VAEs for Anomalous Jet Tagging

Taoli Cheng
Mila, University of Montreal

In collaboration with

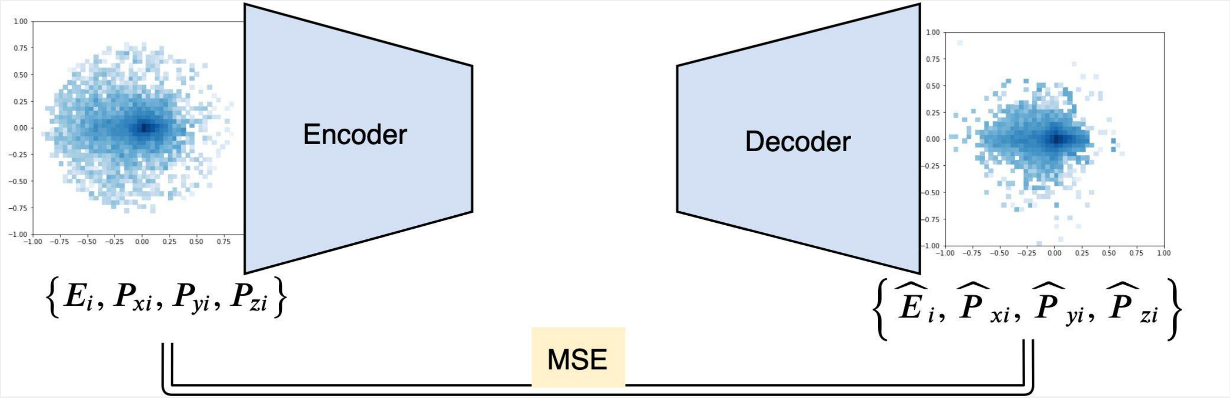
Jean-Francois Arguin, Julien Leissner-Martin, Jacinthe Pilette (Universite de Montreal (CA))
Tobias Golling, Takuya Nobe, Johnny Raine (Universite de Geneve (CH))
Amir Farbin, Debottam Bakshi Gupta (University of Texas at Arlington (US))

Jan. 16, 2020
ML4Jets, NYU

Review: Generative Model for Anomaly Detection

- Explicit (e.g. VAE) or implicit (e.g. GAN) estimation of log p(x)
- Building Blocks
  - Model
  - Embedding Architecture
  - Anomaly Metric

Autoencoder for Anomalous Jet Tagging

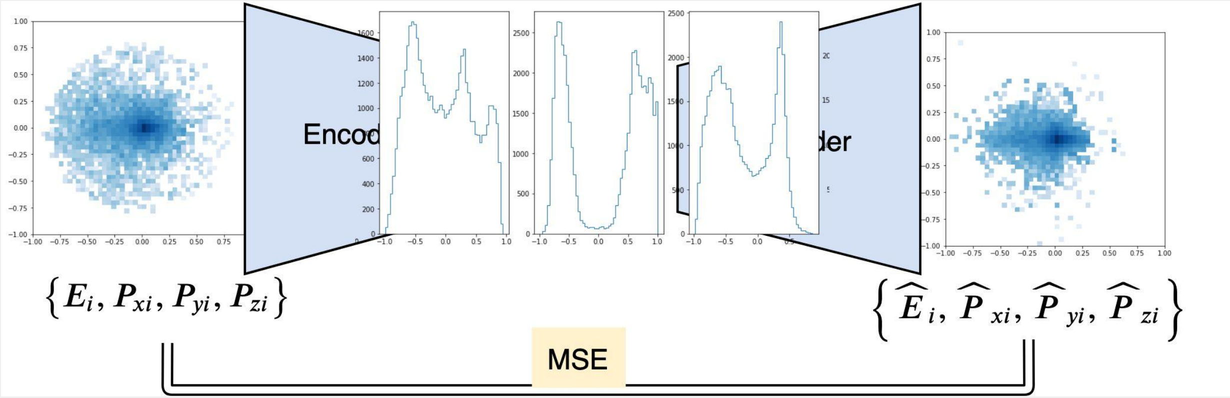


- T. Heimel , G. Kasieczka , T. Plehn , and J. M Thompson. QCD or What? arXiv:1808.08979.
- M. Farina, Y. Nakai, D. Shih. Searching for New Physics with Deep Autoencoders. arXiv: 1808.08992
- Tuhin S. Roya and Aravind H. Vijayb. A robust anomaly finder based on autoencoder. arXiv: 1903.02032

Review: Generative Model for Anomaly Detection

- Explicit (e.g. VAE) or implicit (e.g. GAN) estimation of log p(x)
- Building Blocks
  - Model
  - Embedding Architecture
  - Anomaly Metric

Autoencoder for Anomalous Jet Tagging



- T. Heimel , G. Kasieczka , T. Plehn , and J. M Thompson. QCD or What? arXiv:1808.08979.
- M. Farina, Y. Nakai, D. Shih. Searching for New Physics with Deep Autoencoders. arXiv: 1808.08992
- Tuhin S. Roya and Aravind H. Vijayb. A robust anomaly finder based on autoencoder. arXiv: 1903.02032

From Autoencoder to Variational Autoencoder

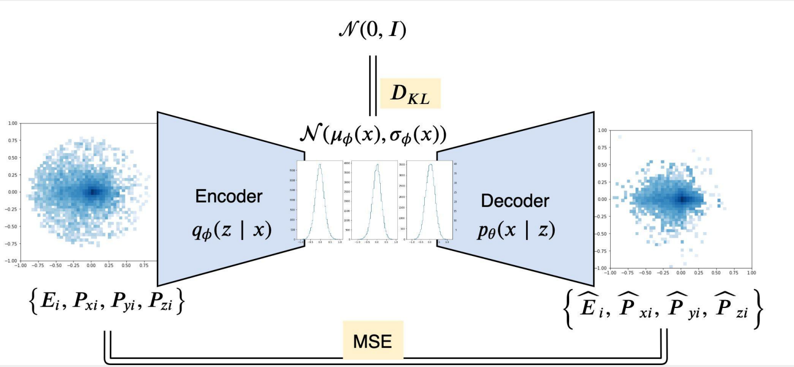
- VAE:
  - enforce a prior distribution in the Latent space through a D_KL (Kullback-Leibler Divergence) term

$$D(q(x)\|p(x)) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

(regularization term to autoencoder)

