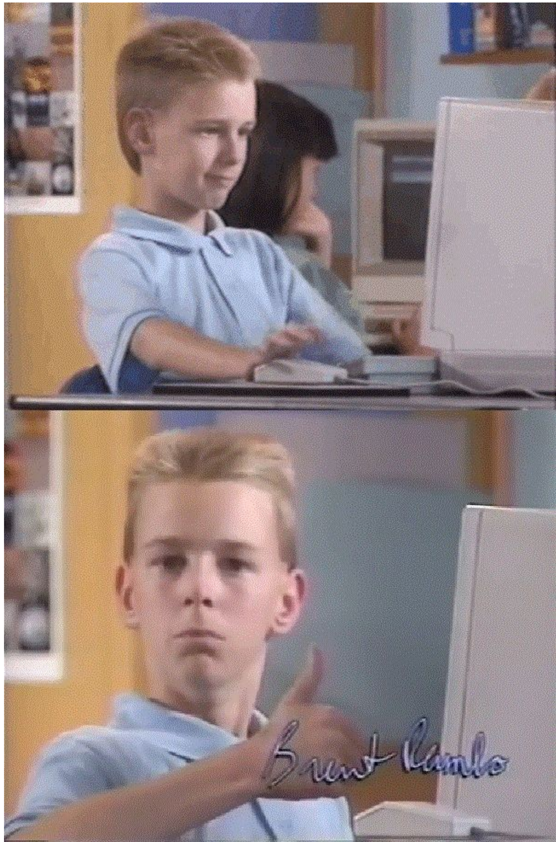
Jet Transformer - Update

A Network for quark/gluon Jet Tagging

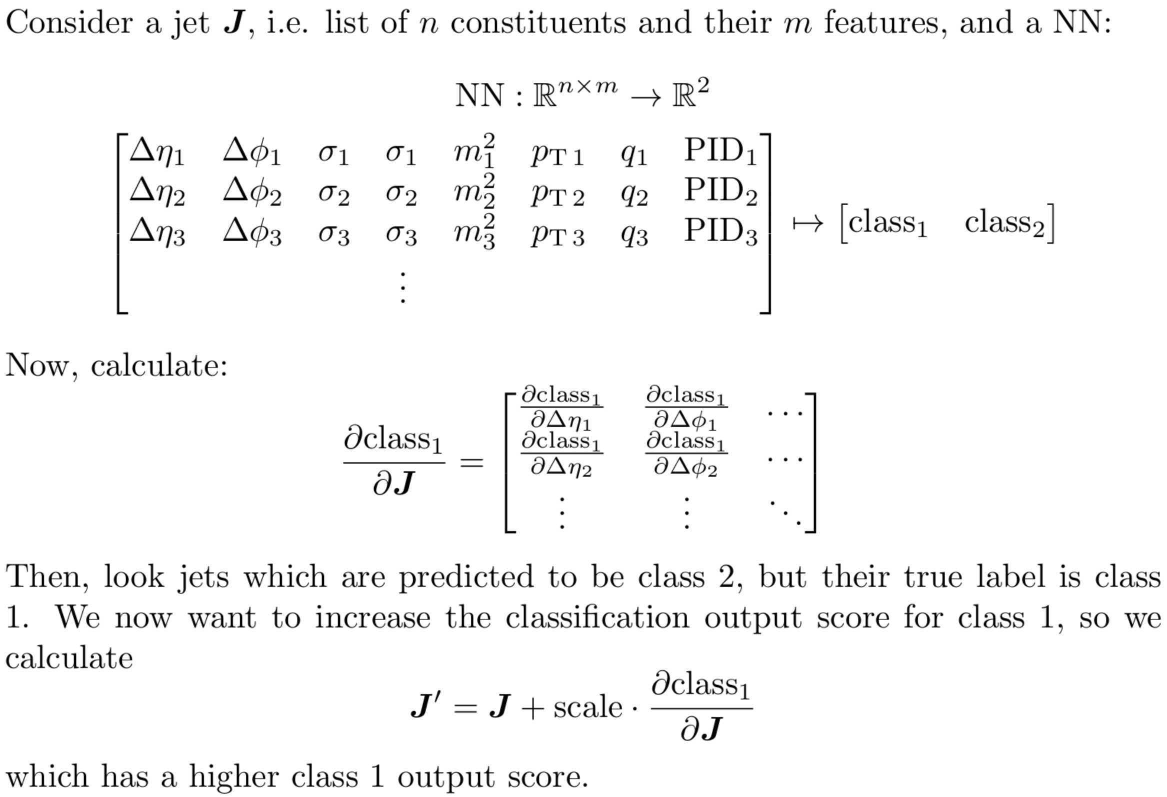
Expectation:

Reality:



Towards Explainable AI

- Froze network parameters



Consider a jet $\boldsymbol{J}$, i.e. list of n constituents and their m features, and a NN:

$$\mathrm{NN} : \mathbb{R}^{n\times m} \to \mathbb{R}^2$$

$$\begin{bmatrix} \Delta\eta_1 & \Delta\phi_1 & \sigma_1 & \sigma_1 & m_1^2 & p_{\mathrm{T}\,1} & q_1 & \mathrm{PID}_1 \\ \Delta\eta_2 & \Delta\phi_2 & \sigma_2 & \sigma_2 & m_2^2 & p_{\mathrm{T}\,2} & q_2 & \mathrm{PID}_2 \\ \Delta\eta_3 & \Delta\phi_3 & \sigma_3 & \sigma_3 & m_3^2 & p_{\mathrm{T}\,3} & q_3 & \mathrm{PID}_3 \\ & & & \vdots & & & & \end{bmatrix} \mapsto \begin{bmatrix} \mathrm{class}_1 & \mathrm{class}_2 \end{bmatrix}$$

Now, calculate:

$$\frac{\partial \mathrm{class}_1}{\partial \boldsymbol{J}} = \begin{bmatrix} \frac{\partial \mathrm{class}_1}{\partial \Delta\eta_1} & \frac{\partial \mathrm{class}_1}{\partial \Delta\phi_1} & \cdots \\ \frac{\partial \mathrm{class}_1}{\partial \Delta\eta_2} & \frac{\partial \mathrm{class}_1}{\partial \Delta\phi_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Then, look jets which are predicted to be class 2, but their true label is class 1. We now want to increase the classification output score for class 1, so we calculate

$$\boldsymbol{J}' = \boldsymbol{J} + \mathrm{scale} \cdot \frac{\partial \mathrm{class}_1}{\partial \boldsymbol{J}}$$

which has a higher class 1 output score.

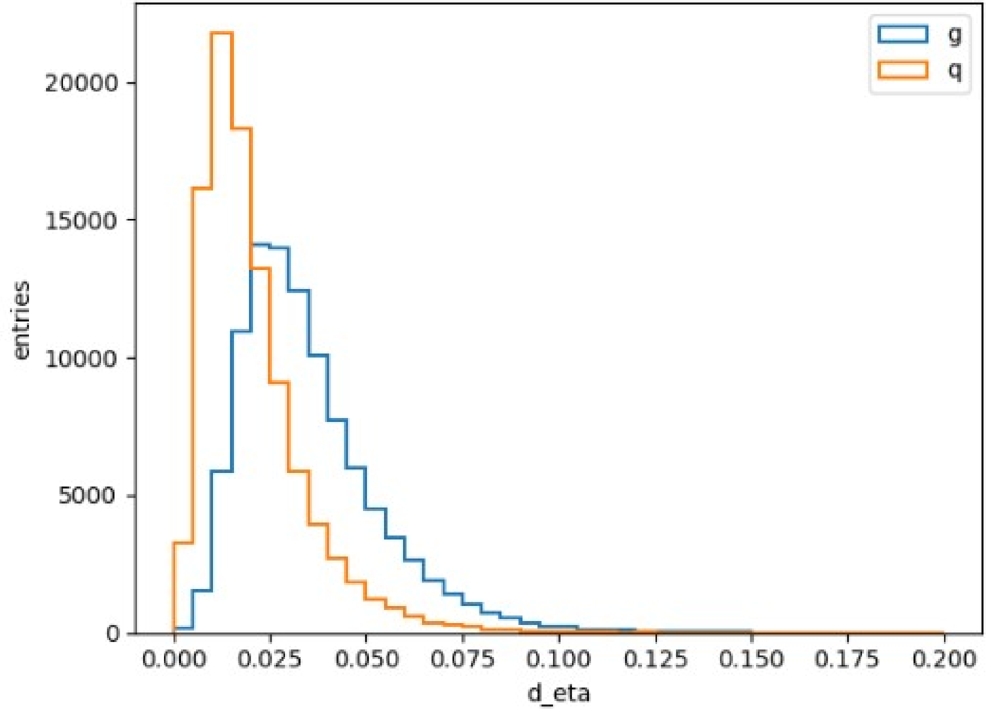
Does the network pick up Characteristics of q/g Jets?

