DEEP SETS FOR ATLAS FLAVOUR TAGGING

Summary of ATL-PHYS-PUB-2020-014

ATLAS PUB Note

ATL-PHYS-PUB-2020-014

May 25, 2020

Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS

The ATLAS Collaboration

This work introduces a new architecture for Flavour Tagging based on Deep Sets, which models the jet as a set of tracks, in order to identify the experimental signatures of jets containing heavy flavour hadrons using the impact parameters and kinematics of the tracks. This approach is an evolution with respect to the Recurrent Neural Network (RNN) currently adopted in the ATLAS experiment, which treats track collections as a sequence. The Deep Sets model comprises a permutation-invariant and highly parallelisable architecture, leading to a significant decrease in training and evaluation time, and thus allowing for much faster turn-around times for optimisation. Additionally, this permutation invariance encoded in the model is more physically motivated than the sequence-based RNN. We compare the Deep Sets algorithm with the RNN benchmark, probe the model to interpret the information learned, and provide studies optimising the Deep Sets algorithm by loosening the track selection and including additional inputs.

ATL-PHYS-PUB-2020-014
25 May 2020

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-014/

OUTLINE

- A new approach based (and replacing) the previous RNN-based algorithm
- As a general remark: FTag algorithms are using either
  - secondary vertices displaced from the primary vertex
  - impact parameters of track in the jet (ATLAS RNN algorithm from 2017)
    - RNN improves the performance of IP2D and IP3D exploiting correlations between tracks:
      - jets are track sequences: *track order matters*
- DIPS: jets are sets of unordered tracks (ATL-PHYS-PUB-2020-014, May)
  - benefits: quicker convergence, faster to train and optimise
- Auxiliary studies:
  - track optimization
  - how to calibrate in data - *skipped here*
  - interpretability of the models, i.e. understand what the network learned.

GENERALITIES

- in ATLAS (run 2) tracking occurs up to $|\eta|<2.5$ in a 2T axial B field
  - typically 4 pixel measurements points, 8 Si microstrip tracker points, and many other in the TRT for $|\eta|<2$.
- Samples: t-tbar with at least one W decaying leptonically (PowhegBox v2+Pythia 8.230 + EvtGen)
- Tracking: general quality requirements
  - >= 7 hits in the silicon layers (pixel and SCT, where dead sensors are not penalised),
  - <= 2 missing hits where expected in the silicon layers,
  - <= 1 hit shared by multiple tracks,
  - >= 1 hit in the pixel detector, and $|\eta| < 2.5$.
  - Primary vertex = highest sum of pt^2

Tracks and Jets

- jets (Antikt4EMPflow, calibrated) must have
- p_T > 20 GeV and $|\eta|<2.5$,
- no overlap with generator level muons or electrons from the W, and must pass the jet vertex tagger optimized for particle flow jets (to suppress pileup)

- tracks are associated to jets based on a ΔR matching (depending on p_T, max Dr =0.45 at p_T=20 GeV, Dr=0.25 at pT=200GeV)
- tracks must have pT > 1GeV, $|d0| < 1$ mm, and $|z0 \sin\theta| < 1.5$ mm.
- Jets are labelled, in order
  - b-jets: at least one b-hadron (from MC truth) with p_T > 5GeV and ΔR with respect to the jet axis < 0.3
  - c-jets, as before
  - τ-jets, as before
  - else light-flavor jet

PAST ALGORITHMS

- The IP3D algorithm uses
  - d_0 and $z_0\sin\theta$ distances (or significance) are used to build templates in 14 non overlapping regions (based on the track hit patter) for b-jets, c-jets, light jets
  - tracks are assigned probabilities of coming from b-jets, c-jets, light jets based on the templates (built with simulation)
  - PDF for the track parameters within a jet are taken as independent:
    - => jet level probabilities are derived

$$D_{\mathrm{IP3D},l} = \log \prod_{i \in \mathrm{tracks}} \frac{p_b^i}{p_l^i} \qquad D_{\mathrm{IP3D},c} = \log \prod_{i \in \mathrm{tracks}} \frac{p_b^i}{p_c^i}.$$

- The RNN based algorithm aims to overcome this overly simplistic assumption of independence + it adds new input features
- RNNs (introduced in ATL-PHYS-PUB-2017-003) uses LSTM (long short term memory) cell to preserve long range correlations between the elements of the *sequence*; this improves performance over IP3D even when using exactly the same input - NOTE: *RNN are sequential non parallelizable algorithms*

