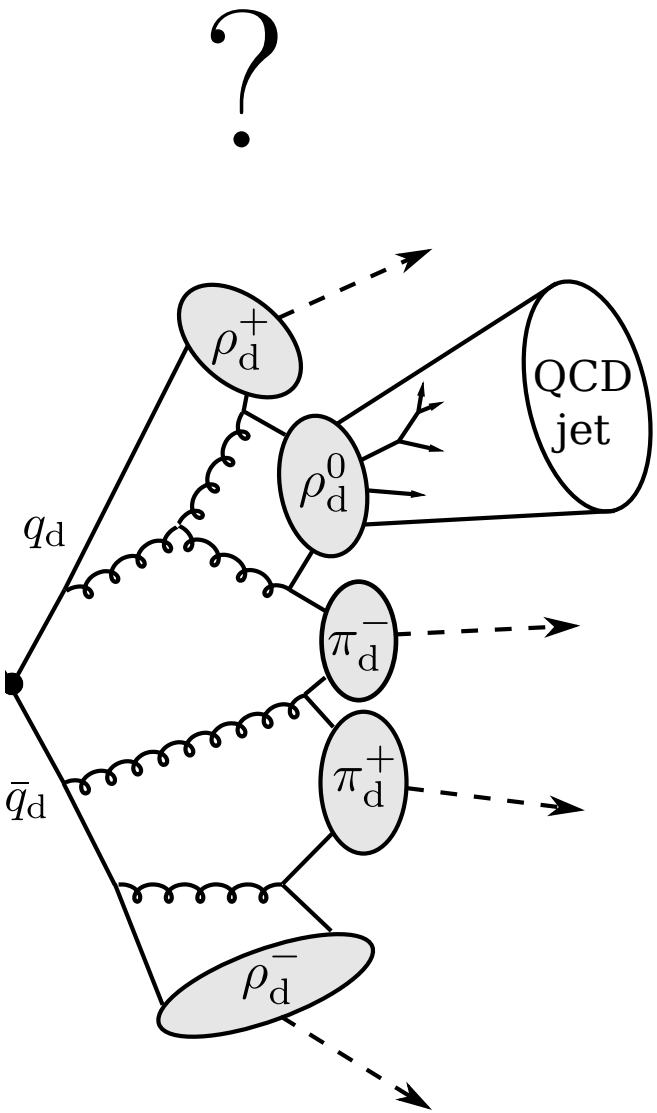
Unsupervised tagging of semivisible jets with energy-based autoencoders in CMS

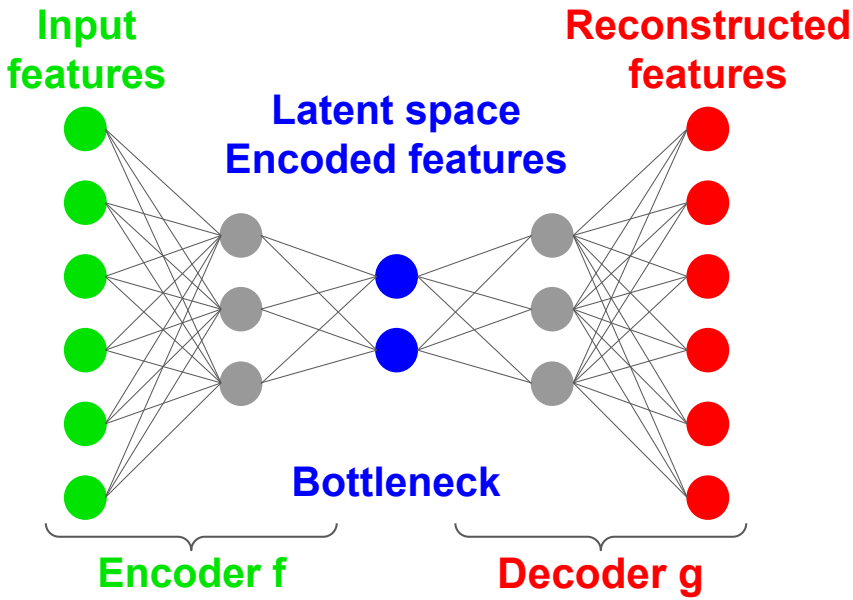
Florian Eble, on behalf of the CMS collaboration

