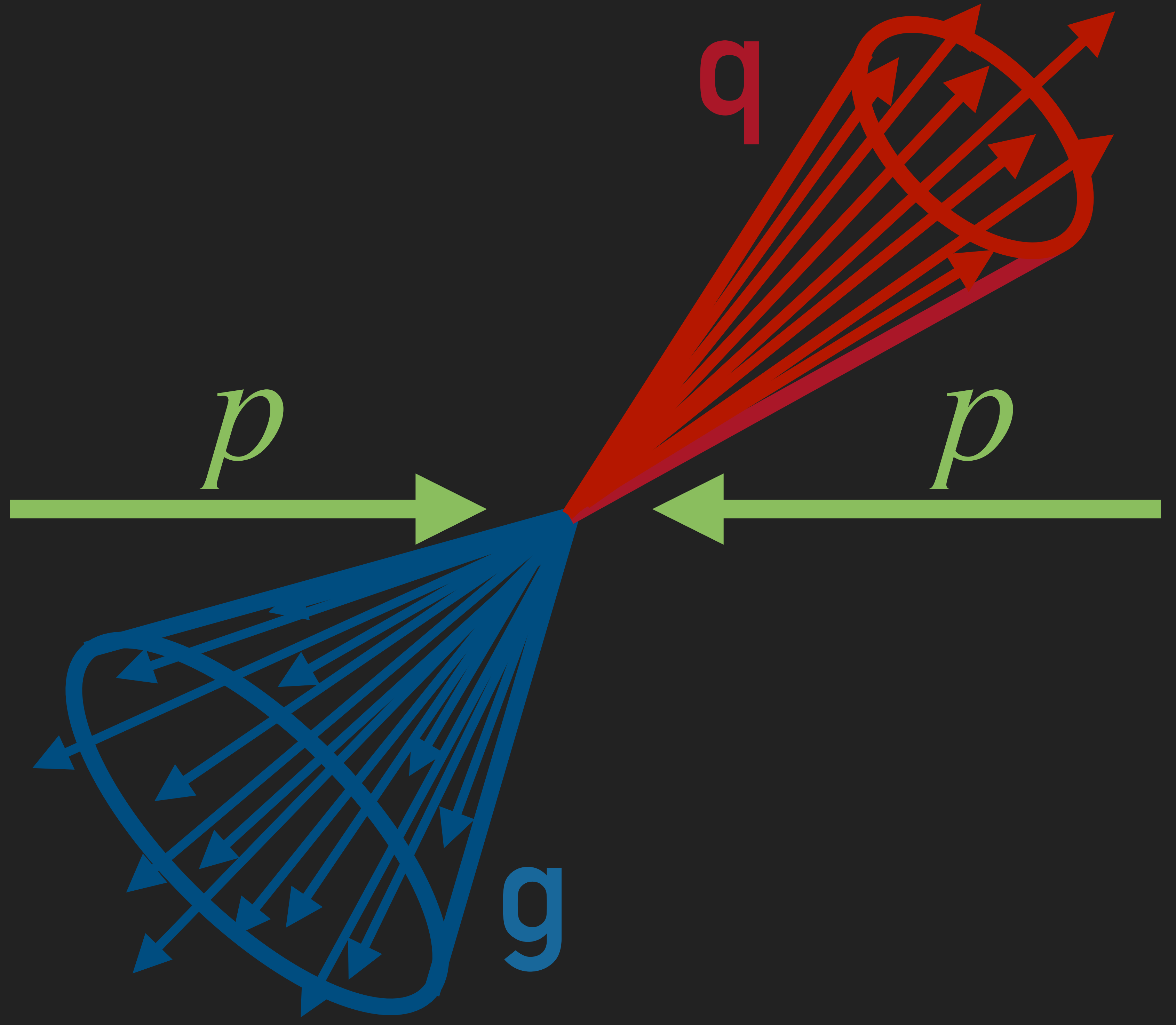
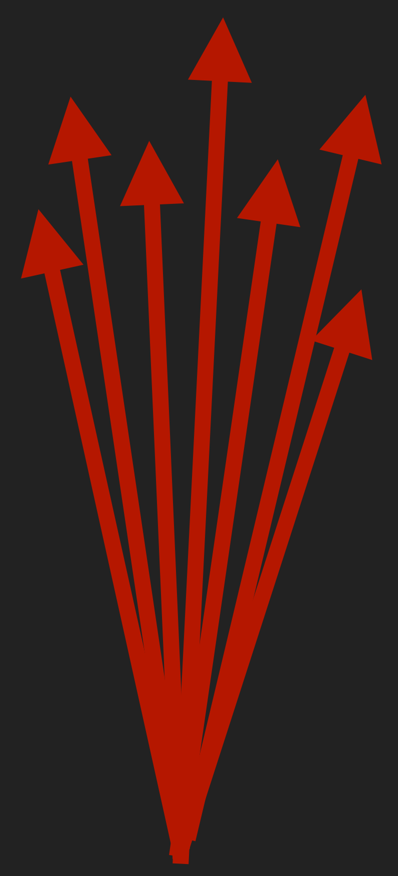
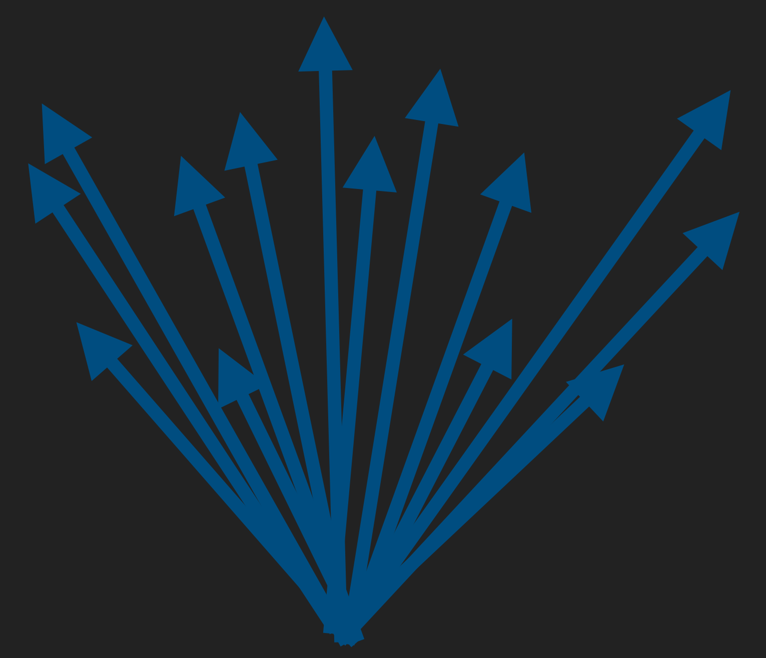


SAMUEL JANKOVYCH

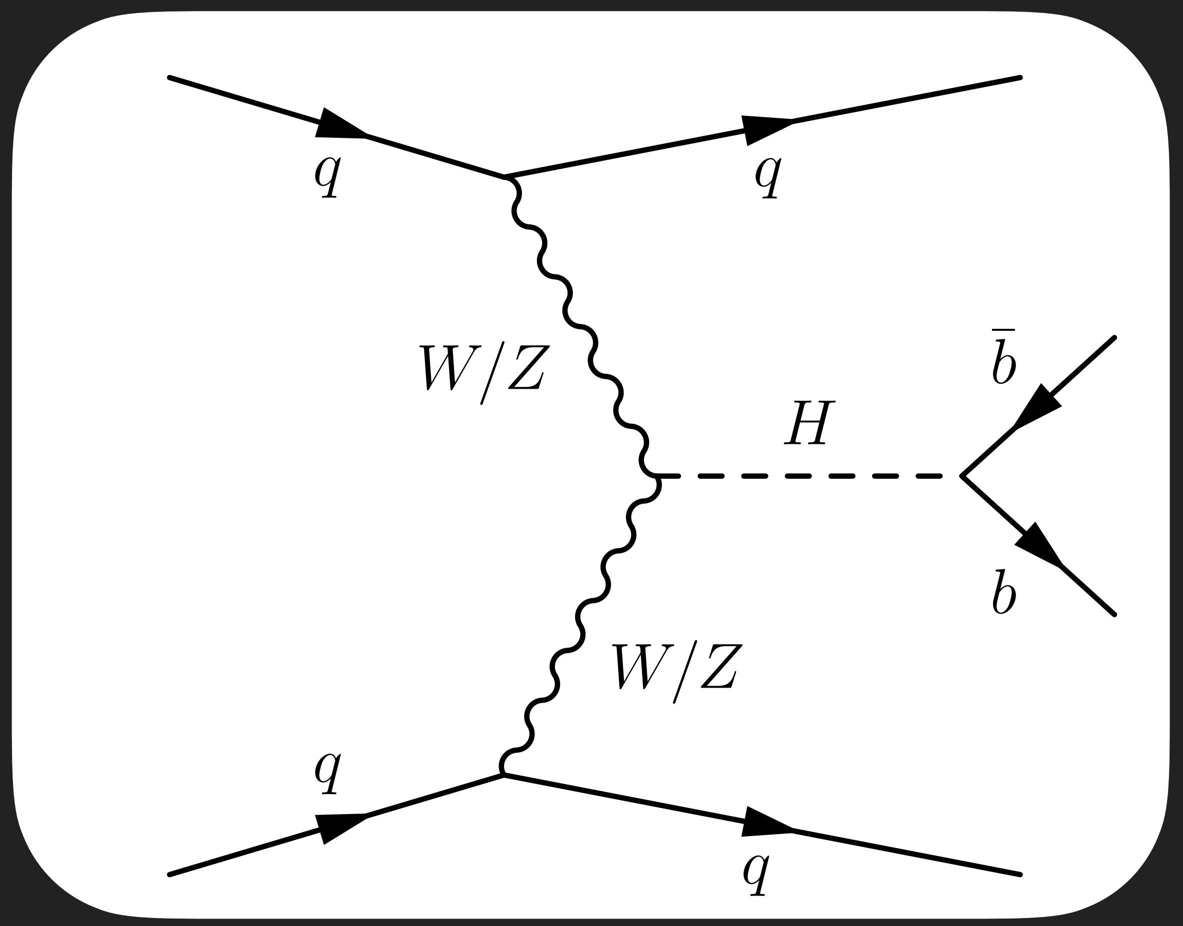
PERFORMANCE AND EFFICIENCY MEASUREMENT OF A TRANSFORMER-BASED QUARK GLUON JET TAGGER IN *ATLAS*

What? Quarks vs. Gluons

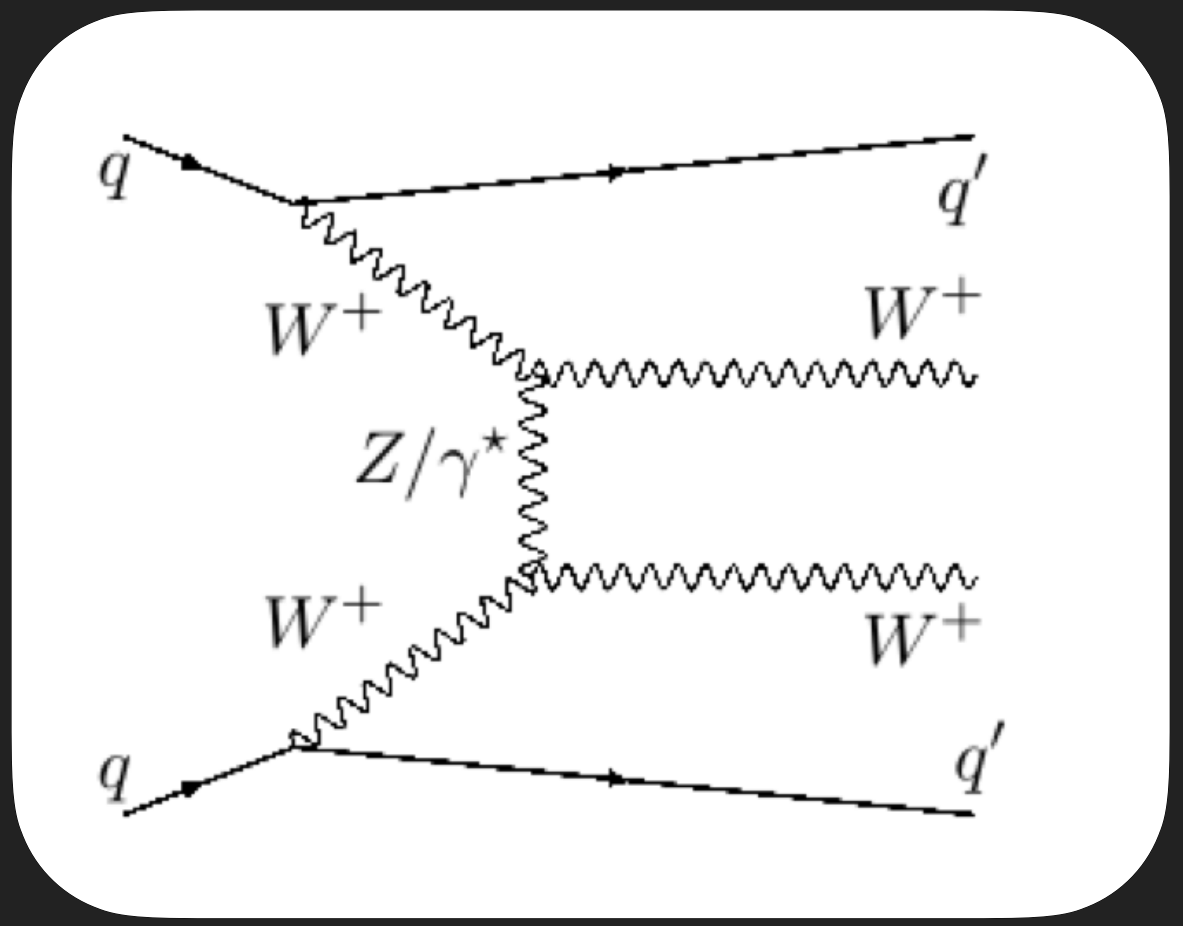




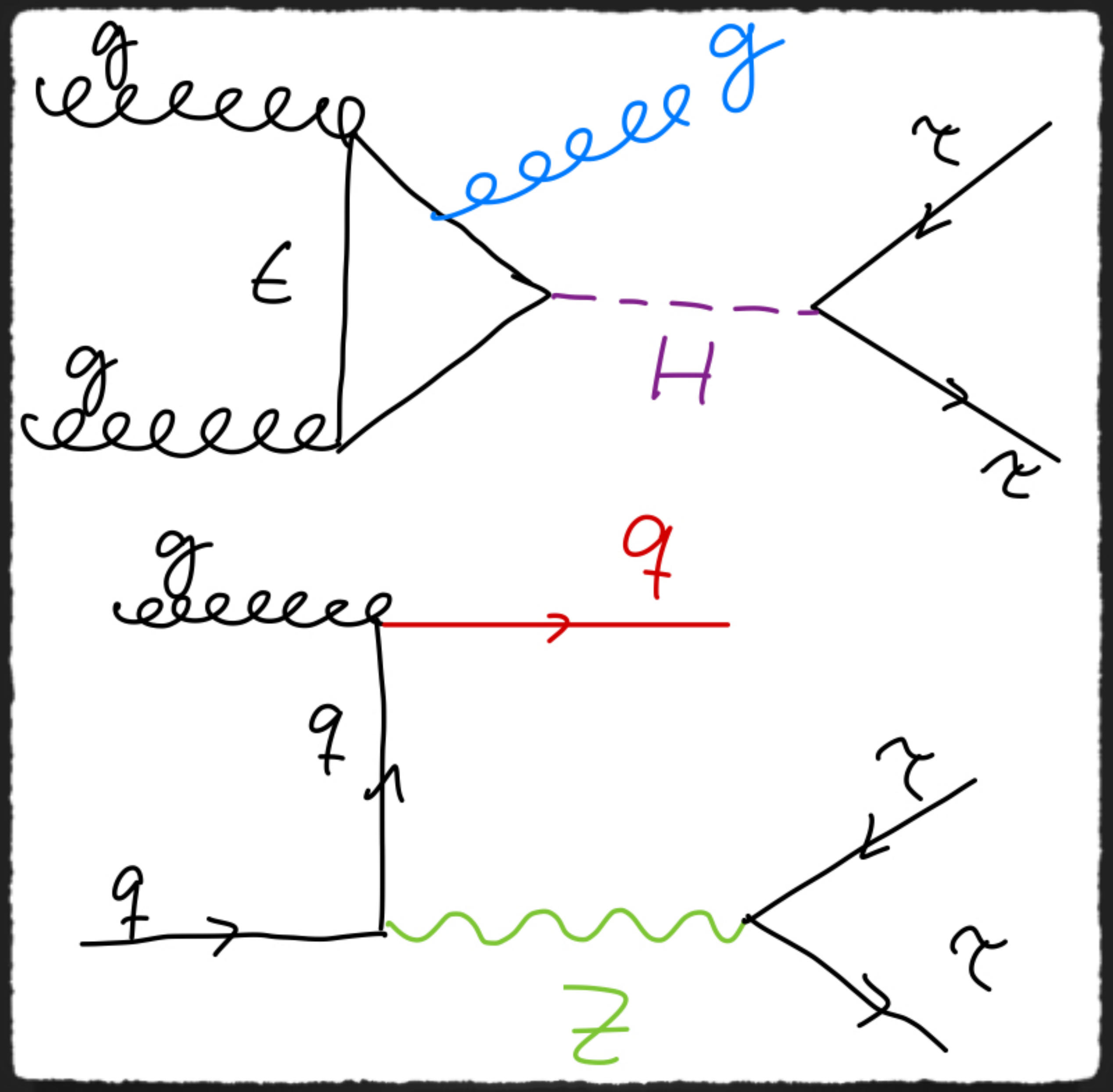
Why?



