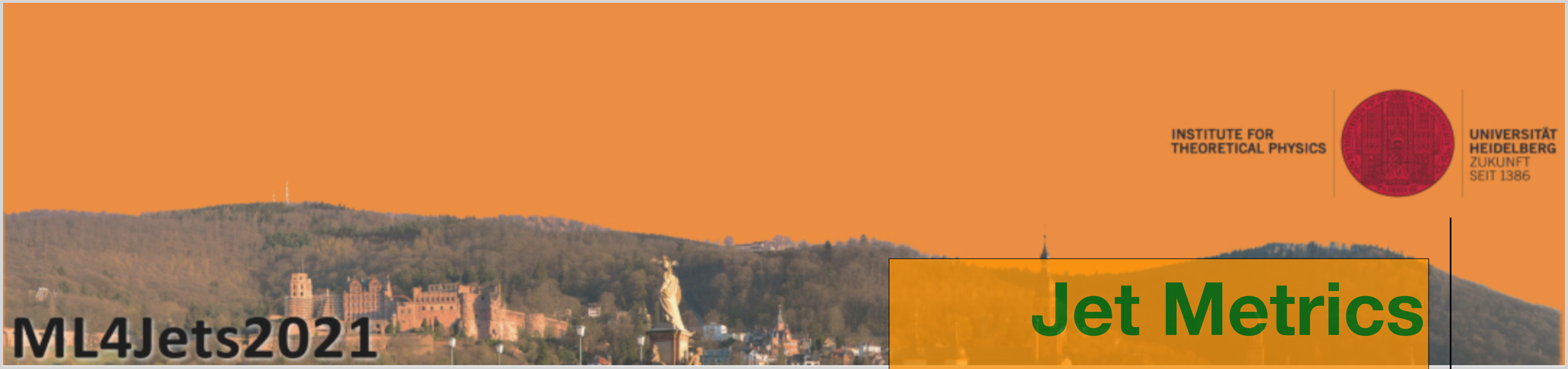


INSTITUTE FOR THEORETICAL PHYSICS

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Jet Metrics and Autoencoders

Rashmish K. Mishra
Harvard

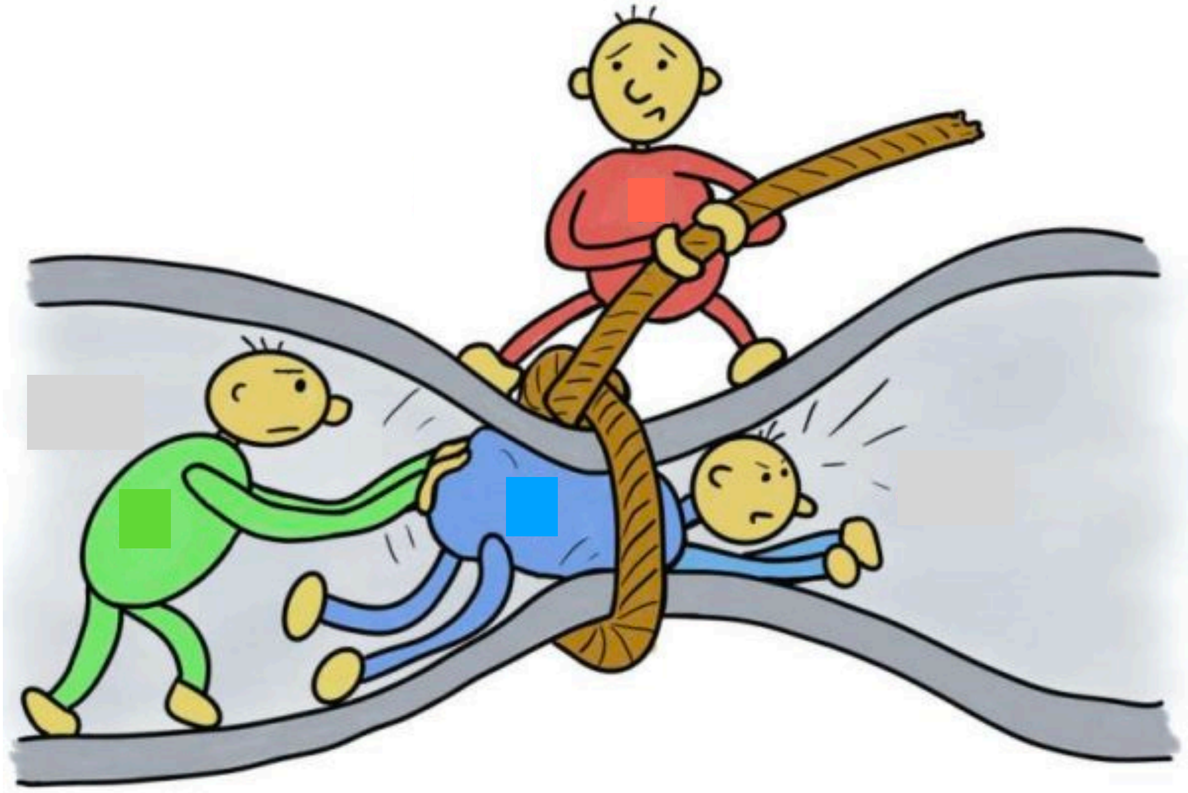
Based on work *in progress* with
K. Fraser, S. Homiller, B. Ostdiek, M. Schwartz

