Expected b-tagging performance with the ATLAS Phase 2 detector and MVA developments for ttH(bb) analysis

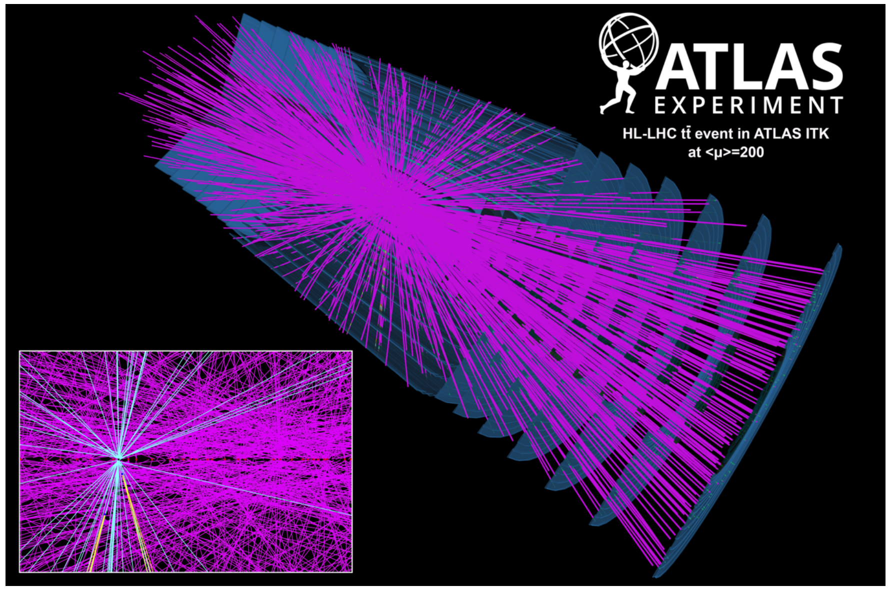
Neelam Kumari

Thesis Supervisors:
Arnaud Duperrin and **Thomas Strebler**

CPPM 3rd year PhD students seminar
Centre de Physique des Particules de Marseille
Aix-Marseille Université / IN2P3-CNRS

December 06th, 2021

Outline



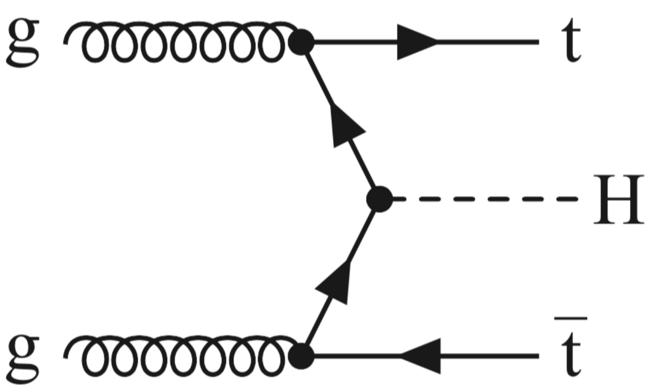
Introduction

- Top-Yukawa Coupling, TTHbbAnalysis
- b-tagging algorithms

Part-I

Expected b-tagging performance with the ATLAS Phase 2 detector

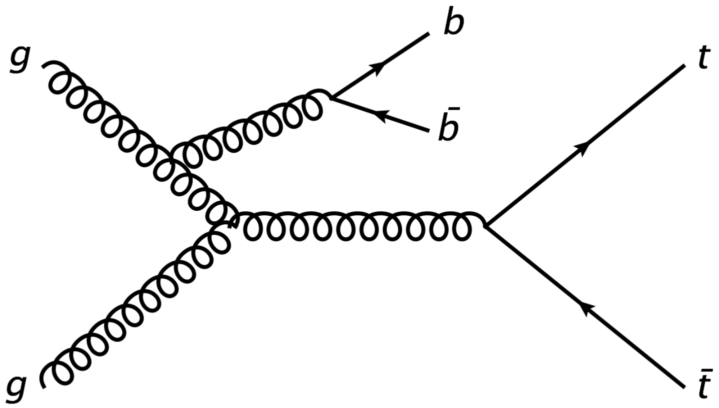
- HL-LHC, ITk
- b-tagging performance with ITk



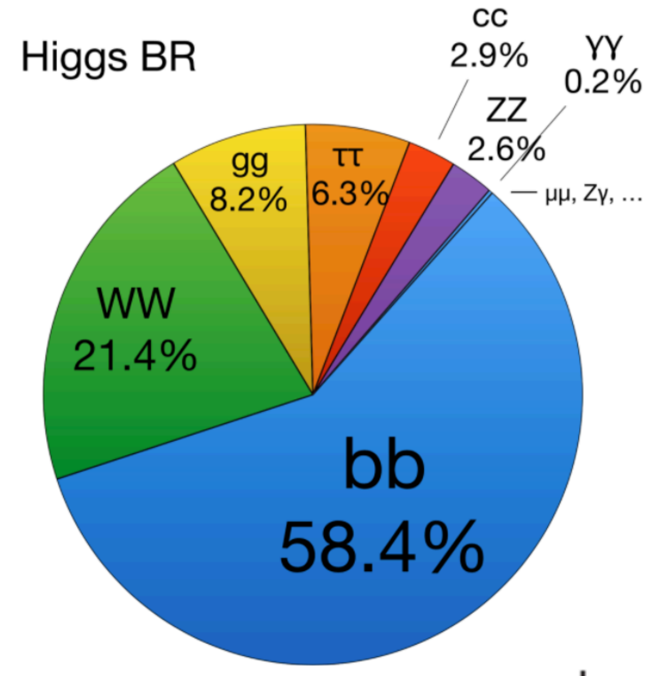
Part-II

MVA developments for ttHbb analysis

- BDT trainings
- DNNs development

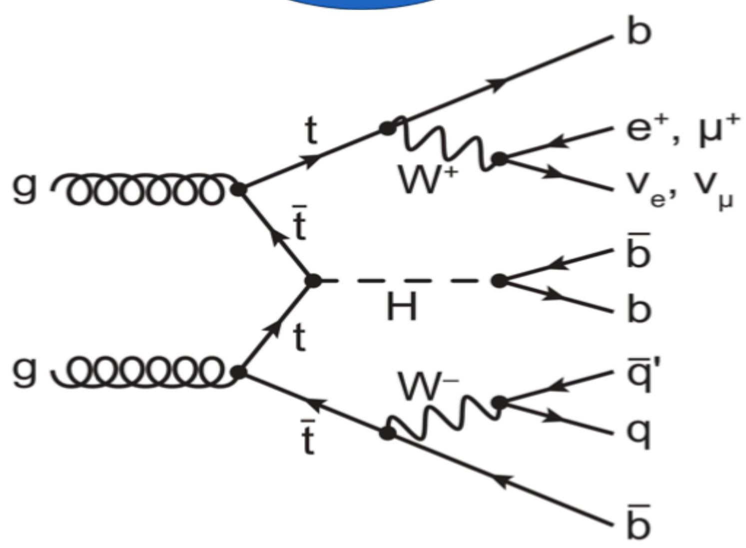


Introduction



TTH and Top Yukawa coupling

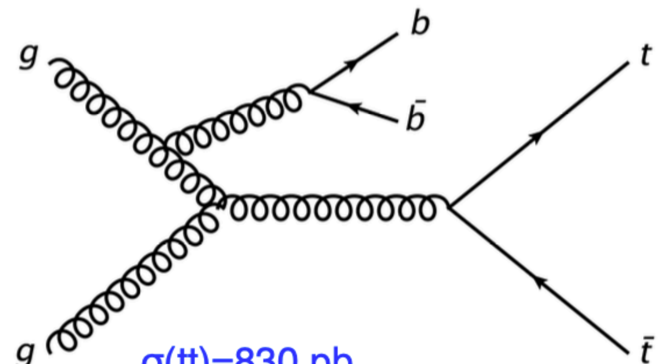
- Probing the **top-Higgs Yukawa y_t**: Largest in Standard Model and sensitive to potential New Physics
- The associated production of a Higgs boson and a top quark-antiquark pair (**ttH production**) provides direct probe of y_t
- First observation was made in 2018, combining Run 1 and partial Run 2 LHC data



σ(ttH)xBR(H->bb) = 0.3 pb

Why ttH (H→bb) ?

- ttH(bb) channel exploits the large branching ratio of **H→bb (58%)** and the leptonic decays of top quarks has distinctive signature
- Two channels based on the number of leptons in the final state: **single-lepton** and **di-lepton**
- **Challenges**:
  - Modeling of tt+HF background due to large irreducible background with big theoretical uncertainty (from tt+bb)
  - Higgs boson reconstruction challenging due to b-jet combinatorics

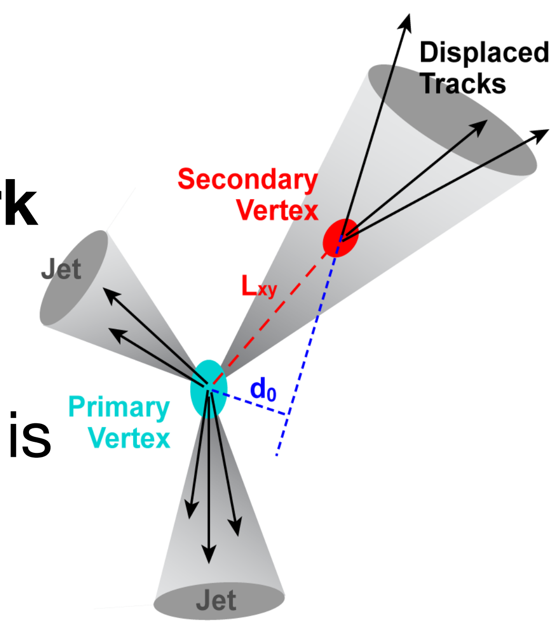


σ(tt)=830 pb
σ(tt+b-jets) ~ O(10) pb

b-tagging

ttH studies depend greatly on b-tagging (used to identify the b-quark content of jets):

- Top decays produce b quarks
- Identification of the H→bb candidate is a vital aspect of the TTH (H→bb) analysis

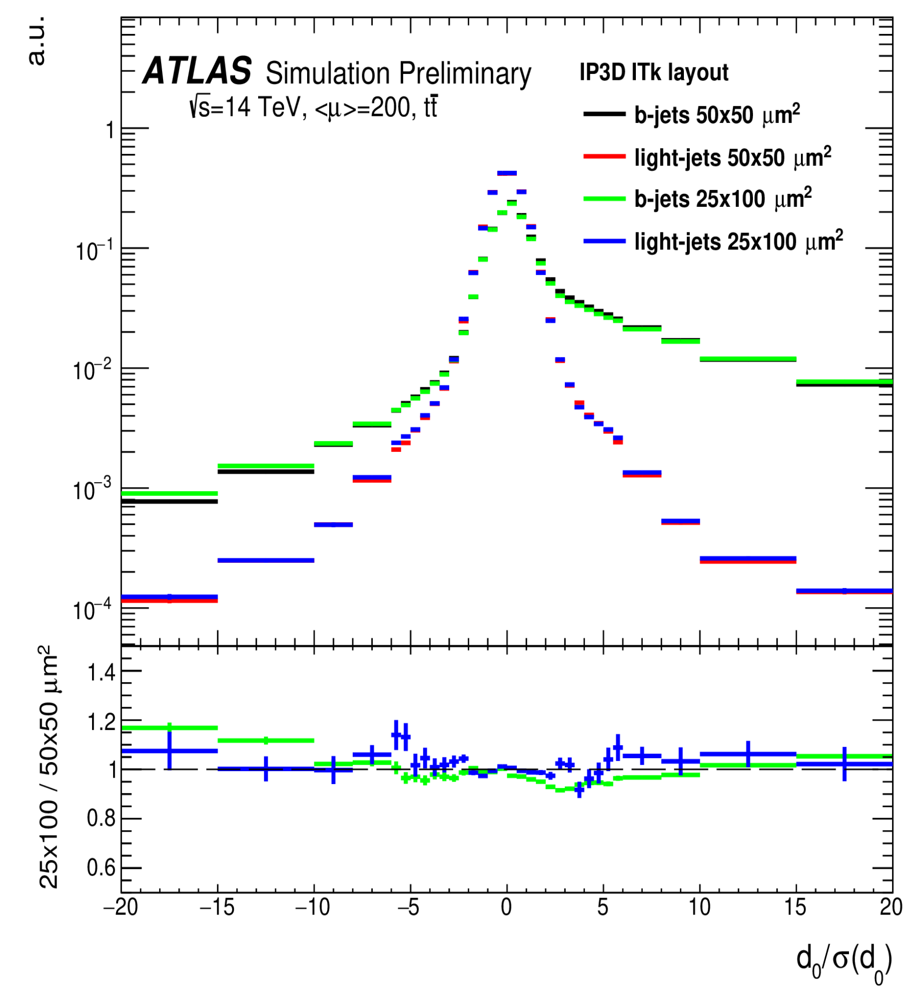


b-tagging is also key to many physics analyses at High Luminosity LHC e.g HH production

b-jet tagging rely on B-hadron properties:

- Displaced vertex from primary vertex (PV) called **secondary vertex (SV)** due to its long life (~1.5ps)
- Tracks from B-decays have **large Impact parameters (IP)**

1. **IP based tagging algorithm: IP3D**
   - Assigns probabilities to tracks based on 2D likelihood templates PDFs, with the **z0sin(θ)** and **d_0** lifetime of signed significances of tracks built from b and light simulated jets
2. **SV based algorithms: SV1**
   - Exploits 4 vertex properties : **Vertex mass, ΔR** (between the jet and PV-SV line), **Energy fraction**, **Number of two-track vertices**
3. **MV2:** High-level BDT-based taggers exploiting outputs of low-level taggers including IP3D and SV1



Expected b-tagging performance with the ATLAS Phase 2 detector

HL-LHC and ATLAS Inner Tracker

The **High-Luminosity Large Hadron Collider** (HL-LHC) project aims to boost up the performance of the LHC in order to increase the potential for discoveries after 2027.

