Machine learning in jet physics beyond jet classification at CMS

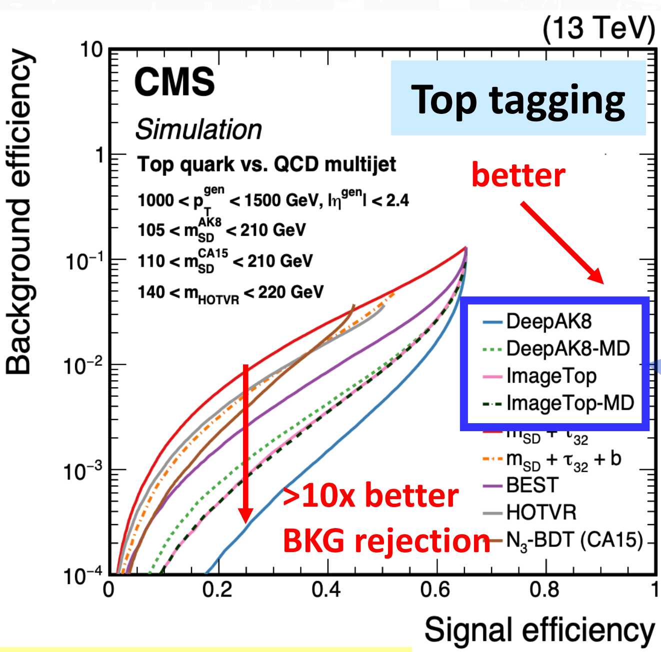
Loukas Gouskos (CERN)
on behalf of the CMS Collaboration

Motivation

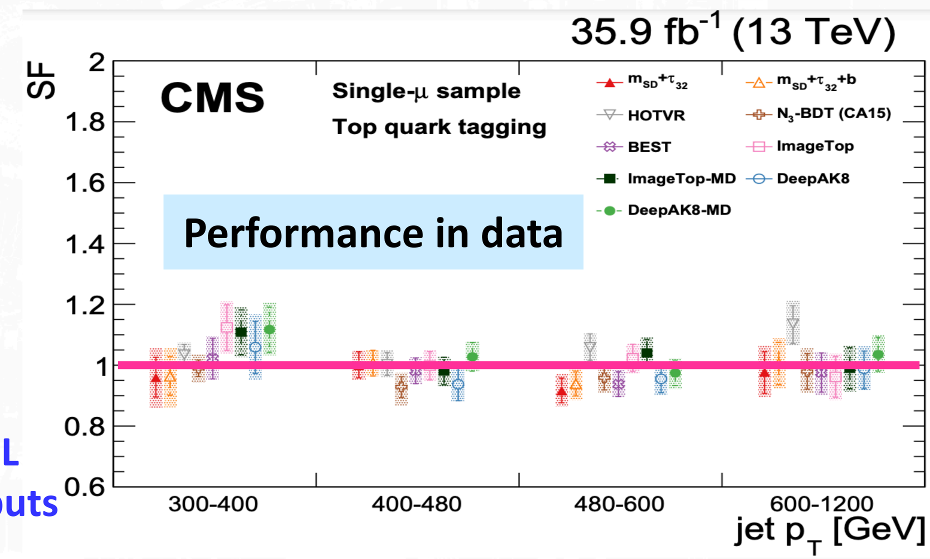
- Jets: essential for the success of the LHC physics program
  - both for **standard model (SM)** and **beyond SM (BSM)** physics analyses

Key for success: Well calibrated jets & constantly improving our “JetToolboox”
→ particularly important these days that LHC integrated luminosity increases only ~linearly with time & do not expect big jumps in collision energy

- Machine learning (ML) tools proven to be very powerful in jet physics
  - Jet classification paved the way:



Advanced ML
+ Low-level inputs



JINST 15 (2020) P06005

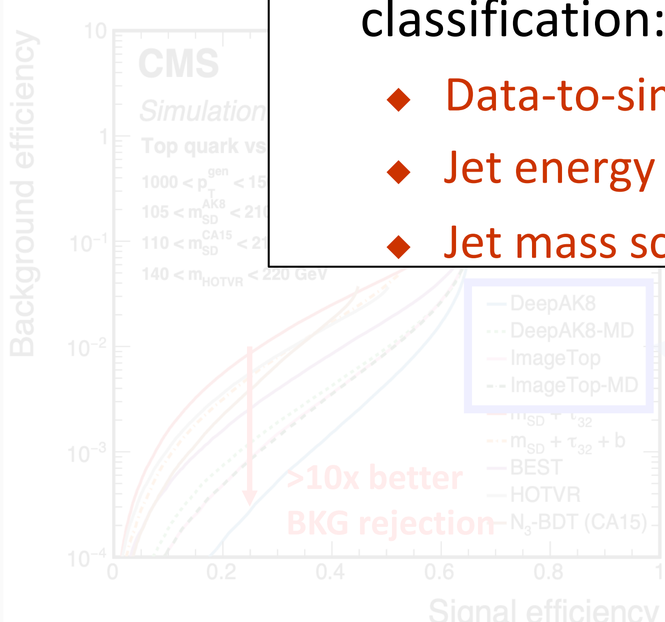
Very significant improvement in performance which translates in data [more details in Congqiao’s talk]

Motivation

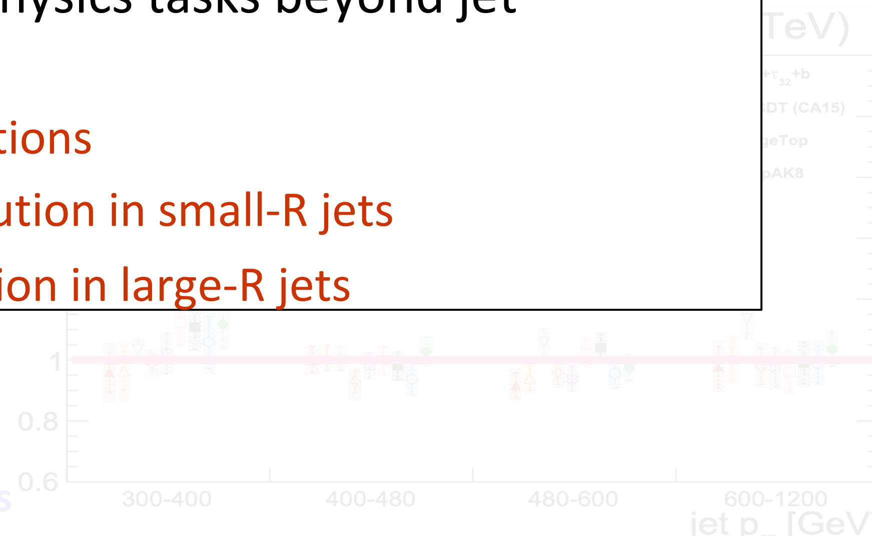
- Jets: essential for the success of the LHC physics program
  - both for **standard model (SM)** and **beyond SM (BSM)** physics analyses

Key for success: Well calibrated jets & constantly improving out "JetToolboox"
→ particularly important these days that LHC integrated luminosity increases only ~linearly with time & do not expect big jumps in collision energy

- Machin[…]
  - Jet c[…]



Advanced ML
+ Low-level inputs



Very significant improvement in performance which translates in data [more details in Congqiao's talk]

- Advanced ML techniques started to be explored in CMS for many other important jet physics tasks beyond jet classification:
  - Data-to-simulation corrections
  - Jet energy scale and resolution in small-R jets
  - Jet mass scale and resolution in large-R jets

Jet tagging calibration

- We are entering an era of extensive and deep understanding of the performance of the new generation tagging tools in data
  - Calibrate taggers → derive "scale factors" [SF] using a suite of data samples
  - Next big bet: Use these results to improve/tune MC generators

- e.g.,: b-tagging discriminant shape calibration

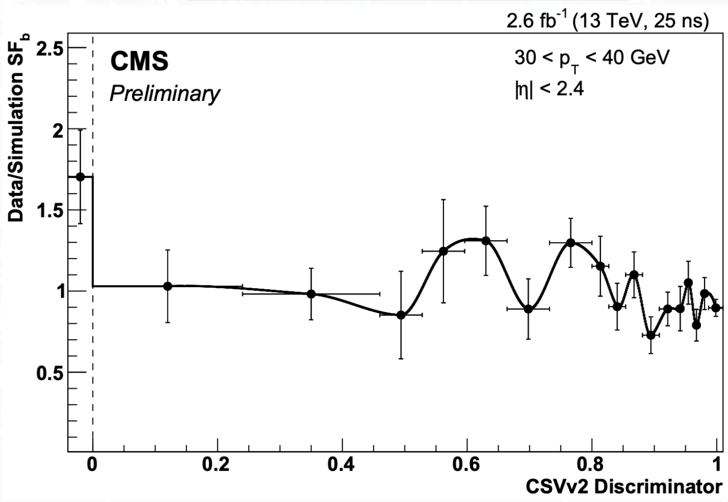
CMS-PAS-BTV-15-001

  - **Traditional approach:** Method relies on a "Tag-&-Probe" method in 2L events
    - **Tag jet:** Define b-enriched (tt2L) and b-depleted (DY) samples based on a selection on the b-tagging discriminant
    - **Probe jet:** Derive SFs following an iterative approach
      - subtract contribution from other flavours [e.g., light contribution in b-enriched region]

$$SF_{f,i+1} = \left. \frac{Data - \omega_i \cdot MC_{\neg f}}{MC_f} \right|_{R=R(f)}$$

