Data-driven extraction of quark and gluon jet substructure in proton-proton and heavy-ion collisions

Kylie Ying

Relativistic Heavy Ion Group

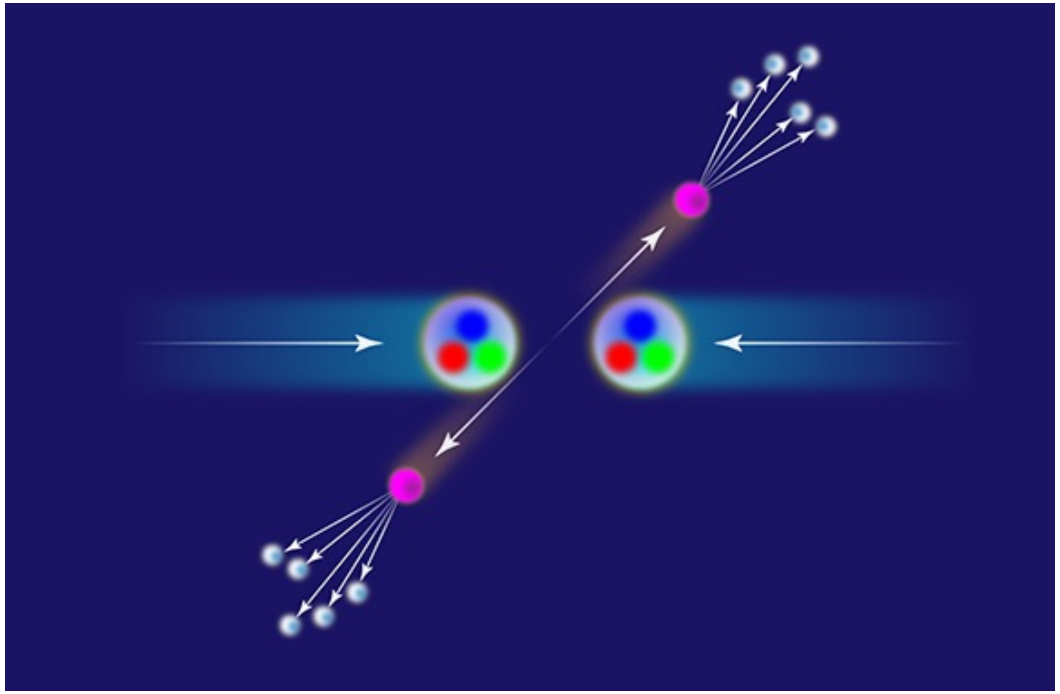
Massachusetts Institute of Technology

Work in collaboration with Yen-Jie Lee, Yi Chen, and Jasmine Brewer

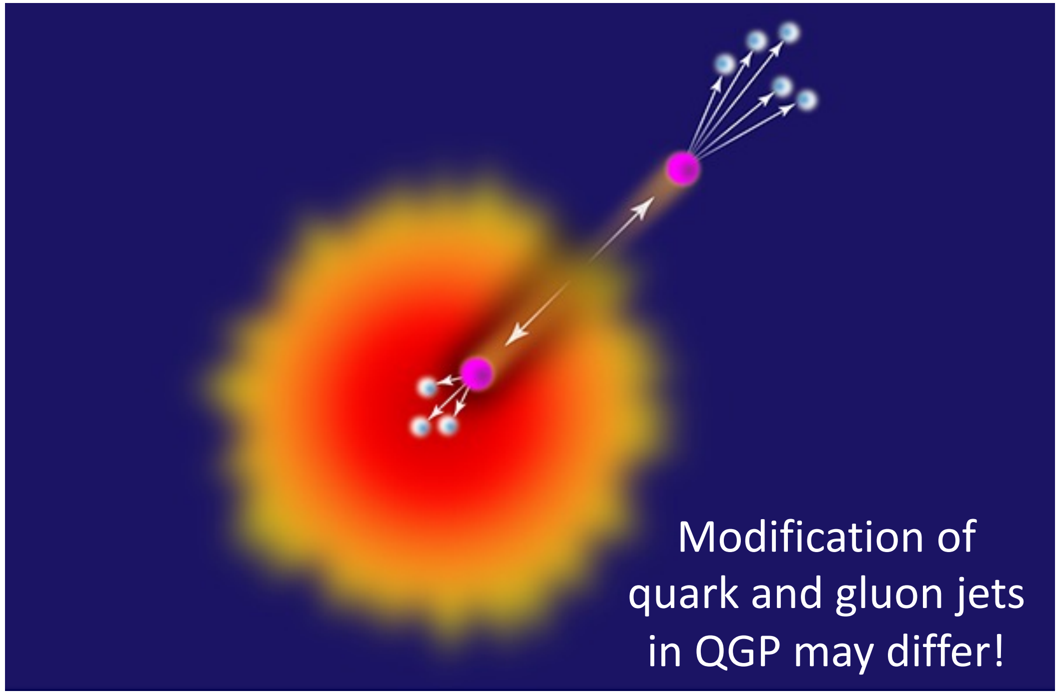
October 12, 2021

What happens during a particle collision?

pp collision



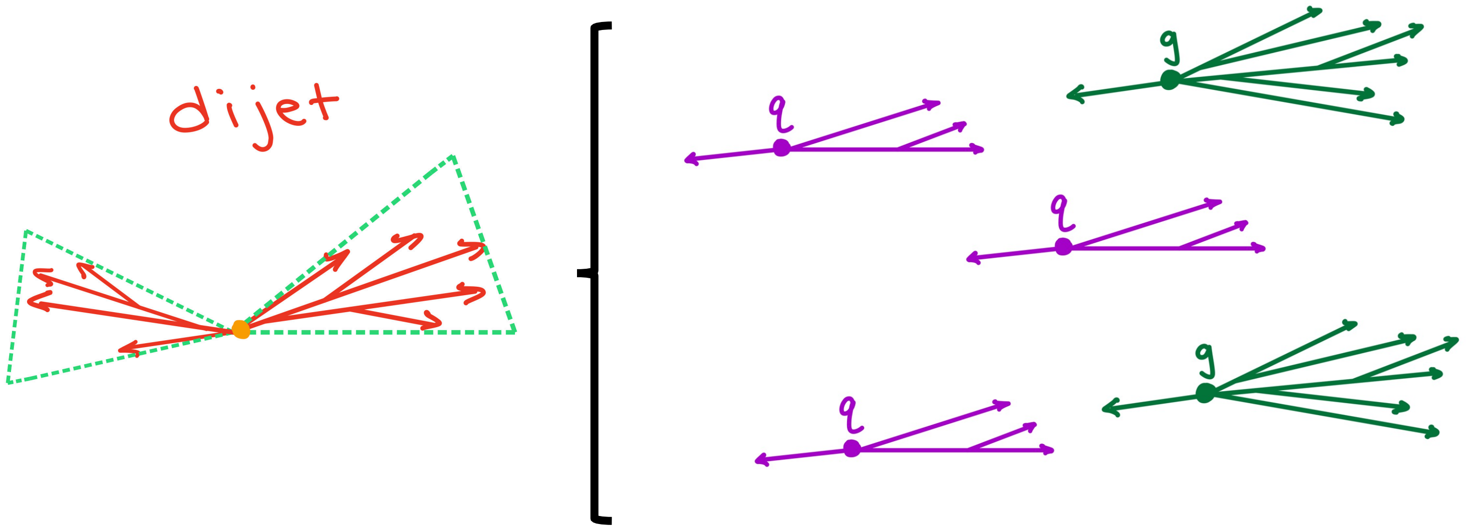
PbPb collision



[Manuel, APS]

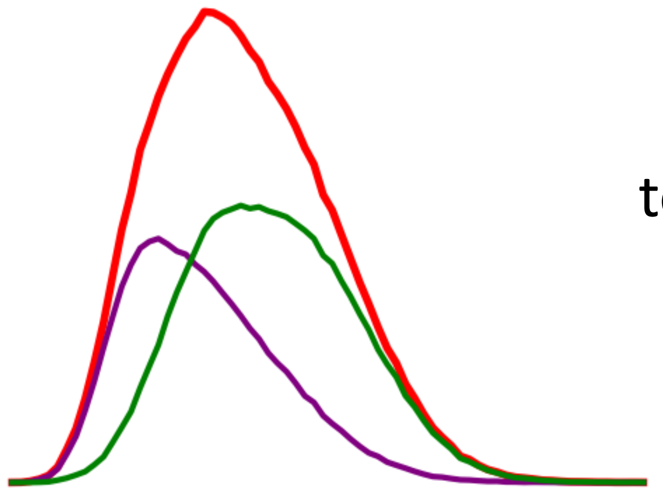
Distinguishing quark and gluon jets

- Collected jet samples are mixtures of quark and gluon jets



Distinguishing quark and gluon jets

- Collected jet samples are mixtures of quark and gluon jets

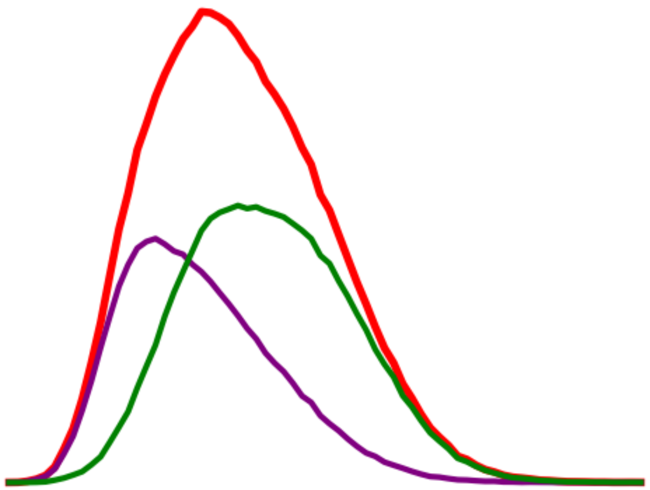


Jet observable

Topic Modeling
to extract jet substructure

Distinguishing quark and gluon jets

- Collected jet samples are mixtures of quark and gluon jets

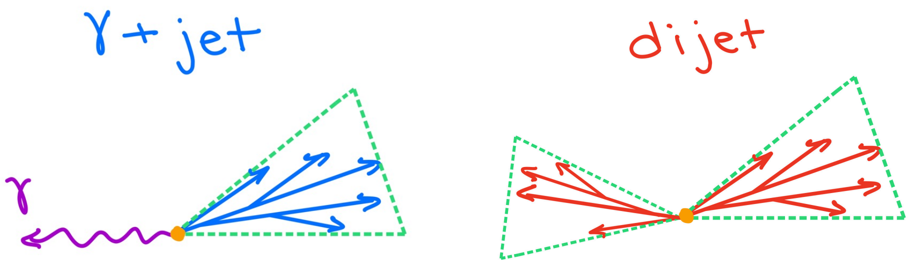


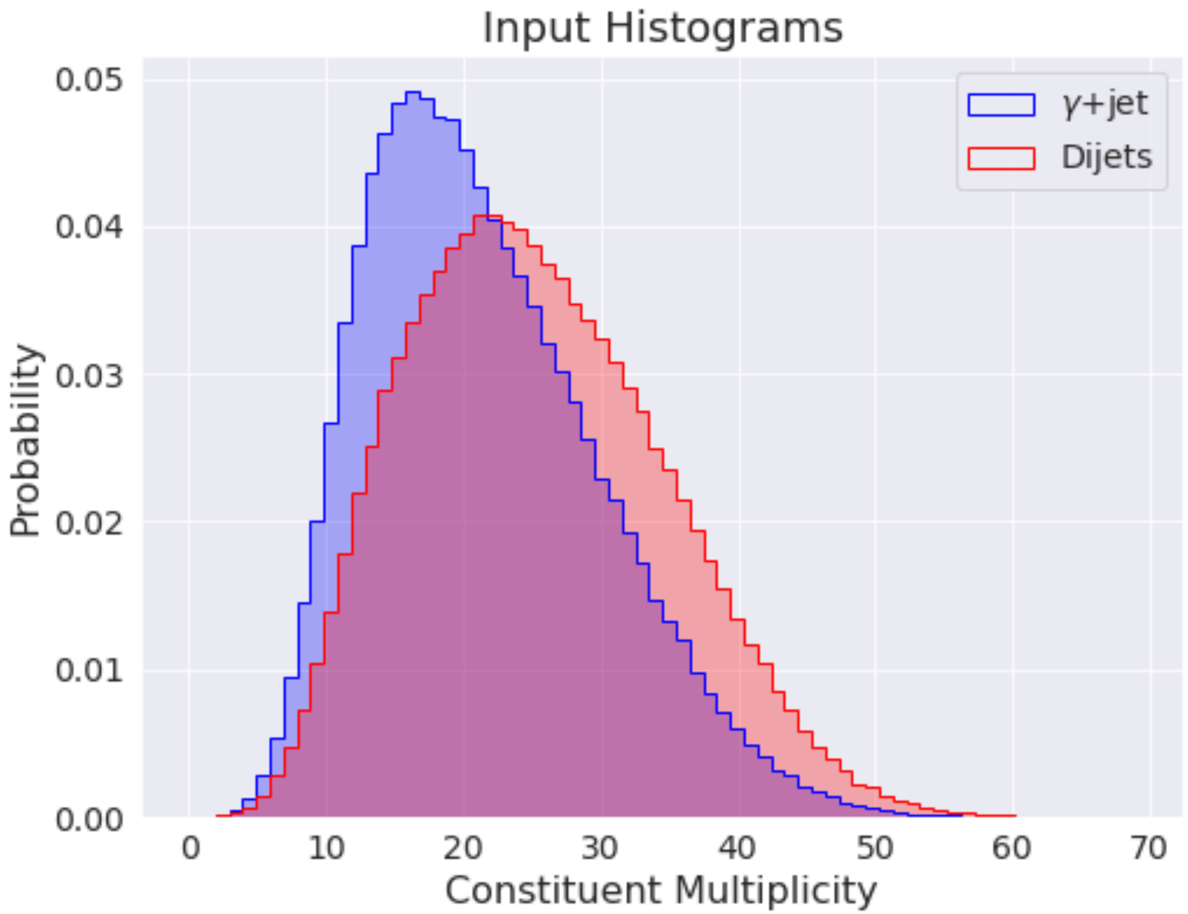
Jet observable

Topic Modeling
with machine learning
observables

Observed samples are mixtures!

- Two input distributions:
  - $p_{\gamma+jet}(x)$ and $p_{dijets}(x)$



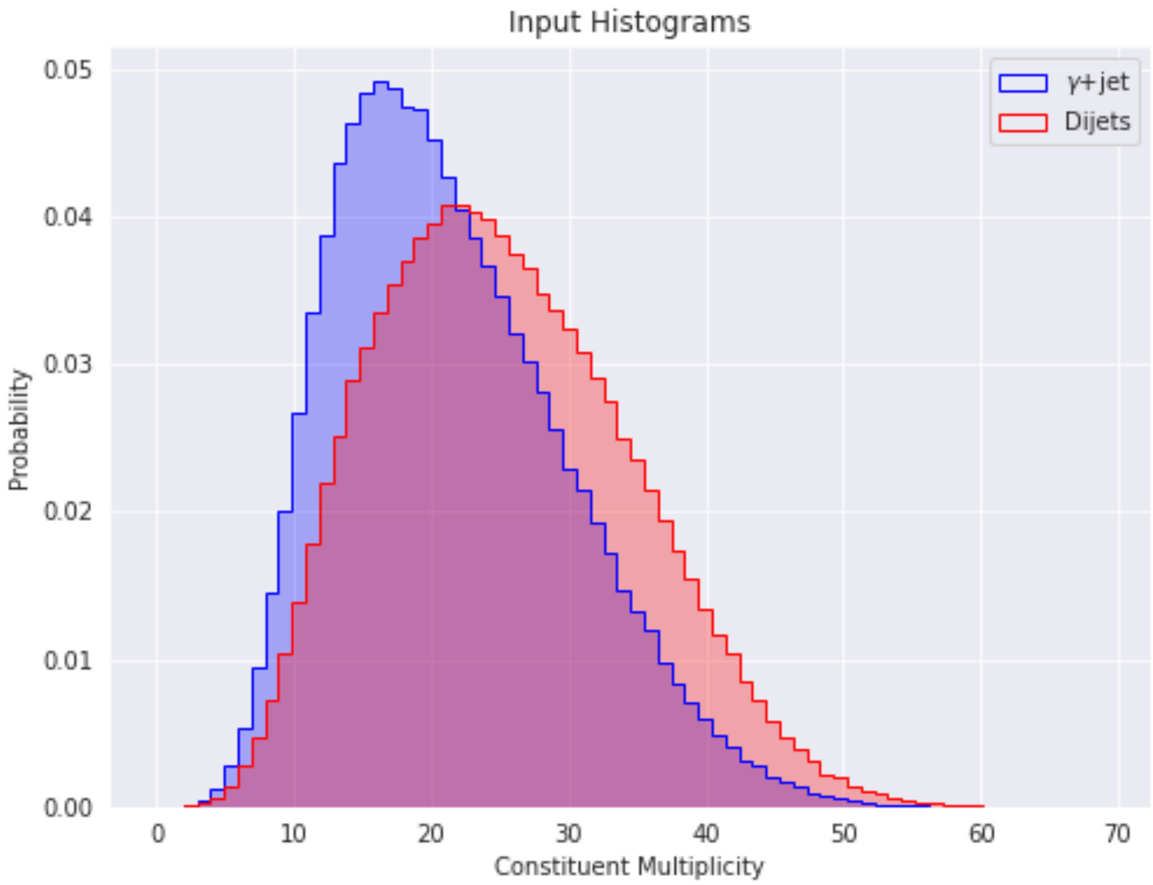


PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Observed samples are mixtures!

