AI-safety for jet flavour tagging at the CMS experiment

Xavier Coubez, Nikolas Frediani, Spandan Mondal,
Andrzej Novak, Alexander Schmidt and Annika Stein

III. Physikalisches
Institut A

RWTH AACHEN
UNIVERSITY

17.3.2021 DPG Spring Meeting, Dortmund21

Outline

1. Introduction to AI safety (general / jet flavour tagging)
2. Adversarial attacks and how they influence the model performance
    1. Adding Gaussian noise
    2. Applying the Fast Gradient Sign Method (FGSM)

Introduction to AI safety

AI safety: example for image classification

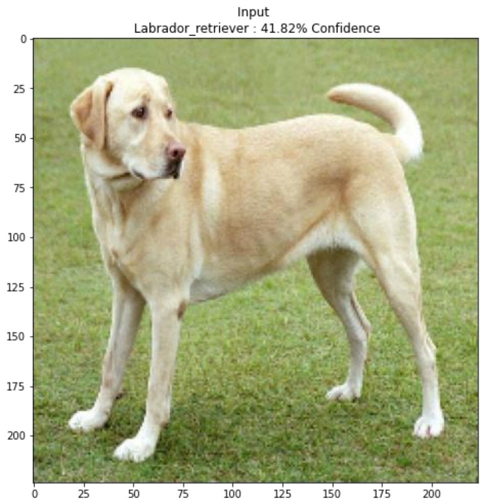


Fig. 1

+ $\epsilon \times$

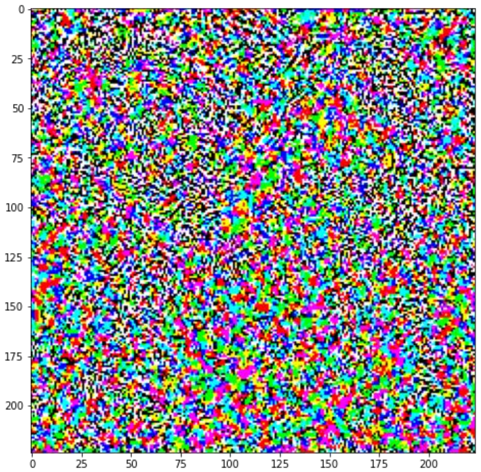


Fig. 2

→

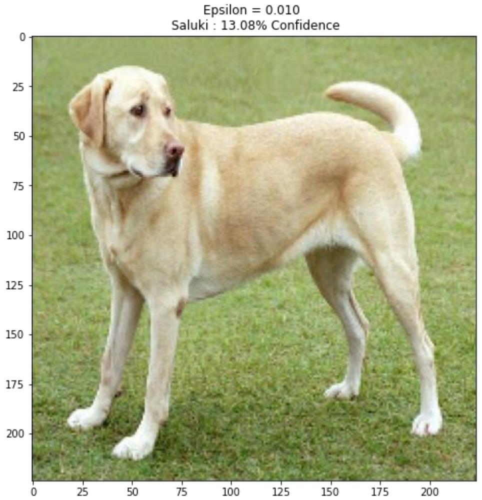


Fig. 3

Classifier: labrador (breed of dog)

Classifier: saluki (breed of dog)

german: „Windhund“

→ Generate adversarial samples with perturbations that are not too easy to identify
→ Check their influence on the model performance

[1]

Application: jet heavy-flavour tagging at CMS

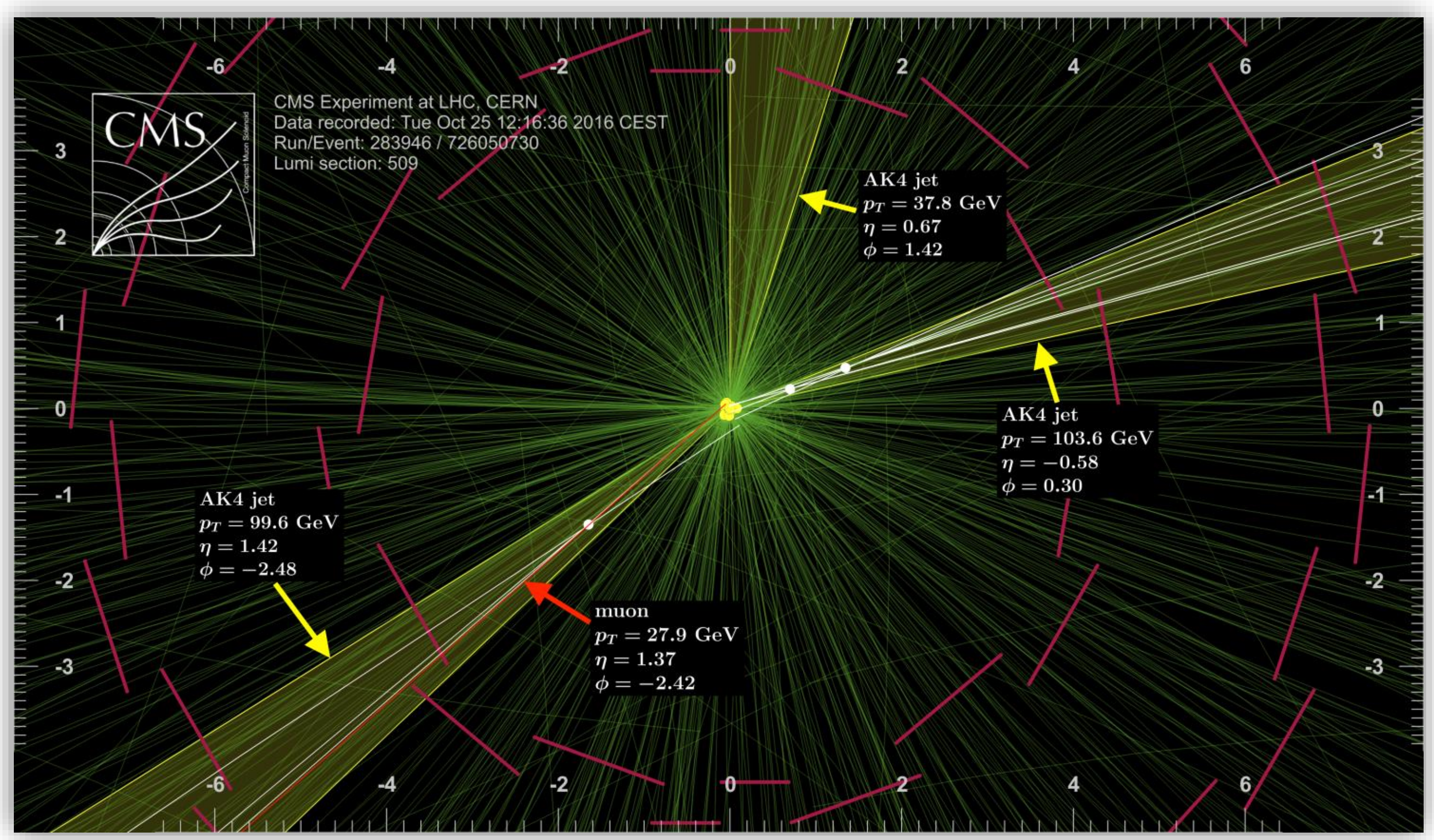


Fig. 4

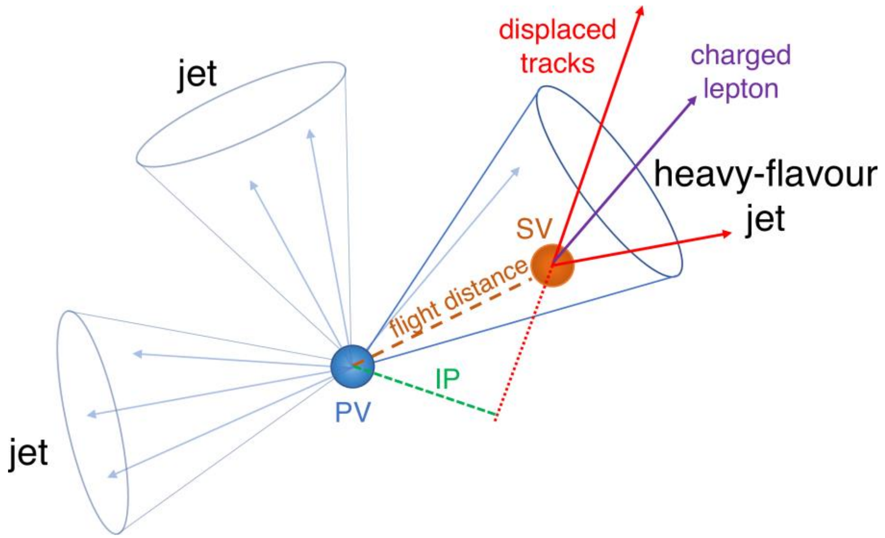


Fig. 5

[2]

AI safety: jet heavy-flavour tagging

Jet,
Track,
Secondary Vertex
properties of a b-jet

Classifier: b-jet

+ $\epsilon \times$

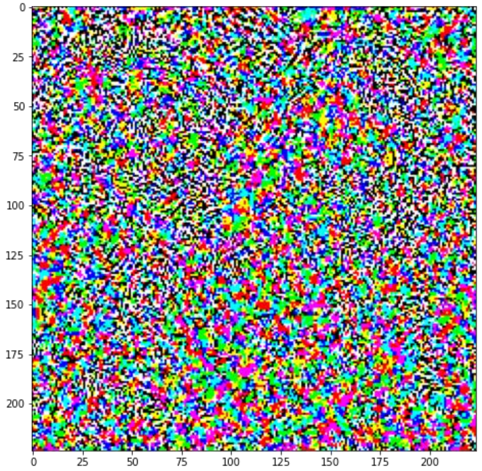


Fig. 6

→

Slightly distorted
Jet,
Track,
Secondary Vertex
properties of a b-jet

Classifier: light-jet

If one pixel alone can fool neural networks [3] for image classification…

…could subtle mismodelings in our simulations cause wrong results in physics analysis?

[4,5]

Inputs

13x Jet

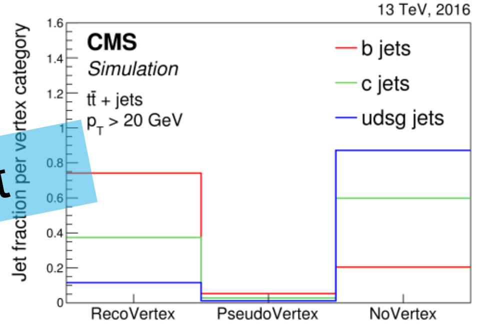


Fig. 7

46x Track
(7x6 + 1x4)

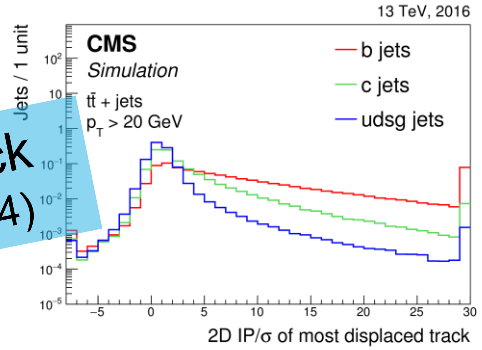


Fig. 8

8x SV

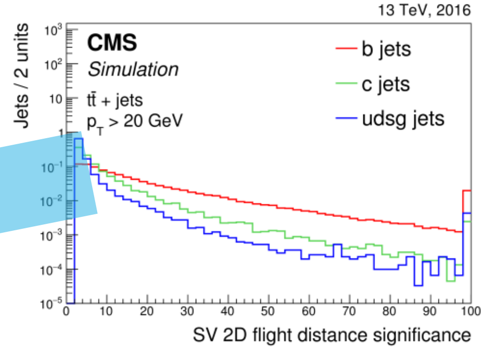


Fig. 9

Deep Neural Network
(DNN)

