Discriminating quark & gluon jets at CMS

Yuta Takahashi (Zurich)

18 July, BOOST 2017

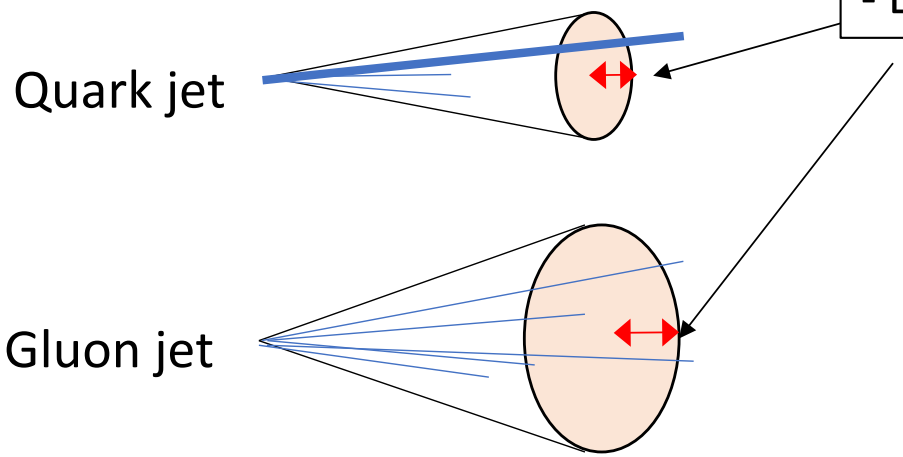
On behalf of CMS Collaboration

Contents

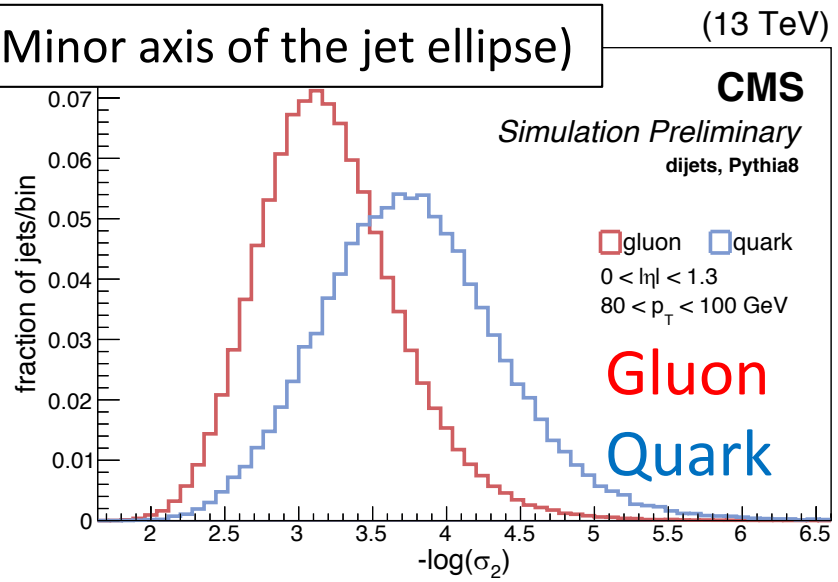
- Quark/gluon likelihood discriminator

- On-going development, using BDT, towards 2017+

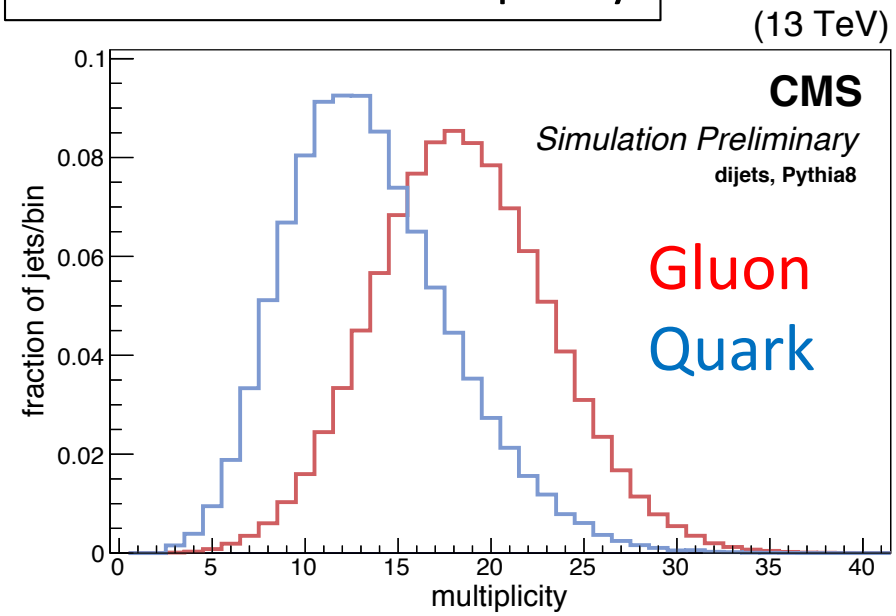
- Jets from light-flavour quarks ≠ jets from gluons
- CMS has developed likelihood-based discriminator, built from **3** kinematic variables



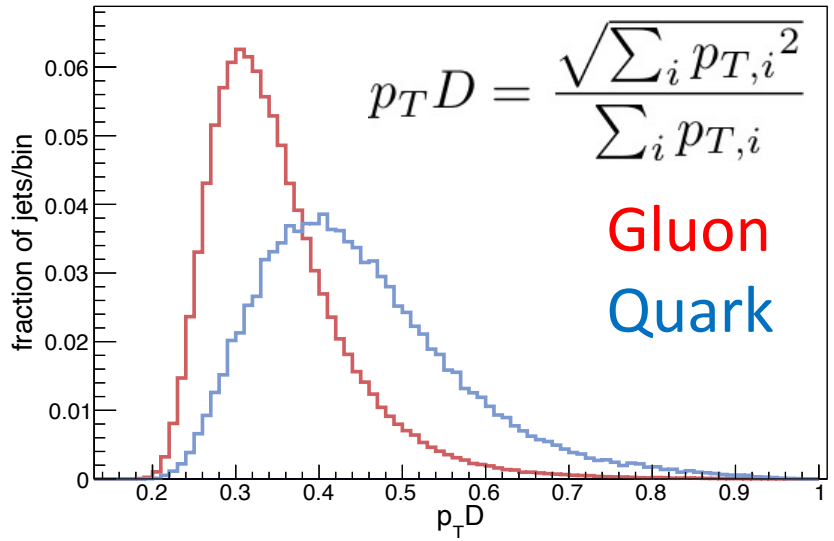
- Log (Minor axis of the jet ellipse)



Jet constituent multiplicity



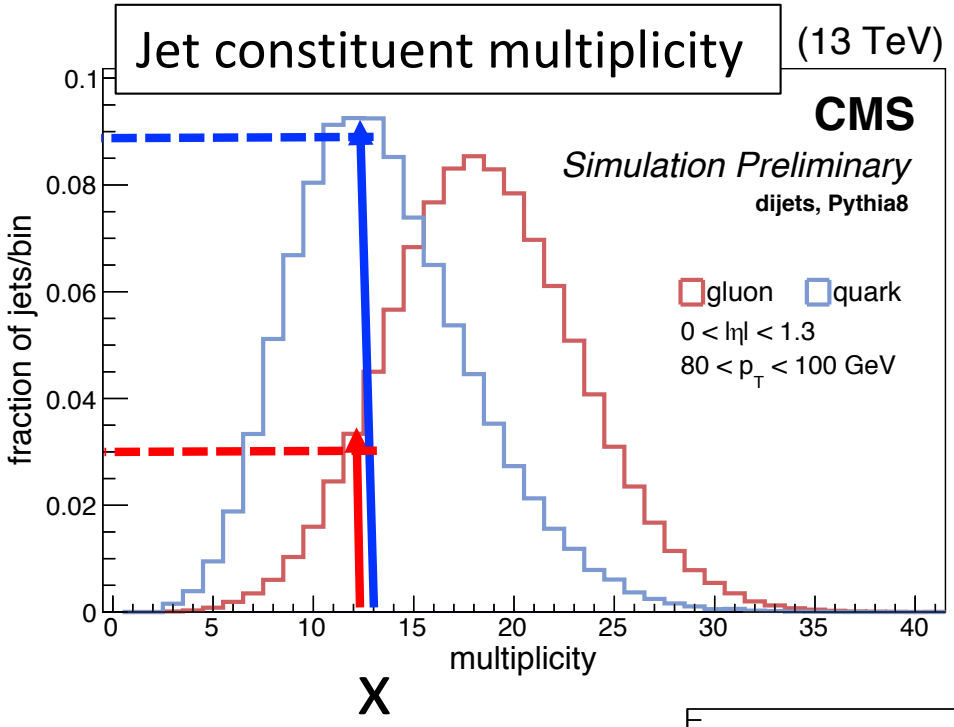
Fragmentation function (ptD)



Building likelihood ...