- Two input distributions:
  - $p_{\gamma+jet}(x)$ and $p_{dijets}(x)$

At LHC energies, γ+jet and dijets have different quark/gluon contributions

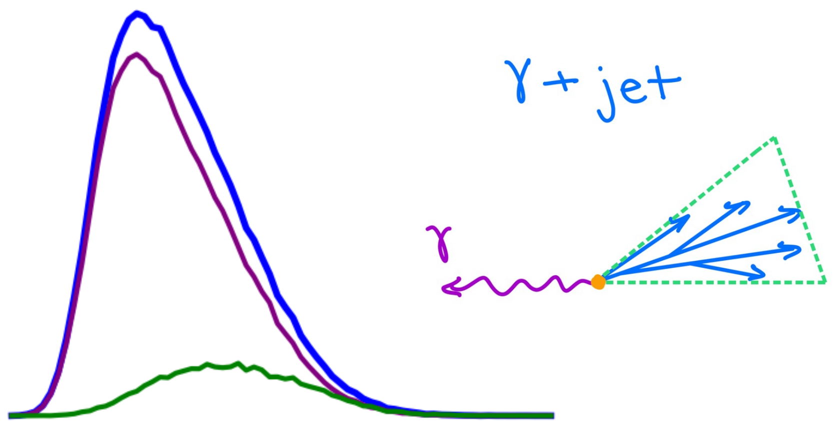


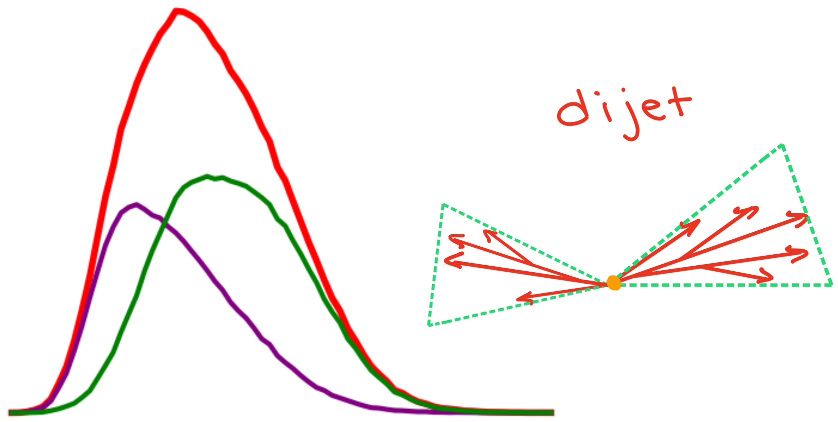
PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Observed samples are mixtures!

- Two input distributions:
  - $p_{\gamma+jet}(x)$ and $p_{dijets}(x)$
- These are mixtures of base distributions:
  - $b_1(x)$ and $b_2(x)$
- In other words:
  - $p_{\gamma+jet}(x) = f_1\, b_1(x) + (1 - f_1)\, b_2(x)$
  - $p_{dijets}(x) = f_2\, b_1(x) + (1 - f_2)\, b_2(x)$

However, there are infinitely many ways to define the fractions and distributions!

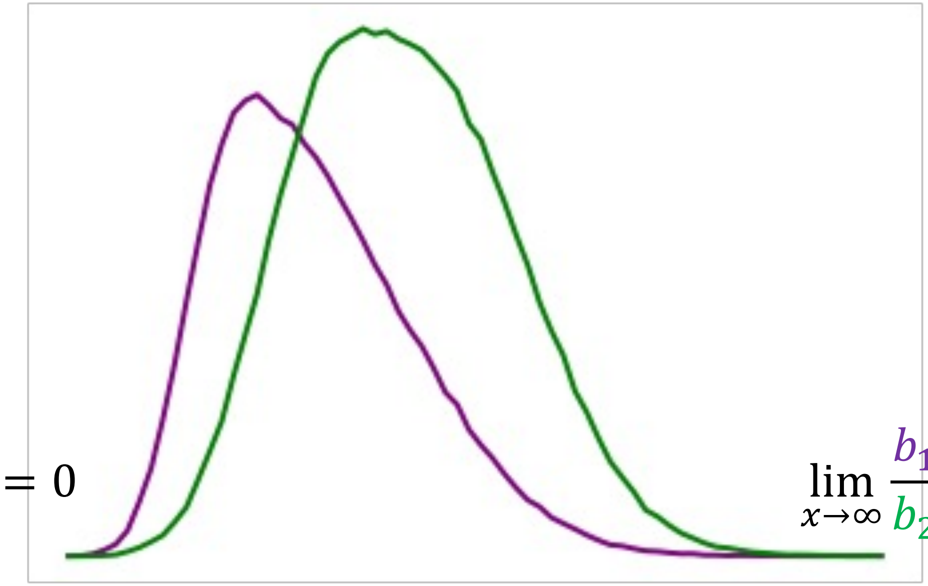




Mutual irreducibility

- If we require the base distributions, $b_1(x)$ and $b_2(x)$, to be **mutually irreducible**, then we can resolve this ambiguity!

$$\lim_{x\to-\infty} \frac{b_1(x)}{b_2(x)} = 0$$



$$\lim_{x\to\infty} \frac{b_1(x)}{b_2(x)} = 0$$

Mutual irreducibility:

- We cannot write $b_1(x) = c\, b_2(x) + (1-c)F$, where F is some probability distribution and $0 < c \leq 1$
- At the limits, the samples are “pure” topic 1 or topic 2

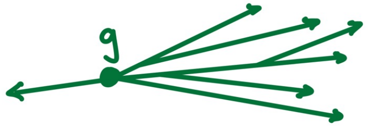
[Komiske, *et al.*, 1809.01140]

Computing base distributions

$$\kappa_1 = \inf \frac{p_{\gamma+jet}(x)}{p_{dijets}(x)} \qquad \kappa_2 = \inf \frac{p_{dijets}(x)}{p_{\gamma+jet}(x)}$$

- We can compute base distributions from the mixtures:

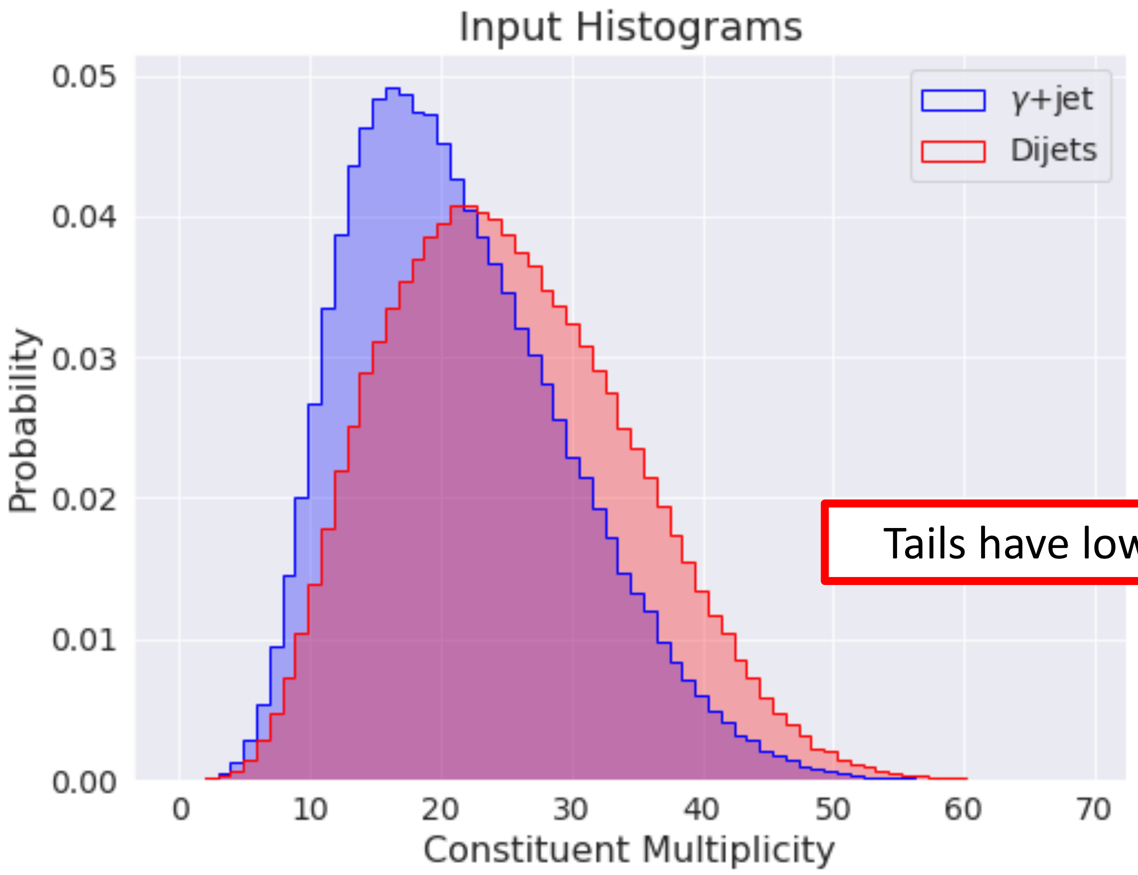
$$b_1(x) = \frac{p_{\gamma+jet}(x) - \kappa_1 p_{dijets}(x)}{1-\kappa_1}$$



$$b_2(x) = \frac{p_{dijets}(x) - \kappa_2 p_{\gamma+jet}(x)}{1-\kappa_2}$$

[Brewer, *et al.*, 2008.08596]

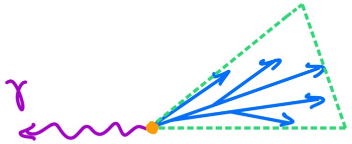
How do we get κ?



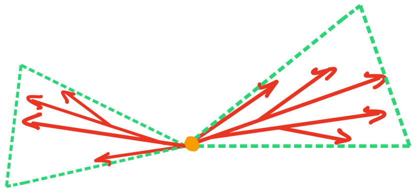
Tails have low statistics → large uncertainties!

PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

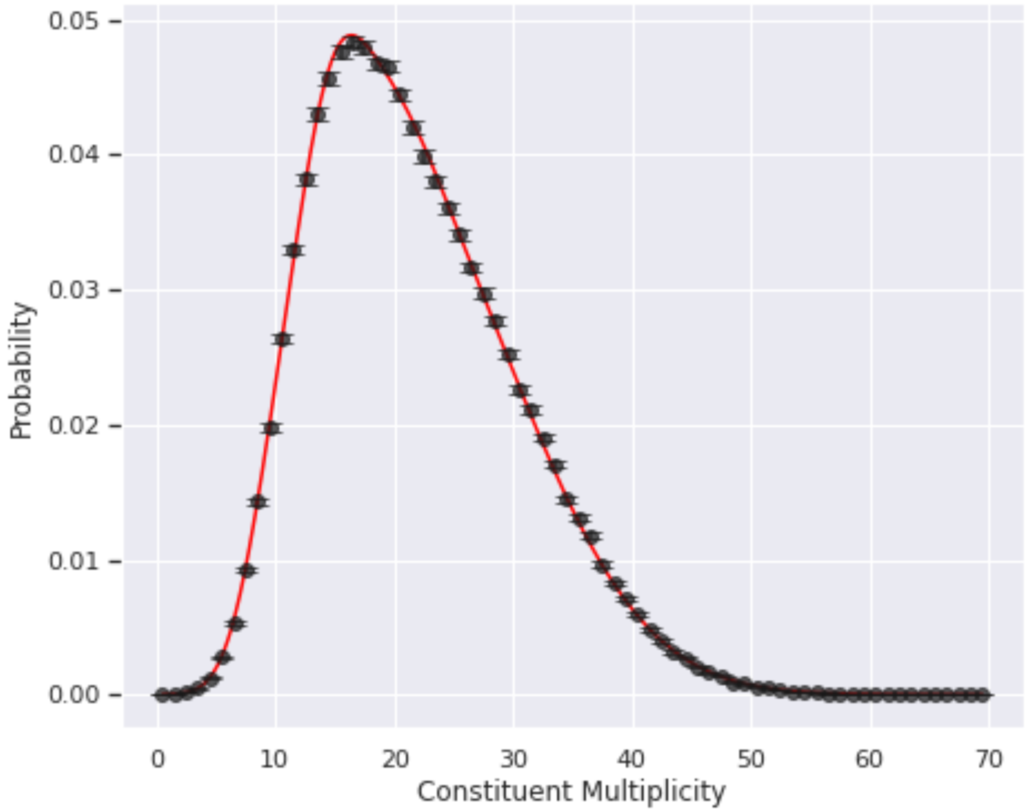
How do we get κ?



