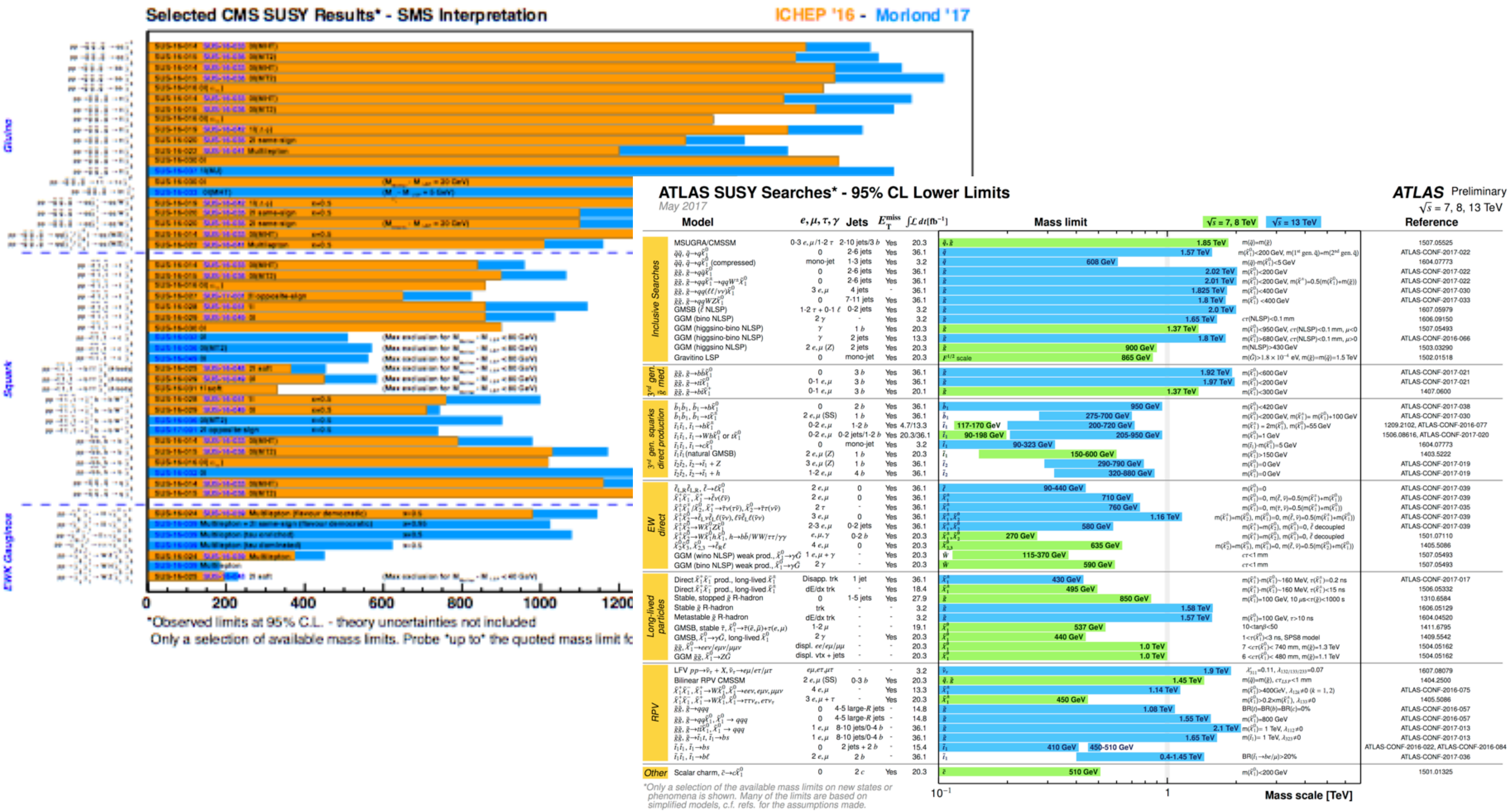
Searching for New Physics with Deep Autoencoders

David Shih
NHETC, Rutgers University

IML Working Group Seminar
Oct 30, 2018

Based on Farina, Nakai & DS 1808.08992

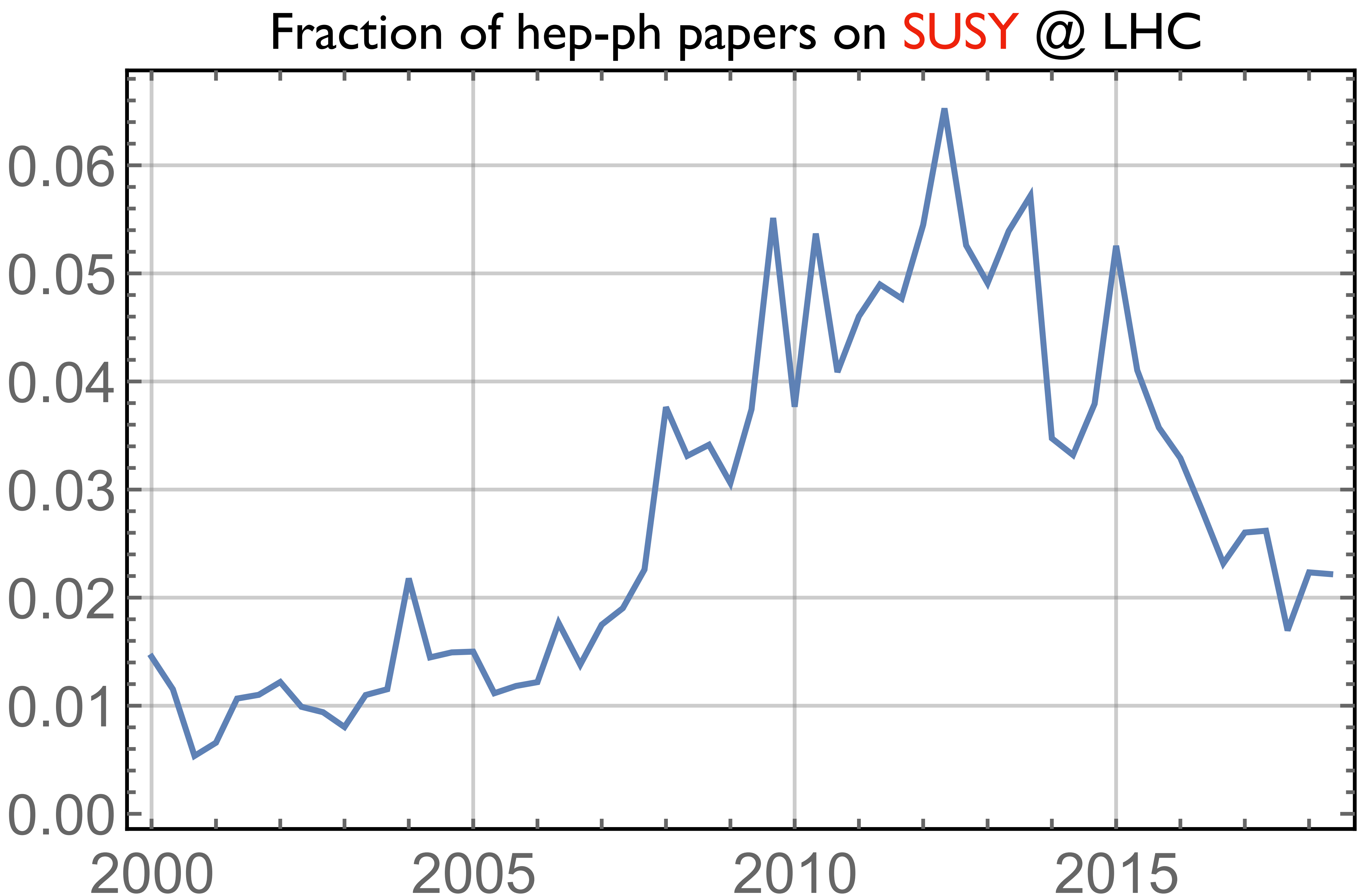


The LHC has searched for new physics in many places.

So far, there has been no evidence of anything.

Many well-motivated models (SUSY, composite higgs, dark matter, …) have not turned up as expected.

People are losing interest in “well-motivated models”…



We need new ideas!

Can we search for new physics in the data without knowing what we're looking for?

Can we find the unexpected?

Can we find a needle in a haystack, without knowing what needles are?

Sounds hopelessly difficult…

Maybe deep learning can help!

Deep learning at LHC

Recently there has been a lot of interest in applications of deep learning to the LHC.

- classification (eg quark/gluon tagging, boosted resonance tagging)
- pile-up removal
- event generation
- triggering
- anomaly detection
- …..

Deep learning at LHC

Recently there has been a lot of interest in applications of deep learning to the LHC.

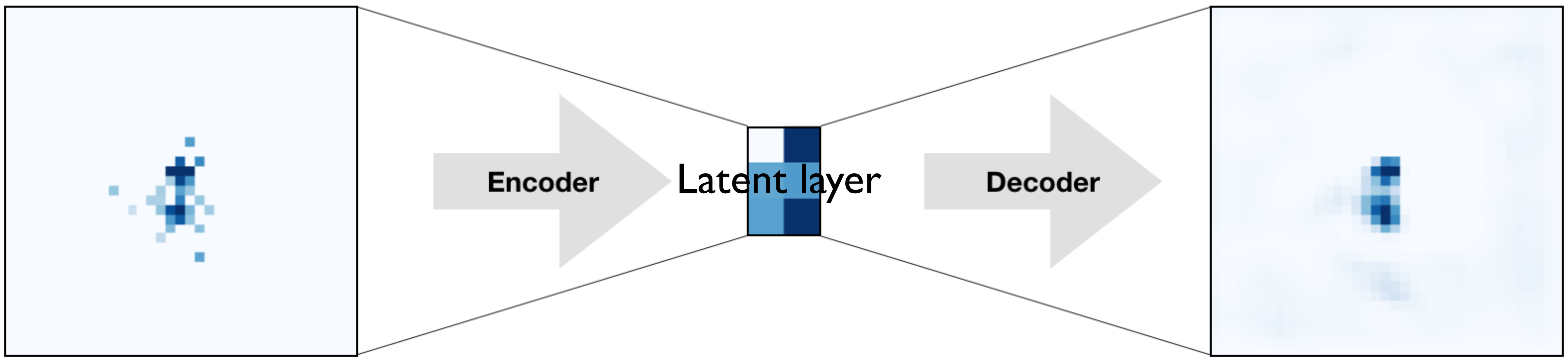
- classification (eg quark/gluon tagging, boosted resonance tagging)
- pile-up removal
- event generation
- triggering
- anomaly detection
- …..

This talk

A promising idea:

Deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992



An autoencoder maps an input into a "latent representation" and then attempts to reconstruct the original input.

The encoding is lossy ("information bottleneck"), so the decoding cannot be perfect.

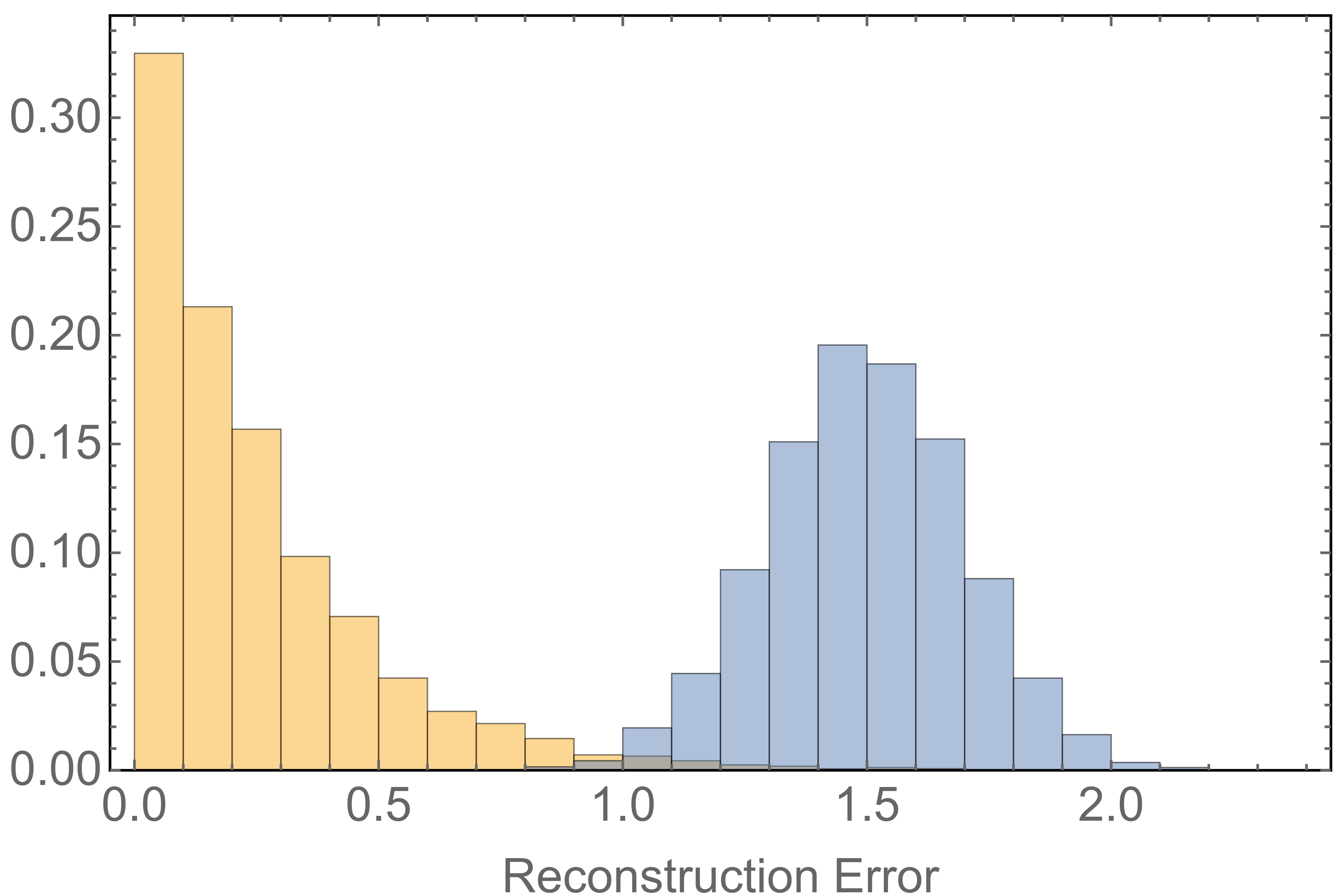
Some previous approaches:
Aguilar-Saavedra et al, "A generic anti-QCD jet tagger" 1709.01087
Collins et al, "CWoLa Hunting" 1805.02664
Hajer et al "Novelty Detection Meets Collider Physics" 1807.10261

Deep autoencoders

Heimel et al 1808.08979;   Farina, Nakai & DS 1808.08992

Quantify AE performance using reconstruction error:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

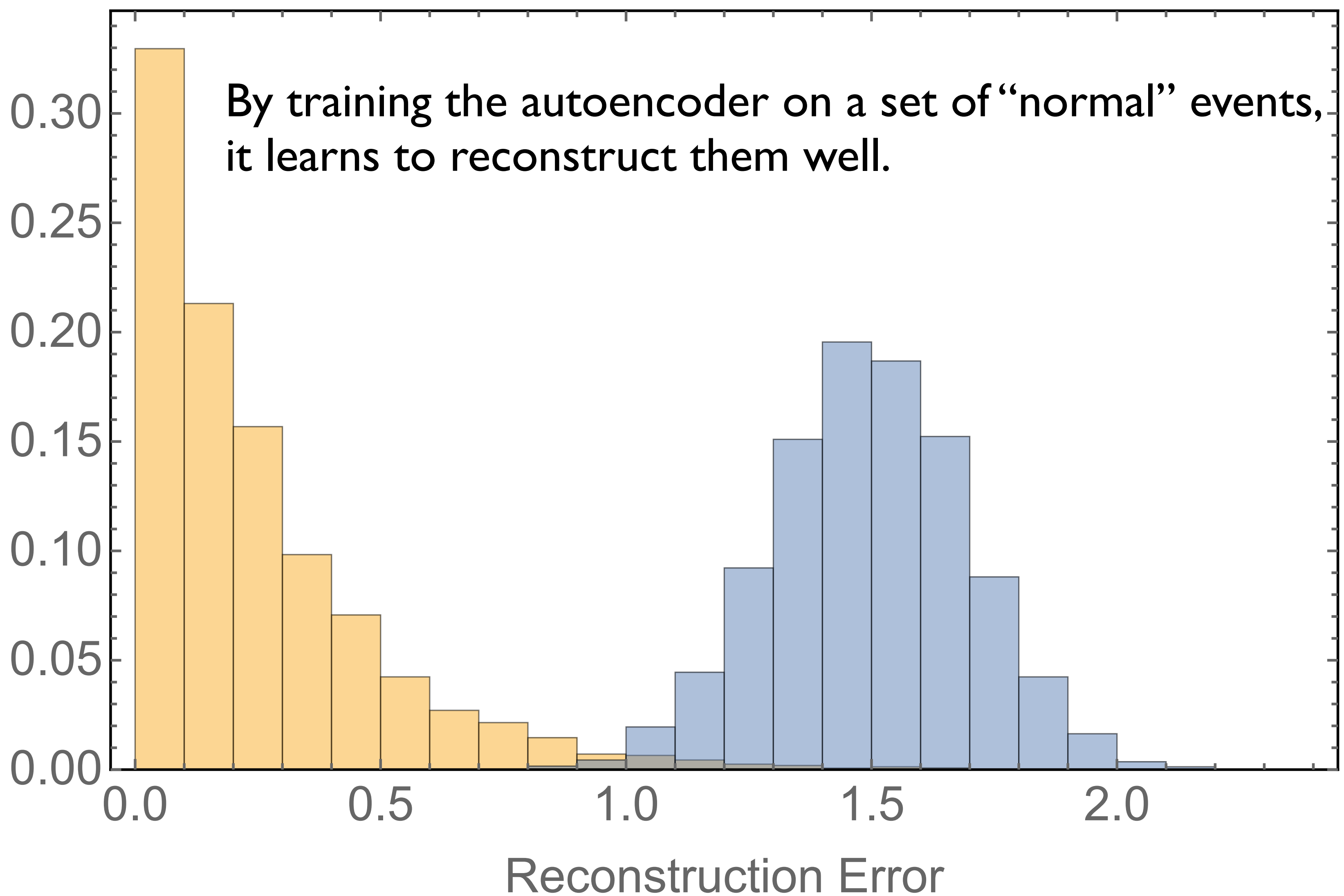


Deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992

Quantify AE performance using reconstruction error:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$

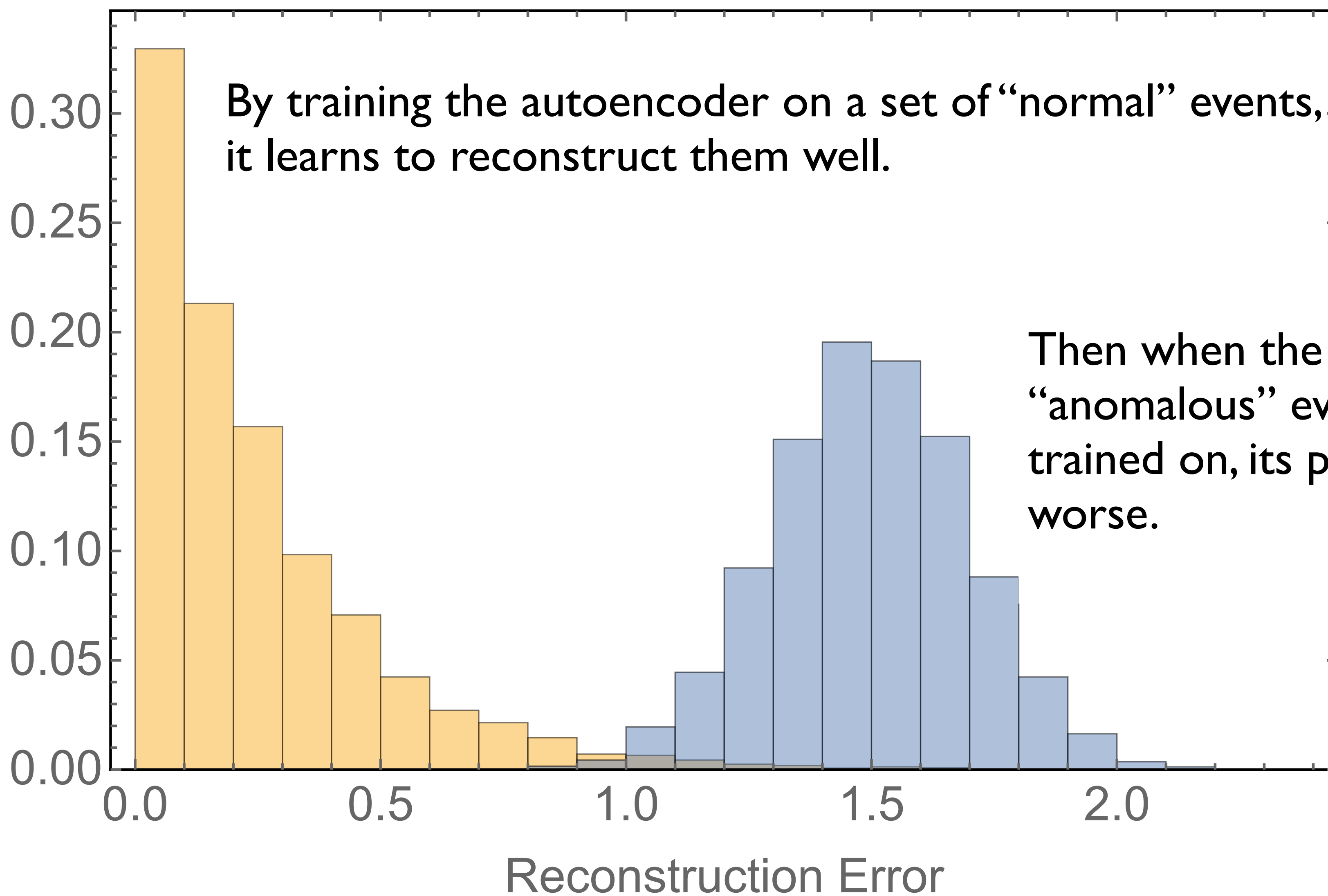


Deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992

Quantify AE performance using reconstruction error:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$



By training the autoencoder on a set of "normal" events, it learns to reconstruct them well.

