Quark-Gluon Discrimination and Pile Up at CMS

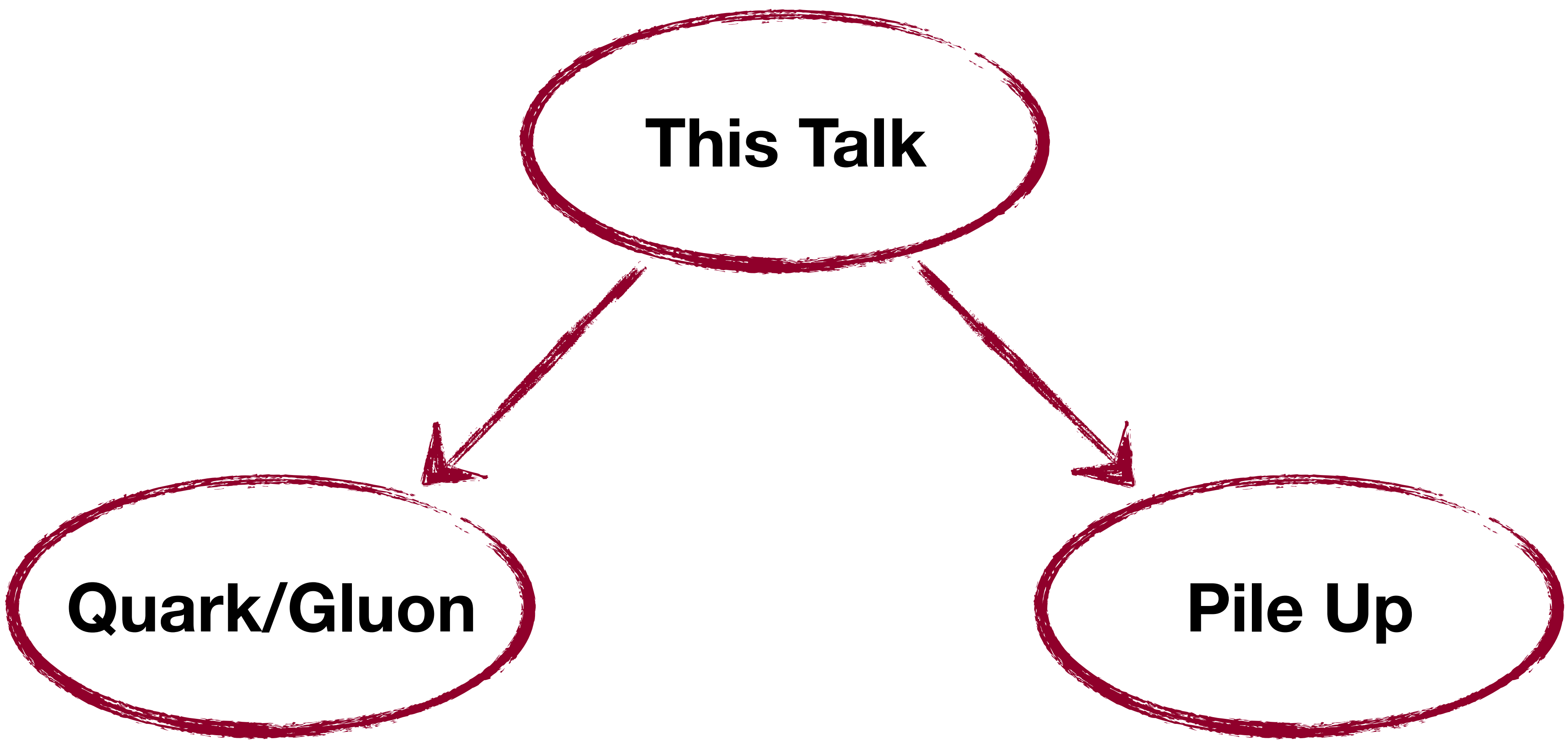
Francesco Pandolfi
ETH Zürich

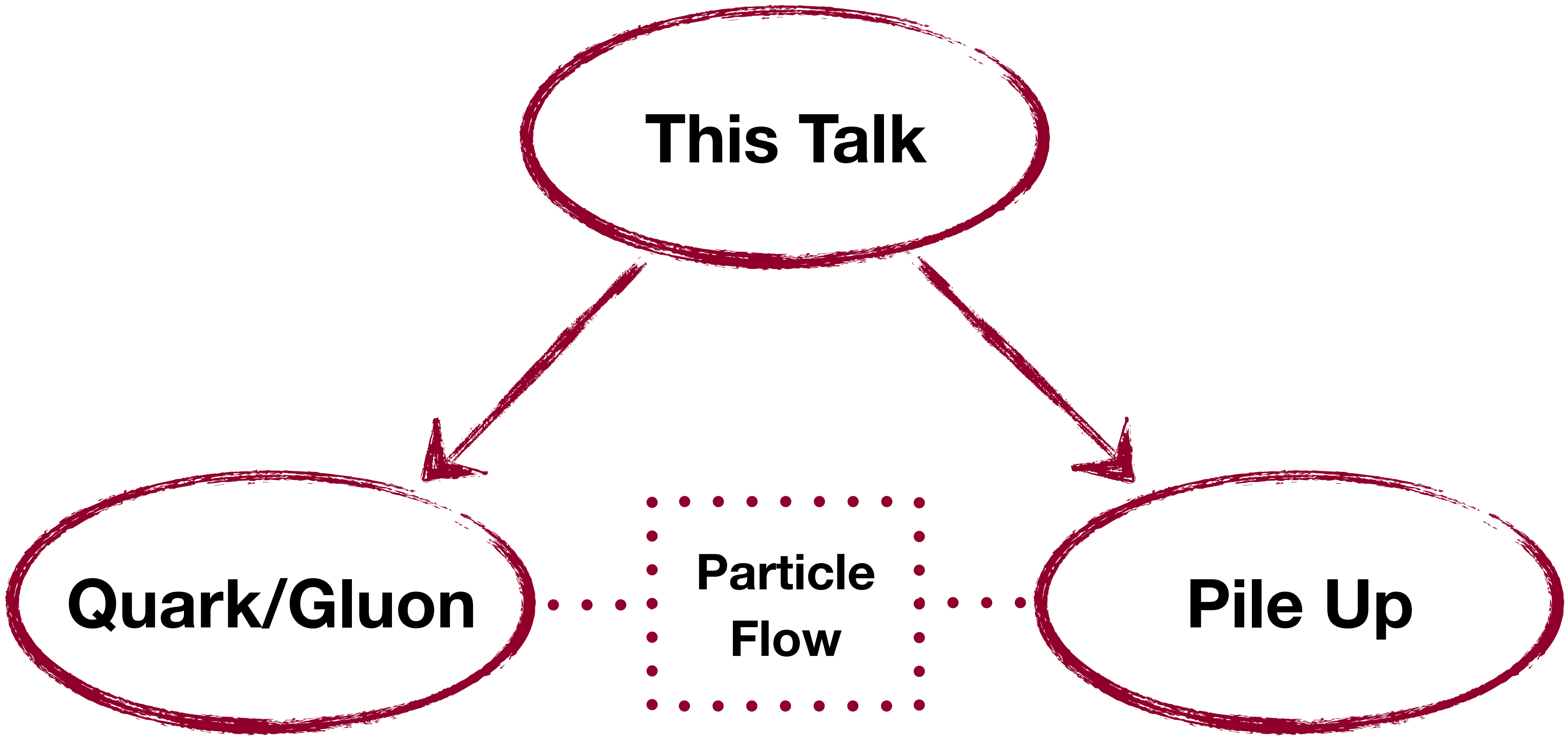
on behalf of the CMS collaboration

ETH Institute for Particle Physics

CMS
Compact Muon Solenoid

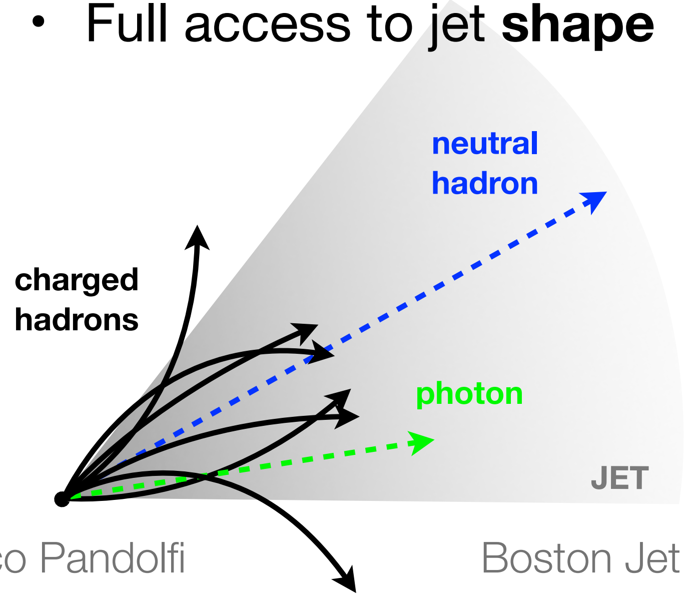
Boston Jet Physics Workshop
MIT, 21.01.2014

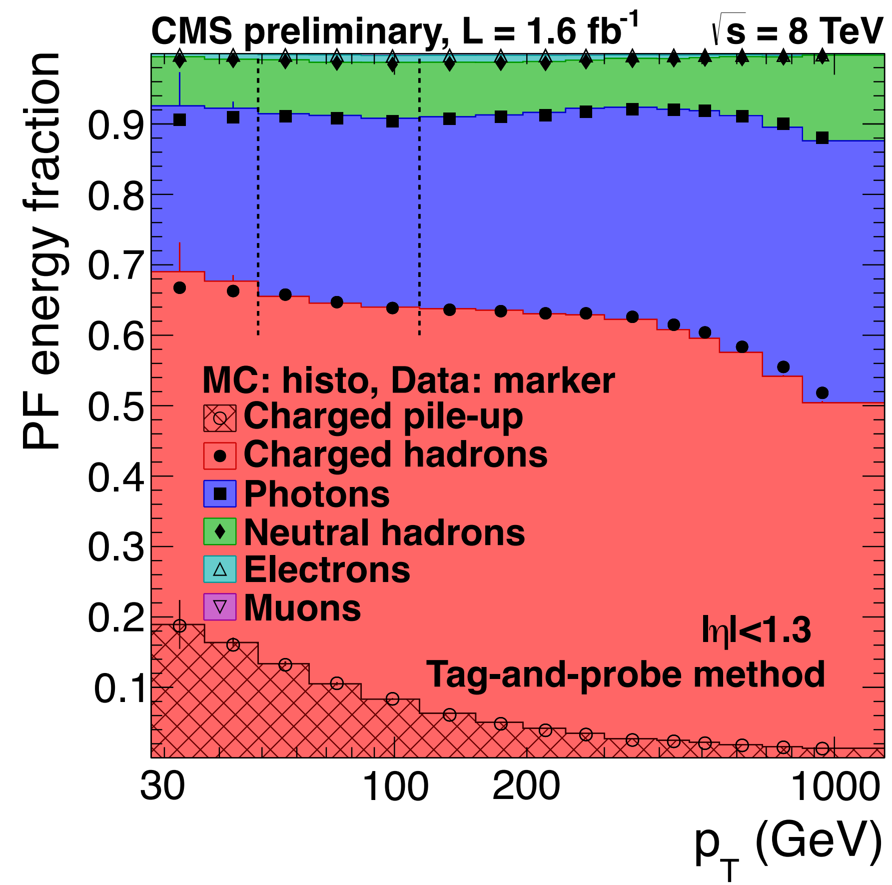

This Talk
Quark/Gluon
Pile Up


This Talk
Quark/Gluon
Particle Flow
Pile Up

Jet Structure Through Particle Flow

- **Particle Flow**: reconstructing all stable particles in event
  - Using all detectors **in unison**
- Powerful tool for **jet structure**
  - Particle-level information
  - Full access to jet **shape**