DIPS vs RNN - 1

- DIPS performance is studied in comparison with RNNIP, an evolution of the RNN algorithm is use within the suite of FTag production algorithms;
  - ***RNNIP architecture:*** To be understood: difference with current standard implementation of the RNN in FTag
    - *100 nodes hidden layer of LSTM*
    - *Dropout layer (dropout fraction 20%)*
    - *20 node fully connected layer for classification*
  - inputs

Input	Description
s_{d0}	d_0/σ_{d0}: Transverse IP significance
s_{z0}	$z_0 \sin\theta/\sigma_{z0 \sin\theta}$: Longitudinal IP significance
$\log p_T^{frac}$	$\log p_T^{track}/p_T^{jet}$: Logarithm of fraction of the jet p_T carried by the track
$\log \Delta R$	Logarithm of opening angle between the track and the jet axis
IBL hits	Number of hits in the IBL: could be { 0, 1, or 2 }
PIX1 hits	Number of hits in the next-to-innermost pixel layer: could be { 0, 1, or 2 }
shared IBL hits	Number of shared hits in the IBL
split IBL hits	Number of split hits in the IBL
nPixHits	Combined number of hits in the pixel layers
shared pixel hits	Number of shared hits in the pixel layers
split pixel hits	Number of split hits in the pixel layers → created by multiple charged particles
nSCTHits	Combined number of hits in the SCT layers
shared SCT hits	Number of shared hits in the SCT layers

Table 1: Track features used as inputs for RNNIP and DIPS algorithms.

