Quark vs. Gluon Discrimination

Giorgia Rauco
on behalf of the **CMS Collaboration**

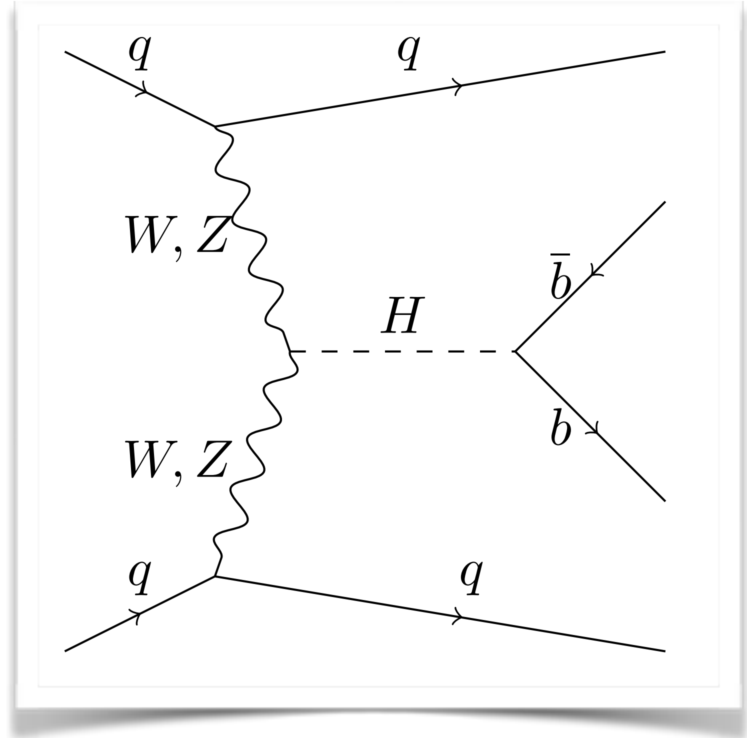
Physik Institut
Universität Zürich

Jet Substructure "Planning for the future" Event at the Fermilab LPC
November 30th, 2016

the motivation

A lot of analysis at the LHC are characterized by **hadronic final states with gluon-induced QCD backgrounds**

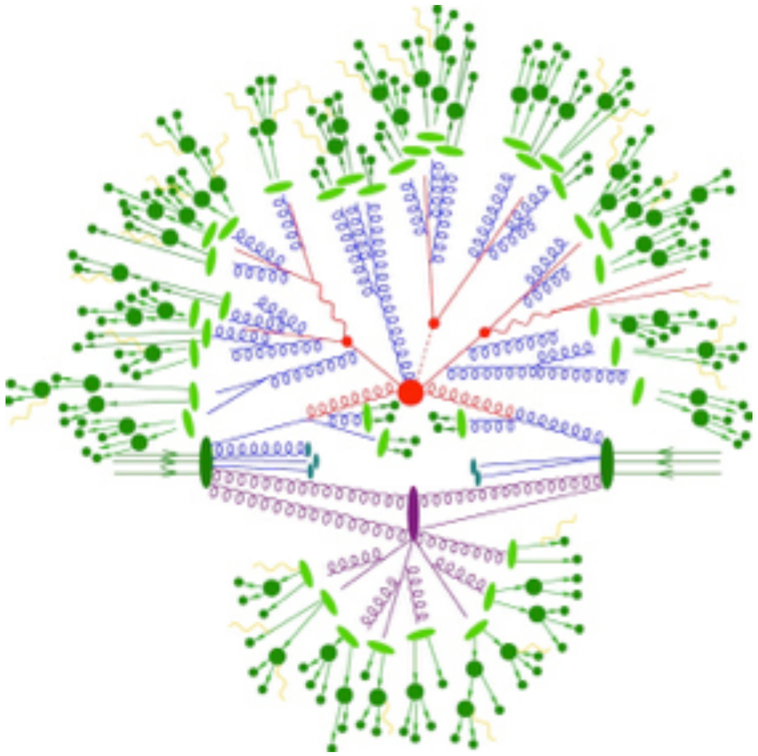
<u>example</u>: search for the Higgs boson produced through Vector Boson Fusion and decaying to a pair of b-quarks



QCD background: mainly composed by gluons
Signal: composed by quarks

↓

So having a tool able to discriminate between gluons and quarks will have a fundamental importance in **enhancing the separation between signal and background**



***Jet**: [noun] A jet is a narrow cone of hadrons and other particles produced by the hadronization of a quark or gluon*

main processes in the hadronization is gluon emission:

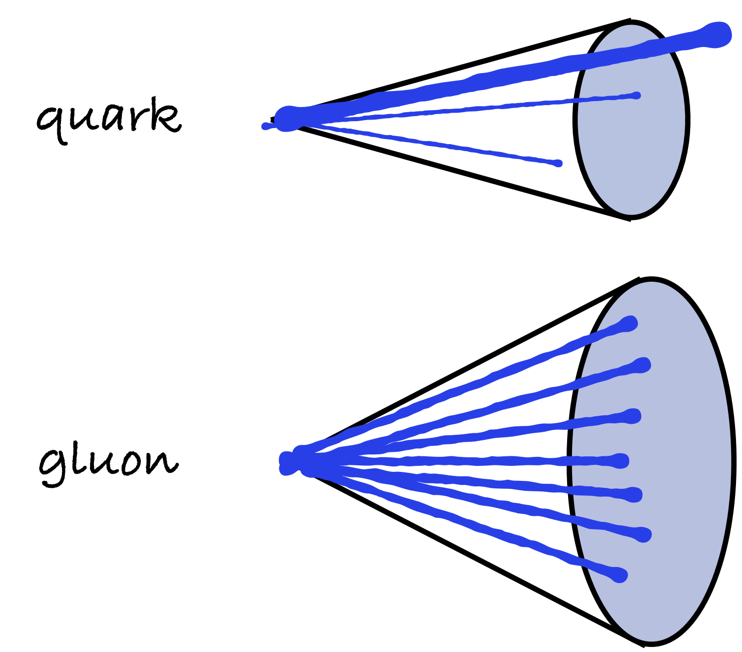
$\propto C_A = 3$ if it is a gluon

$\propto C_F = \frac{4}{3}$ if it is a quark

jets from light-flavor quarks $\neq$ jets from gluons

Main differences are:

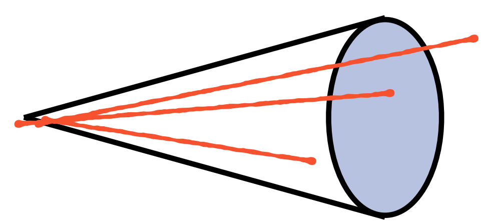
- ✳ the **particle multiplicity** is higher in gluon jets than in light-quark jets;
- ✳ the **fragmentation function** of gluon jets is considerably softer than that of a quark jet;
- ✳ gluon jets are less **collimated** than quark jets.