* sorry for the colour mismatch

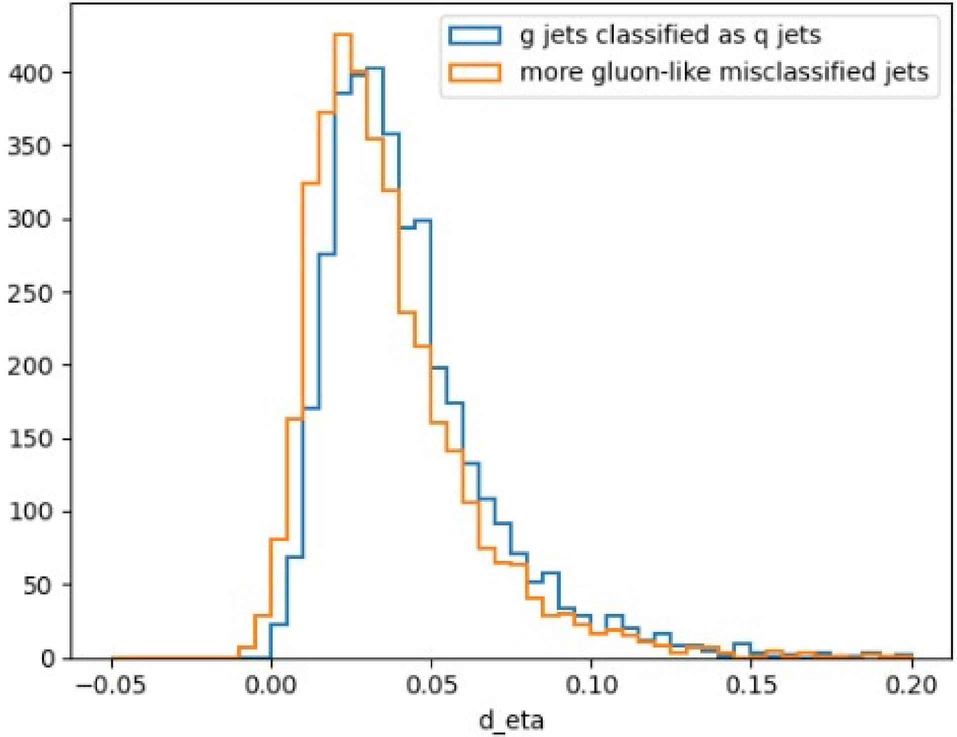
** scale=1 here

*** NN trained for 3 epochs

DATA:



NN:



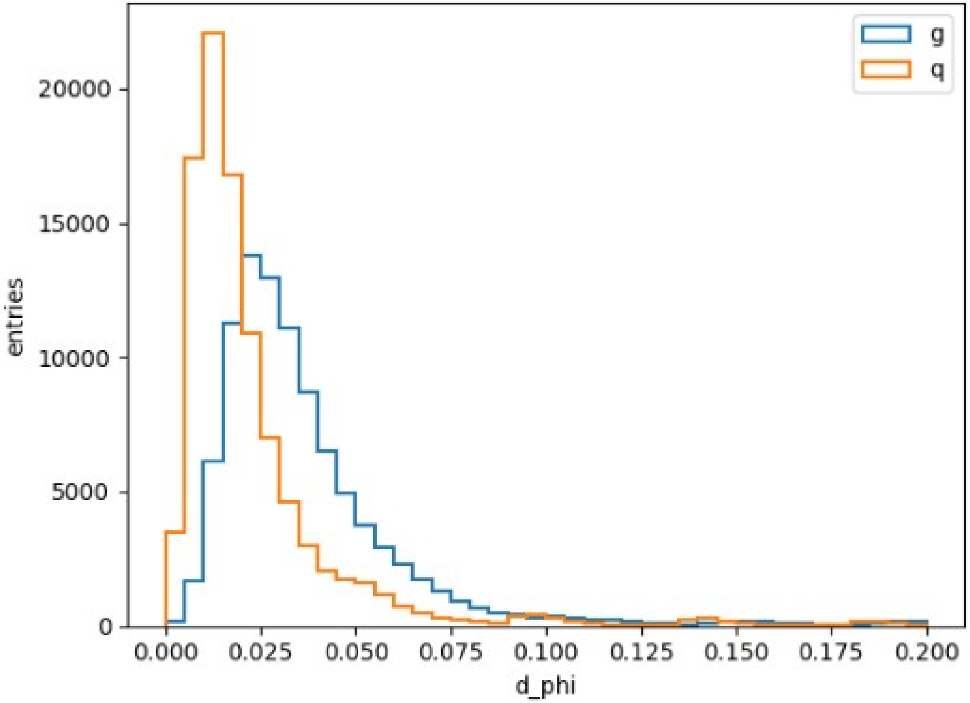
Does the network pick up Characteristics of q/g Jets?