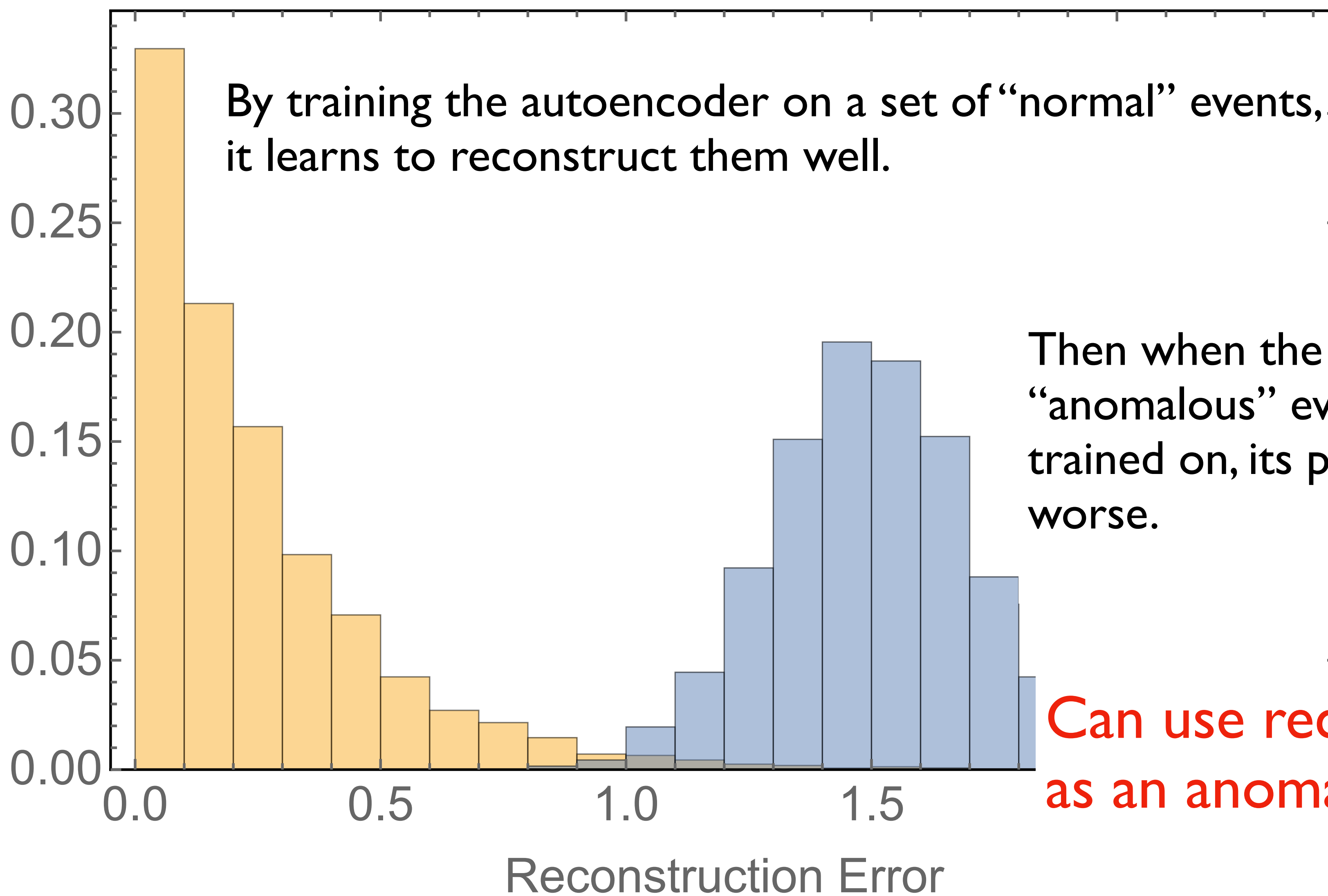
Then when the autoencoder encounters "anomalous" events that it was not trained on, its performance should be worse.

Deep autoencoders

Heimel et al 1808.08979; Farina, Nakai & DS 1808.08992

Quantify AE performance using reconstruction error:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i^{in} - x_i^{out})^2$$



By training the autoencoder on a set of "normal" events, it learns to reconstruct them well.

Then when the autoencoder encounters "anomalous" events that it was not trained on, its performance should be worse.

Can use reconstruction error as an anomaly threshold!

Autoencoder architectures

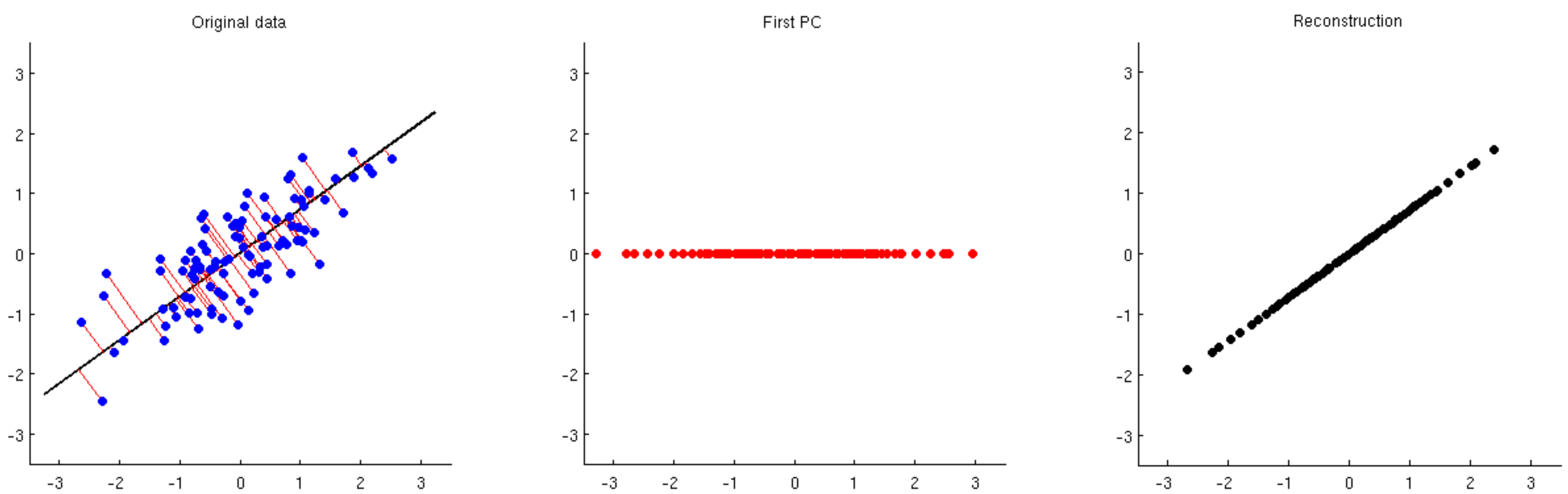
We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convolutional NN

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)



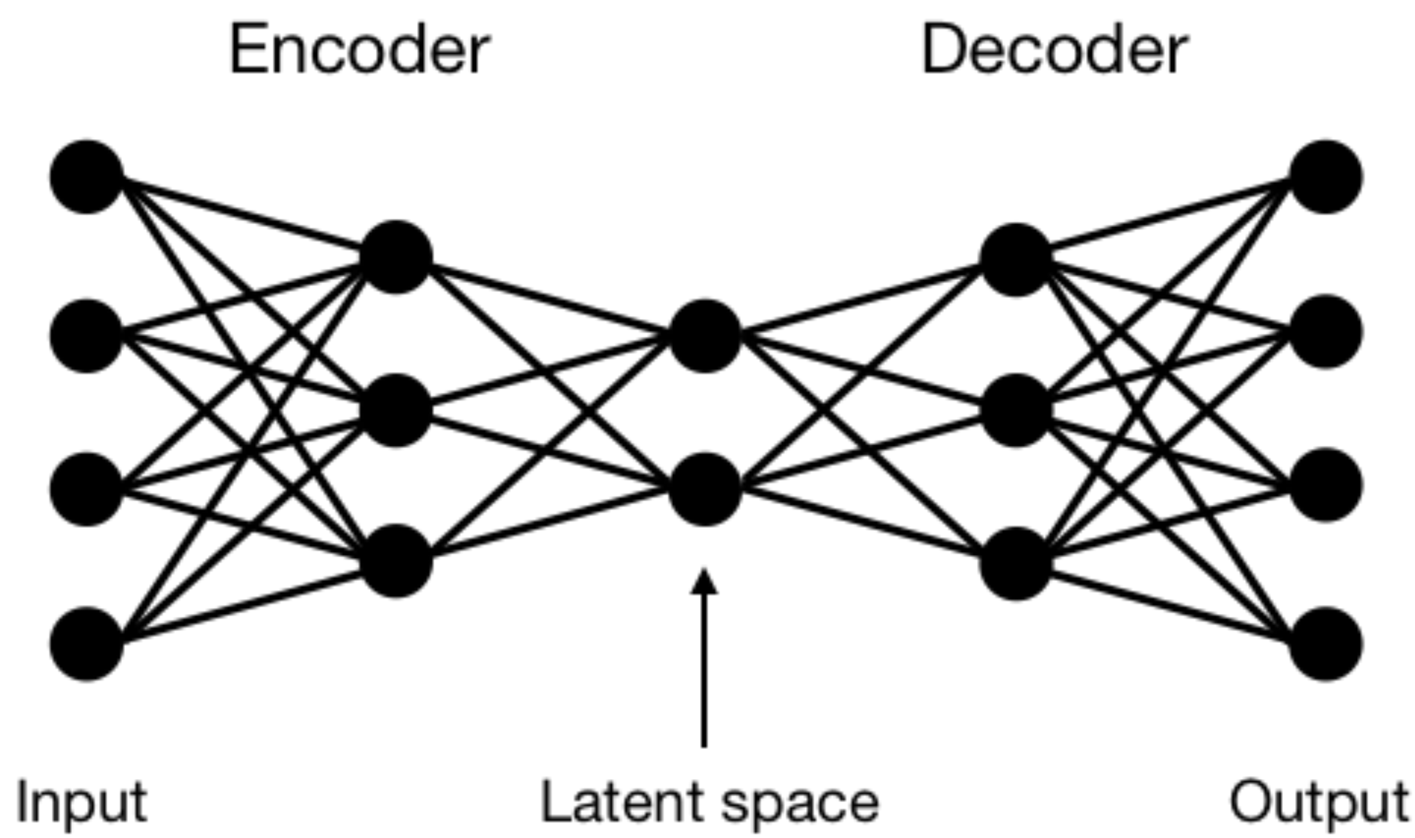
Project onto the first d PCA eigenvectors $z = \mathcal{P}_d x_{in}$

Inverse transform to reconstruct original input $x_{out} = \mathcal{P}_d^T z = \mathcal{P}_d^T \mathcal{P}_d x_{in}$

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convoluti



Flatten input into column vector

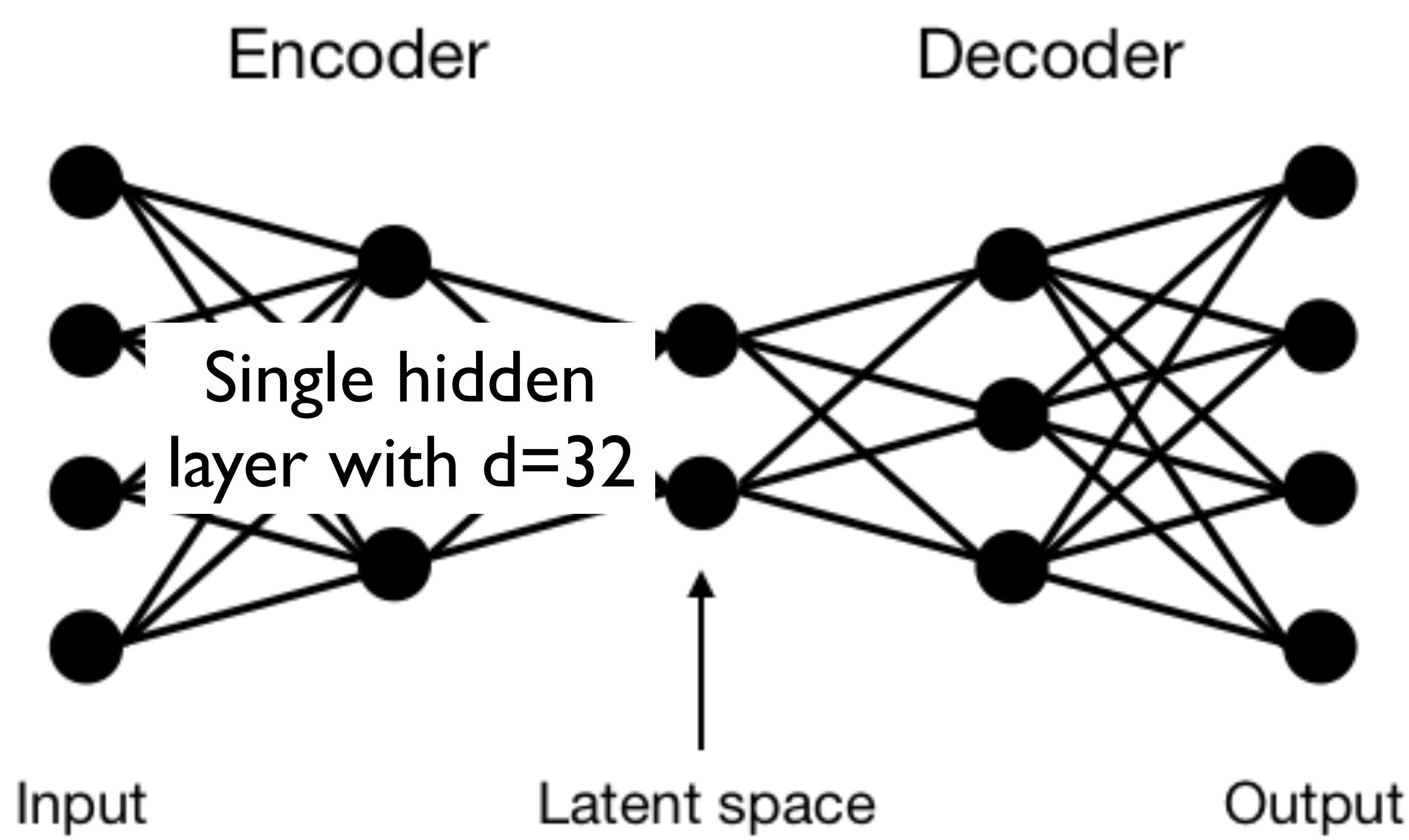
(dimension d)

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convoluti

Flatten input into column vector

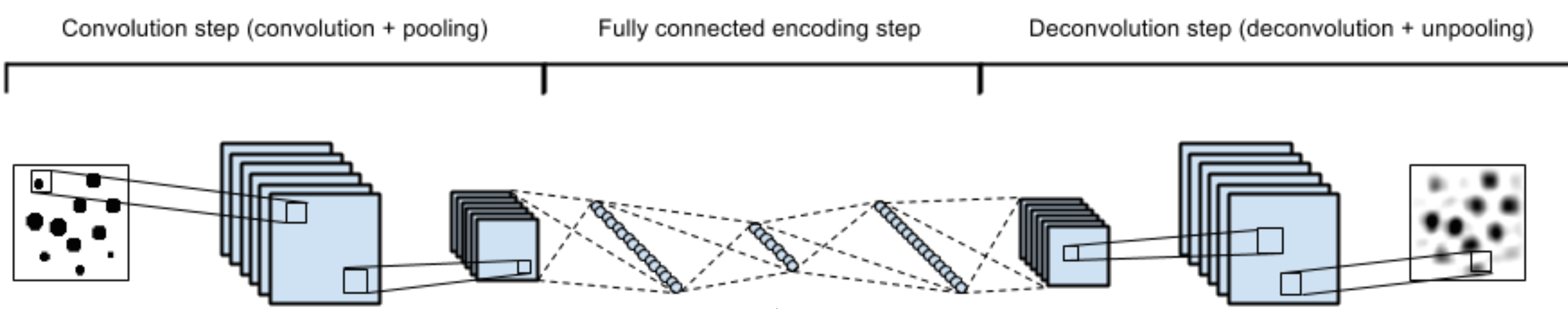


(dimension d)

Autoencoder architectures

We considered three autoencoder architectures (many more are possible):