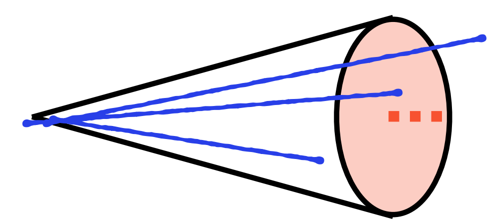
the discriminating variables

From an experimental point of view the differences between quark-like and gluon-like jets are translated into the following observables:

multiplicity

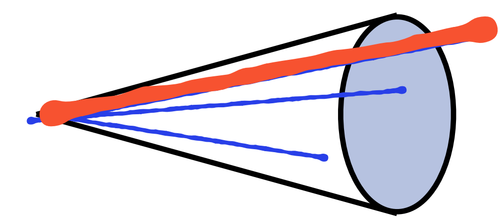


minor axis of the jet ellipse on the eta-phi plane

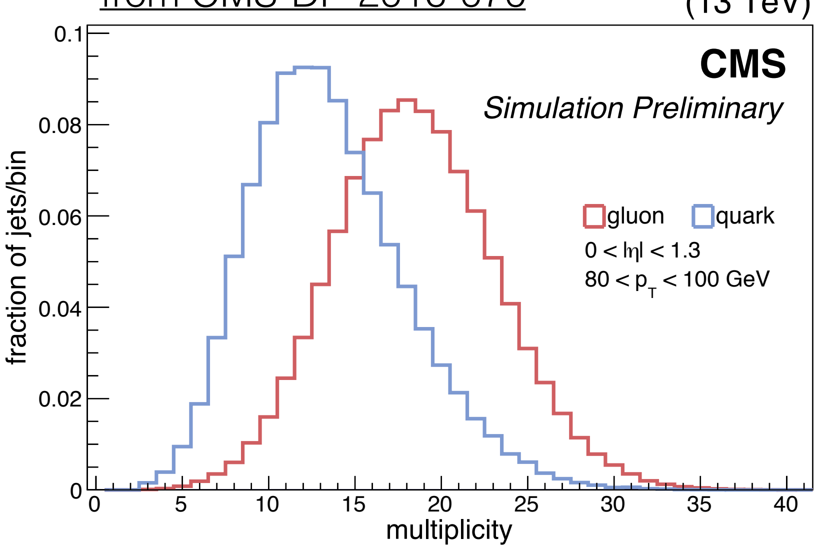


fragmentation function

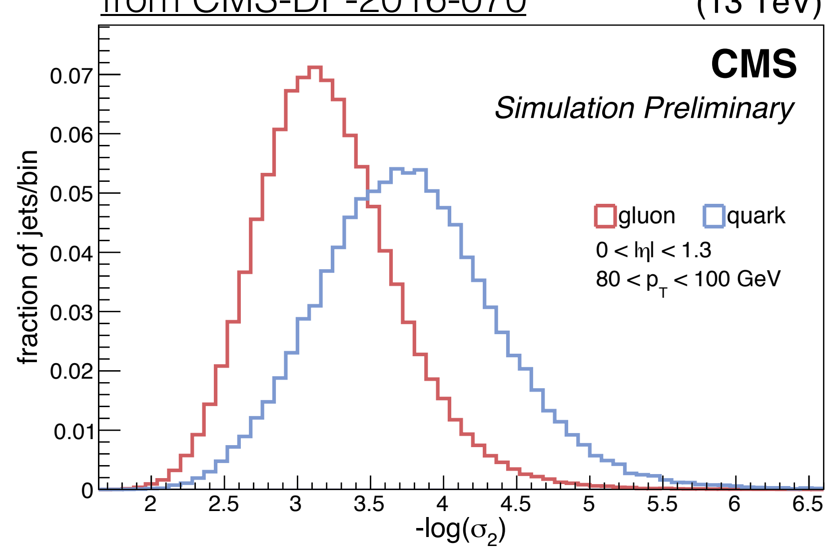
$$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}}$$



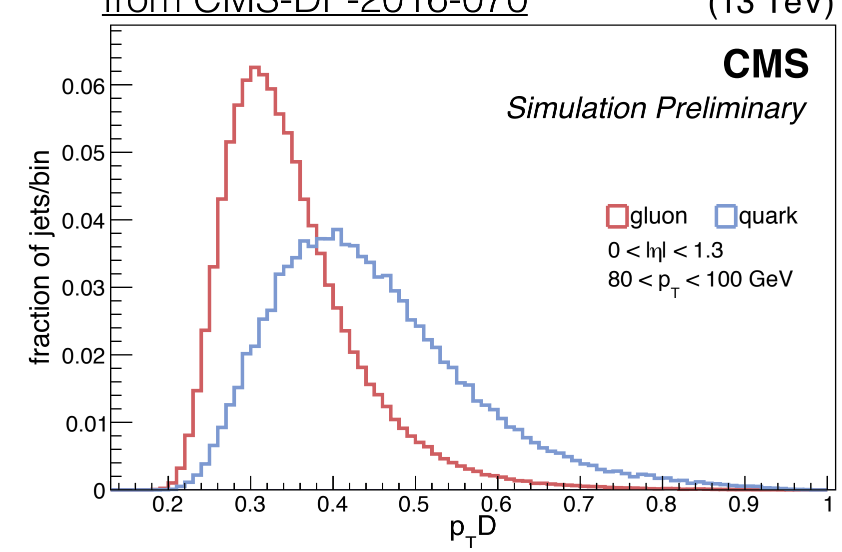
from CMS-DP-2016-070



from CMS-DP-2016-070



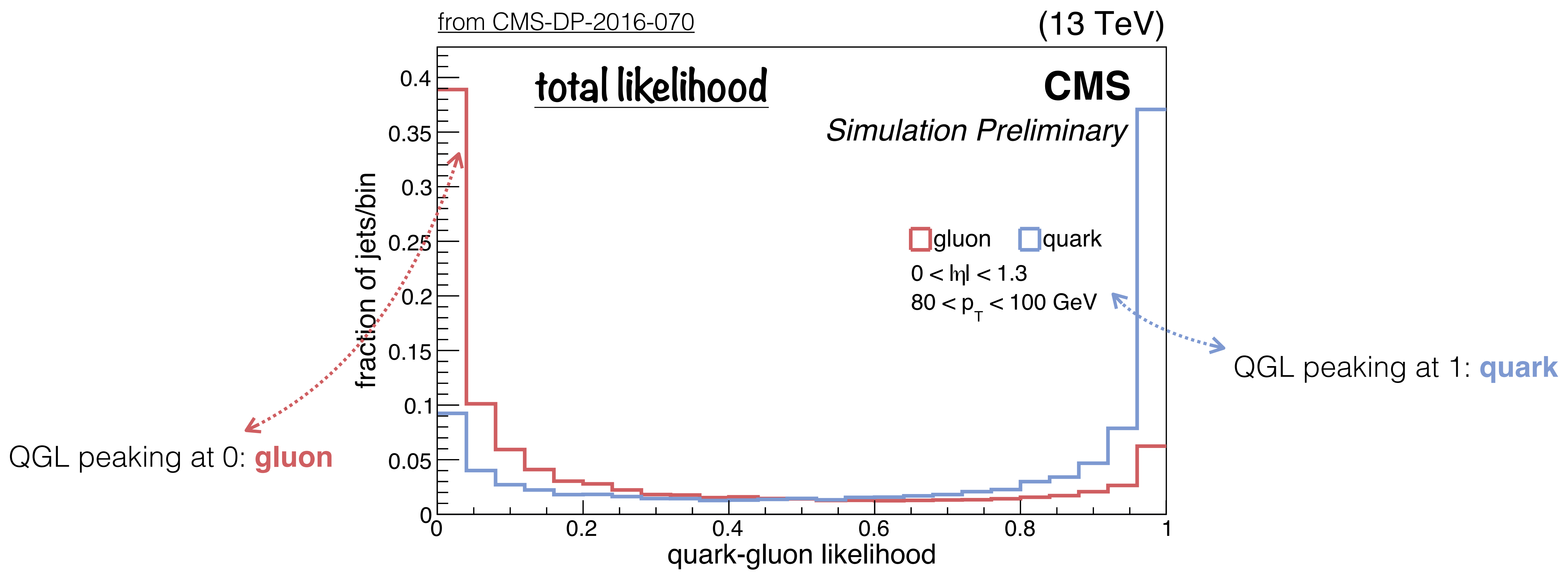
from CMS-DP-2016-070



*initial studies and PDF building on QCD dijet events showered with Pythia8

building the discriminator

- pdf's of the variables are **multiplied** to give the total likelihood
- the likelihood is determined for **several η/pt bins** (from pt > 30 GeV and across the whole η detector acceptance)
- training studies performed in simulated QCD dijet training events (PYTHIA 8)