VBF



VBS

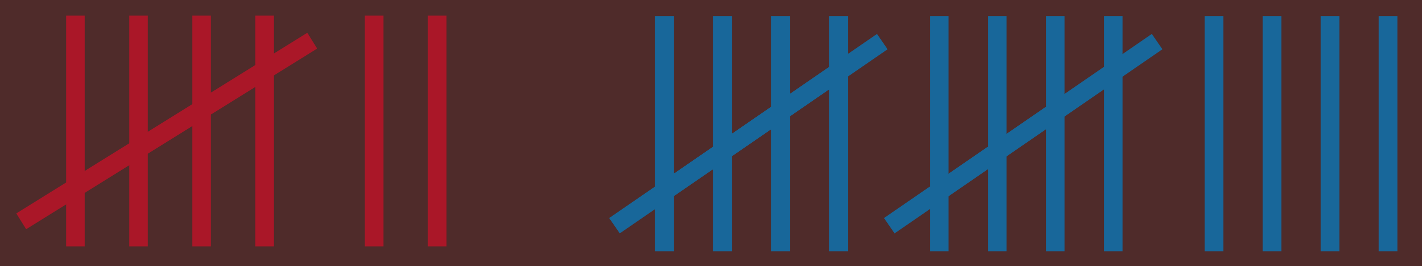


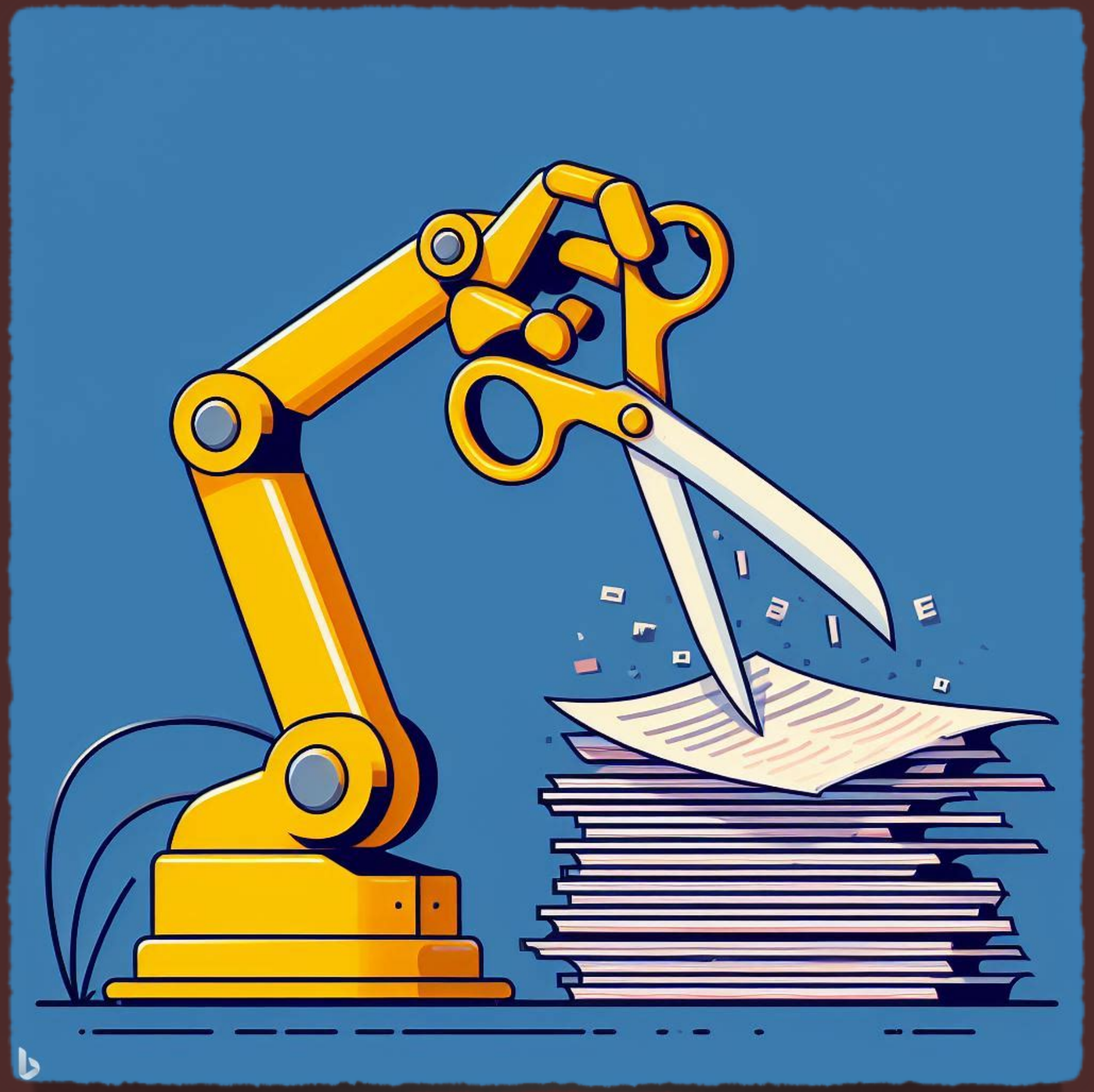
H+j vs Z+j

QUARK GLUON TAGGING AT *ATLAS*

OLD → NEW

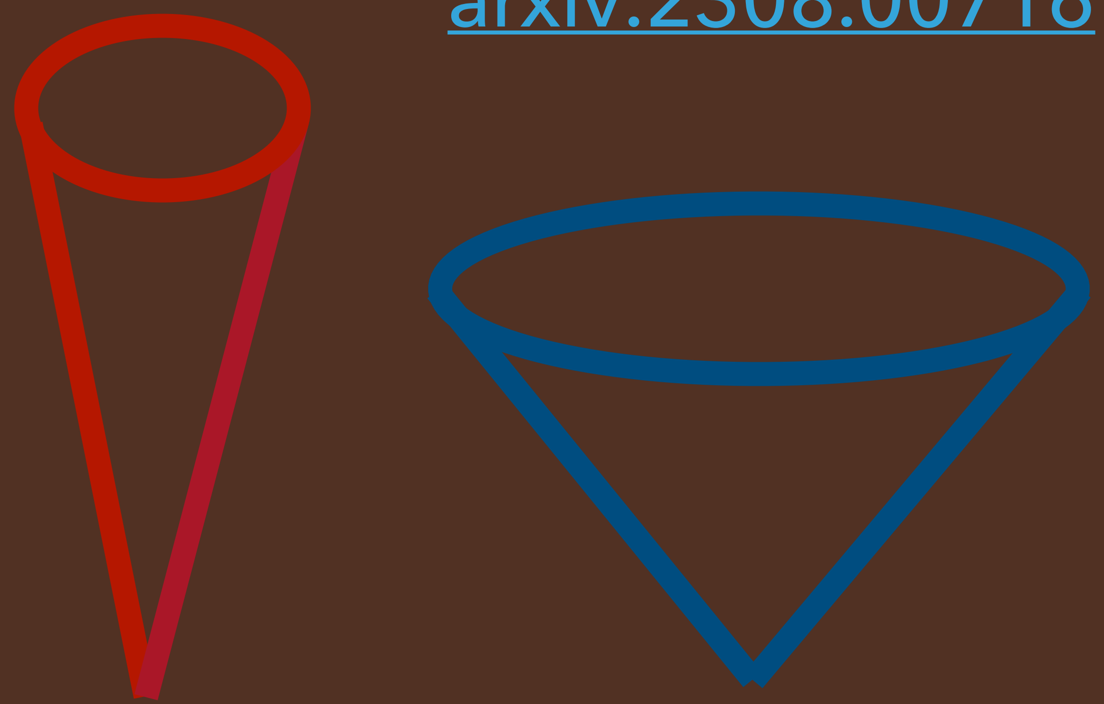
arxiv:1405.6583

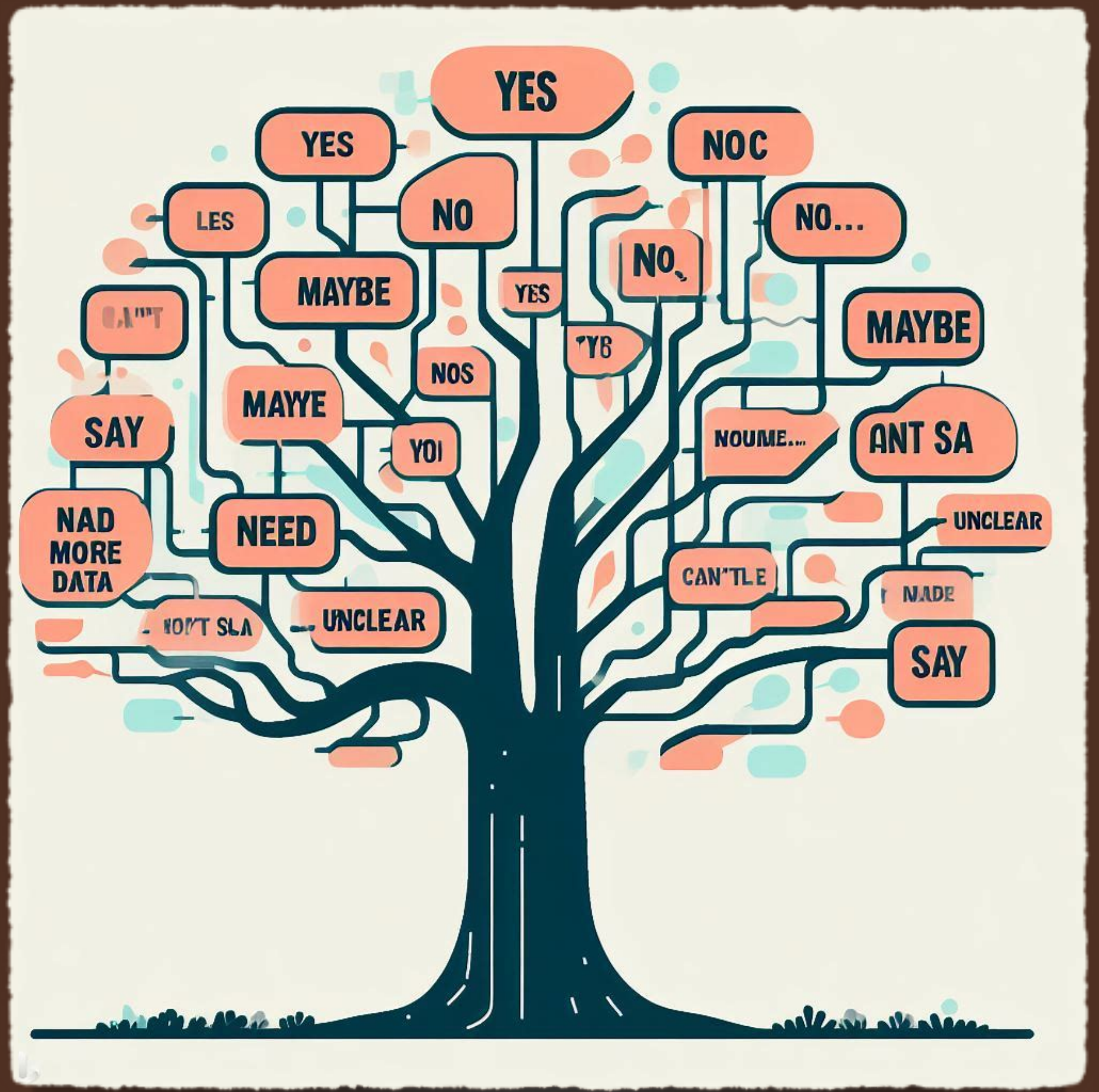




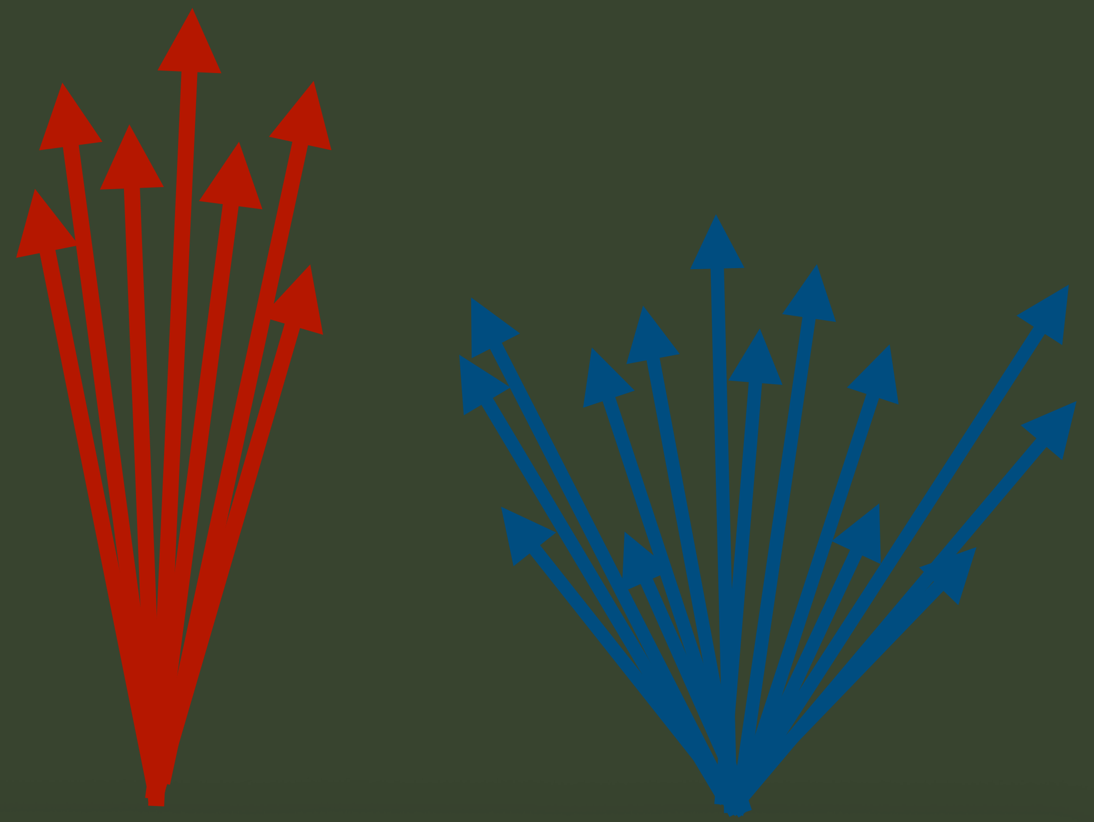
DALL-E's impression of **cut-based** tagger

