# Better latent spaces for better autoencoders

Barry M. Dillon

July 8, 2021

Institute for Theoretical Physics
University of Heidelberg

Universität Heidelberg
Zukunft. Seit 1386.

'Batter latent spaces for better autoencoders', hep-ph/2104.08291

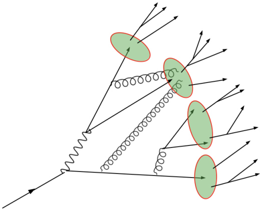
BMD, Tilman Plehn, Christof Sauer, and Peter Sorrenson

# Outline

1. Jet images and VAEs
2. Latent space classification
3. Summary

# Jet images

## Boosted jets

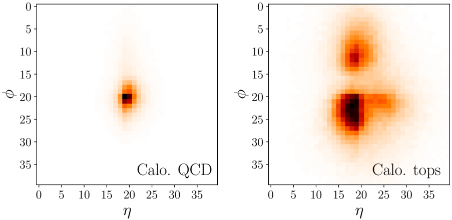


‘ML landscape of top taggers’ Plehn et al

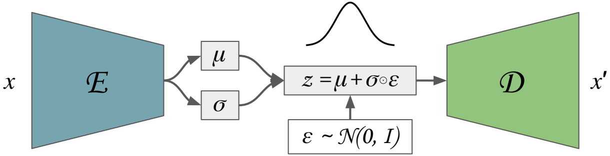
## Pre-processing

- centre in $(\eta, \phi)$
- rotate so that the principle axis points along $\eta = 0$
- pixelise to a 40x40 image
- normalise the $p_T$ in the pixels

## Jet images QCD & top



# The Variational Autoencoder (VAE)




1. Encoding: jet image $\rightarrow q_\phi(z|x)$ (a latent Gaussian distribution)
2. Decoding: $q_\phi(z|x) \rightarrow \vec{z} \rightarrow$ reconstructed jet image

The VAE loss:

$$\mathcal{L} = \Big\langle - \big\langle \log p_\theta(x|z) \big\rangle_{q_\phi(z|x)} + \beta_{\mathrm{KL}} D_{\mathrm{KL}}(q_\phi(z|x), p(z)) \Big\rangle_{p_{\mathrm{data}}(x)}$$

The latent space prior gives the latent space structure.

Standard VAE: Gaussian distribution with $\bar{z} = 0$ and $\sigma_z = 1$.

# The problem with autoencoding..

The complexity-anomaly problem

the reconstruction loss tends to trigger based on the complexity of a jet, not how anomalous it is.

LHCOsummer2020 talk: 'Anomaly detection with convolutional autoencoders and latent space analysis', D. Jaroslawski, D. Shih, K. Nash, M. Tran, Y. Gershtein

arXiv:2104.0829, 'Better latent spaces for better autoencoders', BMD, T. Plehn, C. Sauer and P. Sorrenson

arxiv:2104.09051, 'Autoencoders for unsupervised anomaly detection in high energy physics', T. Finke, M. Krämer, A. Morandini, A Mück, I. Oleksiyuk

arxiv:2106.10164 , 'Rare and Different: Anomaly Scores from a combination of likelihood and out-of-distribution models...', S. Caron, L. Hendriks, R. Verheyen

Examples:

- AEs are good at tagging anomalous top jets, but not anomalous QCD jets!
- Tagging dark-matter jets

Why?..

For reconstruction loss: complexity $\leftrightarrow$ out-of-distribution

But pre-processing is very important arxiv:2104.09051, Finke et al

# Latent space classification

We're not the first to do this:
arxiv:2007.01850, T. Cheng, J. Arguin, J. Leissner-Martin, J. Pilette, T. Golling
arxiv:2010.07940, M. van Beekvelda, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. R. de Austri, M. Santoni and M. White
arxiv:2103.06595, B. Bortolato, BMD, J F. Kamenik, A. Smolkovič

## Why latent space classification?

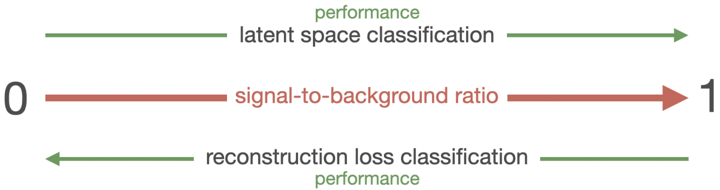
- The latent space encodes physical information about the jet
  $\rightarrow$ Different jets should be separated in latent space
- Limit to very small latent spaces
  $\rightarrow$ We want to encode something related to a class label

# Latent space classification

We're not the first to do this:
arxiv:2007.01850, T. Cheng, J. Arguin, J. Leissner-Martin, J. Pilette, T. Golling
arxiv:2010.07940, M. van Beekvelda, S. Caron, L. Hendriks, P. Jackson, A. Leinweber, S. Otten, R. Patrick, R. R. de Austri, M. Santoni and M. White
arxiv:2103.06595, B. Bortolato, BMD, J F. Kamenik, A. Smolkovič