- Principal Component Analysis (PCA)
- Dense NN
- Convolutional NN



M. Ke, C. Lin, Q. Huang (2017)

Encoder **Latent space** **Decoder**

128C3-MP2-128C3-MP2-128C3-32N-6N-32N-12800N-128C3-US2-128C3-US2-1C3

Jet Images

We focused on jet images as inputs to the autoencoders

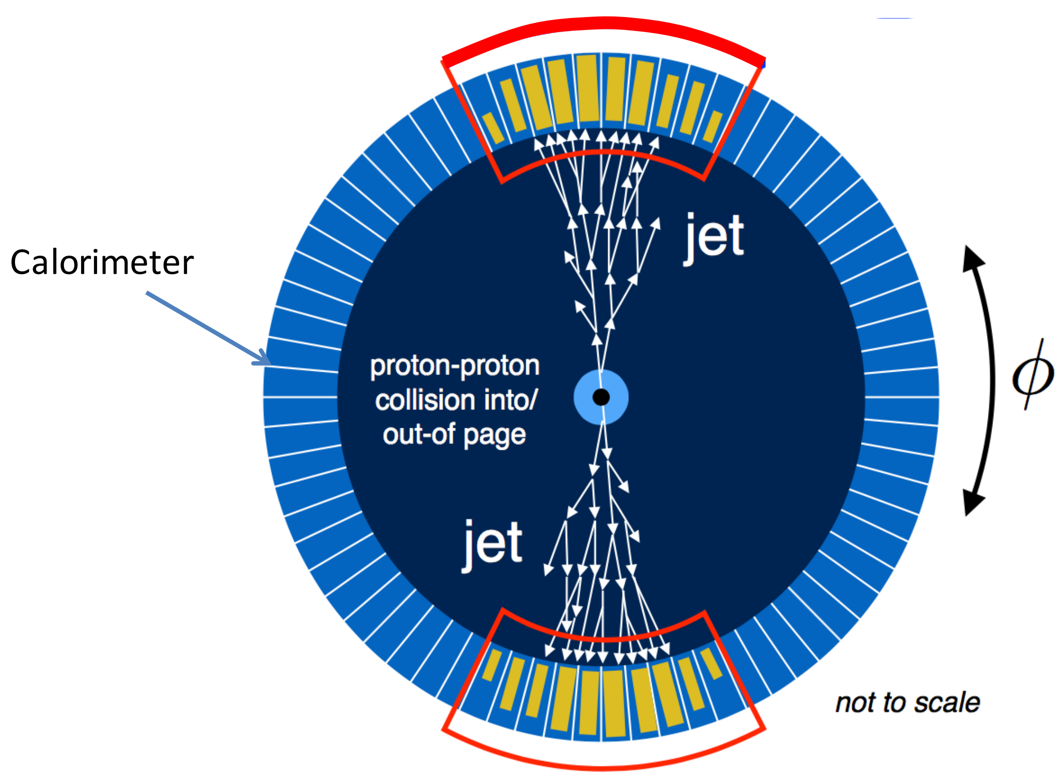
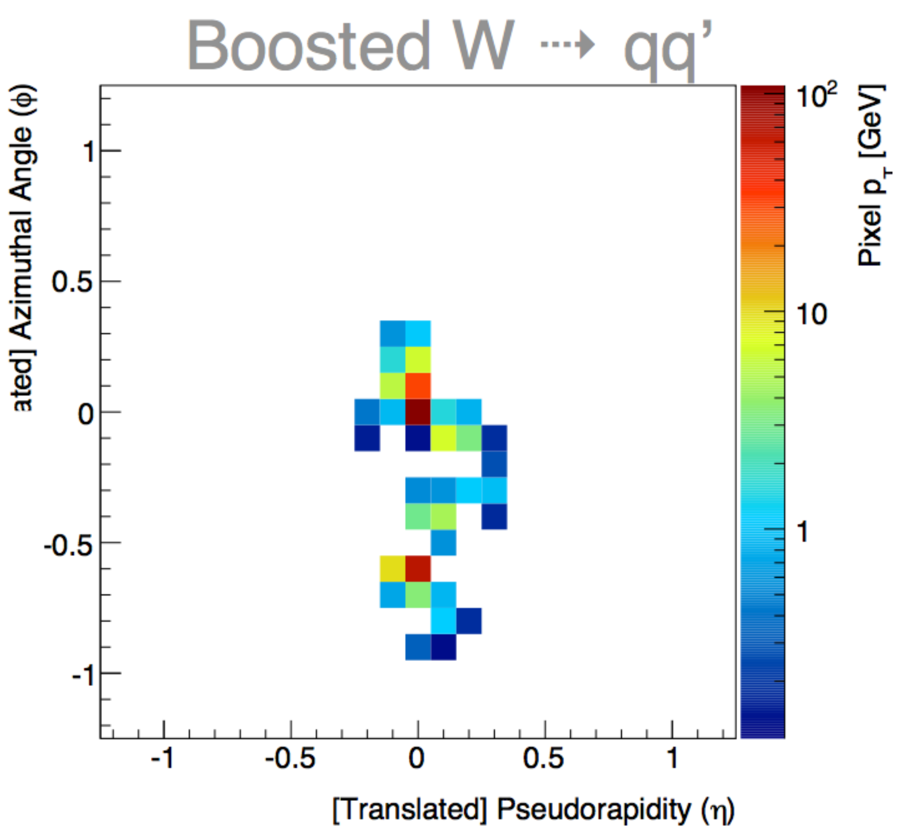


Figure credit:
B. Nachman



Can think of a jet as an image in eta and phi, with

- Pixelation provided by calorimeter towers
- Pixel intensity = pT recorded by each tower

Sample definitions

We took

- QCD jets as background
- tops and 400 GeV gluinos as signals.

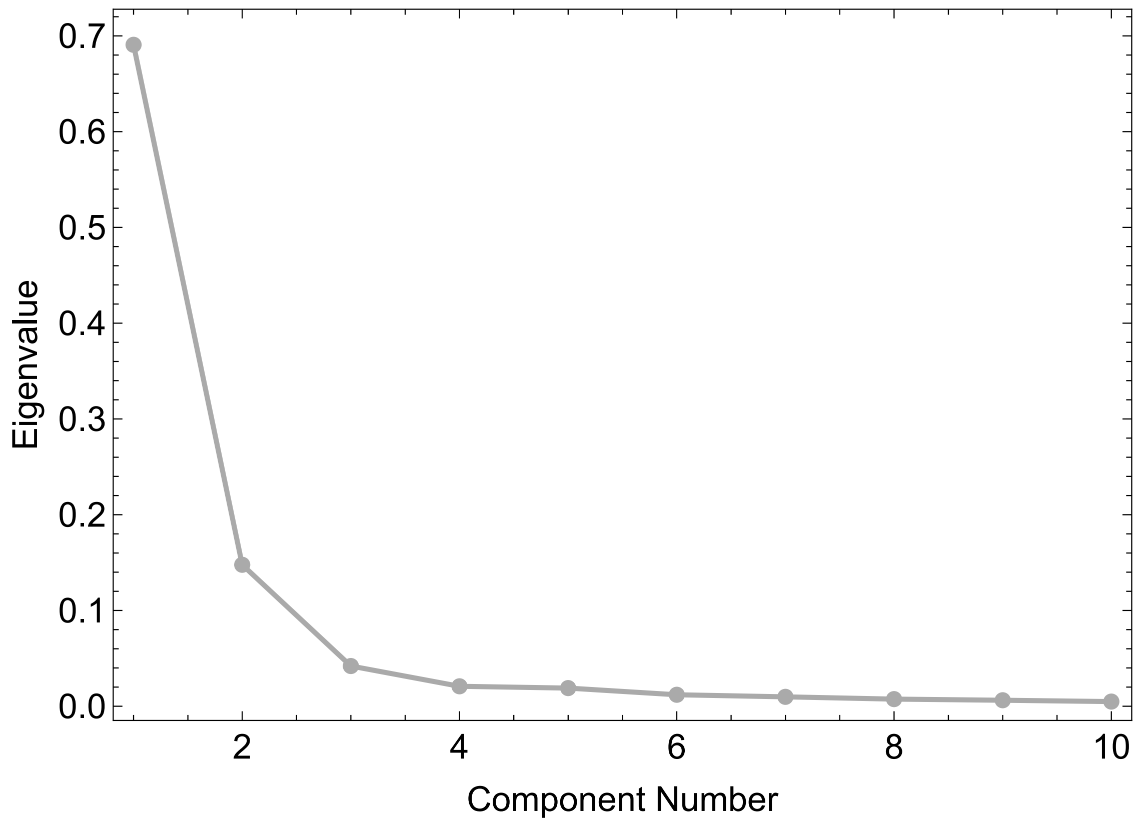
Same jet and image specifications as for recent top tagging study (Macaluso & DS 1803.00107)

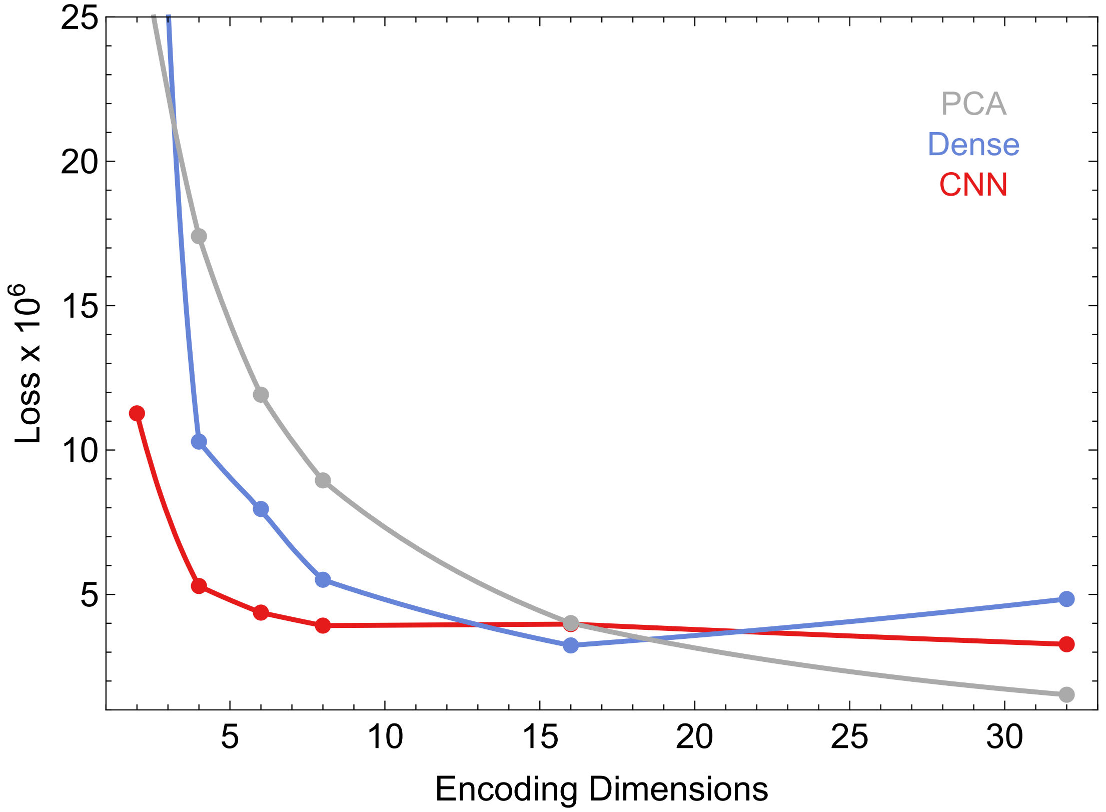
	CMS
Jet sample	13 TeV $p_T \in (800, 900)$ GeV, $\lvert\eta\rvert < 1$ Pythia 8 and Delphes particle-flow match: $\Delta R(t, j) < 0.6$ merge: $\Delta R(t, q) < 0.6$ 1.2M + 1.2M
Image	37×37 $\Delta\eta = \Delta\phi = 3.2$
Colors	$(p_T^{neutral},\ p_T^{track},\ N_{track},\ N_{muon})$

Choosing the latent dimension

d too large → autoencoder becomes identity transform
d too small → autoencoder cannot learn all the features

Should choose the latent dimension in an unsupervised manner
(ie without optimizing on a specific signal)

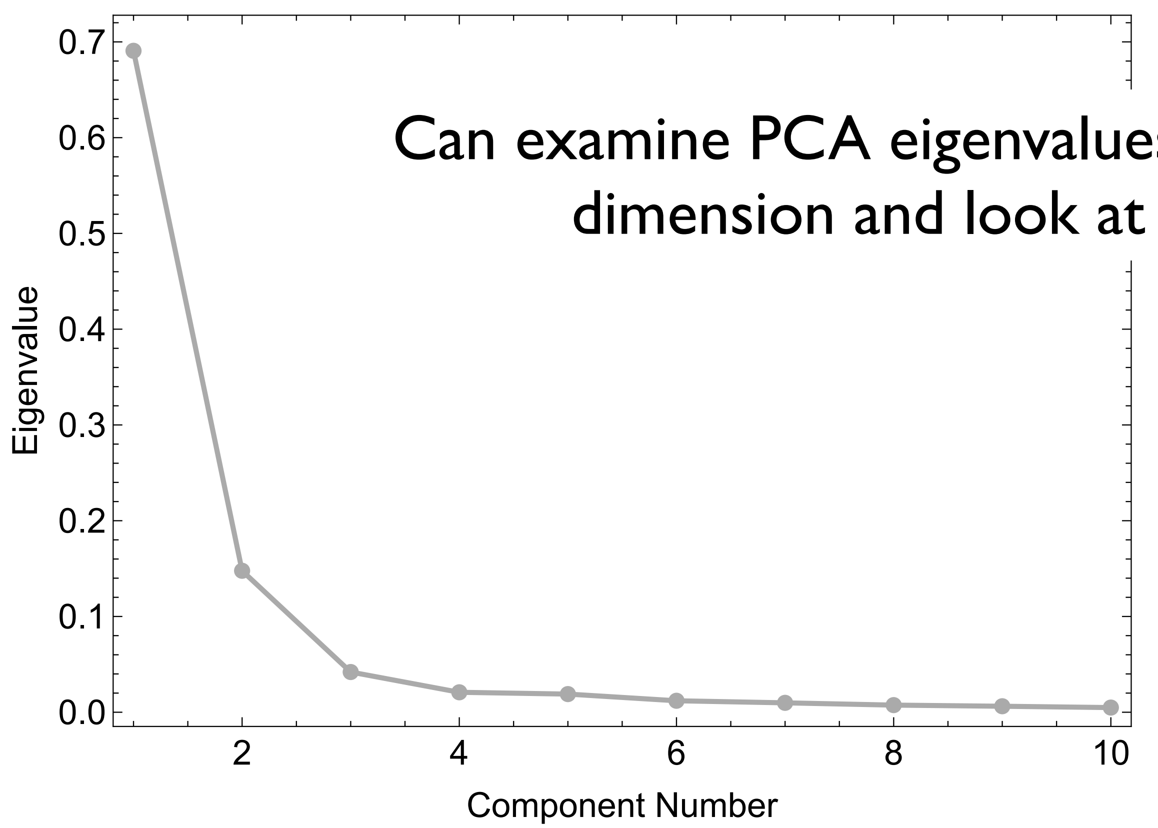


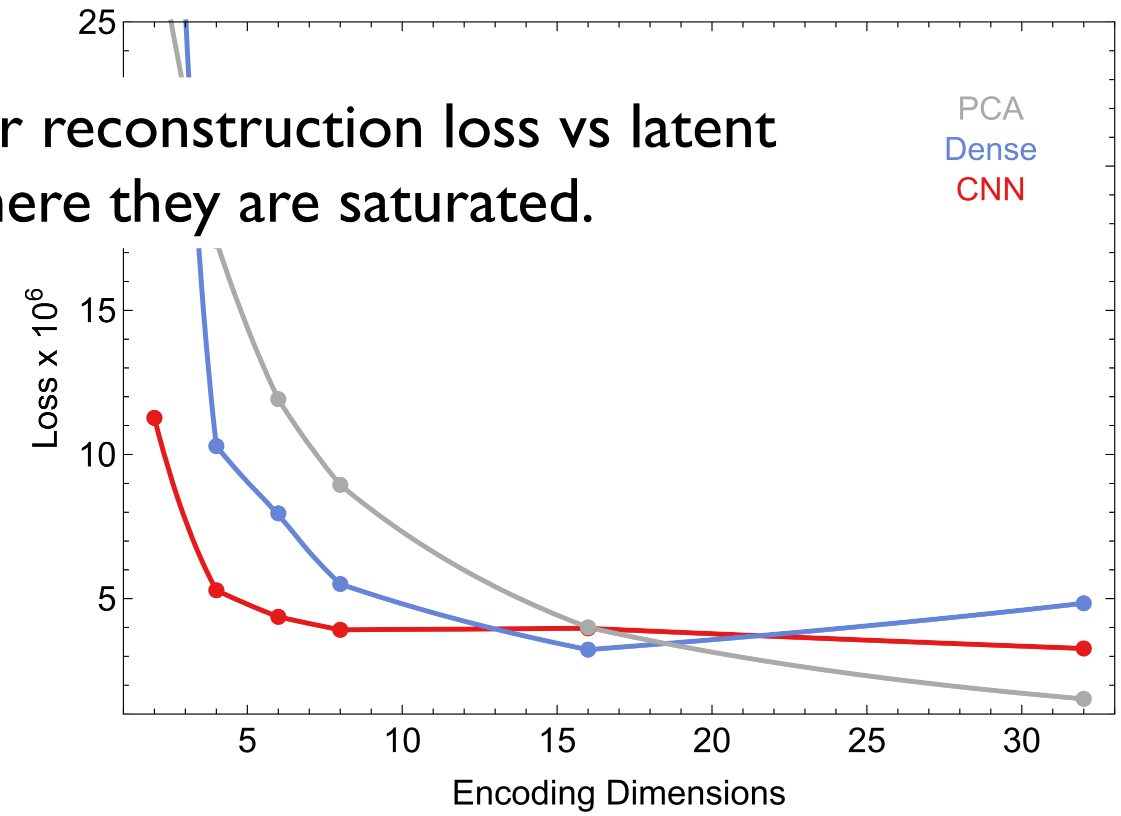


Choosing the latent dimension

d too large → autoencoder becomes identity transform
d too small → autoencoder cannot learn all the features

Should choose the latent dimension in an unsupervised manner
(ie without optimizing on a specific signal)



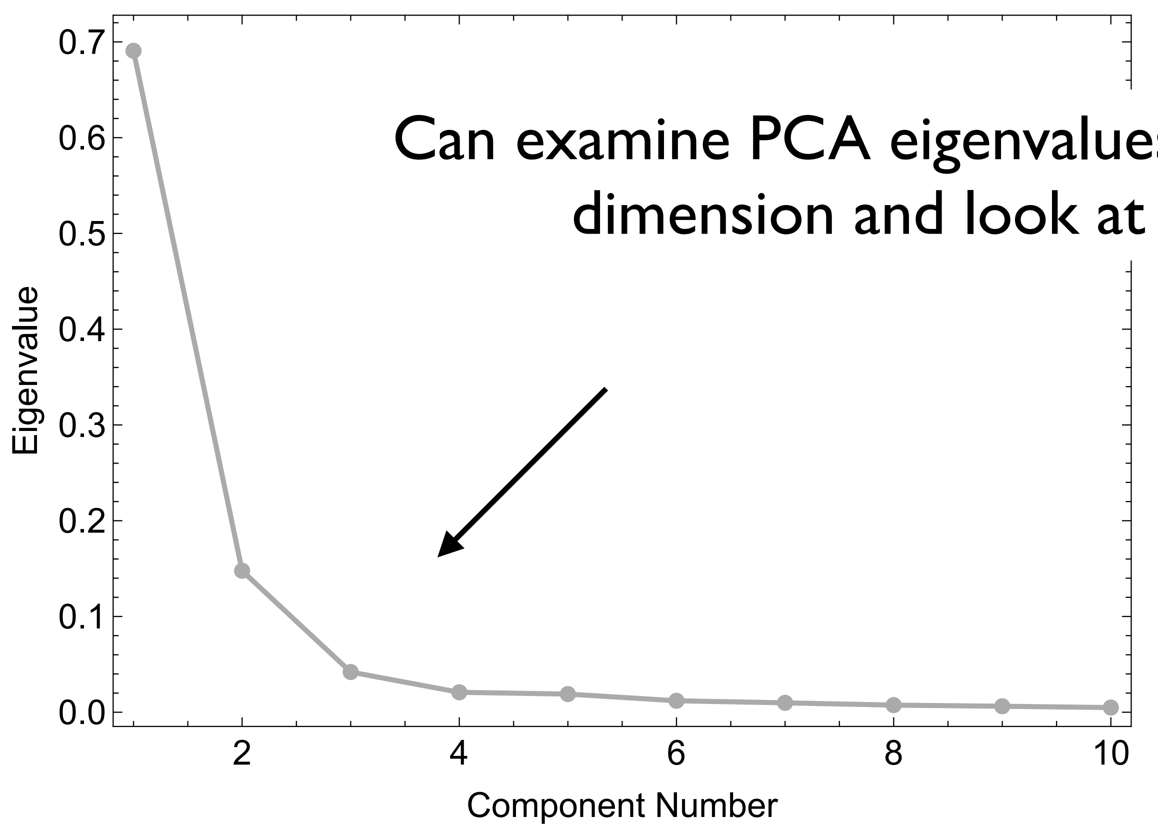


Can examine PCA eigenvalues or reconstruction loss vs latent dimension and look at where they are saturated.