Outputs

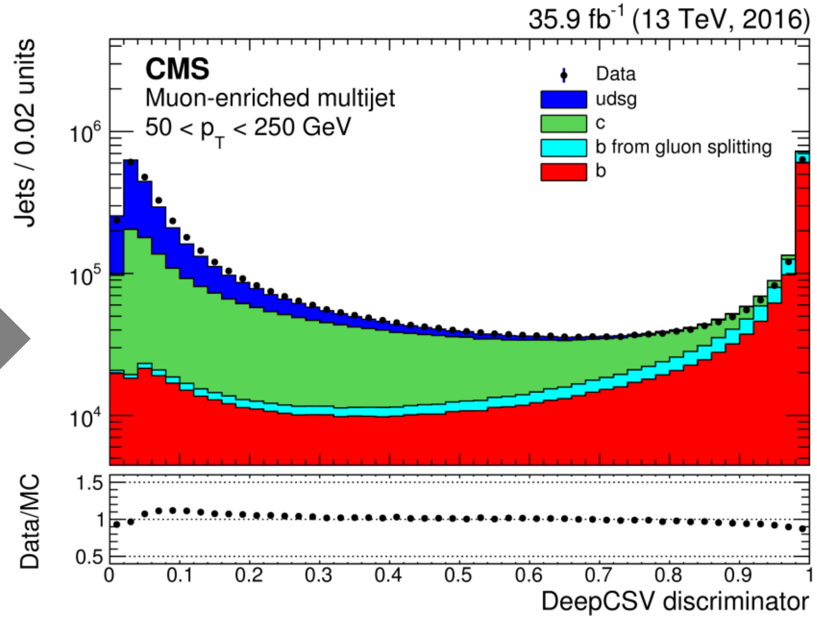
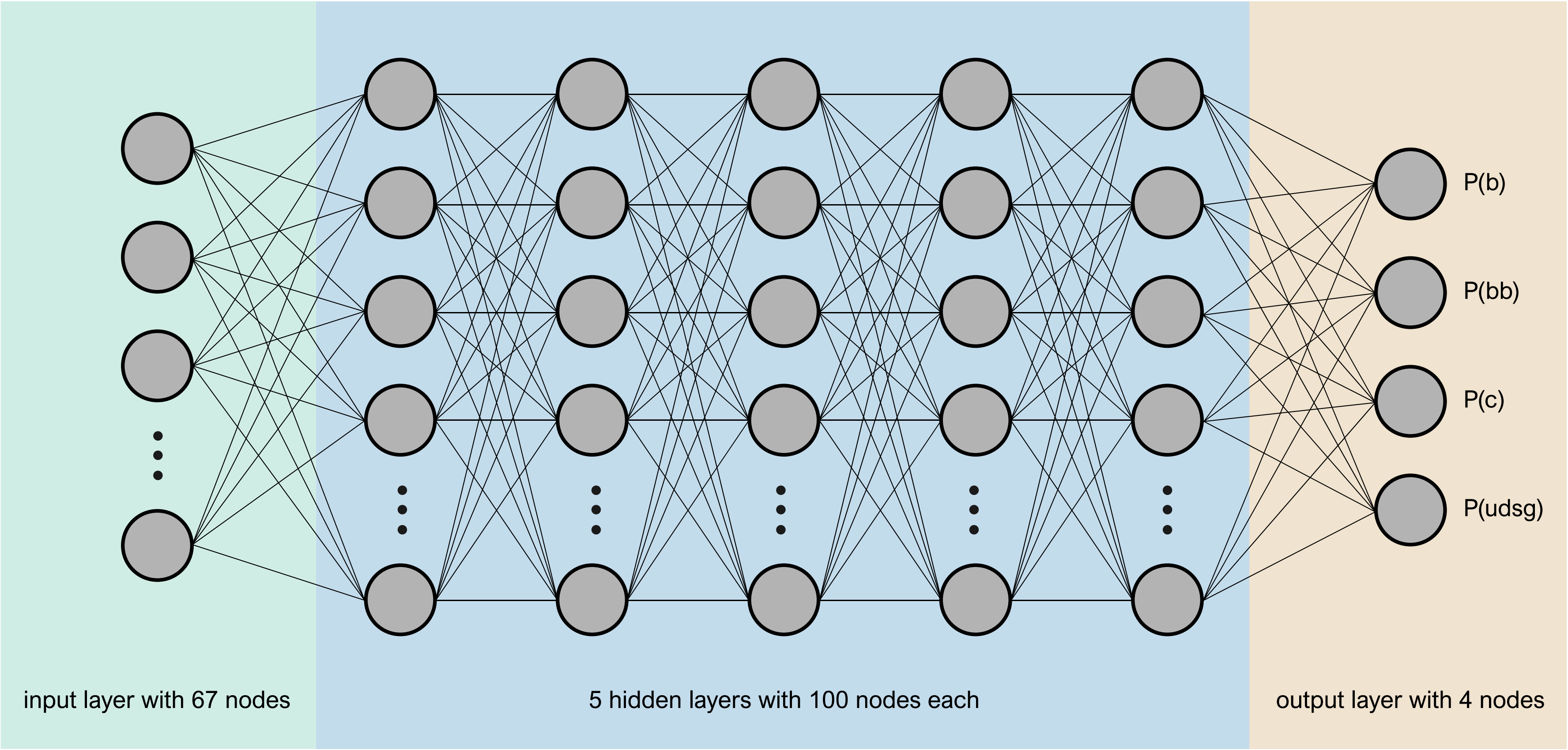


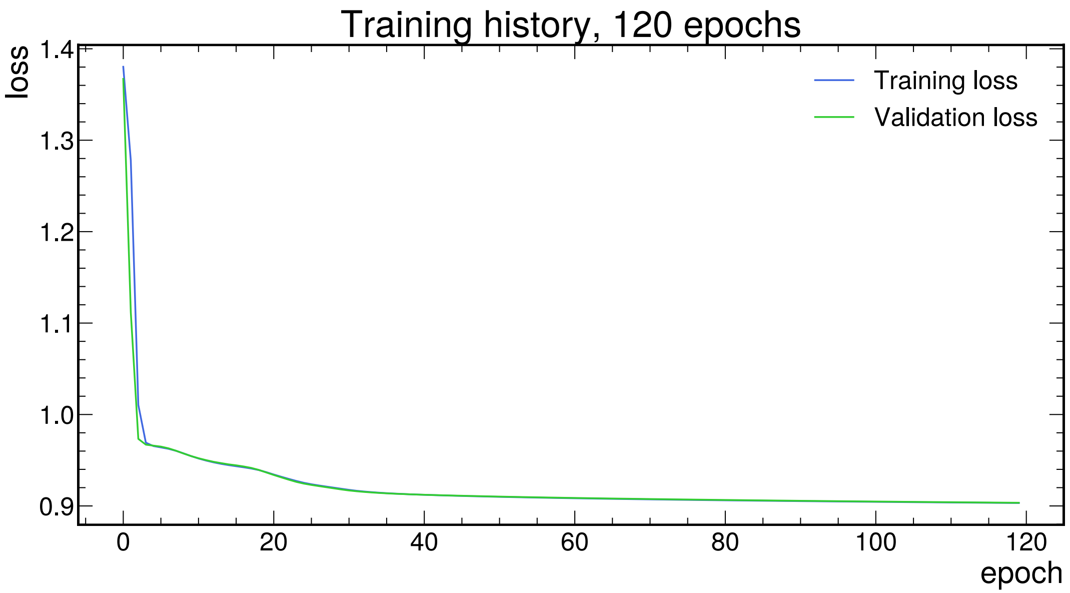
Fig. 10

[2]

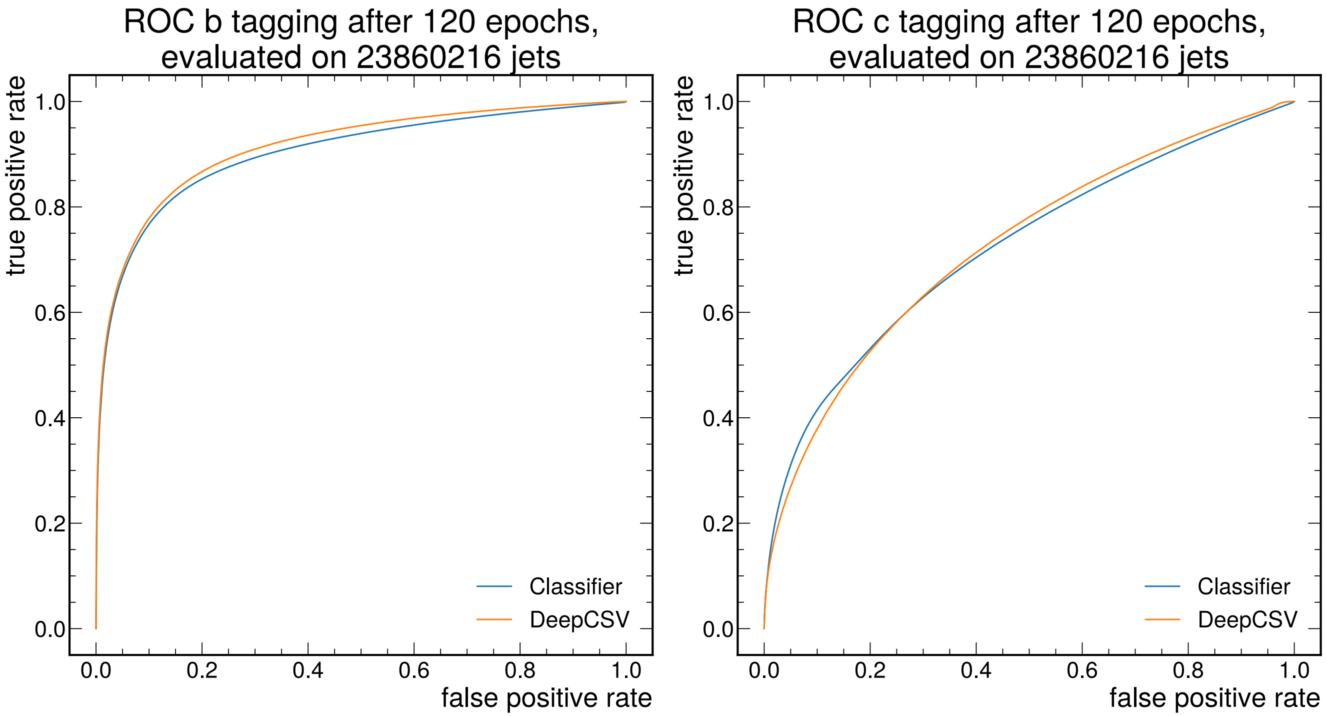
Model architecture


P(b)
P(bb)
P(c)
P(udsg)
input layer with 67 nodes
5 hidden layers with 100 nodes each
output layer with 4 nodes

Testing the model performance



Training and validation loss



Receiver-Operating-Characteristic (ROC) curves

& Area under the ROC curve (AUC)

Adversarial attacks

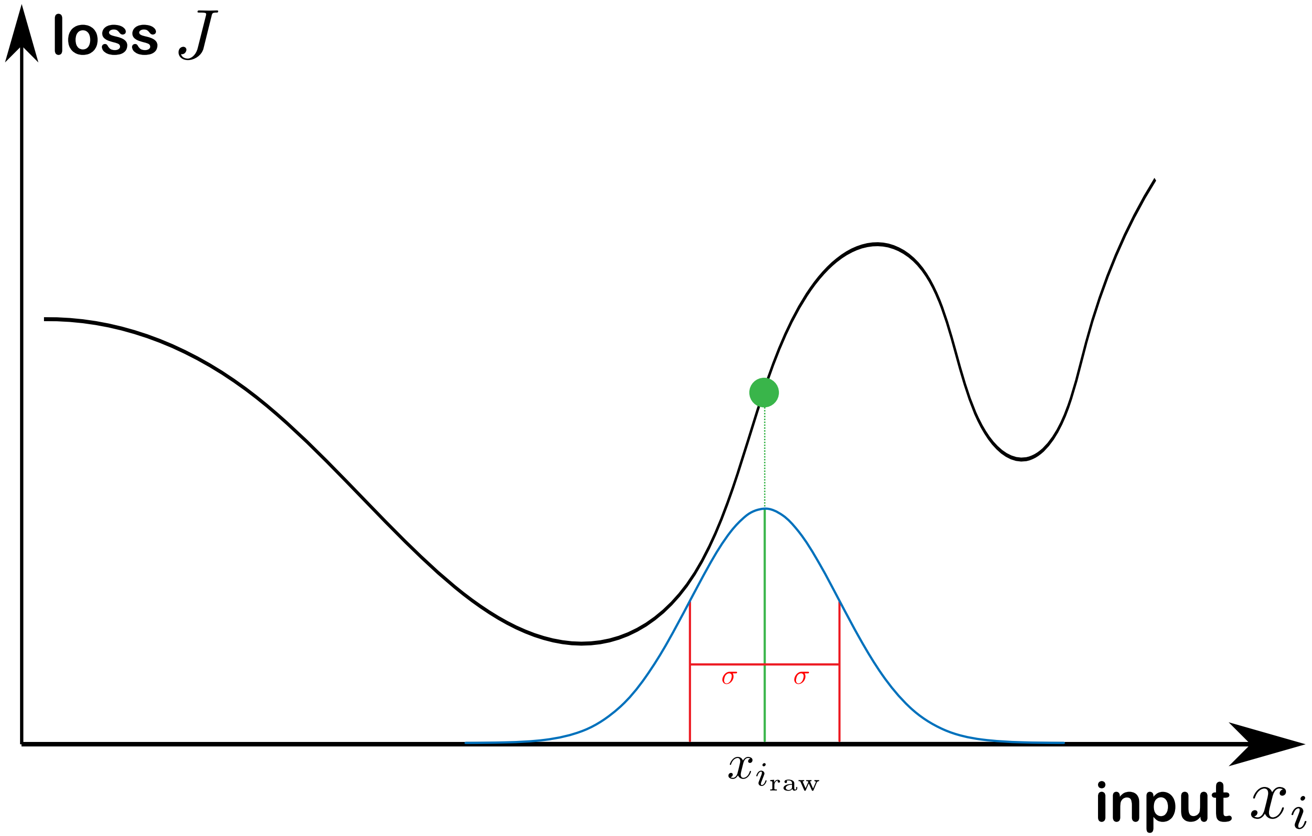
Gaussian noise

Gaussian noise

Adversarial samples x_{noise} are generated by adding a noise term ξ:

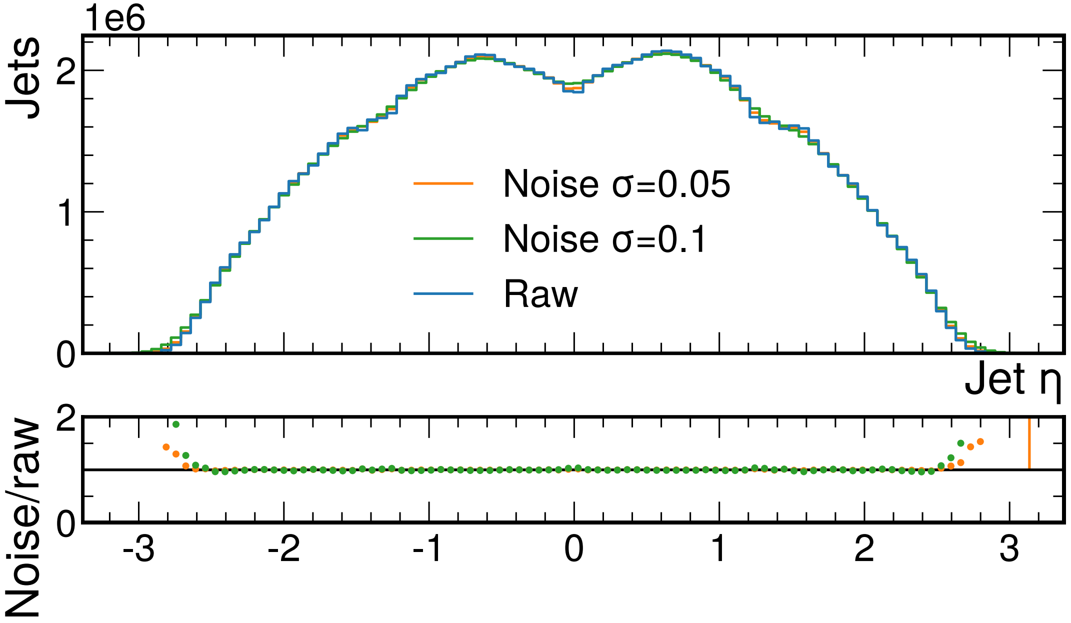
$$x_{noise} = x_{raw} + \xi$$

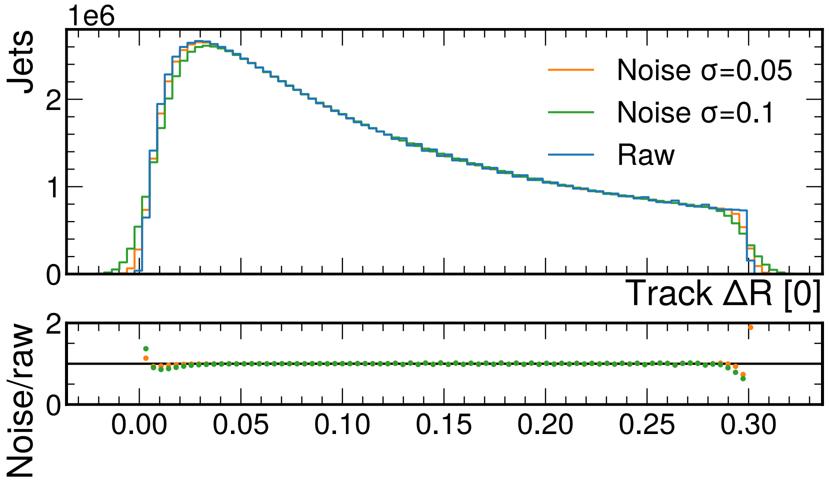
ξ follows a gaussian distribution centred at $\mu = 0$, the standard deviation σ will be varied

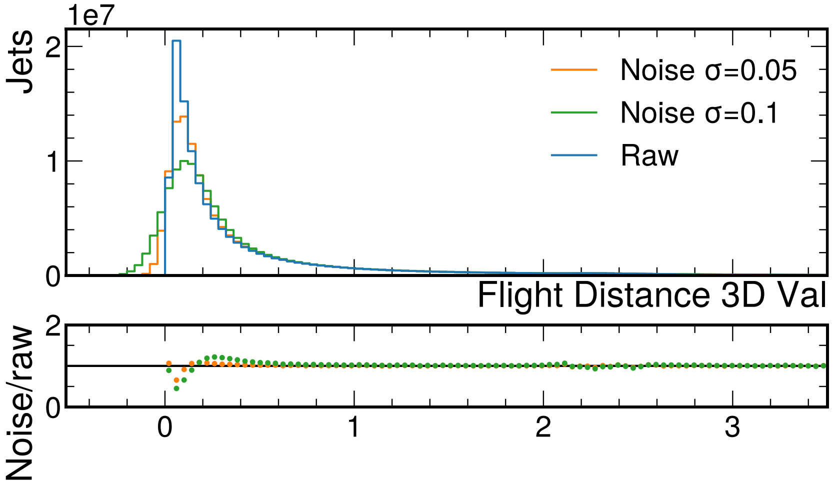


[1,4,6]

