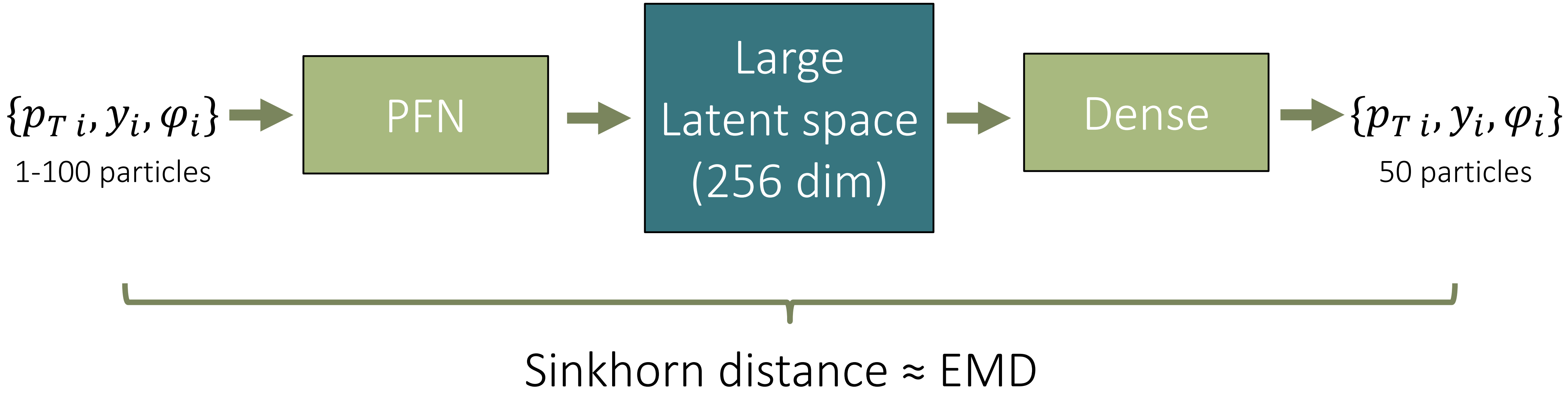
3) $\boldsymbol{\beta}$ is the distance resolution in reconstruction space

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

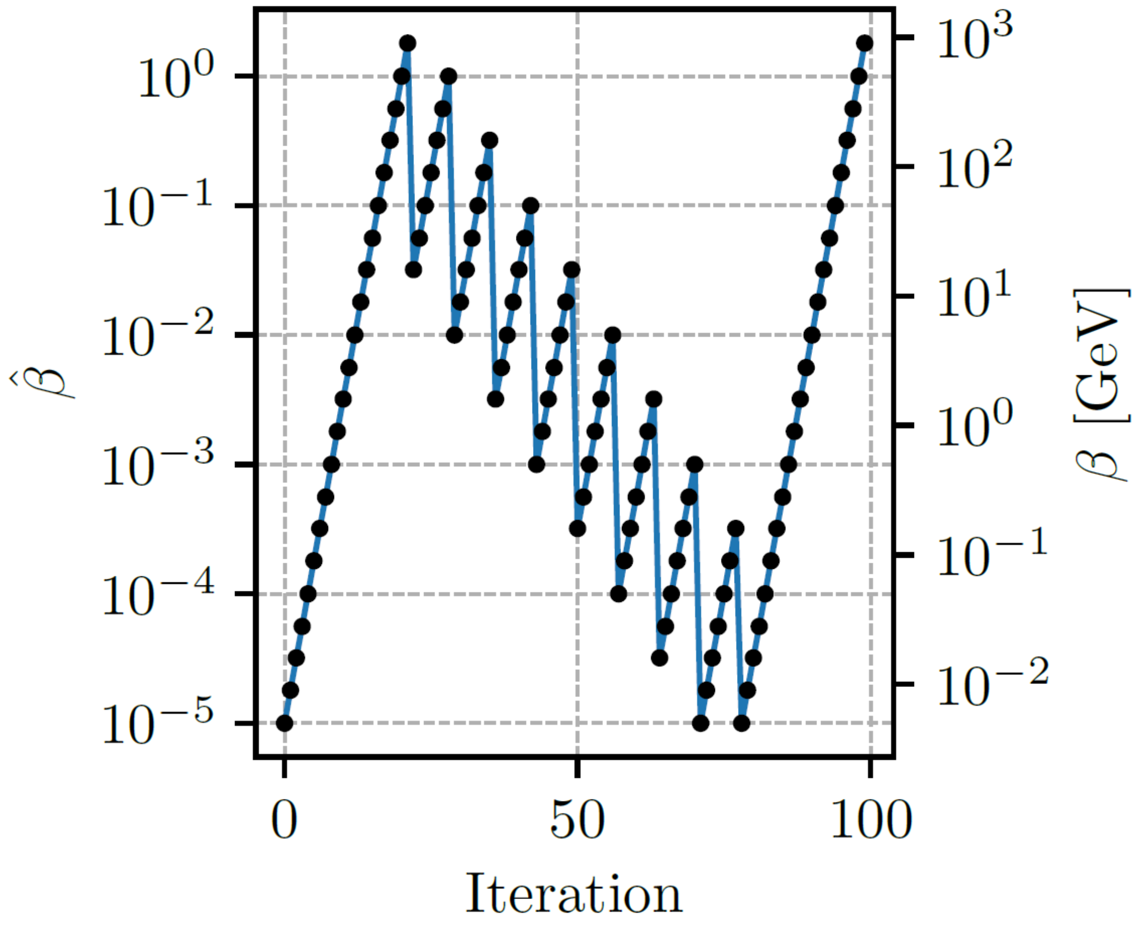
Cheese Course

Application to W Jets

Jet VAE

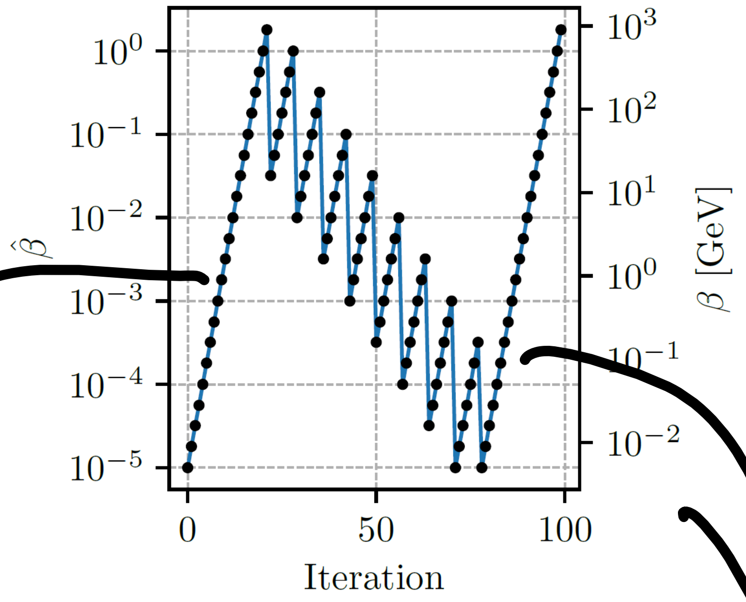

$\{p_{T\,i}, y_i, \varphi_i\}$
1-100 particles
PFN
Large
Latent space
(256 dim)
Dense
$\{p_{T\,i}, y_i, \varphi_i\}$
50 particles
Sinkhorn distance ≈ EMD

Annealing

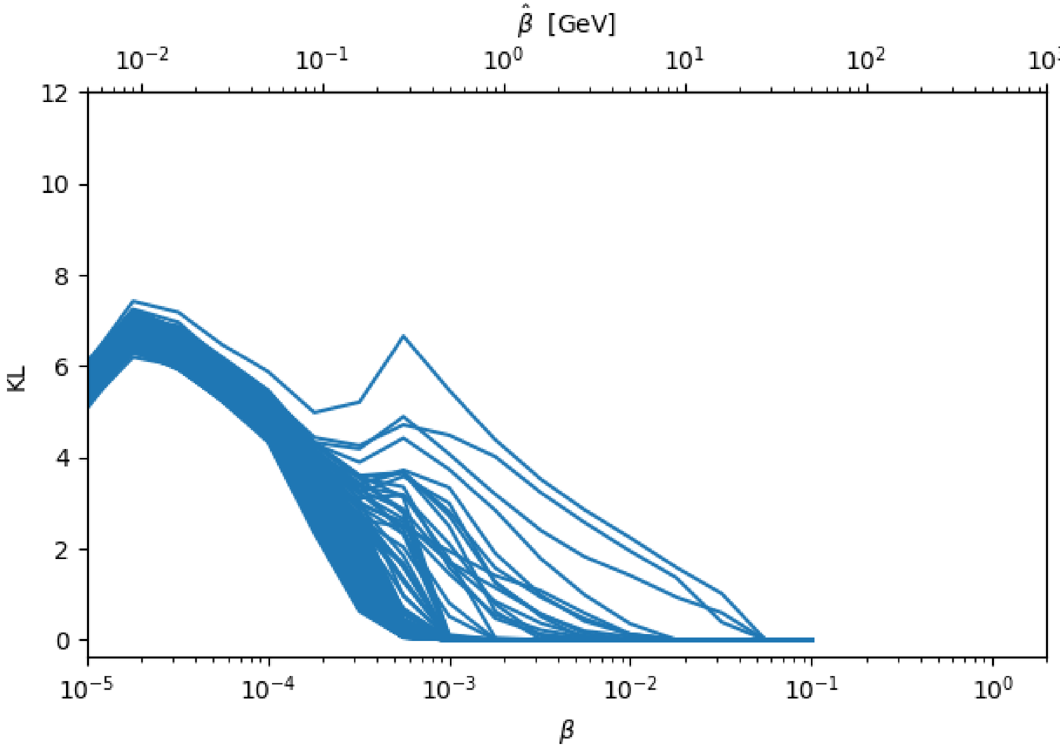


10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\hat{\beta}$
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
β [GeV]
0
50
100
Iteration

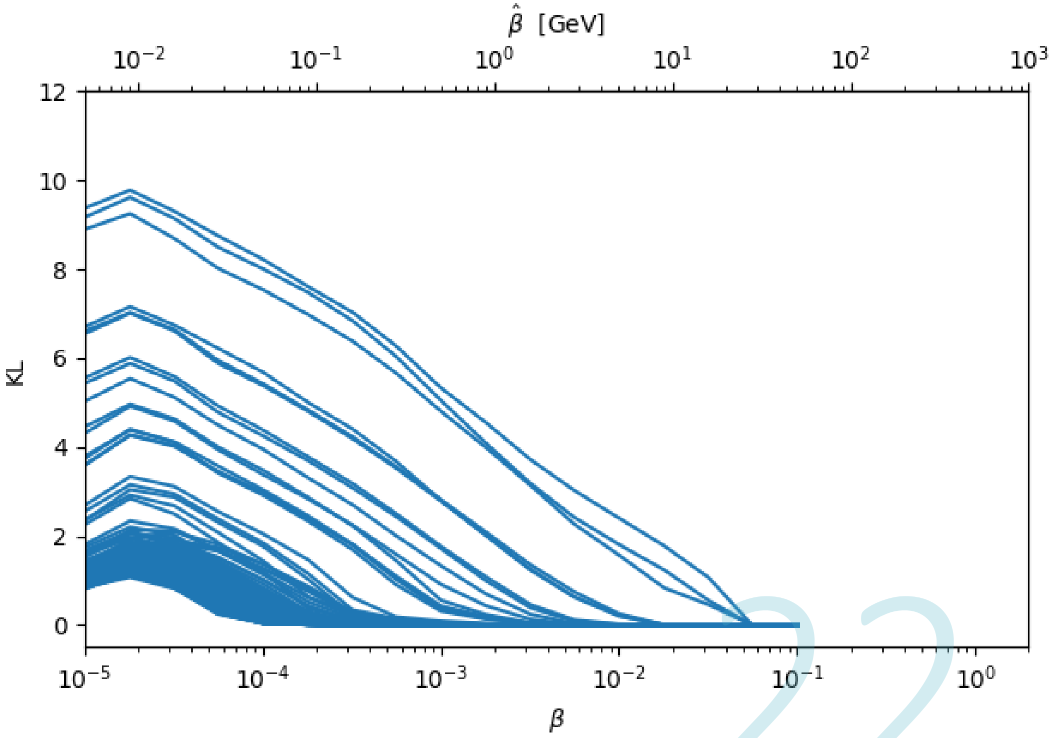
Annealing



10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
$\hat{\beta}$
0
50
100
Iteration
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
β [GeV]



$\hat{\beta}$ [GeV]
10^{-2} 10^{-1} 10^0 10^1 10^2 10^3
KL
0 2 4 6 8 10 12
10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0
β

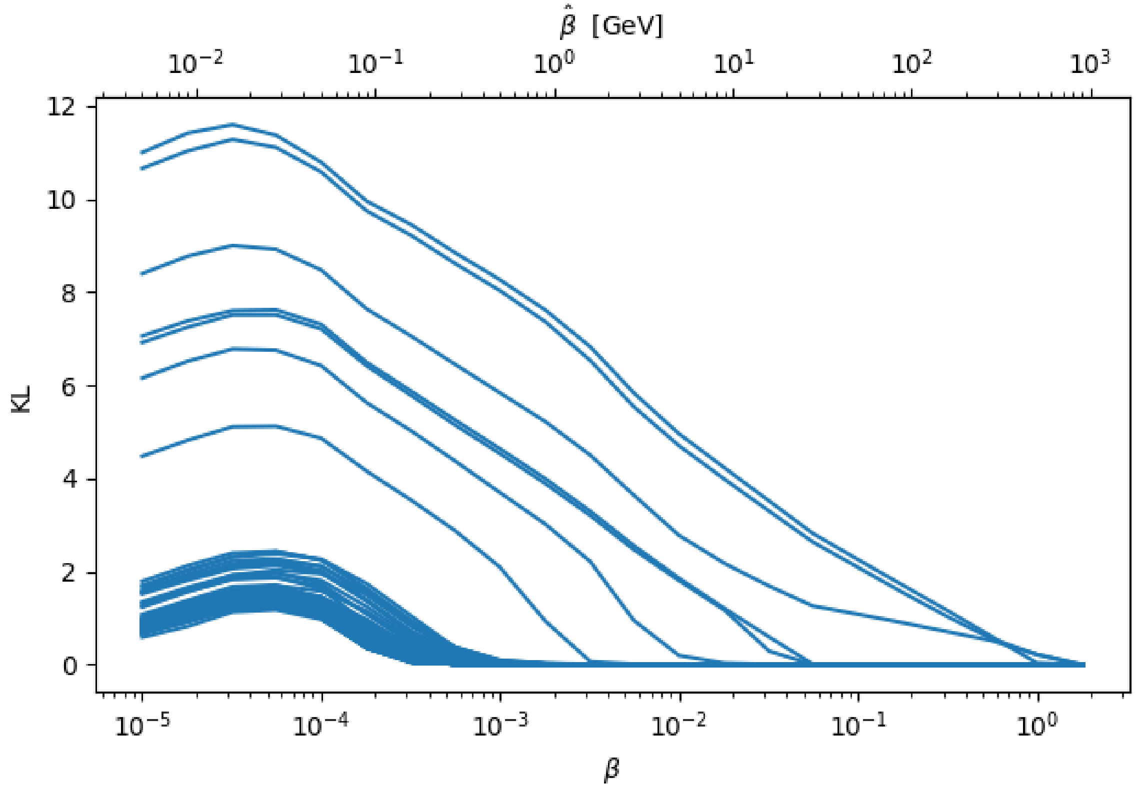


$\hat{\beta}$ [GeV]
10^{-2} 10^{-1} 10^0 10^1 10^2 10^3
KL
0 2 4 6 8 10 12
10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^0
β

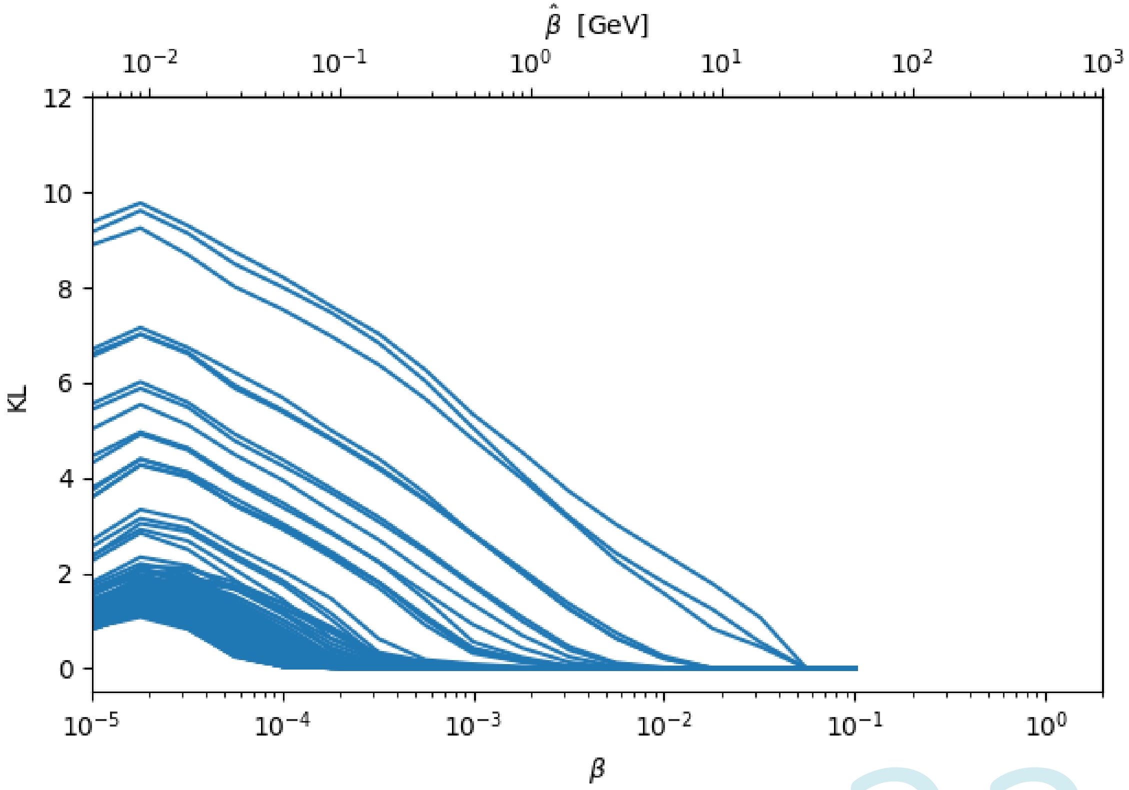
Exploring the Learnt Representation:

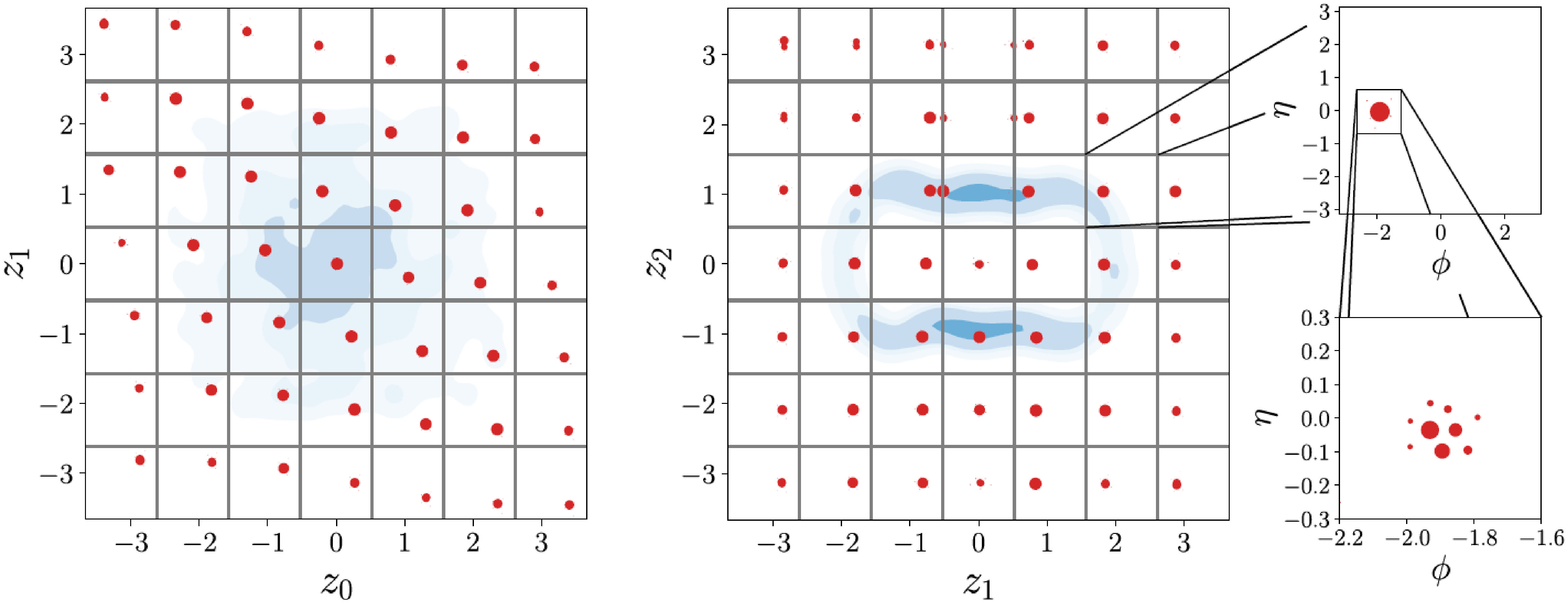
W Jets

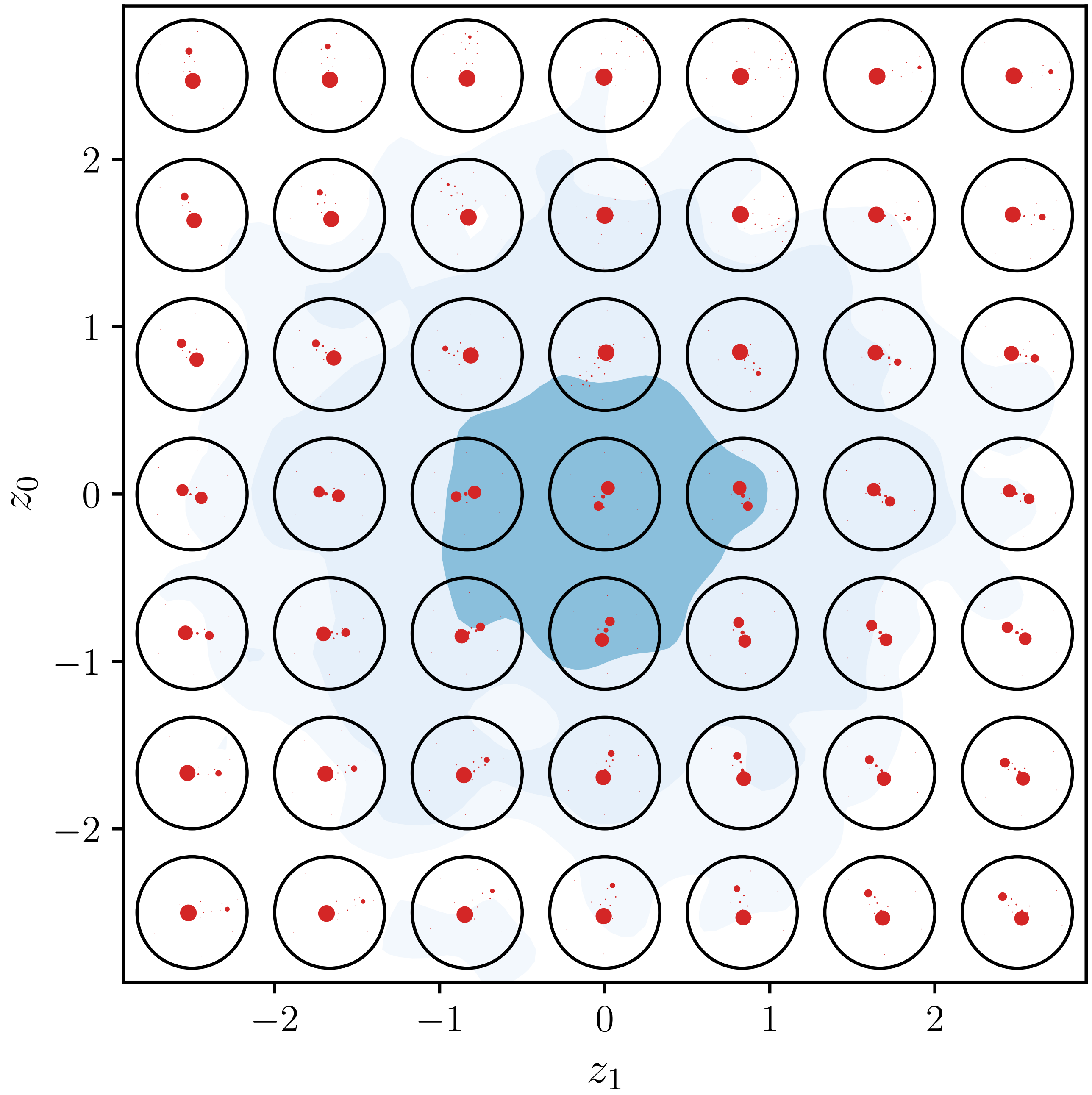
Uncentered W Jets



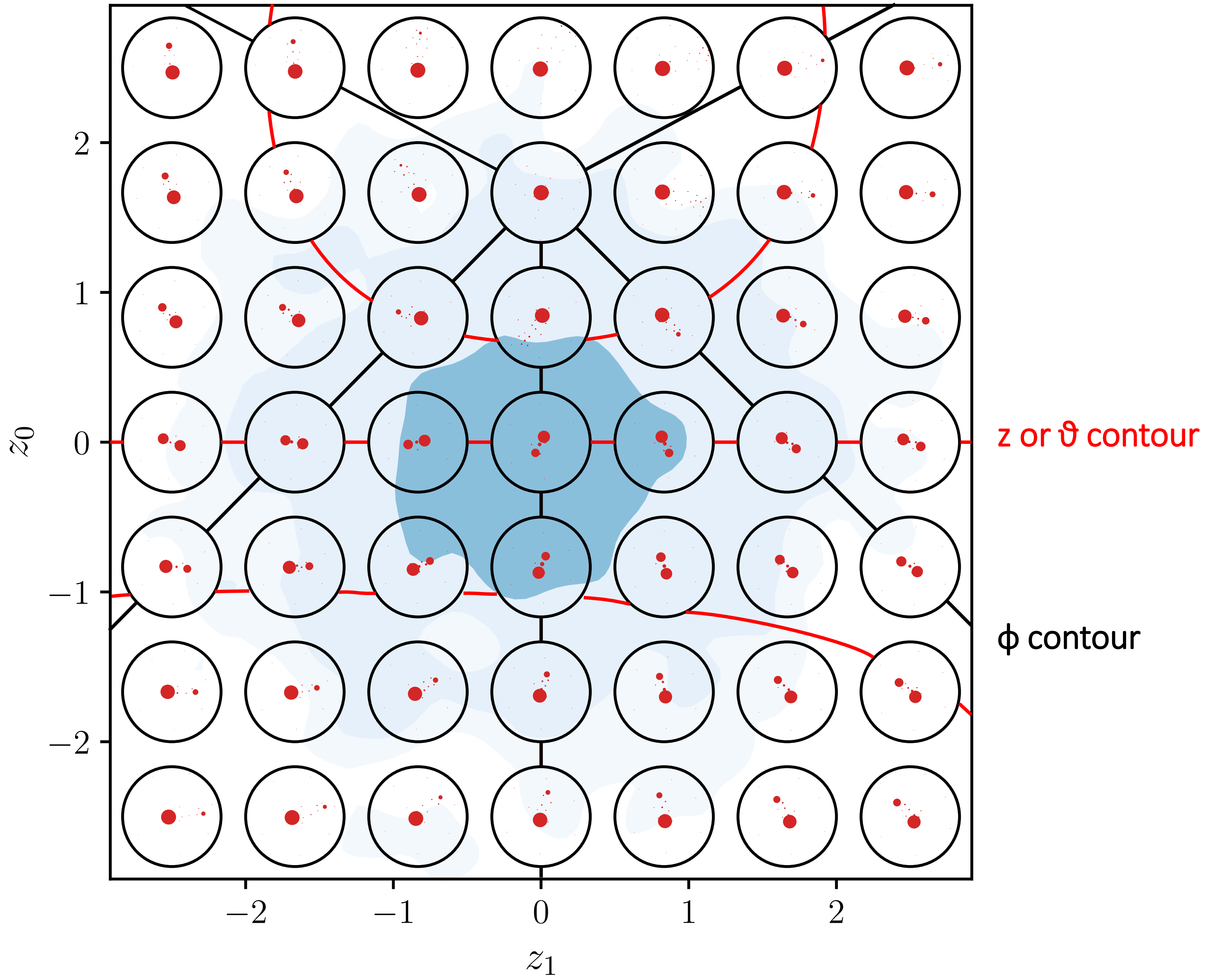
Centered W Jets



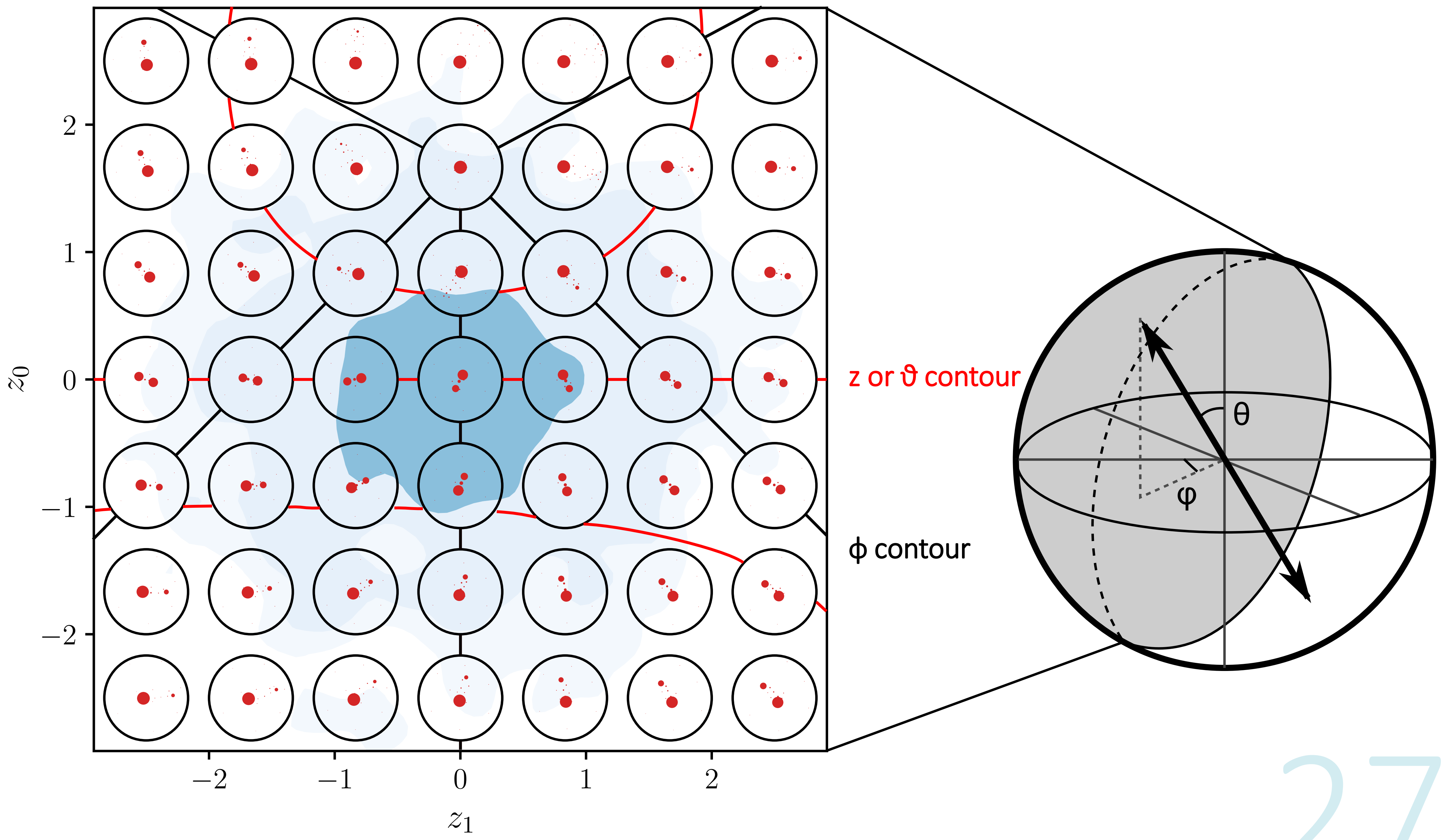

z_0
z_1
z_1
z_2
ϕ
η
ϕ
η



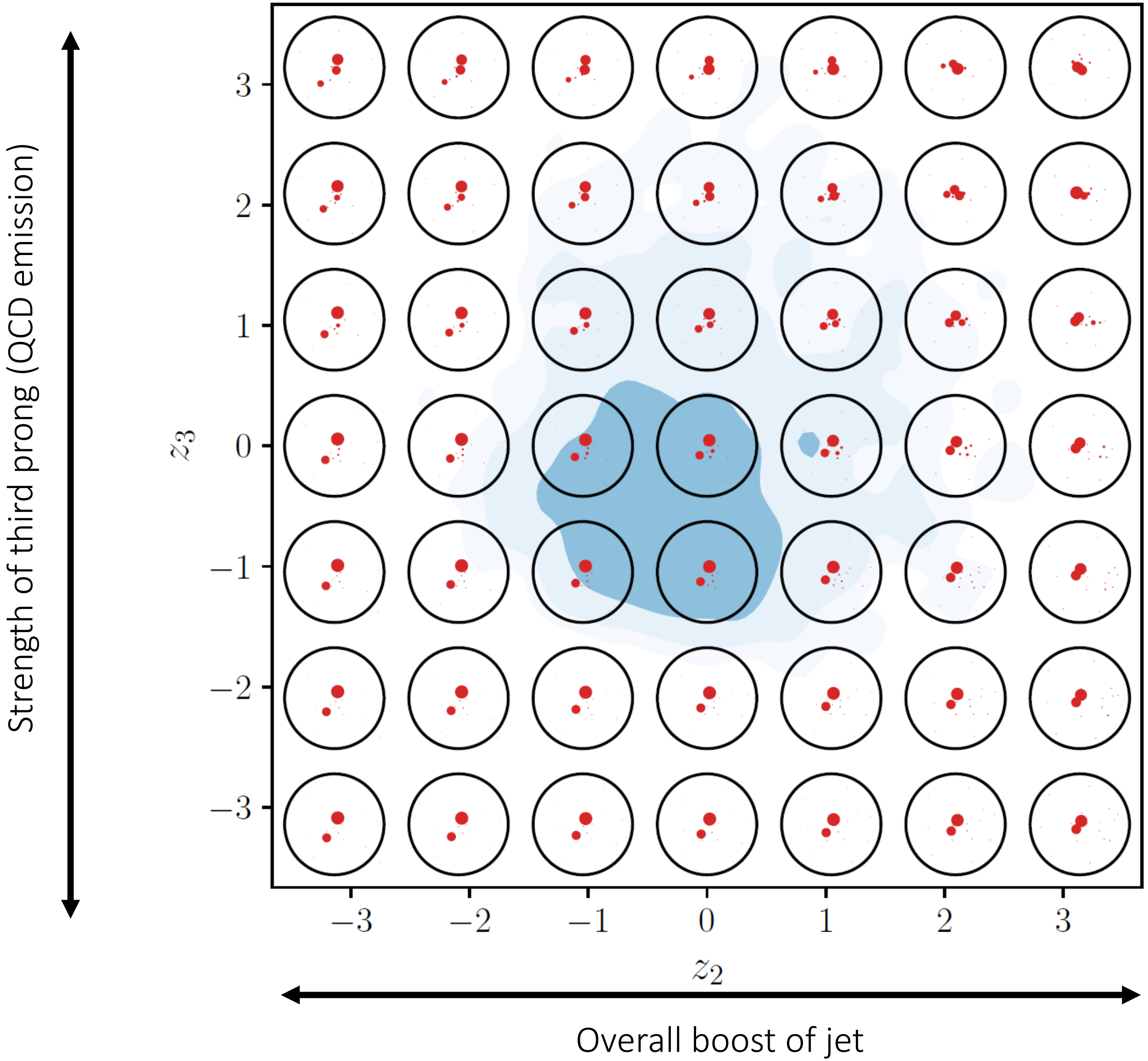
2
1
0
−1
−2
z_0
−2
−1
0
1
2
z_1



2
1
0
−1
−2
z_0
−2
−1
0
1
2
z_1
z or ϑ contour
ϕ contour

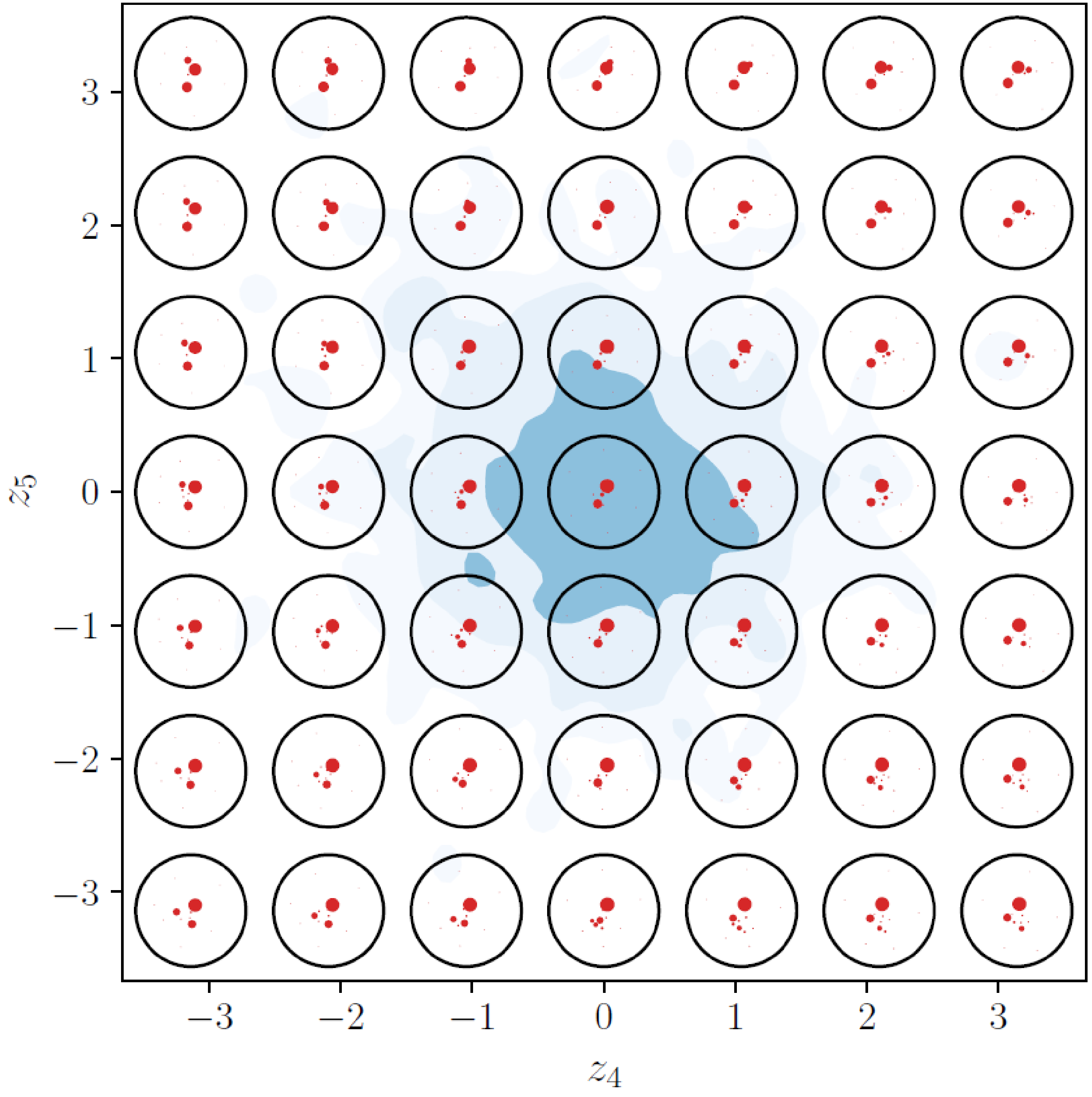


z or ϑ contour
ϕ contour
θ
φ
z_0
z_1
2
1
0
−1
−2
−2
−1
0
1
2



Strength of third prong (QCD emission)
z_3
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
z_2
Overall boost of jet