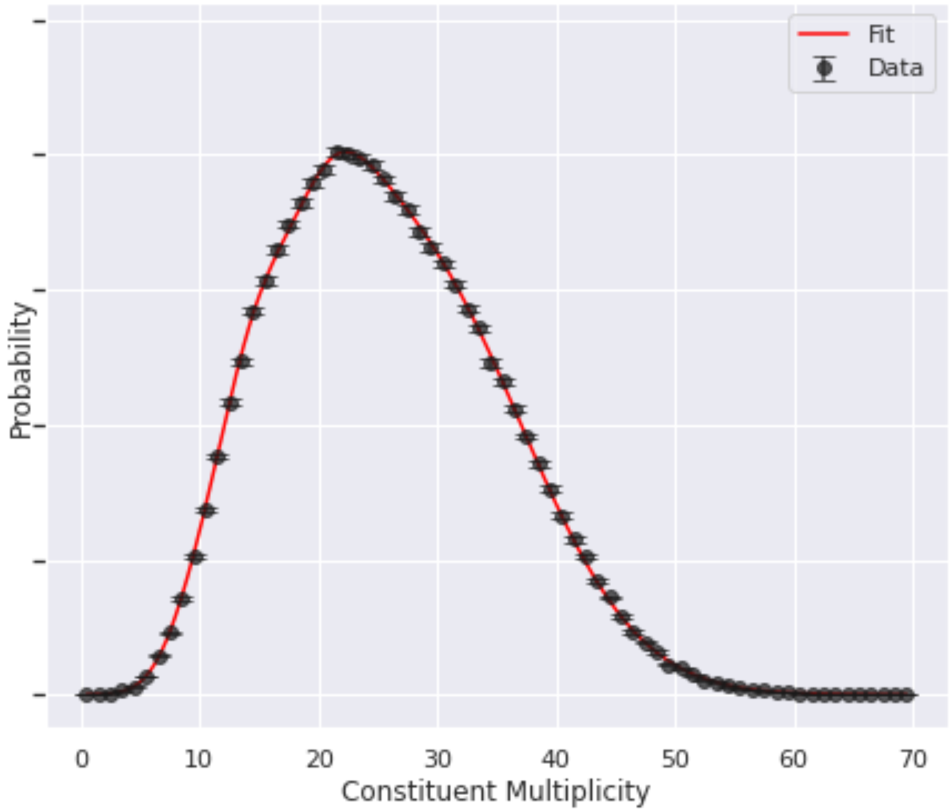
Fit a model to the data!
Model: sum of 4 skew-normal distributions



γ+jet fit

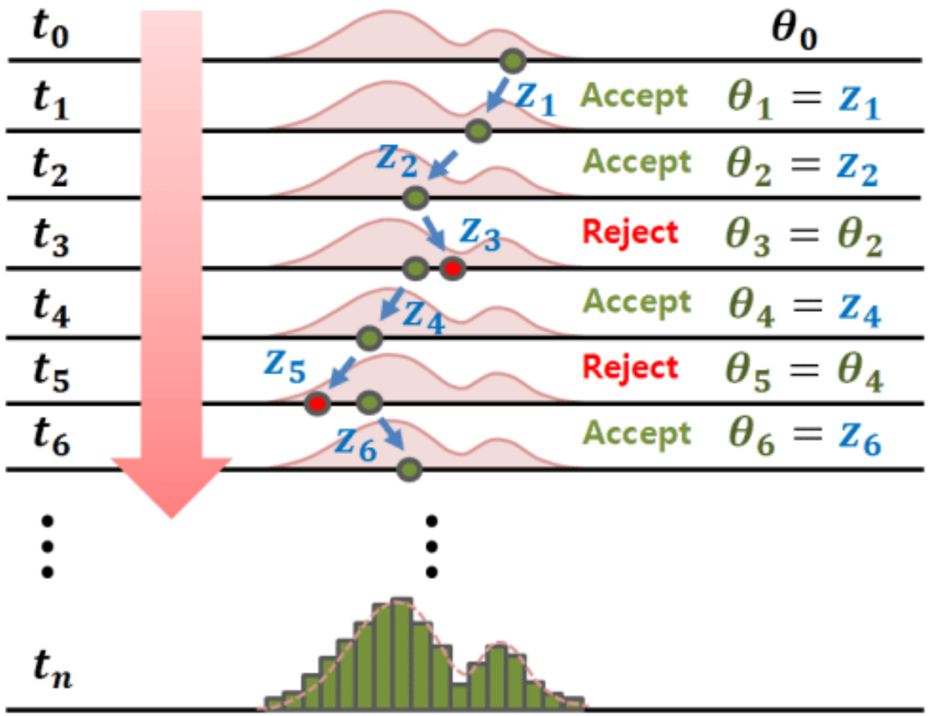


Dijet fit



Markov Chain Monte Carlo

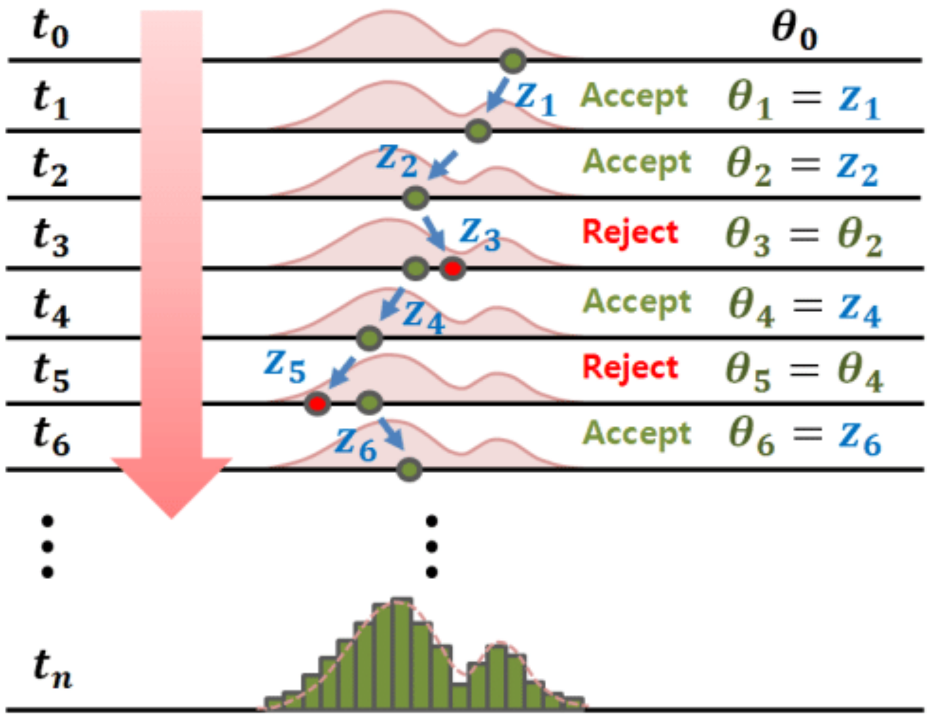
- Each of parameters in our model has a **probability distribution** of its value
  - MCMC attempts to find this through sampling!



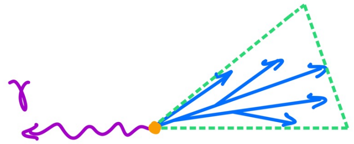
[Jin, *et al.*, 10.1080]

Markov Chain Monte Carlo

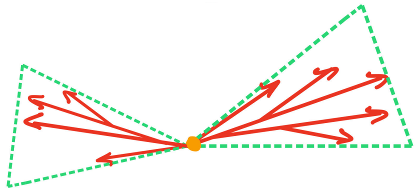
- Initialize θ_0
- For t = 0, 1, 2, …, n:
  - Draw a tentative sample z_t from $Q(\theta|\theta_t)$
  - Accept new z_t with probability A:
    $$A = \min(1, \frac{P(z_t|D)Q(\theta_t|z_t)}{P(\theta_t|D)Q(z_t|\theta_t)})$$
  - If z_t is accepted, then set $\theta_{t+1} \leftarrow z_t$
  - Else set $\theta_{t+1} \leftarrow \theta_t$



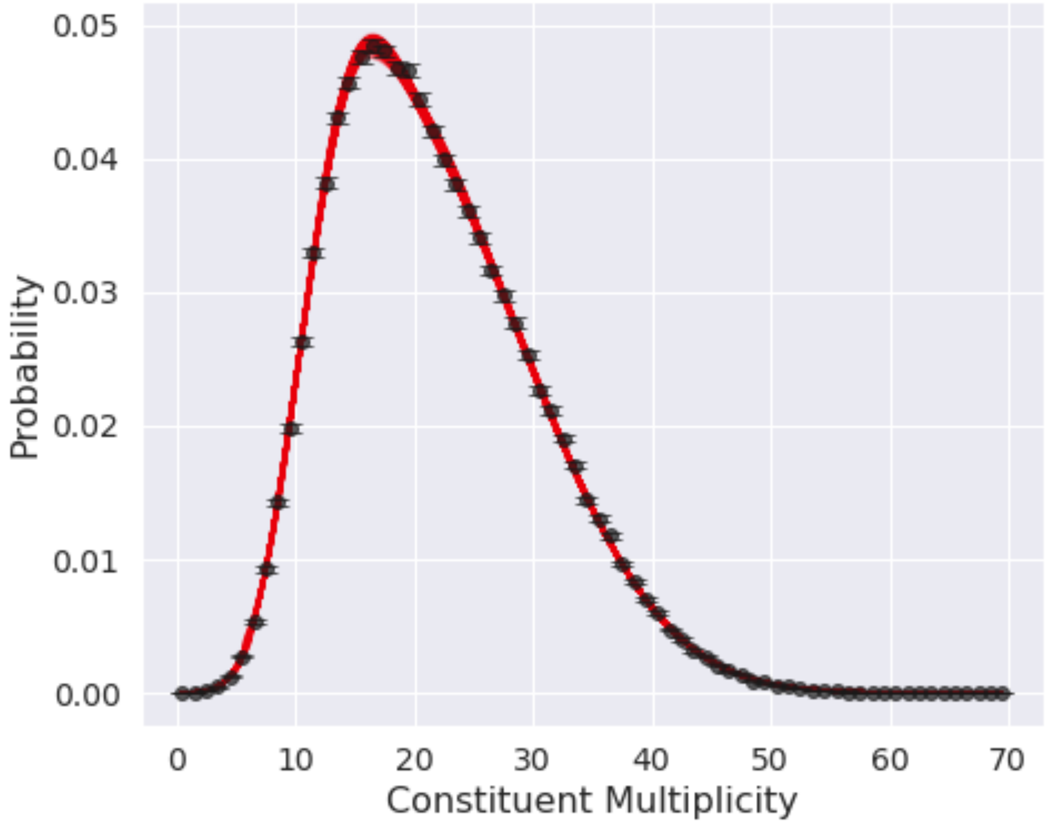
How do we get κ?



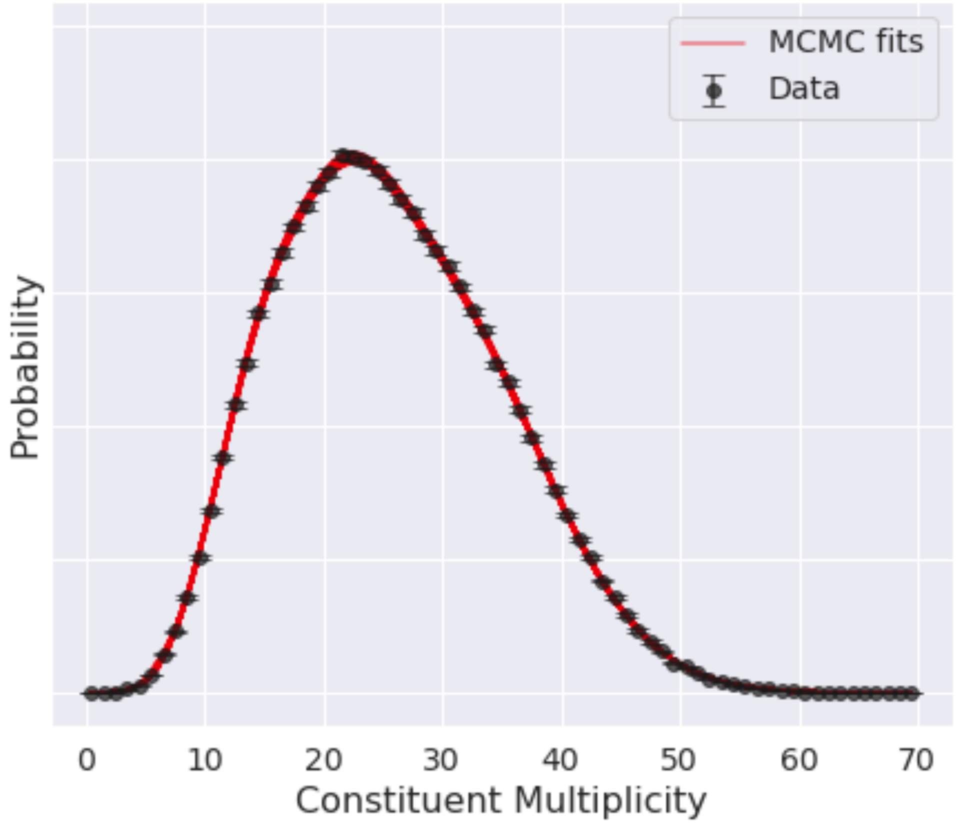
Use Markov Chain Monte Carlo (MCMC) to
sample parameters that fit the data