$p_{quark}(x)$

$p_{gluon}(x)$



x

Same for

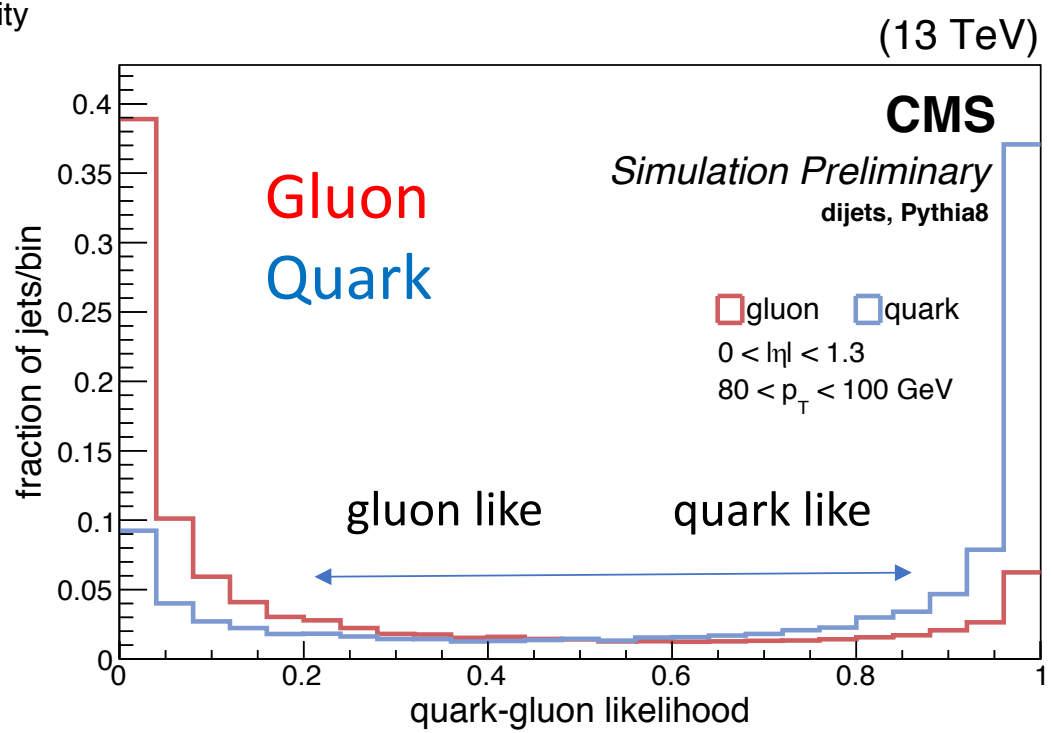
Minor axis of the jet ellipse

Fragmentation function (ptD)

$$L_{\text{quark}} = \Pi p_{i,\text{quark}}(x_i),$$

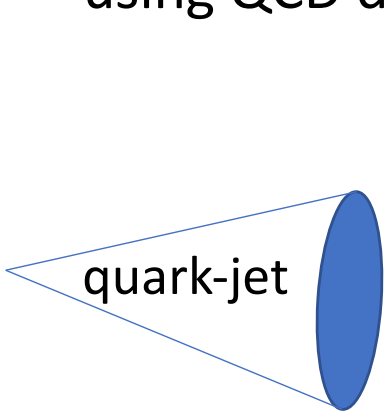
$$L_{\text{gluon}} = \Pi p_{i,\text{gluon}}(x_i)$$

$$\Rightarrow \mathcal{L} = \frac{L_{\text{quark}}}{L_{\text{quark}} + L_{\text{gluon}}}$$

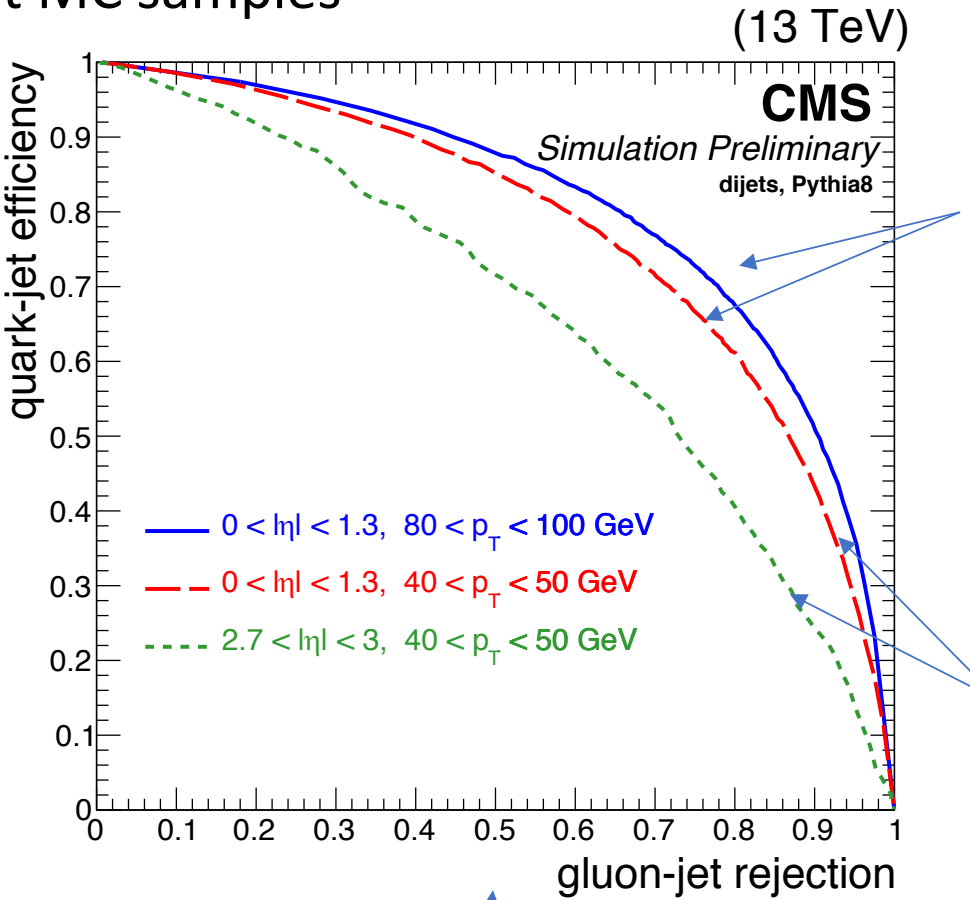


Typical performance (Simulation)

Likelihood is built in bins of jet p_T / η / ρ (average p_T density per unit area), using QCD di-jet MC samples

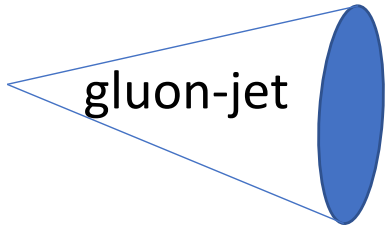


Labeled as quark



Better performance at high p_T (larger discrimination power coming from Jet constituent multiplicity)

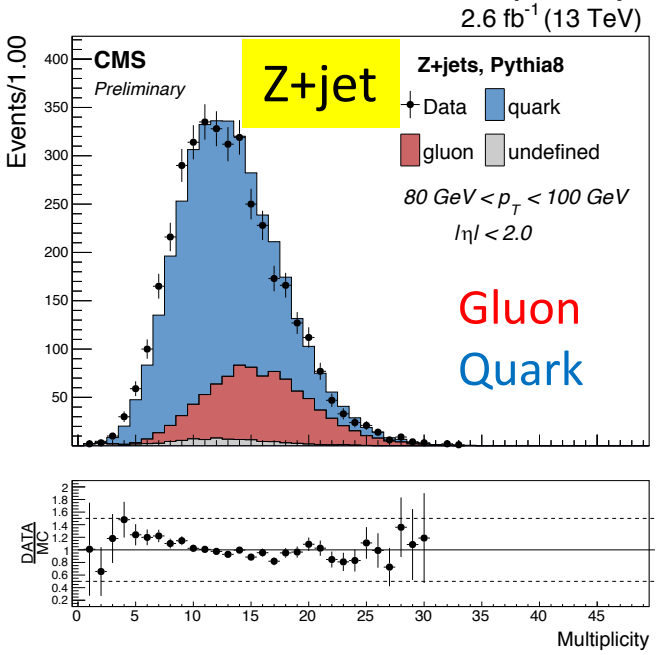
Better performance at central rapidity (detector resolution)



