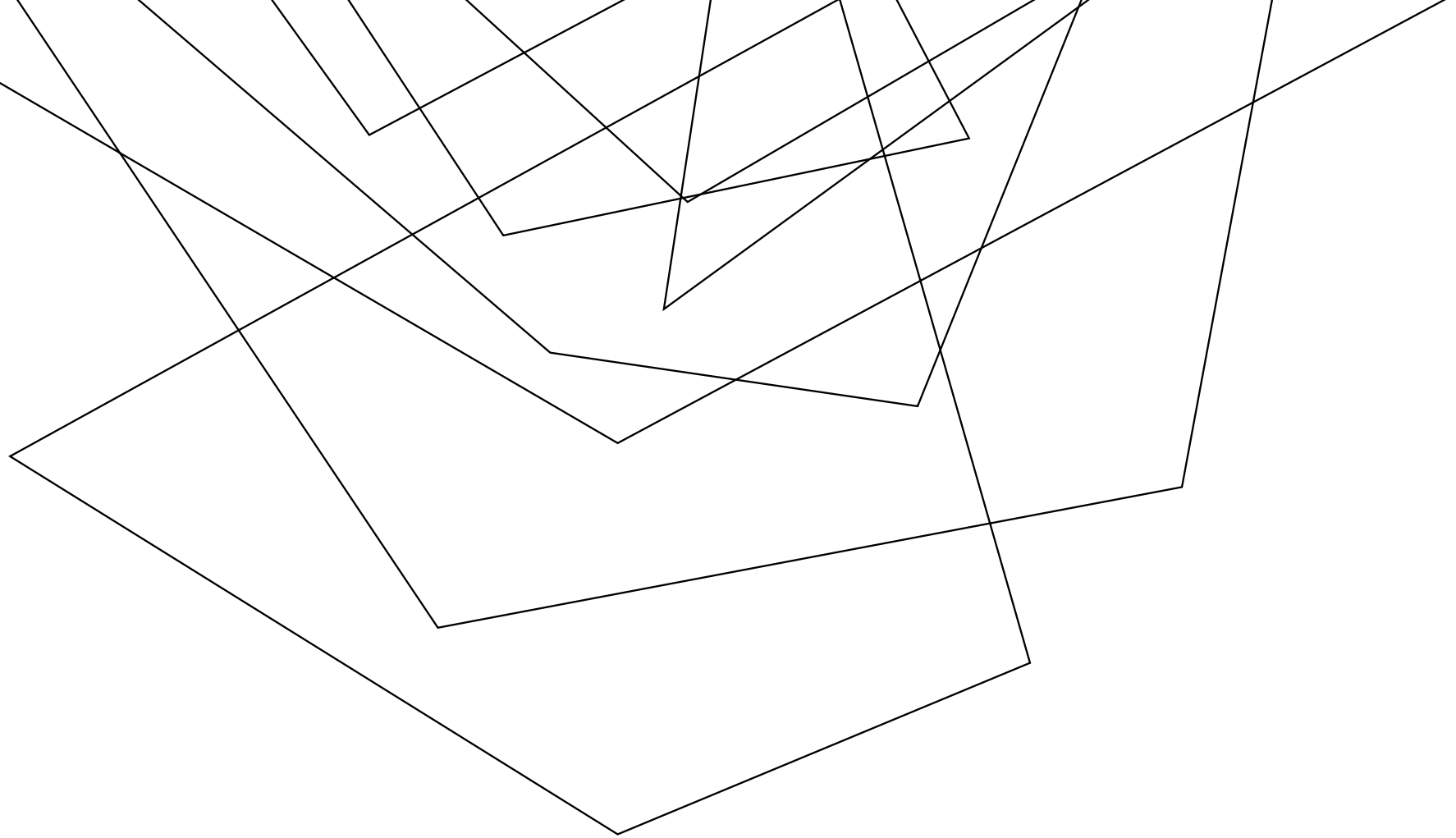


AUTOENCODER

Thorben Finke (RWTH Aachen University)