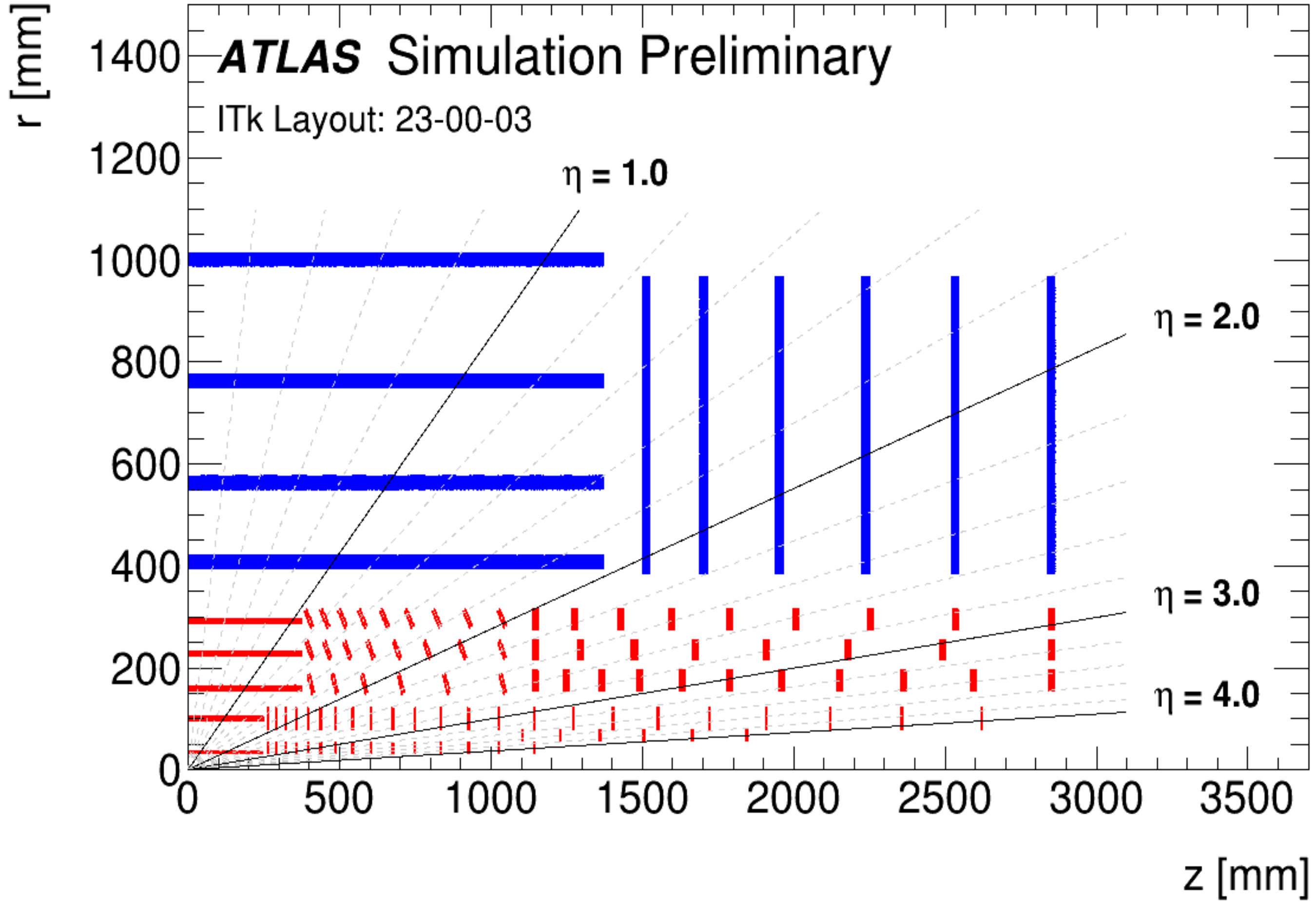
Challenges

Increase in pile-up (3-4x Run 3) will require better tracking

Increase in integrated luminosity (10x wrt Run 1-3) imposes improved **radiation-hardness**

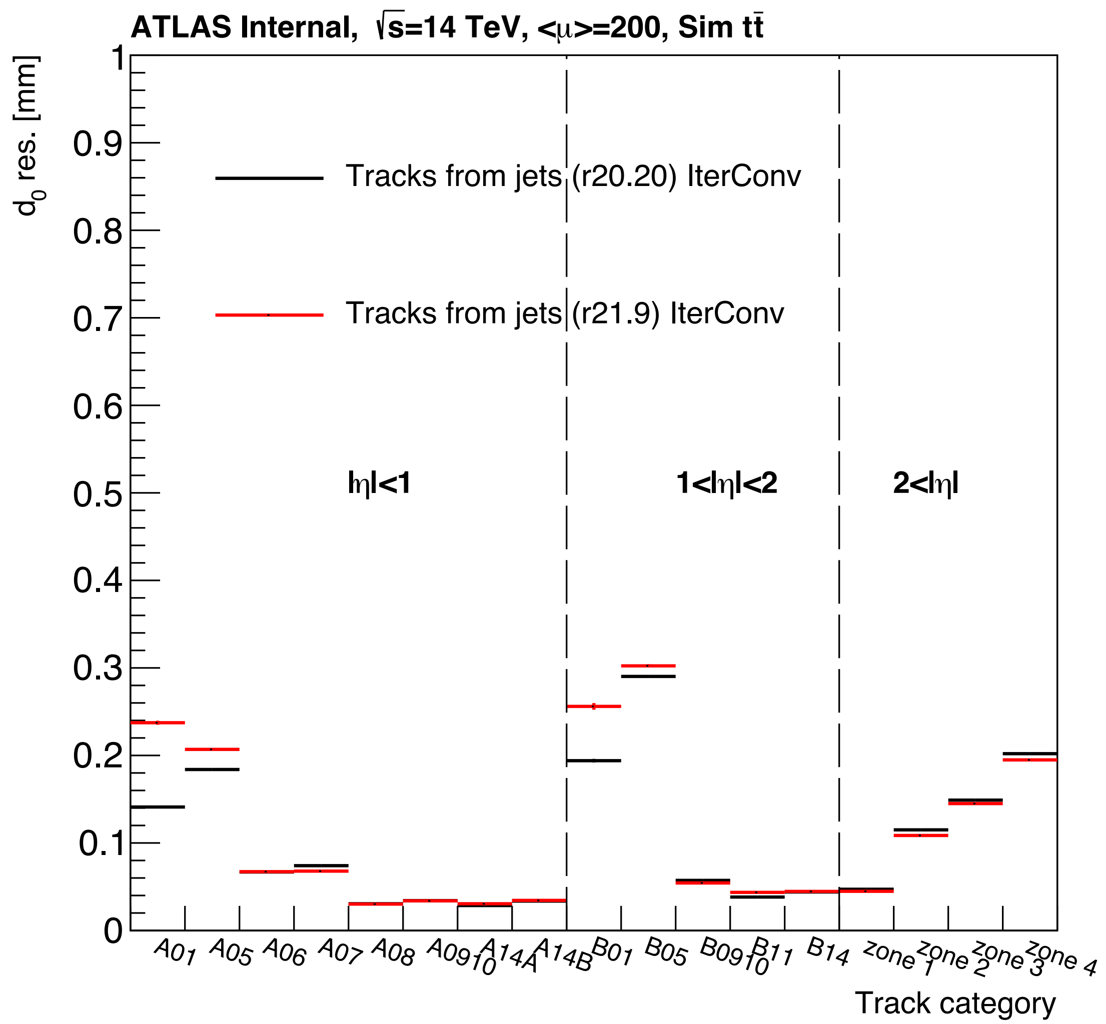
Current ATLAS Inner Detector (ID) replaced by Inner Tracker (ITk) to maintain tracking performance in harsh conditions

- Extended forward pseudo- rapidity from **2.5 to 4** for increased tracking acceptance
- All-silicon design consists of inner **Pixel** (|η|<4) and outer **Strip** (|η|<2.7) sub-detectors
- Latest ITk layout design with innermost pixel layer closer with respect to previous versions (**R=34 mm**)



ITk b-tagging framework

- B-tagging algorithms had already been optimised for ITk with the previous ATLAS software release 20.20 (PUB note [ATL-PHYS-PUB-2020-005])
- Release 20.20 had been phased out in early 2020: **Developments done in switching to latest upgrade software release 21.9 to stay synchronised with latest developments related to ITk simulation and tracking**
- Developments done were used for performances shown in the 2021 PUB note [ATL-PHYS-PUB-2021-024



Track categories:

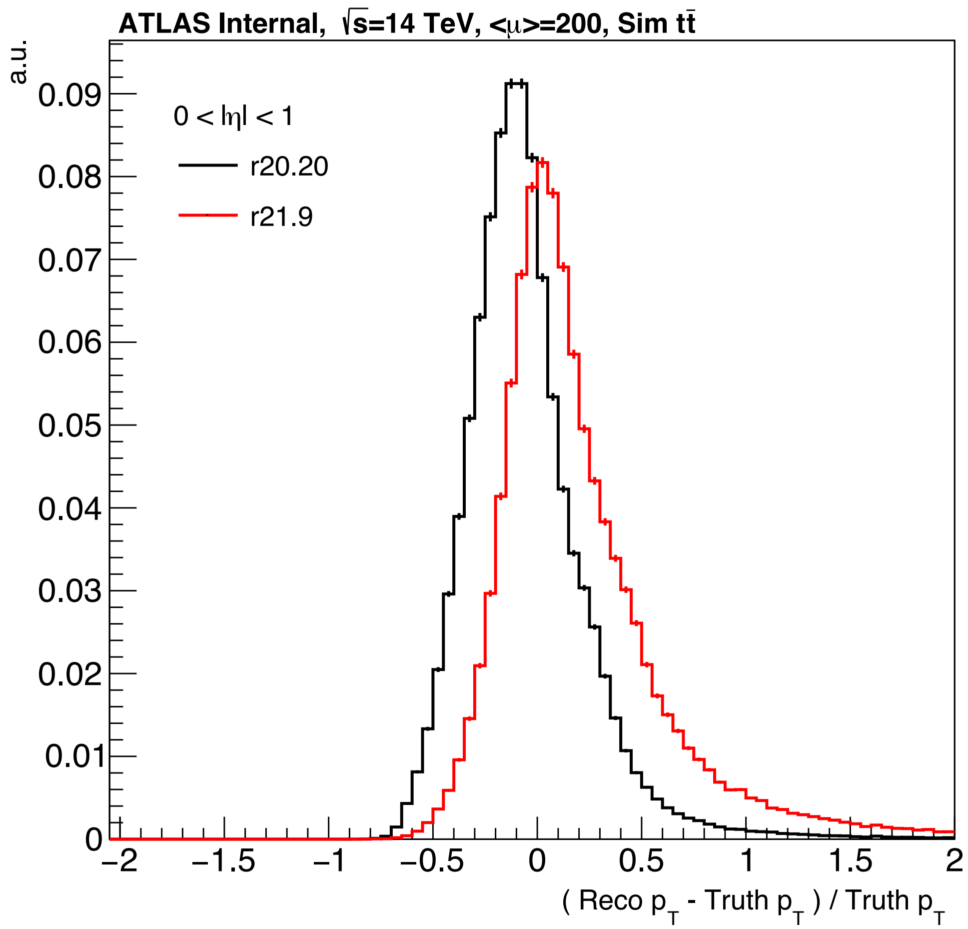
- PDF templates for IP based taggers are obtained in 14 exclusive categories defined by the hit patterns of the tracks
- Track hit content for |η|<1 (region A) + 1<|η|<2 (region B) + track kinematic for |η|>2