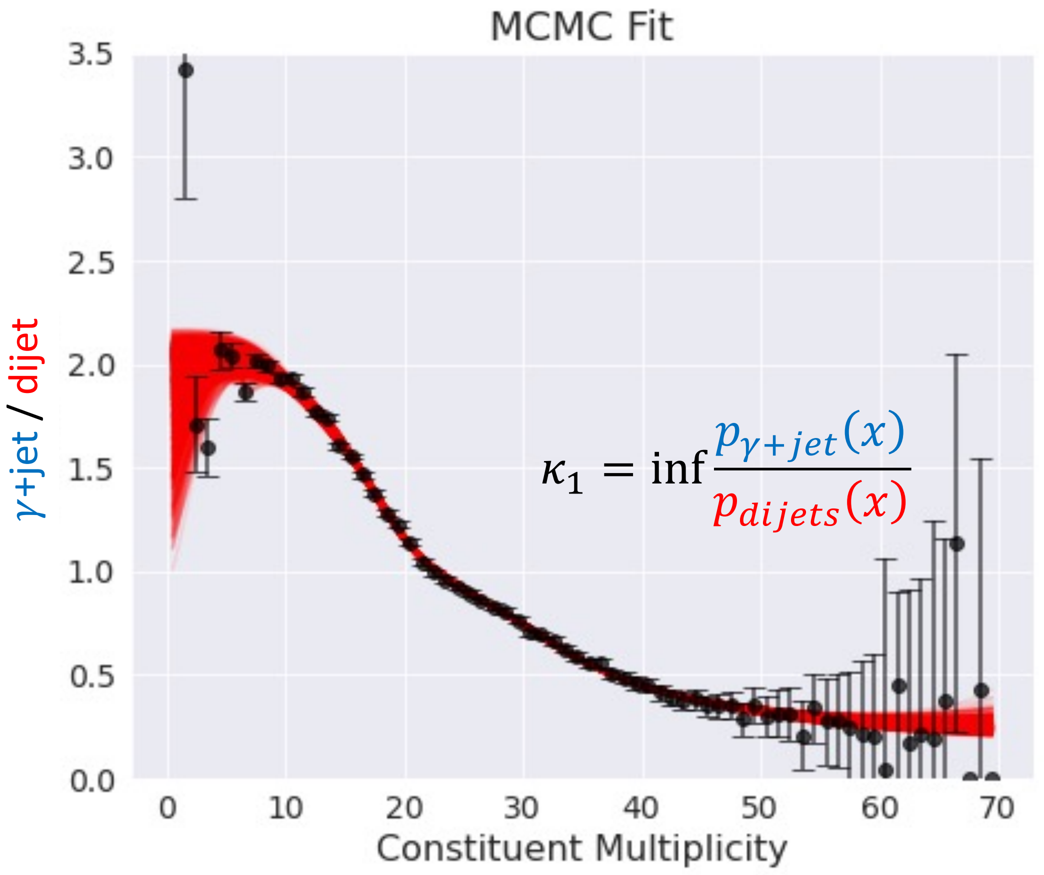
γ+jet MCMC fit

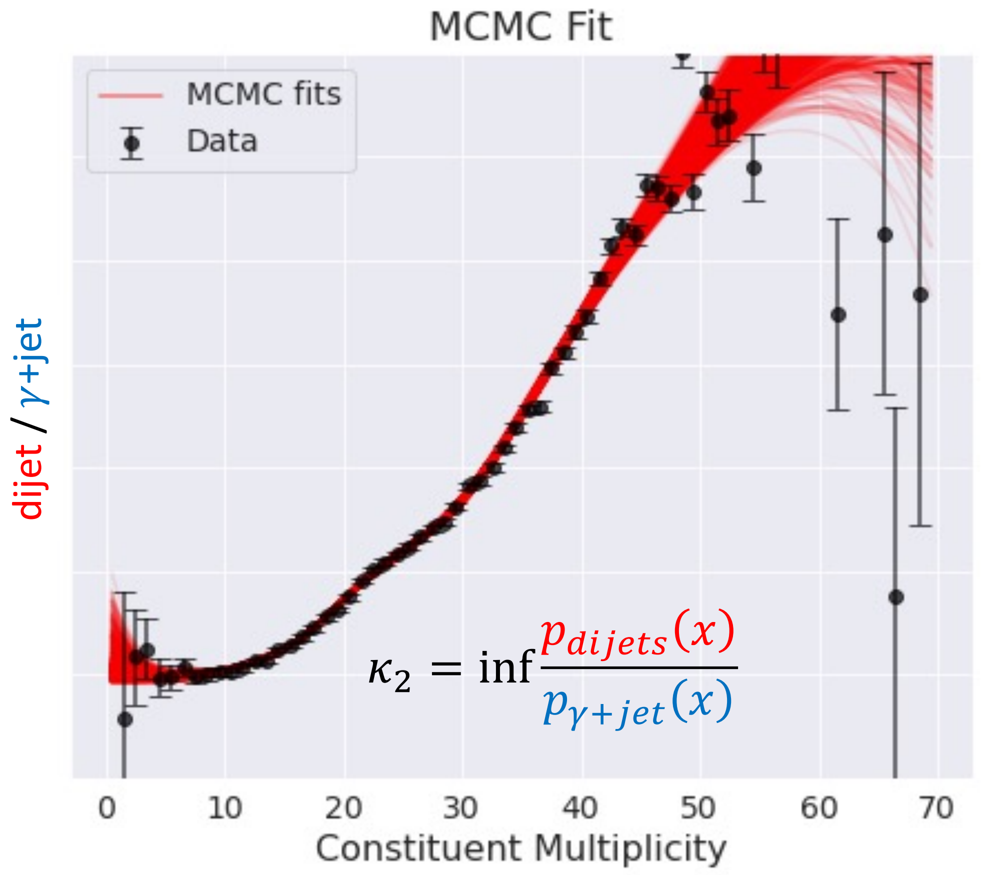


Dijet MCMC fit

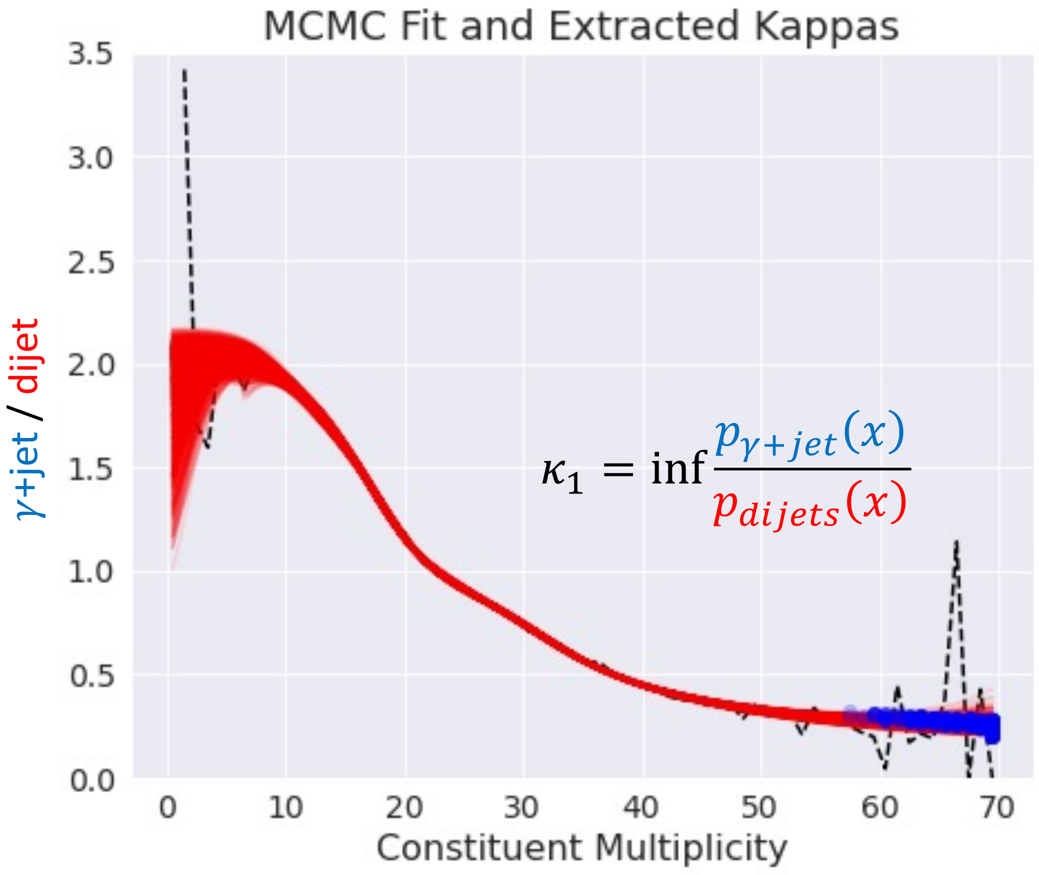


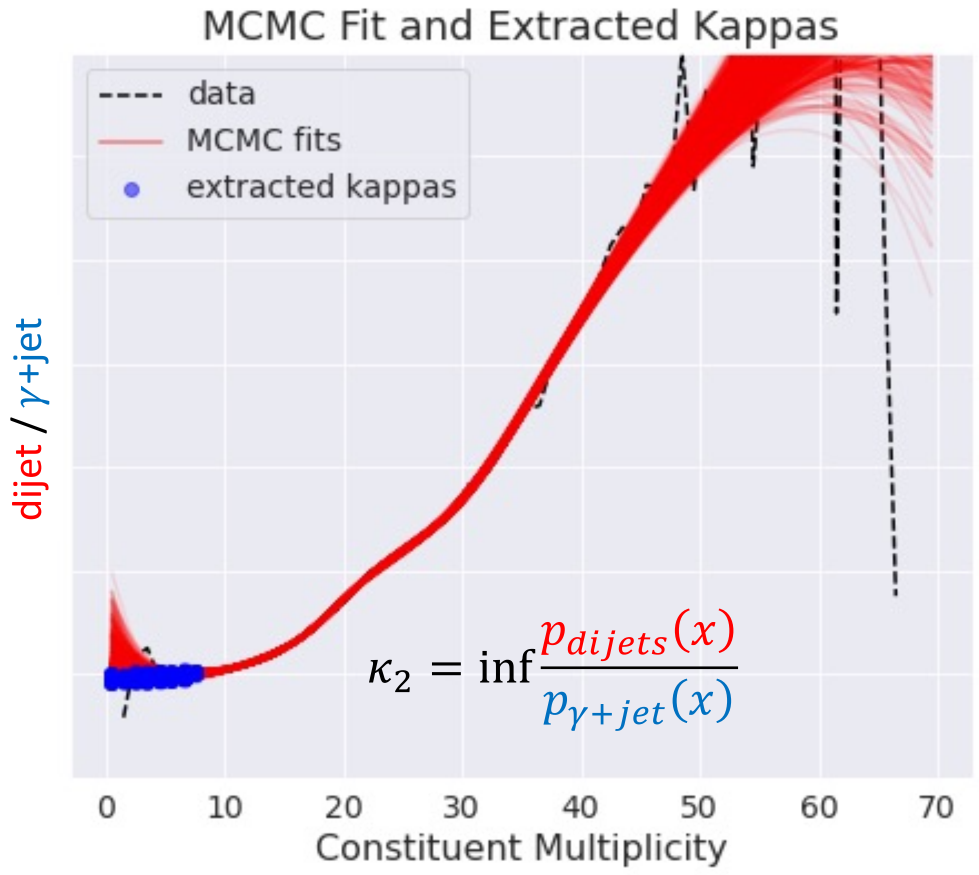
Extracting κ from the MCMC fits



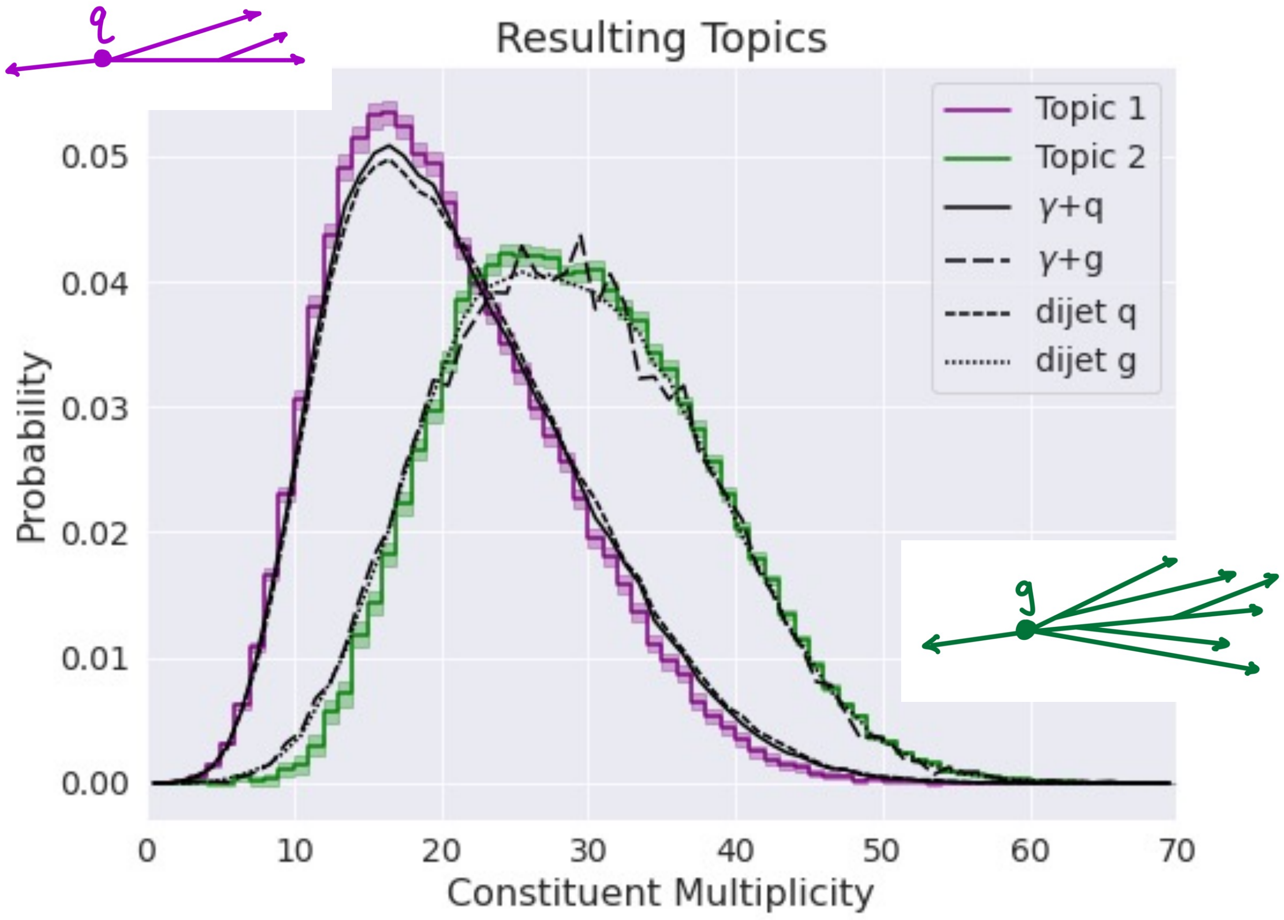


Extracting κ from the MCMC fits





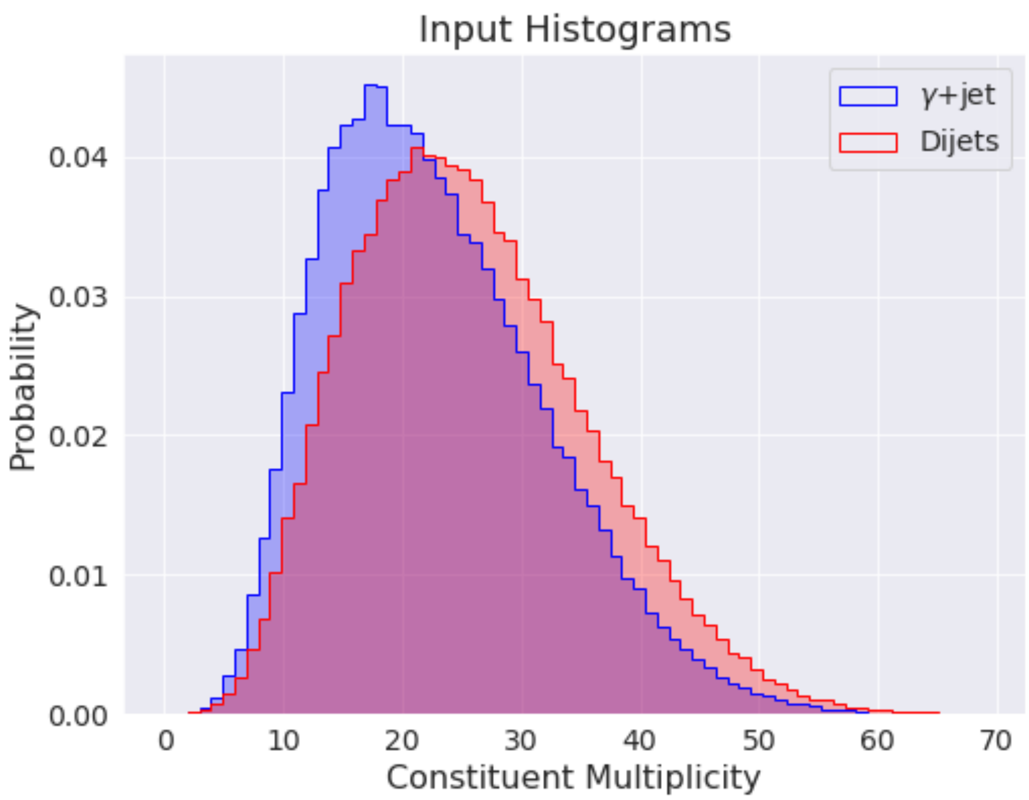
We can use κ to calculate base distributions

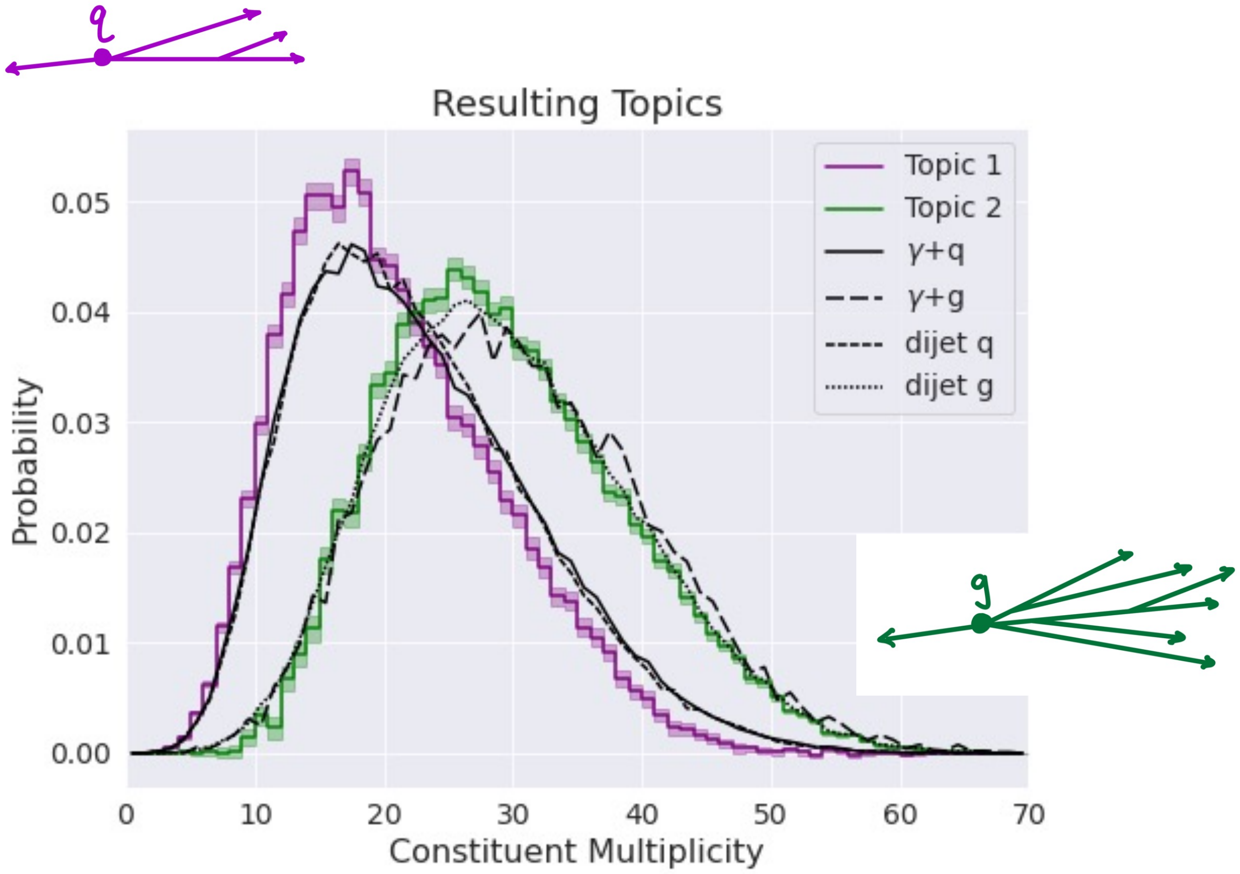


$$b_1(x) = \frac{p_{\gamma+jet}(x) - \kappa_1 p_{dijets}(x)}{1-\kappa_1}$$

$$b_2(x) = \frac{p_{dijets}(x) - \kappa_2 p_{\gamma+jet}(x)}{1-\kappa_2}$$

PbPb results

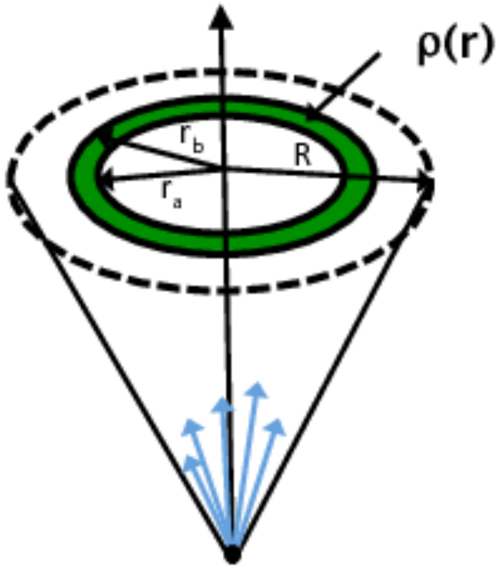




PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Defining jet shape

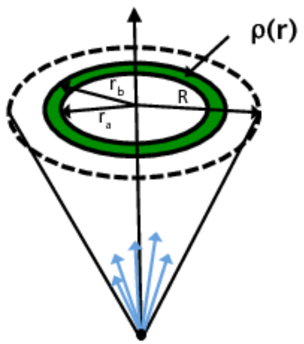
- **Jet shape** describes how the jet transverse momentum is distributed as a function of radial distance from the jet axis