  - Likelihood estimation:
    logp(x) > $-L_{VAE}$: Evidence Lower Bound

$$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \qquad p(z) = \mathcal{N}(0, I)$$



$$L = \frac{1}{4n} \sum_i \|\hat{x}_i - x_i\|_2^2 + \beta \, D_{KL}(q(z|x)\|p(z))$$

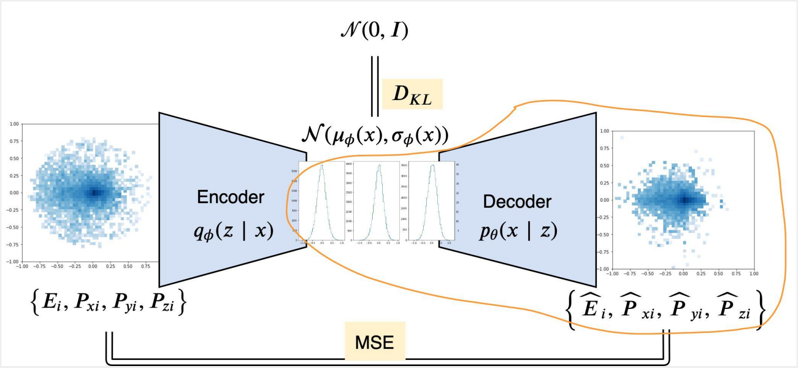
From Autoencoder to Variational Autoencoder

- VAE:
  - enforce a prior distribution in the Latent space through a D_KL (Kullback-Leibler Divergence) term

    $$D(q(x)\|p(x)) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

    (regularization term to autoencoder)
  - Likelihood estimation: logp(x) > $-L_{VAE}$: Evidence Lower Bound
  - Generative model: Sample from gaussians → generate new jets
  - Anomaly Detection: distance in input space (Mean Squared Error (MSE)); latent space (KL divergence)

$$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \qquad p(z) = \mathcal{N}(0, I)$$



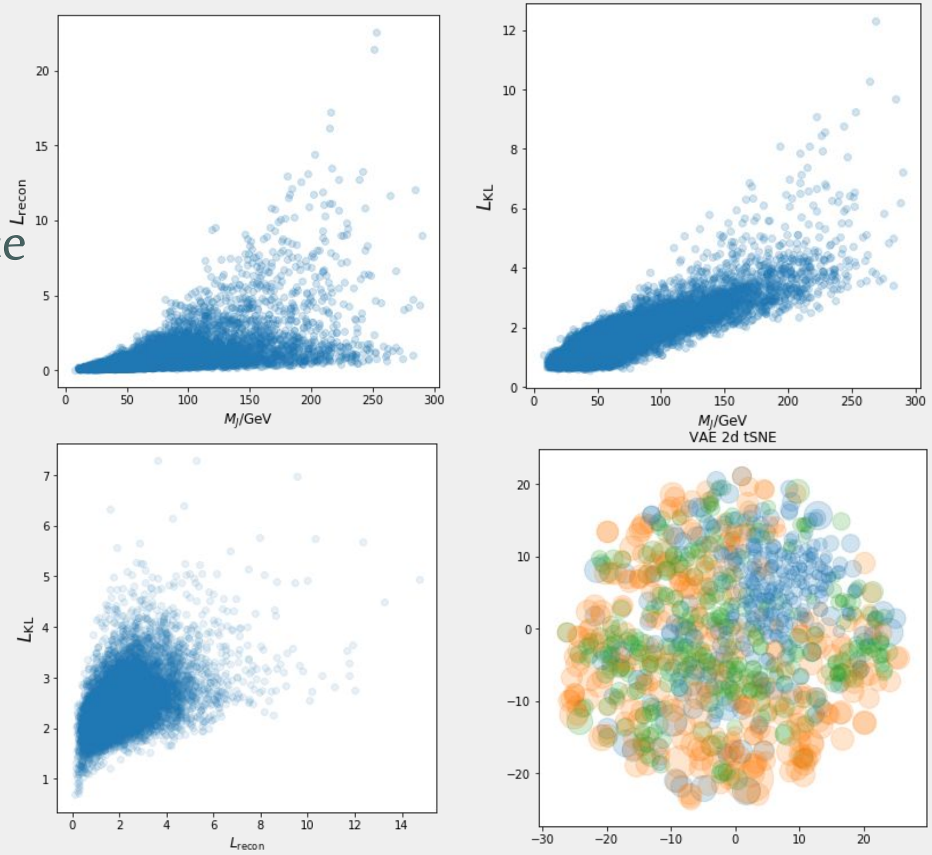
$$L = \frac{1}{4n} \sum_i \|\hat{x}_i - x_i\|_2^2 + \beta\, D_{KL}(q(z|x)\|p(z))$$

Settings

- Simple FCN/LSTM architecture
- Taking the first 20 pt-ordered jet constituents (zero-padded)
  - Inputs: four vectors (E, Px, Py, Pz) of jet constituents (particle flow objects)
  - Preprocessing: Boost to jet rest frame, Centering, Rotating → Principal axis alignment
- Train on 600,000 QCD jets (of which 20% serve as validation set)
  - QCD dijet production: pp → jj
  - ATLAS fatjet trigger: R = 1.0 antikt jets, pT > 450 GeV
  - No trimming applied
- VAE
  - $d_{hidden} = 10$
  - β = 0.1, 0.5, 1, 5

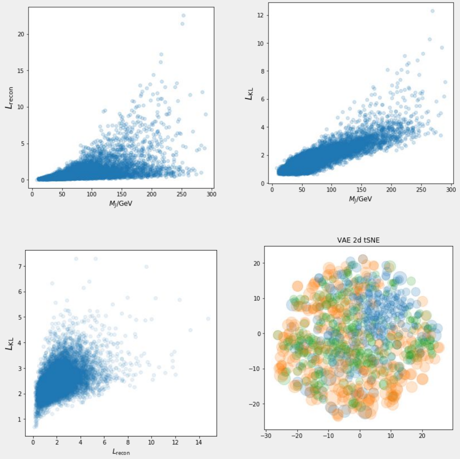
KL Regularization Strength -- beta = 0.1

- MSE and KL both has jet mass correlation
- Very strong mass correlation in latent space