Autoencoders as anomaly detectors

1811.10276, 1905.10384, 1808.08979, 1809.08992, 1903.02032, 2007.01850

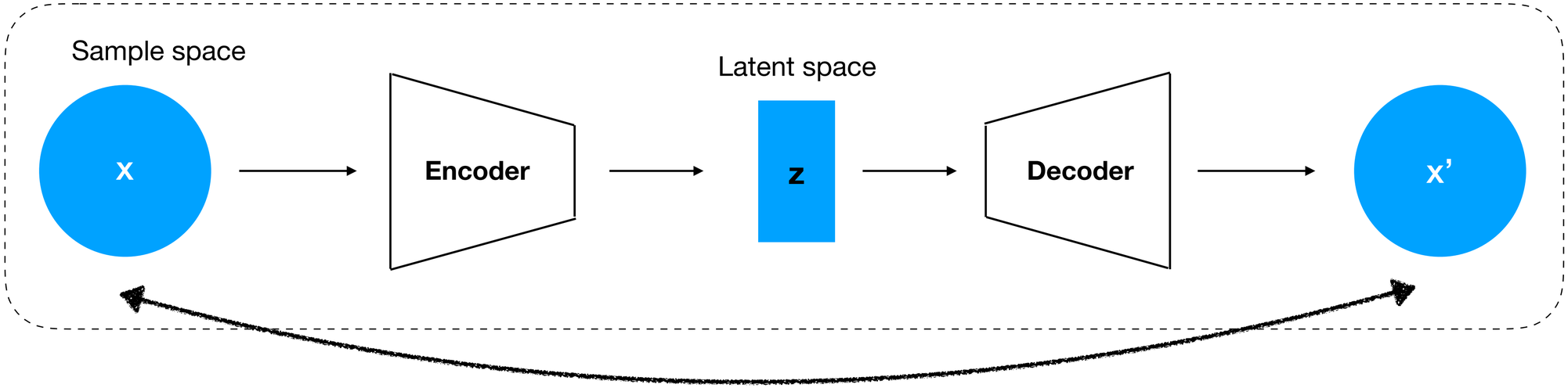
Autoencoders have become very popular in the context of detecting QCD jets from non-QCD jets (that can arise in SM or BSM).

Basic Idea: Create a bottle-neck of information (dimensional reduction), which is enough to successfully create QCD jets but not for other types of jets.



- **Unsupervised**
- **Hyperparameter tuning needed**
- **Claim: universal anomaly detector**

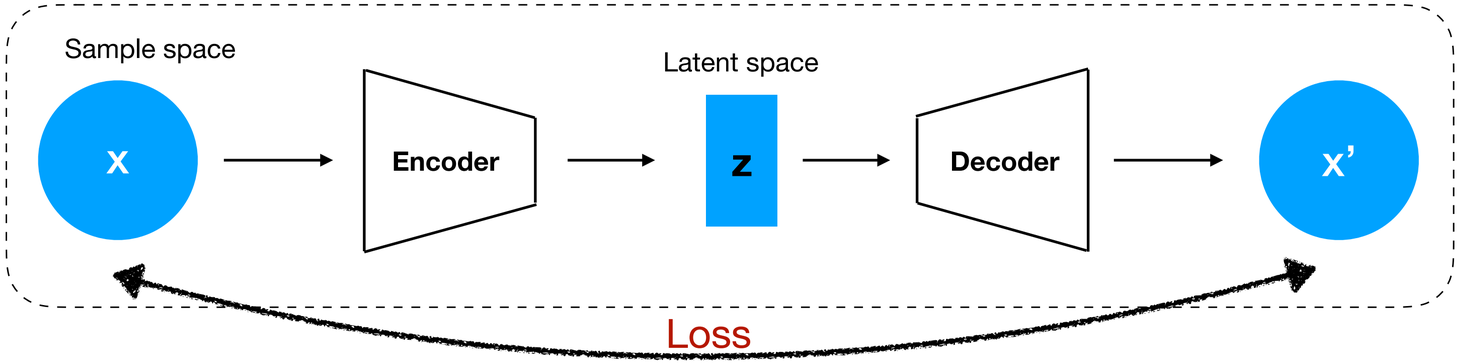
Autoencoder basics



Loss = the “difference” between x and x’

The network is trained to minimize the Loss

Autoencoder and metrics



An autoencoder is necessarily tied to the loss it minimizes, and therefore to the way x and x' are compared.

The correct comparison depends on the task.

In the context of anomaly detection, the correct comparison depends on what is anomalous w.r.t. a given set.

Autoencoder and metrics

Usual choice for loss: Mean Squared Error (MSE)

e.g. in a pixel based image input $$L_{\mathrm{MSE}} = \sum_{\text{pixels i, j}} \left(x_{ij} - x'_{ij}\right)^2$$

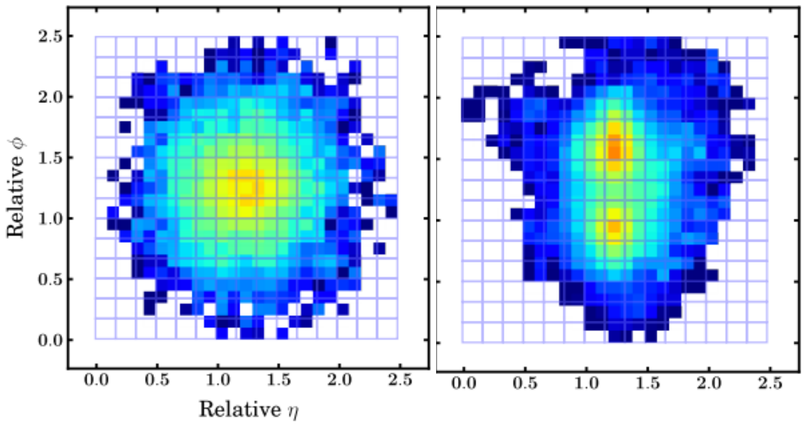
x_{ij}: intensity at location (i, j)