Tracks associated to the jet are ordered by decreasing s_{d0}

The first **15 tracks** are used

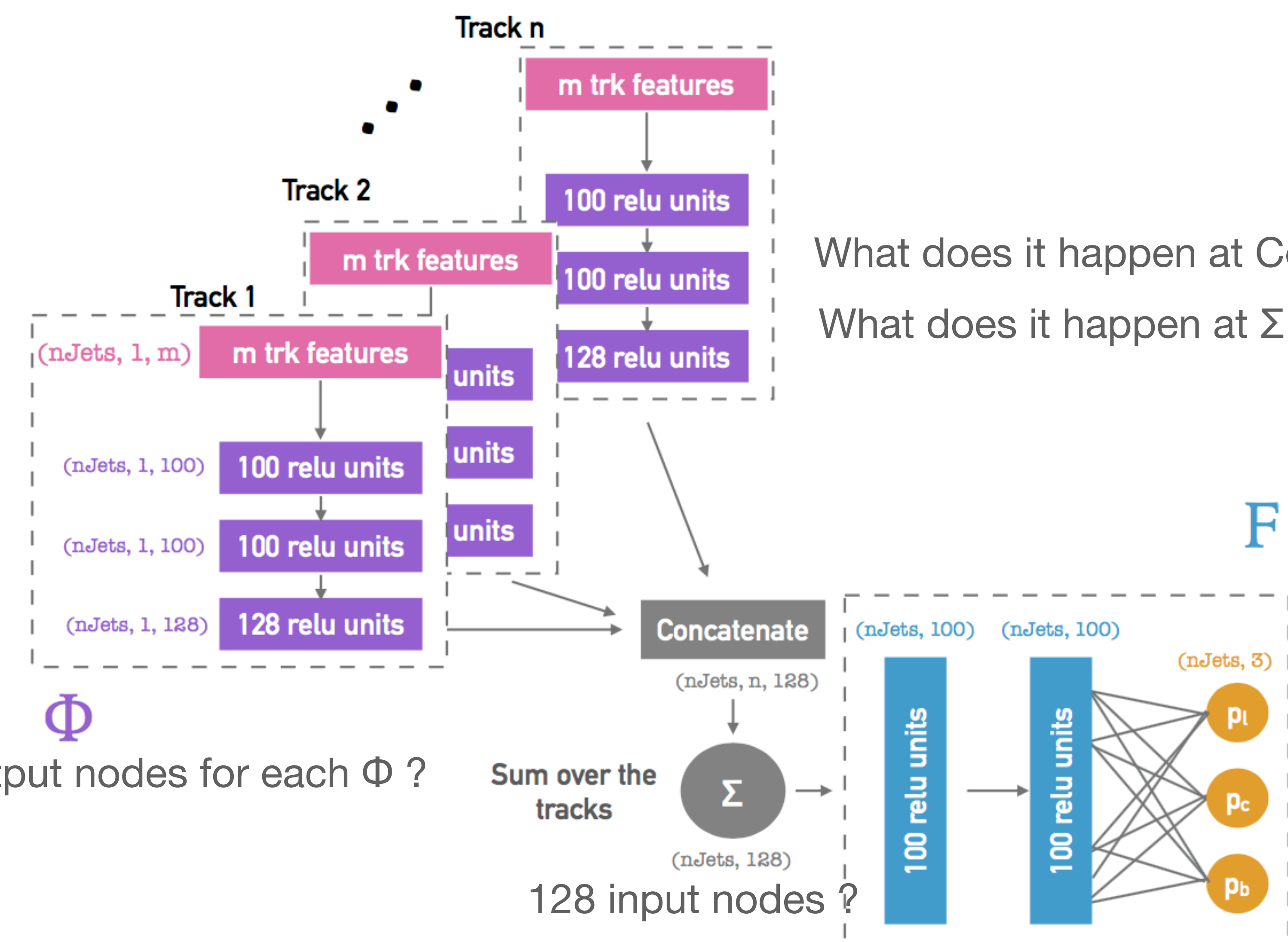
DIPS vs RNN - 2

- DIPS architecture has intrinsically no dependence on the order of the elements of a *set* (or arbitrary size) instead of a *sequence*
  - A NN (Φ) is applied to all inputs (p_i) of a track *[bonus: operation of processing the tracks in the jet with the network can be easily parallelised -> GPU]*
    - Φ (track network) extracts relevant track features
  - The sum of the Φ output is processed with a feed-forward NN (F) *[naturally encodes track permutation invariance]*
    - F (jet network) extracts relevant jet features, giving probabilities for b-, c- light flavour jets
  - The output is a multi-class classification : p_b, p_c, p_l combined into a b-tagging discriminant D_b

$$D_b = \log \frac{p_b}{(1 - f_c)p_l + f_c p_c}$$

    - f_c can be optimized post-training
    - (free parameter) accounting for the fraction of c-jets in the non-b jets

ARCHITECTURE



What does it happen at Concatenate ?

What does it happen at Σ (Sum over tracks) ?

Is nJets the jet index ?

128 output nodes for each Φ ?

128 input nodes ?

Figure 2: Architecture for the DIPS algorithm. The number of hidden units in the different neural network layers correspond to the final optimized architecture.

RESULTS

- DIPS gains 15% (5%) extra rejection power vs light (c-) jets at the same b-jet efficiency vs RNNIP

