Image and Sequence Based Jet Tagging Applications on Experiments

Michael Kagan

SLAC

Jets and Jet Substructure at the LHC Workshop

June 3, 2021

ML Jet-Tagging on Experiments

Model Performance
Accuracy

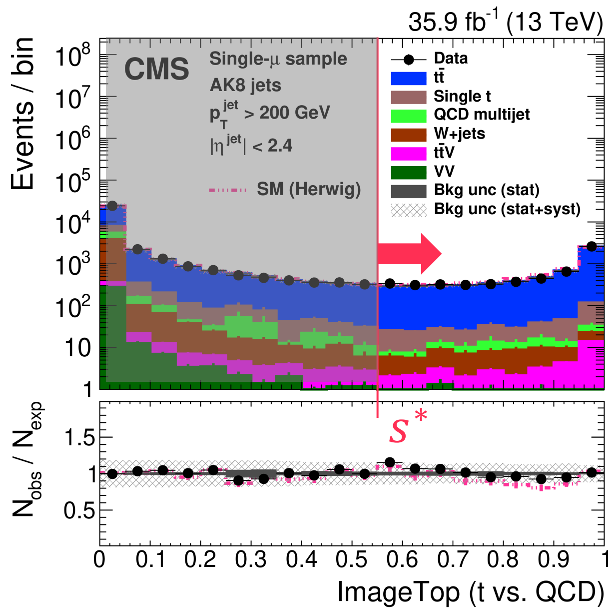
Interpretability
Model & Uncertainties

Calibration
Uncertainties

Calibrating Jet Taggers: Scale Factor

$$SF = \frac{\epsilon_{data}}{\epsilon_{MC}} = \frac{p_{data}(s > s^*)}{p_{MC}(s > s^*)} = \frac{\left(\frac{N^{pass} - N^{pass}_{bkg}}{N - N_{bkg}}\right)_{data}}{\left(\frac{N^{Pass}_{sig}}{N_{sig}}\right)_{MC}}$$

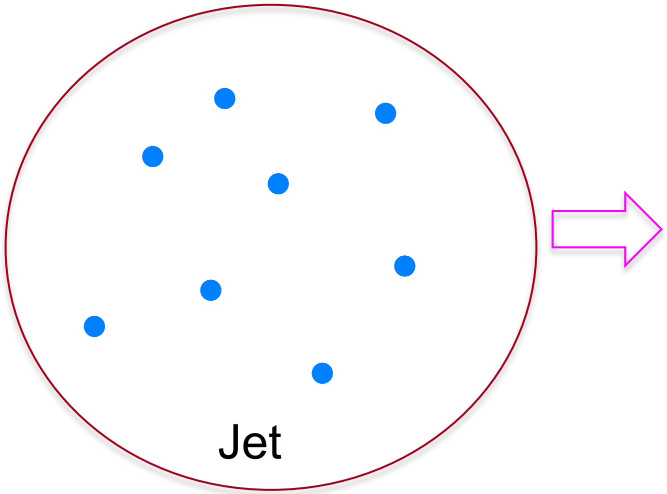
$$p_{data}(s) = \pi_{sig} p_{sig}(s) + (1 - \pi_{sig}) p_{bkg}(s)$$

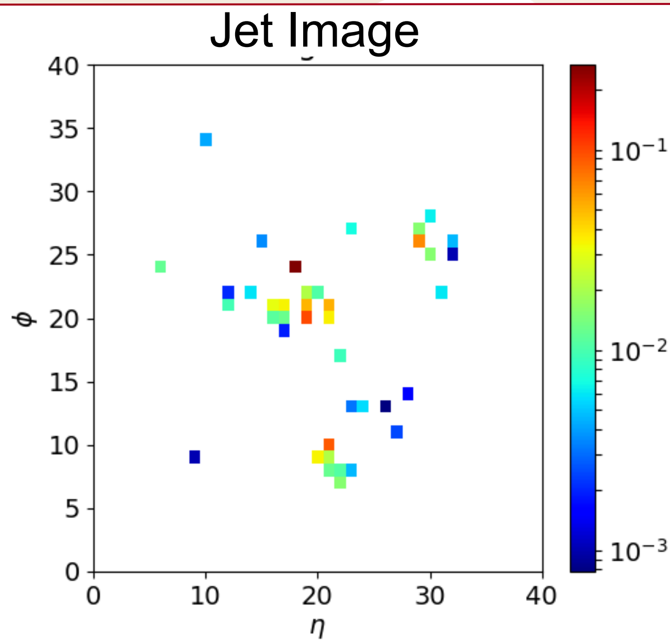


S = Classifier Prediction

- Correct the MC efficiency of a cut on classifier output
  - Little insight into “why” a scale factor deviates from unity

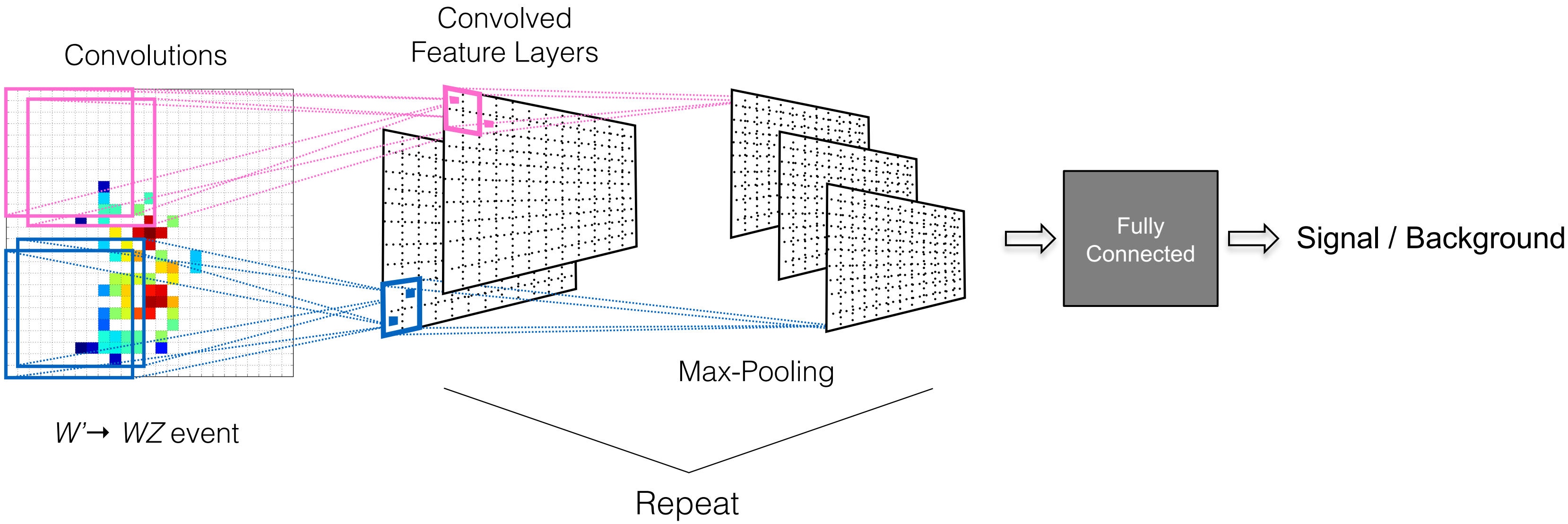
Computer Vision and Jets



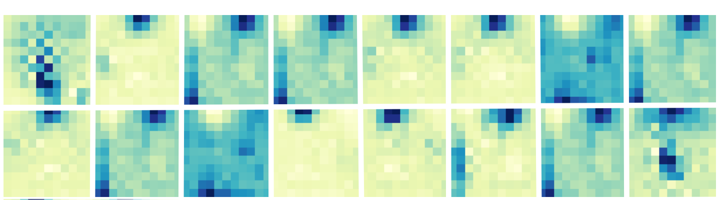


- **Jet-image** –2D representation of jet as distribution of energy over $\eta - \phi$
- Multi-channel "color" jet images – separate images for different detectors (calorimeter / track) or particles (charged / neutral hadrons, muon, etc.)
- How to deal with track images?
  - More pixels may improve performance
  - Cost: larger models and more memory needed

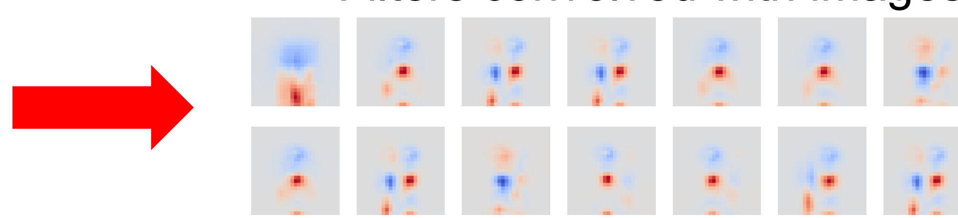
Convolutional Neural Network



Filters

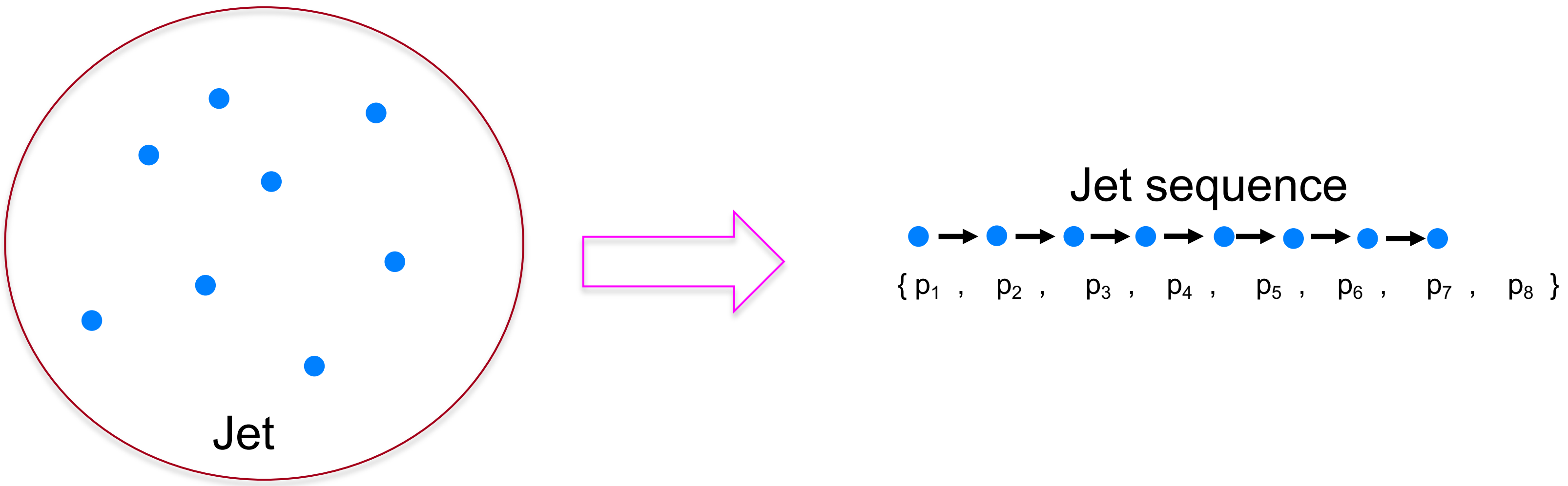


Filters convolved with images



[arXiv:1407.5675]
[arXiv:1511.05190]

Sequence Modeling



- Jets are a grouping of a variable number of particles
- With physically motivated ordering: **jet as a sequence**

Recurrent Neural Network

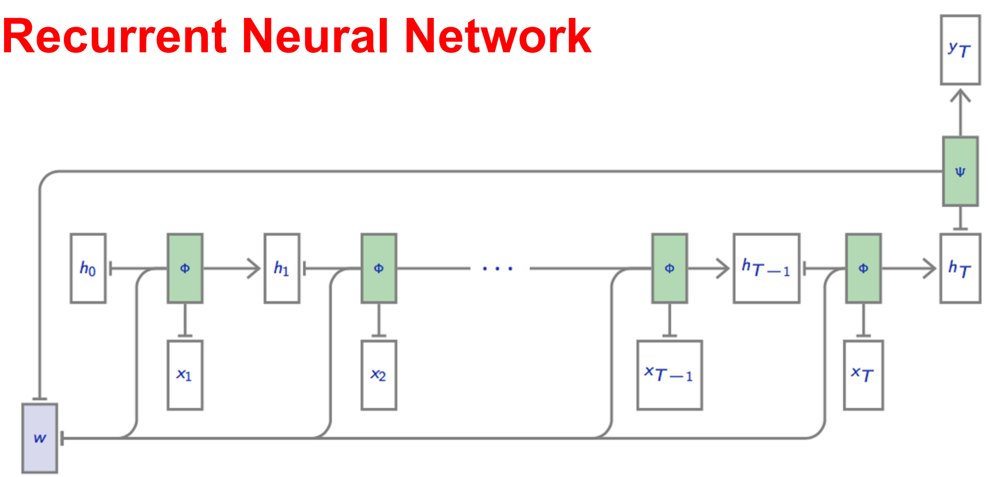
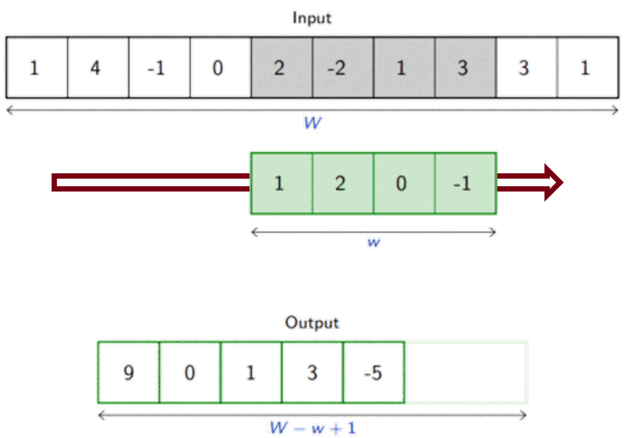
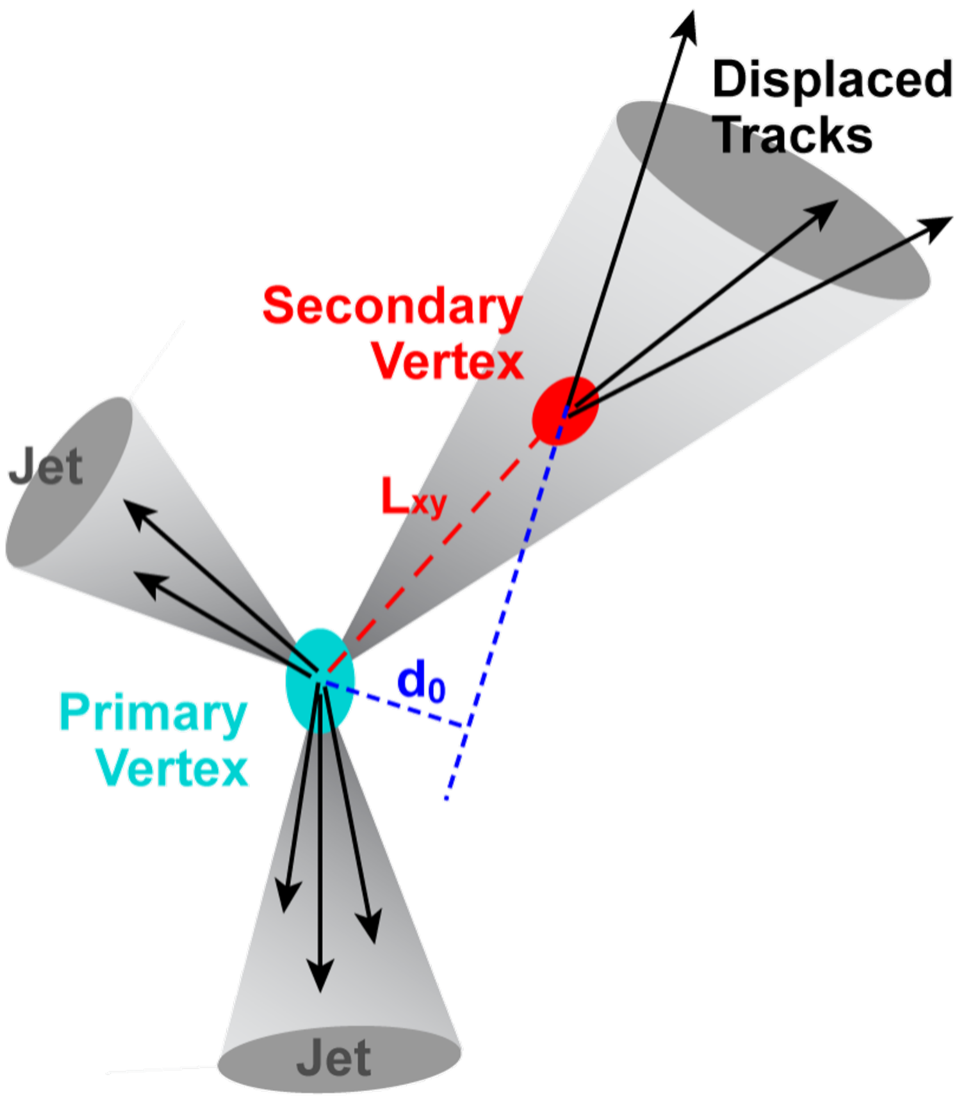


