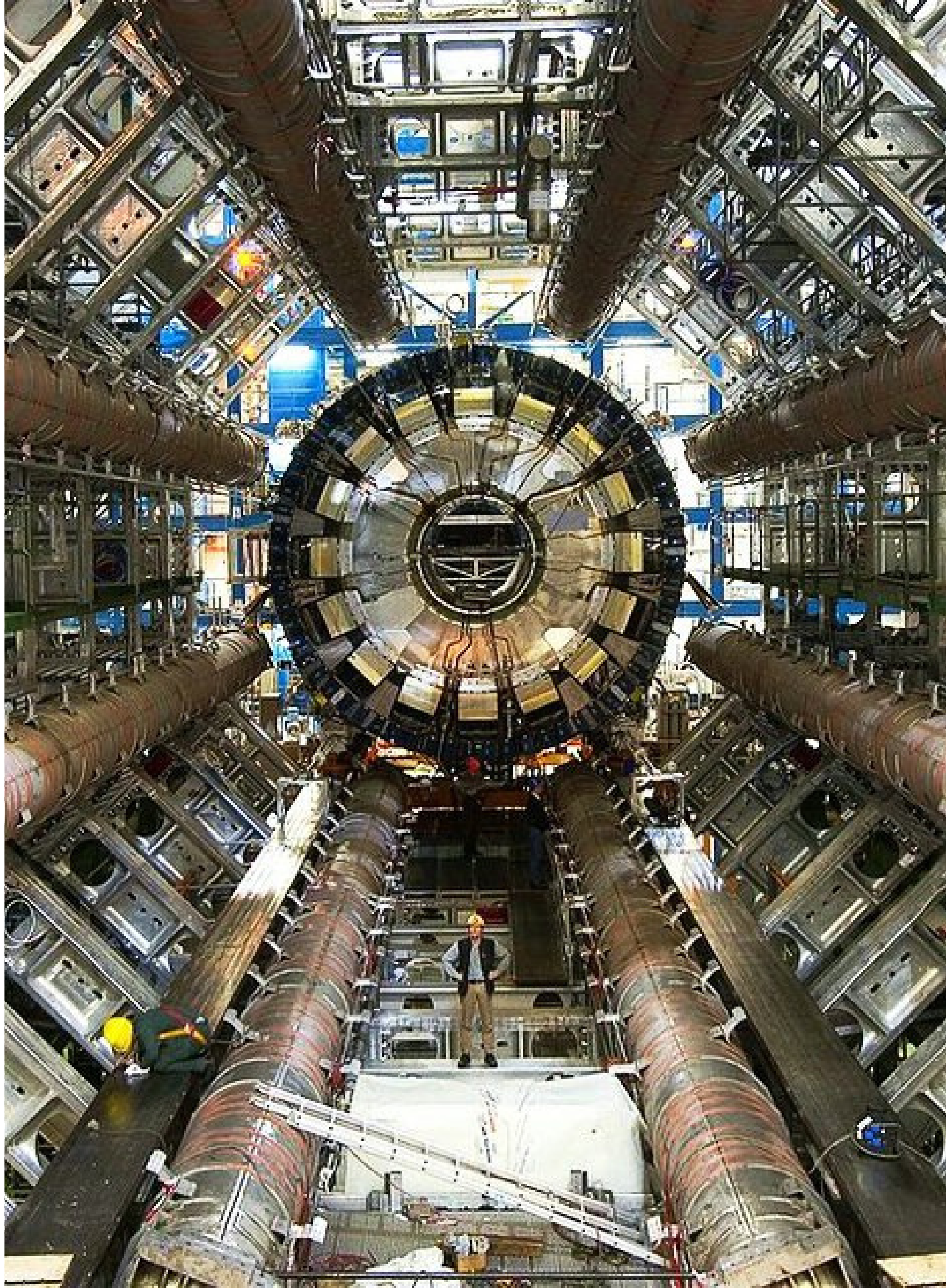
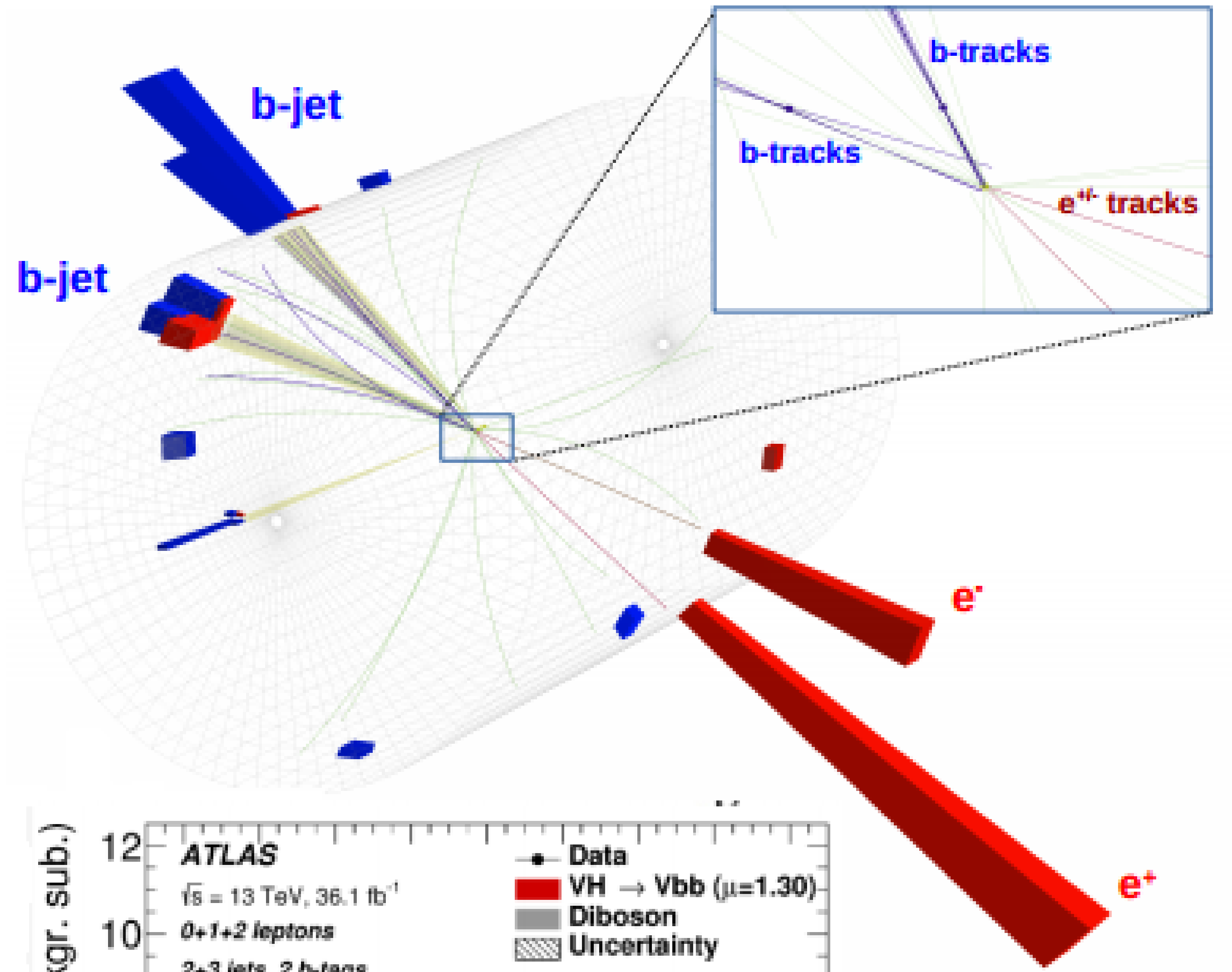
B-TAGGING: A REVIEW