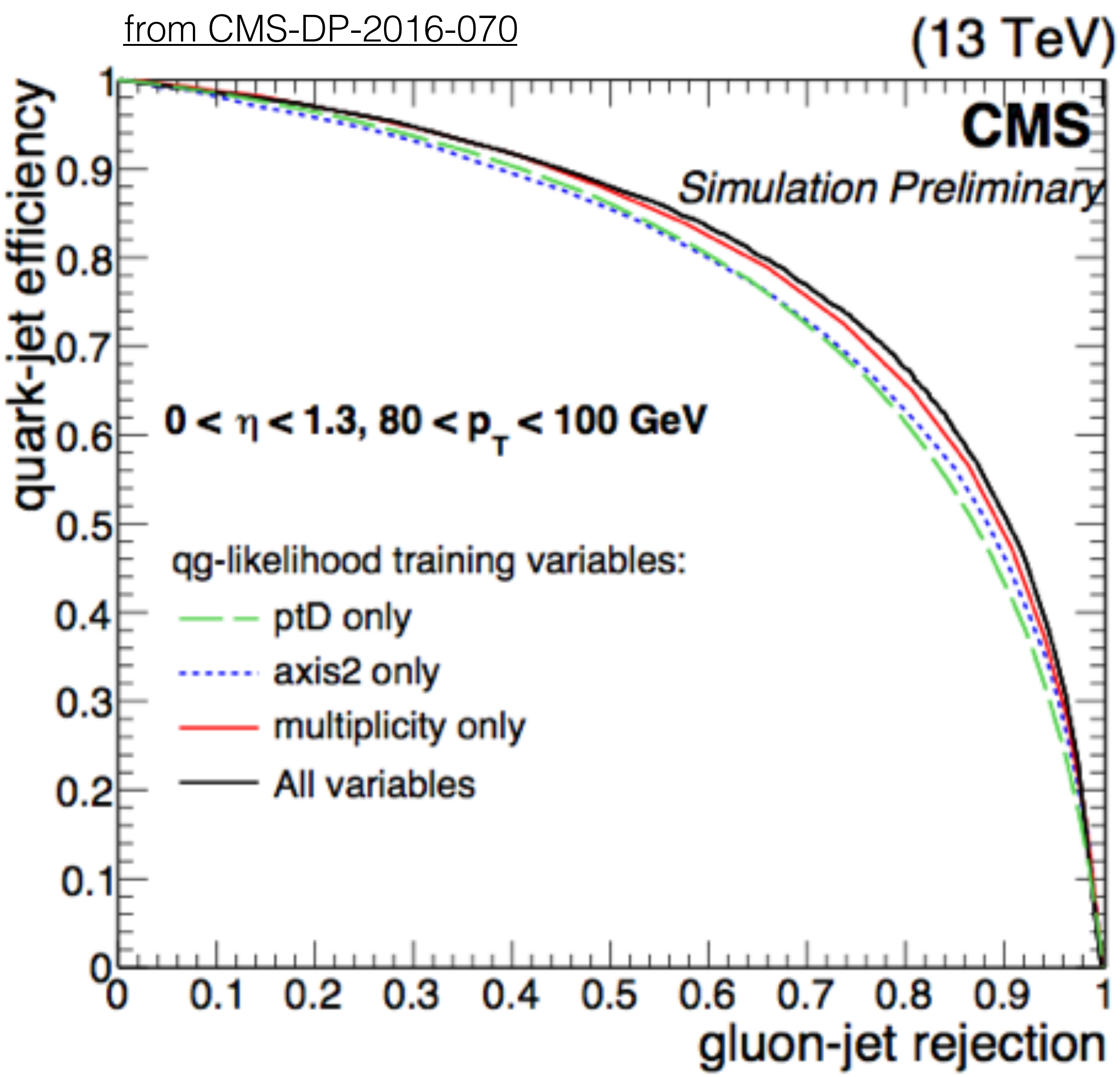
the tagger output indicates if a given jet is more likely to be originated by a quark

performances of the discriminator

discriminator performances studied on QCD simulation , comparing:

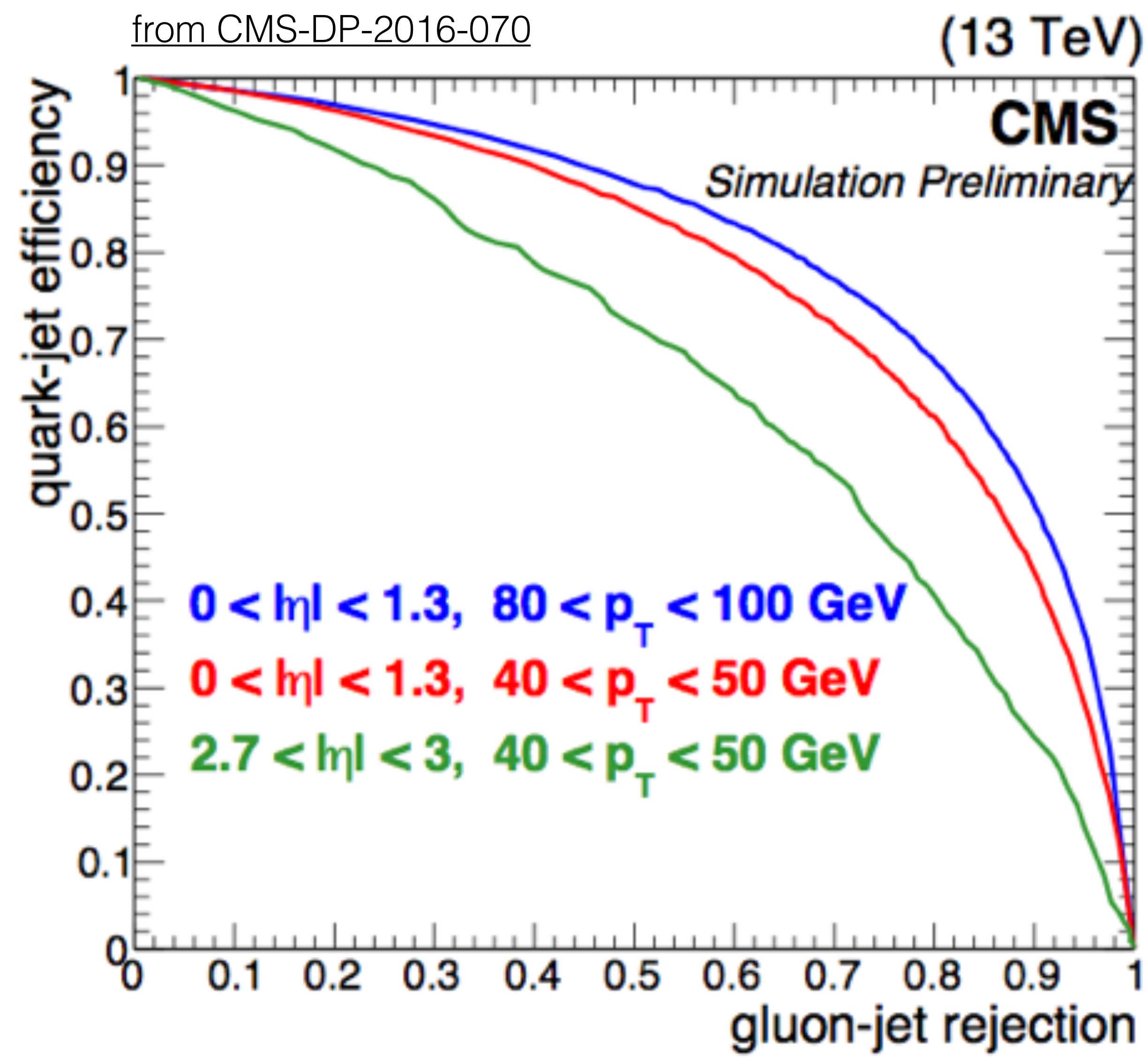
separation power of the single variables used in the training

