Run 3 performance and advances in heavy flavor jet tagging in CMS

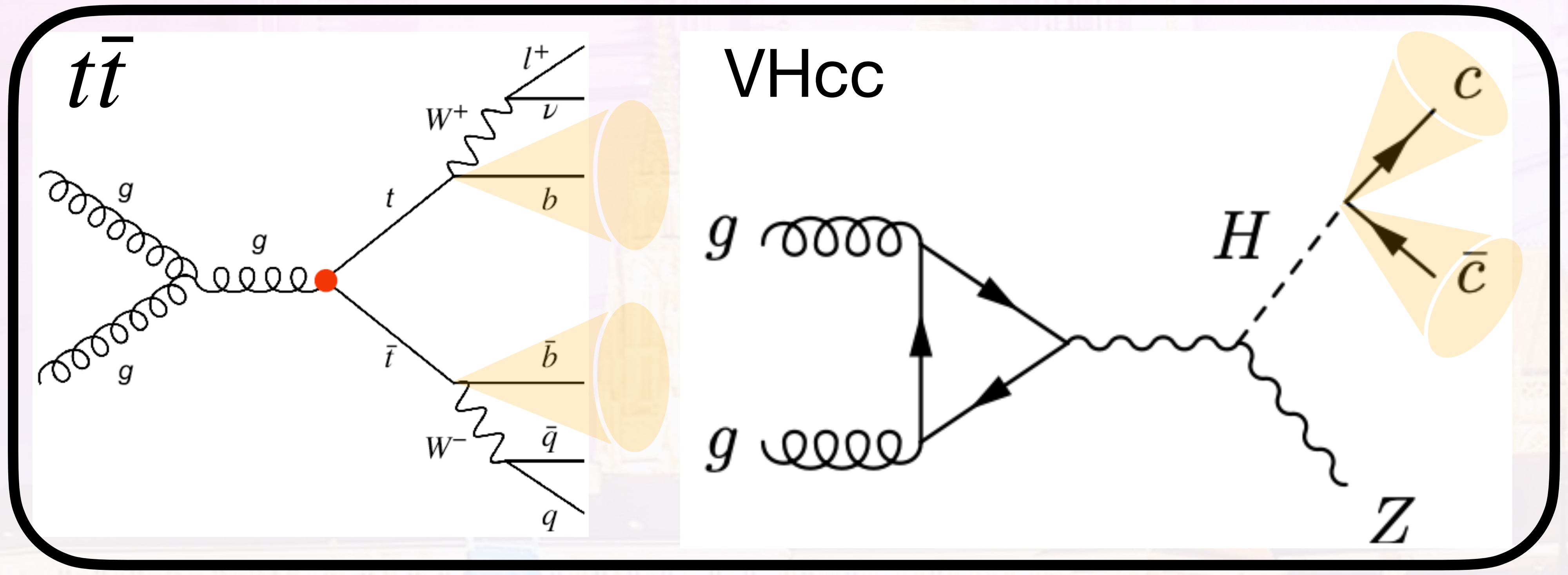
Uttiya Sarkar

on behalf of the CMS collaboration

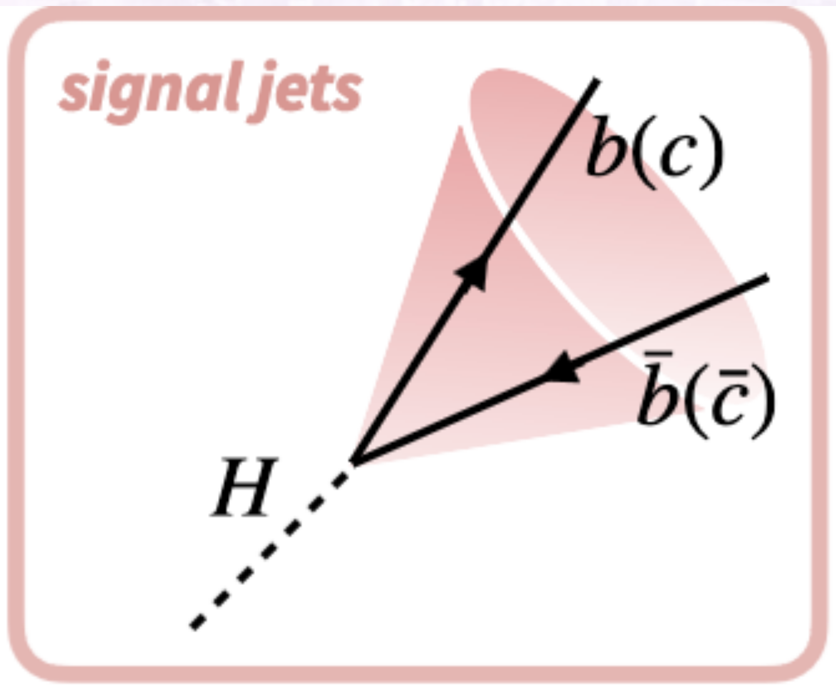
Overview

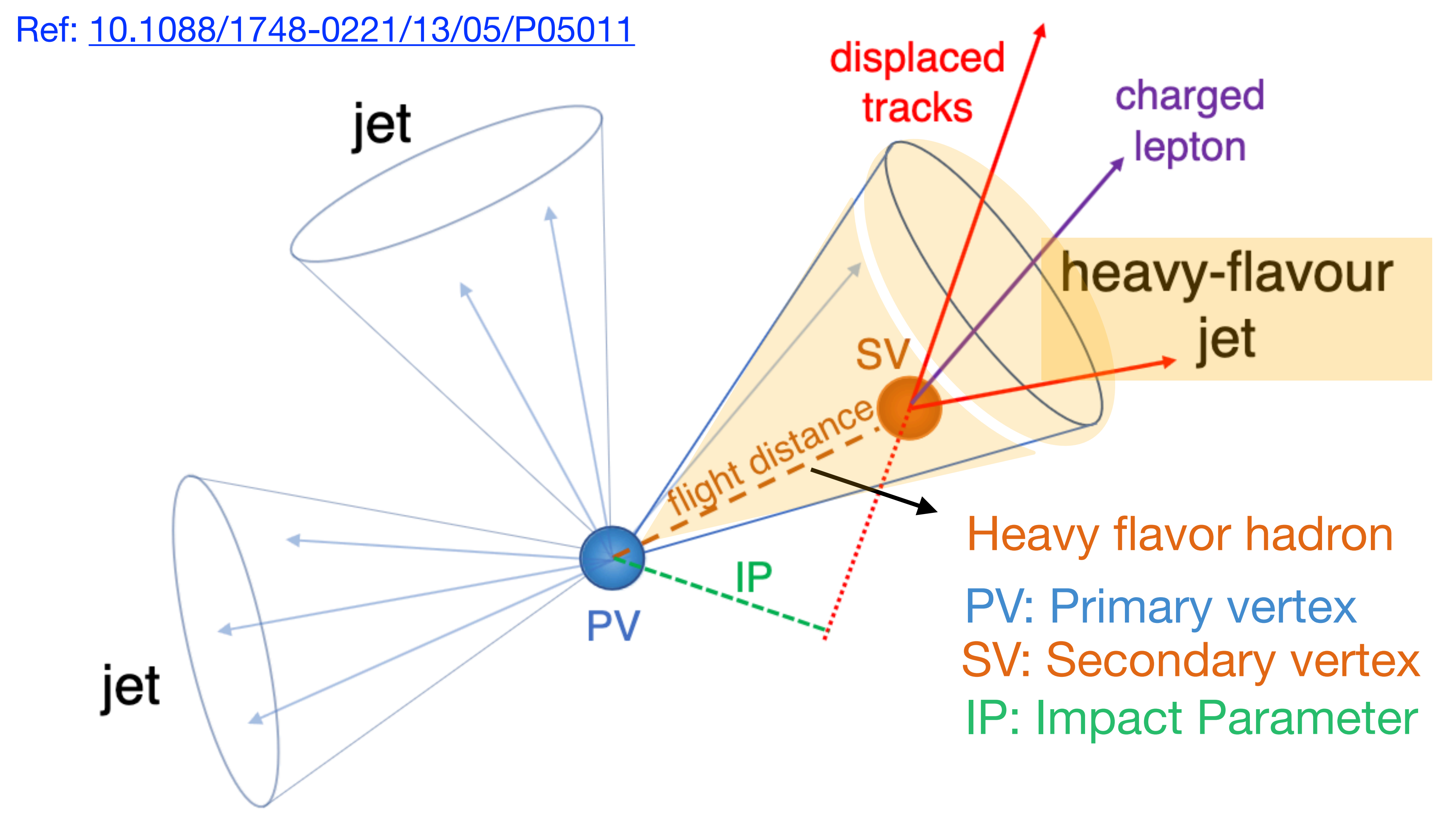
- **Heavy-flavor tagging -** jets originating from b (***b jets***) or c (***c jets***) quarks
- **Physics motivation:** Important in Standard Model (SM), Top, Higgs, BSM and SUSY processes

Resolved jet topology



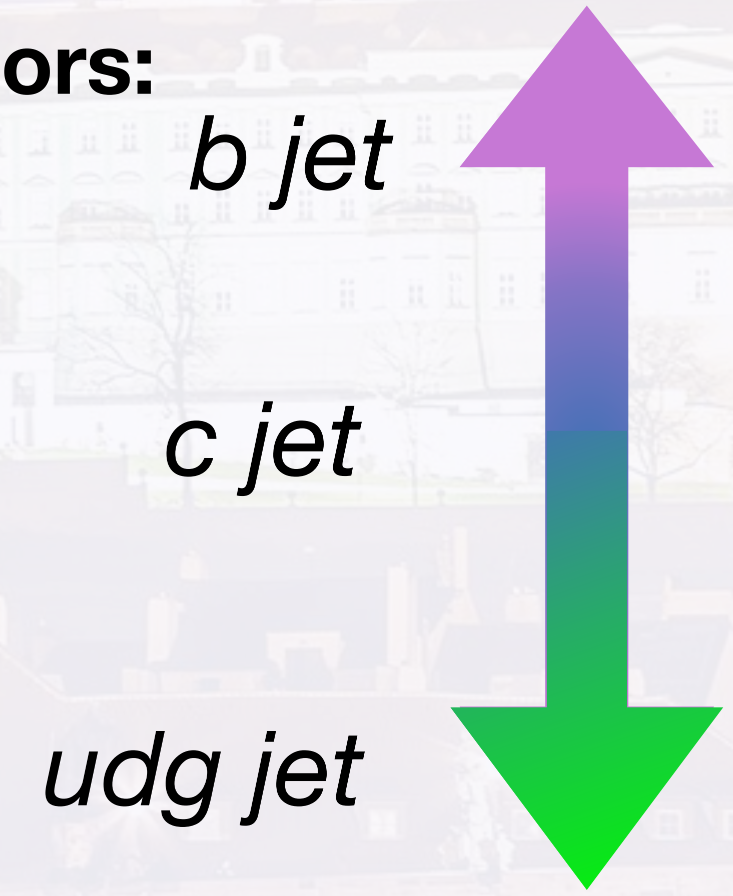
High momentum: merged jet topology




Ref: 10.1088/1748-0221/13/05/P05011

- Within CMS tracker resolution
- **Significant improvement in machine learning based tagging algorithms**

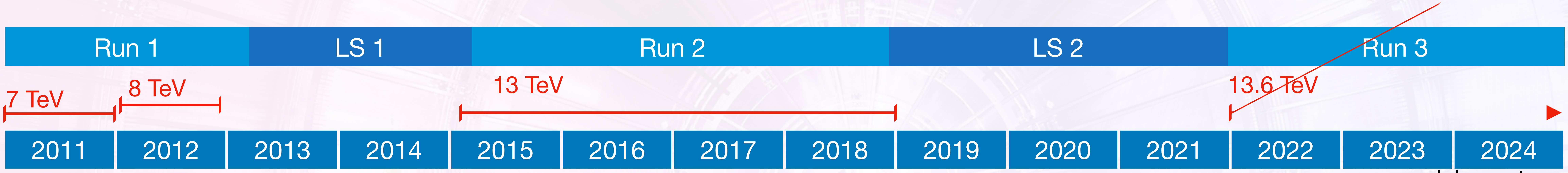
- **Discriminators:**



$$\mathrm{BvsX} = \frac{\mathrm{P(B)}}{\mathrm{P(B)} + \mathrm{P(X)}}$$

Performs beyond the heavy flavor tagging (can tag τ-jets)

Historical evolution of flavor taggers in CMS



combined secondary vertex (CSV)
Likelihood ratio

combined secondary vertex v2 (CSVv2)
MLP

DeepCSV
DNN

DeepJet
CNN + RNN

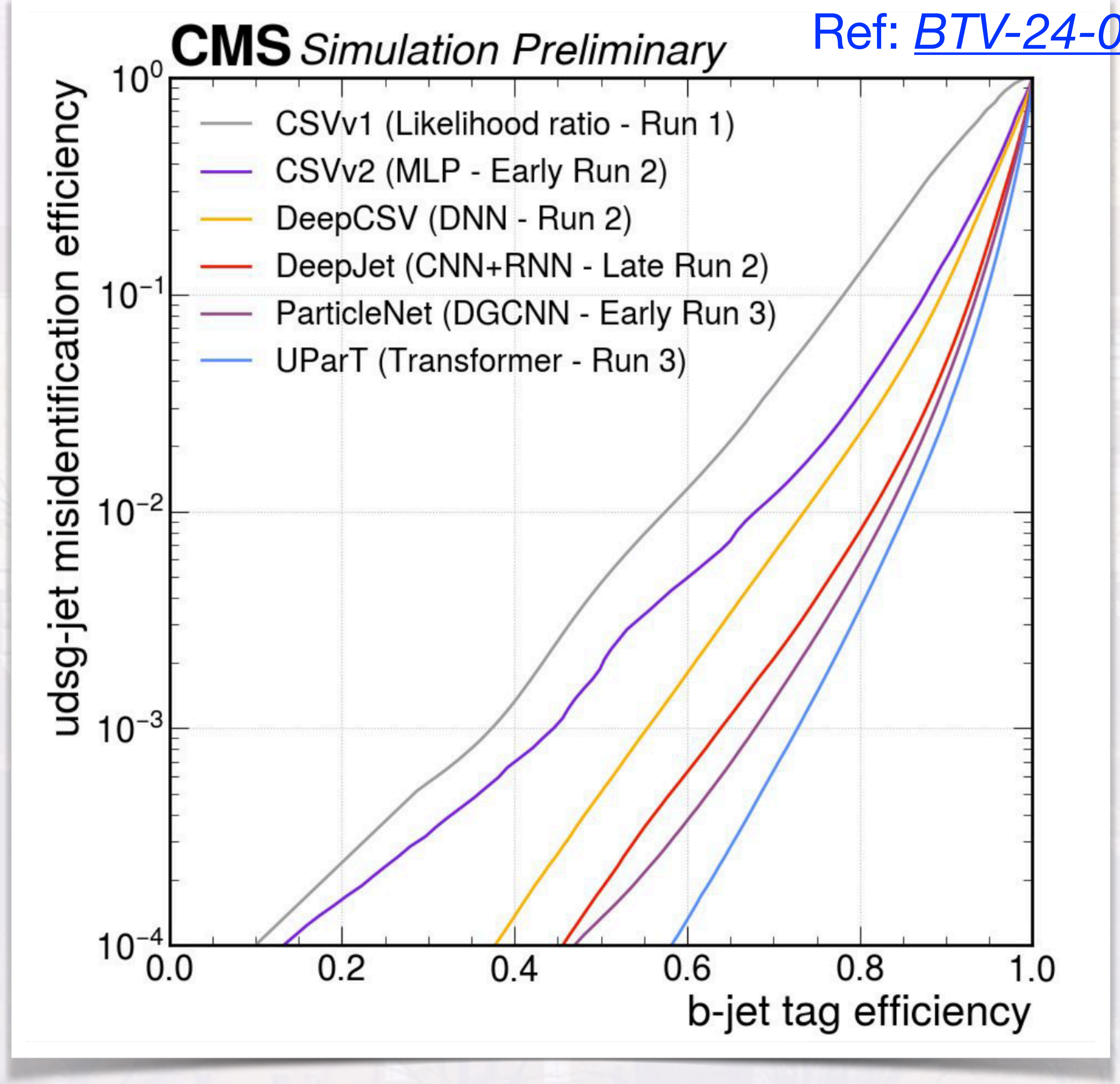
ParticleNet

DGCNN - Treat particles as particle clouds - unordered set of its constituent particle

RobustParT
Multi Head Attention

UParT

Distortion of input features to enhance the robustness

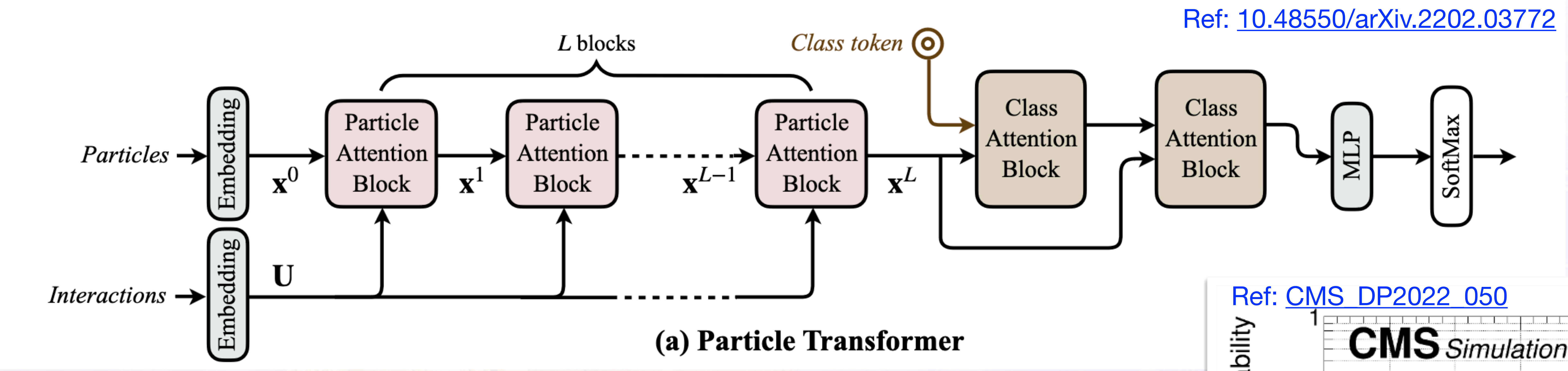


Ref: *BTV-24-066*

Significant improvement in performance over the last decade!