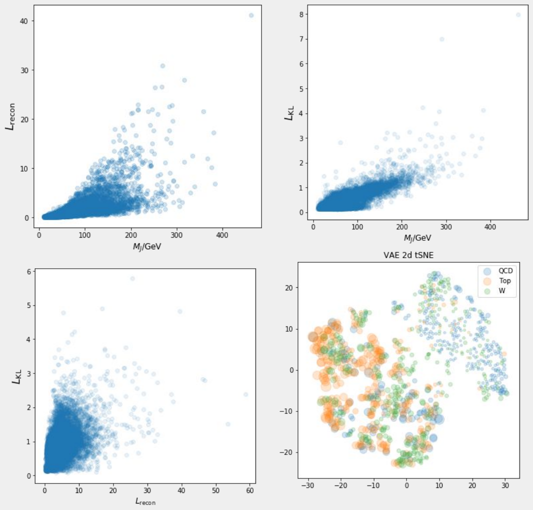


KL Regularization Strength

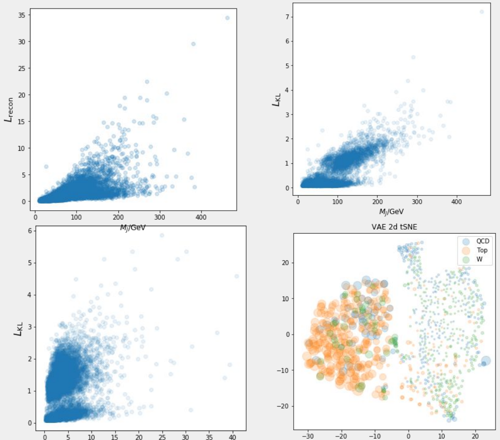
beta=0.1



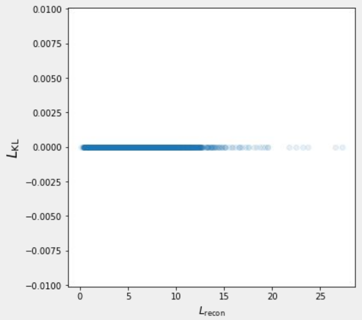
beta=0.5



beta=1



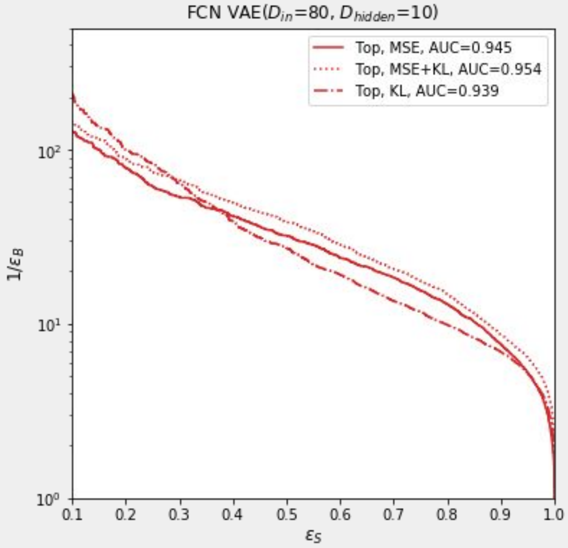
beta>~1



- As beta increases, reconstruction performance decreases. Latents develop different modes (mass modes).
- KL vanishing for beta>~1

Jet Tagging Performance

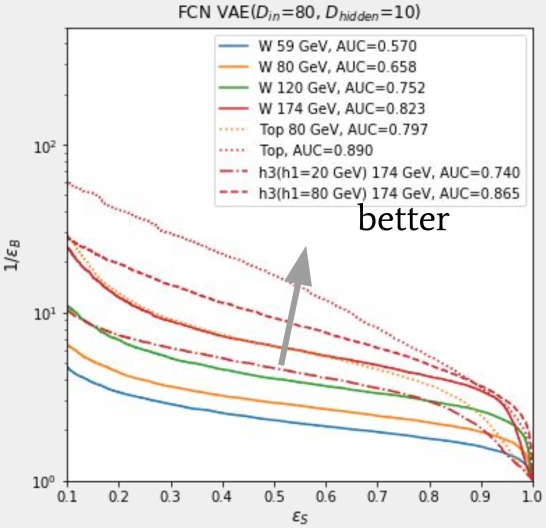
- Datasets:
  - Background: QCD, R=0.1, pT>450 GeV
  - Testing on pT :[550, 650] GeV
    - 2-prong: W (W’ → W Z)
      - M: 59 GeV, 80 GeV, 120 GeV
    - 3-prong: Top (Z’ → t t~)
      - M: 80 GeV, 174 GeV
    - H(->hh->4j), MH=174 GeV
      - Mh = 20 GeV, 80 GeV
- Anomaly Metric
- Examine:
  - Jet mass effects
  - pT effects (training on full pt, test on fixed pt)
  - Jet type (focus on prong-ness)



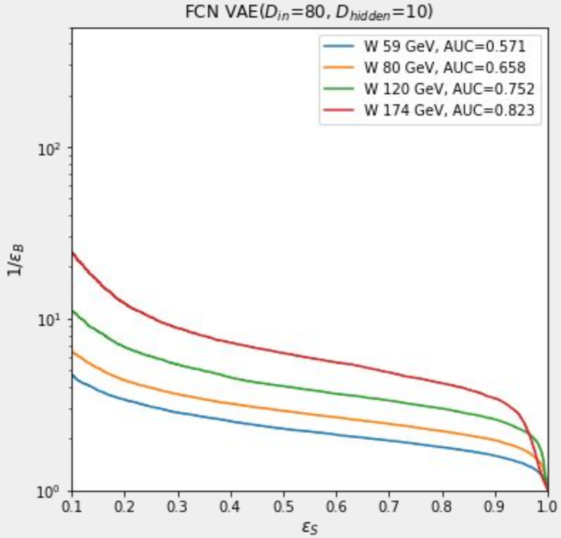
Test results for Top Tagging Reference Data (* T. Heimel , G. Kasieczka , T. Plehn , and J. M Thompson. QCD or What? arXiv:1808.08979.)