from CMS-DP-2016-070



different kinematics regions

from CMS-DP-2016-070



the strategy of the validation on data

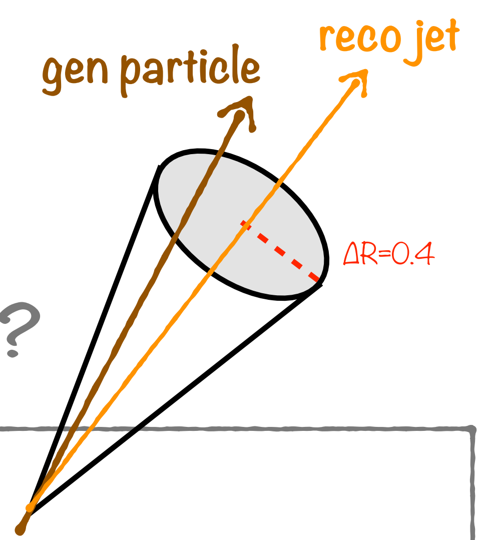
A validation of the discriminator on 13 TeV collision data has been done, using:

- **Z+jets** events, which are quark-enriched
- **dijet** events, which are gluon-enriched

the full 2015 dataset is analyzed

By the simultaneous use of these two control samples, the performance of the discriminator can be verified:

- on both parton flavors
- across the whole phase space



how is the flavor of the reconstructed jet identified?

to tag the jets a **matching strategy** is exploited

- the ΔR closest Monte Carlo generated parton, with status code = 23 (or 11 for Herwig++) to the reconstructed jet is the one giving the jet flavor
- if there is no Monte Carlo generated parton close (in a cone of radius 0.4) to the reconstructed jet, then the jet is considered as undefined