MNIST example:

input image

reconstructed image

This introduces the need to preprocess (e.g. align/rotate/ center) the image (1407.5675)



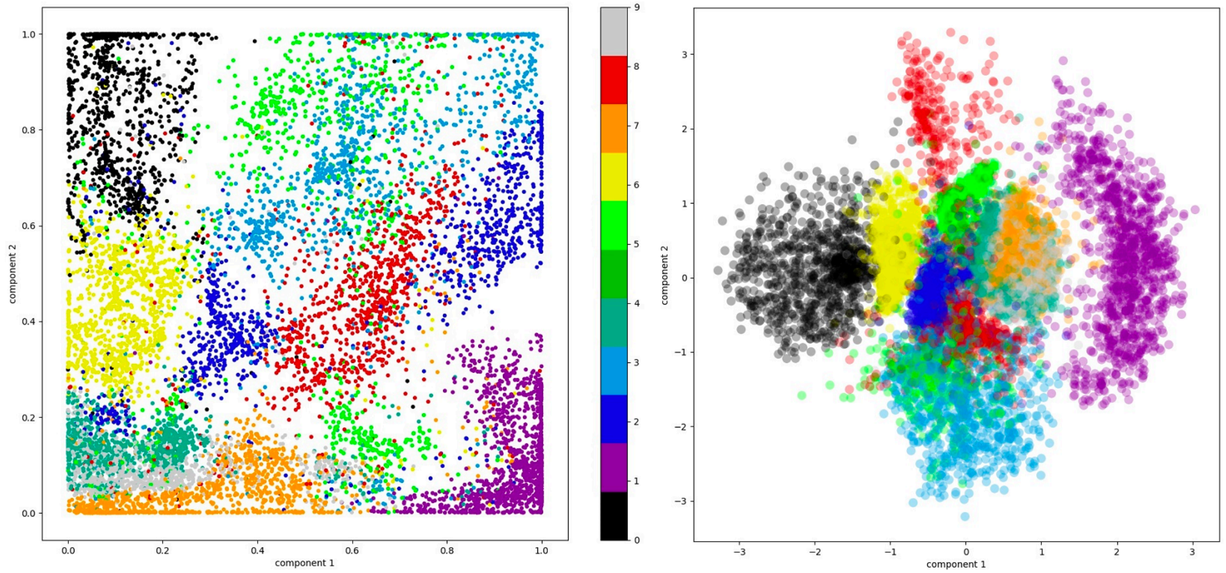
W jet, before and after preprocessing(1407.5675)

Autoencoder and metrics

Requiring a meaningful structure in latent space: β - variational autoencoders

The loss is modified to introduce a term dependent on Latent Space $L = (1-\beta)L_{\mathrm{MSE}} + \beta L_{\mathrm{KLD}}$

An example from MNIST data set: the latent space for an AE and a VAE.



https://pureai.com/articles/2020/05/07/~/media/ECG/pureai/Images/2020/05/autoencoders3.asxh

L_{KLD} : Kullback–Leibler divergence between latent space and a gaussian

The AE has some classes that get placed in many different clumps, but the VAE does not (e.g. blue dots)

The space of metrics

- Regular AE, VAE, β-VAE are Autoencoders with a different "distance" in the loss, depending on the task. The loss has an MSE part, which is hard to understand physically (not IRC safe, depends on details of pixelation, preprocessing etc)

- There are other well motivated metrics: Earth mover's distance (EMD), a metric based on optimal transport has been introduced recently. Has useful features such as a notion of locality and being IRC safe.

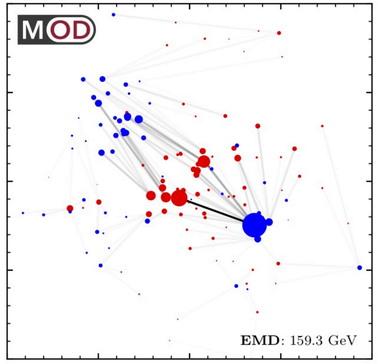
Earth Mover's Distance

EMD is based on the optimal transport to change the radiation pattern of one event to another.
It is IRC safe, desirable from a calculation point of view.

$$\mathrm{EMD}(\mathcal{E}, \mathcal{E}') = \min_{g} \sum_{ij} g_{ij} \left(\frac{\theta_{ij}}{R}\right)$$

i, j index particles in $\mathcal{E}, \mathcal{E}'$

$$g_{ij} > 0, \sum_{j} g_{ij} \leq E_i, \sum_{i} g_{ij} \leq E'_j$$



1902.02346

g_{ij} : transport from ith particle in first event to the jth particle in the second event

E_i : energy of ith particle in the event

θ_{ij} : distance in (η, ϕ) plane

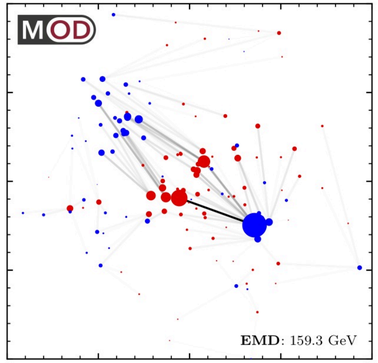
Earth Mover's Distance and its generalisation

EMD is based on the optimal transport to change the radiation pattern of one event to another.
It is IRC safe, desirable from a calculation point of view.