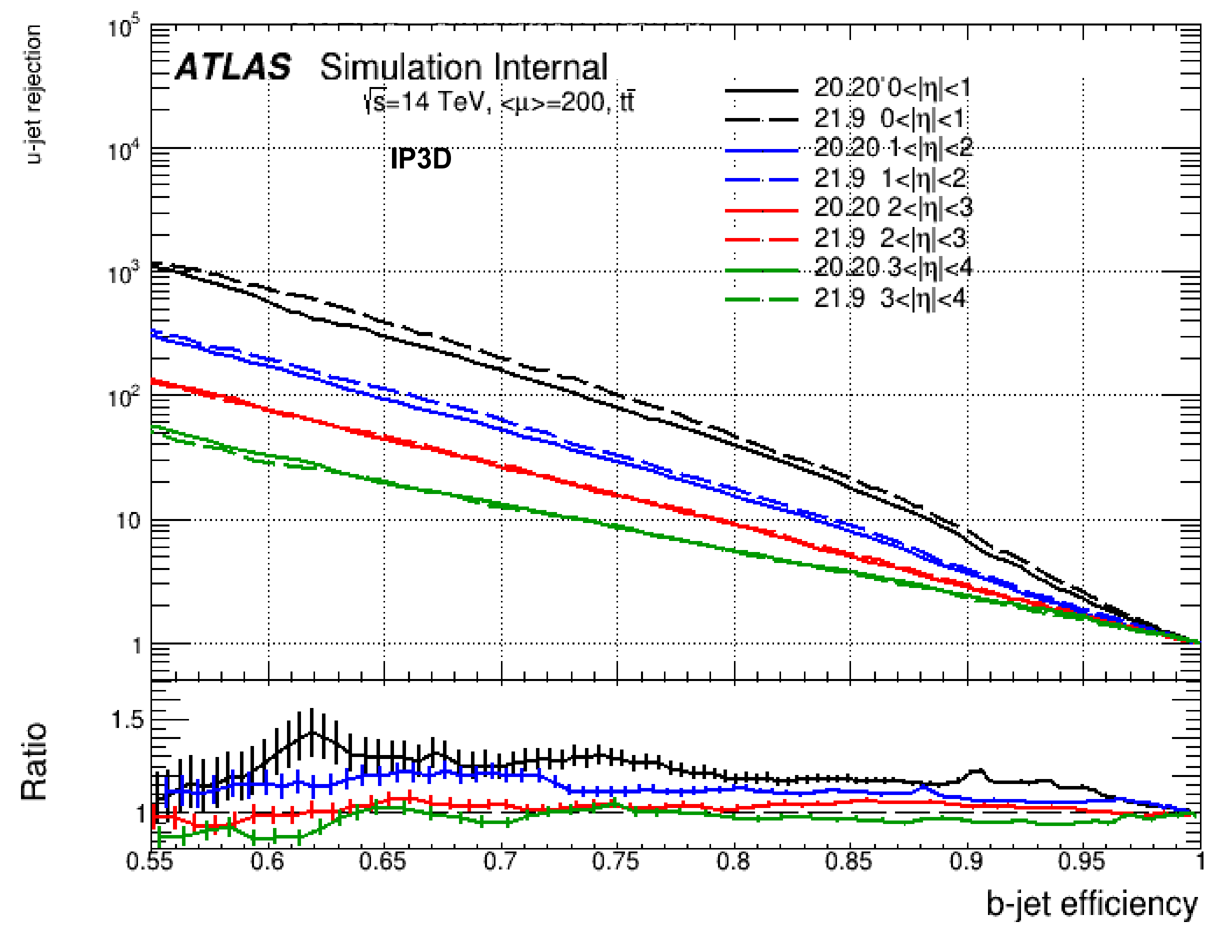
- *d0 Resolution per track categories consistent between the releases*
- *[1]Differences arise due to low statistics of tracks in release 21.9*

[1] Plots in back-up

b-tagging performance

- Reconstructed jet p_T and truth jet p_T (truth p_T = p_T of truth jet matched to reco jet) sizeably different and visible **differences in jet energy response**
- The selection cut on $p_{T\,(Truth)}$ >20 GeV produced similar agreement to the default selection cut on $p_{T\,(Reco)}$ >20 GeV
- **Re-weighted the $p_{T\,(Truth)}$ spectrum to make things more compatible for performance checks**

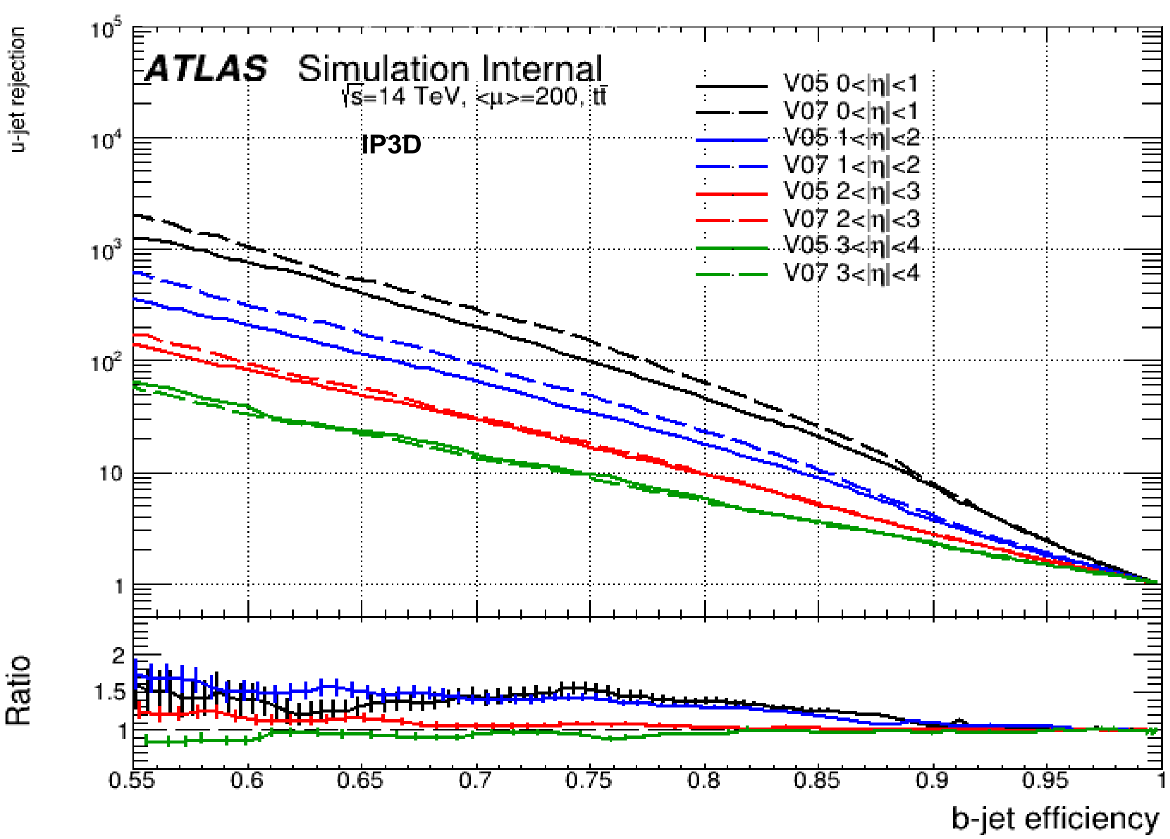




- *IP3D performance consistent between releases in different η and p_T region and is within **20%-30%***
- *SV1 Tagger performance within **30%-40%***

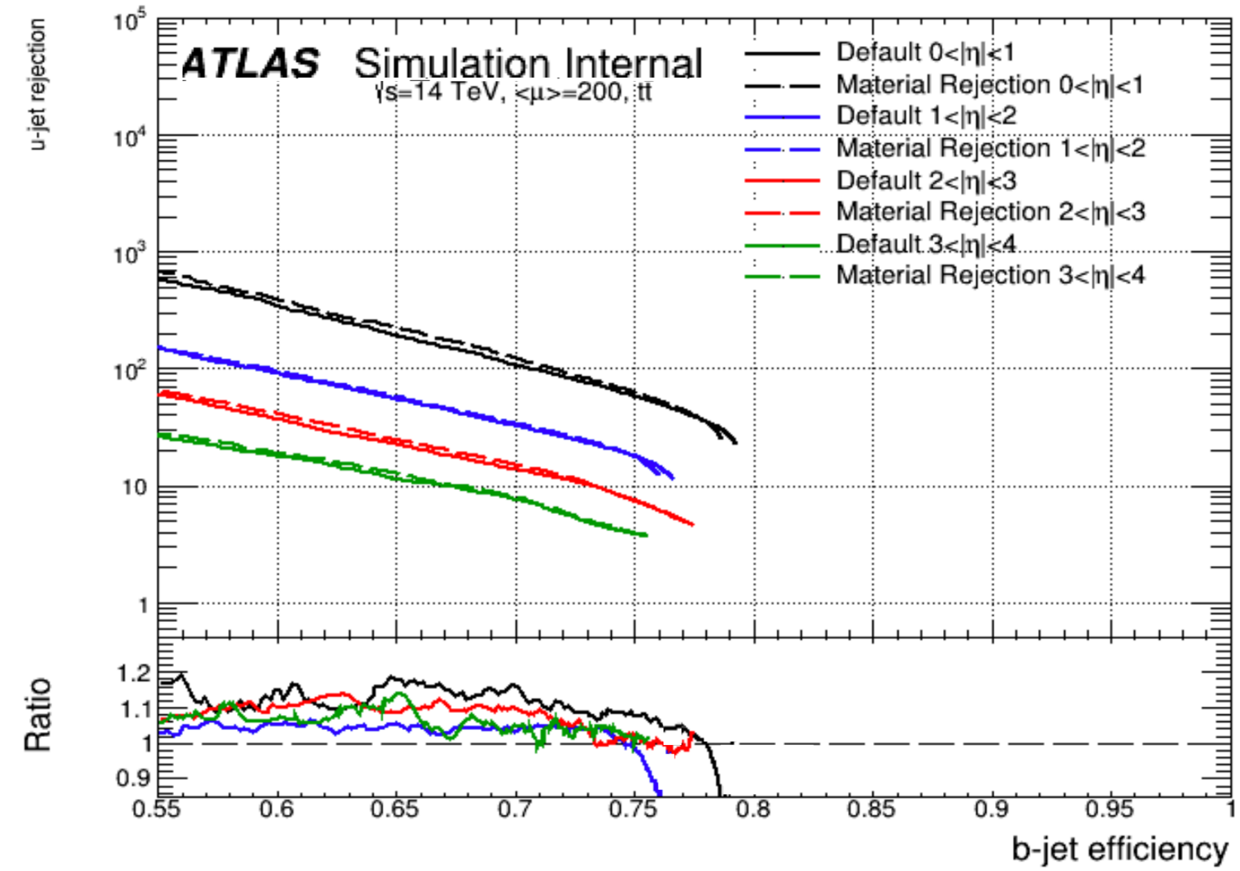
Re-optimised IP3D categorisation & Material rejection

- Studies were already done in release 20.20 to further **optimise Track categories** which highlighted possible improvements when exploiting more detailed categories based on **p_T** or **hit content: Synchronised developments in current release 21.9**