* sorry for the colour mismatch

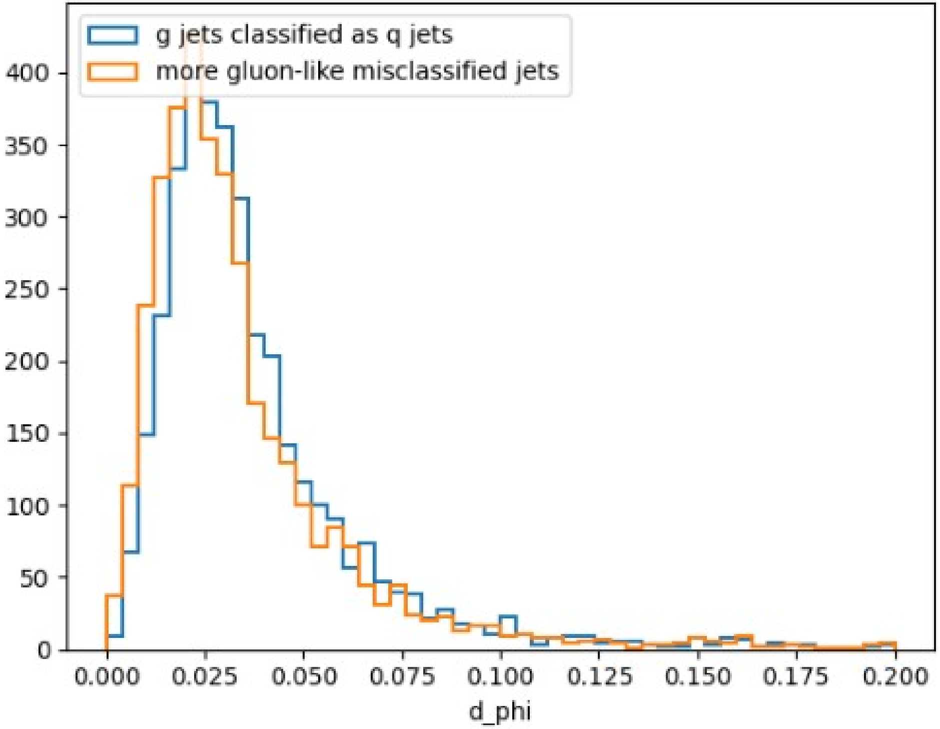
** scale=1 here

*** NN trained for 3 epochs

DATA:



NN:



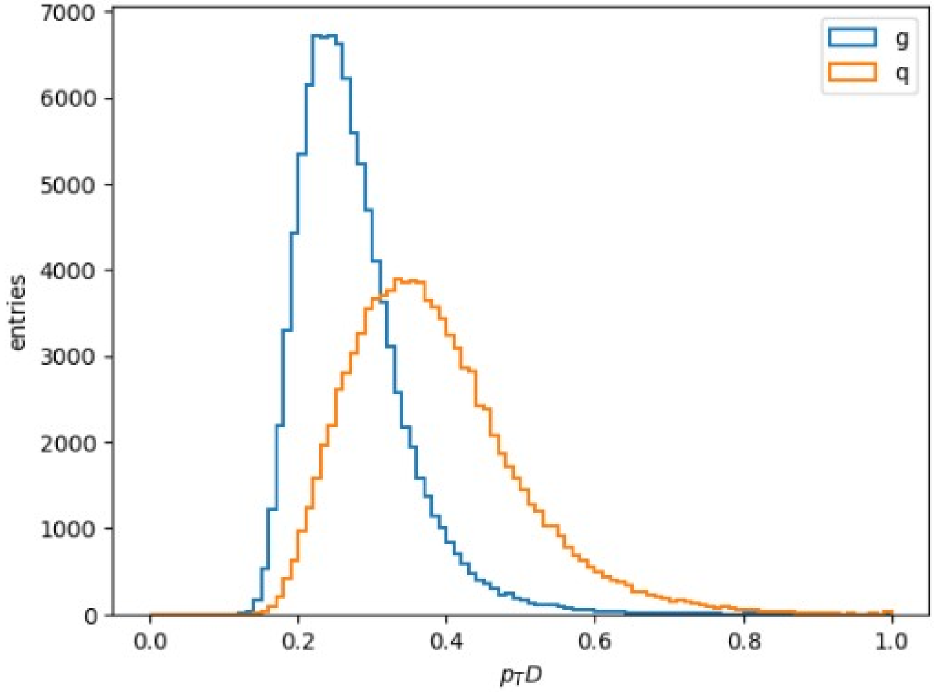
Does the network pick up Characteristics of q/g Jets?