$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_g \sum_{ij} g_{ij} \left(\frac{\theta_{ij}}{R}\right)$$

i,j index particles in $\mathcal{E},\mathcal{E}'$

$$g_{ij} > 0, \sum_j g_{ij} \leq E_i, \sum_i g_{ij} \leq E'_j$$



1902.02346

g_{ij} : transport from ith particle in first event to the jth particle in the second event

E_i : energy of ith particle in the event

θ_{ij} : distance in (η,ϕ) plane

EMD is one example of a class of metrics based on OT (optimal transport).

C. Villani, Optimal Transport, old and new, Springer

General OT based metric

For events $\mathcal{E},\mathcal{E}'$

$$C(\mathcal{E},\mathcal{E}') = \min_g \left(g_{ij}c_{ij}\right)$$

g_{ij} : optimal transport, c_{ij} : cost

If the cost is defined in terms of a distance, the OT is the pWasserstein distance

$$W_p(\mathcal{E},\mathcal{E}') = \min_g \left(g_{ij}d_{ij}^p\right)^{1/p}$$

d_{ij} : distance between events

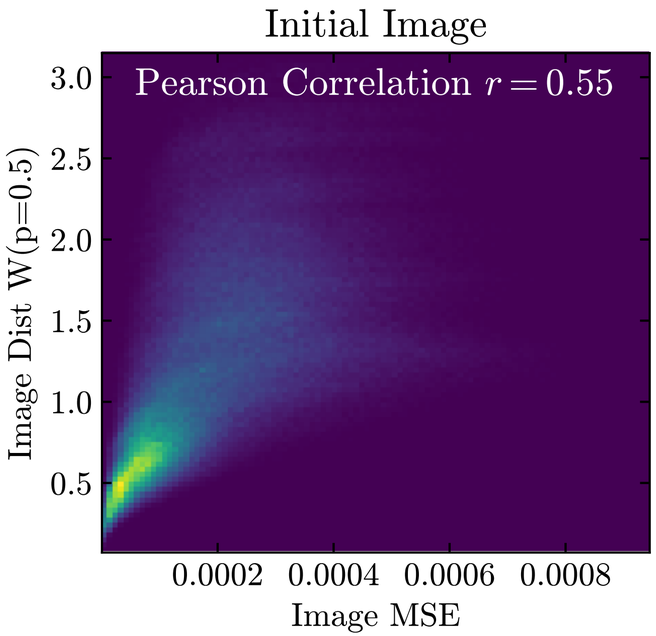
W_1 is the EMD. We can consider general p.
In next slides, we will consider p=1 and p=1/2.

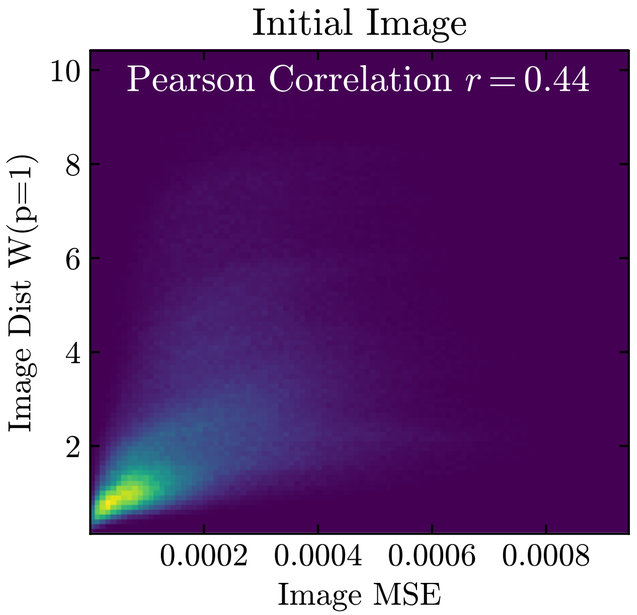
General Motivation

How universal/metric-independent are these methods?

As a first step, is there a relation between these different kinds of metrics - MSE in the context of autoencoders, and EMD and its generalization Wp?