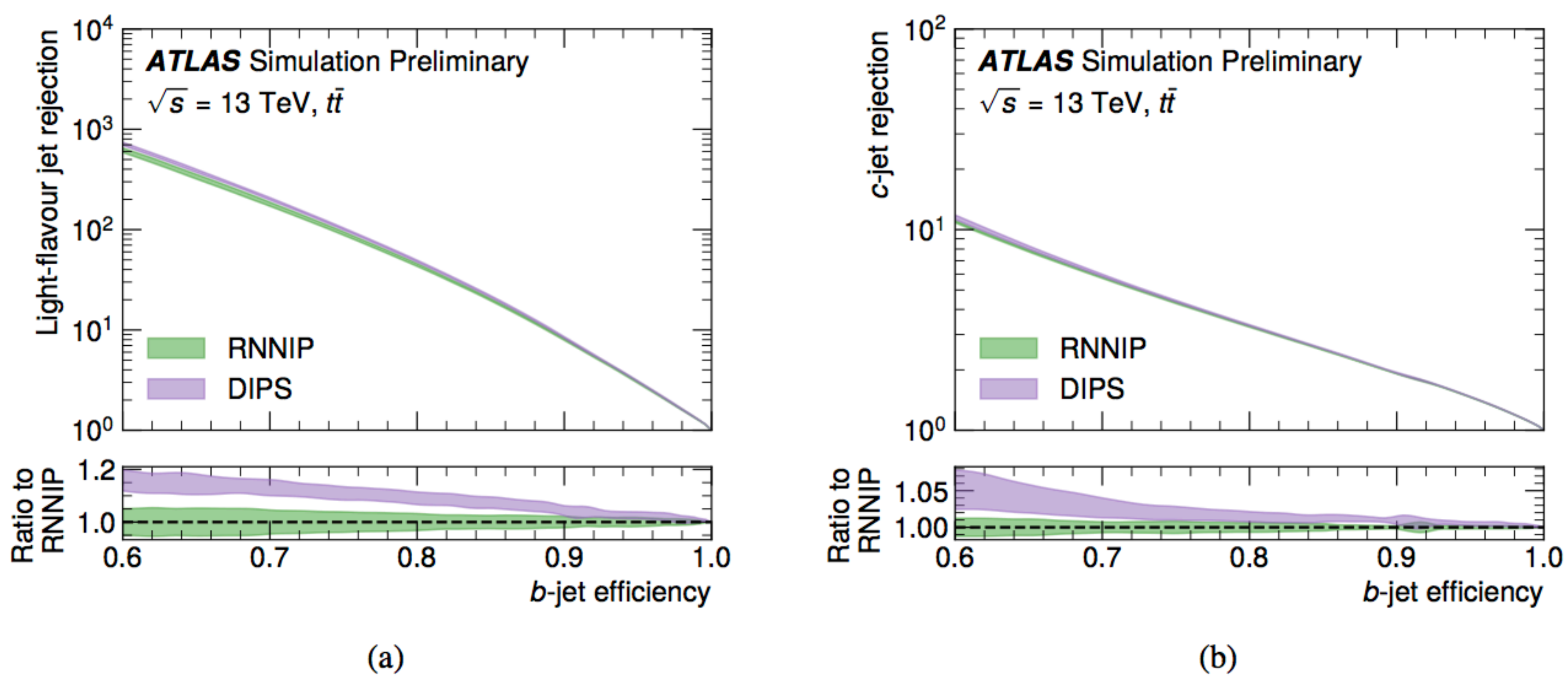


Figure 4: Light-flavour jet rejection as a function of b-jet efficiency (a) and c-jet rejection as a function b-jet efficiency (b) of the RNNIP (green) and DIPS (purple) algorithms. The central curves and error bands show the mean and standard deviation, respectively, of the rejection at each b-jet efficiency for 5 trainings. The ratios are computed with respect to the RNNIP ROC curve.

Results- features learned

$\frac{\partial D_b}{\partial x_{ik}}$ >0 => powerful signal indicator

$\frac{\partial D_b}{\partial x_{ik}}$ <0 => background indicator

$\frac{\partial D_b}{\partial x_{ik}}$ ~0 => not very informative

- In summary:
  - Only the first (max s_{d0}) ~5 tracks matter
  - s_{d0} is the most useful feature, in strong correlation, for the first track, with the p_T (relative to the jet p_T) and ΔR, distance from the jet axis
    - *harder fragmentation of b-quarks w.r.t. c- and light-flavor quarks*
  - Split/shared IBL (and pixel) hits in the tracks with largest s_{d0} are indicative of background