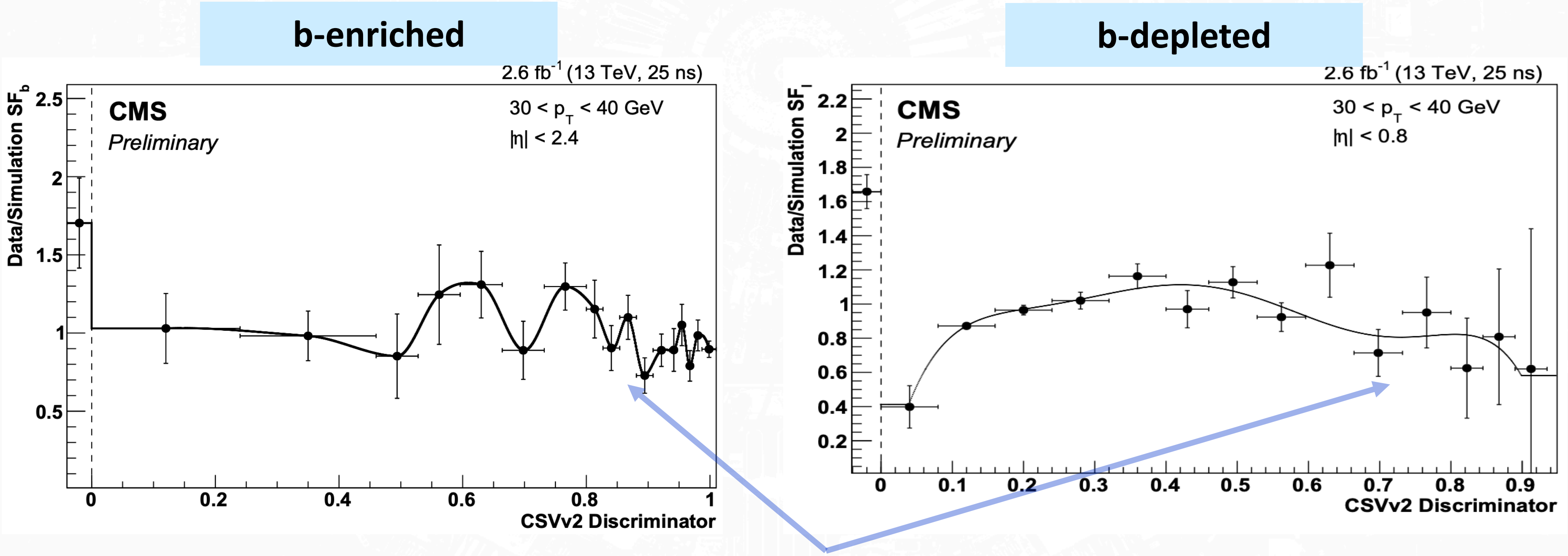
$$\omega = \prod_j^{\text{jets}} SF_j$$

b-tagging shape SFs



Jet tagging calibration (II)

- Yet: lots of room for improvement
  - SFs binned in p_T (jet), η(jet) and b-tagging discriminant
    - Ensure that each bin has sufficient stats (tedious..)
    - Not straight forward to include additional variables
  - Challenging to include additional control regions
    - e.g., finer flavor splitting of the b-depleted region
  - Results could be sensitive to fluctuations; finding the right fit function challenging