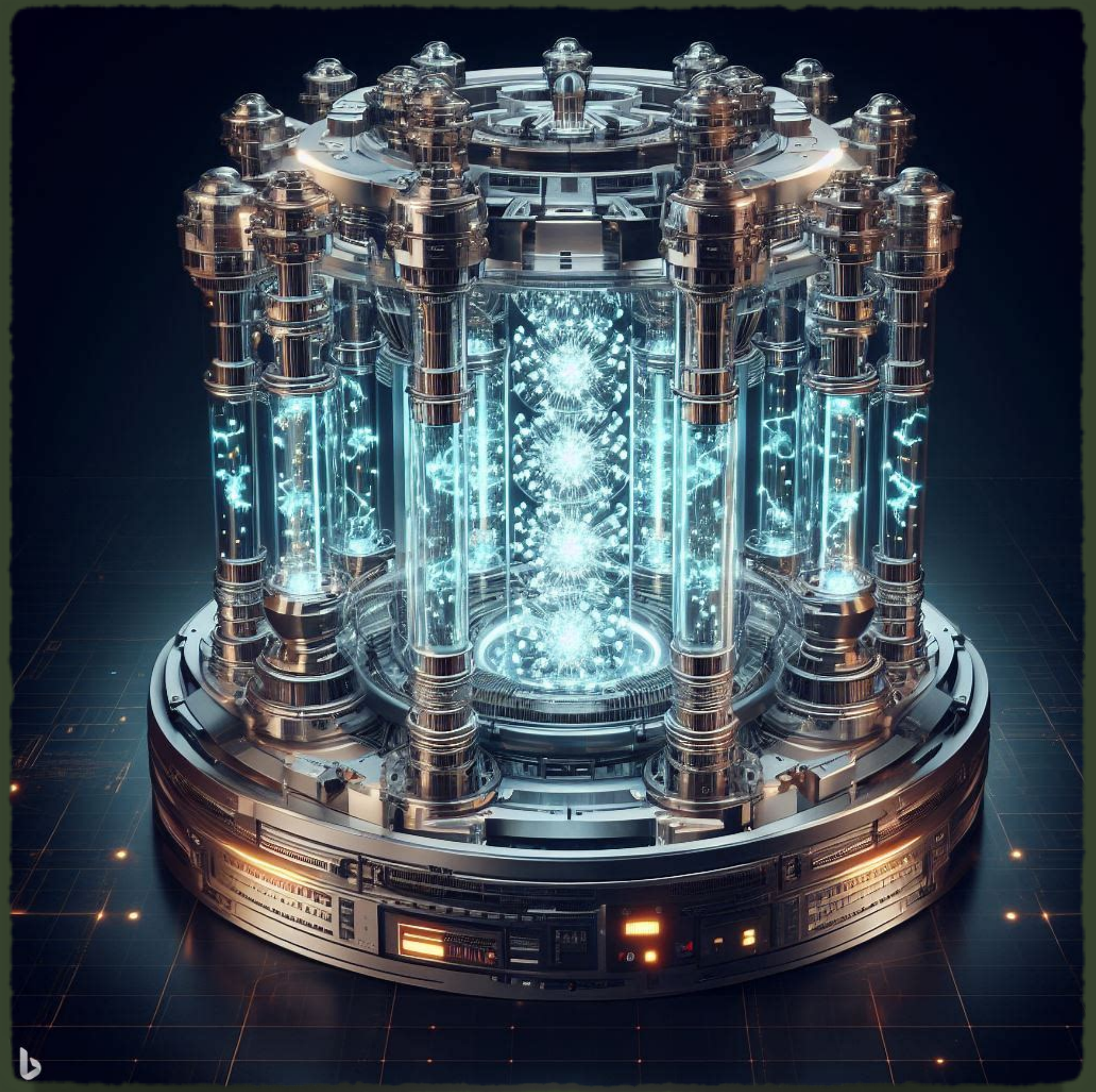
arxiv:2308.00716





DALL-E's impression of **BDT**



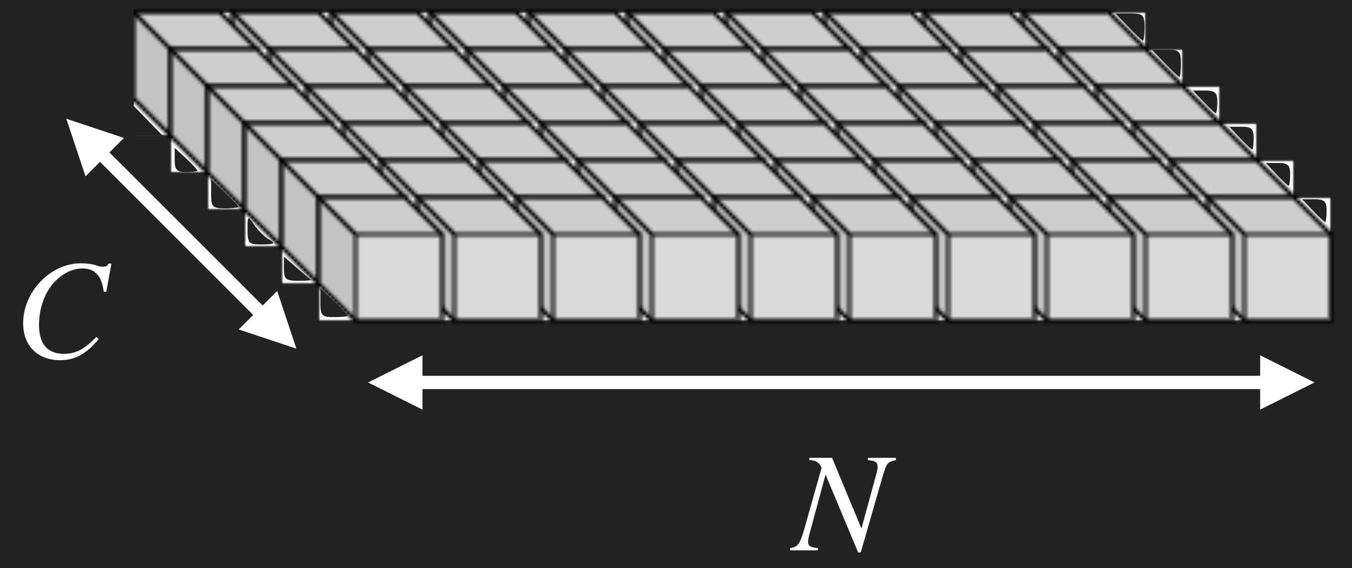


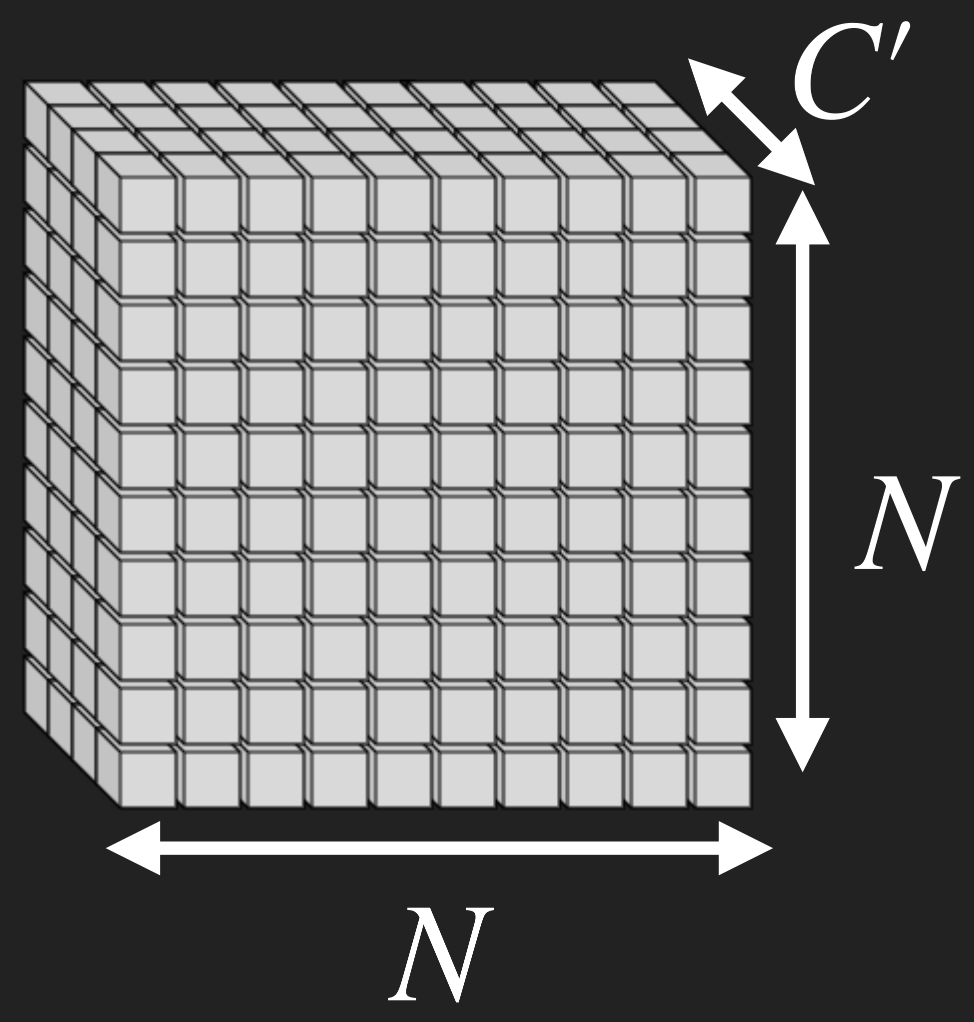
DALL-E's impression of **Transformer**

INPUT VARIALBES

Constituent Variables
$\log \frac{p_T}{p_T^{\text{jet}}}$
$\log \frac{E}{p_T^{\text{jet}}}$
$\Delta\eta = \eta - \eta^{\text{jet}}$
$\Delta\phi = \phi - \phi^{\text{jet}}$
$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$
isCharged
isTopo

Constituent Interaction Variables
$\log \Delta^{ab} = \log \sqrt{(\eta^a - \eta^b)^2 + (\phi^a - \phi^b)^2}$
$\log m^{2,ab} = \log ((p^a + p^b)^2 / (p_T^{\text{jet}})^2)$
$\log k_T^{ab} = \log (\min \left(\frac{p_T^a}{p_T^{\text{jet}}}, \frac{p_T^b}{p_T^{\text{jet}}}\right) \Delta^{ab})$
$z^{ab} = \min (p_T^a, p_T^b) / (p_T^a + p_T^b)$



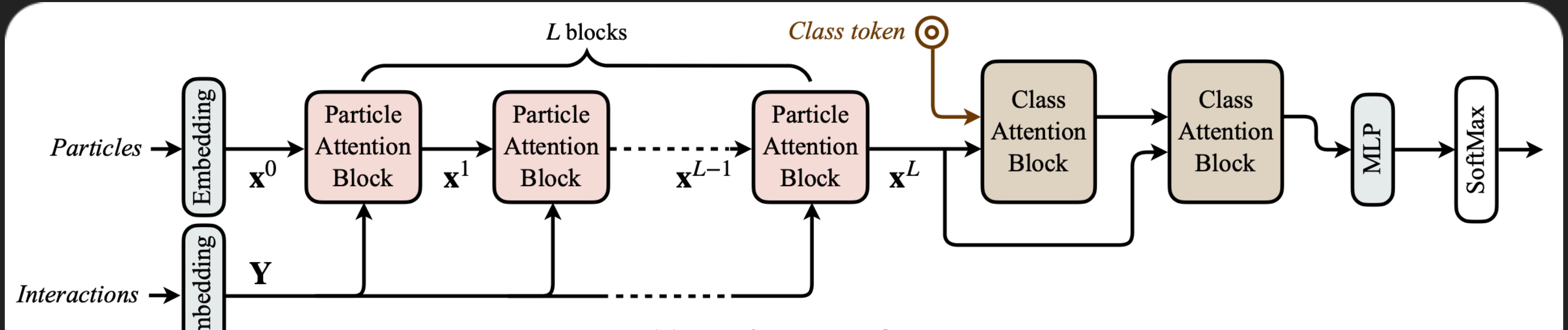


C = # of constituent variables = 7

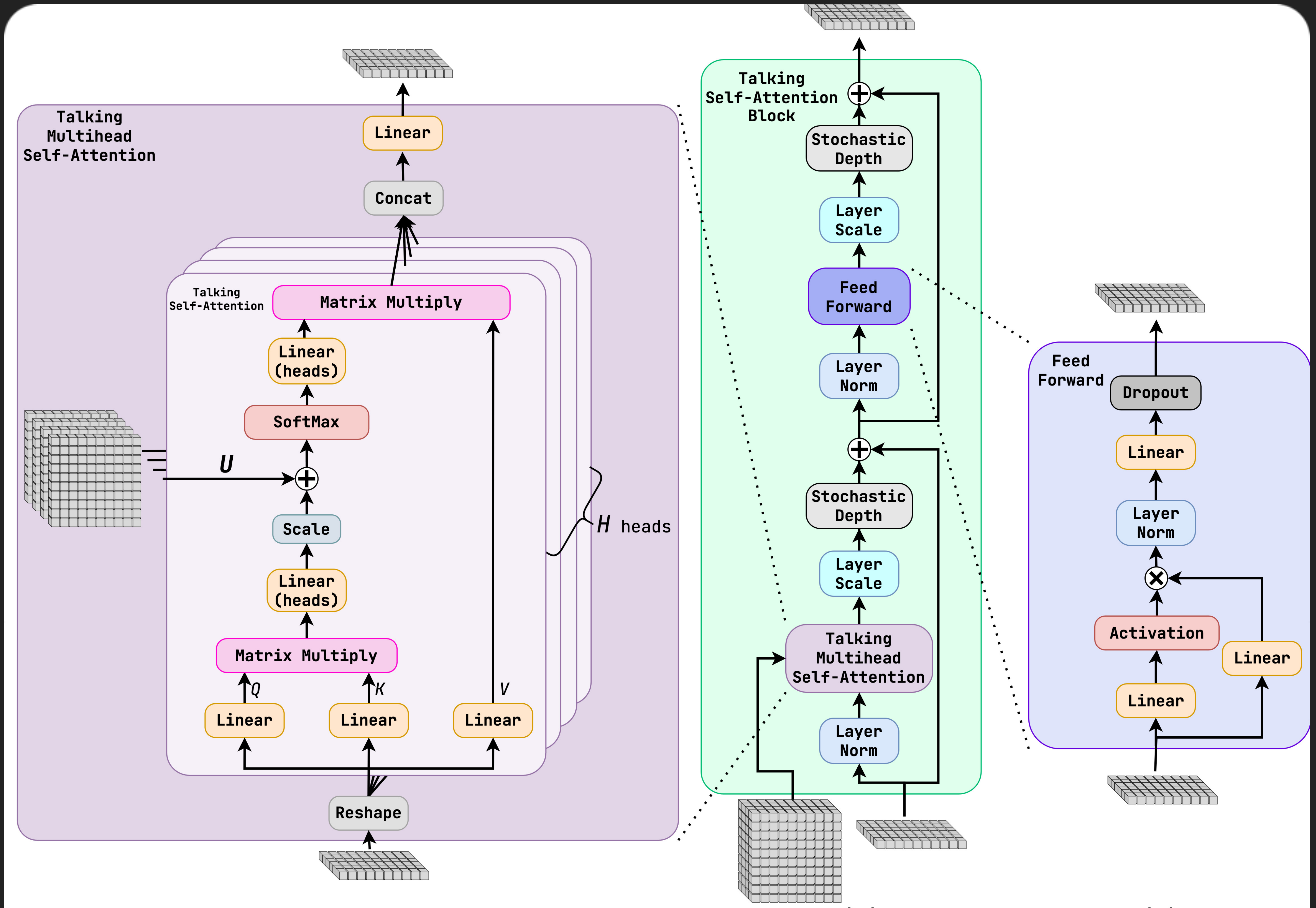
C = # of constituent interaction variables = 4

N = # of constituents (max 60)

DYNAMICALLY ENHANCED PARTICLE TRANSFORMER



(a) Particle Transformer

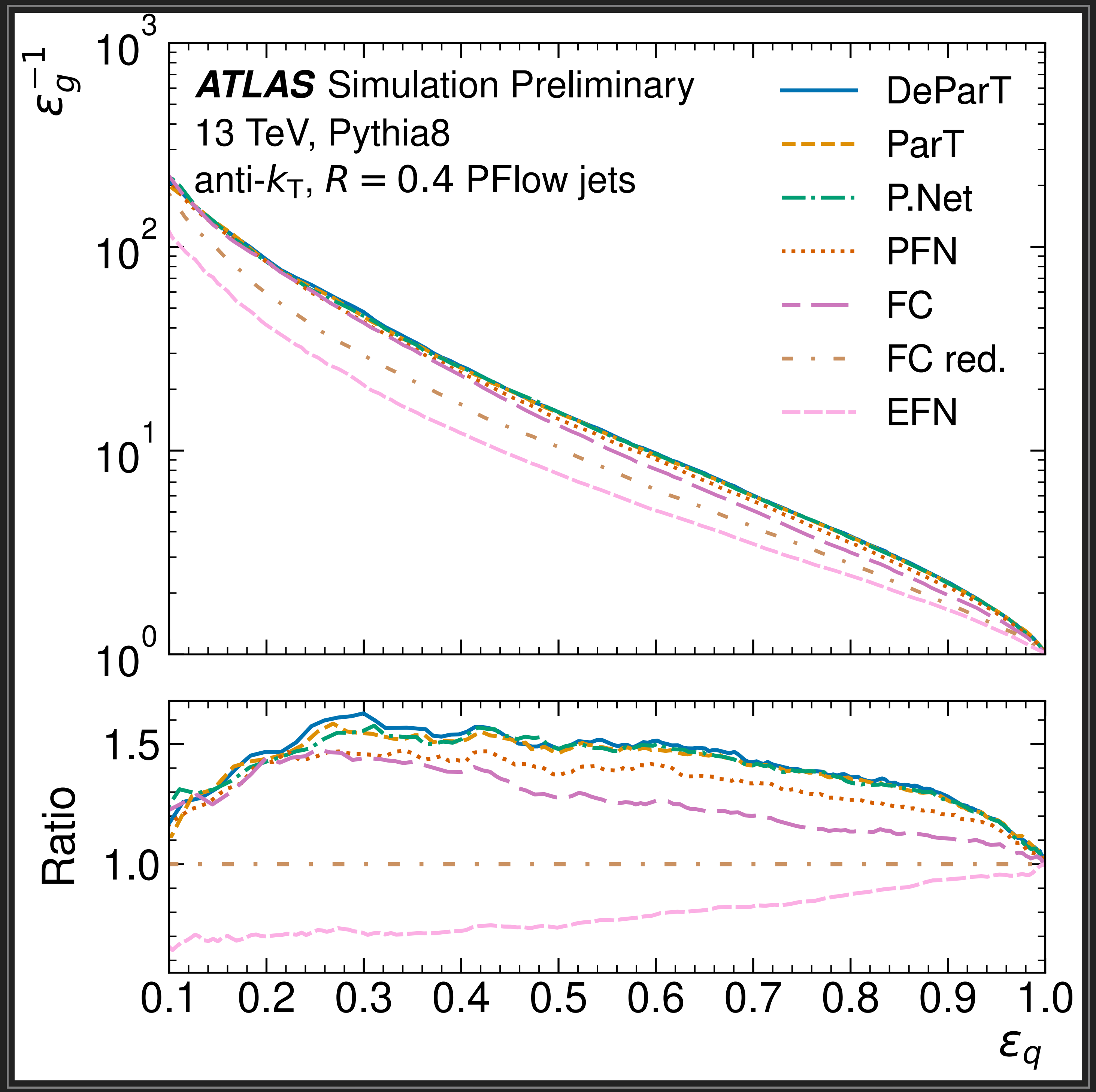


(a) (b) (c)

- enhancement of ParT
- DeiT III - deeper models
  - stochastic depth
  - layer scale
- gated FFN
- **Talking Heads** arxiv:2003.02436
  - more communication