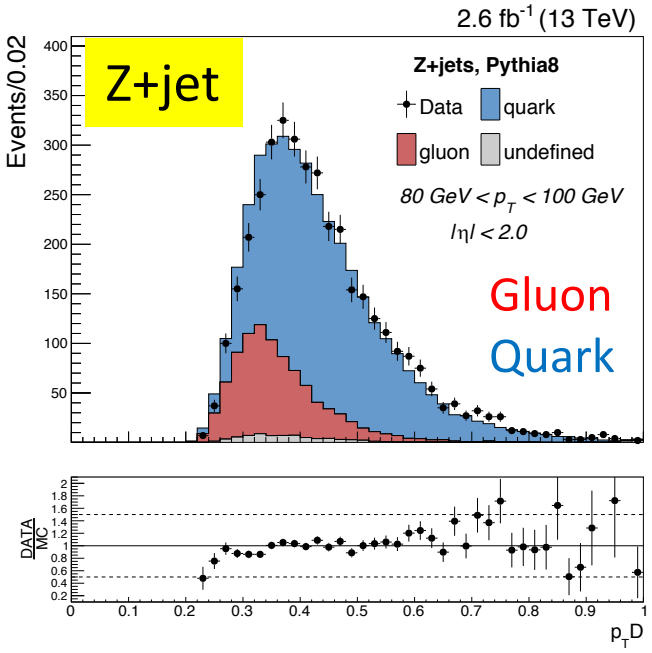
Labeled as gluon

Validation using 13TeV Data

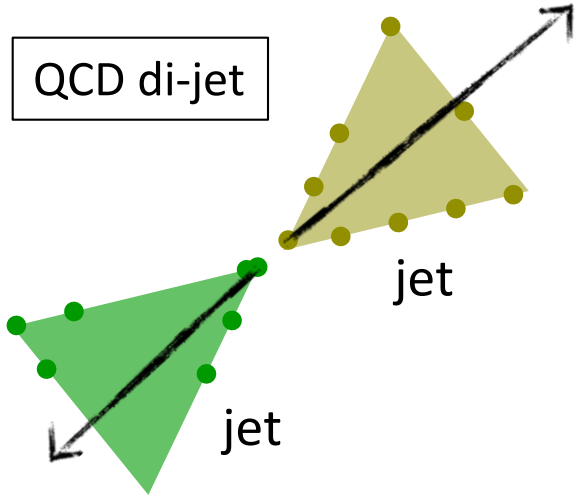
Jet Constituent Multiplicity



Fragmentation (ptD)

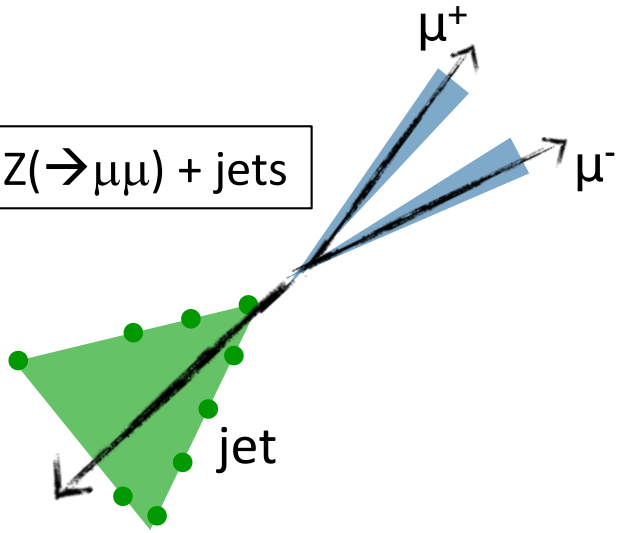


QCD di-jet



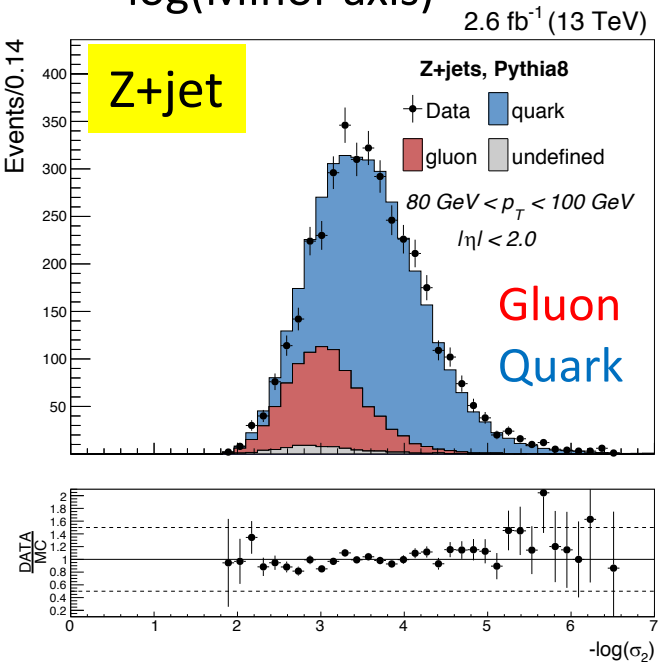
Enriched in Gluon jets (70 - 80%)

Z(→μμ) + jets



Enriched in light-flavour quarks (70 - 80%)

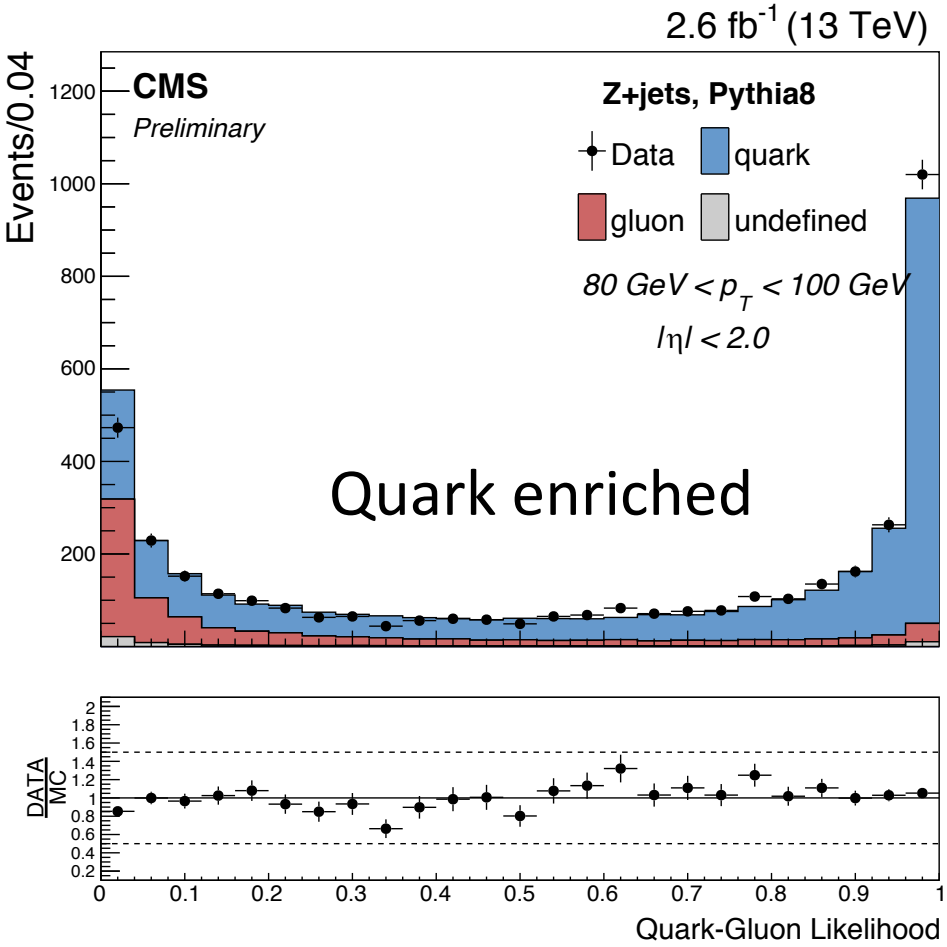
– log(Minor axis)