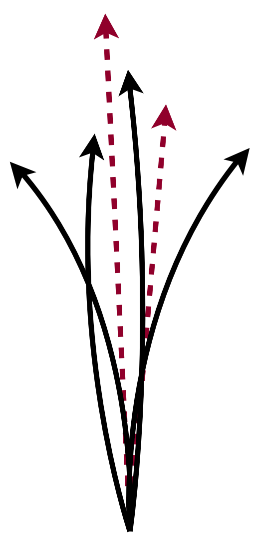
Quark-Gluon Discrimination

JME-13-002

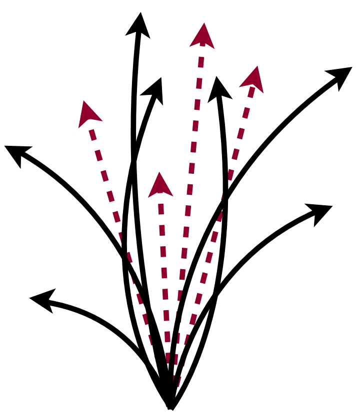
It’s Not Just ‘Jets’

- Most physics analyses **flavour specific**
  - eg. signal is quarks, background is gluons
- Quarks and gluons have **different hadronization**
  - Gluon jets: higher multiplicities, wider, more uniform energy fragmentation

Quark jets:



Gluon jets:



Searching for Discriminating Variables

- Quark jets:
  - Less constituents
  - Narrower cone
  - Asymmetrical energy sharing between constituents

Searching for Discriminating Variables

- Quark jets:
  - Less constituents
  - Narrower cone
  - Asymmetrical energy sharing between constituents

MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)

ENERGY SHARING VARIABLES

- Pull
- R
- p_TD

How the Variables Are Defined

MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)

ENERGY SHARING VARIABLES

- Pull
- R
- $p_T D$

How the Variables Are Defined

MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

from Particle Flow

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)

ENERGY SHARING VARIABLES

- Pull
- R
- $p_T D$

How the Variables Are Defined

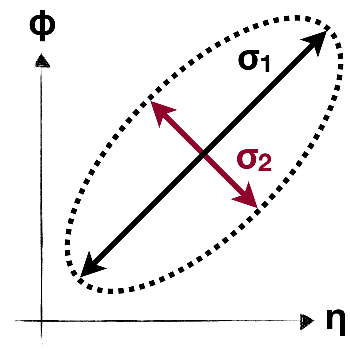
MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

from Particle Flow

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)



and $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$

ENERGY SHARING VARIABLES

- Pull
- R
- $p_T D$

How the Variables Are Defined

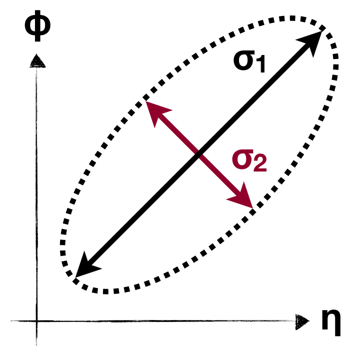
MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

from Particle Flow

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)



and $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$

ENERGY SHARING VARIABLES

- Pull $\quad |\vec{t}| = \left| \frac{\sum_i p_{\mathrm{T},i}^2 |r_i| \vec{r}_i}{\sum_i p_{\mathrm{T},i}^2} \right|$ **with** $\vec{r}_i = (\Delta\eta_i,\ \Delta\phi_i)$
- R = $p_{\mathrm{T,i}}$ (max) / $\sum p_{\mathrm{T,i}}$
- p_TD $\quad p_\mathrm{T}D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$

Three Variables for the Discriminator

MULTIPLICITY VARIABLES

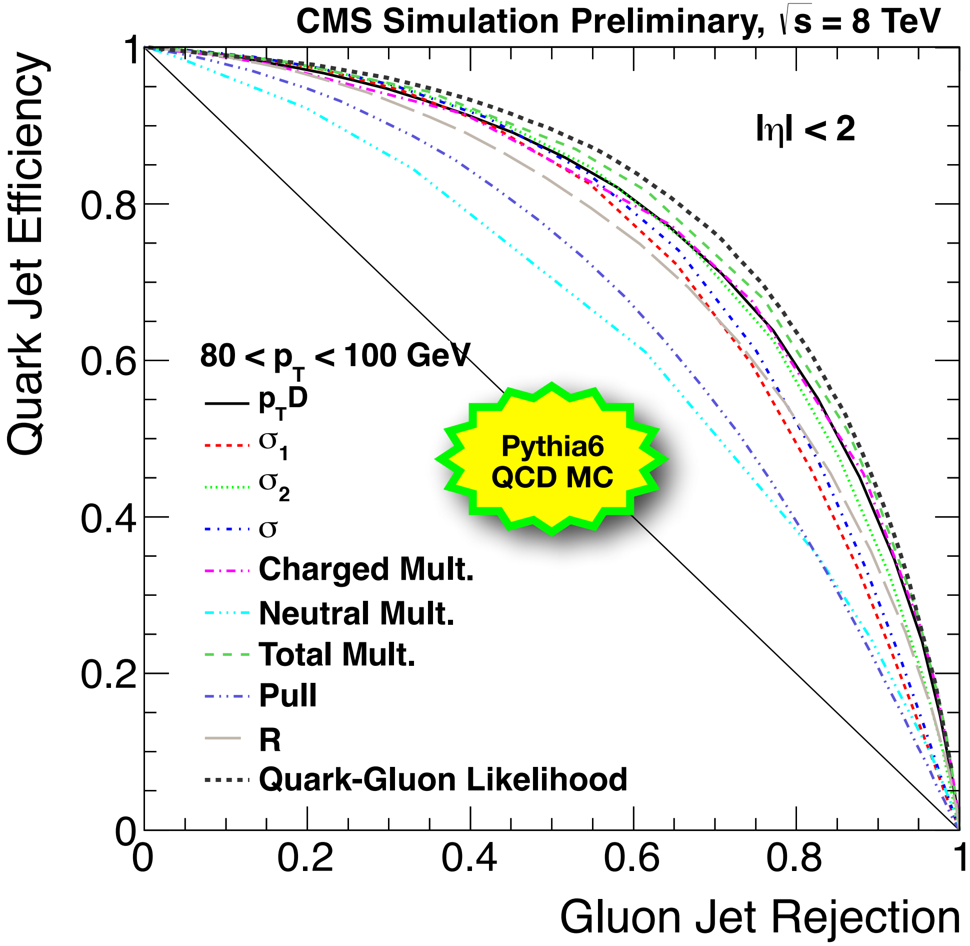
- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)

ENERGY SHARING VARIABLES

- Pull
- R
- $p_T D$



Three Variables for the Discriminator

MULTIPLICITY VARIABLES

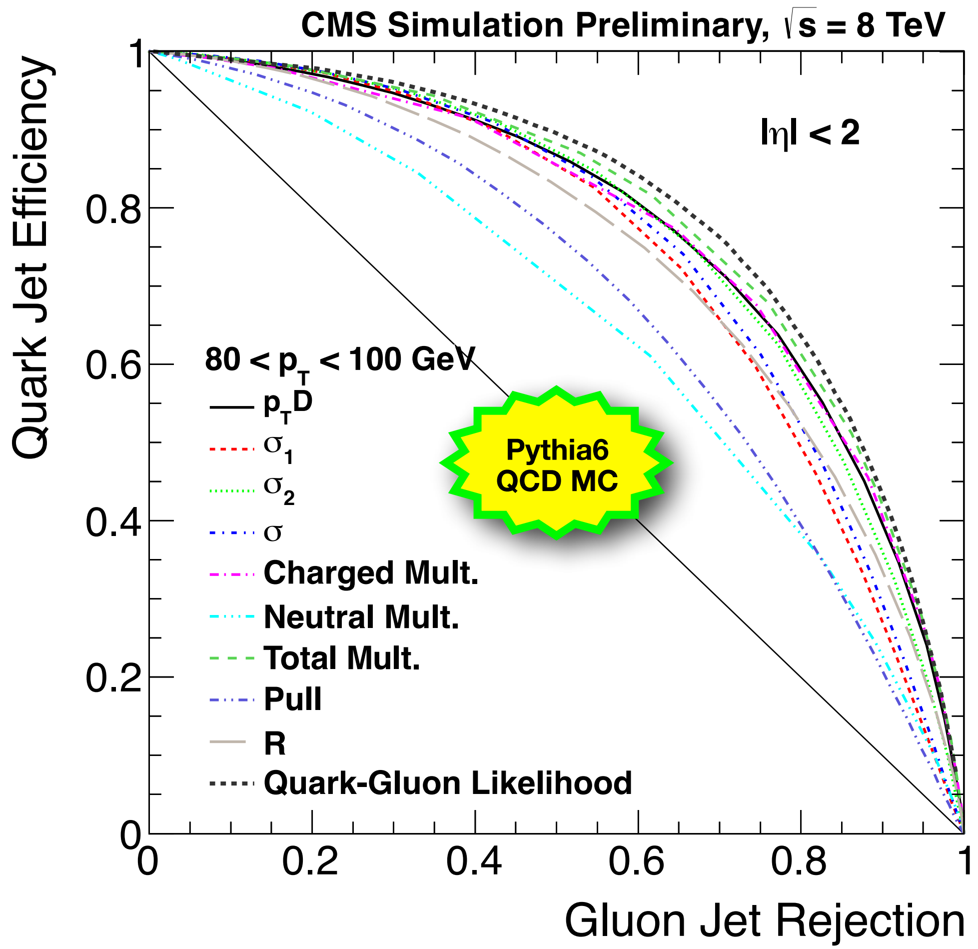
- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

WIDTH VARIABLES

- RMS of PFCandidate η-φ spread (σ)
- Major axis of η-φ matrix (σ_1)
- Minor axis of η-φ matrix (σ_2)