Based on CMS and ATLAS Collaboration papers from Run 1 and 2 at LHC

Speaker: G.Vladlen

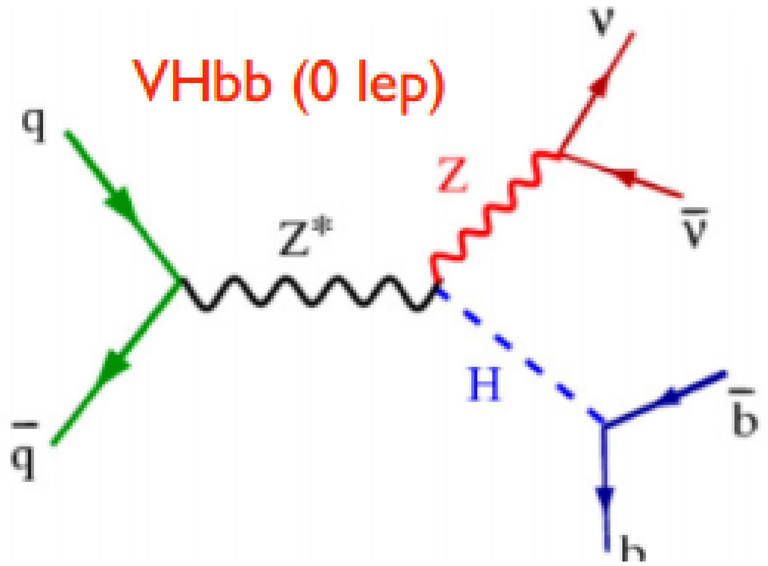


WHY B-TAGGING?