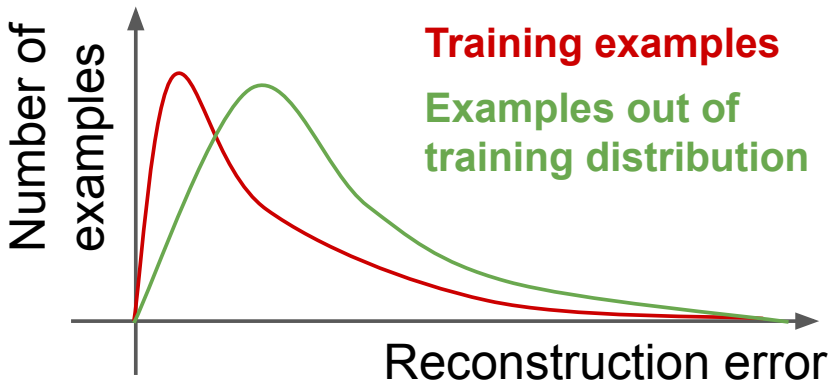
ETH zürich

01/05/2024

EuCAIFCon 2024


?
ρ_{d}^{+}
QCD jet
ρ_{d}^{0}
q_{d}
π_{d}^{-}
π_{d}^{+}
$\bar{q}_{\mathrm{d}}$
ρ_{d}^{-}