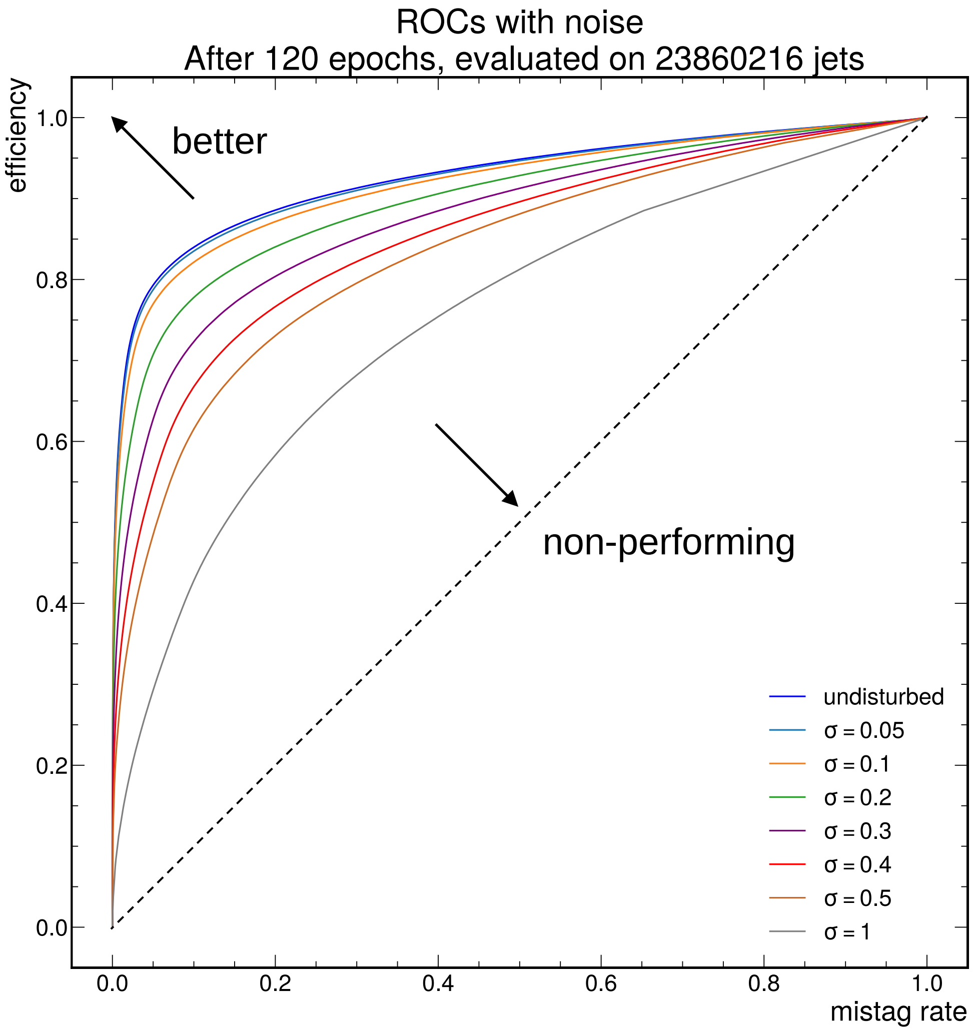
Input shapes

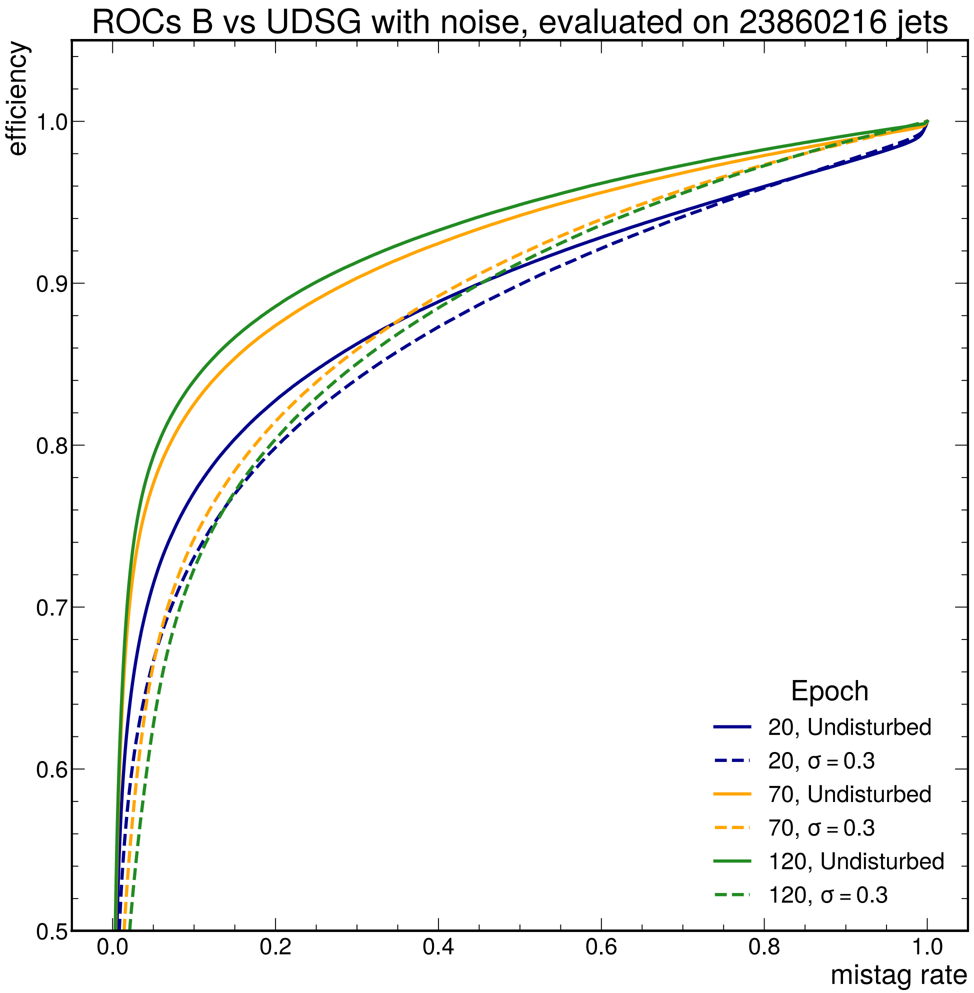




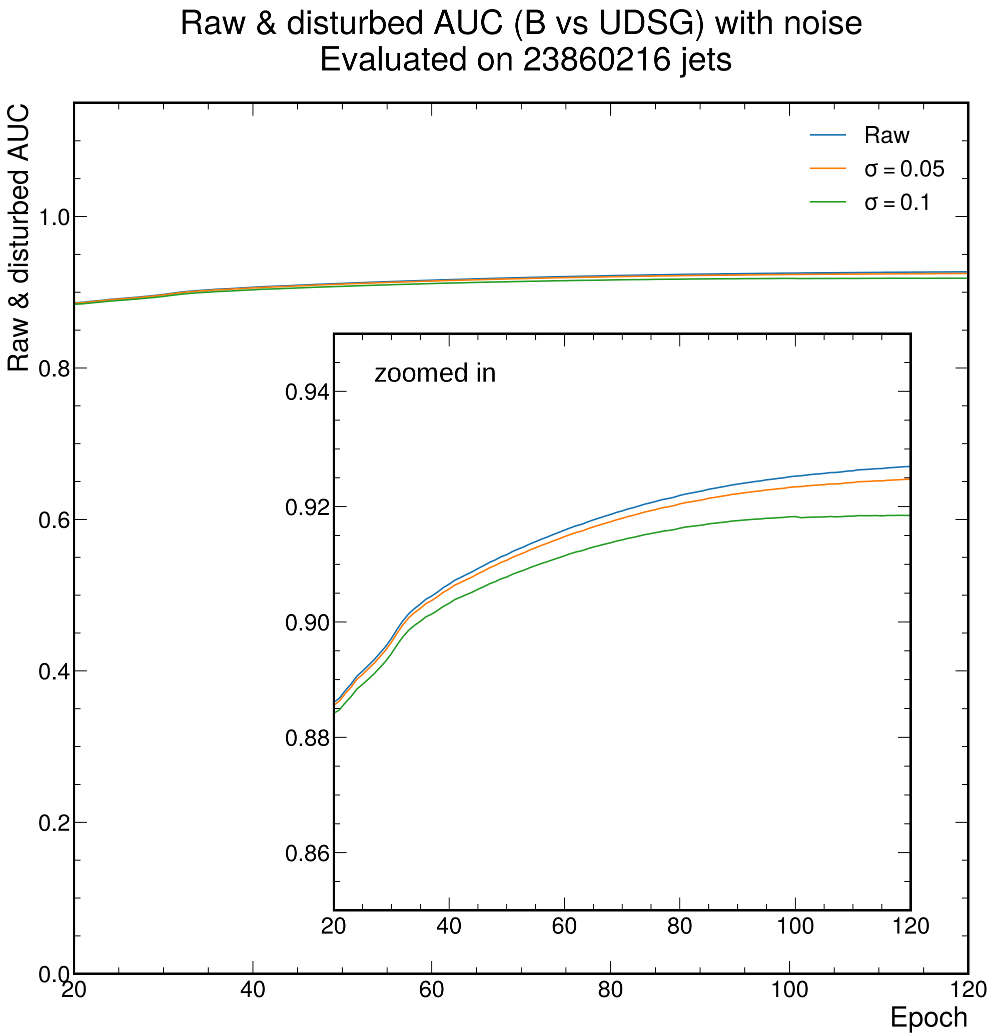


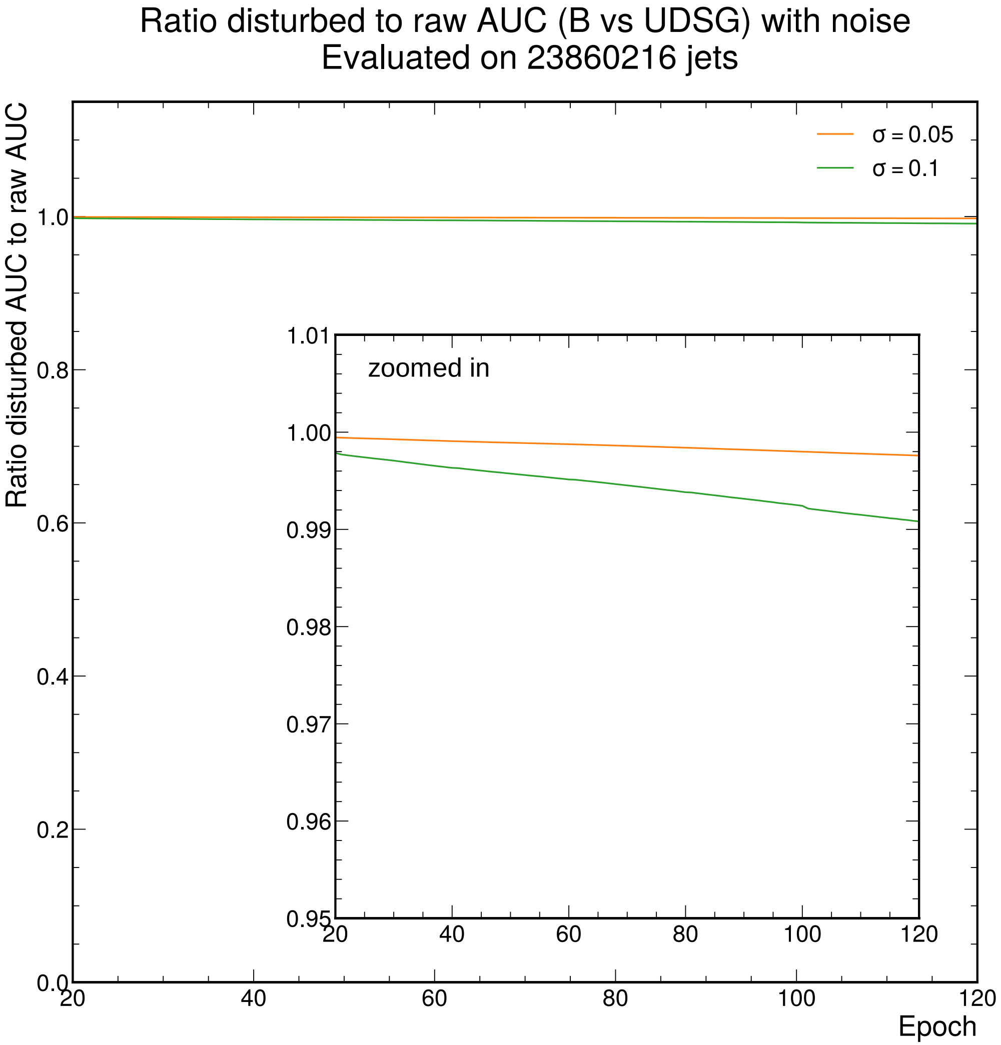
ROC curves (b vs. udsg (light) jets)





Evolution of AUC with number of epochs





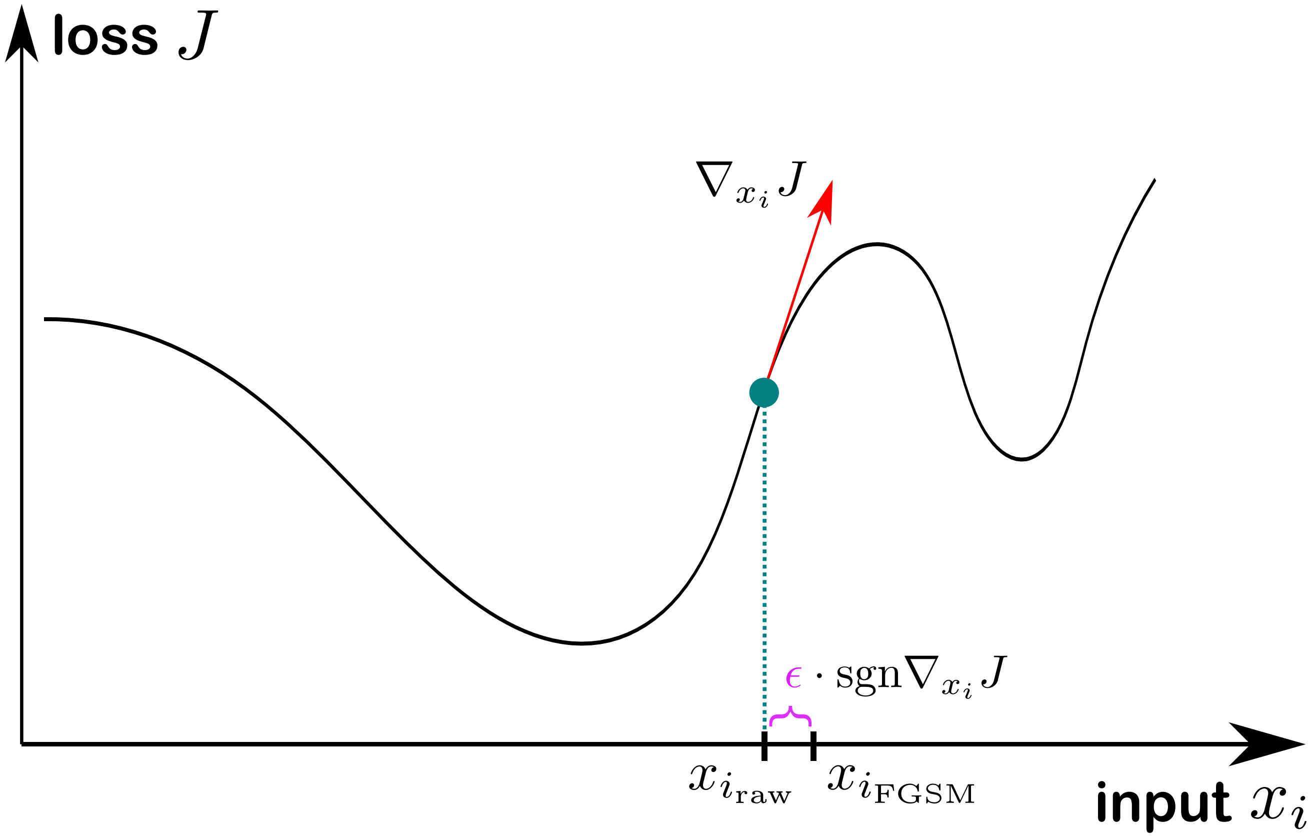
Adversarial attacks

Fast Gradient Sign Method (FGSM)

Fast Gradient Sign Method (FGSM)

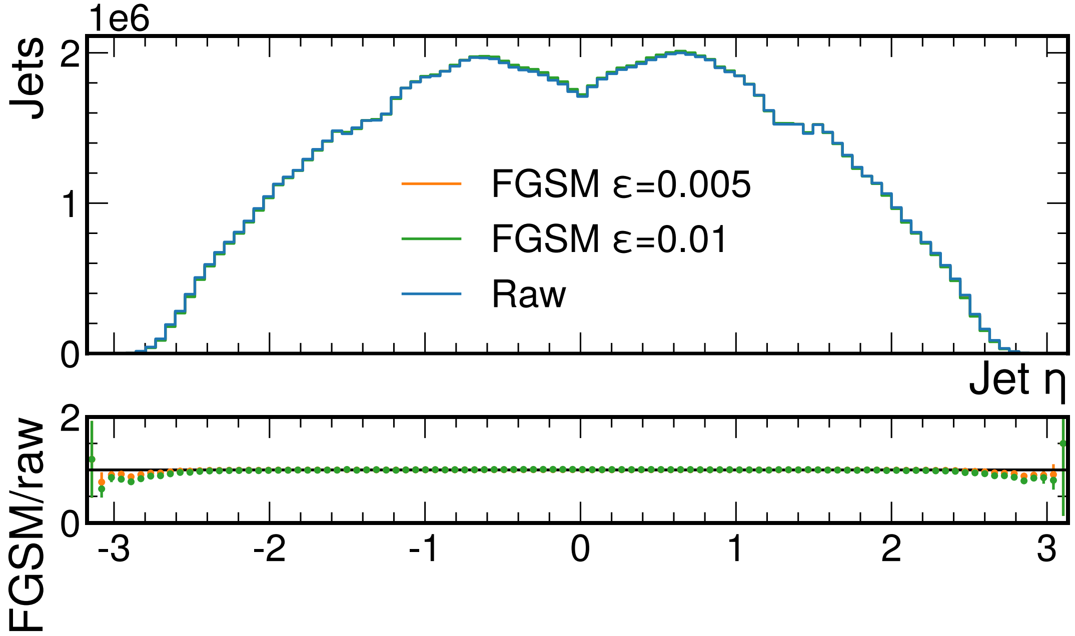
Systematic distortion of the inputs by maximizing the loss function

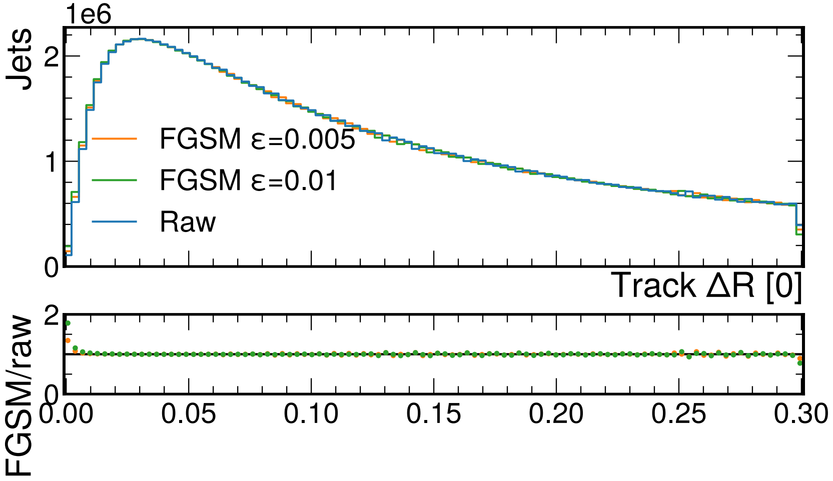
$$x_{FGSM} = x_{raw} + \epsilon \cdot \text{sgn}(\nabla_x J(y, x))$$

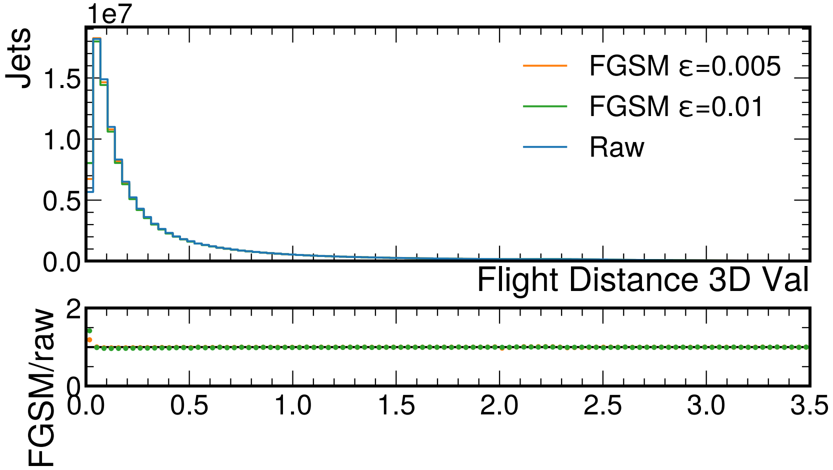


[1,4,5,6]