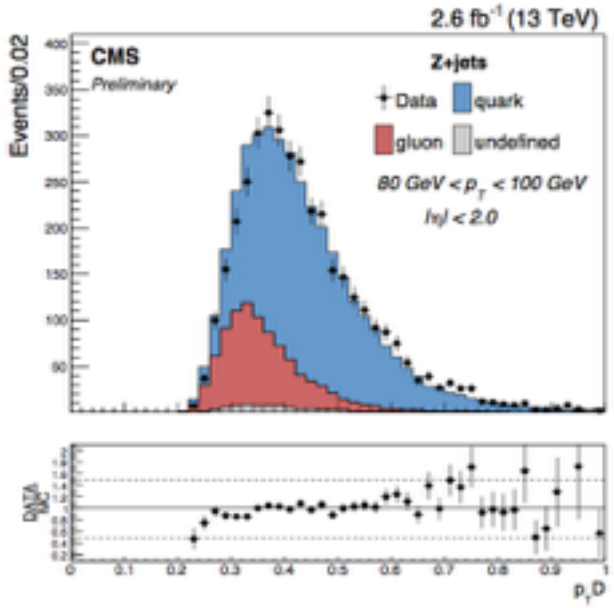
the validation on Data

training variables validation on Z+jets events →

QGL validation on both CRs →

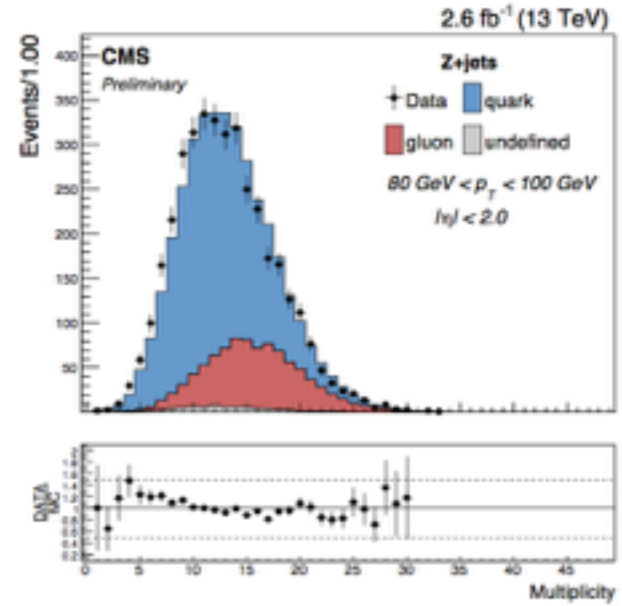
fragmentation

from CMS-DP-2016-070



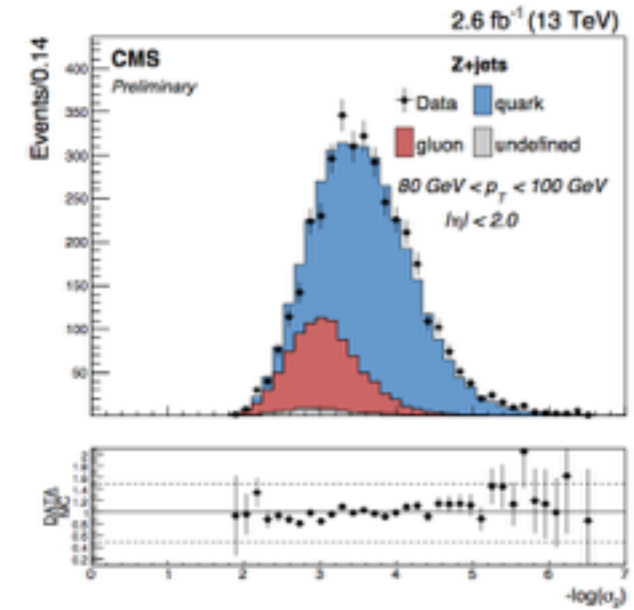
multiplicity

from CMS-DP-2016-070

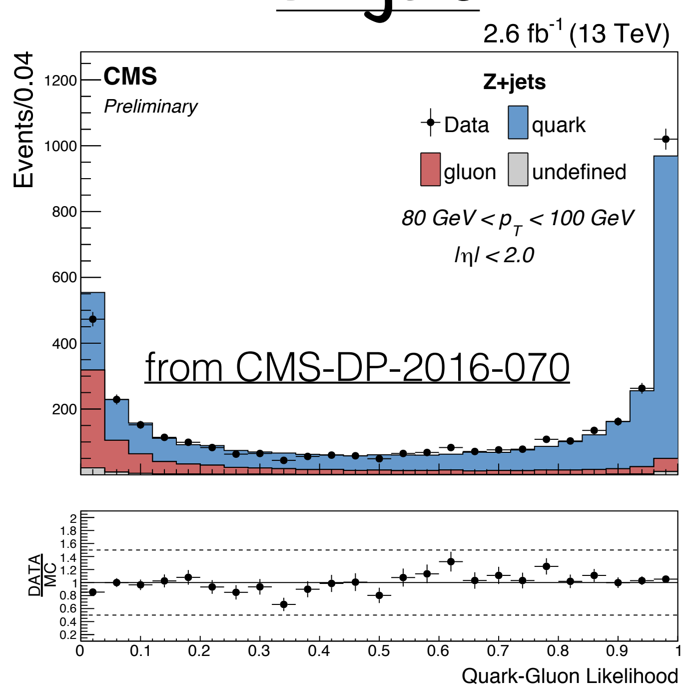


minor axis

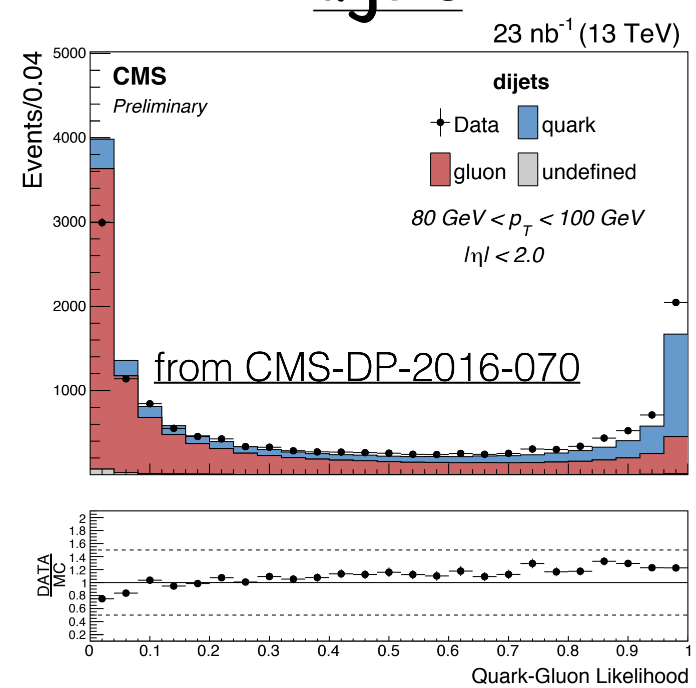
from CMS-DP-2016-070



Z+jets



dijets



fair overall data/MC agreement, but…

the systematics extraction

To account for residuals discrepancies between data and Monte Carlo scale factors are extracted, that will be applied at analysis level by the analyzers wishing to use the Quark Gluon Discriminator

- a reweighting based method has been applied
- solve a 2x2 linear system for each QGL bin (25 bins)
- taking the number of events of data and of the quark and gluon MC components for the two control samples **at the same time**

gluons quarks

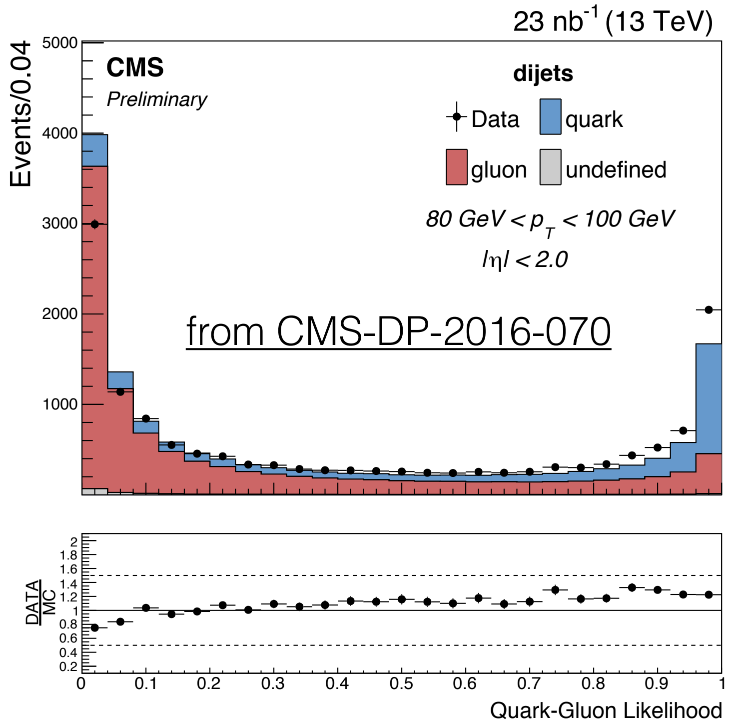
$$N^{DY}_{data} = \alpha_g N^{DY}_{MC,gluons} + \alpha_q N^{DY}_{MC,quarks} + N^{DY}_{MC,undef}$$
$$N^{QCD}_{data} = \alpha_g N^{QCD}_{MC,gluons} + \alpha_q N^{QCD}_{MC,quarks} + N^{QCD}_{MC,undef}$$

where the obtained parameters α_g and α_q are **weights** to be applied to jets

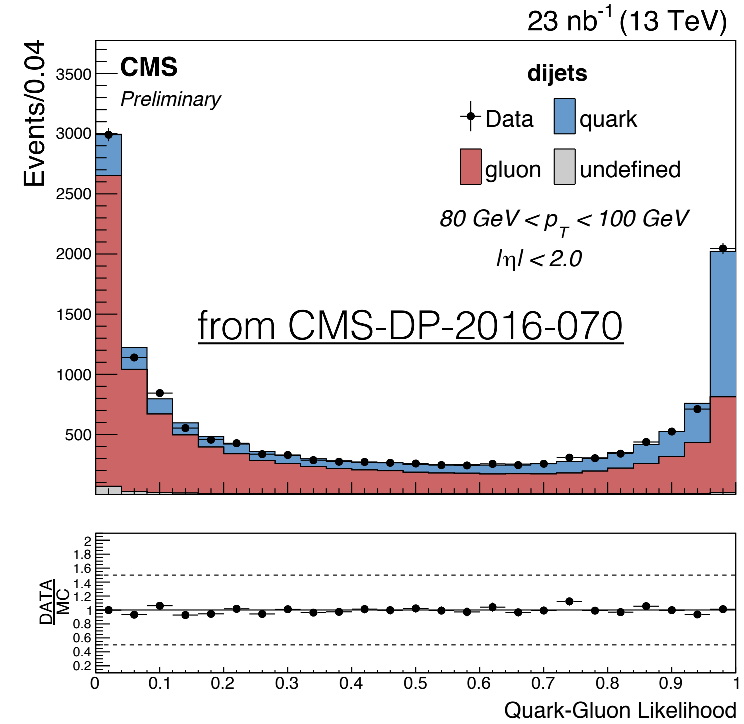
Monte Carlo remapping

the fitted weights functions are then applied on both the QCD and DY samples and the QGL MC distributions are remapped

