Distinguishing quark and gluon jets in the CMS experiment

Giorgia Rauco
Supervisors: Prof. Florencia Canelli and Dr. Annapaola de Cosa

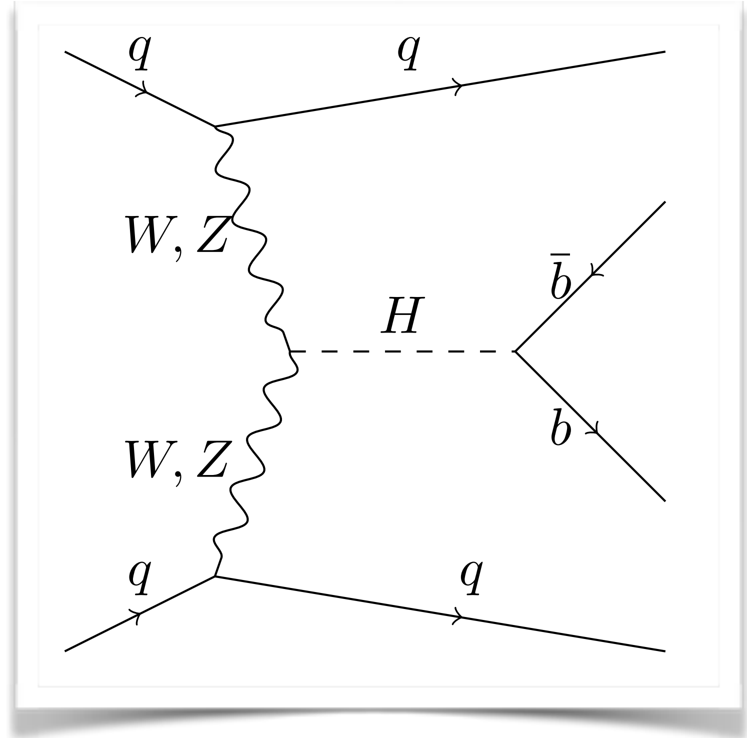
Physik Institut
Universität Zürich

Zürich PhD seminar
November 24th, 2016

the motivation

A lot of analysis at the LHC are characterized by **full hadronic final states** and suffer mainly form **QCD multijets background**

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



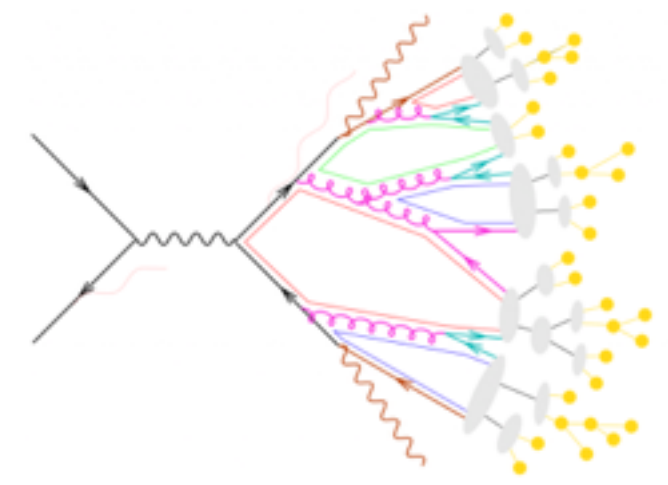
QCD background: mainly composed by gluons
Signal: mainly composed by quarks

↓

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

the theoretical background

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



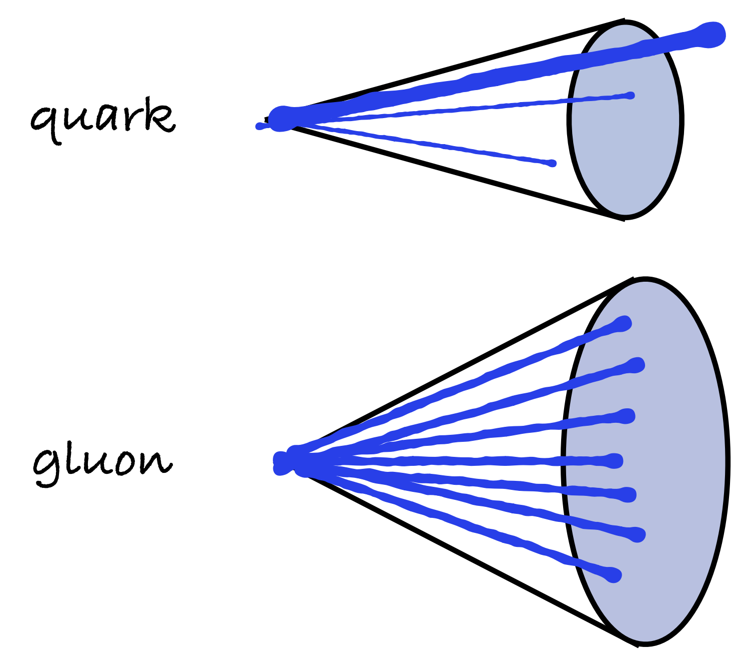
parton shawer in Drell Yan process

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



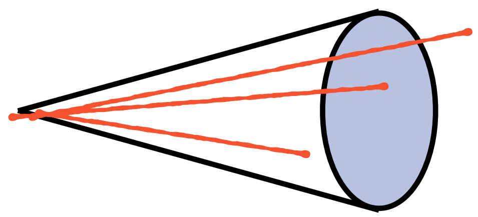
Main differences are:

- ✳ 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.