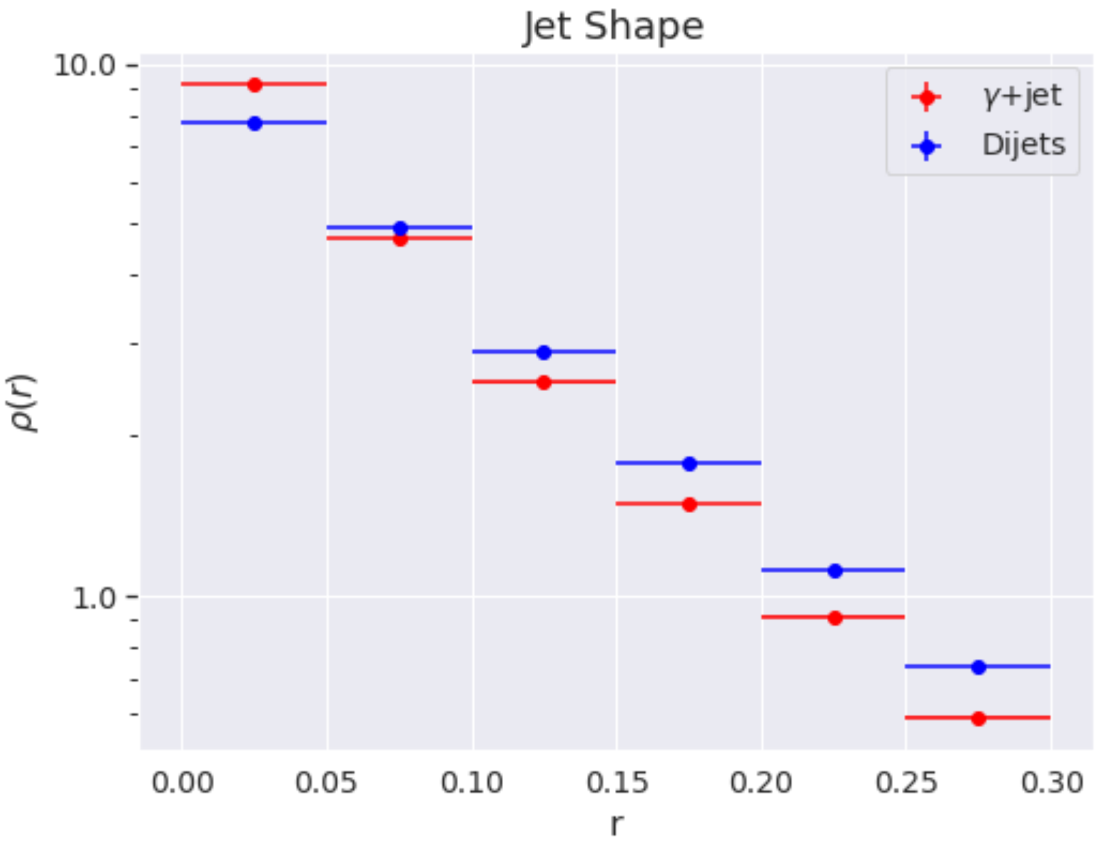


$$\rho(r) = \frac{1}{\delta r}\frac{1}{N_{jet}}\sum_{jets}\frac{\sum_{tracks\in[r_a,r_b)} p_T^{track}}{p_T^{jet}}$$

[CMS, 1809.08602]

Jet shape for pp γ+jet and dijets





PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

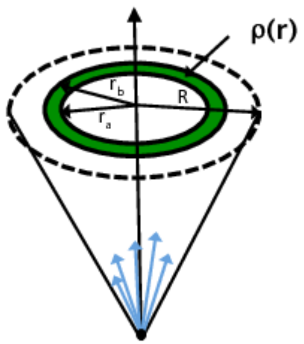
Topic modeling results to extract jet shape

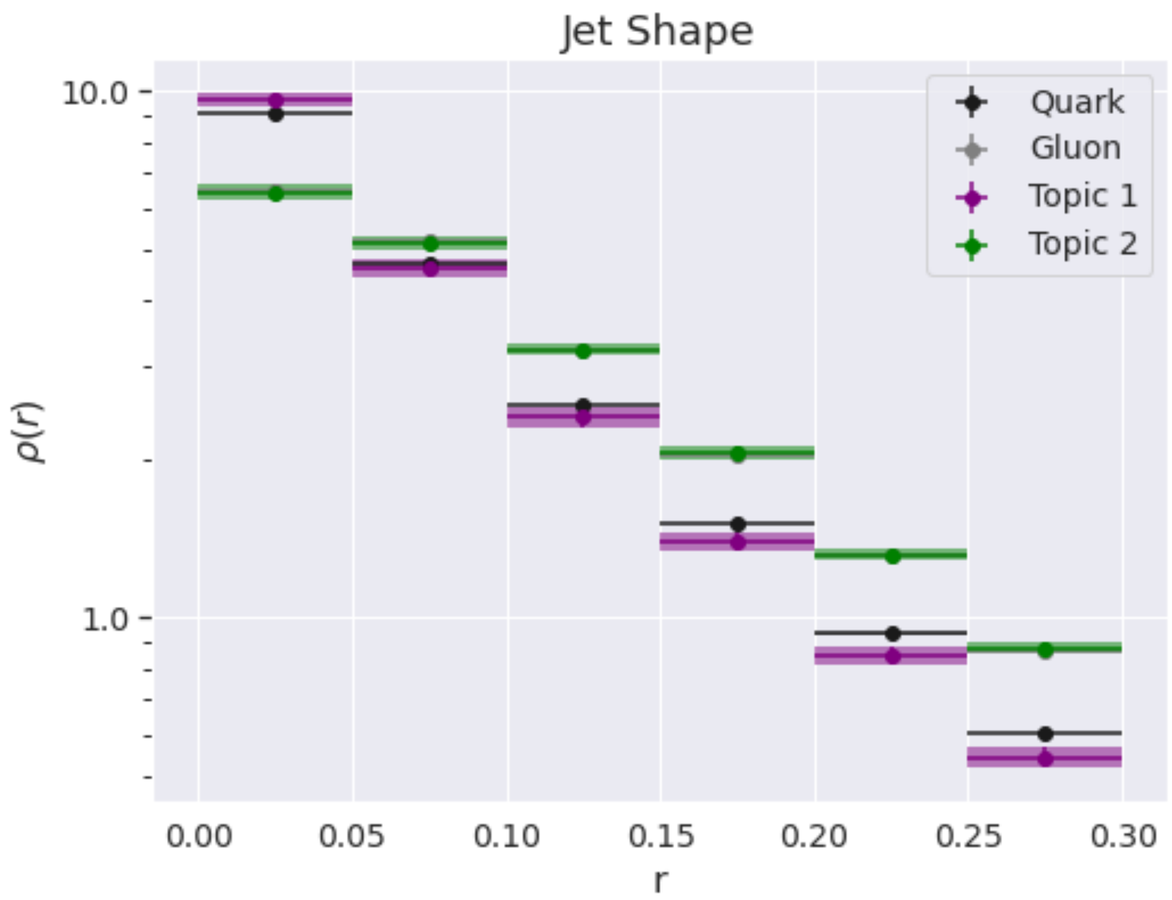
- We can apply a linear combination to extract each bin value in the jet shape:

$$\rho_1(r) = \frac{\rho_{\gamma+jet}(r) - \kappa_1 \rho_{dijets}(r)}{1-\kappa_1}$$

$$\rho_2(r) = \frac{\rho_{dijets}(r) - \kappa_2 \rho_{\gamma+jet}(r)}{1-\kappa_2}$$

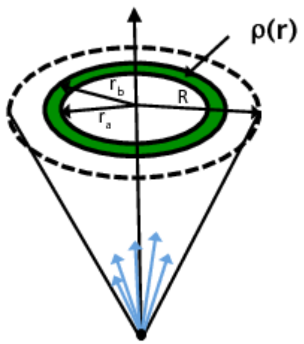
Topic modeling extraction - pp

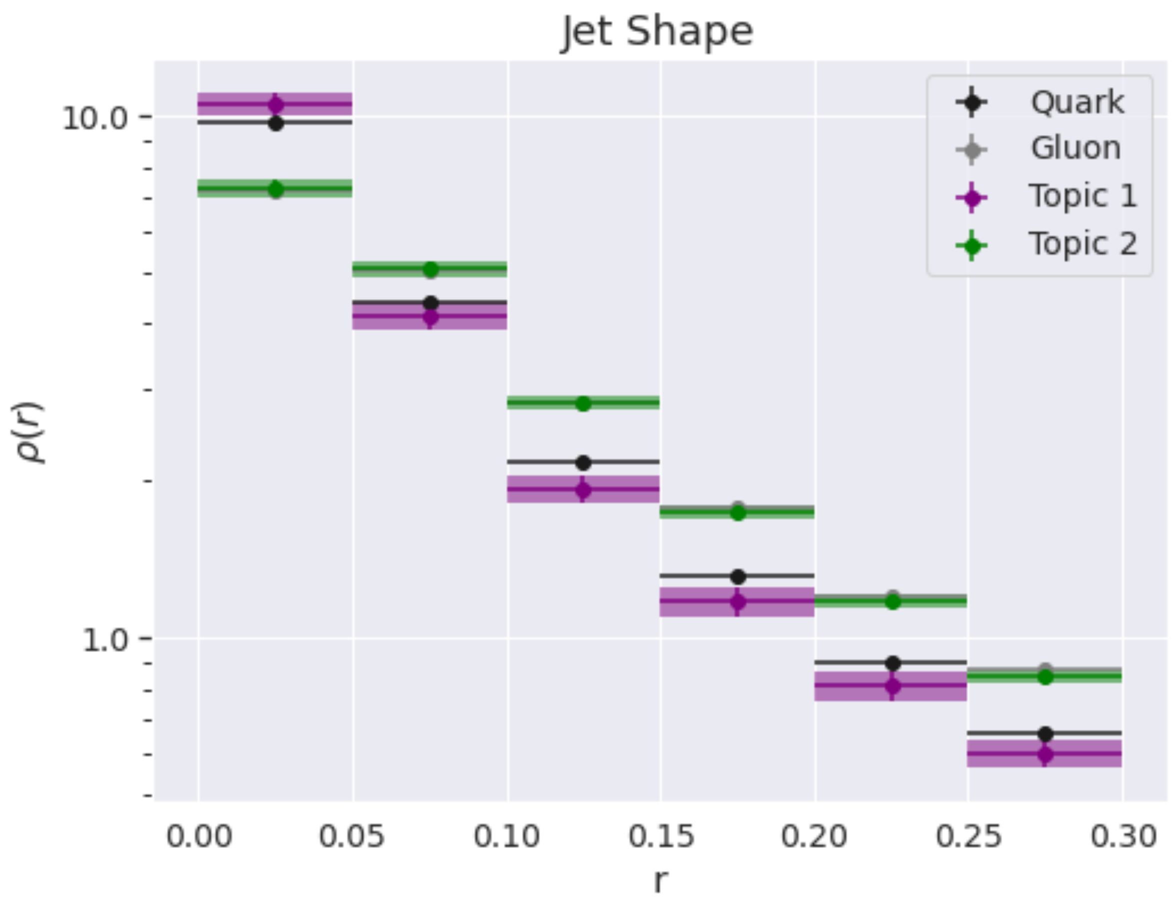




PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Topic modeling extraction - PbPb

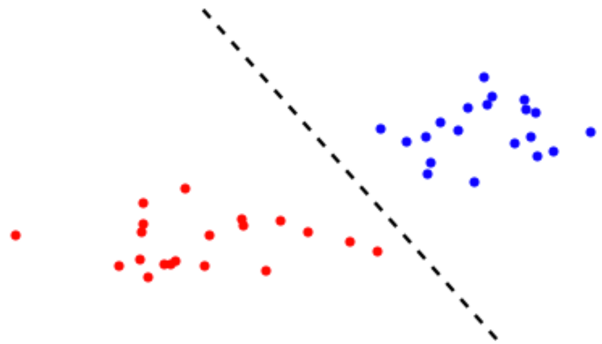


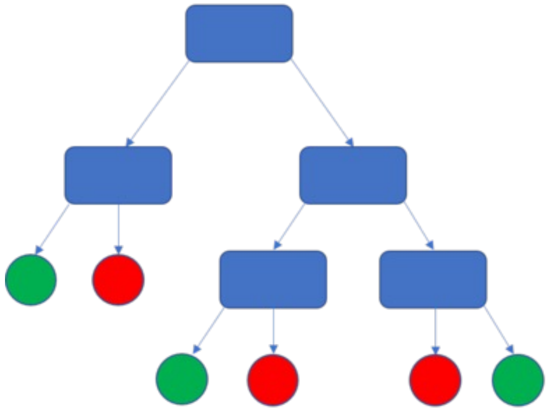


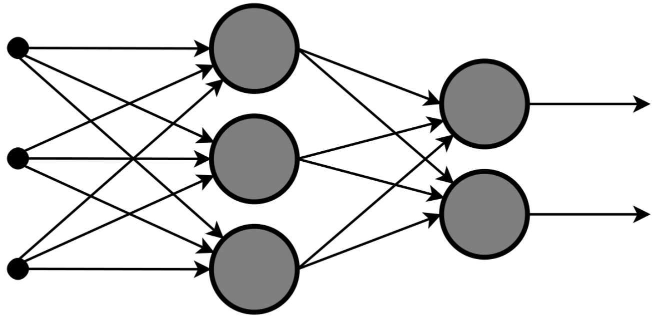
PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Can we find a better observable?