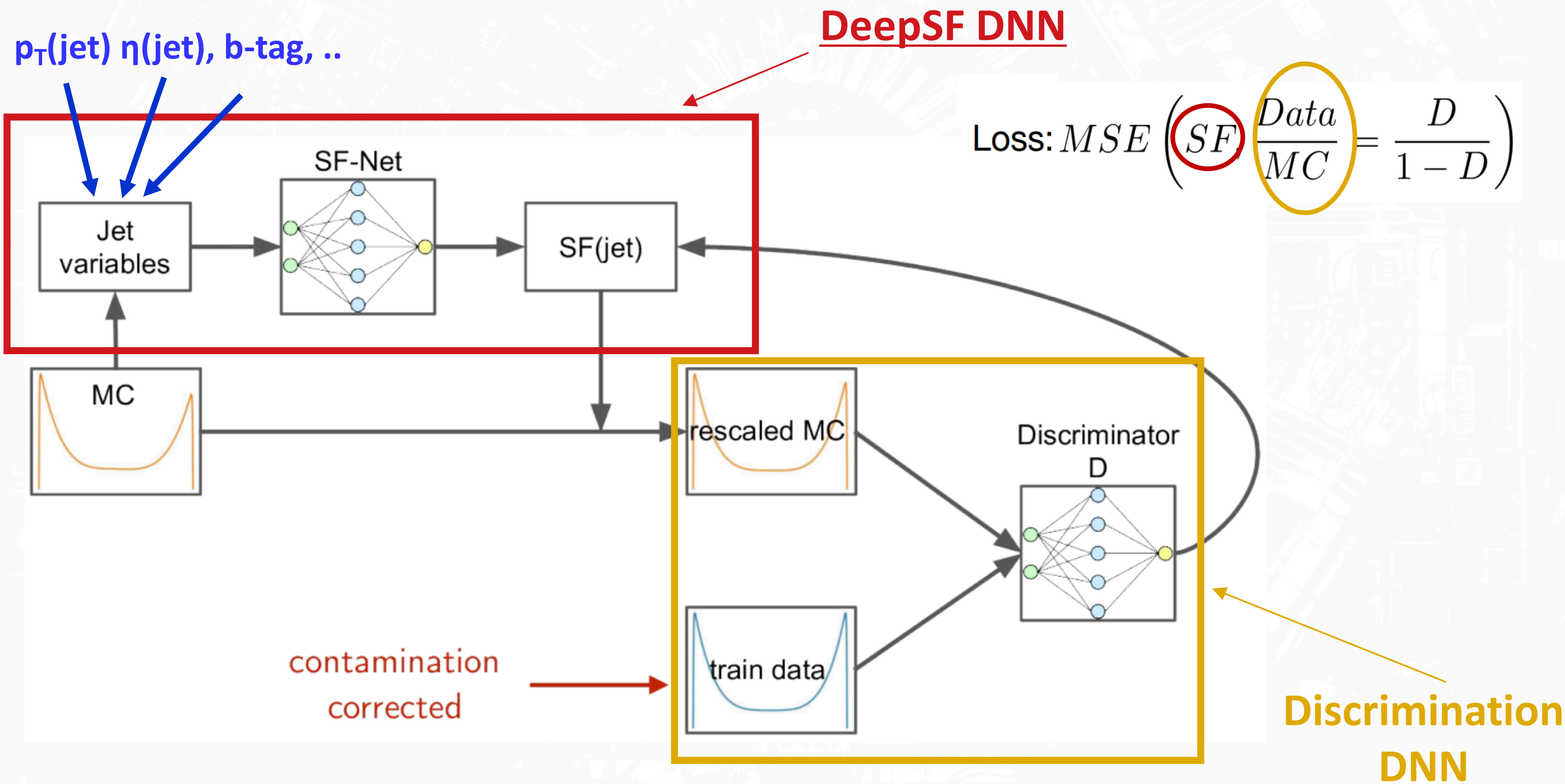


Fluctuations: smoothing challenging

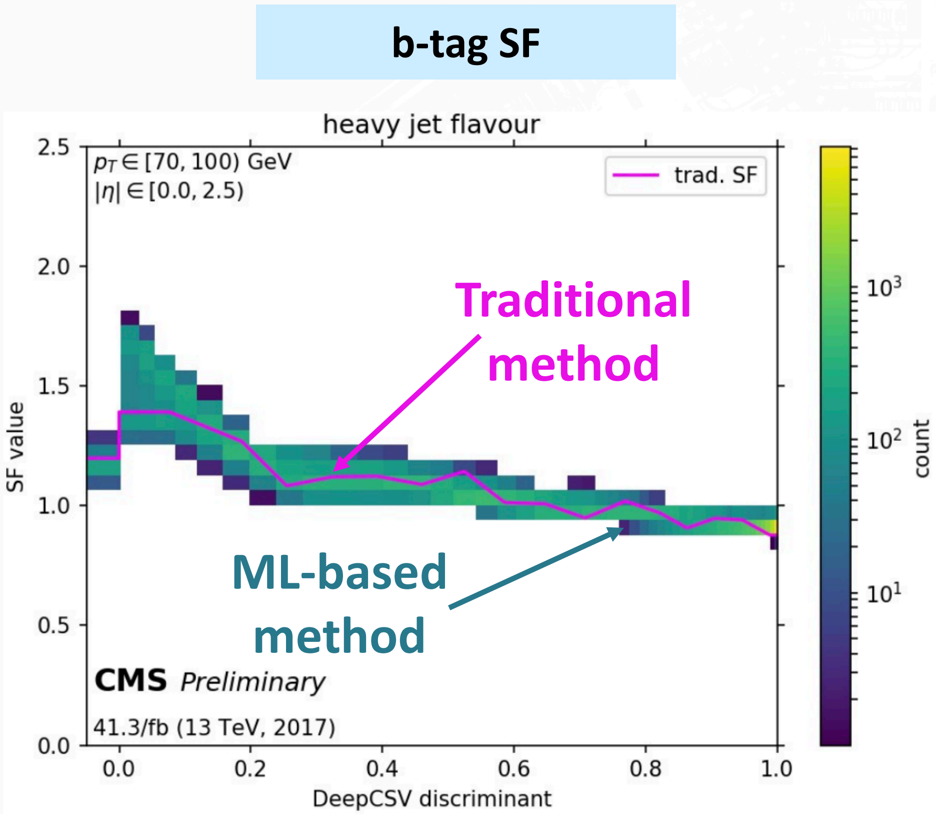
Jet tagging calibration with ML

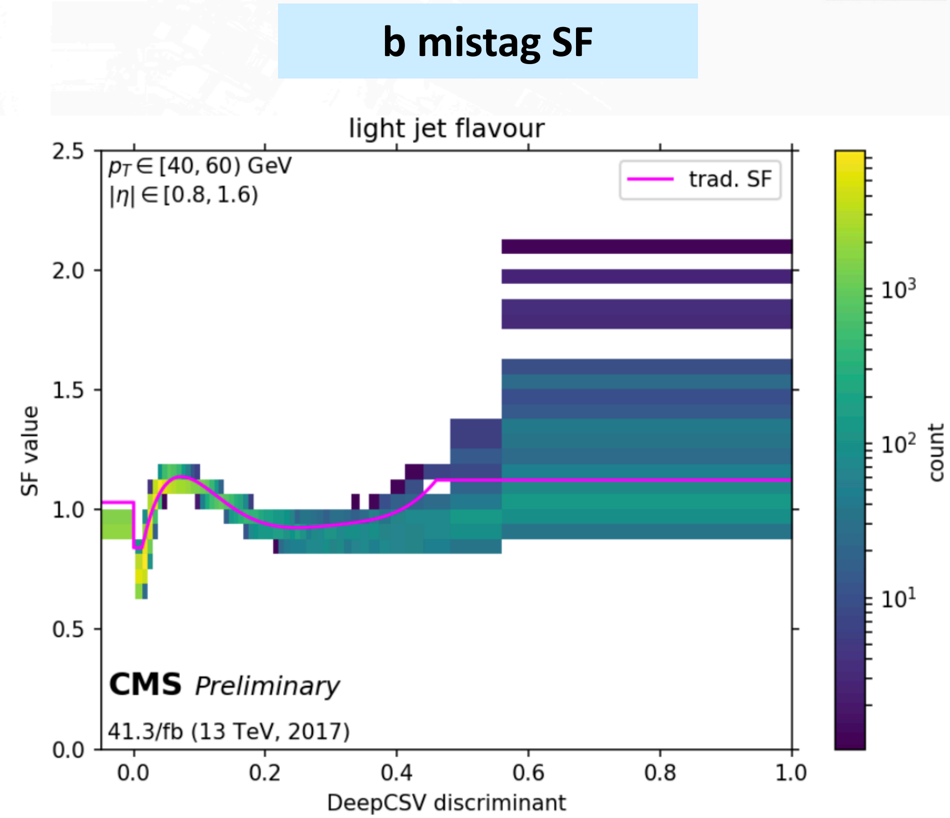
[CMS-DP-2019-003](CMS-DP-2019-003)

- ML can be a powerful tool in this topic
  - simplify procedure
  - improve precision and add flexibility
- **ML-based approach:** Develop a DNN per jet flavor to minimize the diff b/w data and MC



Jet tagging calibration with ML (II)





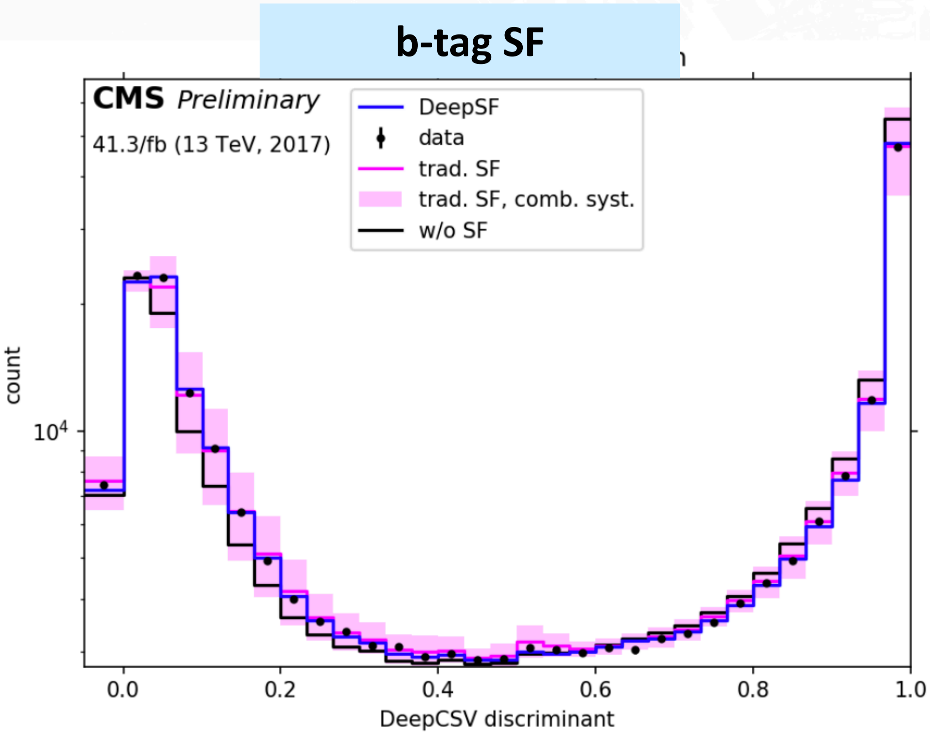
- Traditional SFs: A single value / discriminant bin
- DeepSFs: Different values / discriminant bin due to the direct dependence on p_T(jet), η(jet), etc..
  - Entire procedure repeated 25 times w/ different random seeds
  - **Final DeepSFs:** ensemble of the each of 25 outcomes

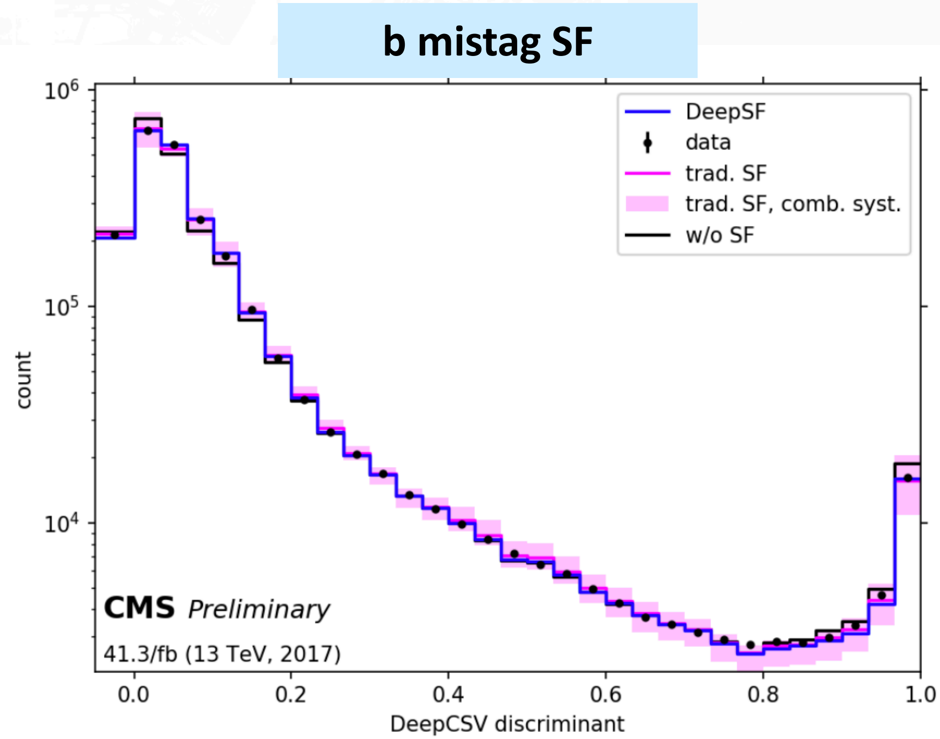
CMS-DP-2019-003

Jet calibration with ML (III)