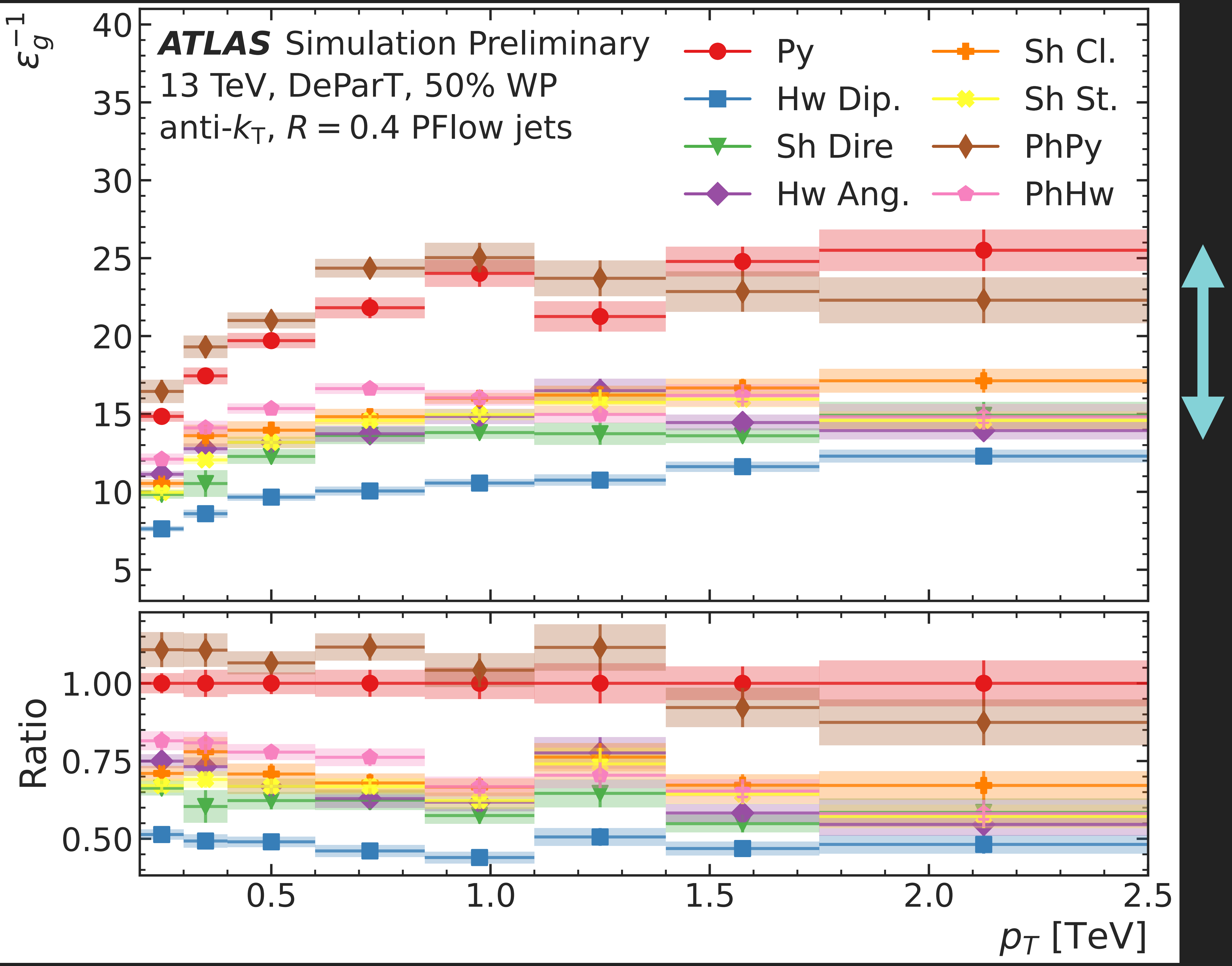
TAGGER PERFORMANCE

Model	AUC	$\varepsilon_g^{-1}@\varepsilon_q = 0.5$	# Params $[10^6]$	Inference Time [ms]	GPU Memory [MB]
DeParT	0.8489	15.4242	2.62	266.51	1684
ParT	0.8479	15.2457	2.62	233.84	1730
ParticleNet	0.8476	15.4402	2.59	768.74	5410
PFN	0.8406	14.2387	2.64	136.93	393
FC	0.8280	13.5199	2.63	65.53	76
FC reduced	0.8038	10.3639	2.63	84.84	47
EFN	0.7761	7.7222	2.60	101.53	337



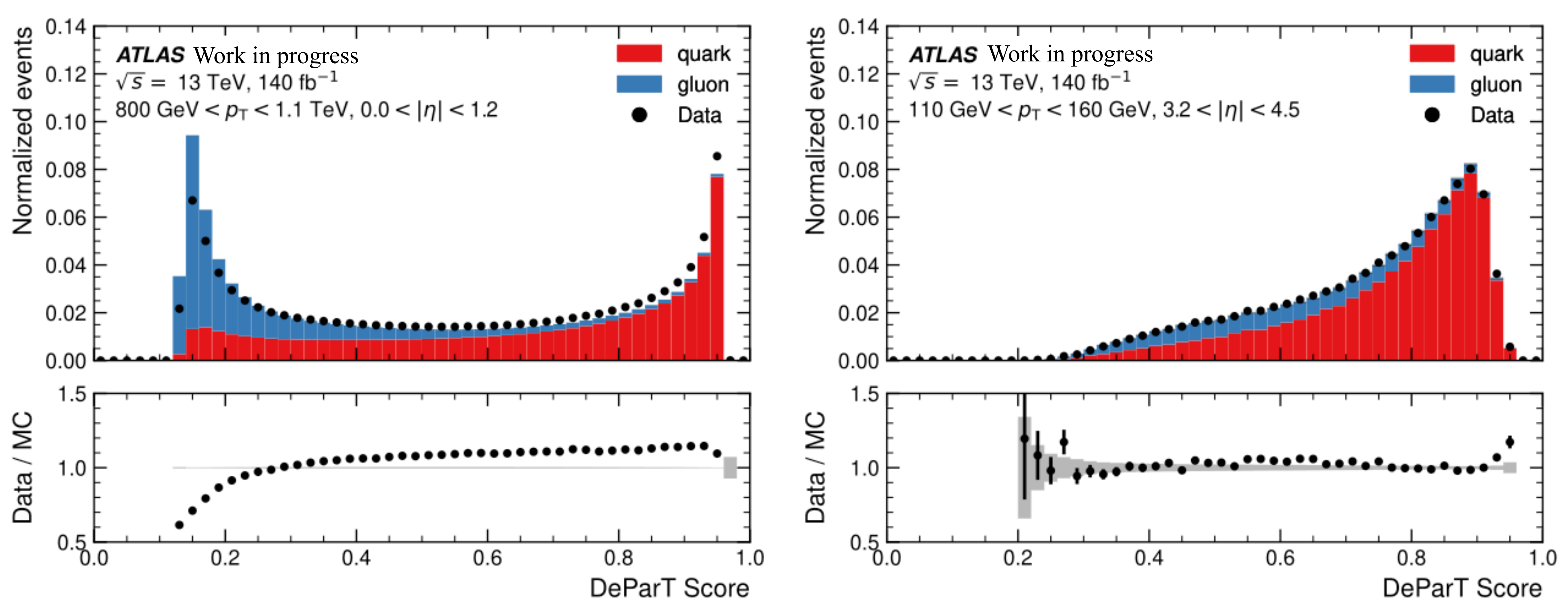
- fixed total # params –> similar performance
- DeParT outperforms ParT
- ParticleNet - expensive at 2.6M params

MONTE CARLO DEPENDENCE



- **big difference** from nominal
- only PhPy is similar
  - the same PS and Had.
- PS and Had. –> big effect

EFFICIENCY MEASUREMENT

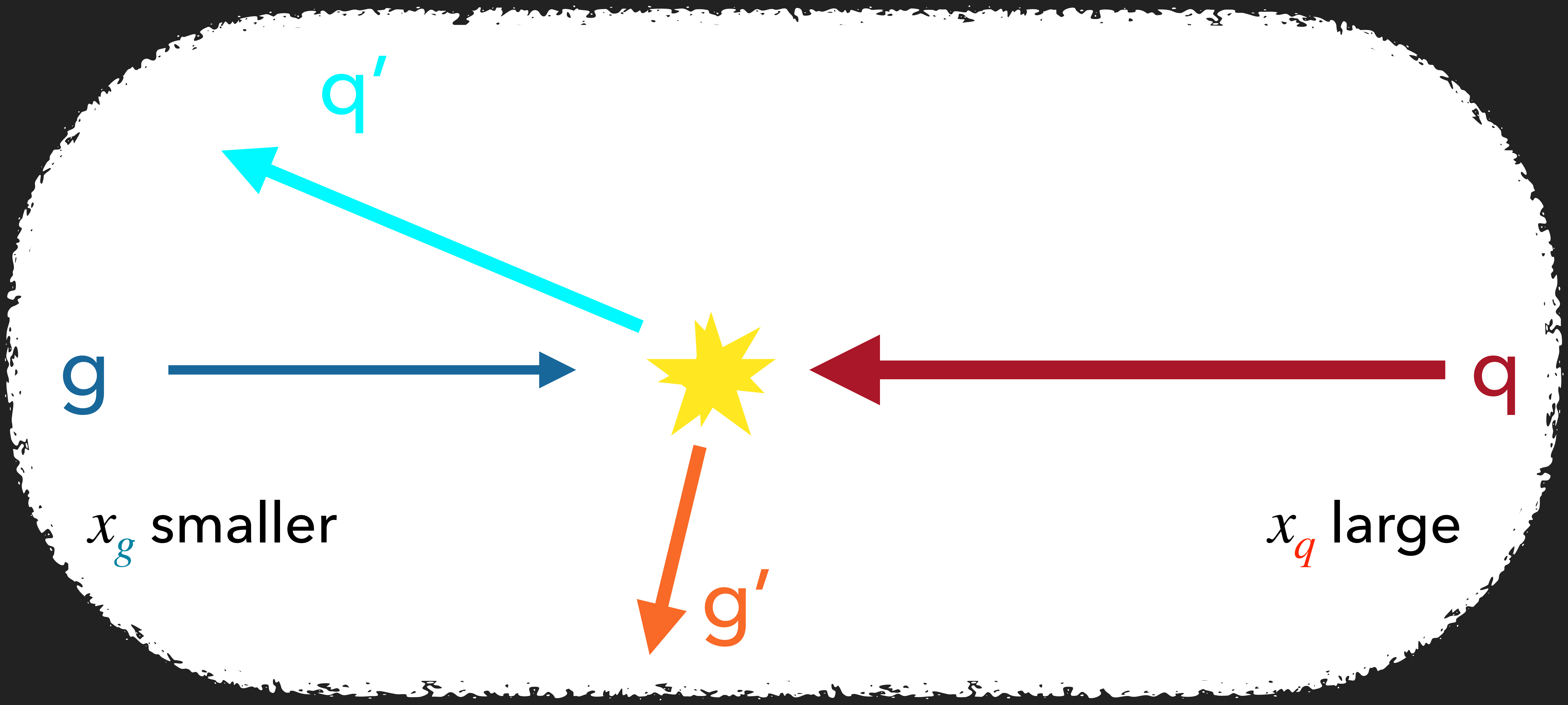


- **measure** efficiency on data
- **calibration** = correct differences between MC and data

Need pure quark and gluon samples in data

MIXED DISTRIBUTIONS

dijet event
- central jet = jet w/ lower eta = gluon enriched
- forward jet = jet w/ higher eta = quark enriched



$$p_F(x) = f_F^q p_q(x) + (1 - f_F^q) p_g(x)$$

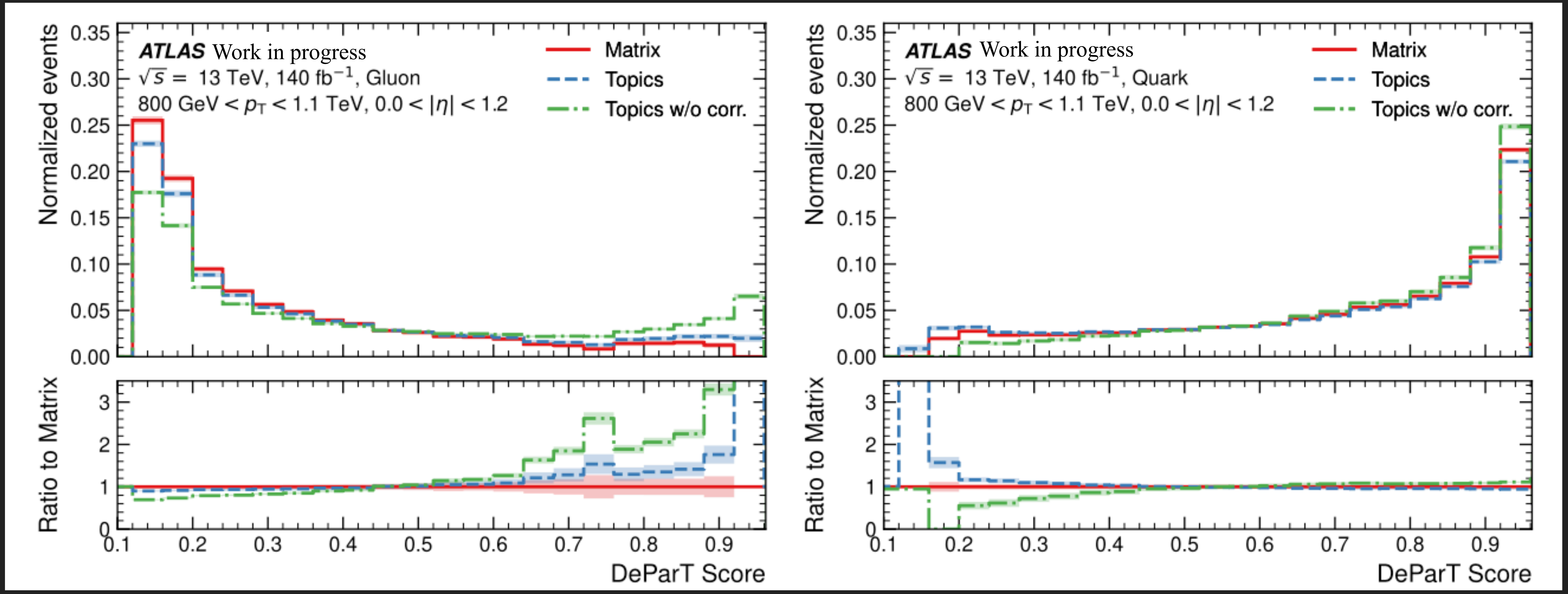
$$p_C(x) = f_C^q p_q(x) + (1 - f_C^q) p_g(x)$$

Matrix Method

- estimate mixing fractions f_F^q, f_C^q from MC
- solve equations for $p_g(x), p_q(x)$
- MC based

Jet Topics

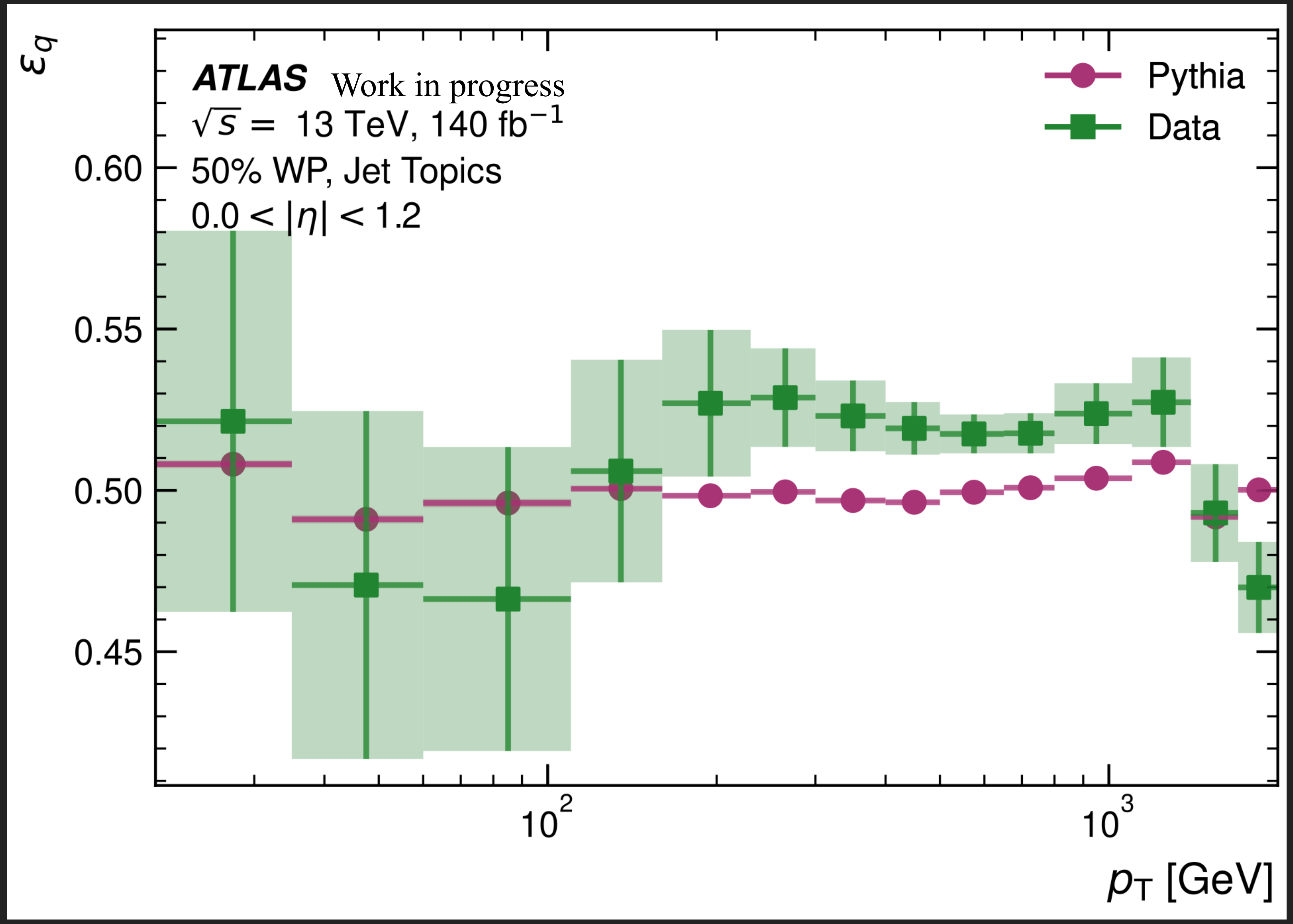
- assume additional condition (mutual irreducibility) on $p_g(x), p_q(x)$
- get $p_g(x), p_q(x)$ directly from $p_F(x), p_C(x)$
- use MC only as a correction
- reduced MC dependance