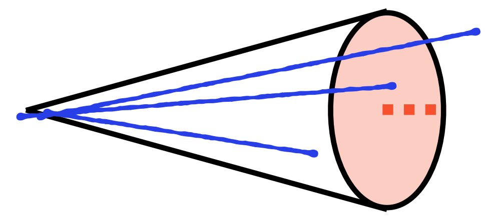
the discriminating variables

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

multiplicity

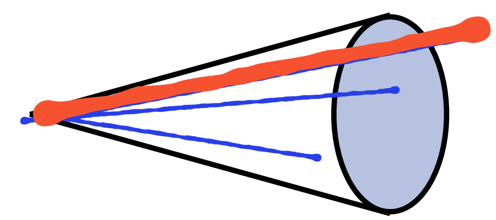


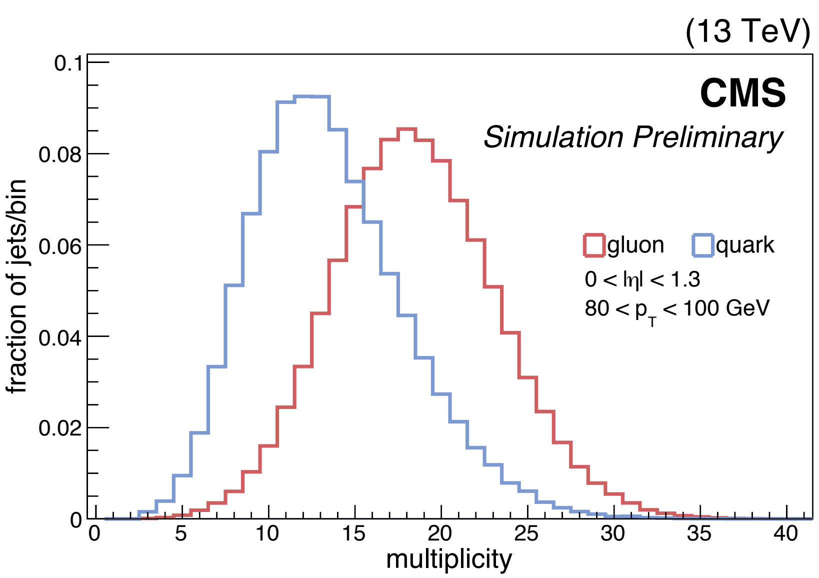
minor axis of the jet ellipse on the eta-phi plane