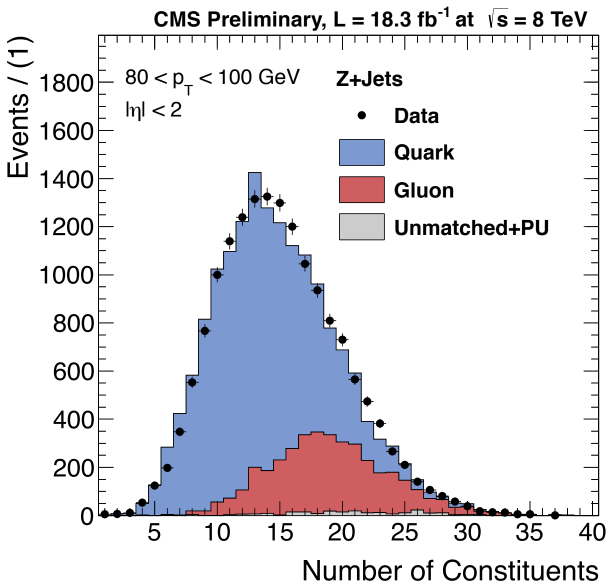
ENERGY SHARING VARIABLES

- Pull
- R
- $p_T D$

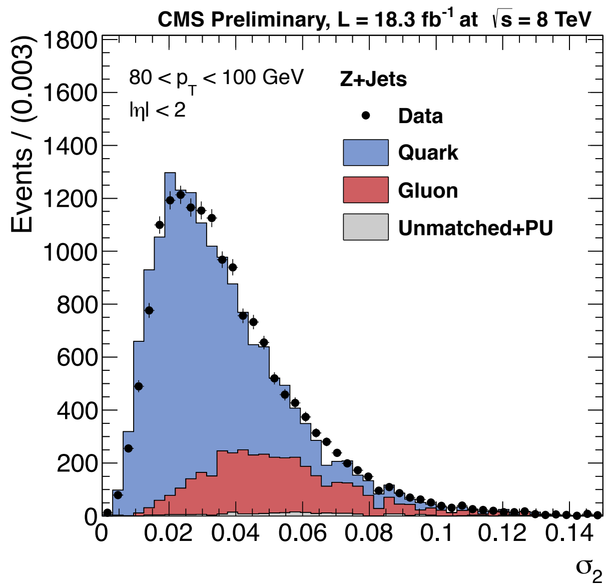


Variables Look Good in the Data

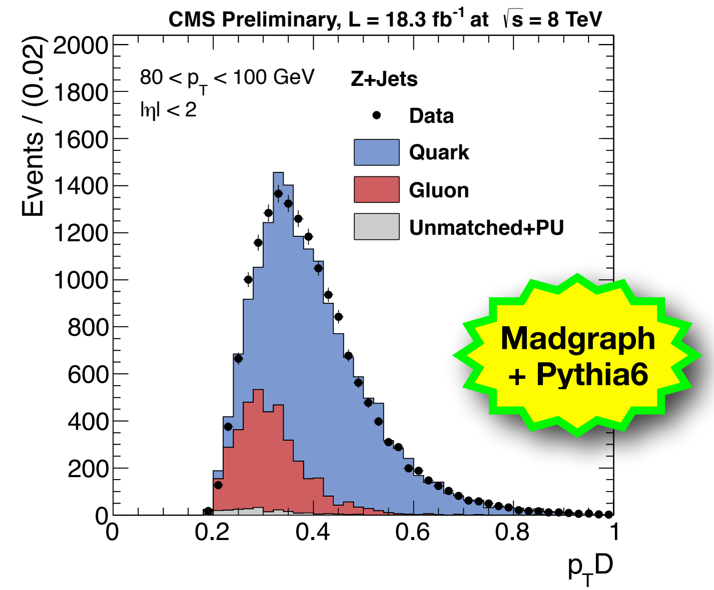
Total Multiplicity



Minor axis (σ_2)

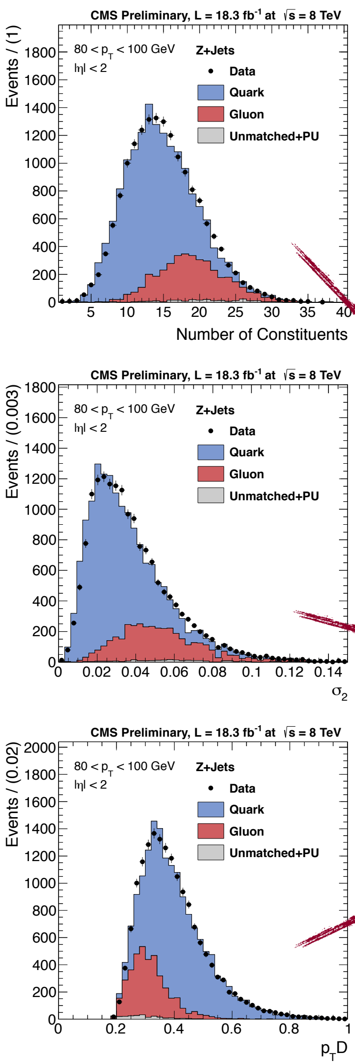


p_TD

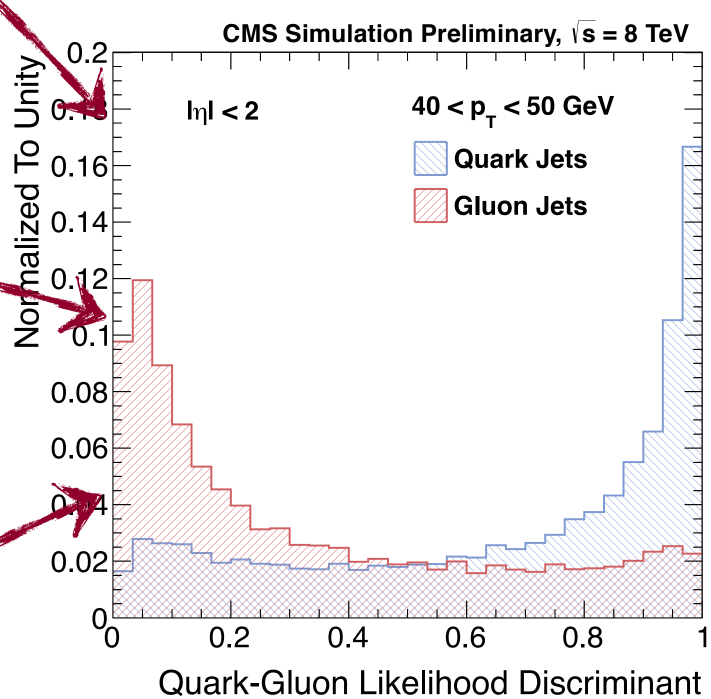


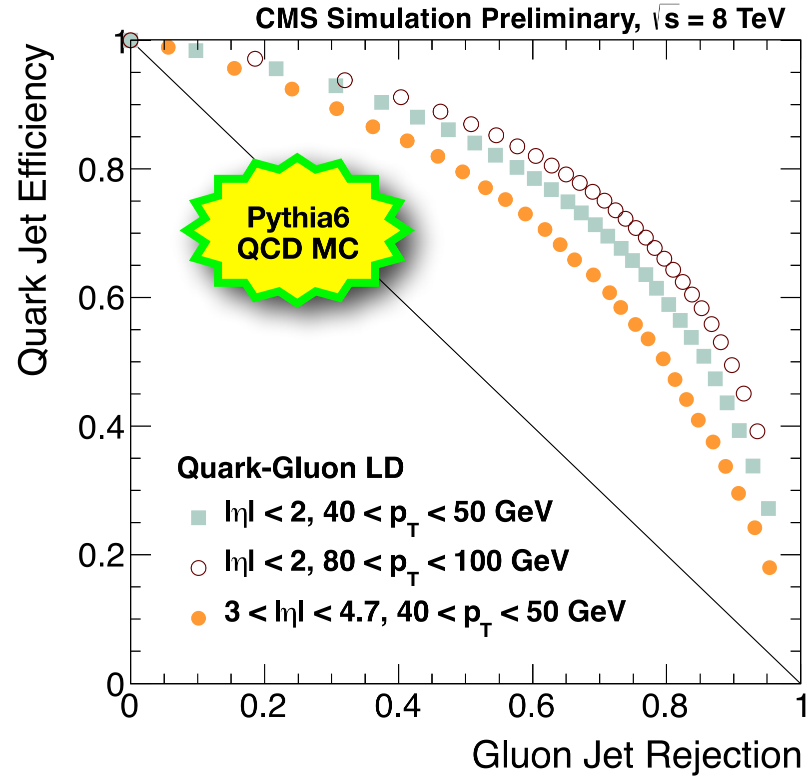
- Control sample: Z + (one) jet
  - **Quark** enriched
  - Good data/MC **agreement**

Building the Discriminator



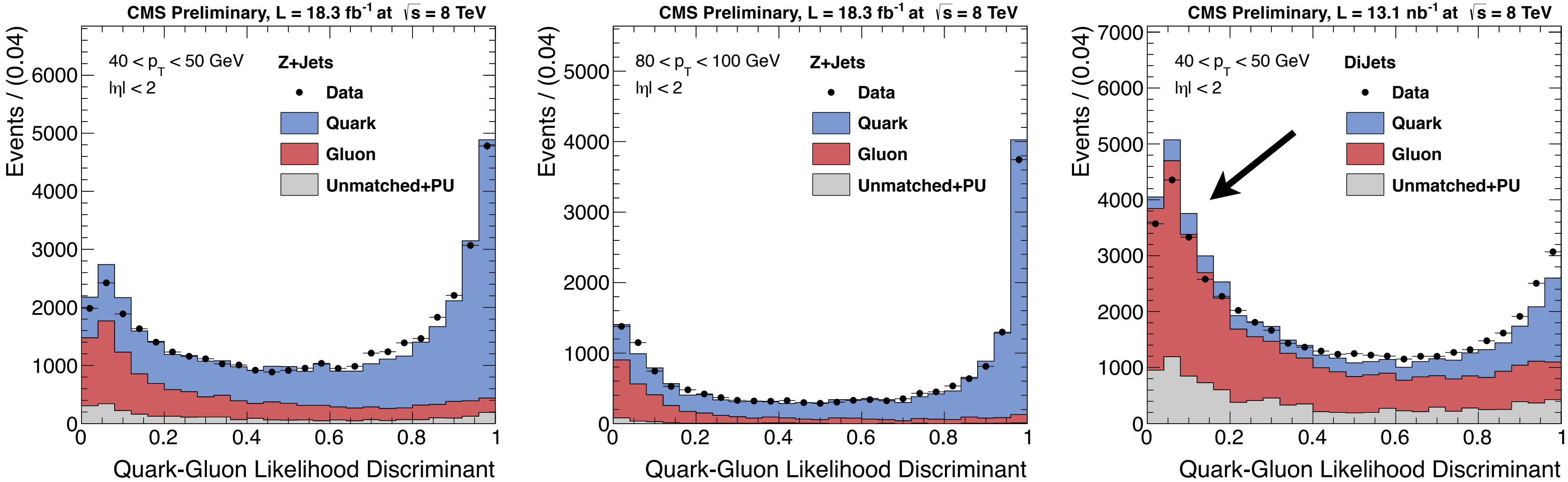
- ❖ Three variables combined into **simple likelihood** product
  - Binned PDFs: in $\boldsymbol{p_T}$ and **pile-up**
  - Discrimination up to |η| = 5