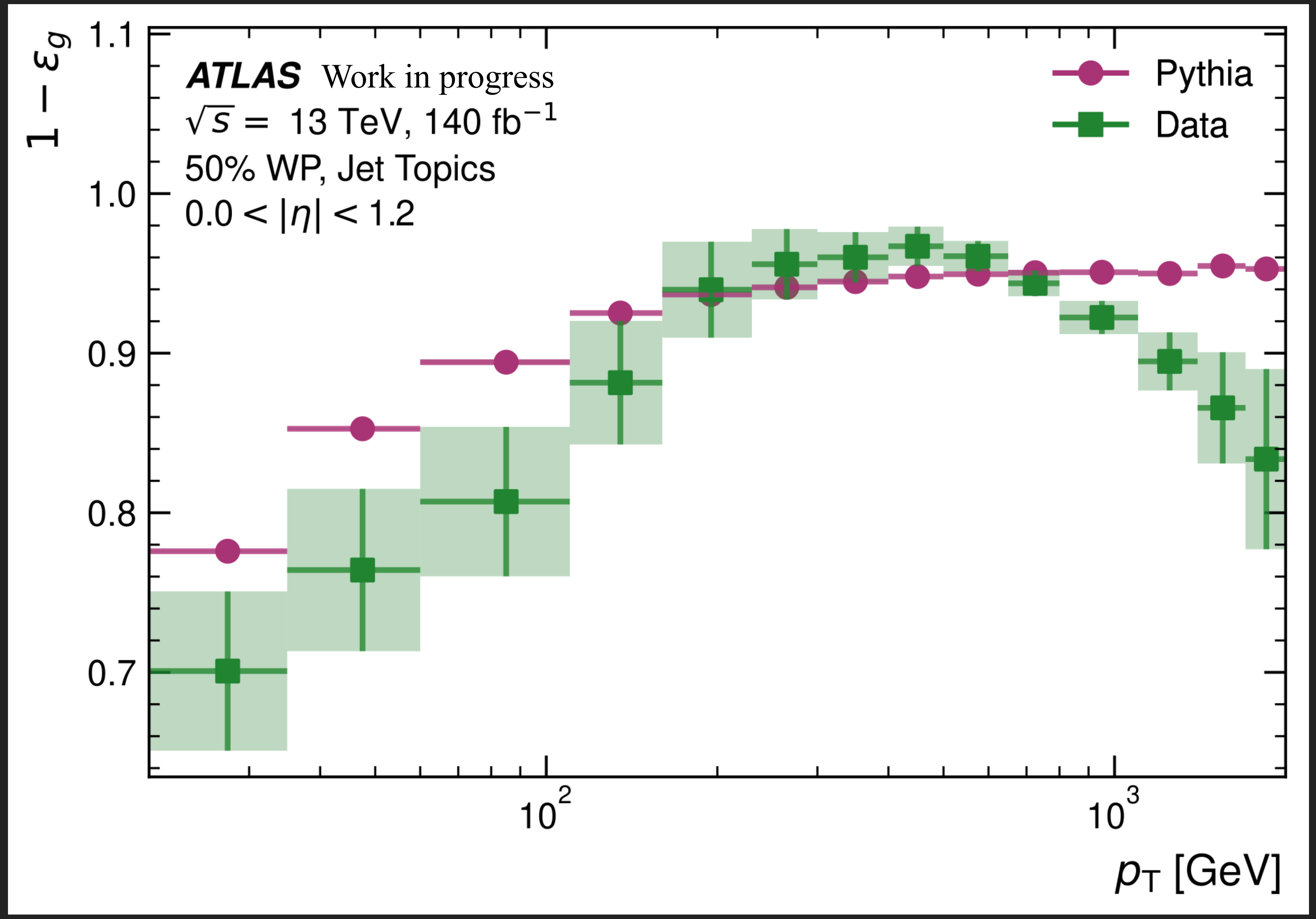
EFFICIENCY MEASUREMENT

$$\varepsilon_q(x_{\mathrm{WP}}) = \int_{x_{\mathrm{WP}}}^{1} p_q(x)\,dx$$

$$\varepsilon_g(x_{\mathrm{WP}}) = \int_{x_{\mathrm{WP}}}^{1} p_g(x)\,dx$$



Quark



Gluon

UNCERTAINITIES

Matrix Method

1. statistical
2. JES/JER, JVT, PU
3. sample independence
4. MC non-closure
5. **theoretical = MC modeling**

Exp.

Meth.

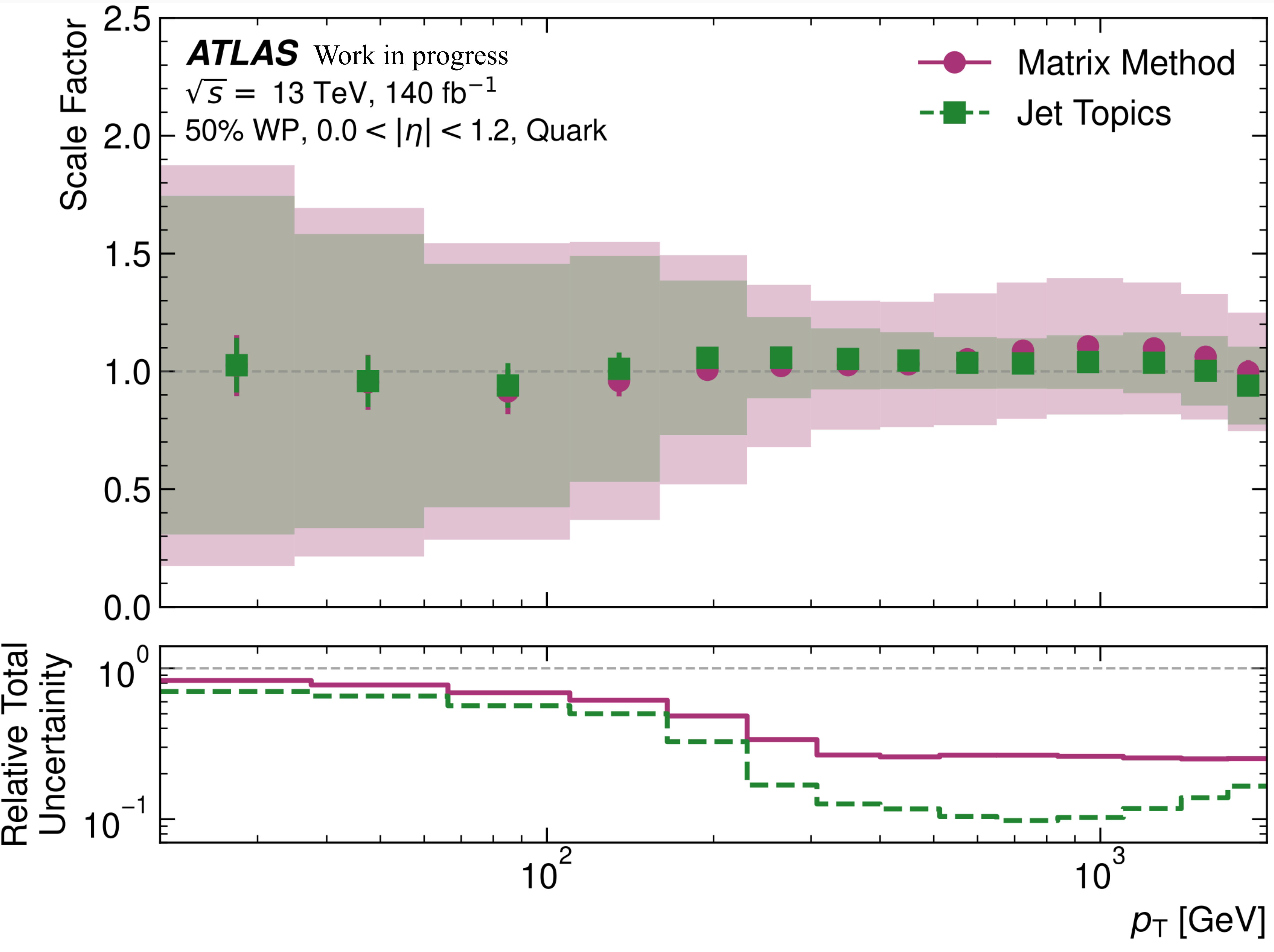
Theo.

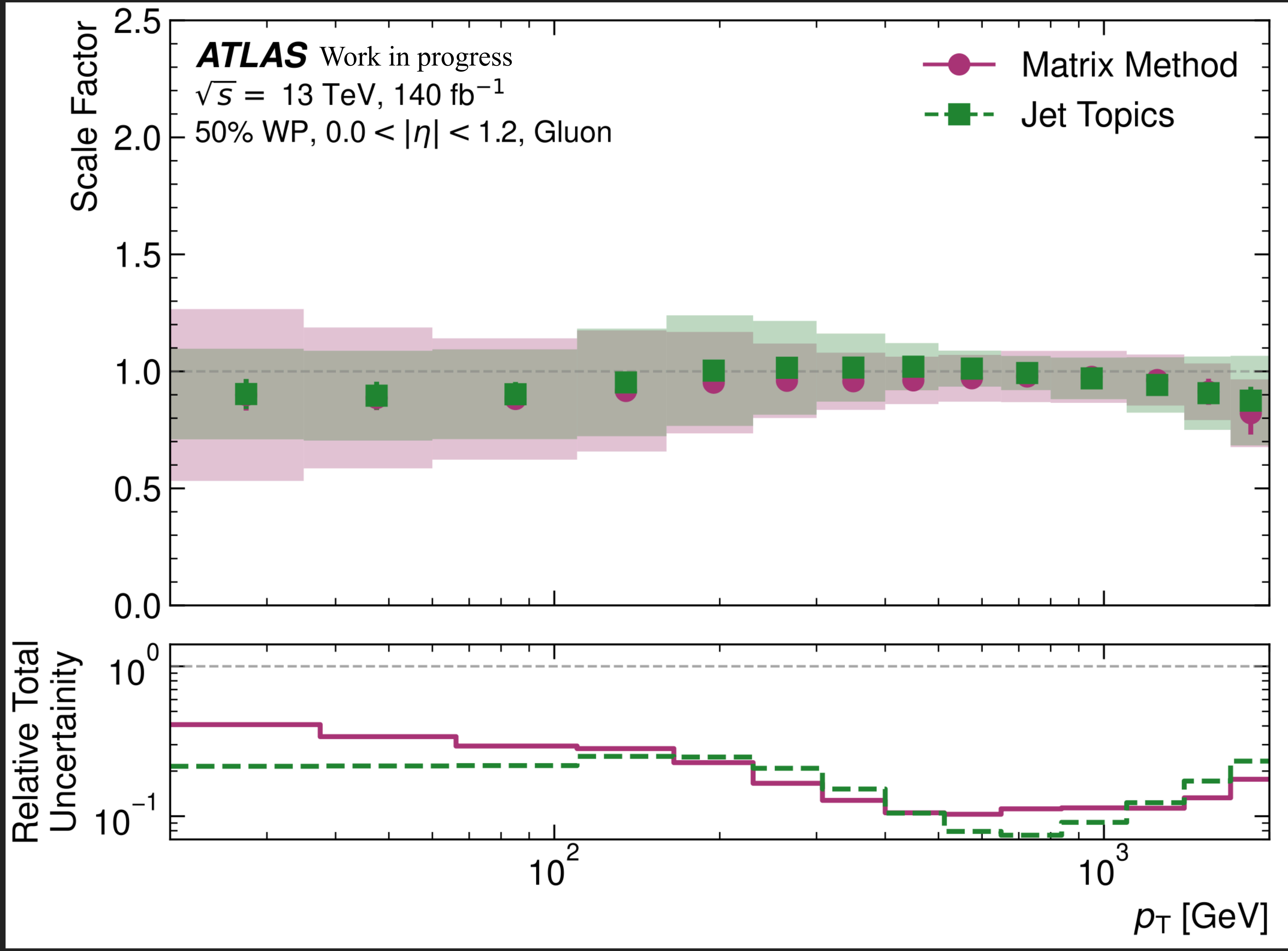
Jet Topics

1. statistical
2. JES/JER, JVT, PU
3. sample independence
4. MC non-closure
5. κ extraction
6. $\kappa(q,g), \kappa(g,q)$ modelling