Train-the-Trainer Workshop, Dortmund 19.06.2023

OUTLINE

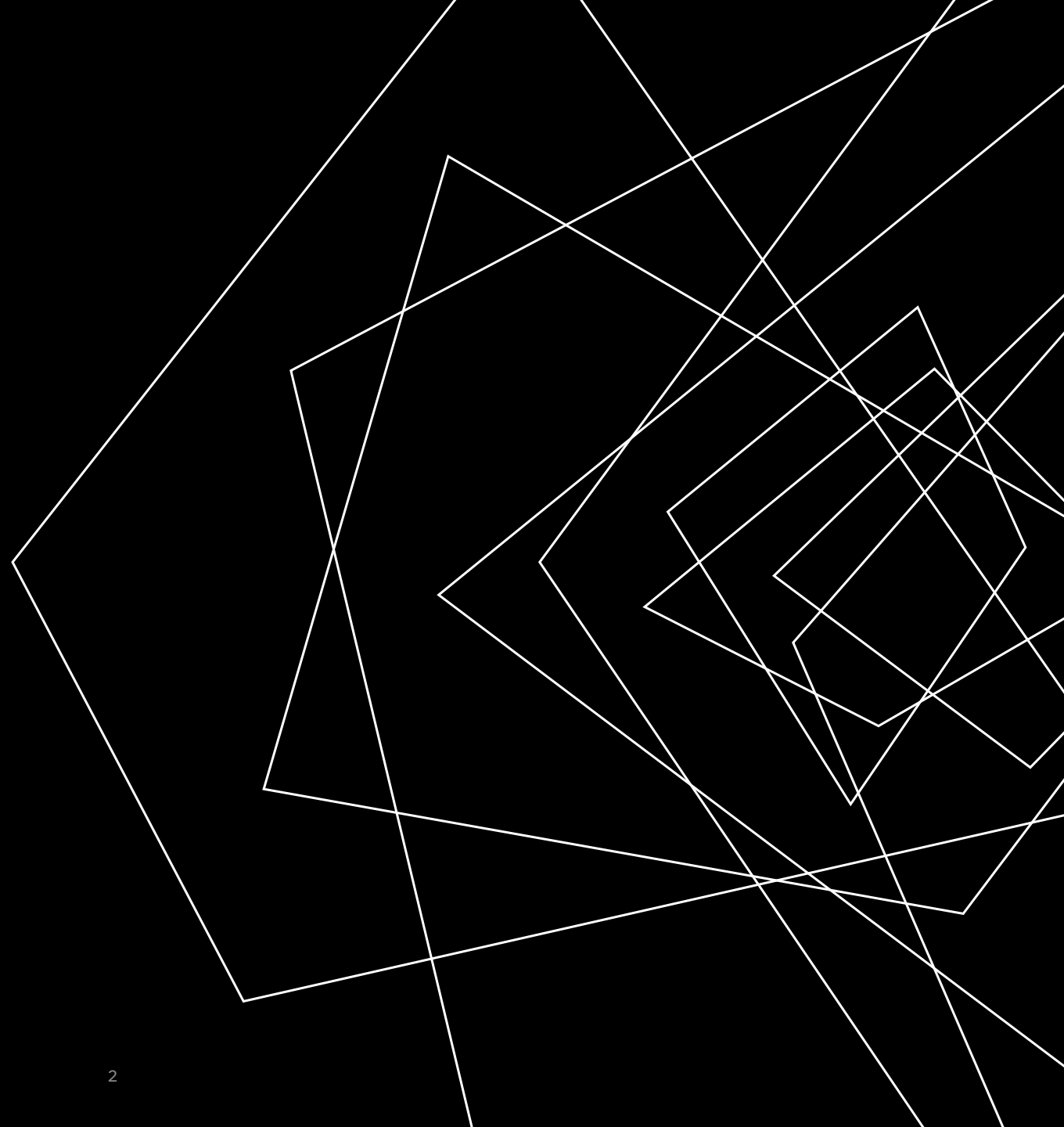
Introduction

Basic setup

Autoencoder variants

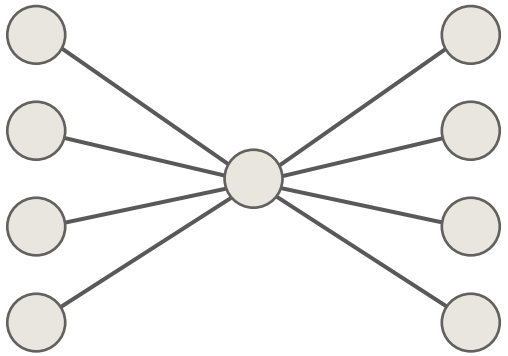
Detecting anomalies

Summary



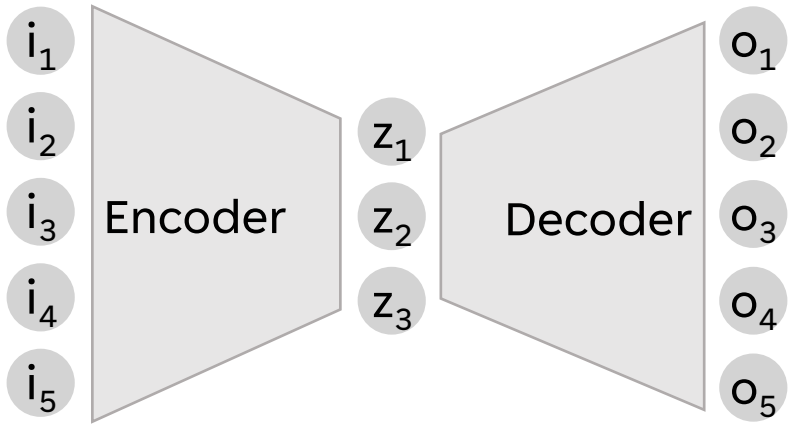
INTRODUCTION

- Autoencoder perform data compression and expansion
- Best linear encoding $\boldsymbol{E} \in \mathbb{R}^{c \times d}$ and decoding $\boldsymbol{D} \in \mathbb{R}^{d \times c}$ into orthonormal basis:
  Principle Component Analysis (PCA)
  $$\vec{x}' = \boldsymbol{DE}\vec{x} \text{ minimize } \|\vec{x} - \vec{x}'\|^2$$
- Results in $\boldsymbol{E} = \boldsymbol{D}^T$ eigenvectors of the covariance matrix of the data
- Visualization for $d = 4, c = 1$



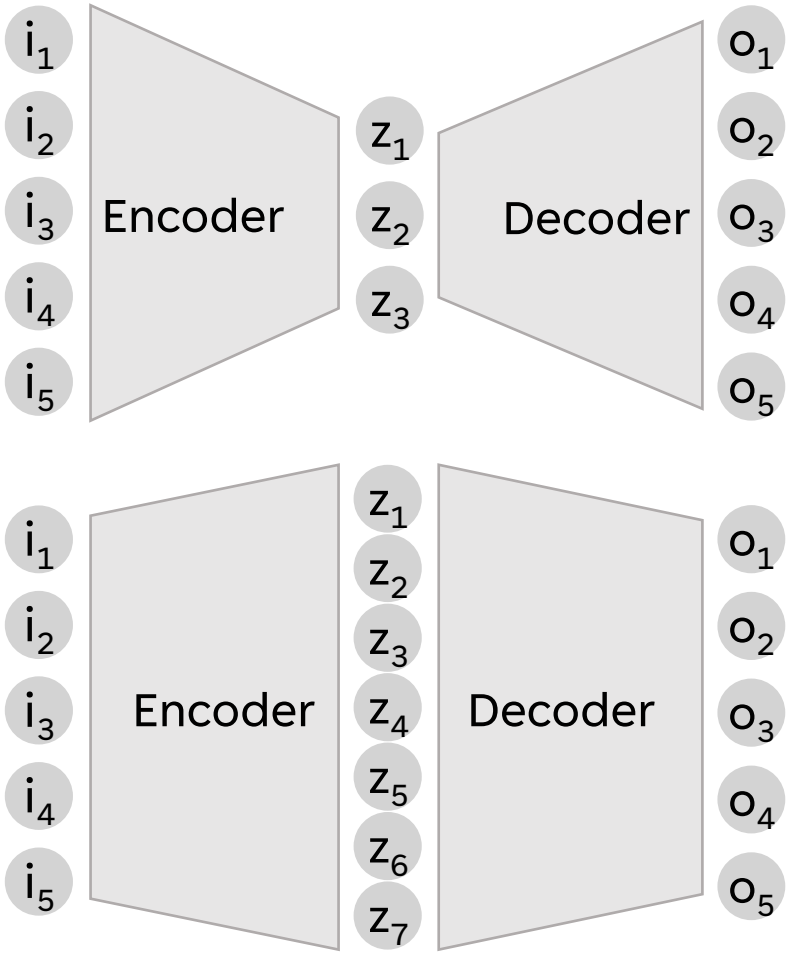
BASIC SETUP

- Autoencoder consist of two parts: Encoder and decoder
- Encoder projects input $\boldsymbol{i}$ to some latent representation $\boldsymbol{z}$
- Decoder projects back to the input dimension
- Train to minimize difference between input and output
- Unsupervised method, since no labels required



BASIC SETUP

- Undercomplete AE: Reduzed latent dimension
- Overcomplete AE: Sparsity constraint
- Applications:
  - Feature extraction
  - Denoising
  - Data generation
  - Anomaly detection