## Why latent space classification?

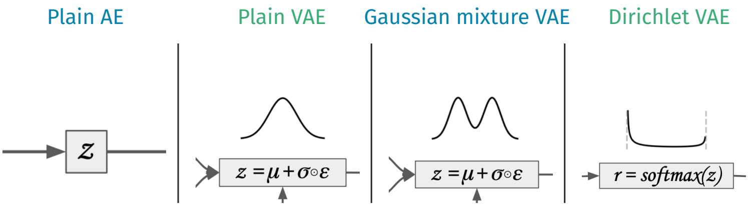
- The latent space encodes physical information about the jet
  $\rightarrow$ Different jets should be separated in latent space
- Limit to very small latent spaces
  $\rightarrow$ We want to encode something related to a class label




Can we retain performance in the low S/B limit?

# A choice of latent spaces

arXiv:2104.0829, ‘Better latent spaces for better autoencoders’, BMD, T. Plehn, C. Sauer and P. Sorrenson




- GMVAE

  Latent space is a Gaussian mixture model.

  Means and variances of the mixtures are learned.

- DVAE

  Latent space is a multinomial mixture model, with jets being assigned mixture weights for each mixture.

  Use a prior to shape latent space, and impose a hierarchy in the mixtures.

  Very similar to Latent Dirichlet Allocation models!

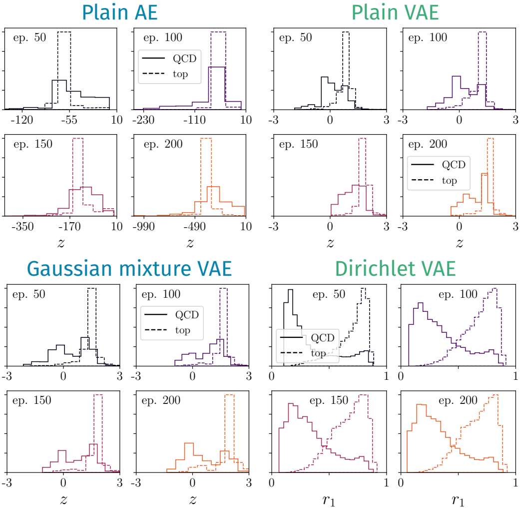
  ‘Uncovering latent jet substructure’, arxiv:1904.04200, BMD, D. A. Faroughy, J. F. Kamenik

  ‘Learning the latent structure of collider events’, arxiv:2005.12319, BMD, D. A. Faroughy, J. F. Kamenik, M. Szewc

# The first test: stability and bi-modality

The test:

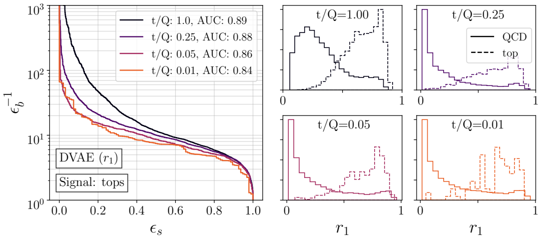
- AE and VAEs have 1D latent spaces ($z$)
- 100k QCD jets
  100k top jets
- train networks for 200 epochs
- loss converges at $\sim$100 epochs



# DVAE: tagging anomalous top jets

Dirichlet VAE with 2 mixtures (1D latent space) a varying t/Q ratio.

Hierarchical Dirichet prior to define a 'background' mixture for t/Q<1, $\alpha = [1.0, 0.25]$.




Prior $\Rightarrow$ top jets are consistently pushed to the second mixture

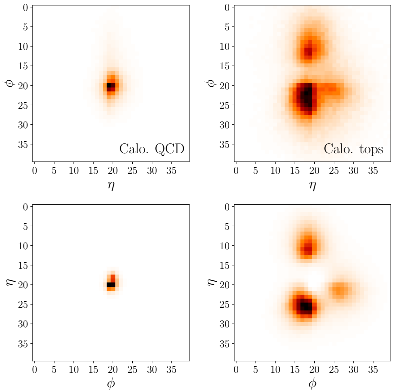
The prior tells us where to look for anomalies in latent space!

# DVAE: an interpretable latent space

The Dirichlet VAE has a mixture model structure in latent space.

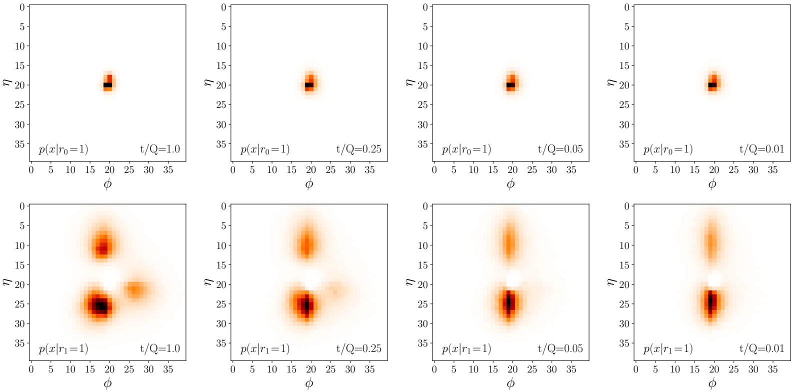
We can interpret what these mixtures represent using the decoder.