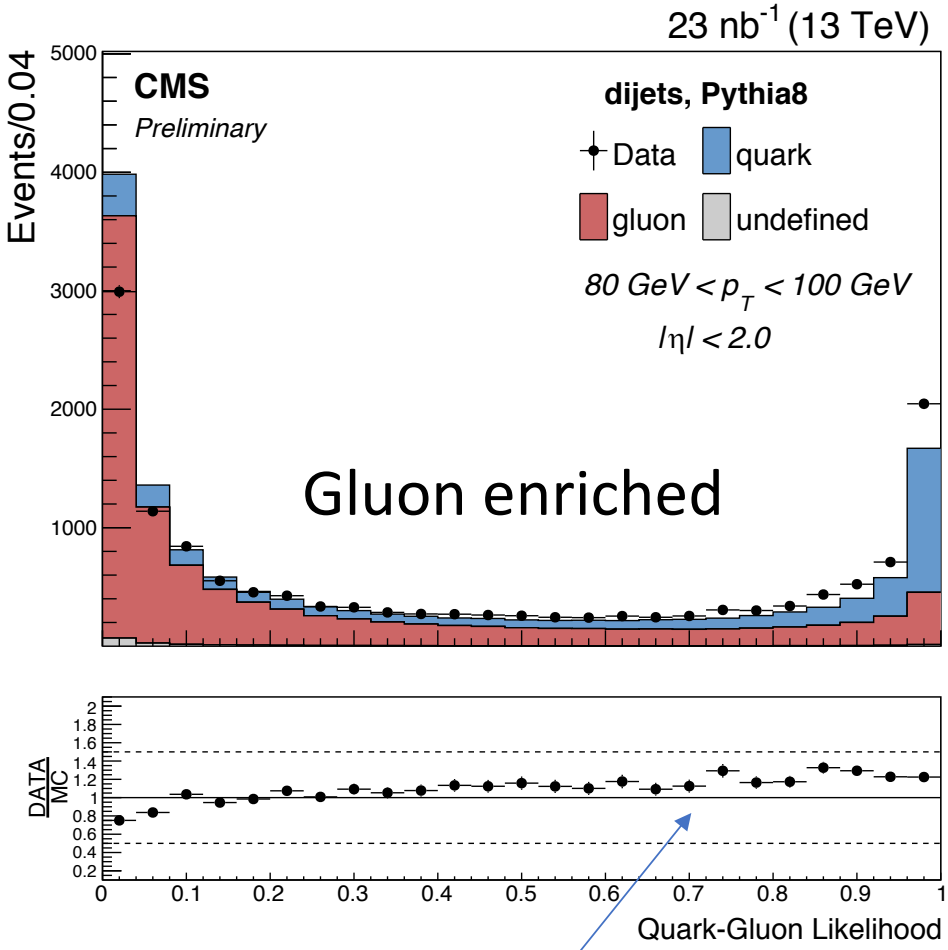


Likelihood score distribution

Z + jet control region

QCD di-jet control region



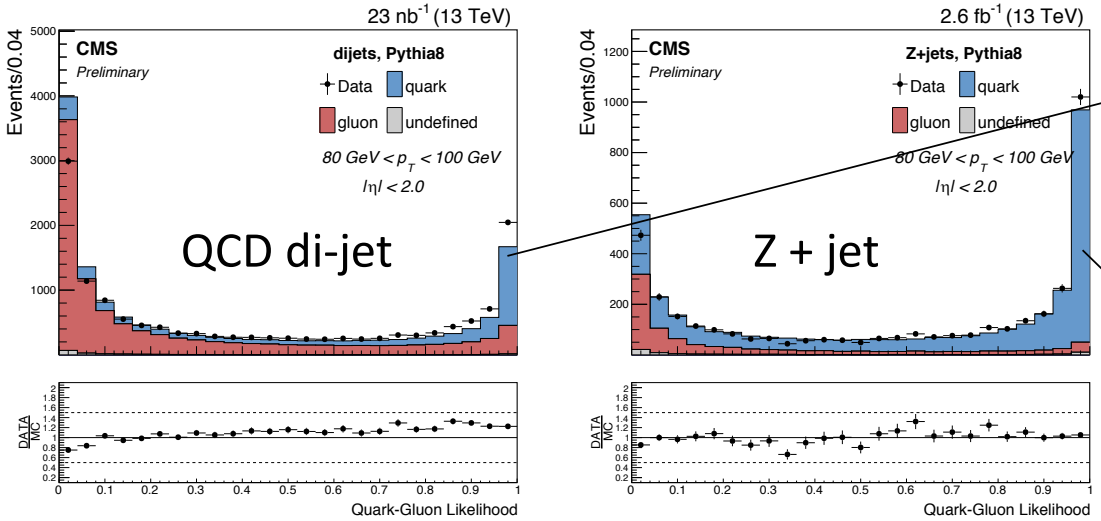


Overall, decent agreement

Some miss-modelings

MC correction

Derive “reweighting function” using two validation regions

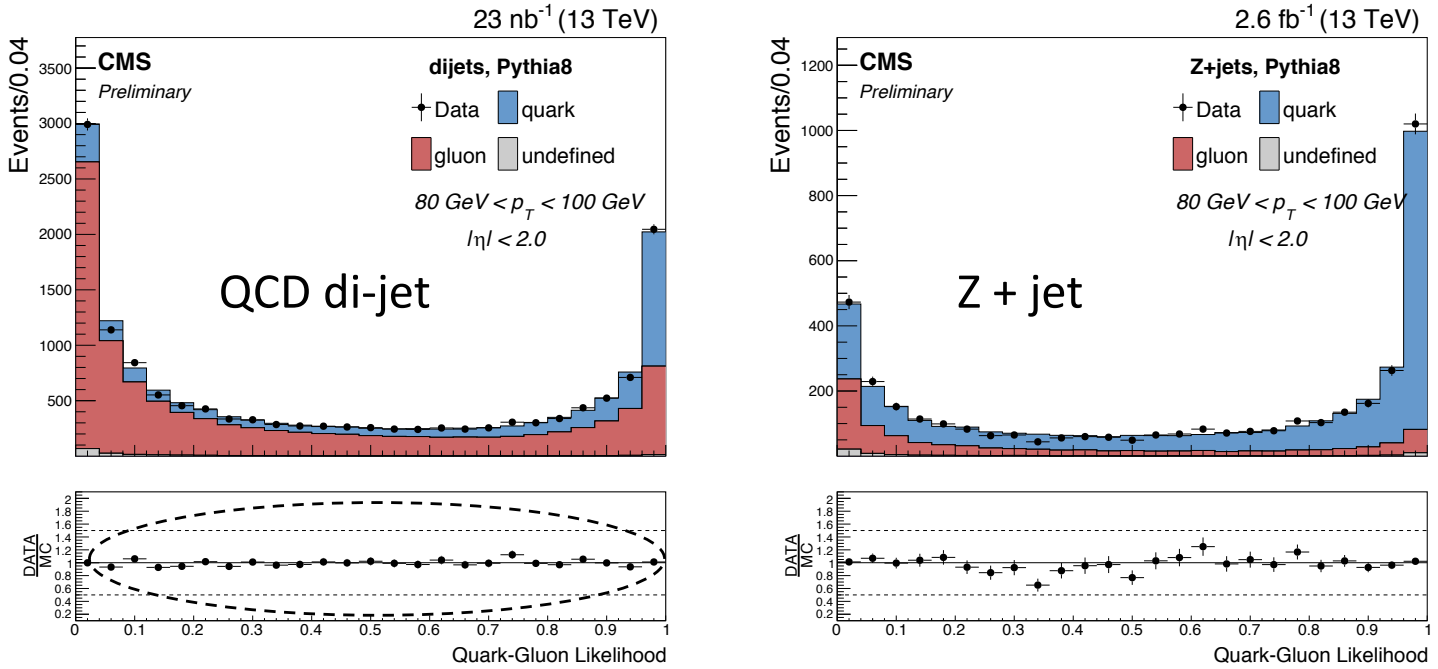


$$N_{data,\,QCD} = w_g \,\square + w_q \,\square$$

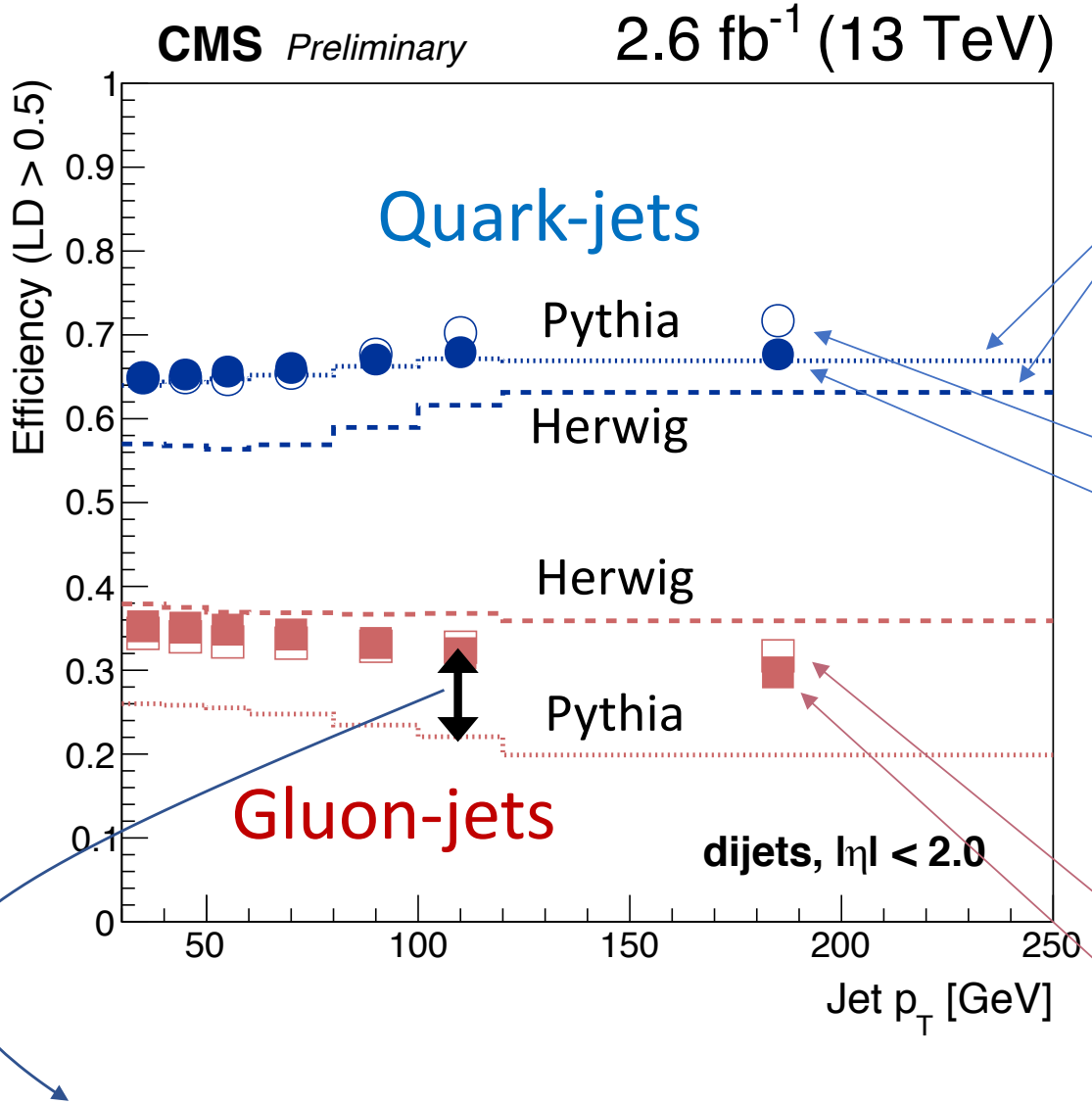
$$N_{data,\,DY} = w_g \,\square + w_q \,\square$$

Obtained w_g, w_q is fit
by using polynomial function

After reweighting



Dependence on the MC generator



Before MC correction

After correction, individually for each MC

Herwig

Pythia

Differences among MC generators is absorbed

After correction

Herwig

Pythia

Systematics: difference between w/ and w/o MC correction (for the given MC generator)

Documentations

8 TeV

CMS-PAS-JME-13-002 (https://cds.cern.ch/record/1599732)

13 TeV

CMS-DP-2016-070 (https://cds.cern.ch/record/2234117)

CMS PAS-JME-16-003 (https://cds.cern.ch/record/2256875)

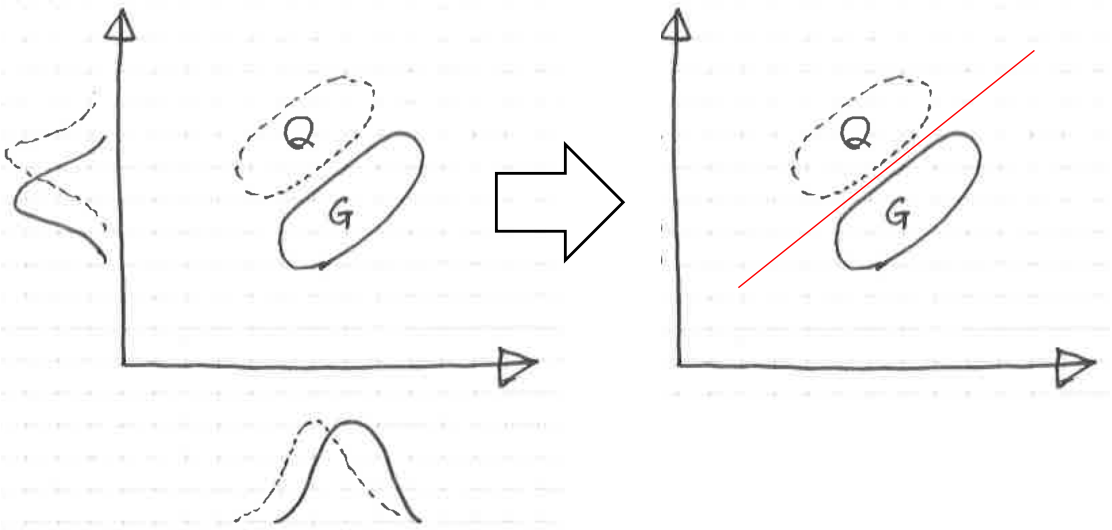