- Use machine learning to enhance separability between quark and gluon jet distributions

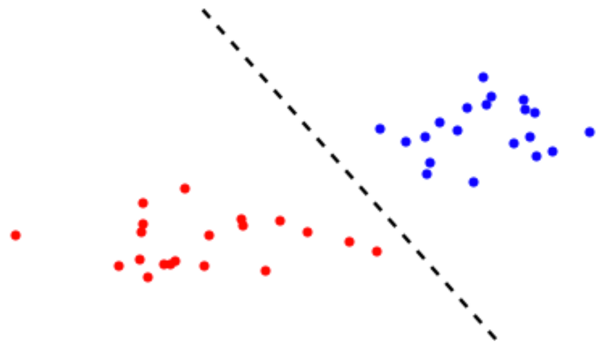






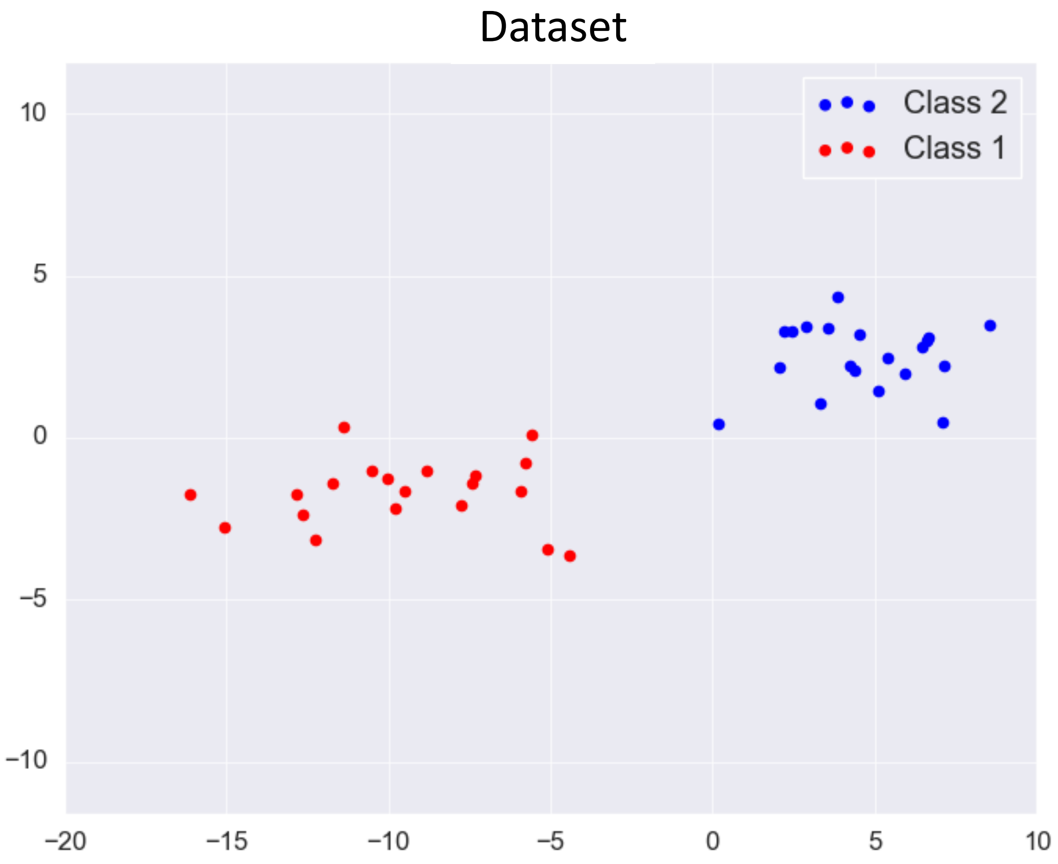
Can we find a better observable?

- Use machine learning to enhance separability between quark and gluon jet distributions

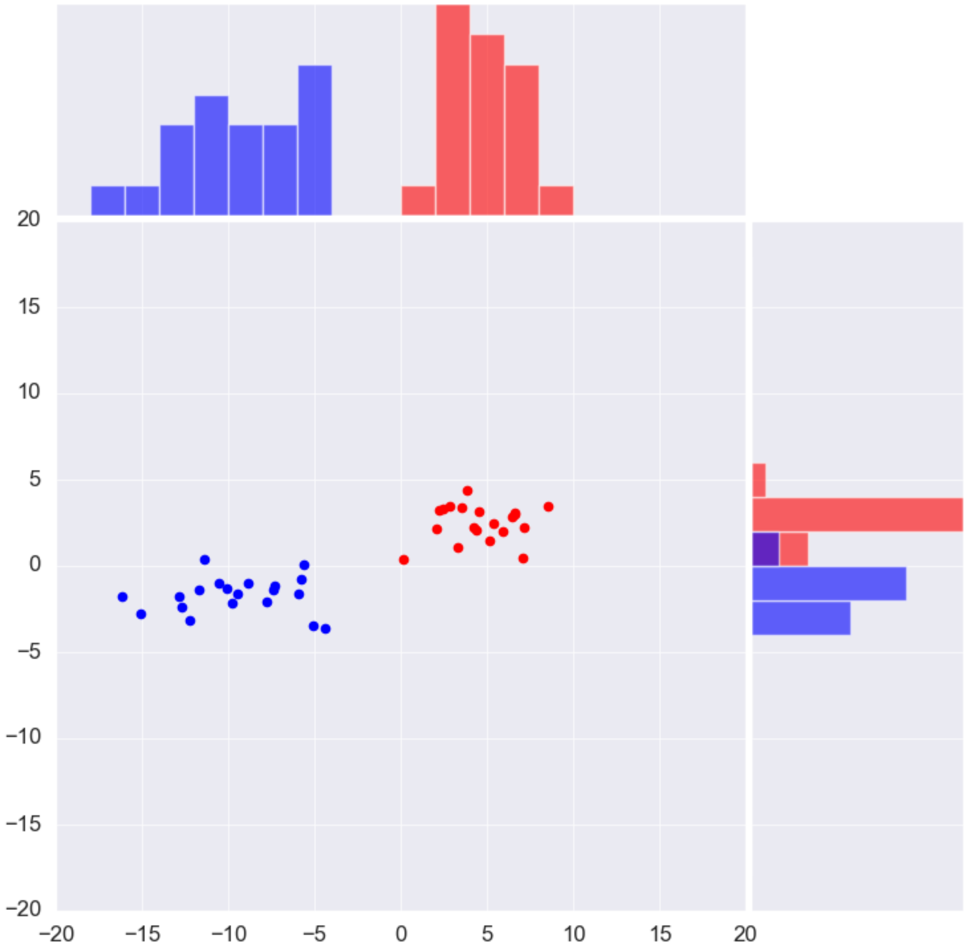


Linear Discriminant Analysis (LDA)
reduces dimensionality while
maximizing class separability

Linear Discriminant Analysis

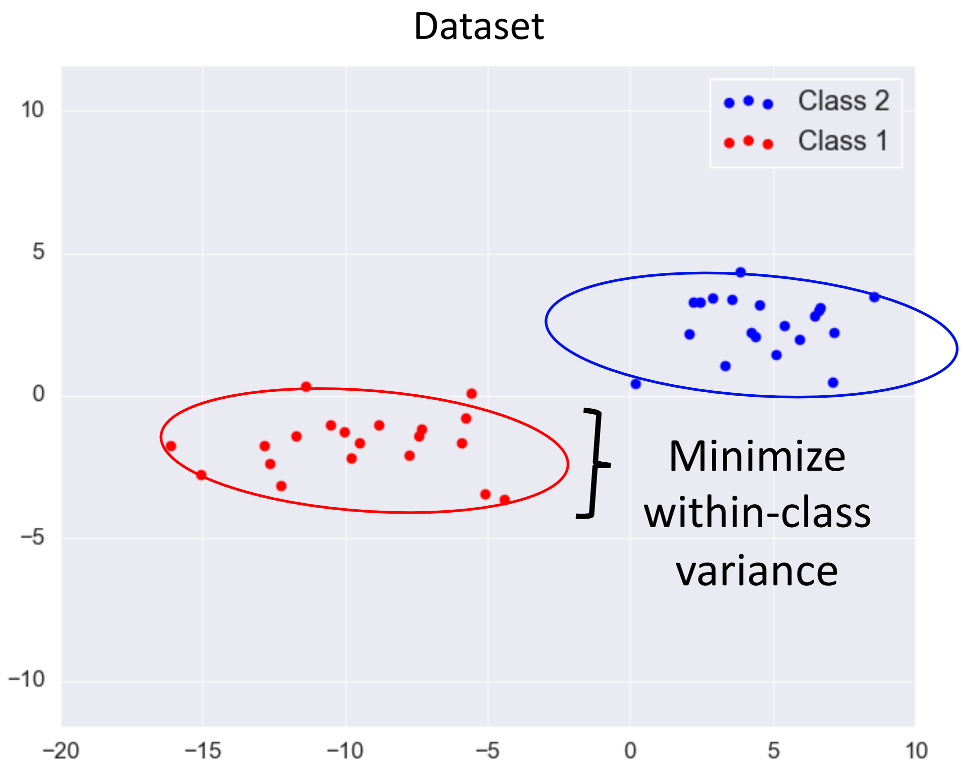


Histograms of X/Y projections of the data



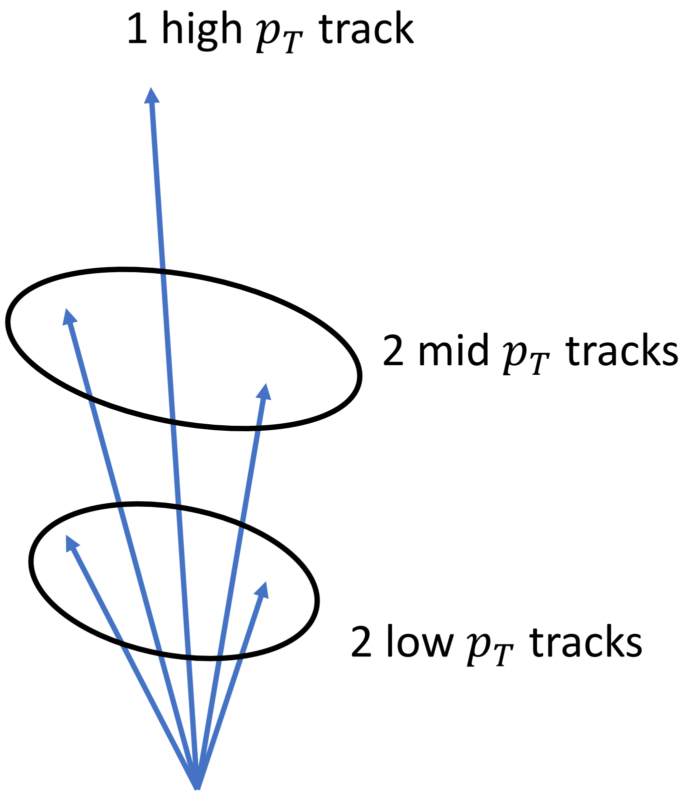
LDA


Dataset
Class 2
Class 1
10
5
0
−5
−10
−20
−15
−10
−5
0
5
10
Maximize
between-class
variance

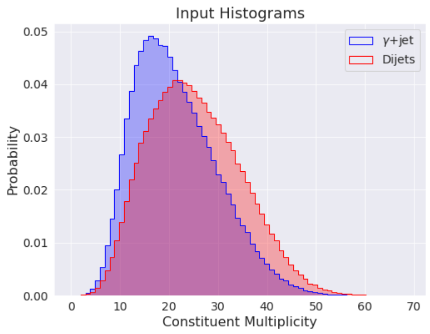

Dataset
Class 2
Class 1
10
5
0
−5
−10
−20
−15
−10
−5
0
5
10
Minimize
within-class
variance

LDA Multiplicity

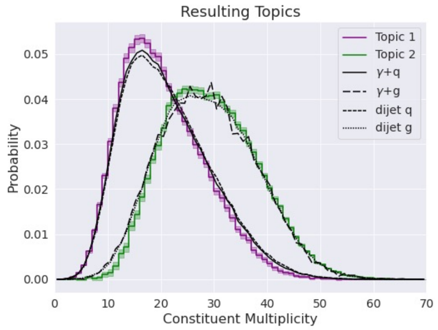
- Partition the multiplicity of a jet into p_T bins
  - Ex: [2, 2, 1] is our 3-dimensional feature vector
- Construct these vectors for quark and gluon jets
- Train LDA and project → **LDA multiplicity**
- This is still a work in progress!



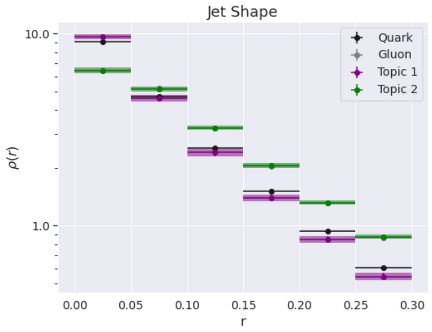
In summary…



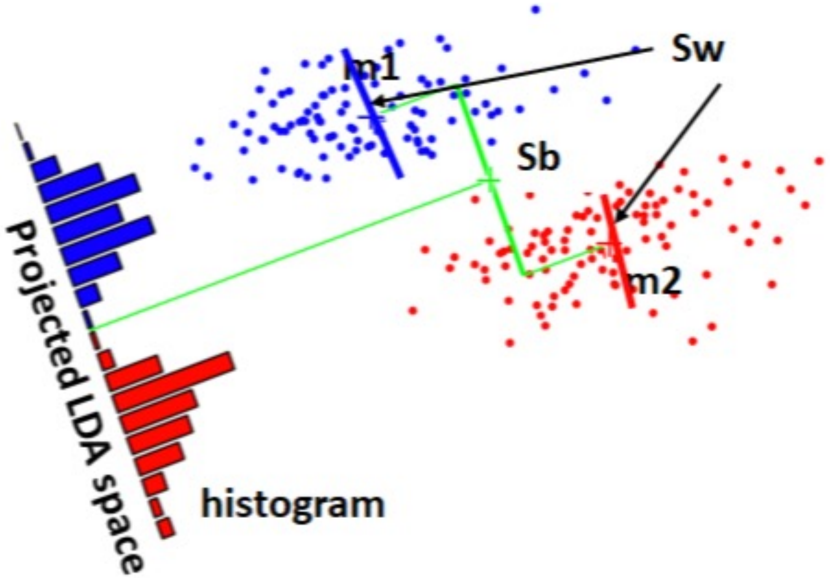
Topic modeling



Jet substructure extraction



Linear Discriminant Analysis



Thank you!

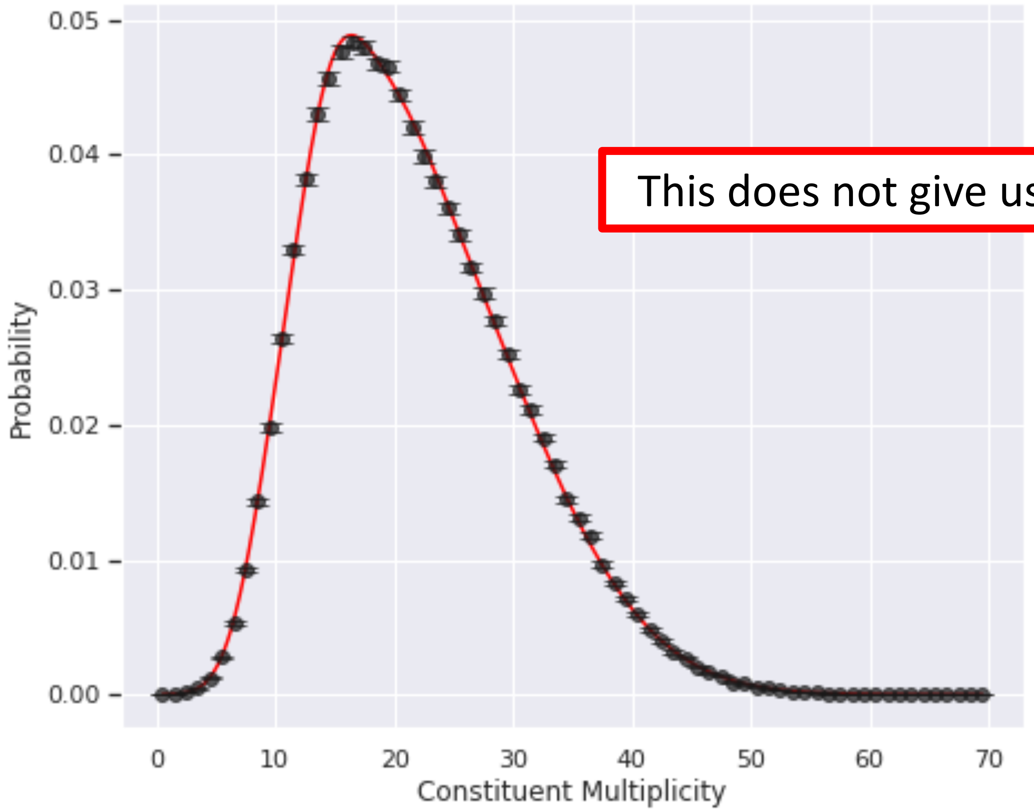
Questions/comments?

kying@mit.edu

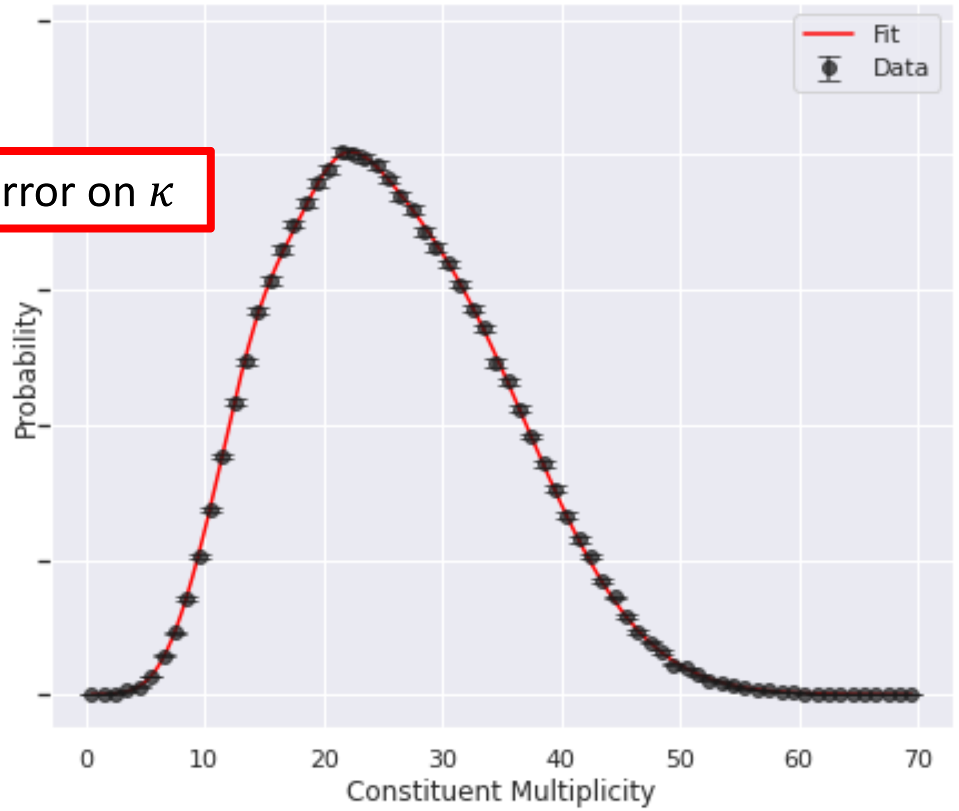
How do we get κ?