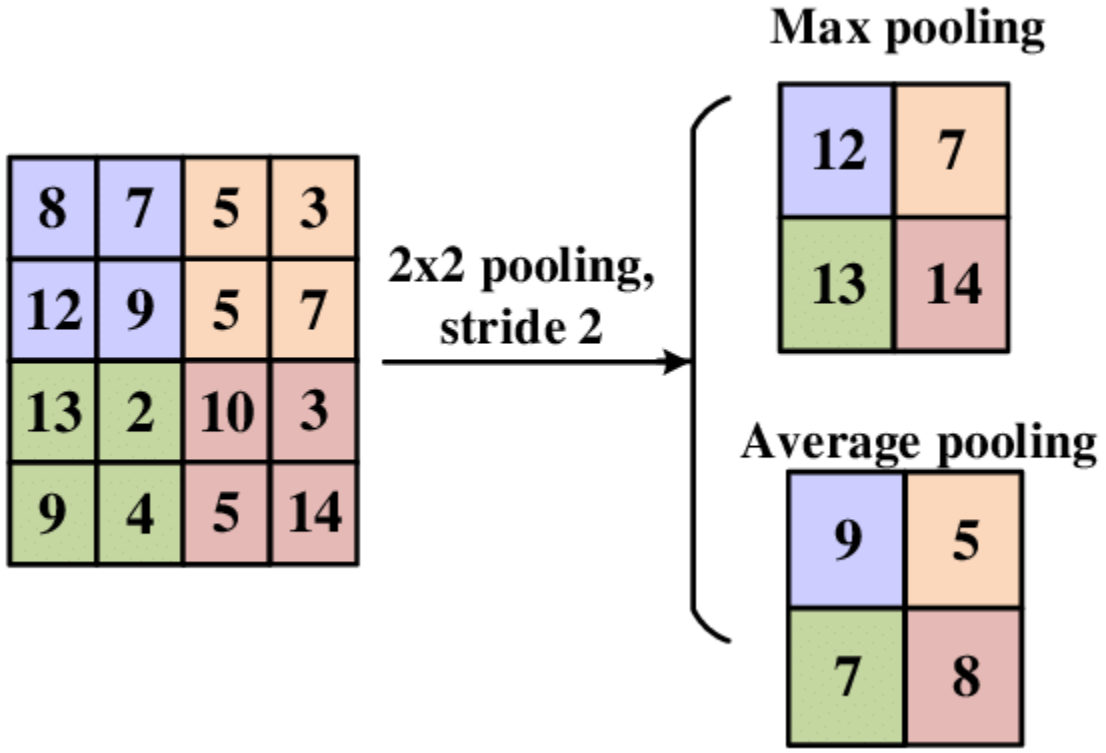


VARIOUS INPUTS

- Features
  - Fully connected architectures
- Sequences
  - Recurrent architectures
- Images
  - Convolutional architectures
- Graphs
  - Graph convolutional architectures

CONVOLUTIONAL AUTOENCODER

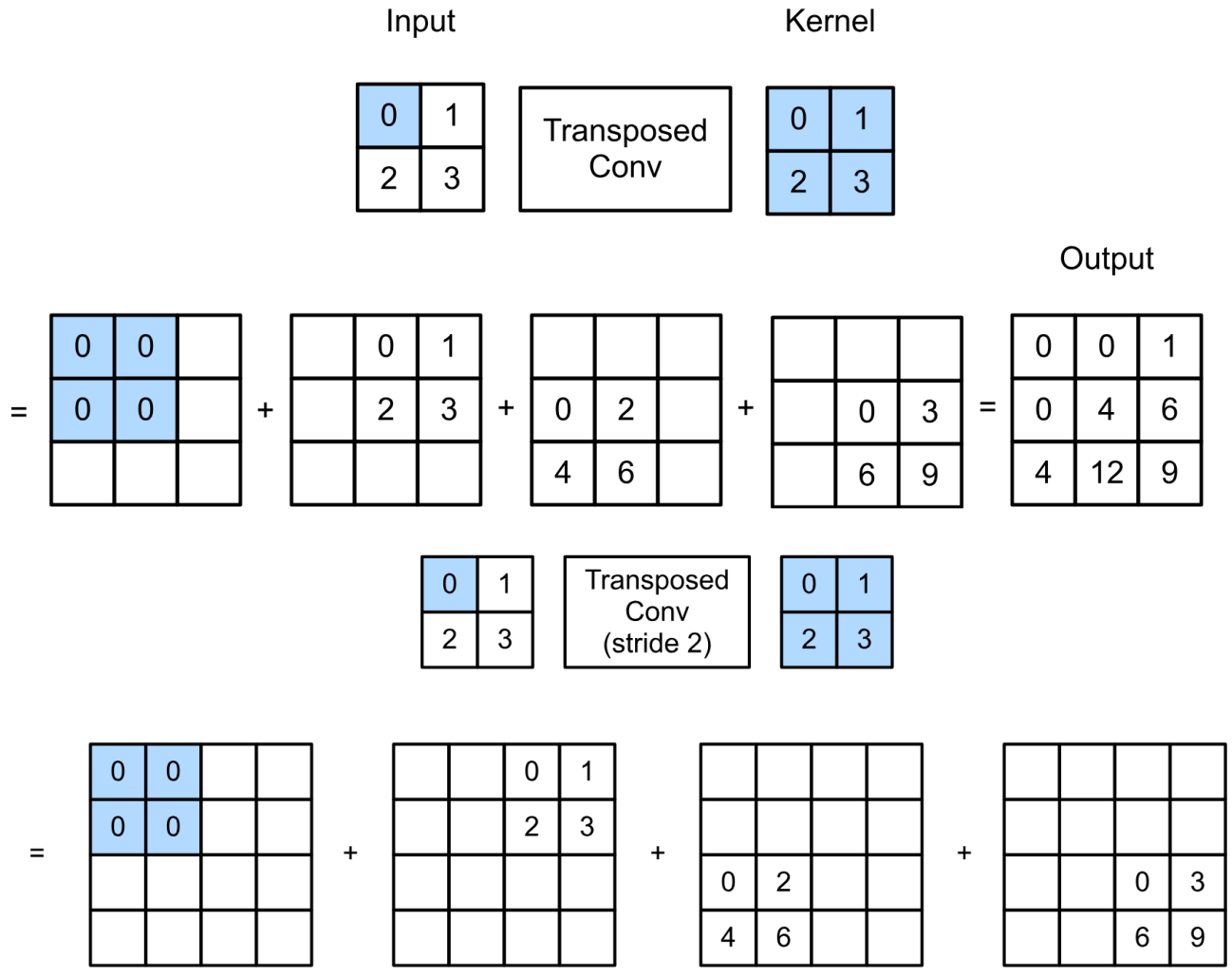
- **Dimensional reduction**
  - Pooling operations
  - Convolutions with stride > 1
- Dimensional inflation
  - Upsampling
  - Transposed convolutions



Source

CONVOLUTIONAL AUTOENCODER

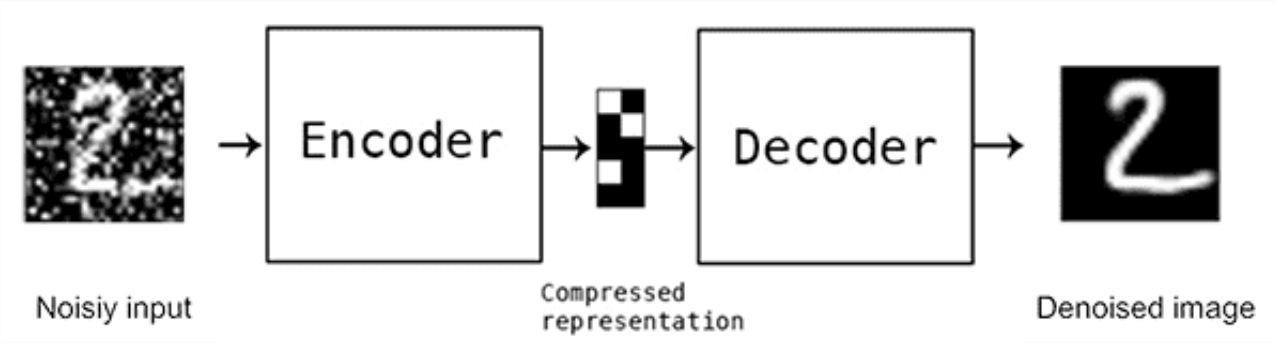
- Dimensional reduction
  - Pooling operations
  - Convolutions with stride > 1
- **Dimensional inflation**
  - Upsampling
  - Transposed convolutions