Image Credit: Fleuret, Deep Learning Course

1D Convolutions

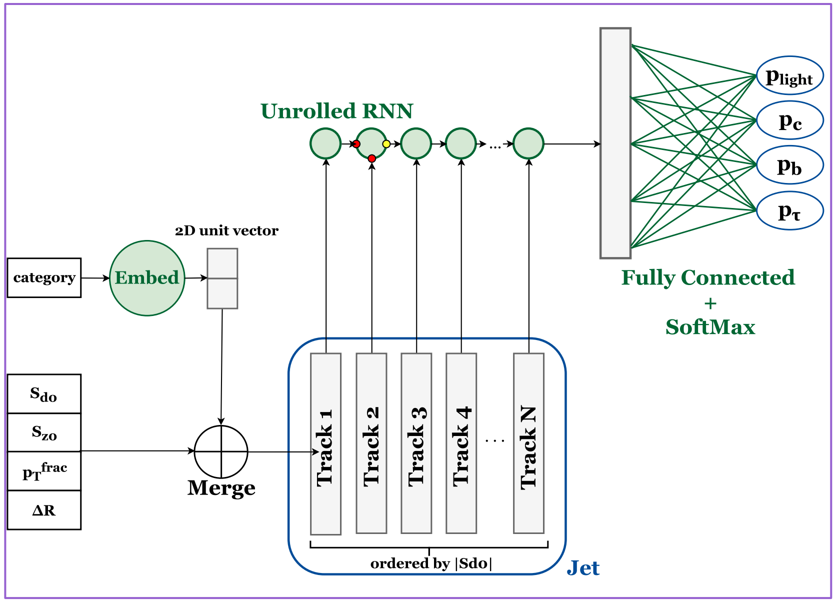


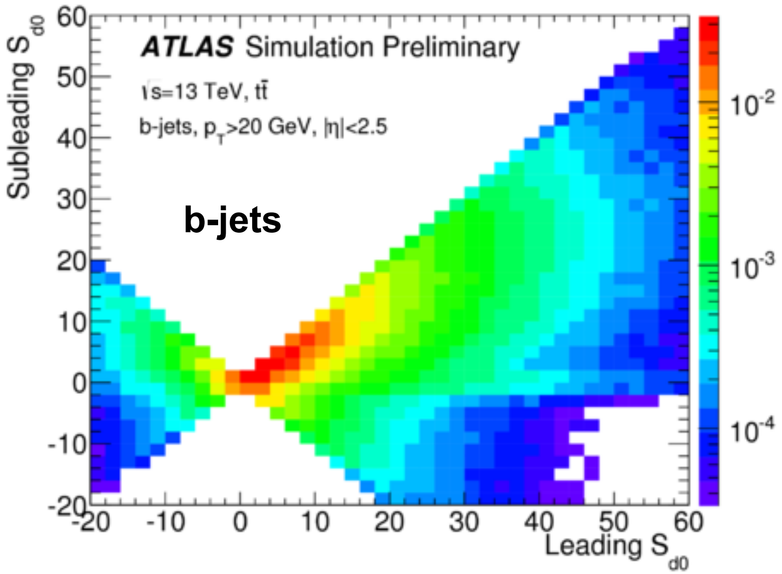
B-tagging

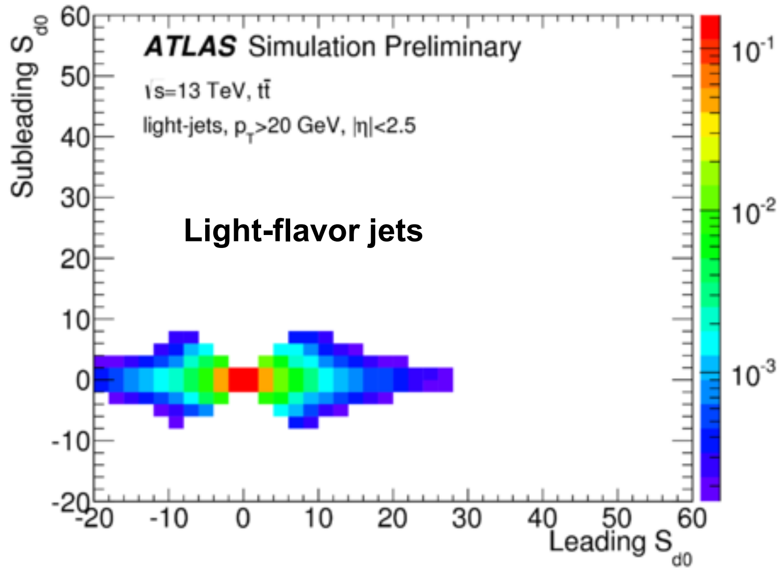

Displaced
Tracks
Secondary
Vertex
Jet
Lxy
d0
Primary
Vertex
Jet

Recurrent Neural Net b-tagging - RNNIP

ATL-PHYS-PUB-2017-003

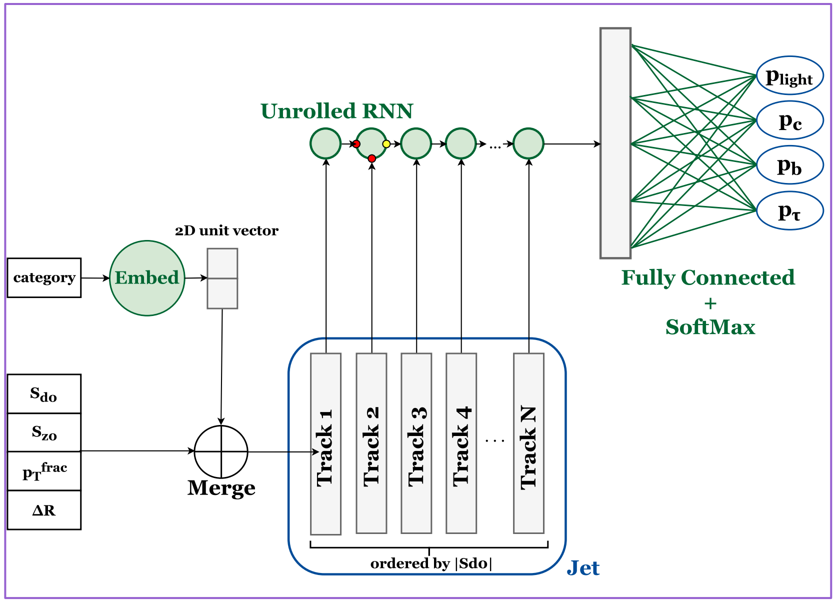


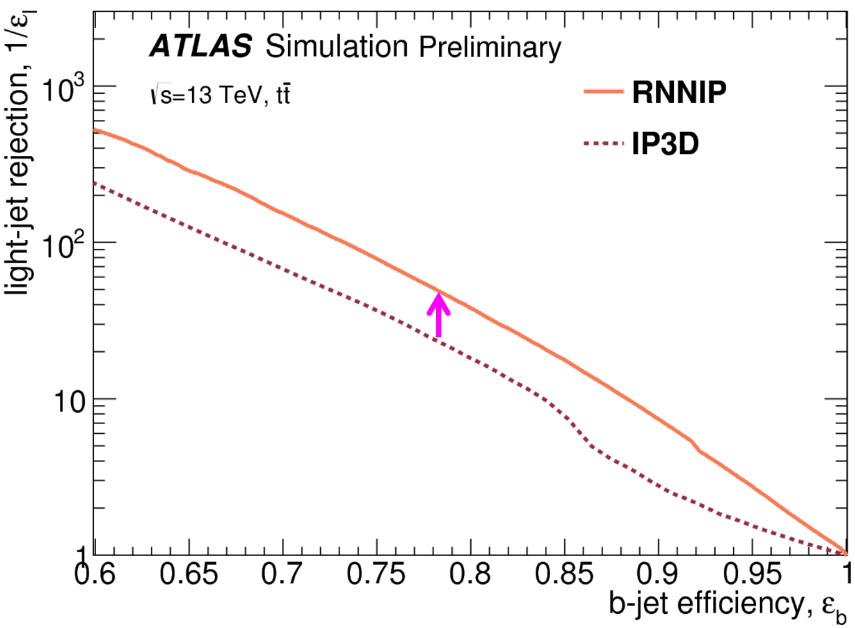


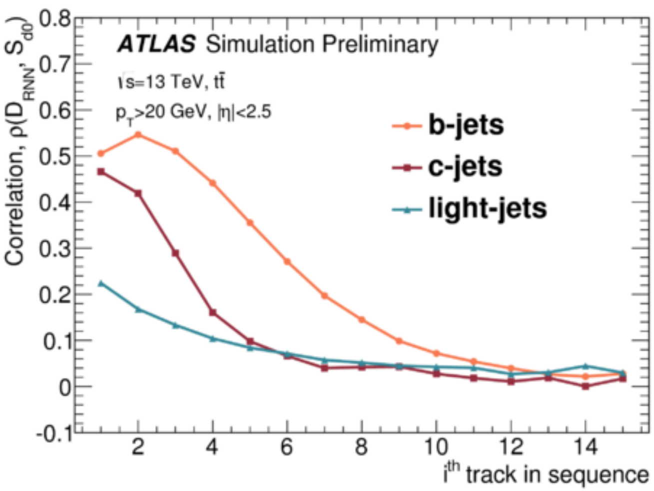


Recurrent Neural Net b-tagging - RNNIP

ATL-PHYS-PUB-2017-003

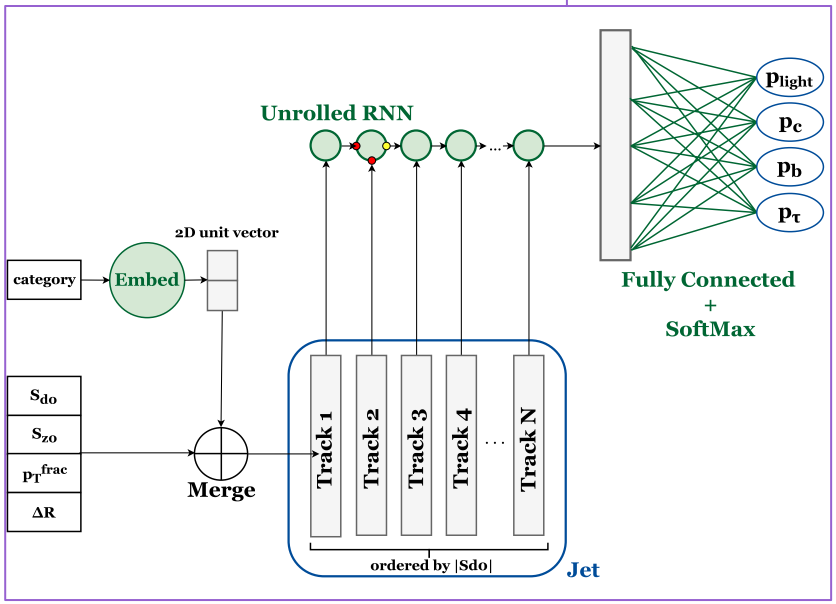


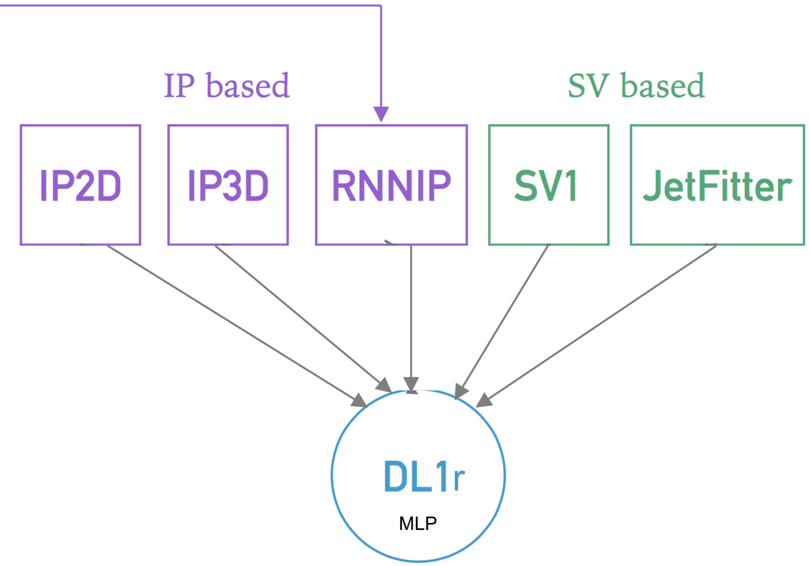


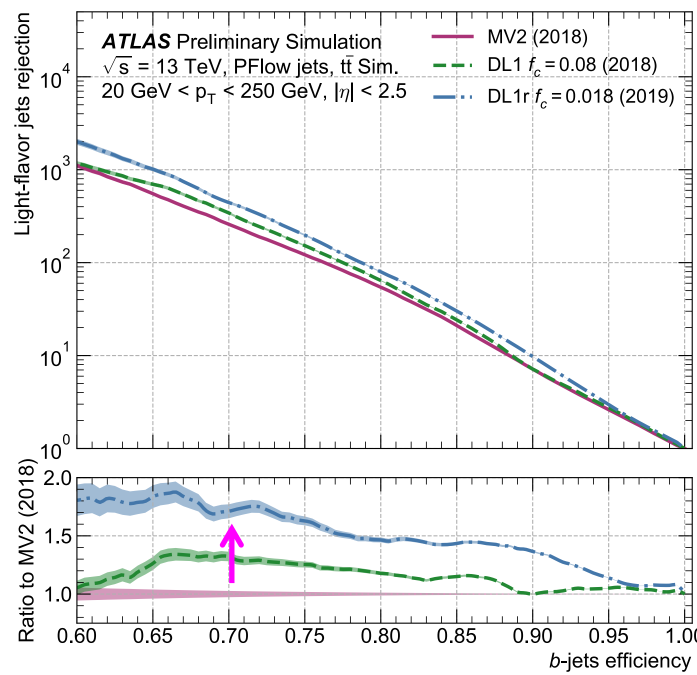


DL1r

ATLAS-PHY-PLOTS-FTAG-2019-005

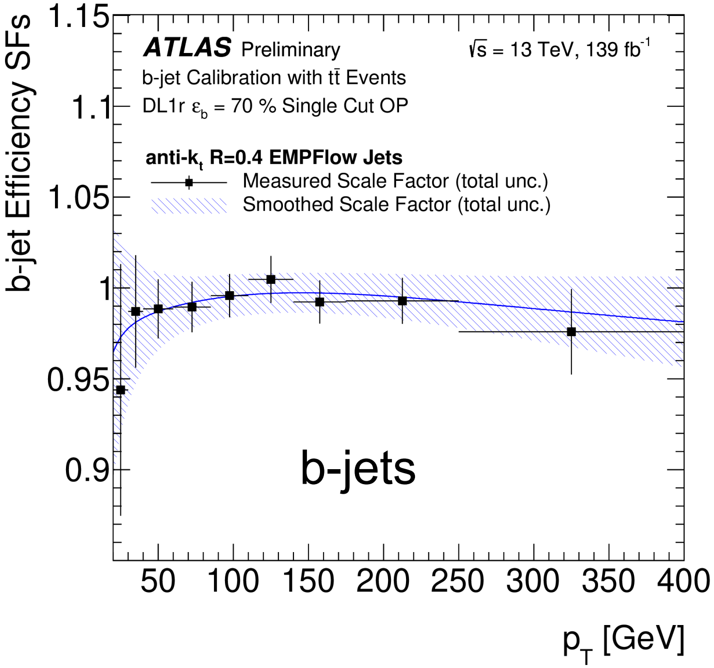


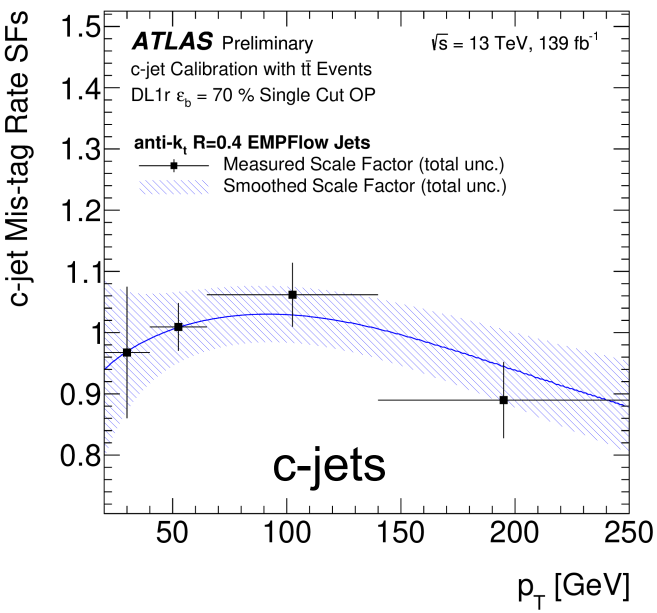


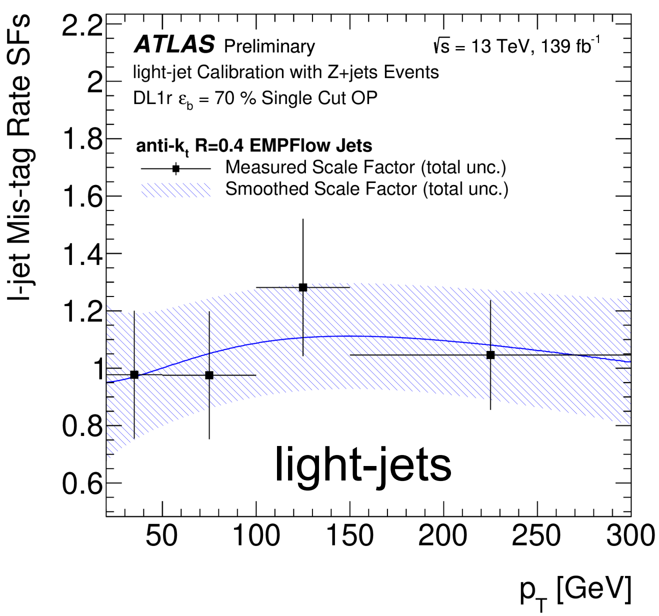


DL1r Calibration

ATLAS-PHY-PLOTS-FTAG-2021-001

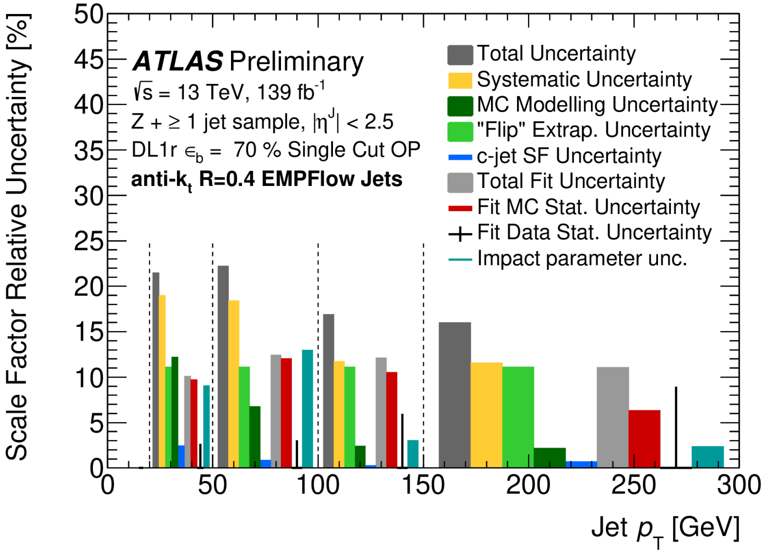






DL1r Calibration Uncertainties

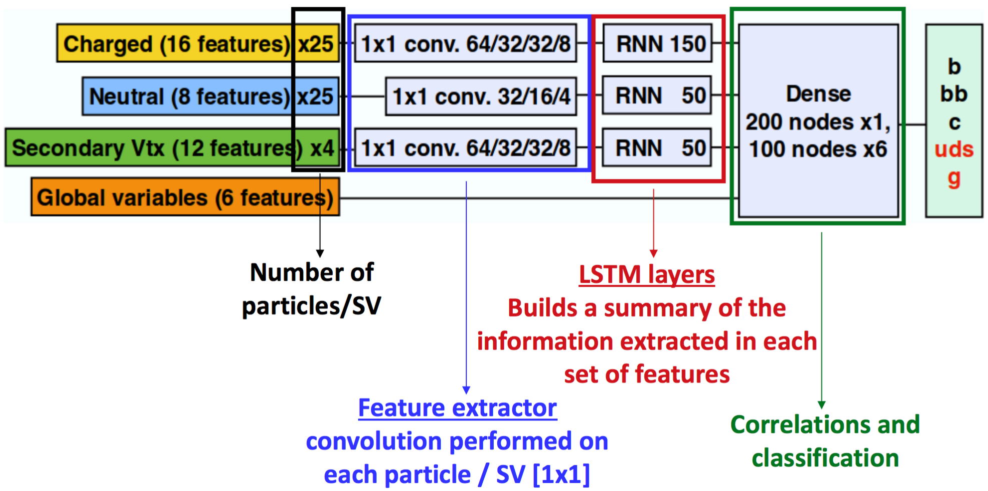
- Theory modeling among largest uncertainties
- Significant recent SF reduction, independent of tagger, e.g. from
  - Improving charge deposition modeling in Silicon
  - Better method to estimate sample flavour composition
- Still difficult to separate "Model is learning bad correlations" from imperfect calibration methods

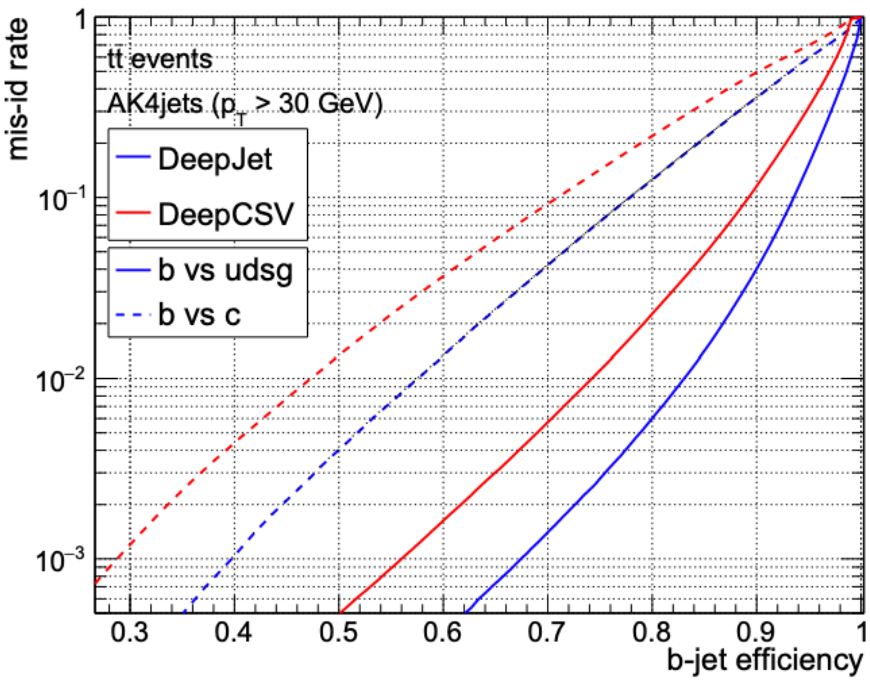


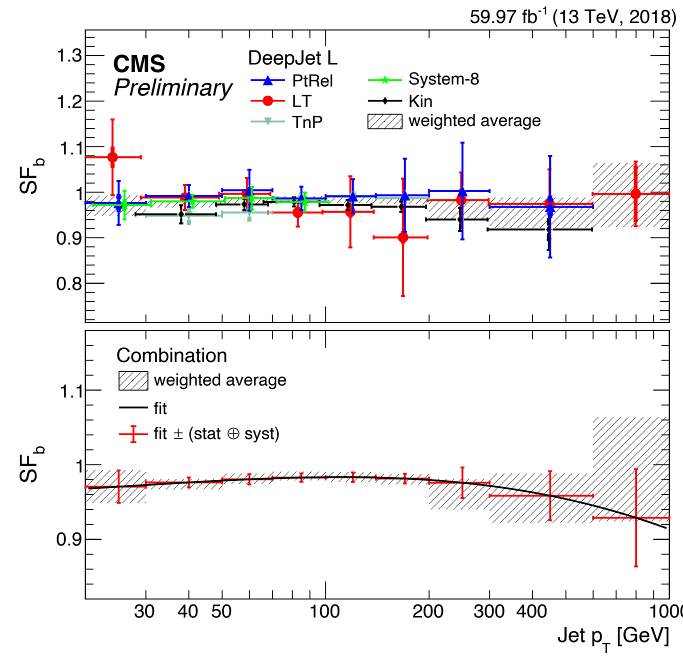
ATLAS-PHY-PLOTS-FTAG-2021-002

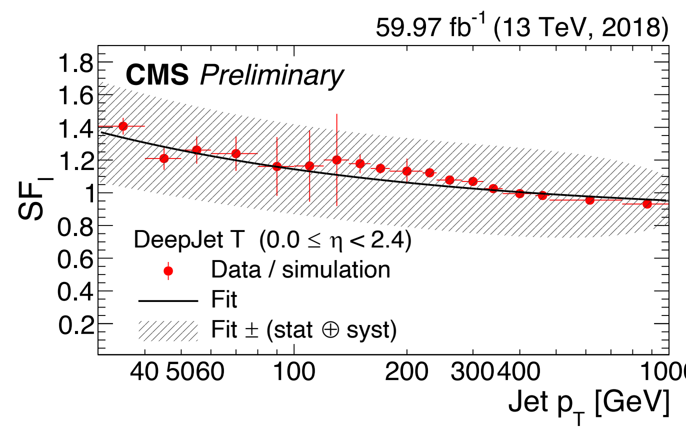
DeepJet

[JINST 15 (2020) P12012](#)
[CMS-DP-2018-058](#)
[CMS-DP-2021-004](#)