e.g. look at the correlation between MSE and Wp distance between QCD jets





We don’t see very high correlation at this level

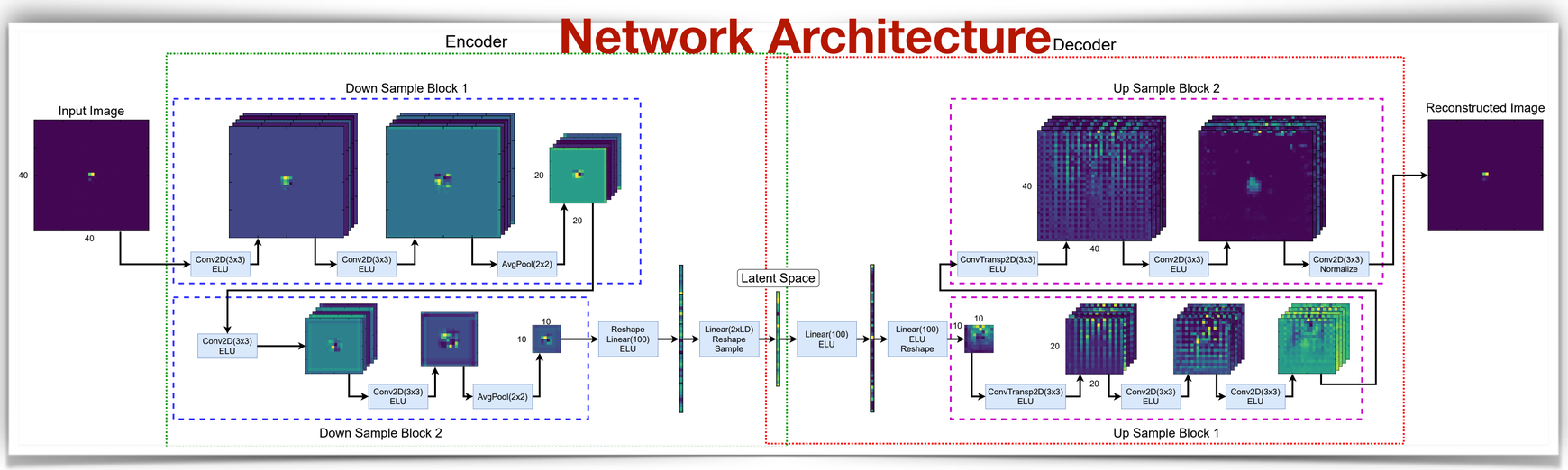
Physics case: QCD vs top

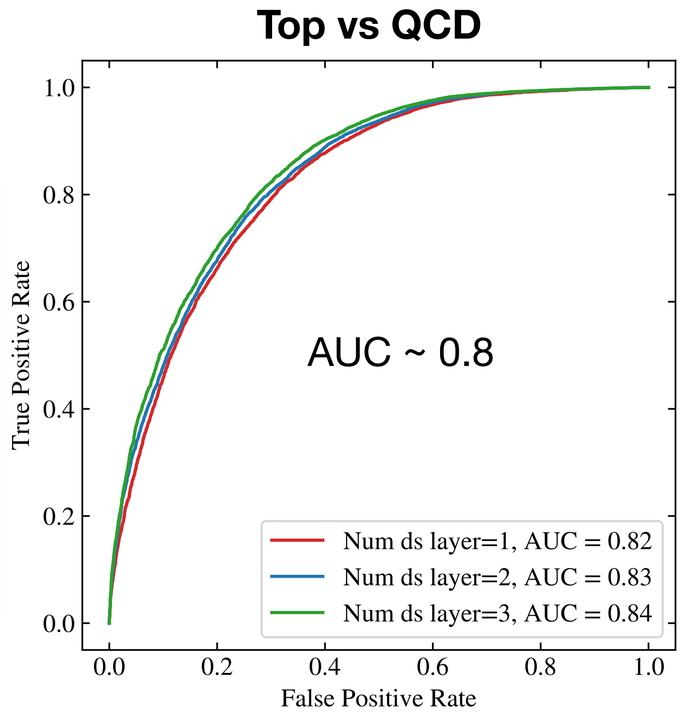
“How far is a top jet from a QCD jet?”

Dataset: https://zenodo.org/record/4614656#.YNSThhNKhsM
from 2007.01850: VAE for anomalous jet tagging

R=1 anti-kt jets with pT > 450. Top jets, pT peak ~ 1.2 TeV

Choice 1: Use a β CVAE trained on QCD images. The anomaly score is the total loss function.





We consider 1, 2, 3 downsampling blocks, 64 dimensional LS,
and vary $\beta \in \{0, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$

Optimal $\beta = 10^{-7}$

Physics case: QCD vs top

Dataset: https://zenodo.org/record/4614656#.YNSThhNKhsM
from 2007.01850: VAE for anomalous jet tagging

"How far is a top jet from a QCD jet?"

Choice 2 : Use EMD distance from representative QCD jets as the anomaly score.

Cai et al, 2008.08604,
Komiske et al 1902.02346

No training needed.
No Hyper-parameter tuning needed.

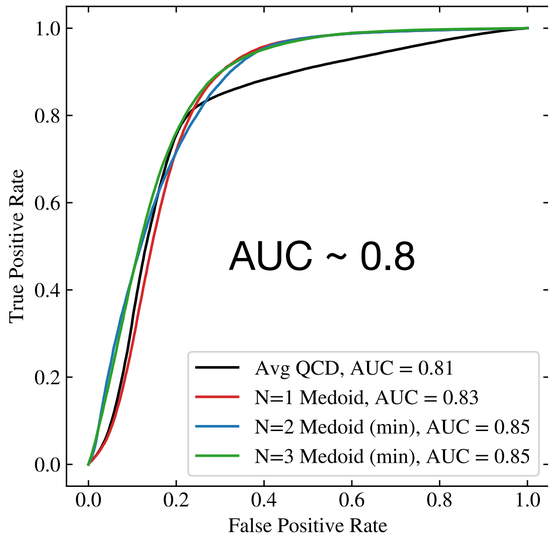
Possible choice of anomaly scores: EMD distance from

- average QCD image
- minimum distance from n medoids (based on EMD itself)

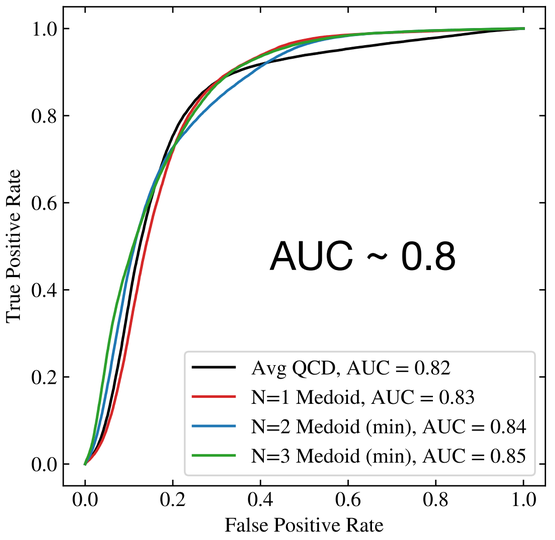
These two entirely different approaches (VAE vs EMD) give very similar ROC curves and similar AUCs.