Source

DENOISING AUTOENCODERS

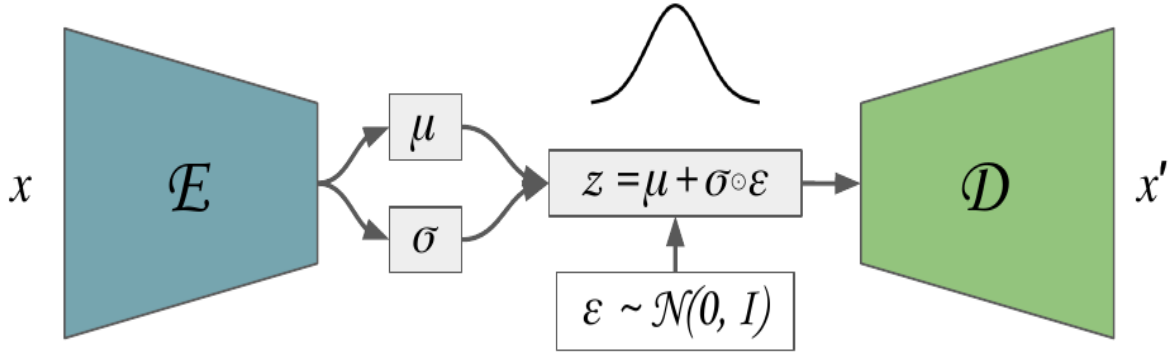
- Add noise to the inputs
- Train to reconstruct input without noise
- Originally intended to learn more robust representations
- Reduced risk of overfitting
- Physics applications
  - Gravitational waves 1711.09919
  - Foreground removal in cosmology 1902.09278
  - Radio astronomy 2110.08618
  - Air shower signal recovery 1901.04079



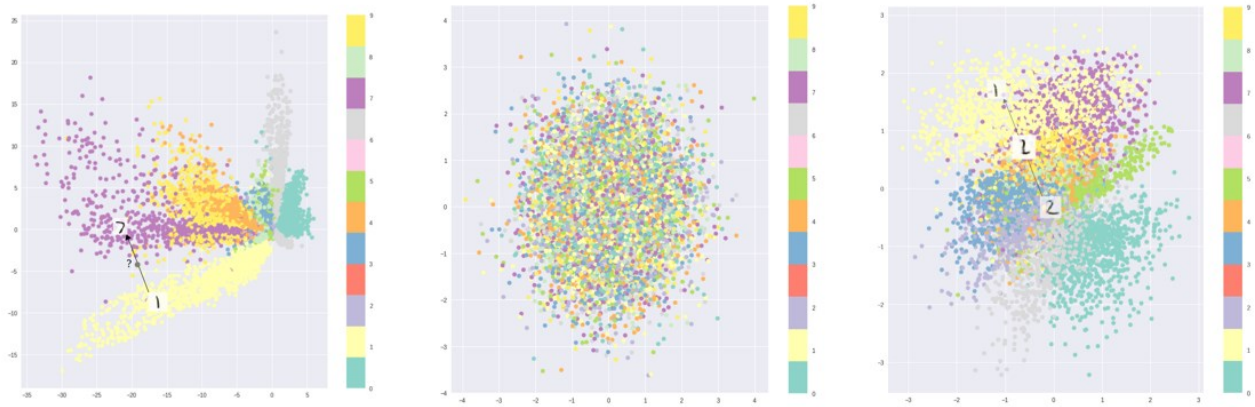
Source

VARIATIONAL AUTOENCODER

- Replace deterministic latent space by trainable distribution
- Reparameterization trick to update the distribution parameters
- Additional term in the loss imposing a given distribution
  - $\mathcal{L} = \beta D_{KL}(p, q) + \|x - x'\|^2$
- Results in smooth and more interpretable latent space