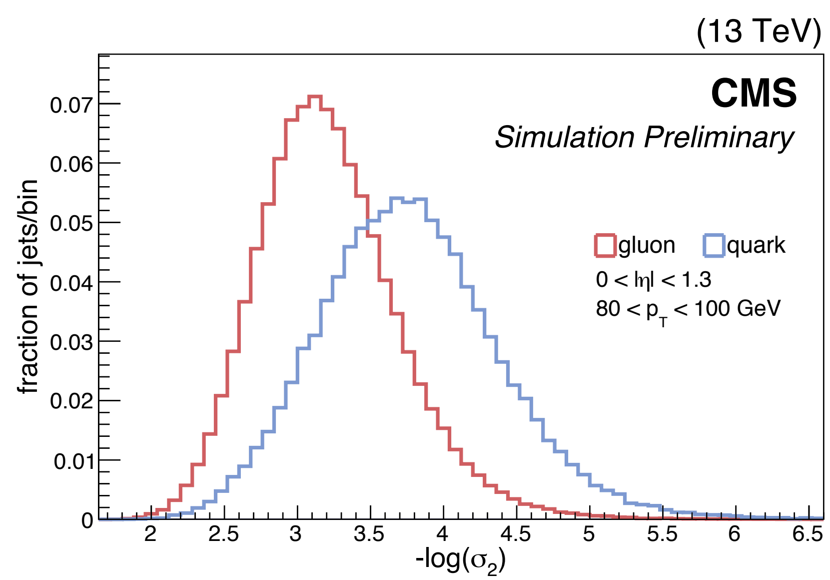


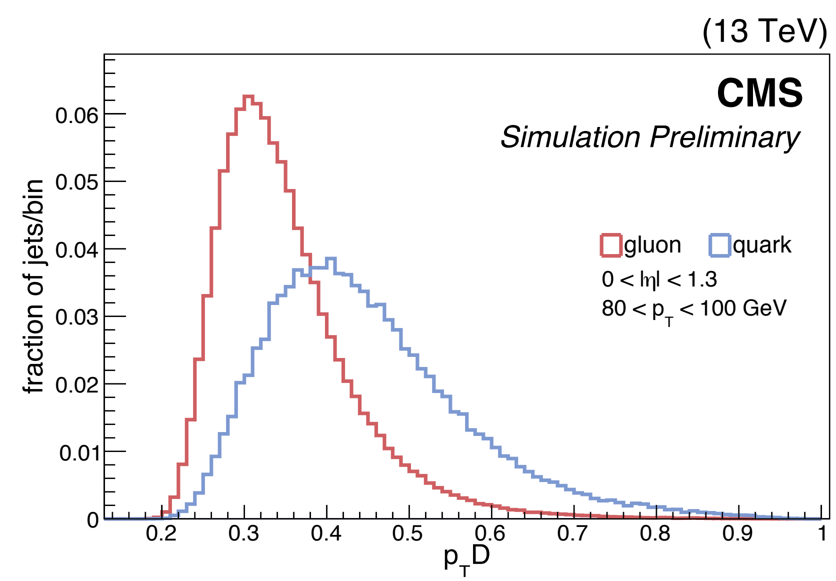
fragmentation function

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





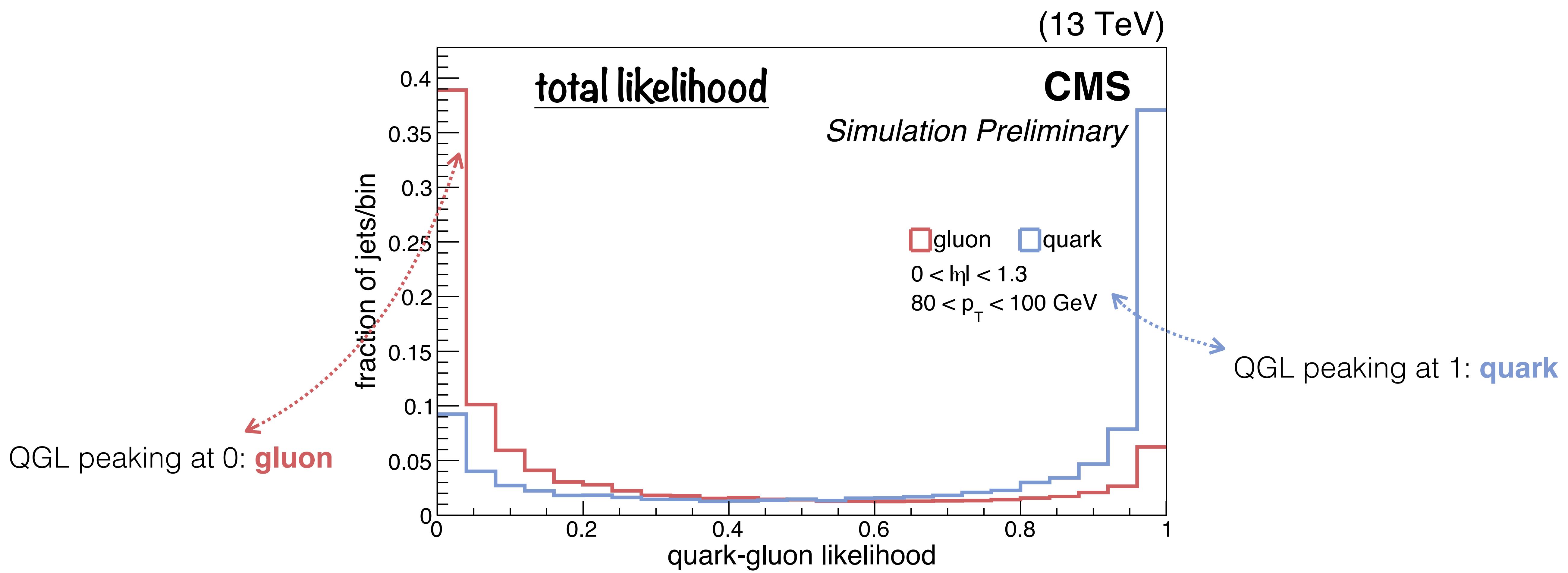




building the discriminator