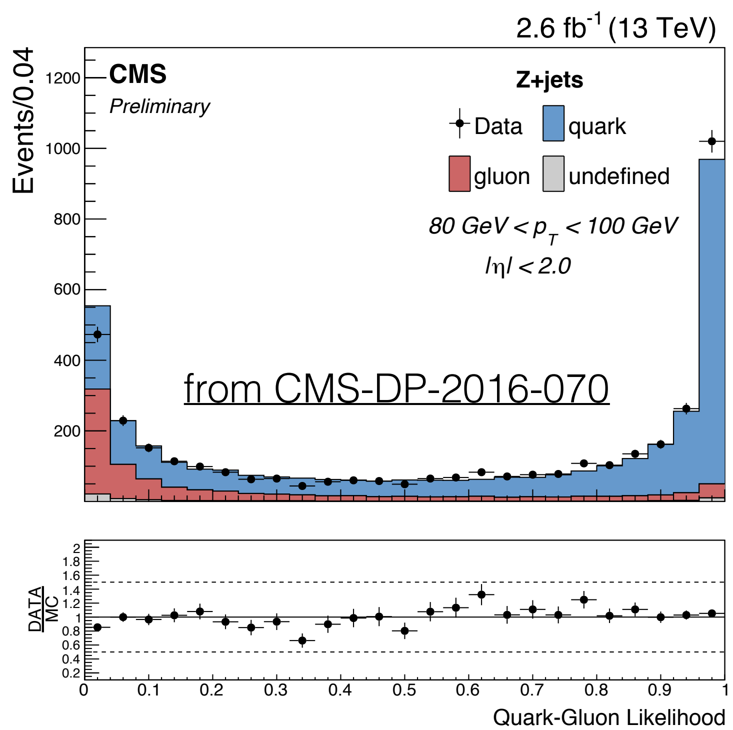
dijets

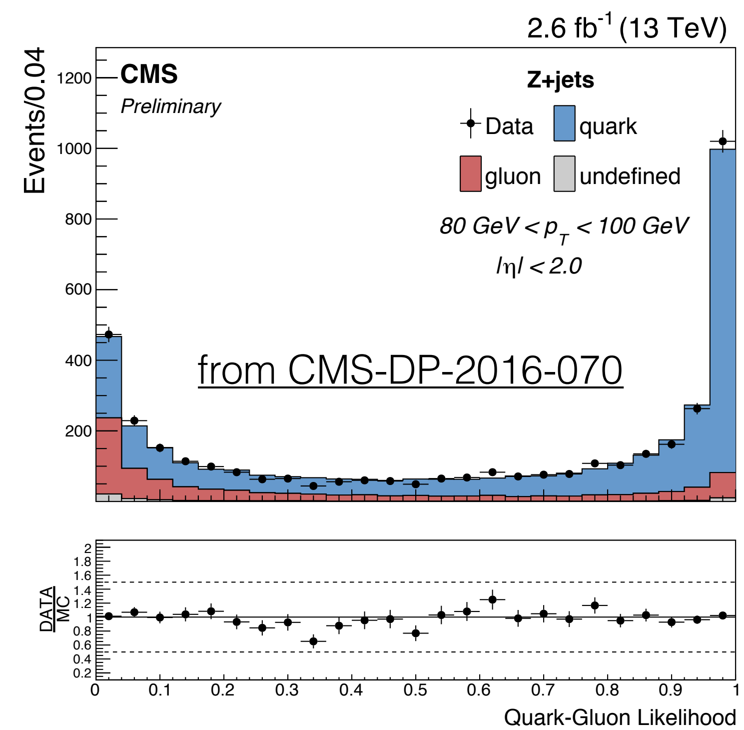


reweighting →



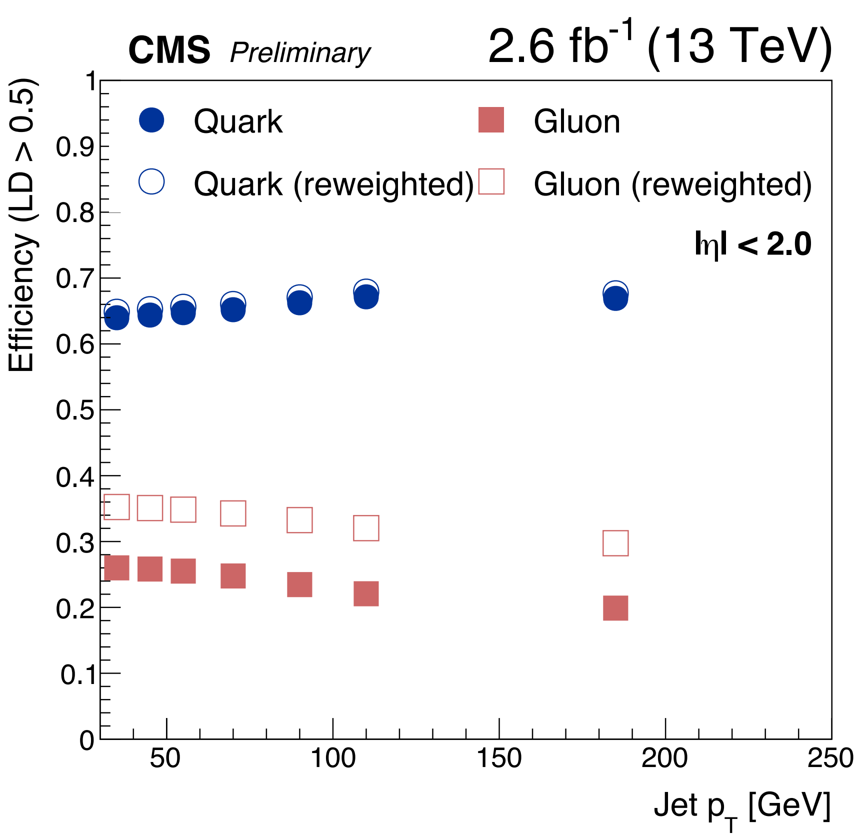
Z+jets



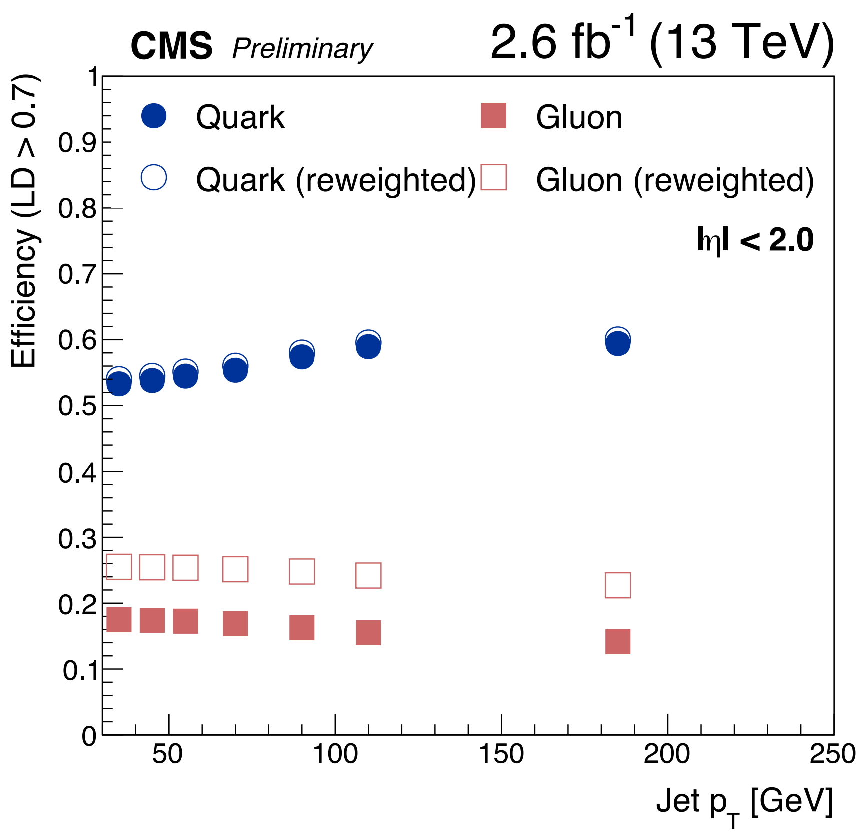


the tagging efficiencies for three Quark-Gluon Likelihood working points have been computed, depending on the jet transverse momentum