Fit a model to the data!
Model: sum of 4 skew-normal distributions

γ+jet fit

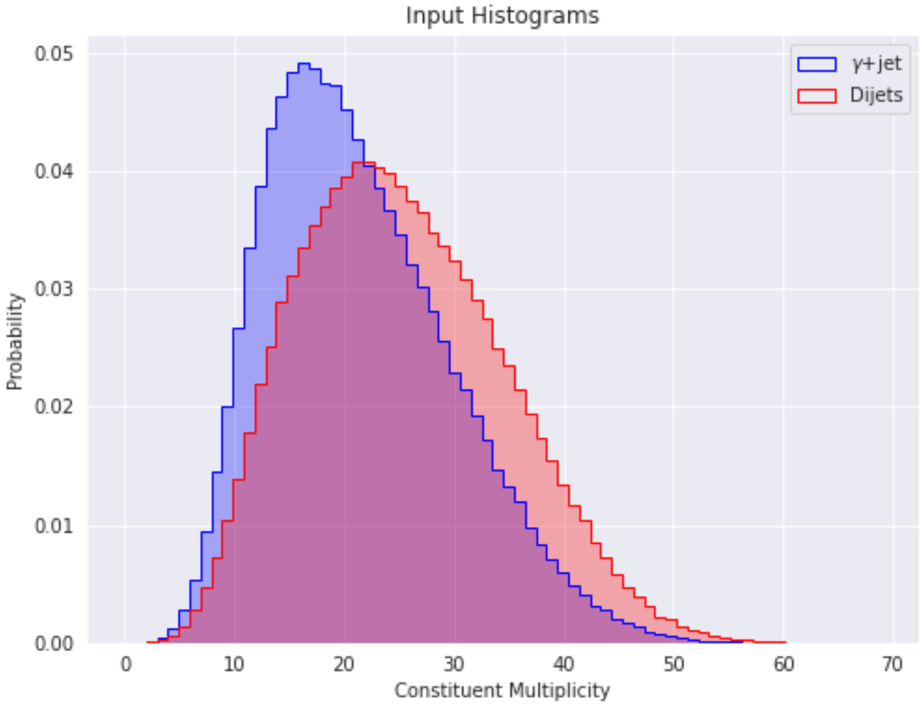


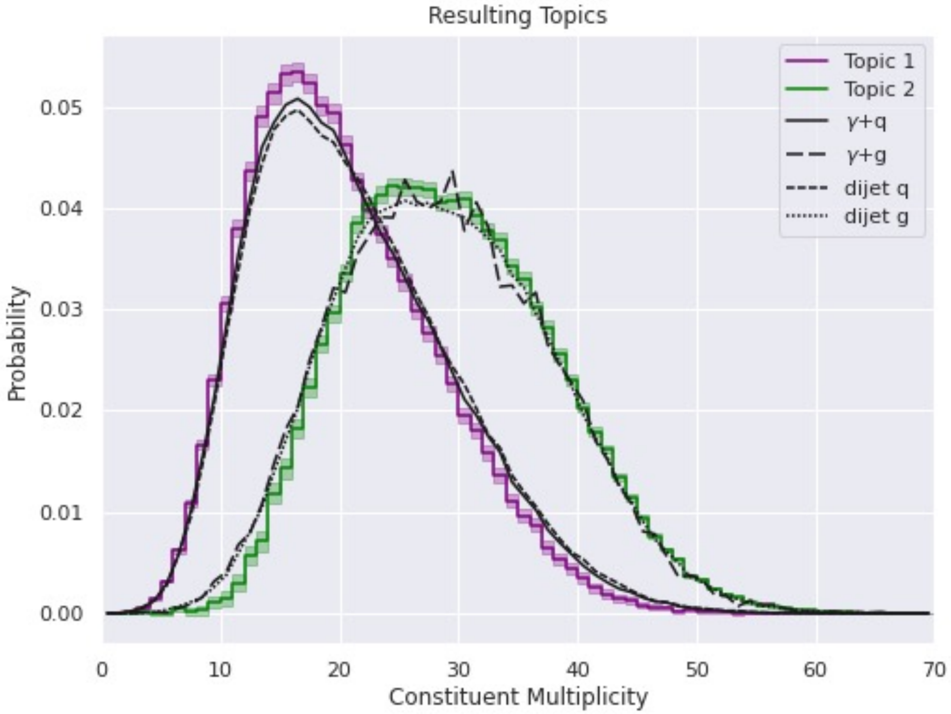
Dijet fit



This does not give us the error on κ

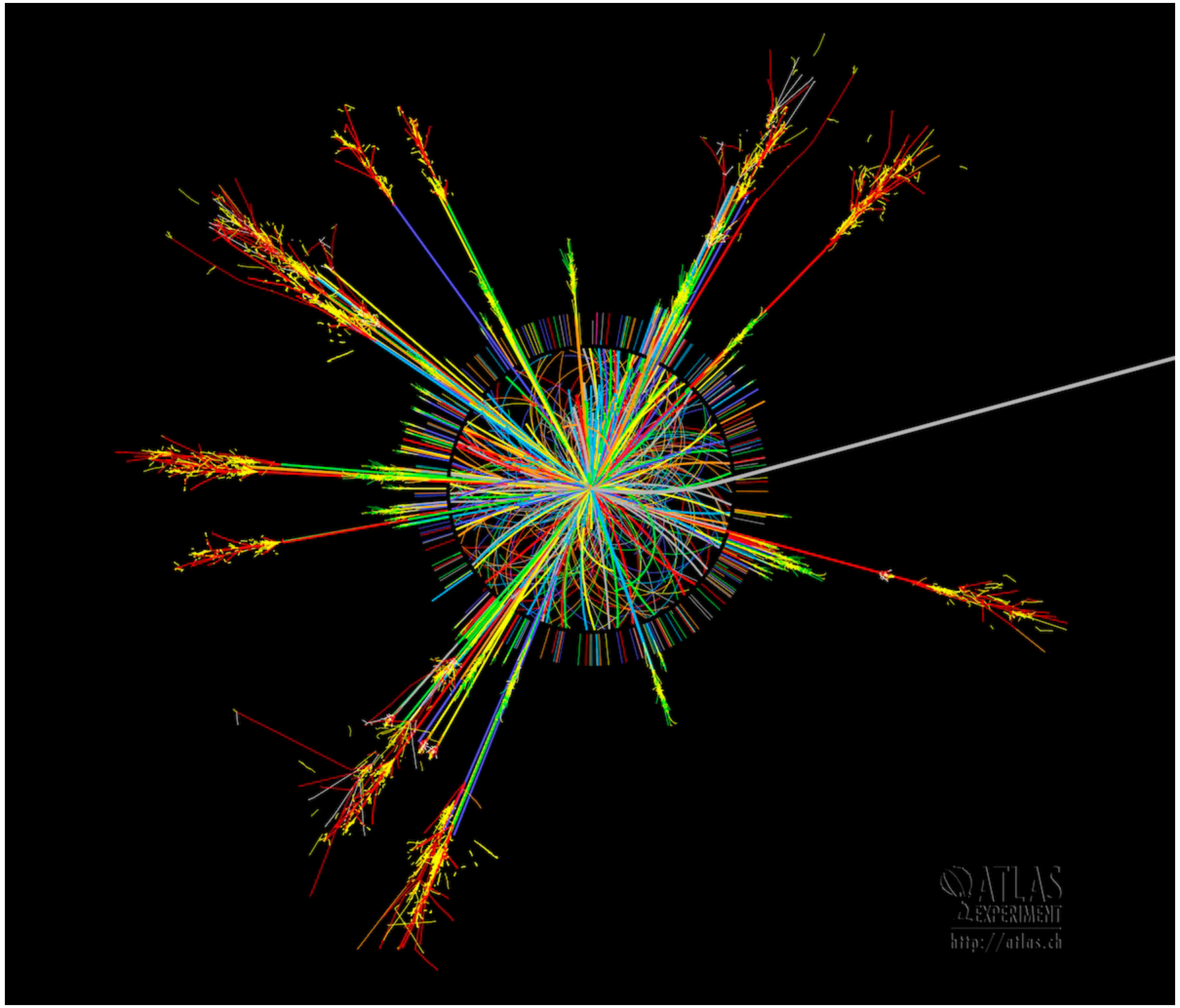
pp results



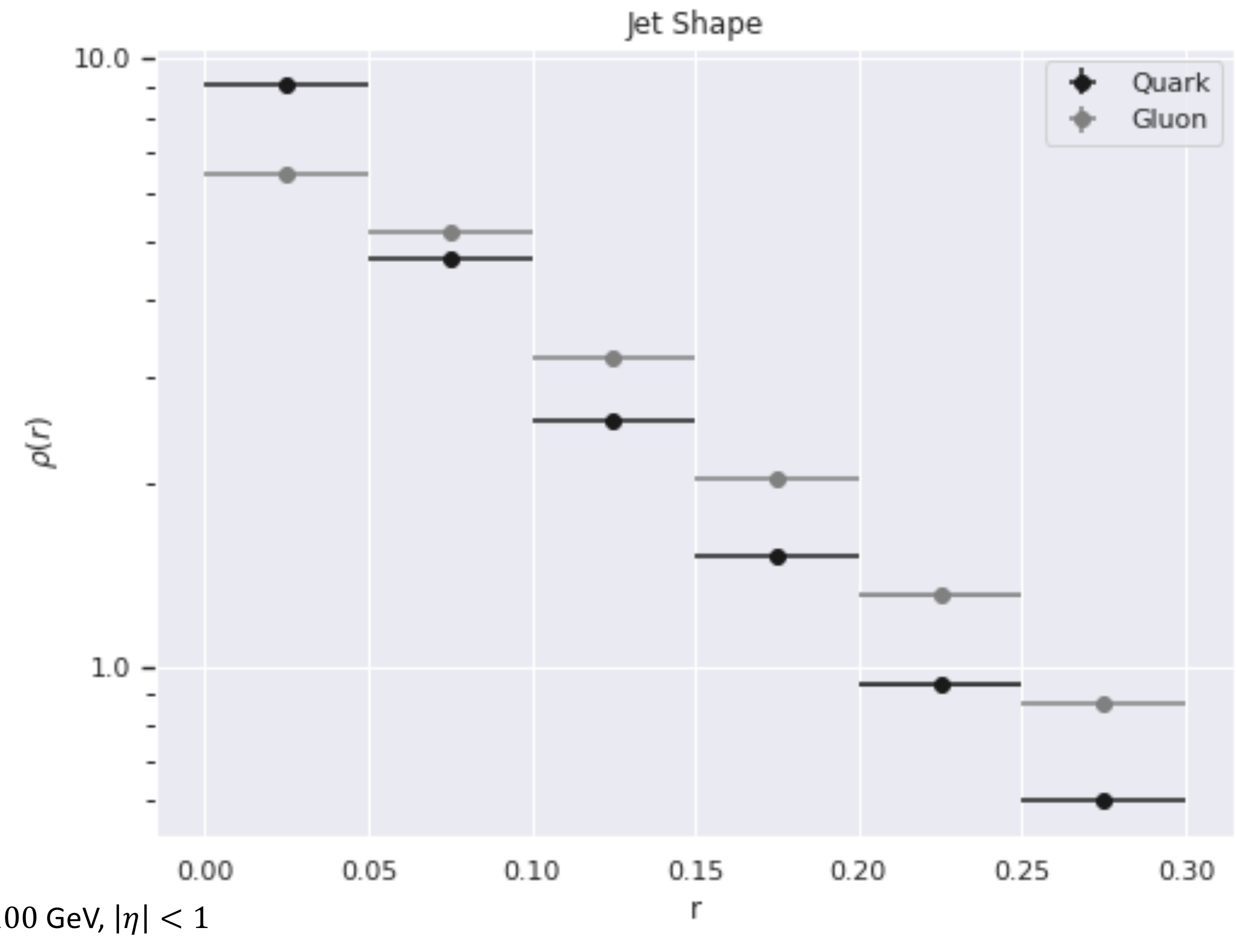


PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Jets in Particle Collisions

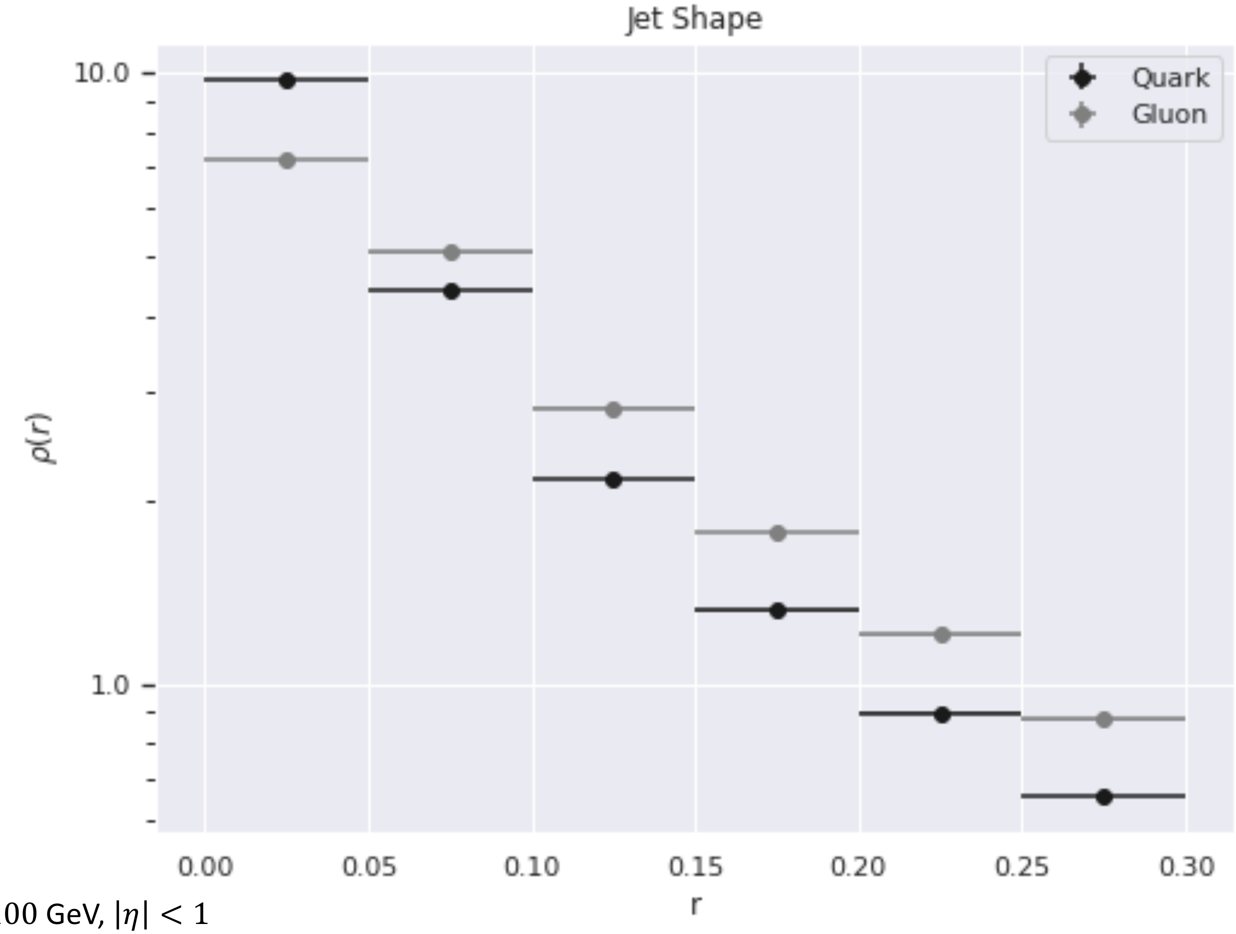


Jet shape for pp quark / gluon truths



PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Jet shape for PbPb quark / gluon truths