- IP3D performance improve with new categorisation in region |η|<2 of 50-60 %

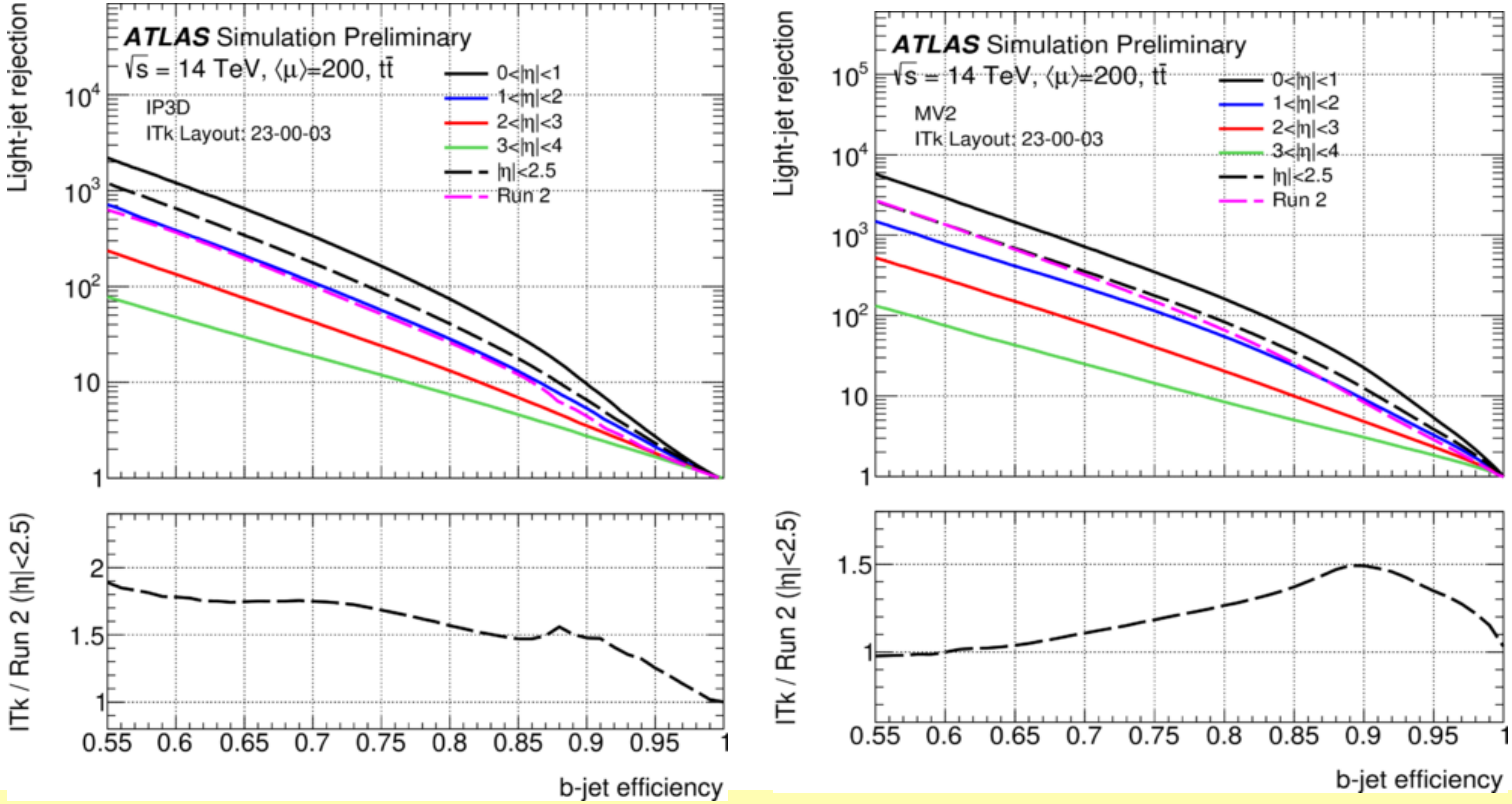
- SV reconstructed when particle has interacted with the detector material becomes a sizeable source of SV for light jets
- **SV reconstructed near pixel layers must be vetoed**



- *Slight better performance observed with Material rejection up-to* ***5-20 %***

PUB note results

- Overall improvements were highlighted in **ITk PUB note**
- b-tagging performance for the ITk compared to the **Run 2** performance



- ***IP3D performance improved***: *Improved IP resolution expected with ITk*
- ***Better performance with MV2***: *Displays for a 77% b-jet efficiency working point a light-jet rejection 20% higher than Run 2 detector performance, driven by the IP3D improved performance*

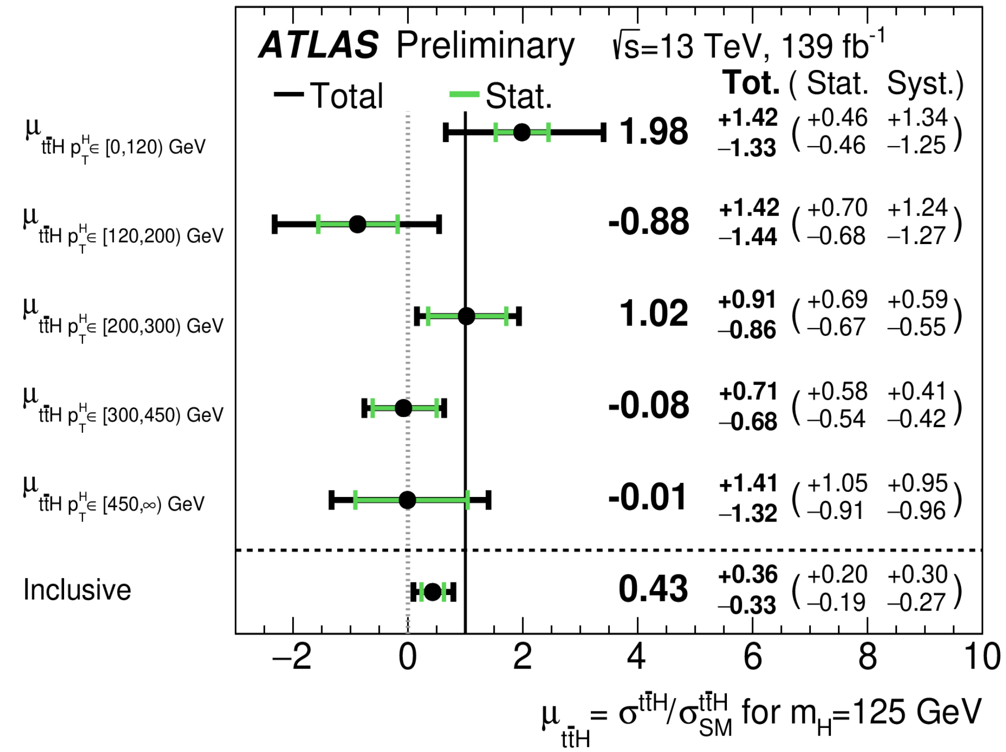
MVA developments for ttH(bb) analysis

ttH (H→bb) leptonic: Analysis strategy

Higgs measurement explored through **Simplified Template Cross Sections** (STXS) formalism where cross-section is measured as a function of the **p_T^H**

- Events categorised in signal regions (**SRs**) are defined by the **#leptons**, **#jets**, **#b-tagged jets** (4 working points) and **#boosted Higgs boson candidates**
- The single-lepton channel is split into boosted and resolved channels
- Events split by p_T^H: (0-120), (120-200), (200-300), (300-450) and (450,∞)
- Define control regions (**CRs**) to constrain tt+≥1b and tt+≥1c

Region		Dilepton				Single-lepton			
		$\mathrm{SR}^{\geq 4j}_{\geq 4b}$	$\mathrm{CR}^{\geq 4j}_{3b\ \mathrm{hi}}$	$\mathrm{CR}^{\geq 4j}_{3b\ \mathrm{lo}}$	$\mathrm{CR}^{3j}_{3b\ \mathrm{hi}}$	$\mathrm{SR}^{\geq 6j}_{\geq 4b}$	$\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$	$\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$	$\mathrm{SR}_{\mathrm{boosted}}$
#leptons		= 2				= 1			
#jets		≥ 4			= 3	≥ 6	= 5		≥ 4
#b-tag	@85%	–				≥ 4			
	@77%	–				–			$\geq 2^{\dagger}$
	@70%	≥ 4	= 3			≥ 4			–
	@60%	–	= 3	< 3	= 3	–	≥ 4	< 4	–
#boosted cand.		–				0			≥ 1
Fit input		BDT	Yield			BDT/Yield	$\Delta R^{\mathrm{avg}}_{bb}$		BDT



***Previous round results**: Measurement of the ttH production with ttH(bb) events using full Run-2 (139 fb⁻¹) luminosity with 1 or 2 leptons (e/μ) in final state [HIGG-2020-23]*

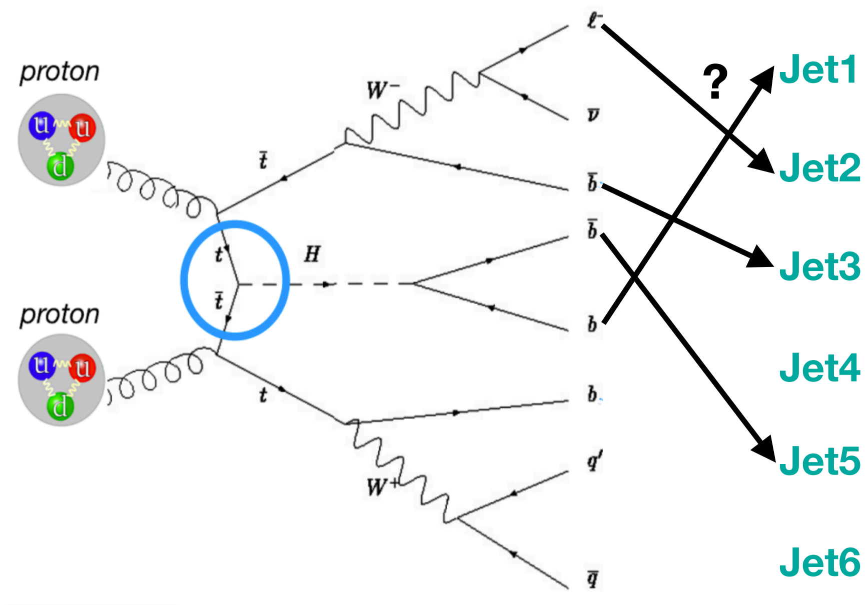
Signal-strength measurements in the individual STXS p_T^H bins and inclusive signal strength

ttH (H→bb) leptonic: MVA

*Multivariate classifiers are used in two parts in the analysis for **reconstructing Higgs boson candidate** objects and **classifying ttH signal events***

Reconstruction:

- **BDT training** per jet-parton combination on signal is used to find all the correct combinations (highest BDT score) and rest treated as background
- **Major emphasis is to efficiently reconstruct the Higgs boson candidate correctly in a given STXS bin + define high-level variables used to discriminate ttH from tt+bb e.g "Higgs mass"**



Each possible combination will end up with a corresponding reconstruction of objects, e.g. Higgs boson, etc.

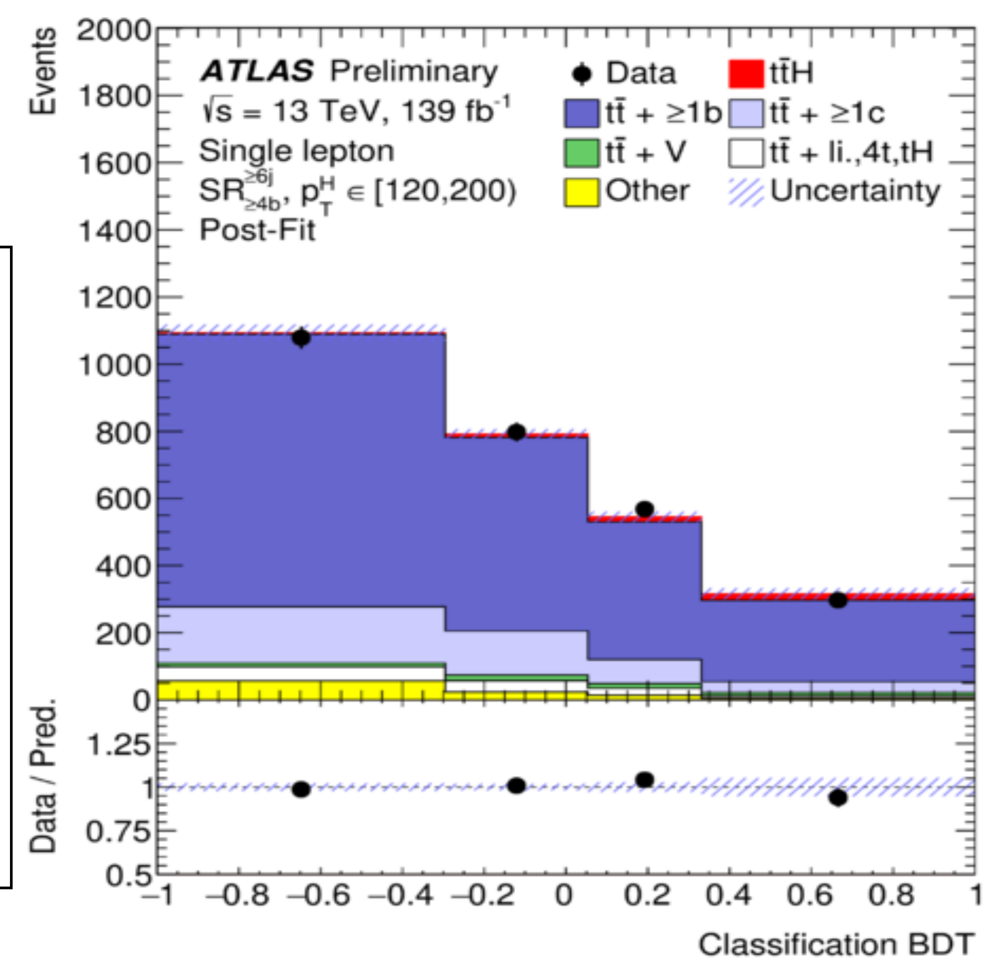
Reconstructed efficiency in STXS bin

p_T^H [GeV]	Dilepton $SR^{\geq 4j}_{\geq 4b}$	Single-lepton $SR^{\geq 6j}_{\geq 4b}$	Single-lepton $SR_{boosted}$
Inclusive	51%	43%	91%
[0, 120)	43%	35%	—
[120, 200)	50%	45%	—
[200, 300)	64%	57%	—
[300, 450)	78%	59%	90%
[450, ∞)			93%