JET TOPICS VS MATRIX METHOD

smaller uncertainities of jet topics

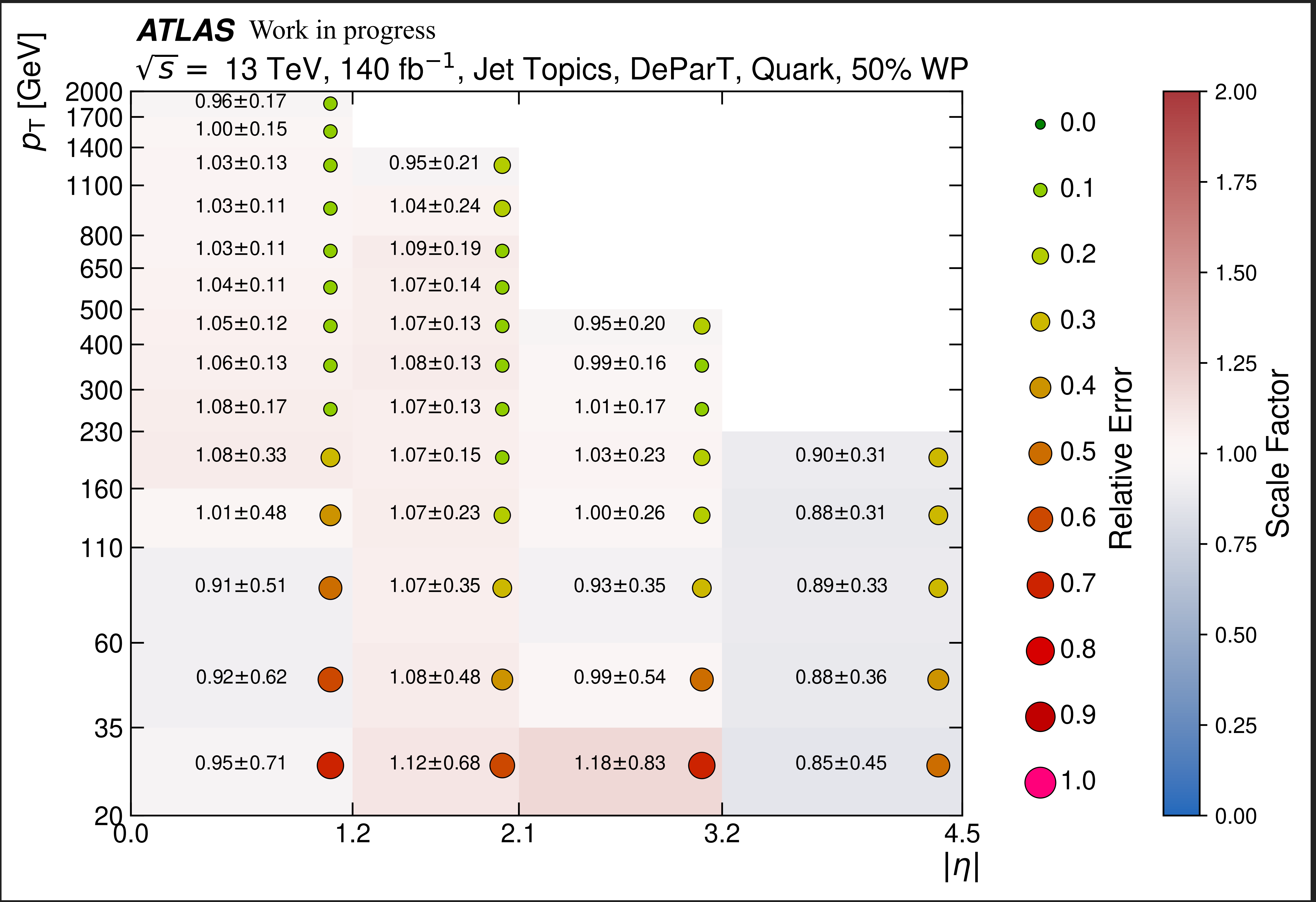


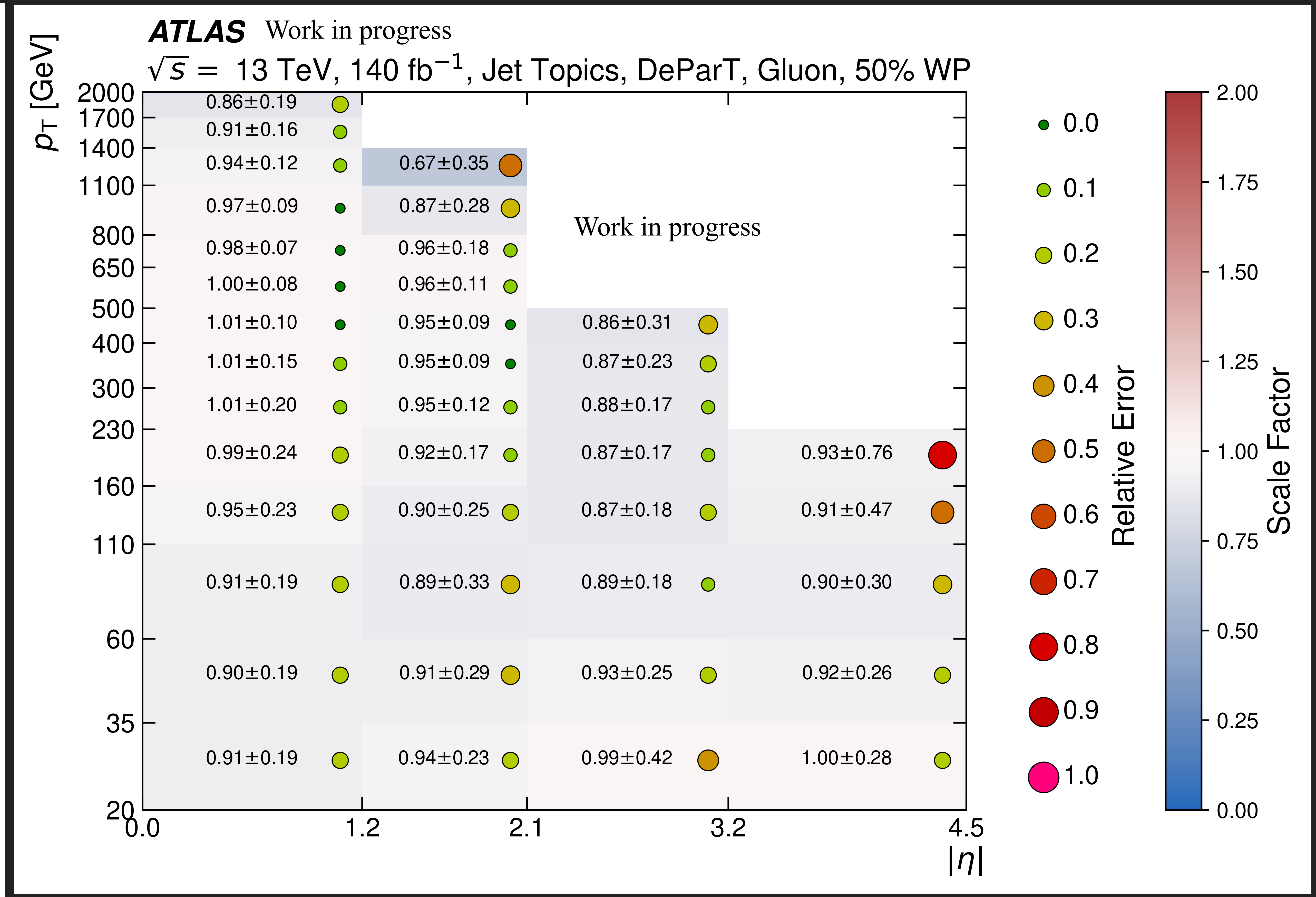


Quark $\mathrm{SF}_g = \dfrac{1 - \varepsilon_g^{\mathrm{Data}}}{1 - \varepsilon_g^{\mathrm{MC}}}$

Gluon $\mathrm{SF}_q = \dfrac{\varepsilon_q^{\mathrm{Data}}}{\varepsilon_q^{\mathrm{MC}}}$

EFFICIENCY CORRECTION FACTORS



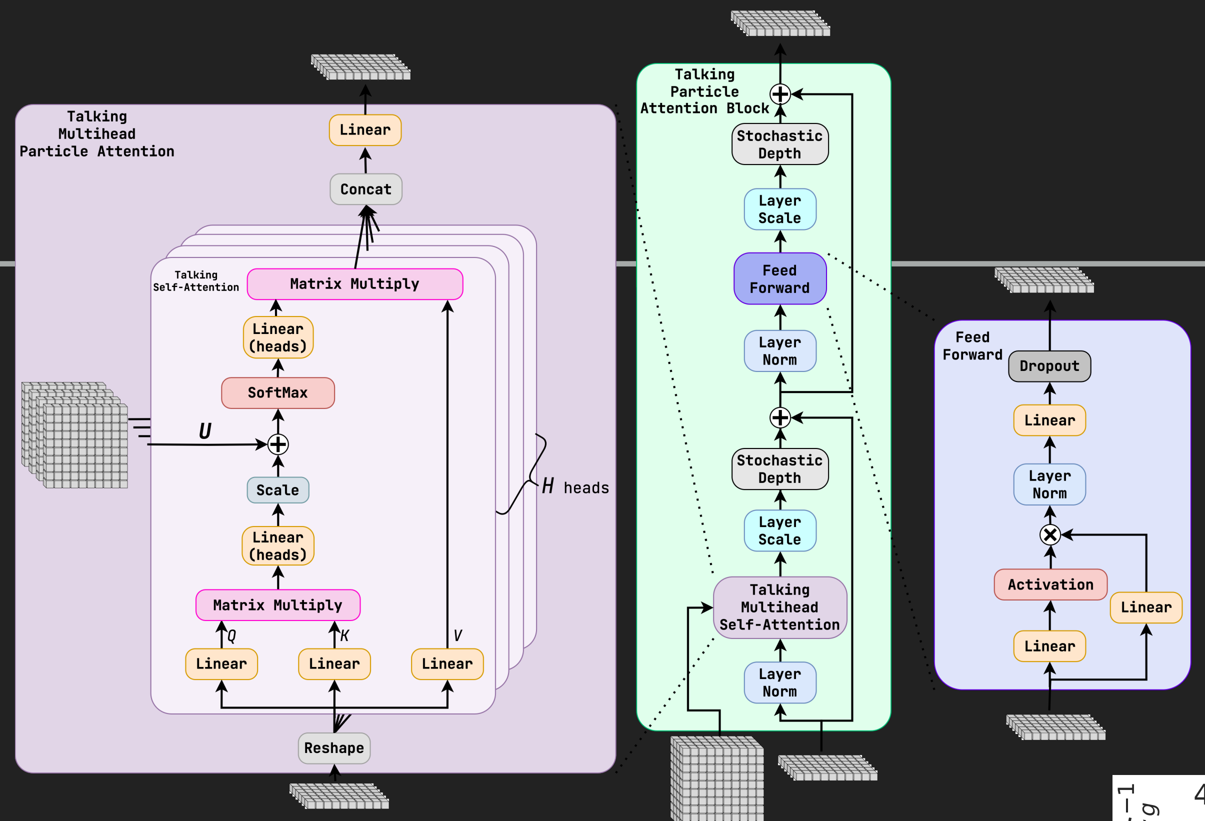


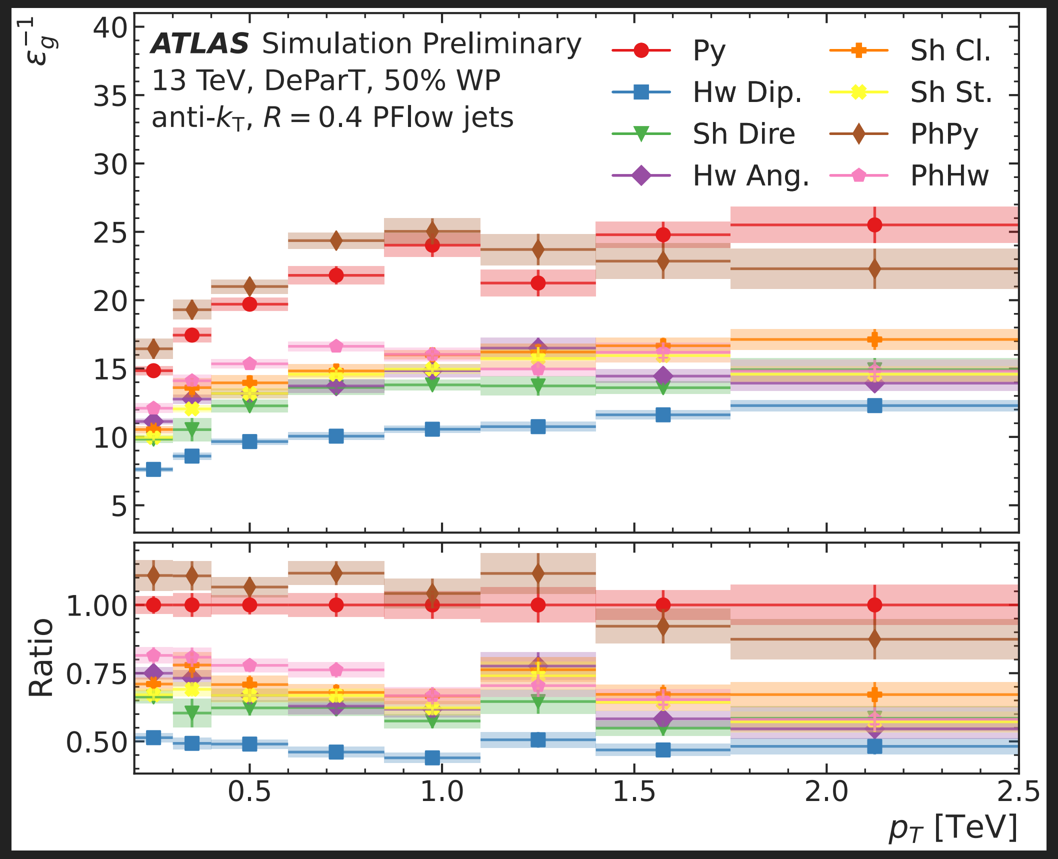
Quark $$\mathrm{SF}_q = \frac{\varepsilon_q^{\mathrm{Data}}}{\varepsilon_q^{\mathrm{MC}}}$$

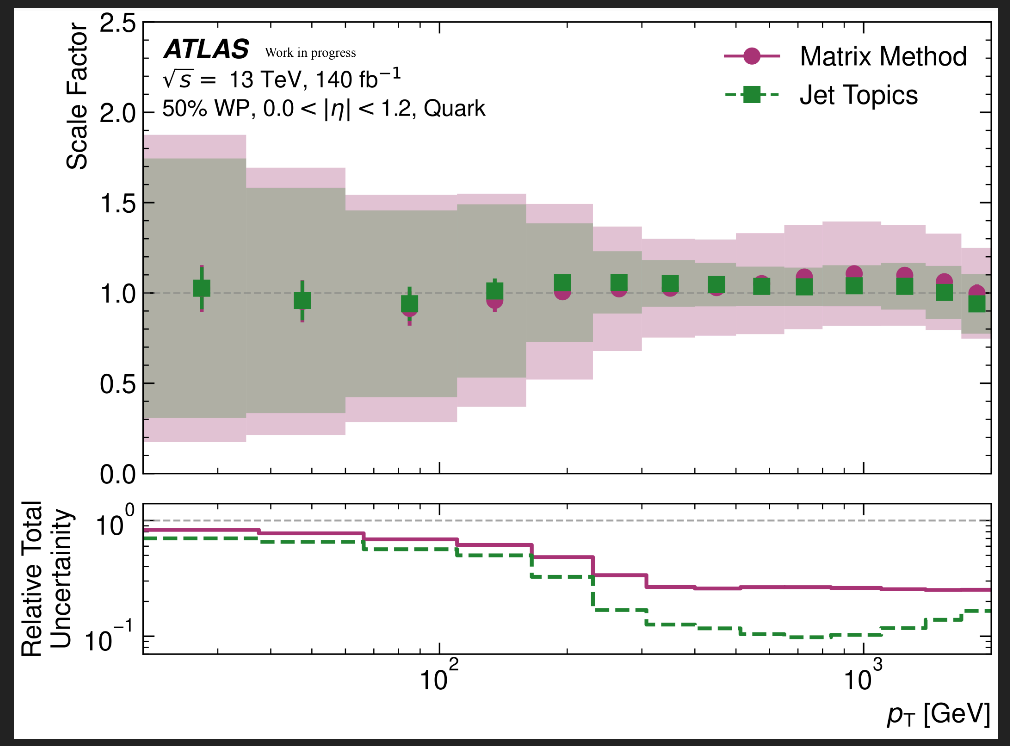
Gluon $$\mathrm{SF}_g = \frac{1-\varepsilon_g^{\mathrm{Data}}}{1-\varepsilon_g^{\mathrm{MC}}}$$

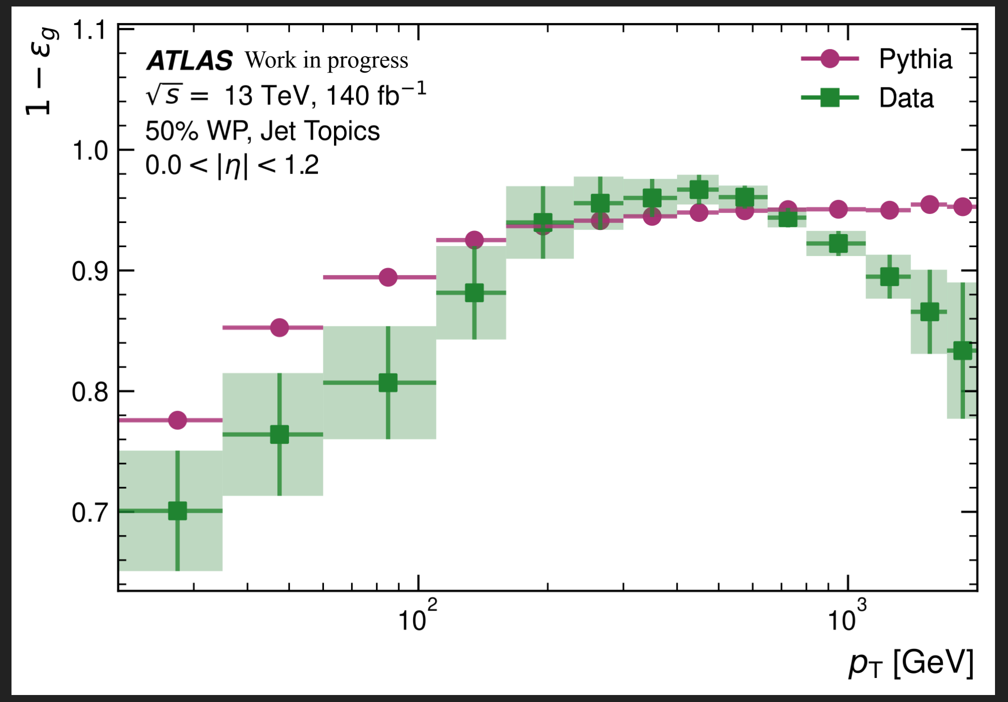
CONCLUSION

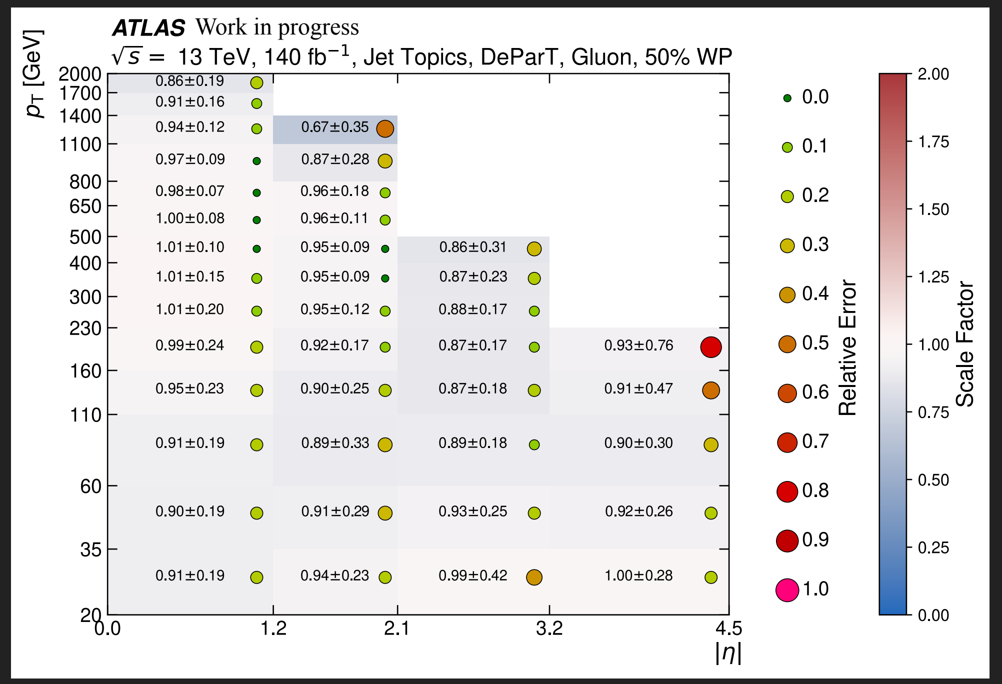
- Developed transformer-based q/g tagger
- MC dependence
- Jet Topics calibration - reduces uncertainties
- Performance measurement in data
- Scale factors for physics analysis applications