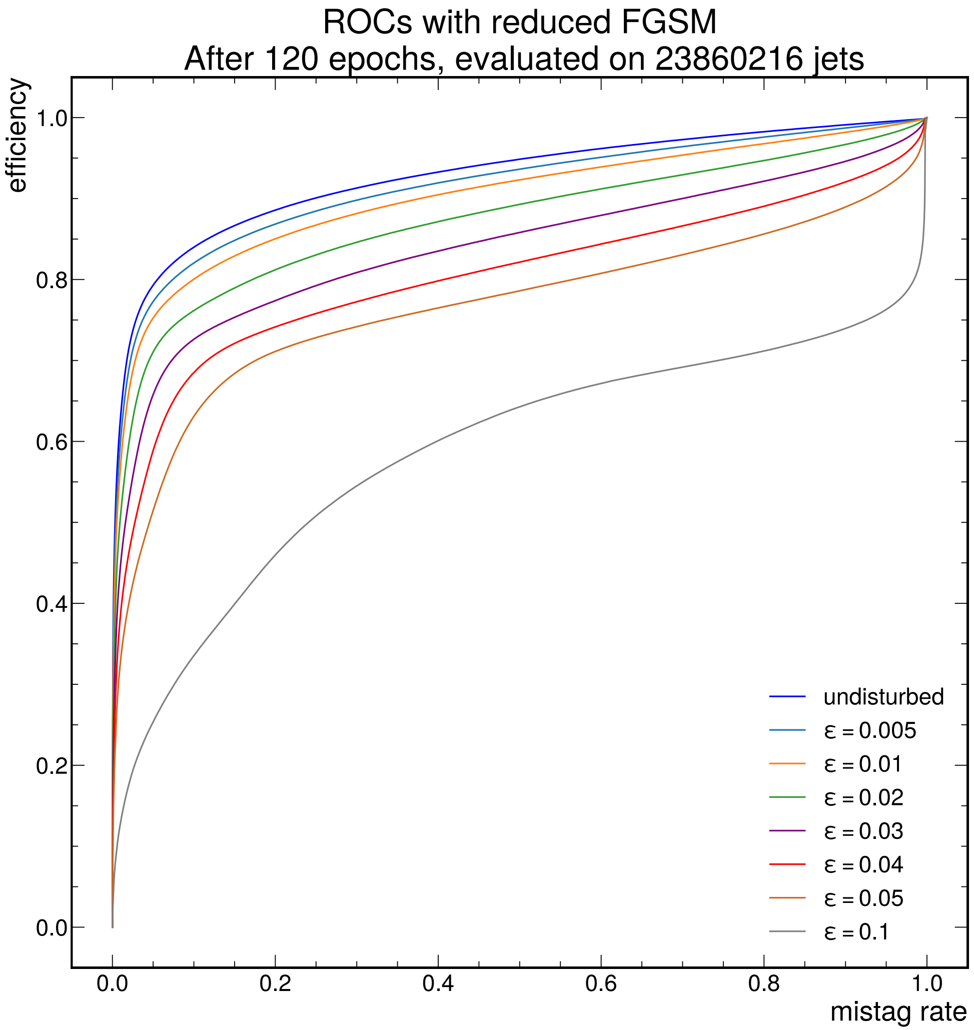
Input shapes

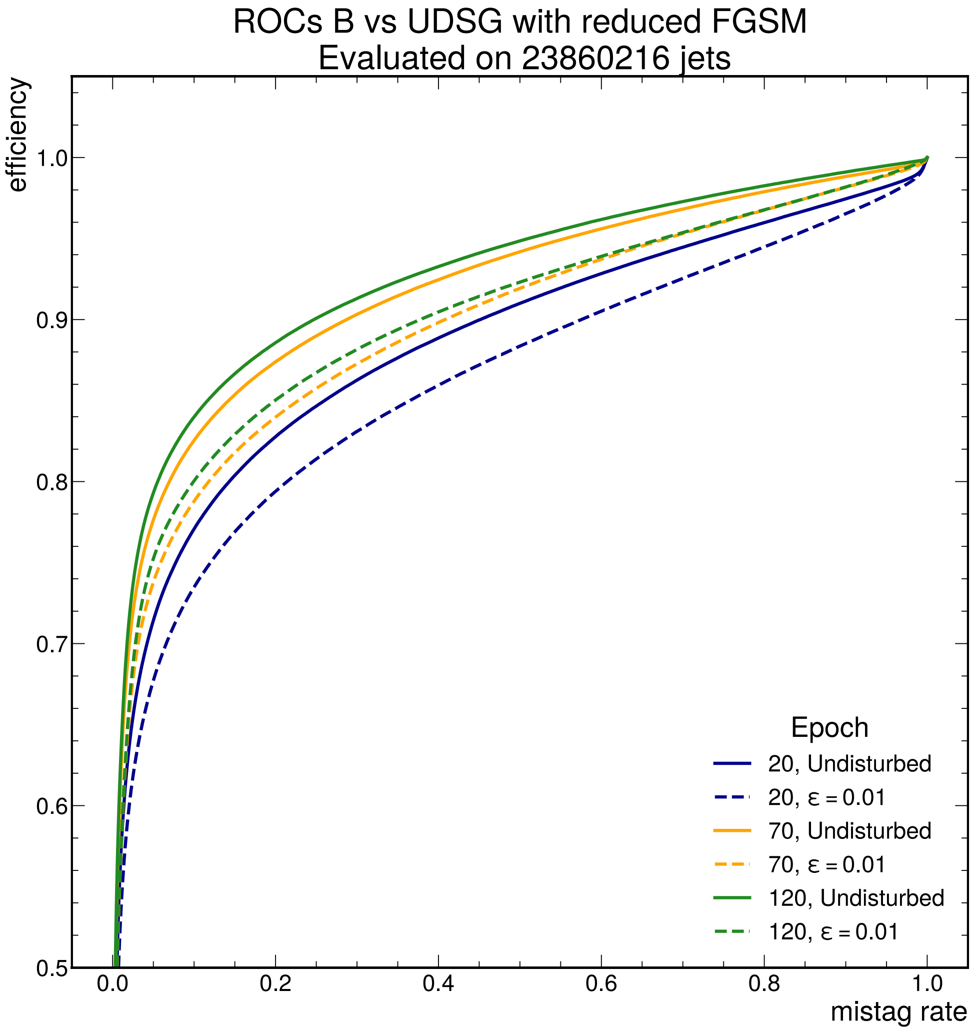




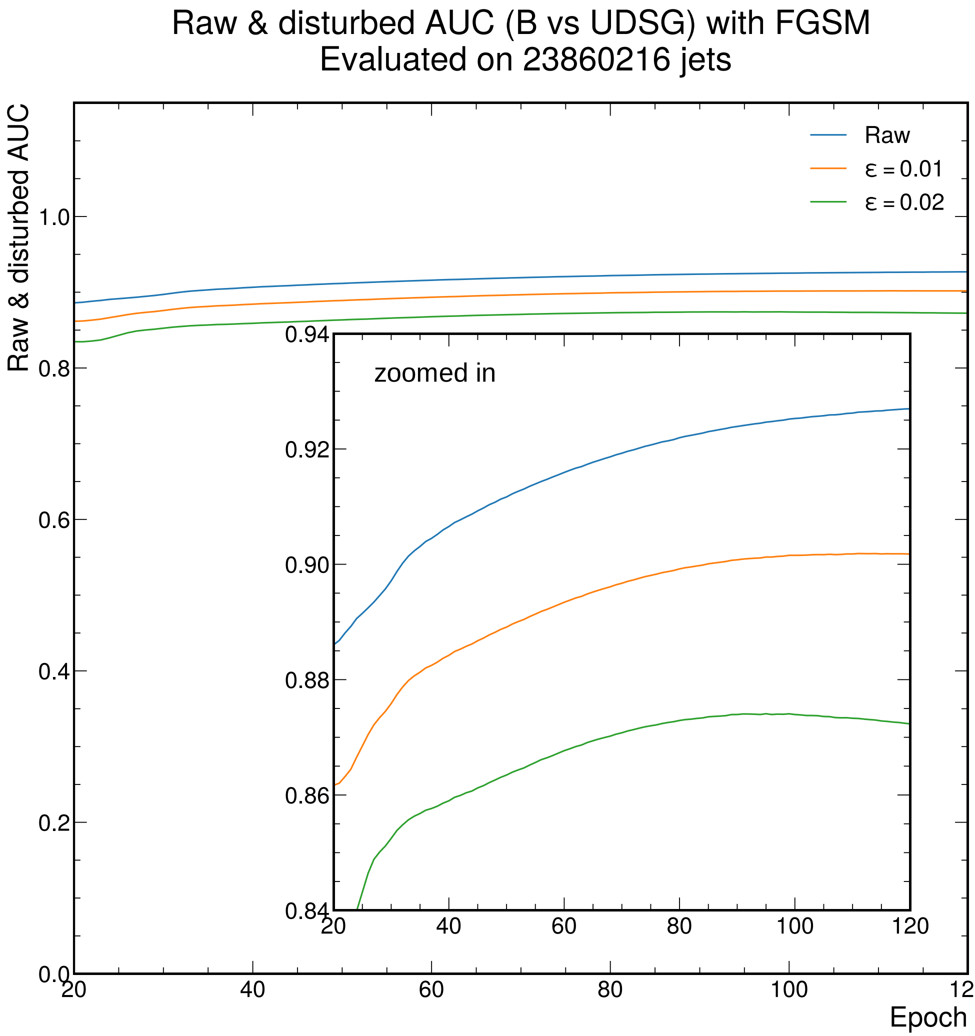


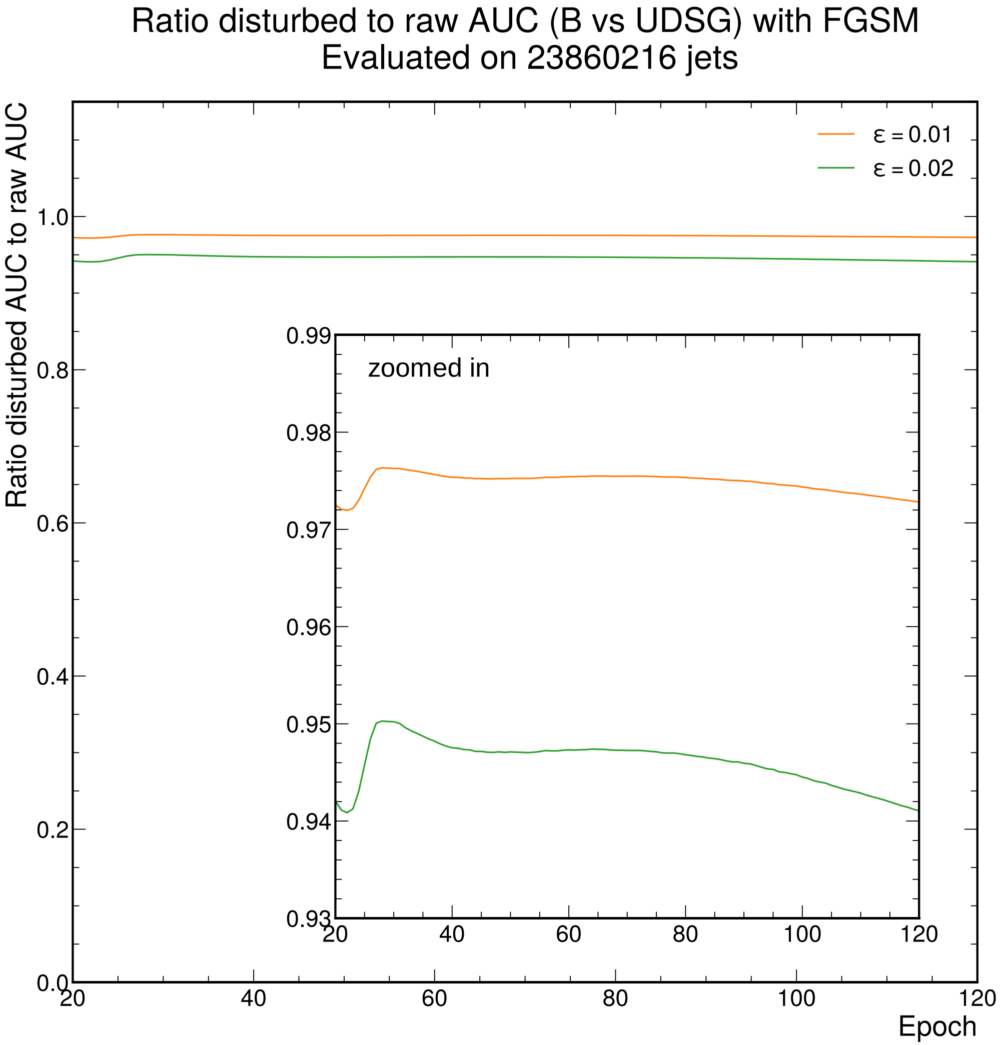
ROC curves (b vs. udsg jets)





Evolution of AUC with number of epochs





Conclusion

- AI safety studies for jet flavour tagging have been done for the first time: almost invisible disturbances of the inputs result in noticable performance drops → applicable & concerning for HEP in general

- Results are consistent with expectations: model performance improves with increasing number of epochs, but susceptibility towards adversarial attacks becomes larger as well

- After studying the impact of adversarial attacks on input shapes and performance, next steps could be:
  - Investigating the influence on the scale factors
  - Improving the resistance of the model against adversarial attacks (e.g. Adversarial Training [7])
  - Applying other attacks of higher complexity

References

1) I. J. Goodfellow, J. Shlens and C. Szegedy, *Explaining and Harnessing Adversarial Examples*, ICLR, (2015), arXiv:1412.6572.

2) The CMS Collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, JINST **13** P05011, (2018), arXiv:1712.07158.

3) Jiawei Su, Danilo Vasconcellos Vargas and Sakurai Kouichi, *One pixel attack for fooling deep neural networks*, (2017), arXiv:1710.08864

4) B. Nachman and C. Shimmin, *AI Safety for High Energy Physics*, (2019), arXiv:1910.08606.

5) C. Shimmin, *ipython notebooks for my MLHEP 2020 tutorial on adversarial attacks on jets*, MLHEP, (2020), https://github.com/cshimmin/advjets-mlhep2020 (last accessed: 01.03.2021)

6) N. Frediani, *First studies in AI-safety for jet flavour tagging at the CMS experiment,* Bachelor thesis, (2020).

7) Anirban Chakraborty, Manaar Alam, Vishal Dey et. al., *Adversarial Attacks and Defences: A Survey*, (2018), arXiv:1810.00069.

Images

Fig. 1, 2, 3 & 6. Reproduced from work created and shared by Google and used according to terms described in the Creative Commons 4.0 Attribution License. (https://www.tensorflow.org/tutorials/generative/adversarial_fgsm). Labrador Retriever by Mirko CC-BY-SA 3.0 from Wikimedia Commons.