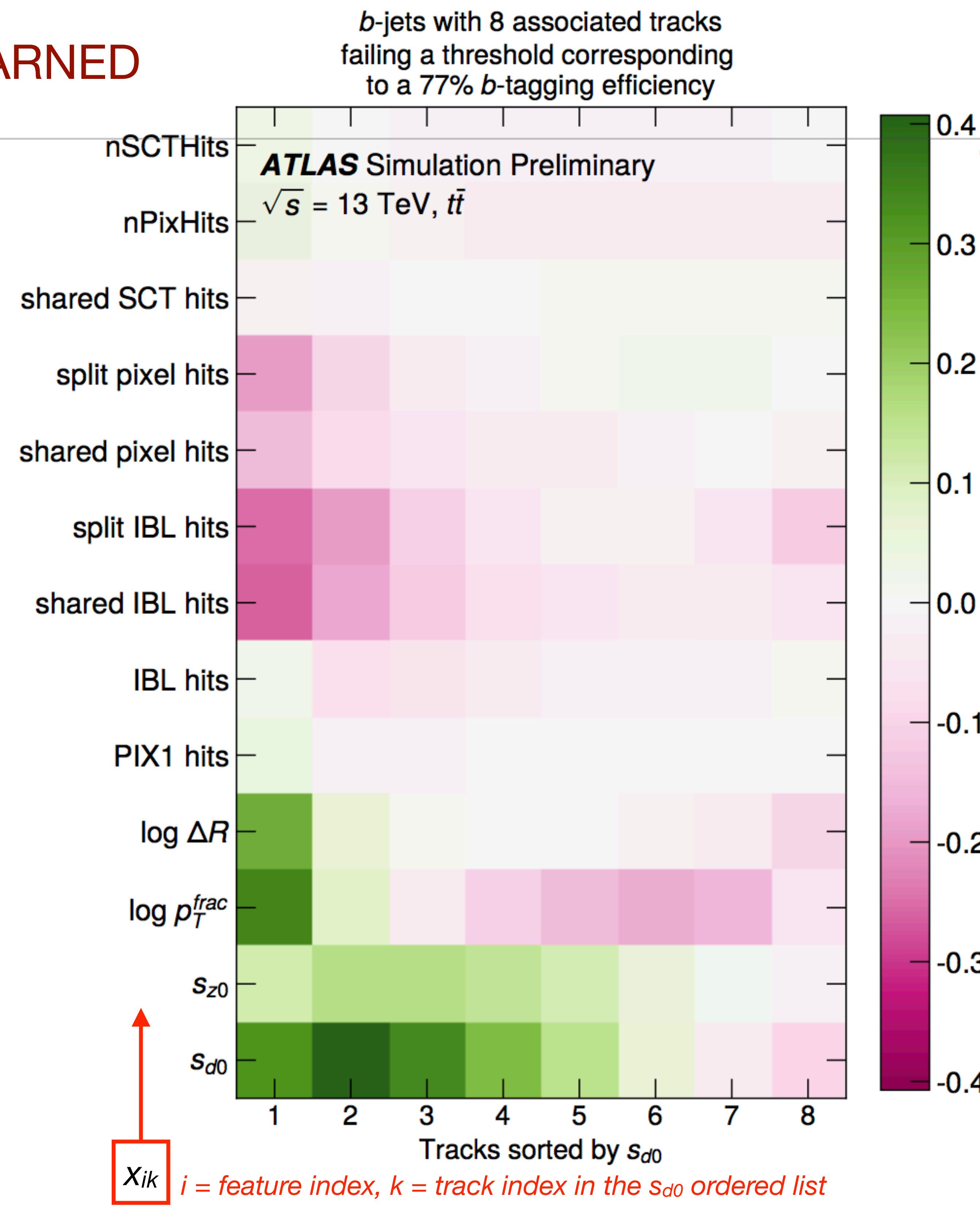


Figure 5: Saliency map for b-jets with 8 tracks. The track features are shown on the y-axis, the tracks (ordered by s_{d_0}) are listed on the x-axis. The colors in each pixel represent the gradient defined in Equation 4.

$\frac{\partial D_b}{\partial x_{ik}}$

N jets

$$\frac{\partial D_b}{\partial x_{ik}} = \frac{1}{N}\sum_{j=1}^{N}\frac{\partial D_b^{(j)}}{\partial x_{ik}^{(j)}}$$

$$D_b = \log \frac{p_b}{(1-f_c)p_l + f_c p_c}$$

Training end execution time

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

Table 2: Timing metrics for trainings performed on Nvidia 2080 Ti GPUs. The nominal value denotes the mean of five independent trainings, while the error bar is the standard deviation.

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

Table 3: Timing metrics for the full test dataset (3 million jets) with GPU evaluations on an NVIDIA Titan X GPU. The nominal value denotes the mean of five independent trainings, while the error bar is the standard deviation.

- Quicker convergence w.r.t. RNNIP

TRACK OPTIMIZATION

- Nominal:
  - tracks (up to 15) must have $p_T > 1$GeV, $|d0| < 1$ mm, and $|z0 \sin\theta| < 1.5$ mm.
- Loose:
  - tracks (up to 25) must have $p_T > 0.5$GeV, $|d0| < 3.5$ mm, and $|z0 \sin\theta| < 5$ mm.
  - >4x more pileup tracks, +25% more fragmentation/hadronization tracks, +20% more b-related tracks
- Loose + new features: d_0 and $z_0\sin\theta$

Jet Flavour	Track selection	n_{trk}	n_{trk}^{HF}	n_{trk}^{hadr}	n_{trk}^{other}
b-jets	*nominal*	5.9 ± 2.7	3.4 ± 1.8	2.0 ± 1.9	0.4 ± 0.8
	loose	8.1 ± 3.2	3.9 ± 1.8	2.5 ± 2.1	1.7 ± 1.7
c-jets	*nominal*	5.1 ± 2.5	1.7 ± 1.0	2.9 ± 2.2	0.4 ± 0.8
	loose	7.1 ± 3.1	1.8 ± 1.0	3.6 ± 2.4	1.7 ± 1.7
Light-flavour jets	*nominal*	4.6 ± 2.6	-	4.1 ± 2.5	0.5 ± 0.9
	loose	6.8 ± 3.3	-	5.0 ± 2.7	1.8 ± 2.0

Table 4: The average per jet total number of tracks (n_{trk}), the number of tracks from heavy flavour decays (n_{trk}^{HF}), the number of tracks from hadronisation, excluding those from heavy flavour decays (n_{trk}^{hadr}), and the number of tracks from mismeasurement, material interactions, and pile-up (n_{trk}^{other}), are shown for the *nominal* and *loose* selections for each jet flavour.