2211.01421



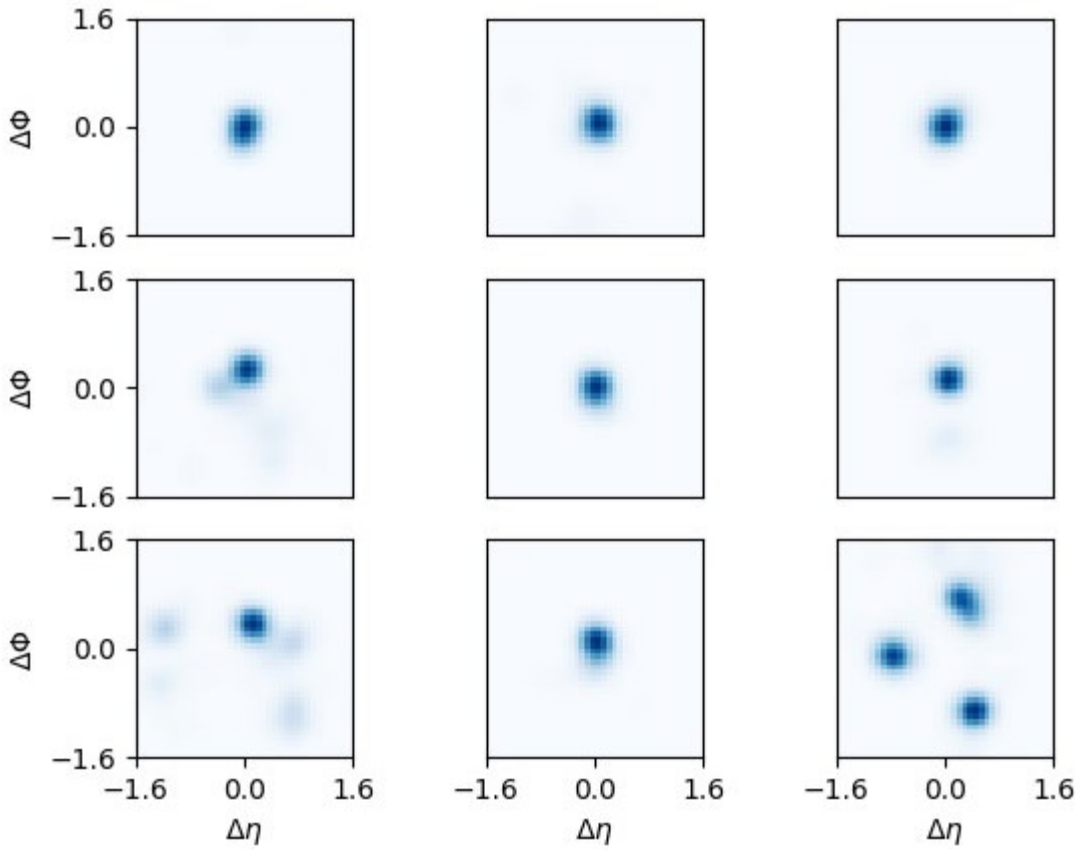
Source

ANOMALY DETECTION USING AUTOENCODERS

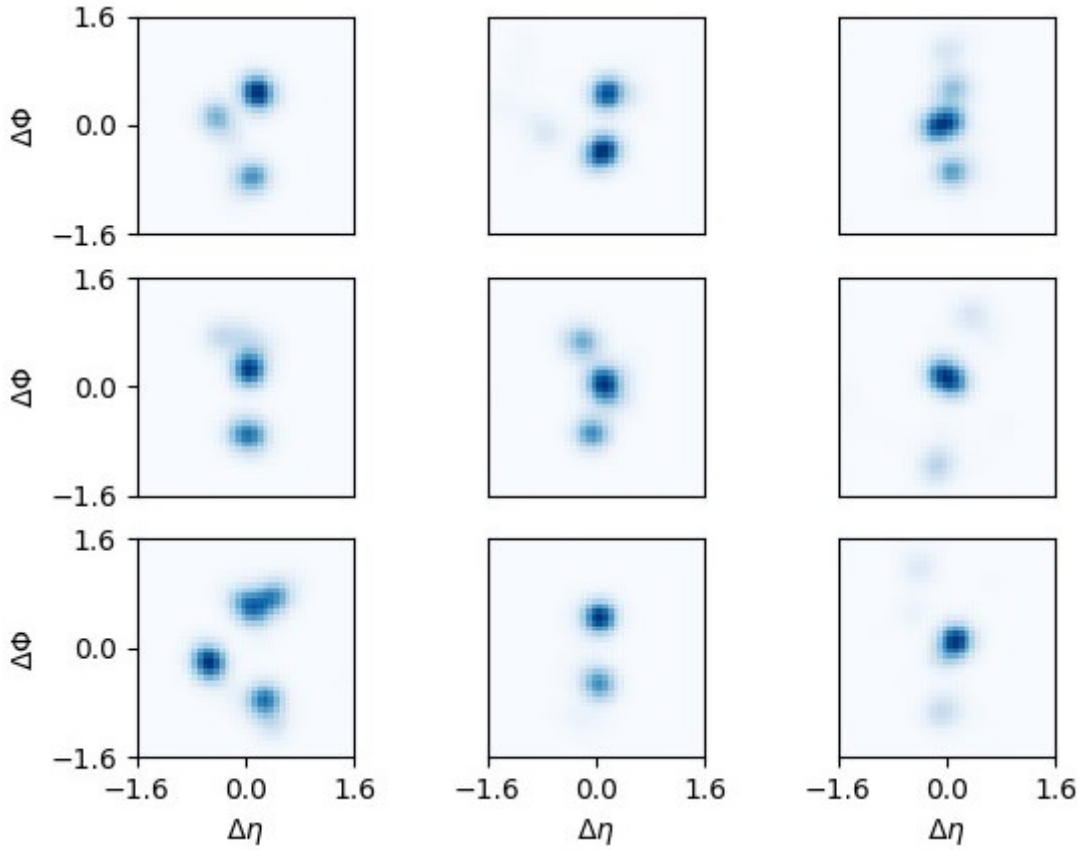
- Learn what's normal to find anomalies
- During training learn meaningful representation of normal data
- Idea: reconstruction of anomalous data is worse than for normal data
  → Use the reconstruction loss as anomaly score during inference
- Example: anomalous jet tagging on the top tagging benchmark dataset
  10.5281/zenodo.2603256

ANOMALY DETECTION USING AUTOENCODERS – JET IMAGE EXAMPLE

Light QCD jets

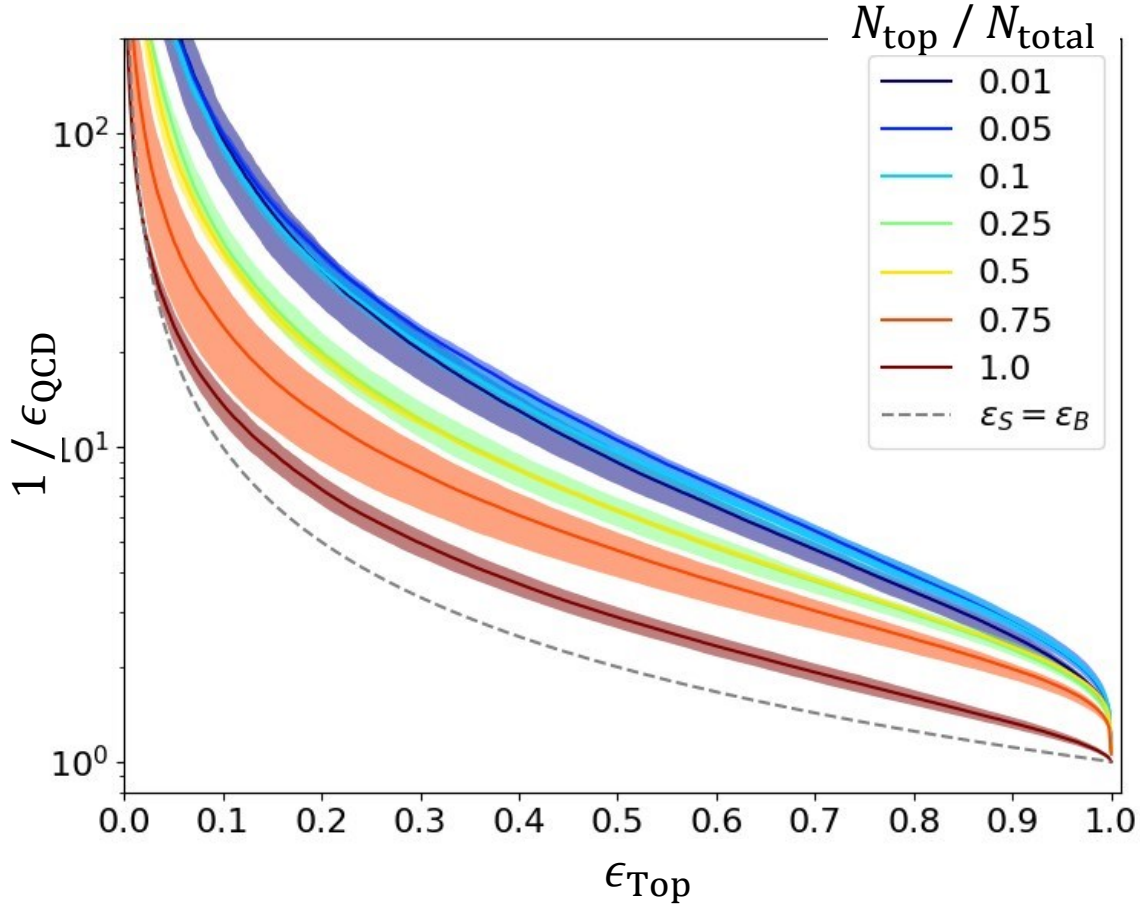


Top jets



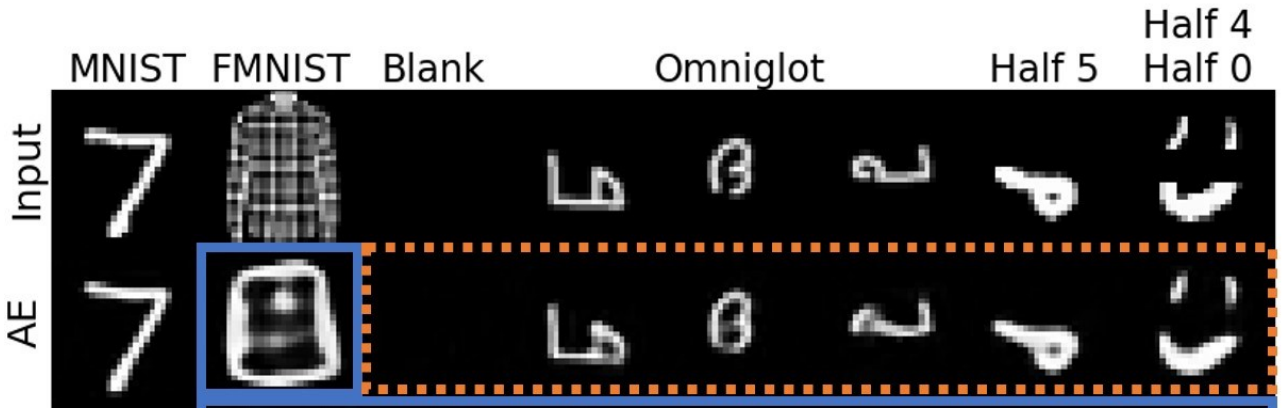
ANOMALY DETECTION USING AUTOENCODERS – JET EXAMPLE

- Need to check signal contamination during training
- Expectation: at ~0.5 this should become random
- Outcome: even for training on top only, top jets are reconstructed worse