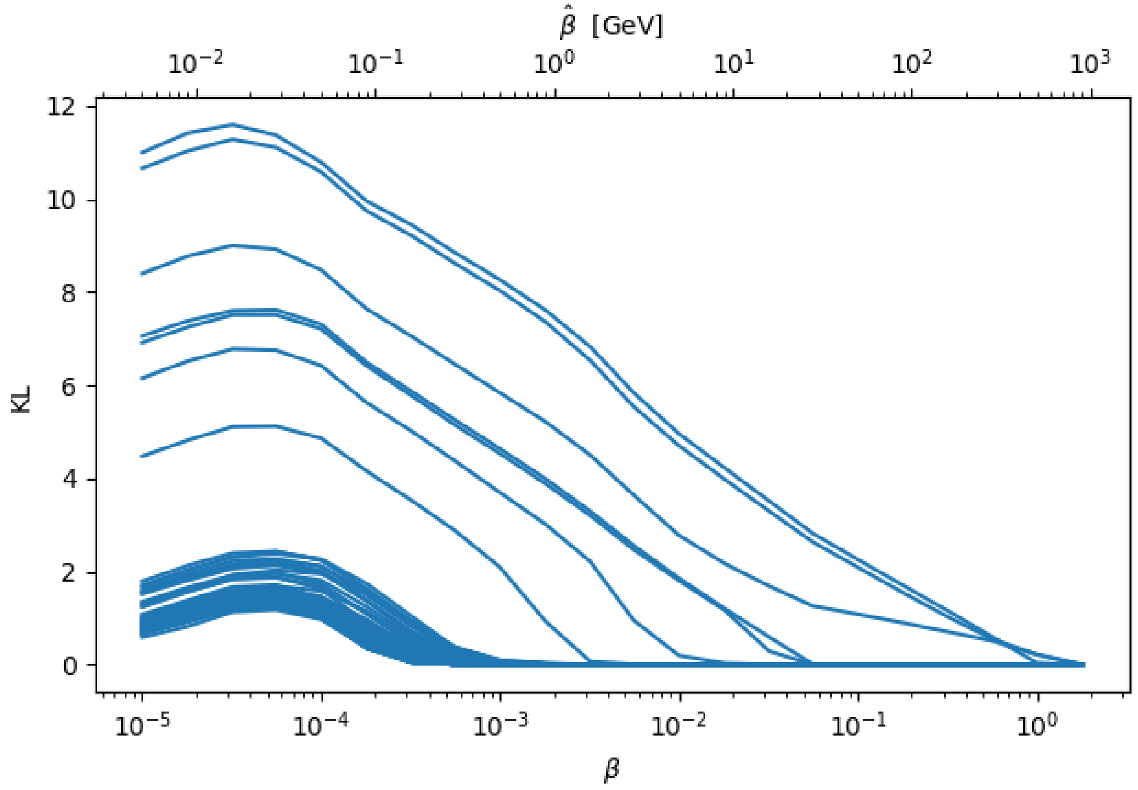
Orientation of third prong (QCD emission)



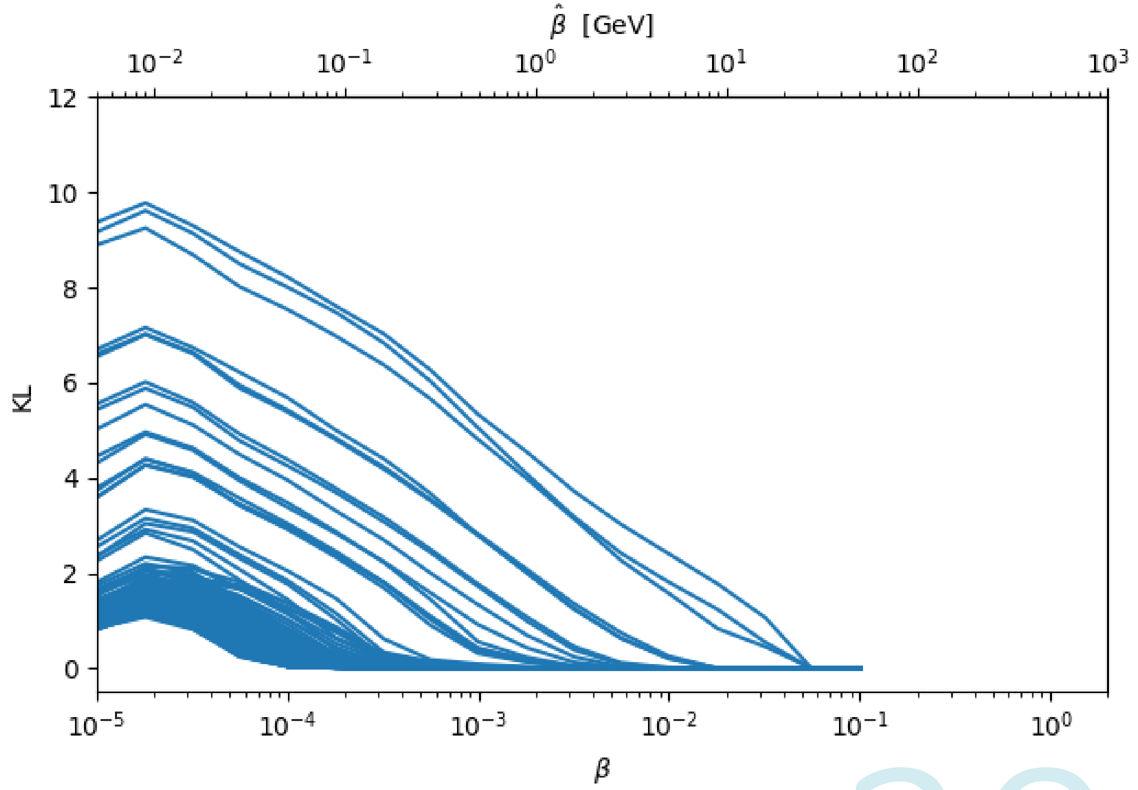
Exploring the Learnt Representation:

W Jets

Uncentered W Jets



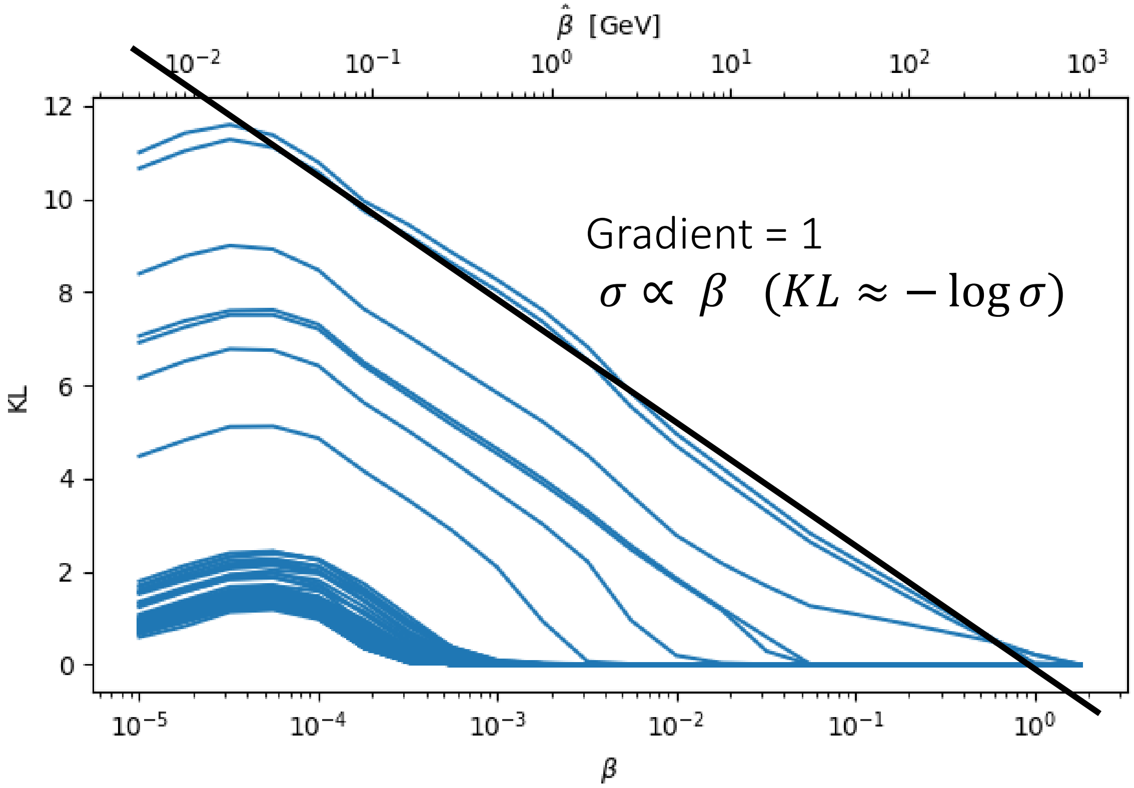
Centered W Jets



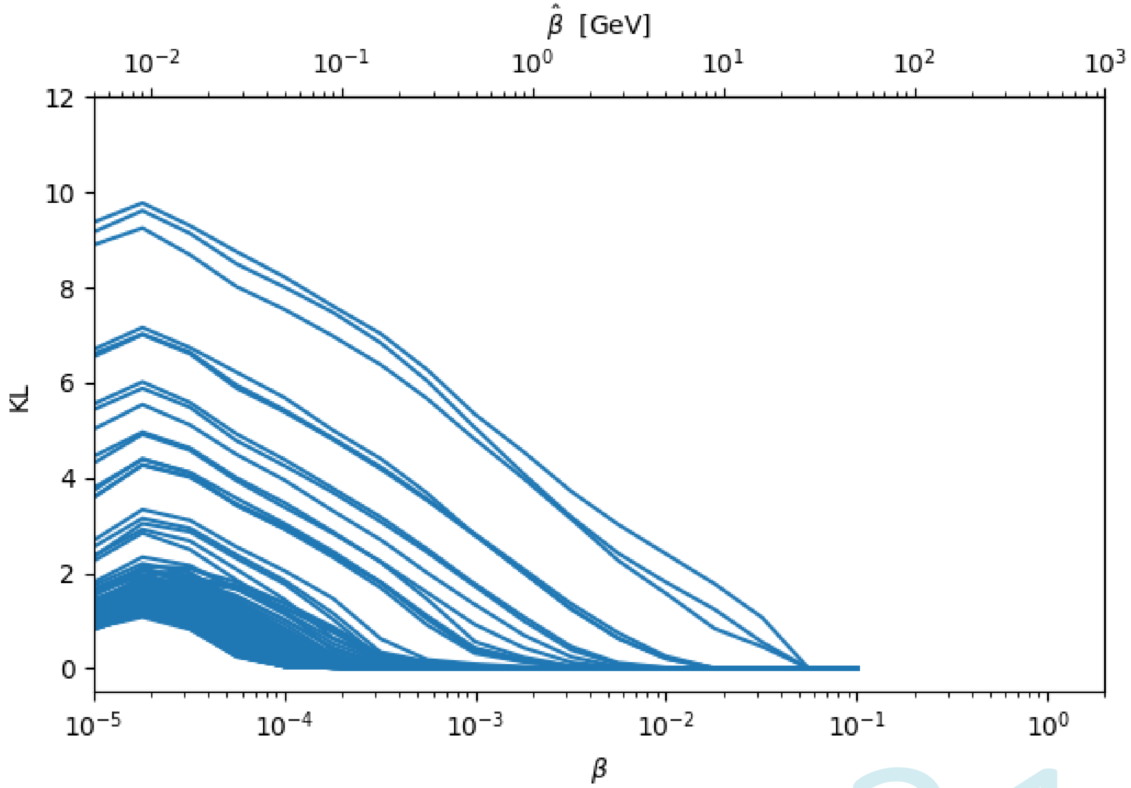
Exploring the Learnt Representation:

W Jets

Uncentered W Jets



Centered W Jets



Dimensionality

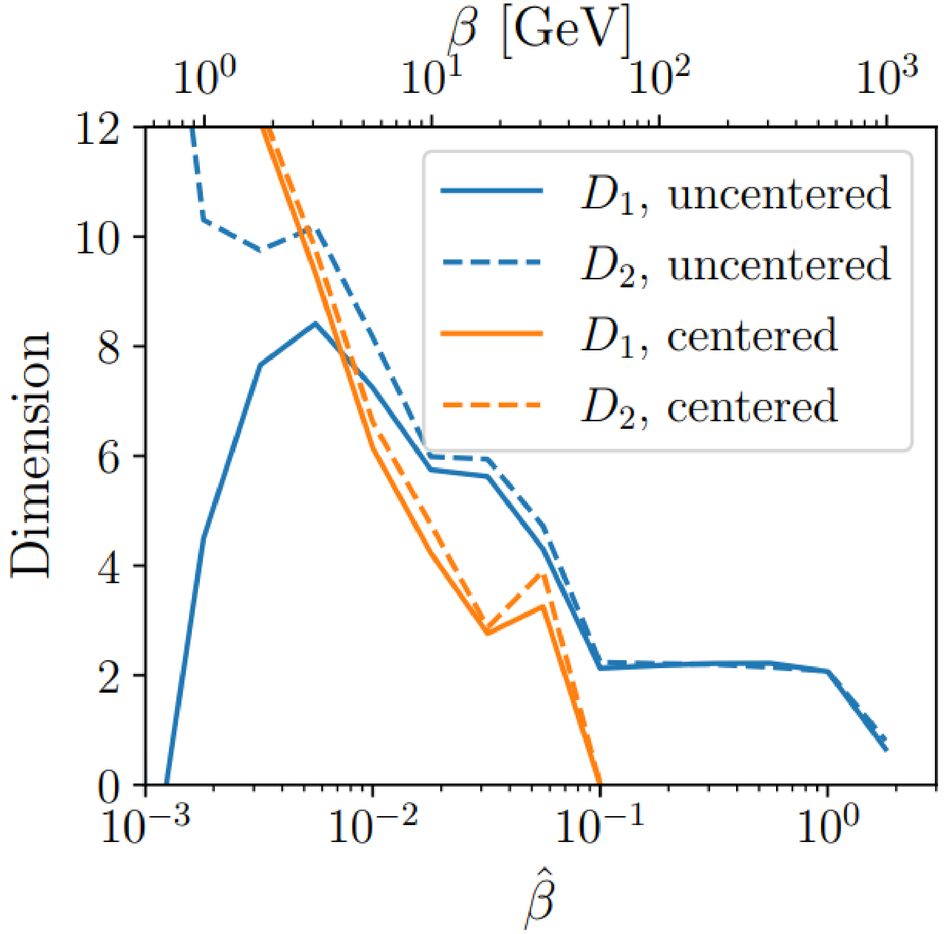
$$D_{corr} \equiv \frac{d \log N}{d \log r}$$

$$D_1 \equiv -\frac{d\;KL}{d \log \beta} \cong \sum_i \frac{d \log \sigma_i}{d \log \beta}$$

Variation of information with scale.