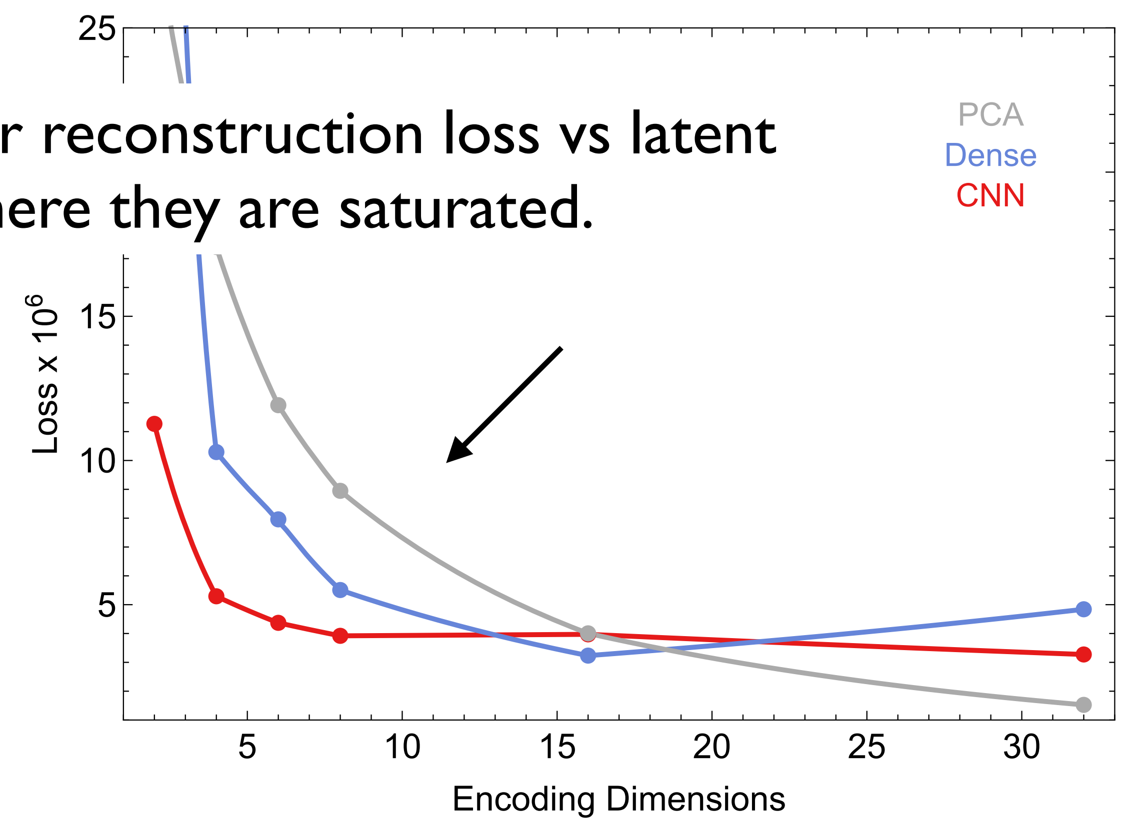
Choosing the latent dimension

d too large → autoencoder becomes identity transform
d too small → autoencoder cannot learn all the features

Should choose the latent dimension in an unsupervised manner
(ie without optimizing on a specific signal)



Can examine PCA eigenvalues or reconstruction loss vs latent dimension and look at where they are saturated.

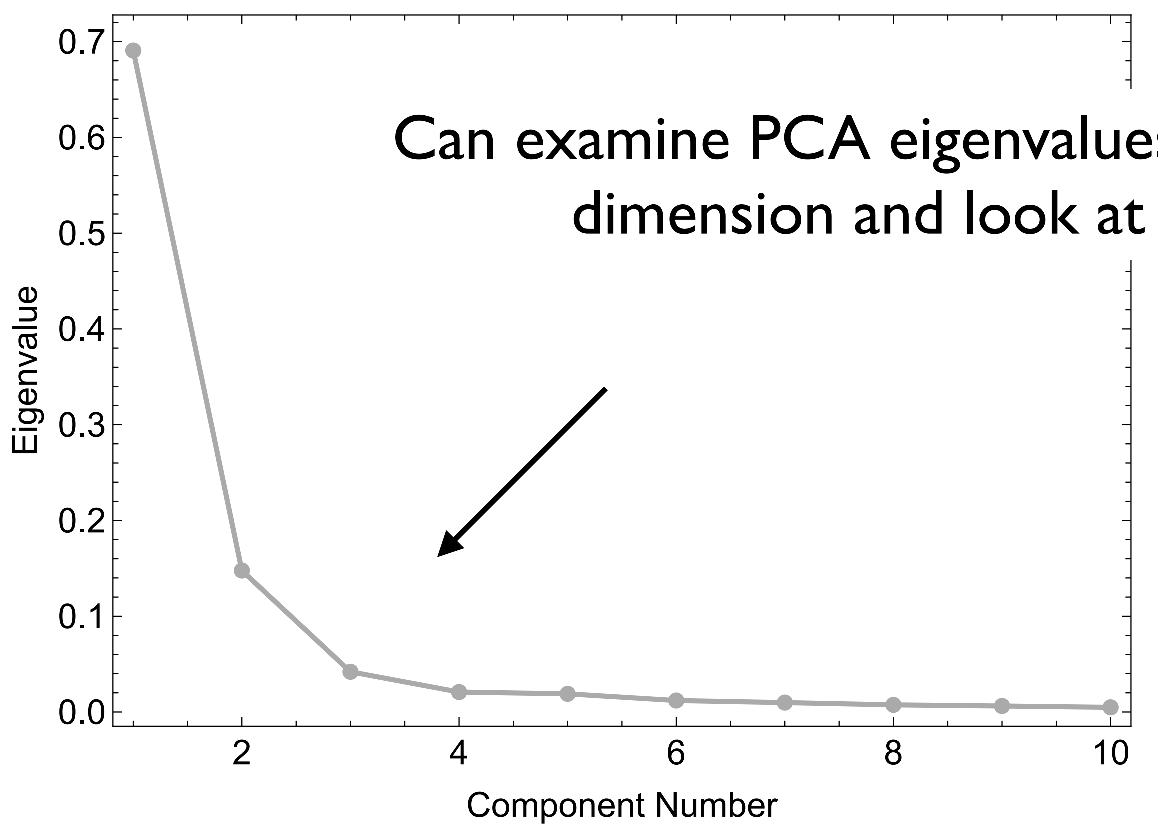


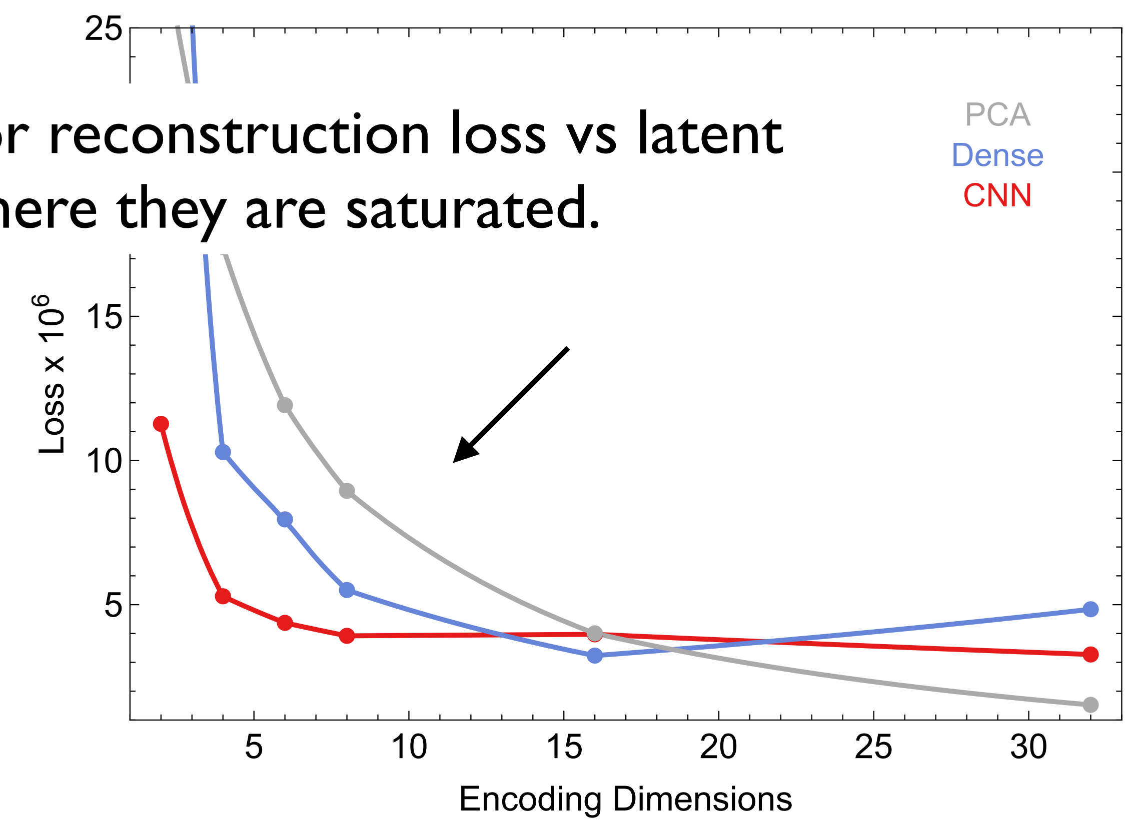
Choosing the latent dimension

d too large → autoencoder becomes identity transform
d too small → autoencoder cannot learn all the features

Should choose the latent dimension in an unsupervised manner
(ie without optimizing on a specific signal)

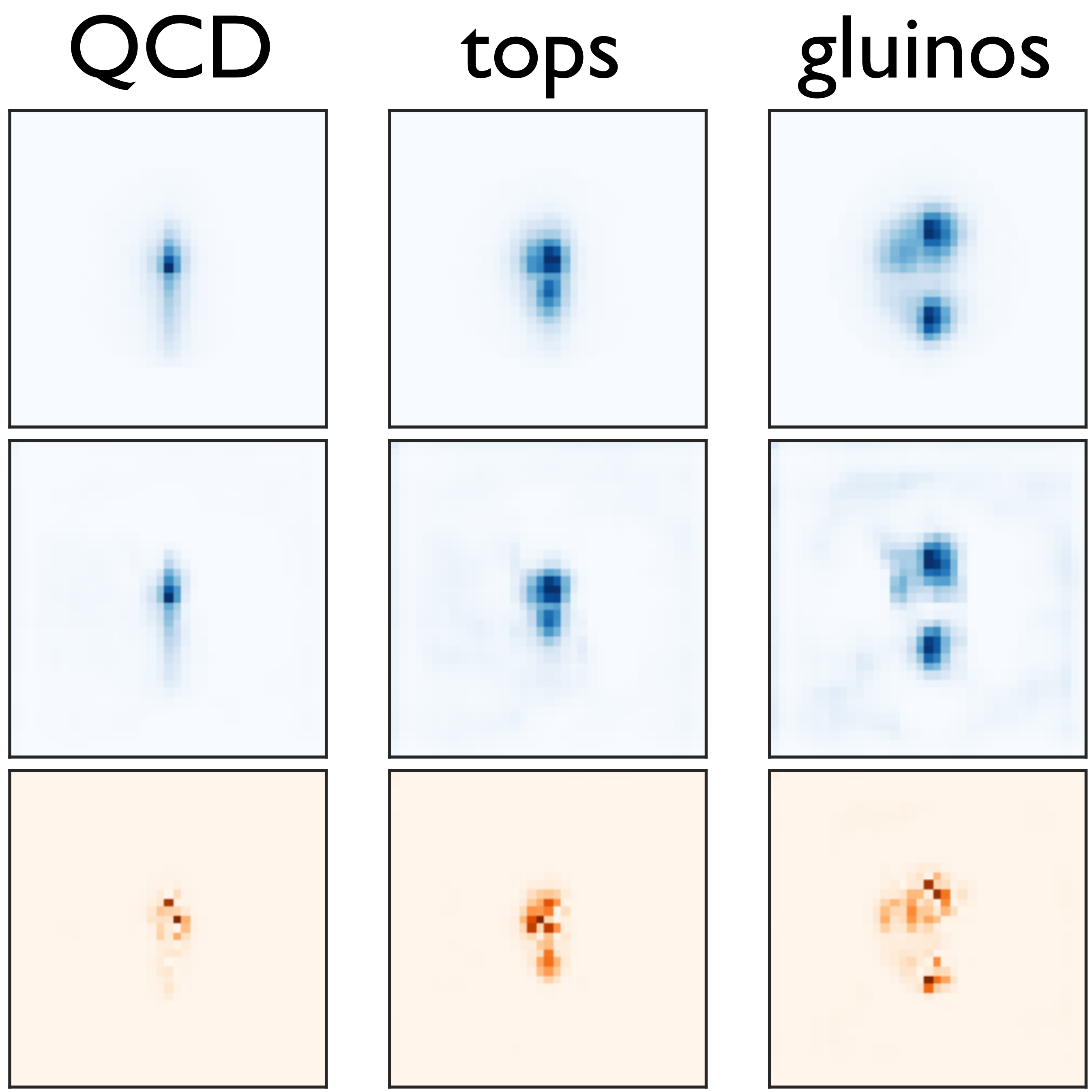
Can examine PCA eigenvalues or reconstruction loss vs latent dimension and look at where they are saturated.





We chose d=6

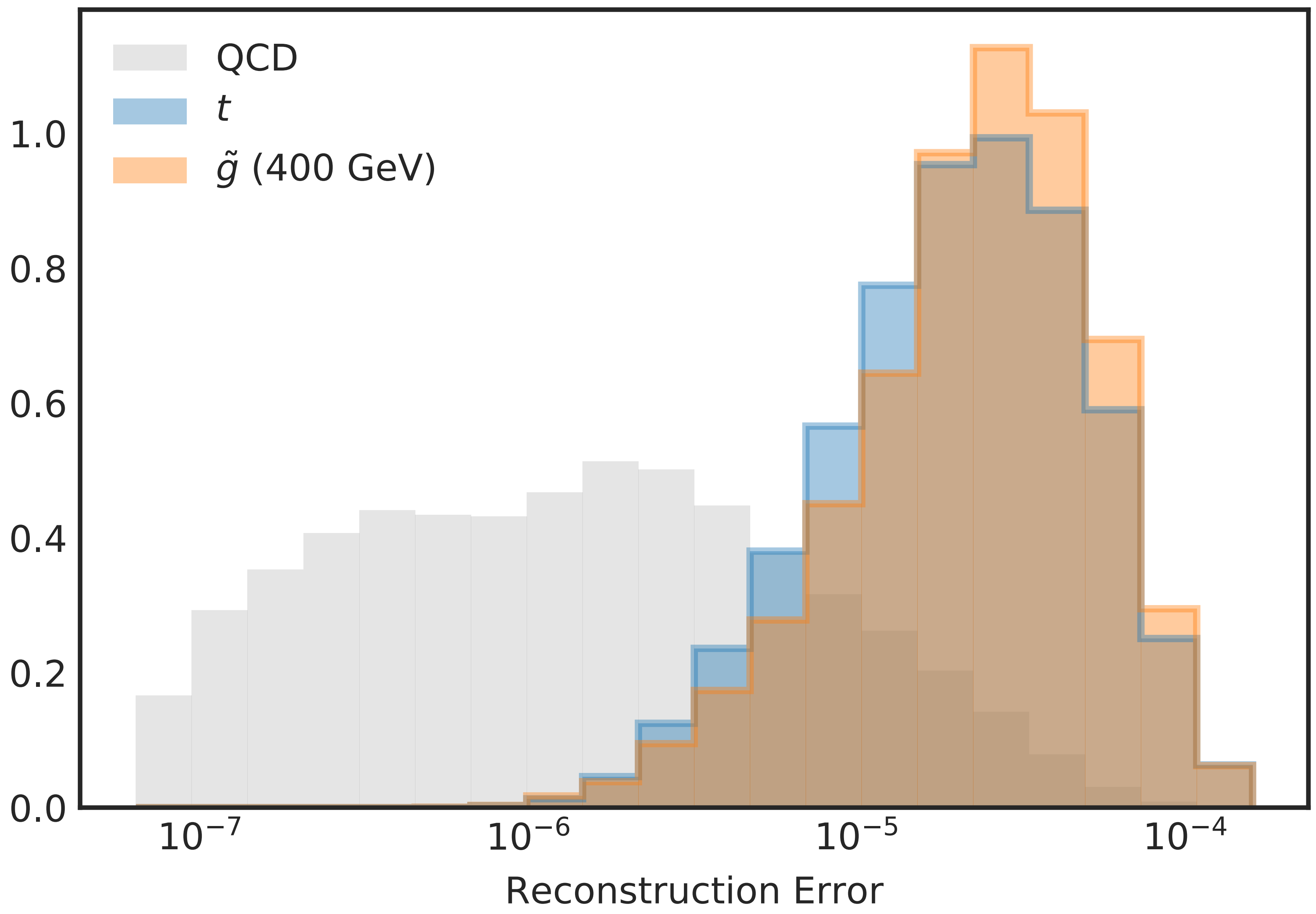
Performance should be worse on "anomalous" events that autoencoder was not trained on.



(CNN AE)

The algorithm works when trained on QCD backgrounds!