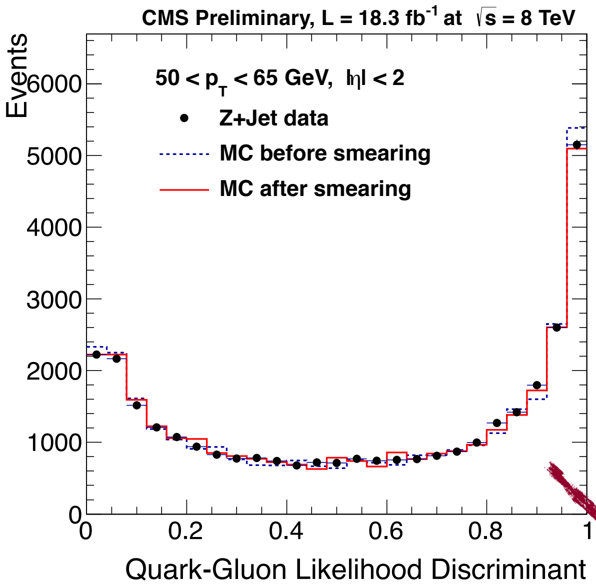
Some Discrepancies in the Gluons

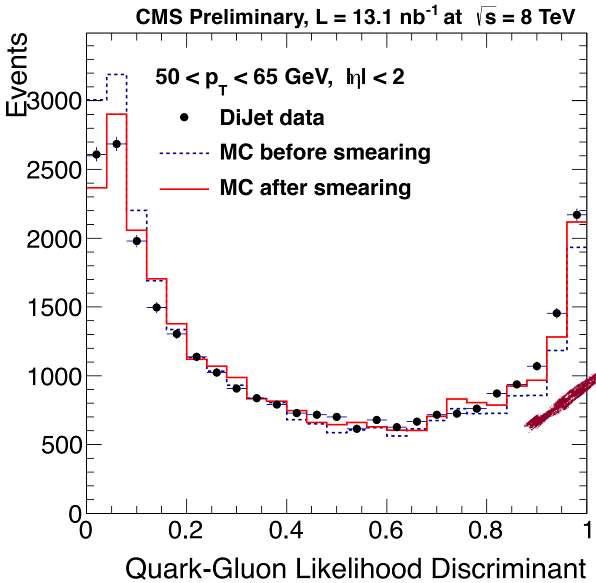
Z+Jets 40 < p_T < 50 GeV, |η| < 2	Z+Jets 80 < p_T < 100 GeV, |η| < 2	DiJets 40 < p_T < 50 GeV, |η| < 2



- Observe **discrepancy** in dijet control sample
  - With **high gluon fraction**

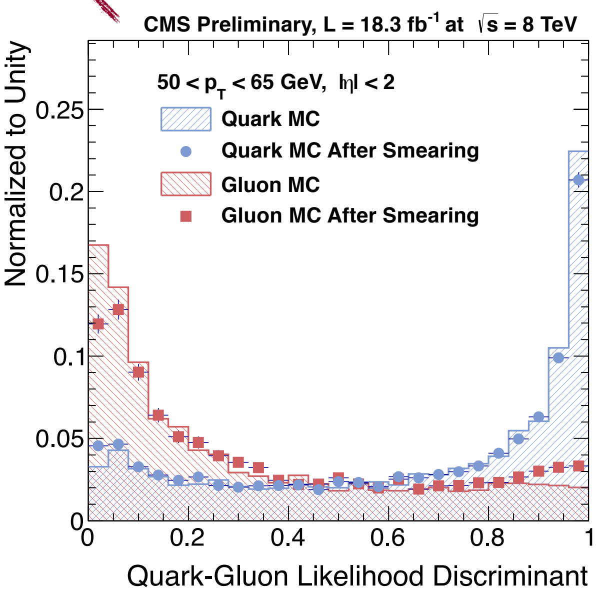
Smearing the MC to Match Data





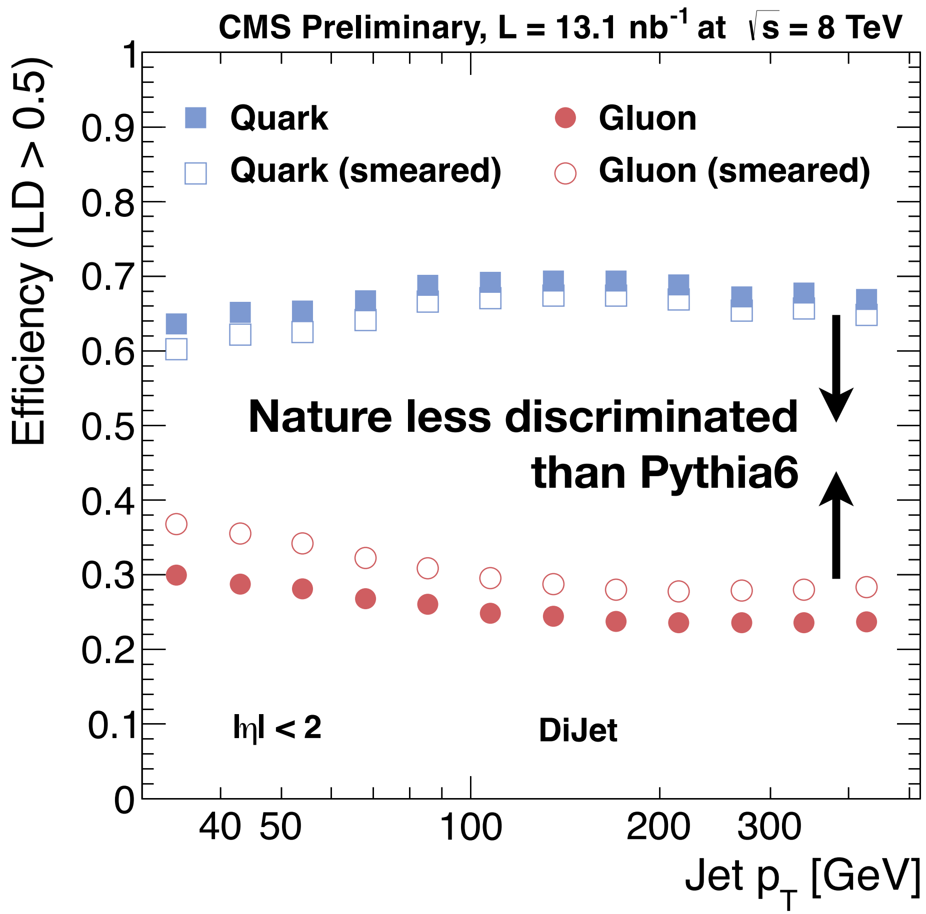
- **Smear MC** to take into account discrepancy
  - **Not a reweighting**: a re-mapping
  - Change tagger value **jet-by-jet**
- **Two-step** procedure:

1. **Derive** on Z+jets (>70% quarks)
   - **Separately** for quarks/gluons
   - Until agreement with data
2. **Apply** on QCD (>60% gluons)
   - They work!



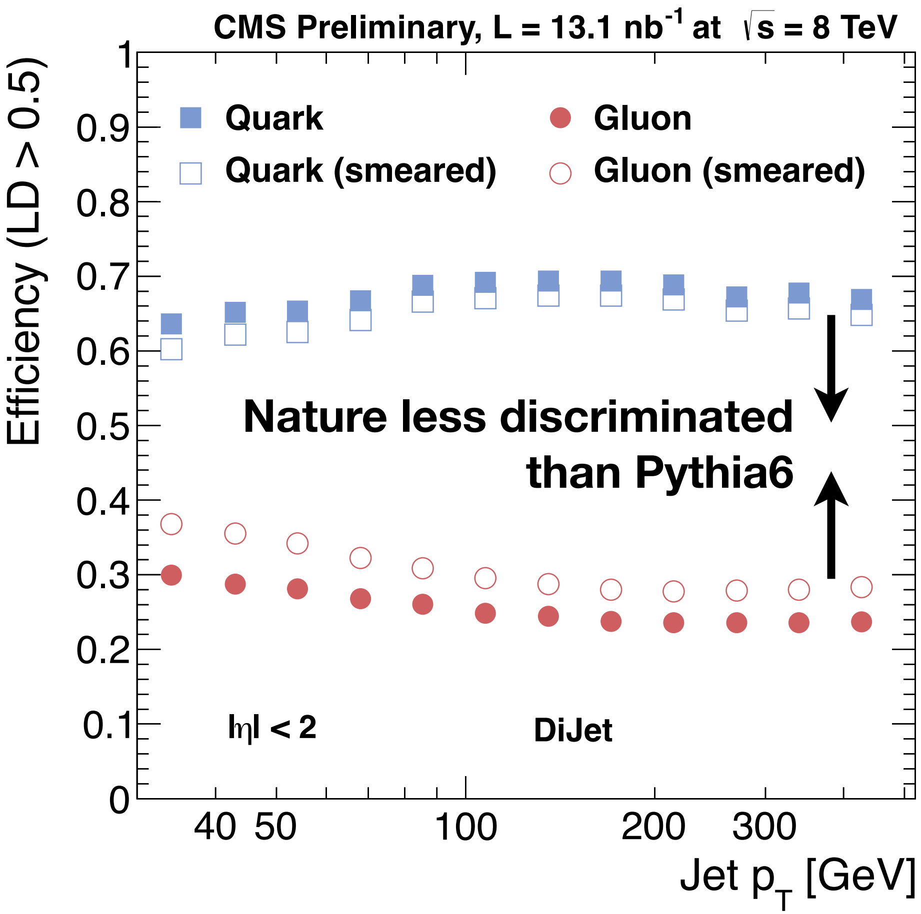
Nature Lies Between Pythia6 and Herwig++