$$D_2 \equiv \frac{d\langle |\Delta \boldsymbol{x}|^2 \rangle}{d\;\beta^2}$$

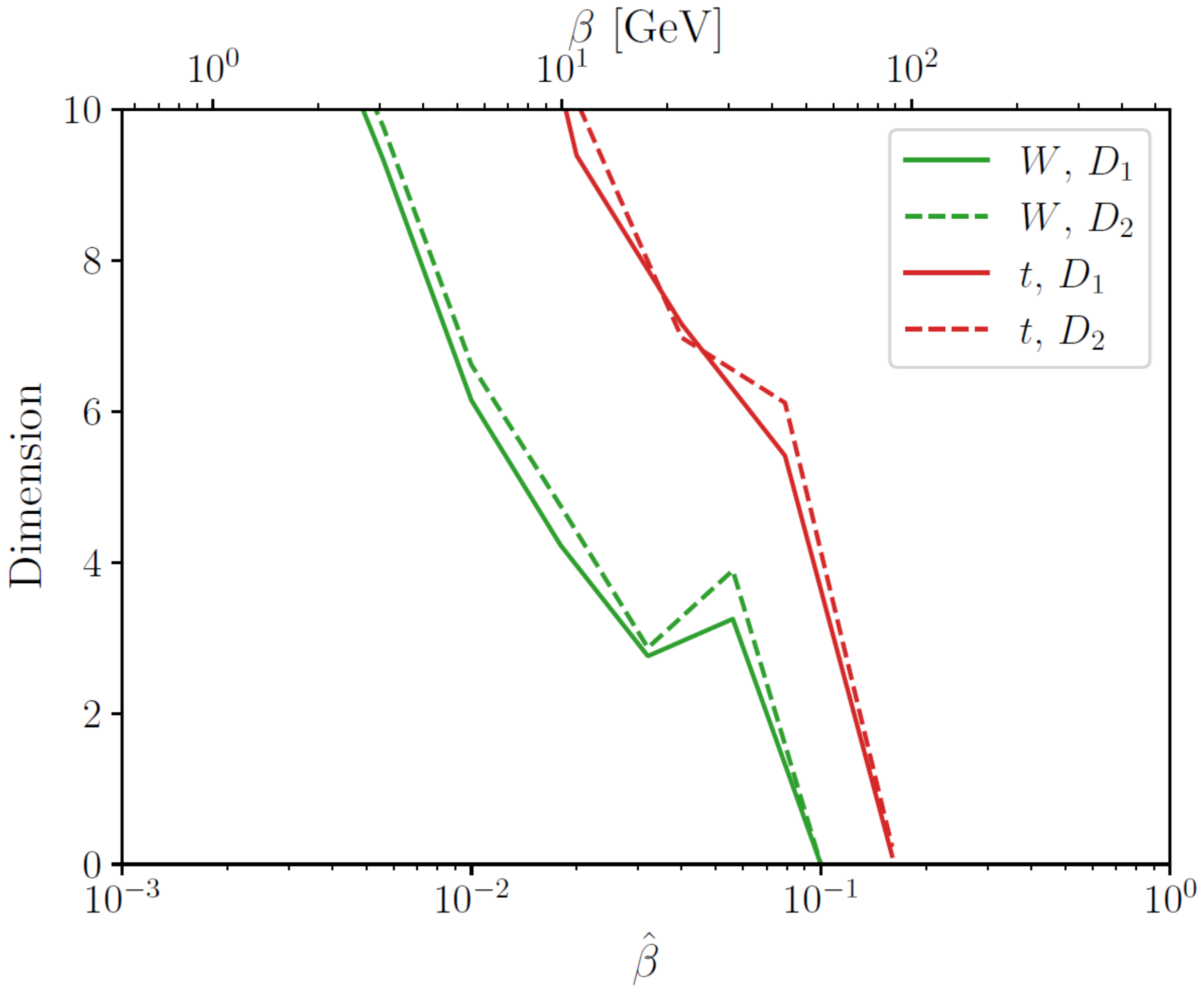
Variation of resolution with scale (think $\langle r^2 \rangle = D\;\sigma^2$ for D-dimensional Gaussian).



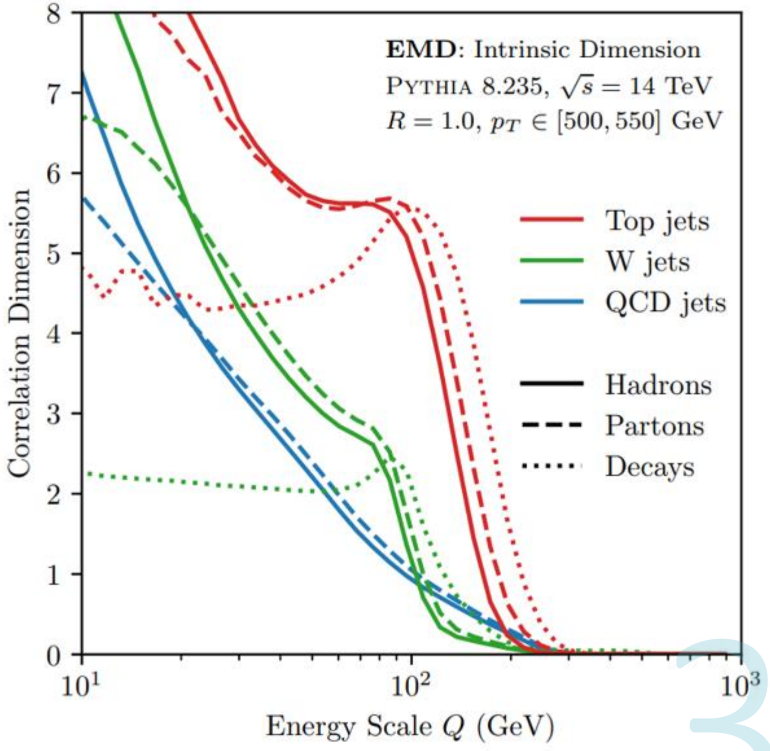
I am still trying to work out formally the meaning of these expressions, but they have an air of truthiness about them and empirically give sensible results.

See also
1810.00597 Danilo Jimenez Rezende, Fabio Viola

Dimensionality

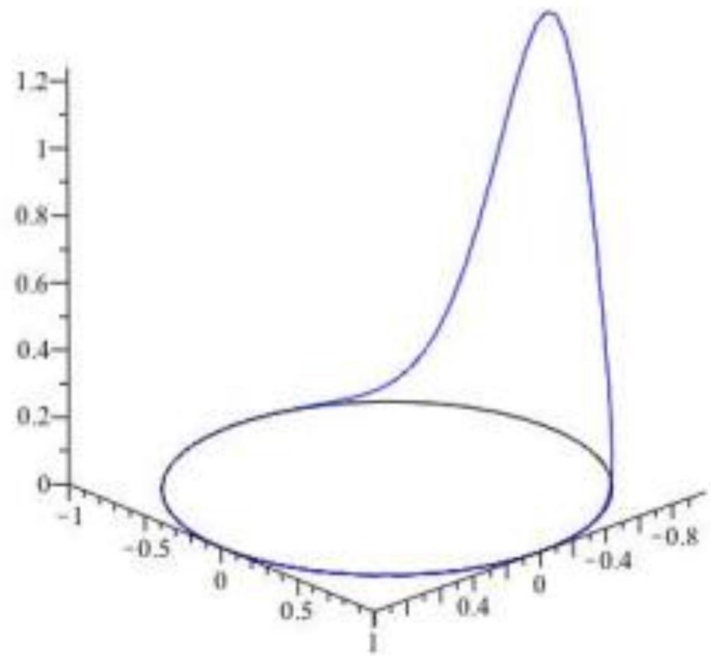


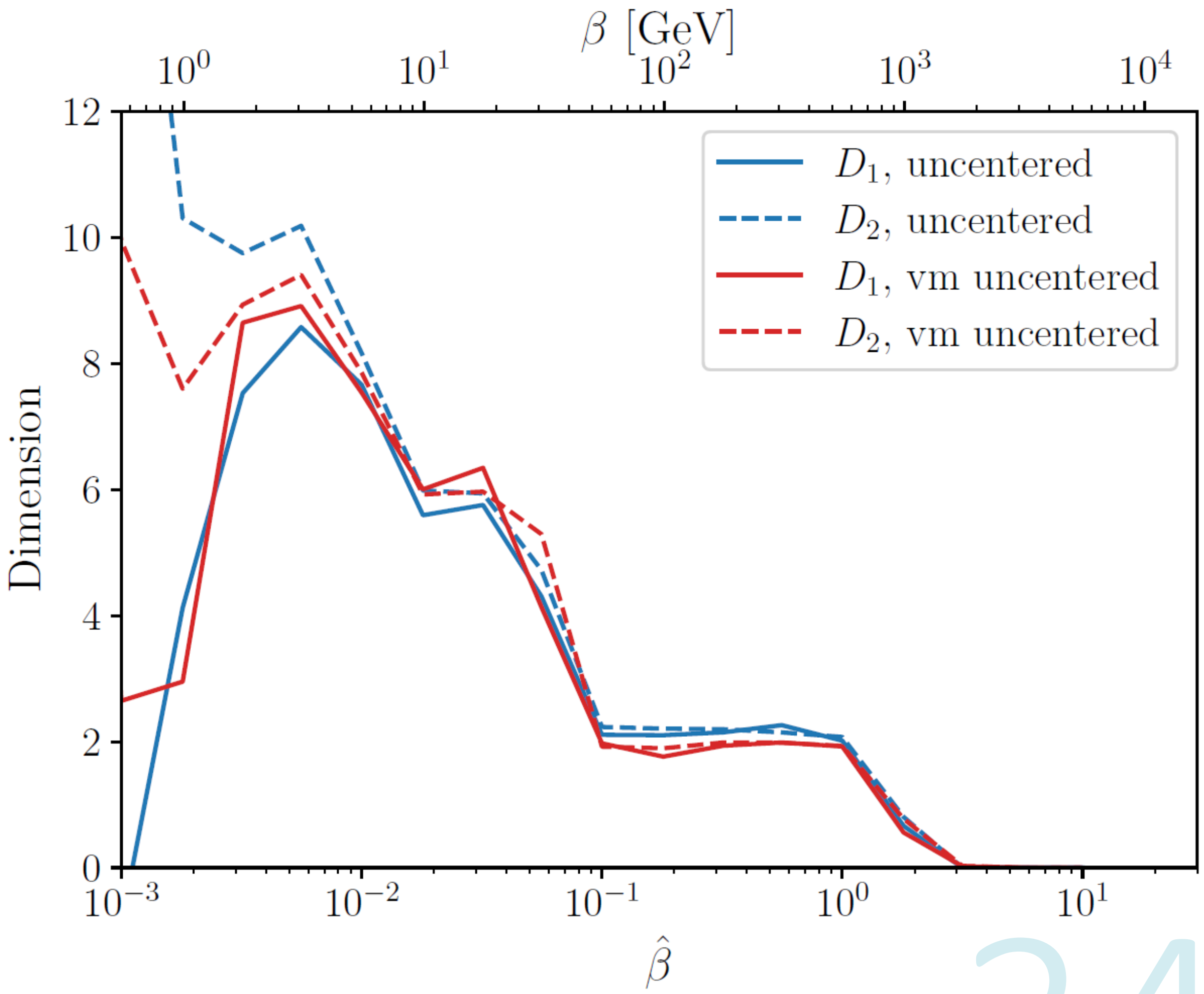
[1902.02346] Komiske, Metodiev, Thaler



Dimensionality

Von Mises Distribution





Dessert

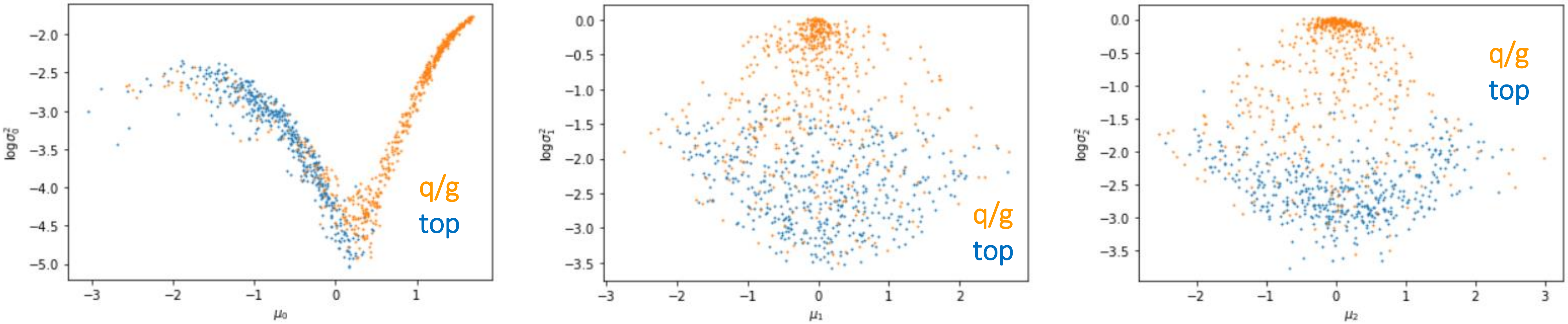
Unsupervised Classification

Mixed Samples

Top and light g/q

Decoder learns:

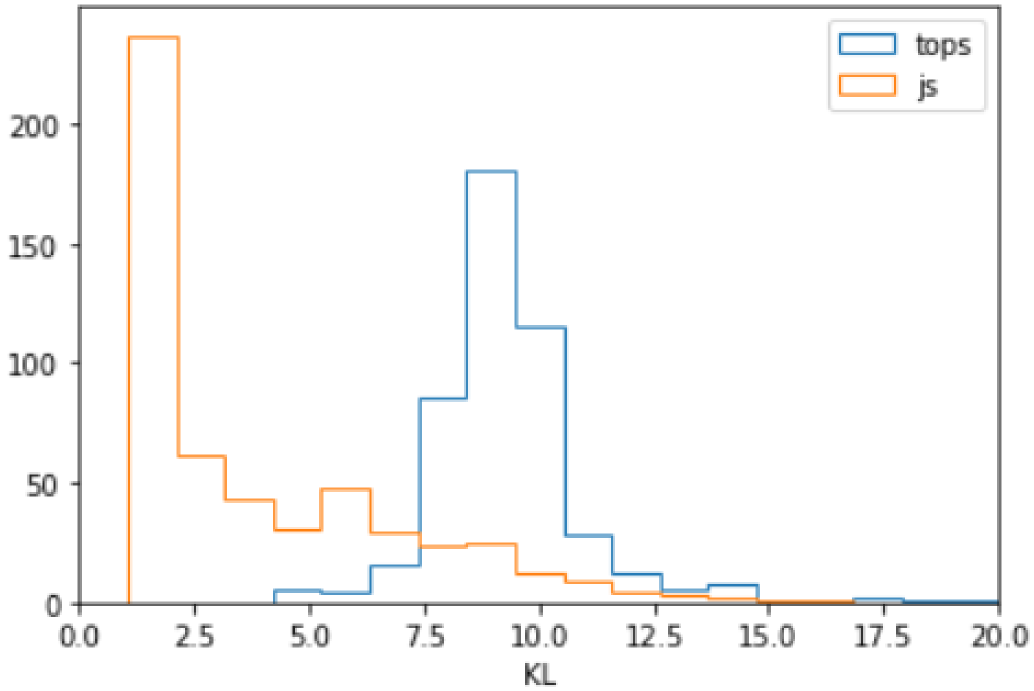
1. If $z_0 > 0$, then it is a light jet and ignore the substructure information in z_1, z_2, etc.
2. If $z_0 < 0$, then it is a top jet, and get three-prong substructure from z_1, z_2, etc.

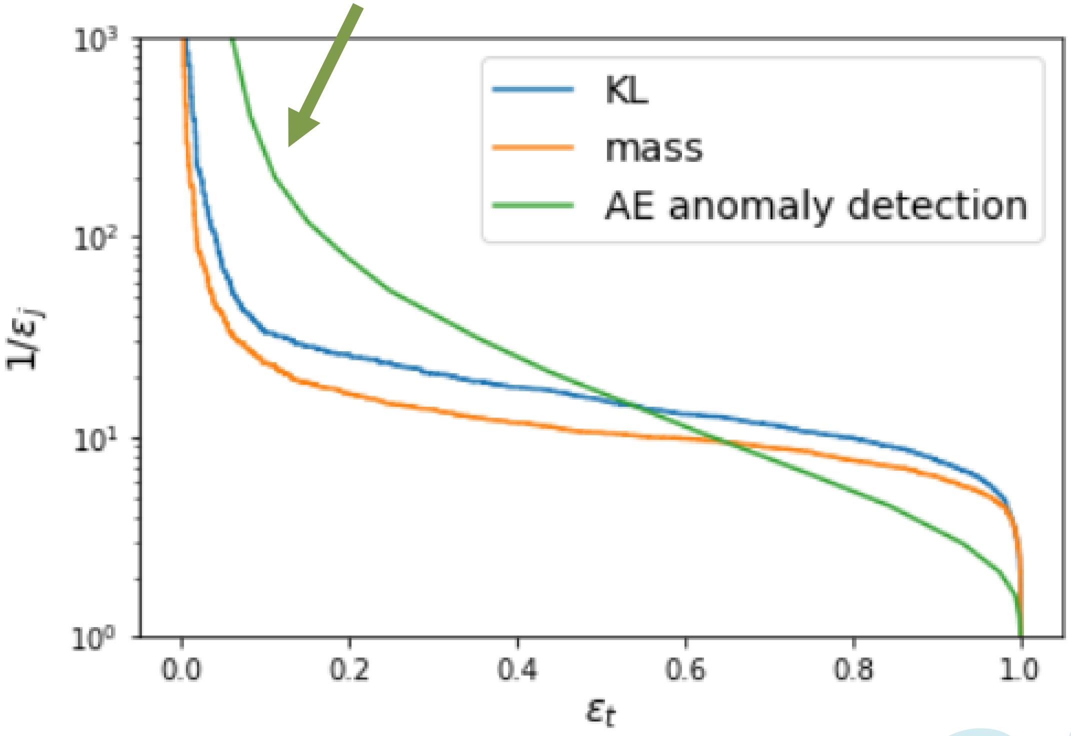


Mixed Samples

Top and light g/q

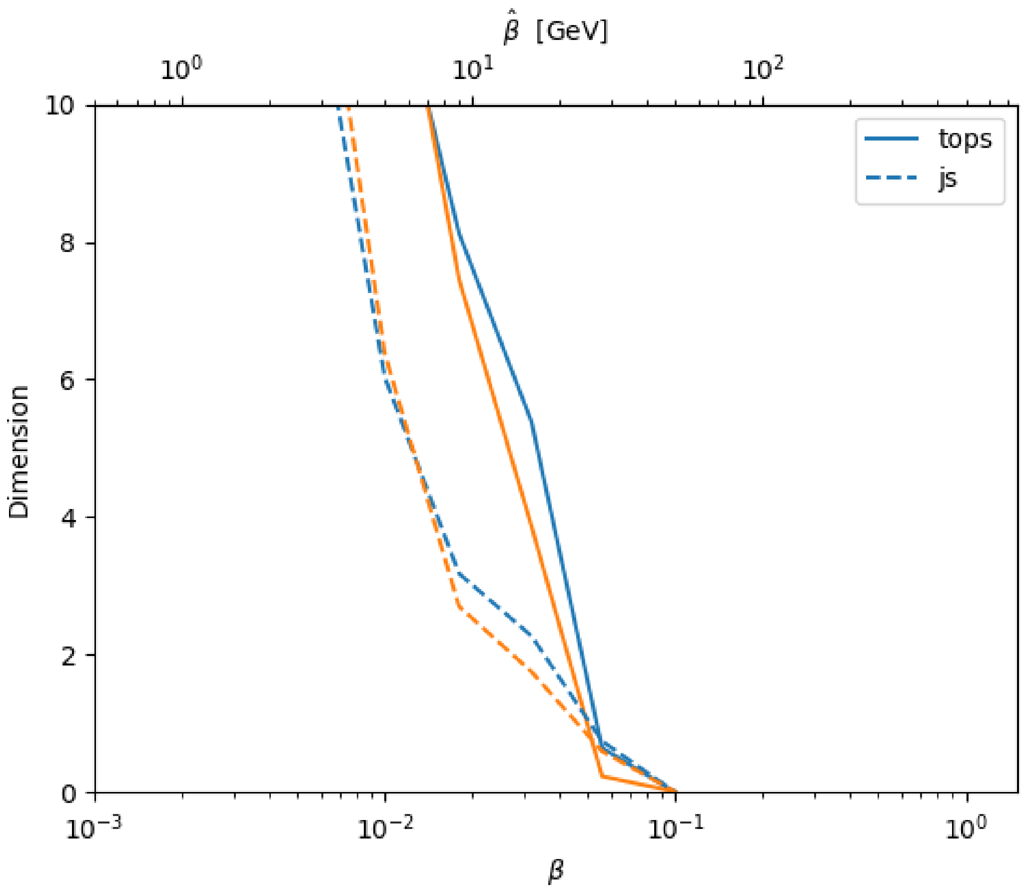
[1808.08979] T. Heimel, G. Kasieczka, T. Plehn, J. M. Thompson