Boosted Jet Tagging

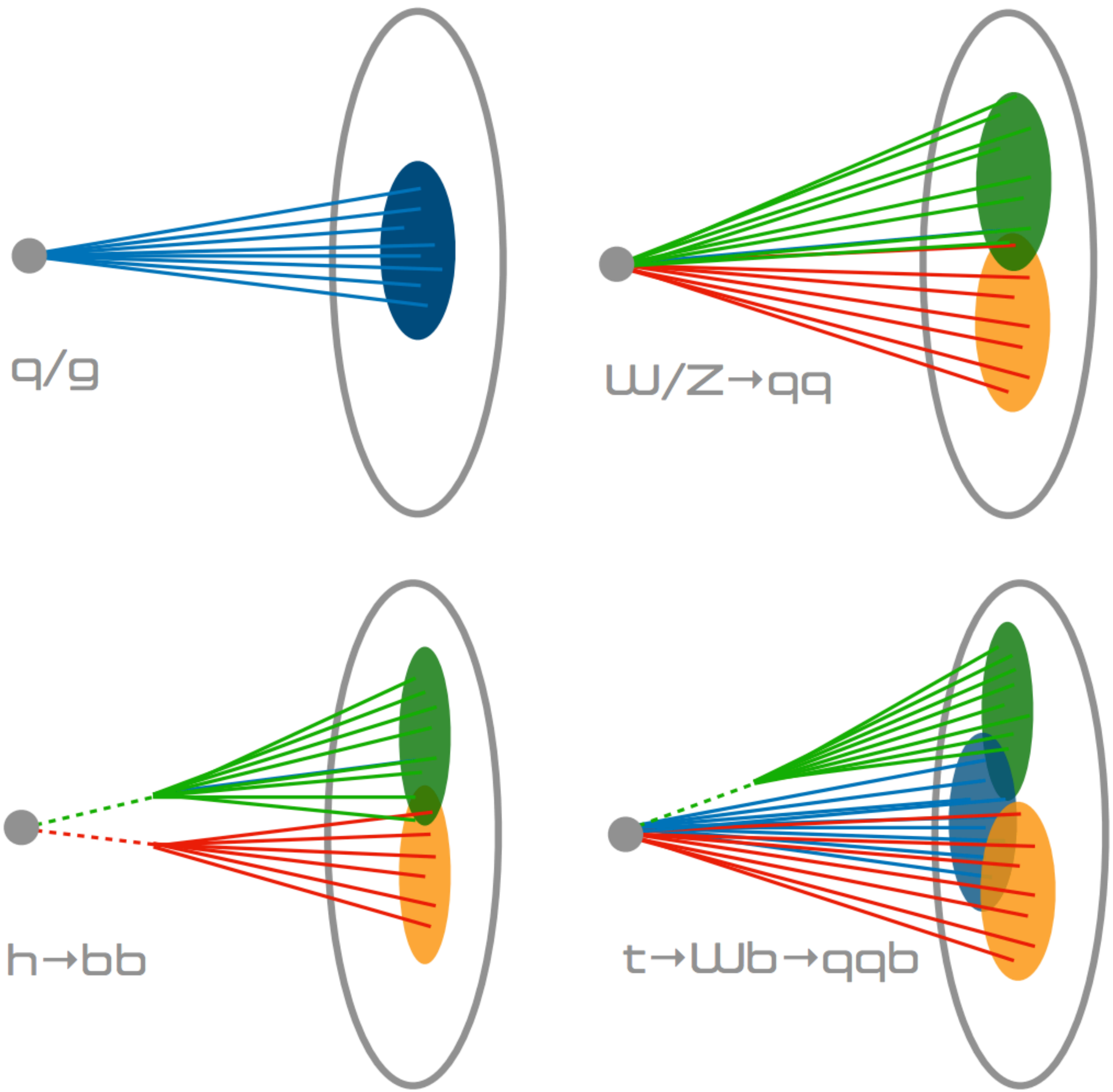
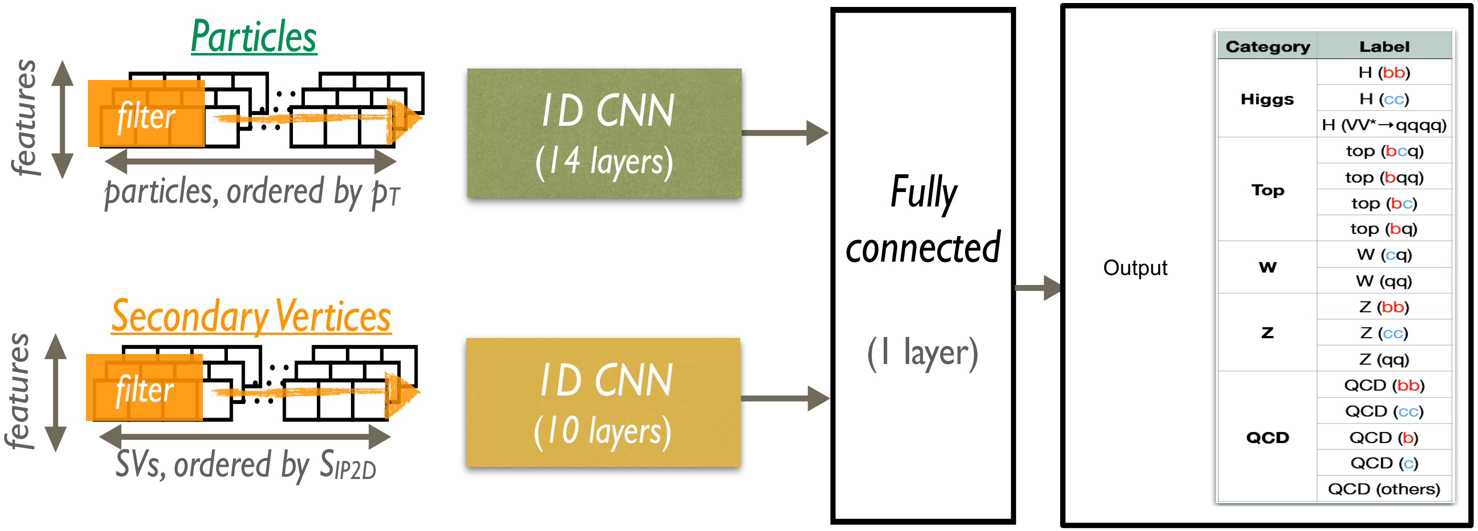


Image Credit: arXiv:1909.12285

ML Boosted Jet Taggers on CMS

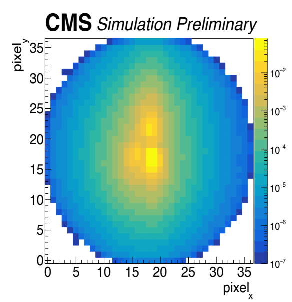
JINST 15 (2020) P06005

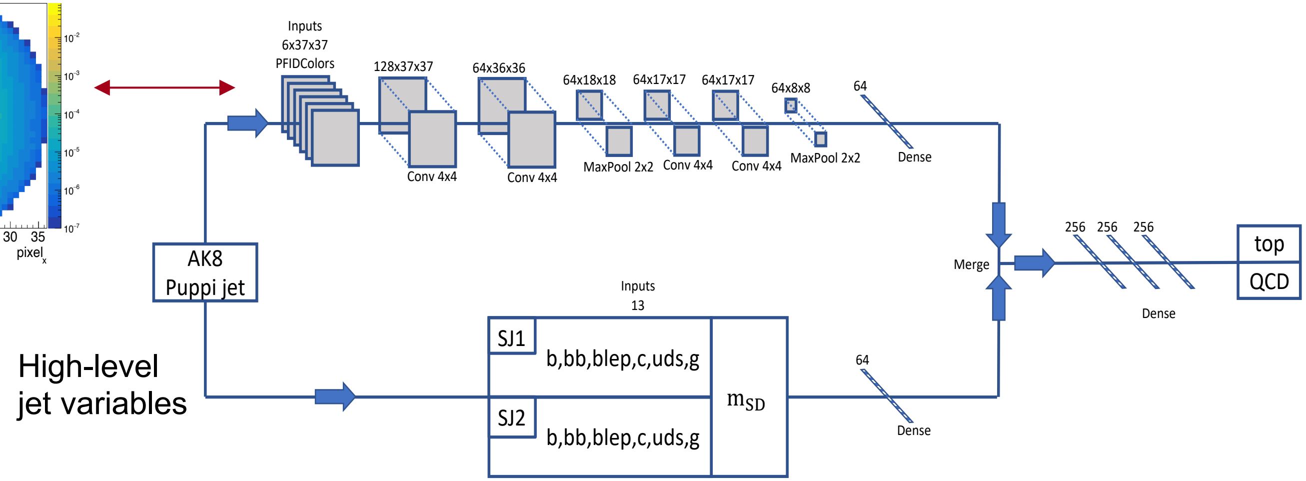
Deep AK8



ImageTop

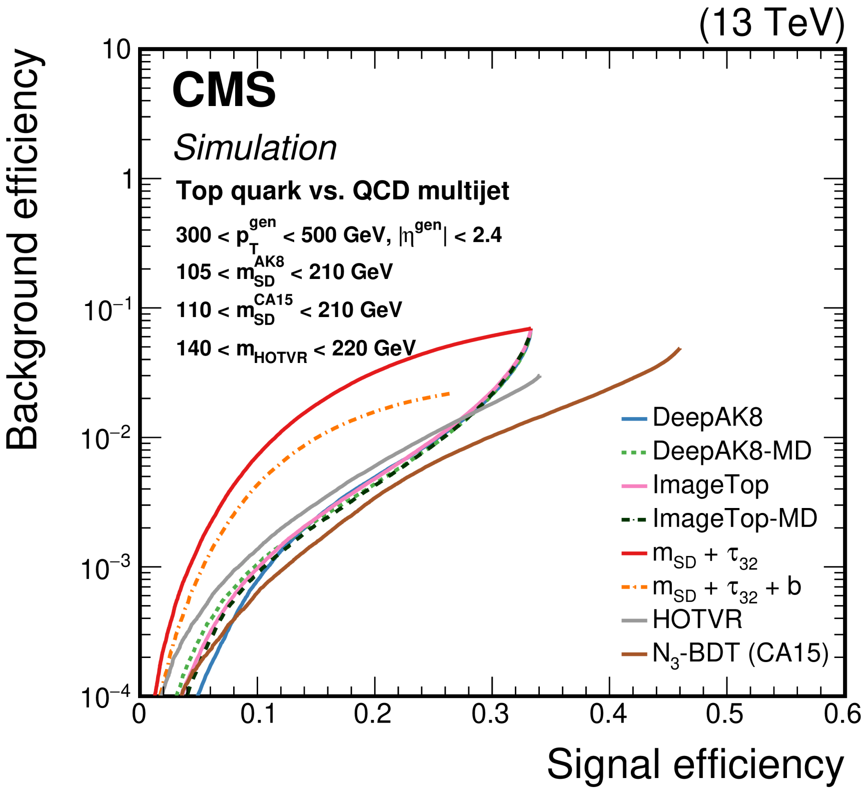
- 6 channels
  - All PF candidates
  - Charged hadron
  - Neutral hadron
  - Photon
  - Electron
  - Muon