Transformer models: ParticleTransformer

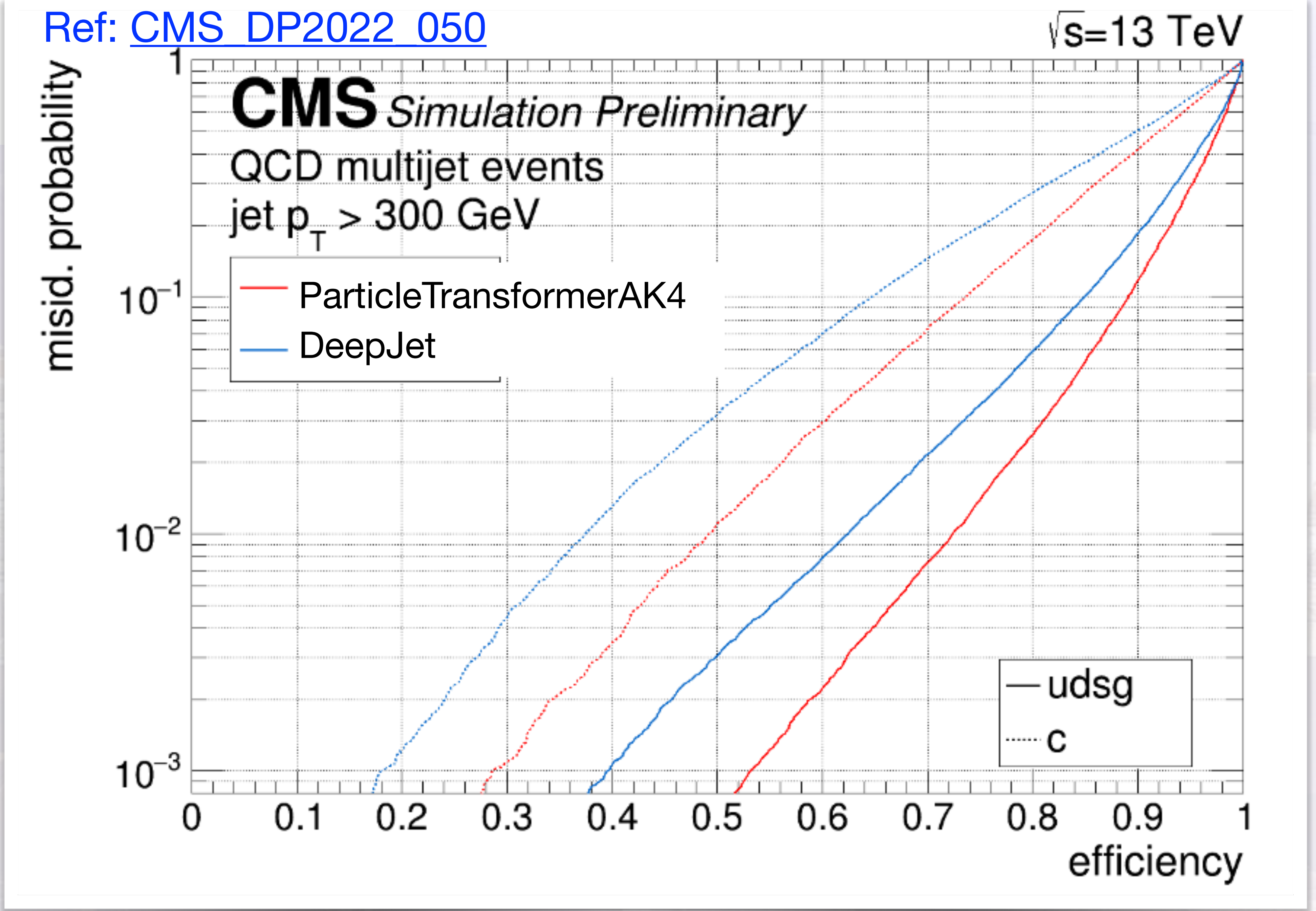
Ref: 10.48550/arXiv.2202.03772 ← See more details here



(a) Particle Transformer

- Based on the “Attention” model designed for particles
- Input embedding:
  - Not only inject single particle information, but also include pair-wise features
- Multi-Head Attention (MHA) Pair-wise feature

Ref: CMS_DP2022_050



Significant improvement as compared to DeepJet!

Transformer model: UParT

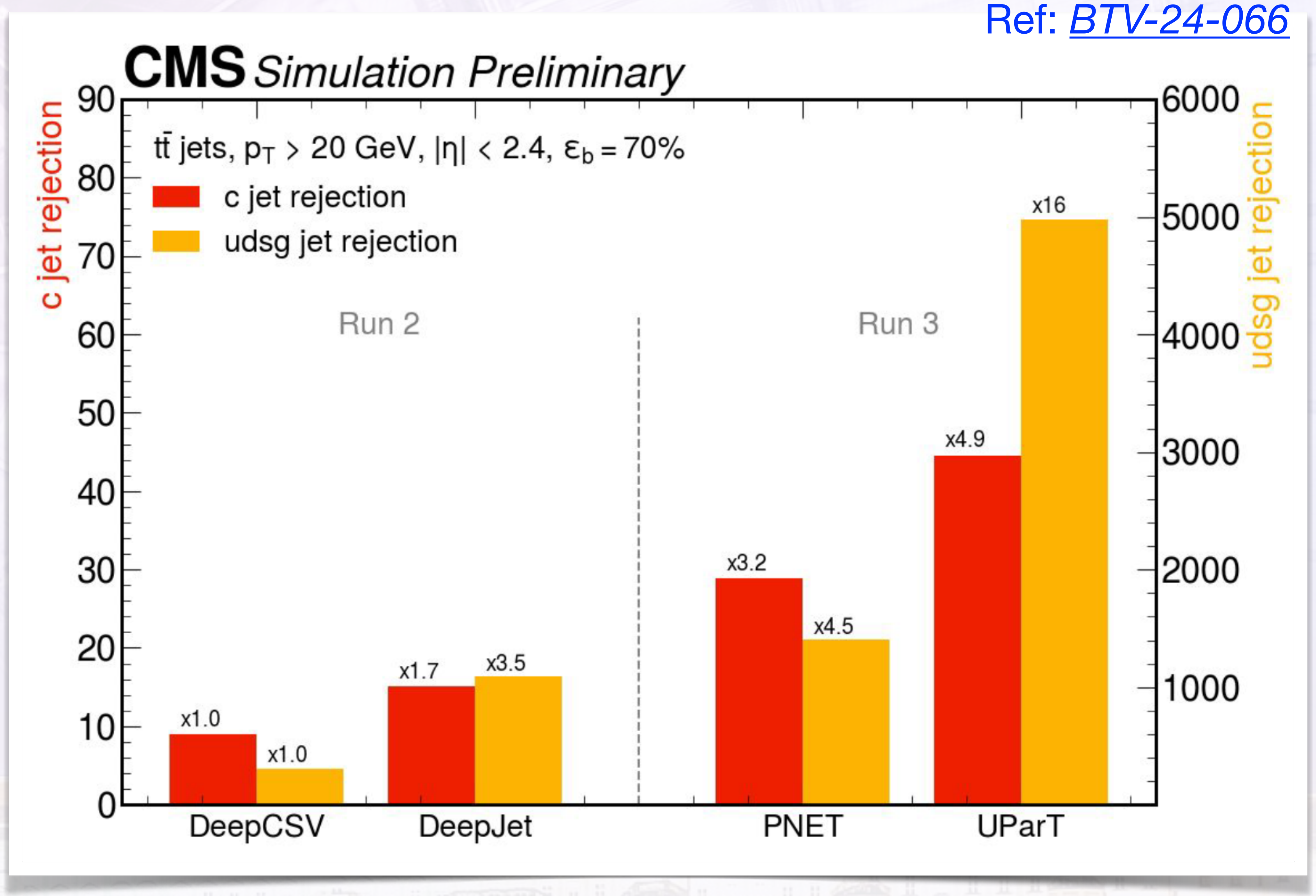
- Extension of ParticleTransformer
  - Extended class: extending from b and c jet identification to s and hadronic tau (one per final state) identification
  - Extended regression: simultaneous flavor aware jet energy and resolution regression

$$L = \overbrace{\mathrm{CatEntropy}(\mathrm{x}, \mathrm{x}_{\mathrm{truth}})}^{\textit{Classification}} + \overbrace{\lambda \times \log(\cosh(y - y_{truth})}^{\textit{Regression}}) + \overbrace{\gamma \times [\rho_{0.16}(\mathrm{z} - \mathrm{z}_{\mathrm{truth}}) + \rho_{0.84}(\mathrm{z} - \mathrm{z}_{\mathrm{truth}})]}^{\textit{Quantile regression (resolution estimation)}}$$

- Input variable distortion:
  - Reduce the observed differences prior to any calibration
  - Improve robustness of the classifier against injected mismodelings

➡Distortions of UParT: Preserving the Particle Cloud representation and the feature importance mapping

Ref: BTV-24-066

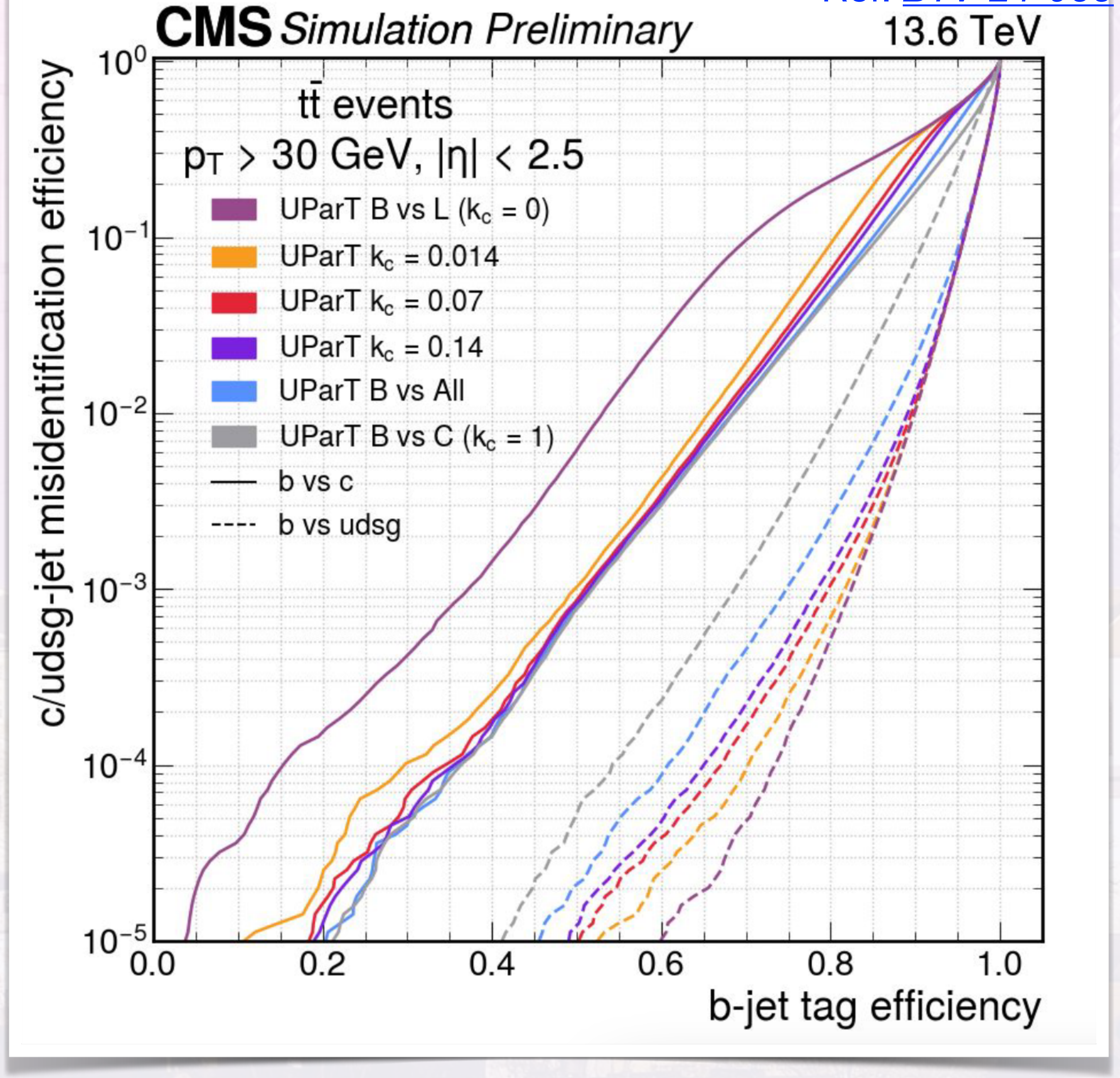


Most performant heavy flavor tagging algorithm so far in CMS!

Flavor tagging Performance: UParT

b-tagging

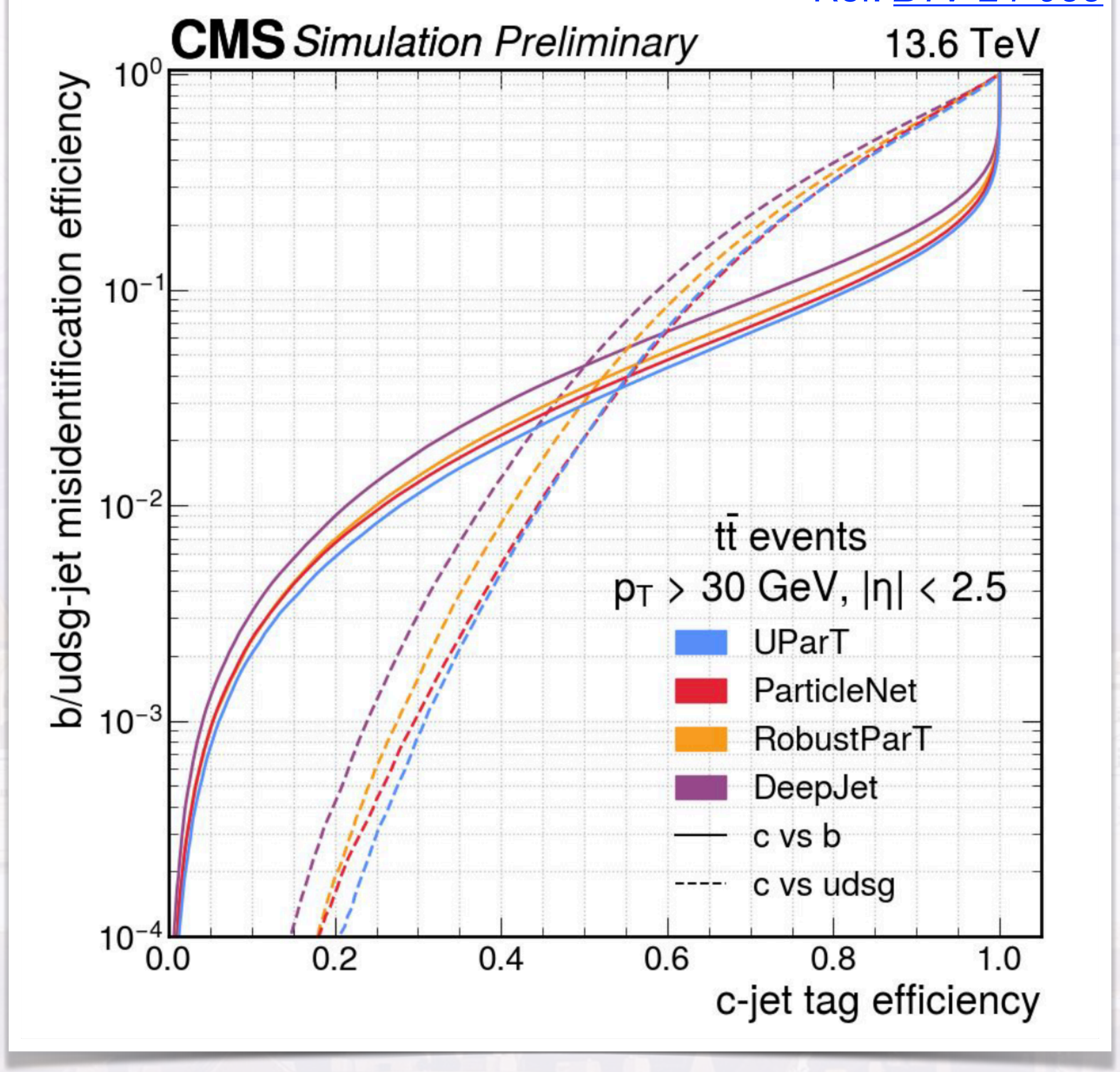
Ref: *BTV-24-066*



Significant improvement in b-tagging efficiency!