Previous round results: [HIGG-2020-23]

Classification:

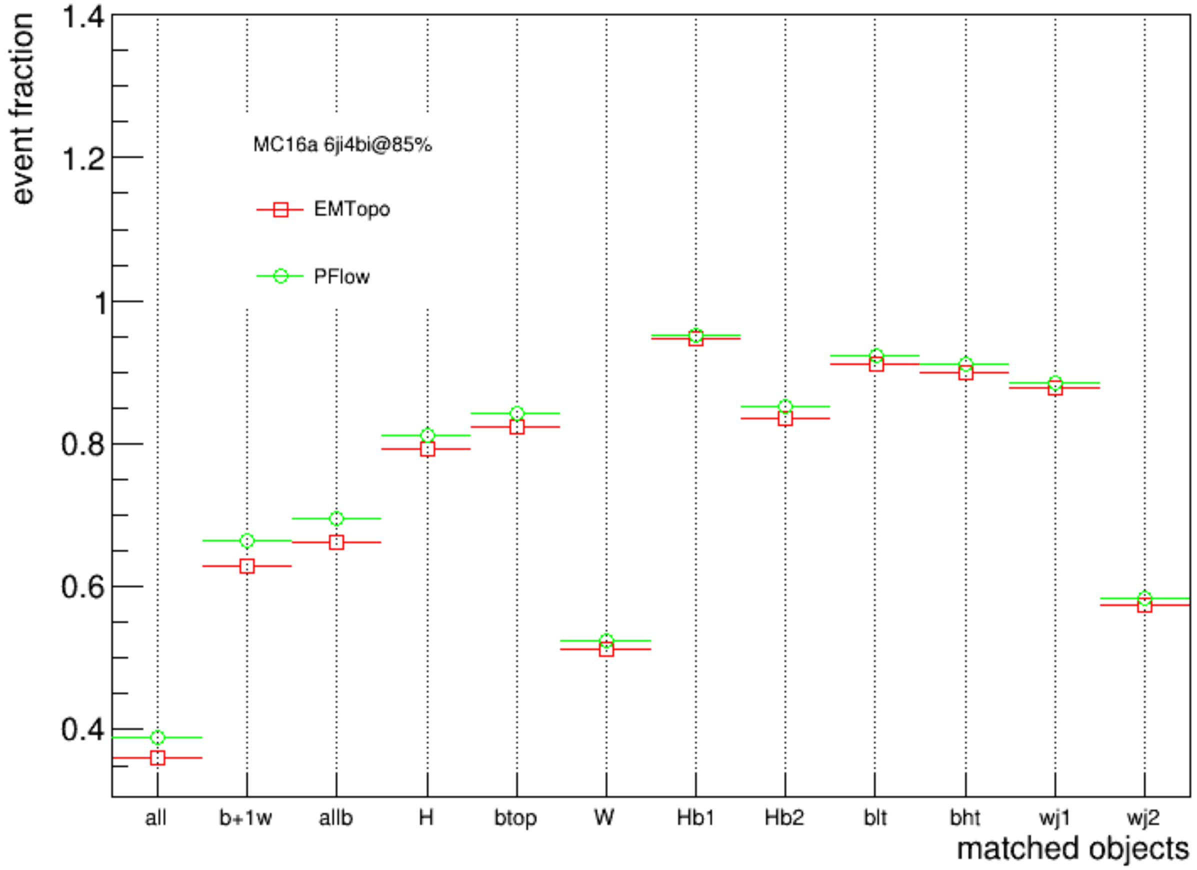
- Discriminate signal from background, using kinematic properties as well as reconstruction BDTs and output is used in signal+background fits in signal regions

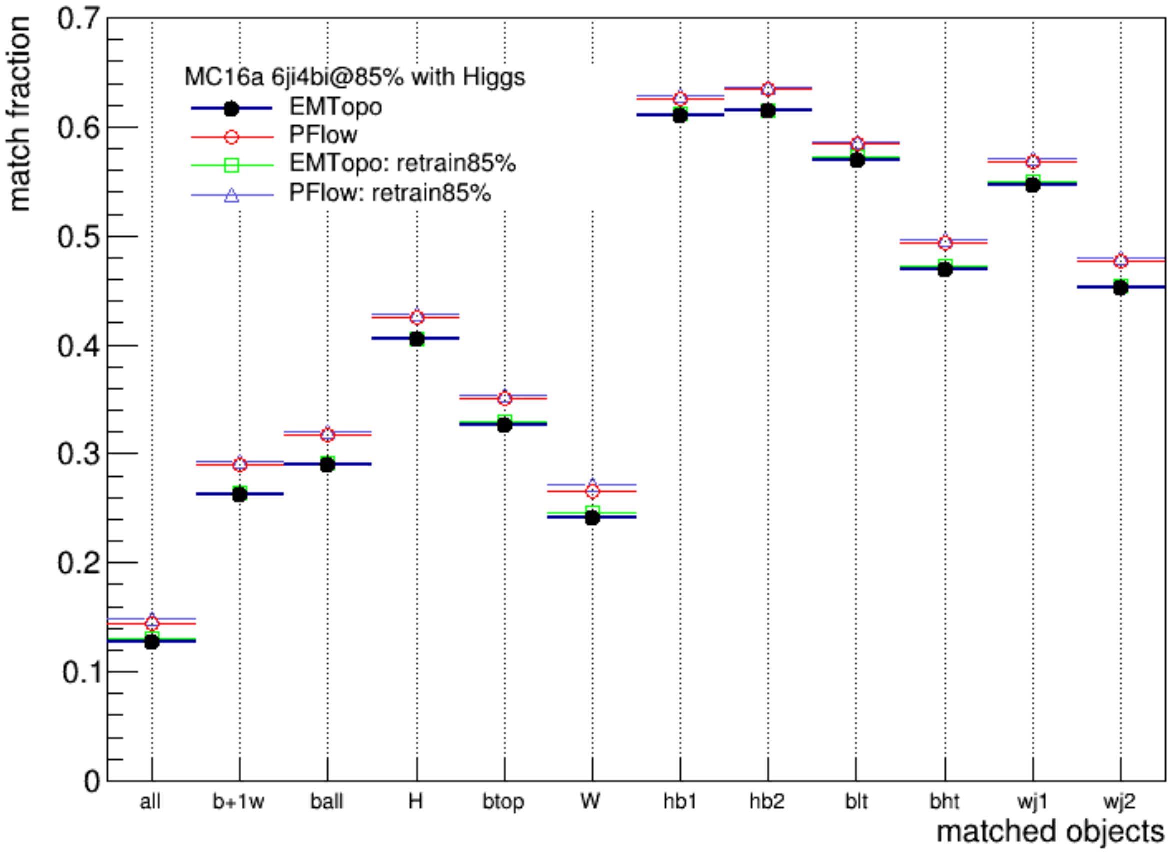


Reconstruction BDT performance

*BDTs are retrained in next round of TTH(bb) Analysis exploiting the recent jet analysis and more performant b-tagging algorithms (**DL1r vs MV2c10**)*

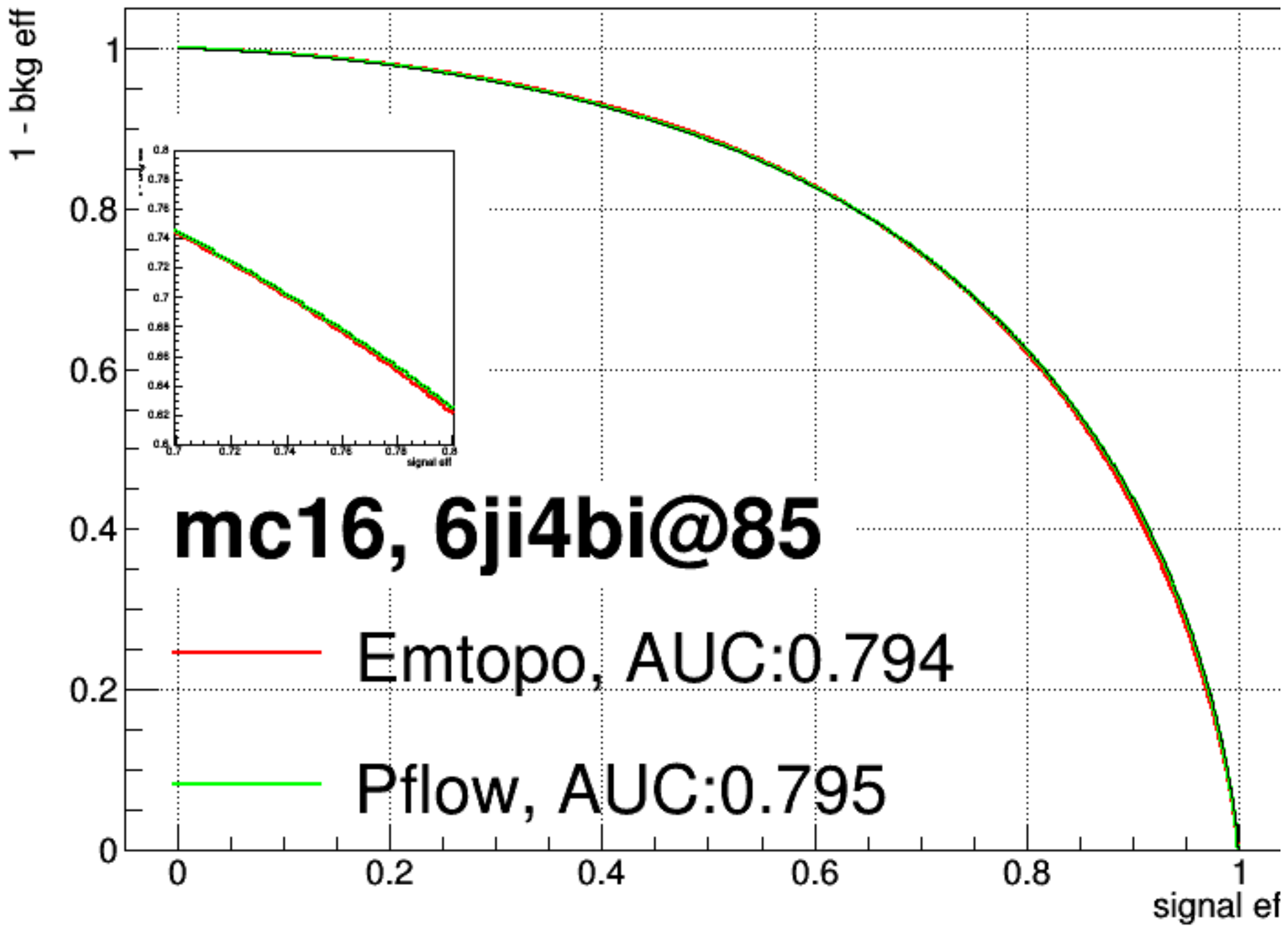
- **Event fraction**: Fraction of events where the truth object is reconstructed
- **Match fraction**: Fraction of events where the truth object is reconstructed + selected the permutation with best BDT score