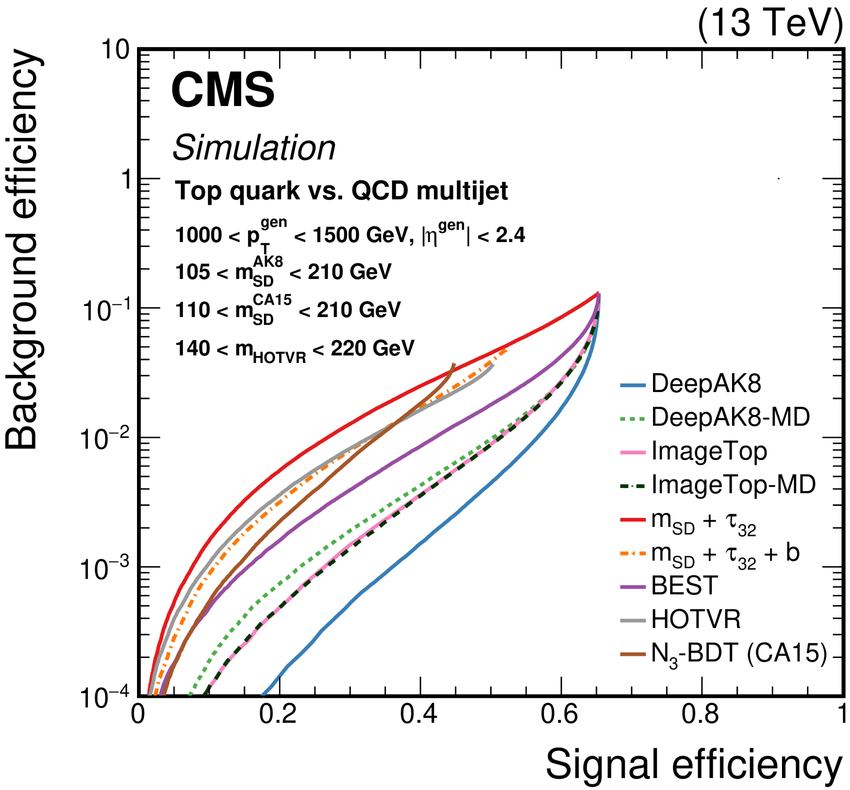




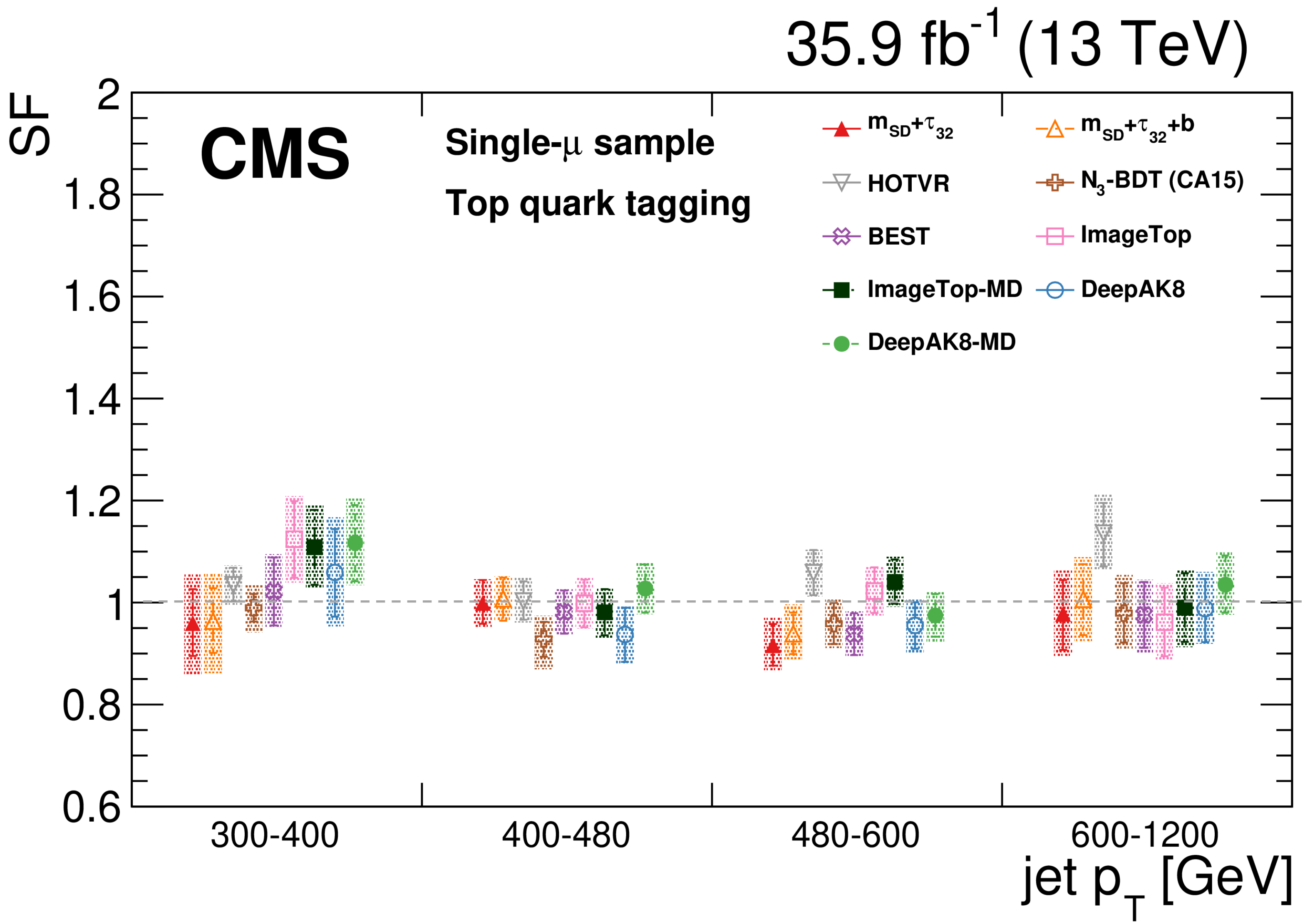
Tagger Performance

JINST 15 (2020) P06005



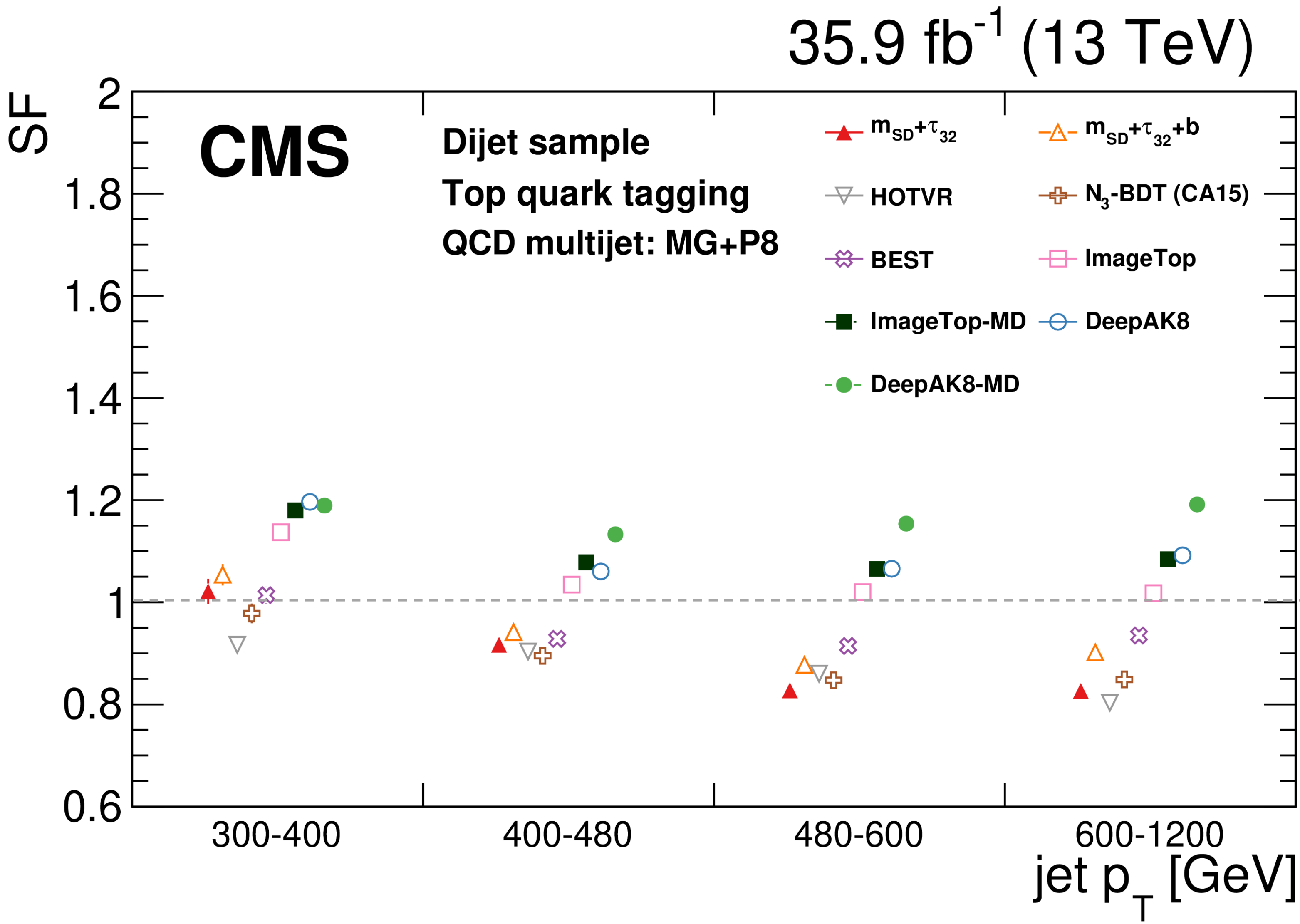


Top Tagging Scale Factors



JINST 15 (2020) P06005

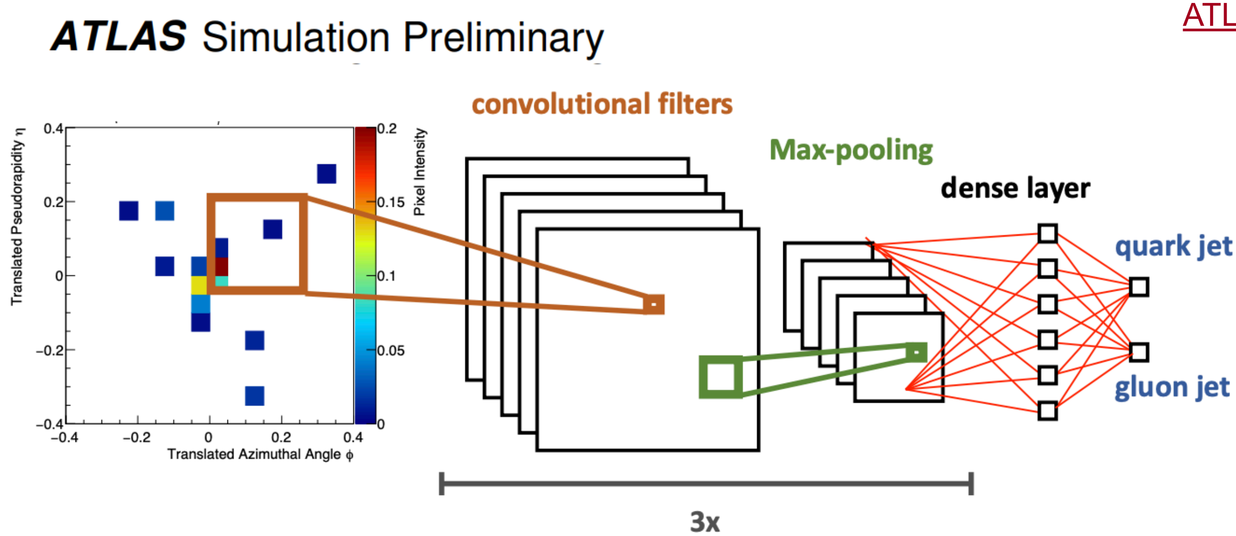
Top Misidentification Scale Factors



JINST 15 (2020) P06005

Quark versus Gluon with Jet Images

ATL-PHYS-PUB-2017-017

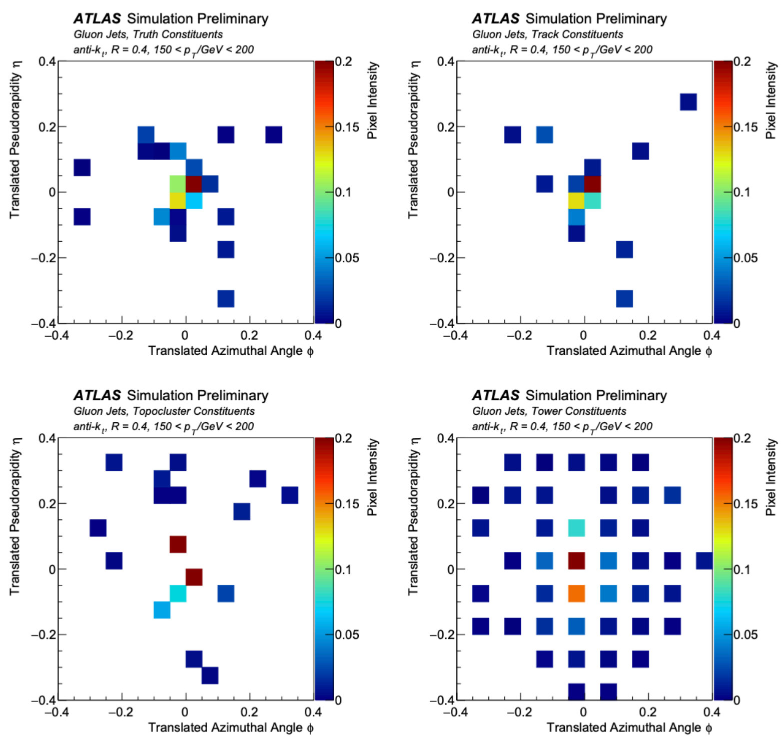


Truth

Tracks

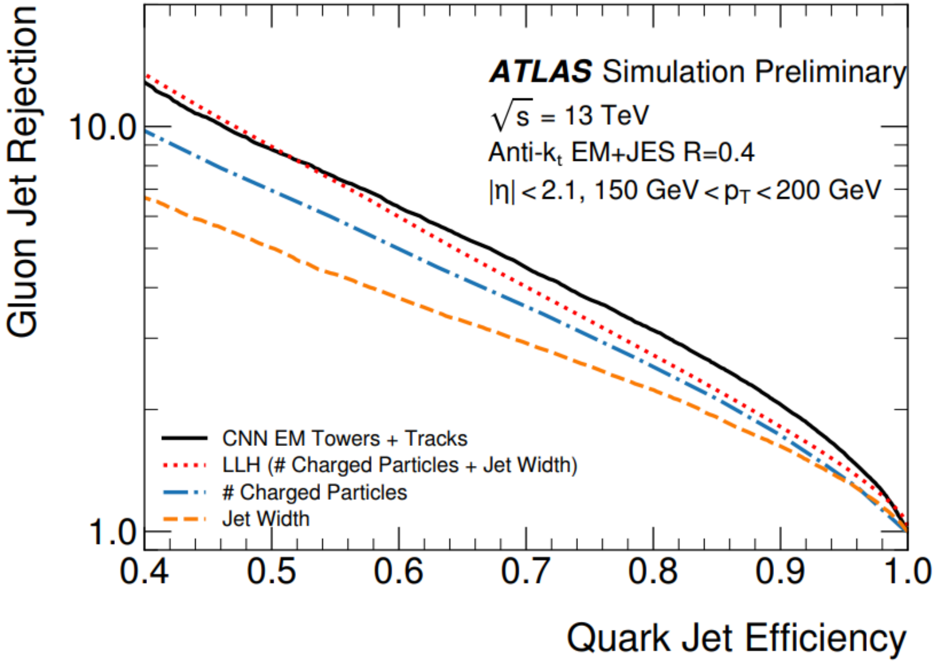
Topo Clusters

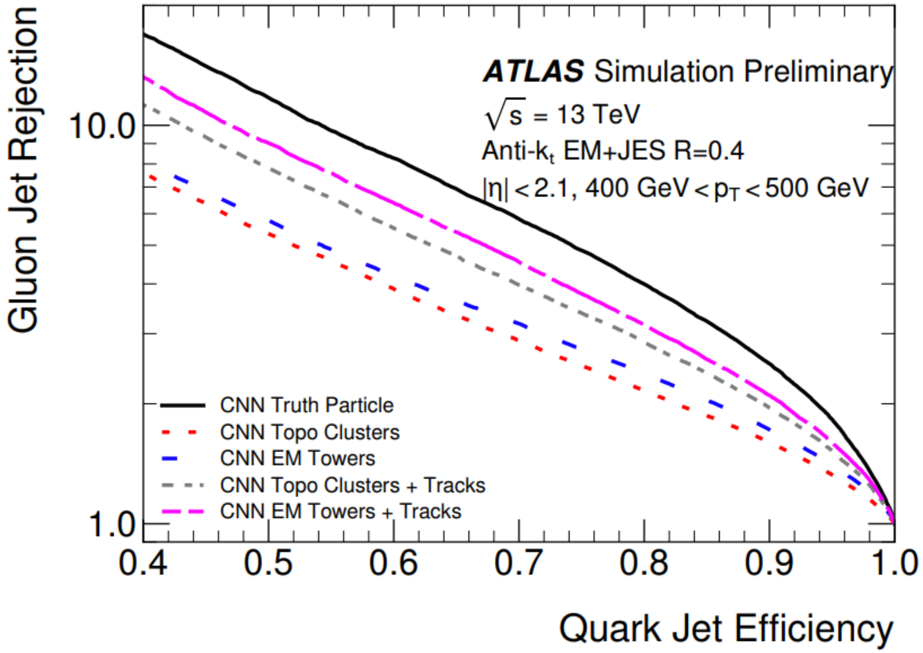
Towers



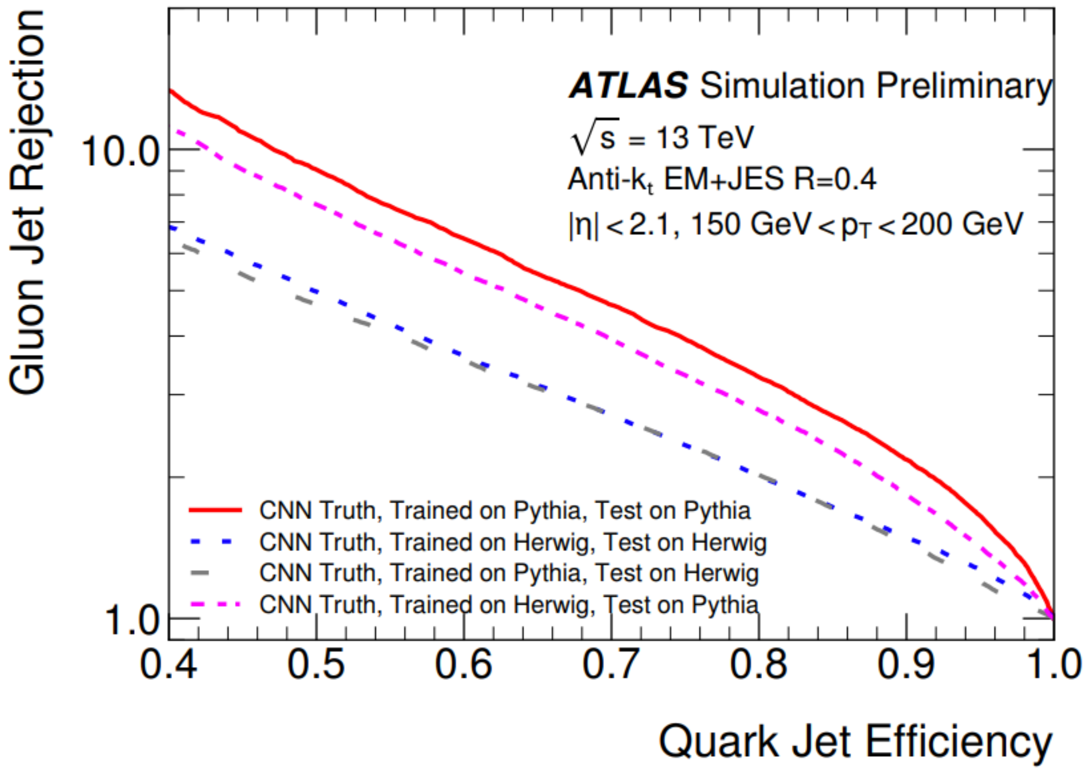
Quark versus Gluon with Jet Images

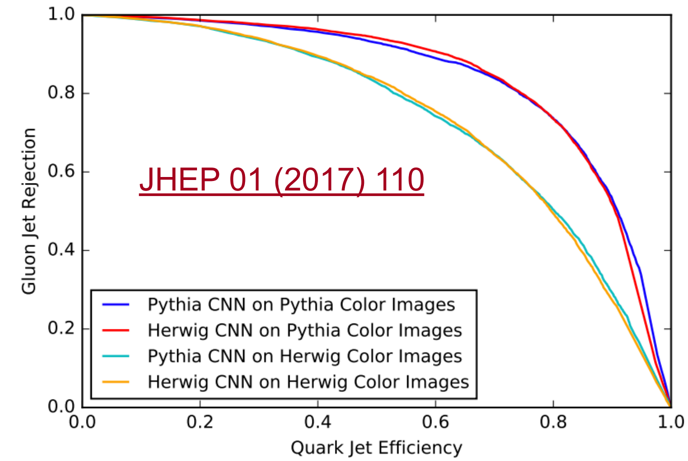
ATL-PHYS-PUB-2017-017

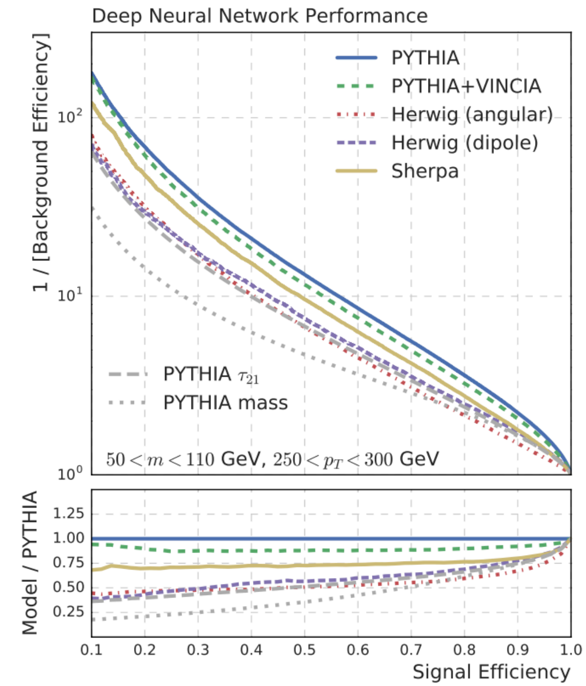




Sensitivity to Generators → Representation Learning







- What you train on seems to have smaller impact than what you test on
- Robustness of the learned representations?

Phys. Rev. D 95, 014018 (2017

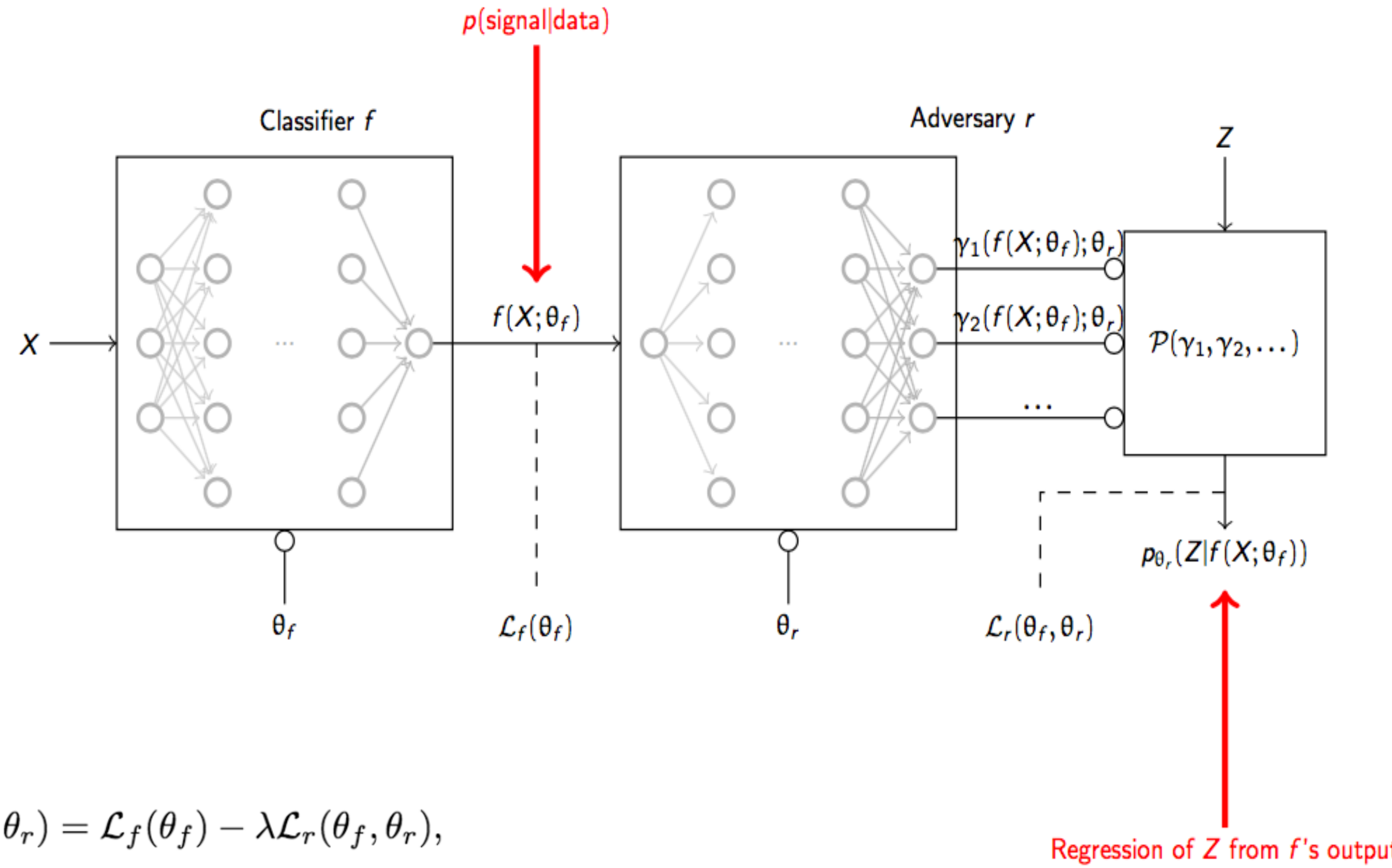
Mitigating Dependencies

Mitigating Dependencies

- With flexibility comes complexity:
  - Hard to control how models learn and utilize information
  - Potentially unwanted sensitivity to poorly modeled aspects of simulation
  - Potentially unwanted sculpting of key physics distributions like mass

- Decorrelation methods
  - Reweighting training distributions
  - DDT: Designing decorrelated taggers JHEP 05 (2016) 156
  - DisCo: Distance Correlation regularization Phys. Rev. Lett. 125, 122001 (2020)
  - Adversarial Learning NeurIPS 2017, 981-990, Phys. Rev. D 96, 074034 (2017)

Adversarial Learning



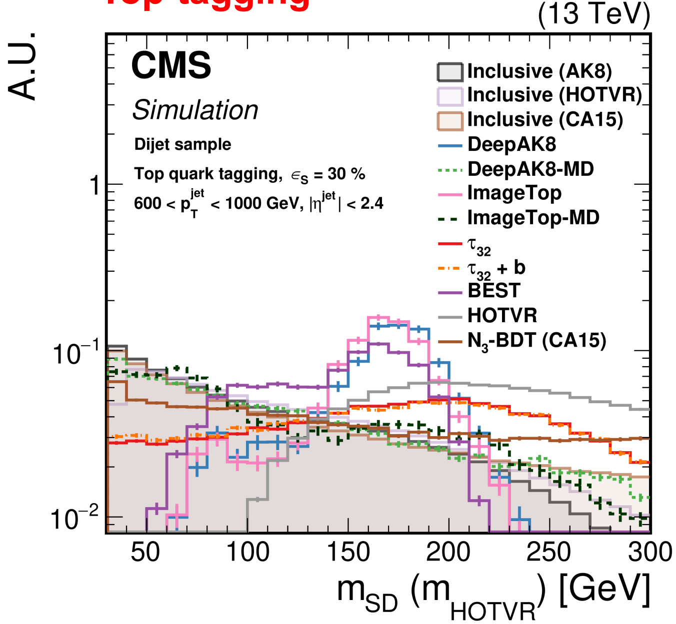
$$E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r),$$