ANOMALY DETECTION USING AUTOENCODERS

- Train an autoencoder on MNIST data
- Check the reconstruction of outliers
  - Generalization beyond the original data manifold
- Solution: Autoencoding under normalization constraints 2105.05735



2105.05735

NORMALIZED AUTOENCODER

- Energy based model

$$p_\theta(x) = \frac{1}{\Omega_\theta} \exp\left(-\frac{\mathrm{E}_\theta(x)}{T}\right)$$

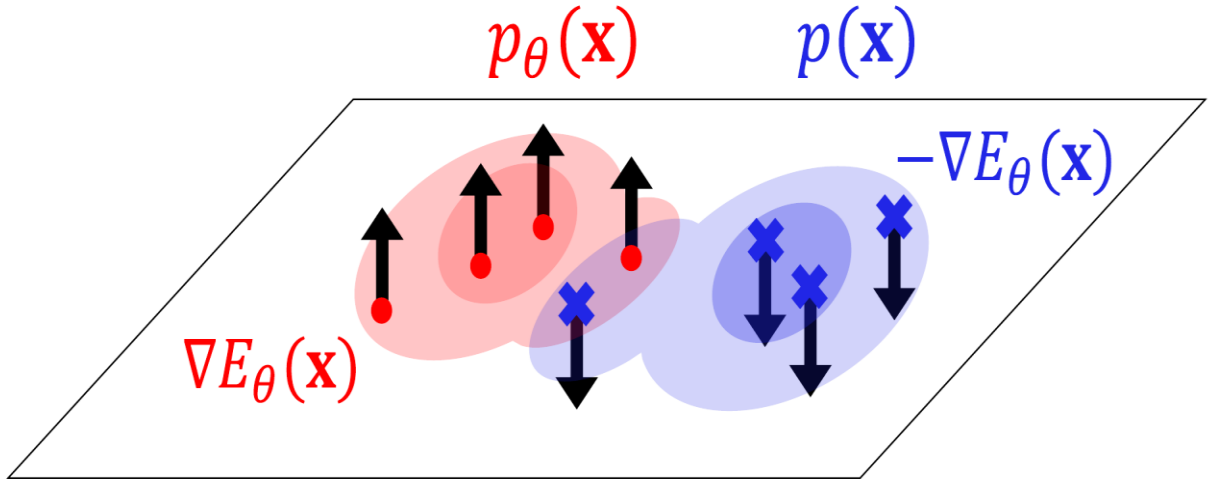
- Normalization constraint

$$\Omega_\theta = \int \exp\left(-\frac{\mathrm{E}_\theta(x)}{T}\right) dx$$

- Maximize $p_\theta(x)$:

$$\mathcal{L} = \mathbb{E}_{x\sim p(x)}\left[\frac{E_\theta(x)}{T}\right] - \mathbb{E}_{x\sim p_\theta(x)}\left[\frac{E_\theta(x)}{T}\right]$$

positive energy negative energy



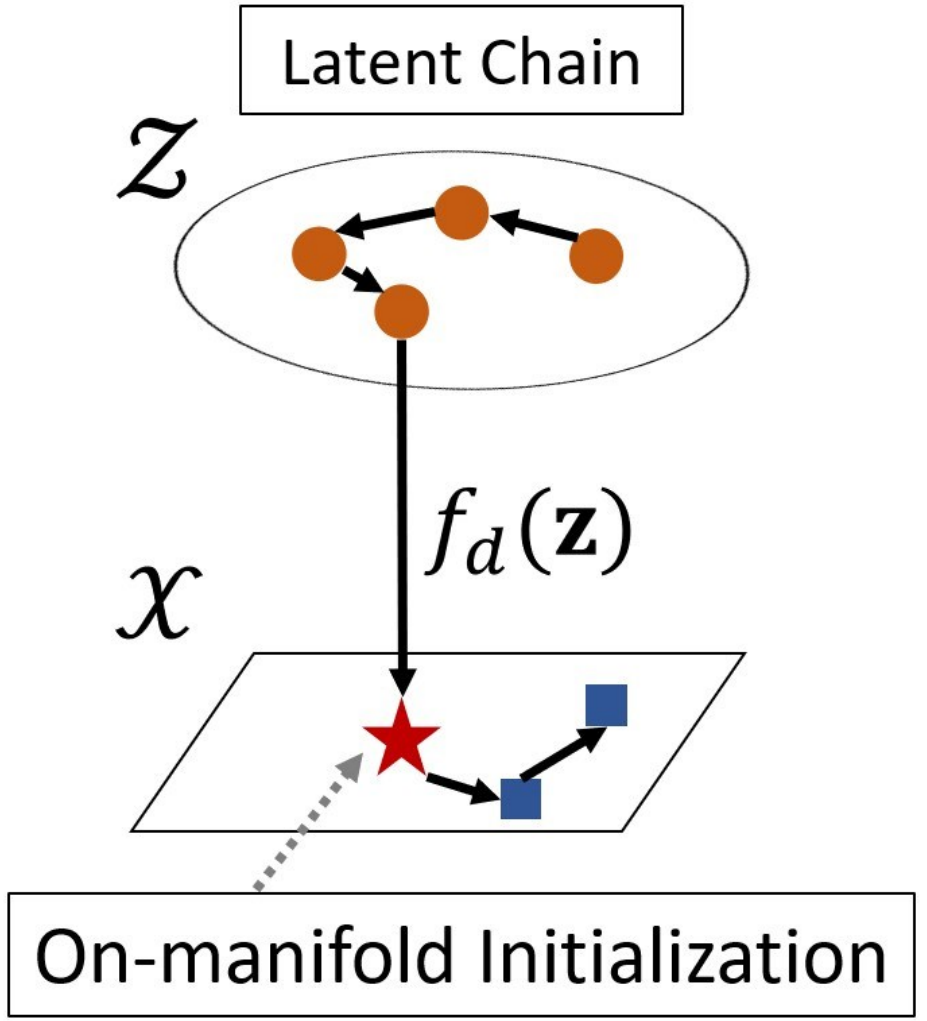
NORMALIZED AUTOENCODER SAMPLING FROM THE MODEL

- Energy defined as reconstruction error
- Negative samples obtained using Langevin Monte Carlo

$$x_{t+1} = x_t + \lambda_x \nabla_x \log\left(\mathrm{p}_\theta(\mathrm{x_t})\right) + \sigma_x \epsilon_t$$