To understand this, we can look at the correlation between latent space Euclidean distance and EMD distance between QCD images

Top vs QCD, using $W_{p=1}$

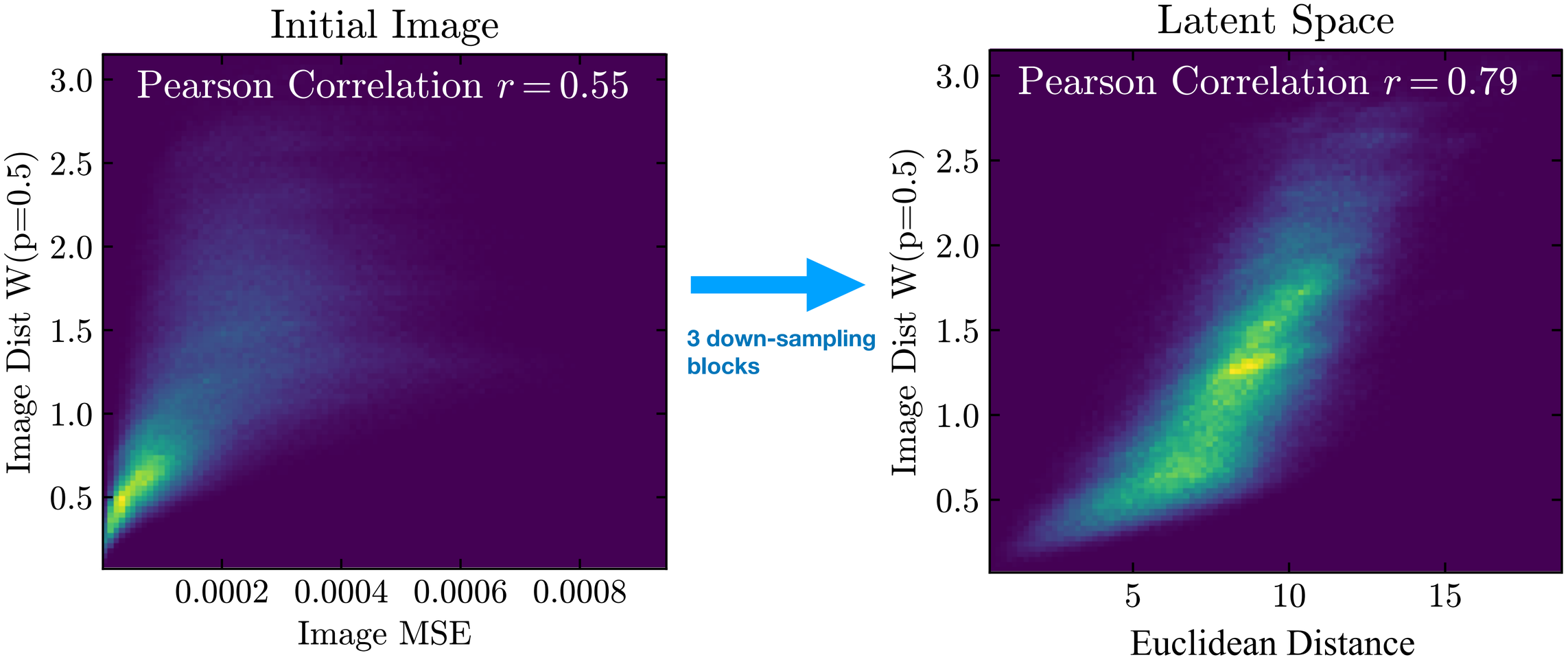


Top vs QCD, using $W_{p=1/2}$