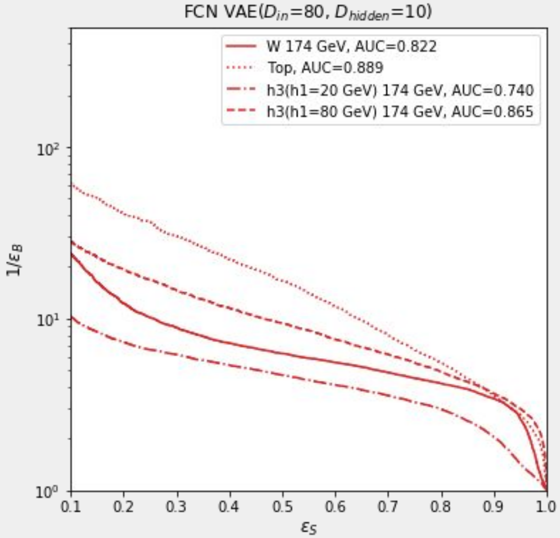
Jet Tagging Performance

- Jet mass dependence: discriminative power decreases when jet mass decreases (mass correlation) ---> works well for Top, but low significance for W jets
- Jet complexity dependence



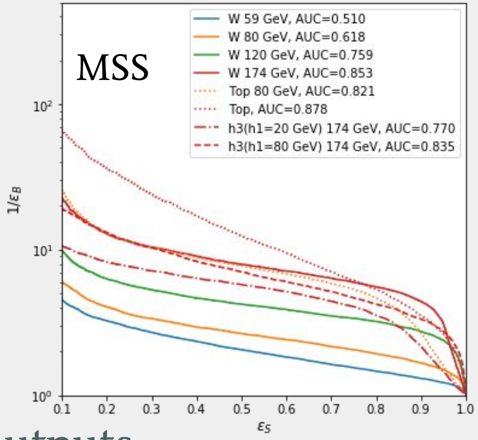
beta=0.1

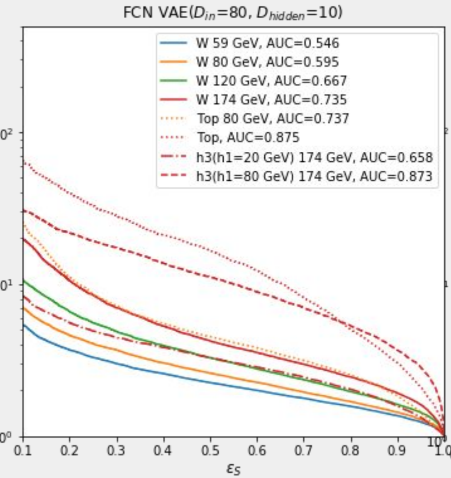




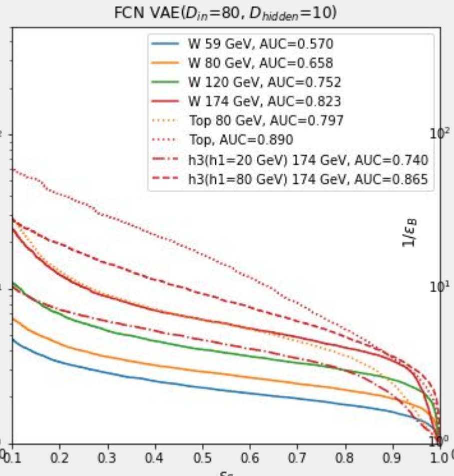
Jet Tagging Performance

- Anomaly Metric
  - Reconstruction error: MSE
  - Negative Log Likelihood: MSE+KL
  - Latent space: KL divergence
  - EMD(Energy Mover Distance) between inputs and outputs
  - MSS: l2 norm the input feature vector (chi2)

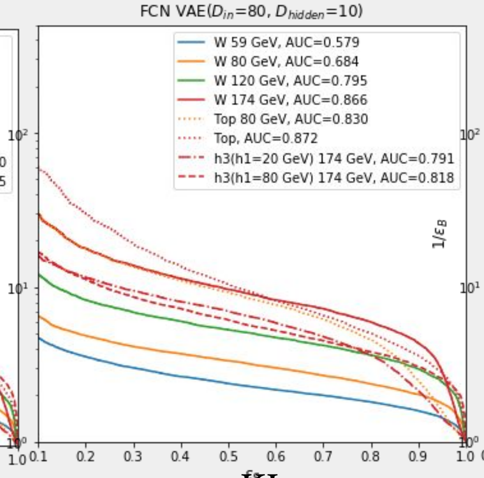




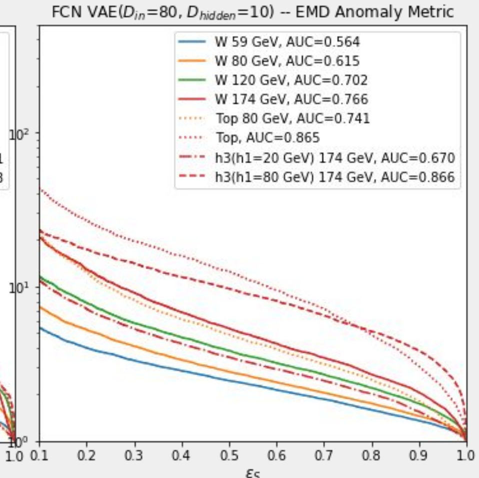
MSE



MSE+KL



KL



EMD

Anomaly Detection can Fail

- Outliers can be assigned higher probability sometimes, this happens in a general scope of anomaly detection using generative models
- Quick example: MSE based anomaly metric has intrinsic mass dependence → naive VAE assigns higher probability to lower mass jets