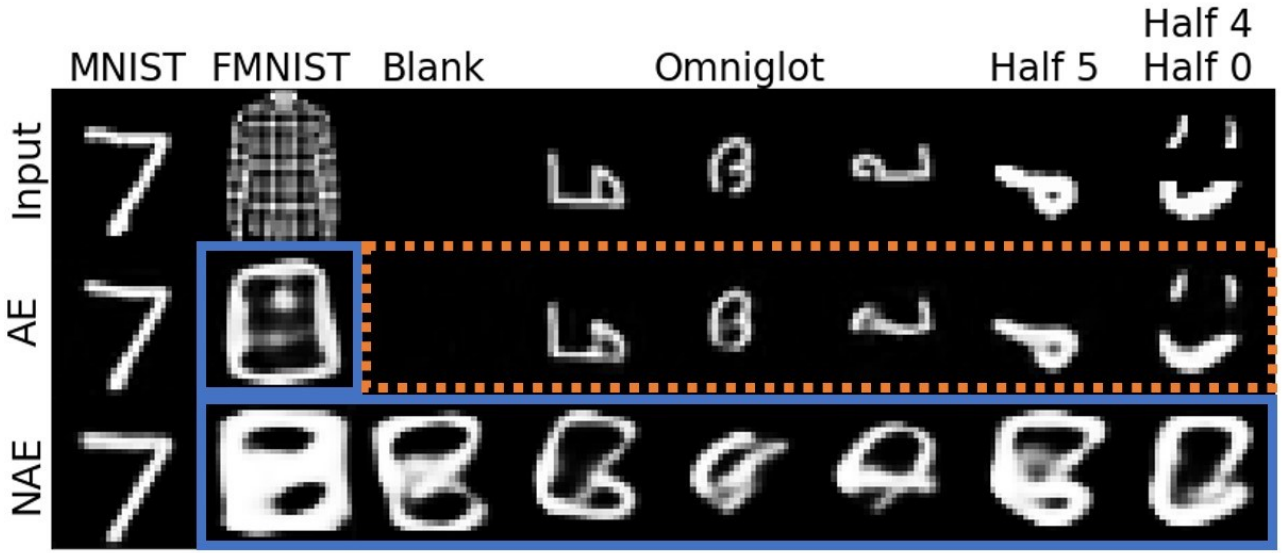
- Use on-manifold initialization (OMI) using the decoder of the autoencoder
  - Introduces second MCMC in the latent space

$$q_\theta(z) = \frac{1}{\Psi_\theta} \exp\left(-\frac{H_\theta}{T_z}\right), H_\theta = E_\theta\left(D(z)\right)$$



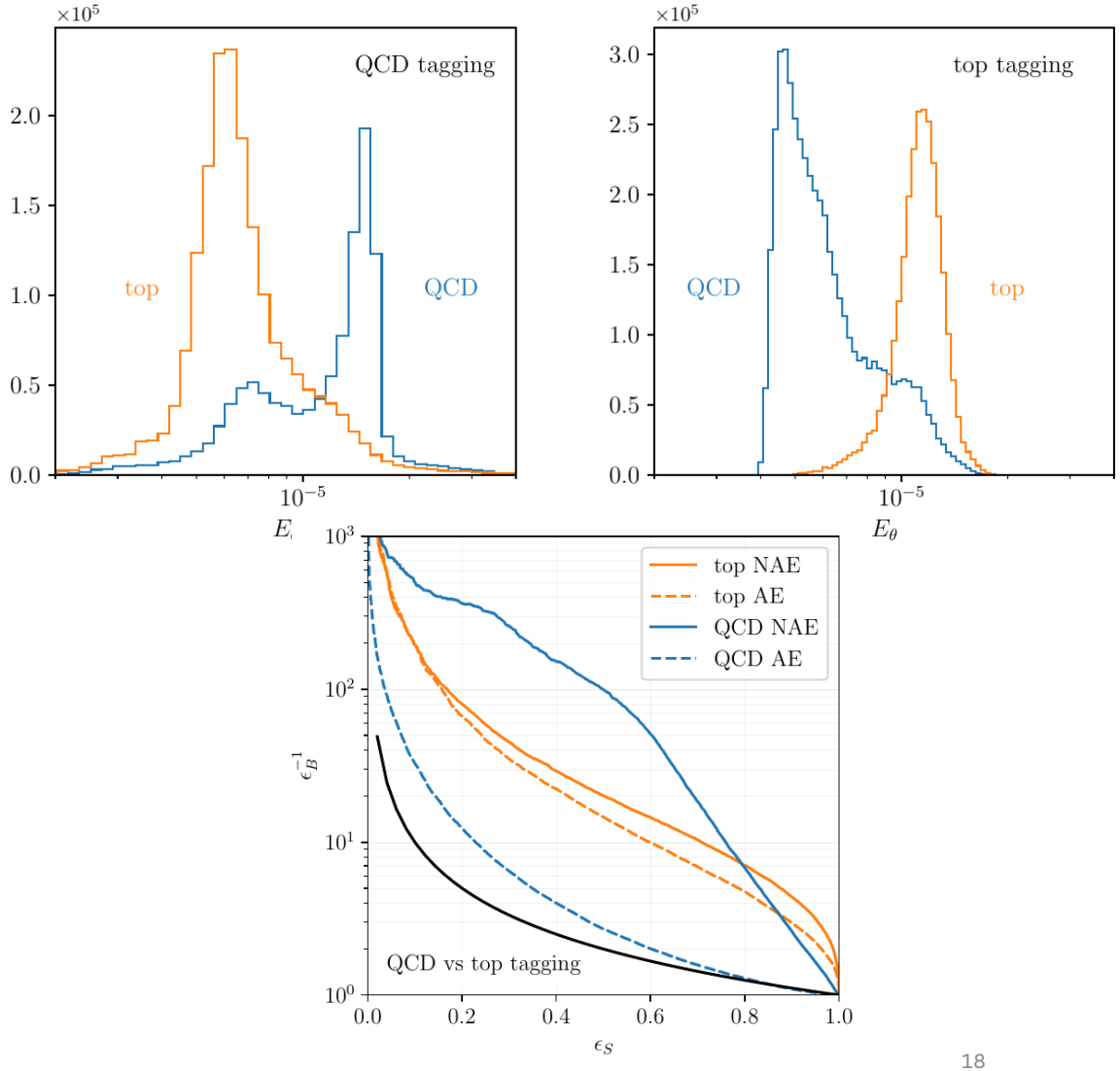
NORMALIZED AUTOENCODER RESULTS

- Results on the MNIST Example from earlier:



NORMALIZED AUTOENCODER IN HEP

- Autoencoder for anomaly detection in HEP show similar issues
- Example QCD and top jets
- Normalized autoencoder shows promising results [2206.14255]