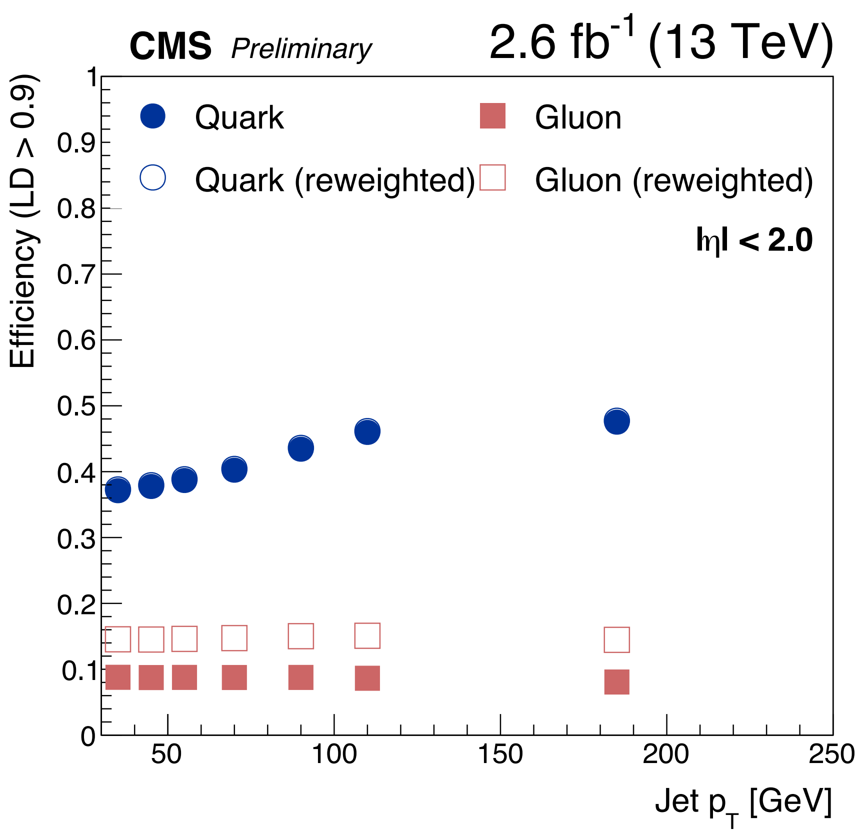
from CMS-DP-2016-070



from CMS-DP-2016-070



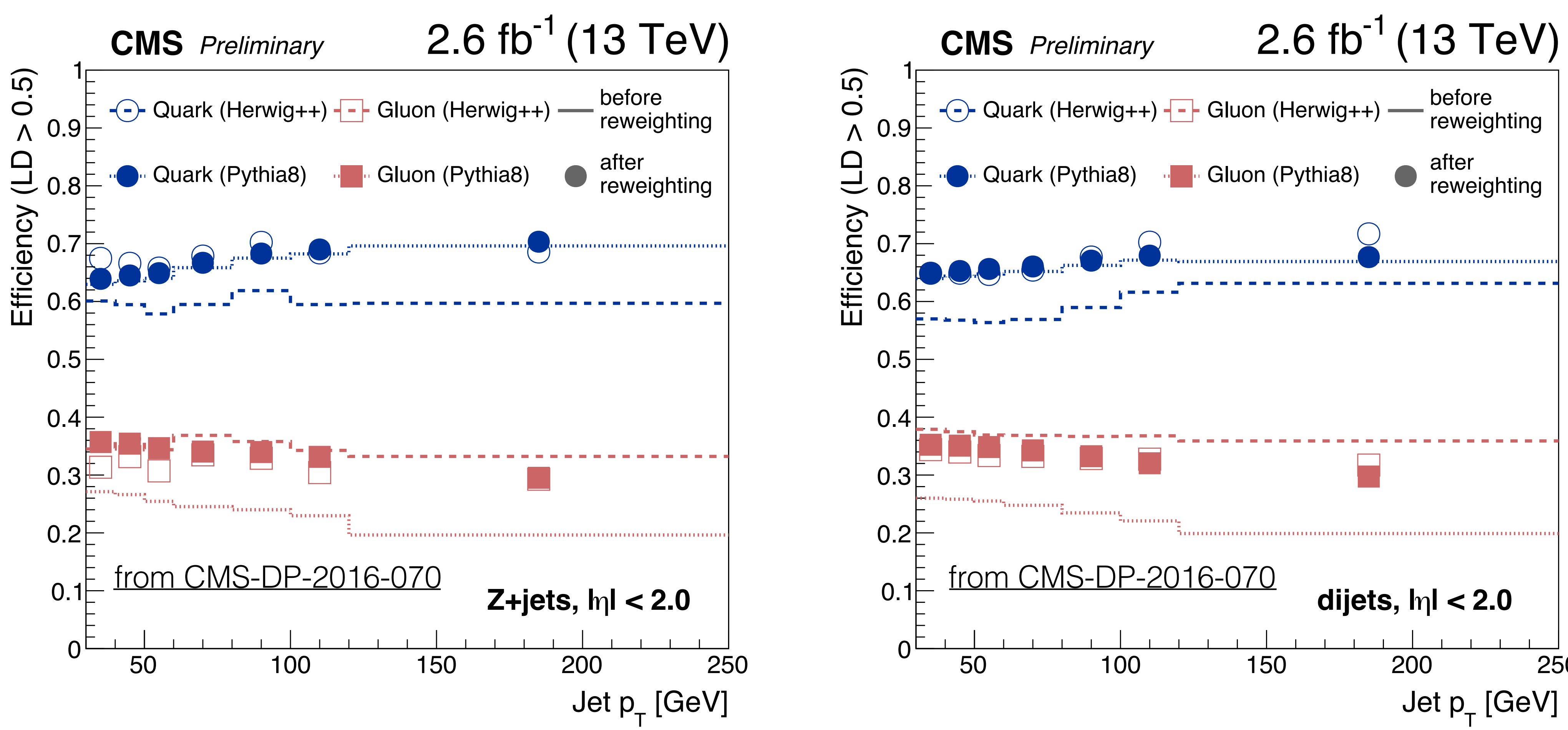
from CMS-DP-2016-070



the reweighting approach used to fix the residual Data/MC disagreements **doesn't affect the quark tagging efficiencies**, while the efficiency on **gluons change of about the 10%**

reweighting efficiencies and generator comparison

- generator comparison has also been method (Pythia8 VS Herwig++)
  - shape of the tagger
  - systematics have been rederived
- reweighting method has good performances on both parton shower
- selection efficiencies after the reweighting is very close for both generators



conclusions

- the capability to distinguish between quark-like and gluon-like jets is important for CMS analysis to improve the discrimination between signal and background
- a tool has been built based on the likelihood product between the pdf of three highly discriminating variables
- this tool provides a unique output expressing the probability for a given jet to come from the hadronization of a quark
- a validation on two control regions has been performed, to ensure a correct functioning of the tagger
- weights and systematics have been extracted to improve the shape agreement between data and Monte Carlo
- a final comparison between the performances obtained on MadGraph+Pythia8 and Herwig++ has been done

CMS-DP-2016-070

- **results presented are really new**, as they been published last week

Giorgia Rauco, giorgia.rauco@cern.ch

Thank you for the attention!