- 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)

a **likelihood-discriminator** is built

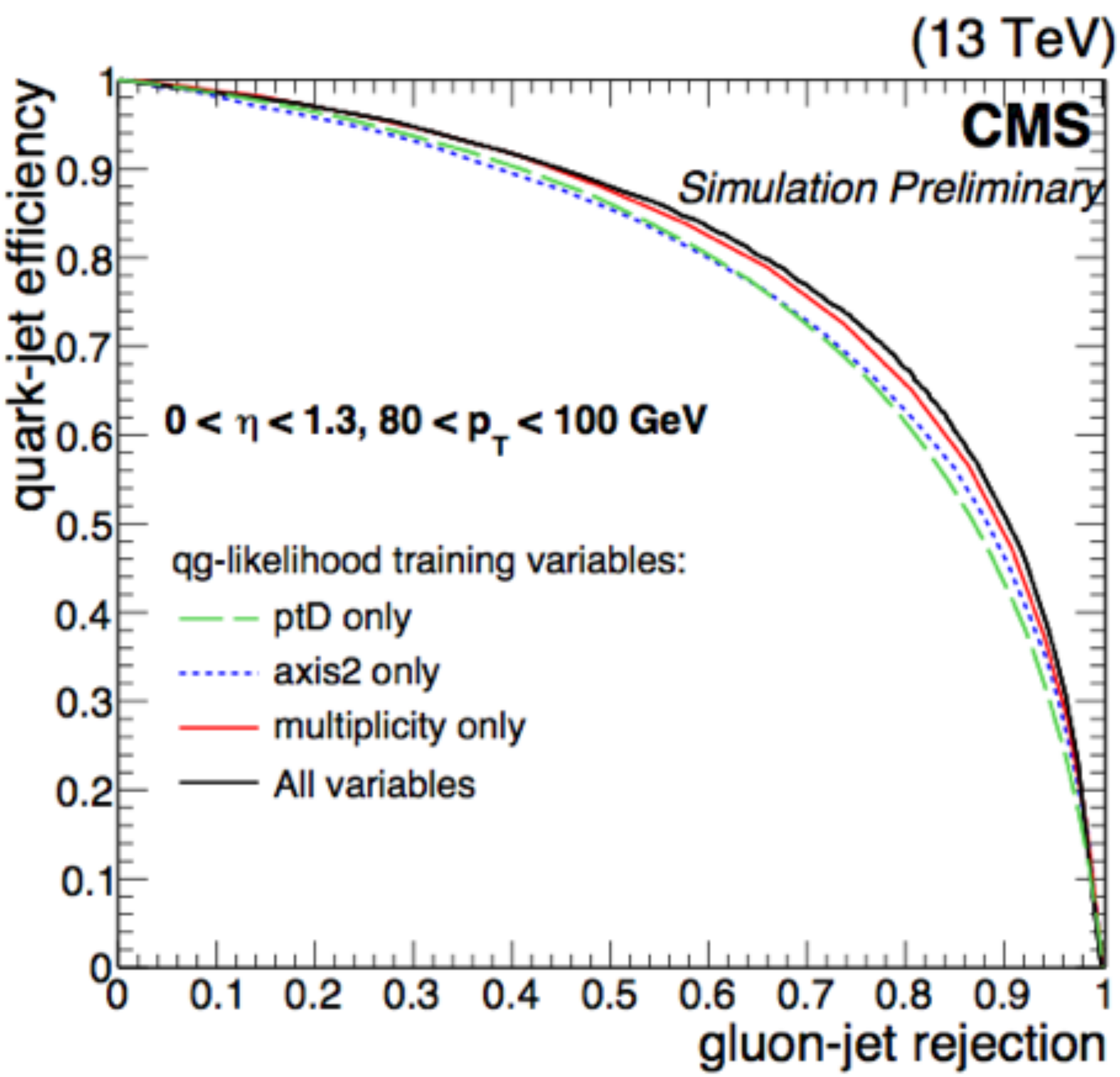


the tagger expresses the probability for a given jet to be most likely originated by a quark