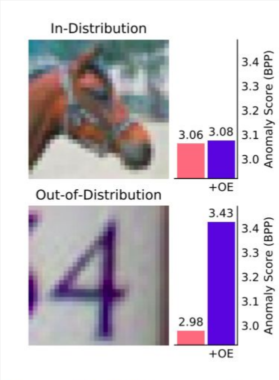
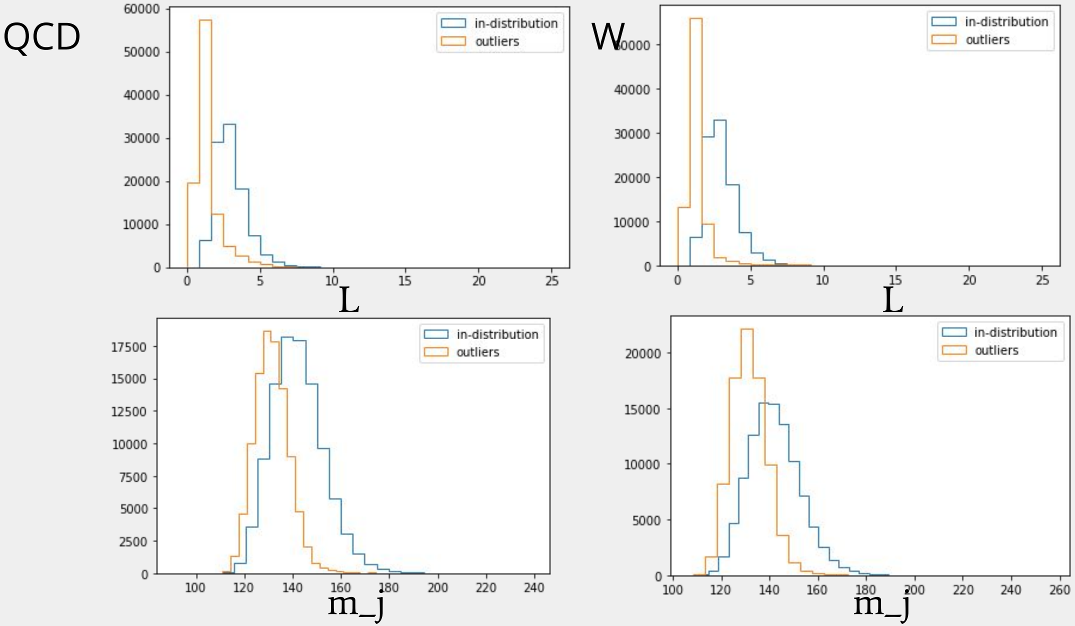


Figure 2: OOD scores from PixelCNN++ on images from CIFAR-10 and SVHN.



- D. Hendrycks, M. Mazeika, T. Dietterich. Deep Anomaly Detection with Outlier Exposure. arXiv: 1812.04606

Outlier Exposure (OE)

- Is the VAE learning useful enough representations?
  - Restricted by the format of loss function
  - Need extra information to guide directions for better anomaly detection
- Semi-supervised Learning: encourage specific directions in the loss landscape
  - Relative weight lambda controls the OE strength $L = L_{VAE} - \lambda L_{OE}$
  - Restricting reconstruction error strength between Out-of-distribution (OoD) and In-distribution (InD) samples

$$L_{OE} = sigmoid(MSE(OoD) - MSE(InD))$$

  - Restricting KL divergence in latent space

$$L_{OE} = min\{0, KL(OoD) - KL(InD) - margin\}$$

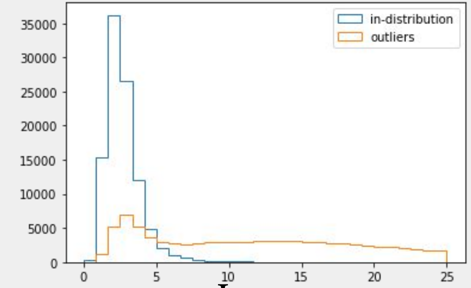
- Training Scenarios:
  - Fine-tuning using outlier exposure
  - Train with outlier exposure from scratch (results shown for this scenario)

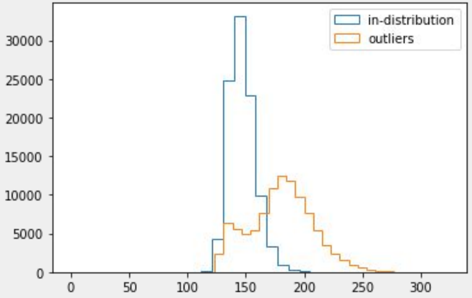
Quick Test -- Top Tagging Reference Data -- OE(QCD) Training on Top

- Quick test on training VAE with top jets, using QCD as outlier exposure samples → test on QCD and W jets

AUC = 0.920

QCD



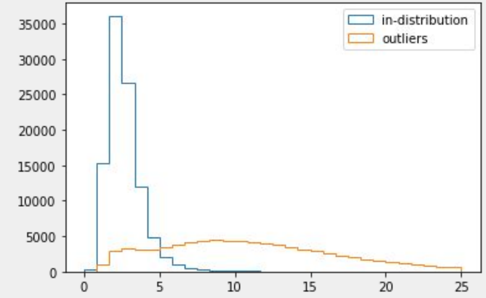


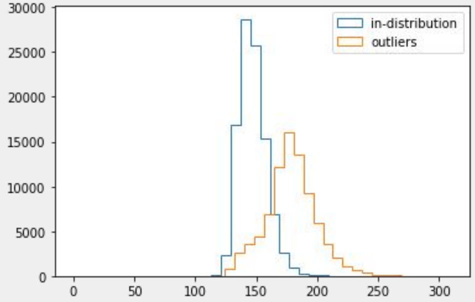
L

m_j

AUC = 0.939

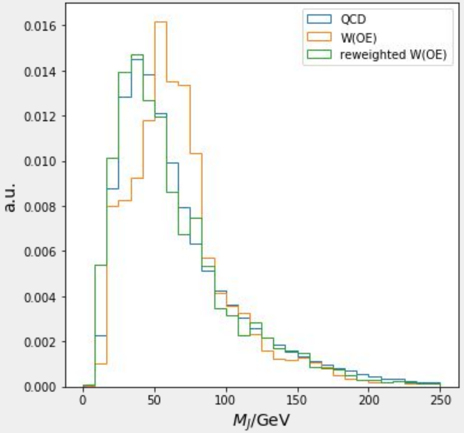
W





Outlier Exposure -- Results

- Outlier samples: W (mass rescaled) jets with mass distribution reweighted to match QCD jets (→ mass decorrelation)
- Annealing training of OE weight lambda (cyclically annealing, from 0 - 2)
- Results:
  - Test on different jet type and jet mass
  - Mass decorrelation