c-tagging

Ref: *BTV-24-066*

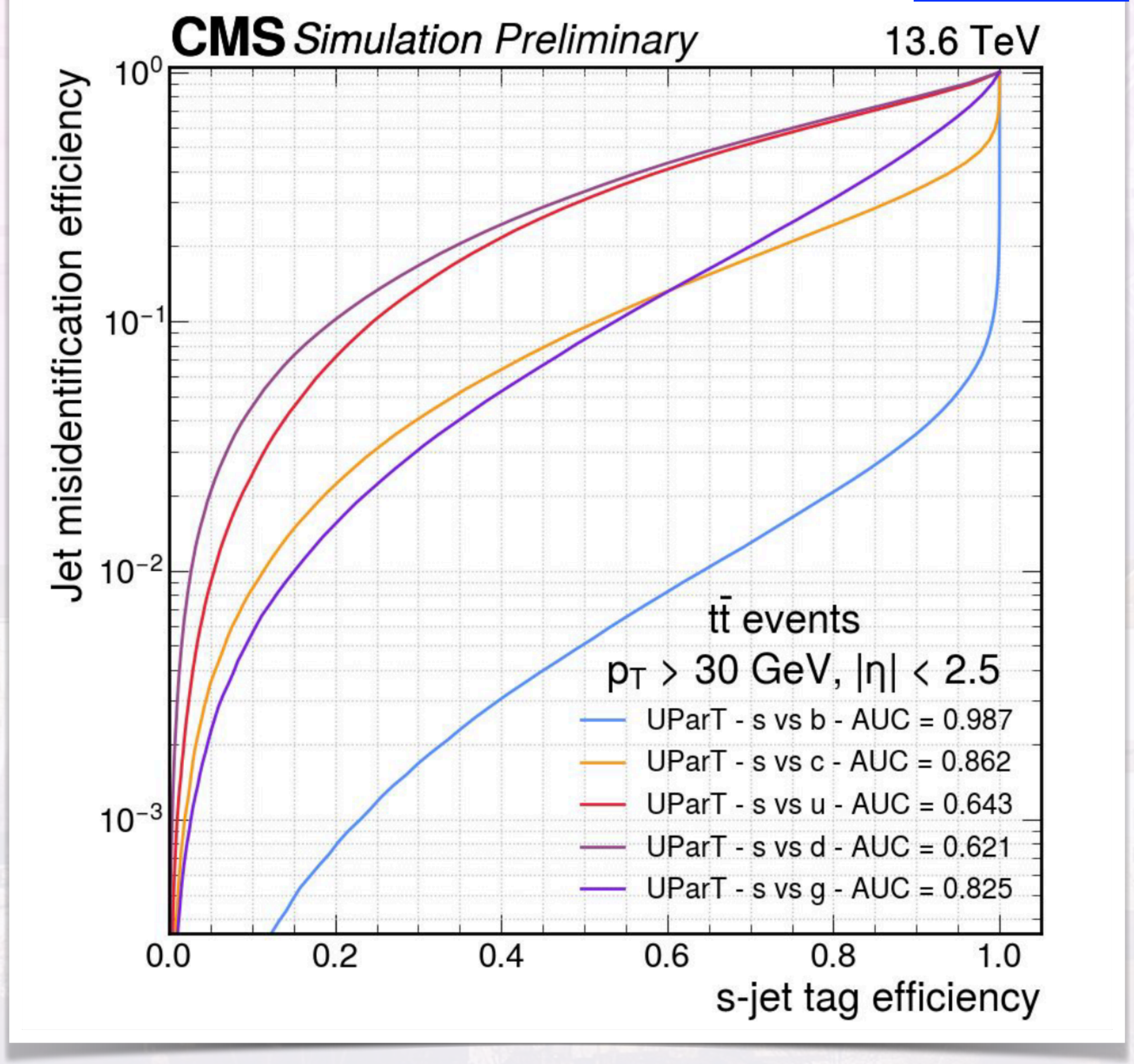


Improvement in c-tagging efficiency and c vs b discrimination

Flavor tagging Performance: UParT

s-tagging

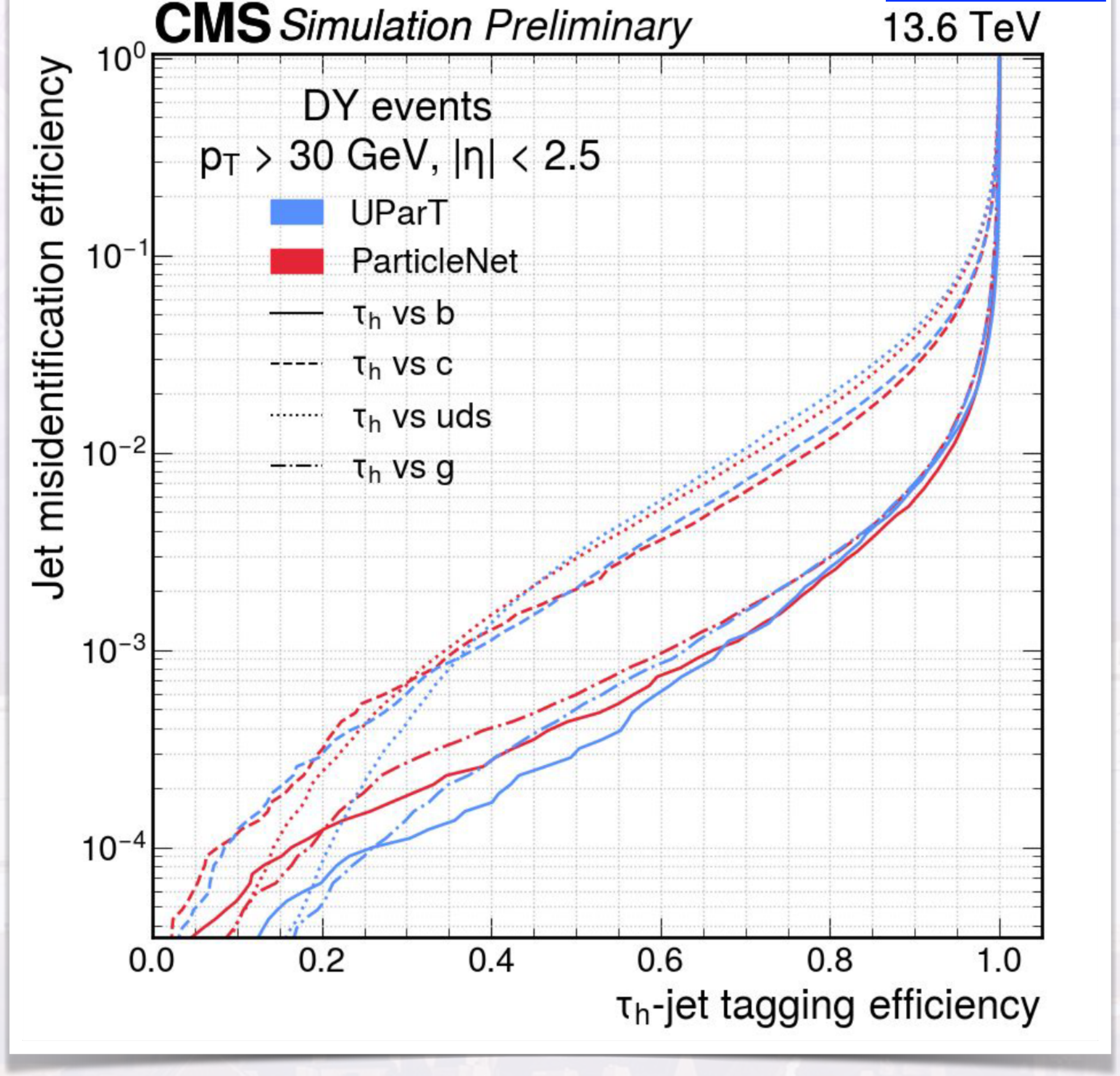
Ref: BTV-24-066



First attempt of s-tagging in CMS!

τ-tagging

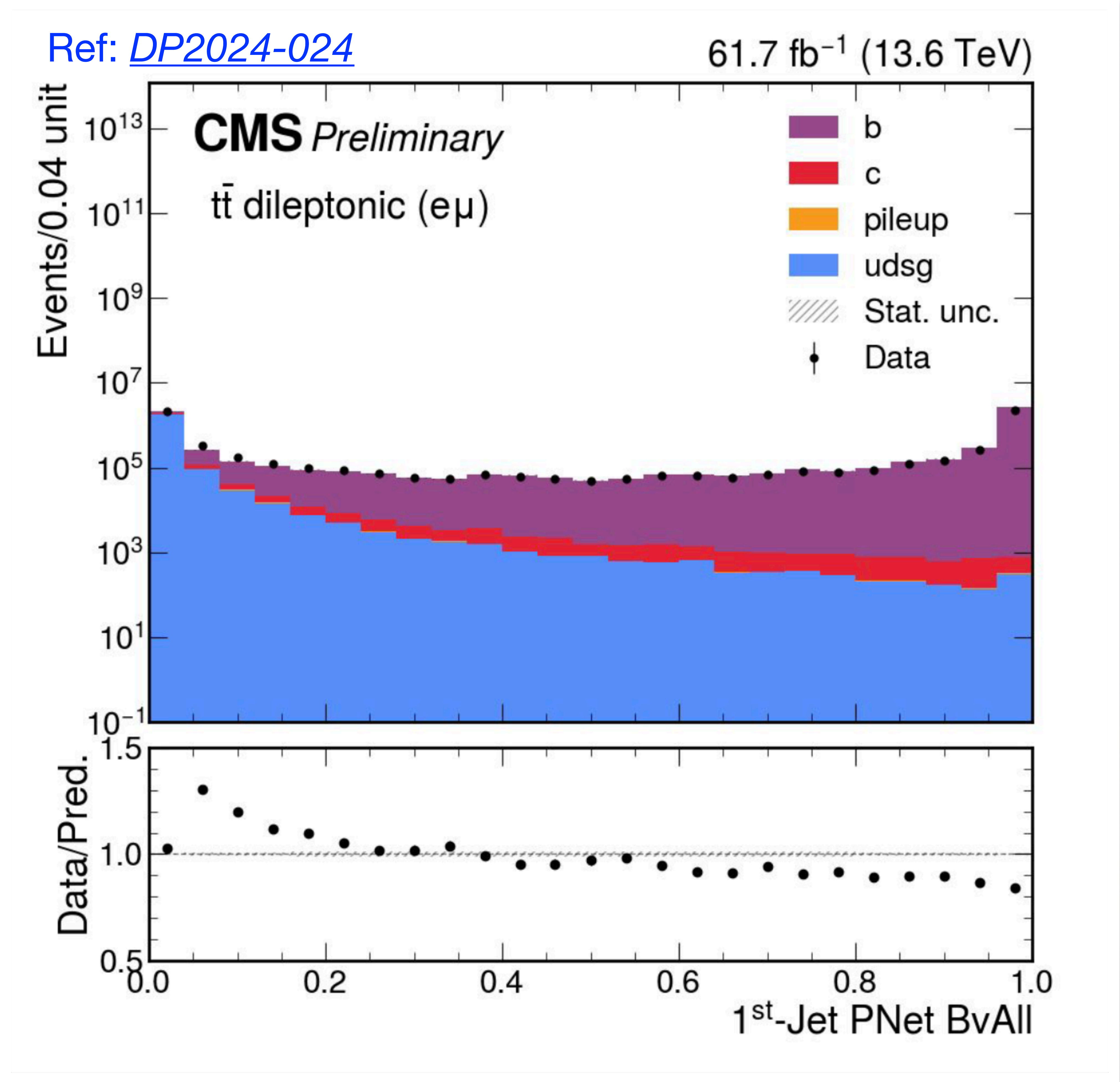
Ref: BTV-24-066



Improvement in τ-tagging performance

Data vs. Simulations Mismodeling

- Observed **differences in Data and Simulations**
  - Imperfect modeling of the input variables affects the modeling of the output discriminator distributions
- Prone to changes in the calibration of the detector alignment
- Different estimation in simulations as compared to data



Ref: DP2024-024

Data/simulation disagreement due to mismodeling in simulations

Requires calibrations

Calibration and Scale Factors

- Define Scale-Factors $SF_f = \epsilon_f^{\text{data}}(p_T, \eta) / \epsilon_f^{\text{sim}}(p_T, \eta)$

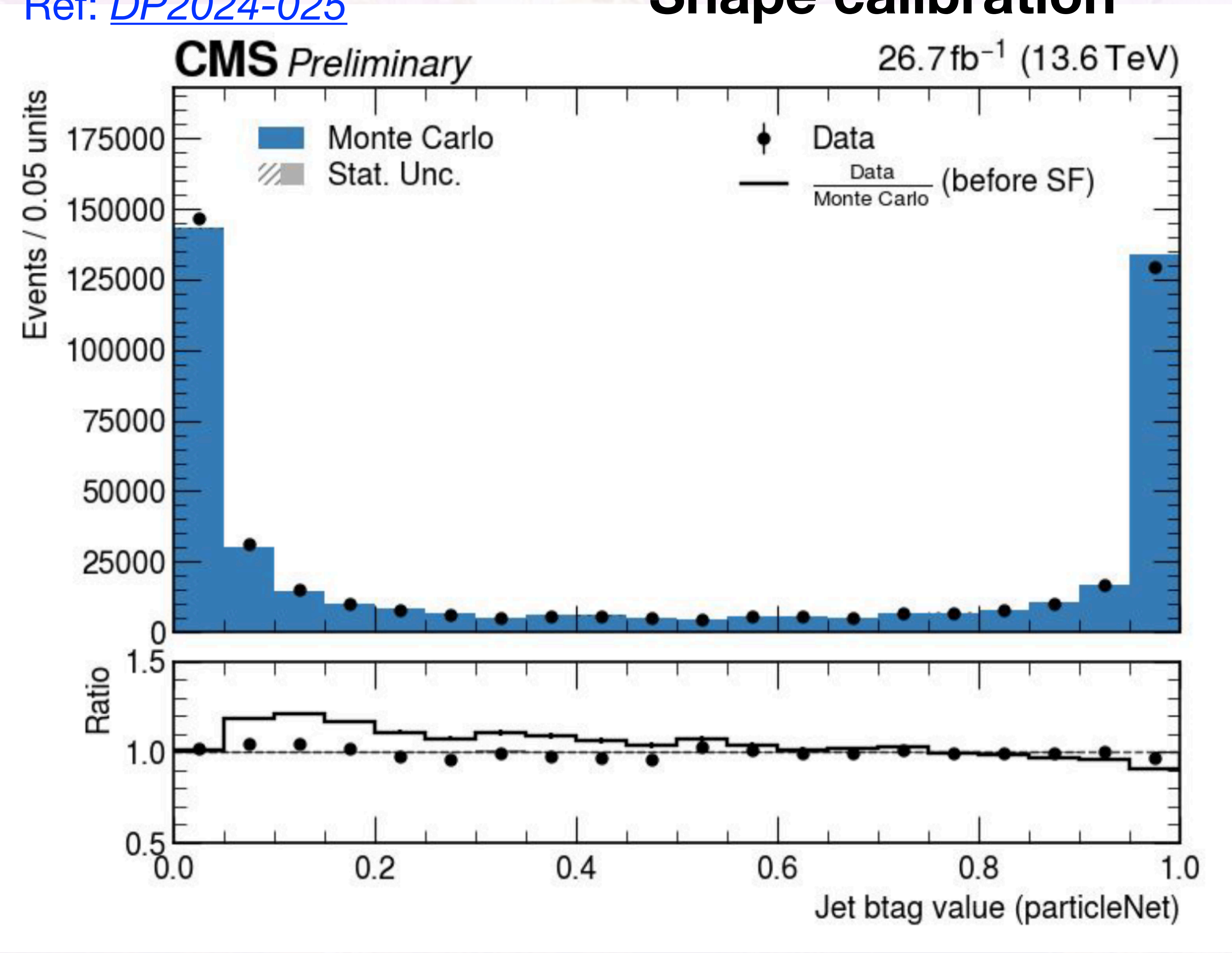
$\epsilon_f^{\text{data}}(p_T, \eta)$ efficiencies for a jet with flavor f in data

$\epsilon_f^{\text{sim}}(p_T, \eta)$ efficiencies for a jet with flavor f in simulation

- Define heavy flavor enriched data sample to derive the SF
- For c-tagging, calibration of full discriminator shape, simultaneously for CvsL and CvsB arXiv:2111.03027

Shape calibration

Ref: *DP2024-025*

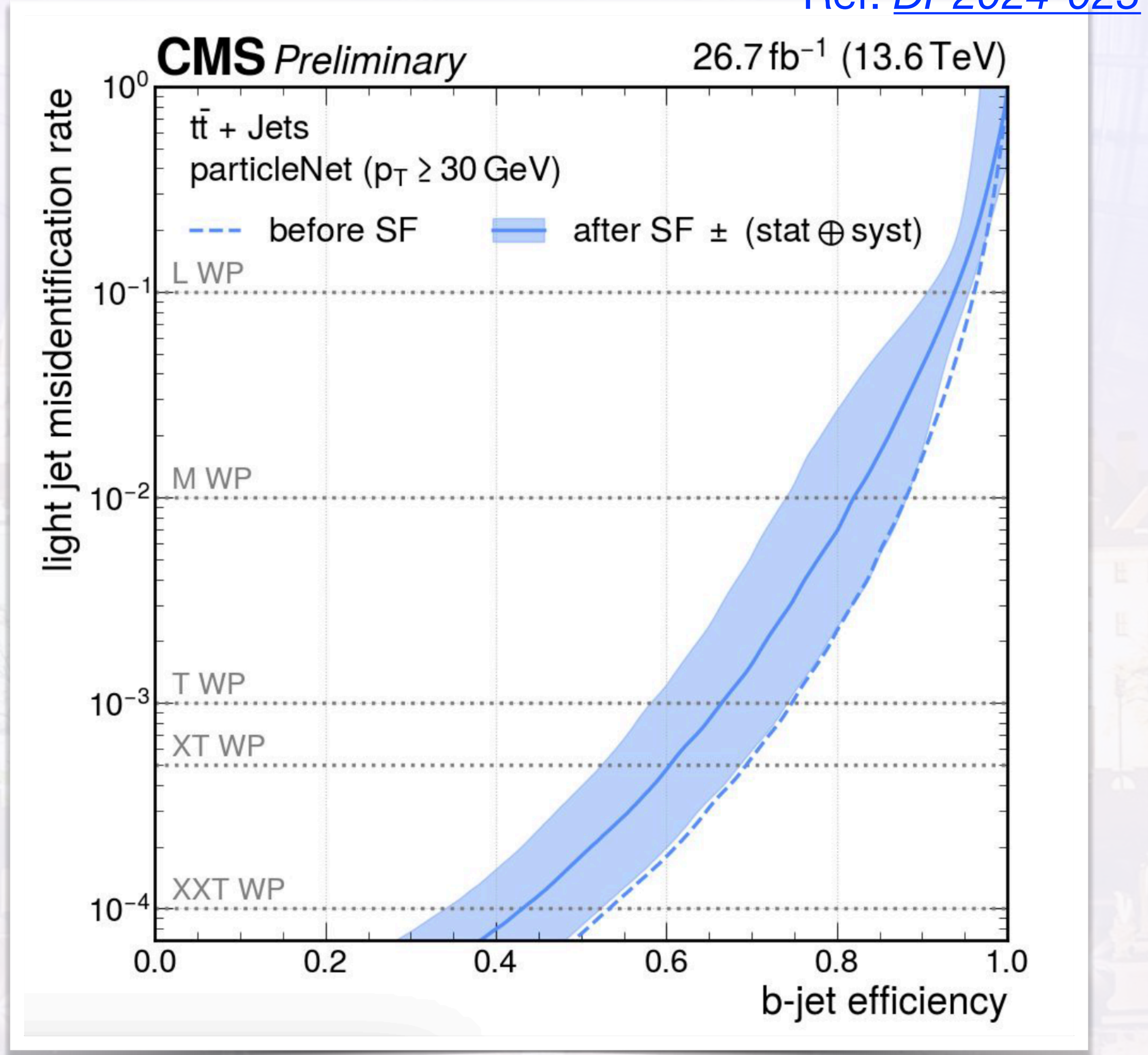


Good Data vs. Simulations closure after applying the SFs

First Run-3 SFs derived from 2022 data

Working-point calibration

Ref: *DP2024-025*



Change in performance before and after applying the SFs

Boosted Object tagging

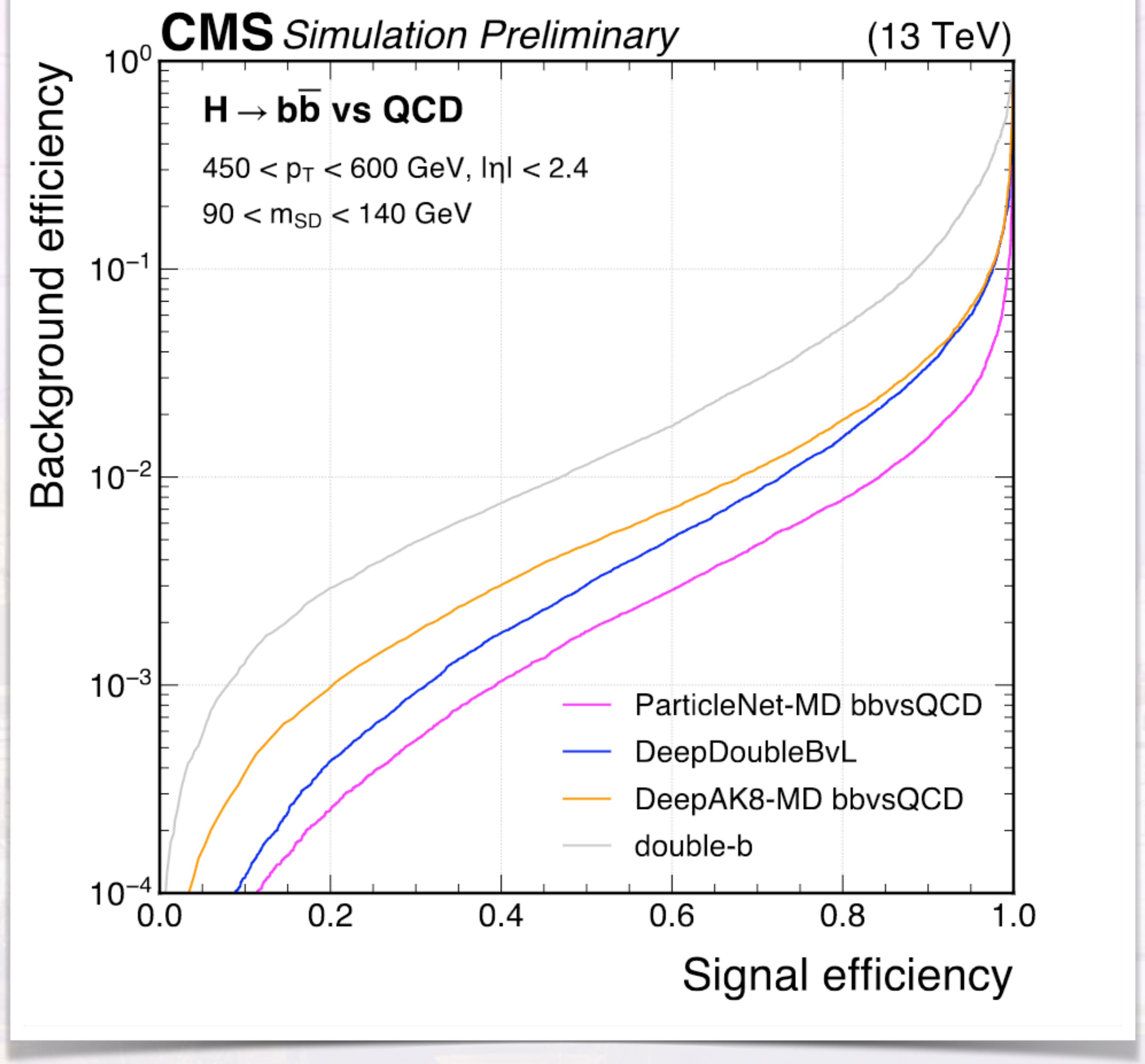
- Merged jet analyses (H→bb, H→cc) can leverage from the innovative tagging techniques of boosted objects