Calorimeter images **vs** the learned mixture distributions for t/Q=1:



# DVAE: an interpretable latent space

What does the network learn as we vary t/Q?



# DVAE: tagging anomalous QCD jets?

We now go to a latent space with 3 mixtures (2D latent space)

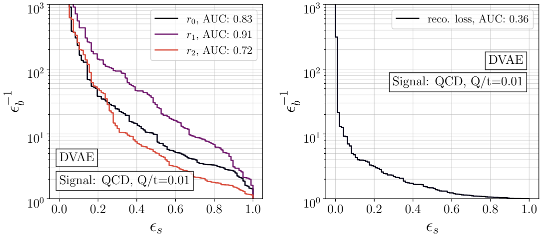
Hierarchical Dirichet prior to separate features, $\alpha = [1.0, 0.25, 0.1]$.

# DVAE: tagging anomalous QCD jets?

We now go to a latent space with 3 mixtures (2D latent space)

Hierarchical Dirichet prior to separate features, $\alpha = [1.0, 0.25, 0.1]$.

We find:



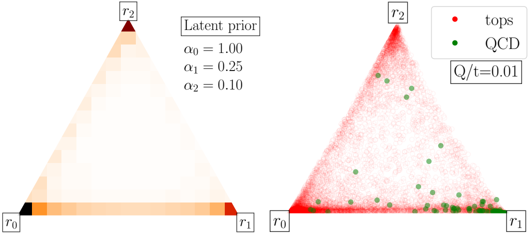

Now we find excellent performance with classification in latent space!

... the difficulty is in knowing which latent space direction to use.

# Latent space interpretation

The latent space has 3 mixtures (2D latent space)

We can visualise the embedding on the Gibbs triangle:

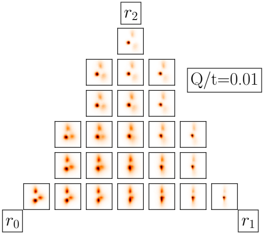



- the latent space is structured hierarchically
- extracts features at different levels of prevalence in the dataset
- organises the jets accordingly

# Latent space interpretation

The latent space has 3 mixtures (2D latent space)

We can visualise the embedding on the Gibbs triangle:




- the latent space is structured hierarchically
- extracts features at different levels of prevalence in the dataset
- organises the jets accordingly

1. Jet images and VAEs

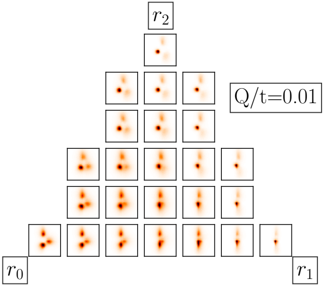
2. Latent space classification

3. Summary

# Summary

- AutoEncoders are still the best tool we have for out-of-distribution anomaly detection
- reconstruction-loss $\Rightarrow$ complexity-anomaly problem
- Latent-space classification is an interesting alternative!
- .. but Gaussian latent spaces don't make good classifiers..

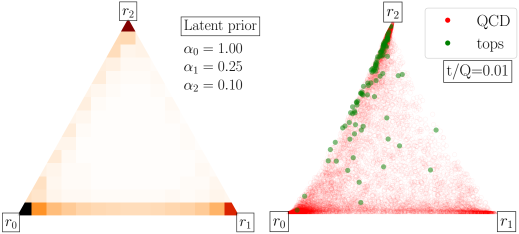
- The Dirichlet-VAE seems like an improvement.
- But it's only an example.
  There's a lot to explore with latent space classification!



# Additional slides

Anomalous top tagging with three mixtures (2D latent space)

We can visualise the embedding on the Gibbs triangle:

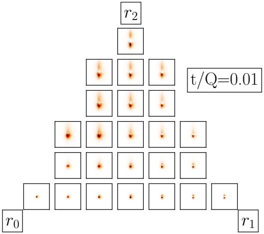



- When we have mostly QCD jets, the DVAE learns jet images from one and 2-prong QCD jets with different angular separation
- top jets mapped to the wide angle 2-prong region of latent space

# Additional slides

Anomalous top tagging with three mixtures (2D latent space)

We can visualise the embedding on the Gibbs triangle:




- When we have mostly QCD jets, the DVAE learns jet images from one and 2-prong QCD jets with different angular separation
- top jets mapped to the wide angle 2-prong region of latent space