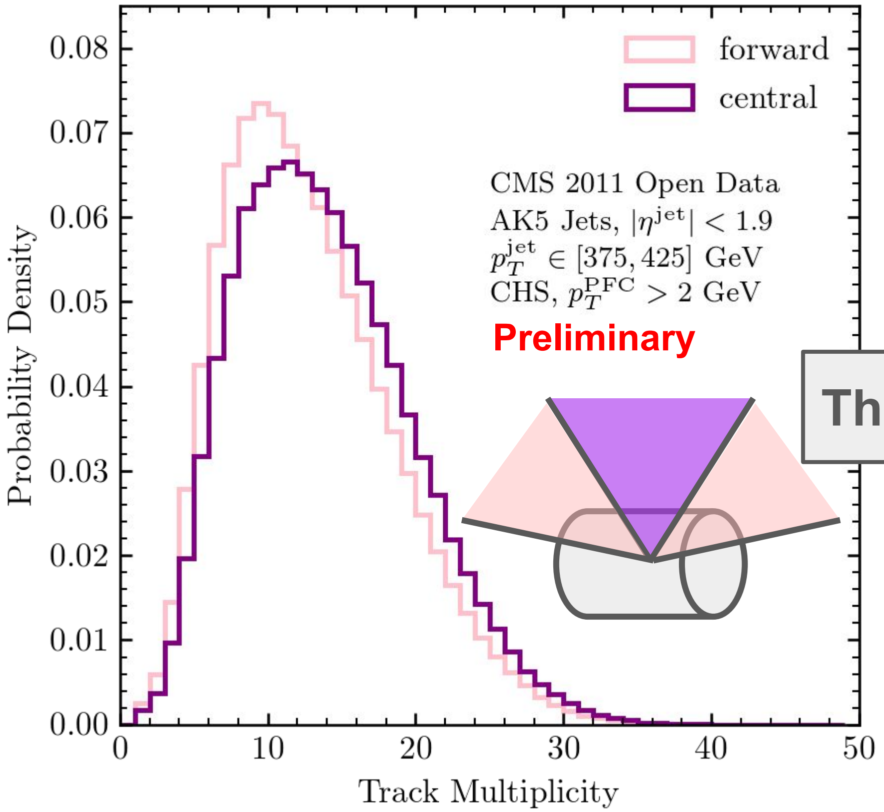
Analyzing CMS Open Collider Data through Topic Modeling

Radha Mastandrea

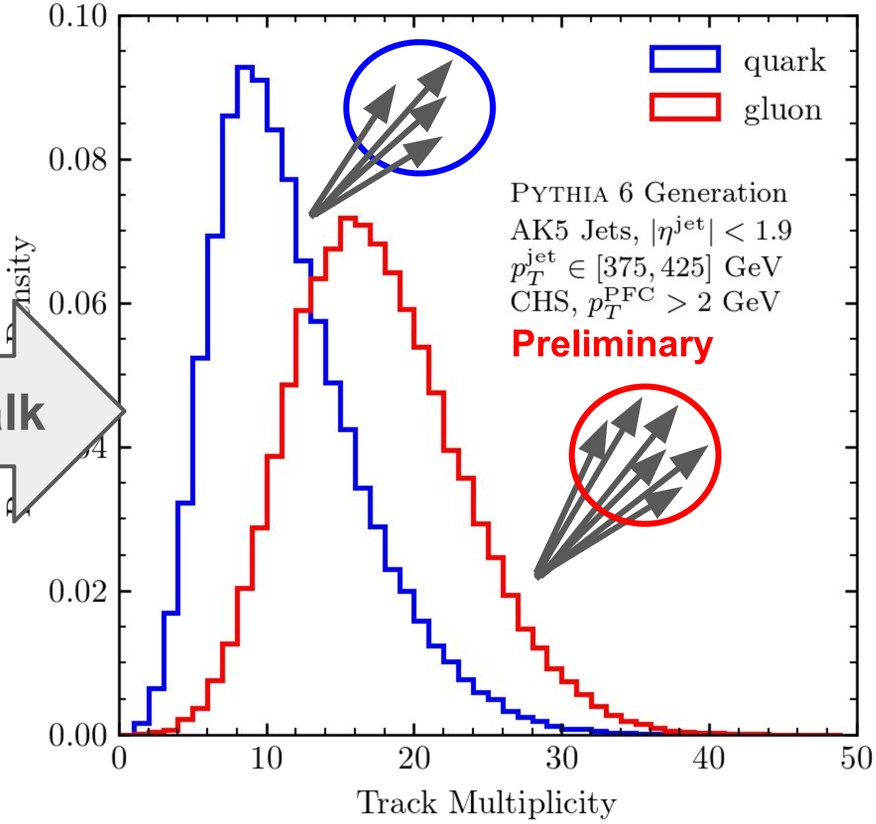
in collaboration with Patrick Komiske, Eric Metodiev, Preksha Naik, and Jesse Thaler

7/25/2019

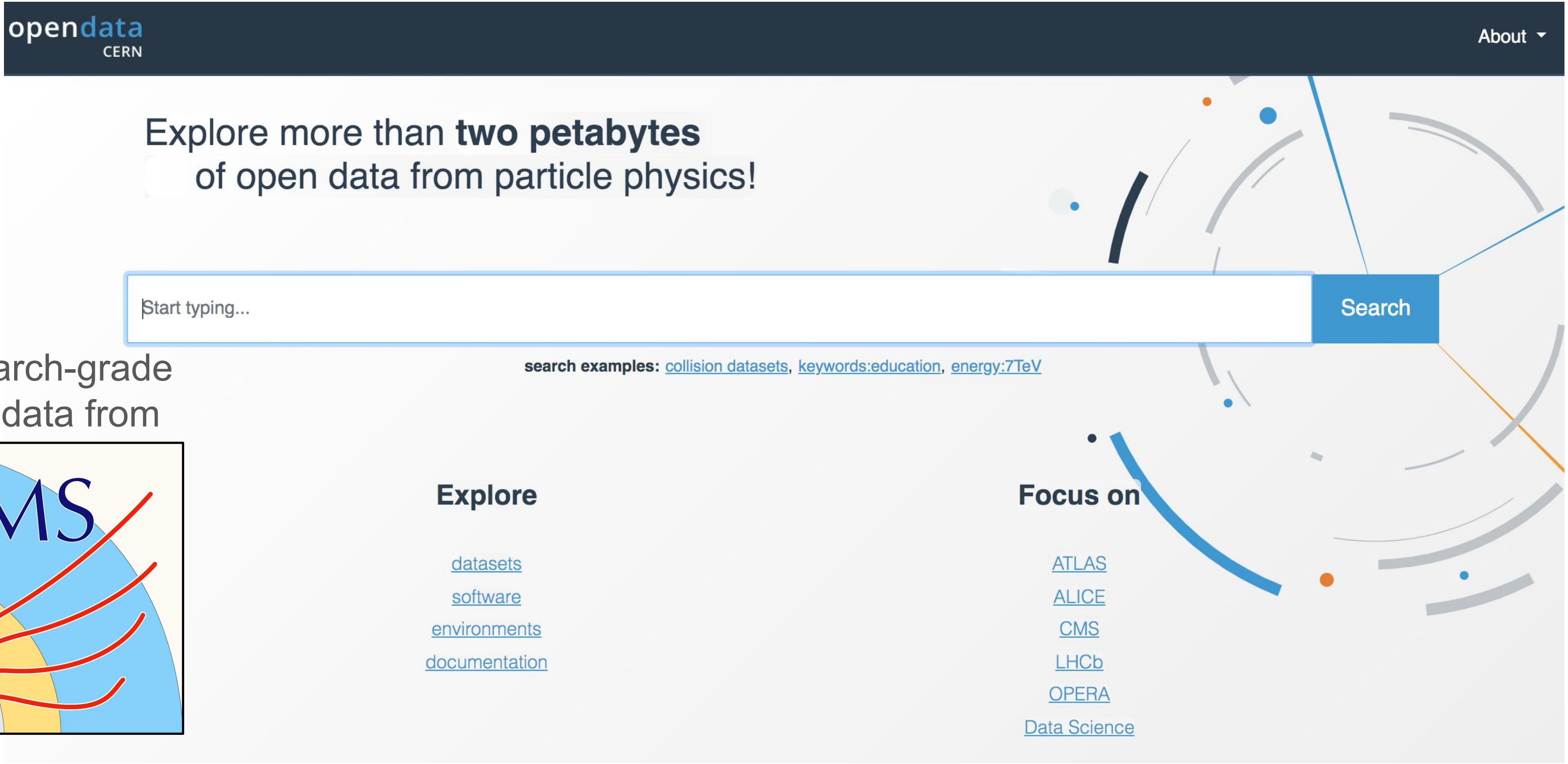
Can we decompose a measured sample of jets into its components?



