QCD or What?: Using Autoencoders in HEP

Jennifer Thompson

Universität Heidelberg

30.10.2018

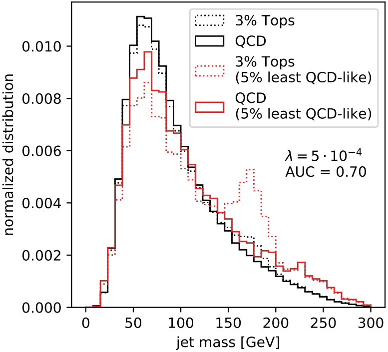
Theo Heimel[1], Gregor Kasieczka[2], Tilman Plehn[1],
arXiv:1808.08979
Jennifer Thompson[1]

[1] ITP Universität Heidelberg [2] Institut für Experimentalphysik Universität Hamburg

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

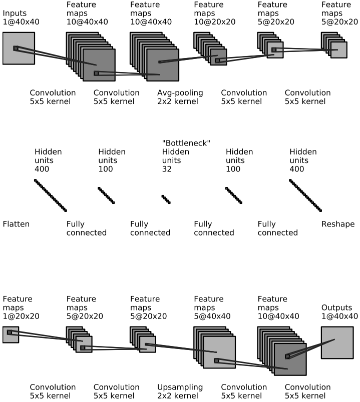
The Autoencoder

- **Data-driven anomaly detector**
- Model-independent approach to new physics searches
- Can be train on a background-dominated signal region
- Possible application in a bump-hunt



The Autoencoder

- **Data-driven anomaly detector**
- Attempt to encode and reconstruct the input
- Learn an efficient compression of QCD
- Reconstruction fails for arbitrary signals
- We consider jet constituents and images as inputs



Tops vs. QCD: bottleneck size

Samples are available: https://goo.gl/XGYju3

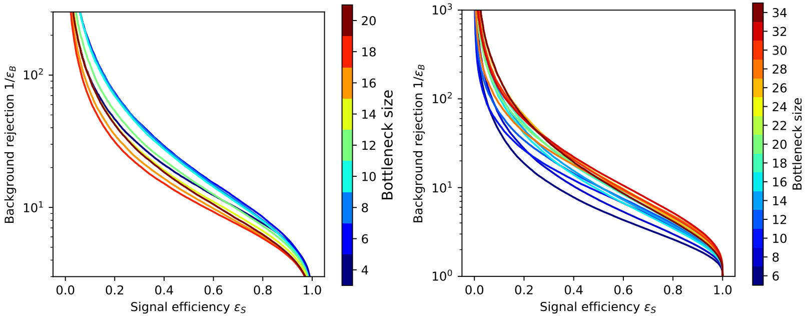
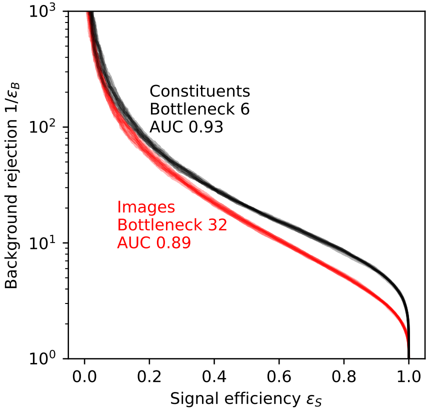


Figure: Dependence on the bottleneck size. Left: constituents. Right: Images.

$\longrightarrow$ Large dependence on bottleneck size

$\longrightarrow$ Constituents prefer lower bottleneck sizes than images

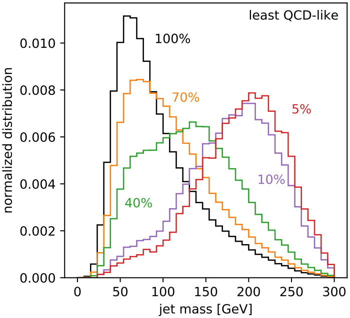
Tops vs. QCD: ROC curve



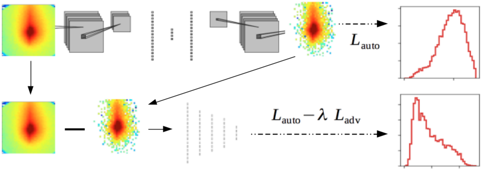
- **AUC**$\sim \mathcal{O}(0.9)$ **without knowing what to look for**
  $\longrightarrow$ AUC$\sim$0.98 for fully supervised
- Constituent approach outperforms images

Jet Mass and the Autoencoder

- Top jets have a much higher jet mas than QCD jets
- The autoencoder is sensitive the jet mass.
- It is learning typical signal v background features.
- It is not necessary to use ML tools just for this.