CMS-DP-2019-003

- Impact on the b-tagging discriminant

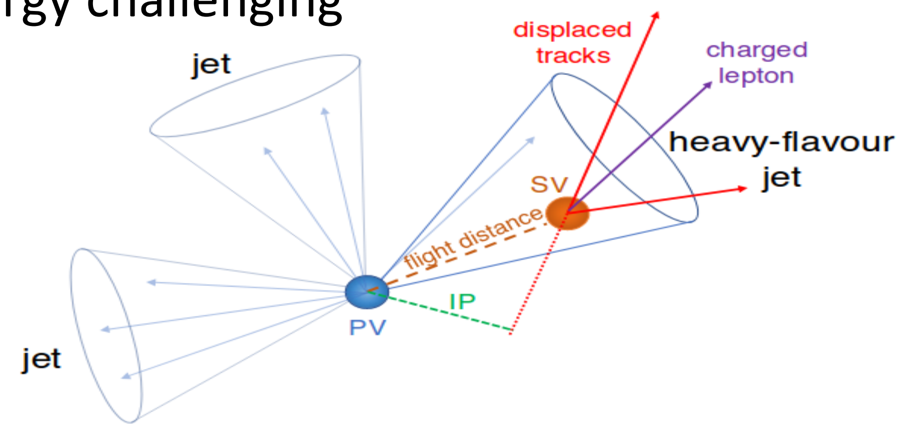




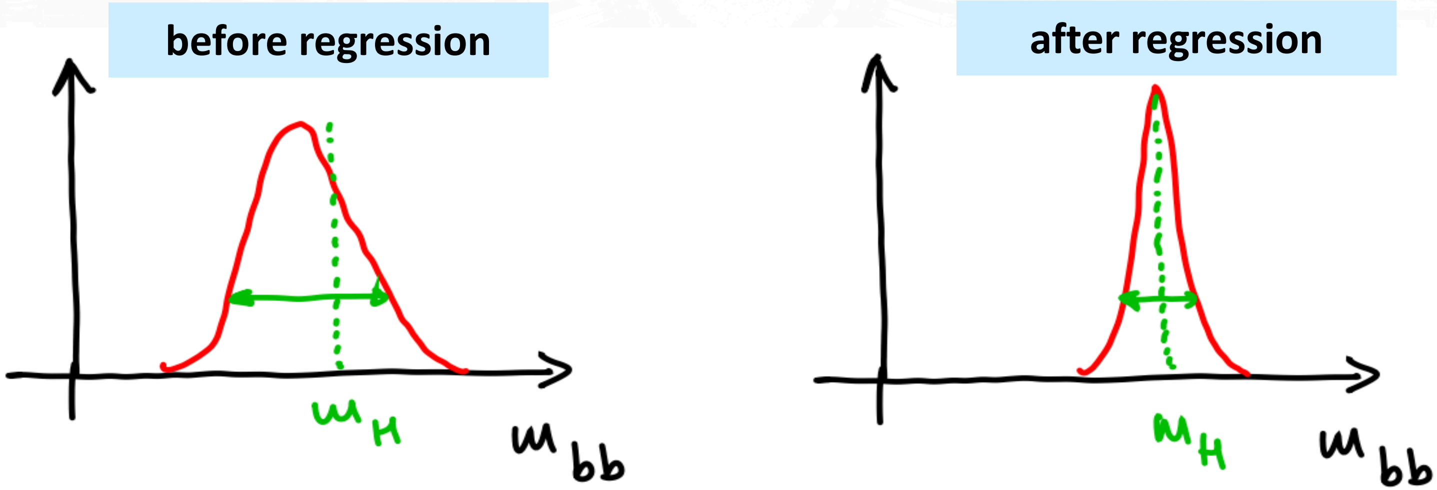
- Consistent results between traditional and ML-based approach
- Work in progress: evaluate full list of systematics + application on physics analyses

Improving b-jet energy scale and resolution

- Precise estimation of the b-quark energy challenging
  - mainly due to energy loss via (undetected) neutrinos from semileptonic decays (~20%)
  - mis-reconstructed tracks and/or tracks outside jet cone, etc..



- **Goal:** Improve b-jet energy scale & resolution
  - Enhanced sensitivity in analyses with b-jets [Higgs, DiHiggs, BSM, etc..]



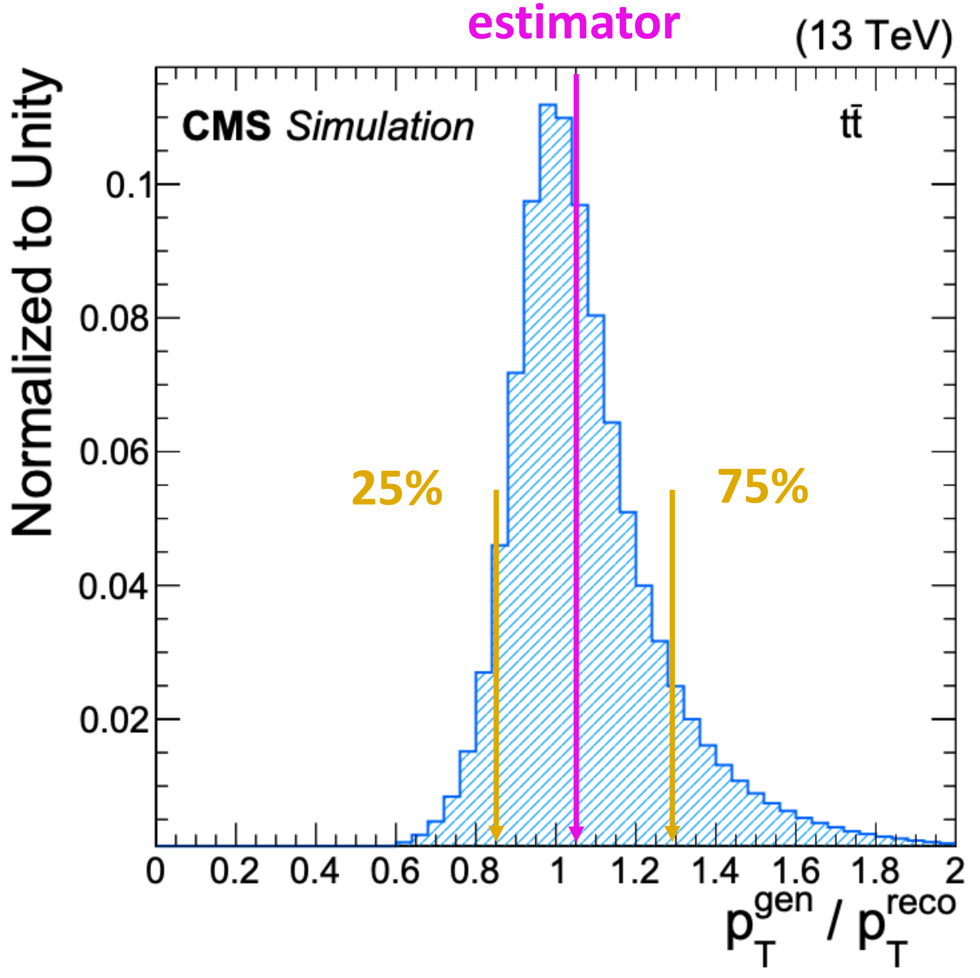
- Previous attempts: shallow ML [BDT] and limited number of inputs
  - New algorithm exploits Deep Neural Networks (DNN)

Algorithm design

CSBS 4 (2020) 10

- Provide an energy scale correction and a jet-by-jet energy resolution:

mean estimator



resolution estimator: (75%-25%)/2

mean estimator: Huber loss

$$H_\delta(z) = \begin{cases} \frac{1}{2}z^2, & \text{if } |z| < \delta; \\ \delta|z| - \frac{1}{2}\delta^2, & \text{otherwise,} \end{cases}$$

resolution estimator: Quantile loss

$$\rho_\tau(z) = \begin{cases} \tau z, & \text{if } z > 0; \\ (\tau - 1)z, & \text{otherwise,} \end{cases}$$

Algorithm design + results

CSBS 4 (2020) 10

Inputs (43 in total):