Fig. 4. © CERN, 2017, for the benefit of the CMS Collaboration. (https://cds.cern.ch/record/2280025/?ln=en)

Fig. 5, 7, 8, 9 & 10. © CERN, 2018, for the benefit of the CMS Collaboration. (Paper: arXiv:1712.07158, Figures: http://cms-results.web.cern.ch/cms-results/public-results/publications/BTV-16-002/)

Other figures: own work

Backup

Variable names

Inputs:

```
'Jet_eta','Jet_pt’,
'Jet_DeepCSV_flightDistance2dSig’,'Jet_DeepCSV_flightDistance2dVal','Jet_DeepCSV_flightDistance3dSig','Jet_DeepCSV_flightDistance3dVal’,
'Jet_DeepCSV_trackDecayLenVal_0', 'Jet_DeepCSV_trackDecayLenVal_1','Jet_DeepCSV_trackDecayLenVal_2','Jet_DeepCSV_trackDecayLenVal_3','Jet_DeepCSV_trackDecayLenVal_4’,'Jet_DeepCSV_trackDecayLenVal_5’,
'Jet_DeepCSV_trackDeltaR_0','Jet_DeepCSV_trackDeltaR_1','Jet_DeepCSV_trackDeltaR_2','Jet_DeepCSV_trackDeltaR_3','Jet_DeepCSV_trackDeltaR_4','Jet_DeepCSV_trackDeltaR_5’,
'Jet_DeepCSV_trackEtaRel_0','Jet_DeepCSV_trackEtaRel_1','Jet_DeepCSV_trackEtaRel_2','Jet_DeepCSV_trackEtaRel_3’,
'Jet_DeepCSV_trackJetDistVal_0','Jet_DeepCSV_trackJetDistVal_1','Jet_DeepCSV_trackJetDistVal_2','Jet_DeepCSV_trackJetDistVal_3','Jet_DeepCSV_trackJetDistVal_4','Jet_DeepCSV_trackJetDistVal_5’,
'Jet_DeepCSV_trackJetPt’,
'Jet_DeepCSV_trackPtRatio_0','Jet_DeepCSV_trackPtRatio_1','Jet_DeepCSV_trackPtRatio_2','Jet_DeepCSV_trackPtRatio_3','Jet_DeepCSV_trackPtRatio_4','Jet_DeepCSV_trackPtRatio_5’,
'Jet_DeepCSV_trackPtRel_0','Jet_DeepCSV_trackPtRel_1','Jet_DeepCSV_trackPtRel_2','Jet_DeepCSV_trackPtRel_3','Jet_DeepCSV_trackPtRel_4','Jet_DeepCSV_trackPtRel_5’,
'Jet_DeepCSV_trackSip2dSigAboveCharm’,
'Jet_DeepCSV_trackSip2dSig_0','Jet_DeepCSV_trackSip2dSig_1','Jet_DeepCSV_trackSip2dSig_2','Jet_DeepCSV_trackSip2dSig_3','Jet_DeepCSV_trackSip2dSig_4','Jet_DeepCSV_trackSip2dSig_5’,
'Jet_DeepCSV_trackSip2dValAboveCharm’,
'Jet_DeepCSV_trackSip3dSigAboveCharm’,
'Jet_DeepCSV_trackSip3dSig_0','Jet_DeepCSV_trackSip3dSig_1','Jet_DeepCSV_trackSip3dSig_2','Jet_DeepCSV_trackSip3dSig_3','Jet_DeepCSV_trackSip3dSig_4','Jet_DeepCSV_trackSip3dSig_5’,
'Jet_DeepCSV_trackSip3dValAboveCharm’,
'Jet_DeepCSV_trackSumJetDeltaR','Jet_DeepCSV_trackSumJetEtRatio’,
'Jet_DeepCSV_vertexCategory','Jet_DeepCSV_vertexEnergyRatio','Jet_DeepCSV_vertexJetDeltaR','Jet_DeepCSV_vertexMass’,
'Jet_DeepCSV_jetNSecondaryVertices','Jet_DeepCSV_jetNSelectedTracks','Jet_DeepCSV_jetNTracksEtaRel','Jet_DeepCSV_vertexNTracks’,
```

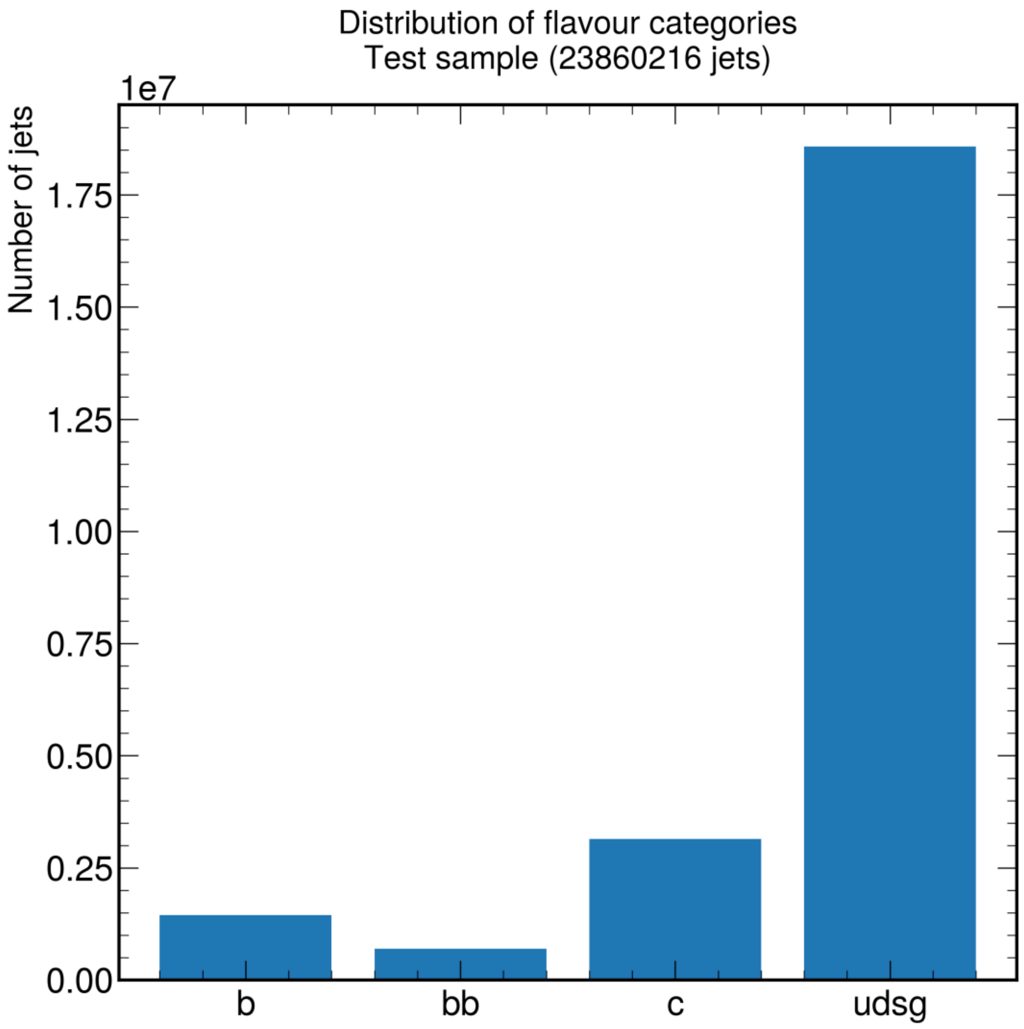
For comparison with DeepCSV:

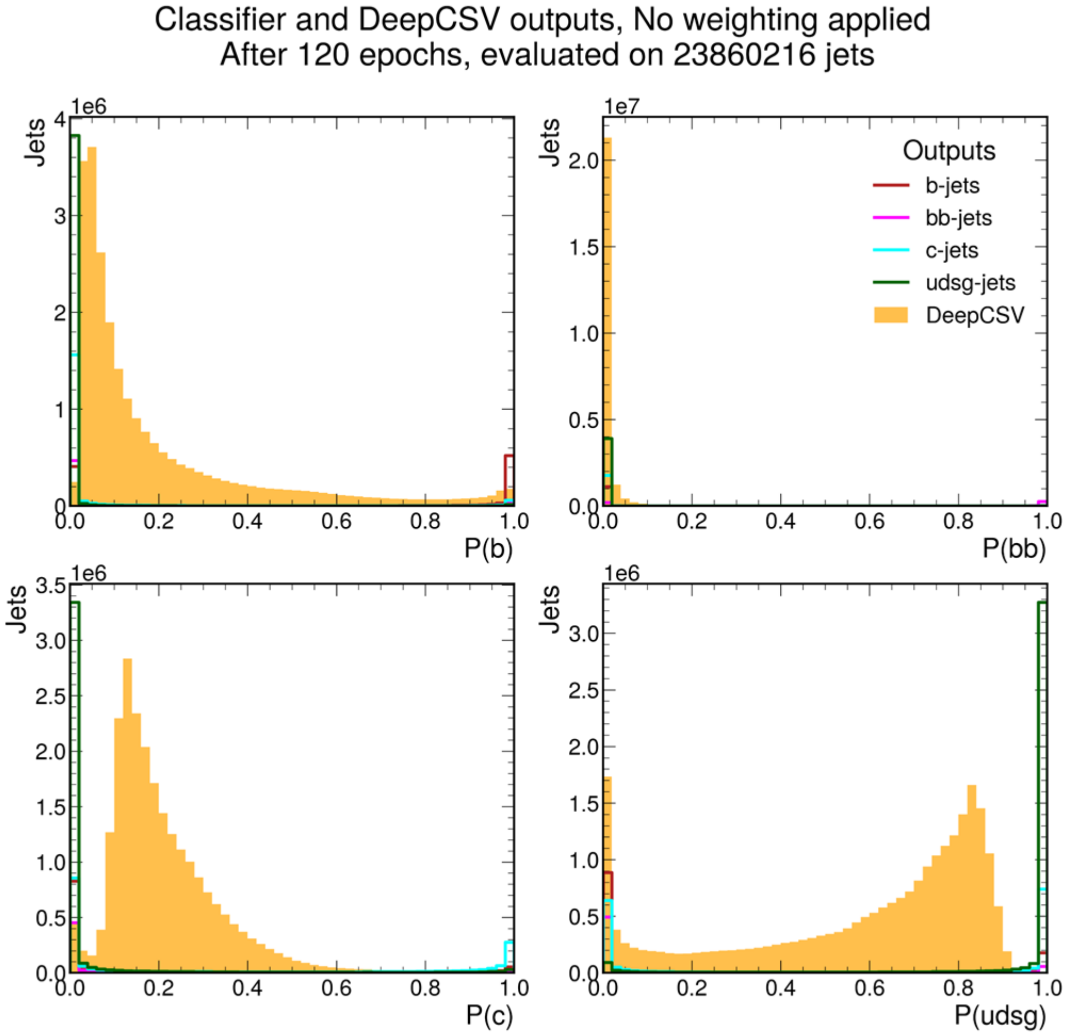
```
'Jet_btagDeepB_b','Jet_btagDeepB_bb','Jet_btagDeepC','Jet_btagDeepL’,
```

Creating the truth outputs was done with:

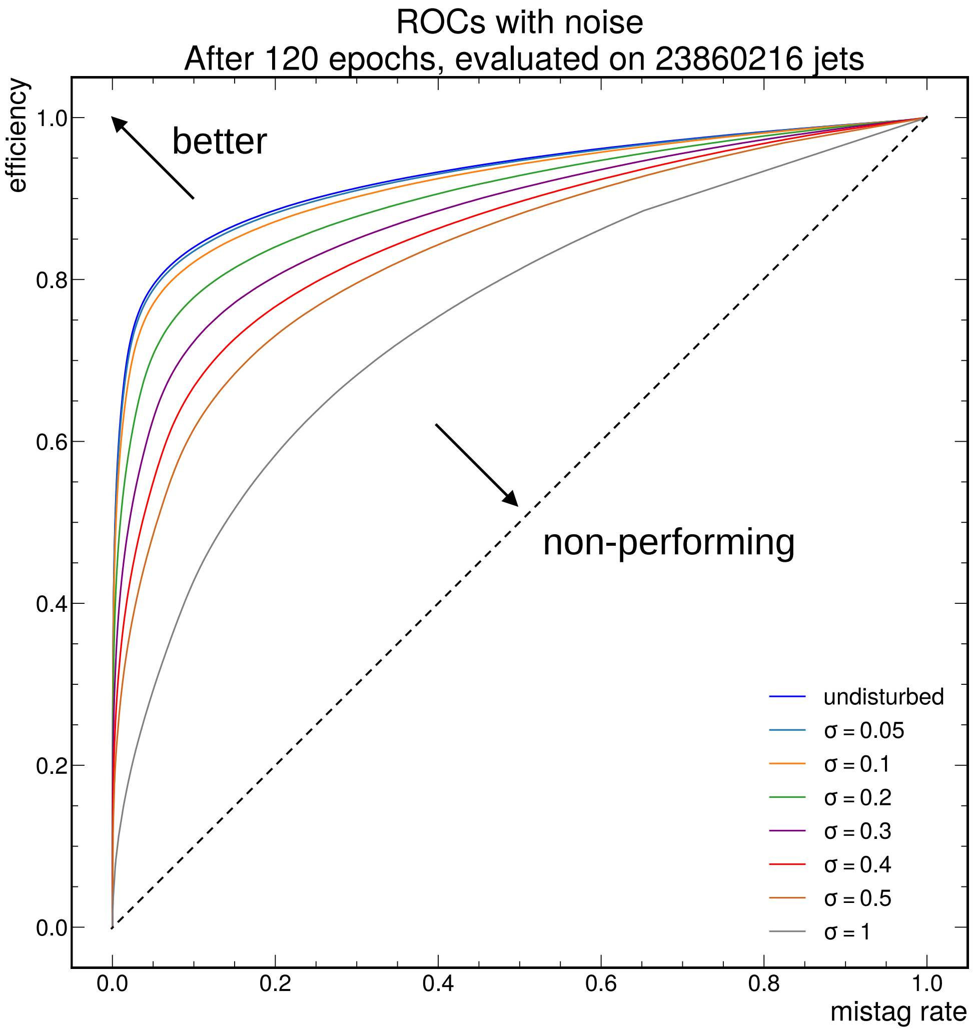
```
'Jet_nBHadrons','Jet_hadronFlavour'
```