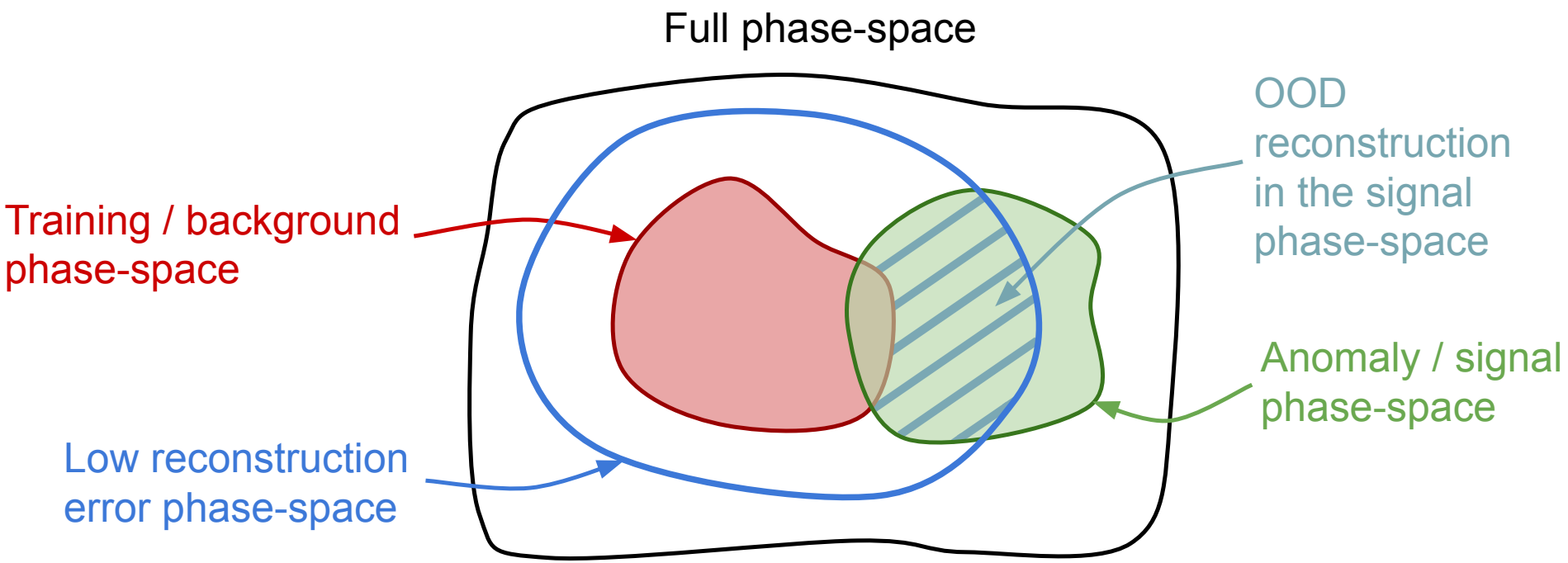

Input features
Latent space
Encoded features
Reconstructed features
Bottleneck
Encoder f
Decoder g


Number of examples
Training examples
Examples out of training distribution
Reconstruction error

The problem of out-of-distribution (OOD) reconstruction

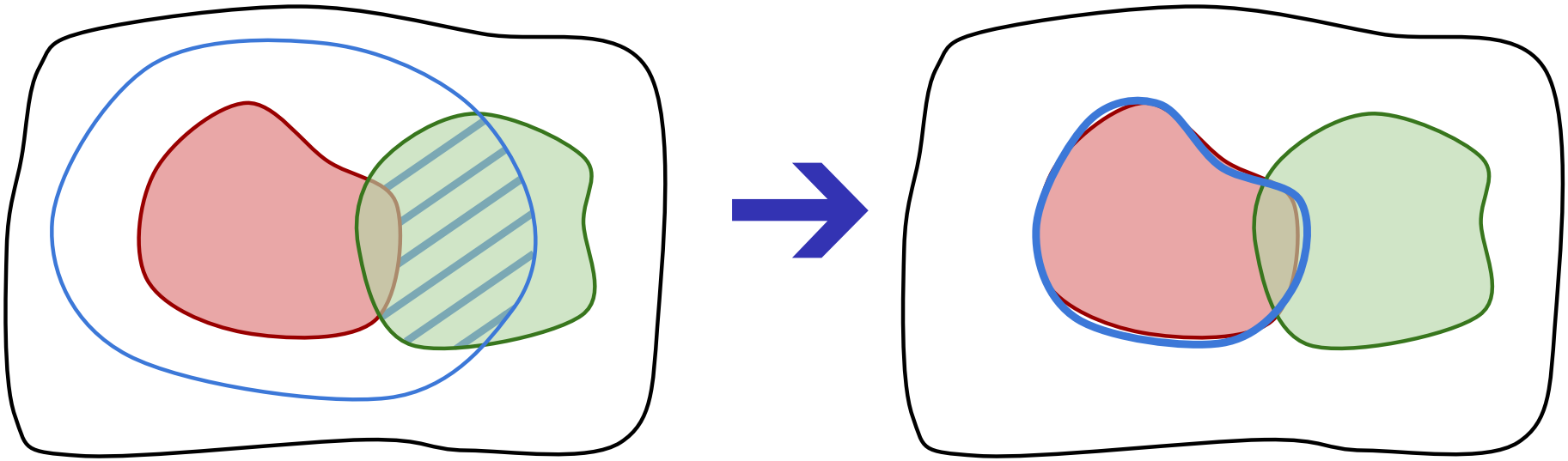
- **AEs are free to minimize reco error outside the background phase space!** including the unknown signal phase space...

➔ This is the problem of **OOD reconstruction**:



Normalized Autoencoders (NAE)

- NAE features a mechanism to suppress OOD reconstruction
- First introduced in arXiv:2105.05735 and used in HEP in arXiv:2206.14225



- NAE paradigm:
  - Define a probability distribution p_θ so that high probability regions have low reco error
  - Sample from p_θ via a MCMC
  - Minimize the distance between the background and p_θ probability distributions
- We propose a different metric to measure this distance, using the Earth Mover's Distance (a.k.a Wasserstein distance) and train NAEs in a fully signal-agnostic fashion