Pythia6

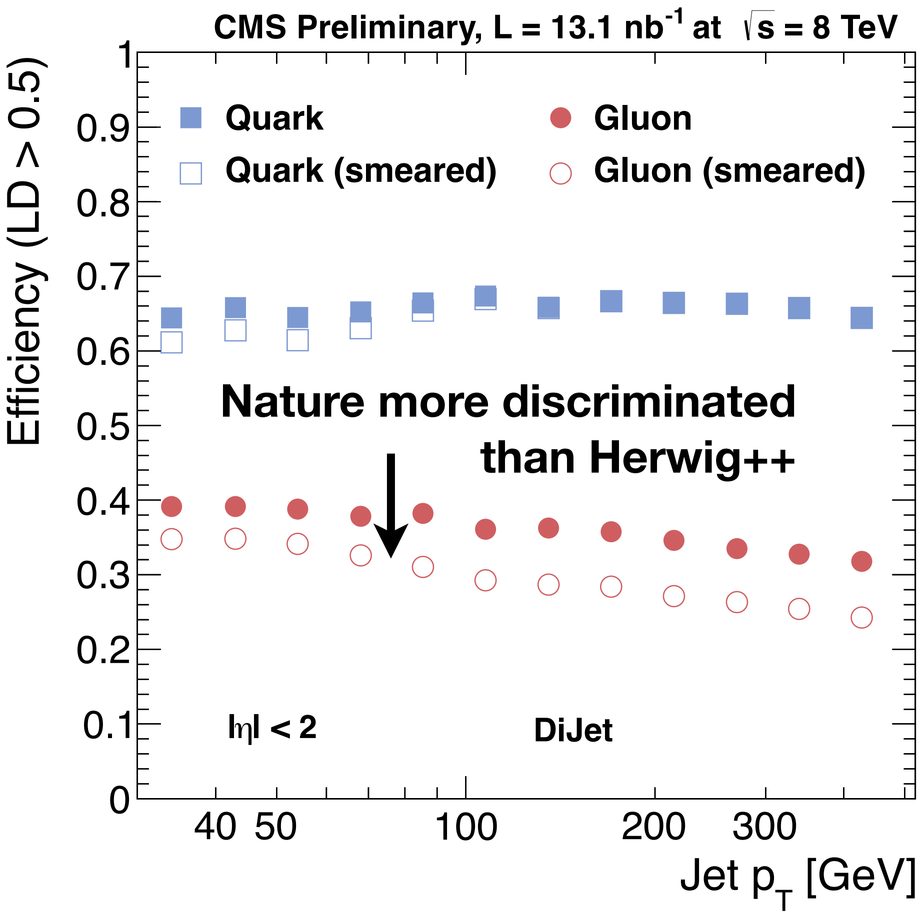


Nature Lies Between Pythia6 and Herwig++

Pythia6



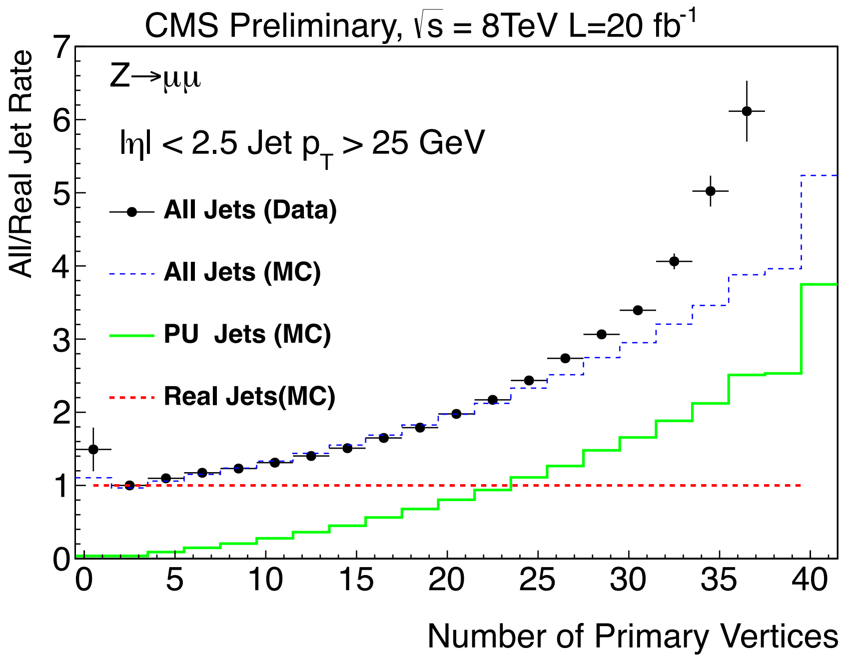
Herwig++



Pile-Up Identification

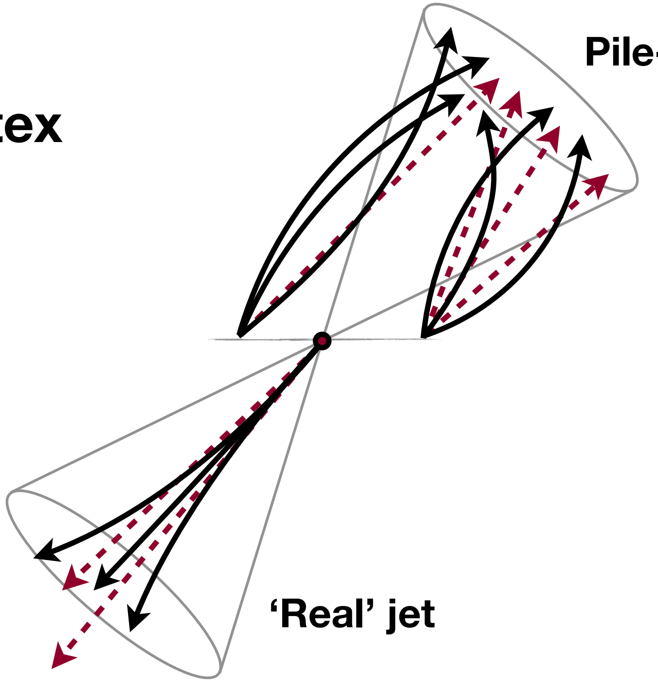
A Lot of Pile Up at the LHC

- Current LHC running: **high-pileup** conditions
  - Average of 23 extra interactions (and up to 40!)
- Additional collisions produce **soft** jets
  - But **can overlap** (combinatorics!)
  - Resulting jets can have $p_T > 25$ GeV
- Pile-up jet ID
  - **Crucial** for analyses with low-p_T jets



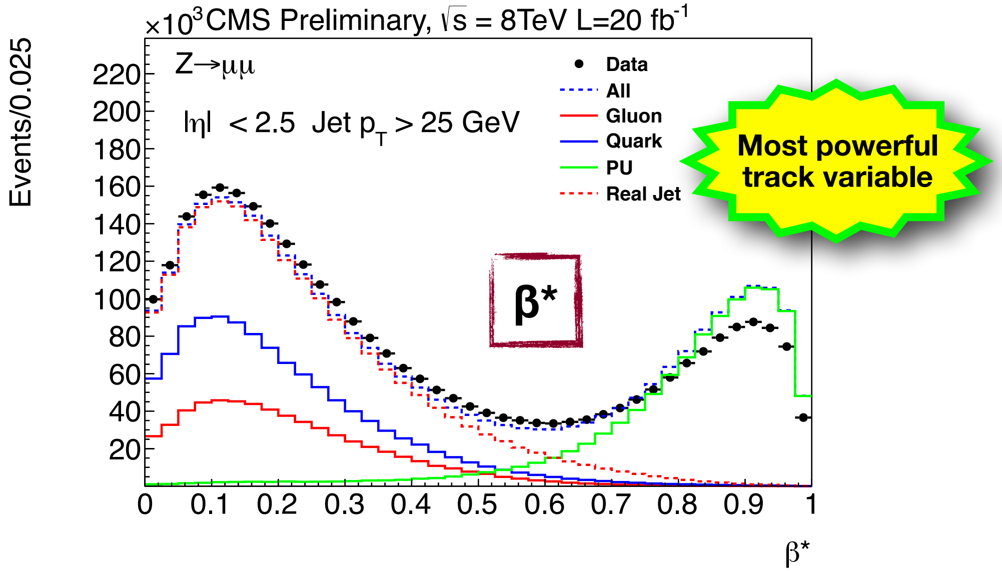
Identifying Pile Up Jets

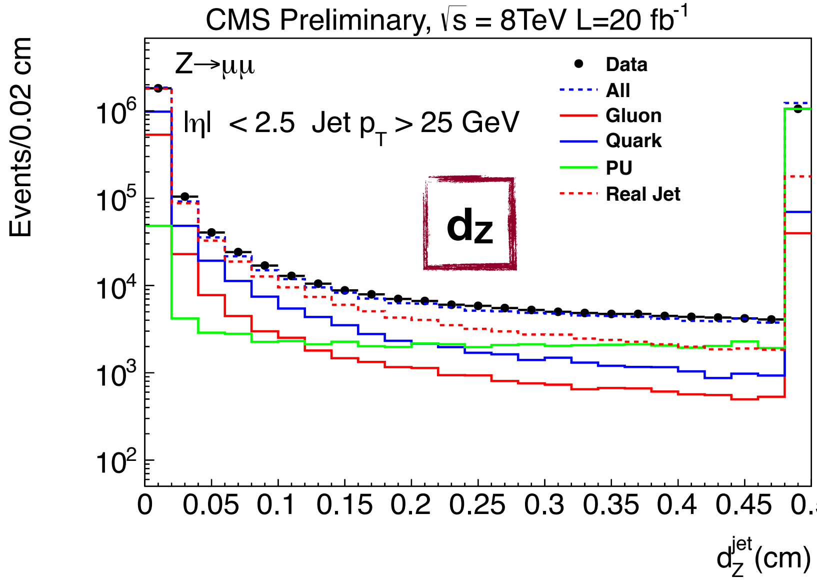
- Pile up jets mainly **overlap of soft jets** from extra interactions
- **Two** main characteristics:
  - Tracks **incompatible with primary vertex**
  - Clustered particles **more diffuse**
- Selected **12 variables** (scanned >80)
  - 4 **track** variables ($|\eta| < 2.5$)
  - 8 **shape** variables ($|\eta| < 5$)
- Again making use of powerful **Particle Flow** information



The Track-Based Variables

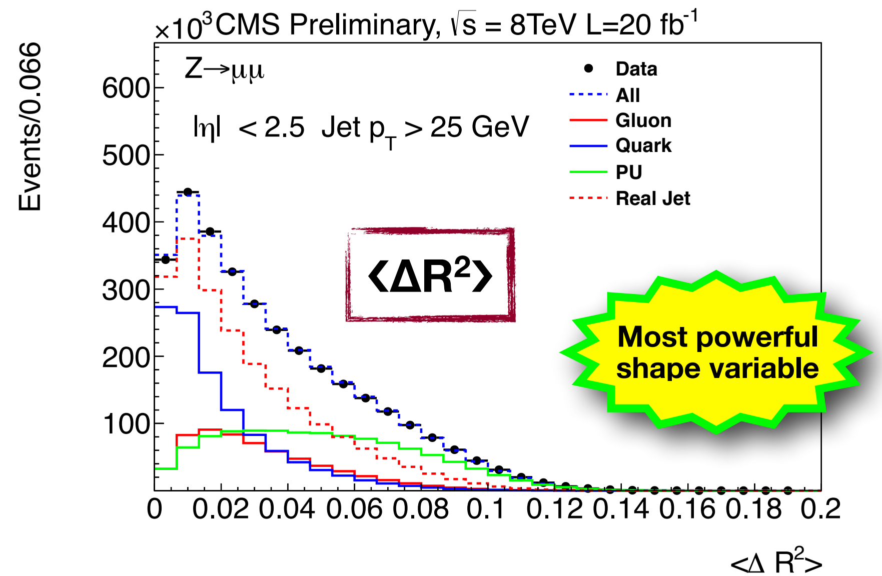
- Four track variables ($|\eta| < 2.5$):
  - **β**: energy fraction of tracks from primary vertex
  - **β***: energy fraction of tracks from other vertexes
  - **d_Z**: z-distance between primary vertex and hardest track
  - **$n_{vertexes}$**
- Some disagreement in β* (and β)
  - Due to known issue in simulation