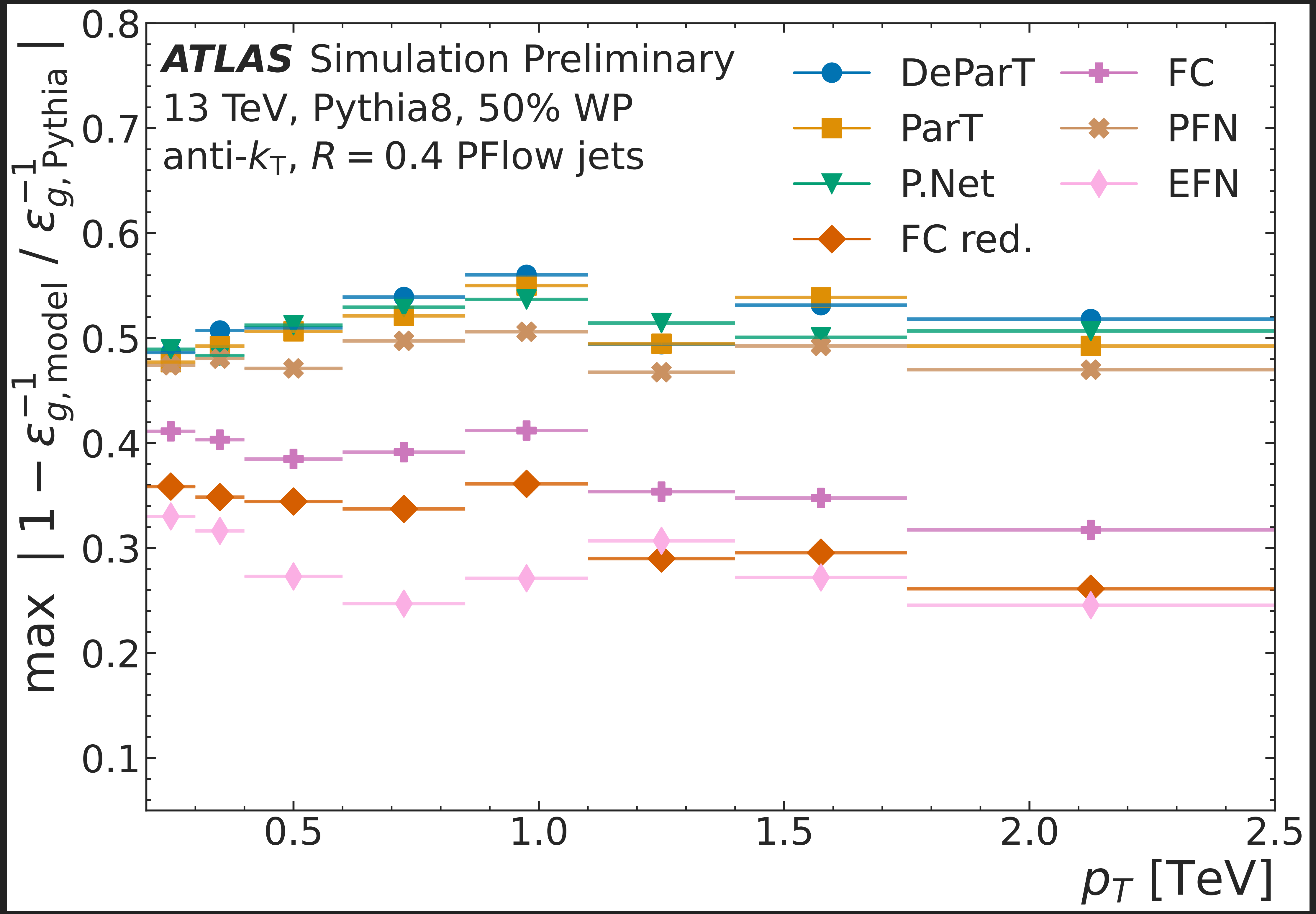
THANK YOU!

QUESTIONS?

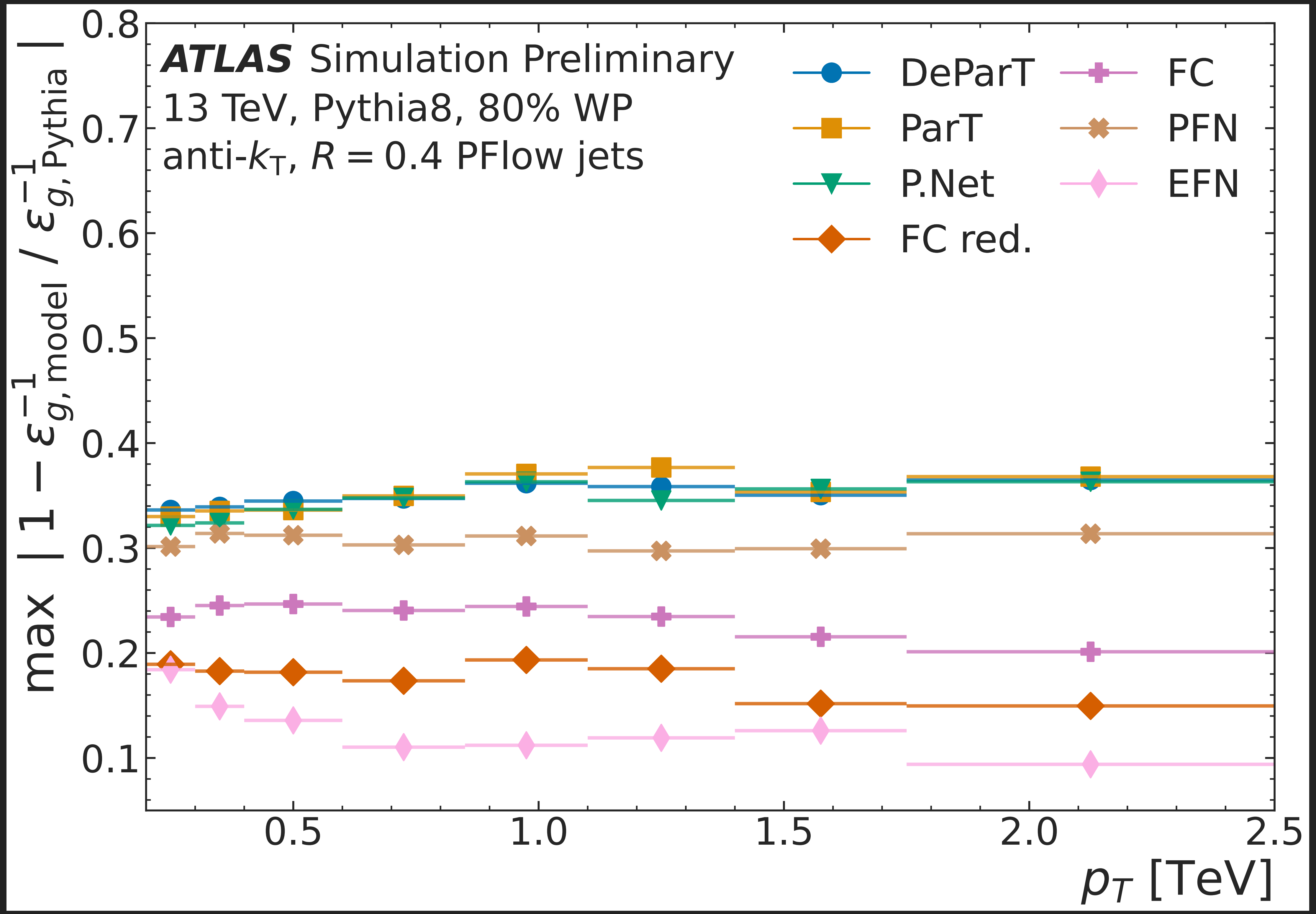
BONUS

MONTE CARLO DEPENDENCE

50% WP



80% WP



- higher = bigger MC dep.
- EFN - lowest MC dep.
- const. based - significantly bigger

EXTRACTION OF PURE DISTRIBUTIONS

$$p_C(x) = f_C^q p_q(x) + (1 - f_C^q) p_g(x)$$
$$p_F(x) = f_F^q p_q(x) + (1 - f_F^q) p_g(x)$$

Matrix Method

$$\begin{pmatrix} p_F(x) \\ p_C(x) \end{pmatrix} = \underbrace{\begin{pmatrix} f_F^q & 1 - f_F^q \\ f_C^q & 1 - f_C^q \end{pmatrix}}_{\equiv F} \begin{pmatrix} p_q(x) \\ p_g(x) \end{pmatrix}$$

$$\begin{pmatrix} p_q(x) \\ p_g(x) \end{pmatrix} = F^{-1} \begin{pmatrix} p_F(x) \\ p_C(x) \end{pmatrix}$$

Jet Topics

$$p_q(x) = \frac{p_F(x) - \kappa(F, C) p_C(x)}{1 - \kappa(F, C)}$$

$$p_g(x) = \frac{p_C(x) - \kappa(C, F) p_F(x)}{1 - \kappa(C, F)}$$

$$\kappa(F, C) = \min_x \frac{p_F(x)}{p_C(x)} \quad \kappa(C, F) = \min_x \frac{p_C(x)}{p_F(x)}$$

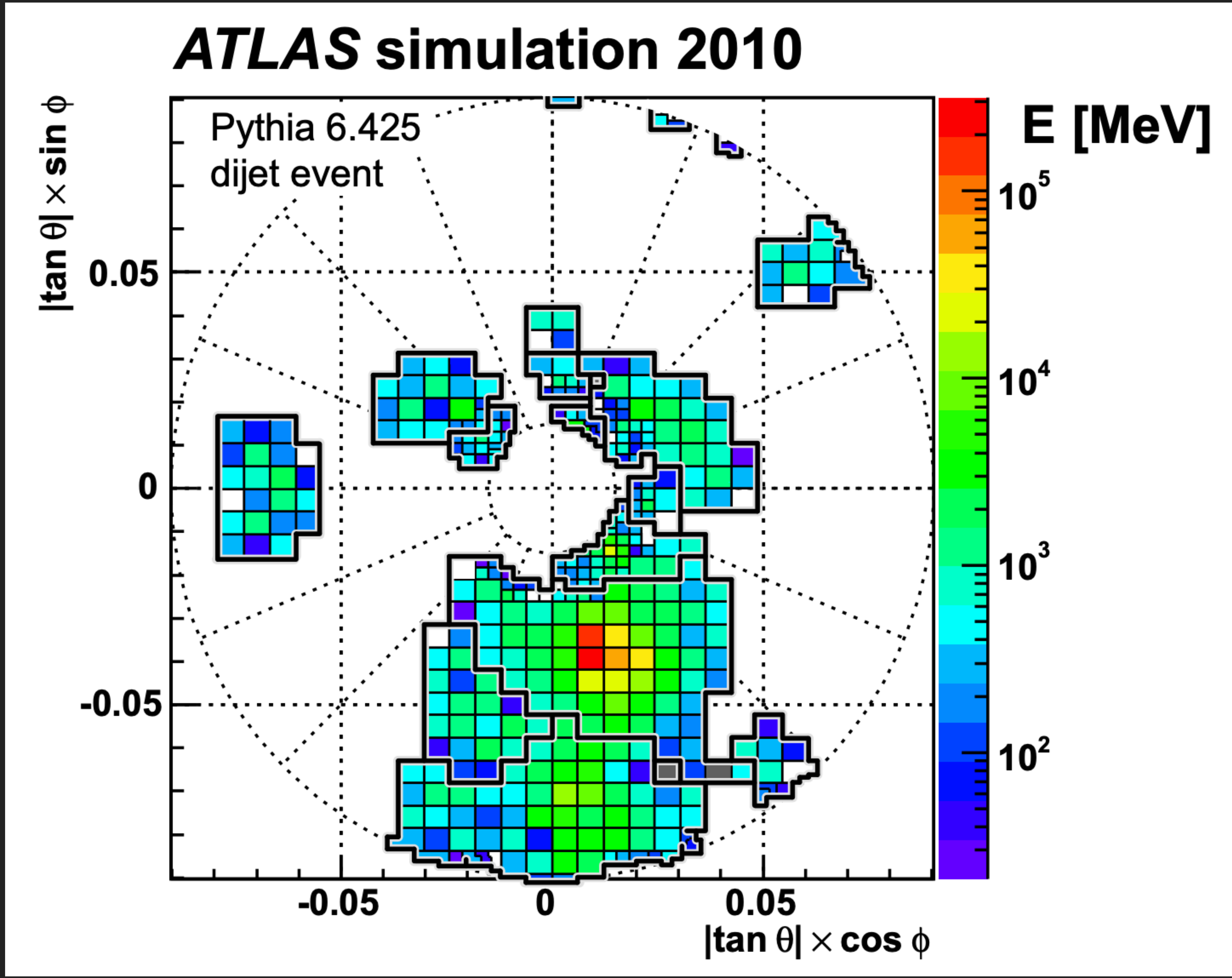
EXTRACTION OF PURE DISTRIBUTIONS

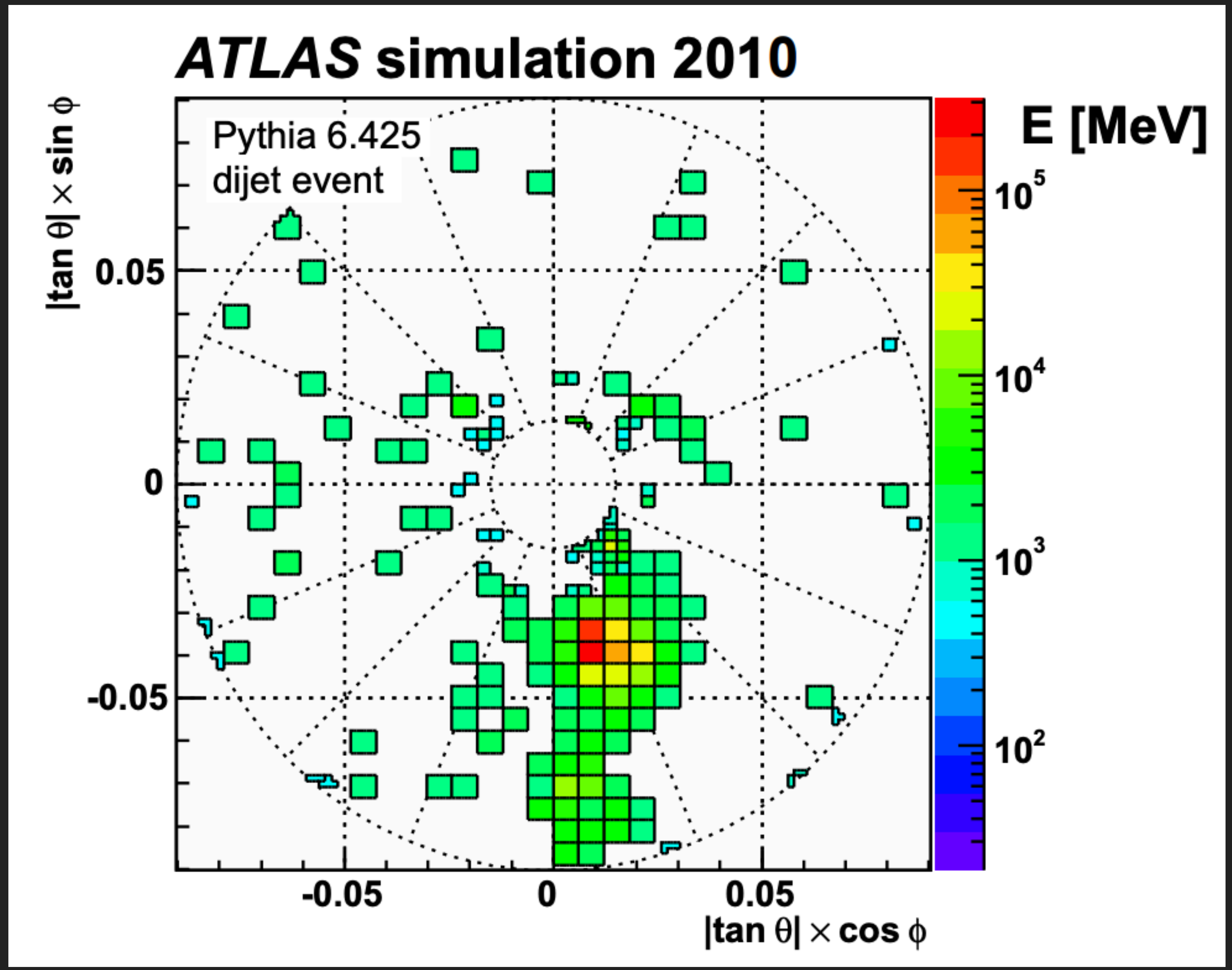
- mutual irreducibility not satisfied $\kappa(q,g) \neq 0, \kappa(g,q) \neq 0$ –> MC correction

$$p_q(x) = \frac{p_{q|g}(x)(1-\kappa(q,g)) + (1-\kappa(g,q))\kappa(q,g)p_{g|q}(x)}{1-\kappa(g,q)\kappa(q,g)}$$

$$p_g(x) = \frac{p_{g|q}(x)(1-\kappa(g,q)) + (1-\kappa(q,g))\kappa(g,q)p_{q|g}(x)}{1-\kappa(q,g)\kappa(g,q)}$$