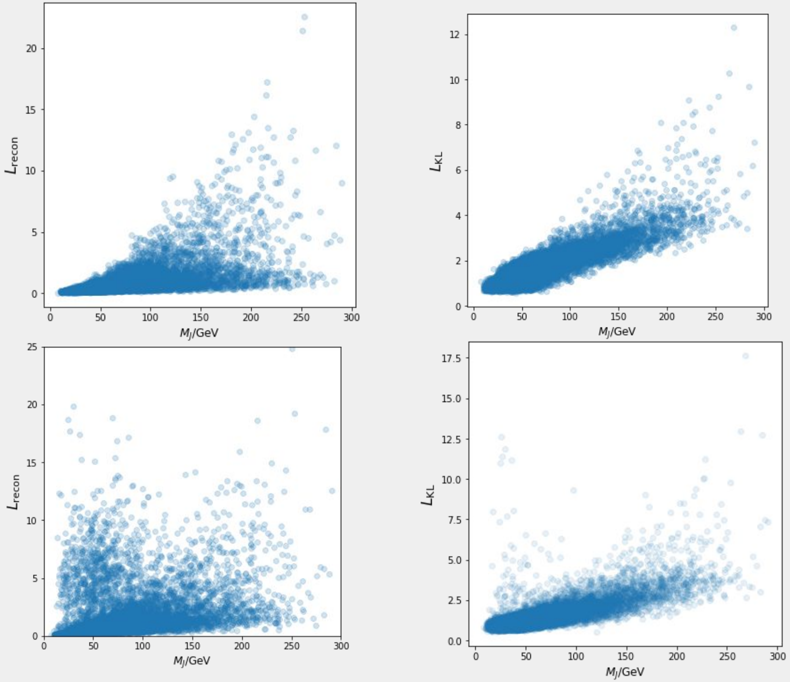


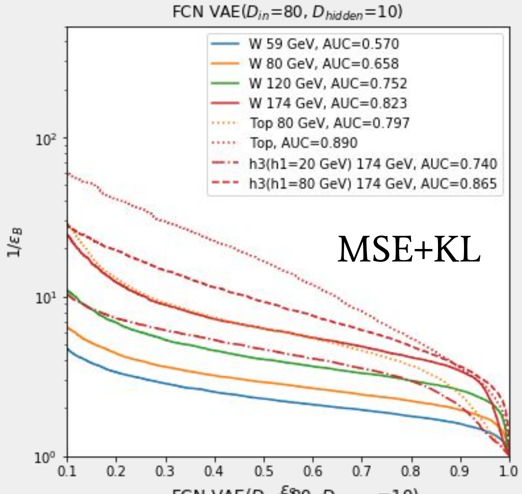
Outlier Exposure -- Results -- MSE-OE

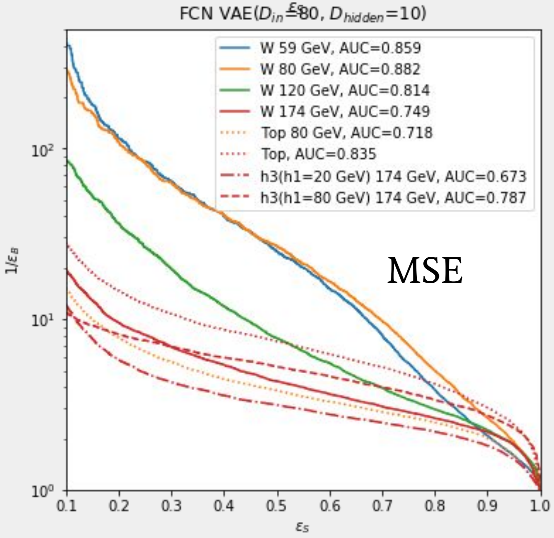
- Train using MSE outlier exposure loss term
- Anomaly metric: MSE

before

after

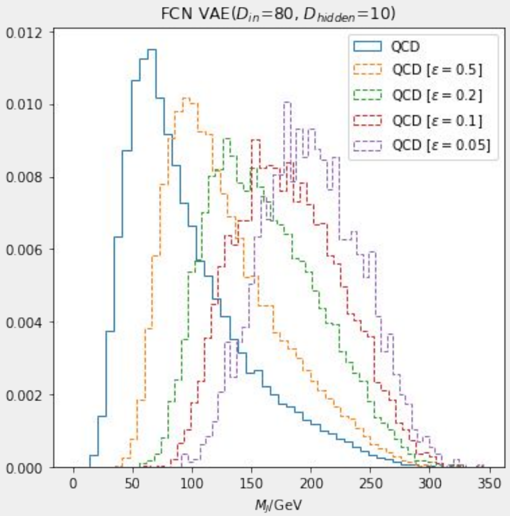






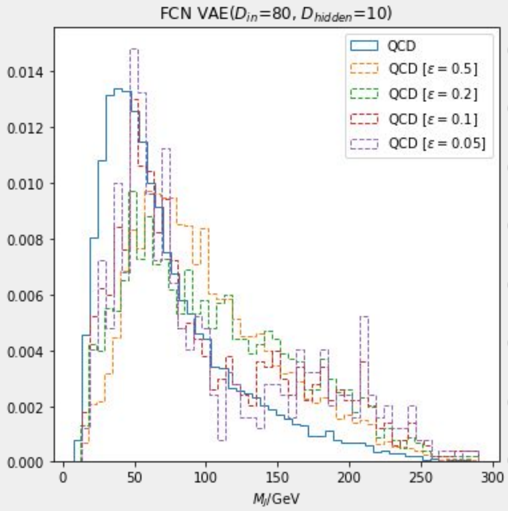
Outlier Exposure -- Results -- MSE-OE

- Train using MSE outlier exposure loss term
- Anomaly metric: MSE+KL (since KL and MSE are correlated; MSE+KL better in MSE-OE case)