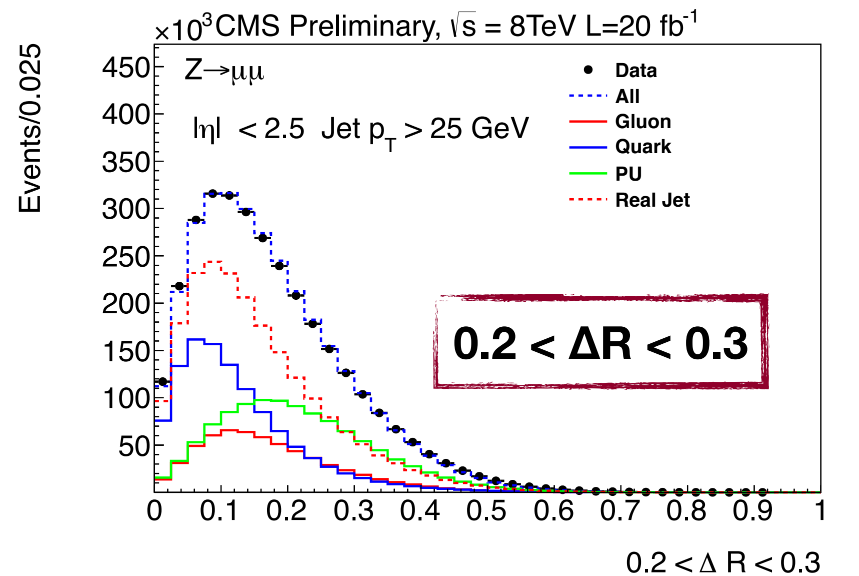




The Shape Variables

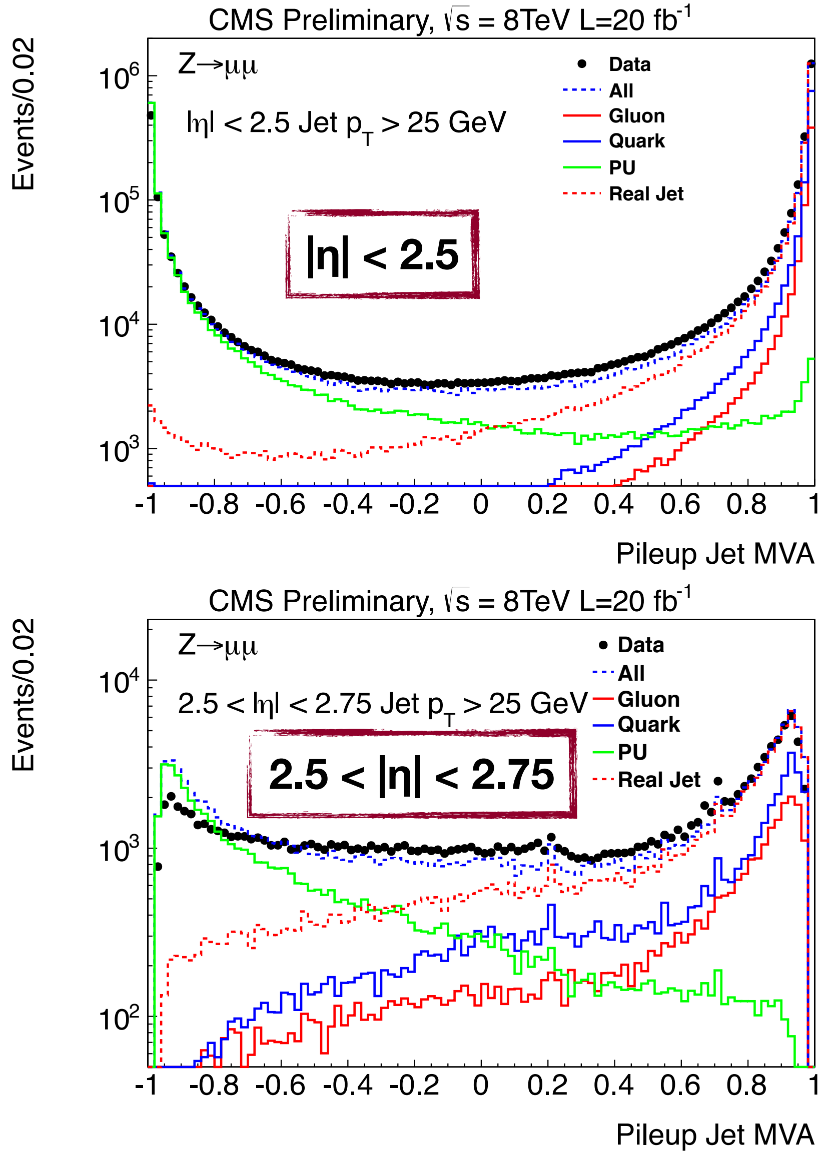
- Twelve track variables ($|\eta| < 5$):
  - **$\langle \Delta R^2 \rangle$**: η-ϕ RMS of candidates
    (similar to σ in quark-gluon)
  - **$A < (\Delta R) < A+0.1$**: fractional p_T sum of particles in a given annulus
    (all **five** annuli from 0 to 0.5 used)
  - **$N_{charged}$**: charged multiplicity
  - **$N_{neutrals}$**: neutral multiplicity
  - **$p_T D$**: as in quark-gluon





Building the Pile-Up Discriminator

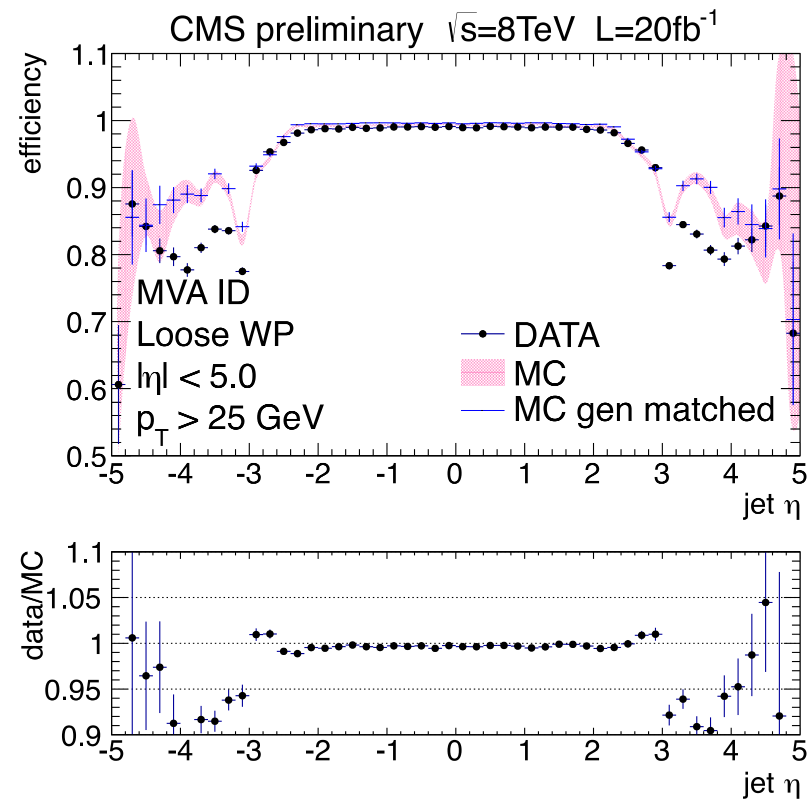
- Twelve variables fed to an MVA (BDT)
  - Trained separately in **four |η| bins**
  - Tested on Z→μμ data
- Best discrimination in **central region**
- Some **disagreements** in endcap region
  - From incorrect simulation of **out-of-time pile-up**
  - Known problem, **fixed** in later releases
  - Will take into account with **scale factors**

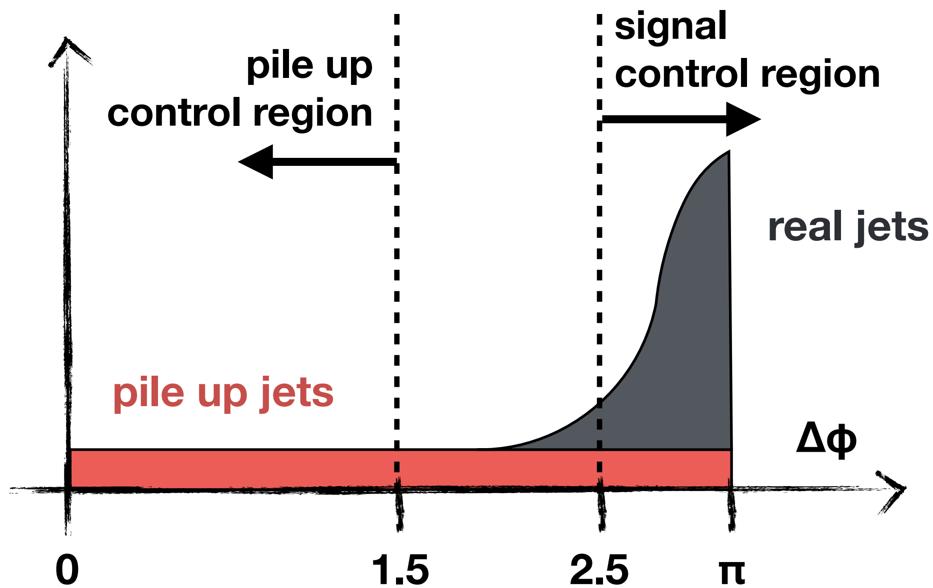


Measuring the Efficiency in the Data

- Derived on $Z \to \mu\mu$ **control sample**

- Use $\Delta\phi$(Z-jet) distribution
  - Real jets **peak** at $\Delta\phi = \pi$
  - Pile up is **flat** in $\Delta\phi$

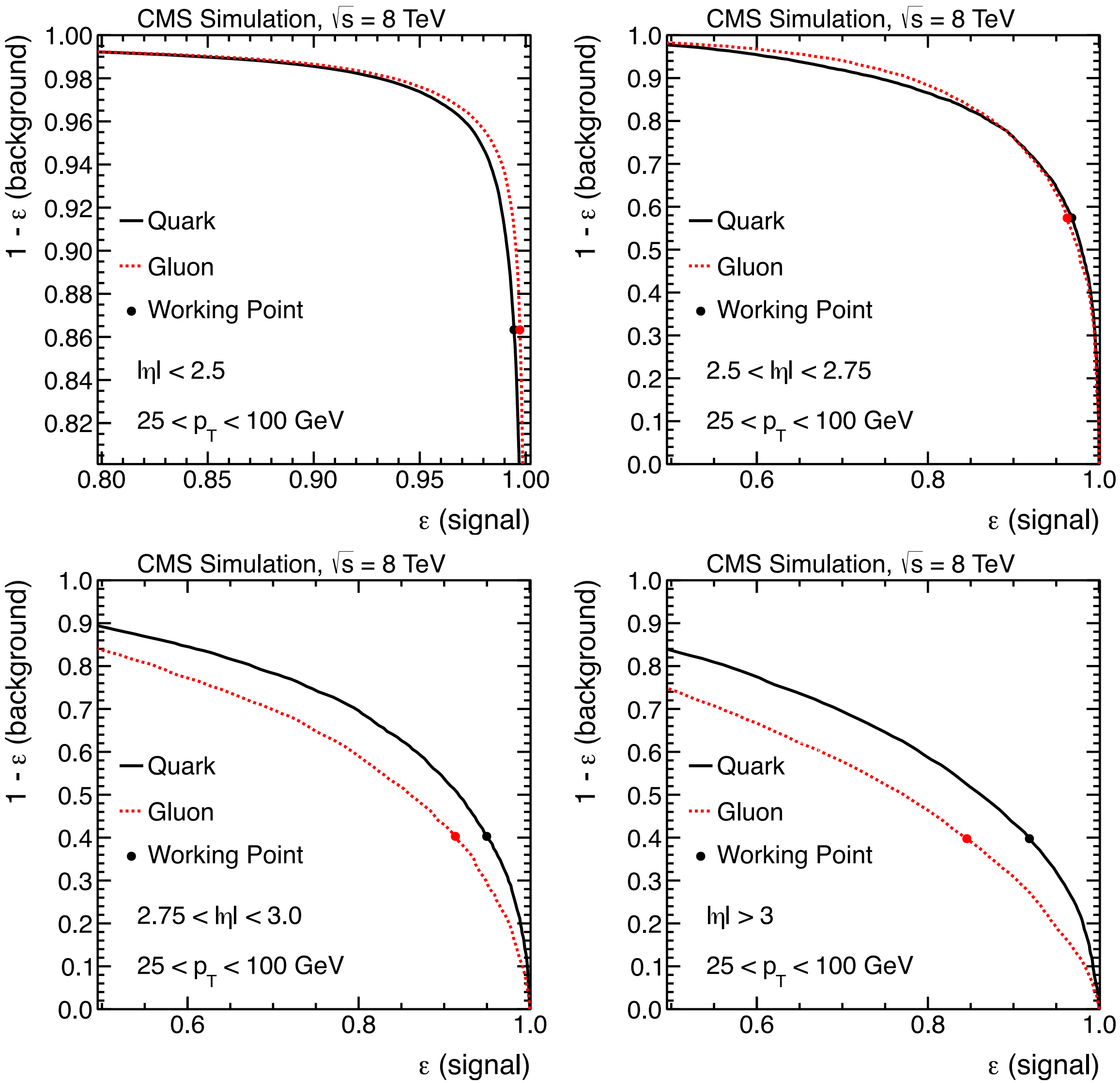




$$\text{efficiency} = \frac{N_{pass}(sig) - k \cdot N_{pass}(BG)}{N_{all}(sig) - k \cdot N_{all}(BG)}$$