This talk

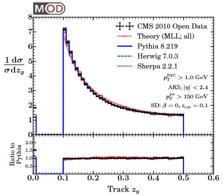


The CERN Open Data portal went live in 2014... opendata.cern.ch



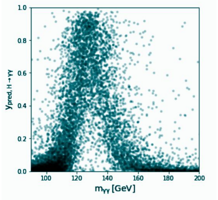
Research-grade LHC data from CMS

...and since then, several exploratory studies have been conducted on the data



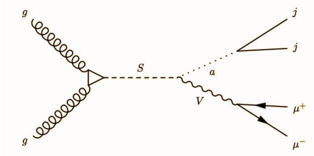
Jet Substructure Studies

Larkoski, Marzani, Thaler, Tripathee, Xue [arxiv:1704.05066]
Tripathee, Xue, Larkoski, Marzani, Thaler [arxiv:1704.05842]



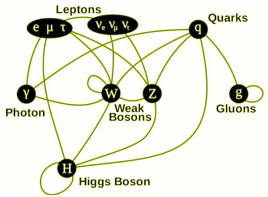
Machine Learning Studies (on simulated data)

Madrazo, Cacha, Iglesias, Marco de Lucas [arxiv:1708.07034]
Andrews, Paulini, Gleyzer, Poczos [arxiv:1807.11916]
Andrews et al. [arxiv:1902.08276]



New Physics Searches

Cesarotti, Soreq, Strassler, Thaler, Xue [arxiv:1902.04222]
Lester, Schott [arxiv:1904.11195]



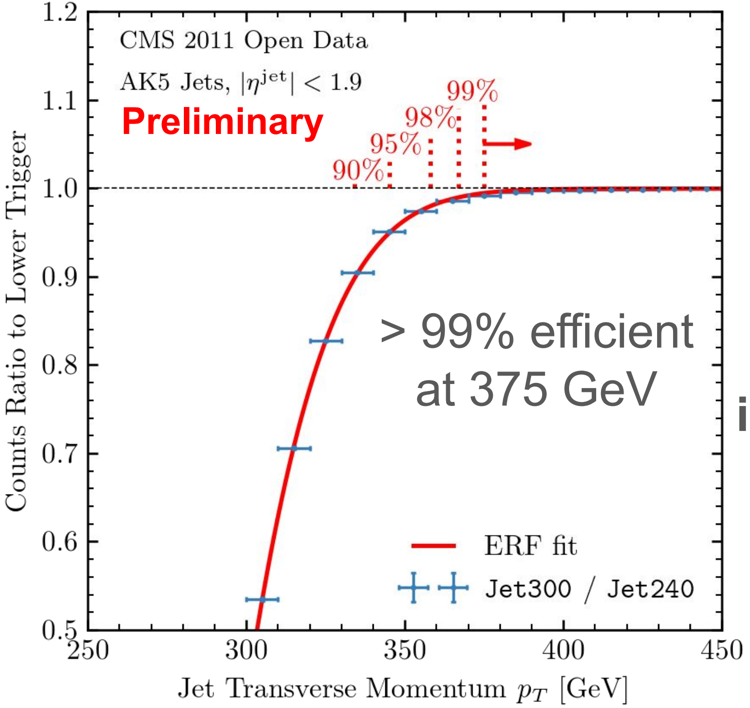
Standard Model Studies

Apyan et al. [arxiv:1907.08197]
Mehdiabadi, Fahim [arxiv:1907.08842]

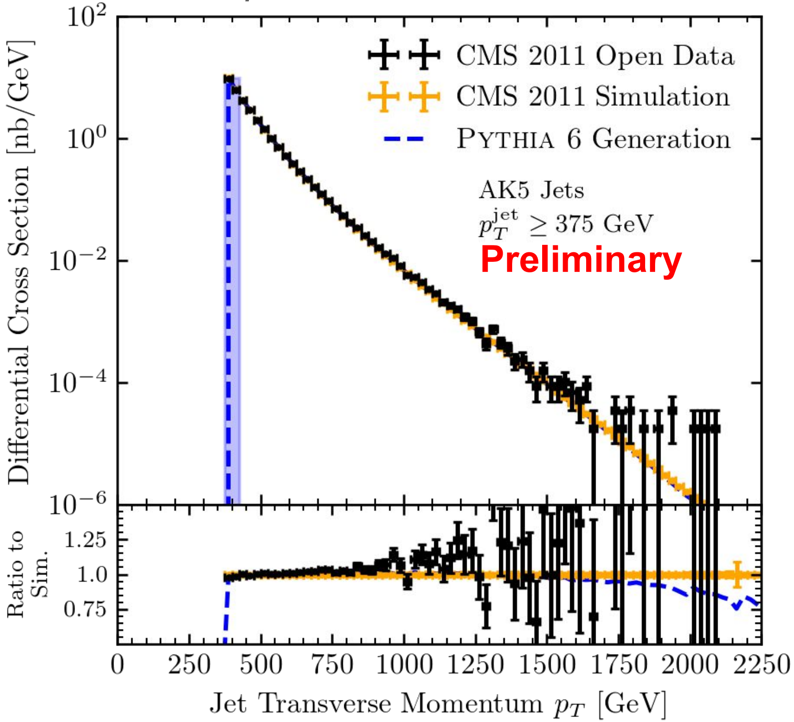
This analysis focuses on moderate-p_T jets

CMS Run 2011A || Jet Primary Dataset || Jet300 Trigger

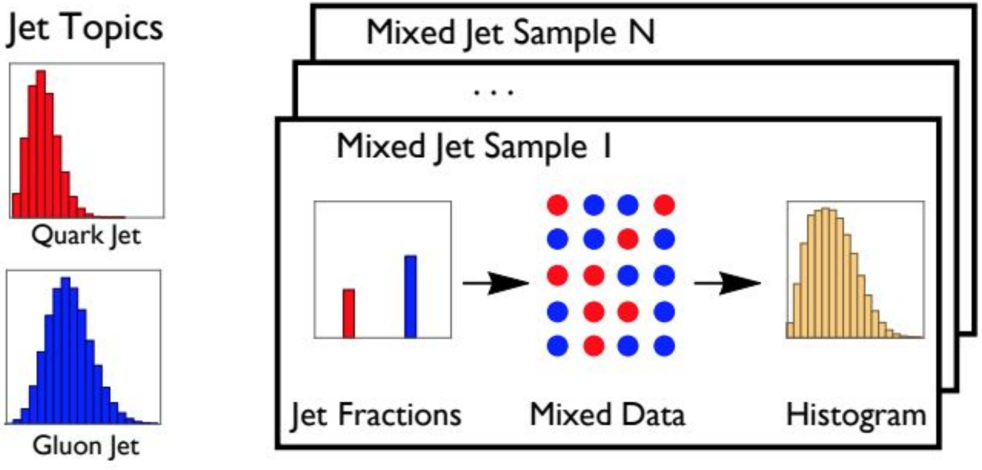
2.3 fb^{-1} || 7 TeV *pp* collisions || anti-kT || R = 0.5 || $p_T \in [375, 425]$ GeV