- Boosted-jet tagging algorithms:
  - double-b ***DNN***
  - DeepAK8MD ***CNN***
  - DeepDoubleX ***CNN, RNN***
  - ParticleNetMD ***DGCNN***

MD = Mass decorrelated

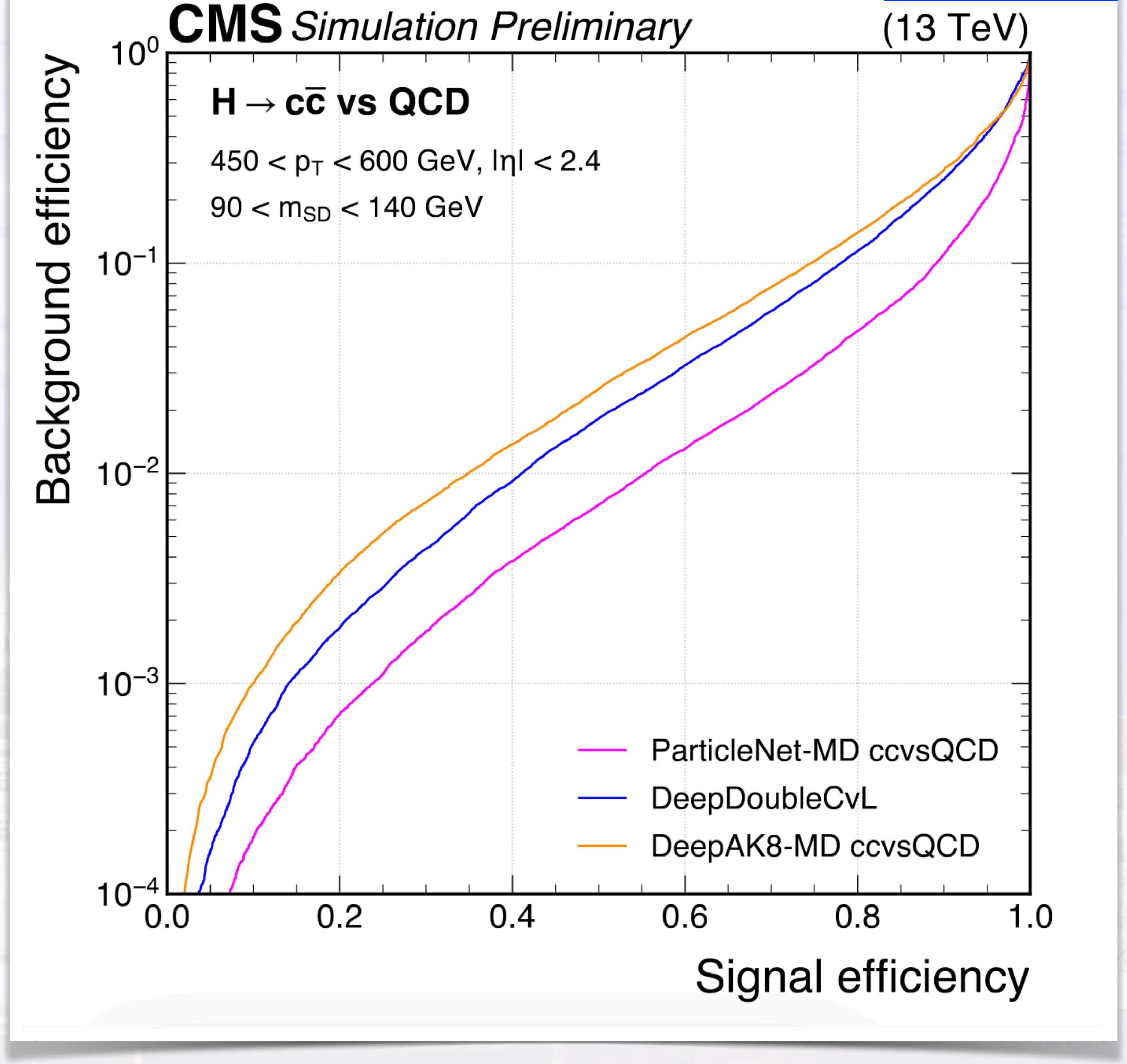
Run-2

Ref: *BTV-22-001*



Run-2

Ref: *BTV-22-001*

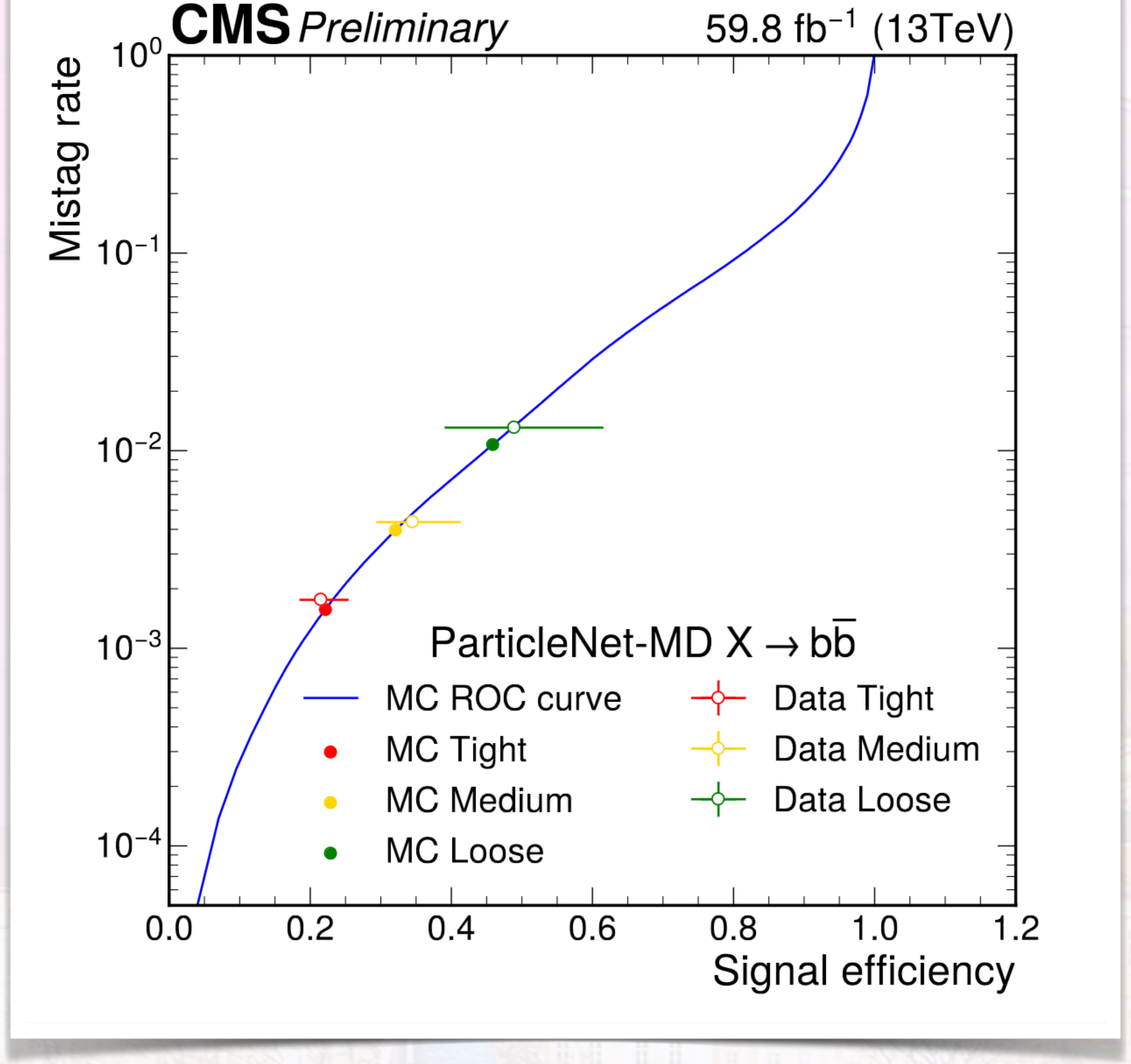


Improved discriminating power with the implementation of newer and better taggers

Calibration and Validation

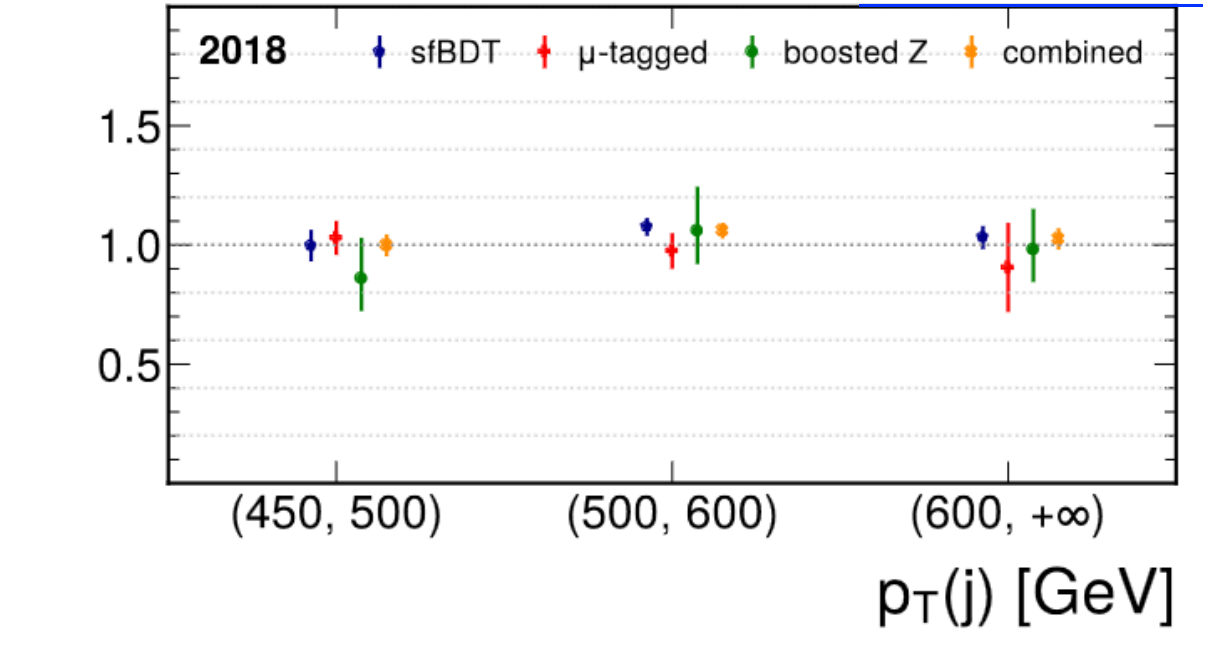
Run-2

Ref: *BTV-22-001*



- Calibration performed with Run-2 data
- Shows the data vs. simulations ROCs of $X\rightarrow b\bar{b}$ for different working points

Ref: *BTV-22-001*



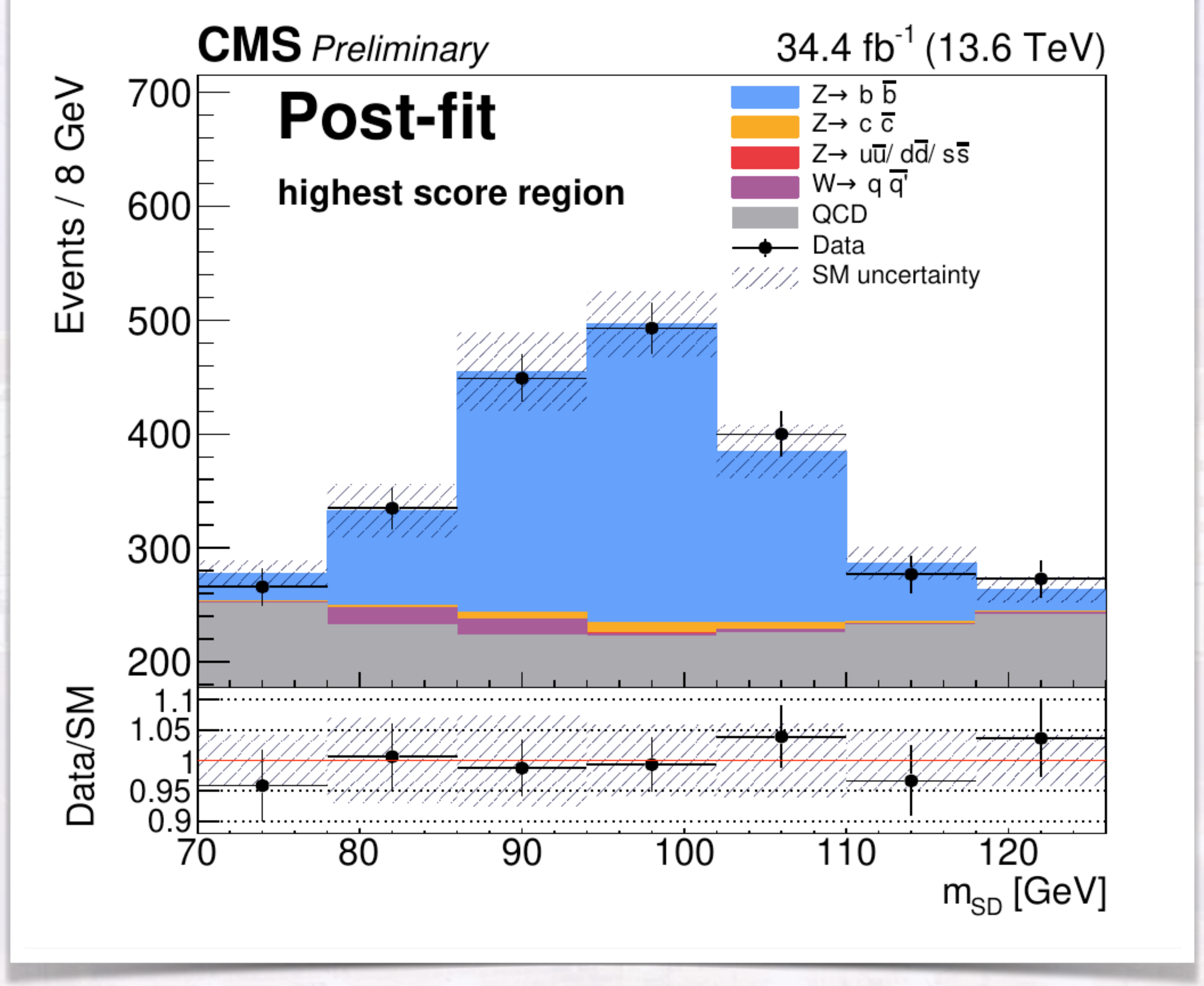
SFs for 2018 data taking period

Validation in $Z\rightarrow b\bar{b}$ tagging in Run-3

- $W\rightarrow q\bar{q}$ and QCD as main background
- Result shows the mass distribution in the highest score region (0.988 < ParticleNetMDbbvsQCD ≤ 1) of ParticleNetMD tagger

Run-3

Ref: *DP2024-055*



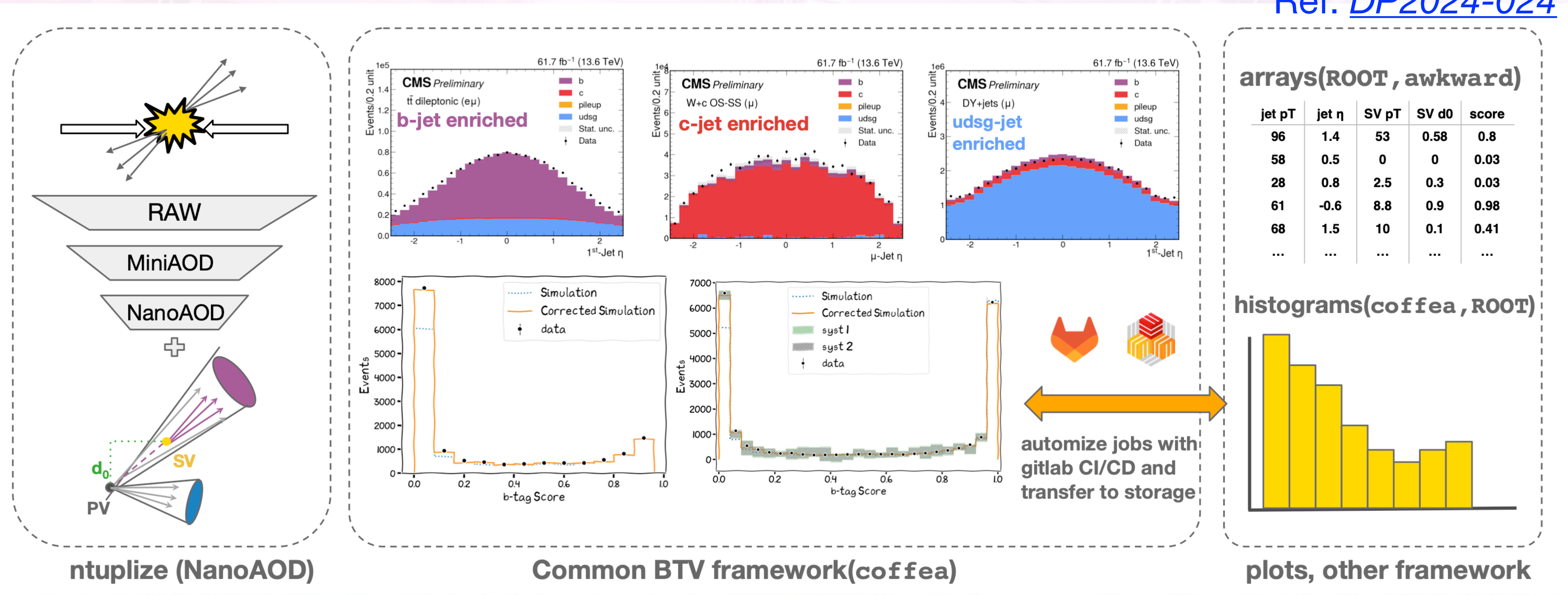
Related talk in ICHEP: *“Jet performance and pileup mitigation in Run3 in CMS”*

Frameworks

Commissioning Workflows

BTVNanoCommissioning

Ref: *DP2024-024*



arrays(ROOT,awkward)