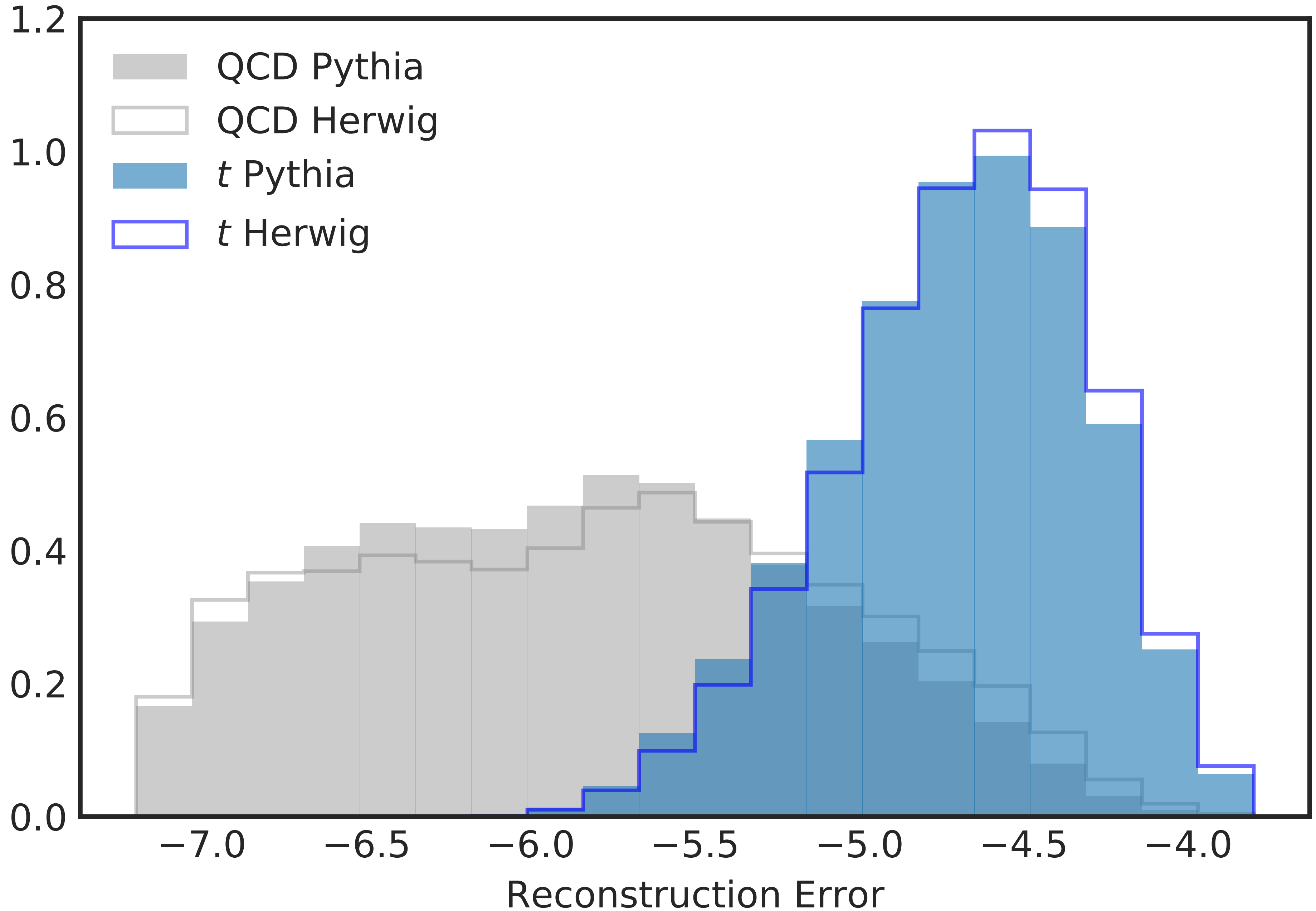
Can use reconstruction error as an anomaly threshold.



(CNN AE)

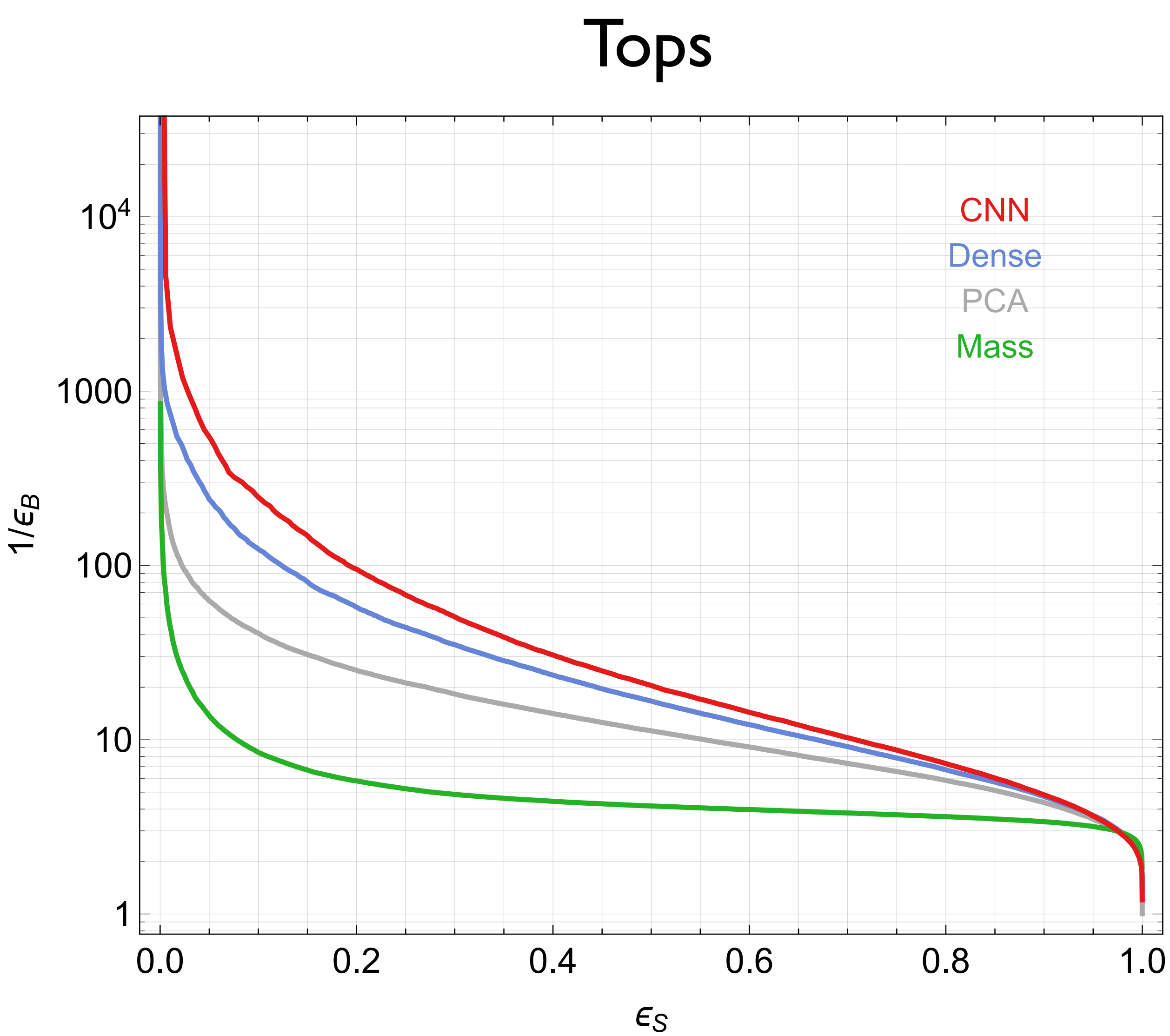
The algorithm works when trained on QCD backgrounds!

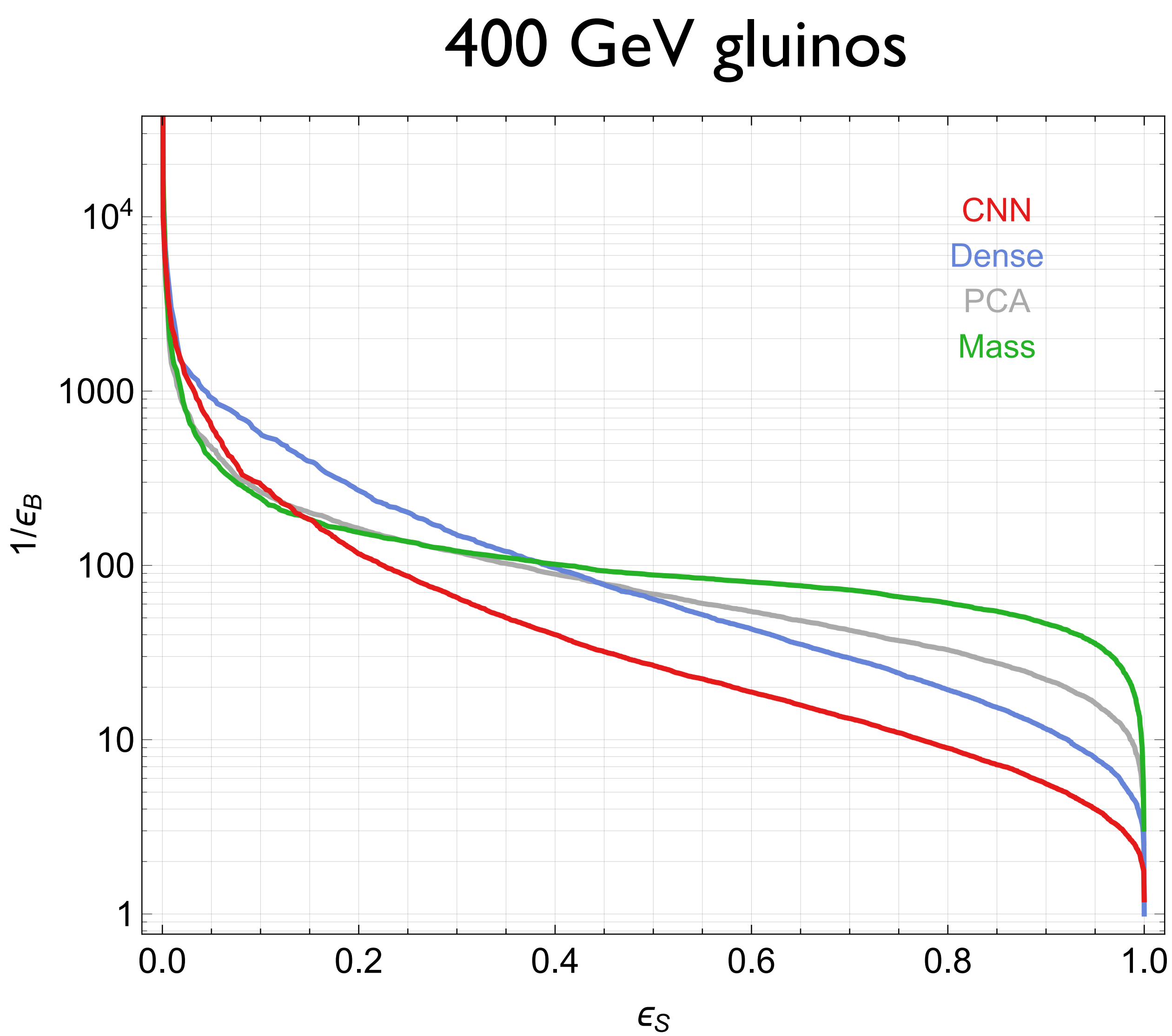
Robustness with other Monte Carlo



Comparison vs jet mass

How do our fancy autoencoders compare against a simpler anomaly detection method: jet mass bump hunt?

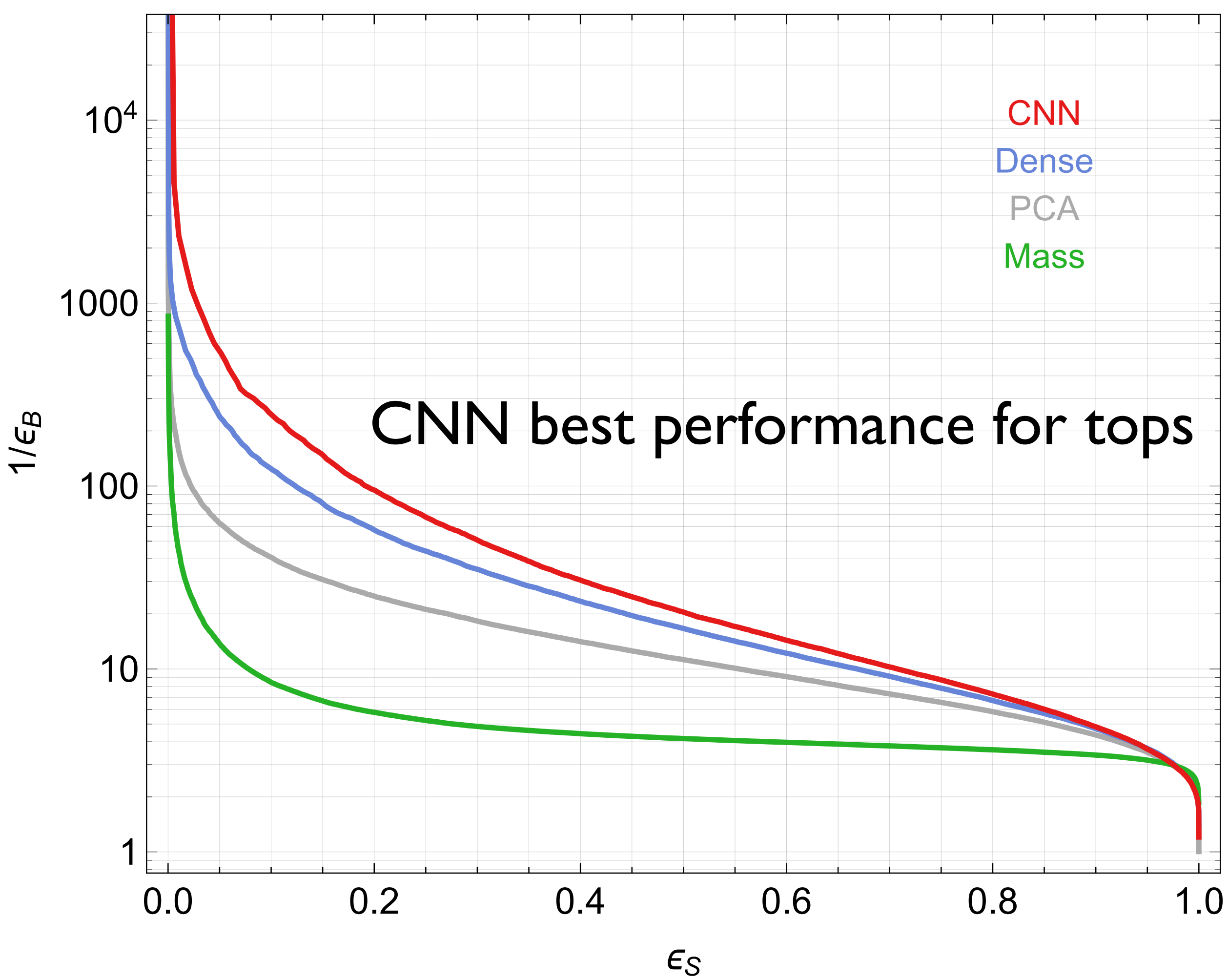




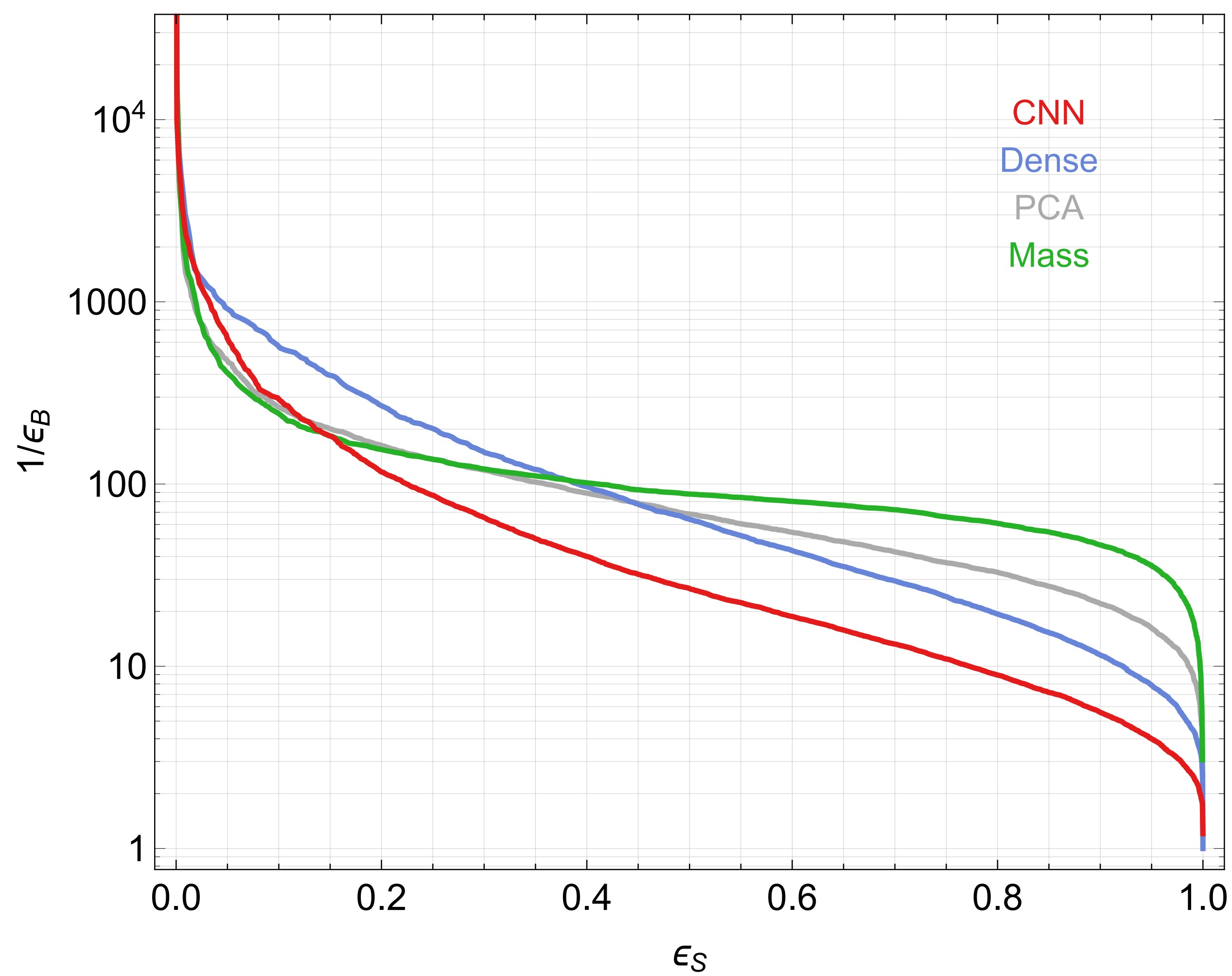
Comparison vs jet mass

How do our fancy autoencoders compare against a simpler anomaly detection method: jet mass bump hunt?