* sorry for the colour mismatch

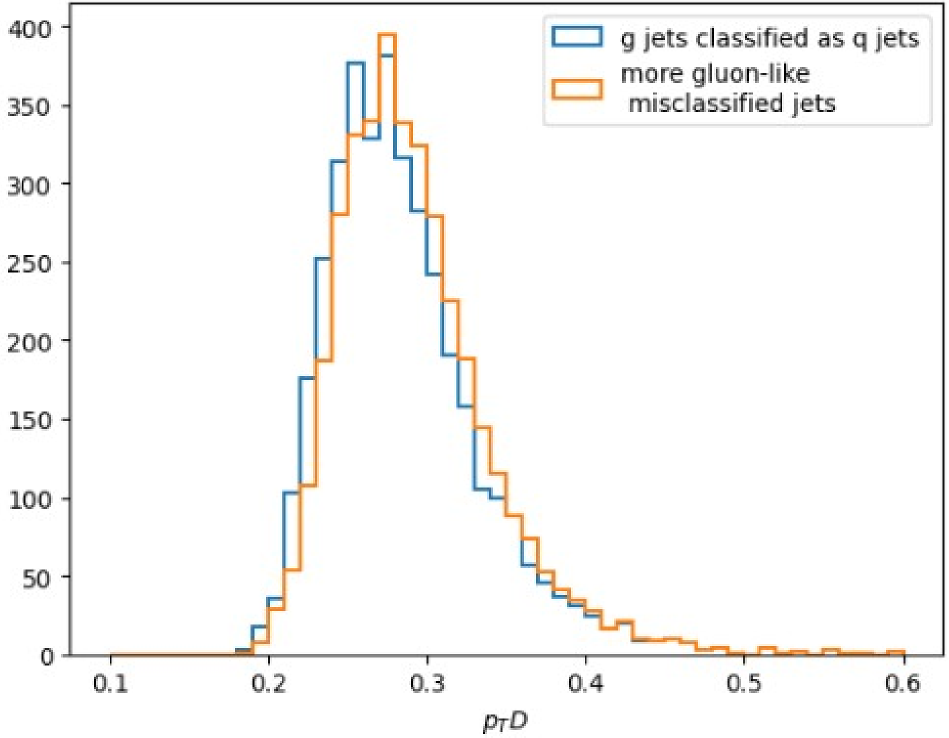
** scale=1 here

*** NN trained for 3 epochs

DATA:



NN:



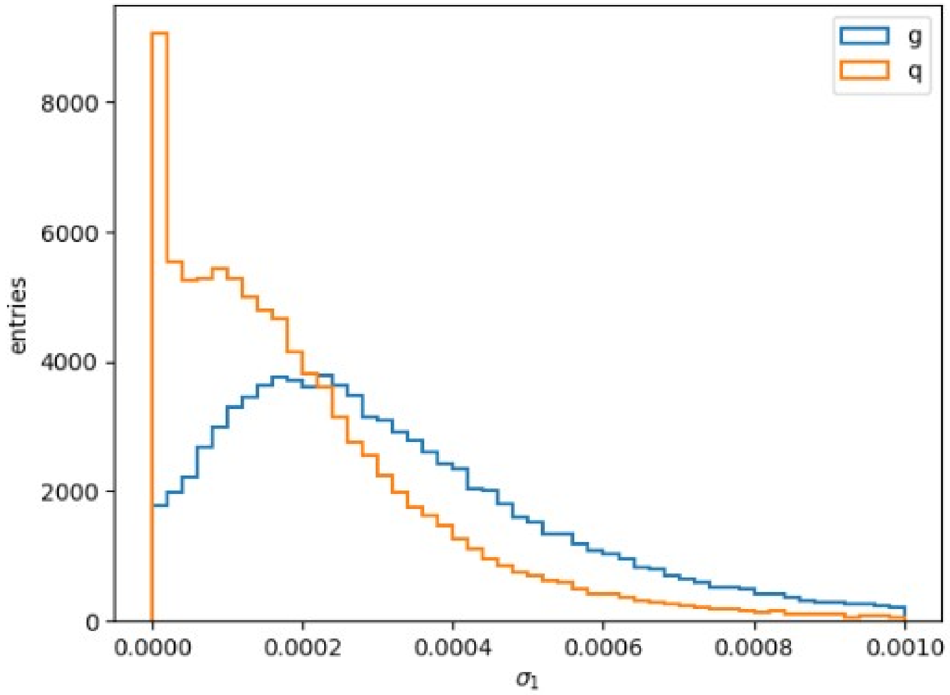
* sorry for the colour mismatch

** scale=1 here

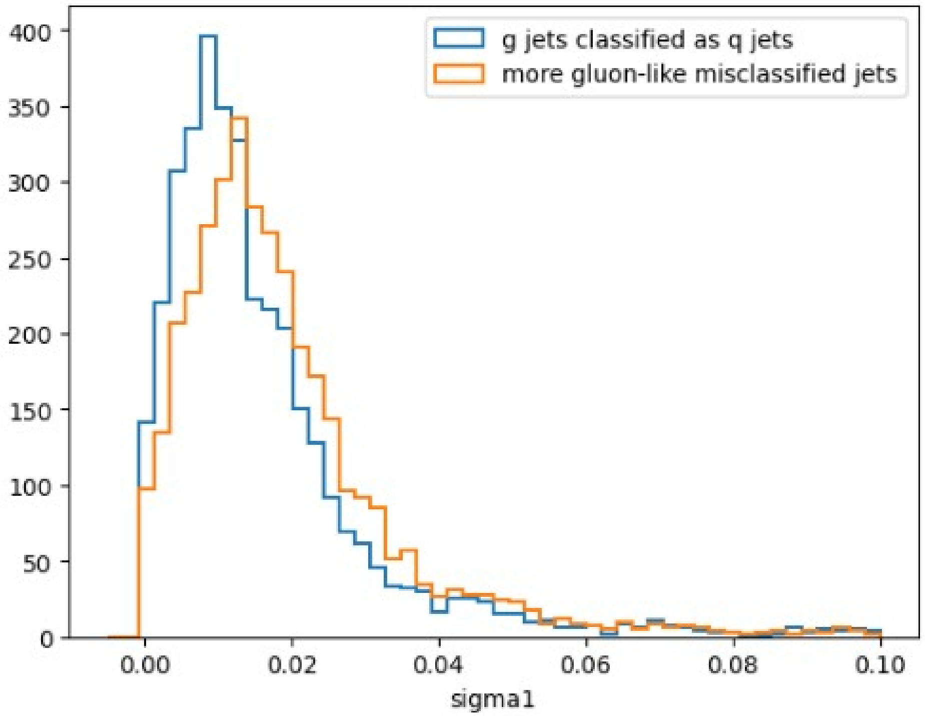
*** NN trained for 3 epochs

Does the network pick up Characteristics of q/g Jets?