Hardest two jets treated **independently**



Mixtures of jets can be decomposed using topic modeling



Relevant Studies

Demix [[arxiv:1710.01167]]

- Metodiev, Thaler [arxiv:1802.00008]
- Komiske, Metodiev, Thaler [arxiv:1809.01140]
- ATLAS Collaboration [arxiv:1906.09254]

LDA

- Dillon, Faroughy, Kamenik [arxiv:1904.04200]

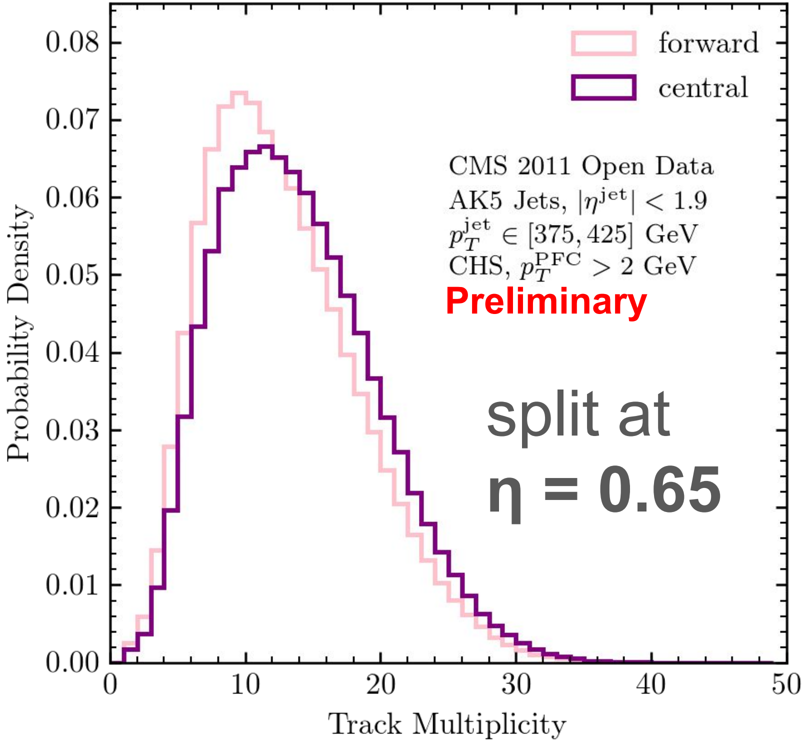
Requirements:

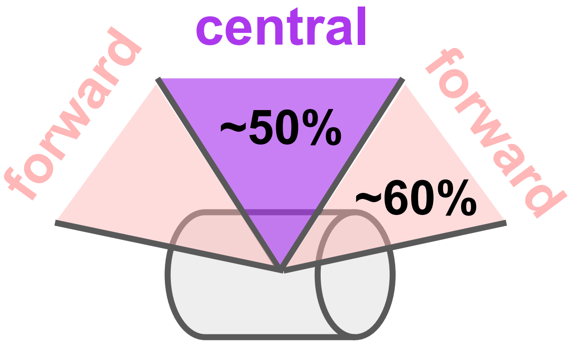
1. **different quark fractions**
2. **sample independence**
3. **anchor bins (AKA mutual irreducibility)**

*see backup slides for explicit formulas

Jet rapidity is an effective quark lever arm

different quark fractions || sample independence || anchor bins



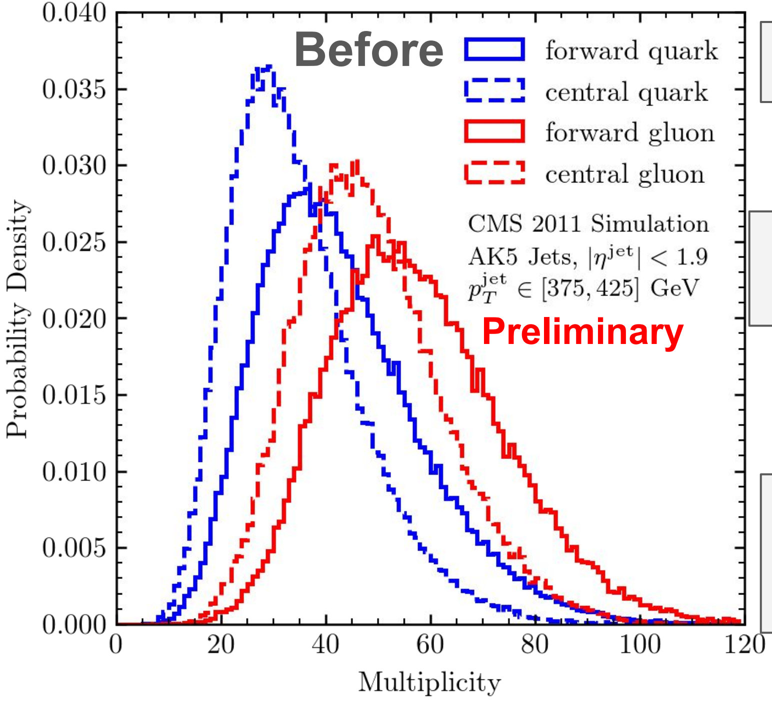


$|\eta_{max}| < 1.9$

for full jet to stay within tracker coverage

Substructure of track-only quark-gluon jets is rapidity-invariant

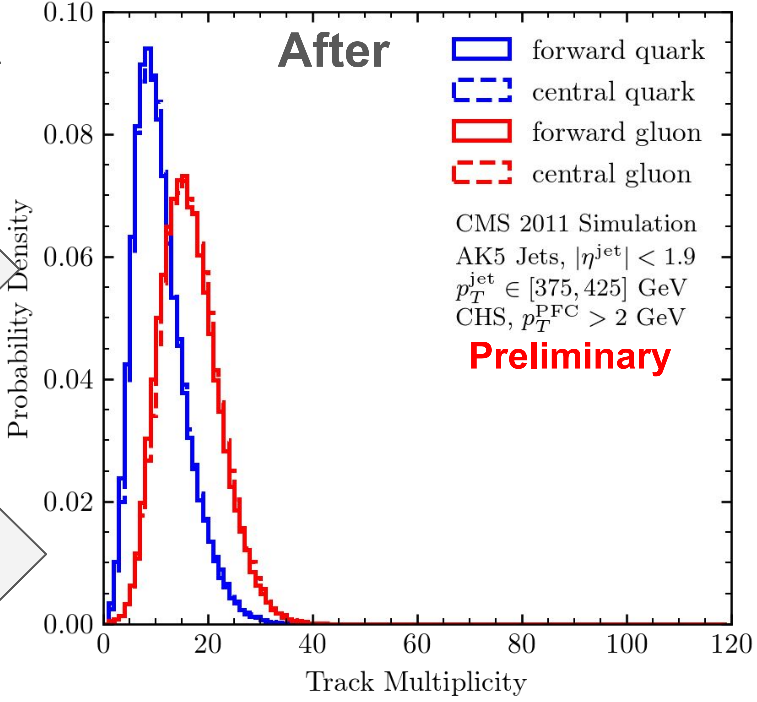
different quark fractions || **sample independence** || anchor bins