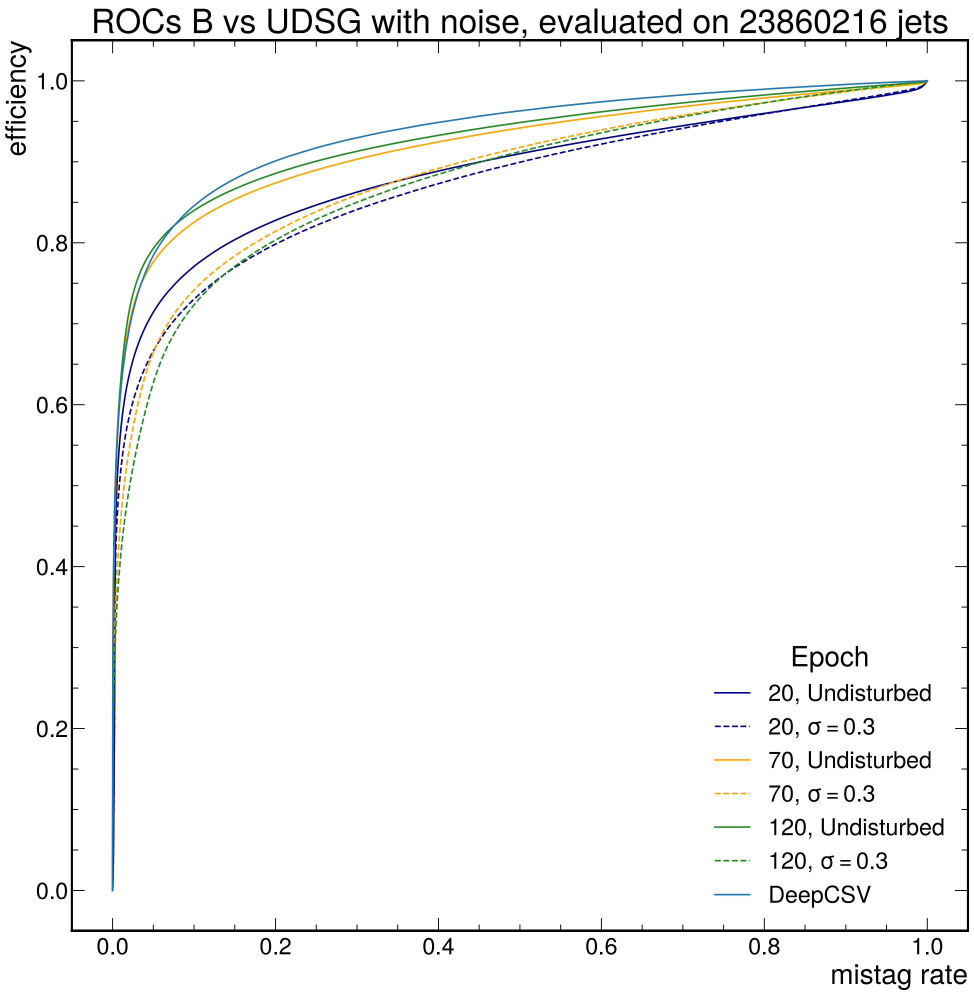
True flavour distribution & Outputs



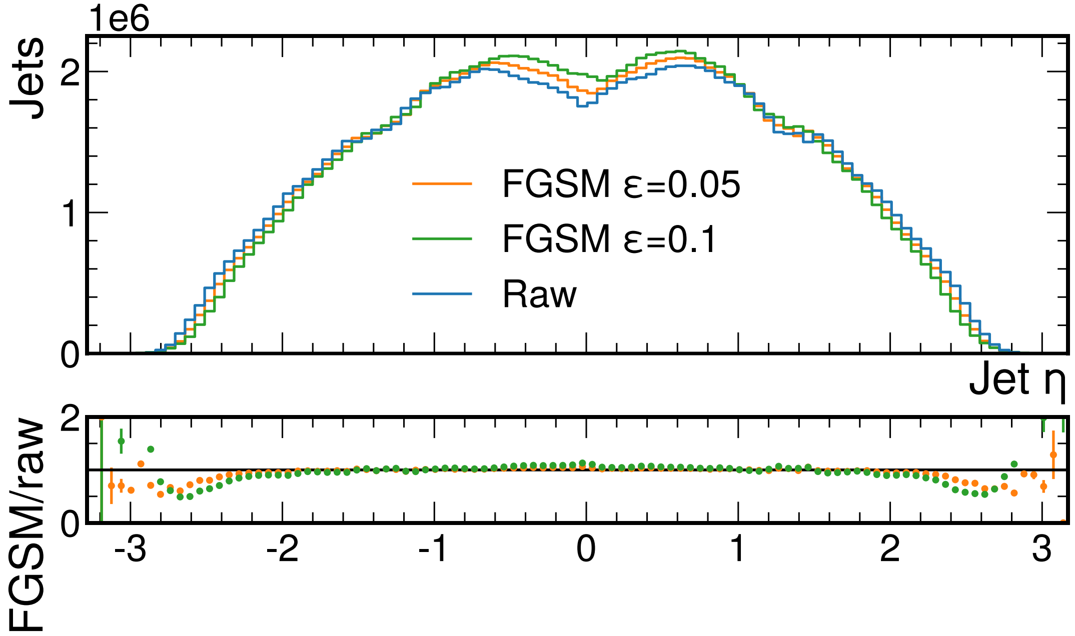


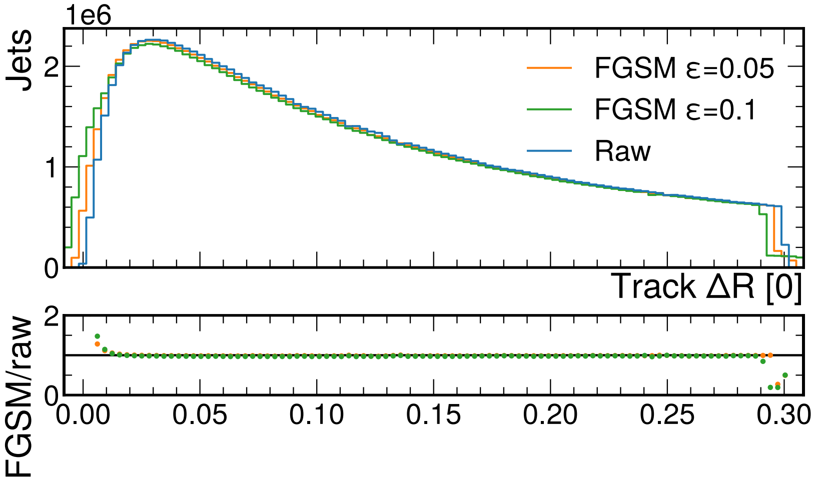
ROC curves (b vs. udsg (light) jets)

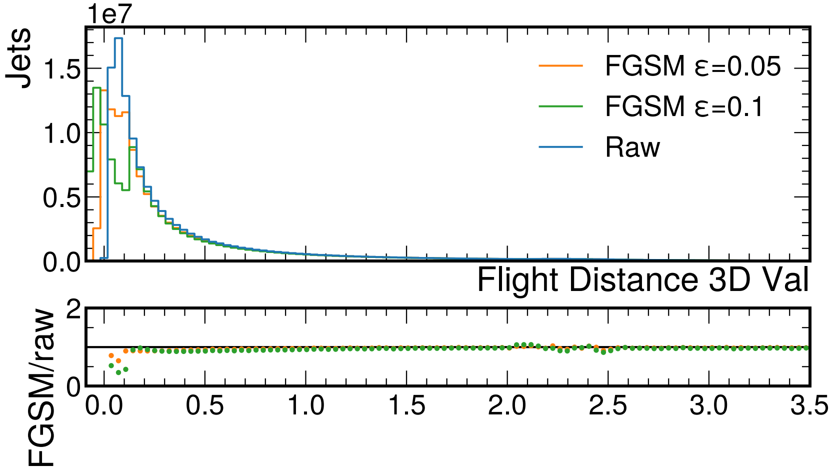




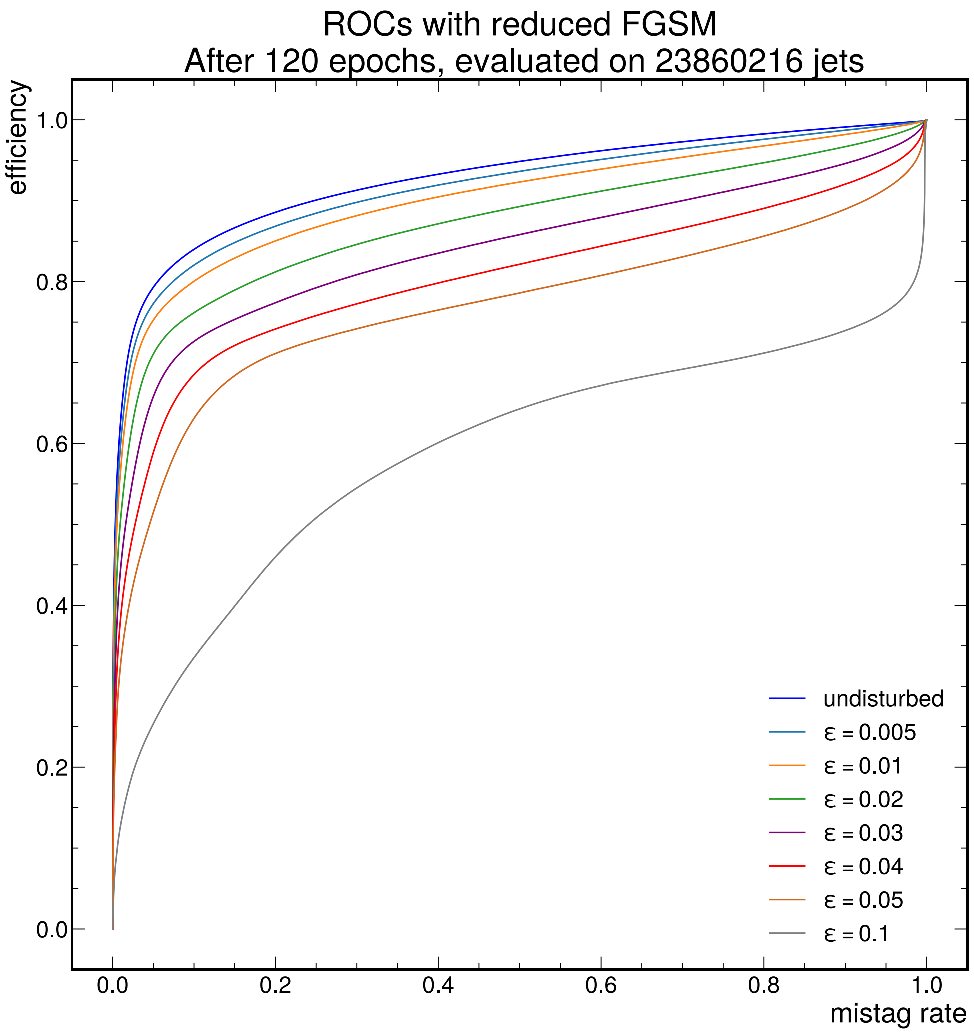
Input shapes (larger ϵ)