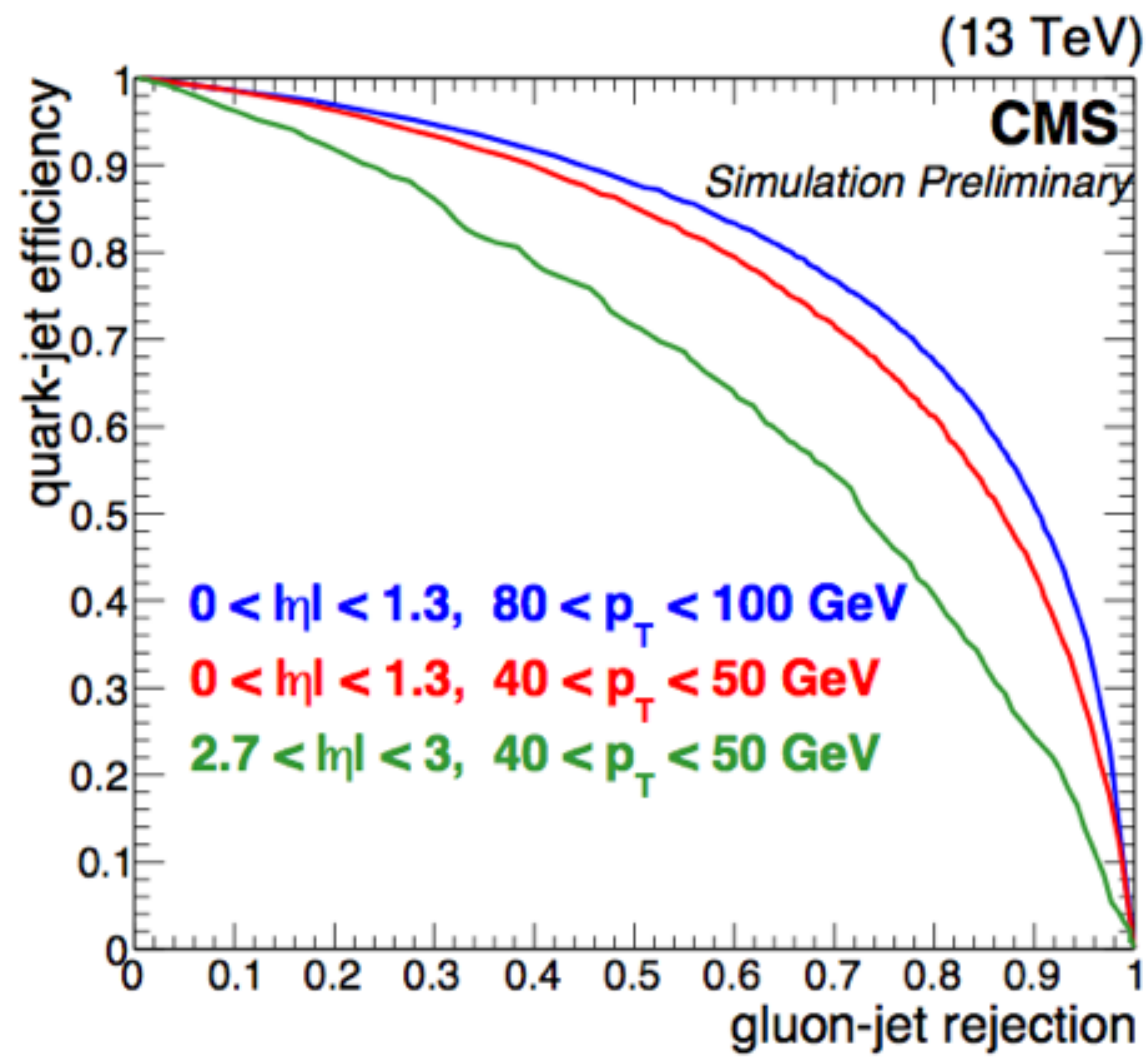
performances of the discriminator

discriminator performances studied on QCD simulation , comparing:

strength of the single variables used in the training



different kinematics regions



the strategy of the validation on data

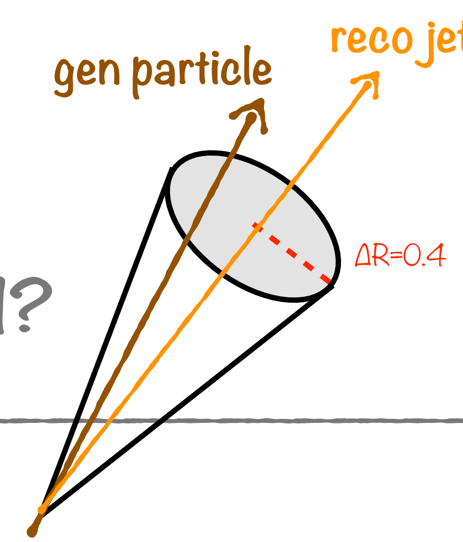
It was important to verify the correct functioning of the discriminator on:

- on both parton flavors
- across the whole phase space

the full 2015 dataset is analyzed

Two control regions are used at the same time:

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



how is the flavor of the reconstructed jet identified?

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

- the closest Monte Carlo generated parton 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

Selections

Z+jets

- two muons of opposite charge 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

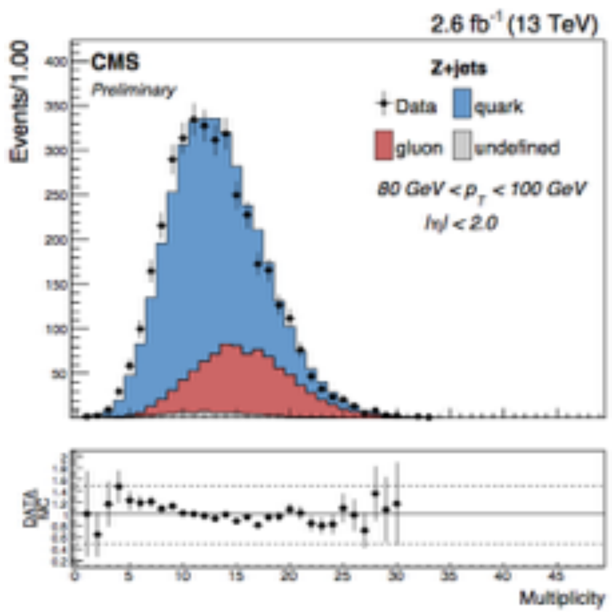
dijets

- 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
- Drell-Yan MadGraph/pythia simulation are used