Track only

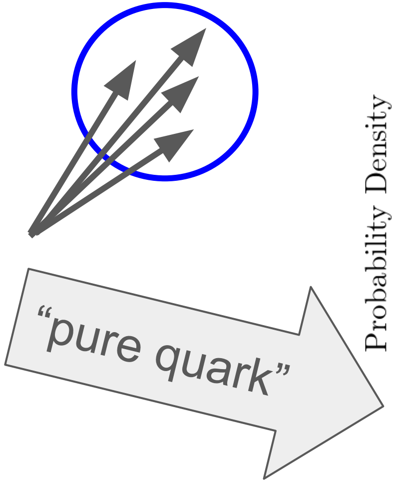
PFC p_T > 2 GeV

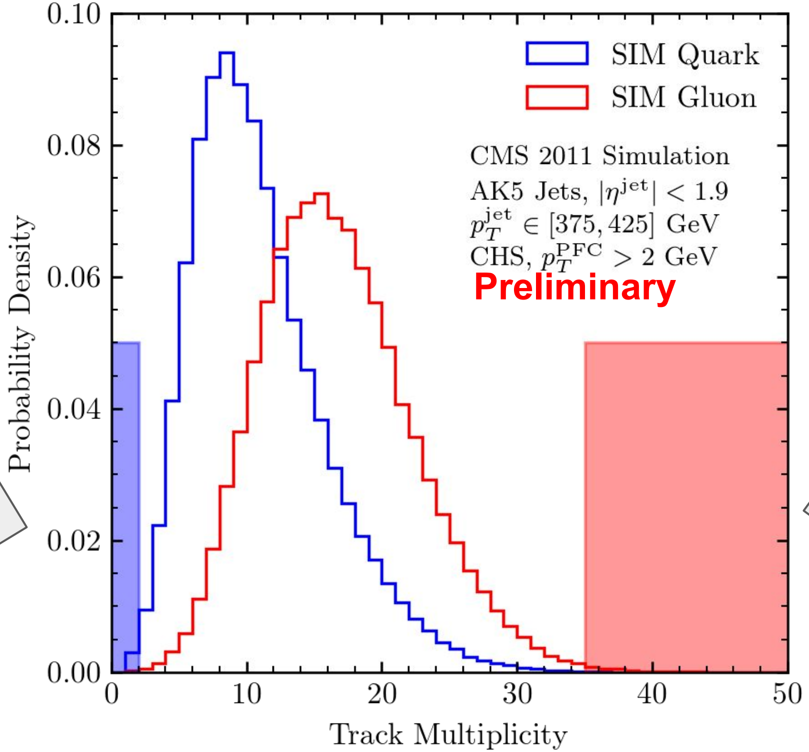
Charged hadron subtraction

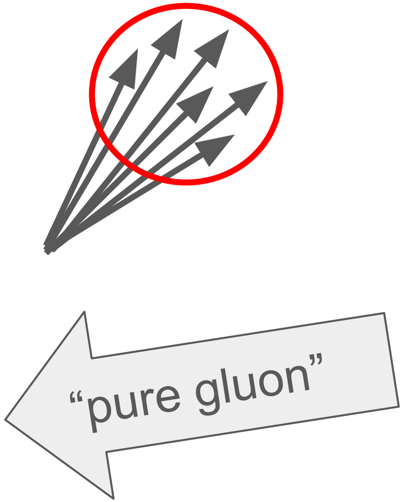


"Anchor bins" define pure quark and gluon phase space regions

different quark fractions || sample independence || **anchor bins**



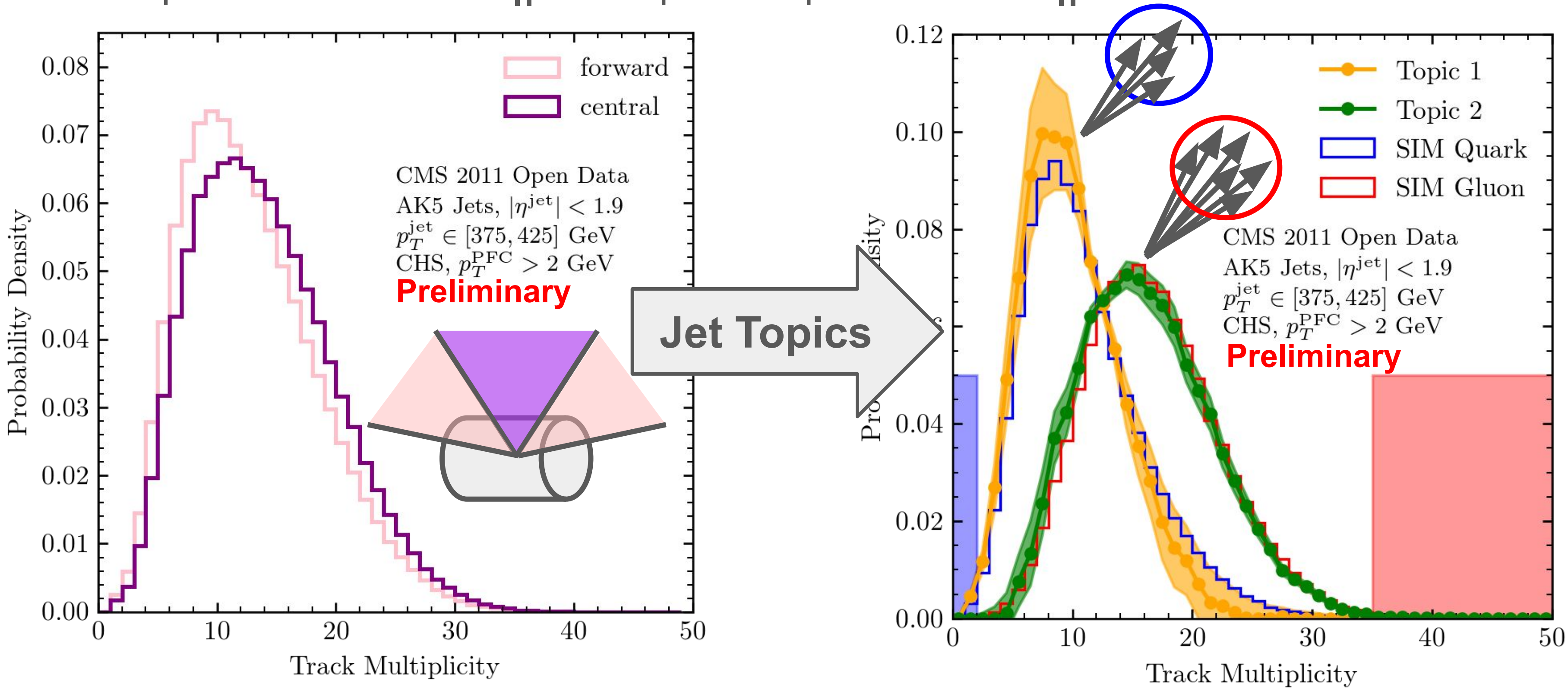




*may not correspond to Pythia parton labels

The topics algorithm recovers quark and gluon jet observable distributions

different quark fractions || sample independence || anchor bins

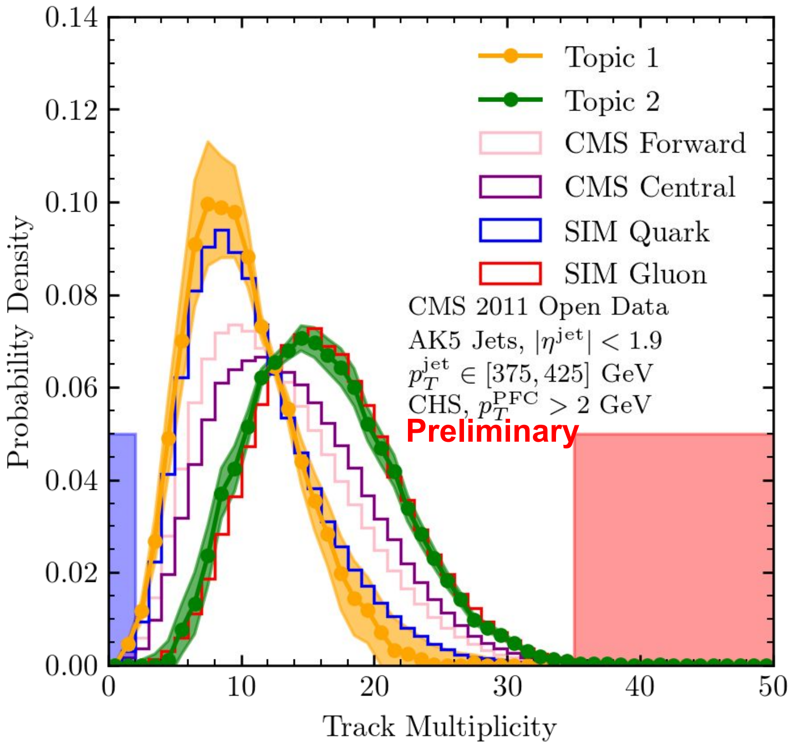


Summary

- Topic modeling has proven itself to be an effective unsupervised machine learning algorithm for decomposing jet mixtures through substructure

- Open data is a valuable tool for exploratory studies, and HEP research is just starting to scratch its surface

- There is incredible potential for both established researchers and new scientists to learn from the CMS open data