TRACK OPTIMIZATION

- Nominal:
  - tracks (up to 15) must have p_T > 1GeV, |d0| < 1 mm, and |z0 sin θ| < 1.5 mm.
- Loose:
  - tracks (up to 25) must have p_T > 0.5GeV, |d0| < 3.5 mm, and |z0 sin θ| < 5 mm.
  - >4x more pileup tracks, +25% more fragmentation/hadronization tracks, +20% more b-related tracks
- Loose + new features: d_0 and $z_0 \sin\theta$

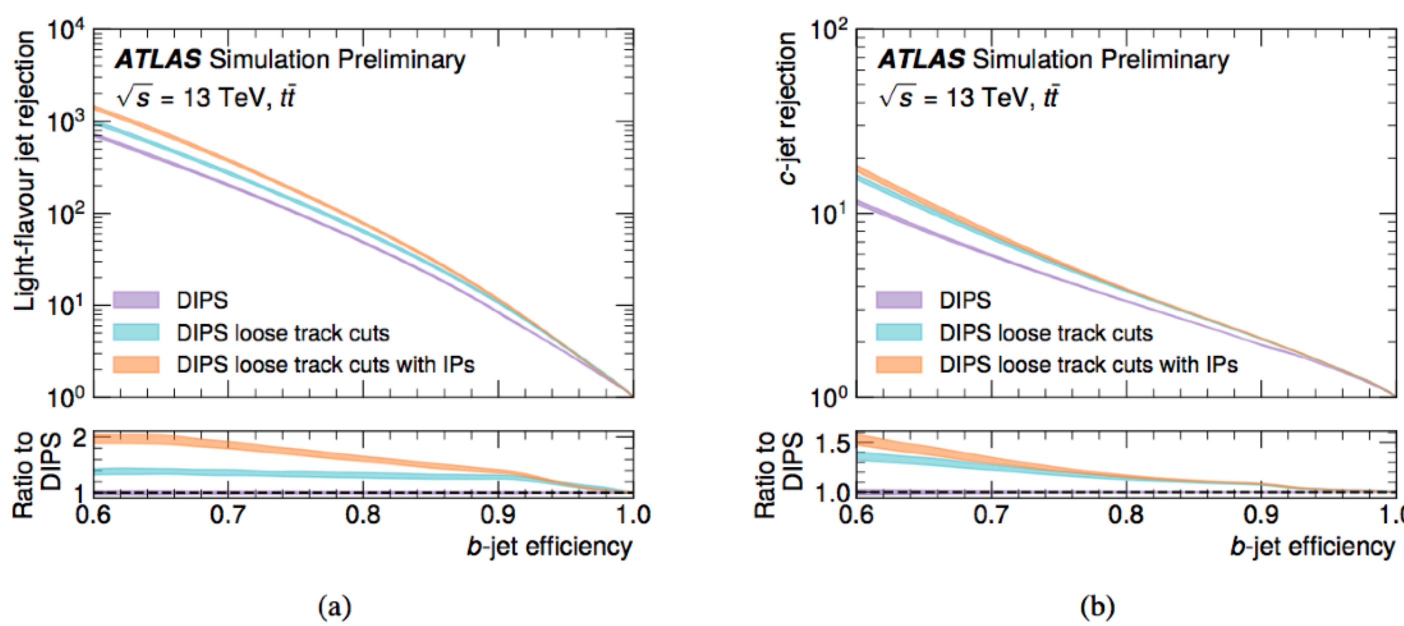


Figure 8: Light-flavour jet rejection as a function of *b*-jet efficiency (a) and *c*-jet rejection as a function of *b*-jet efficiency (b) of the nominal DIPS setup, DIPS with *loose* track selection, and Optimised DIPS with the *loose* track selection and additional IP inputs. The central curves and error bands show the mean and standard deviation, respectively, of the rejection at each *b*-jet efficiency for 5 trainings. The ratios are computed with respect to the DIPS ROC curve.

77% Working Point

b-jet eff.

c-jet rejection

l-jet rejection