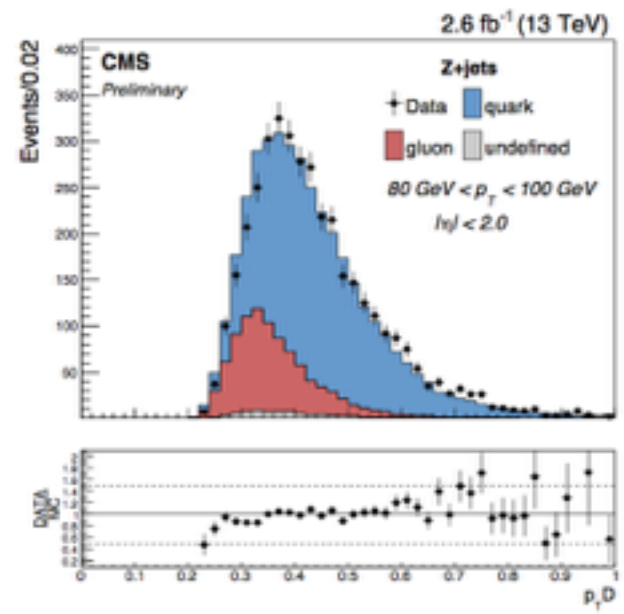
the validation on Data

training variables validation on Z+jets events →

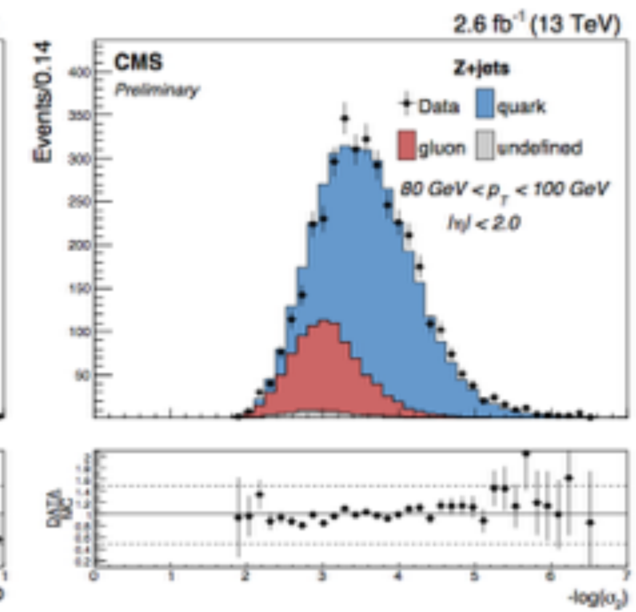
multiplicity



fragmentation

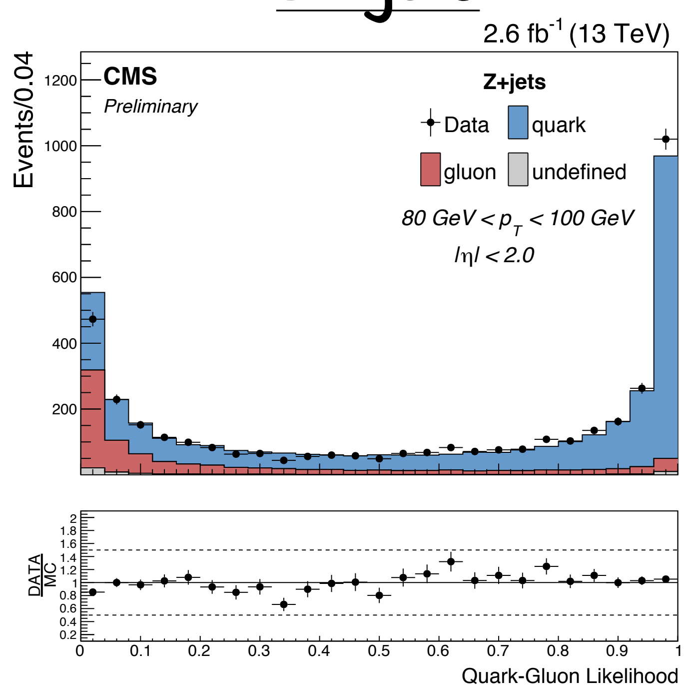


minor axis

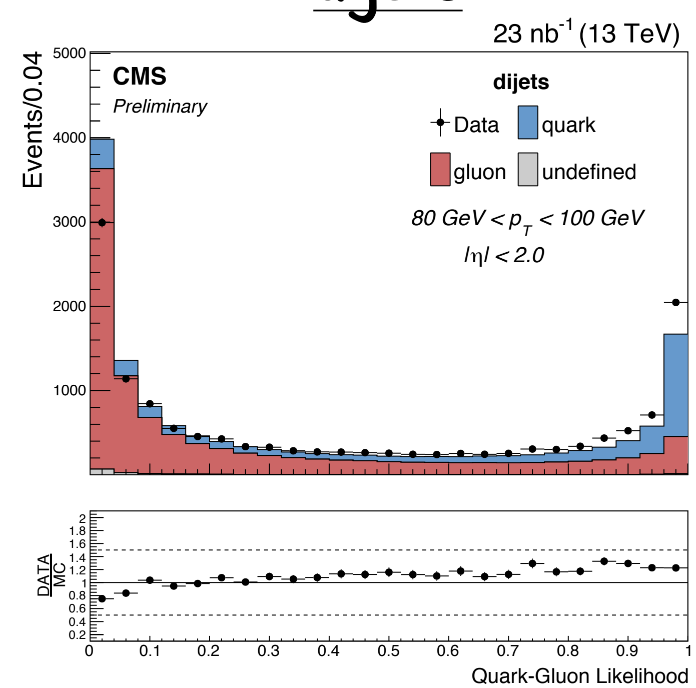


QGL validation on both CRs →

Z+jets



dijets



good overall data/MC agreement, but…

the systematics extraction

… to account for residuals discrepancies between data and Monte Carlo scale factors are extracted