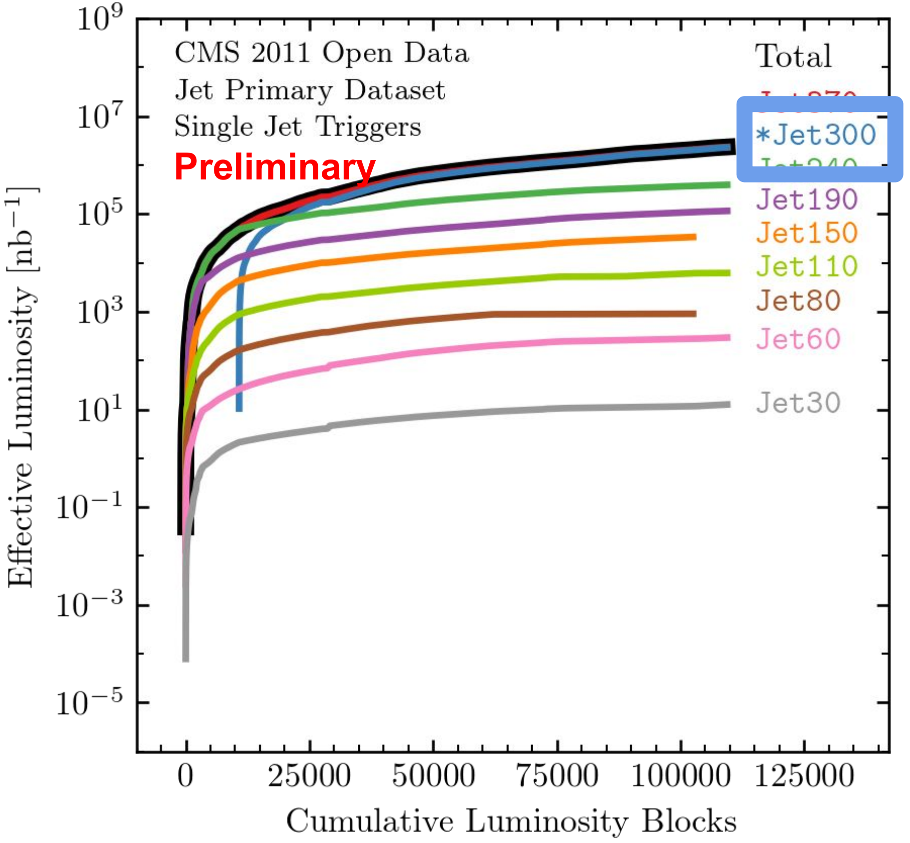


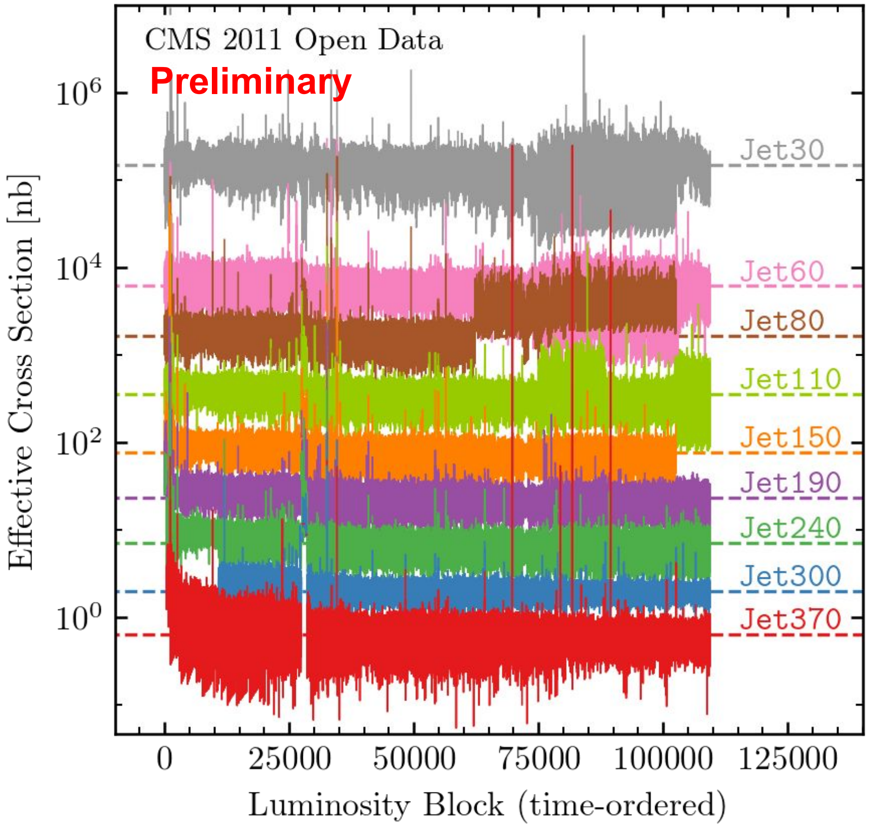
Backup slides

CMS AOD files have been translated to MOD files

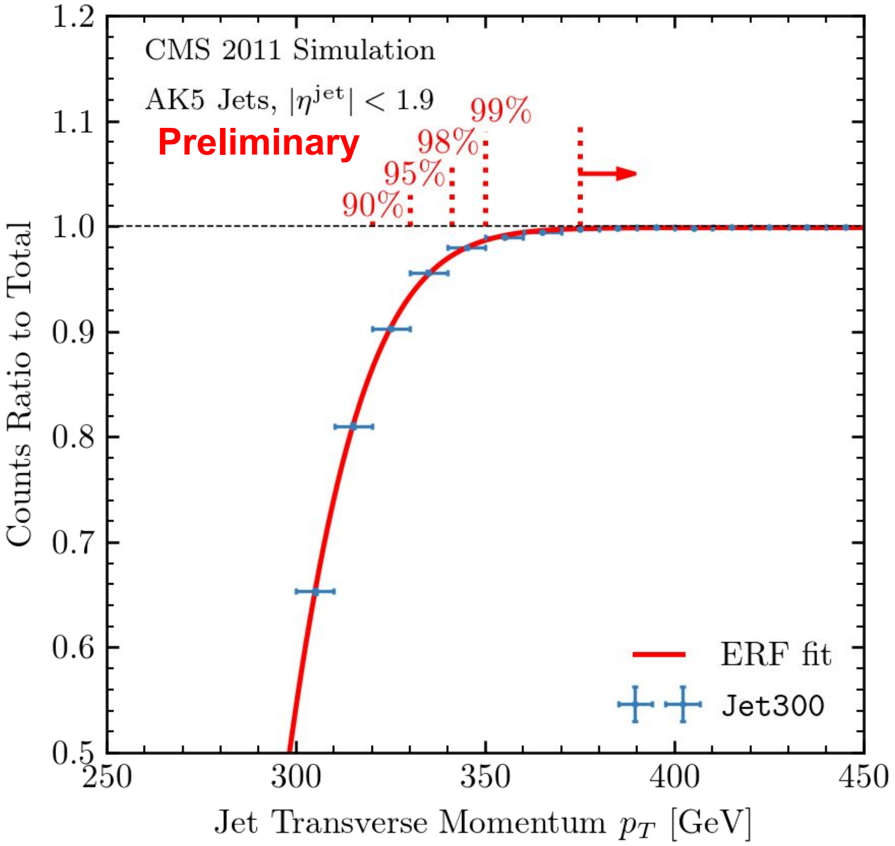
```
BeginEvent Version 6 CMS_2011A Data Jet

# /home/cms-opendata/MITOpenDataProject/eos/opendata/cms/Run2011A/Jet/MOD/12Oct2013-v1/20000/000D4260-D23E-E311-A850-02163E008D77.mod
#    Cond          RunNum        EventNum       LumiBlock             NPV      Timestamp        msOffset
     Cond          160578        38142433             366               4     1300254008          84656

#    Trig                                Name          Prescale_1       Prescale_2        Fired?
     Trig                    HLT_DiJetAve30U_v4                 1               15             0
     Trig                    HLT_DiJetAve50U_v4                 1                3             1
     Trig                    HLT_DiJetAve70U_v4                 1                1             0
     Trig                          HLT_Jet110_v1                1                1             1
     Trig                          HLT_Jet150_v1                1                1             0

#     AK5           px            py              pz          energy              jec   no_of_const   chrg_multip
      AK5  -48.53112195   91.23529327    922.46206960    928.25796767      1.15373647             3             0
      AK5   27.14014056  -27.95814987   -176.24652474    180.60830433      1.11999369            14             8
      AK5    6.87947531  -27.39585642   -127.71244347    130.89105131      1.13558543            10             3
      AK5   -1.21714232   -9.77158690    -26.87058049     28.71690560      1.14206147             8             2

#     PFC           px            py              pz          energy           pdgId           PV?
      PFC    3.05231479   -2.27686970    -18.08729449     18.48433020            211             1
      PFC    7.15976356   -7.56236808    -46.86929997     48.01232034             22             0
      PFC    1.88167876   -1.89435884    -12.60399834     12.88371393            130             0
      PFC    0.40022073   -0.42509065     -2.47631023      2.54420735             22             0
      PFC    5.19161920  -18.85569567    -84.84289283     87.06793964           -211             0
      PFC    0.41414809   -0.59229172     -2.27328073      2.38540005            130             0
      PFC   -0.35573217   -0.03071949     -1.14696234      1.20933513           -211             3
      PFC    0.18477403   -0.39789019     -3.45412354      3.48187127            130             0

EndEvent
```