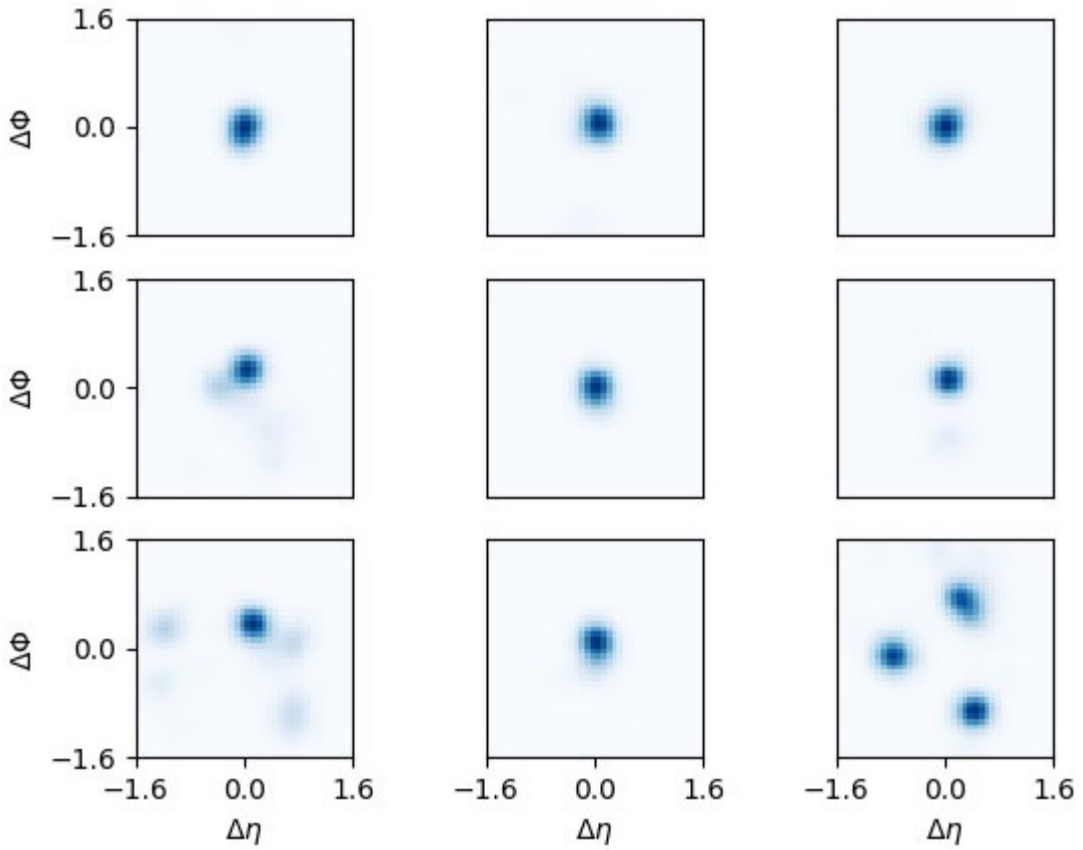


SUMMARY

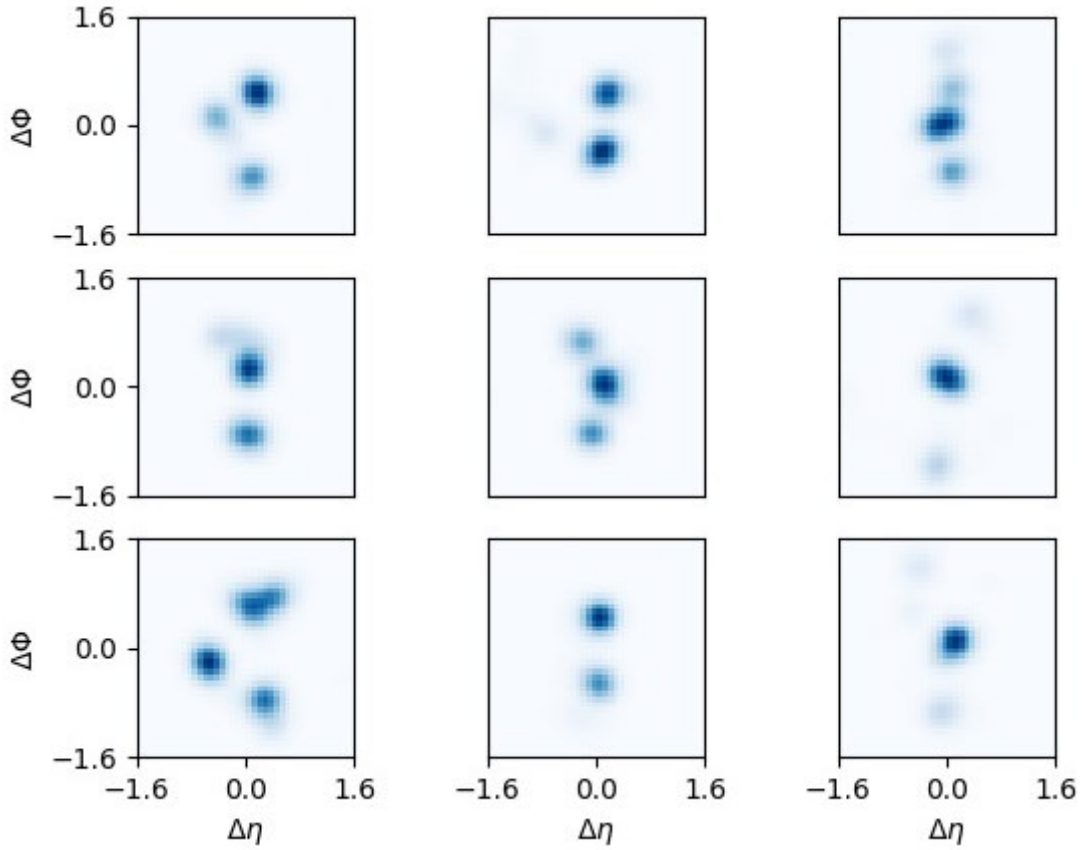
- Autoencoder provide an unsupervised learning algorithm
- Various applications: feature extraction, denoising, data generation, anomaly detection
- Anomaly detection with autoencoders is a difficult task because of generalization capabilities of regular autoencoder

ANOMALY DETECTION USING AUTOENCODERS – JET EXAMPLE

Light QCD jets

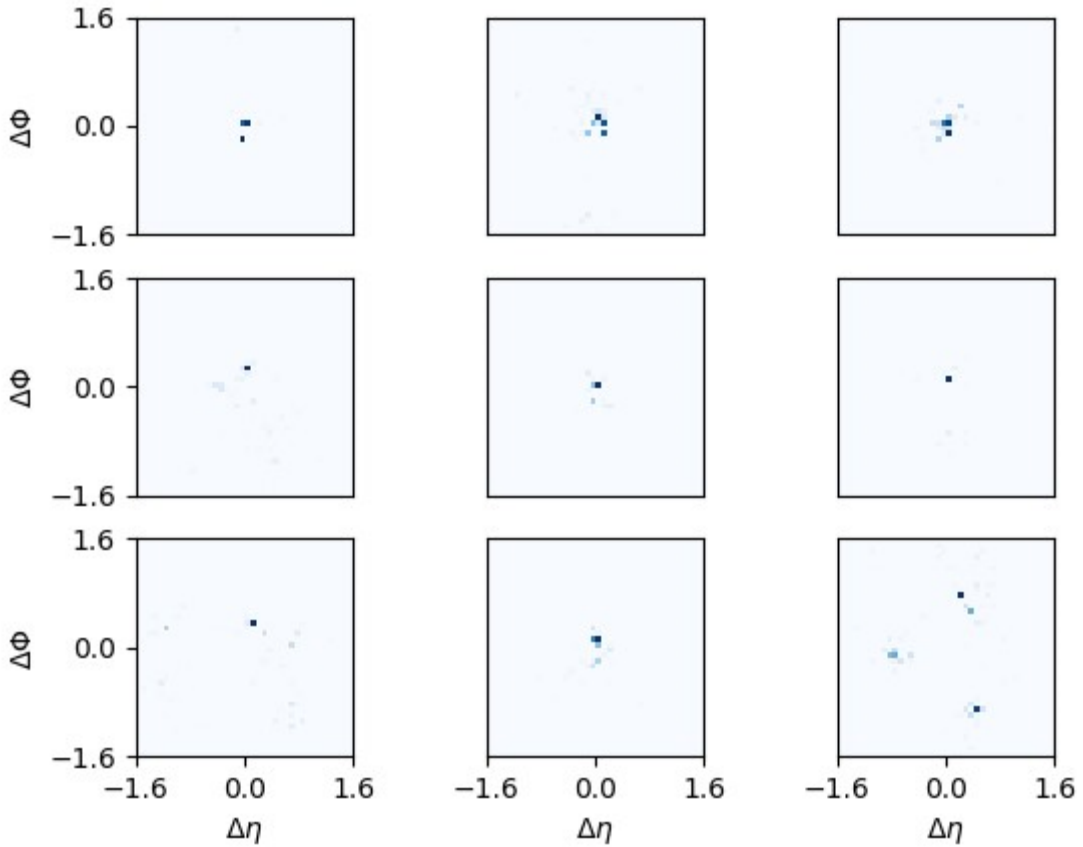


Top jets



ANOMALY DETECTION USING AUTOENCODERS – JET EXAMPLE

Light QCD jets



Top jets