On-going development towards 2017+

In summary:

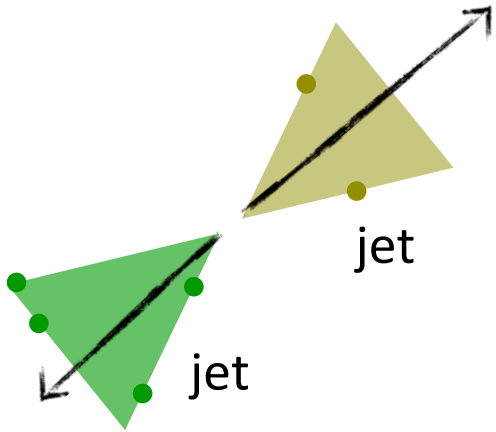
- Start investigating BDT-based discriminator
- Revise the list of input variables

Likelihood → BDT

Benefit: better handling of the correlation among input variables



- Trained BDT based on the QCD di-jet Monte-Carlo simulations
  - Same p_T / eta binning as likelihood
  - To avoid over-training, each bin should contain > 100k events



- Training is performed jet-by-jet basis
  - Each jet should originate from either quark or gluon
  - Each jet should match to the generator-level parton in a unique way

Revised Input variables for BDT

- Examined ~15 possible input variables → use 5 of them
- Decision taken based on the discrimination power, correlations among variables and to be robust for pile-up

Likelihood

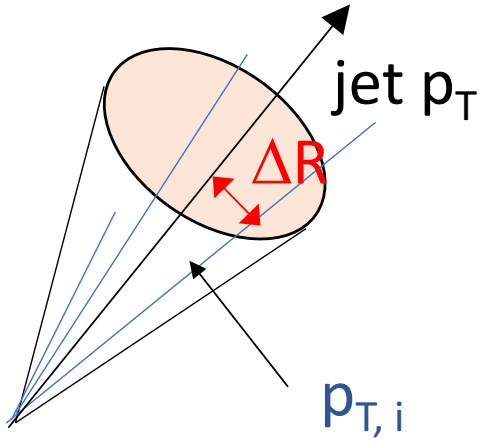
- Minor axis
- ptD
- Jet constituent multiplicity

BDT

- Minor axis
- ptD
- Charged particle multiplicity
- Major axis of the jet
- $\Sigma_i (\log(p_{T,i}/\Delta R)) / \text{jet } p_T$