The Jet 2011 dataset contains many single-jet triggers

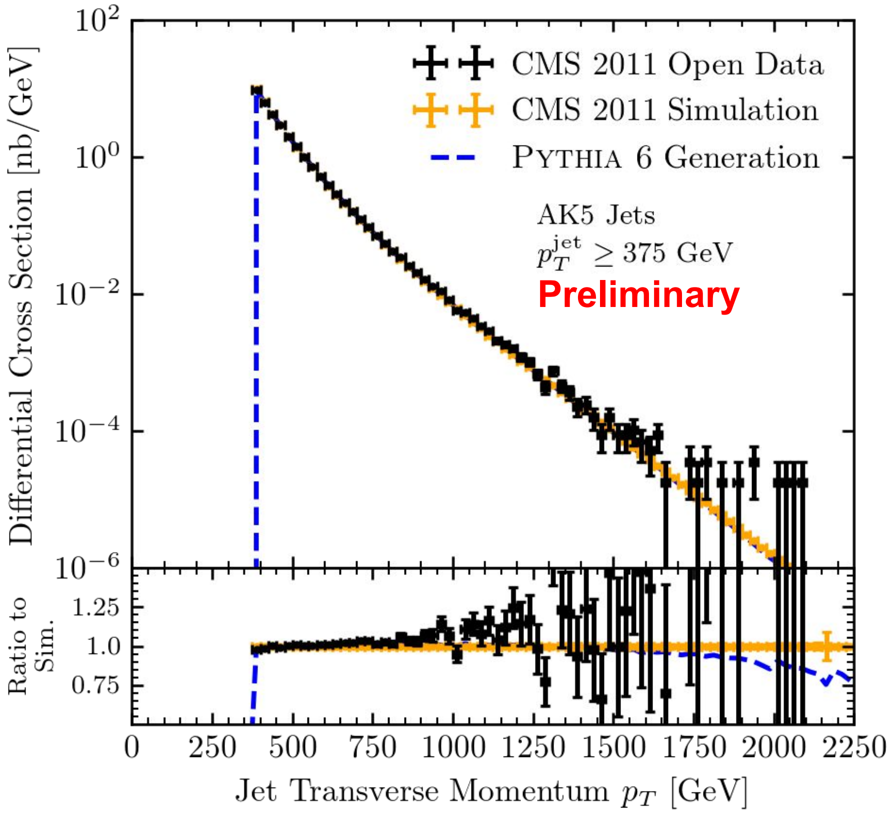


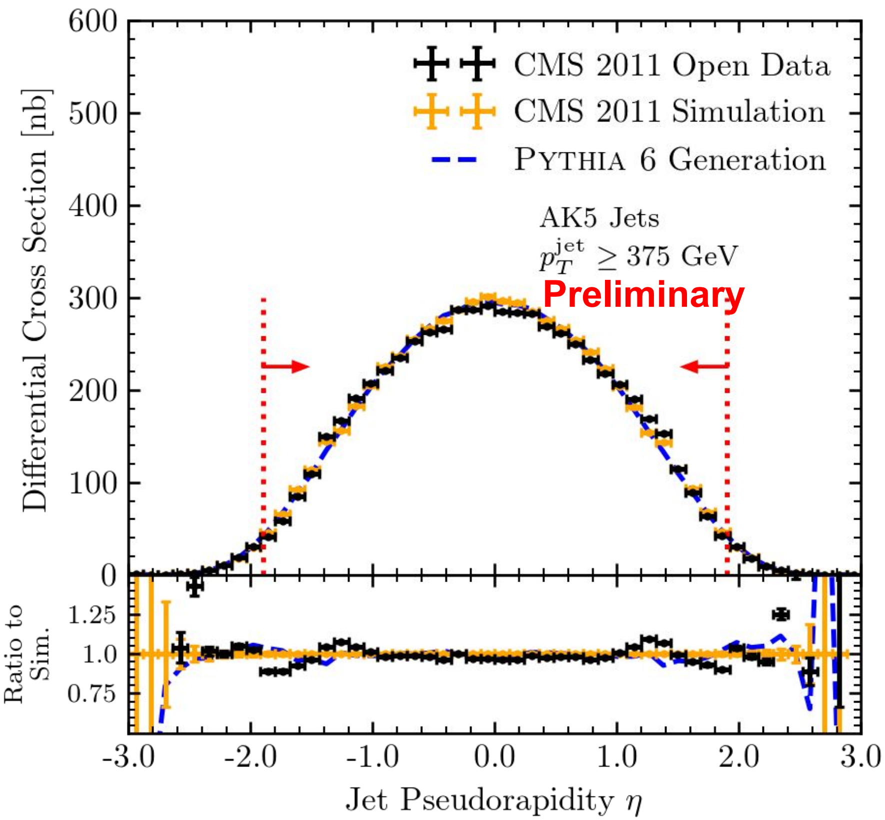


We can cross-check trigger efficiencies with CMS simulated data



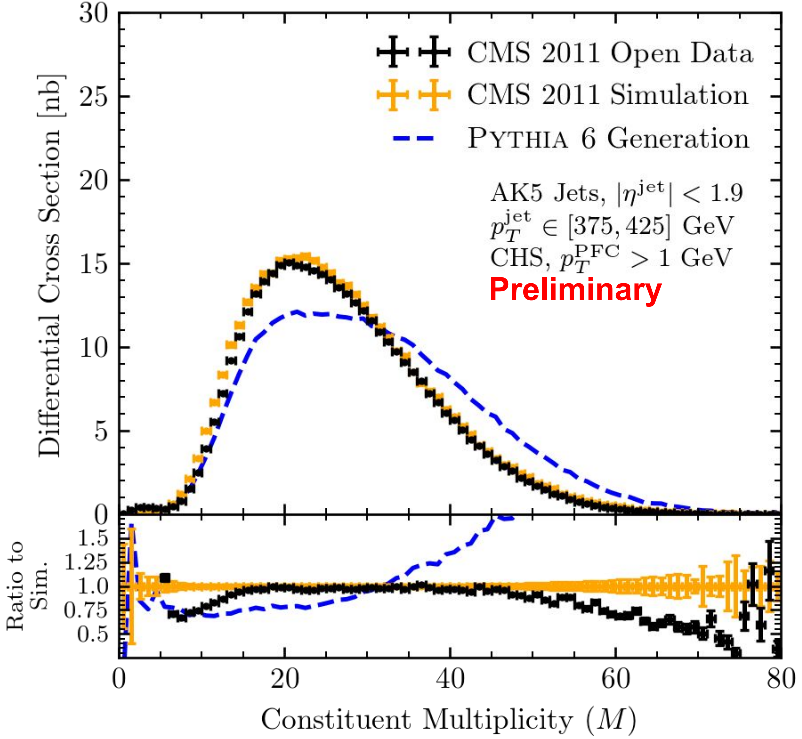
There is good agreement between detected and simulated CMS data