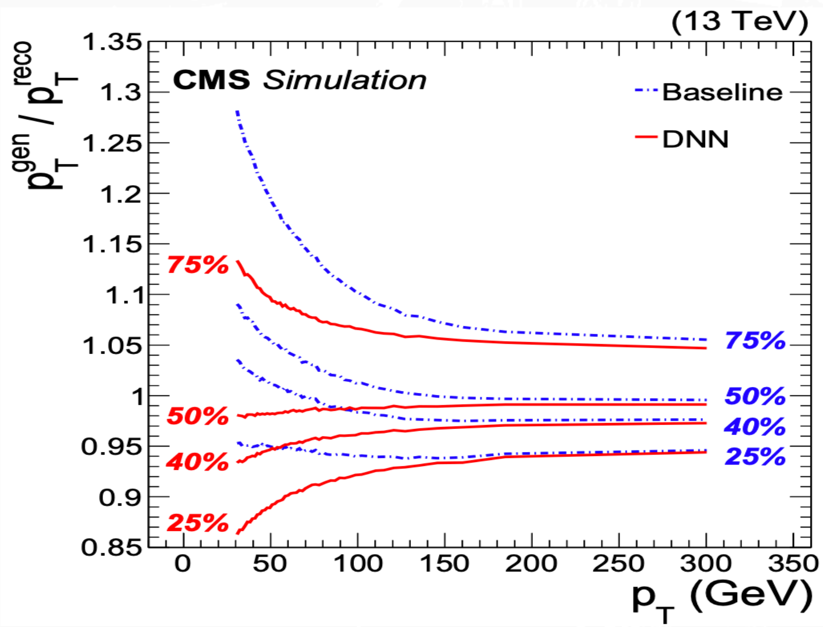
- Jet kinematics
- Jet composition
- PU information
- Info about semi-leptonic decays
- Secondary vertex (SV) properties

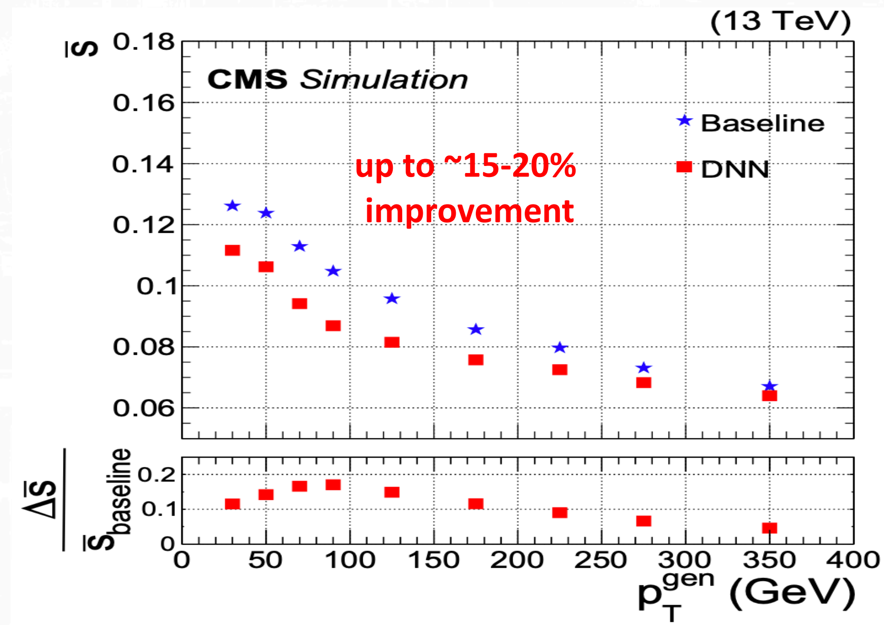
DNN architecture

- Fully connected network
- 6 Layers

Outputs:

- scale correction
- 25% quantile
- 75% quantile



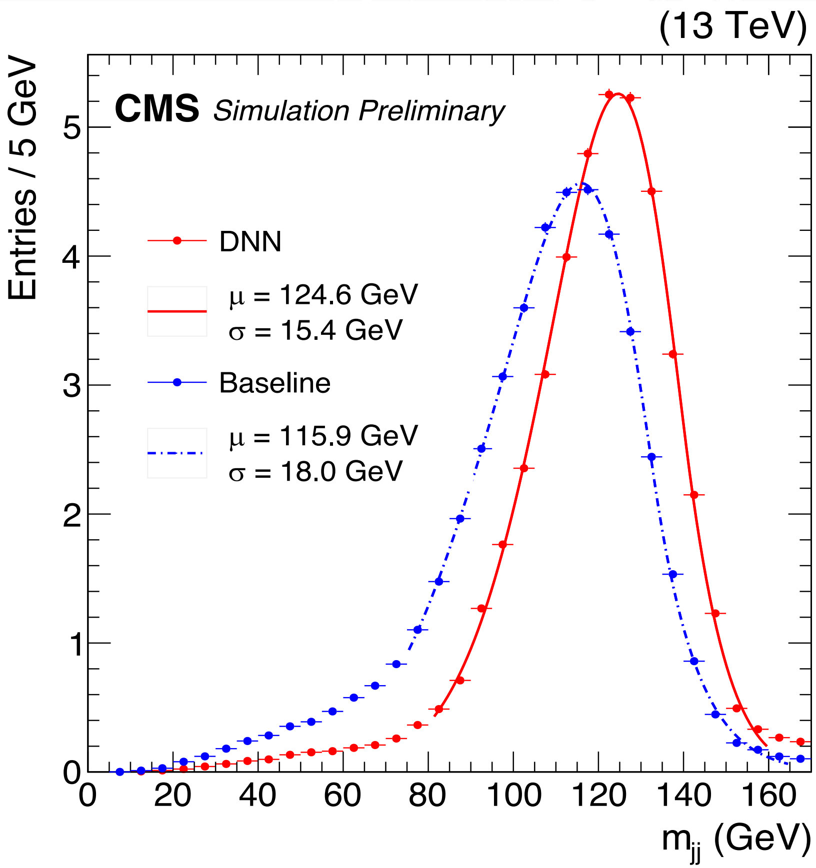


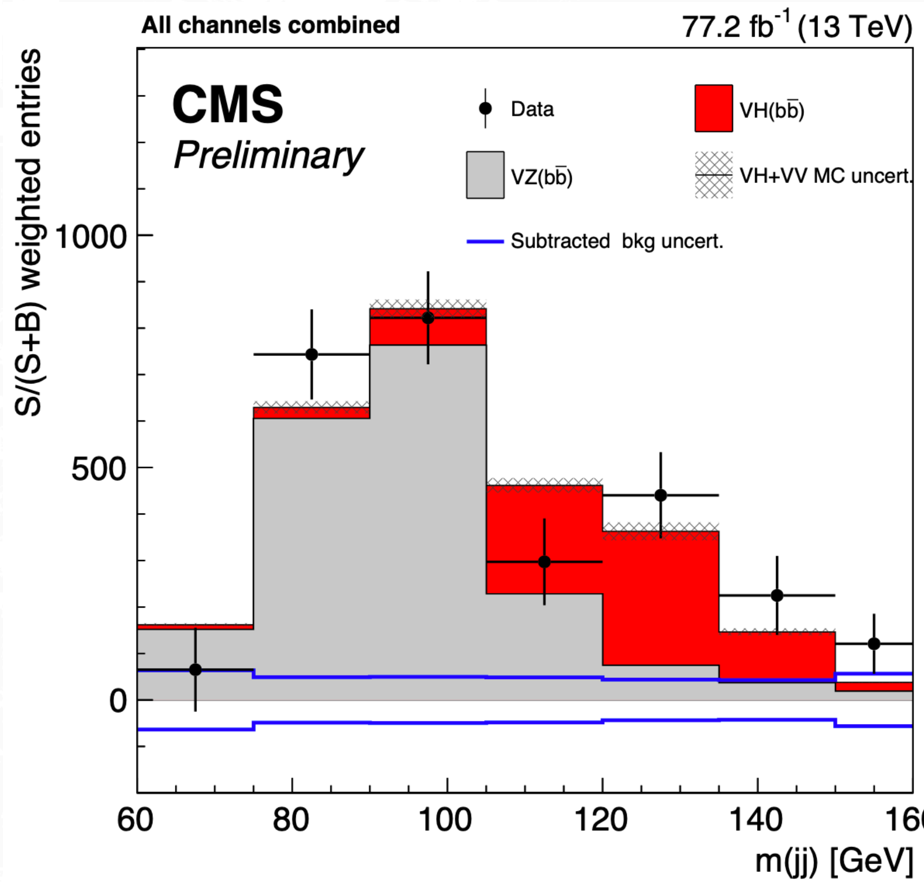
Much flatter jet response and narrower width

Impact on physics analyses

PRL 121, 121801 (2018)

- First application: Observation of the Higgs boson decaying a b-quark pair





Improvement ~20-25% in mass resolution

~10% gain in sensitivity with the DNN-based b-jet energy regression tool

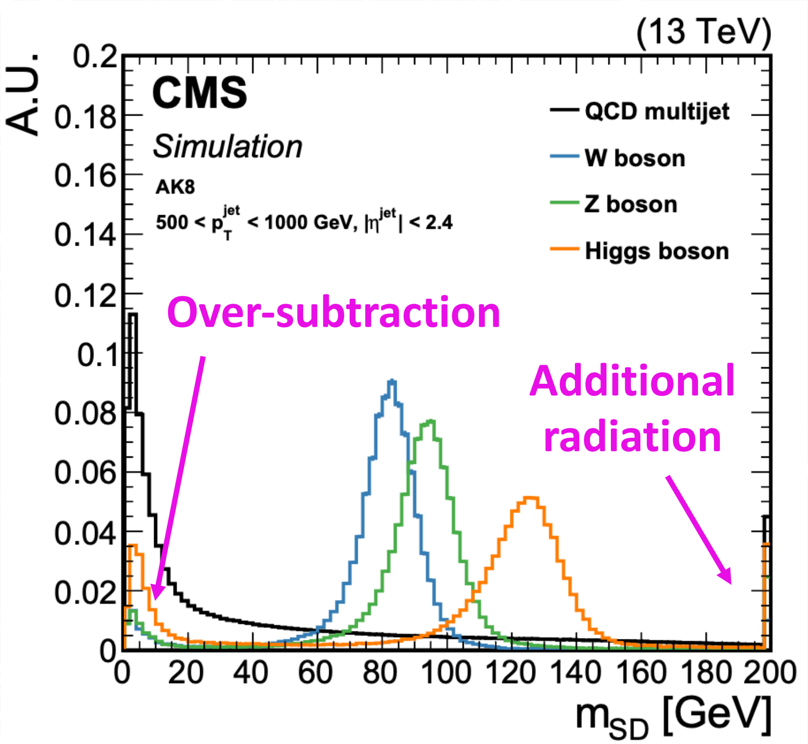
- Similar algorithm developed for c-jets

Large-*R* jet mass regression

New!