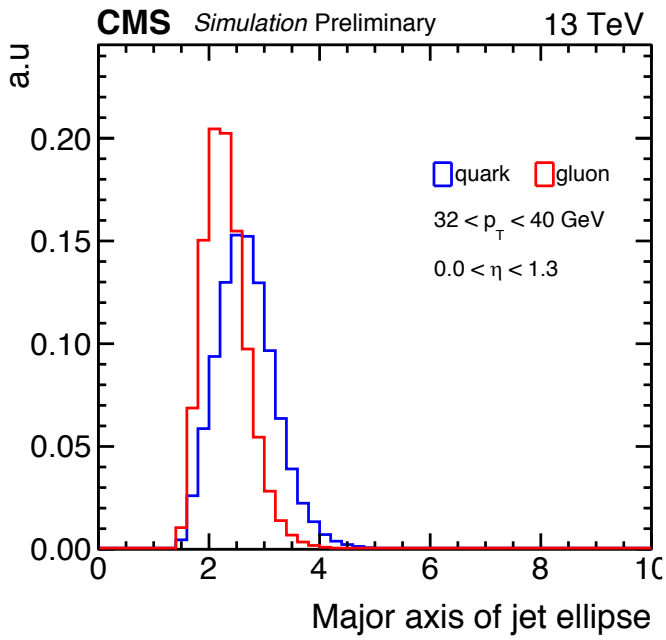
Newly added

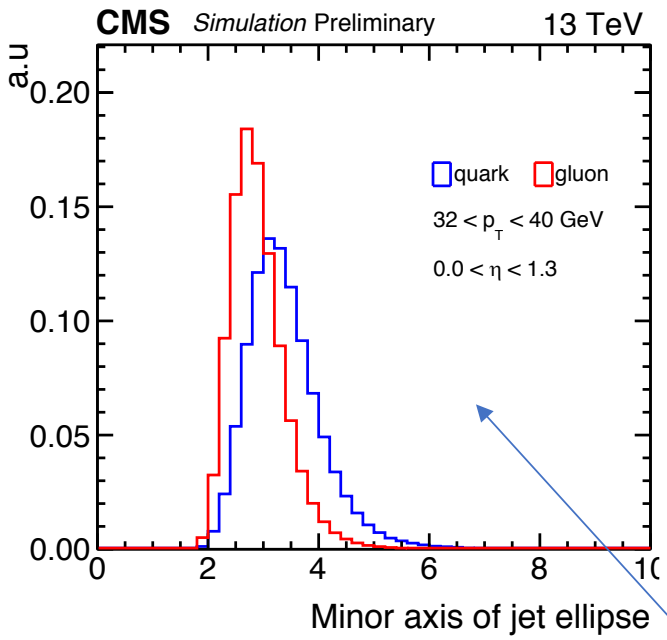
Neutral multiplicity
has a worse detector resolution
and degrades the performance
(and not robust against pileup)