- a reweighting based method have 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**

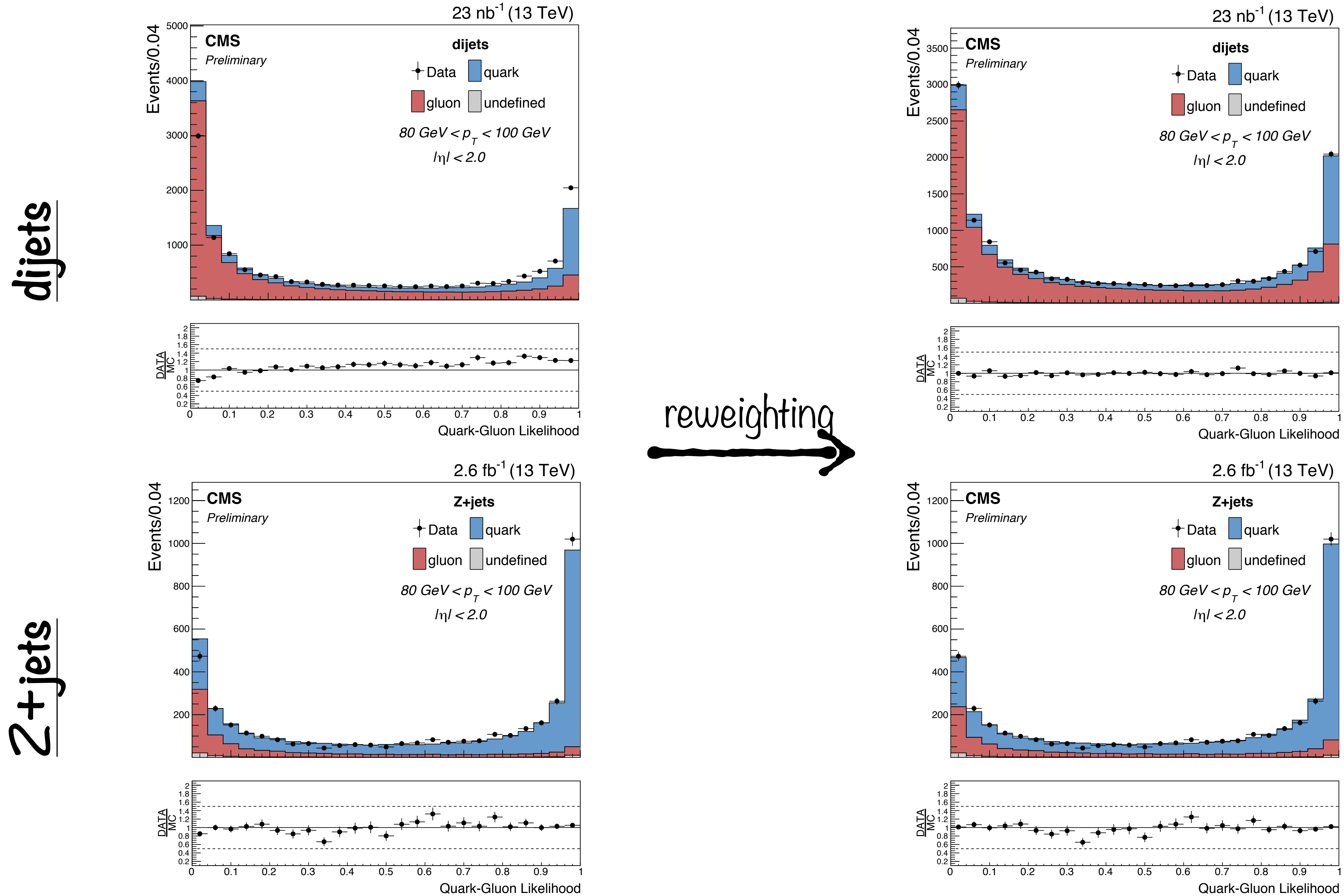
gluons quarks

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

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

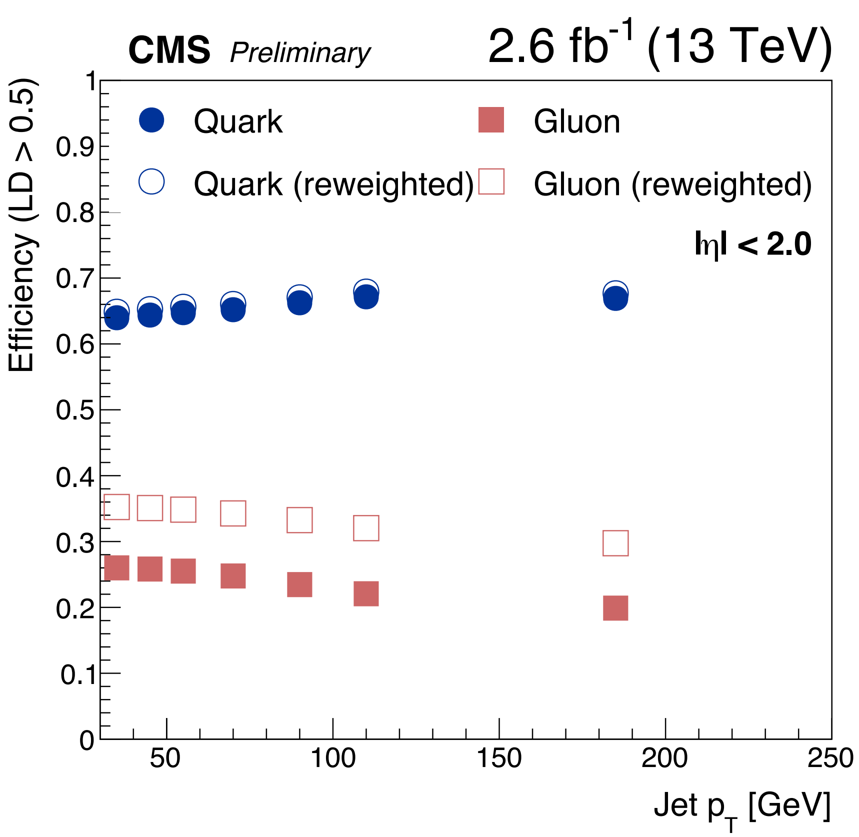
Monte Carlo remapping

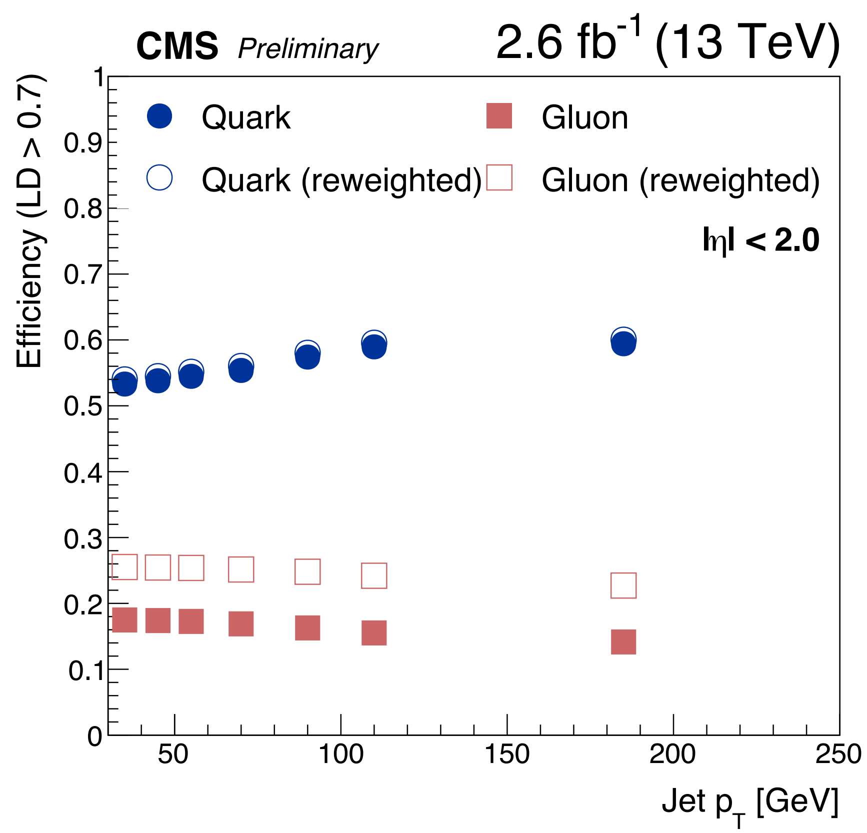
the fitted weights functions are then applied **at the same moment** on both the QCD and DY samples and the QGL MC distributions are remapped