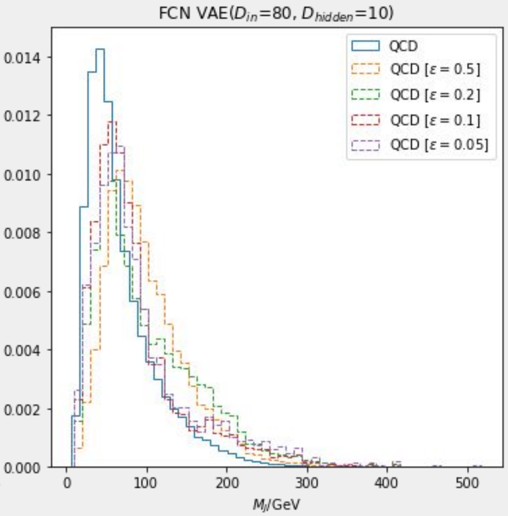
before

reweighted



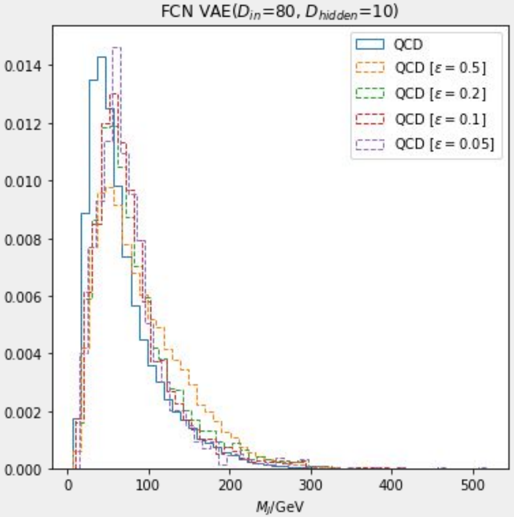
after

Unreweighted

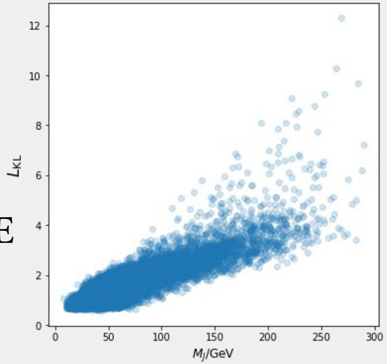


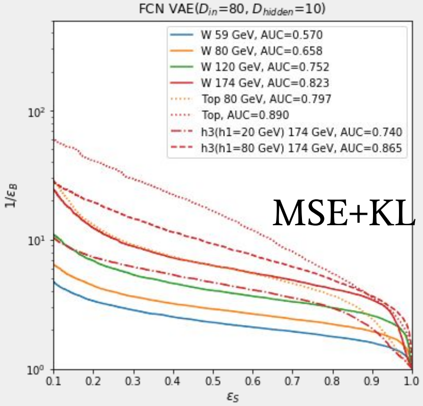
Outlier Exposure -- Results -- KL-OE

- Training using latent KL OE loss
- Reconstruction performance unchanged
- Using latent KL divergence directly as anomaly metric → very good mass decorrelation

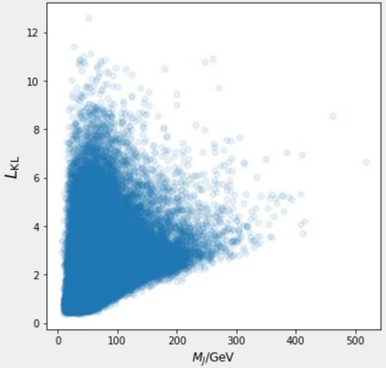


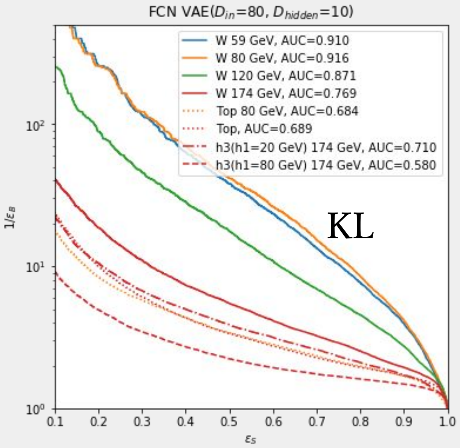
Without OE





With OE





*results shown here are trained using OE unreweighted samples

Summary

- Explored Generative Model (VAE) for anomaly detection
- KL Regularization
- Anomalous Jet Tagging
    - Different jet masses
    - Different jet types
    - Anomaly metrics
- Outlier exposure to increase sensitivity to out-of-distribution samples
    - Especially in latent space
- Mass correlation affected by outlier exposure ← mass sculpting

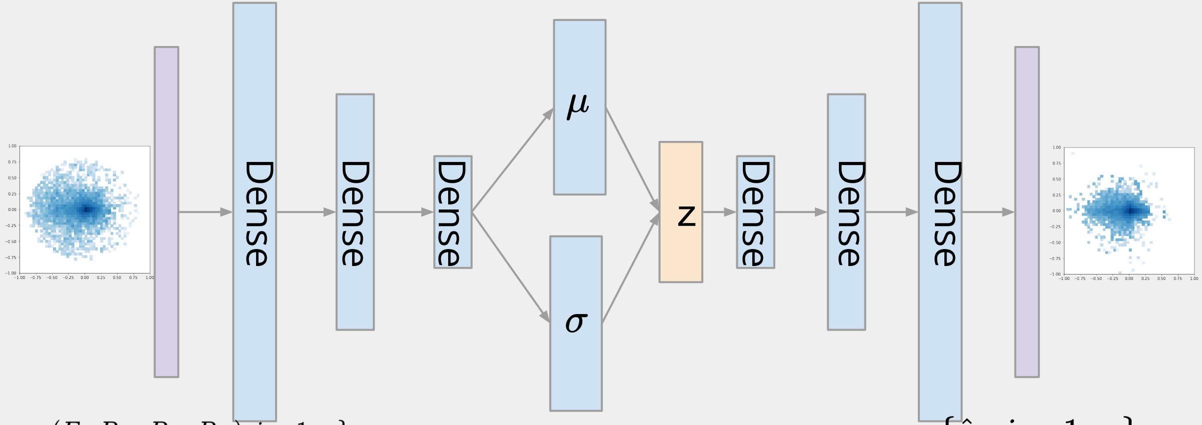
Outlook

- Architecture and input representation
- Reconstruction loss

Backup

VAE Architecture -- FCN

- Latent dimension = 10 (best: 6 - 20)



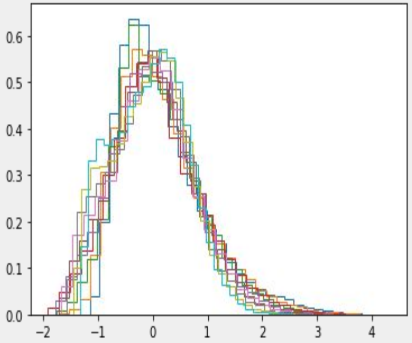
$\{x_i = (E_i, P_{x\,i}, P_{y\,i}, P_{z\,i}), i = 1..n\}$

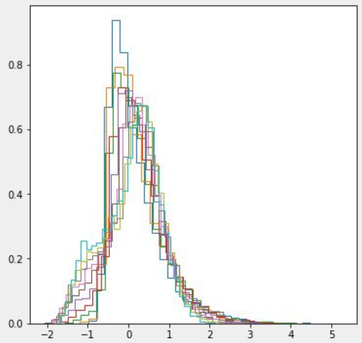
$\{\hat{x}_i, i = 1..n\}$

$$L = \frac{1}{4n} \sum_i \|\hat{x}_i - x_i\|_2^2 + \beta\, D_{KL}(q(z|x) \| p(z))$$