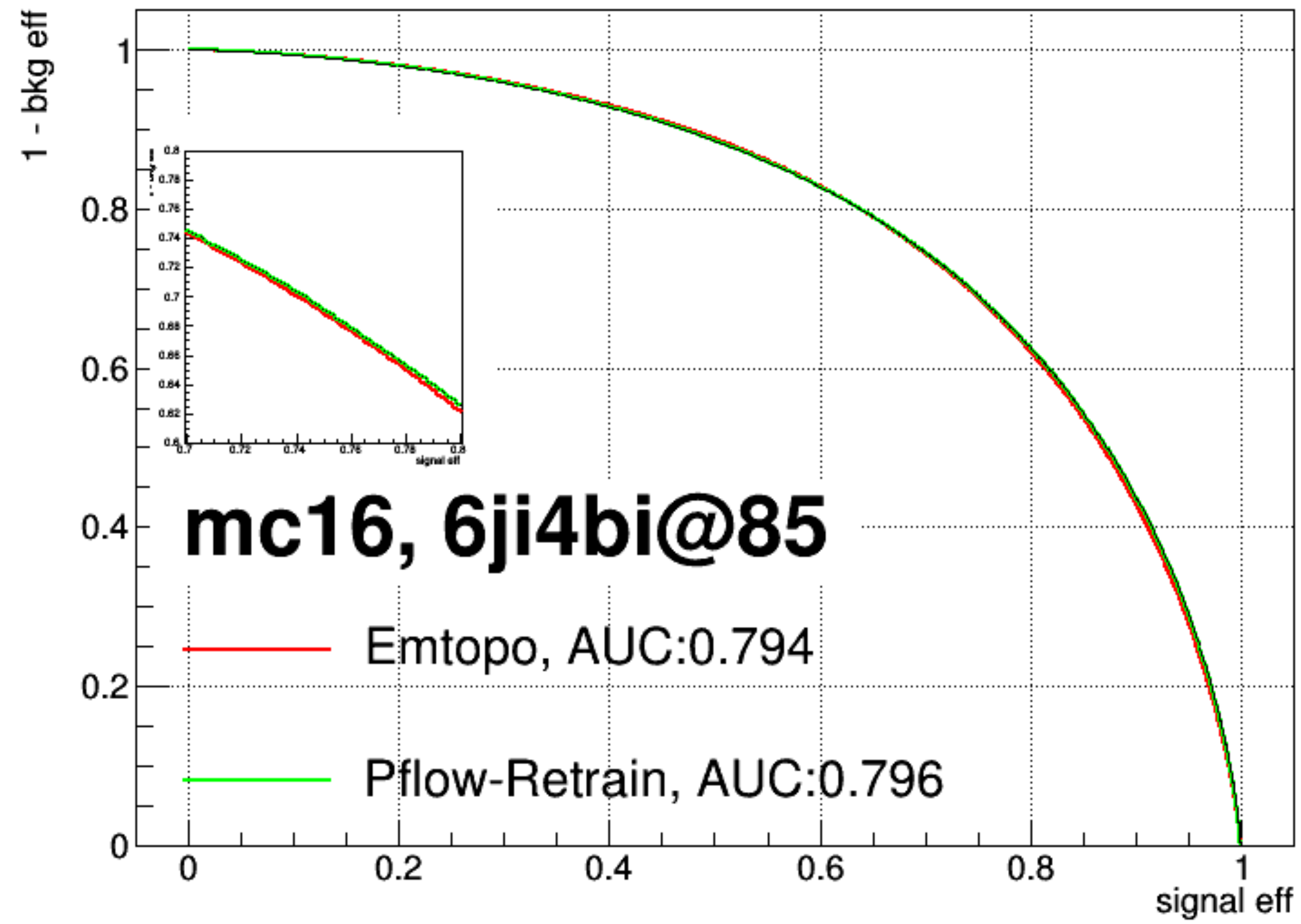




- *Slightly larger fraction of events where the truth objects are reconstructed with PFlow*
- *Reconstruction performance using PFlow/DL1r jets is slightly better than EMTopo/MV2c10 jets performance and retraining showed negligible impact on the performance*

Classification BDT performance





CLassBDT performed using old weights using EMtopo/MV2c10 similar performance

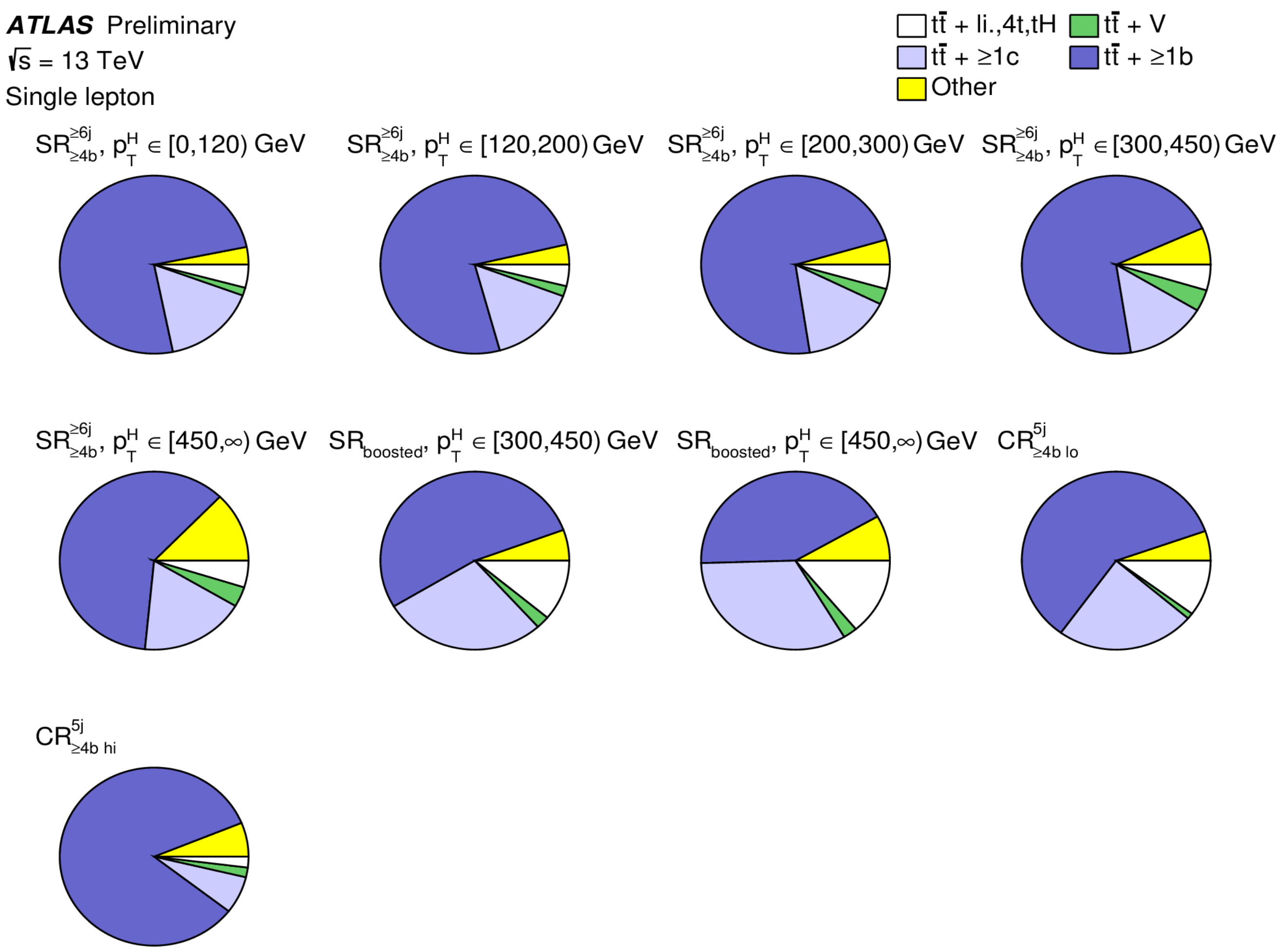
CLassBDT performed with new weights using PFlow/DL1r (re-training) shows negligible impact on the performance

The study also provided a similar baseline using PFlow jets in order to further approach to new MVA architectures using **Deep Neural Networks (DNNs)**

Implement DNNs which is used for binary classification are expected to perform better than BDTs

DNNs developments

- Targeting to perform multi-classification focusing on separating signal from $t\bar{t}$ background but also building **signal vs $t\bar{t}$+light vs $t\bar{t}$+cc vs $t\bar{t}$+b vs $t\bar{t}$+2b vs $t\bar{t}$+bb** discriminant with a probability associated with each background
- Gives the possibility to build Control Regions naturally dominated by a single background component→**Constrain the uncertainties of the different processes and increases the overall sensitivity**



Previous round results: [HIGG-2020-23]

DNN architectures offers flexibility to compute with a single tool:

Perform regression on Higgs kinematics variables to be used for differential STXS measurements

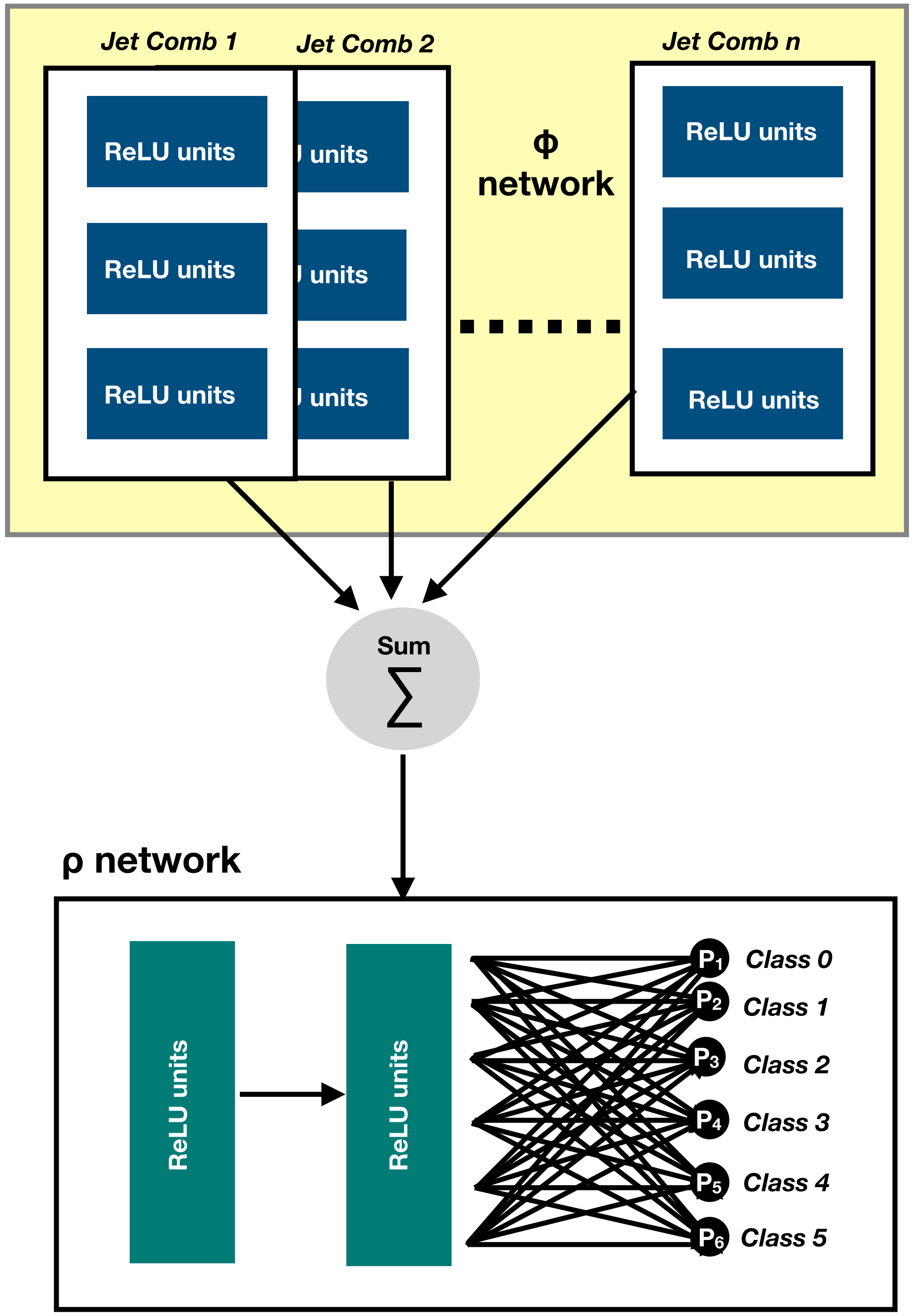
↓

Multi-process classification with discriminants to be used to build CR + perform shape fits for signal extraction

DNNs: Deep-sets

Deep Sets setup:

- treat each element as a set without a specific order (**permutation invariance**)
- Each jet combination is processed by **ϕ** network (of several connected layers)
- These are then **summed Σ** up and the output is processed using **ρ** network consisting of similar connected hidden layers



Inputs:

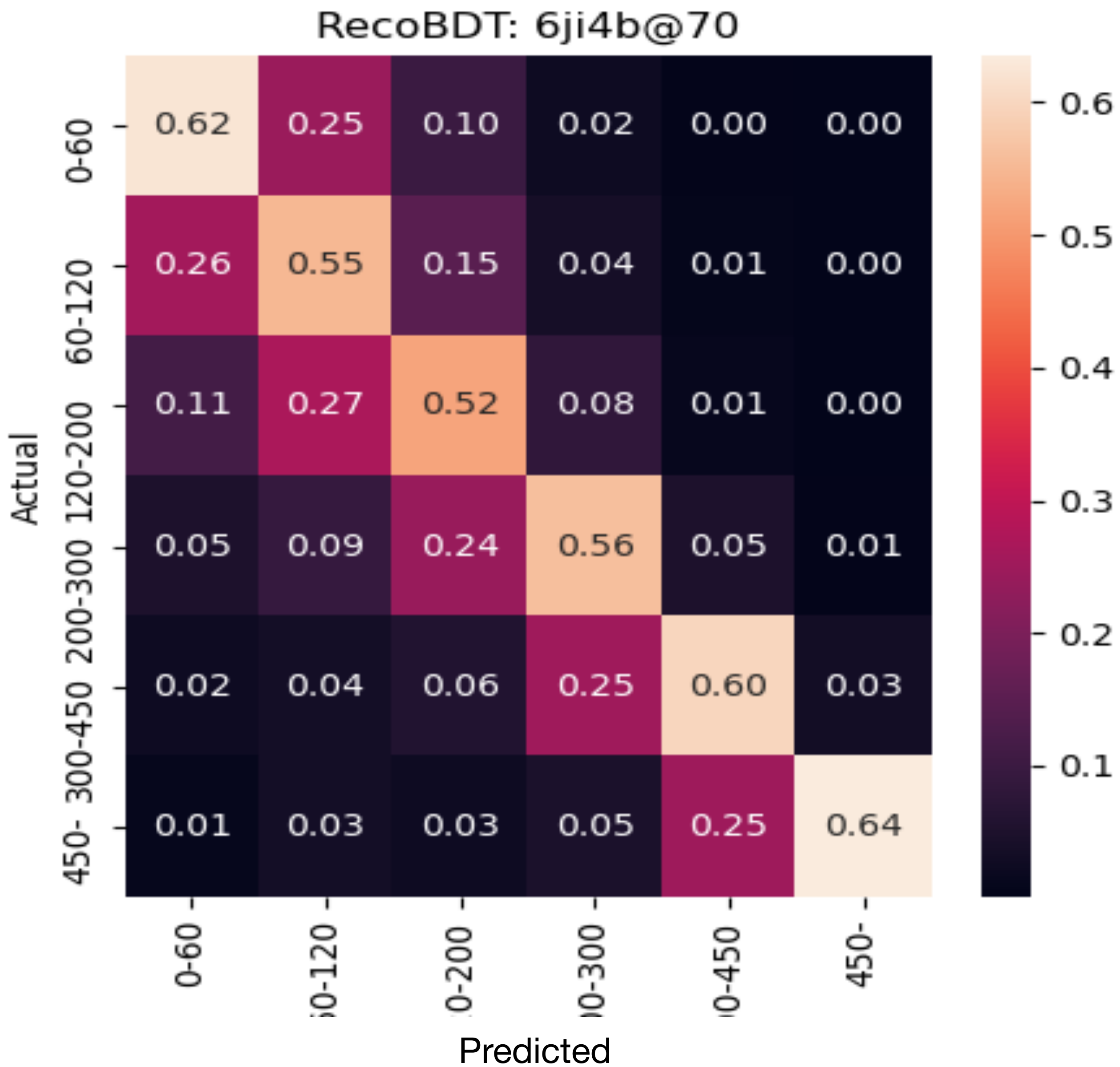
- Both the **multiple jet combinations** as well as the **correlation between features** are considered as inputs to the machine learning model
- Trained models are tested for events with ≥ **6jets**, **≥ 4b-jets:**
- The same set of inputs[2] as in RecoBDT were used in the training (reconstruction)

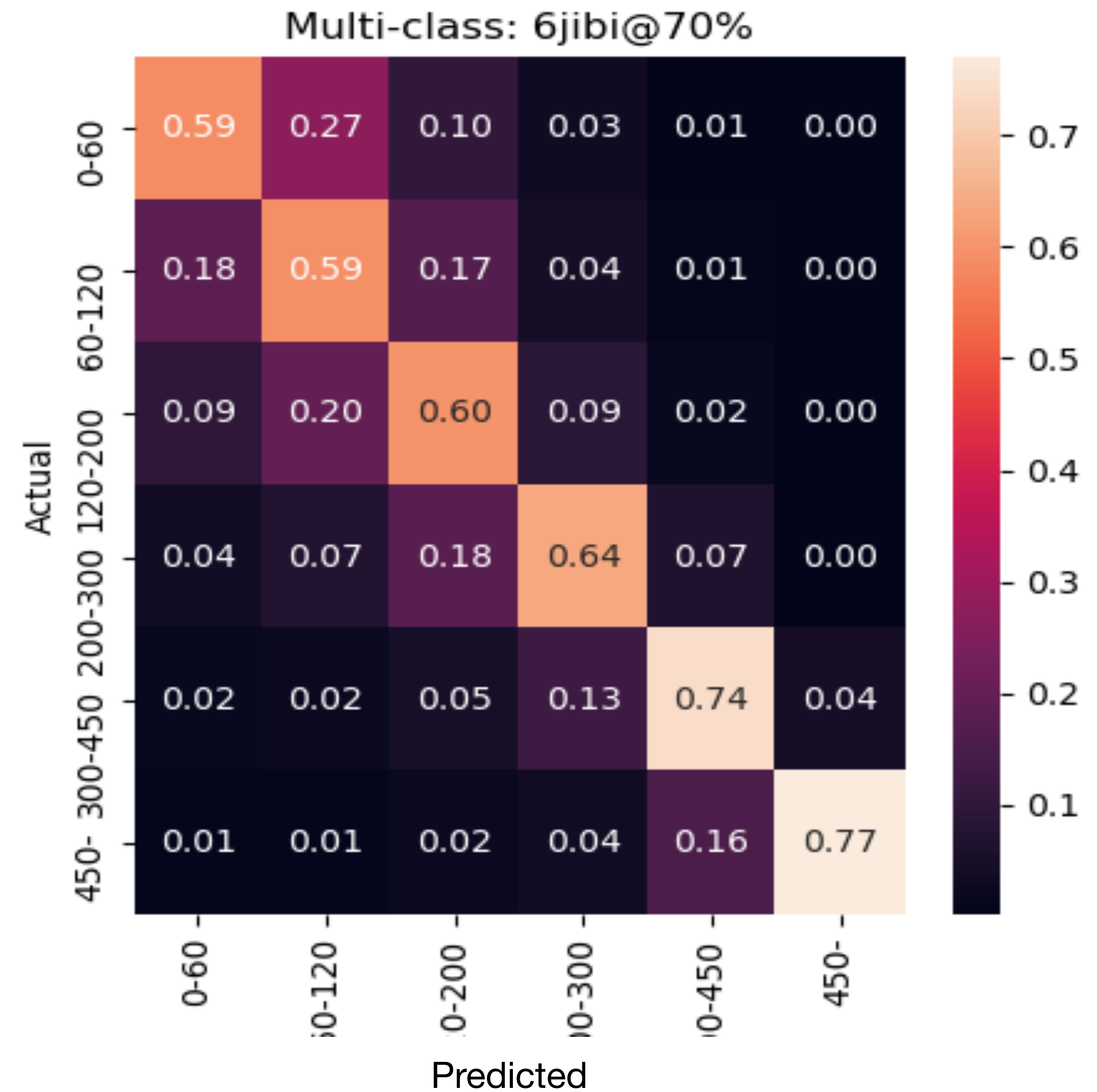
The performance with the new trainings is compared with the performance of recoBDT

[2] List in the back-up

Deep-set reconstruction performance

Training done on multi-classier network using Deep-sets: **STXS bins are taken as the different classes [(0,60 , (60,120), (120,200), (200,300), (300,45), (450,∞)] GeV**

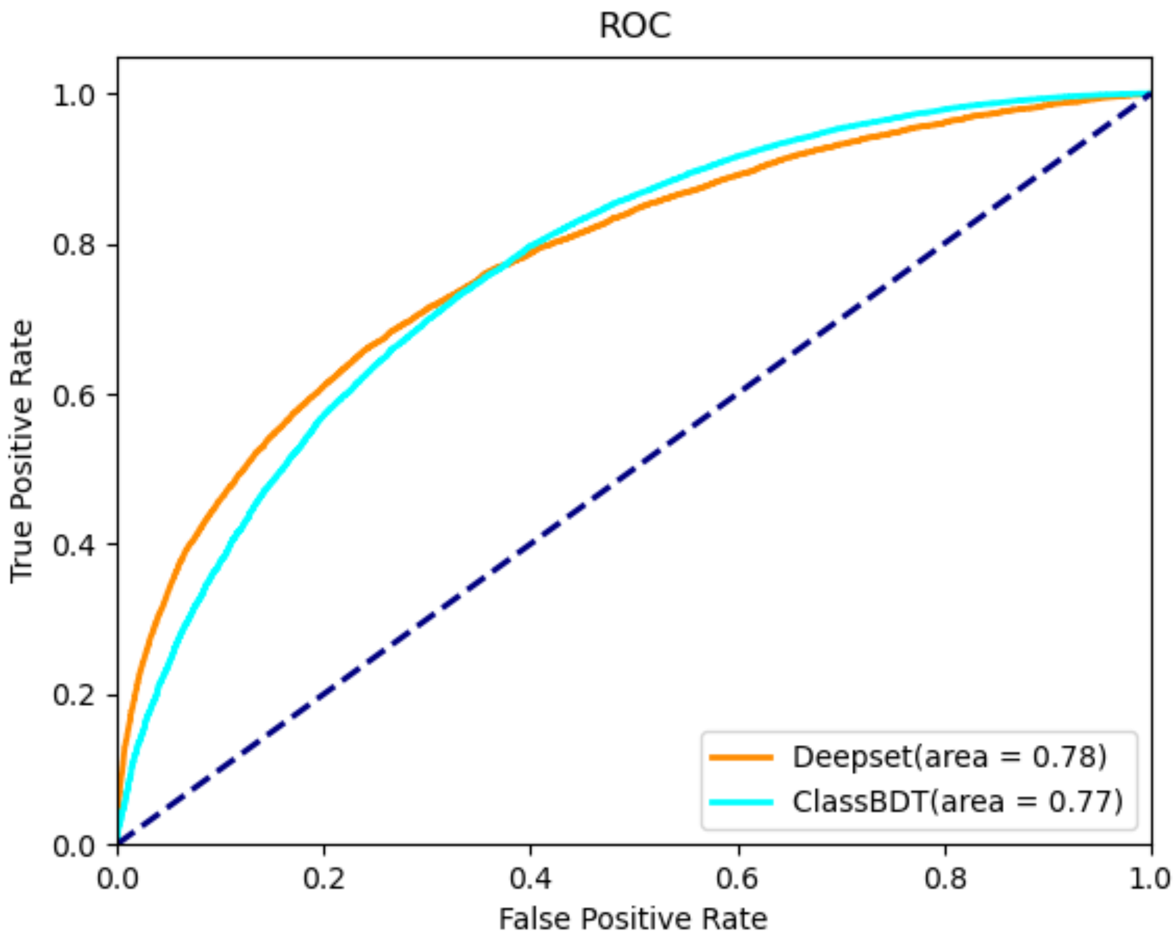


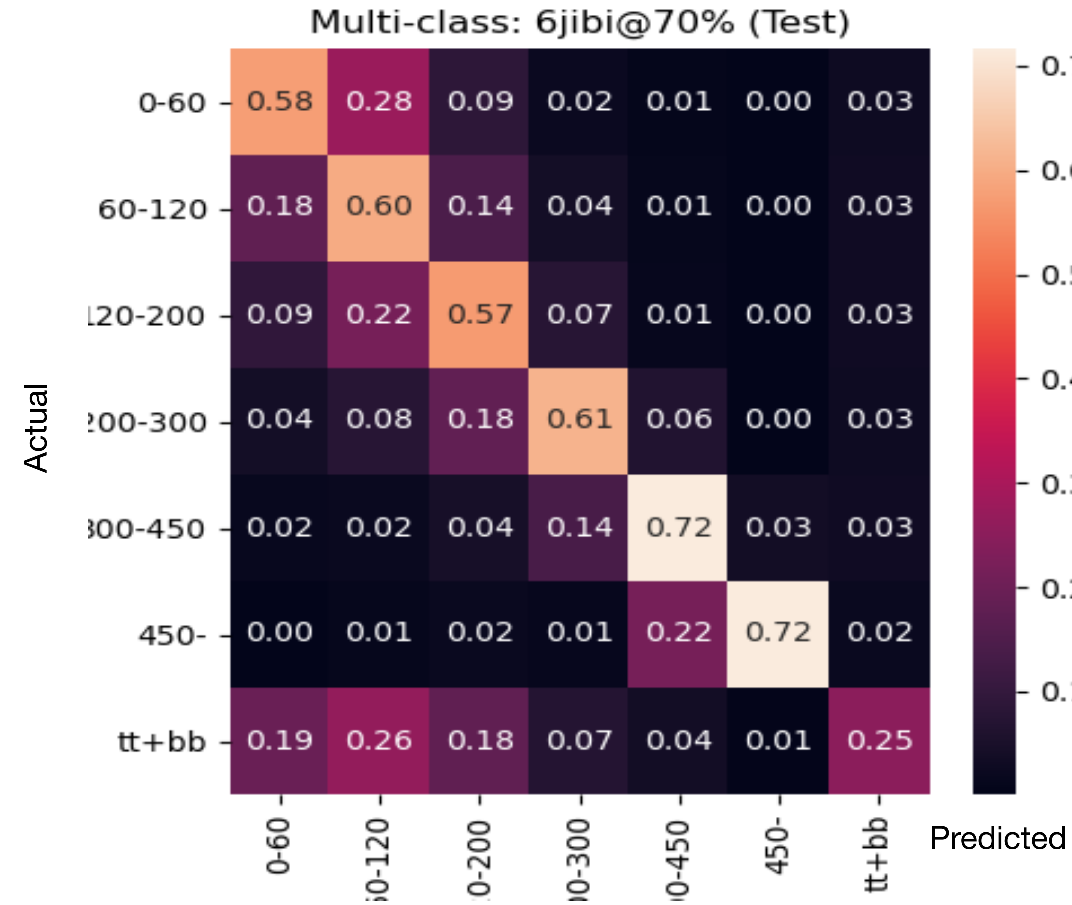


*Deep-set multi-classifier reconstruction performance overall **shows good improvement** when compared to recoBDT*