From Input space to Latent space

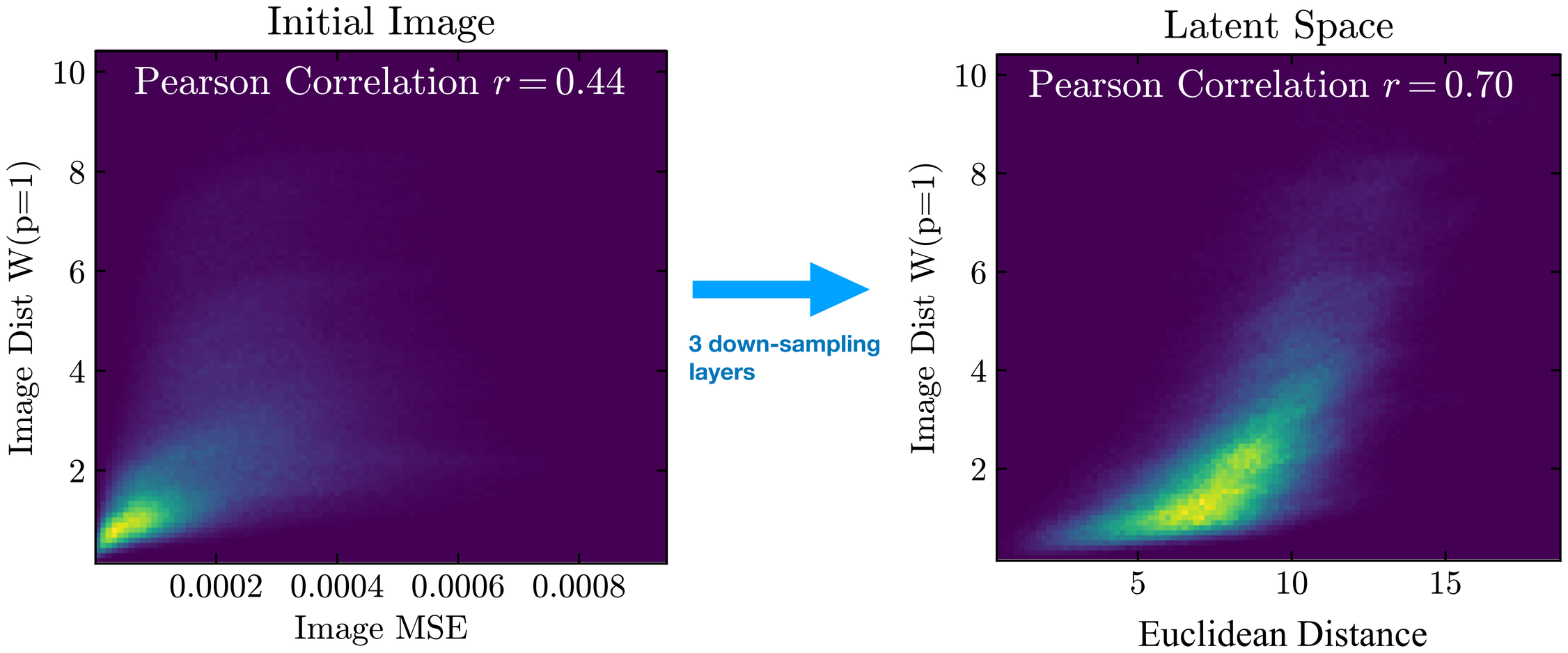
$W_{p=0.5}$



Correlation increases (0.55 to 0.79)

From Input space to Latent space

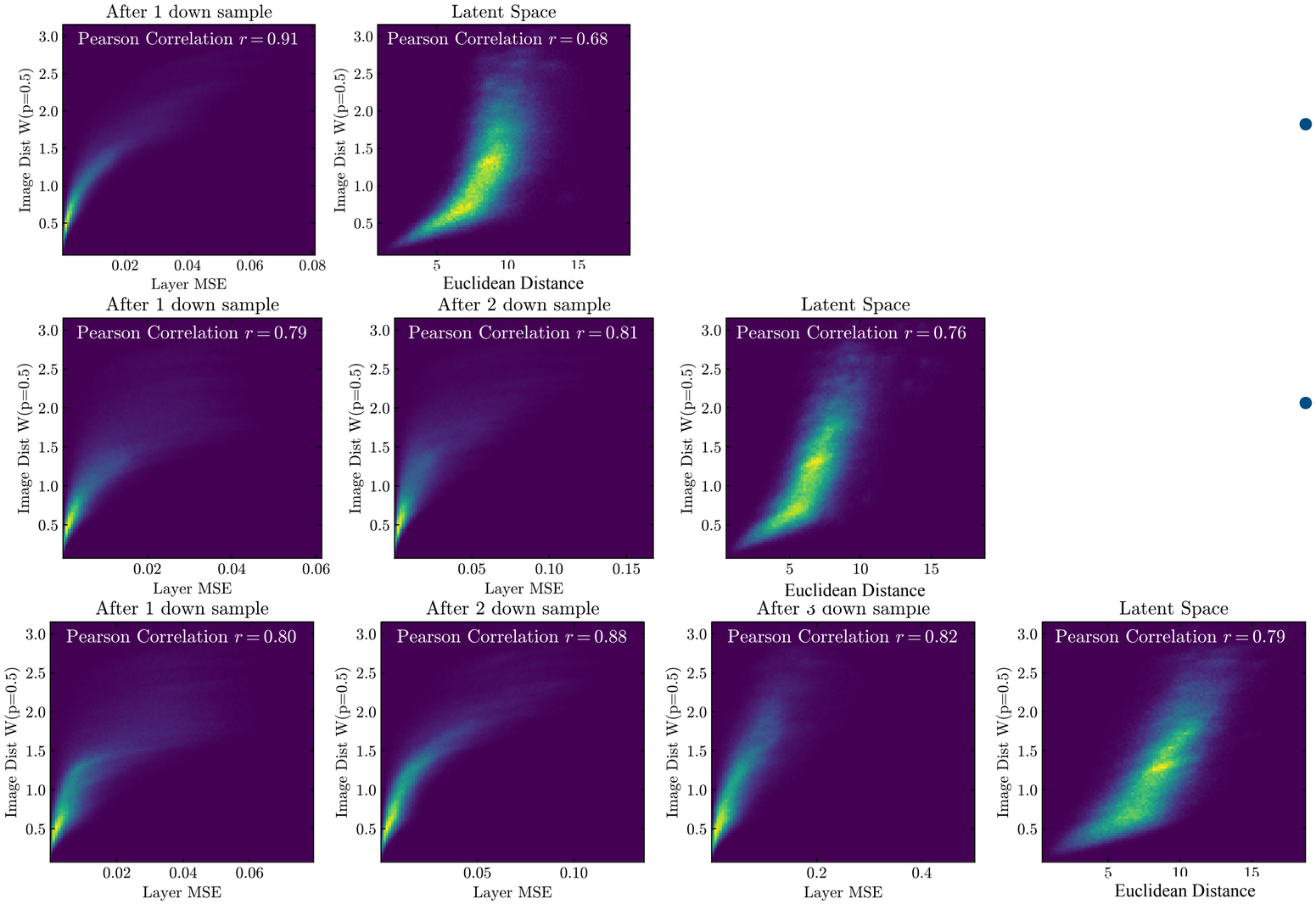
$W_{p=1}$



Correlation increases (0.44 to 0.70)