- Jet mass: powerful observable for tagging boosted objects
  - but very sensitive to soft radiation, pileup, …

- Grooming techniques [e.g., SoftDrop] have been developed to mitigate this effect:
  - Iteratively decluster the jet and remove constituents that are:
    - soft and/or wide angle
  - Pros: simple and well tested in data
  - Cons: some inefficiency
    - e.g., some two prong jet identified as 1-prong
      - m(jet)-> 0 GeV; Higgs jet ~ QCD jet

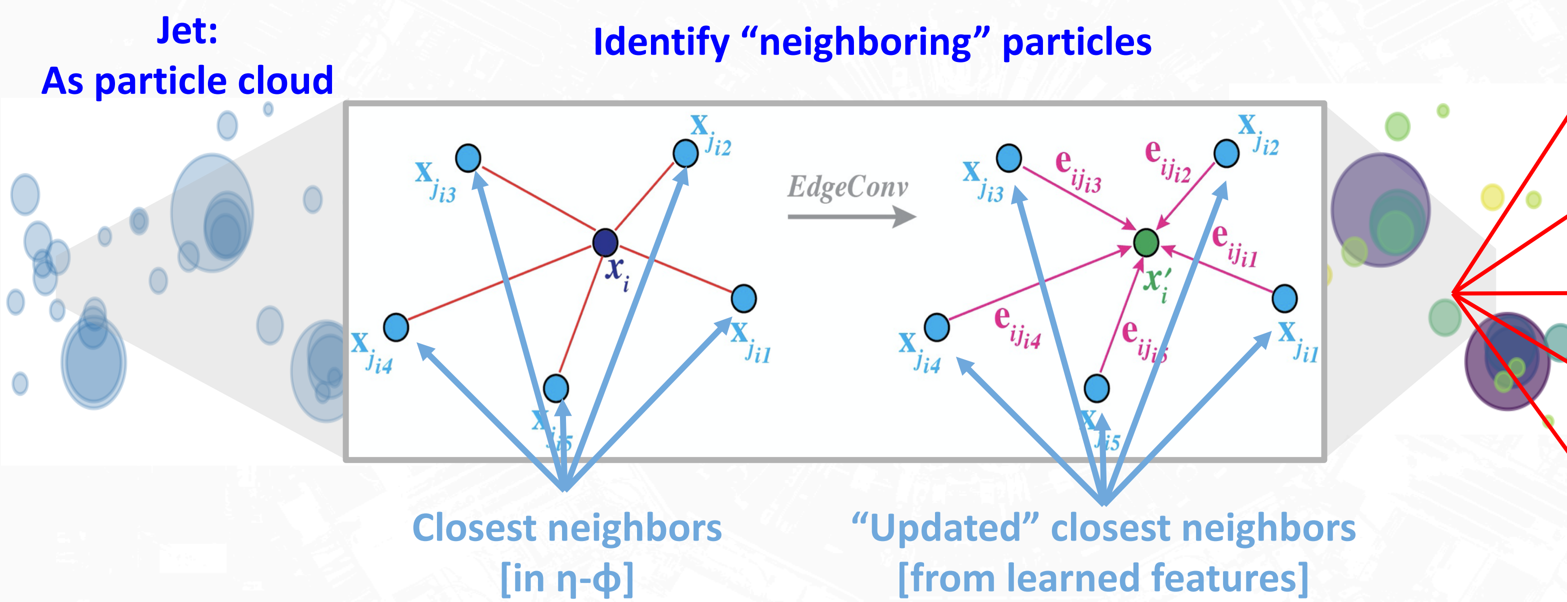


- **Goal:** Develop an algorithm able to reconstruct jet mass with best possible scale and resolution
  - Meanwhile: avoid "sculpting" QCD jet mass distribution

Large-*R* jet mass regression

- Exploit ParticleNet architecture to predict m(jet) directly from jet constituents
  - Same inputs (PF candidates + SV) and same samples as for ParticleNet jet tagging

ParticleNet for Jet Classification



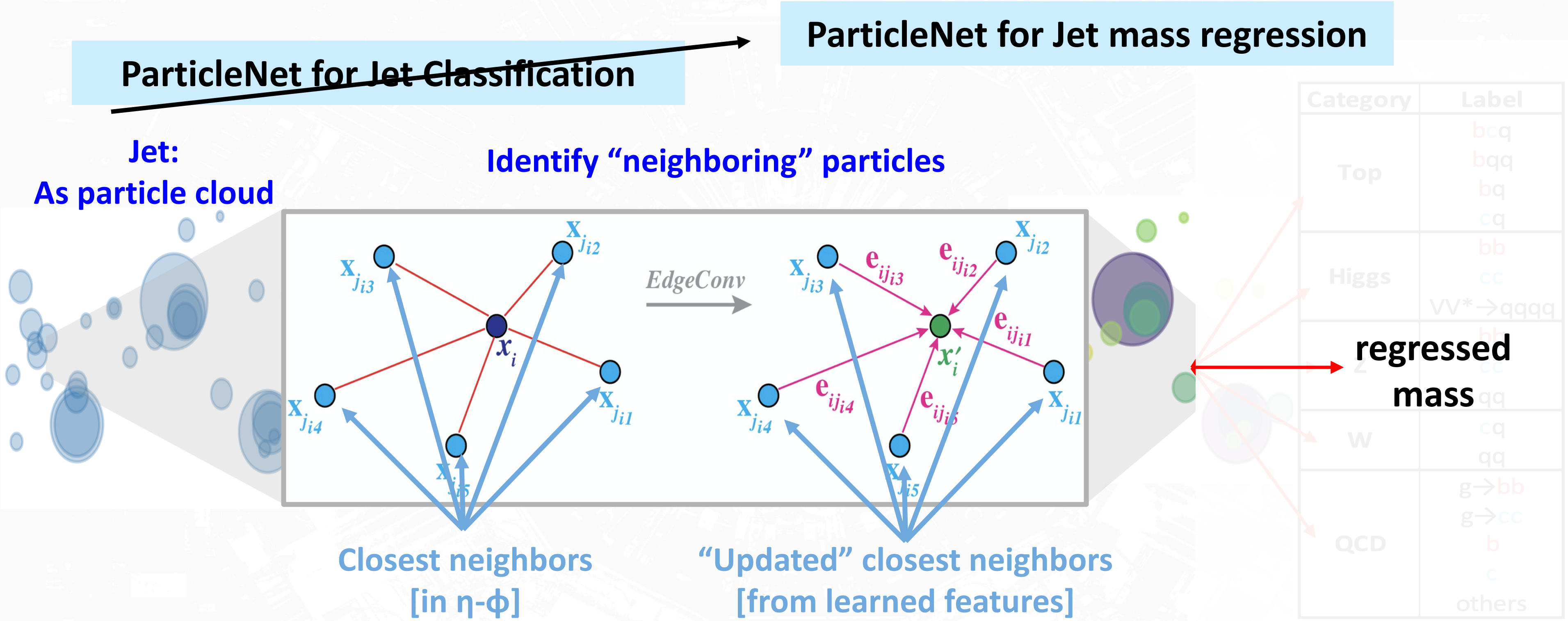
Category	Label
Top	bcq bqq bq cq
Higgs	bb cc VV*→qqqq
Z	bb cc qq
W	cq qq
QCD	g→bb g→cc b c others

PRD 101, 056019 (2020)
CMS-DP-2020-002

Large-R jet mass regression

New! CMS-DP-2021/017

- Exploit ParticleNet architecture to predict m(jet) directly from jet constituents
  - Same inputs (PF candidates + SV) and same samples as for ParticleNet jet tagging