TOWERS VS CLUSTERS





TRAINING REQUIREMENTS

1. p_{T} independence
2. η independence

→ train on 2D flat distribution

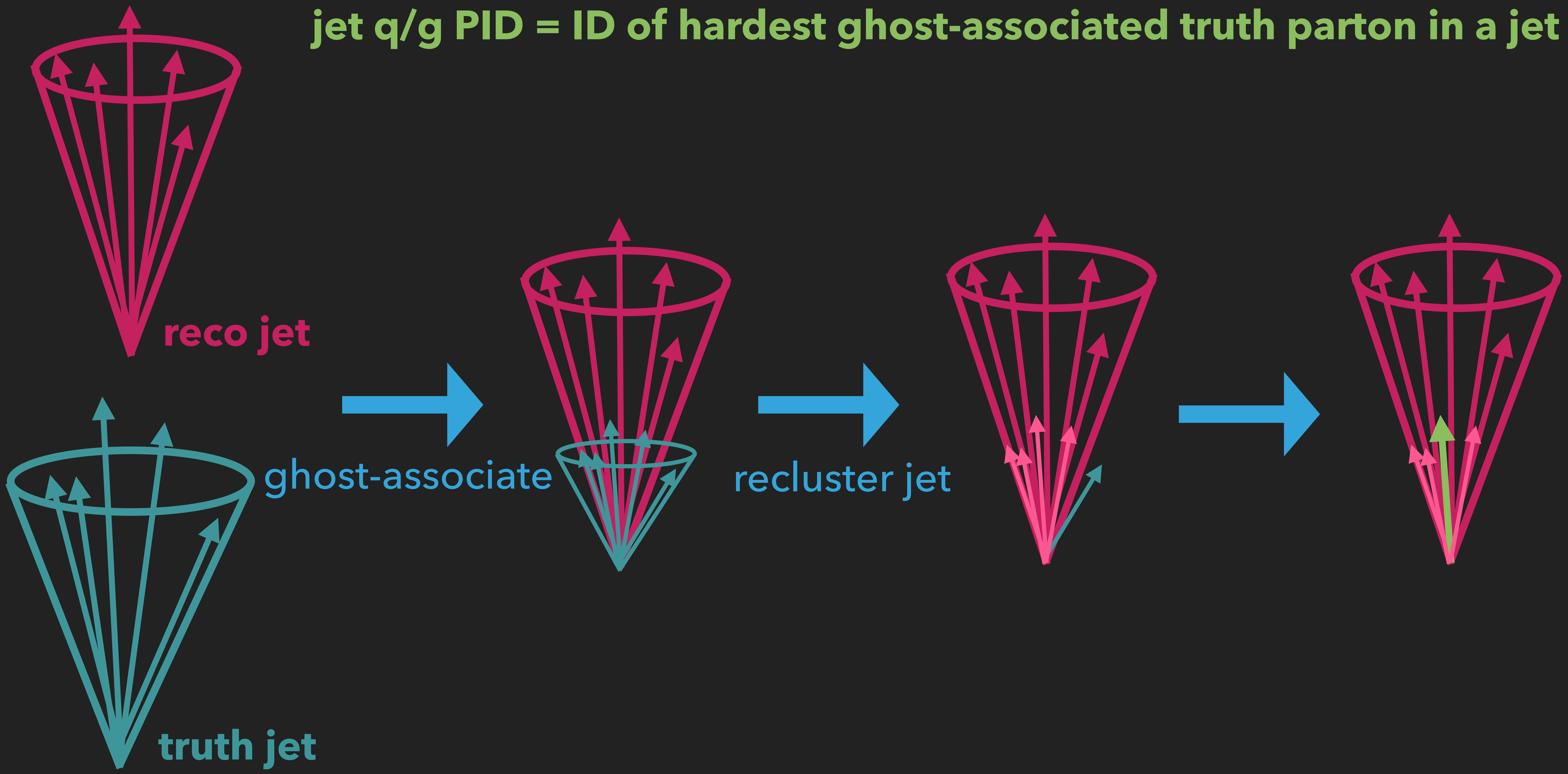
training MC amount limitation

→
1. flattening weights
2. union of 3 2D flat regions

Index	p_{T} range	η range	Total size	Relative size	Num. of η bins	Num. of p_{T} bins
1.	$20 < p_{\mathrm{T}} < 160$ GeV	$\lvert\eta\rvert < 4.5$	50M	49 %	10	10
2.	$160 < p_{\mathrm{T}} < 1300$ GeV	$\lvert\eta\rvert < 2.1$	50M	49 %	6	10
3.	$1300 < p_{\mathrm{T}} < 2000$ GeV	$\lvert\eta\rvert < 1.2$	2.1M	2 %	4	10

PARTON TRUTH LABEL ID

jet q/g PID = ID of hardest ghost-associated truth parton in a jet

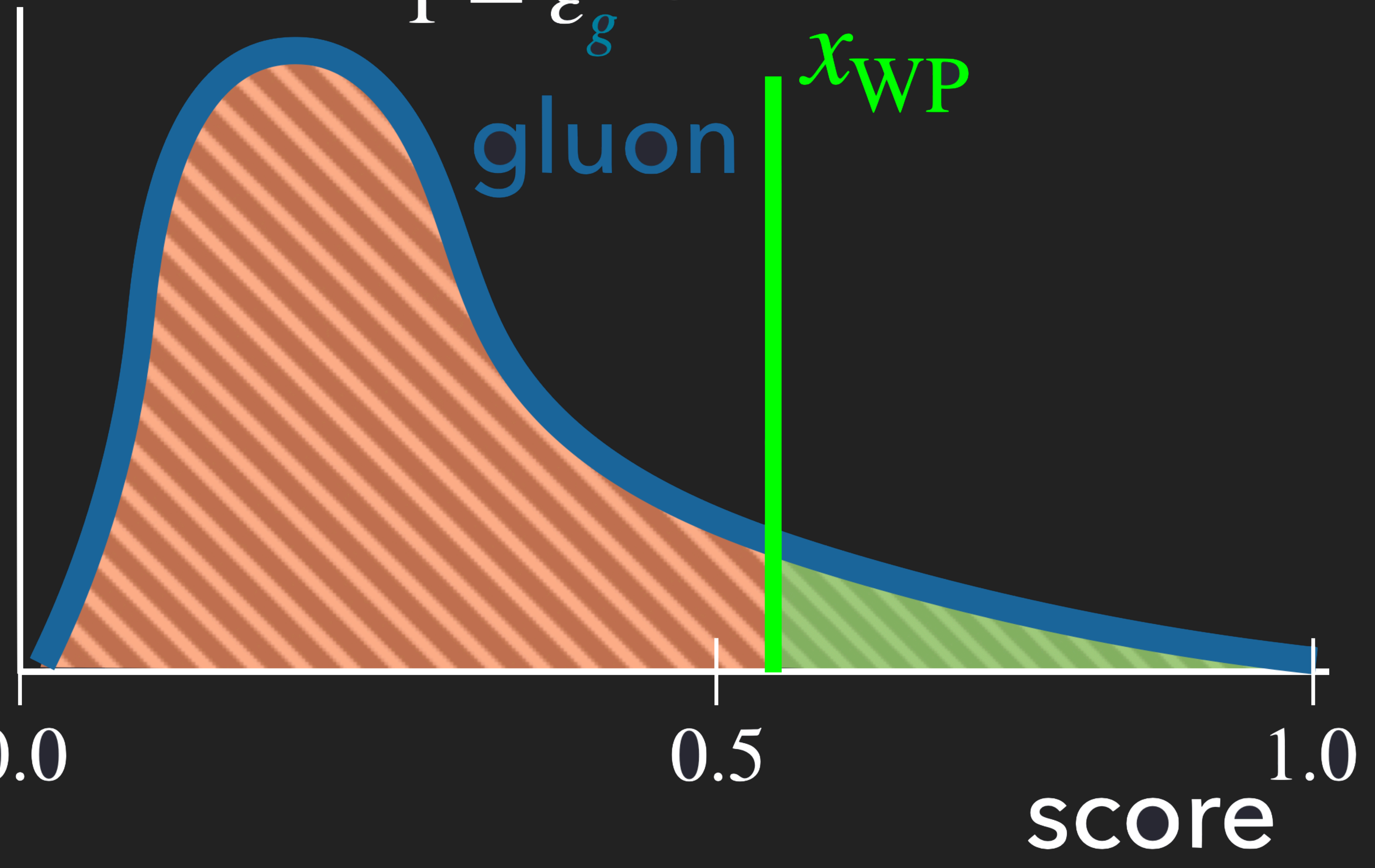


TAGGING EFFICIENCY MEASUREMENT STRATEGY

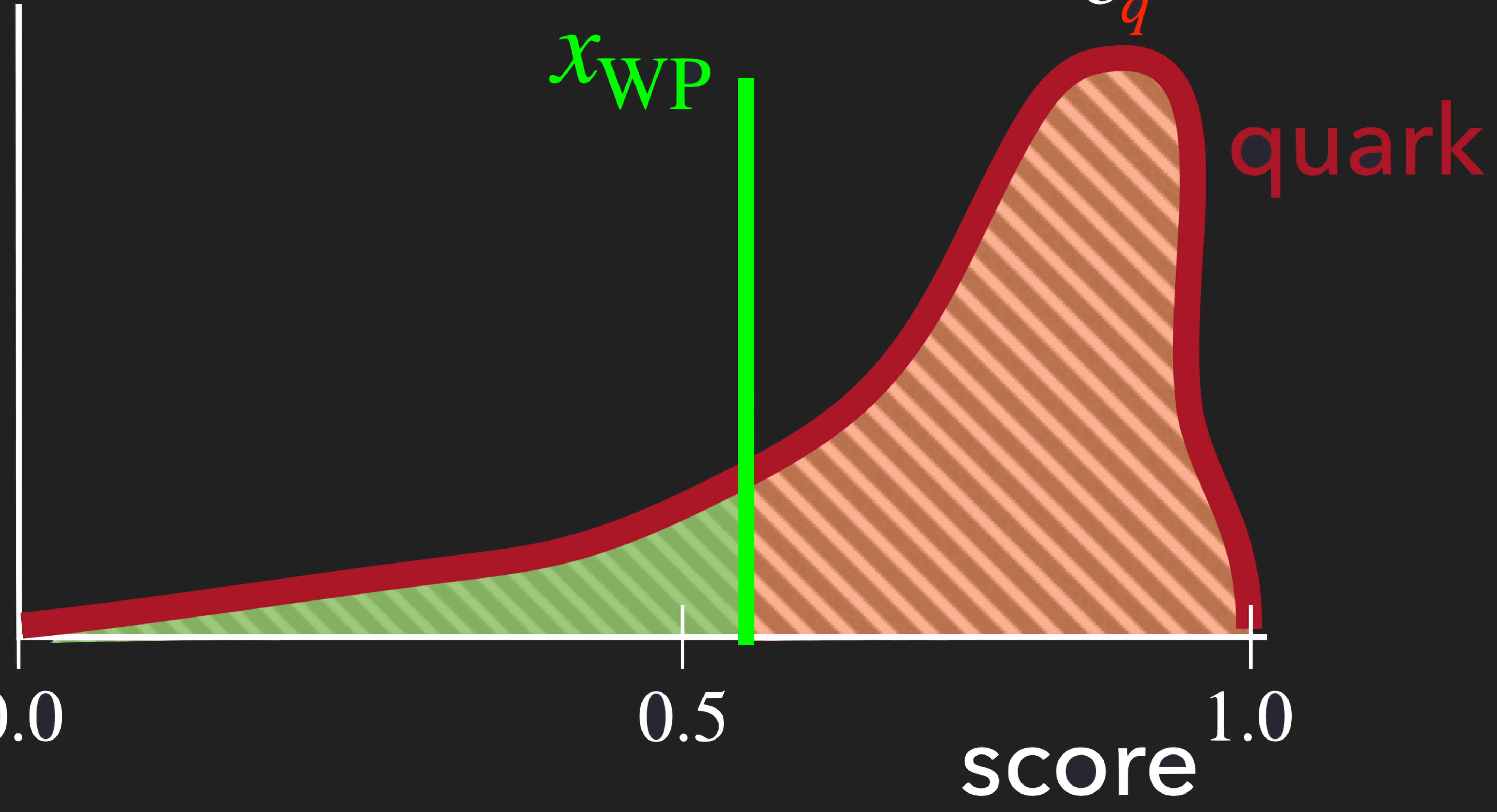
$$\varepsilon_q(x_{\mathrm{WP}}) = \int_{x_{\mathrm{WP}}}^{1} p_q(x)\, dx$$

$$\varepsilon_g(x_{\mathrm{WP}}) = \int_{x_{\mathrm{WP}}}^{1} p_g(x)\, dx$$

$$\mathrm{effSF}_g = \frac{1 - \varepsilon_g^{\mathrm{Data}}}{1 - \varepsilon_g^{\mathrm{MC}}}$$



$$\mathrm{effSF}_q = \frac{\varepsilon_q^{\mathrm{Data}}}{\varepsilon_q^{\mathrm{MC}}}$$



REWEIGHTING

$$p_F(x) = f_F^q p_q(x) + (1 - f_F^q) p_g(x) \qquad p_C(x) = f_C^q p_q(x) + (1 - f_C^q) p_g(x)$$

differences fwd/cntrl for quark and gluons due to detector –> reweight

0th order unfolding

$$p_C(x) \to w^q(x) \cdot p_C(x) = \frac{p_F^q(x)}{p_C^q(x)} \cdot p_C(x)$$

- $p_q(x)$ / $p_g(x)$ can be mathematically extracted from $p_C(x)$ & $p_F(x)$:

$$p_q(x) = \frac{p_F(x) - \kappa(F,C)p_C(x)}{1-\kappa(F,C)} \qquad p_g(x) = \frac{p_C(x) - \kappa(C,F)p_F(x)}{1-\kappa(C,F)}$$

- reducibility factor κ = largest substracted amount possible

$$p_F(x) - \kappa(F,C)p_C(x) \geq 0 \qquad \kappa(F,C) = \min_x \frac{p_F(x)}{p_C(x)}$$

QUANTILE REBINNING, ANCHOR BIN QUANTILED

combine bins to reduce statistical uncertainty

try to obtain an equal amount of counts in each bin

vary the bin size

each bin is characterized with cumulative sum fraction

use the formula for κ directly on rebinned ditributions

hyperparam. - num of quantile bins, utilize multiple bins

$$\kappa(F, C) = \exp\left(- \max_q \ln \frac{p_F(q)}{p_C(q)} \right) \qquad q = \frac{\int_0^{x_q} p_F(x) + p_C(x)\ dx}{\int_0^1 p_F(x) + p_C(x)\ dx}$$

CONDITIONS

1. ***sample independence*** – $p_q(x)$ and $p_g(x)$ are same in all mixtures

2. ***different purities*** – $f_C^q \neq f_F^q$

same with Matrix Method

3. ***mutual irreducibility*** – existence of pure quark and pure gluon phase space regions.

This implies:

$$\kappa(q,g) = \min_x \frac{p_q(x)}{p_g(x)} = \kappa(g,q) = \min_x \frac{p_g(x)}{p_q(x)} = 0$$

MUTUAL IRREDUCIBILITY

- mutual irreducibility not satisfied in MC

- if we know $\kappa(q, g)$ and $\kappa(g, q)$ than pure quark and gluon distributions are

$$p_q(x) = \frac{p_{q|g}(x)(1 - \kappa(q, g)) + (1 - \kappa(g, q))\kappa(q, g)p_{g|q}(x)}{1 - \kappa(g, q)\kappa(q, g)}$$

$$p_g(x) = \frac{p_{g|q}(x)(1 - \kappa(g, q)) + (1 - \kappa(q, g))\kappa(g, q)p_{q|g}(x)}{1 - \kappa(q, g)\kappa(g, q)}$$

- the $\kappa(q, g)$ and $\kappa(g, q)$ are given from MC and:

$$\kappa(q, g) = 0 \longrightarrow p_q(x) = \frac{p_F(x) - \kappa(F, C)p_C(x)}{1 - \kappa(F, C)}$$

$$\kappa(q, g) \neq 0 \longrightarrow p_{q|g}(x) = \frac{p_F(x) - \kappa(F, C)p_C(x)}{1 - \kappa(F, C)}$$

TRAINING HYPERPARAMETERS

Index	p_T range	η range	Total size	Relative size	Num. of η bins	Num. of p_T bins
1.	$20 < p_T < 160$ GeV	$\lvert\eta\rvert < 4.5$	50M	49 %	10	10
2.	$160 < p_T < 1300$ GeV	$\lvert\eta\rvert < 2.1$	50M	49 %	6	10
3.	$1300 < p_T < 2000$ GeV	$\lvert\eta\rvert < 1.2$	2.1M	2 %	4	10

2.62M parameters

Parameter	**DeParT**
Embedding Dimension	128
Self-Attention Block Layers	11
Class Attention Block Layers	2
Heads	8
Expansion	4
Dropout	0.1
Stochastic Depth Drop Rate	0.2
Layer Scale Initialization	$5 \cdot 10^{-3}$
Number of Embedding Layers	3
Size of Embedding Layers	128
Number of Interaction Embedding Layers	3
Size of Interaction Embedding Layers	64
Activation	GELU

Data	
Training Dataset	102M jets
Validation Dataset	3M jets
Number of Epochs (*)	10
Batch Size (*)	1024
Normalization Layer	adaptation on 150 batches
Optimizer	
Optimizer	Adam
Learning Rate (*)	0.0001
β_1	0.9
β_2	0.999
ϵ	10^{-6}
Learning Rate Scheduling	Cosine Decay
Minimum Learning Rate	10^{-6}
Warmup (*)	Linear (20k steps)
Clip Norm	0.8
Loss	
Loss	Binary Cross Entropy
Label Smoothing	None
Weight Decay	None