Deep-set classification performance

- Extended the deep-set multi-classifier to perform both reconstruction and signal vs bkg classification in a single tool
- **Class 0-5 are the ttH STXS classes** and **class 6 is background (here tt+bb used)**
- Adding individual output probabilities of “0-5 classes” will be the final probability being signal (ttH) and “class 6” probability for background





- *Reconstruction performance not affected by extended architecture to include classification of signal vs background*
- *This is the preliminary result for classification performance to establish a reasonable baseline.*

Summary

Part-I

- b-tagging performance in release 21.9 in line with performance in release 20.20
- New IP3D categories relying on pT-based categorisation for |η|<2 + detailed hit content for |η|>2 are available and expected to improve performance
- SV1 performance studied with ITk Material rejection implementation which showed slight improvement in performance
- Overall improvements were highlighted in **ITk PUB note**

Part-II

- Focused on retraining the Multi-Variate Analysis discriminant using Particle Flow + DL1r b-tagging jets
- Slightly better performance for recoBDT and similar performance for classBDT
- Efforts ongoing to implement DNNs targeting to perform the reconstruction and multi-background classification in a single step
- Reconstruction performance with Deep-set multi classifier shows better performance than RecoBDT
- DNN architecture extended to classify signal vs background

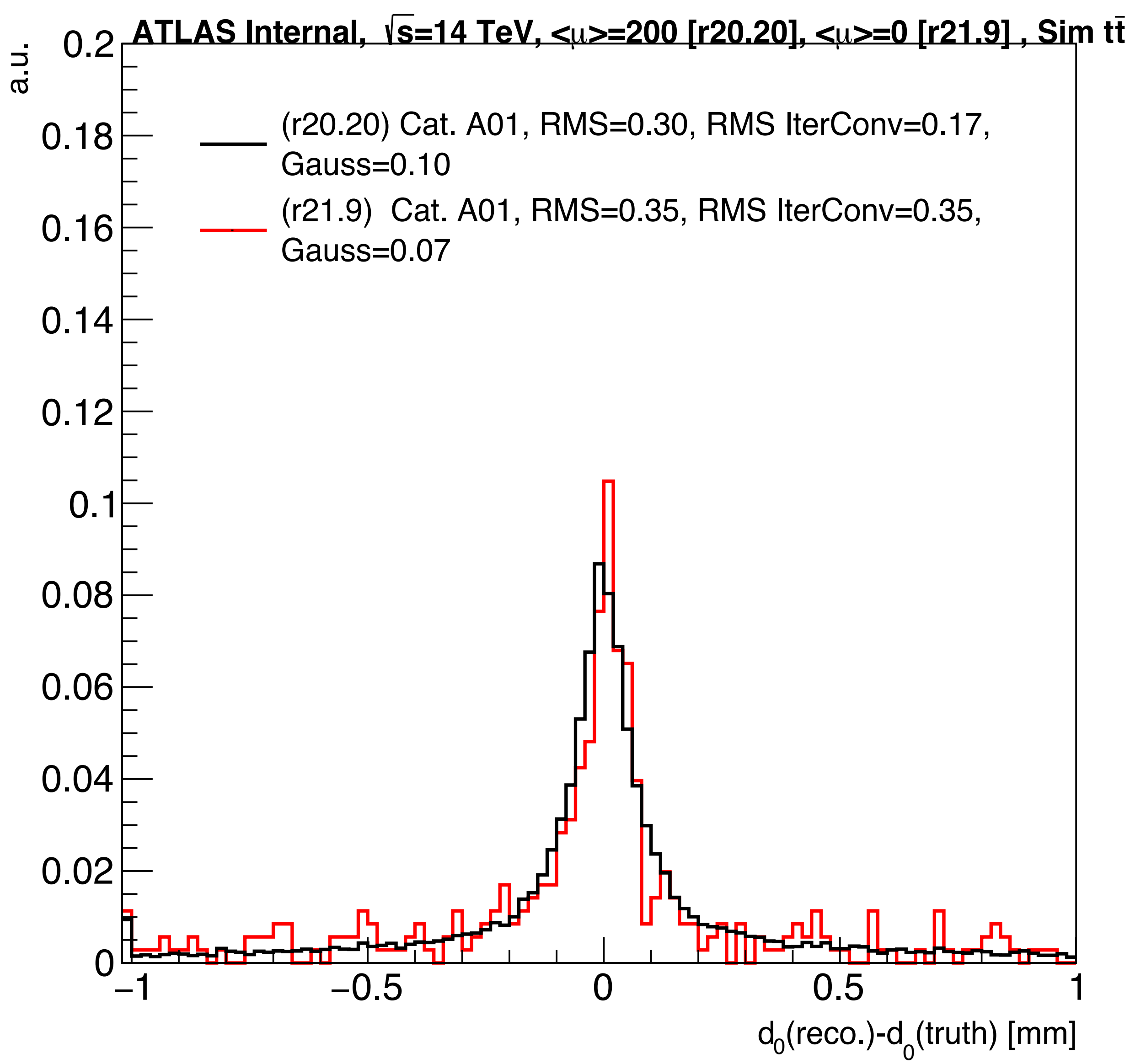
Next →

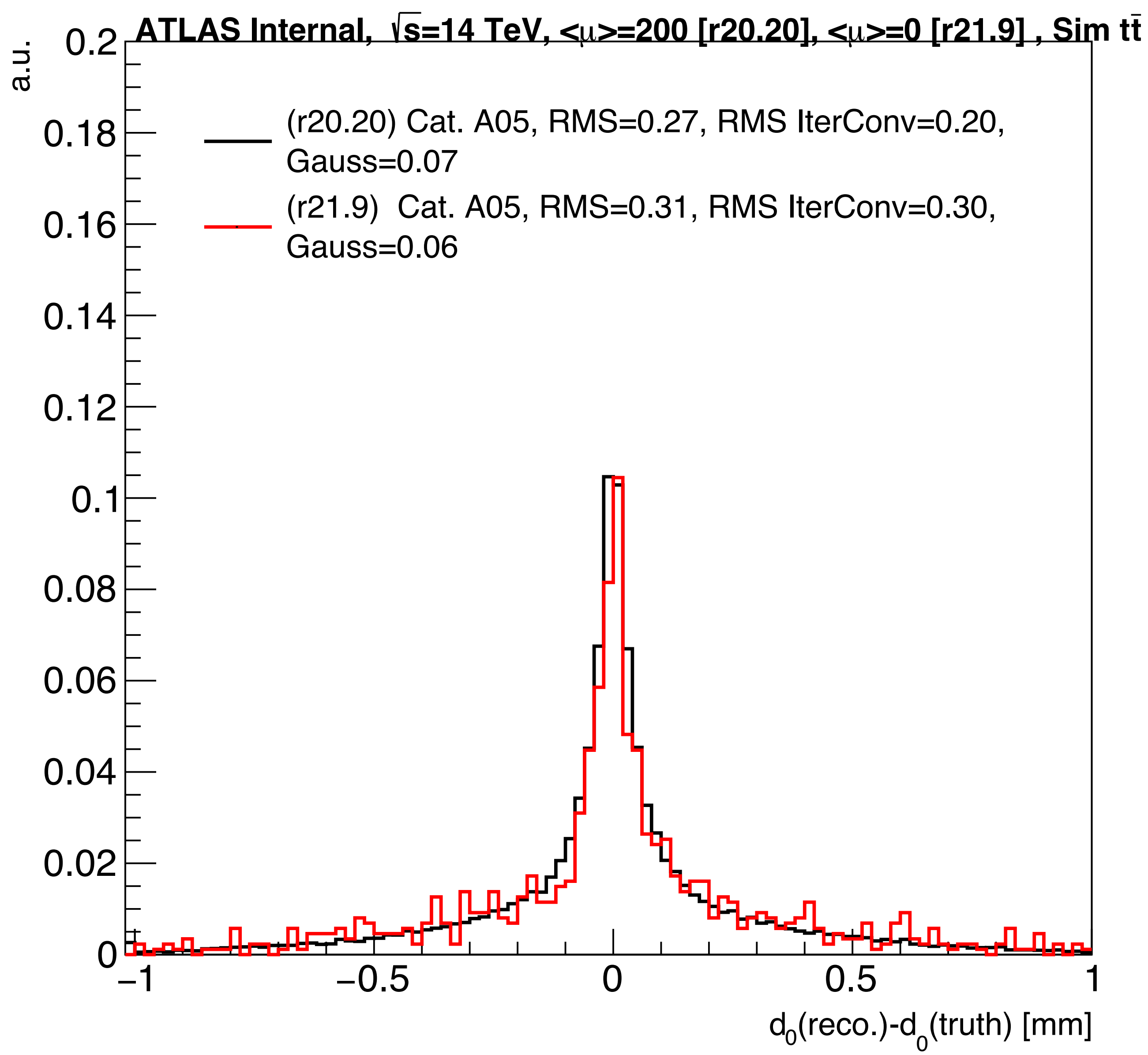
- *Further develop the network to classify signal vs tt background (tt +light vs tt+cc vs tt+b vs tt+2b vs tt+bb)*
- *Study integration in analysis to assess impact on sensitivity and study data/MC agreement*

Thank you for your time!

Back-up

d0 resolution for category A01 and A05





Inputs for reconstruction

- **List of input variables used in the training models:**

Variable	Region ≥ 6j	Region 5j
Topological information from $t\bar{t}$		
t_{lep} mass	✓	✓
t_{had} mass	✓	–
Incomplete t_{had} mass	–	✓
W_{had} mass	✓	–
Mass of W_{had} and b from t_{lep}	✓	✓
Mass of W_{lep} and b from t_{had}	✓	✓
$\Delta R(W_{\mathrm{had}}, b \text{ from } t_{\mathrm{had}})$	✓	✓
$\Delta R(W_{\mathrm{had}}, b \text{ from } t_{\mathrm{lep}})$	✓	✓
$\Delta R(\mathrm{lep}, b \text{ from } t_{\mathrm{lep}})$	✓	✓
$\Delta R(\mathrm{lep}, b \text{ from } t_{\mathrm{had}})$	✓	✓
$\Delta R(b \text{ from } t_{\mathrm{lep}}, b \text{ from } t_{\mathrm{had}})$	✓	✓
$\Delta R(q_1 \text{ from } W_{\mathrm{had}}, q_2 \text{ from } W_{\mathrm{had}})$	✓	–
$\Delta R(b \text{ from } t_{\mathrm{had}}, q_1 \text{ from } W_{\mathrm{had}})$	✓	–
$\Delta R(b \text{ from } t_{\mathrm{had}}, q_2 \text{ from } W_{\mathrm{had}})$	✓	–
min. $\Delta R(b \text{ from } t_{\mathrm{had}}, q \text{ from } W_{\mathrm{had}})$	✓	–
min. $\Delta R(b \text{ from } t_{\mathrm{had}}, q \text{ from } W_{\mathrm{had}}) - \Delta R(\mathrm{lep}, b \text{ from } t_{\mathrm{lep}})$	✓	✓
Topological information from Higgs boson candidate		
Higgs candidate mass	✓	✓
Mass of Higgs candidate and q_1 from W_{had}	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b_2 \text{ from Higgs candidate})$	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate, lep})$	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b \text{ from } t_{\mathrm{lep}})$	–	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b \text{ from } t_{\mathrm{had}})$	–	✓