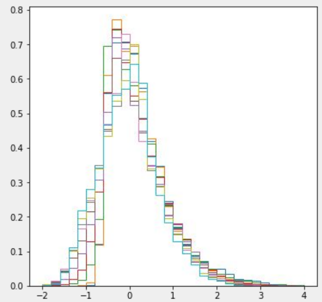
Reconstruction Performance

- Input features





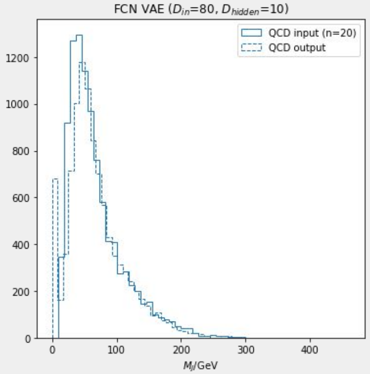
generation

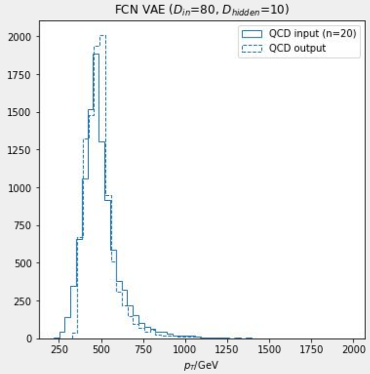


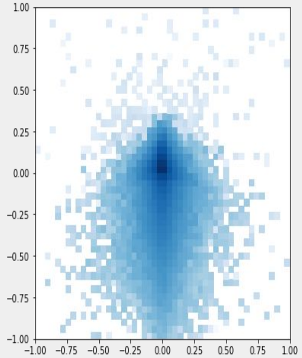
Sampling from latent priors → decoder

- High Level features

beta=0.1

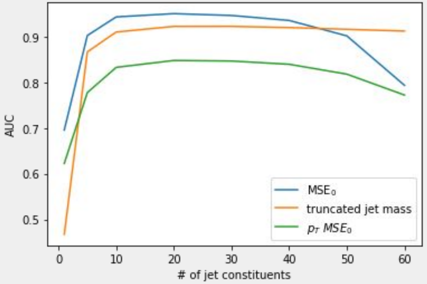




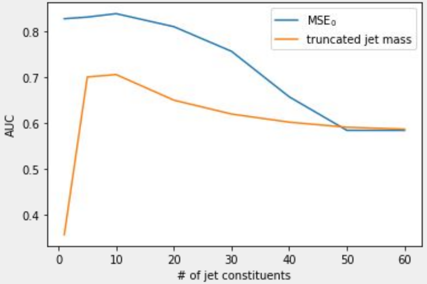


Investigation on MSE Anomaly Metric

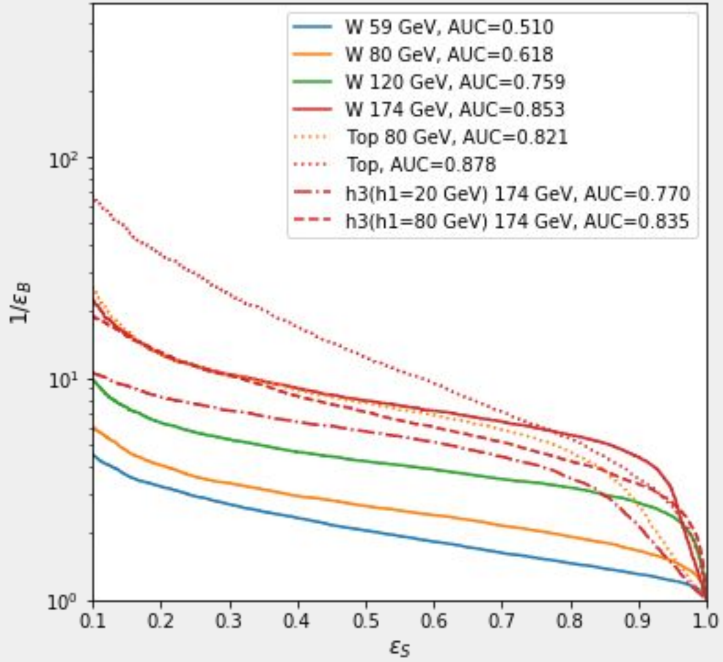
- Pure non-ML metric: $L = \sum_i \|x\|^2 \sim \chi^2 \quad M_J = \sum_i z_i \theta_i^2$
- MSE-based anomaly metric doesn't require perfect reconstruction



Top Tagging Reference

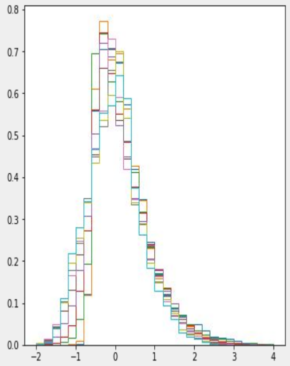


W test jets

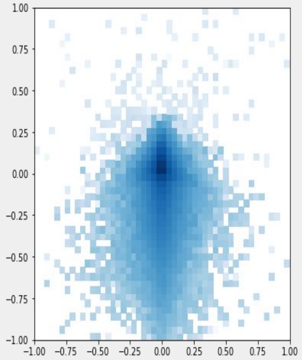


Generator

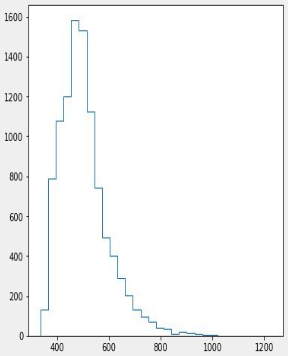
- Sample from prior latent distribution $\mathcal{N}(0, I)$
- Specific dimensions more correlated with mass



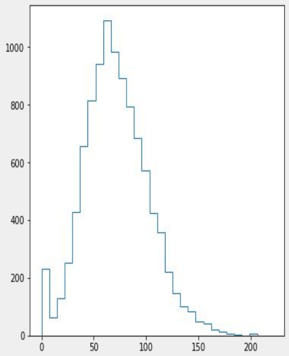
Input space



Jet Images



Jet pt



Jet mass