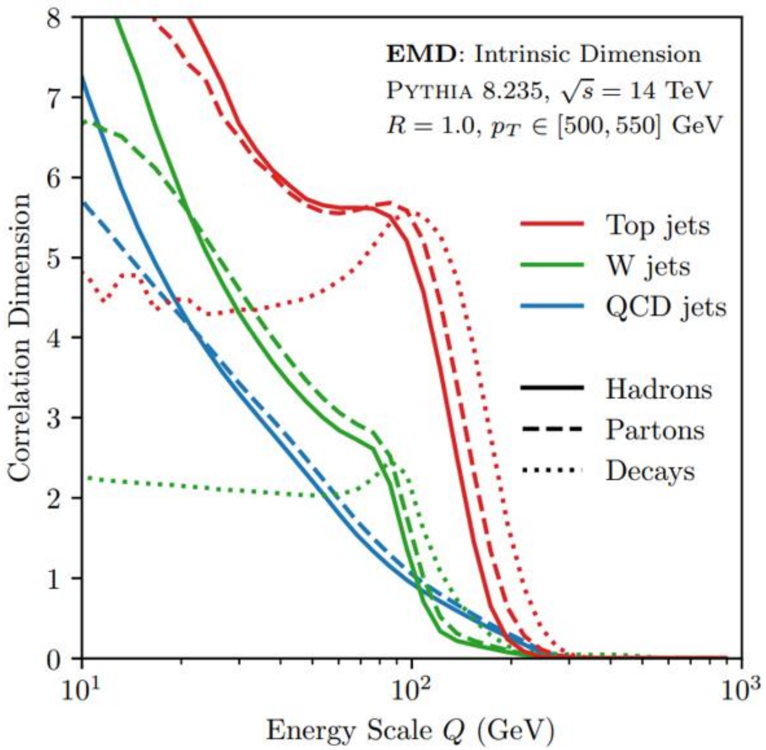




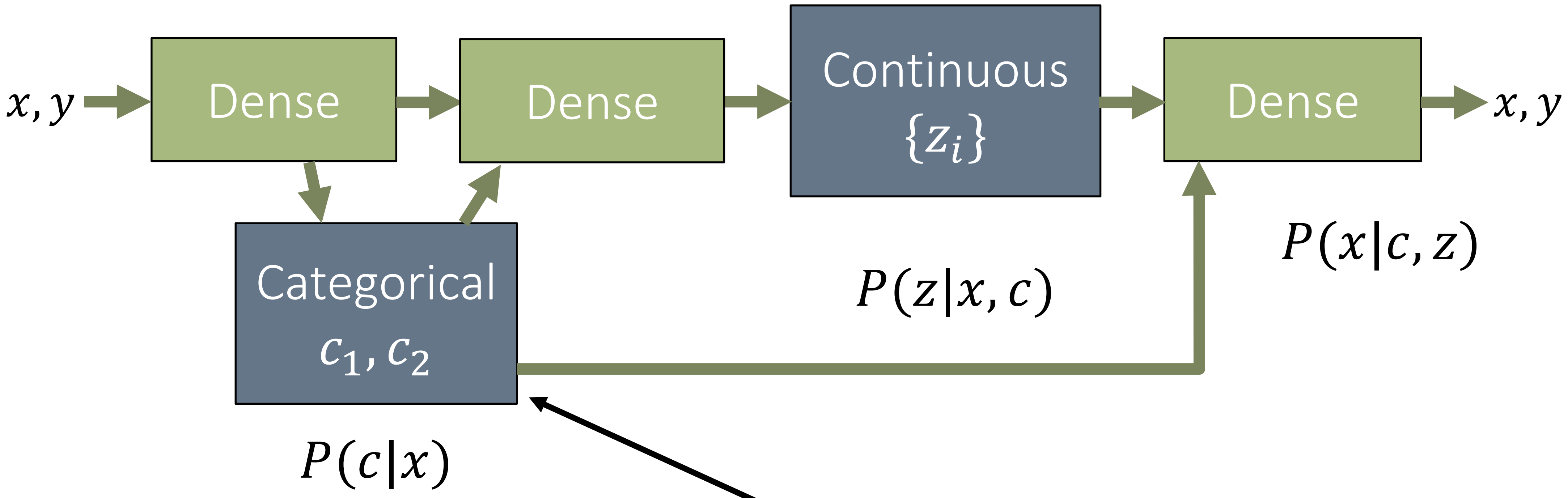
Mixed Samples

Top and light g/q





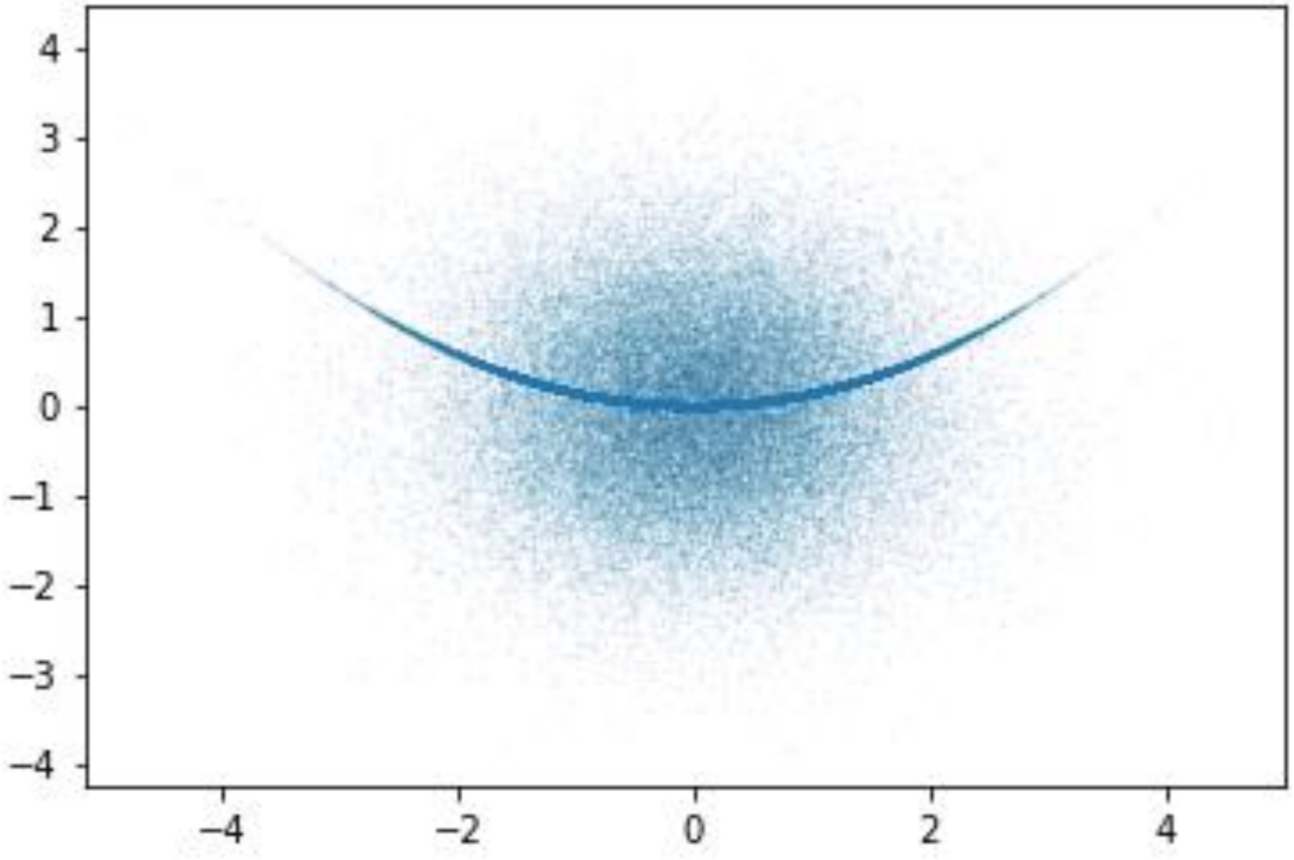
Mixed samples: Another idea



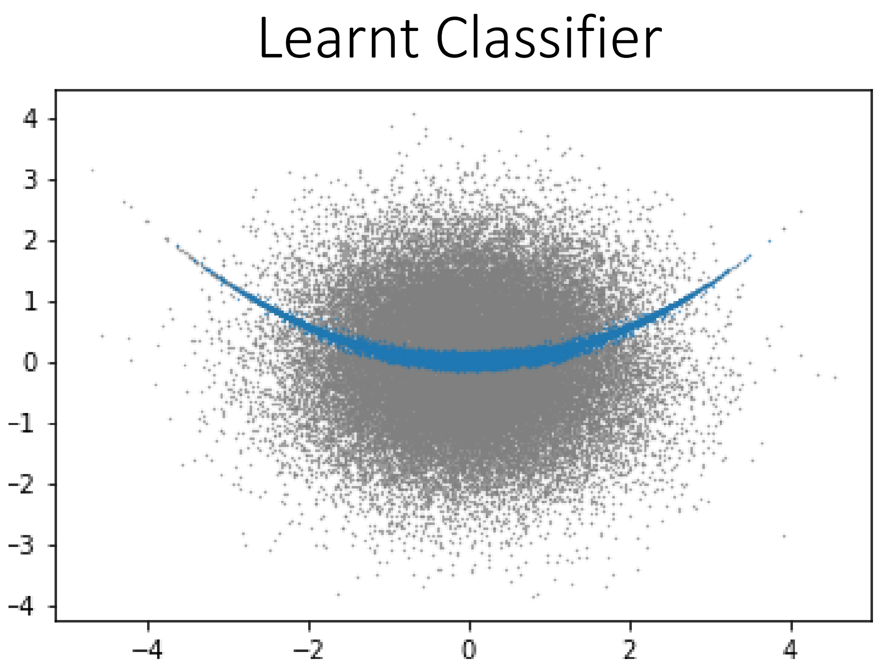
$$Loss = \frac{|\Delta \boldsymbol{x}|^2}{2\,\beta^2} + KL_{\text{Gauss}} + \alpha KL_{\text{Cat}}$$

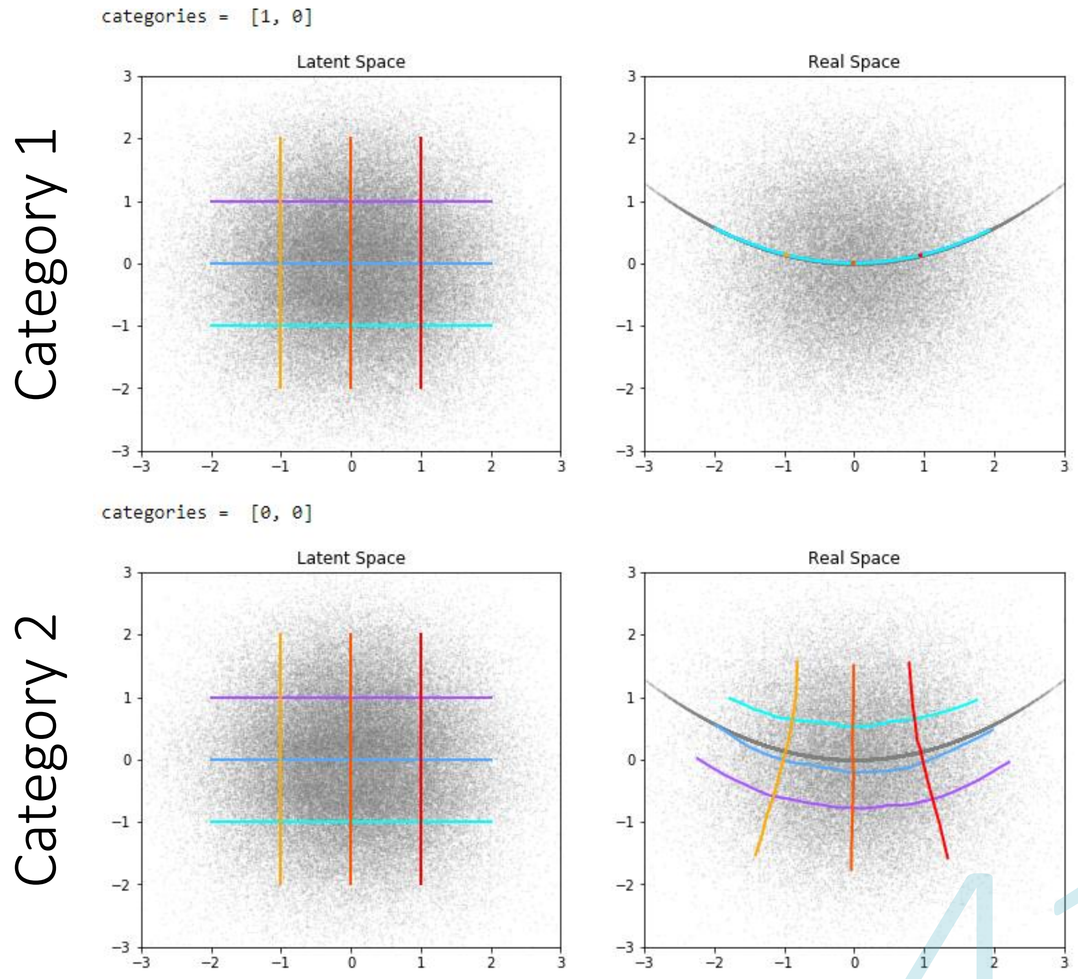
[1611.01144 Eric Jang, Shixiang Gu, Ben Poole]
[1611.00712 Chris J. Maddison, Andriy Mnih, Yee Whye Teh]

Mixed samples: Another idea

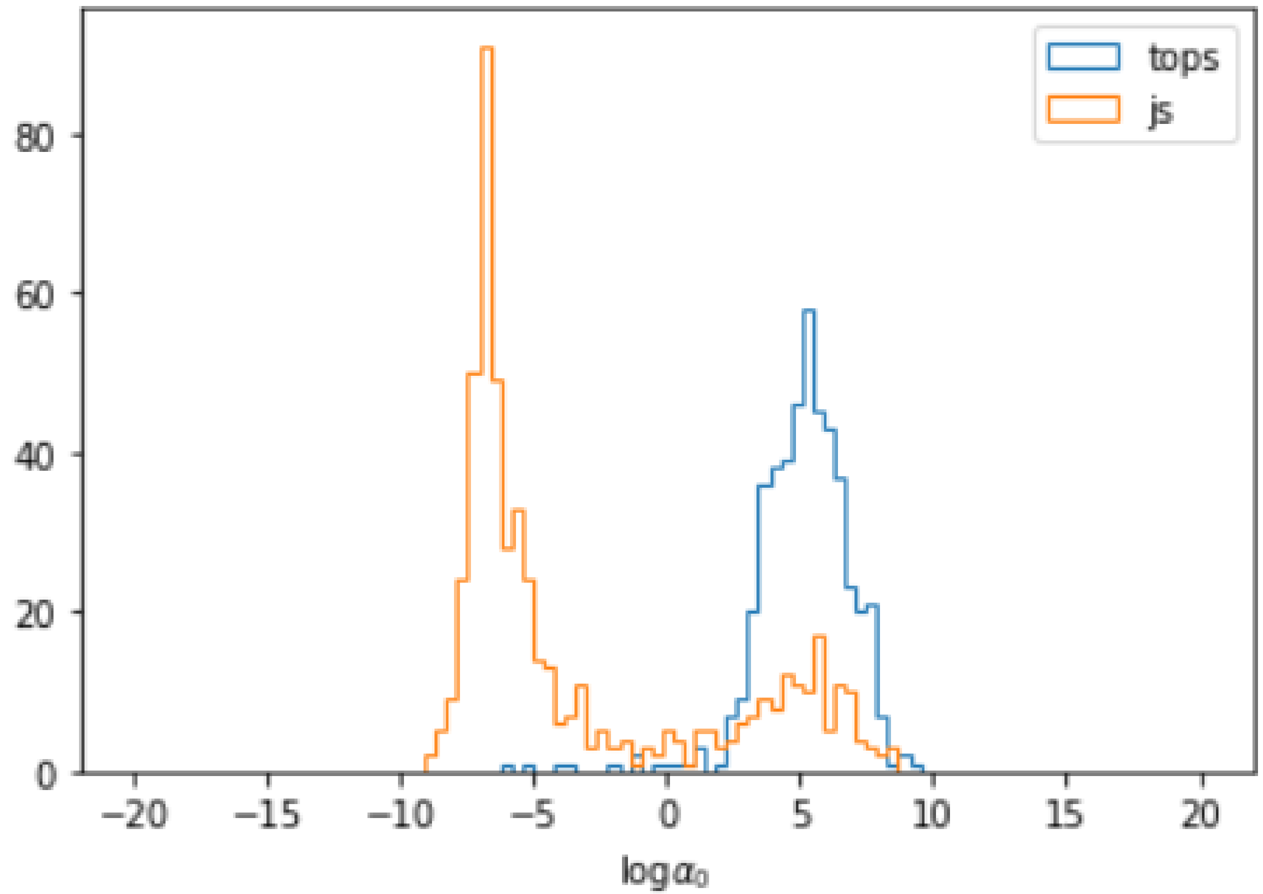


Mixed samples: Another idea





Mixed samples: Another idea



Digestif

Conclusions

VAE latent spaces learn concrete representations of the manifolds on which they are trained.

A meaningful distance metric which encodes interesting physics at different scales leads to a meaningful learnt representation which encodes interesting physics at different scales.