- Build loss that encodes performance of classifier and an adversary
- Classifier penalized when adversary does well predicting Z
- Training is a min-max game targeting saddle point solution

NeurIPS 2017, 981-990

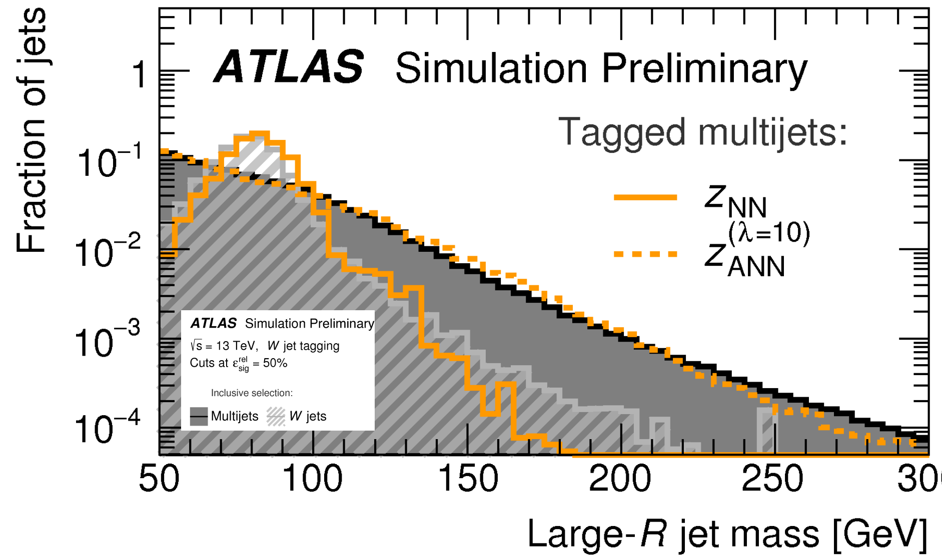
Mass Decorrelation

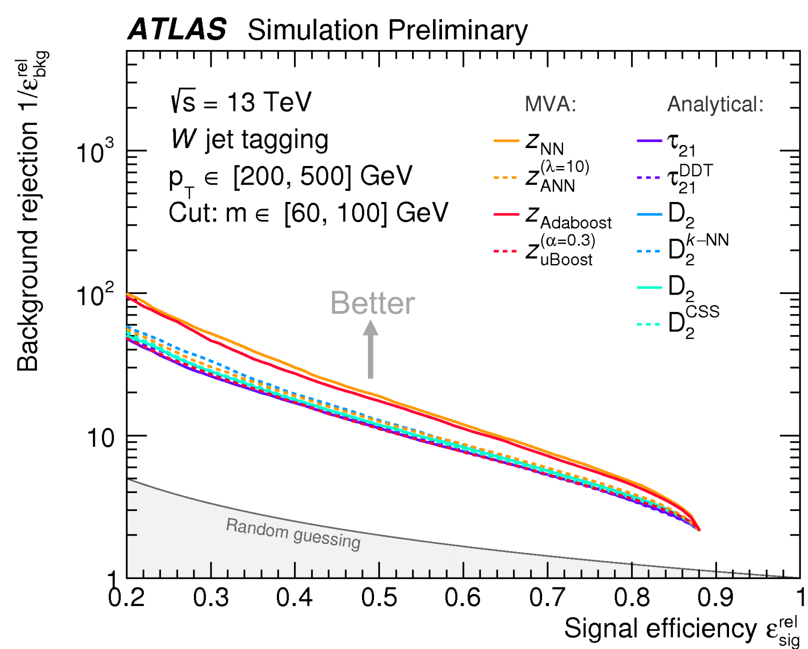
Top-tagging



W-tagging

ATL-PHYS-PUB-2018-014

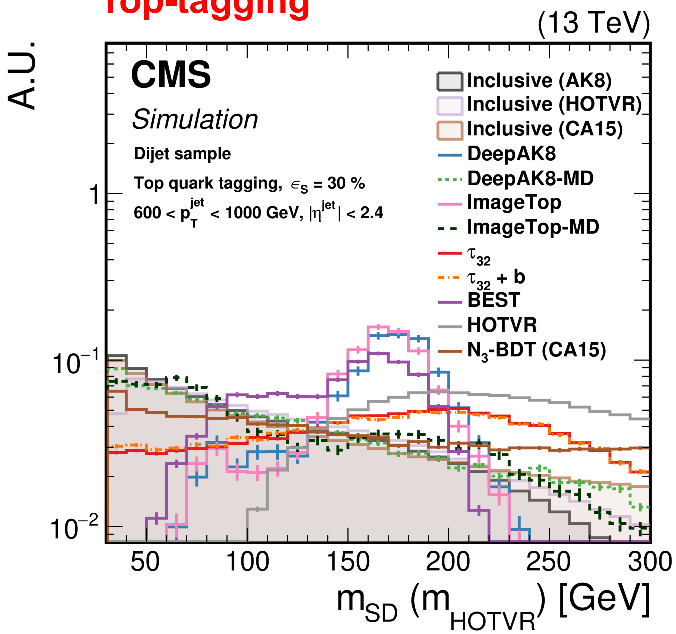


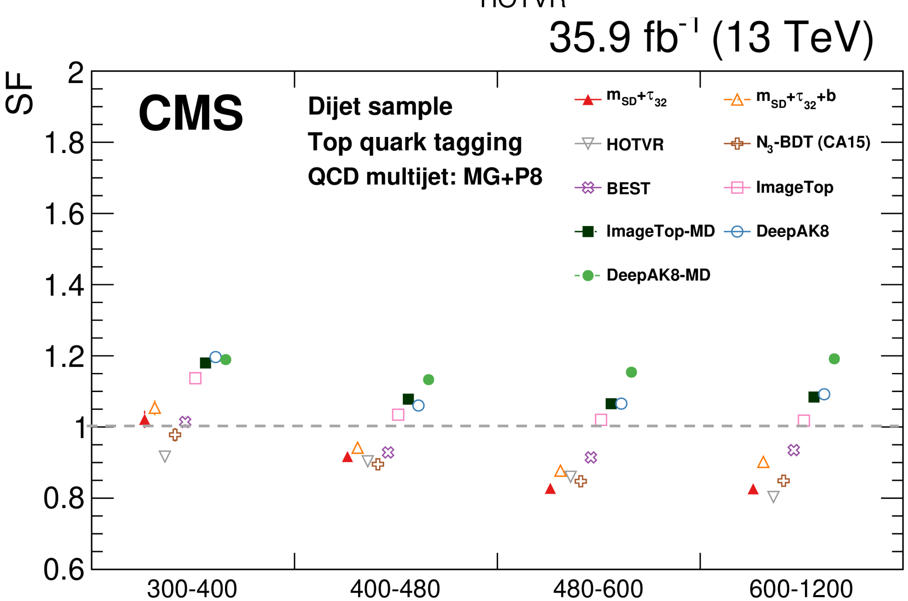


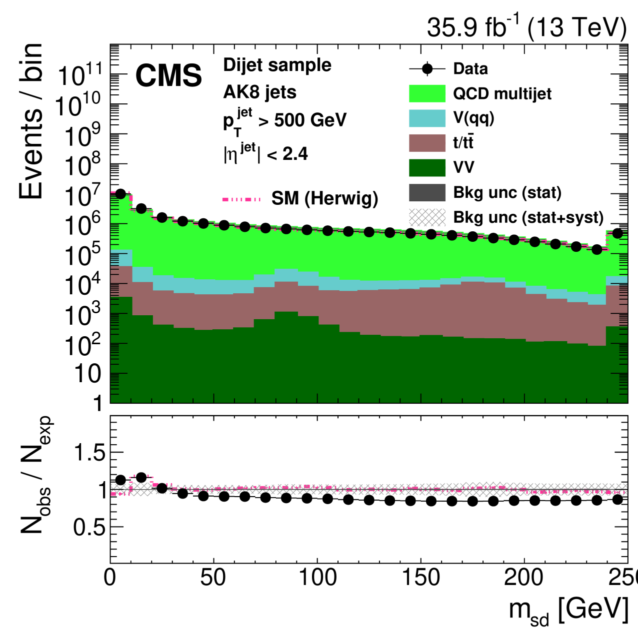
JINST 15 (2020) P06005

Mass Decorrelation

Top-tagging







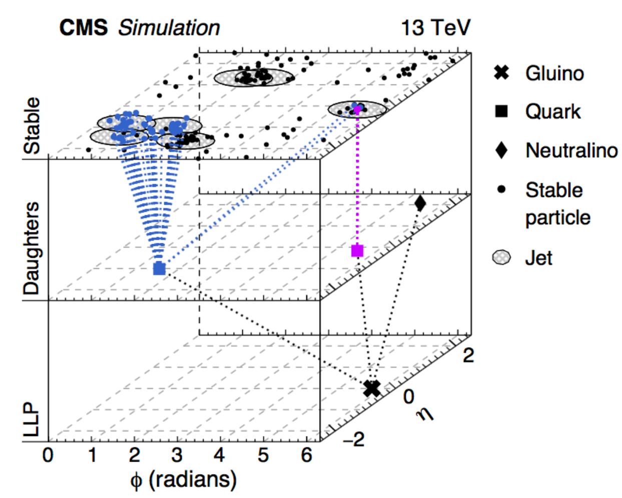
JINST 15 (2020) P06005

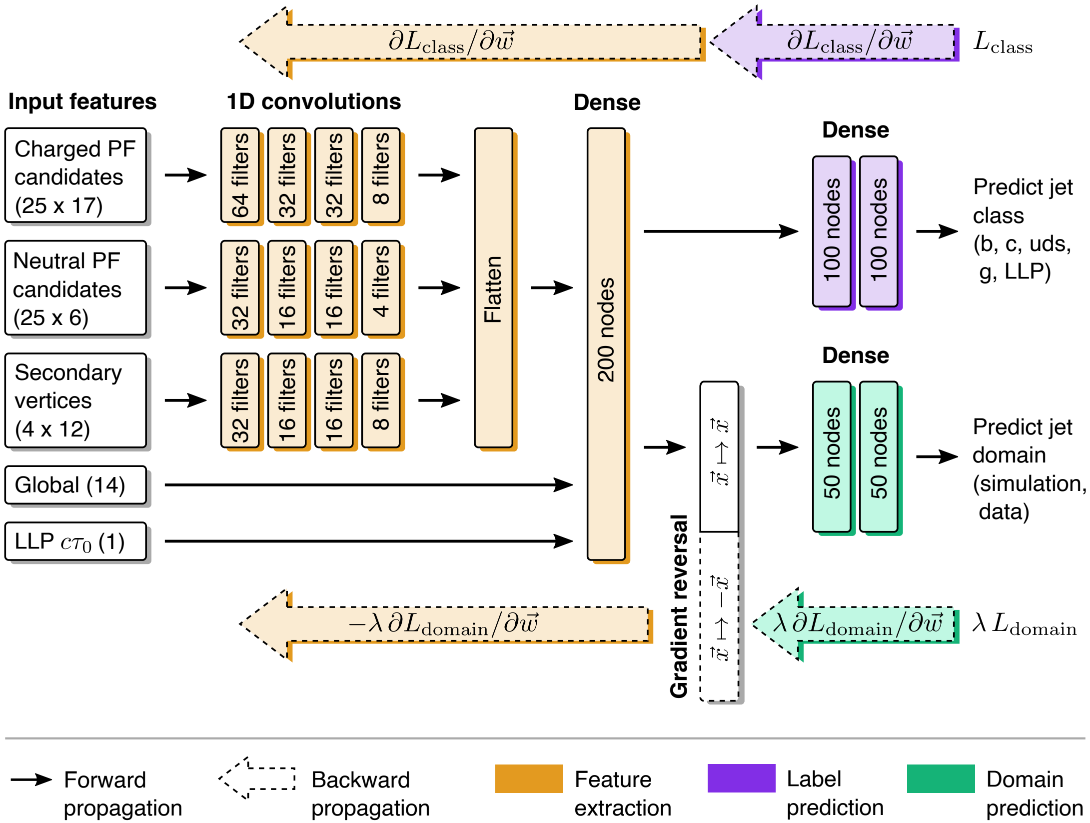
Mitigating Data / MC Differences in LLP Jet Tagging

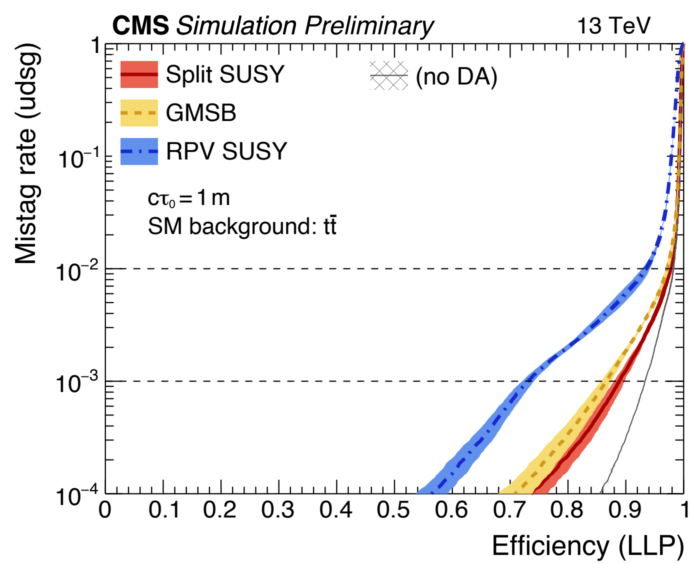
CMS-PAS-EXO-19-011

- Modified version of DeepJet

- Adversarial training to penalize differences in performance on MC vs Data

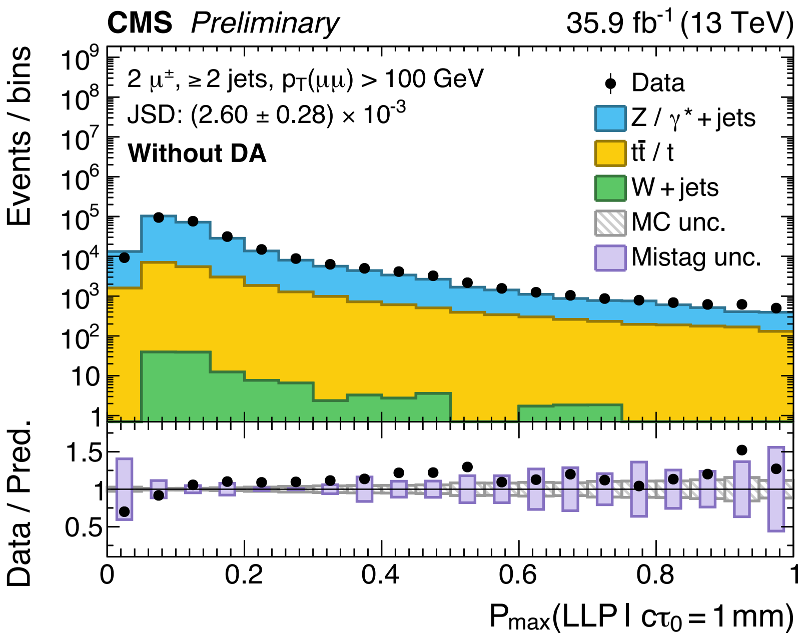


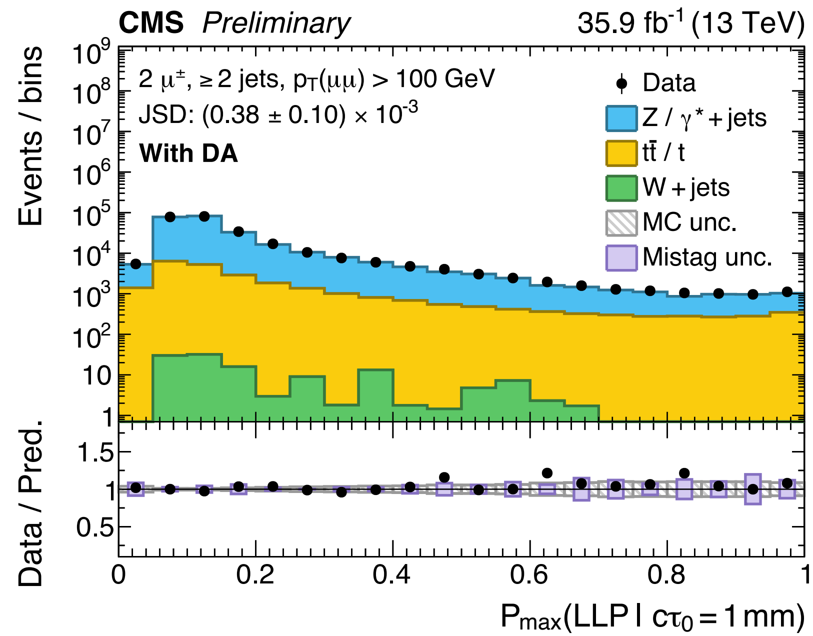




Mitigating Data / MC Differences in LLP Jet Tagging

CMS-PAS-EXO-19-011





Conclusion

- Image and Sequence taggers deployed for Boosted jet tagging and b-tagging → show many of the expected performance gains

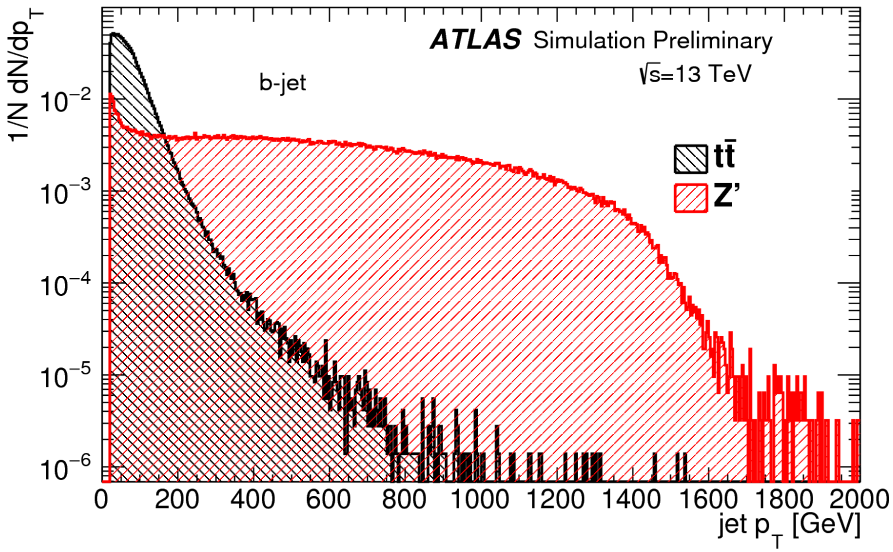
- Scale factors are reasonable
  - Mismodeling is not out of control
  - Interesting potential for mitigating Data /MC differences

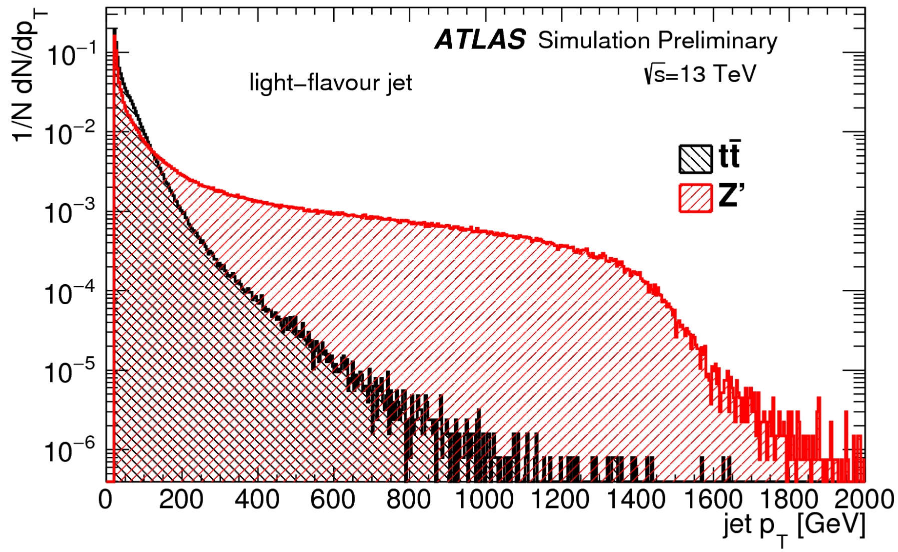
- SF uncertainties worse in samples with more background / flavour fraction uncertainty
  - Must separate uncertainties from calibration method and from learning mismodeled features

- Intriguing questions open about learned representations and how they are expressed