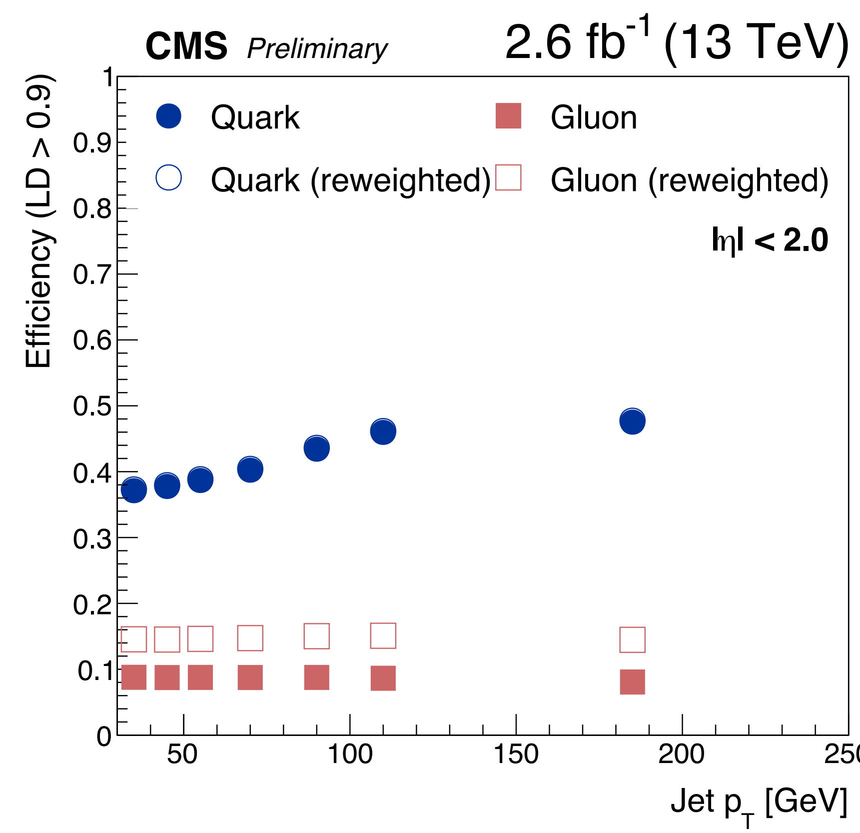


tagging efficiencies

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



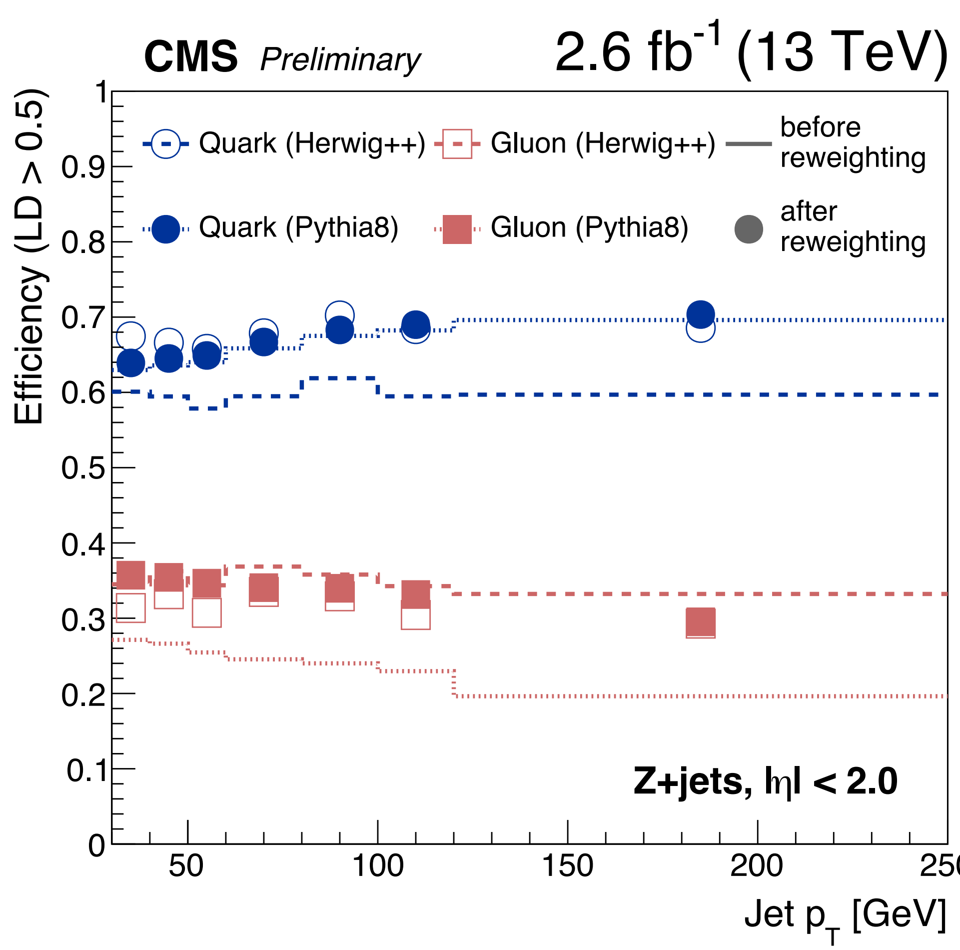


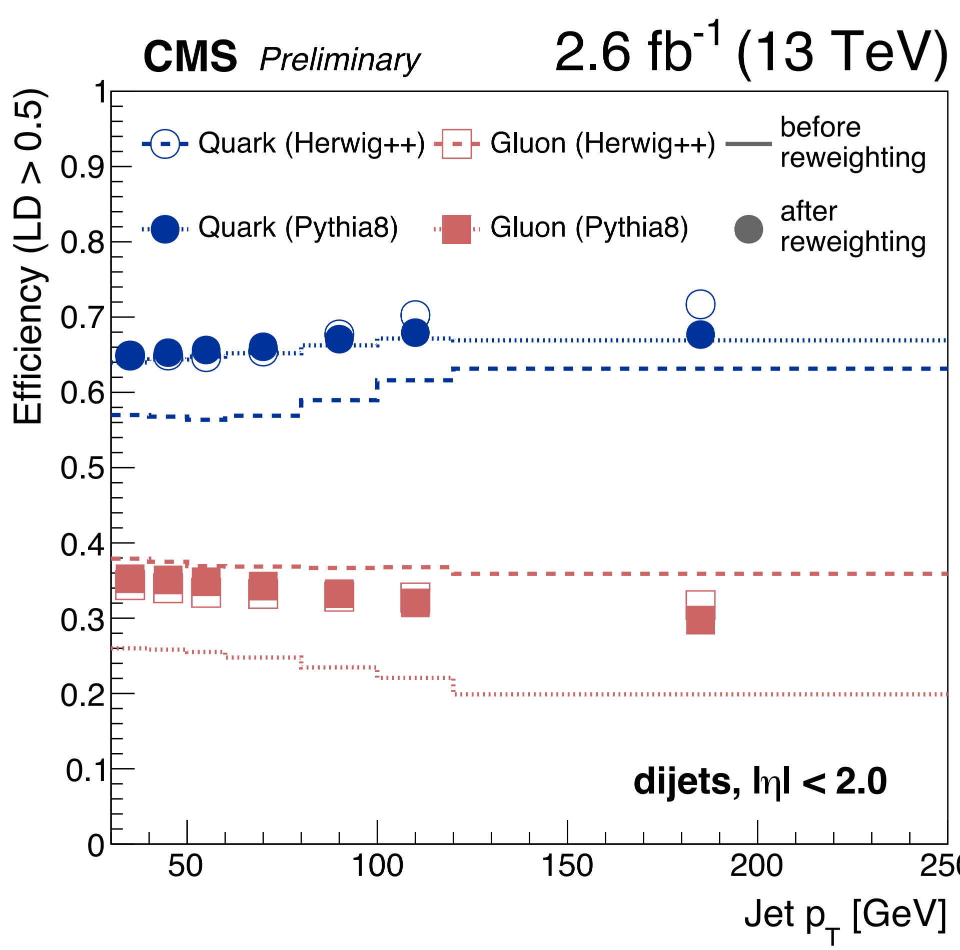


the reweighting approach used to fix the residual Data/MC disagreements **doesn't affect the quark tagging efficiencies**, while the efficiency on **gluons change of about the 10%**

reweighting efficiencies and generator comparison

- generator comparison has also been performed (Pythia8 VS Herwig++)
  - shape of the tagger
  - systematics have been rederived
- some disagreements with *out-of-the-box* MC between the generators, mostly for gluons
- good closure for both parton flavor and parton showers after the reweighting





Status and Plans

- a qg tagger has been trained and validated, with complete studies using the full 2015 data collected by LHC
- we are now working on a new qg discriminator, investigating additional variables and alternative training approaches that may results in better performances
- this tool is already being used by several analysis in CMS!

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 few days ago

Thank you for the attention!