Tops



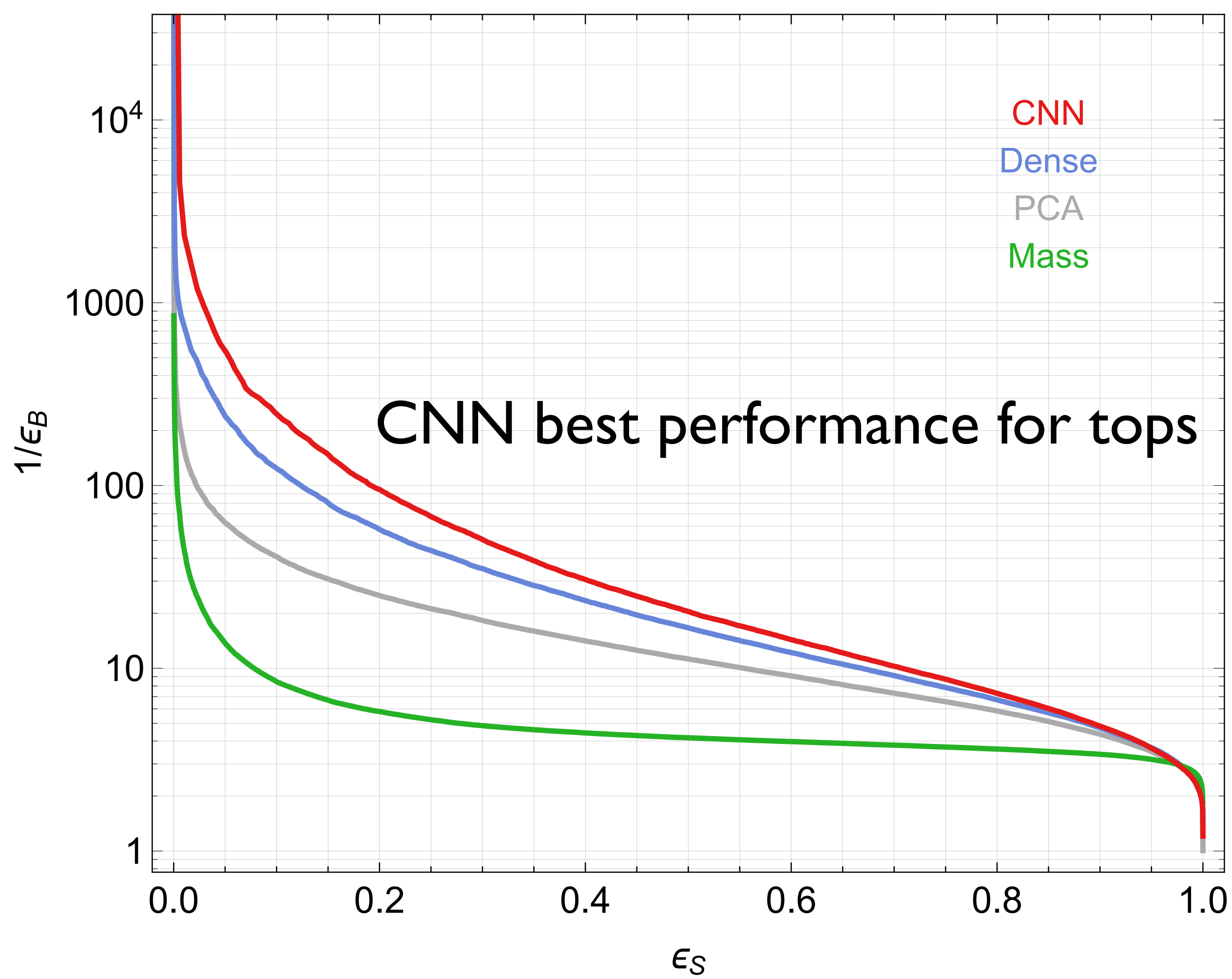
400 GeV gluinos



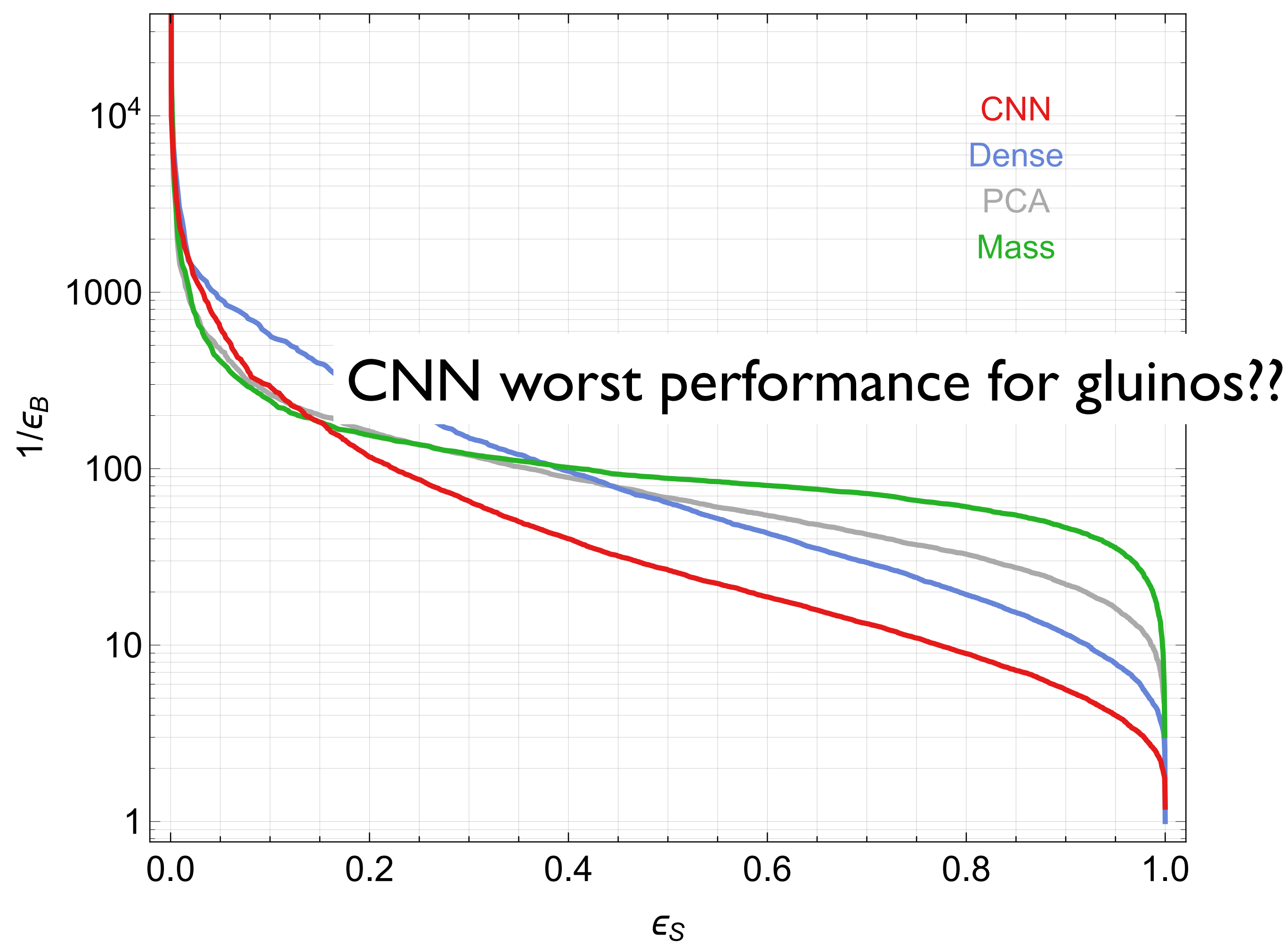
Comparison vs jet mass

How do our fancy autoencoders compare against a simpler anomaly detection method: jet mass bump hunt?

Tops



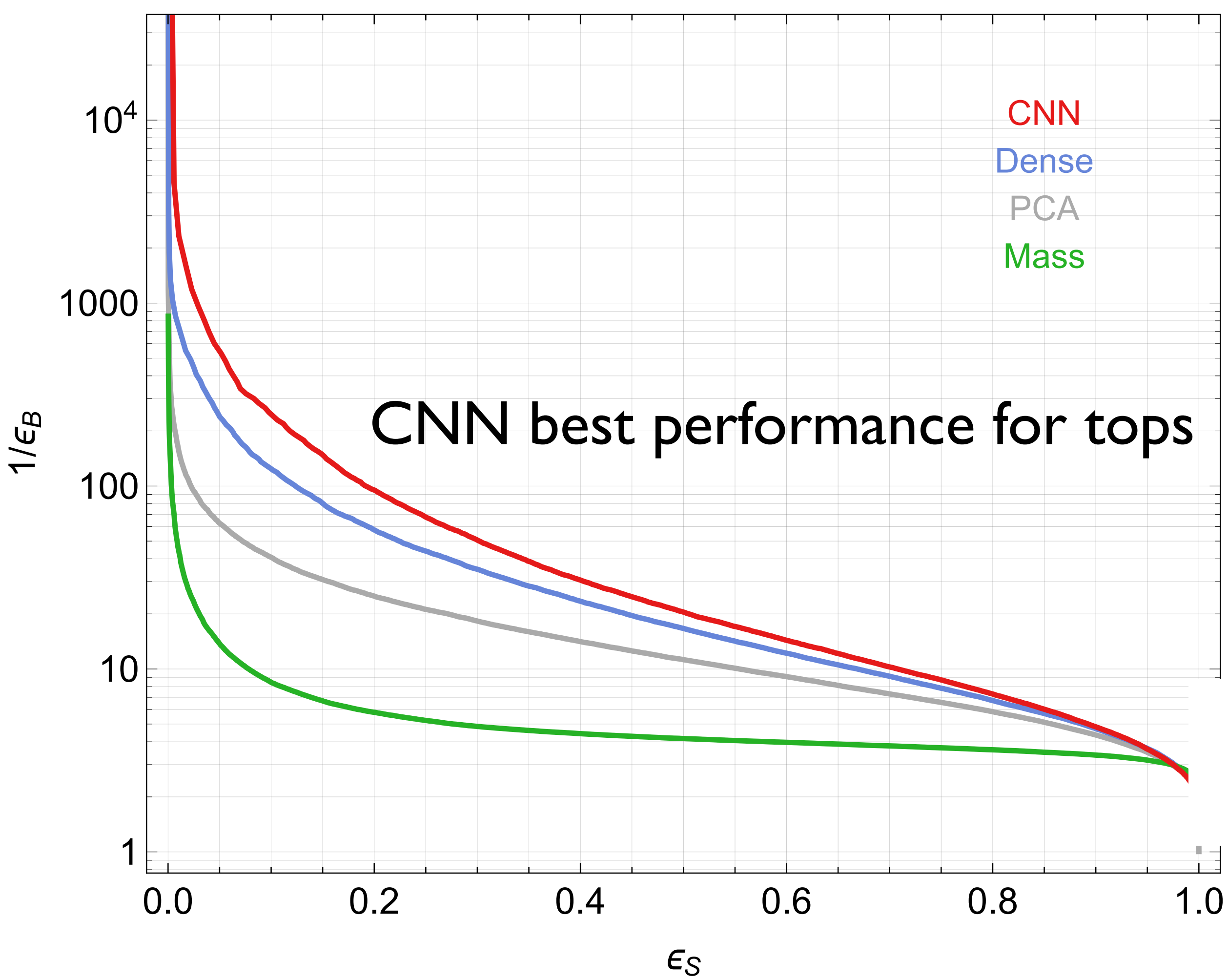
400 GeV gluinos



Comparison vs jet mass

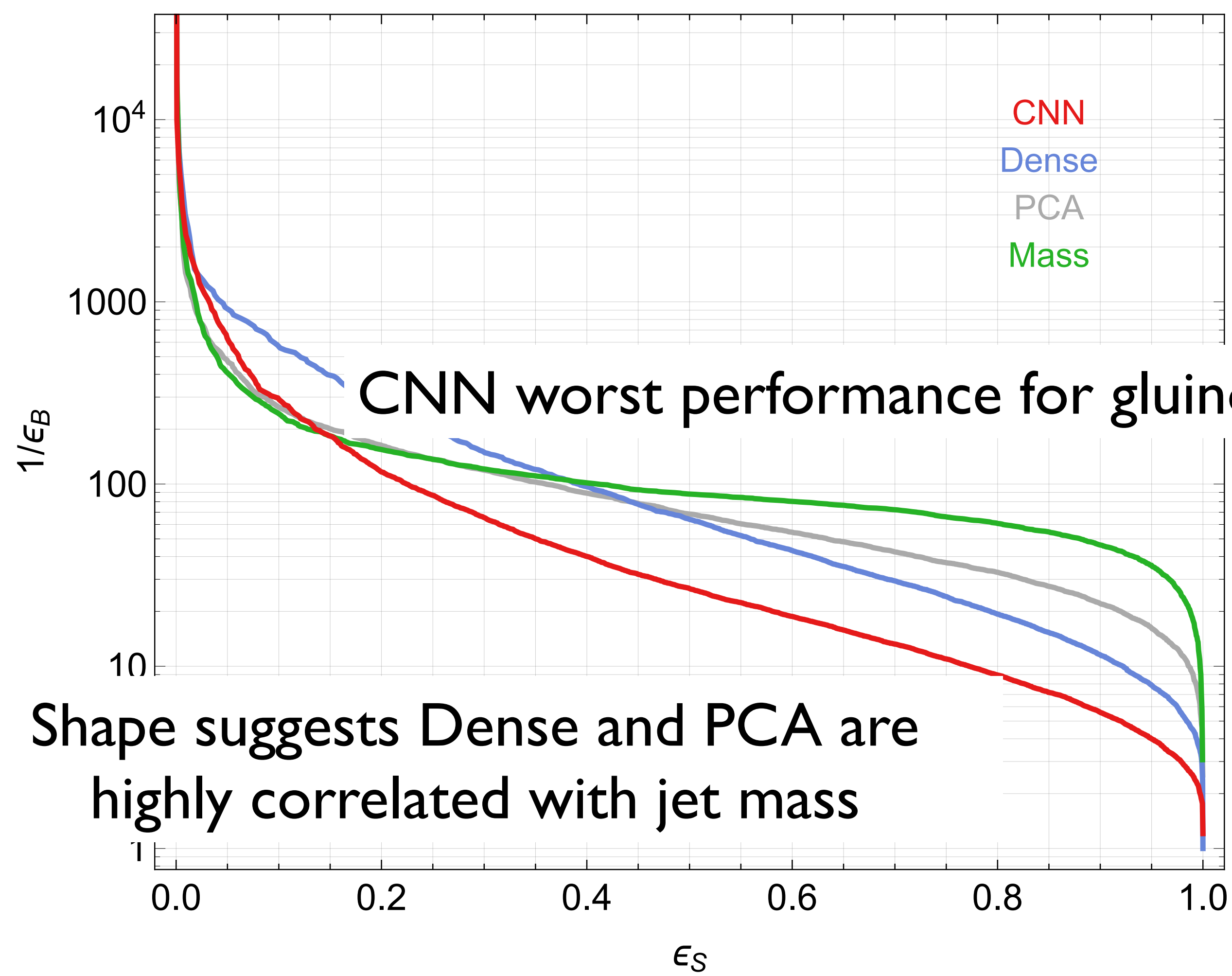
How do our fancy autoencoders compare against a simpler anomaly detection method: jet mass bump hunt?

Tops



CNN best performance for tops

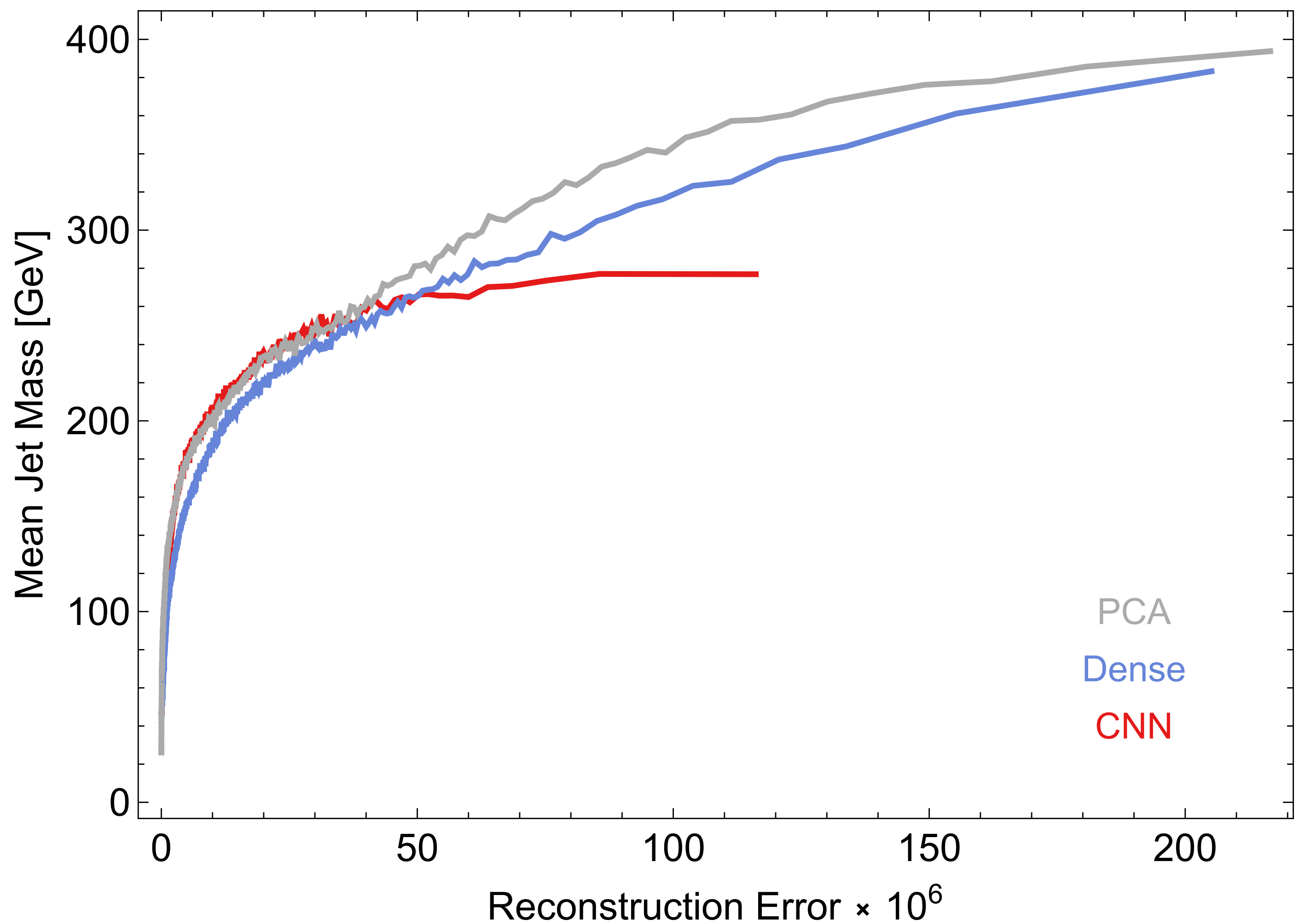
400 GeV gluinos



CNN worst performance for gluinos??