Backup

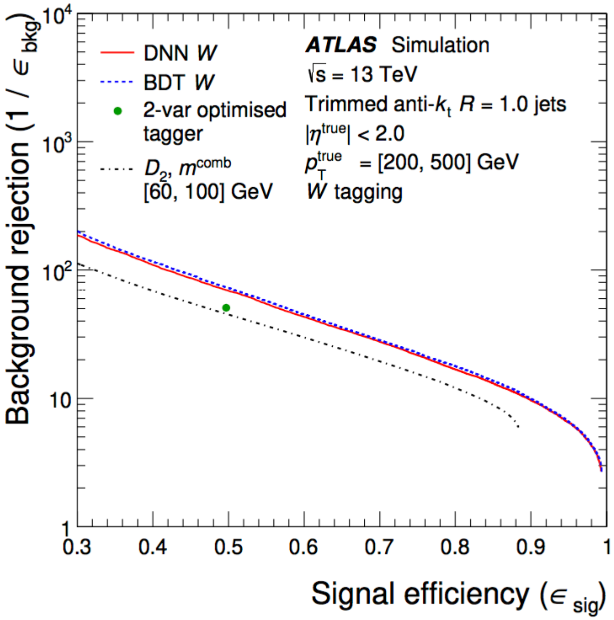
Mitigating Sample Kinematic Bias in Training

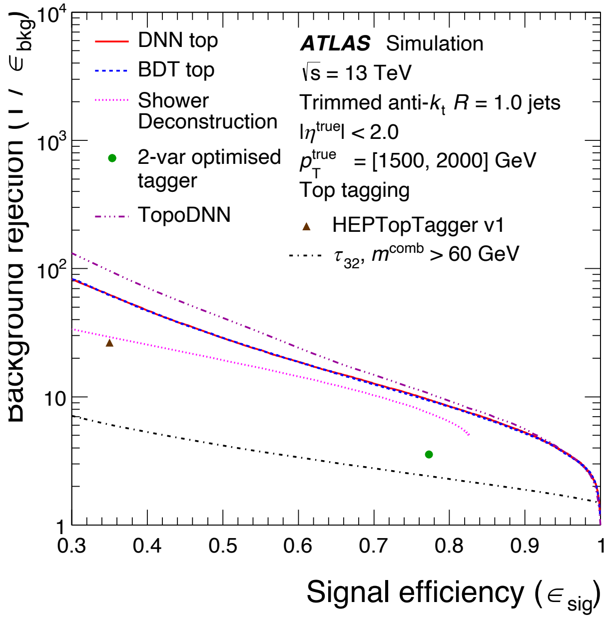




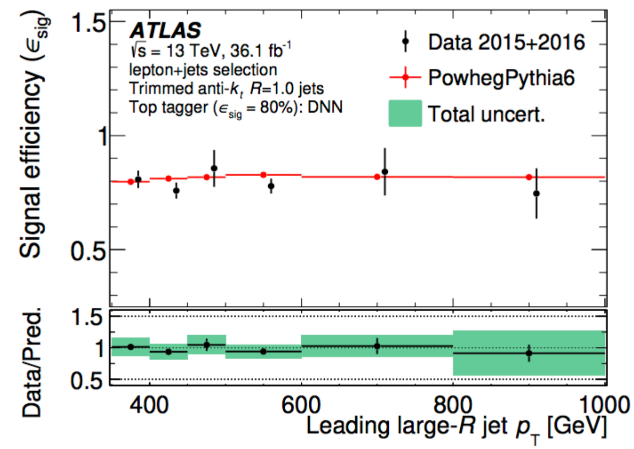
- Want tagger to understand how features change with kinematics
- Don't want to be sensitive to training distribution of kinematics
  - p_T is a pretty good discriminant! But distribution changes in analysis!
- Match key kinematic distributions between Signal / Background
  - Reweighting
  - Down sampling → ATLAS b-tag found this more stable for training

Combining Substructure Variables





arXiv:1808.07858

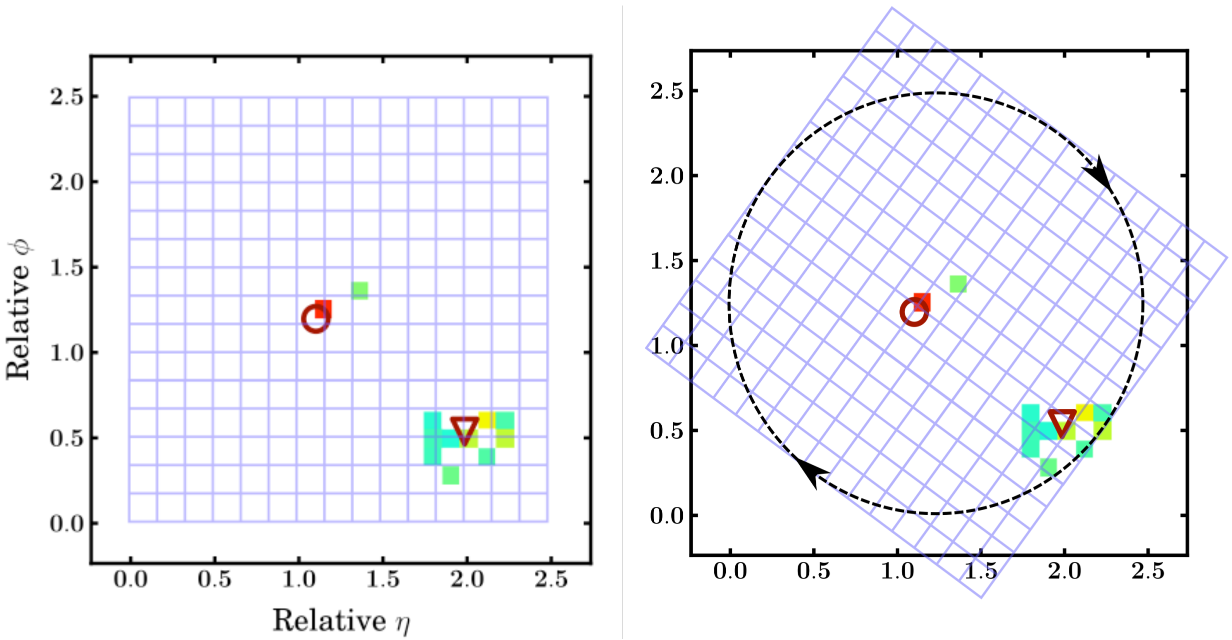


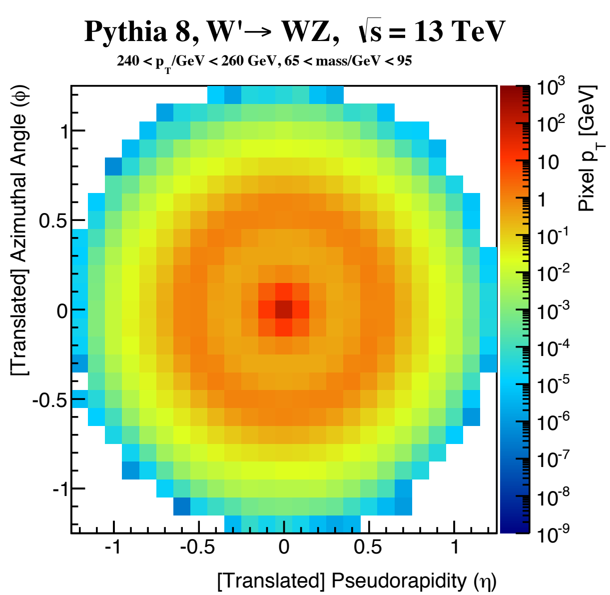
- Wide array of physics insight has gone into developing jet substructure observables

- Direct application of ML for combining power of multiple partially correlated substructure features

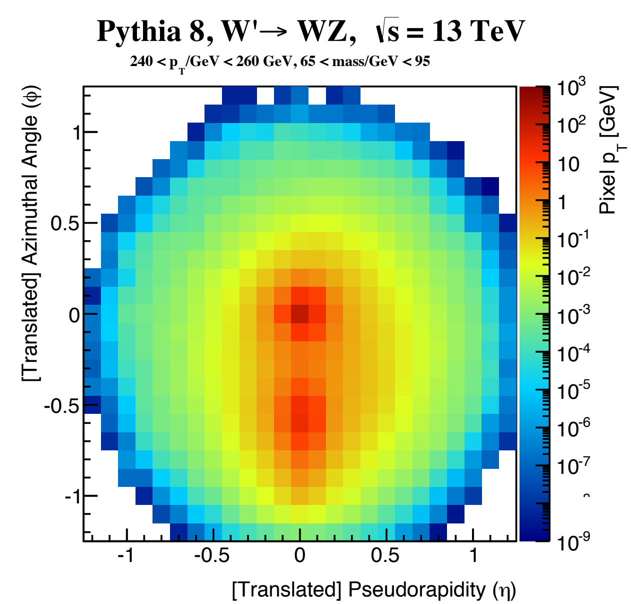
- First calibrations look quite reasonable!

Jet Image Pre-Processing

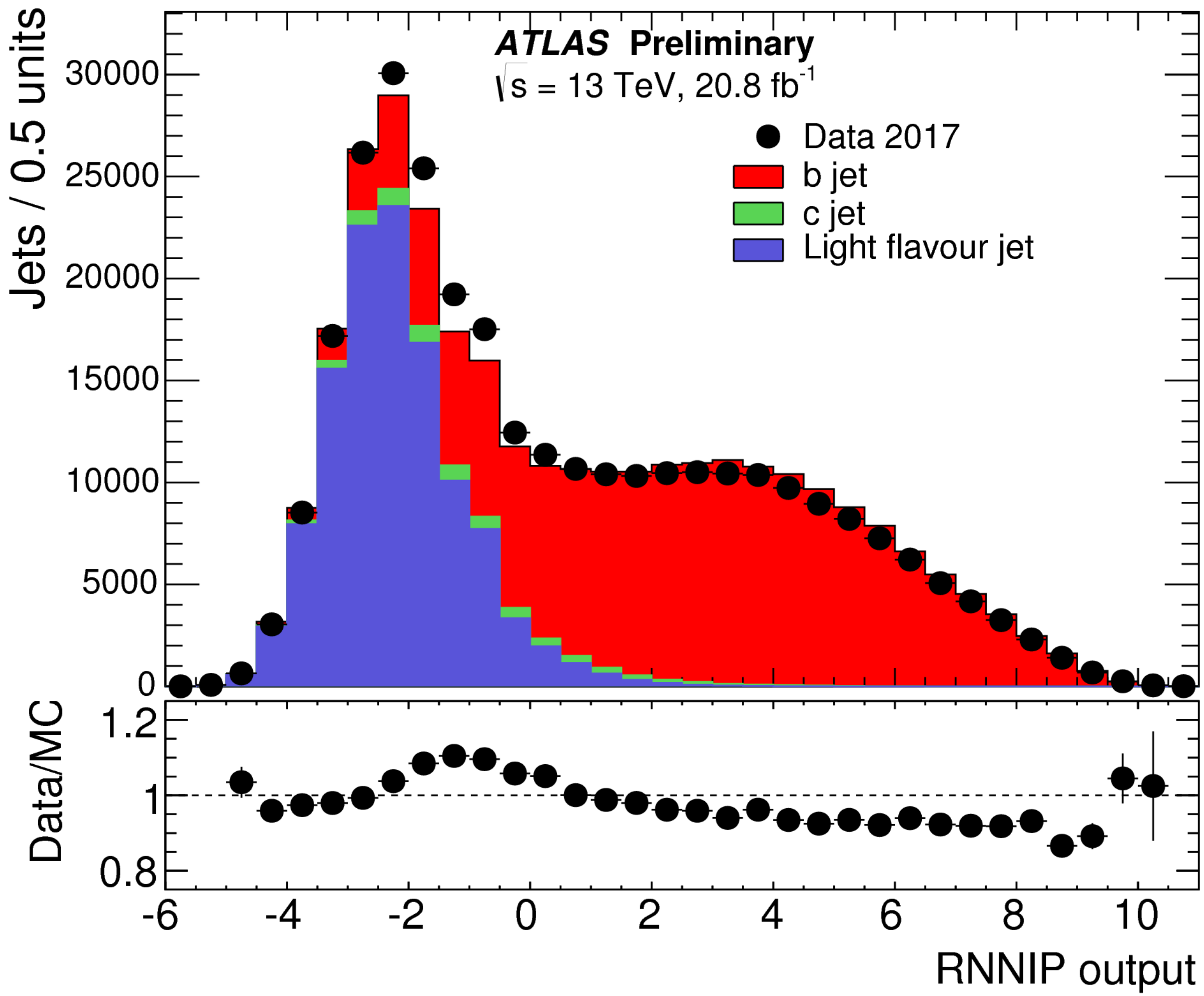




Translate → Rotate → Flip

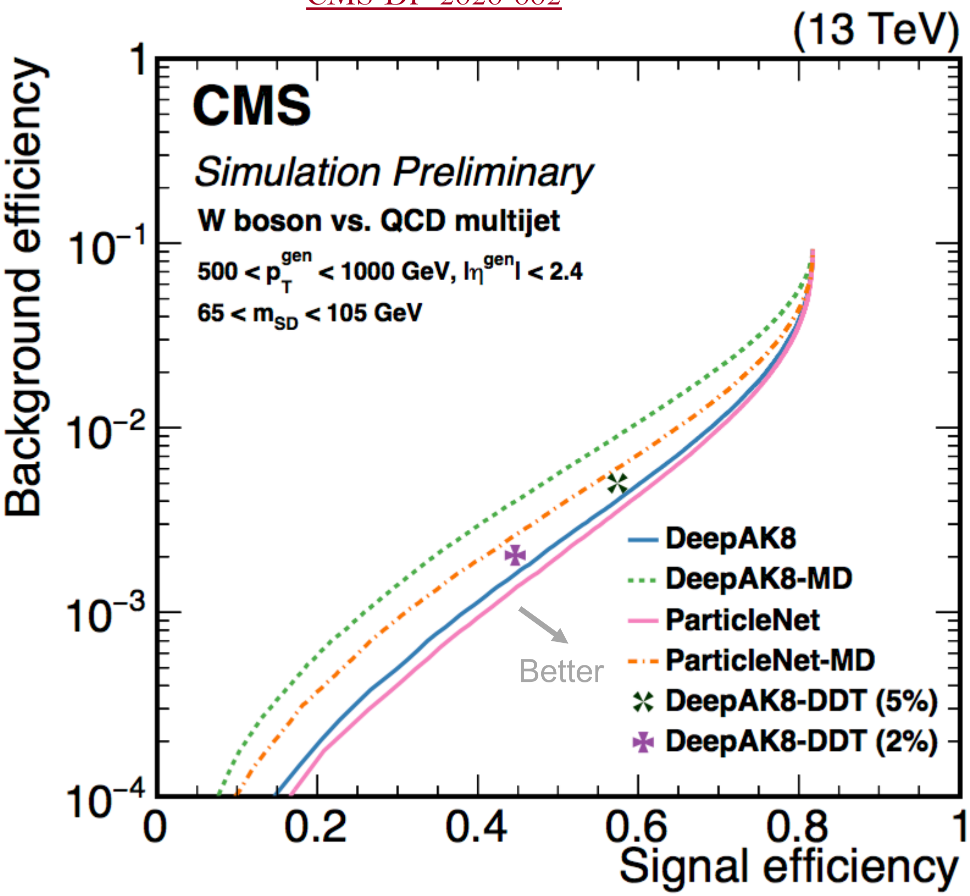


RNNIP

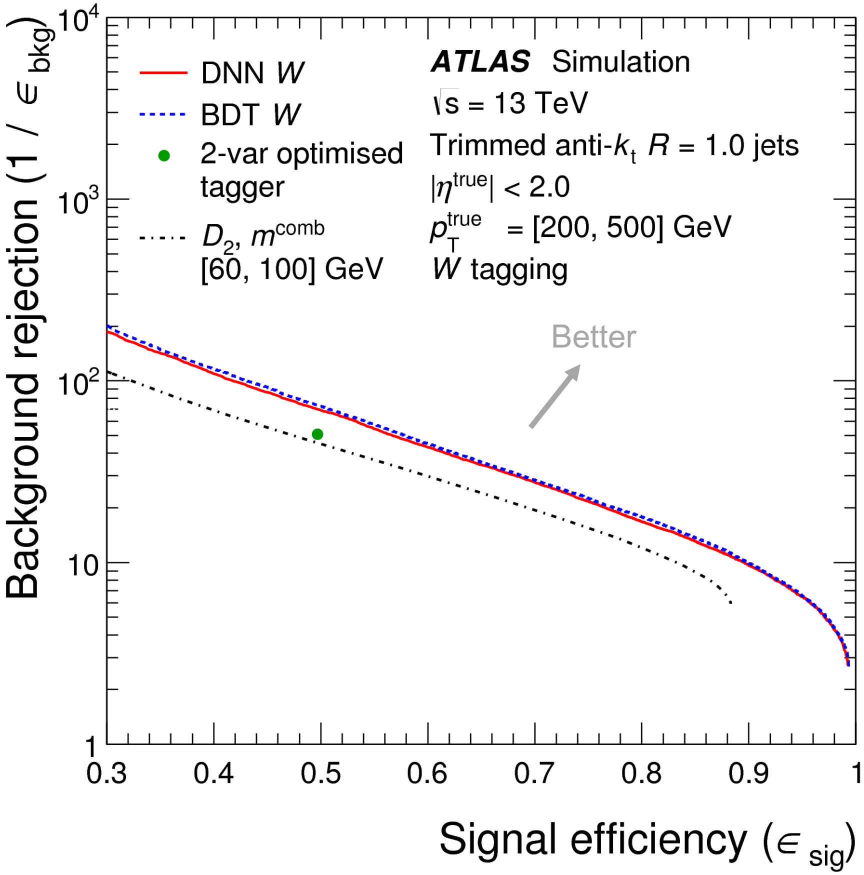

ATLAS Preliminary
$\sqrt{s}$ = 13 TeV, 20.8 fb^{-1}
Data 2017
b jet
c jet
Light flavour jet
Jets / 0.5 units
30000
25000
20000
15000
10000
5000
0
Data/MC
1.2
1
0.8
-6
-4
-2
0
2
4
6
8
10
RNNIP output