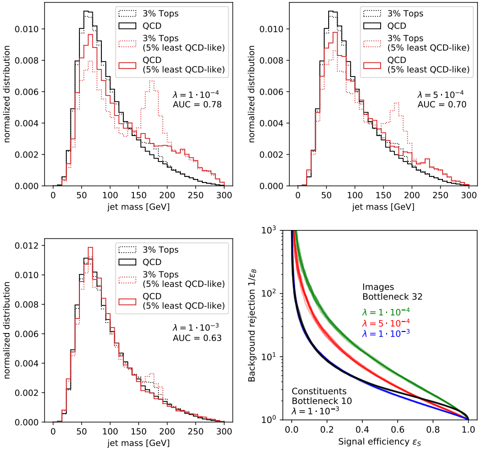


What Else Does the Network Learn?



- We want to stop the network from learning the jet mass.
- Adversarial training:
  $\longrightarrow$ adversary (lower) predicts the jet mass from the autoencoder output.
- Need to balance learning rates/relative contributions to total loss.
  $\longrightarrow$ Best parameter choice depends on QCD p_T slice.
  $\longrightarrow$ But only dependent on the background.

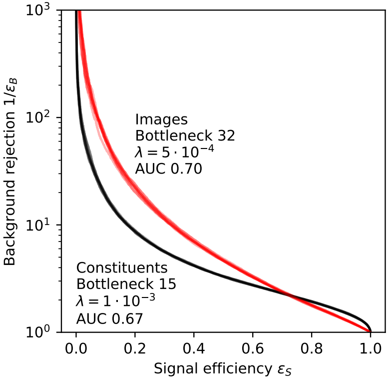
Tops vs. QCD: Adversarial results



$\longrightarrow$ Tradeoff: more mass shaping $\leftrightarrow$ better performance.

$\longrightarrow$ λ is the prefactor to the adversarial loss.

Tops vs. QCD: ROC curves

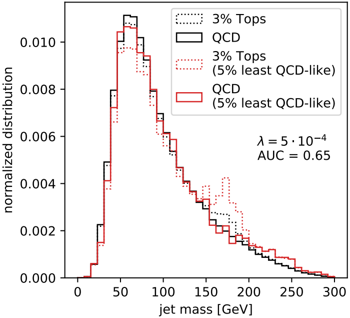


- Still see discrimination power
  $\longrightarrow$ The network learns more than the jet mass.
- Images now outperform constituents
  $\longrightarrow$ CoLa/LoLa approach explictly encodes the mass.
- Move to jet images for the adversarial autoencoder.

Tops vs. QCD: Training on Mixed Samples

⟶ So far we have considered a pure background training region

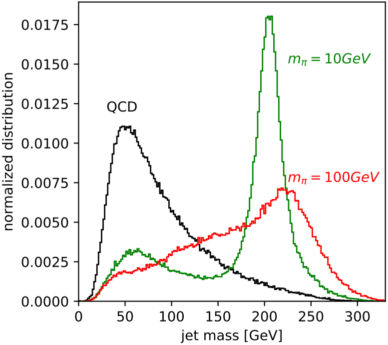
⟶ Now: train on sample with signal+background



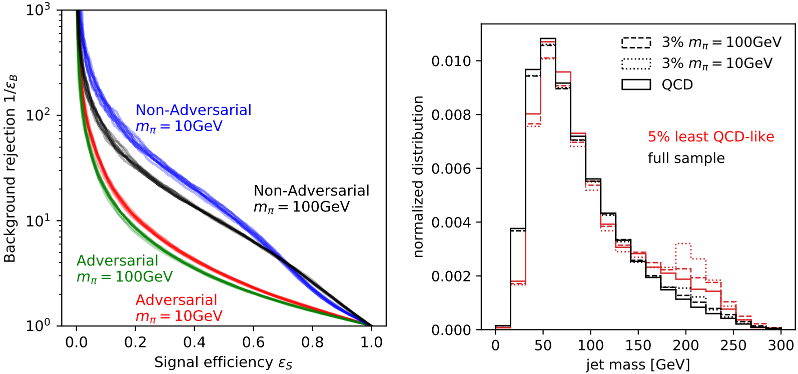
- For background dominated samples, the autoencoder picks out QCD features
- Bottleneck does not have enough information for both tops and QCD
- **Can train and test on same region of phase space**

Dark Showers

- We consider a dark SU(3) symmetry
- 2 points chosen for 200GeV dark quark mass
  $\longrightarrow$ 100GeV dark meson mass
  $\longrightarrow$ 10GeV dark meson mass
- Dark meson can decay to SM via inverted production mechanism



Dark Showers: Adversarial results



$\longrightarrow$ The adversarial autoencoder has discrimination power for a QCD-like signature

Conclusions

https://goo.gl/XGYju3

- Autoencoders are a powerful tool for a generic anomaly search.
- Only a background-dominated signal region is required.
- Adversarial autoencoders can decorrelate the results from an observable
  $\longrightarrow$ possible application in a bump hunt