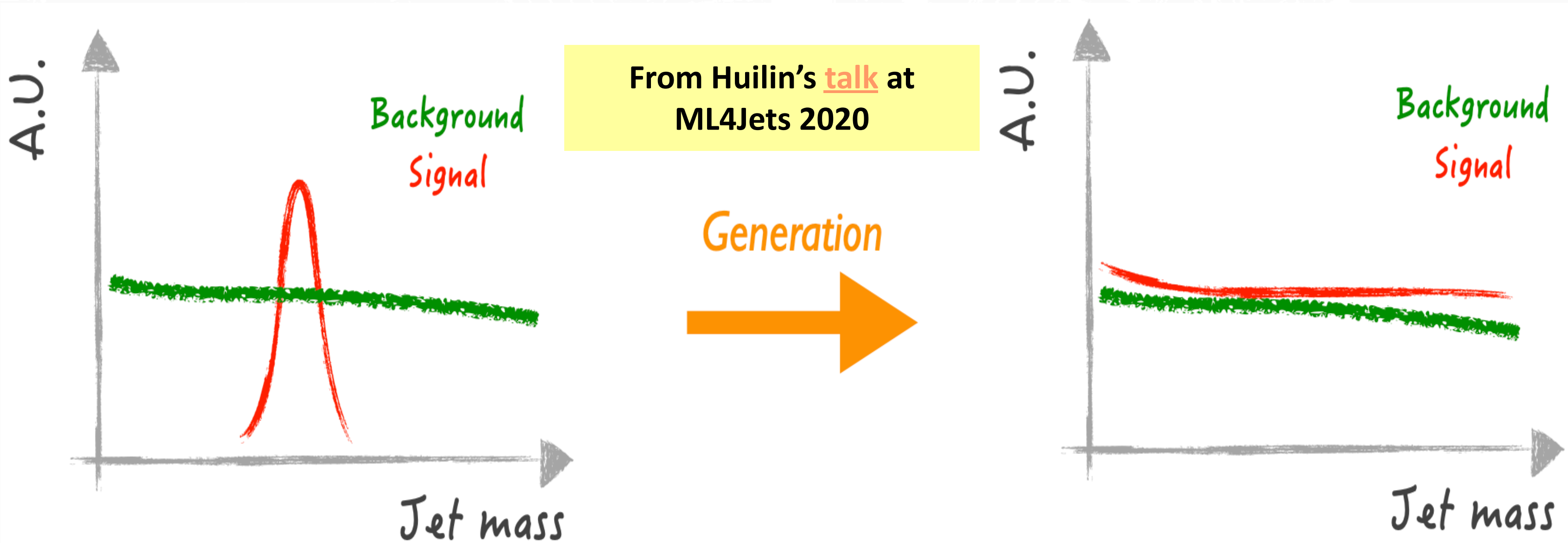


- Focus on Higgs (or generally 2-prong jets) for two jet sizes (R=0.8, R=1.5)

Training details

New!
CMS-DP-2021/017

- **Samples:** Dedicated samples to populate the full mass range
  - equal amount of QCD, X->bb, X->cc, X->qq jets [X: scalar w/ different masses]



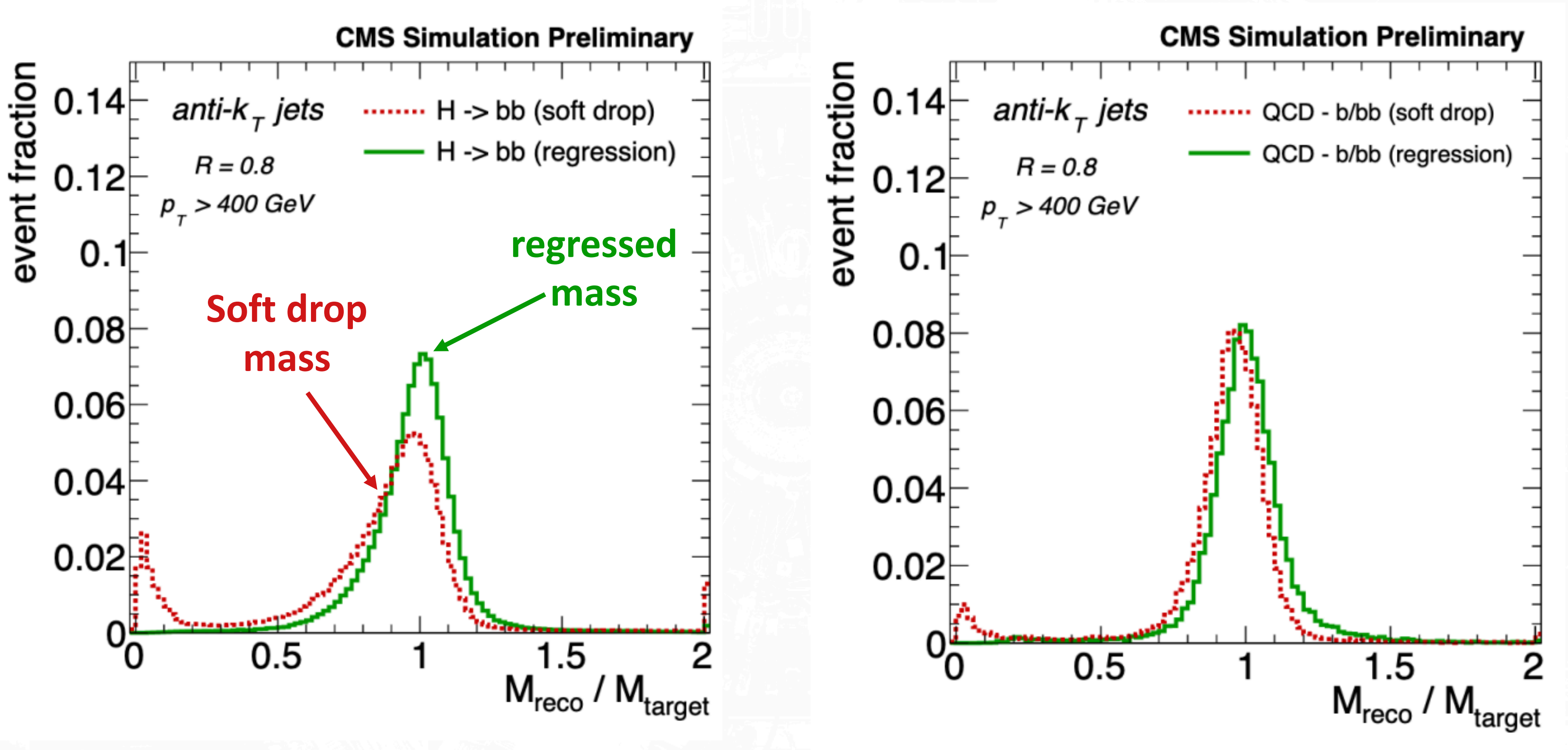
- Target mass:
  - Signal: X pole mass [15-250 GeV]
  - Background: Generated softdrop mass

- Loss function: $$L(y, y^p) = \sum_{i=1}^{n} \log(\cosh(y_i^p - y_i))$$

Performance

New!
CMS-DP-2021/017

Signal jets: H->bb | **Background jets: QCD**

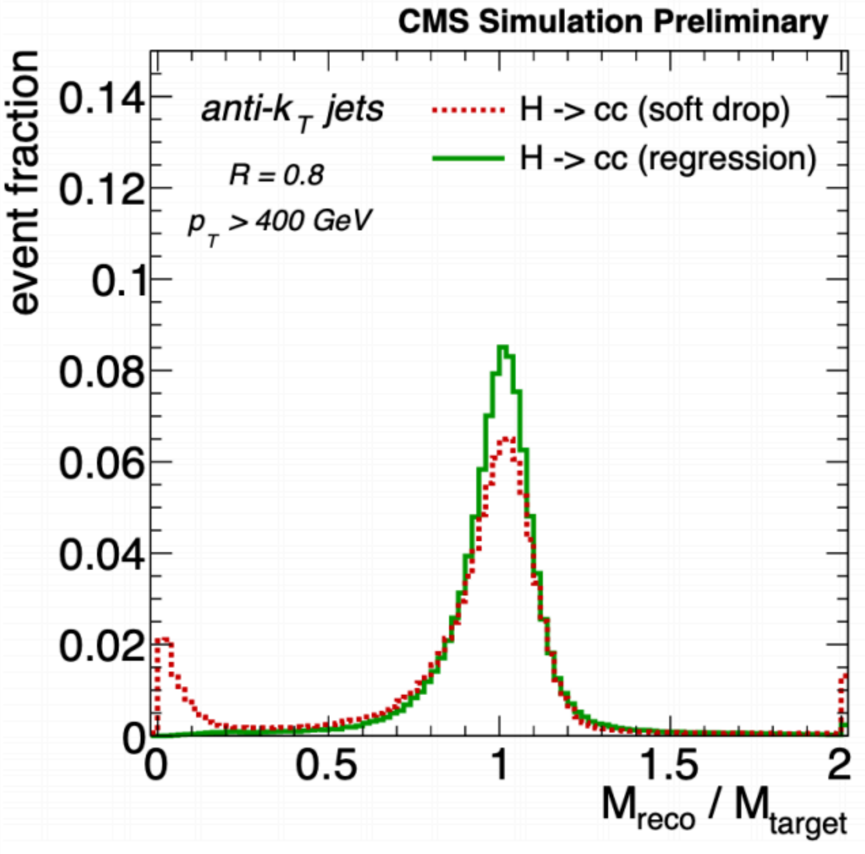


- **Substantial improvement in both mass scale & mass resolution**
- **Tails in m(SD) significantly reduced**