with $k = (\pi - 2.5) / 1.5$

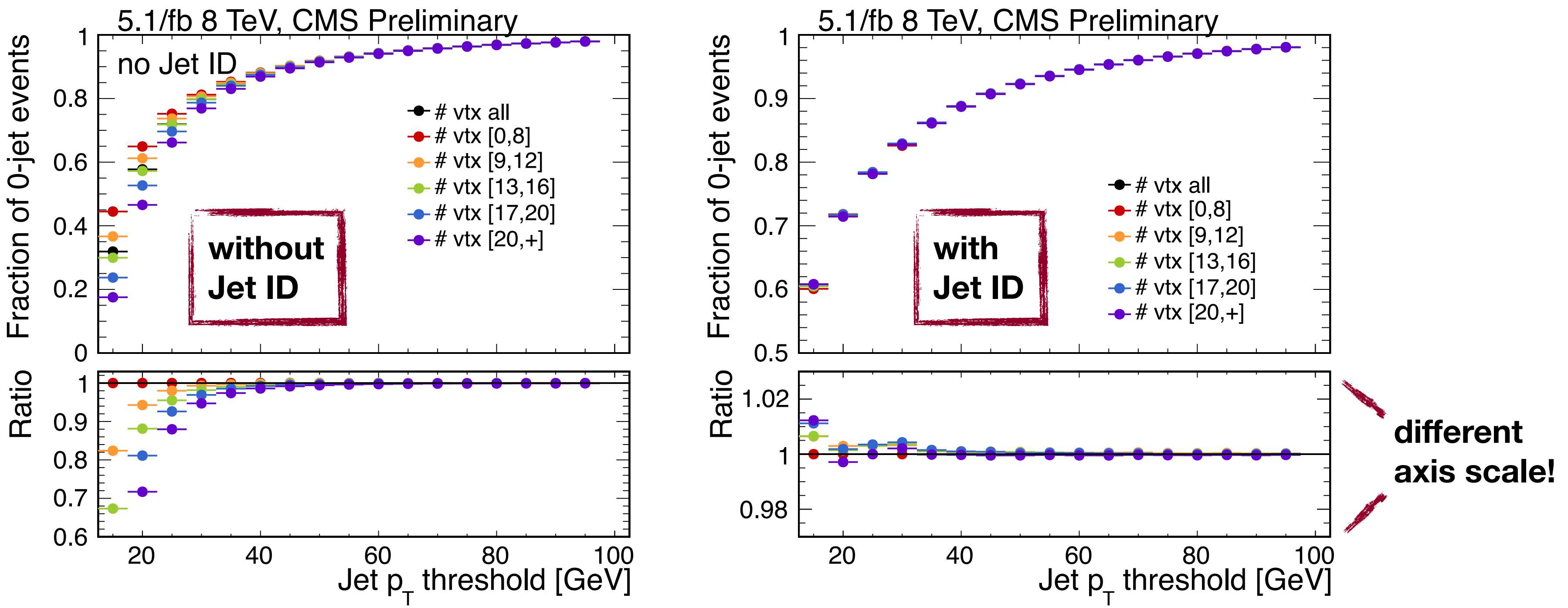
Performance on Quarks and Gluons



- **Very efficient** discrimination in tracker-covered region
  - >99% signal with ~90% background rejection
- Gluons in tracker: higher efficiency
  - More particles → **more tracks** (→ better vertexing)
- Gluon in forward: lower efficiency
  - **Diffuse shape**, like pile up

An Application: Jet Veto

- An example: **jet veto** in Z→μμ analysis
  - Without pile up ID: large dependence on pile up for p_T < 40 GeV
  - With pile up ID: **no dependence**



Conclusions

- **Successfully employed** quark-gluon separation and pile up ID in CMS analyses

- **Quark-gluon** separation
  - Multiplicity, width, energy sharing
  - **Simple likelihood** with three inputs, up to $|\eta| = 5$
  - **Quarks well modeled** by MC, gluons not so well

- **Pile up** jet identification
  - **Track**- and **shape**-based variables
  - Implemented as a **BDT**, also up to $|\eta| = 5$
  - In $|\eta|<2.5$: >99% signal efficiency, ~90% BG rejection

Backup Slides

Quark-Gluon: Variable Definitions

MULTIPLICITY VARIABLES

- Charged Multiplicity
- Neutral Multiplicity
- Total Multiplicity

from Particle Flow

USED JET CONSTITUENTS

- Charged: from primary vertex
- Neutral: $p_T > 1$ GeV

WIDTH VARIABLES

- RMS of PFCandidate η-ϕ spread (σ)
- Major axis of η-ϕ matrix (σ_1)
- Minor axis of η-ϕ matrix (σ_2)

$$\Delta\eta = \eta - \bar{\eta} \qquad \Delta\phi = \phi - \bar{\phi}$$

$$\begin{pmatrix} \sum_i p_{\mathrm{T},i}^2 \Delta\eta_i^2 & -\sum_i p_{\mathrm{T},i}^2 \Delta\eta_i \Delta\phi_i \\ -\sum_i p_{\mathrm{T},i}^2 \Delta\eta_i \Delta\phi_i & \sum_i p_{\mathrm{T},i}^2 \Delta\phi_i^2 \end{pmatrix}$$

$\lambda_{1,2}$ are the eigenvalues of this matrix

$$\sigma_1 = \left(\lambda_1 / \sum_i p_{\mathrm{T},i}^2\right)^{1/2}$$

$$\sigma_2 = \left(\lambda_2 / \sum_i p_{\mathrm{T},i}^2\right)^{1/2}$$

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$$

ENERGY SHARING VARIABLES

- Pull
- $R = \dfrac{\max(p_{\mathrm{T},i})}{\sum_i p_{\mathrm{T},i}}$
- $p_T D$

$$|\vec{t}| = \left| \frac{\sum_i p_{\mathrm{T},i}^2 |r_i| \vec{r}_i}{\sum_i p_{\mathrm{T},i}^2} \right|$$

$$\vec{r}_i = (\Delta\eta_i,\ \Delta\phi_i)$$

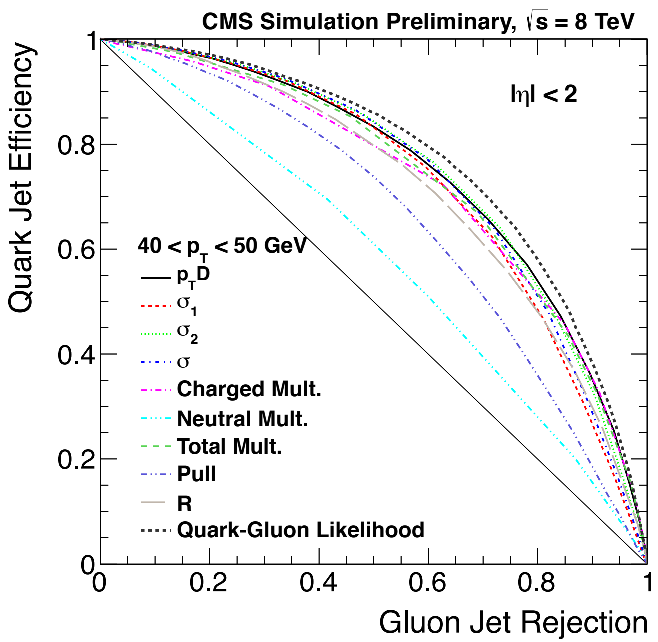
$$p_T D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$$