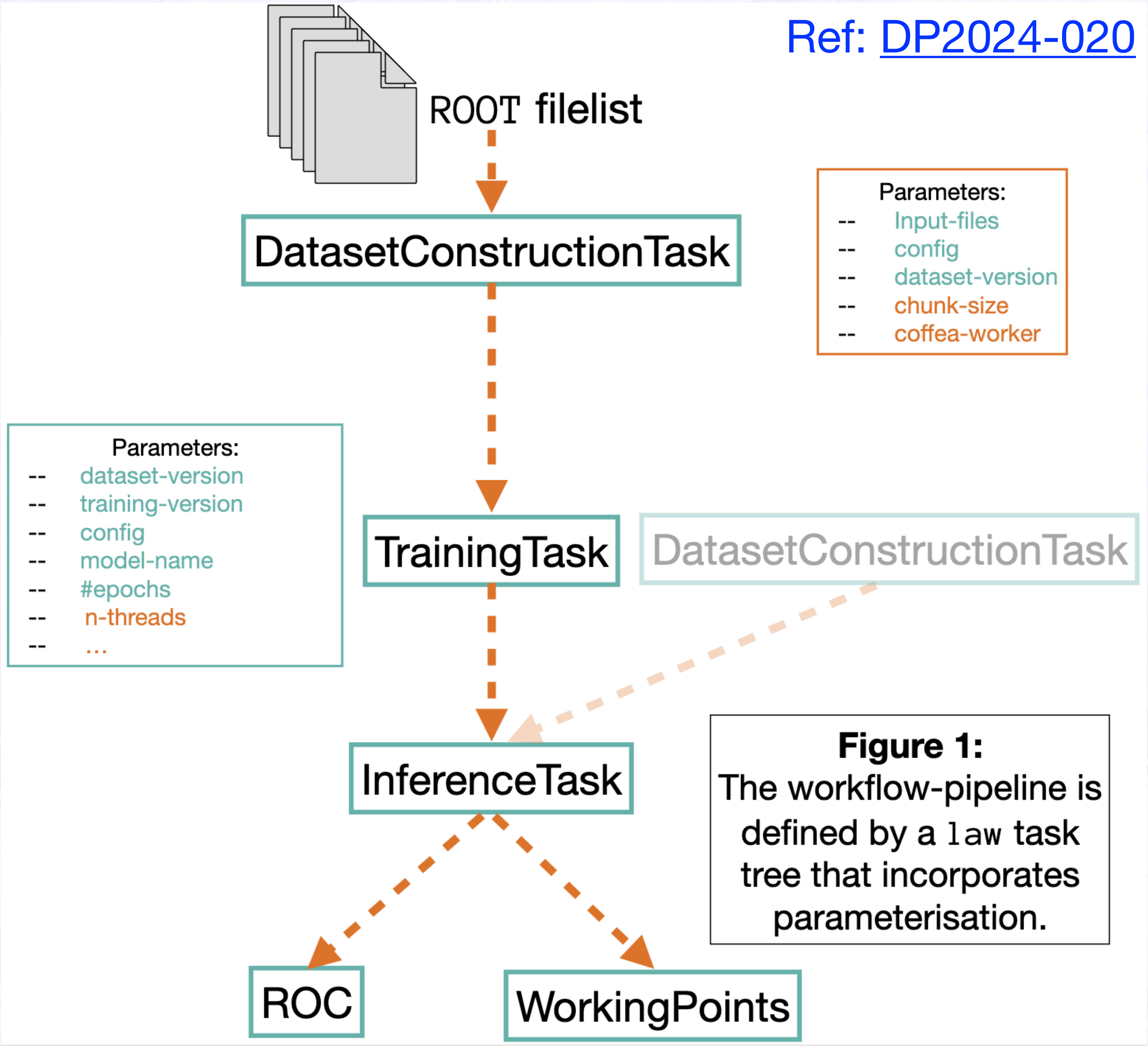
jet pT	jet η	SV pT	SV d0	score
96	1.4	53	0.58	0.8
58	0.5	0	0	0.03
28	0.8	2.5	0.3	0.03
61	-0.6	8.8	0.9	0.98
68	1.5	10	0.1	0.41
...	...	...	...	...

- Fast and efficient
- Pythonic array based manipulation instead of loops
- Automatized using Gitlab Continuous integration - monitor performance in regular intervals

Training

B-Hive

Ref: DP2024-020



Parameters:
-- Input-files
-- config
-- dataset-version
-- chunk-size
-- coffea-worker

Parameters:
-- dataset-version
-- training-version
-- config
-- model-name
-- #epochs
-- n-threads
-- ...

Figure 1: The workflow-pipeline is defined by a `law` task tree that incorporates parameterisation.

- Framework dedicated for training
- Easily customizable to introduce your own model

Summary

- **Rapid Development of Machine Learning Architectures (2017-2023)**:
  - Evolution from ***CSVv1*** to ***UParT* -** increasing complexity and capability of models
- **Innovative Calibration Techniques in Run-2**

- **Improvement in Jet Tagging Performance:**
  - ParticleNet algorithm deployed in the online High Level Trigger (Ref: DP2023-021)
  - Notable advancements from Run-1 to the present
  - Demonstrates significant gains in accuracy and efficiency

- **Ongoing Enhancement in Tagging and Calibration Methods:**
  - Continued development expected in Run-3
  - Aim to further refine and optimize performance
  - New taggers and many more, stay tuned!

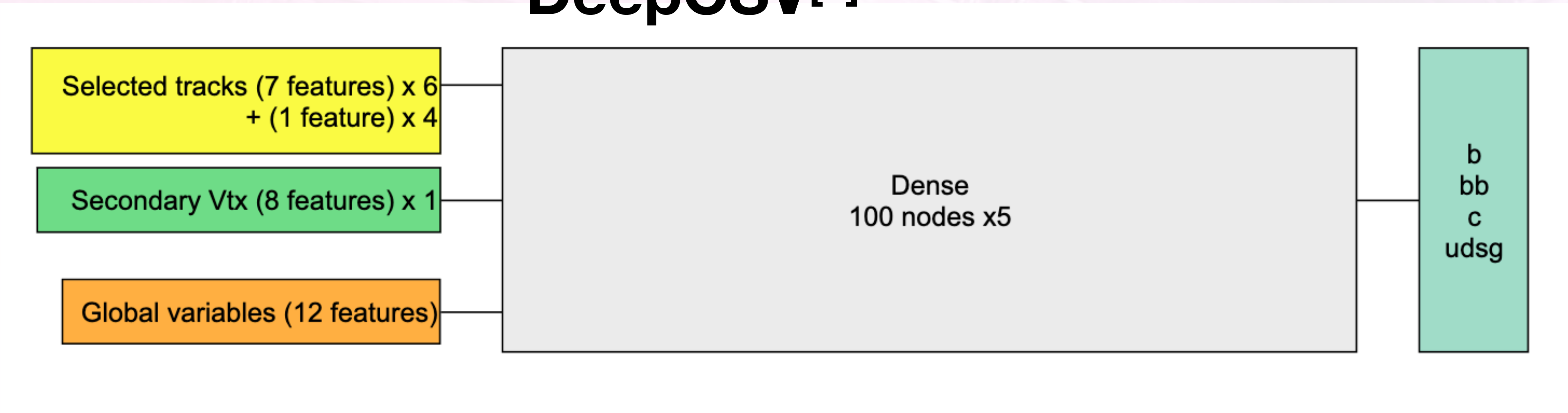
Thank you for listening!

Related poster from Donato Troiano: *Identification of Lorentz-boosted jets in the CMS experiment*

BACKUP

DeepCSV and DeepJet

DeepCSV[1]



Selected tracks (7 features) x 6 + (1 feature) x 4

Secondary Vtx (8 features) x 1

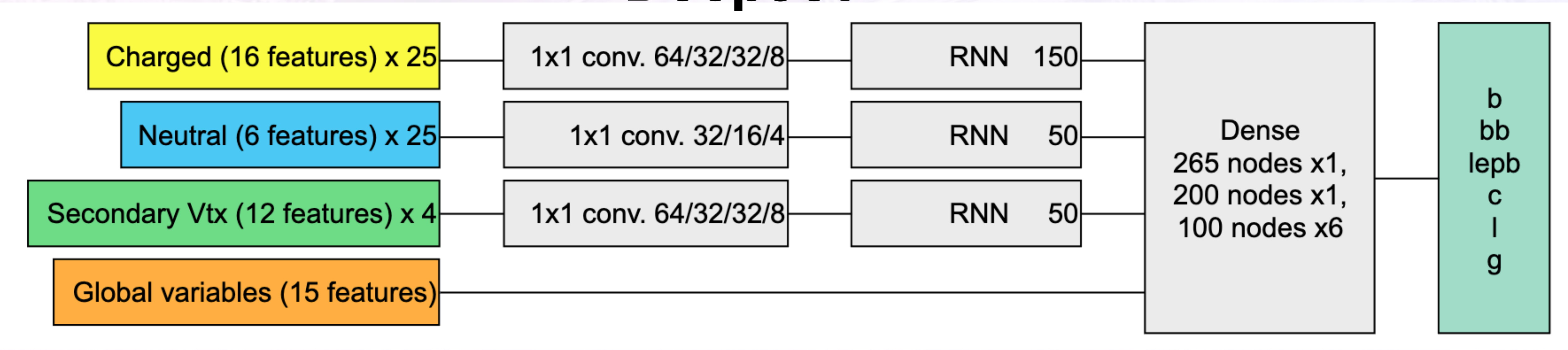
Global variables (12 features)

Dense
100 nodes x5

b
bb
c
udsg

- **Fully connected neural network (dense)**
- Combines properties from **selected** tracks, secondary vertices and global variables directly (66 features)
- Only a small subset of the charged jet constituents pass stringent quality criteria
  - clean and simple environment for the classifier
  - **information loss - potential performance degradation**

DeepJet[2]



Charged (16 features) x 25

Neutral (6 features) x 25

Secondary Vtx (12 features) x 4

Global variables (15 features)

1x1 conv. 64/32/32/8

1x1 conv. 32/16/4

1x1 conv. 64/32/32/8

RNN 150

RNN 50

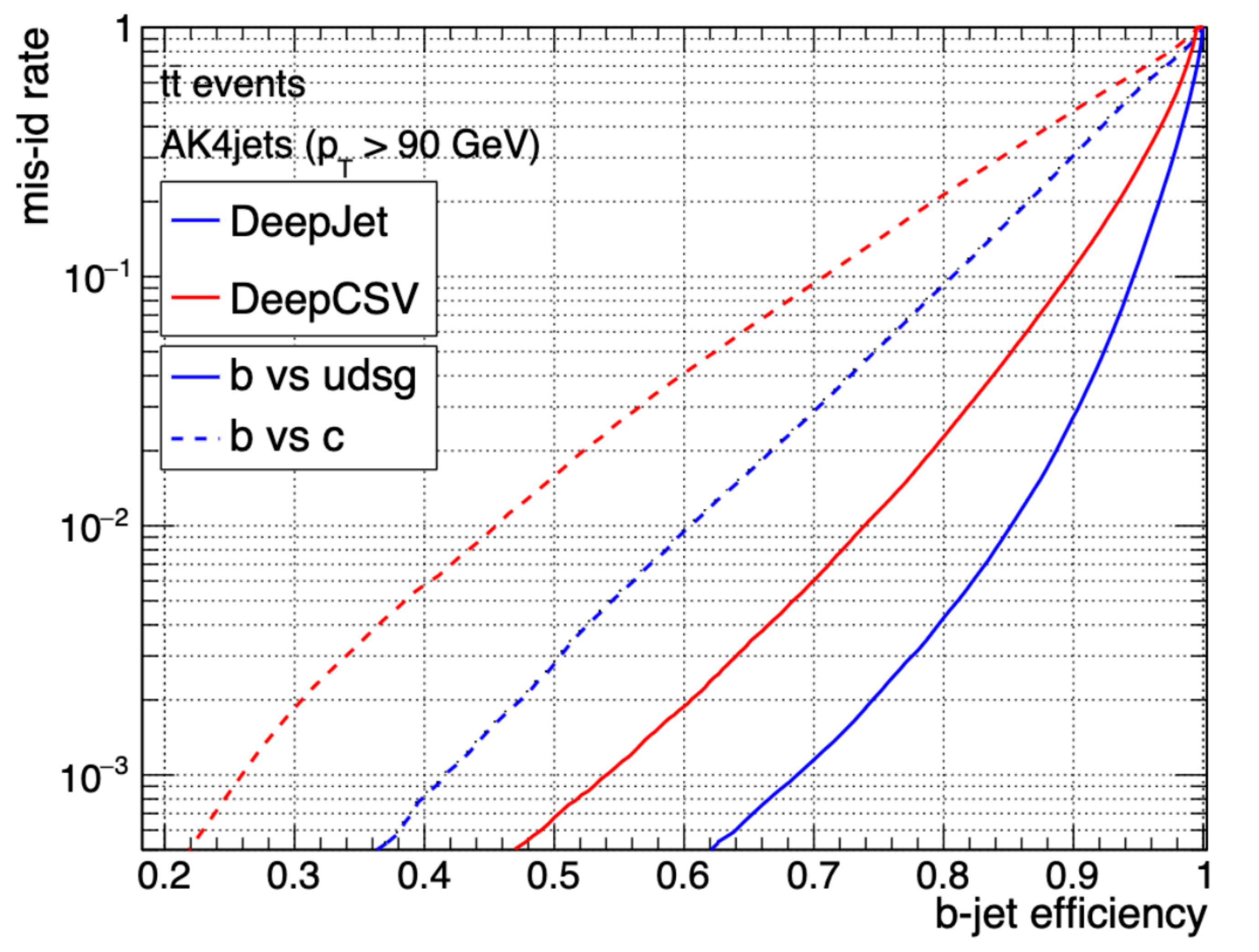
RNN 50

Dense
265 nodes x1,
200 nodes x1,
100 nodes x6

b
bb
lepb
c
l
g

- **Convolution, RNN,** and **Dense** layers
- **Does not rely on a selection** of the jet constituents
  - better purity, more number of inputs
- Full information of all **jet constituents, charged and neutral particles, secondary vertices, and global event variables simultaneously**

~20% gain in efficiency at 10^{-3} misidentification probability[2]



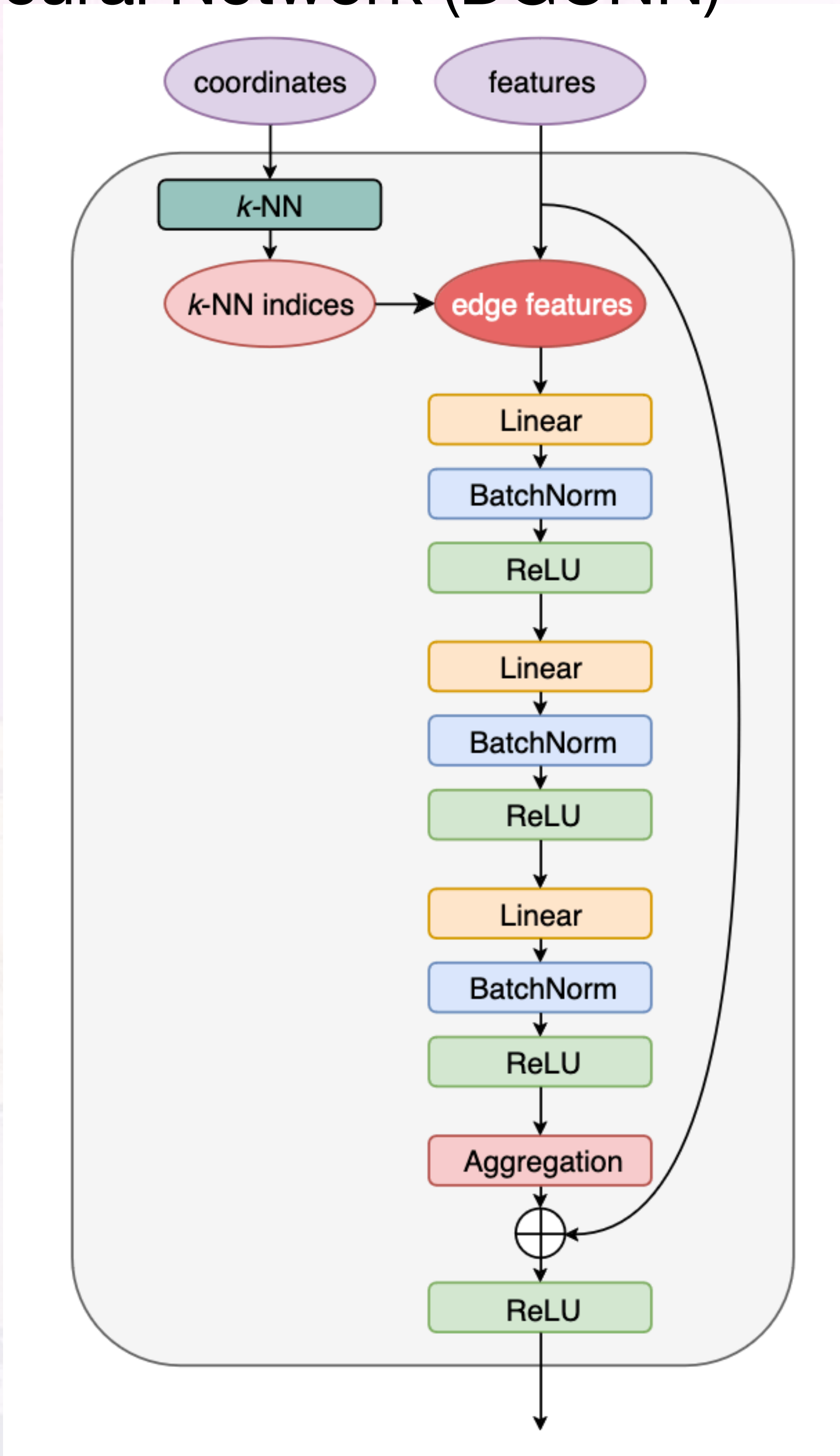
[1] 10.1103/PhysRevD.94.112002

[2] 10.1088/1748-0221/15/12/P12012

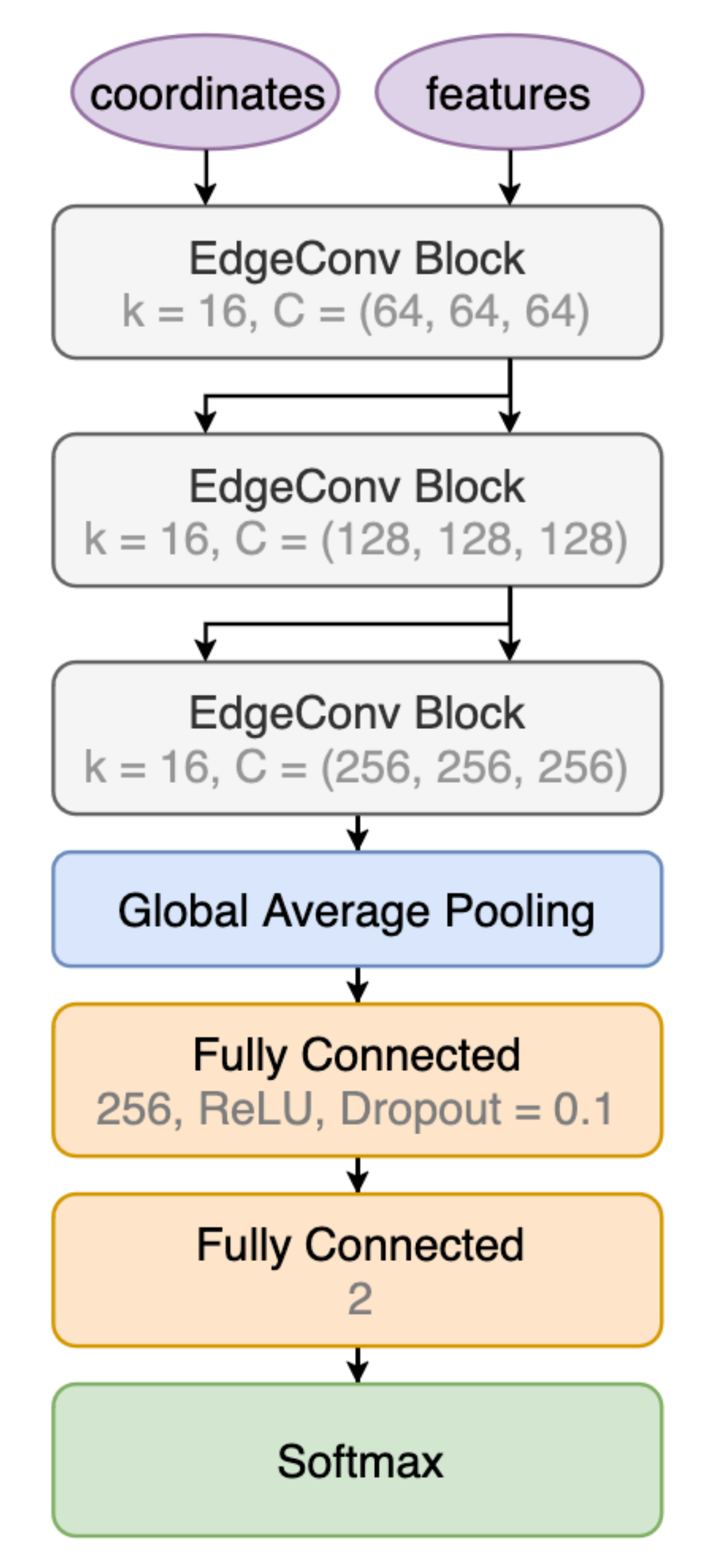
ParticleNet [1]