DATA:



NN:



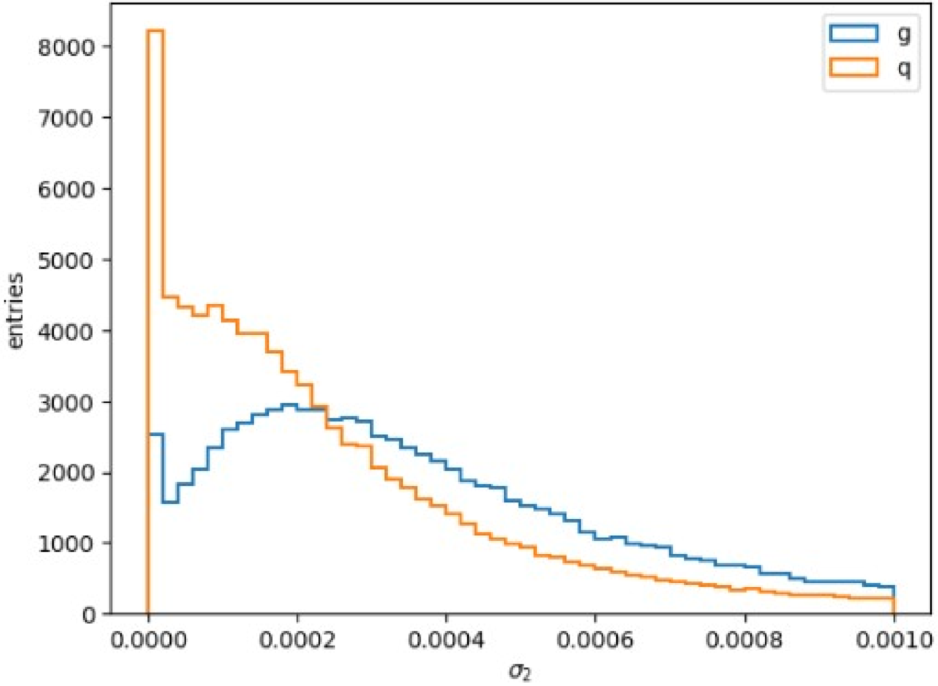
* sorry for the colour mismatch

** scale=1 here

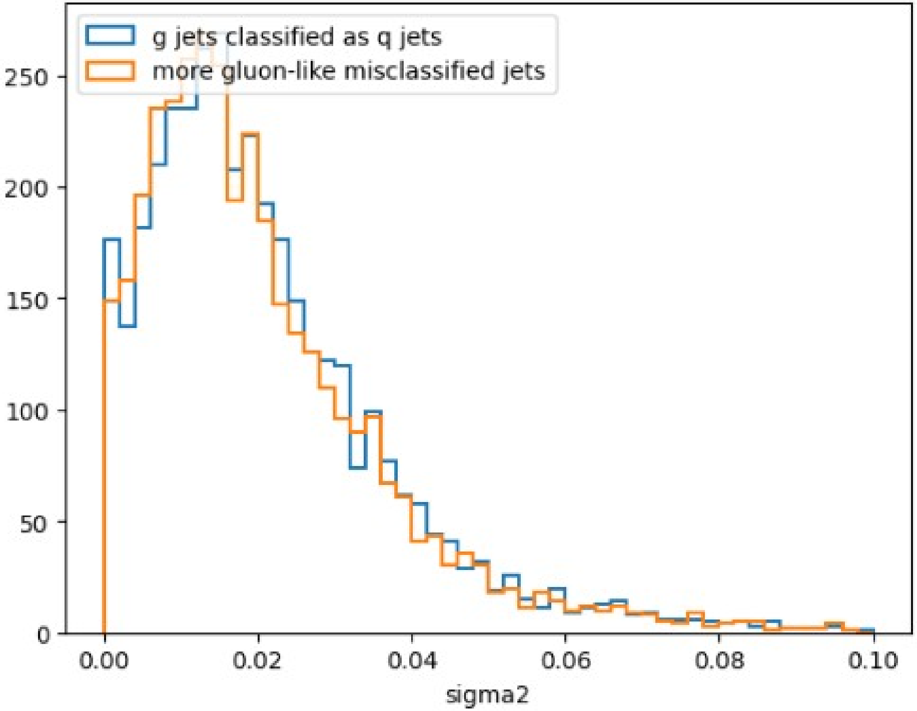
*** NN trained for 3 epochs

Does the network pick up Characteristics of q/g Jets?