Shape suggests Dense and PCA are highly correlated with jet mass

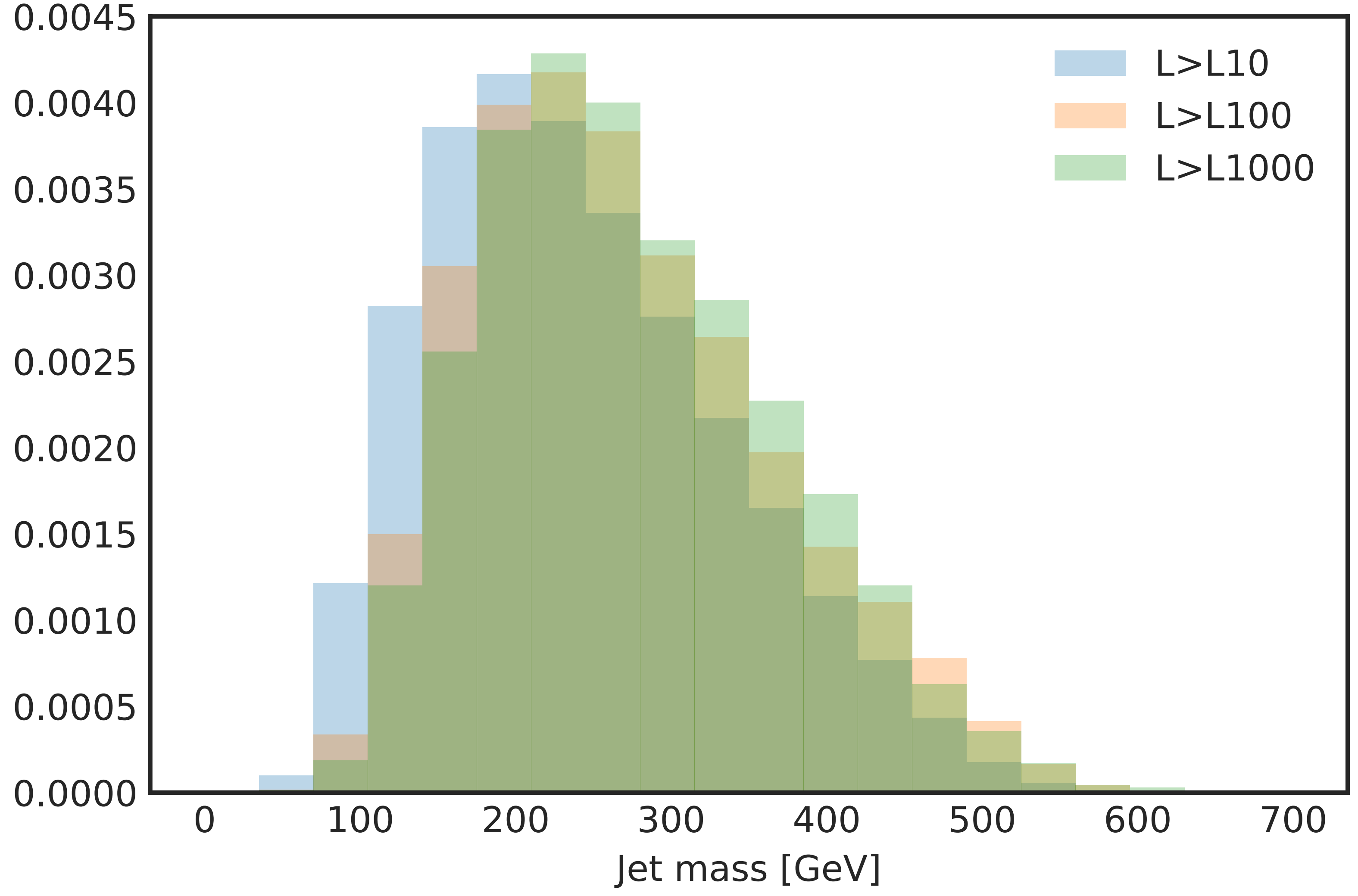
Correlation with jet mass



Indeed, this is confirmed by looking at mean jet mass in bins of reconstruction error for the QCD background.

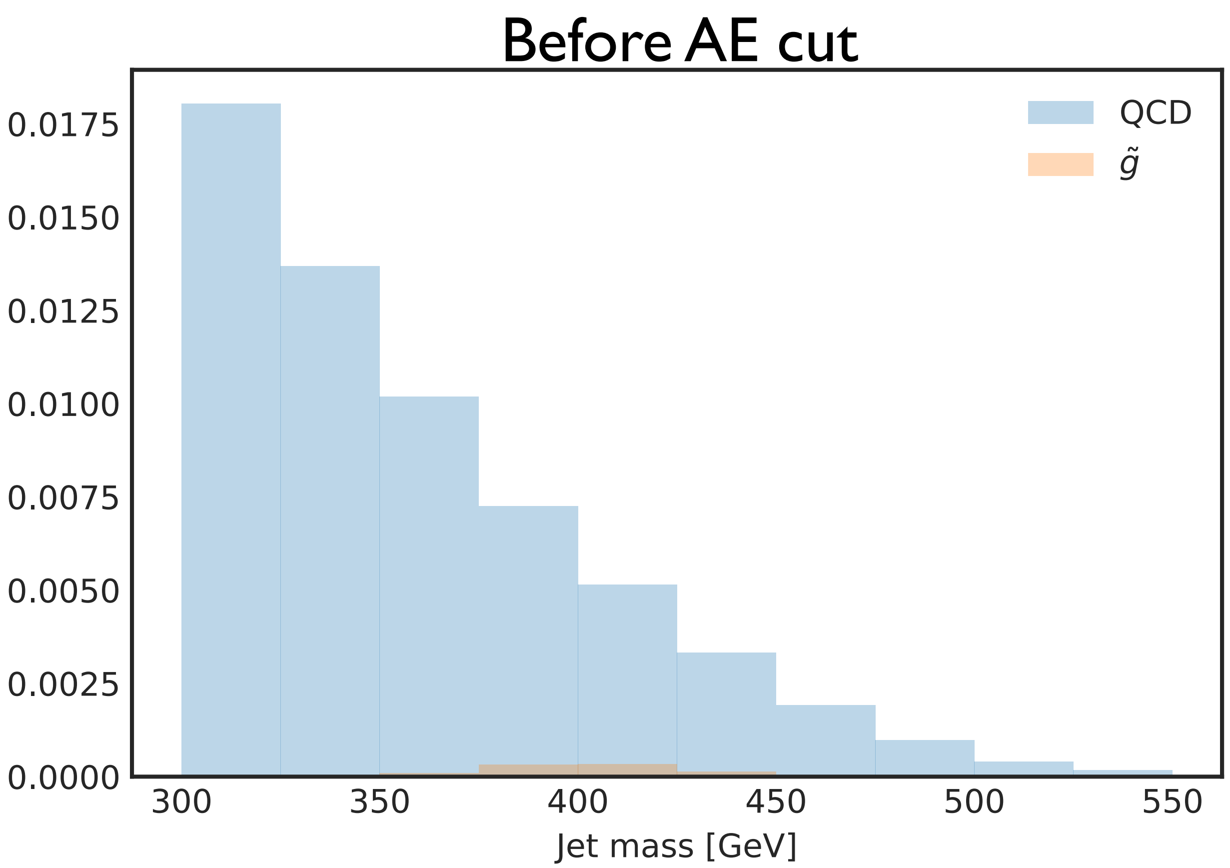
CNN is no longer correlated with jet mass for m≳250 GeV

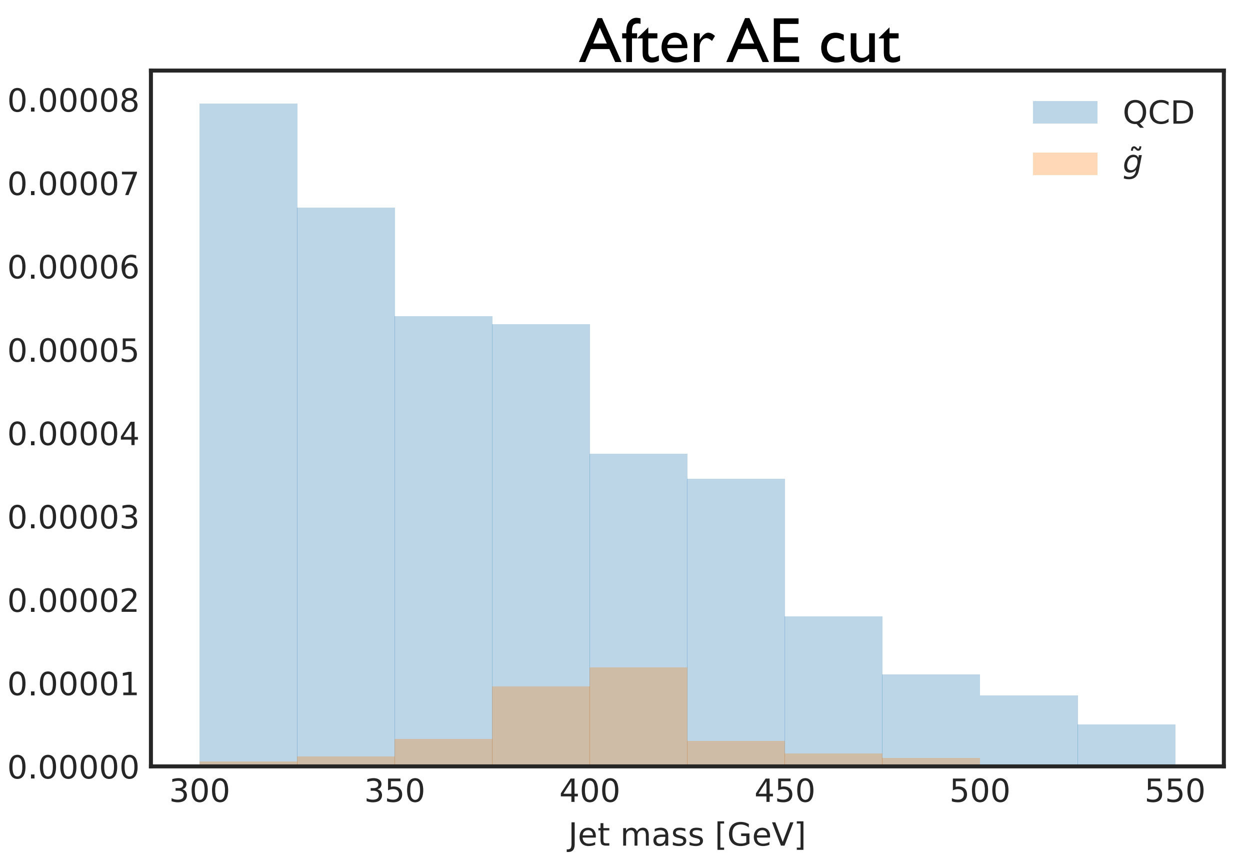
Correlation with jet mass



The QCD jet mass distribution is stable against harder cuts on the reconstruction error, for the CNN autoencoder.

Bump hunt with deep autoencoder





Can combine the CNN autoencoder with a bump hunt in jet mass.

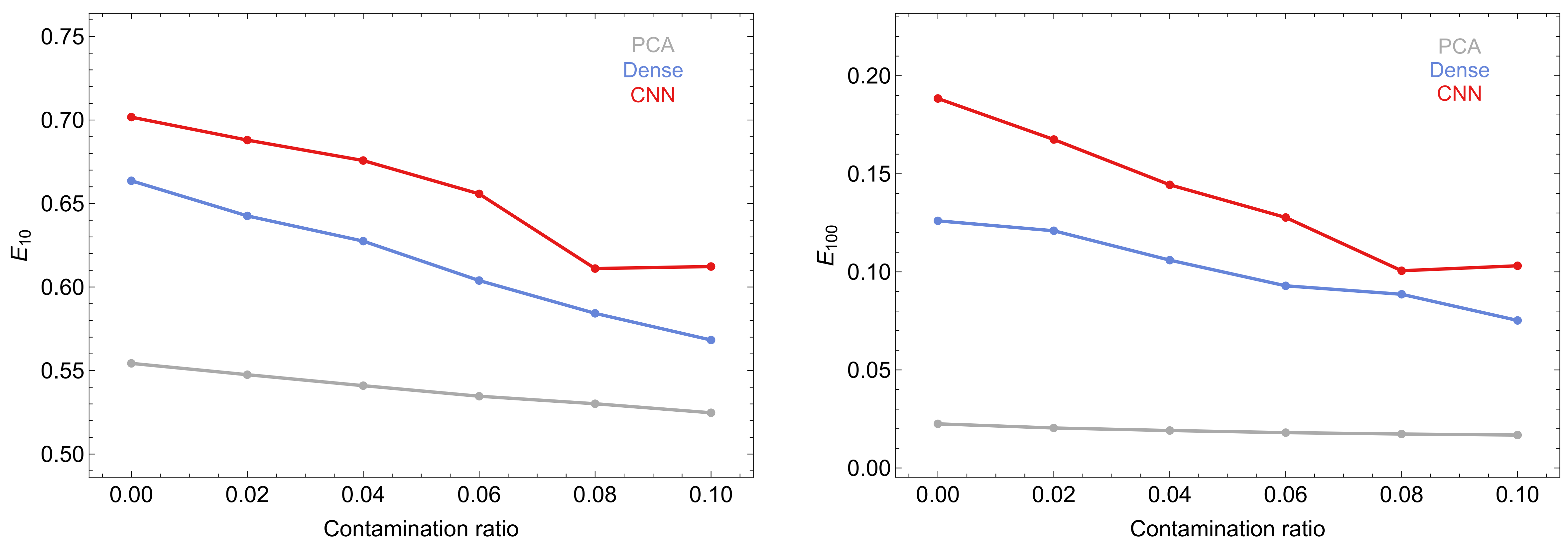
Use the AE to first clean away "boring" QCD jets.

Improve S/B by a lot!

Fully unsupervised learning

Train on sample of QCD background "contaminated" with a small fraction of signal.

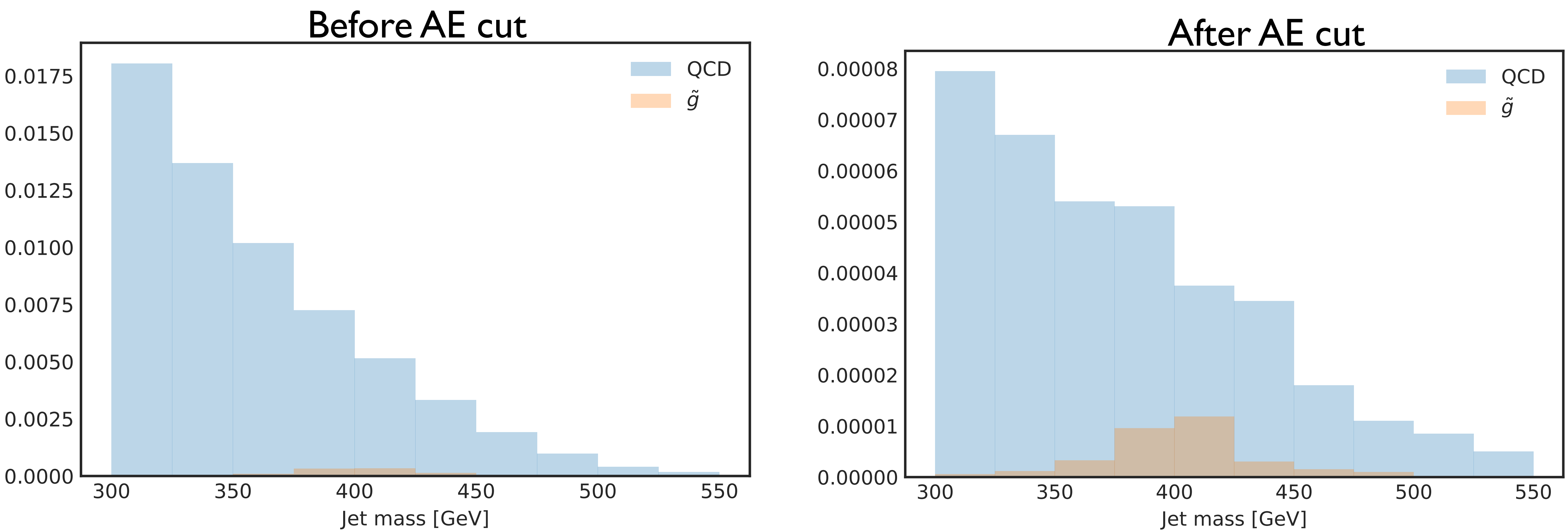
Representative of actual data.



(E_x = signal efficiency at bg rejection = x)

Performance of AE surprisingly robust even up to 10% contamination!

Bump hunt with deep autoencoder



Can train directly on data that contains 400 GeV gluinos,
and still enhance the bump hunt.

Could really discover new physics this way!

Thank you for your attention!

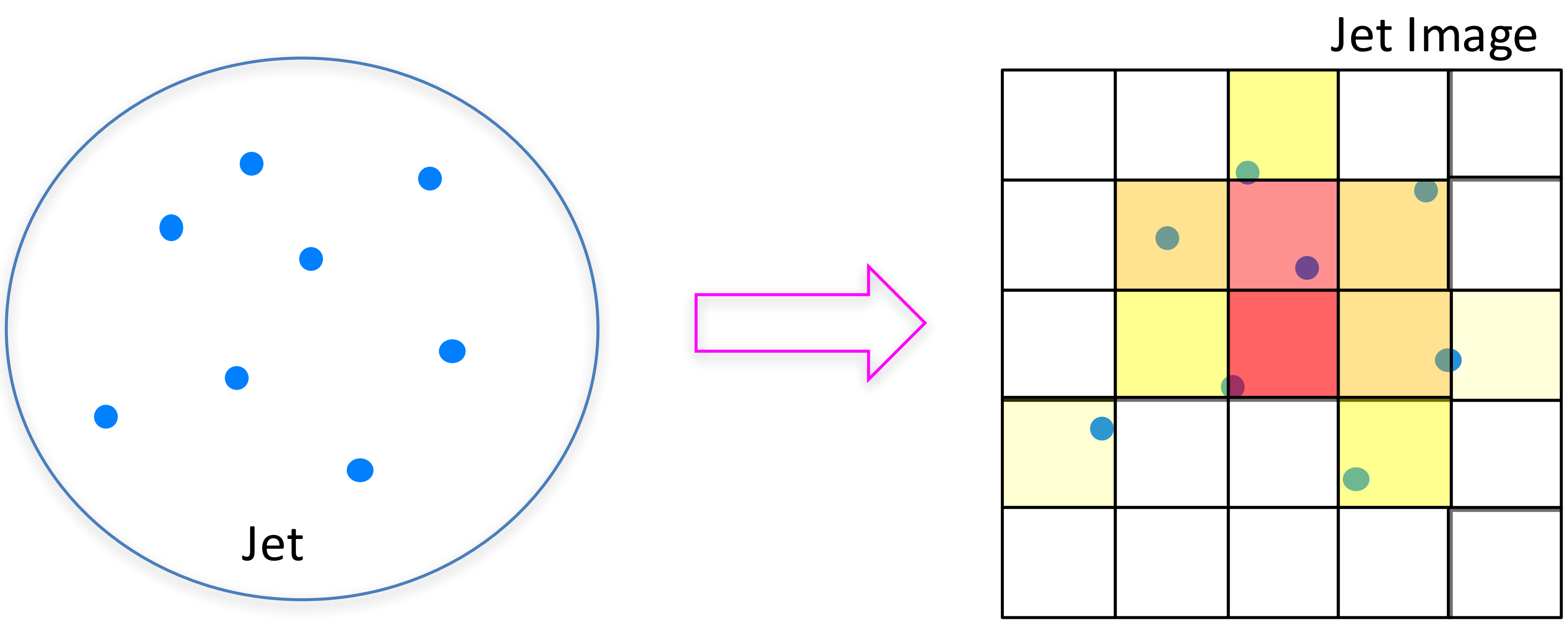
Backup material

CNN Top Tagger Details

	DeepTop minimal	Our final tagger
Training	SGD $\eta = 0.003$ minibatch size=1000 MSE loss	AdaDelta $\eta = 0.3$ with annealing schedule minibatch size=128 cross entropy loss
CNN architecture	8C4-8C4-MP2-8C4-8C4- 64N-64N-64N	128C4-64C4-MP2-64C4-64C4-MP2- 64N-256N-256N
Preprocessing	pixelate→center → normalize	center→rotate→flip → normalize→pixelate
Sample size	150k+150k	1.2M+1.2M
Color	$p_T^{calo} = p_T^{neutral} + p_T^{track}$	$(p_T^{neutral}, p_T^{track}, N_{track}, N_{muon})$