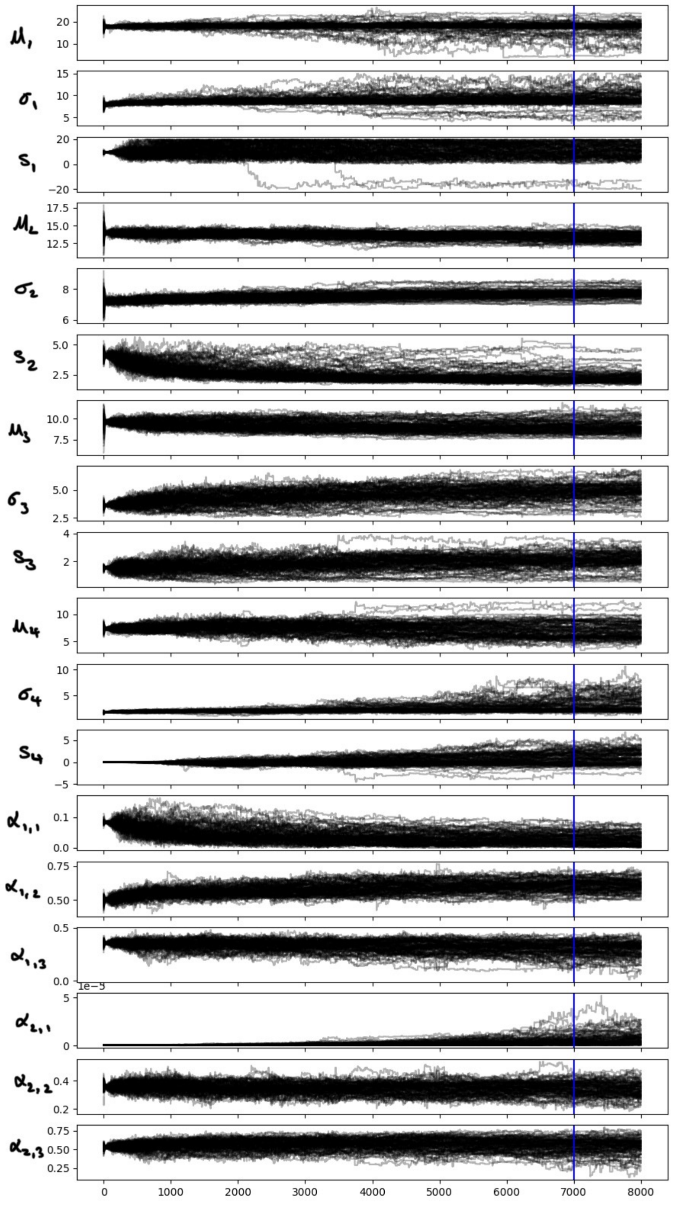
PYQUEN samples, $80 < p_T < 100$ GeV, $|\eta| < 1$

Metropolis-Hastings

- Initialize θ_0
- For t = 0, 1, 2, …, n:
  - Draw a tentative sample z_t from $Q(\theta|\theta_t)$
  - Accept new z_t with probability A:

$$A = \min(1, \frac{P(z_t|D)Q(\theta_t|z_t)}{P(\theta_t|D)Q(z_t|\theta_t)})$$

  - If z_t is accepted, then set $\theta_{t+1} \leftarrow z_t$
  - Else set $\theta_{t+1} \leftarrow \theta_t$



LDA Multiplicity

- Each jet in the PYQUEN simulation is composed of many track particles, each with some p_T^{track}
- We can construct a vector of multiplicities from these tracks, representing the multiplicity in a specified p_T^{track} range
- Here we used bins:
  - $0 < p_T^{track} \leq 1$
  - $1 < p_T^{track} \leq 4$
  - $4 < p_T^{track} \leq 10$
  - $p_T^{track} > 10$

Finite-sampled distributions

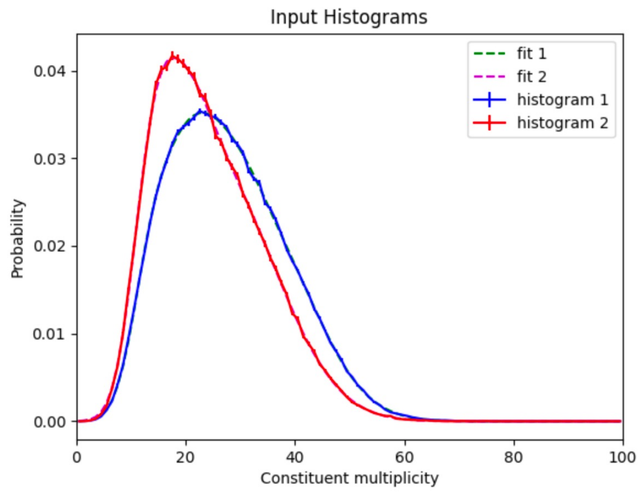
- If we directly try to calculate κ_{ij}, then we may run into large uncertainties
- Model input distribution as skew-normal with parameters θ
- Use least-squares and MCMC to estimate θ then compute κ_{ij}

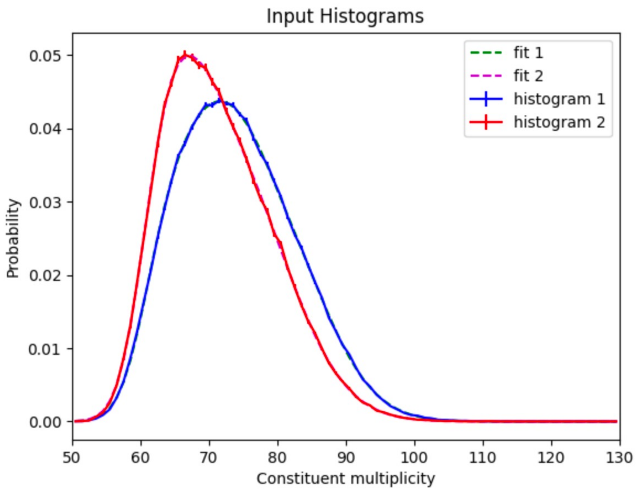
[Brewer, *et al.*, 2008.08596]

PYQUEN samples
100 < pT < 120

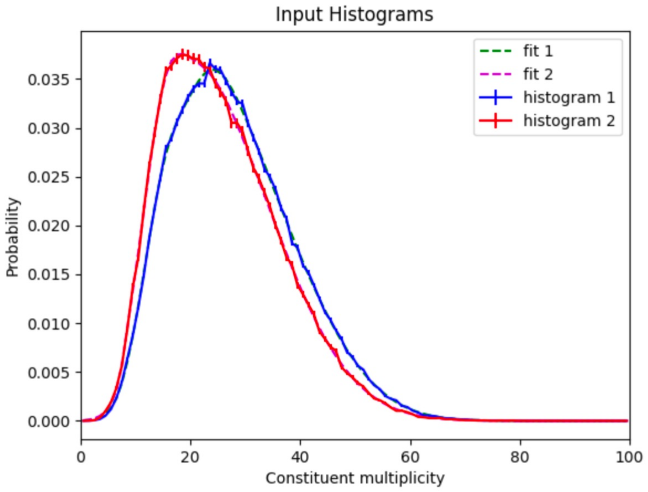
LDA Multiplicity

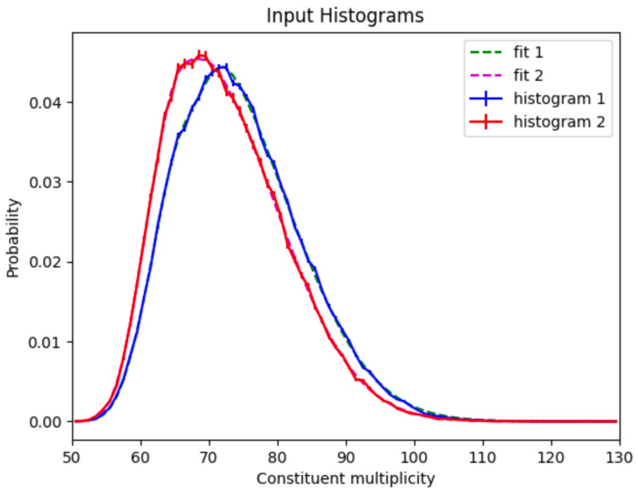
pp





PbPb

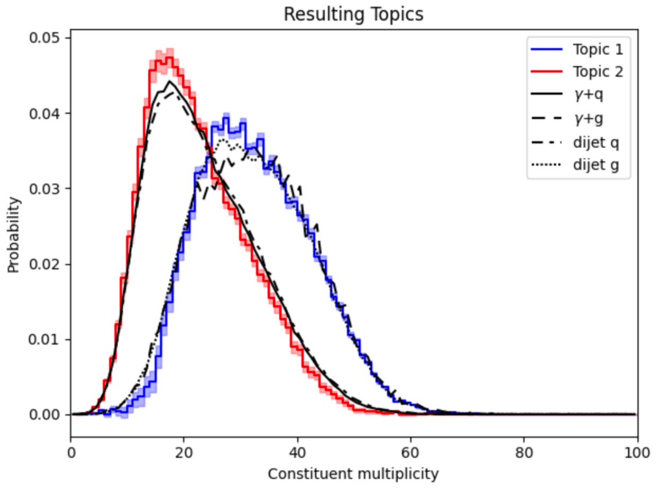


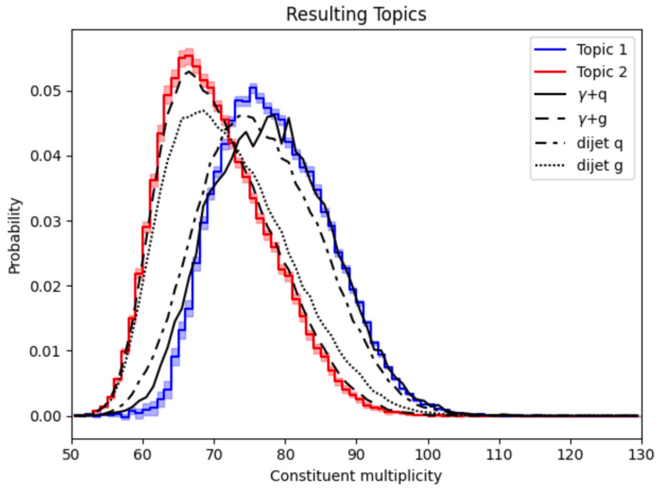


LDA Multiplicity Results

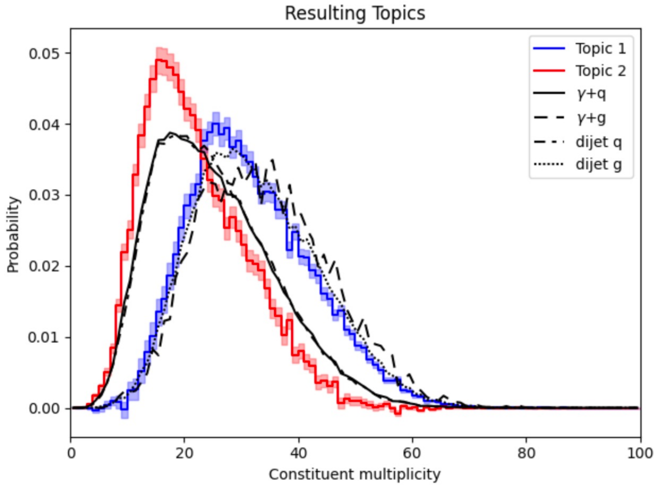
PYQUEN samples
100 < pT < 120

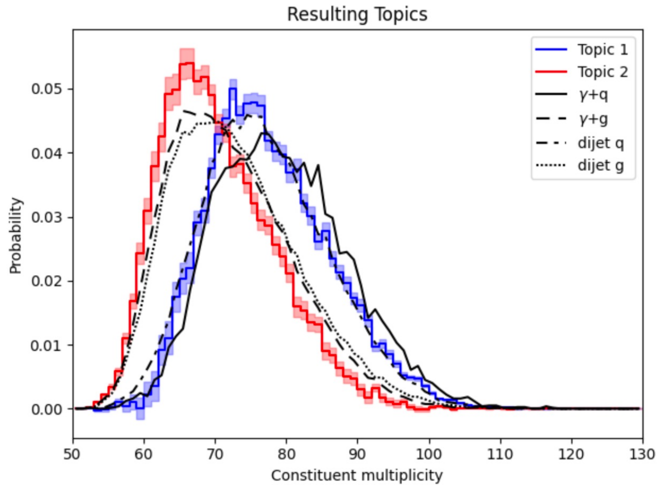
pp





PbPb

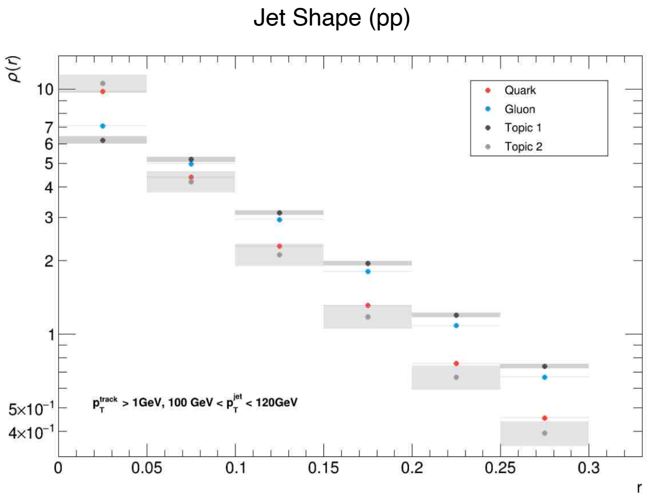


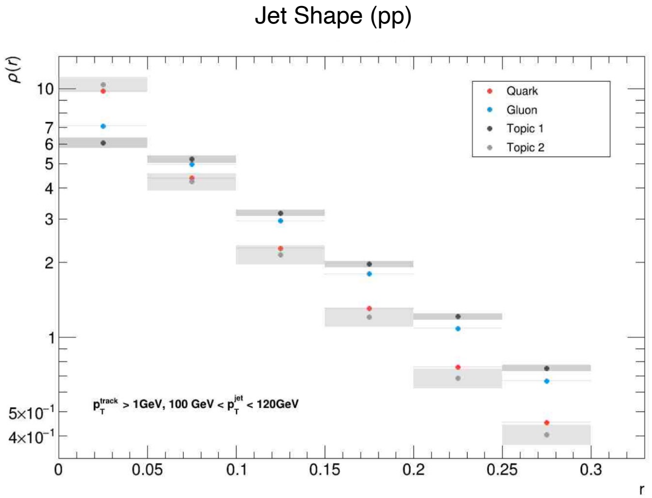


PYQUEN samples
100 < pT < 120

LDA Multiplicity Results on Jet Shape

pp





PbPb