DATA:



NN:



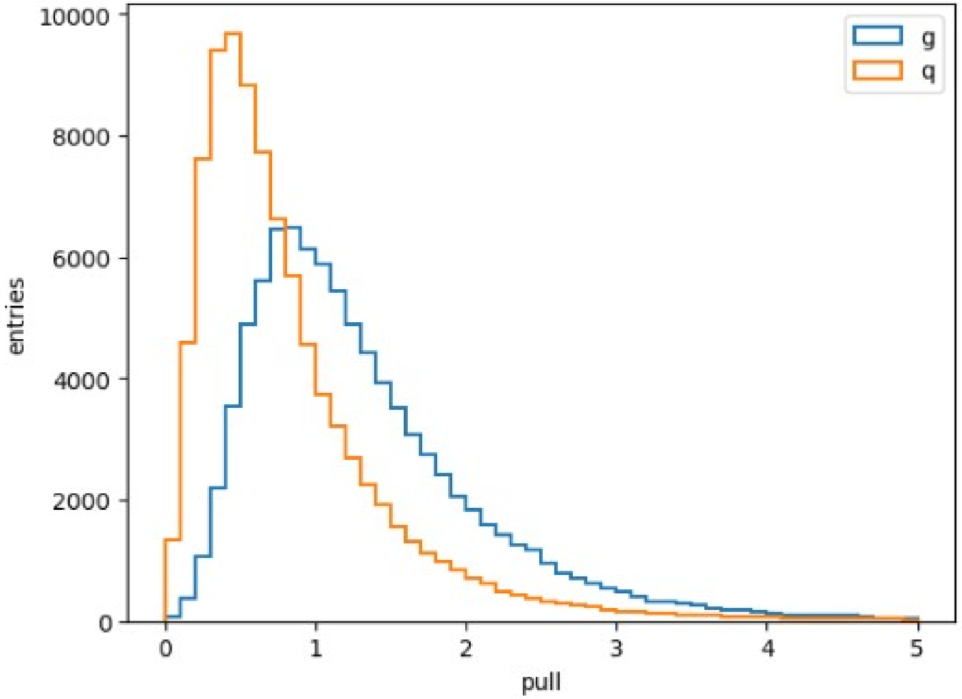
* sorry for the colour mismatch

** scale=1 here

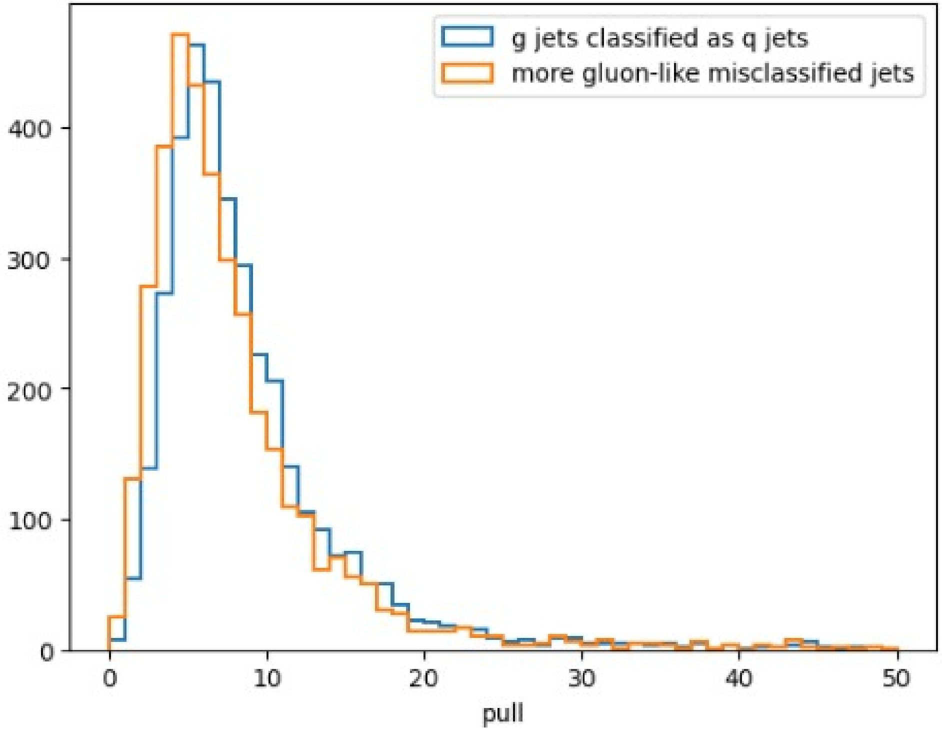
*** NN trained for 3 epochs

Does the network pick up Characteristics of q/g Jets?