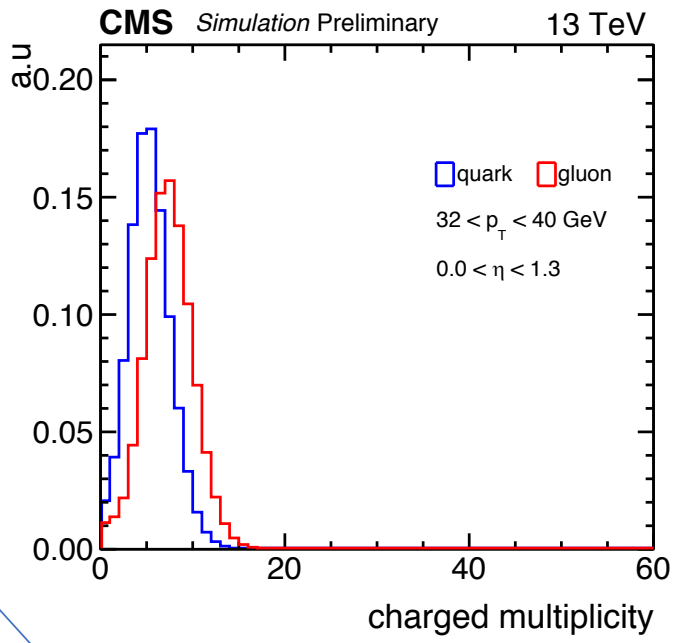


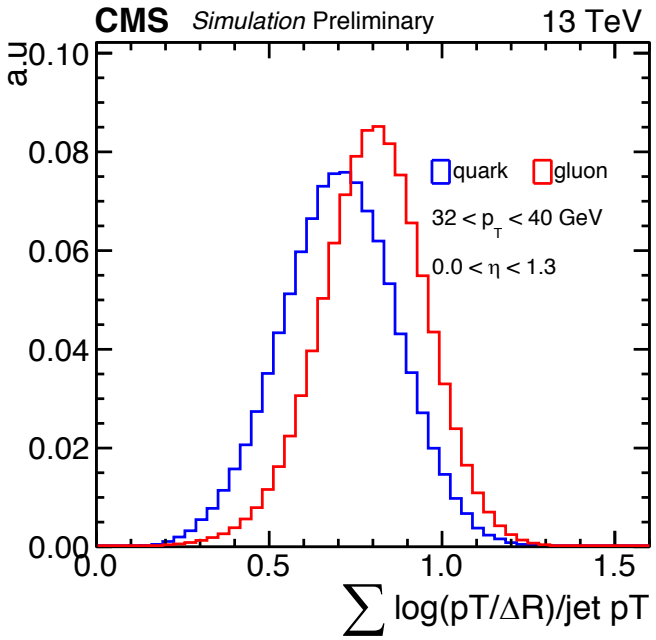
Input variables @ low p_T

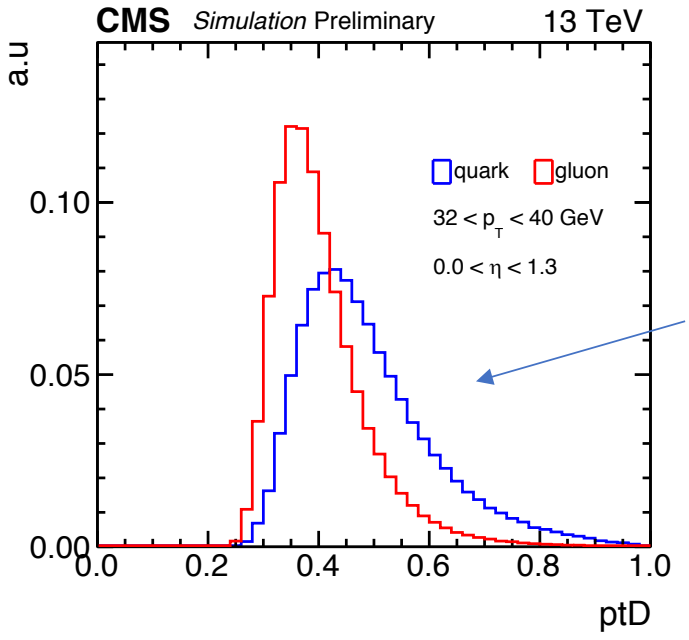
32 – 40 GeV
0 < eta < 1.3







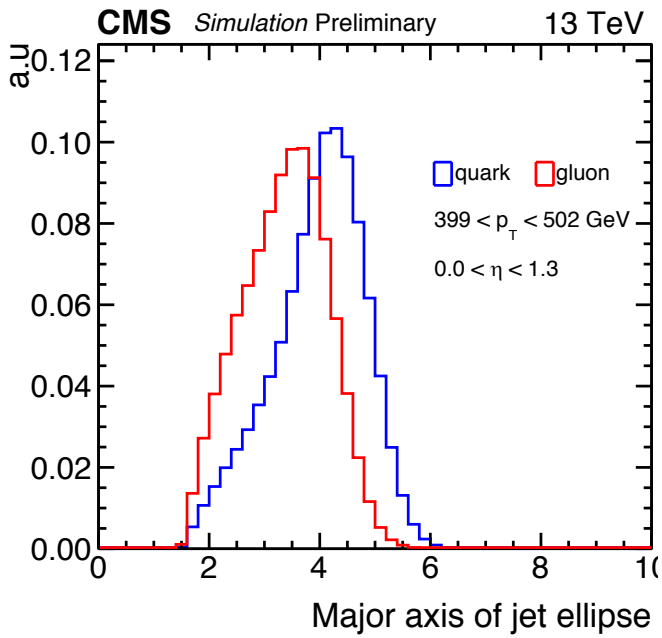


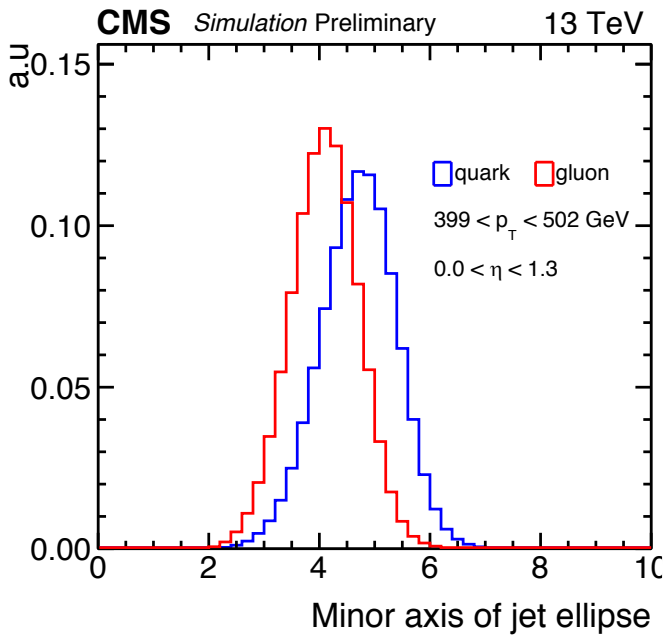


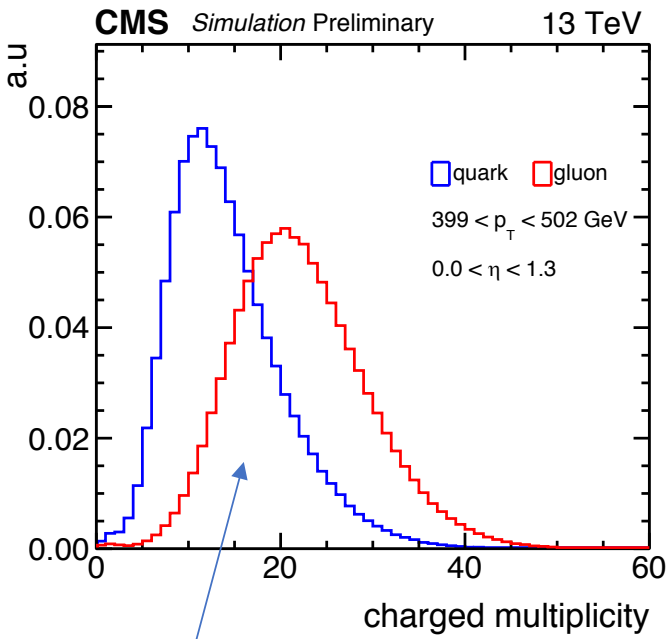
At low p_T, shape variables bring nice discrimination power