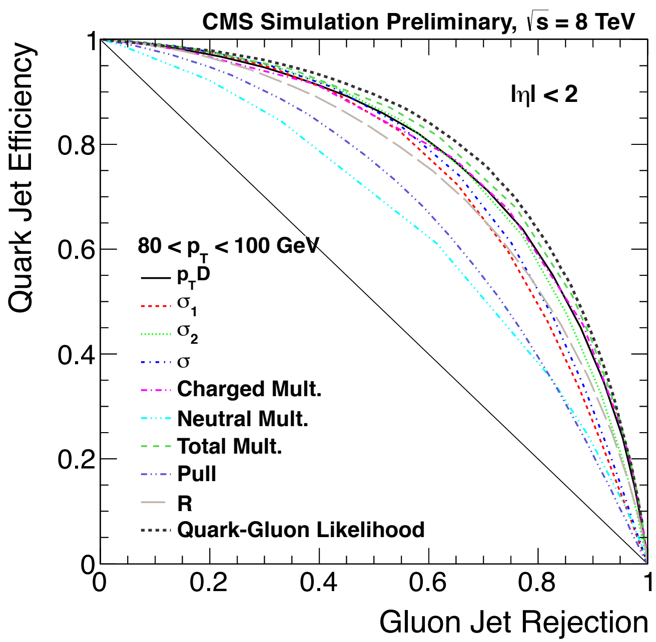
The Other RoC Curves

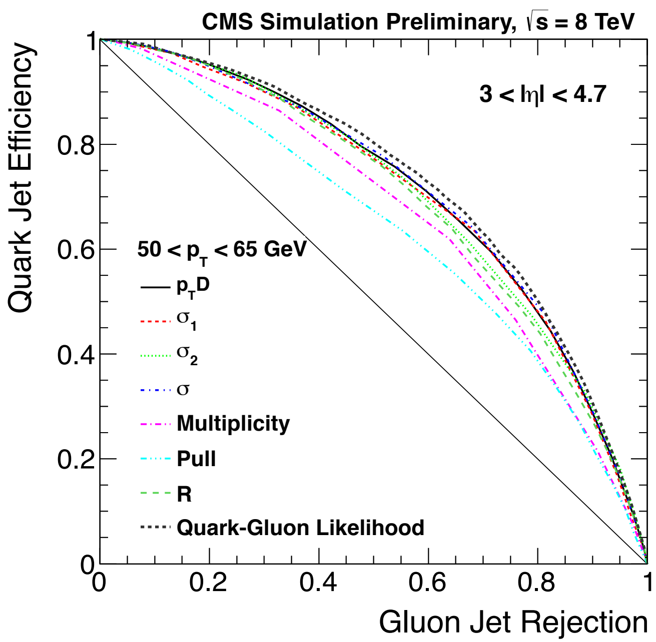
40 < p_T < 50 GeV
|η| < 2

80 < p_T < 100 GeV
|η| < 2

50 < p_T < 65 GeV
3 < |η| < 4.7



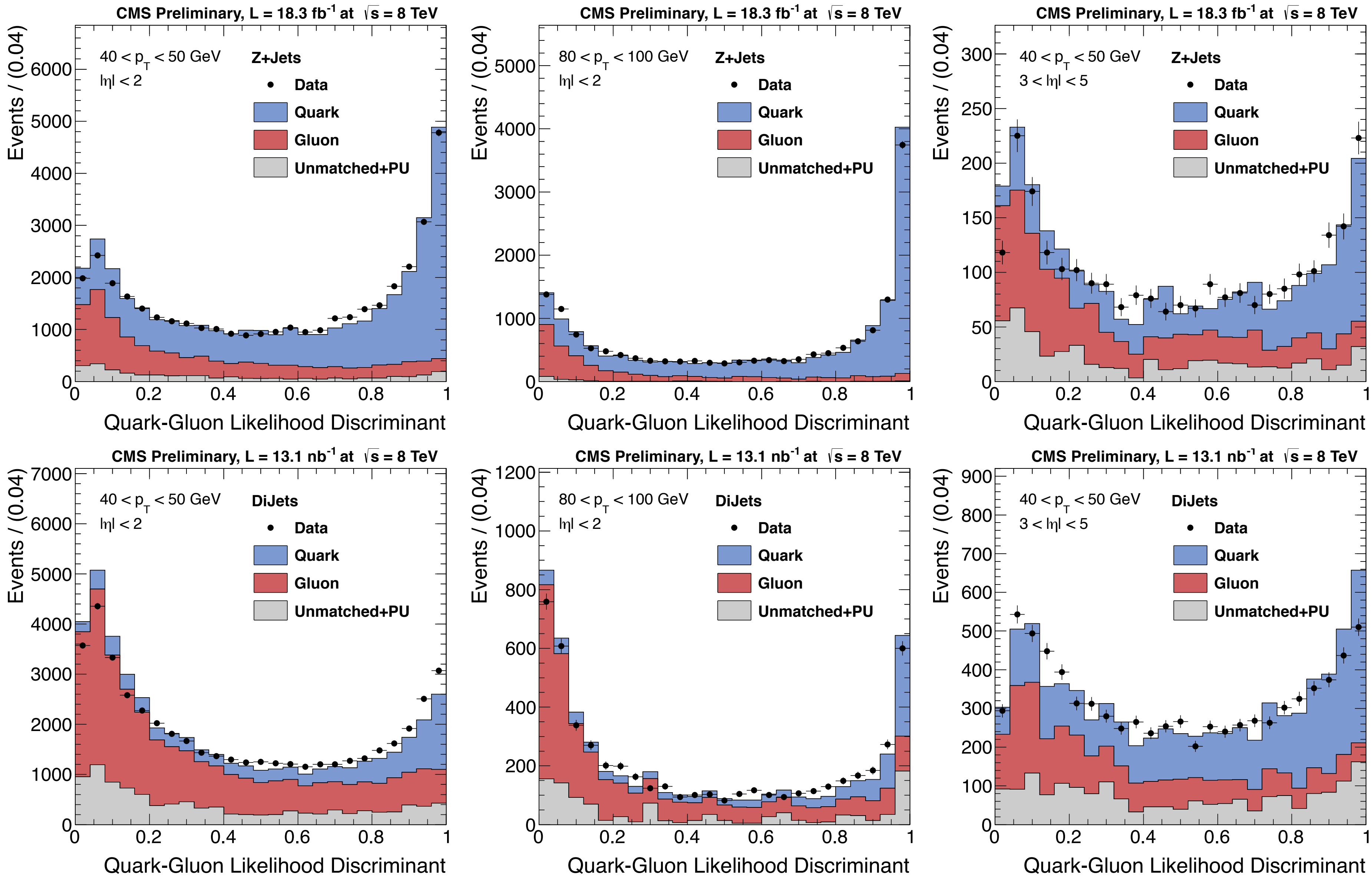




The Other Data/MC Comparisons

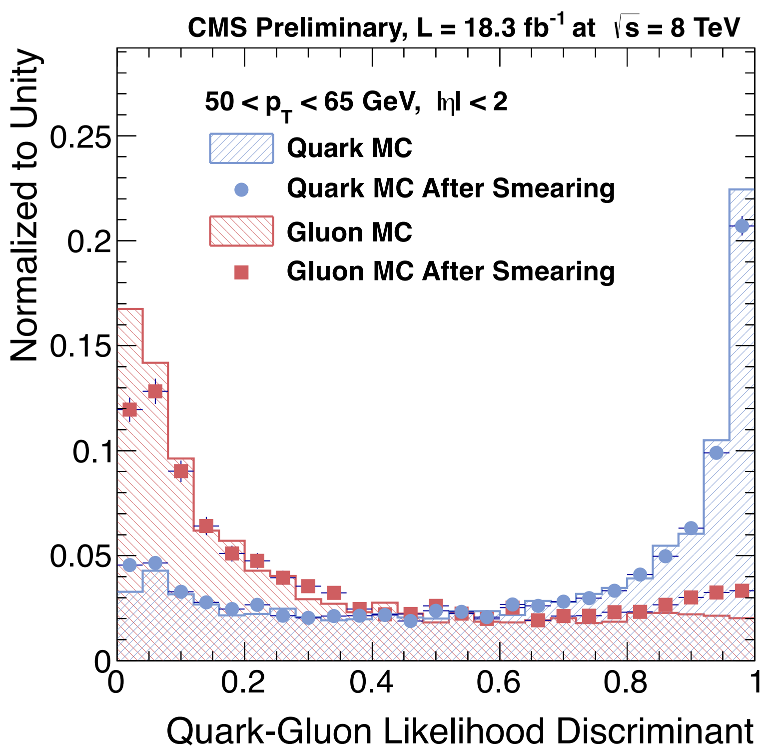
Z + Jets

DiJets

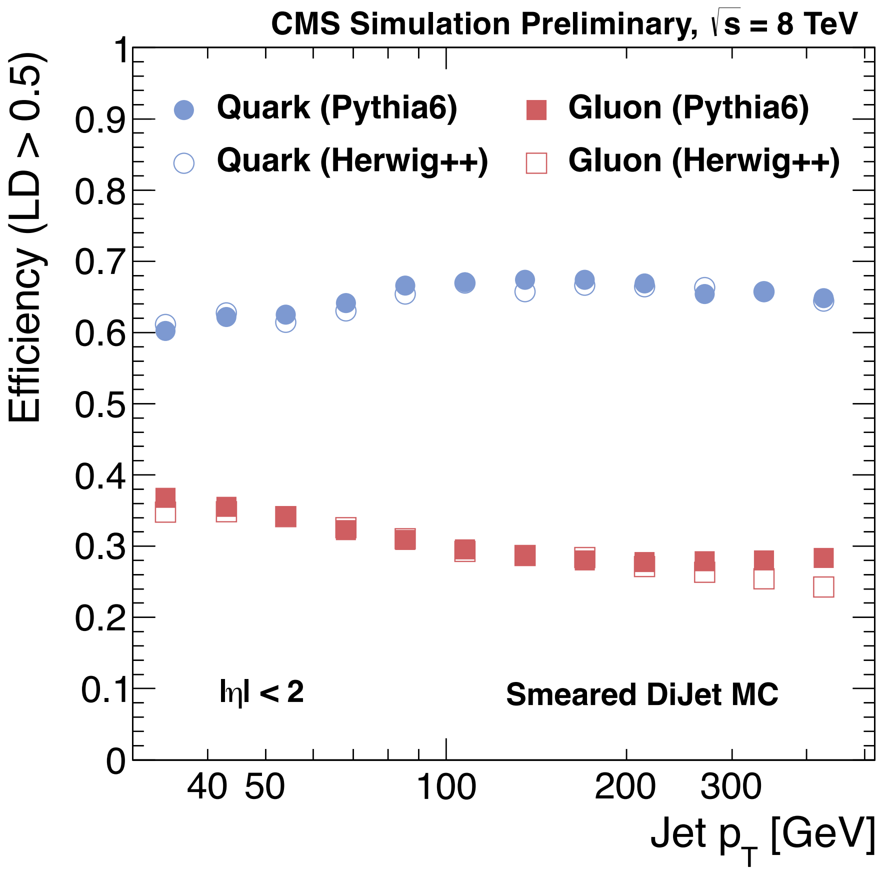


How the QG Smearing Really Works

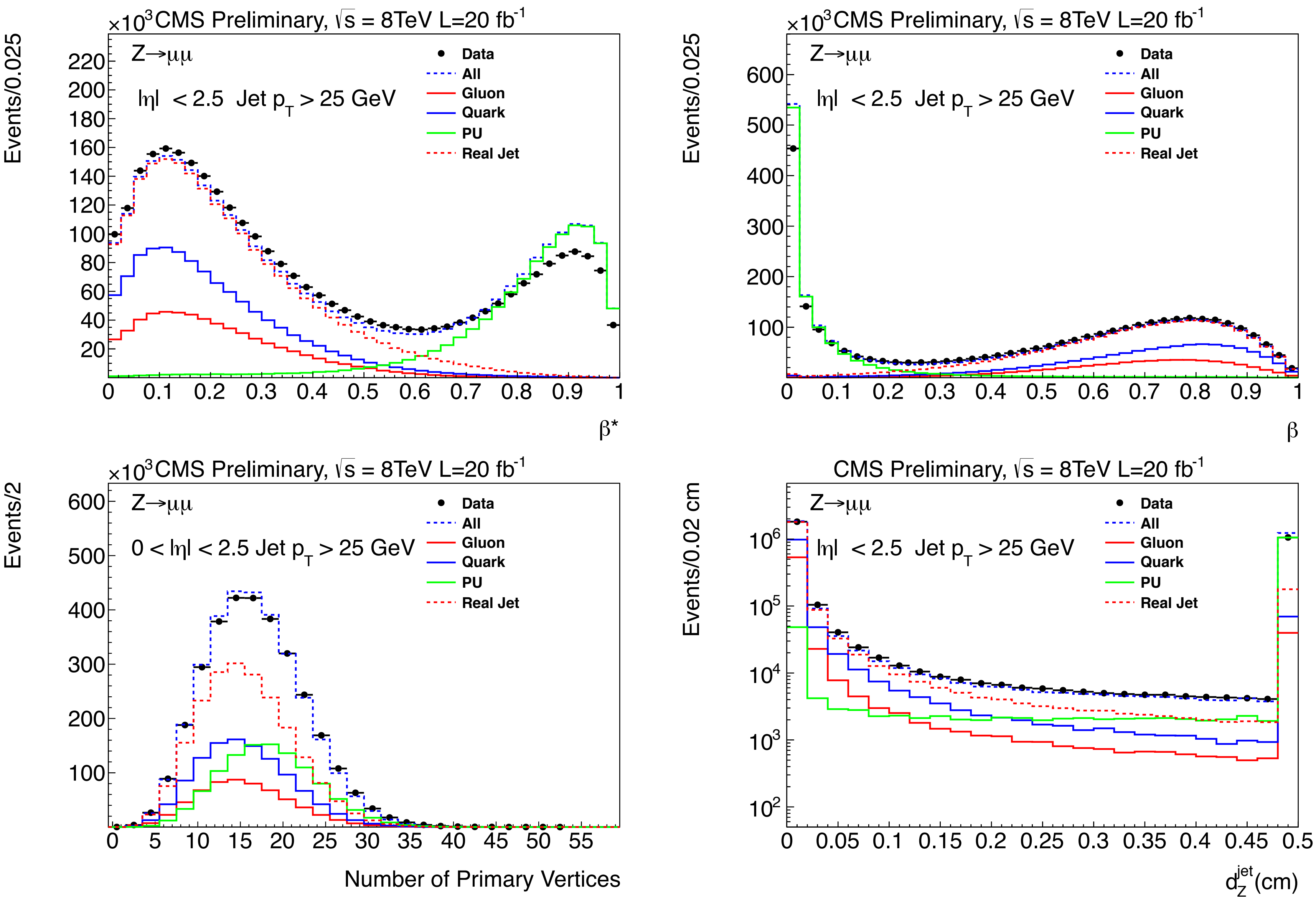
- Take likelihood (LD) distributions for quarks and gluons separately
- Define smearing function: $g(x,a,b) = \tanh(a \operatorname{arctanh}(2x-1)+b)/2+\frac{1}{2}$
  - Changes value of LD on jet-per-jet basis
  - Not a reweighting
  - Reduces LD discrimination
- Smear until data and MC agree
  - χ^2 minimization



Smeared Pythia6 and Herwig++ Agree



Pile Up ID: the Four Track Variables



Pile Up ID: the Other Shape Variables