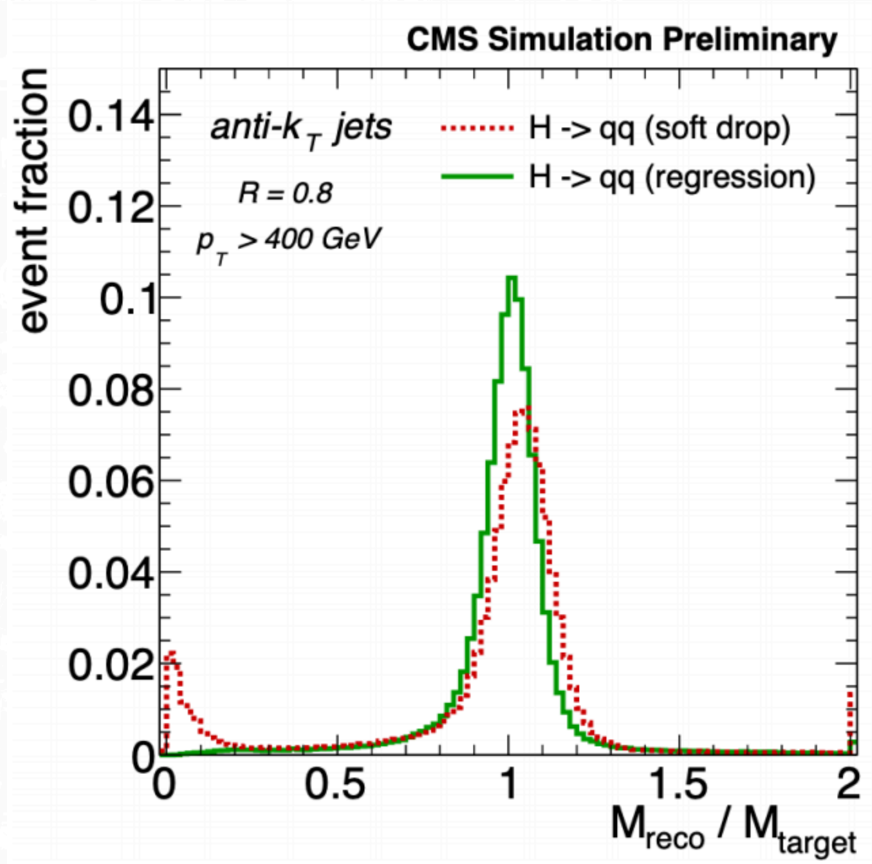
Performance (II)

New!
CMS-DP-2021/017

Signal jets: H->cc



Signal jets: H->qq

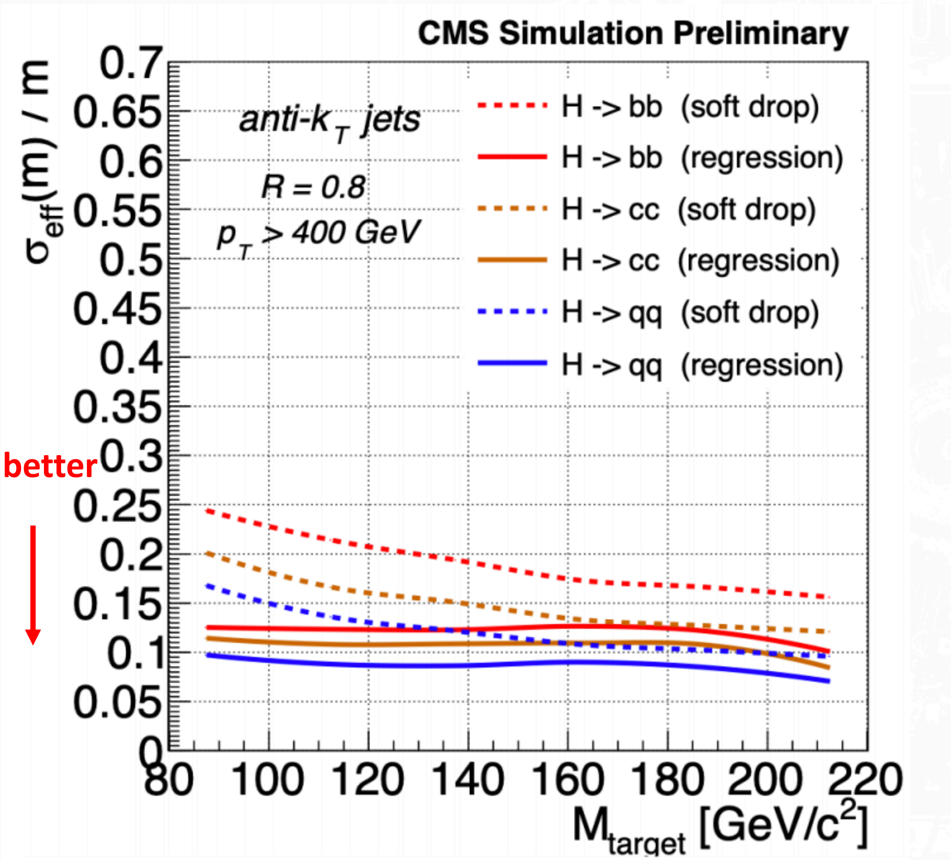


- Improvement for all jet flavours

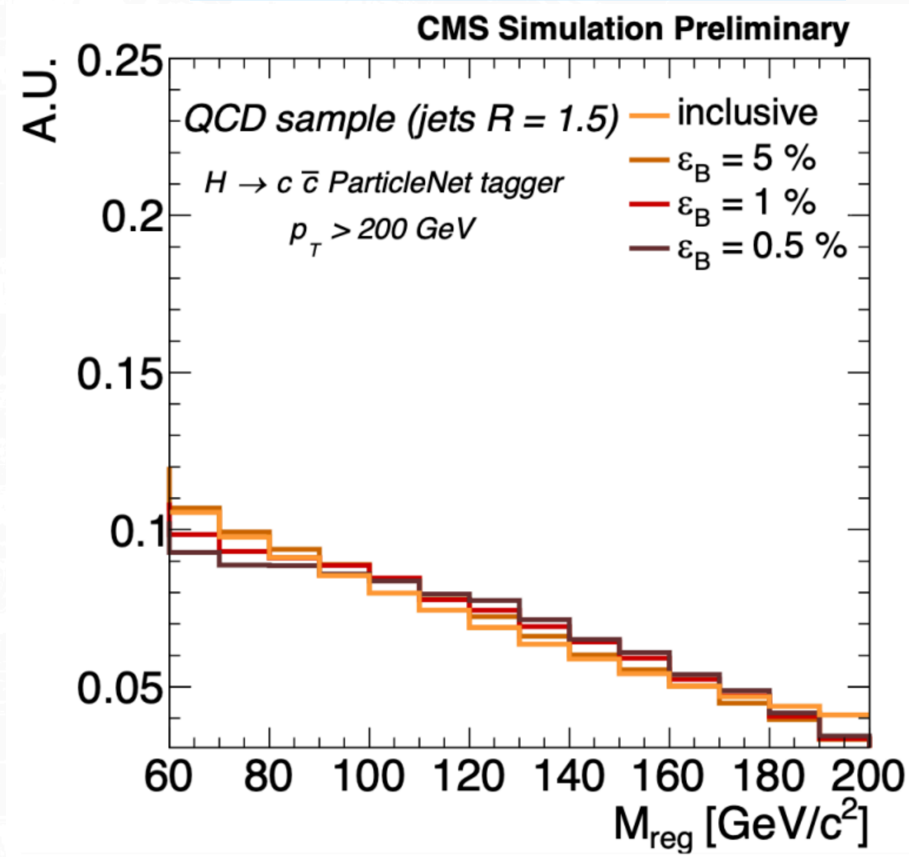
Performance (III)

New!
CMS-DP-2021/017

Mass resolution vs. m(X)



Regressed Mass vs. Tagger WP



- **Mass resolution stable across m(X)**
- **No indication of mass sculpting – even for very tight WPs**
- **Up to ~20-25% improvement in analysis sensitivity with H->bb/cc**

Summary

- Physics with jets essential for the success of the LHC physics program
  - Large effort in both Experiment and Theory communities to improve/extend jet tools

- Major role in these developments: Advanced ML techniques
  - Started with jet flavour tagging
    - showing impressive improvement in performance
  - ML-based application extend to other important jet physics tasks
    - jet calibration, improvement of jet energy and mass reconstruction

- These efforts pays off yielding substantial improvements in physics analyses
  - Still lots of room for improvement / new ideas to try