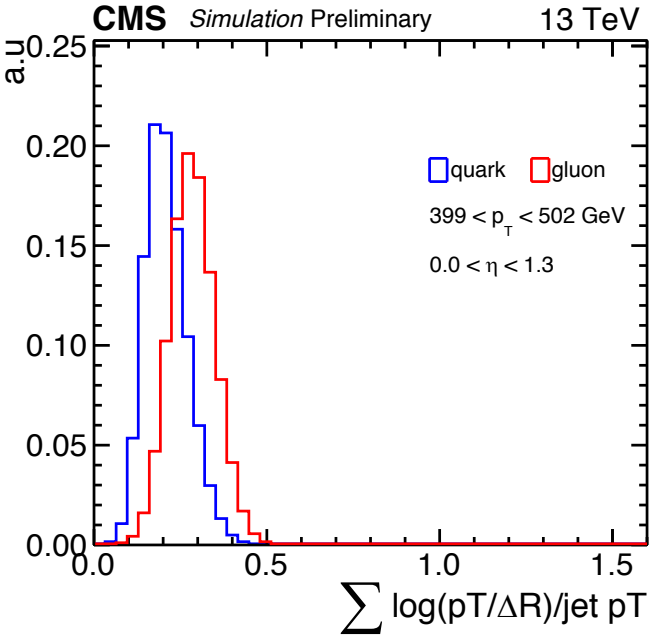
Input variables @ high p_T

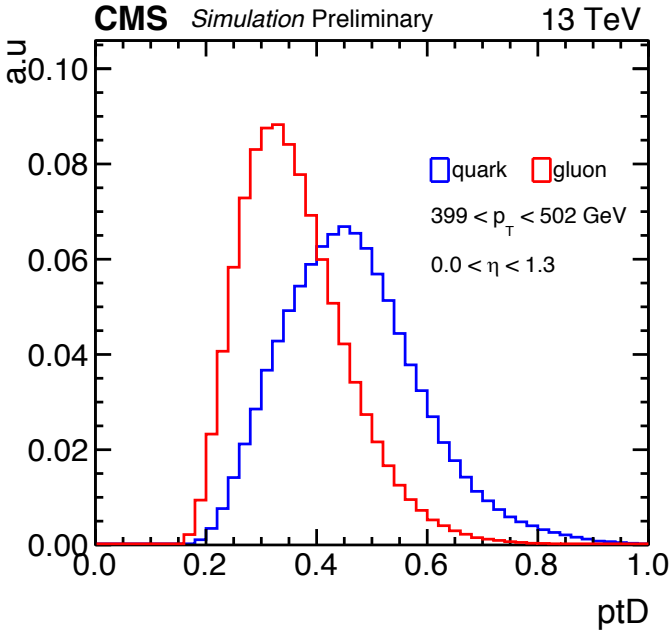
400 – 500 GeV
0 < eta < 1.3





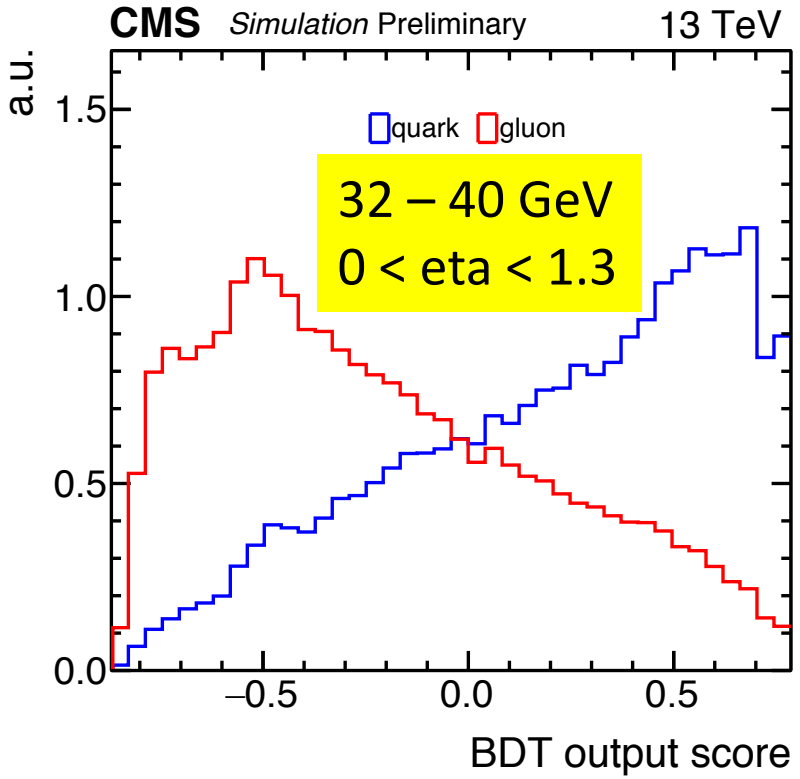


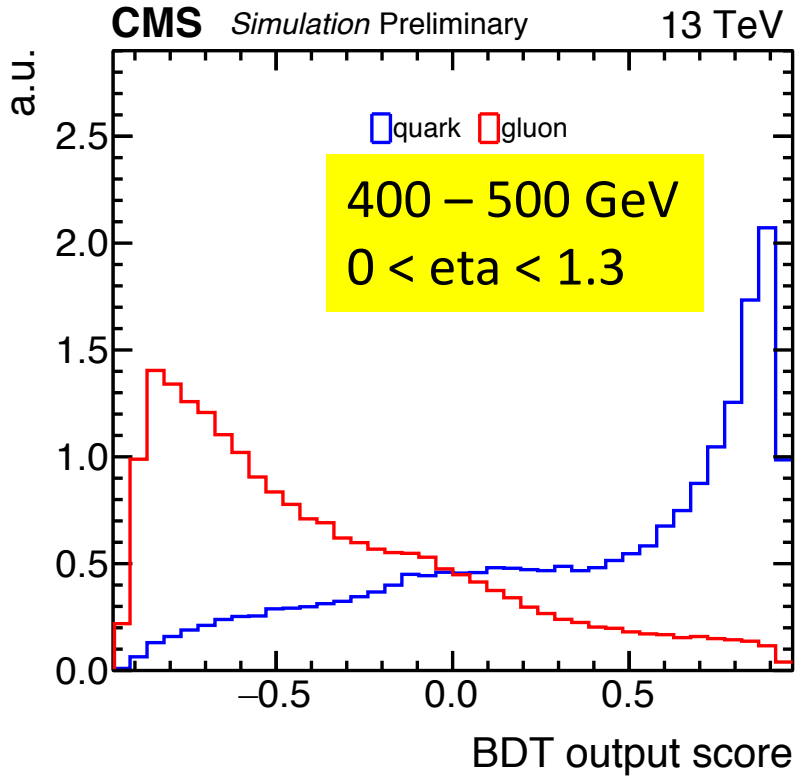




At high p_T, charged particle multiplicity brings nice discrimination power

BDT output distributions



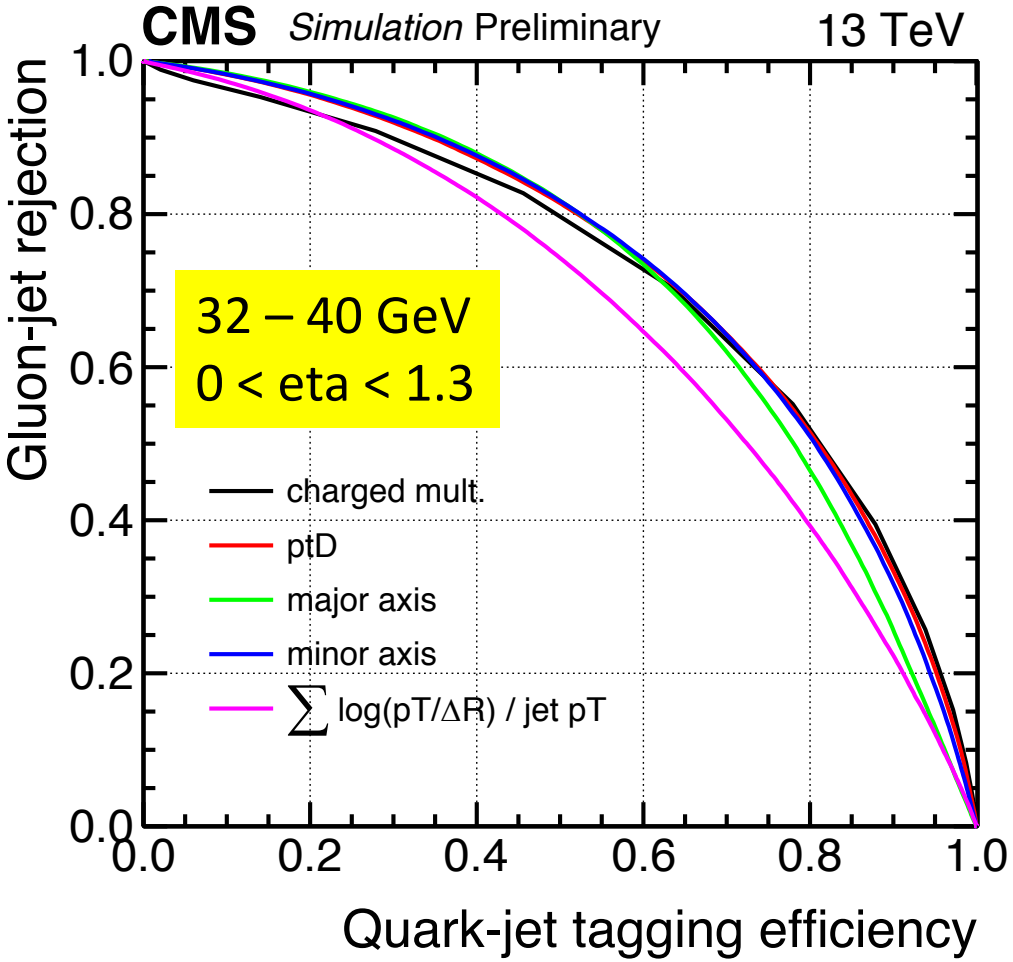


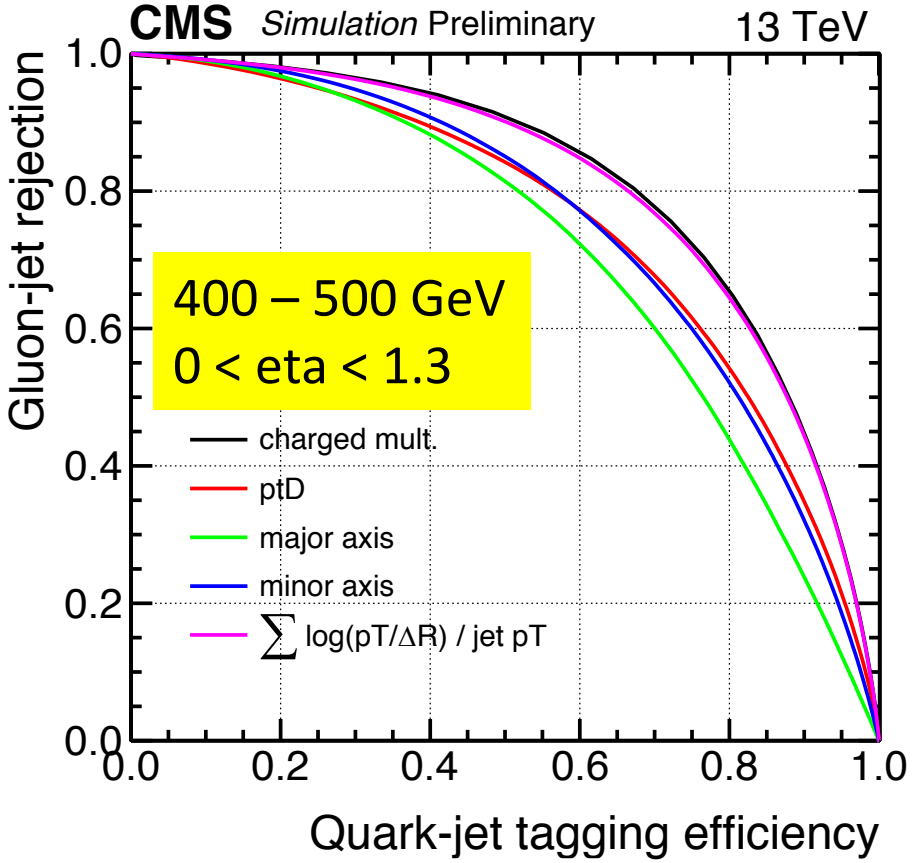
- Initial study shows that BDT has a superior performance compared to likelihood discriminator
  - Up to 10% additional rejection for given quark-tagging eff.
  - (5% coming from likelihood → BDT, 5% from revised input variables)
- Validation on-going using 2016 full dataset
  - Robust against soft emissions ? Systematics ? Comparison of MC generators

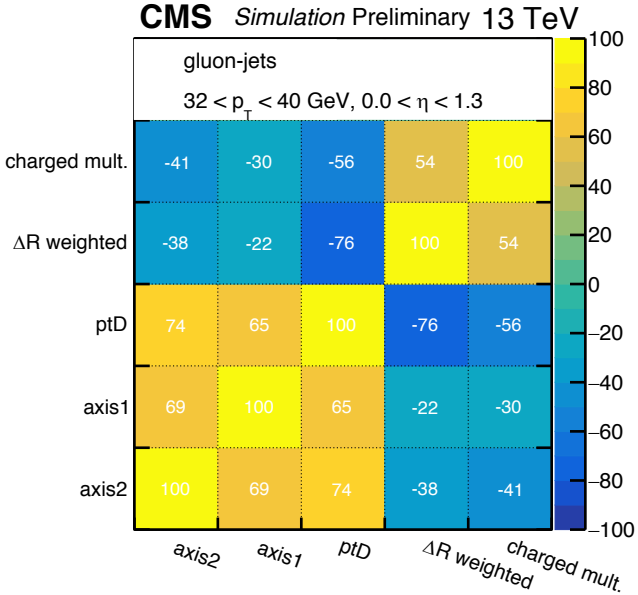
Summary

- “Classic” likelihood-based approach has been fully validated and used among CMS
  - MC corrections available for both Pythia and Herwig MC; absorb differences between different MC generators (as well as systematic uncertainties)
- On-going development for the BDT based quark-gluon tagger
  - Validation on-going with 2016 dataset
- More advanced application of machine learning : See Jan’s talk on Wednesday

ROC curves of individual variable







Linear Correlation Coefficient

Good discrimination power by itself