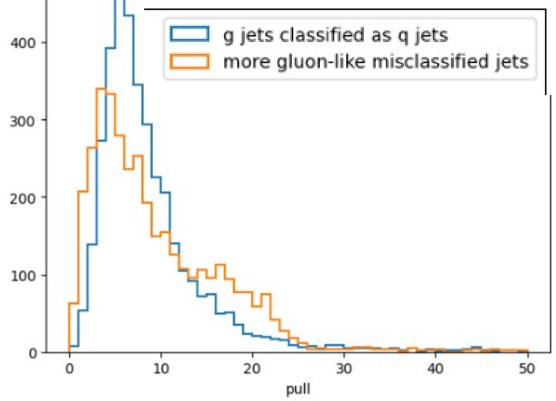
DATA:



NN:



To Do:



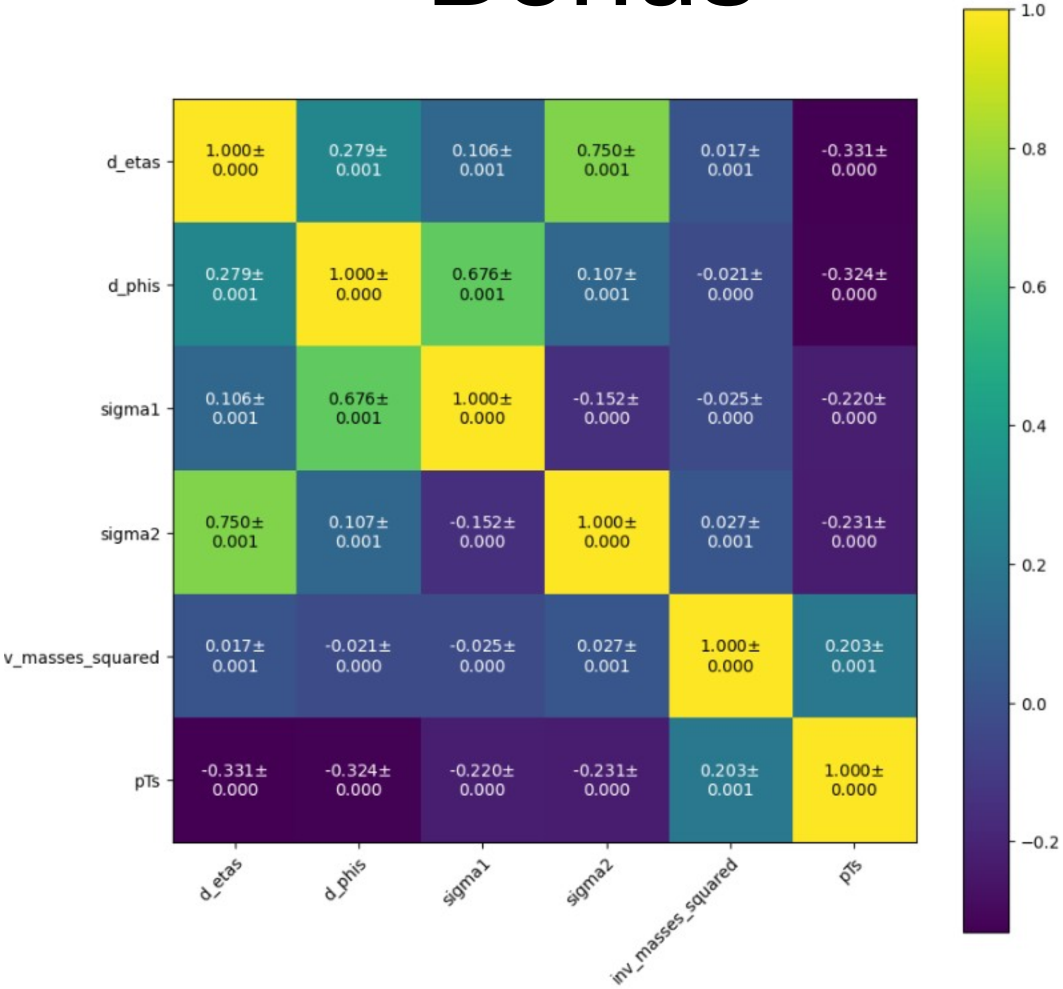
scale=10

- What is the optimal value for ‘scale’
- How does the classification output score vary with the ‘scale’
- What feature does the network ‘want to change the most’?
- Apply this to TOF Transformer

→ plot vector field of gradients for showers!

Bonus

Pearson product-moment correlation coefficients



	d_etas	d_phis	sigma1	sigma2	inv_masses_squared	pTs
d_etas	1.000± 0.000	0.279± 0.001	0.106± 0.001	0.750± 0.001	0.017± 0.001	-0.331± 0.000
d_phis	0.279± 0.001	1.000± 0.000	0.676± 0.001	0.107± 0.001	-0.021± 0.000	-0.324± 0.000
sigma1	0.106± 0.001	0.676± 0.001	1.000± 0.000	-0.152± 0.000	-0.025± 0.000	-0.220± 0.000
sigma2	0.750± 0.001	0.107± 0.001	-0.152± 0.000	1.000± 0.000	0.027± 0.001	-0.231± 0.000
v_masses_squared	0.017± 0.001	-0.021± 0.000	-0.025± 0.000	0.027± 0.001	1.000± 0.000	0.203± 0.001
pTs	-0.331± 0.000	-0.324± 0.000	-0.220± 0.000	-0.231± 0.000	0.203± 0.001	1.000± 0.000