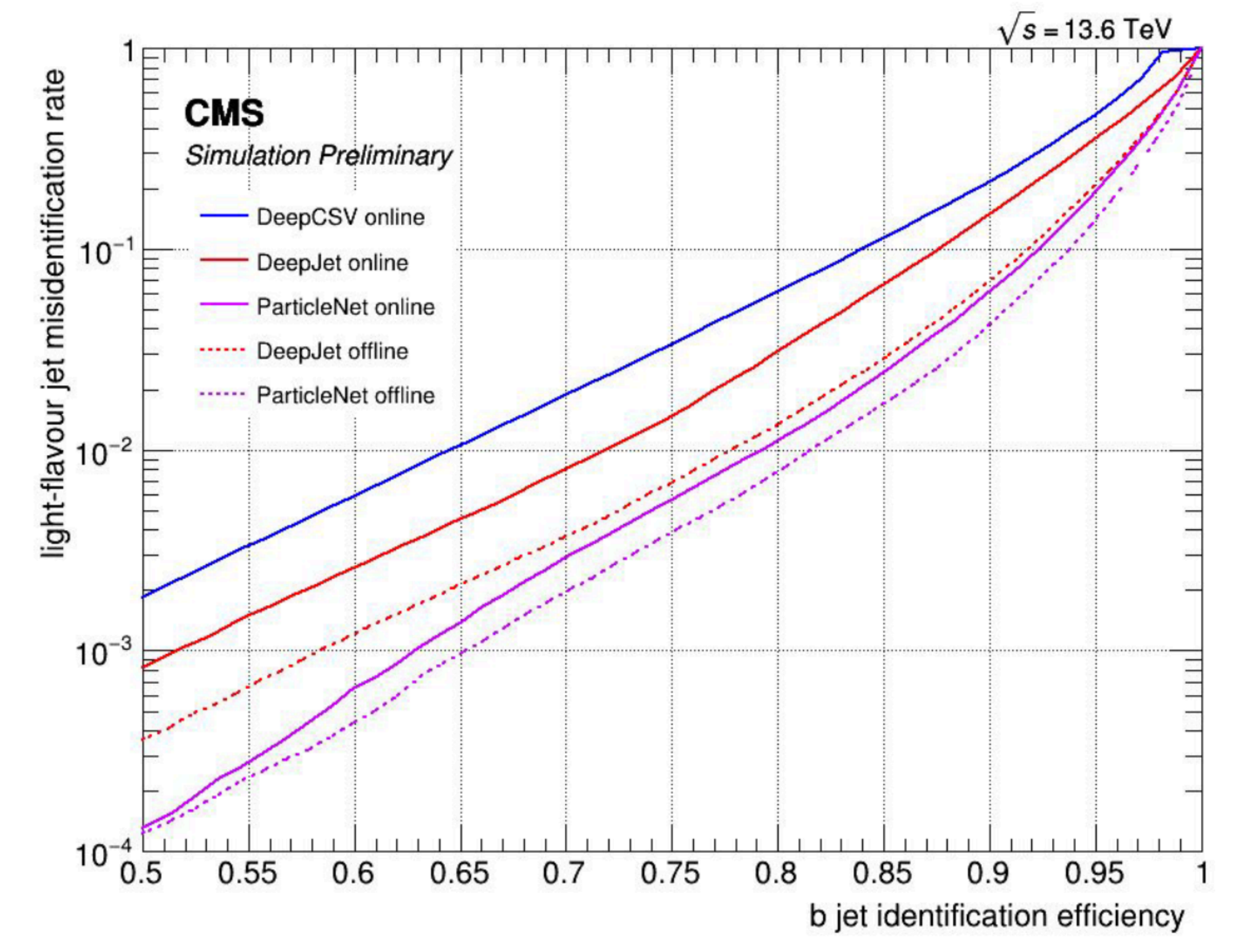
Treat a jet as unordered sets of constituent particles

Perform Edge-convolution and Dynamic Graph Convolutional Neural Network (DGCNN)



One edge-convolution block





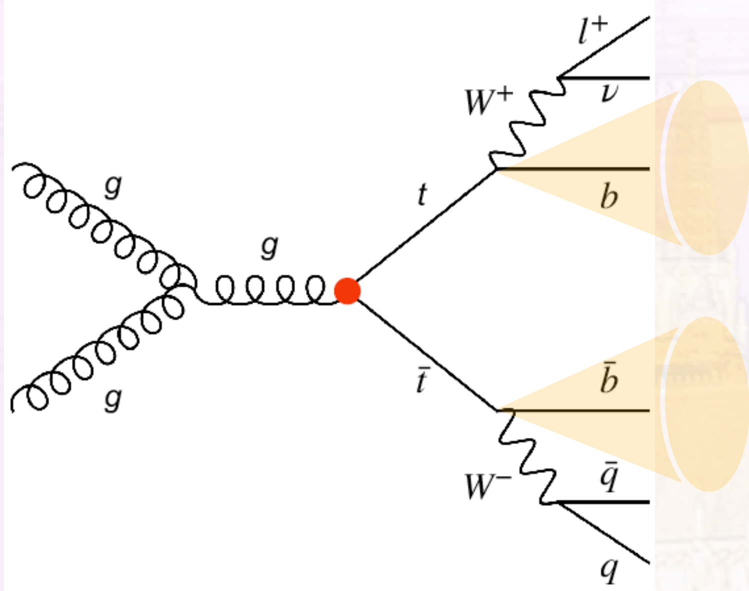
~10% gain in efficiency at 10^{-3} misidentification probability[2]

[1] arxiv1902.08570

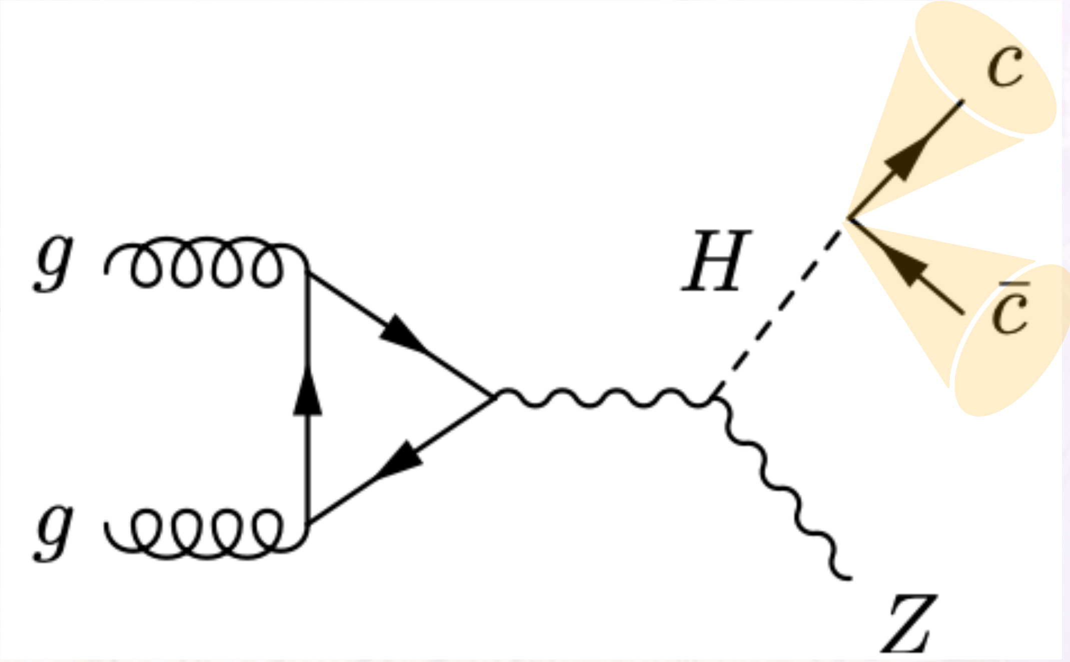
[2] CMS_DP2023_021

Theory

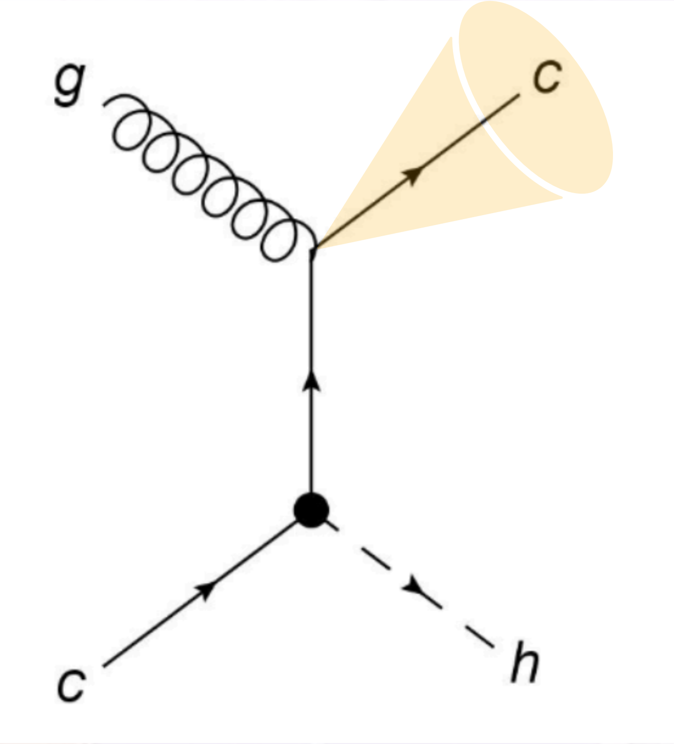
- **Heavy flavor jets =** jets originating from b (***b jets***) or c (***c jets***) quarks arising from the process of hadronization
- Important in Standard Model (SM), Top, Higgs(H->bb,cc), BSM and SUSY processes



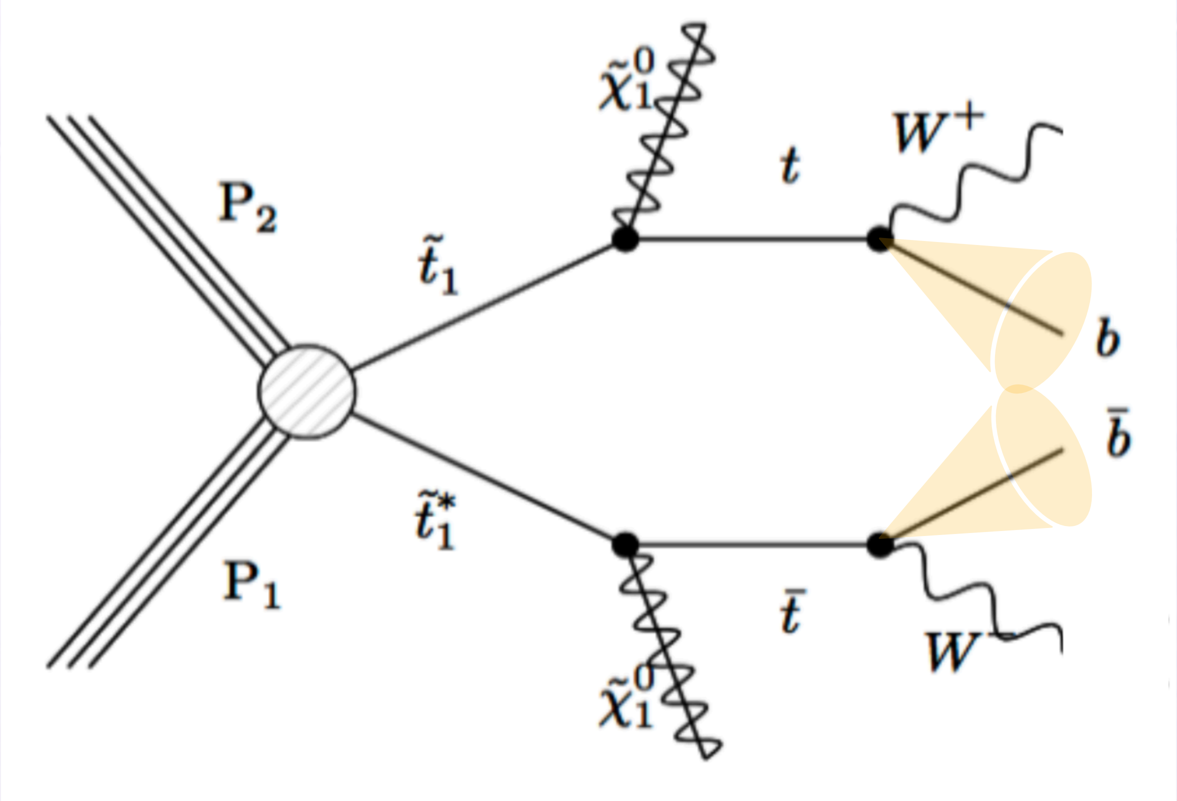
$t\bar{t}$



VHcc[1]

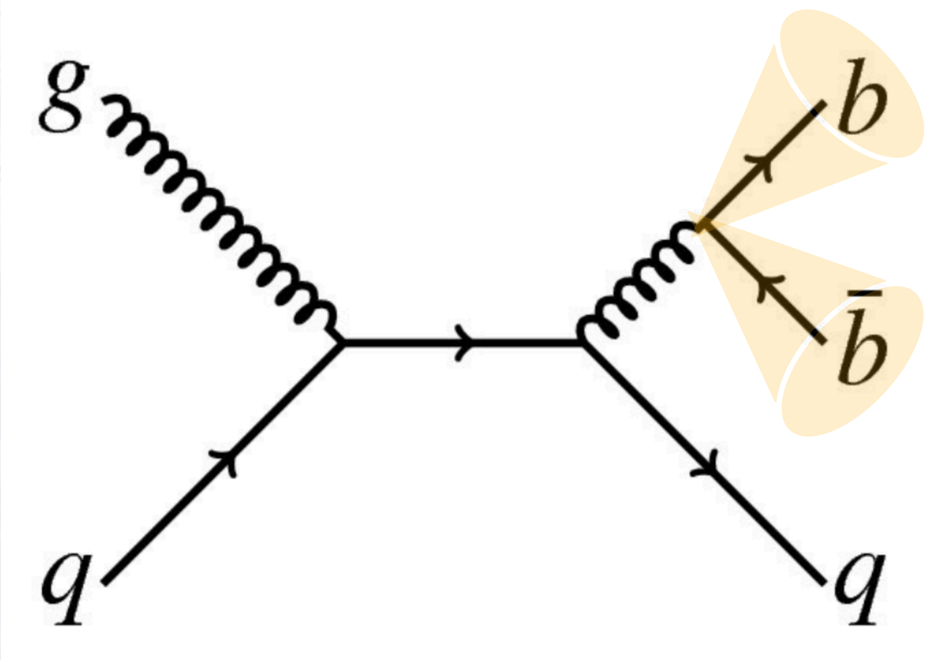


Higgs+c[2]