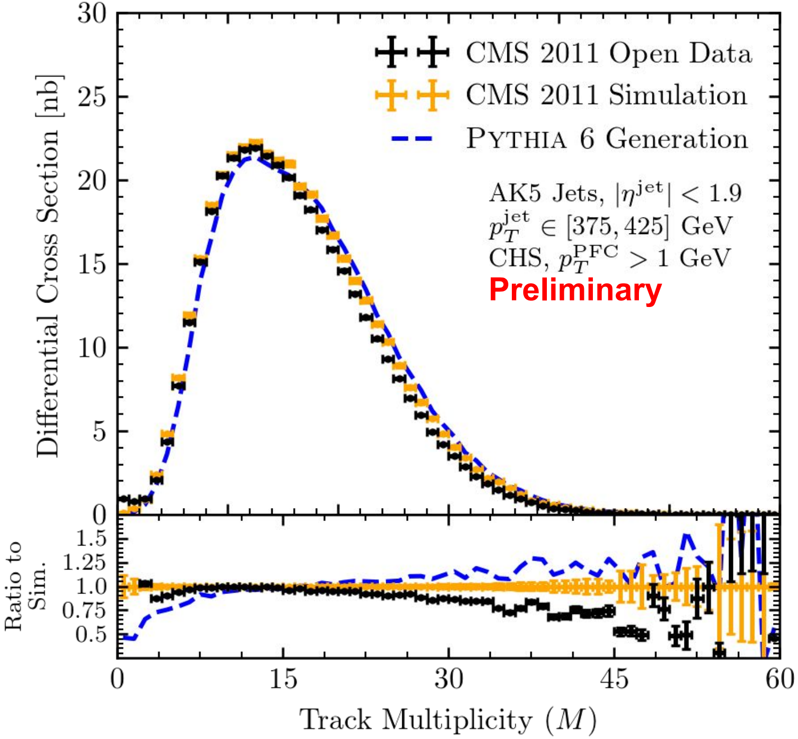


We can pick out the most robust jet observables using simulated data

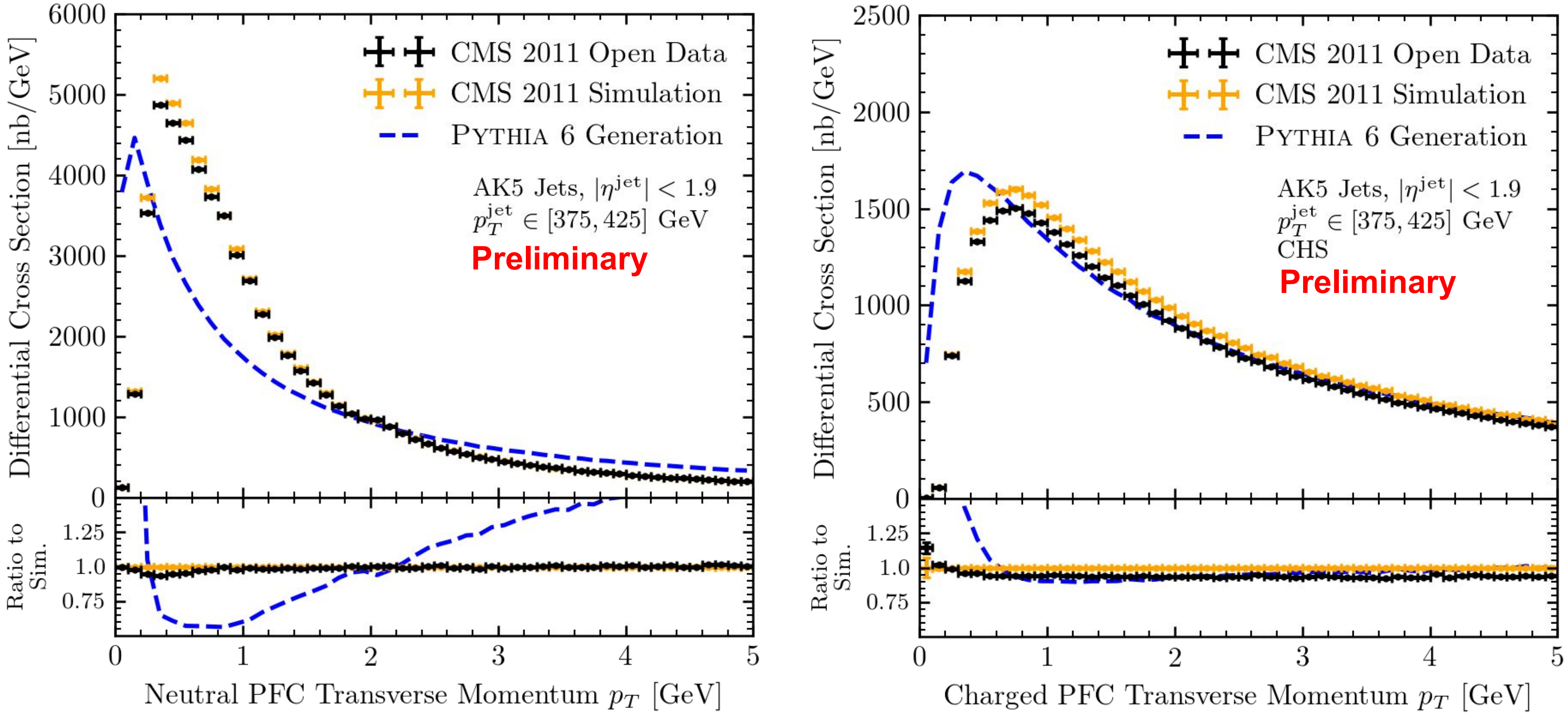
Data and Sim agree, but not with Gen



Track-based observables minimize detector effects



Charged hadron subtraction is necessary to mitigate pileup



The topics algorithm is summarized in a few equations

Definitions

$$\mathbf{x} = \text{jet observable}$$

$$p(\mathbf{x}) = \text{jet observable distribution}$$

$$f_q = \text{quark fraction}$$

$$f_g = \text{gluon fraction}$$

$$\mathcal{L} = \text{log-likelihood ratio}$$

Setup

$$p(\mathbf{x}) = f_q p_q(\mathbf{x}) + f_g p_g(\mathbf{x})$$