Supporting Material

Selections

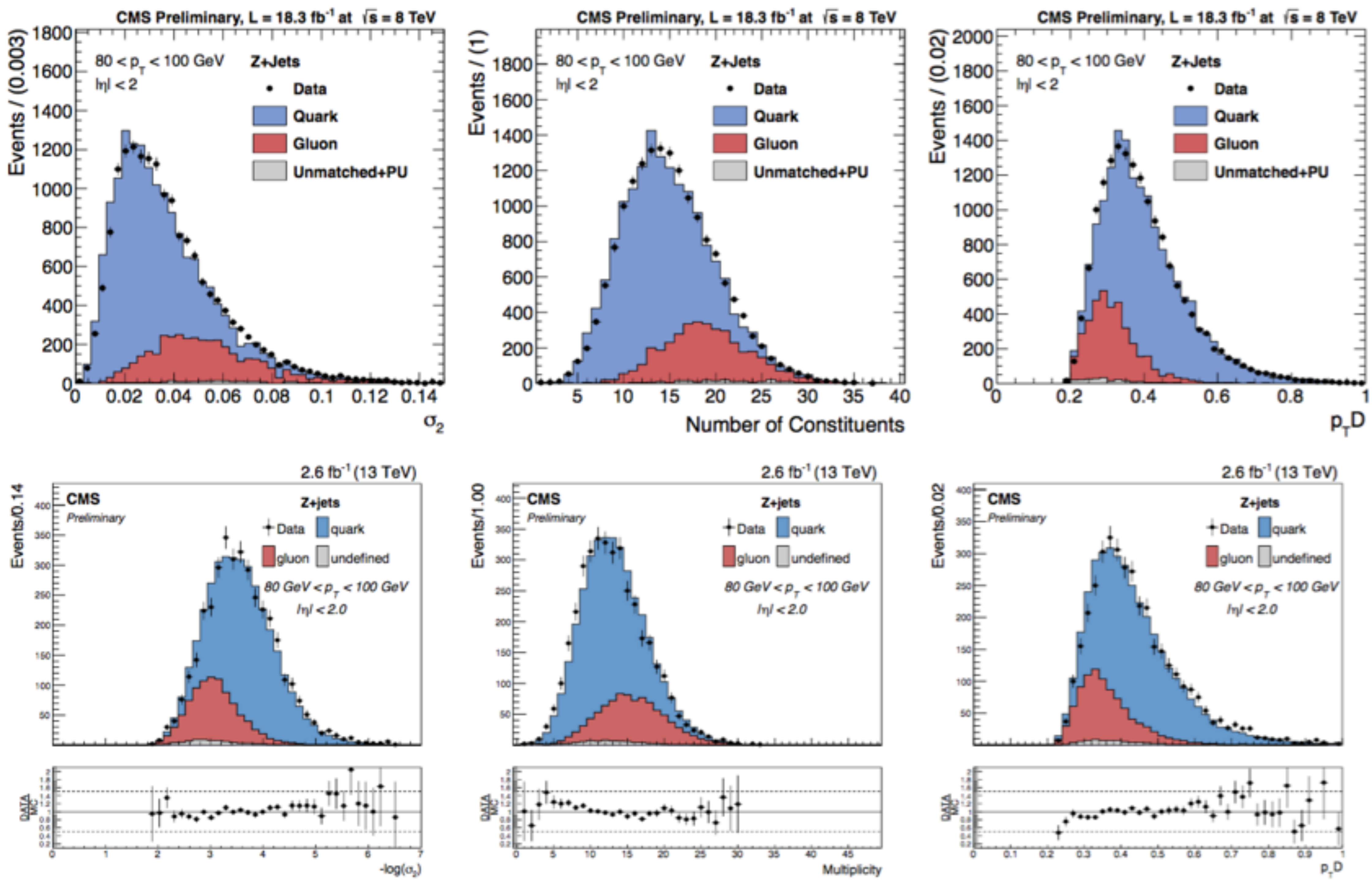
Z+jets

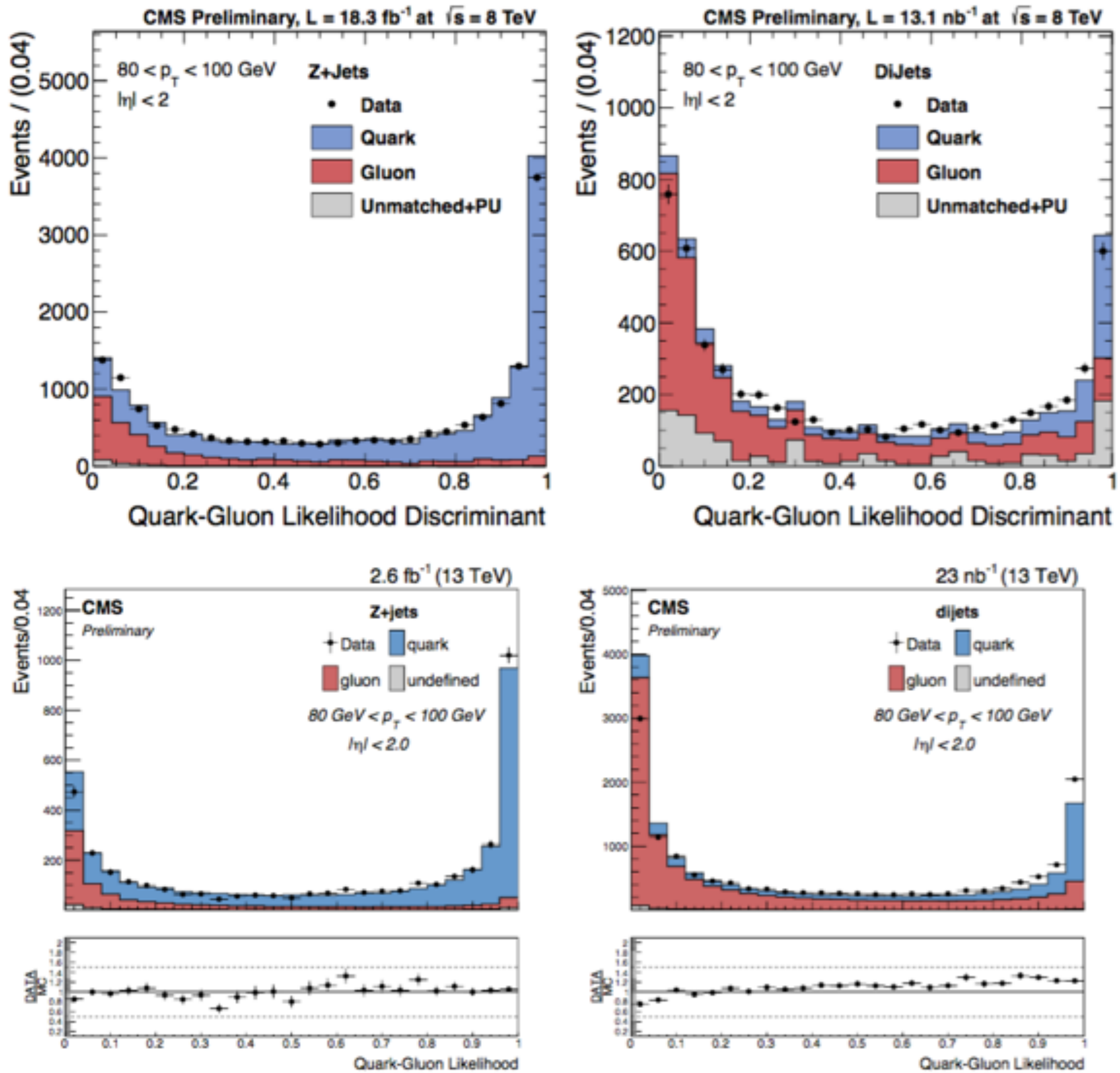
- online selection requesting two isolated muons with p_T >20 GeV
- the dimuon invariant mass to fall in the 70-110 GeV range
- the dimuon system and the (p_T) leading jet to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.1 rad
- the subleading jet in the event to have a p_T smaller than 30% of that of the dimuon system
- Drell-Yan MadGraph/pythia simulation are used

dijets

- prescaled zero bias triggers deployed
- two jets with p_T > 30 GeV
- the two p_T-leading jets to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.5 rad
- the third jet in the event to have a p_T less than 30% of the average p_T of the two leading jets
- dijet tag-and-probe approach is pursued
- QCD MadGraph/pythia simulation are used

Run1 vs Run2: Observables





CMS Preliminary, L = 18.3 fb^{-1} at $\sqrt{s}$ = 8 TeV
80 < p_T < 100 GeV
|η| < 2
Z+Jets
Data
Quark
Gluon
Unmatched+PU
Events / (0.04)
Quark-Gluon Likelihood Discriminant
CMS Preliminary, L = 13.1 nb^{-1} at $\sqrt{s}$ = 8 TeV
80 < p_T < 100 GeV
|η| < 2
DiJets
Data
Quark
Gluon
Unmatched+PU
Events / (0.04)
Quark-Gluon Likelihood Discriminant
2.6 fb^{-1} (13 TeV)
CMS
Preliminary
Z+jets
Data
quark
gluon
undefined
80 GeV < p_T < 100 GeV
|η| < 2.0
Events/0.04
Quark-Gluon Likelihood
23 nb^{-1} (13 TeV)
CMS
Preliminary
dijets
Data
quark
gluon
undefined
80 GeV < p_T < 100 GeV
|η| < 2.0
Events/0.04
Quark-Gluon Likelihood

Run1 vs Run2: ROCs