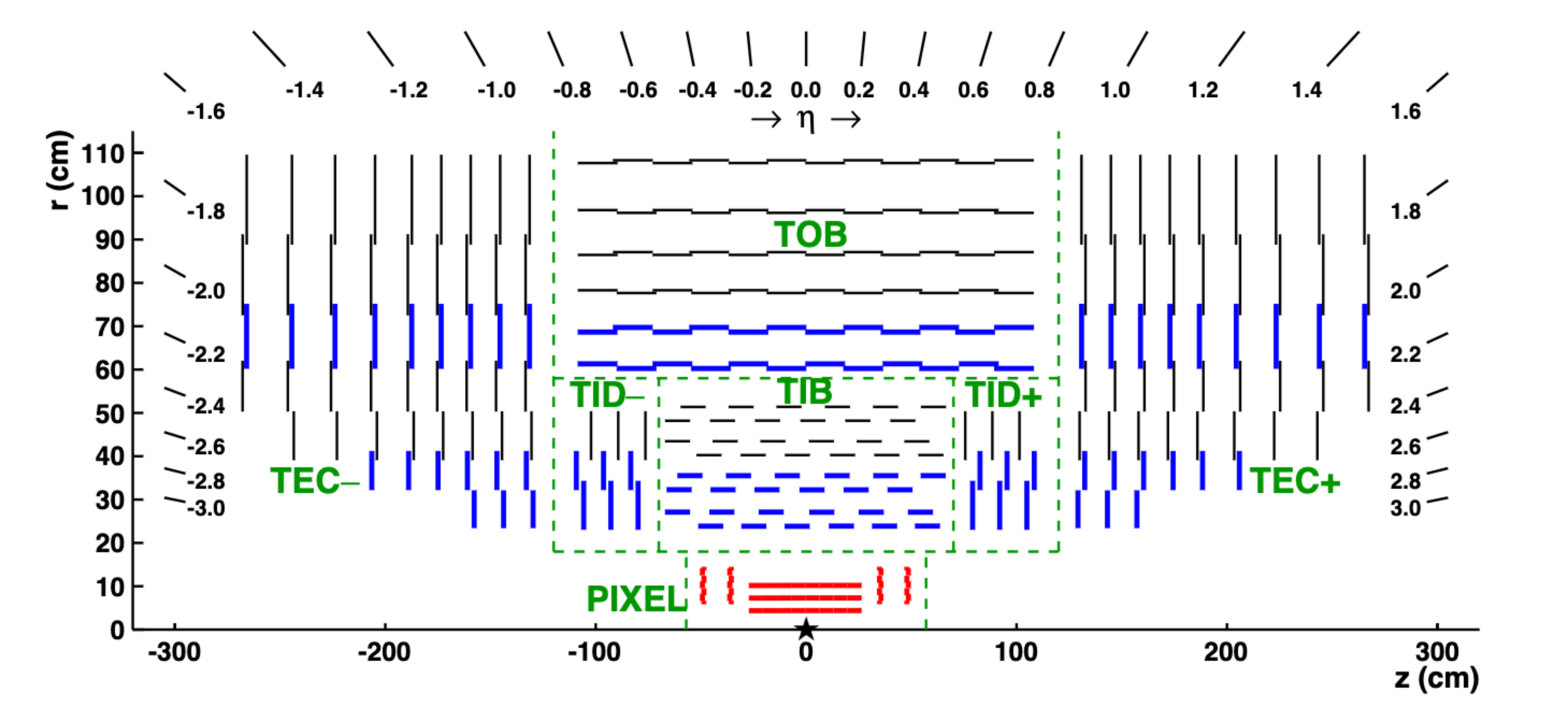
SUSY stop-> SM + MET

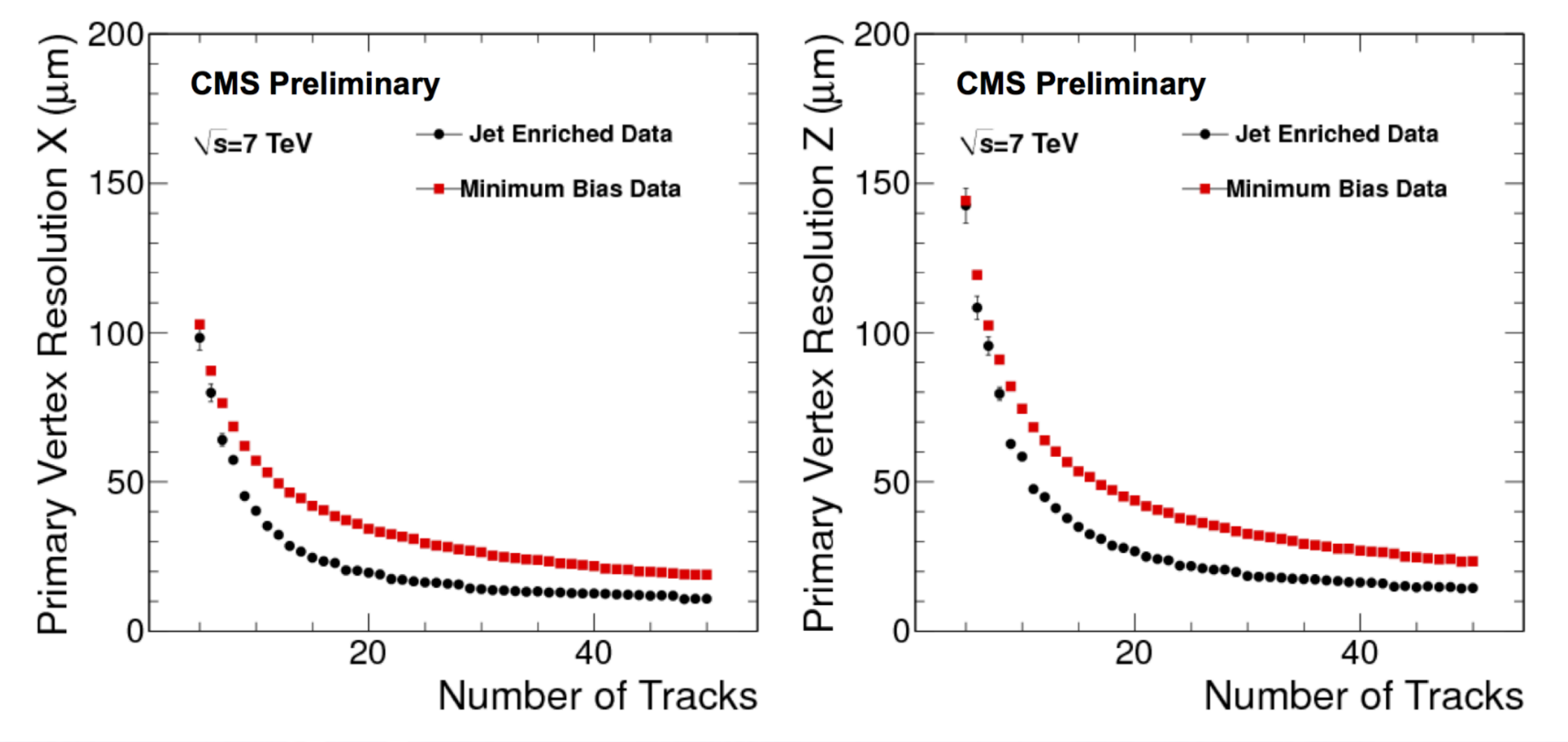
- **QCD:** understanding heavy-parton effects

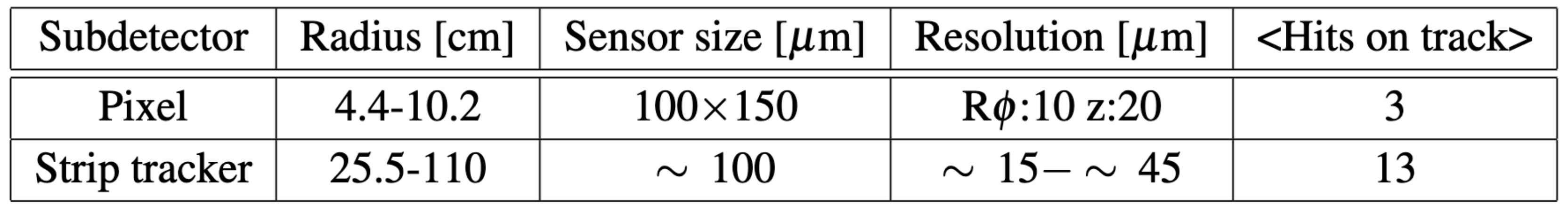


[1] https://www.dpg-verhandlungen.de/year/2024/conference/karlsruhe/part/t/session/20/contribution/6
[2] https://www.dpg-verhandlungen.de/year/2024/conference/karlsruhe/part/t/session/71/contribution/7

CMS Tracking







Subdetector	Radius [cm]	Sensor size [μm]	Resolution [μm]	<Hits on track>
Pixel	4.4-10.2	100×150	Rϕ:10 z:20	3
Strip tracker	25.5-110	$\sim$ 100	$\sim$ 15– $\sim$ 45	13

https://pos.sissa.it/190/041/pdf

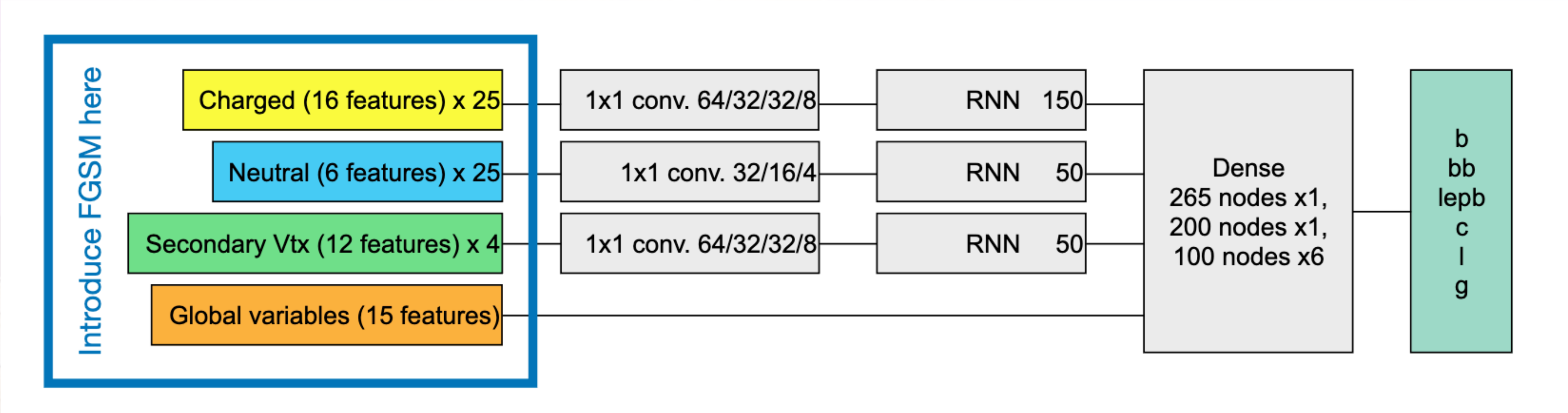
Ideal to observe in CMS, though challenging!

[1] http://cds.cern.ch/record/1279383/files/TRK-10-005-pas.pdf

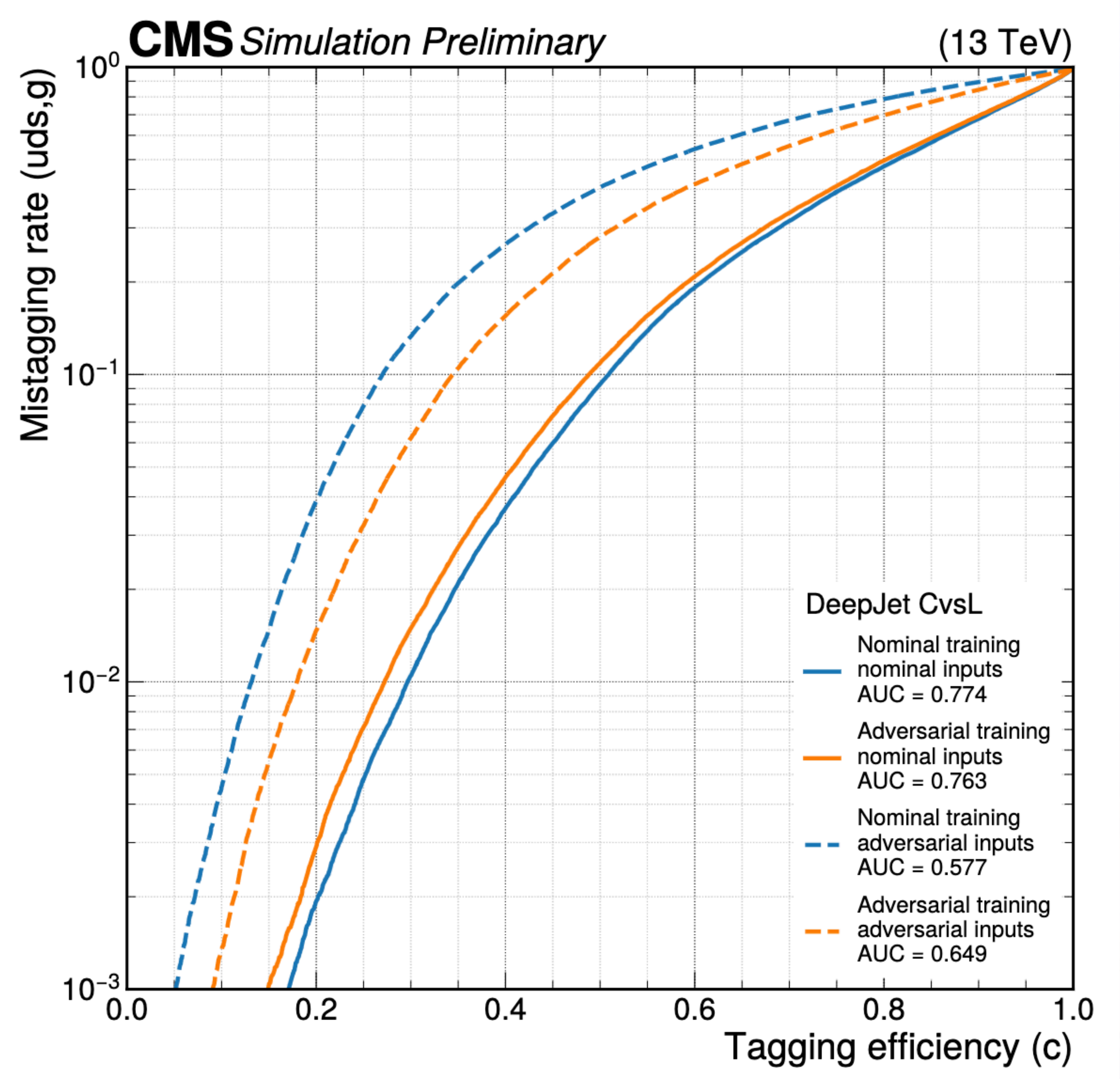
Adversarial Attacks

- Adversarial training strategy:
  - Reduce the observed differences prior to any calibration,
  - Improve robustness of the classifier against injected mismodelings

The Fast Gradient Sign Method (FGSM) is used to systematically distort inputs



DeepJet architecture



https://cds.cern.ch/record/2839919/files/DP2022_049.pdf

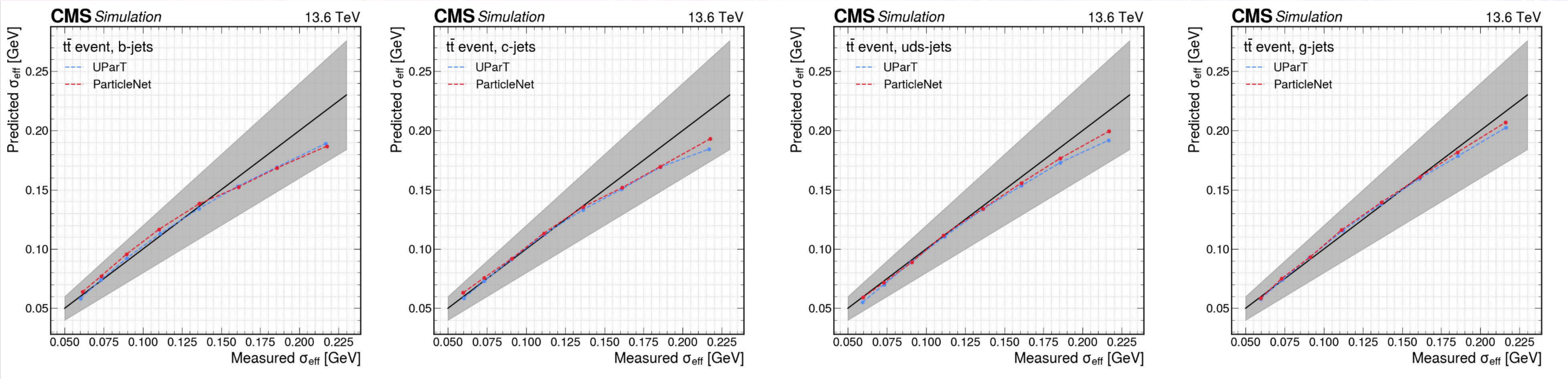
Jet energy regression: UParT

Ref: *BTV-24-XXX*

CMS Simulation Preliminary
13.6 TeV
$t\bar{t}$ events
$p_T > 30$ GeV
UParT
ParticleNet
$0.0 < |\eta| < 1.0$
$1.0 < |\eta| < 1.5$
$1.5 < |\eta| < 2.0$
Median (p_T^{corr}/p_T^{gen})
1.15
1.10
1.05
1.00
0.95
0.90
0.85
100
200
300
400
p_T^{gen} [GeV]

Jet energy resolution: UParT

Ref: *BTV-24-XXX*

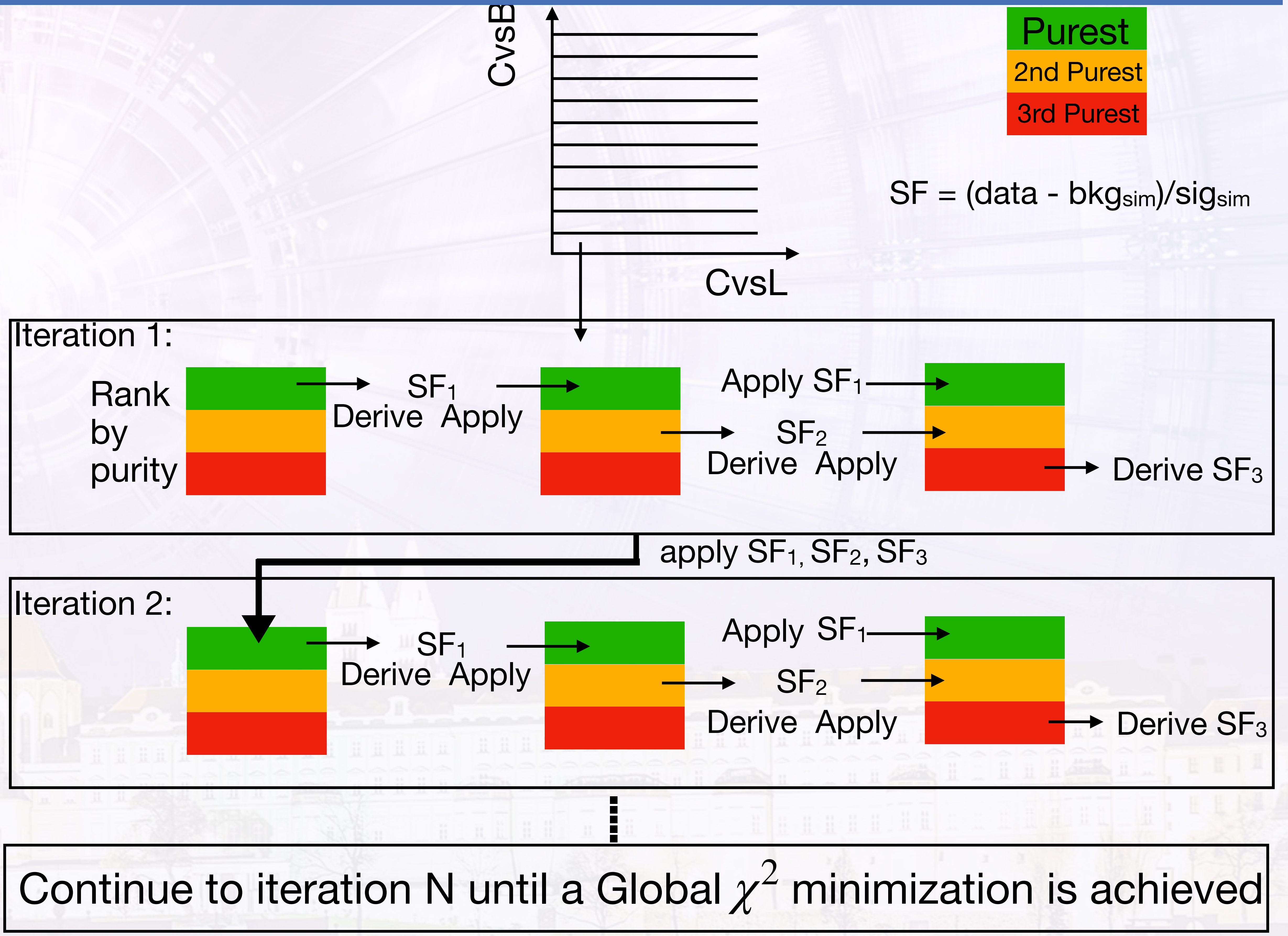


Iterative Fit: c tagging

- Fit iteratively in the 2D plane of CvsB-CvsL over three selections allowing each flavour component to vary independently
  - Divide CvsB-CvsL plane into 10 “bin slices” along the CvsB axis
  - Variable binning along CvsL axis - perform fit and derive SF in each bin

Apply SF$_c$ to the same distributions they were derived from

- Propagate uncertainties (from data):
  - For each systematic source, redo iterative fit with selections shifted by 1σ on each side



CvsB

CvsL

🟩 Purest

🟧 2nd Purest

🟥 3rd Purest

SF = (data - bkg$_{sim}$)/sig$_{sim}$

Iteration 1:

Rank by purity

SF$_1$ Derive Apply

Apply SF$_1$

SF$_2$ Derive Apply

Derive SF$_3$

apply SF$_1$, SF$_2$, SF$_3$

Iteration 2:

SF$_1$ Derive Apply

Apply SF$_1$

SF$_2$ Derive Apply

Derive SF$_3$

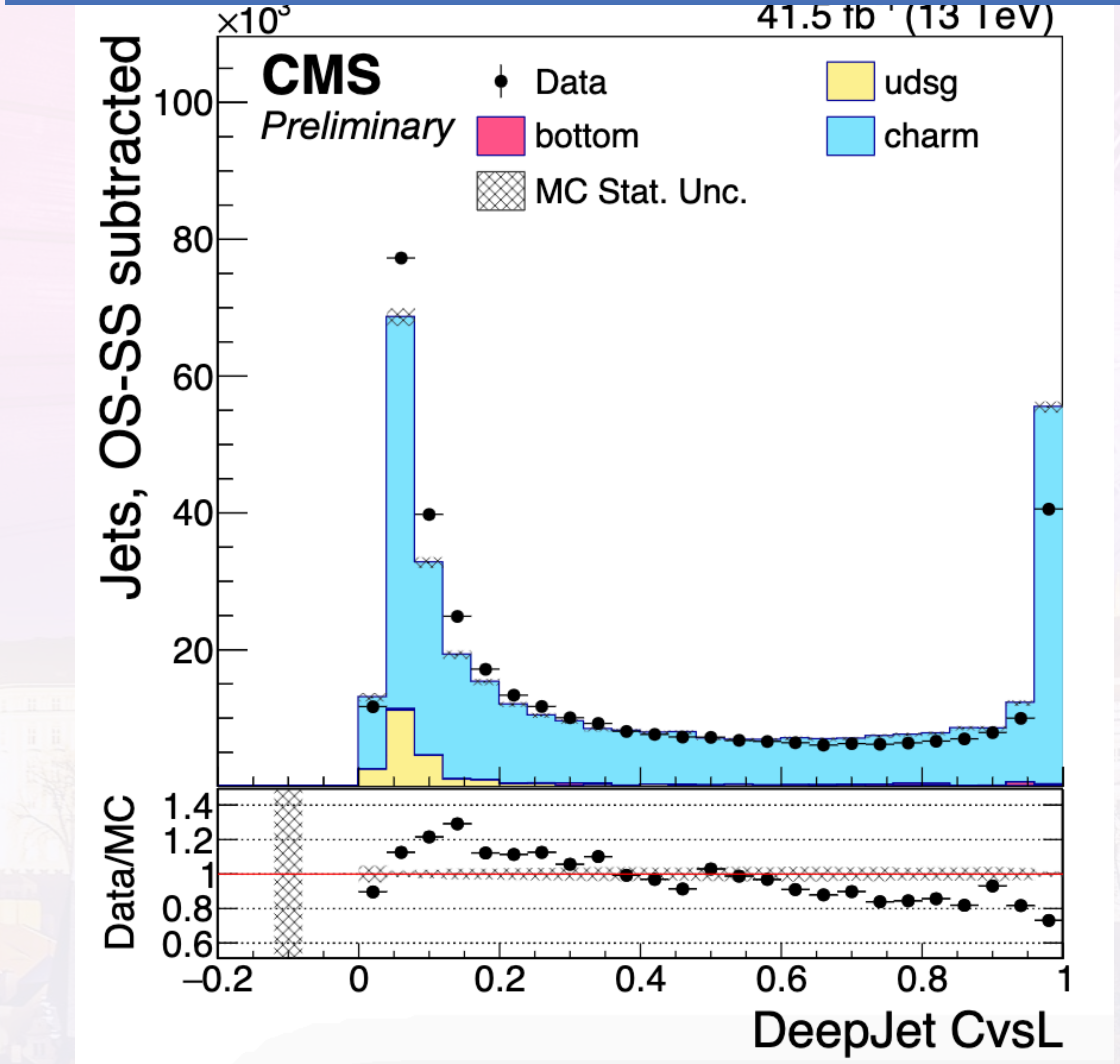
Continue to iteration N until a Global χ^2 minimization is achieved

$$\chi^2_{sel} = \sum_{i=1}^{nBins} \frac{\left(N_{sim,i} - N_{data,i}\right)^2}{\sigma^2_{sim,i} + \sigma^2_{data,i}}$$

$$\text{Global } \chi^2 = \chi^2_{Wc} + \chi^2_{t\bar{t}} + \chi^2_{DY}$$

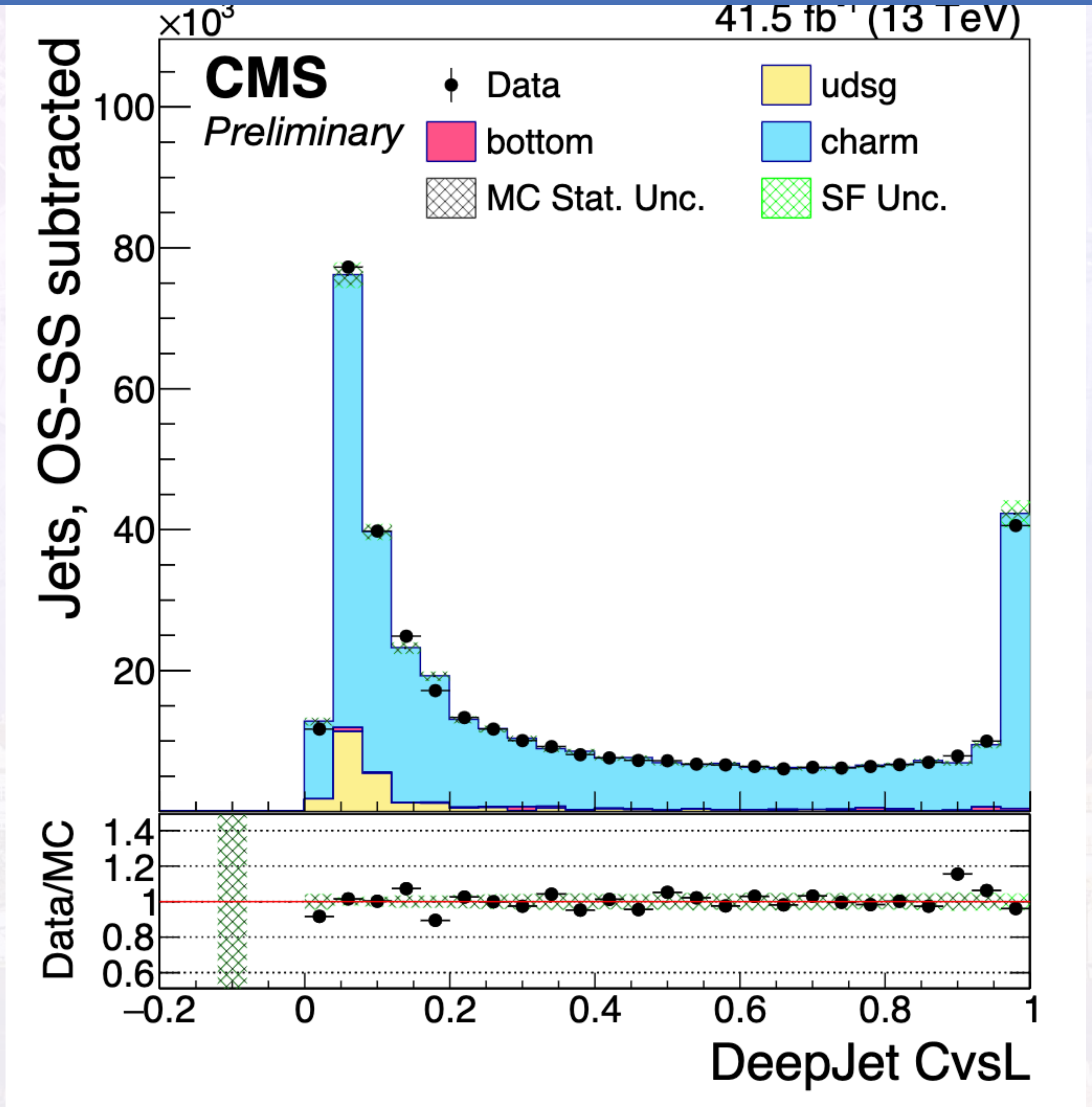
10.1088/1748-0221/17/03/P03014

Performance: c tagging



Before applying SF_c

Apply SF and Uncertainties



After applying SF_c

10.1088/1748-0221/17/03/P03014

Good agreements post calibration!

Calibration

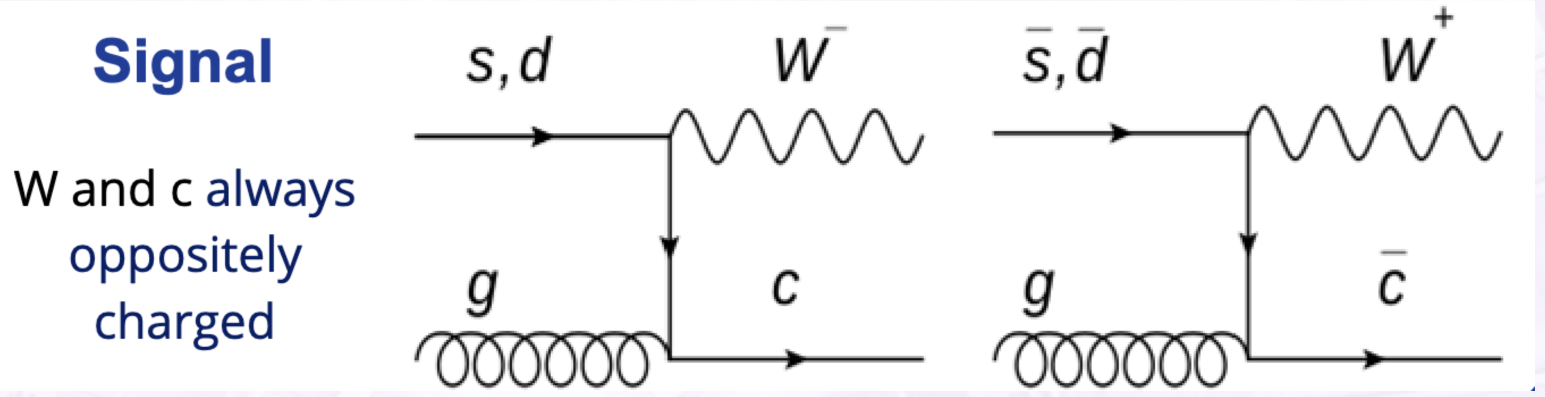
Variable binning along CvsL axis

- Binning for the SFs for each flavour is determined based on:
  - **Stats available:** Set the width to the minimum width required to reach a target stat uncertainty (ϵ_{max}) per bin. Practically, we start with a minimum width, b_{min}, and increase in steps of b_{min}.
  - **An upper limit:** If stat unc < ϵ_{max} is not satisfied even with width >= b_{max}, set bin width to b_{max}.
- Parameters to be preset: ϵ_{max}, b_{min}, b_{max}.

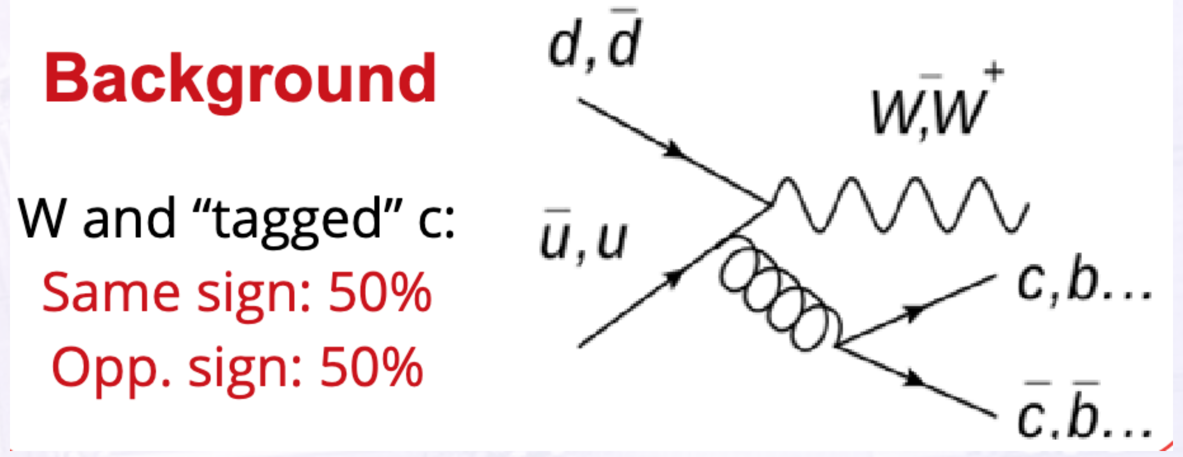
- Final choice of parameters: $\epsilon_{max} = 2\%$, $b_{min} = 0.02$, $b_{max} = 0.10$.

Calibration: Event selection for c SF

- **Define three regions:**
  - **c enriched:** W($l\nu$) and at least one jet with a soft, non-isolated mu,
    - OS – SS selection gives pure W + c distributions

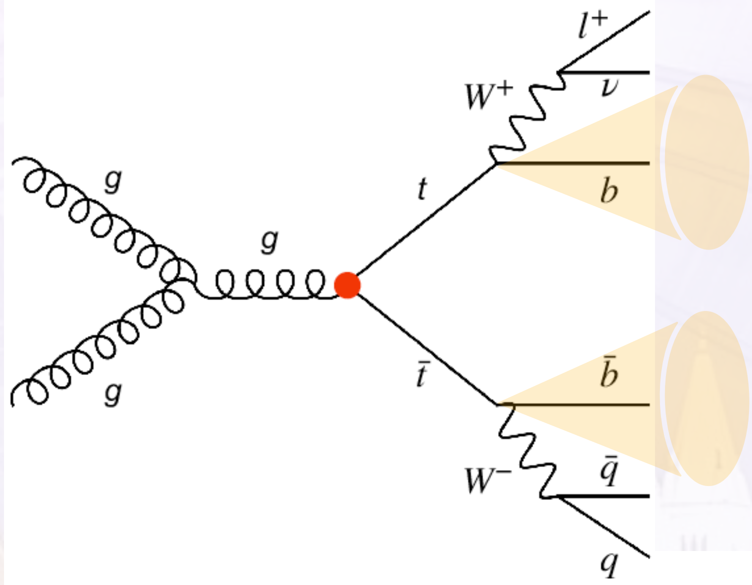

Signal

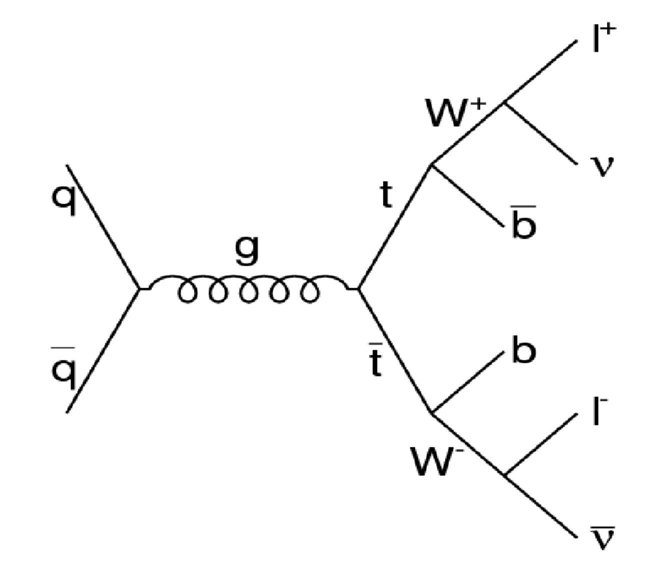
W and c always oppositely charged


Background

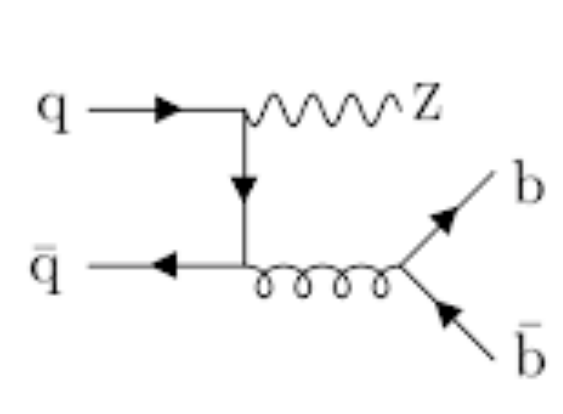
W and “tagged” c:
Same sign: 50%
Opp. sign: 50%

  - **b enriched:** Semileptonic and dileptonic $t\bar{t}$

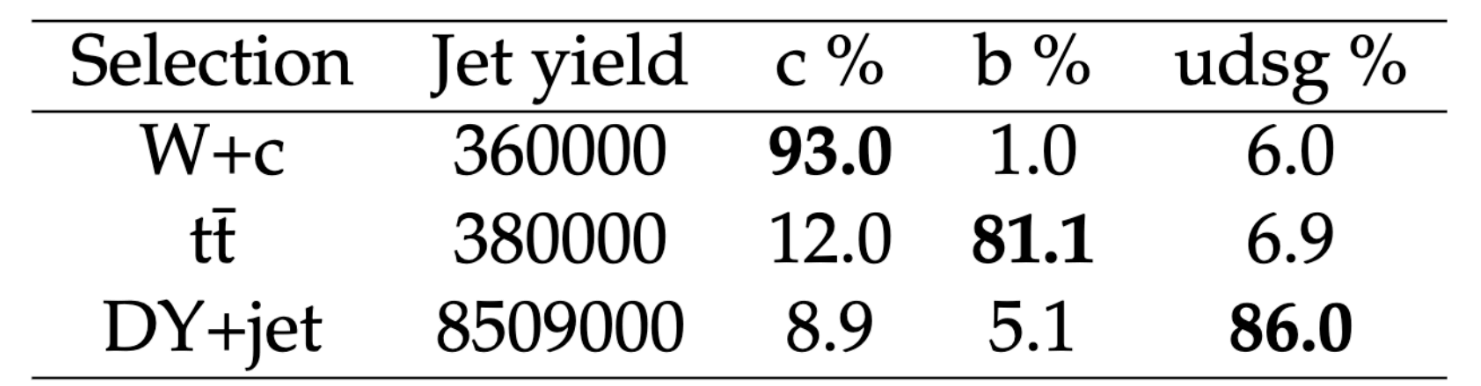




  - **Light-flavor jet enriched:** DY(ll)+jets(udsg)



No heavy-flavor tagging in the selections



Selection	Jet yield	c %	b %	udsg %
W+c	360000	**93.0**	1.0	6.0
tt̄	380000	12.0	**81.1**	6.9
DY+jet	8509000	8.9	5.1	**86.0**

Reasonably pure regions enriched in respective flavors

10.1088/1748-0221/17/03/P03014

Uncertainties: c-taggging

- **Statistical uncertainties:** Evaluated as the statistical uncertainty in the ratio of Data and MC in each bin

- **Systematics:**

  We consider the following sources:

✓ Electron ID

✓ Muon ID

✓ PileUp

✓ JES (total)

✓ JER

✓ Factorisation scale

✓ Renormalisation scale

✓ Parton shower: ISR

✓ Parton shower: FSR

Cross section:

✓ W+Jets

✓ DY+Jets

✓ $t\bar{t}$

✓ Single Top

[1] https://cds.cern.ch/record/2866276