W Tagging

CMS-DP-2020-002

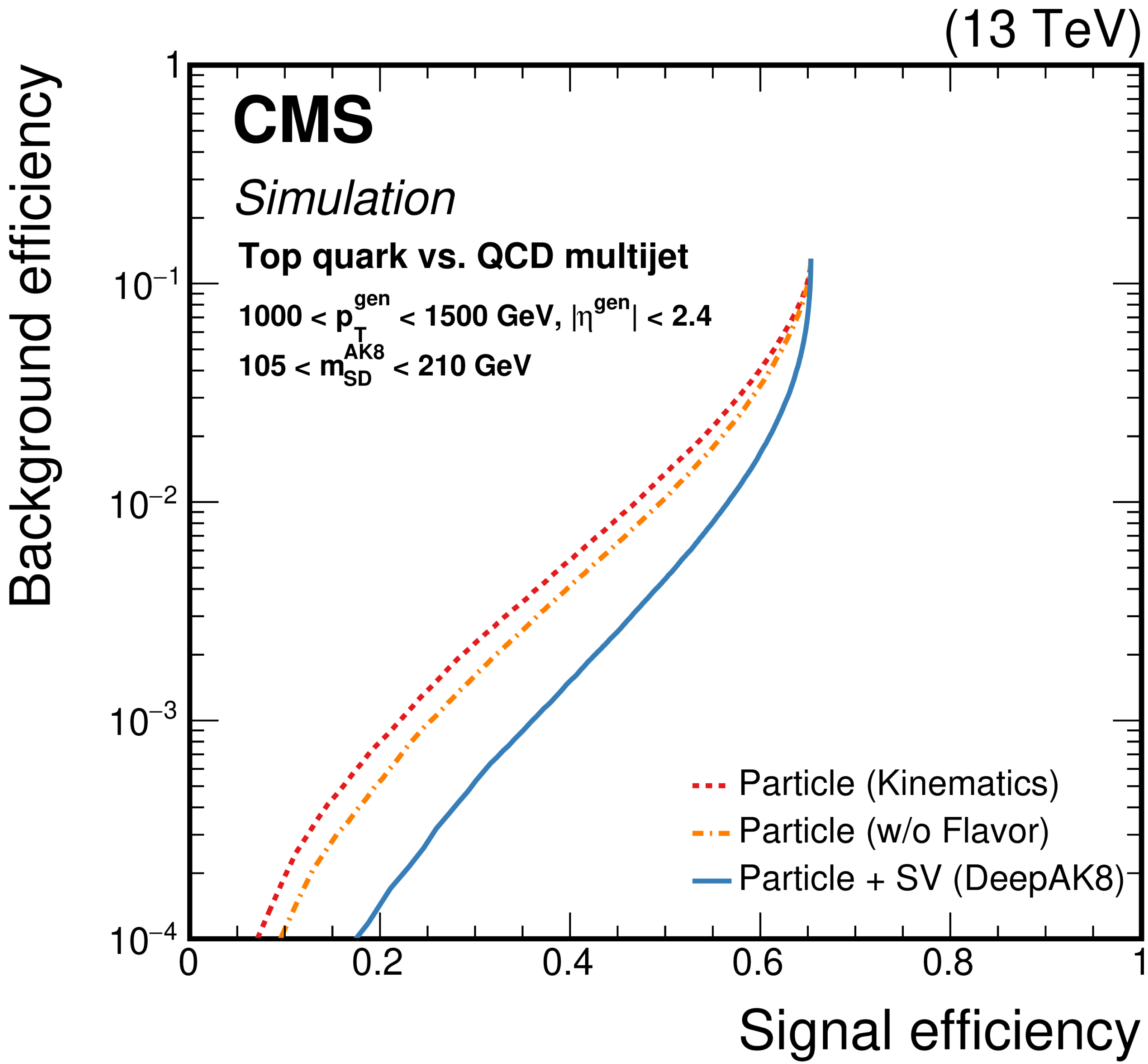


EPJC 79 (2019) 375



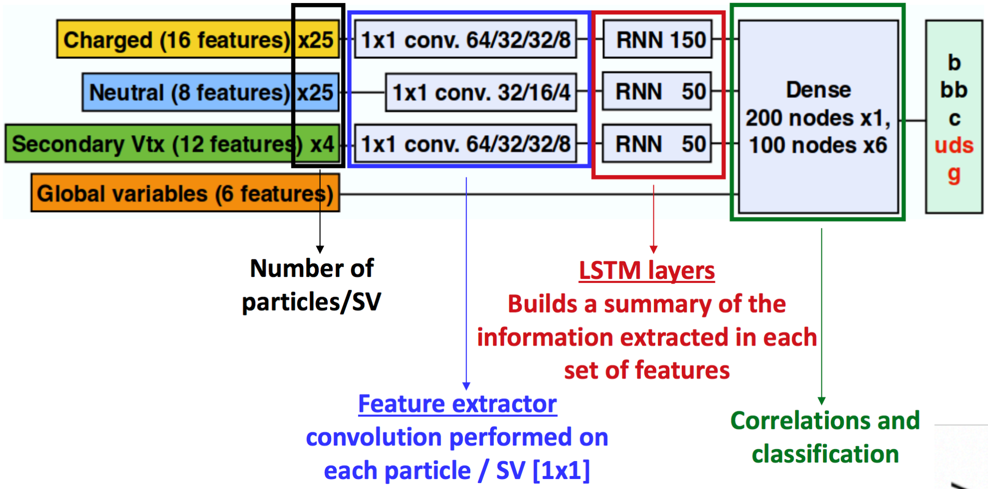
NOTE: different p_T ranges

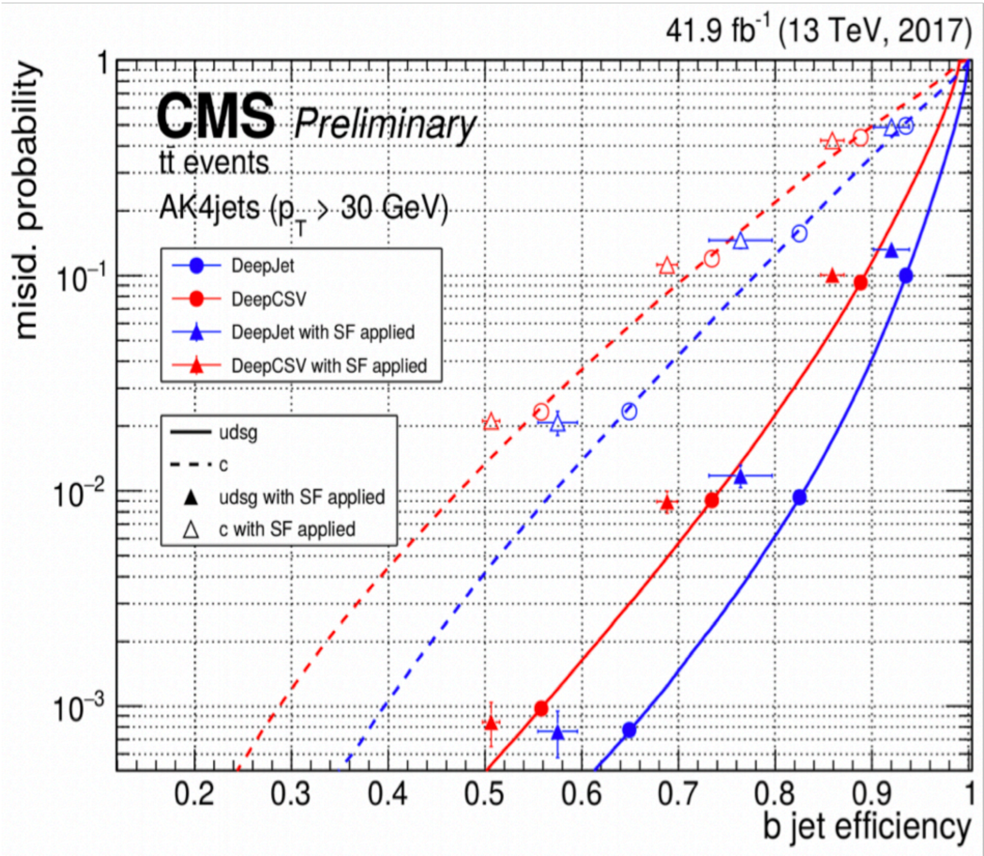
DeepAK8 variations



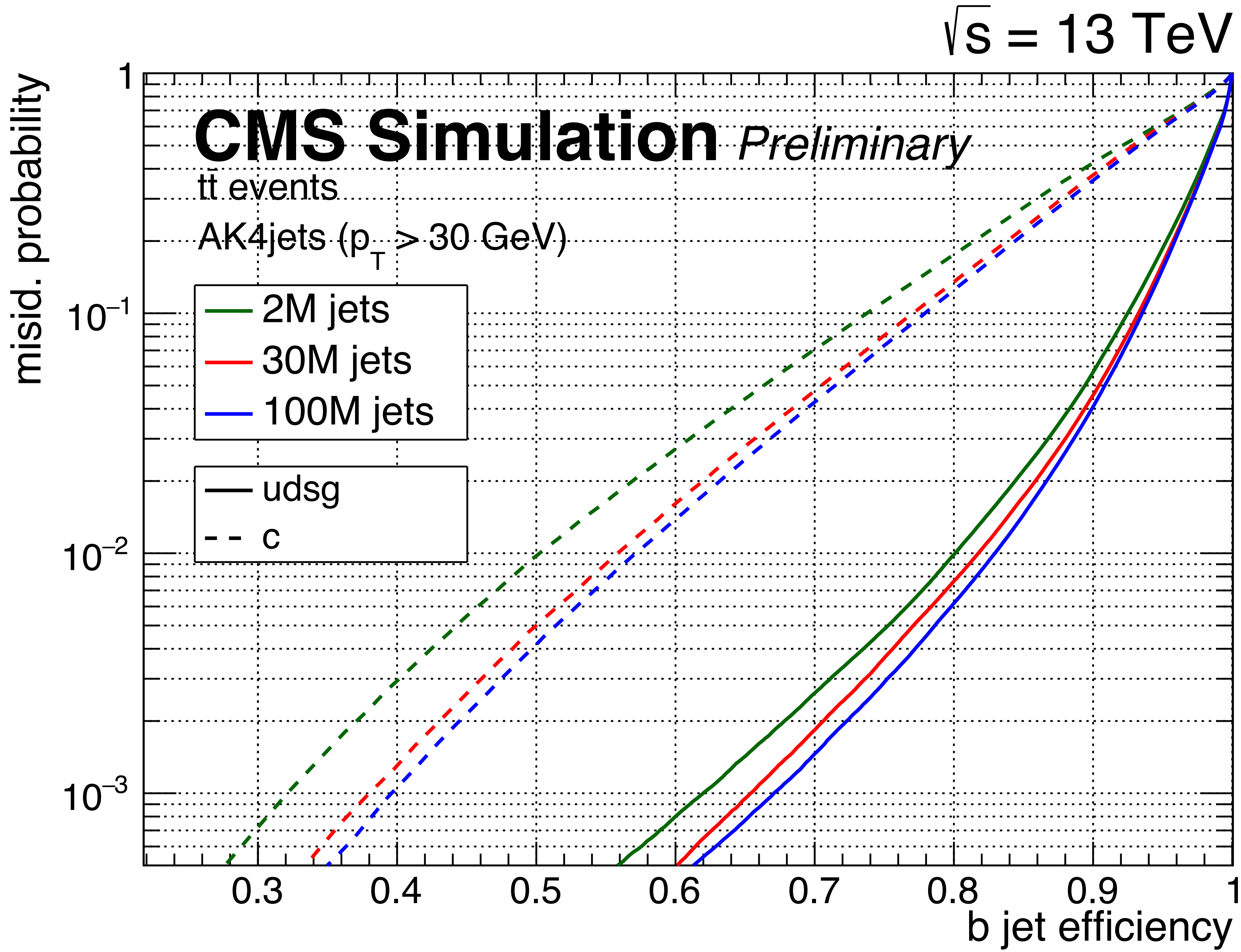
DeepJet

JINST 15 (2020) P12012
CMS-DP-2018-058
CMS-DP-2021-004

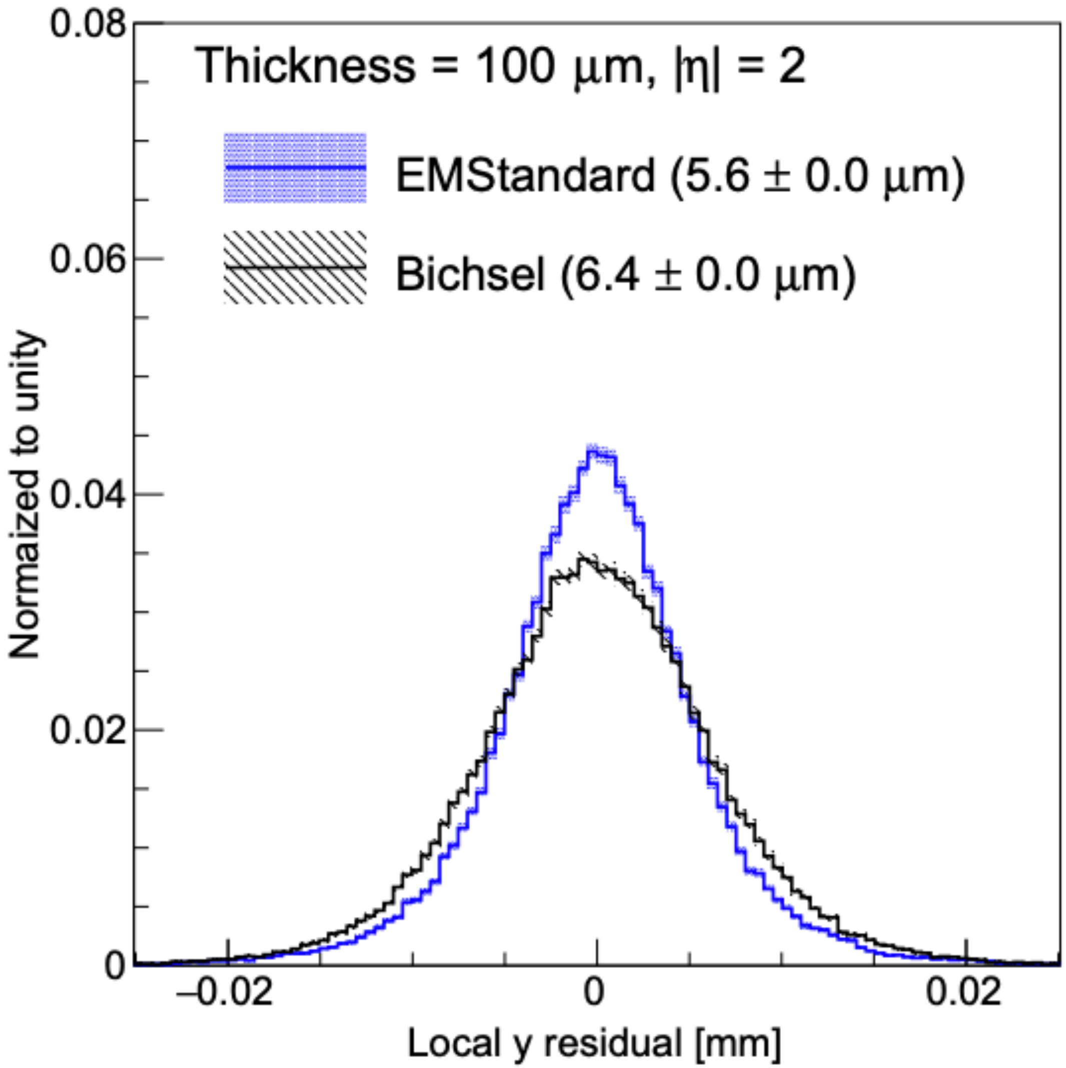




DeepJet and Training Size



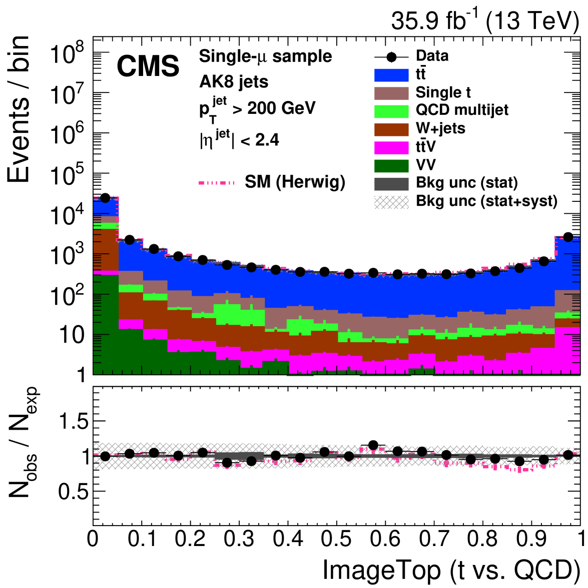
Energy deposition in Si modeling

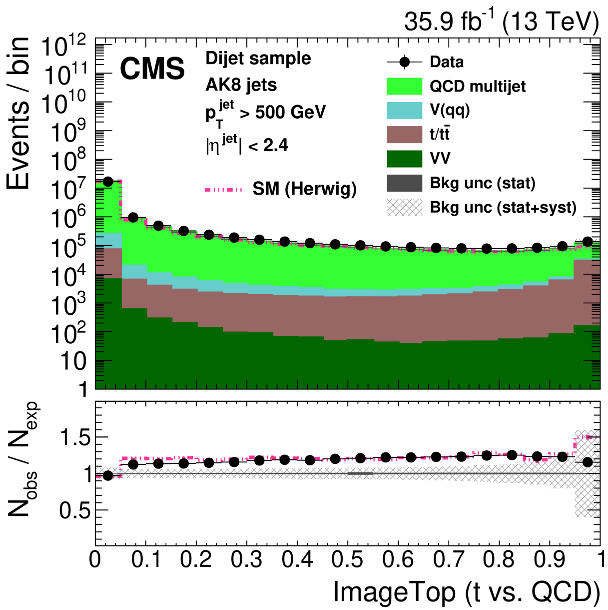


NIM A 899 (2018)