Correlation develops during intermediate layers

Baseline: Initial correlation between MSE and $W_{p=1/2}$: 0.55



For both $W_{p=0.5}$ and $W_{p=1}$

(Only p=1/2 shown here)

- **Correlation in LS increases from 0.68->0.76->0.79 as more down sampling layers are added.**
- **Network with 1 DS layer has higher correlation in the DS layer but lower in the LS, as compared to network with 2 DS layer.**

Strong evidence of the network learning being correlated with optimal transport based distances

Robust against:

- **Number of downsampling layers (1,2,3)**
- **Choices of OT metrics considered (p=0.5, 1)**

Next steps?

- Can we use this to learn about the scope of autoencoders and of signals that they can detect?
- Does this give any general lessons towards looking for anomalous objects at colliders?

Summary and Future Directions

- **Clear correlation seen between convolutional VAEs and OT based metrics: the network is learning an OT based metric. Anomaly scores are comparable for QCD vs top.**

- **The choice of OT based metrics is useful to explore. The right metric will depend on the application. (e.g. pileup, emerging jets, semi-visible jets, different pronged jets, different mass resonances, etc).**

Backup: W_p for p < 1, pros and cons

Metric: $\min_g \left[g_{ij} \left(\frac{\theta_{ij}}{R} \right)^p \right]$

$$g_{ij} > 0, \sum_j g_{ij} \leq E_i, \sum_i g_{ij} \leq E'_j$$

IRC safe for p < 1

- Addition of a zero energy particle does not change the cost
- Collinear splitting also does not change the cost

Triangle inequality is satisfied for p < 1, in the *present* form (no 1/p power).
With 1/p power, the inequality is *reversed*, but a *different* triangle like inequality is satisfied

- Geometrization should be possible, but more care is needed.
- The metric is in L^p space which are well defined space for p < 1

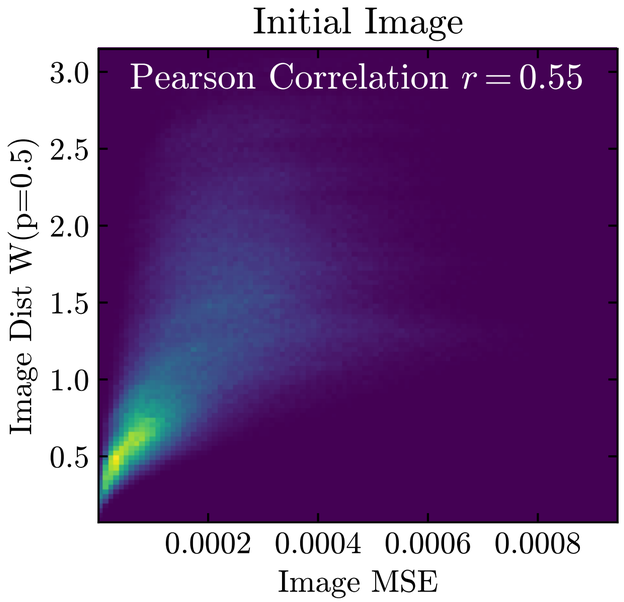
Keith Conrad, L^p spaces for p < 1
https://kconrad.math.uconn.edu/blurbs/analysis/lpspace.pdf

Backup: Using EMD to train the network

(work in progress)

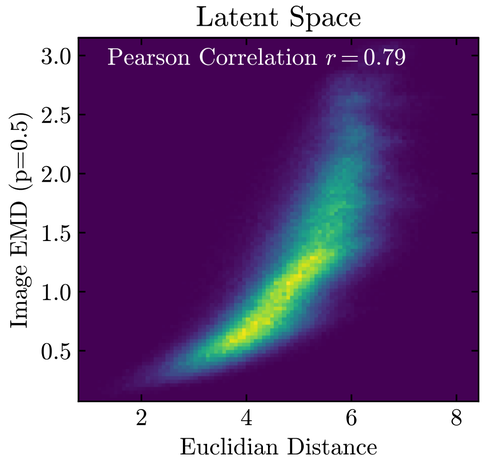
An MSE trained CVAE showed strong correlations between Latent space/Downsampling Layers and Input Images

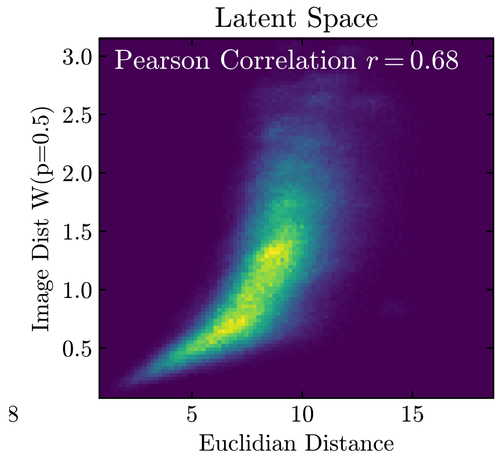
How about using EMD to train the network itself?



EMD trained

MSE trained





Higher correlation in latent space for EMD trained networks, as compared to MSE trained networks