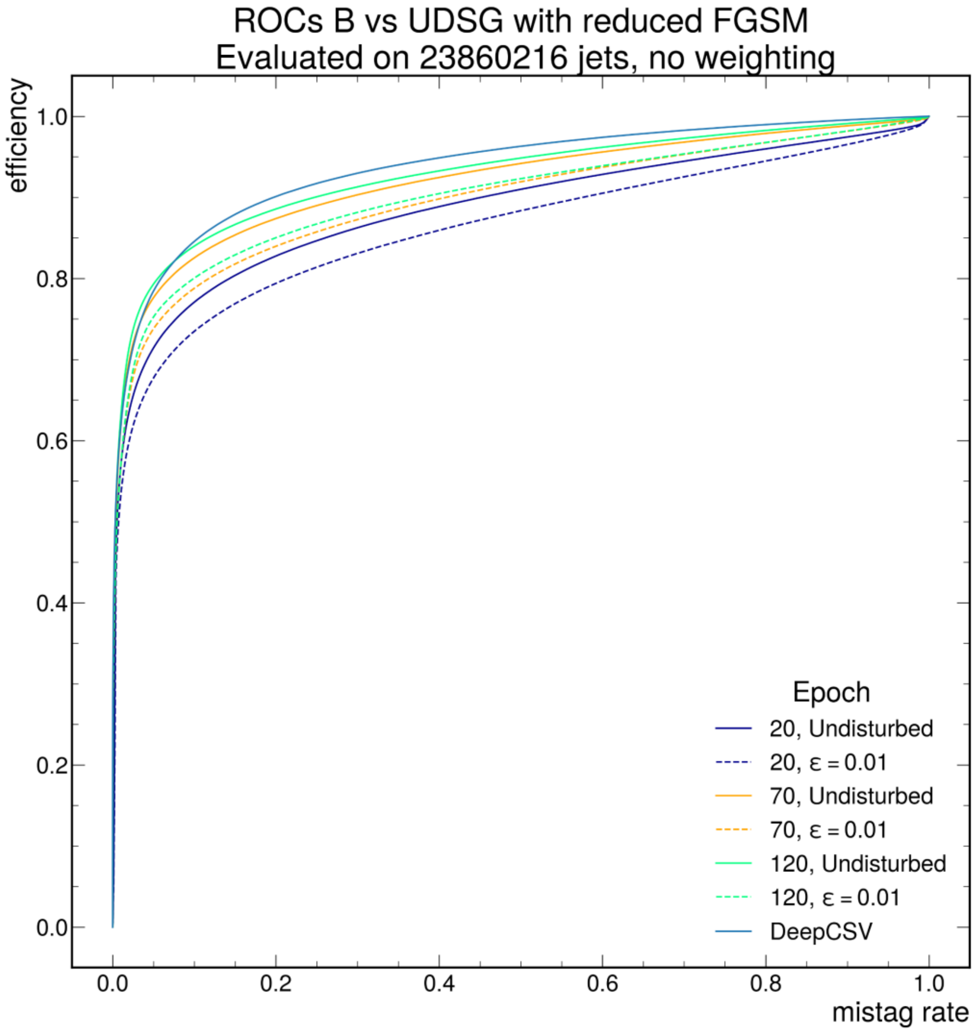




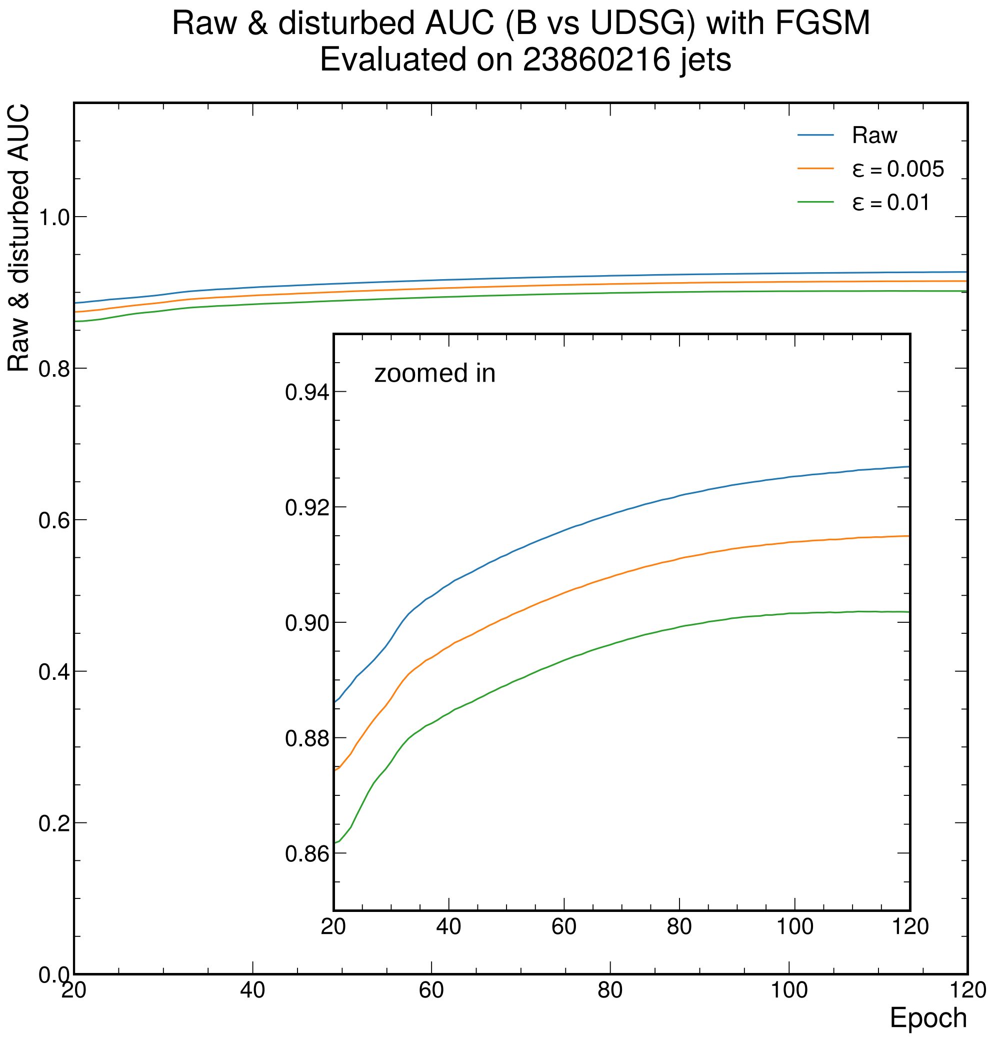


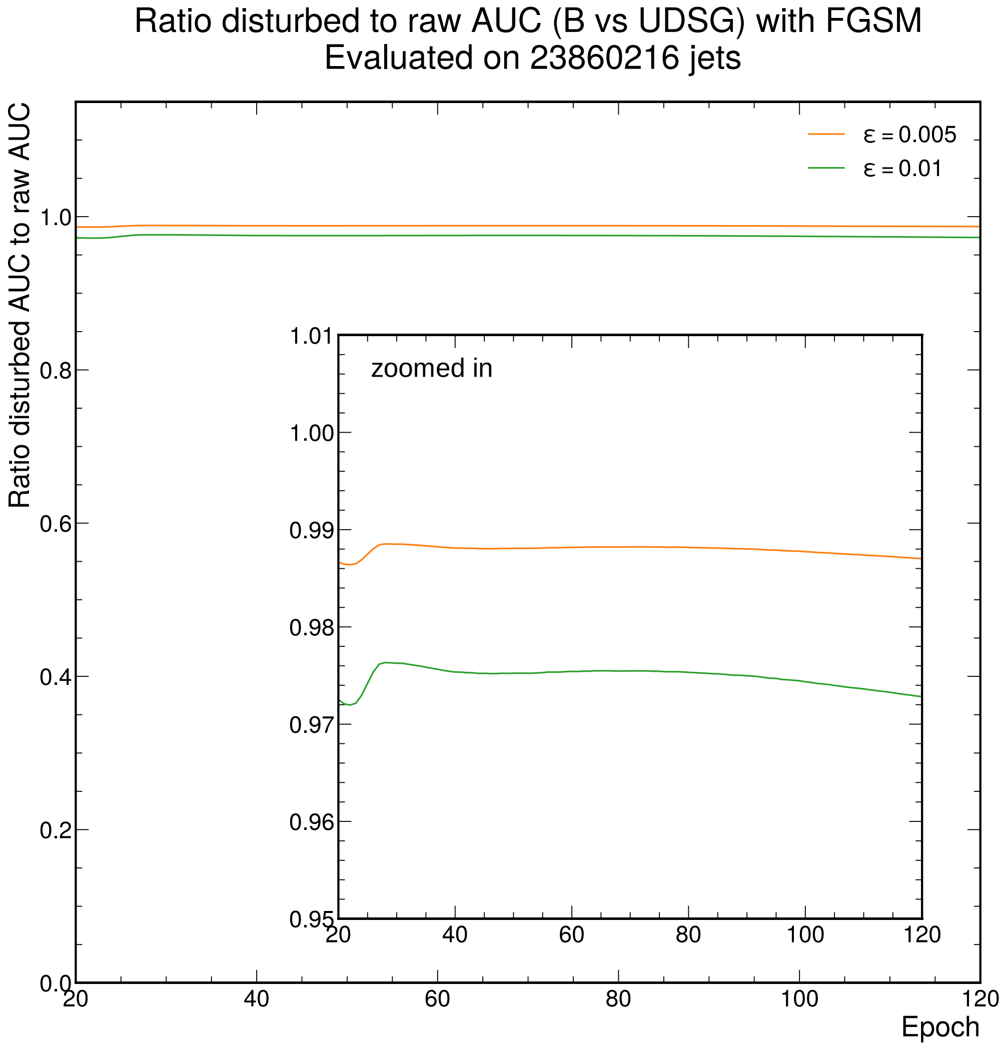
ROC curves (b vs. udsg jets)



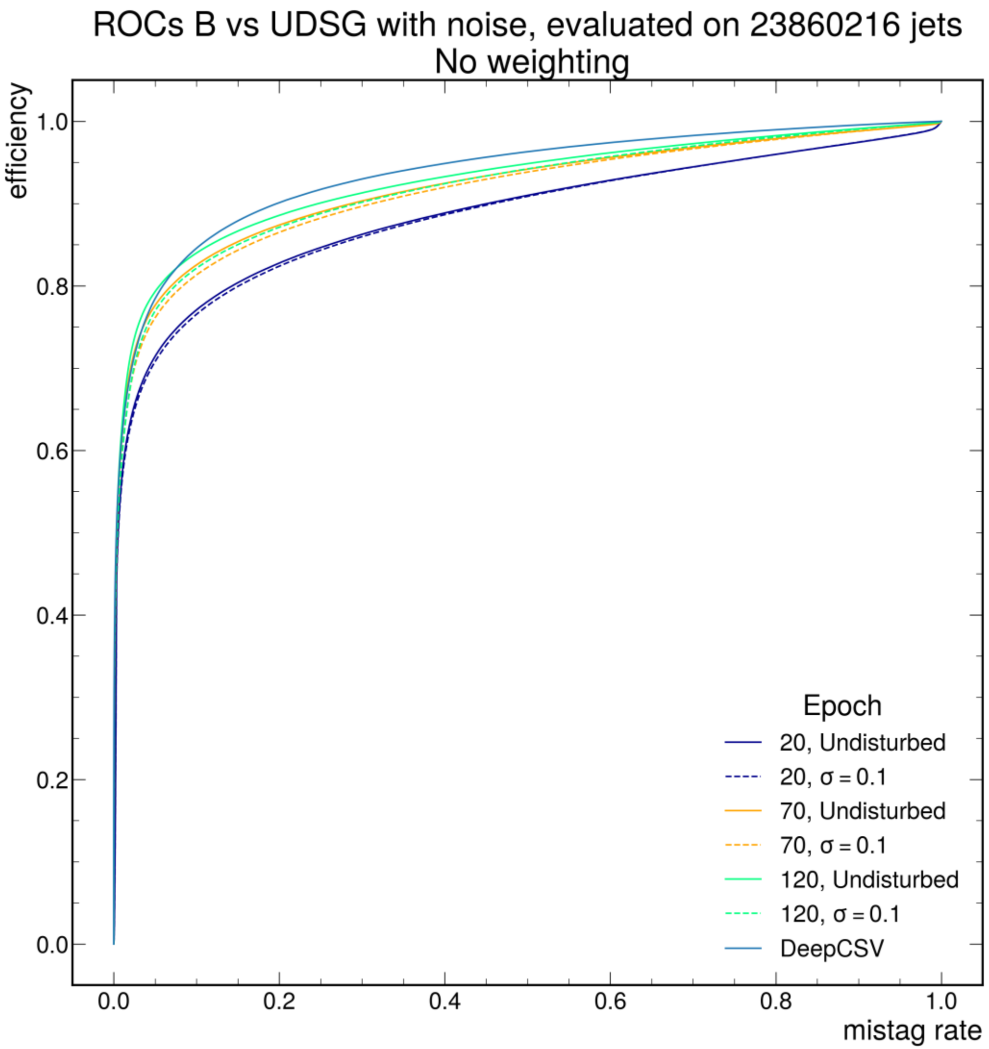


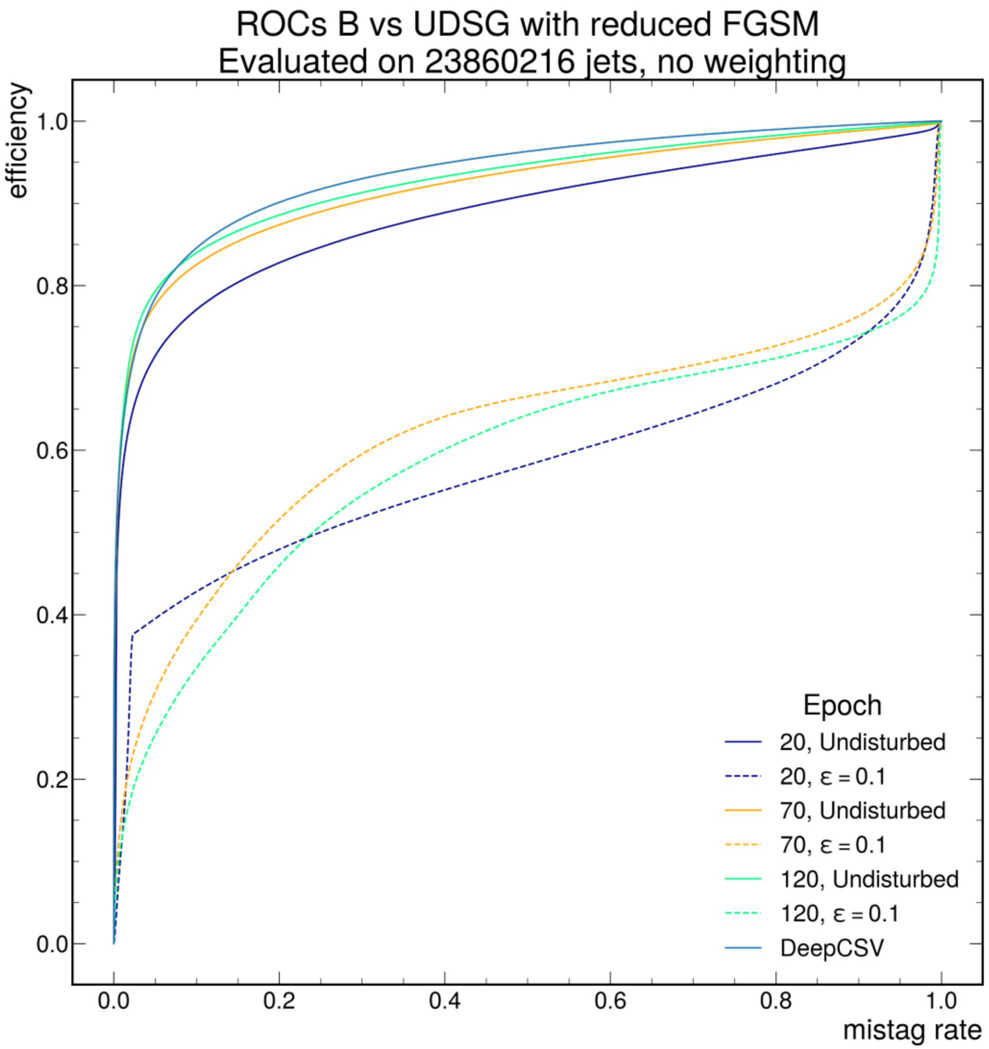
Evolution of AUC with number of epochs





More ROC-curves





AI-safety for jet flavour tagging at the CMS experiment

Xavier Coubez, Nikolas Frediani, Spandan Mondal,
Andrzej Novak, Alexander Schmidt and Annika Stein

17.3.2021

DPG Spring Meeting, Dortmund21

III. Physikalisches
Institut A

RWTH AACHEN
UNIVERSITY