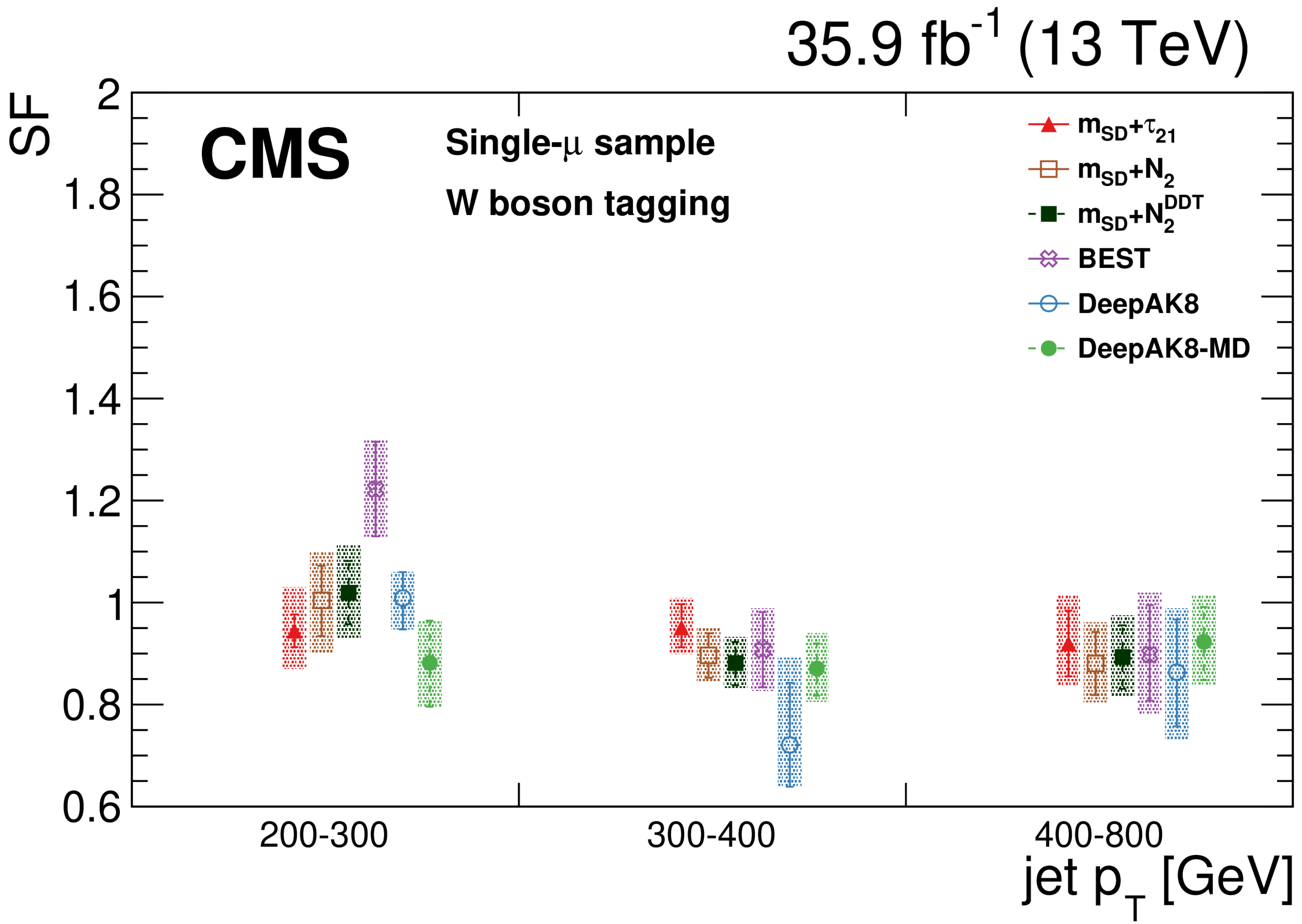
Top Misidentification Scale Factors

JINST 15 (2020) P06005





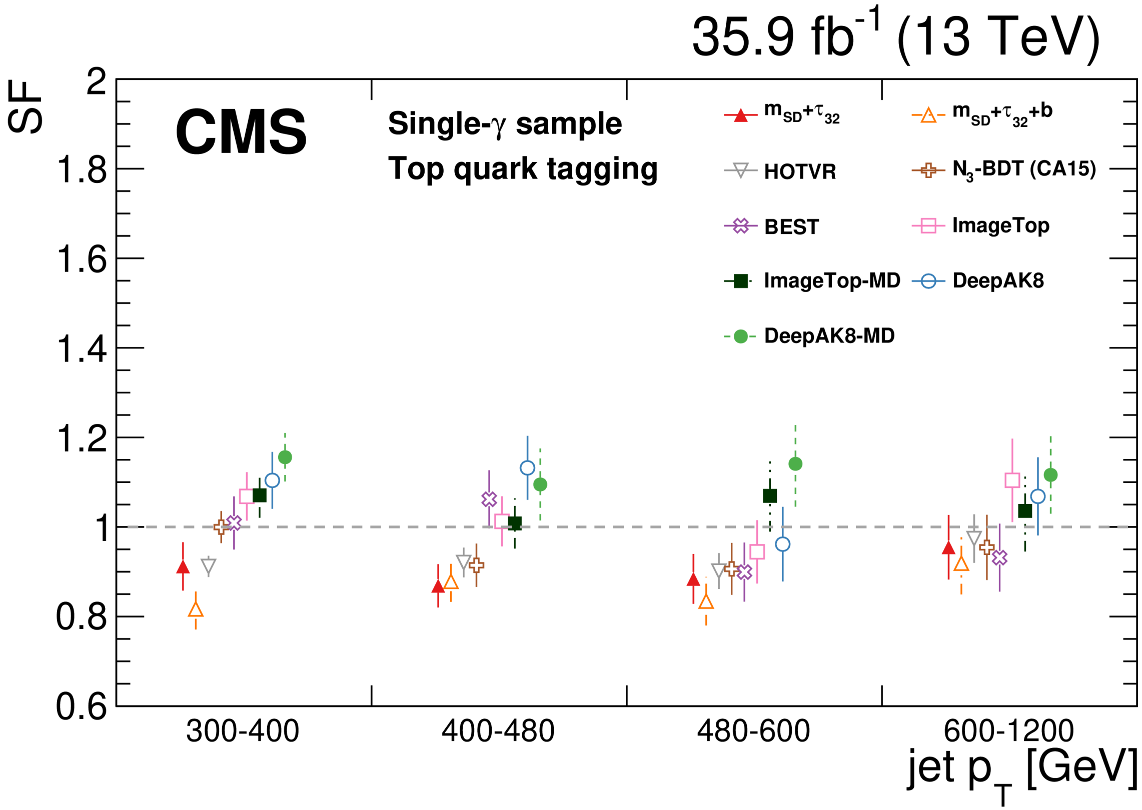
W-tagging Scale Factors

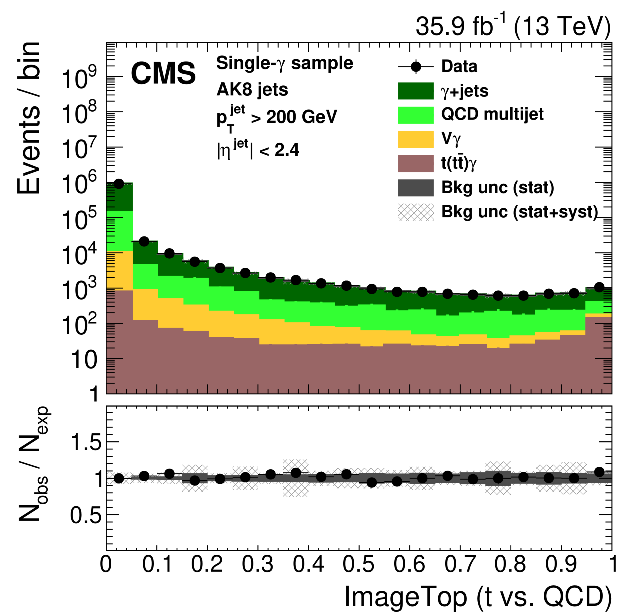


JINST 15 (2020) P06005

Top Misidentification Scale Factors

JINST 15 (2020) P06005



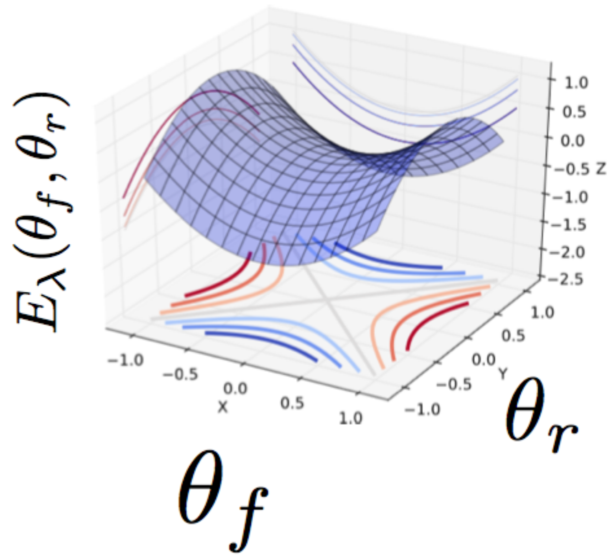


Adversarial Networks

[arXiv:1611.01046]

$$\hat{\theta}_f, \hat{\theta}_r = \arg \min_{\theta_f} \max_{\theta_r} E(\theta_f, \theta_r).$$

$$E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda \mathcal{L}_r(\theta_f, \theta_r),$$



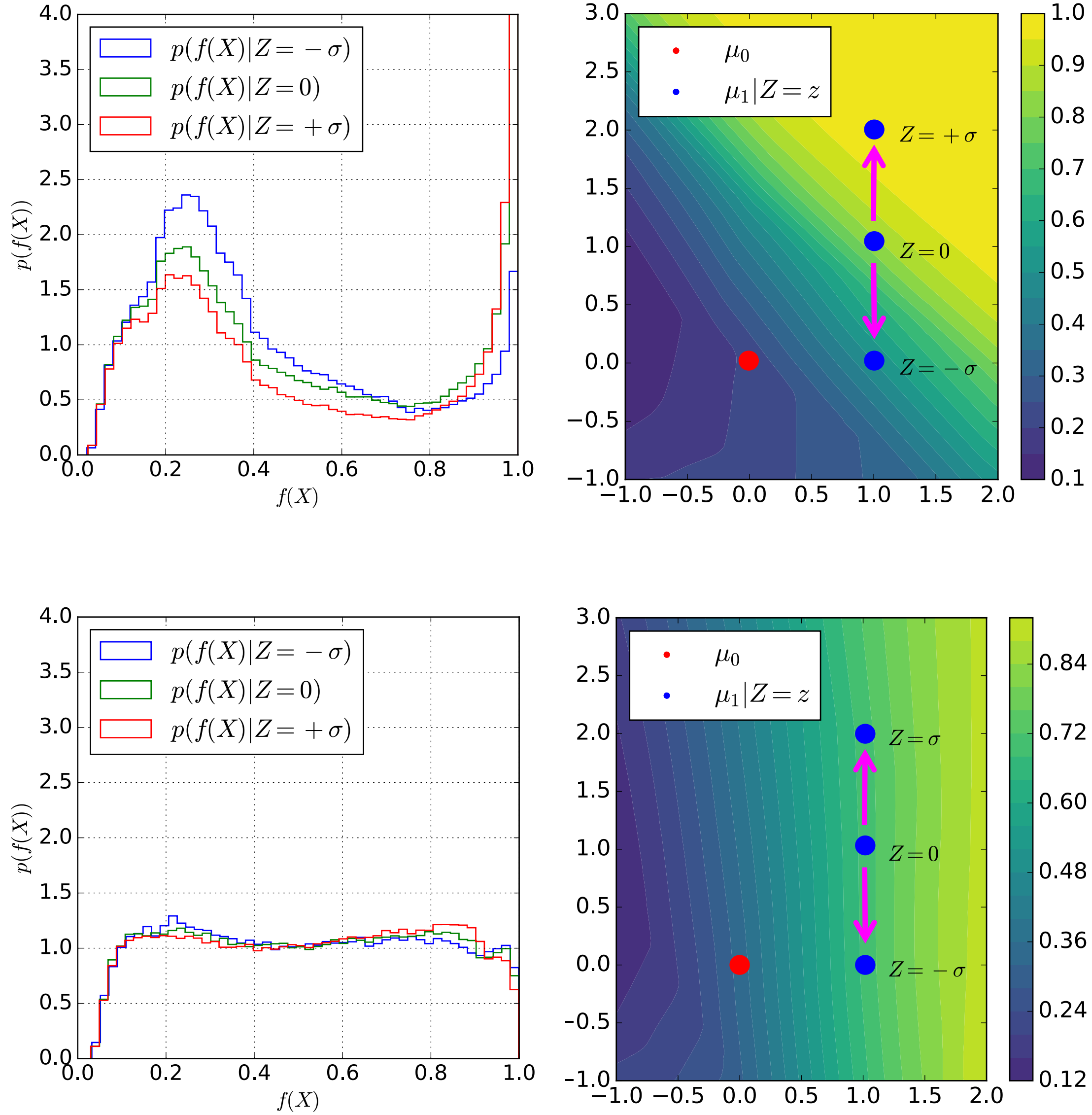
- Loss encodes performance of classifier and adversary
  - Classifier penalized when adversary does well at predicting Z
- Hyper-parameter λ controls trade-off
  - Large λ enforces f(…) to be pivotal, e.g. robust to nuisance
  - Small λ allows f(…) to be more optimal

Learning to Pivot: Toy Example

2D example

$$x \sim \mathcal{N}\left((0,0), \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}\right) \quad \text{when } Y = 0,$$

$$x \sim \mathcal{N}\left((1,1+Z), \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad \text{when } Y = 1.$$

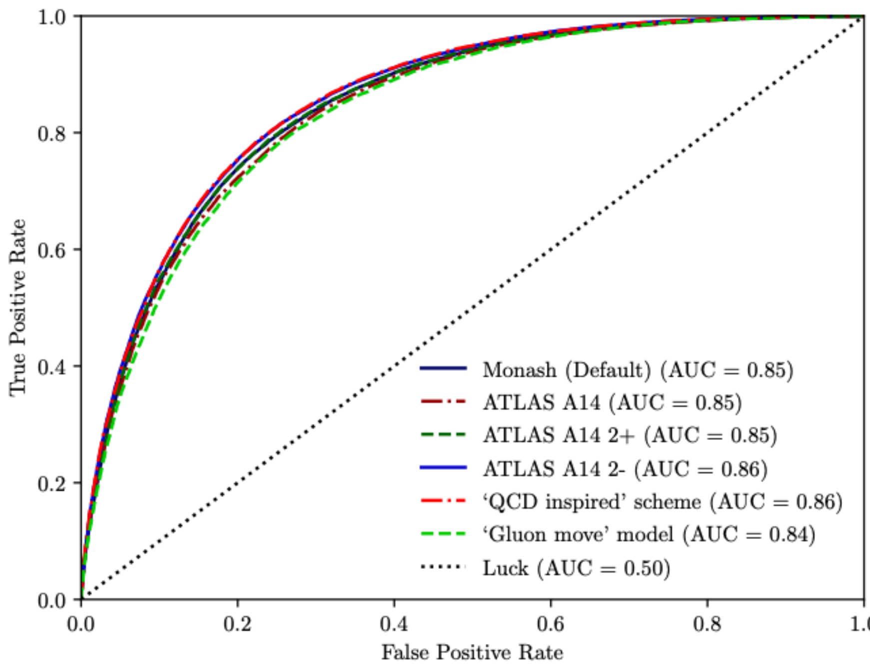


- Without adversary (top) large variations in network output with nuisance parameter
- With adversary (bottom) performance is independent!

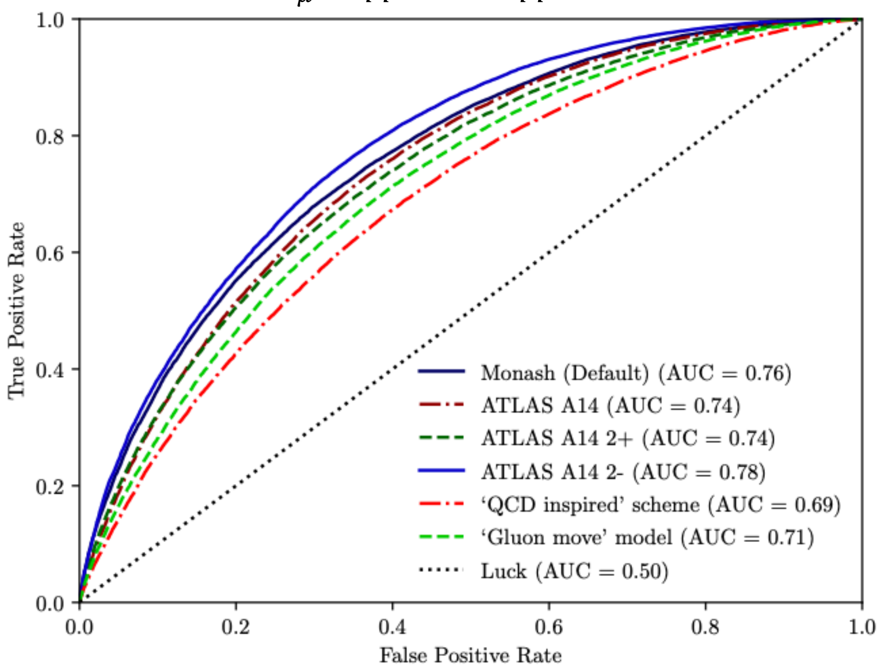
Modeling Comparisons

arXiv:2105.04582

$C_\mu \to qq$ vs $H \to gg$



$C_\mu \to qq$ vs $H \to qq$

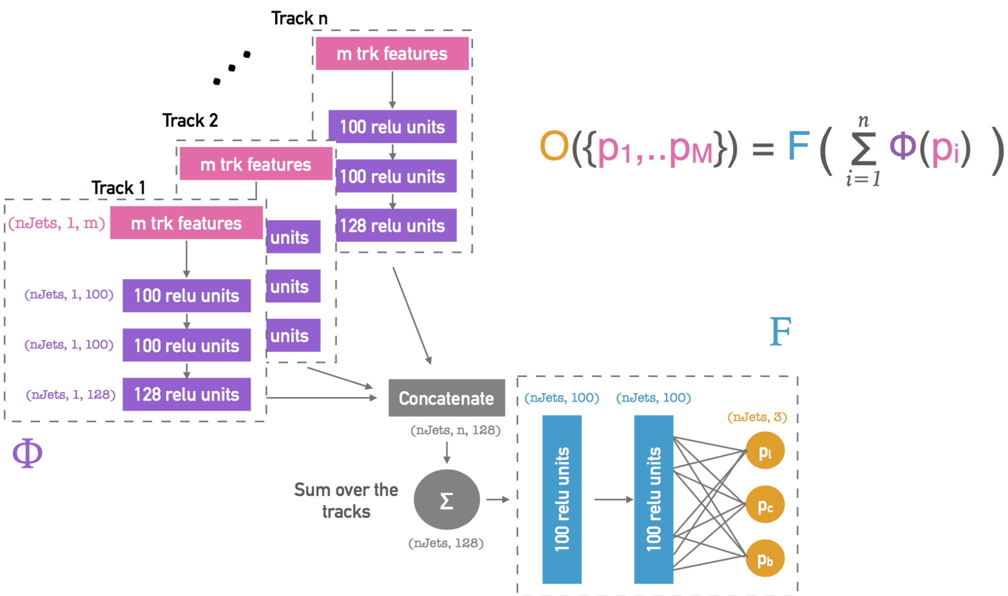


Next Generation Taggers

Next Generation: DIPS – Deep Impact Parameter Sets

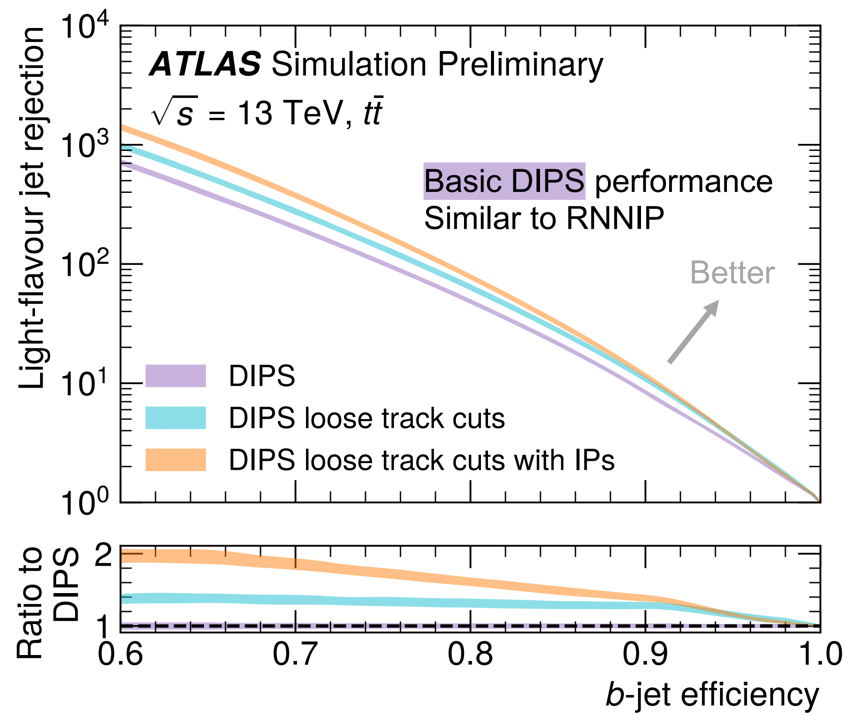
- Challenges of RNN Tagging
  - Must choose sequence ordering, not inherent, which is best?
  - Requires iteration over tracks, can't be parallelized

- Deep Sets: permutation invariant and parallelizable model



Model	Parameters	Training time [min]	Time / epoch [s]
RNNIP	47k	86 ± 13	241 ± 14
DIPS	49k	44 ± 4	78 ± 4

Model	Parameters	GPU Evaluation time [s]	CPU evaluation time [s]
RNNIP	47k	170 ± 2	685 ± 84
DIPS	49k	46 ± 2	206 ± 98



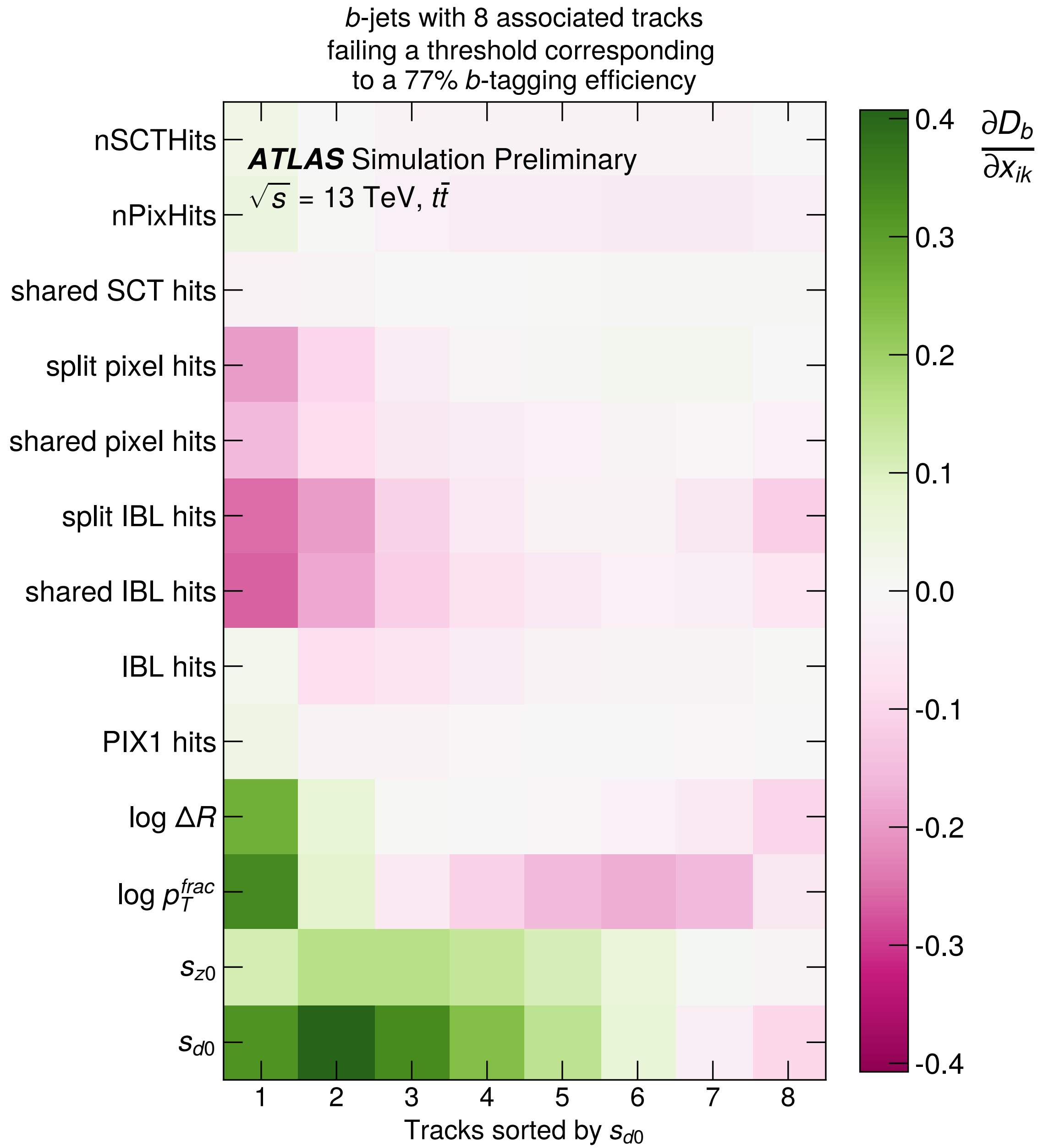
ATL-PHYS-PUB-2020-014 JHEP 01 (2019) 121

Sanity Checks

Salience Map

$$S_{jk} = \sum_{i=1}^{N_{jets}} \frac{dD_b(x^{(i)})}{dx_{jk}^{(i)}}$$

$x^{(i)}$ = all tracks/features of i[th] jet
i = jet
j = feature
k = track



Next Generation: ParticleNet with Graph Neural Networks

Phys. Rev. D 101, 056019 (2020)
CMS-DP-2020-002