For a sufficiently simple manifold, the VAE learnt representation is:

- *Orthogonalized*
- *Hierarchically organized*
- *Has a scale-dependent fractal dimension which directly relates to that of the true data manifold*

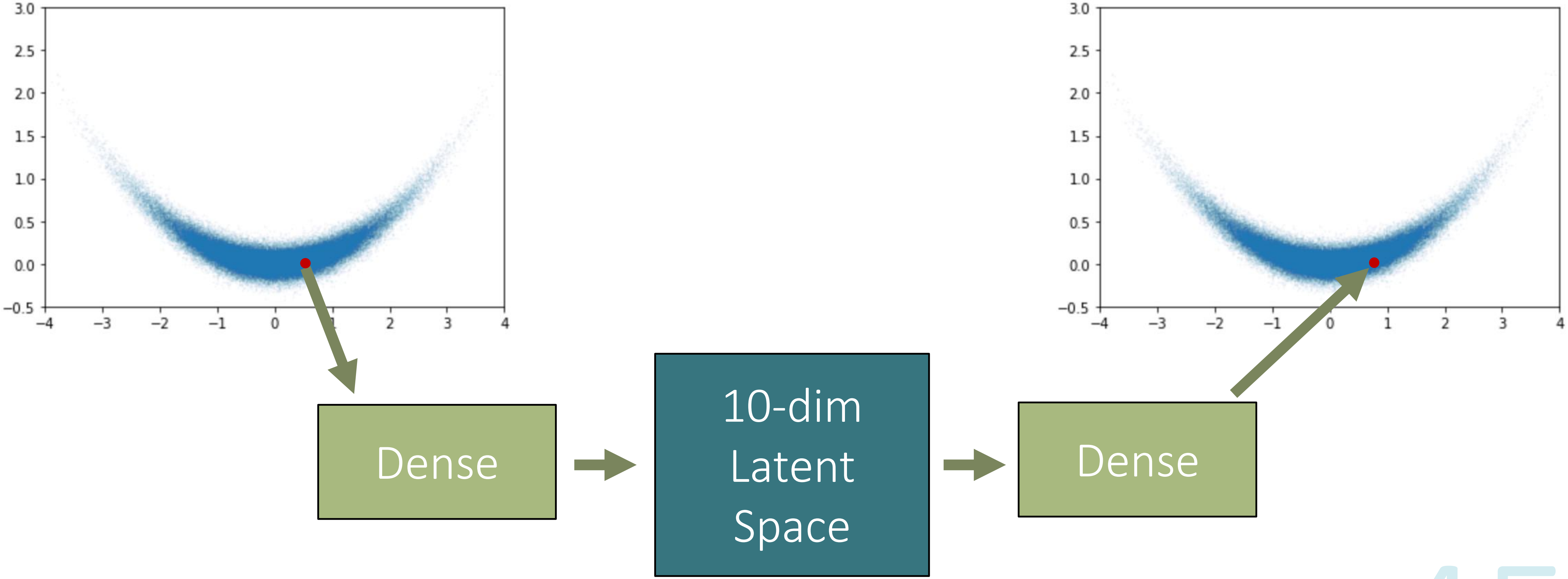
These properties are due to the demand to be ***parsimonious*** with information.

Special thanks to



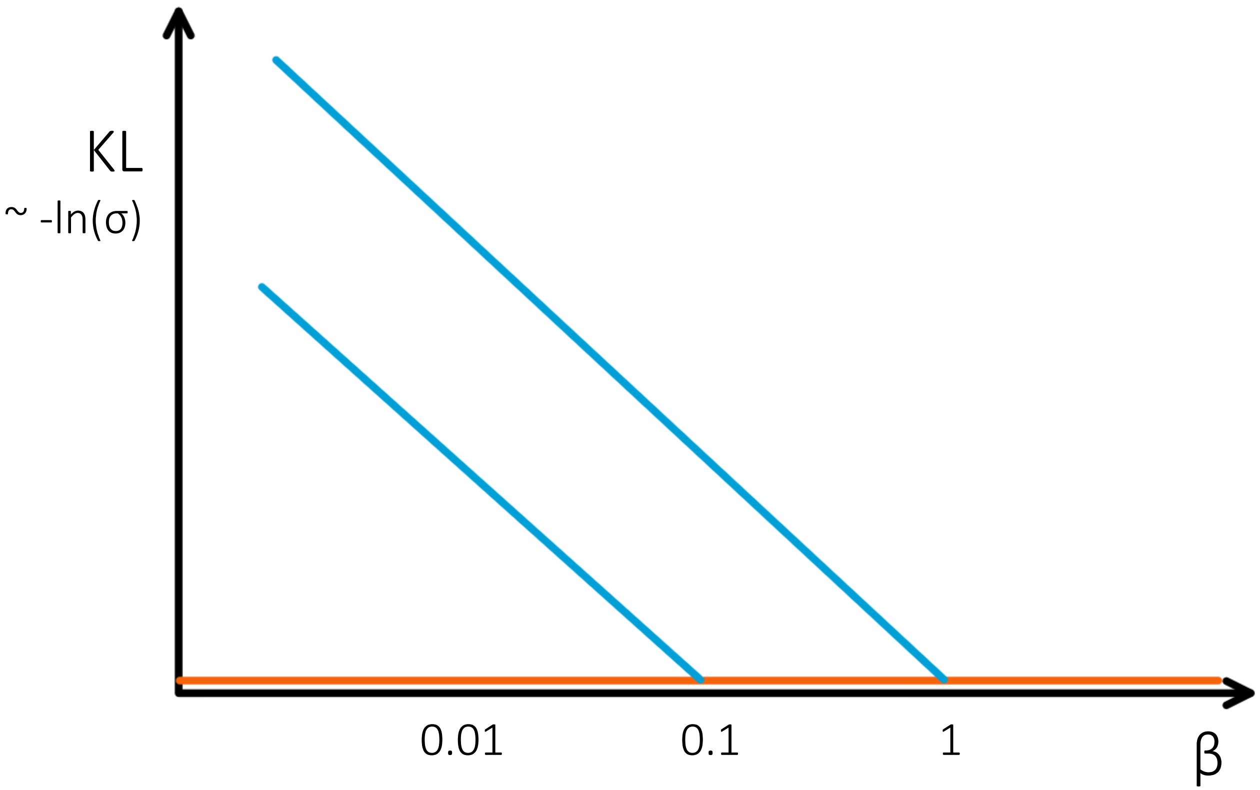

The Variational Autoencoder

Bananas



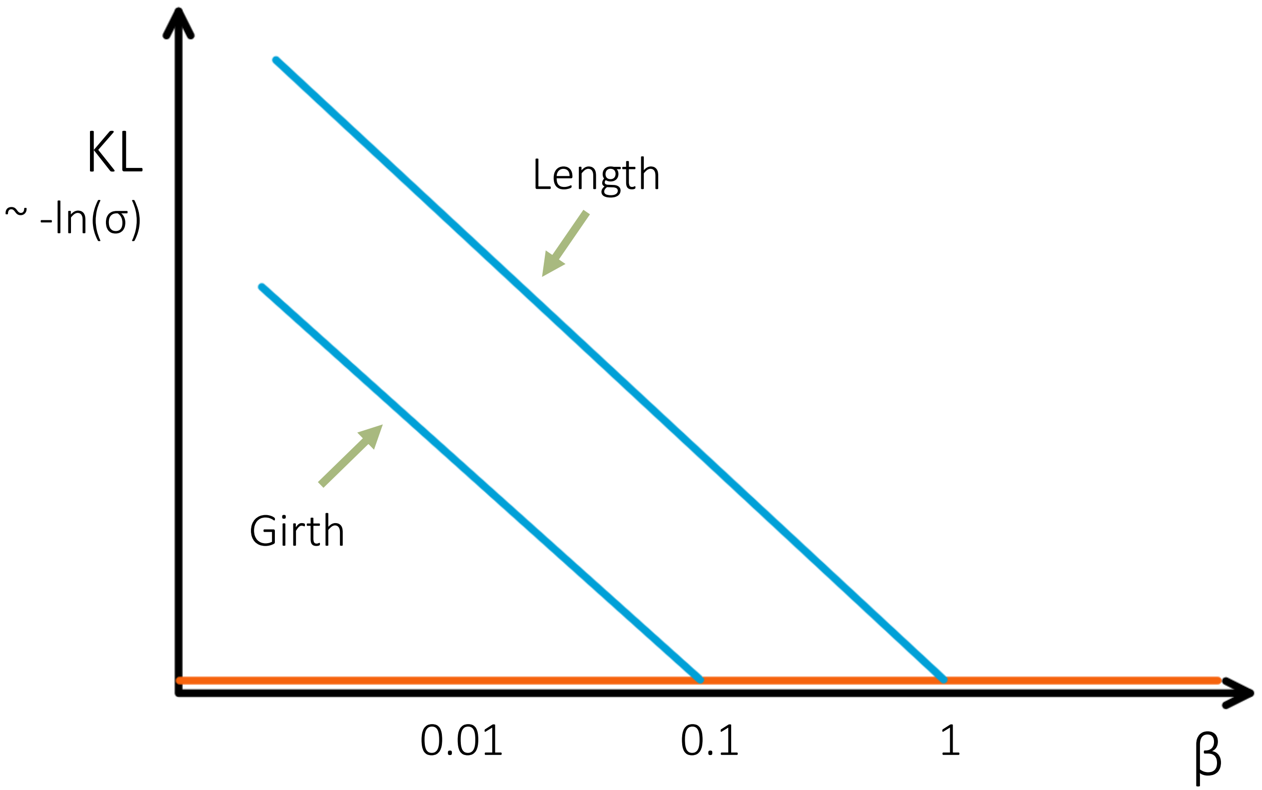
The Variational Autoencoder

Bananas



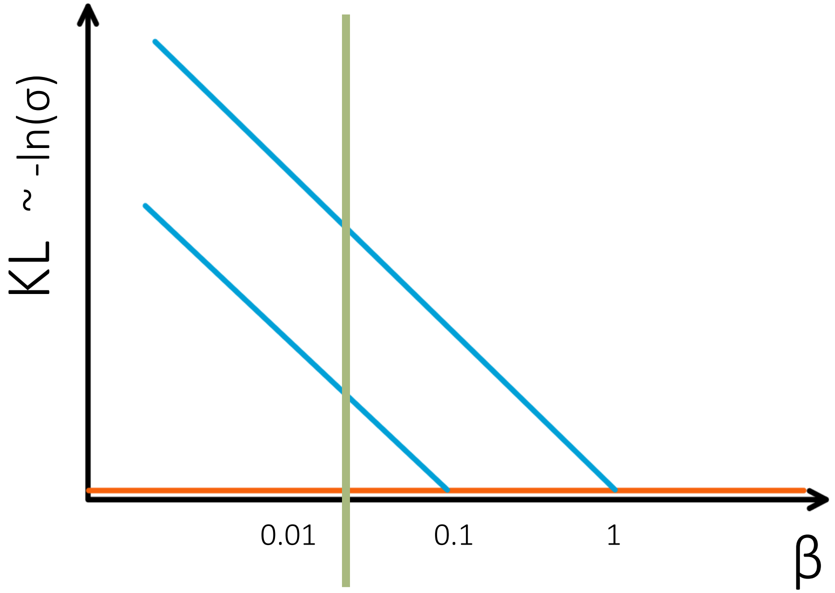
The Variational Autoencoder

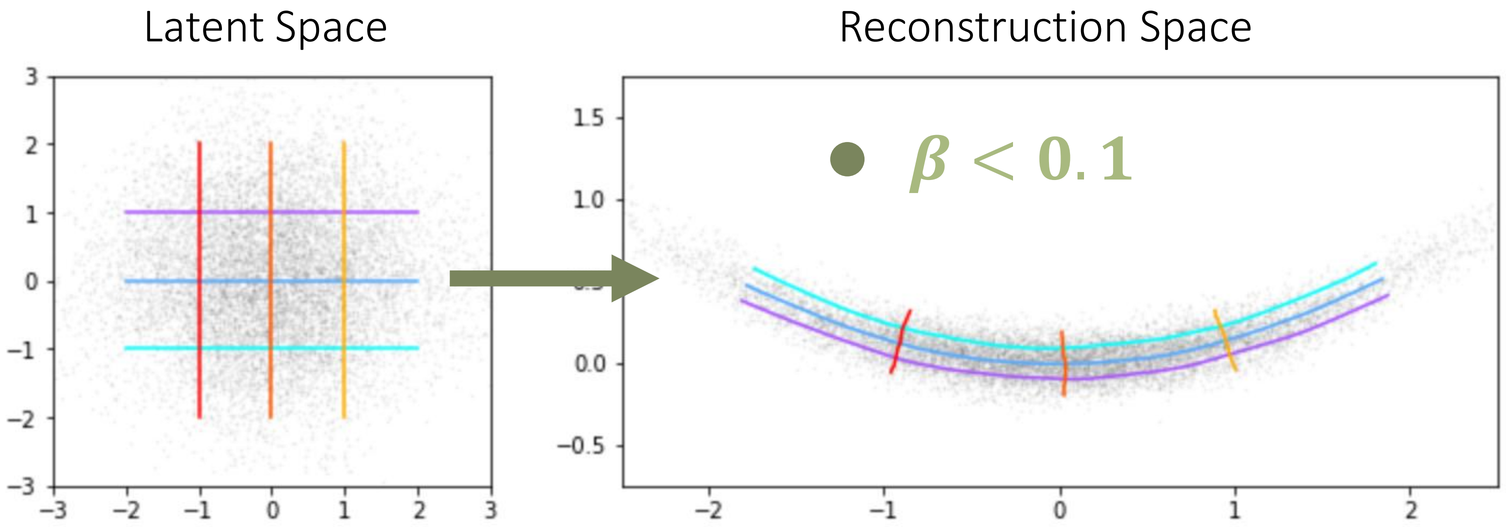
Bananas



The Variational Autoencoder:

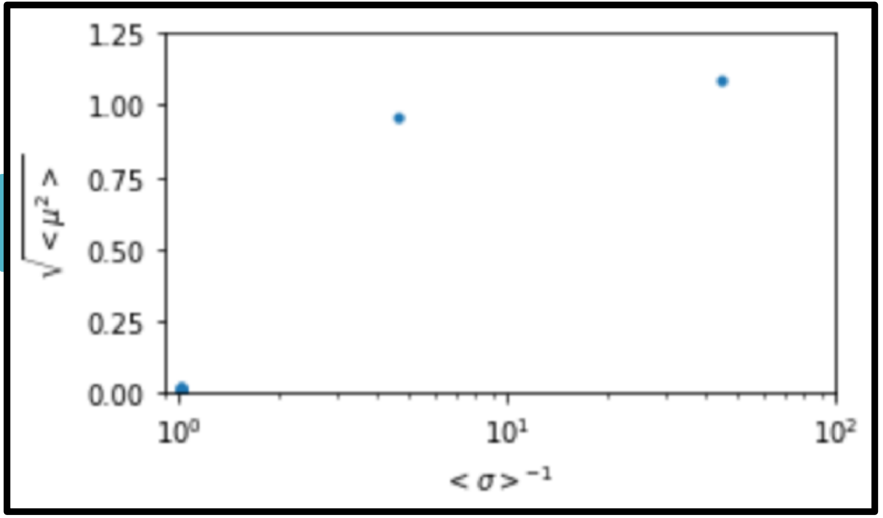
Bananas

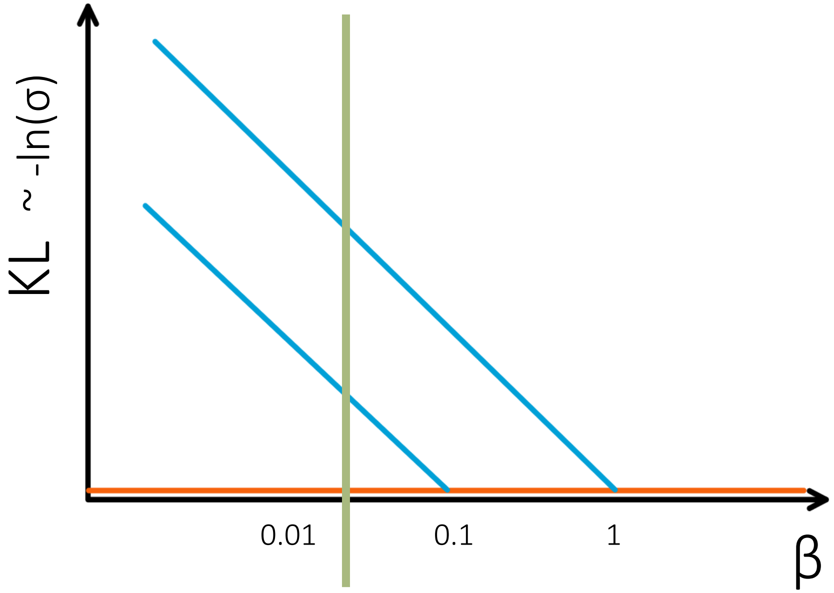




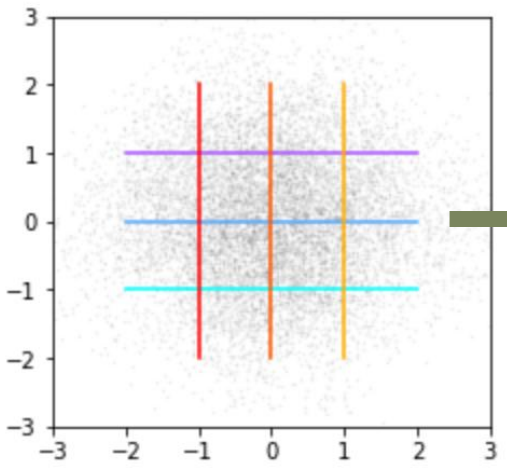
The Variational Autoencod

Bananas

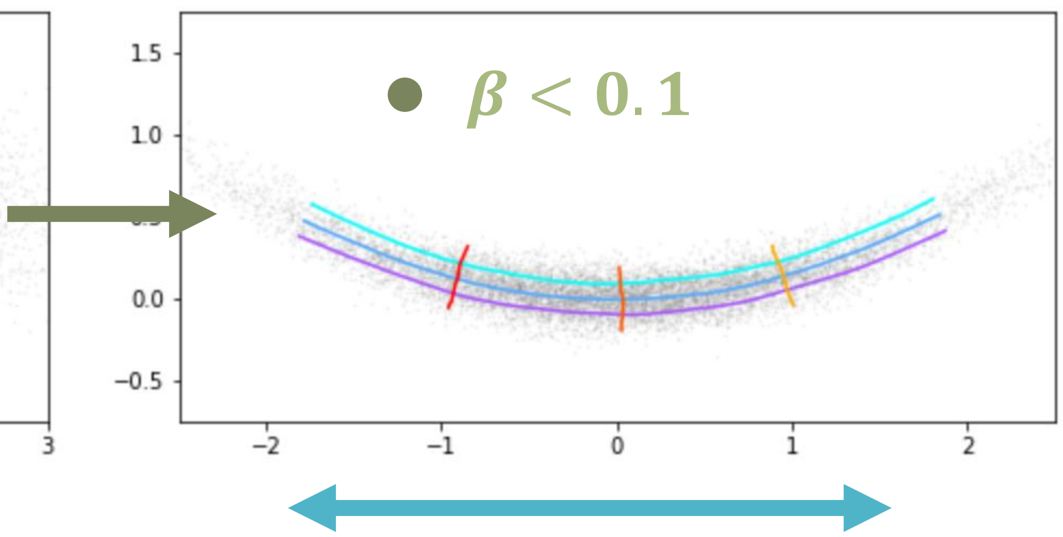




Latent Space



Reconstruction Space

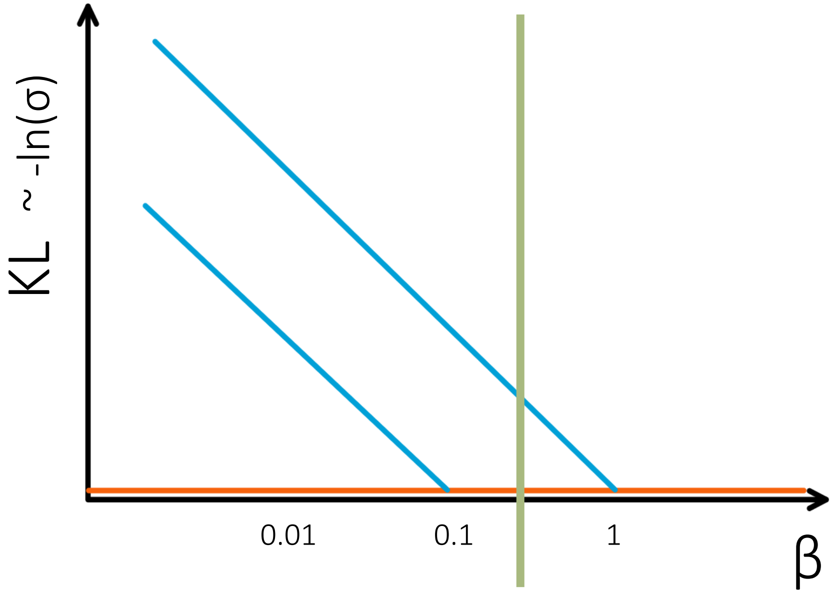


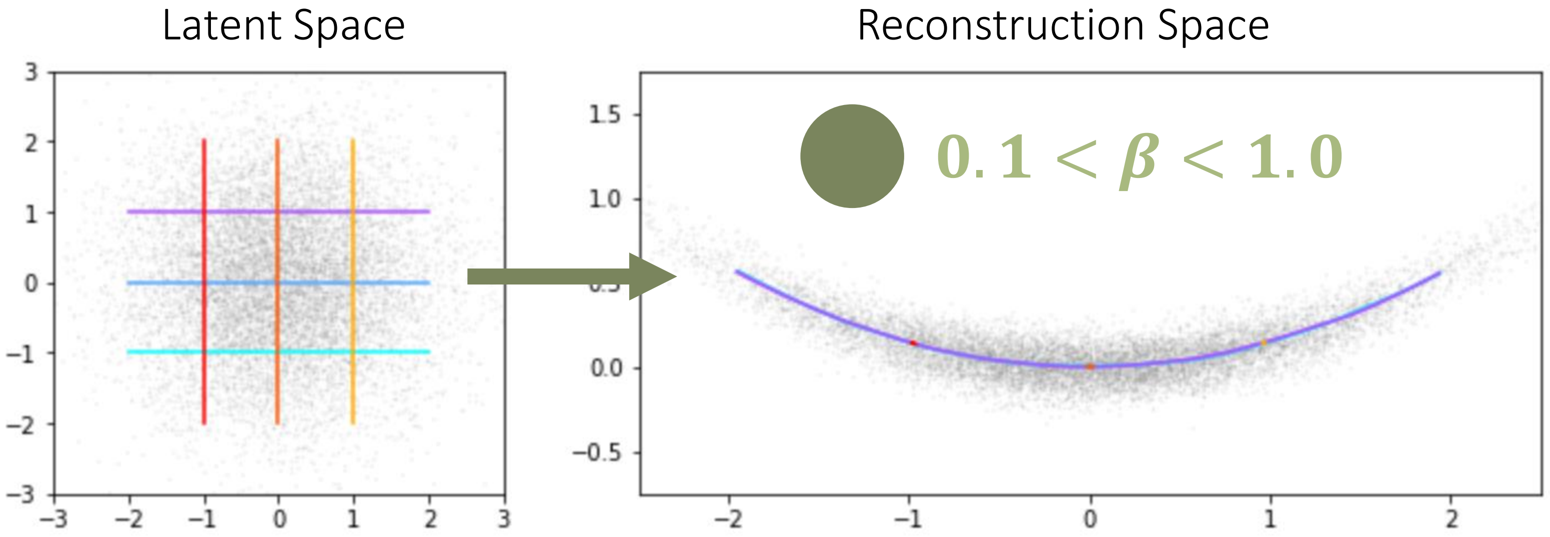
Size = $^{\beta}/_{\sigma}$

The VAE is doing non-linear PCA

The Variational Autoencoder:

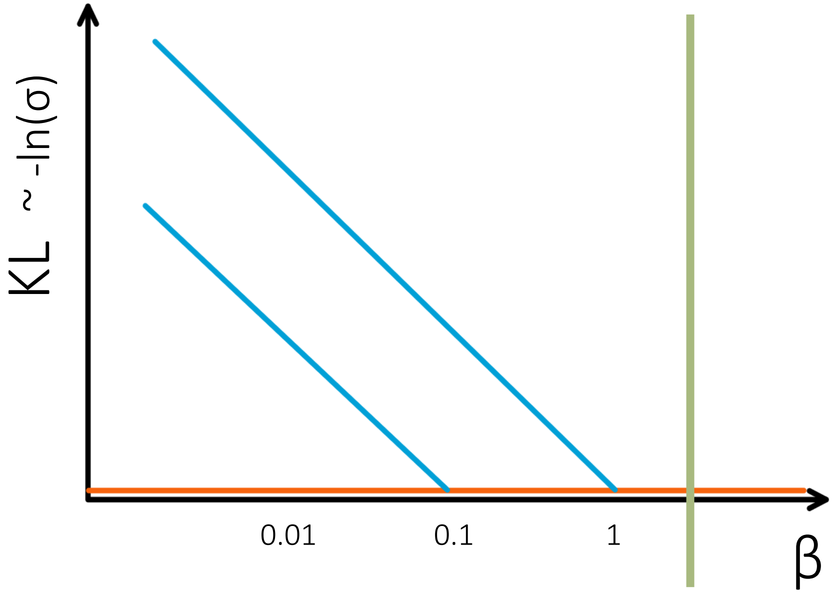
Bananas

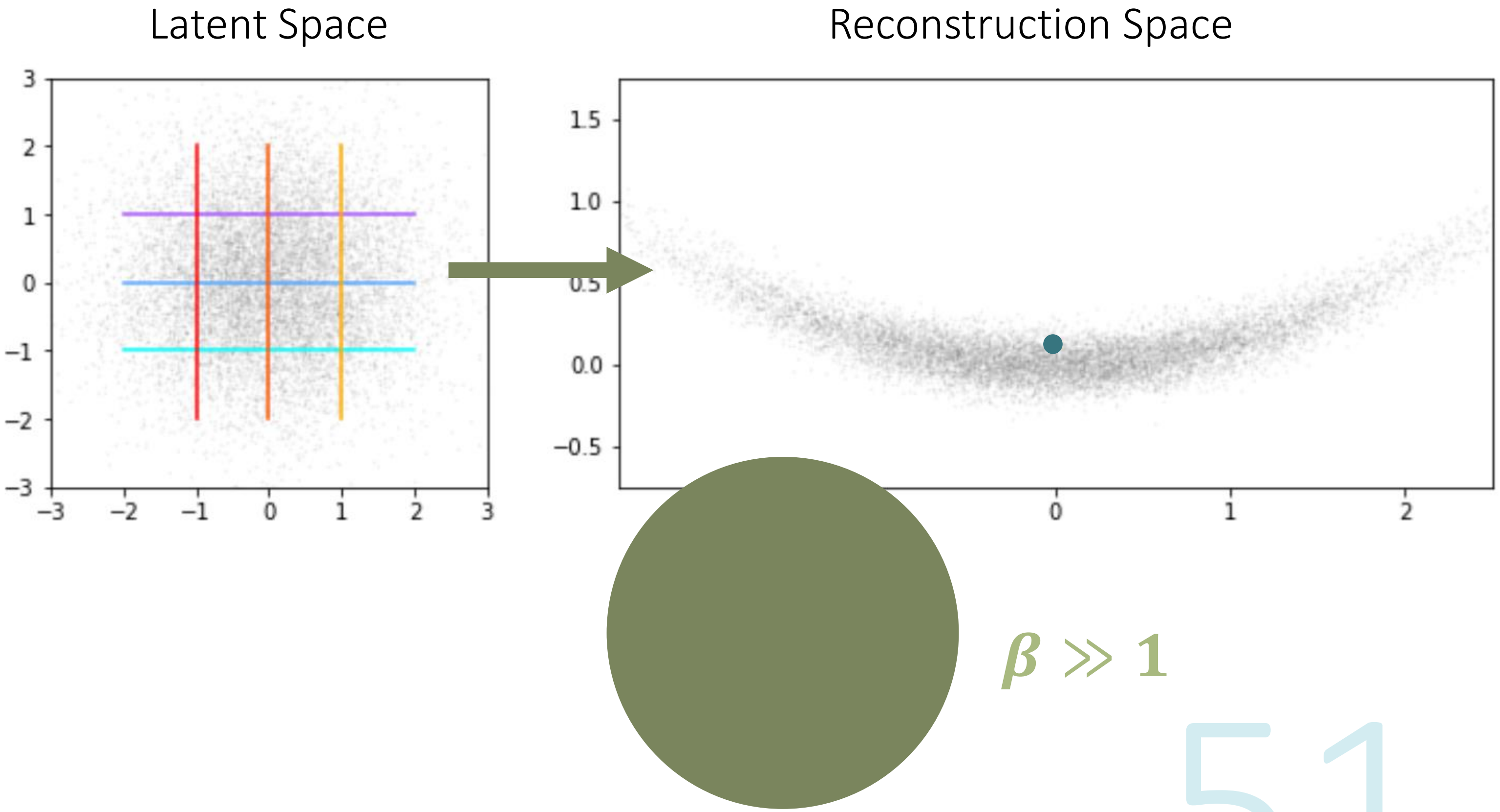




The Variational Autoencoder:

Bananas





$\beta \gg 1$

The Variational Autoencoder

Dimensionality

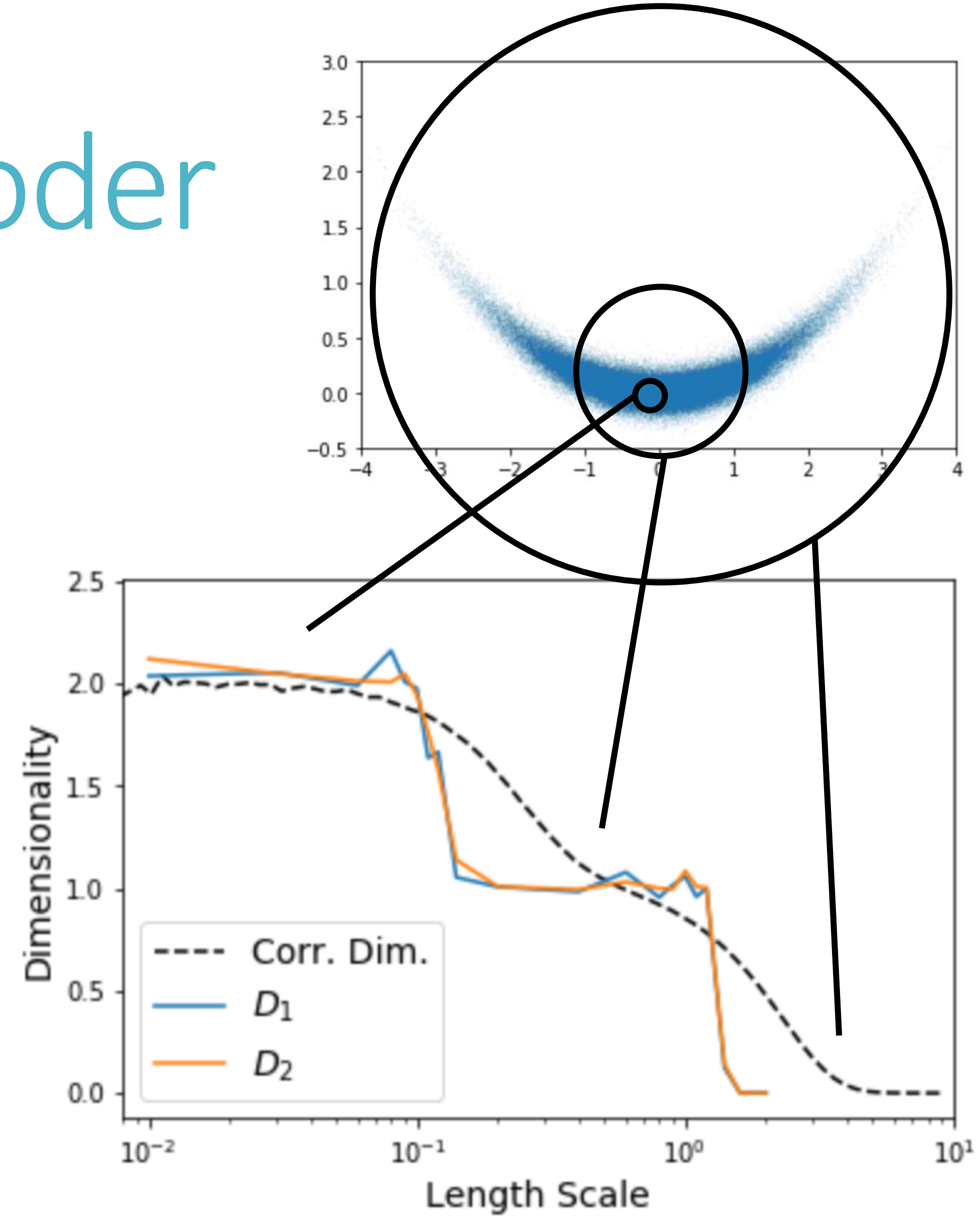
$$D_{corr} \equiv \frac{d\,N}{d\log r}$$

$$D_1 \equiv \frac{d\langle|\Delta \boldsymbol{x}|^2\rangle}{d\,\beta^2}$$

Variation of resolution with scale (think $\langle r^2 \rangle = D\,\sigma^2$ for D-dimensional Gaussian).

$$D_2 \equiv -\frac{d\,KL}{d\log\beta} \cong \frac{d\log\sigma}{d\log\beta}$$

Variation of information with scale.

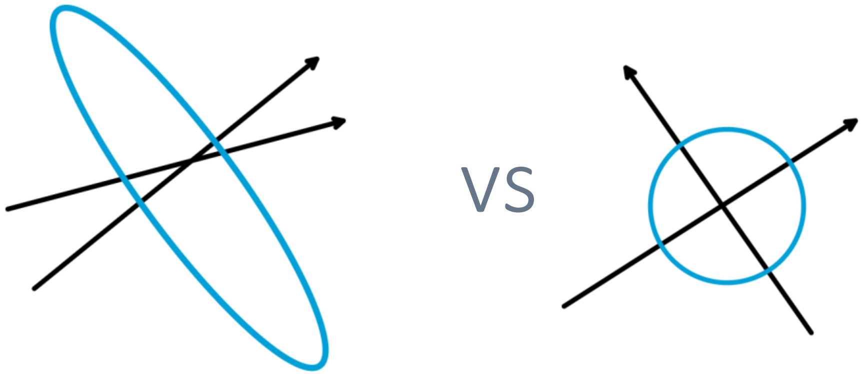


I am still trying to work out formally the meaning of these expressions, but they have an air of truthiness about them and empirically give sensible results.

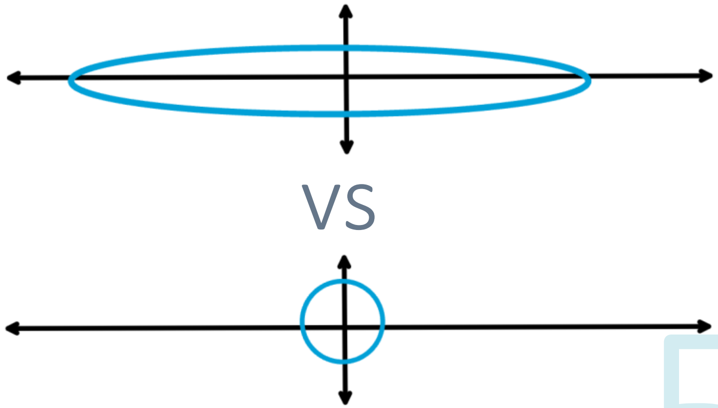
The Variational Autoencoder

Orthogonalization and Organization is Information-Efficient

Orthogonalization:

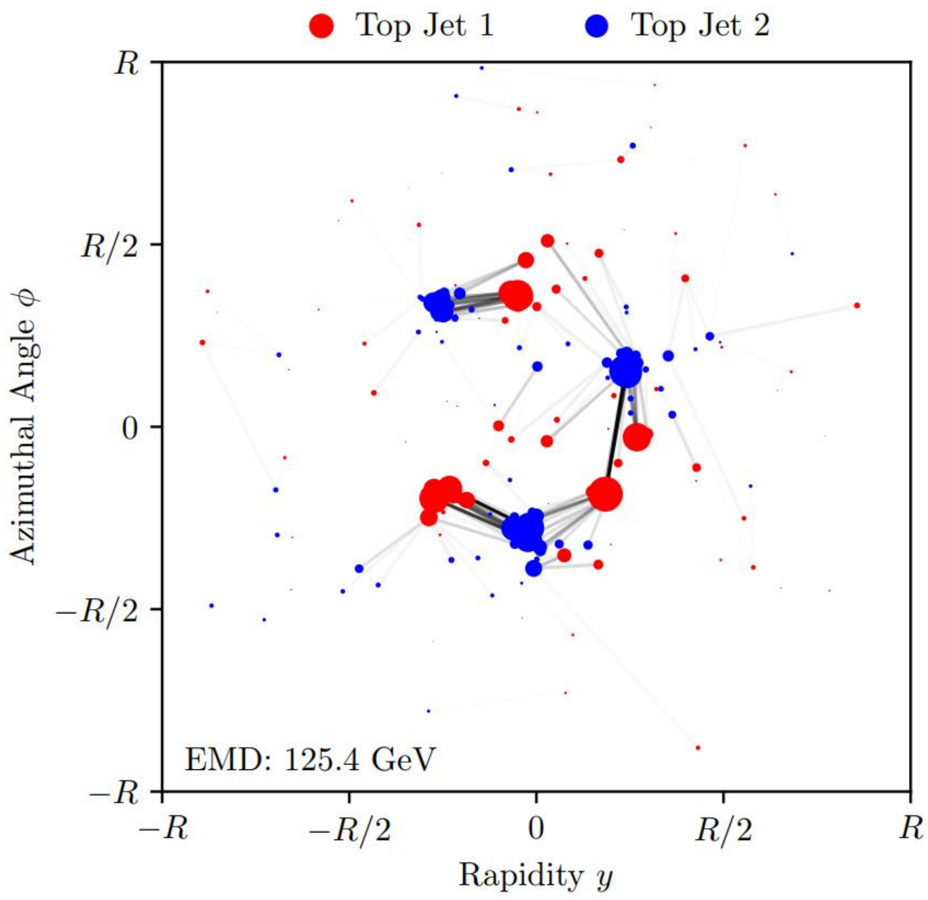


Organization:



Reconstruction Error

Sinkhorn Distance ≈ EMD



arXiv:1902.02346

Sinkhorn's algorithm; start with $\Delta R_{ij}, p_{Ti}, p_{Tj}$ then:

$K_{ij} = \exp(\Delta R_{ij}/\tau)$
$u_i = \mathbf{1}_i$
$v_i = \mathbf{1}_j$

Repeat N times:

$u_i = p_{Ti}/(K.v)_i$
$v_i = p_{Tj}/\left(K^T.u\right)_j$

Return $T_{ij} = u_i K_{ij} v_j$

The Variational Autoencoder:

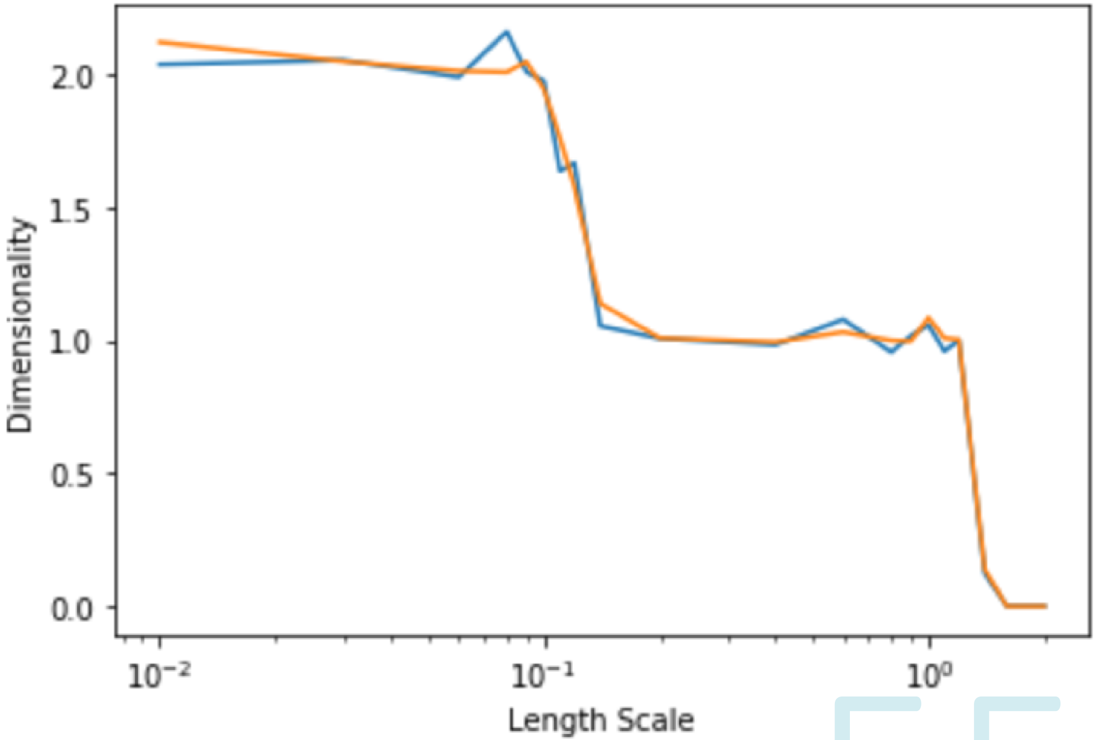
Dimensionality

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

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

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

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



The Variational Autoencoder

Doesn't suffer from curse of dimensionality

Toy data generated from:

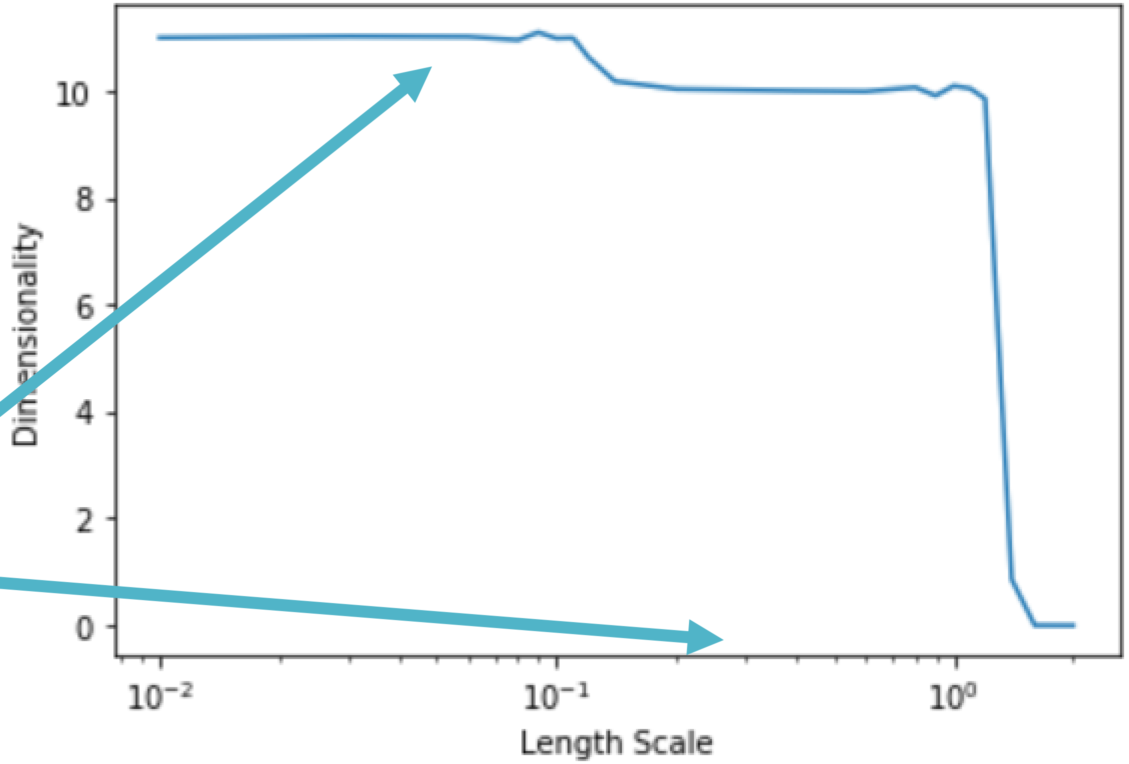
$$P(\vec{x}) = [\prod_{i=1}^{10} N_i(\mu = 0, \sigma = 1)]\ N_{11}\ (\mu = 0, \sigma = 0.1)$$

With $N_{tot} = 5 * 10^5$ points

Typical distance to neighbour $\sim N_{tot}^{-1/10} \sim 0.3$

Correlation dimension runs into sparsity limit before the small dimension is even discovered!

The VAE finds the small dimension.



The Plain Autoencoder

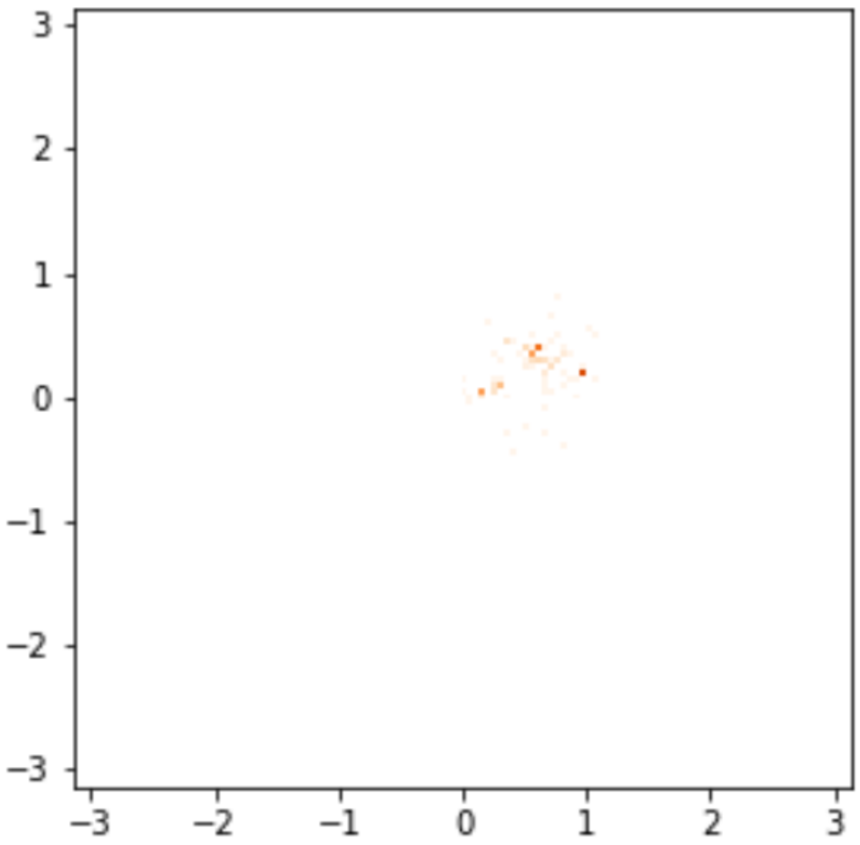
Garbage

My old plan:

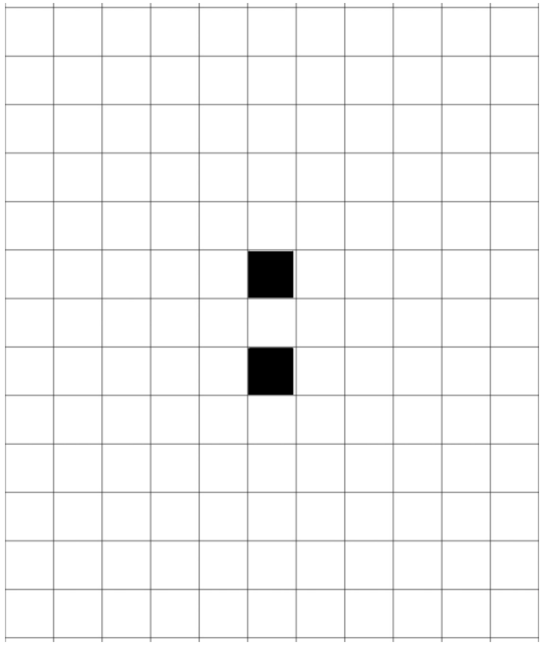
- Train AE on jet images using different latent space sizes N
- Study reconstruction quality as a function of N
- … Learn something about ‘jet information’?

Flaws:

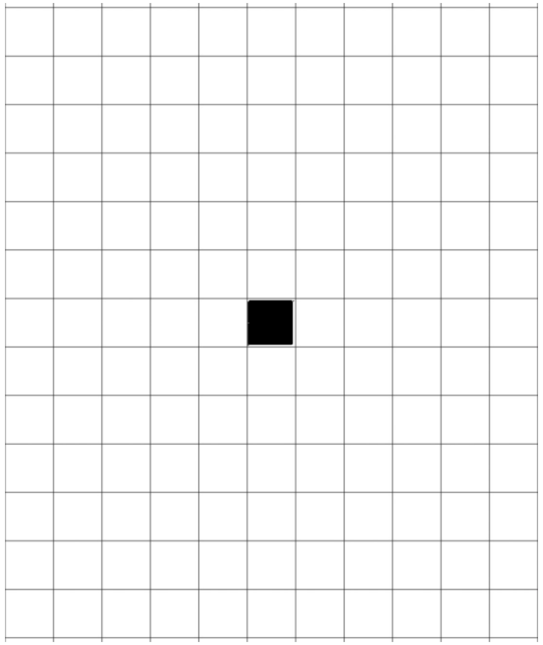
1) Jet images are garbage
2) Autoencoders are garbage



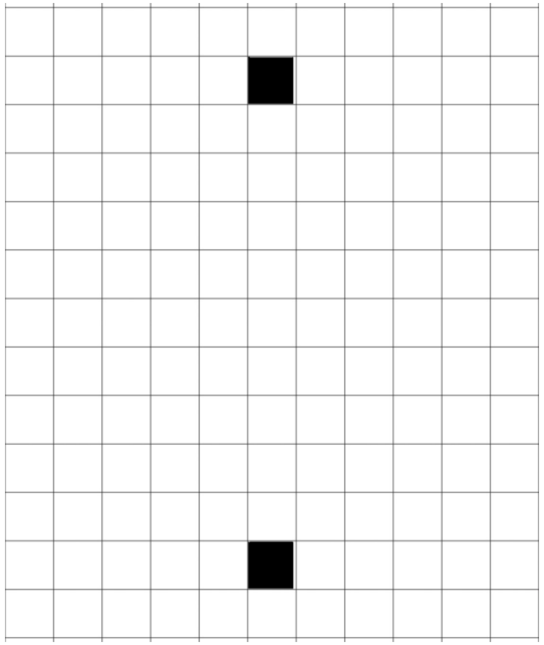
“Jet Images are Garbage”



(a)



(b)



(c)

All three of these jet images are maximally different from eachother according to summed pixel intensity difference, but (a) and (b) are more physically similar than are (b) and (c).

Future Directions

1. What is the point?

2. Alternative latent priors?

3. Alternative metrics?

The Variational Autoencoder

ML Engineer:

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

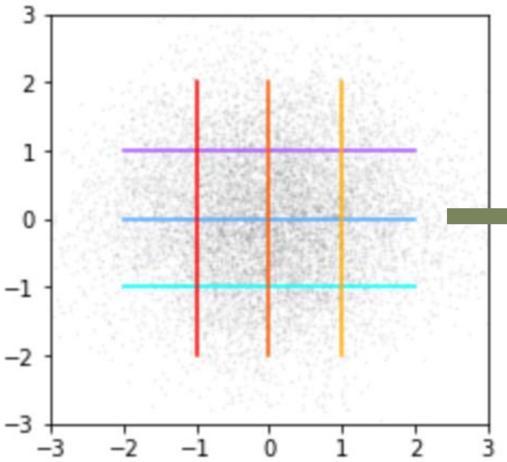
Bayesian statistician:

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

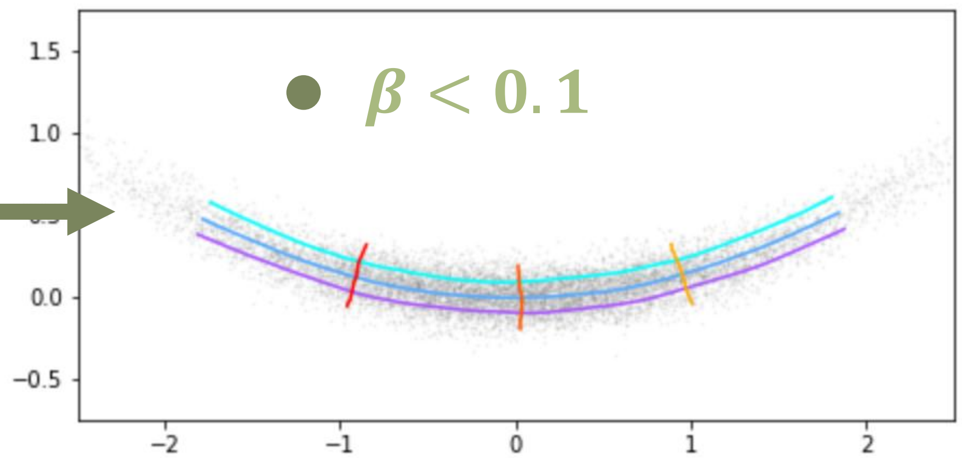
The Variational Autoencoder:

Bananas

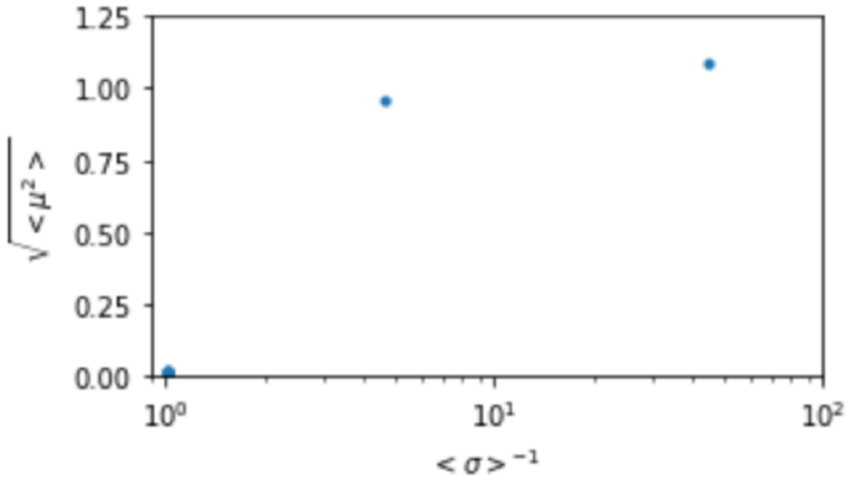
Latent Space

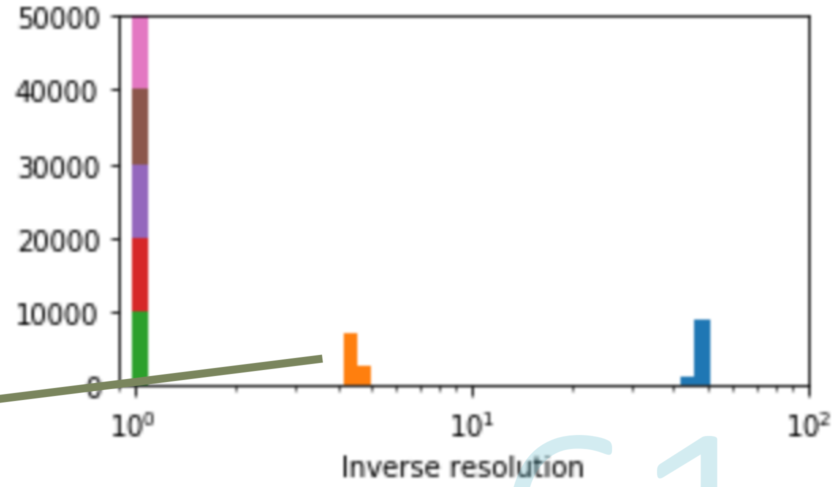


Reconstruction Space



- $\beta < 0.1$





The VAE is doing non-linear PCA

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