$$f_q = 1 - f_q \qquad f_q^{(1)} > f_q^{(2)}$$

Theory

$$\kappa(M_1|M_2) = \frac{1 - f_q^{(1)}}{1 - f_q^{(2)}}$$

$$\kappa(M_2|M_1) = \frac{f_q^{(2)}}{f_q^{(1)}}$$

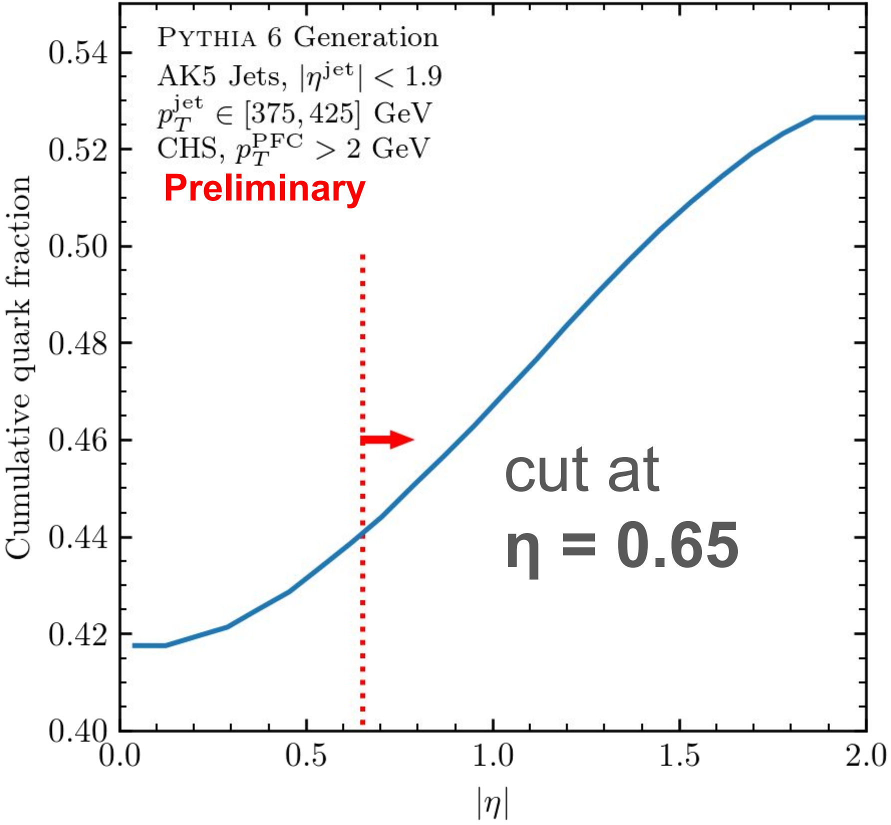
Topic determination

$$\kappa(M_1|M_2) = \exp(-\mathcal{L}(\text{upper anchor bin}))$$

$$\kappa(M_2|M_1) = \exp(\mathcal{L}(\text{lower anchor bin}))$$

$$p_{T_1}(\mathbf{x}) = \frac{p_{M_1}(\mathbf{x}) - \kappa(M_1|M_2)p_{M_2}(\mathbf{x})}{1 - \kappa(M_1|M_2)}$$

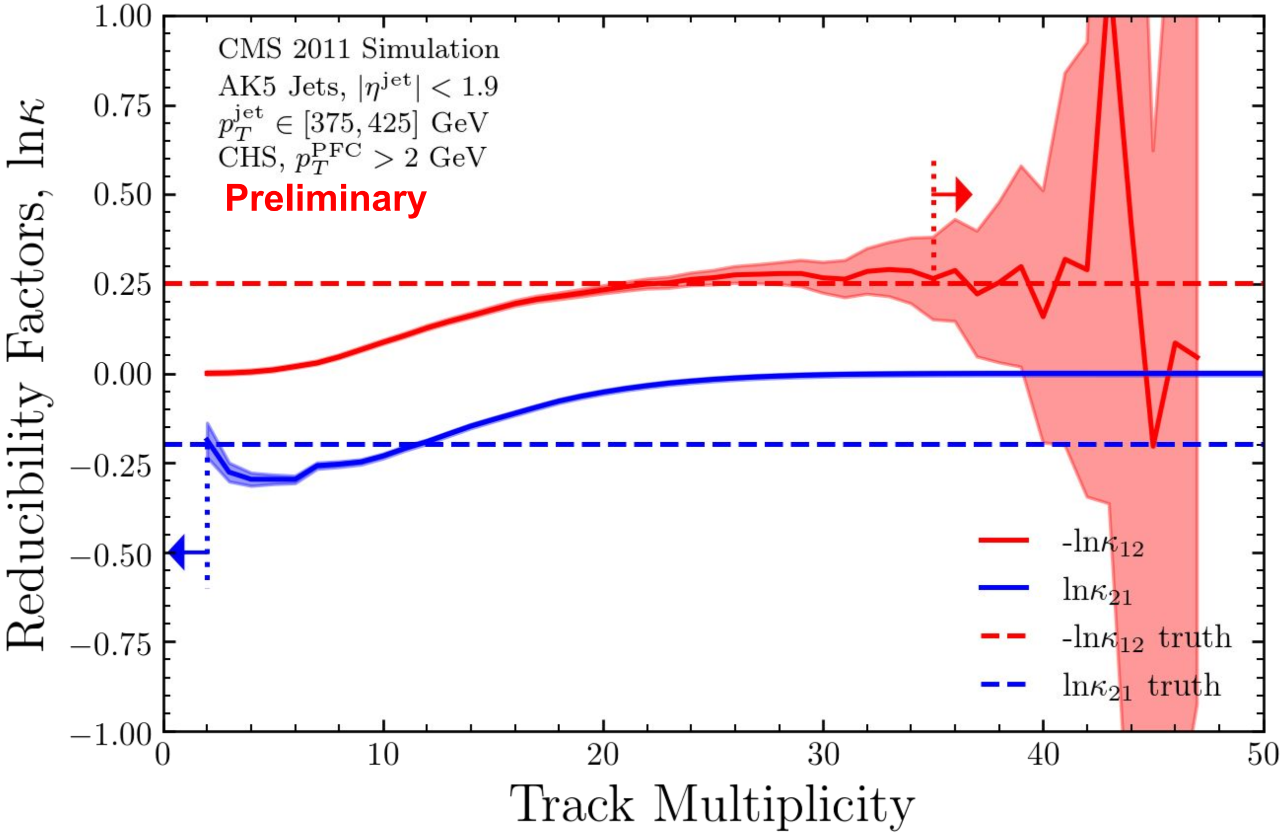
Jet quark fraction is dependent on rapidity



cut at

η = 0.65

Quark and gluon anchor bins are determined quantitatively



Topic recovery is robust with respect to an η gap