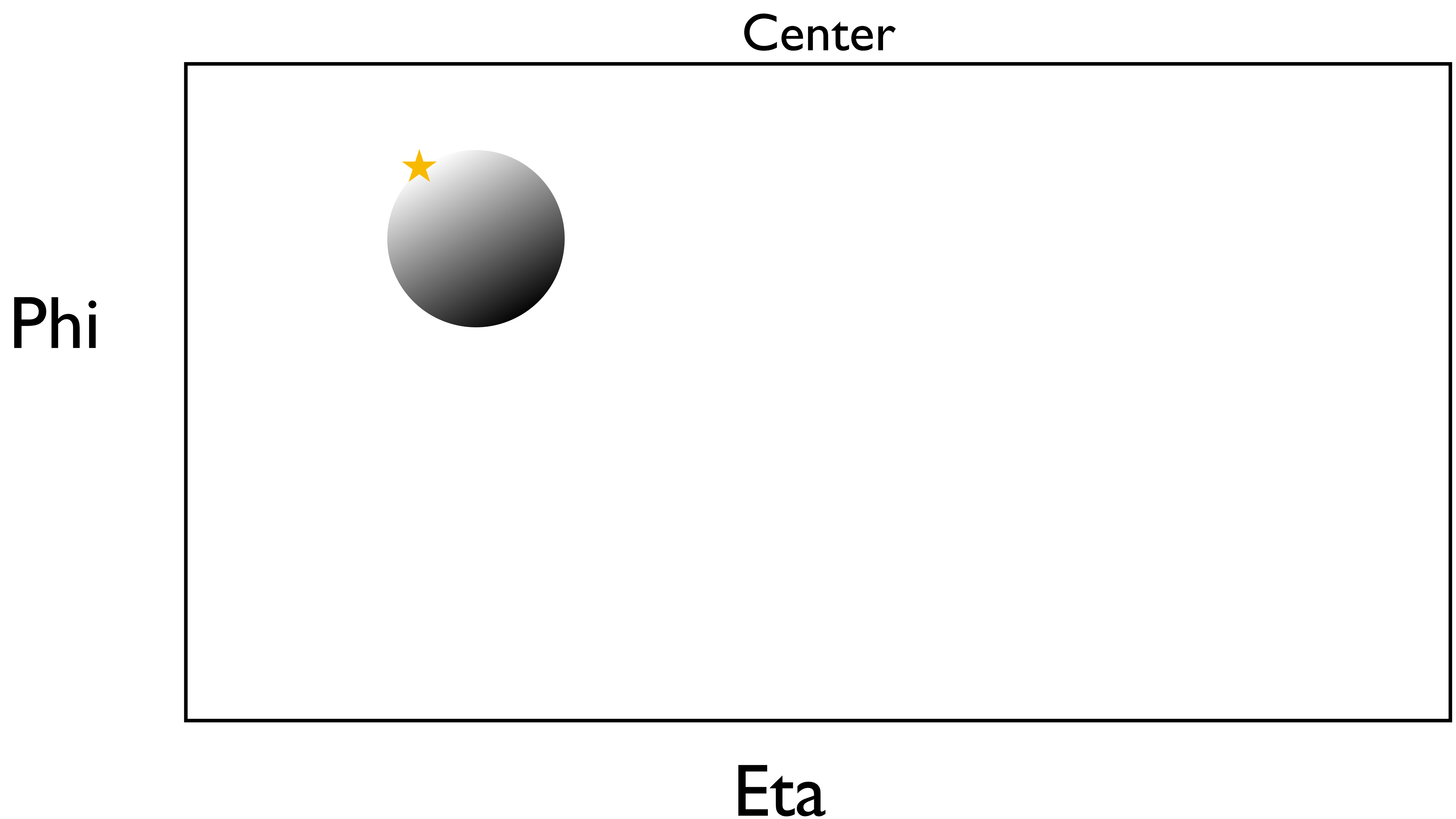
	t	$\tilde{g}$
PCA	0.51 / 0.04	0.98 / 0.36
Dense	0.66 / 0.13	0.90 / 0.39
CNN	0.70 / 0.19	0.77 / 0.23

Jets as images



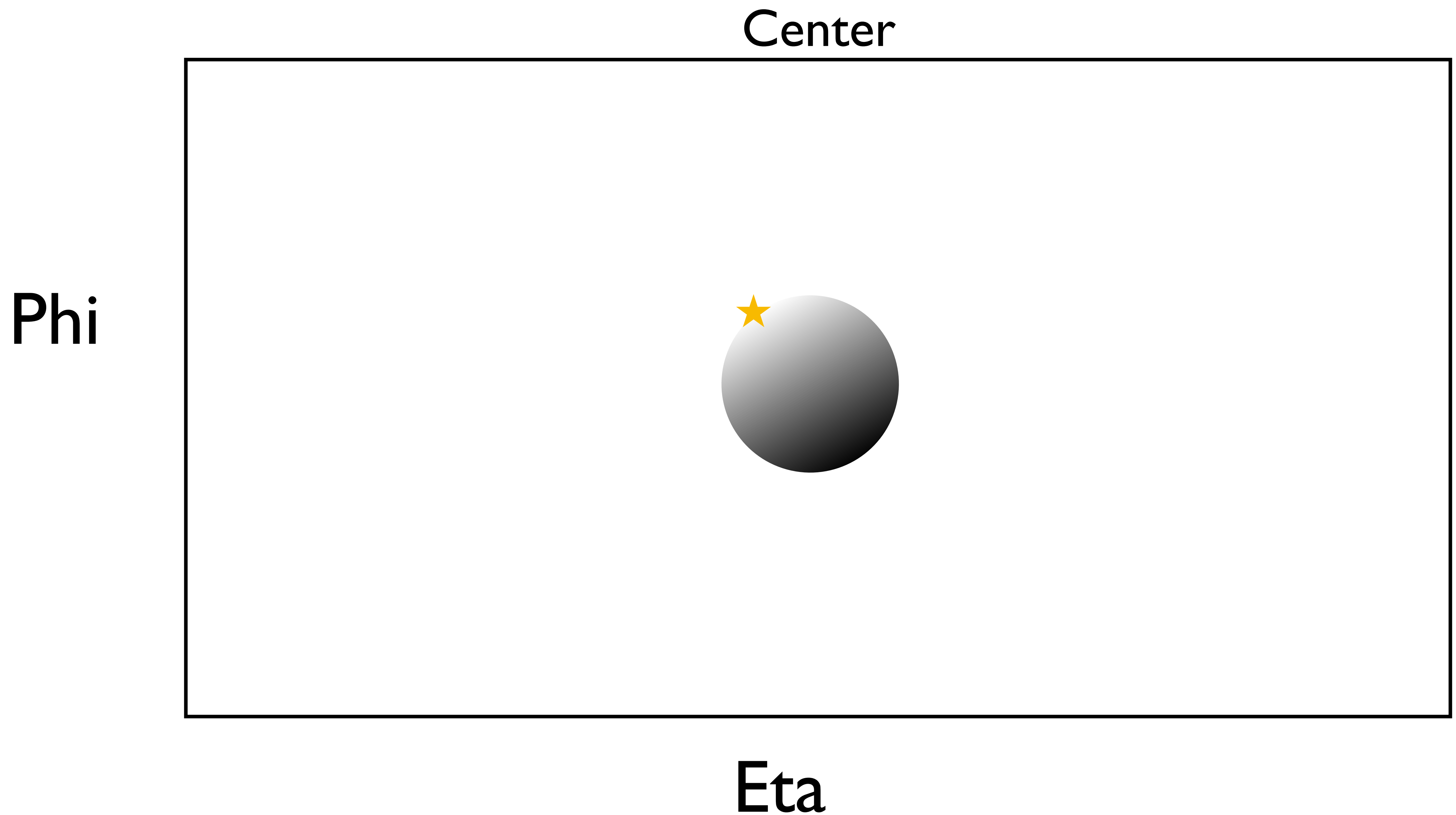
Jet Images — Preprocessing

For training the neural network, it is very useful to uniformize the jet images as much as possible, consistently with the physics.



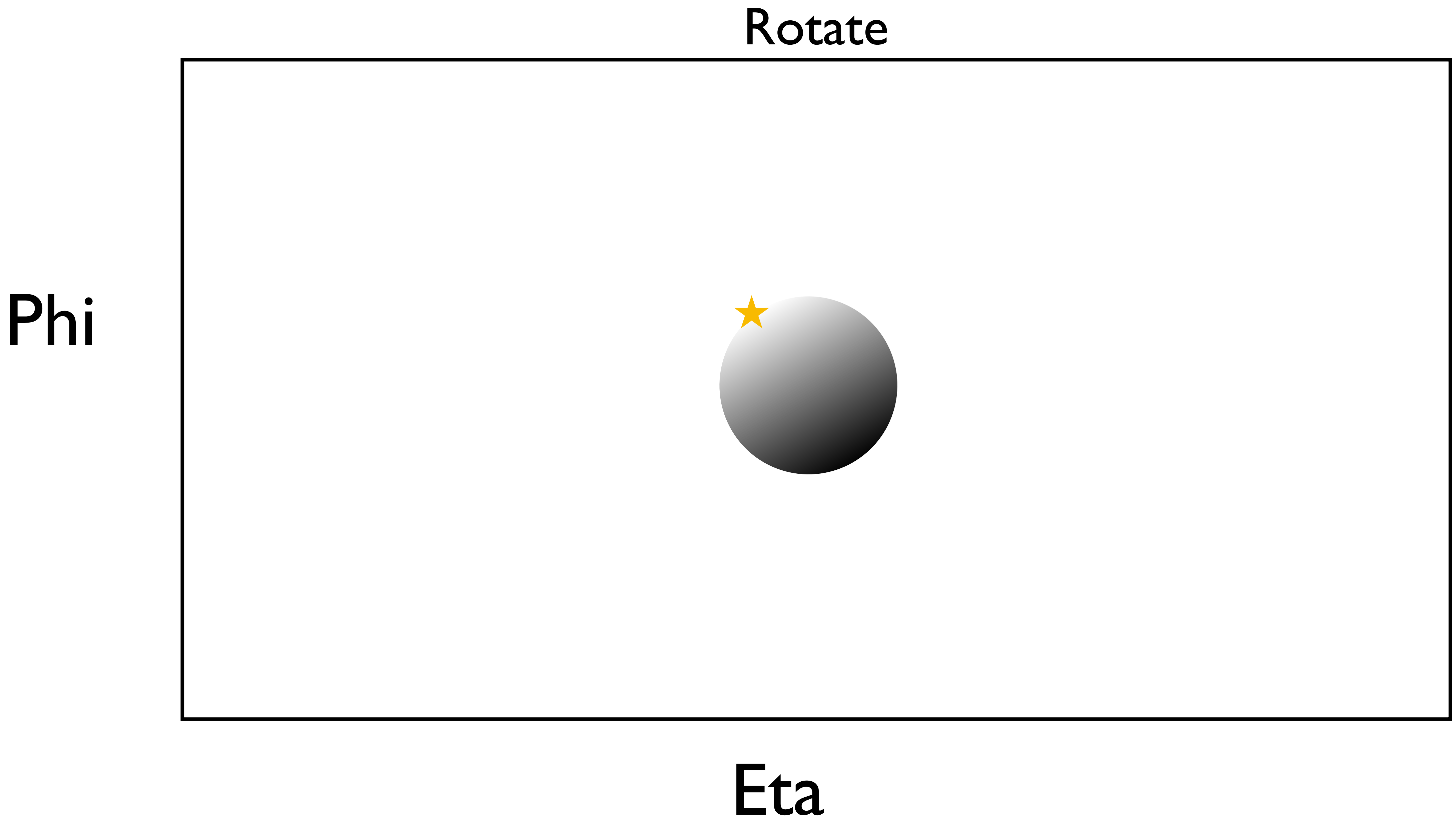
Jet Images — Preprocessing

For training the neural network, it is very useful to uniformize the jet images as much as possible, consistently with the physics.



Jet Images — Preprocessing

For training the neural network, it is very useful to uniformize the jet images as much as possible, consistently with the physics.

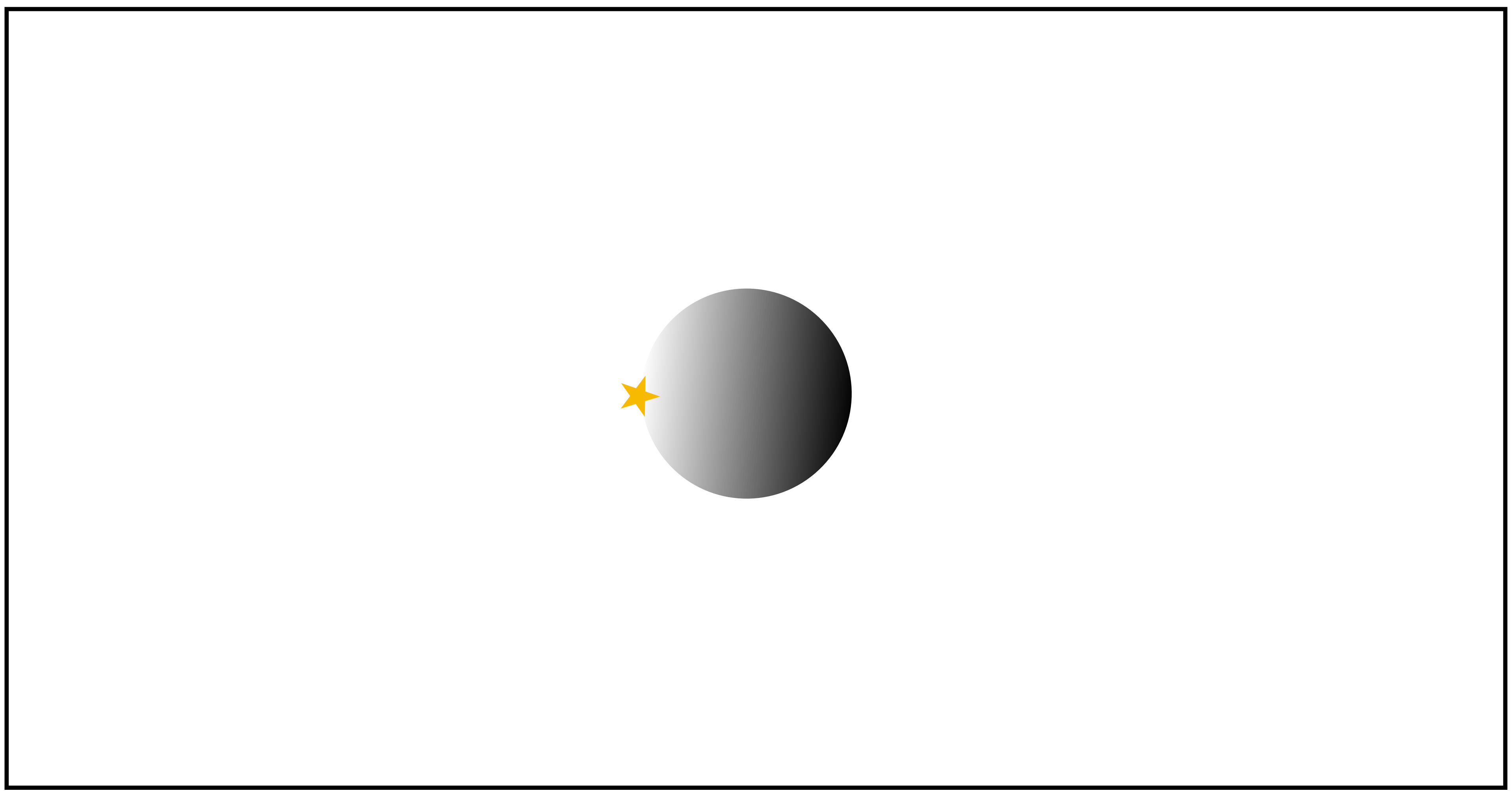


Jet Images — Preprocessing

For training the neural network, it is very useful to uniformize the jet images as much as possible, consistently with the physics.

Rotate

Phi

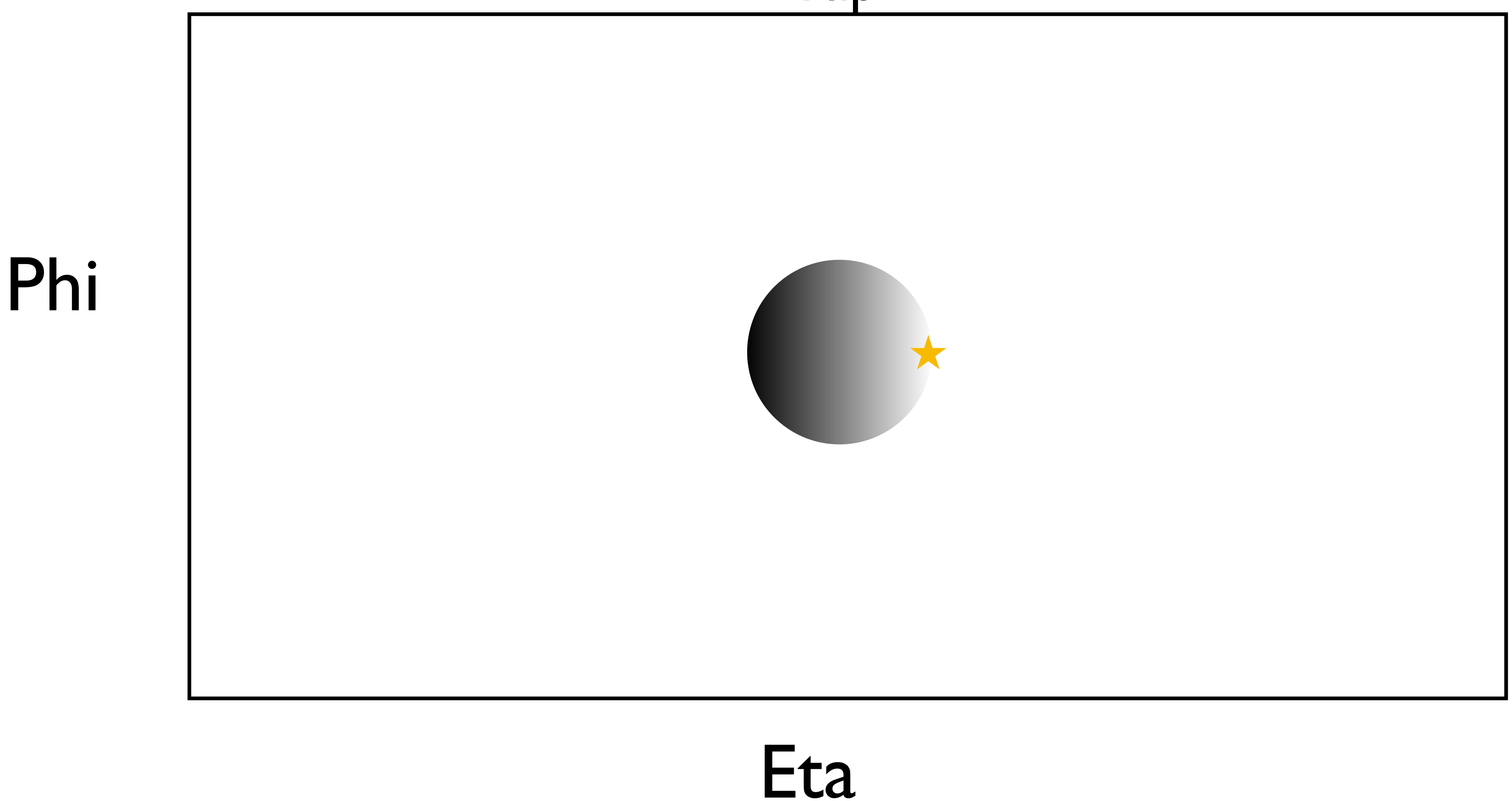


Eta

Jet Images — Preprocessing

For training the neural network, it is very useful to uniformize the jet images as much as possible, consistently with the physics.

Flip



Supervised vs Unsupervised ML

Top tagging is a prime example of "supervised machine learning" — training with labeled datasets.

Supervised learning is great if you know what you're looking for.

But we are interested in searching for the unexpected.

If data has a small, **unknown** signal in it, can we train a NN to find it?

We need "unsupervised learning": training on unlabeled datasets.

Supervised vs Unsupervised ML

Supervised Learning	Unsupervised Learning
Training on labeled data	Training on unlabeled data
Need separate training set	Train directly on entire input dataset
Used for prediction	Used for analysis
Classification, regression	Clustering, density estimation, dimensionality reduction