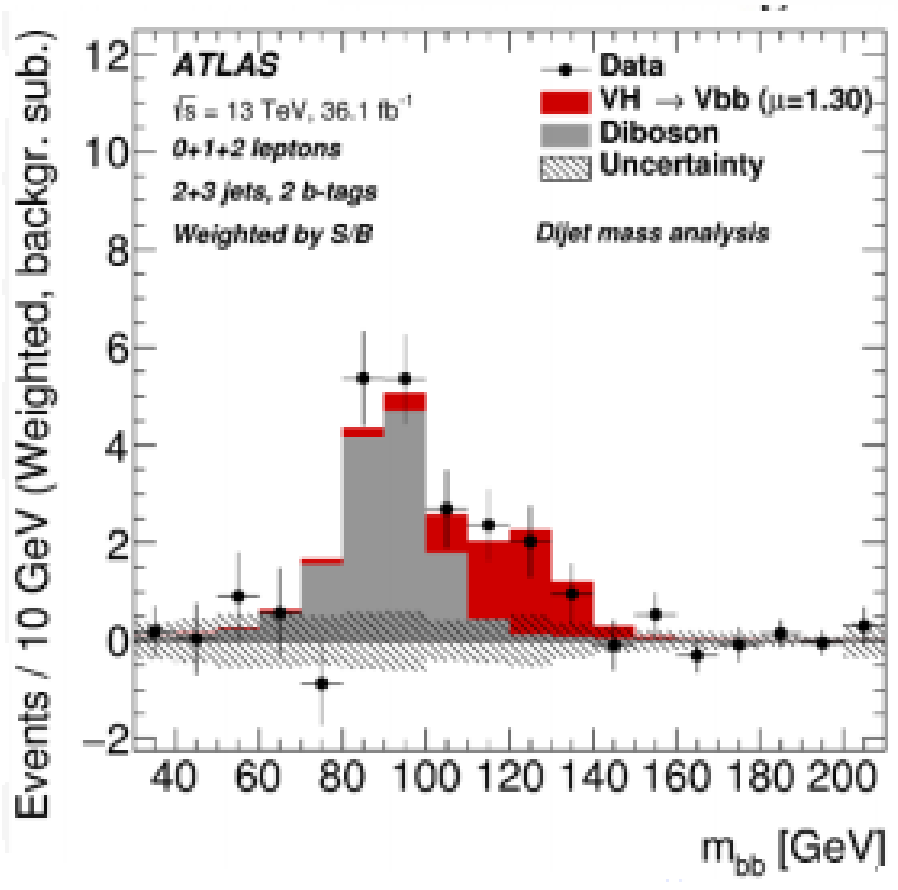


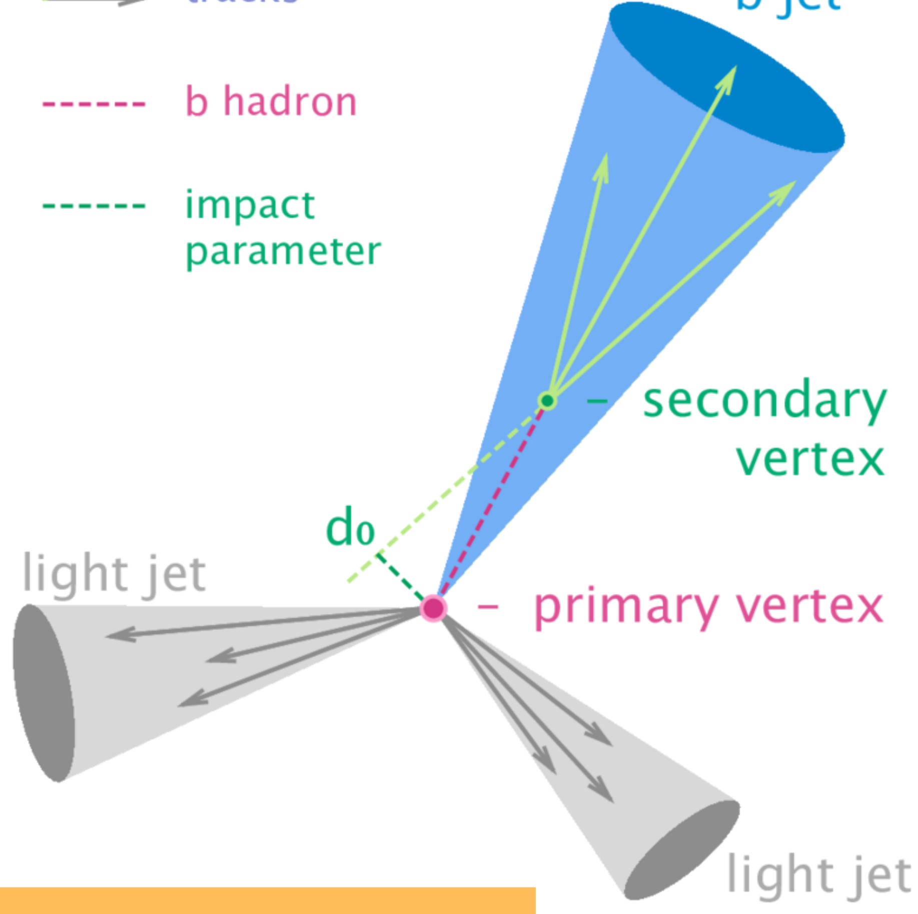
B-tagging is an essential tool to be exploited to study physics processes with b-jets in their final state:

- SM Higgs sectors **(H→bb, HH→bbbb,...)**
- Top physics **(t→Wb)**
- BSM searches **(X → bY)**
- Also used as veto for many backgrounds **(H→WW)**



H->bb (cmb)	Exp.	Obs.	μ
ATLAS	5.5	5.4	1.01 ± 0.20
CMS	5.6	5.5	1.04 ± 0.20

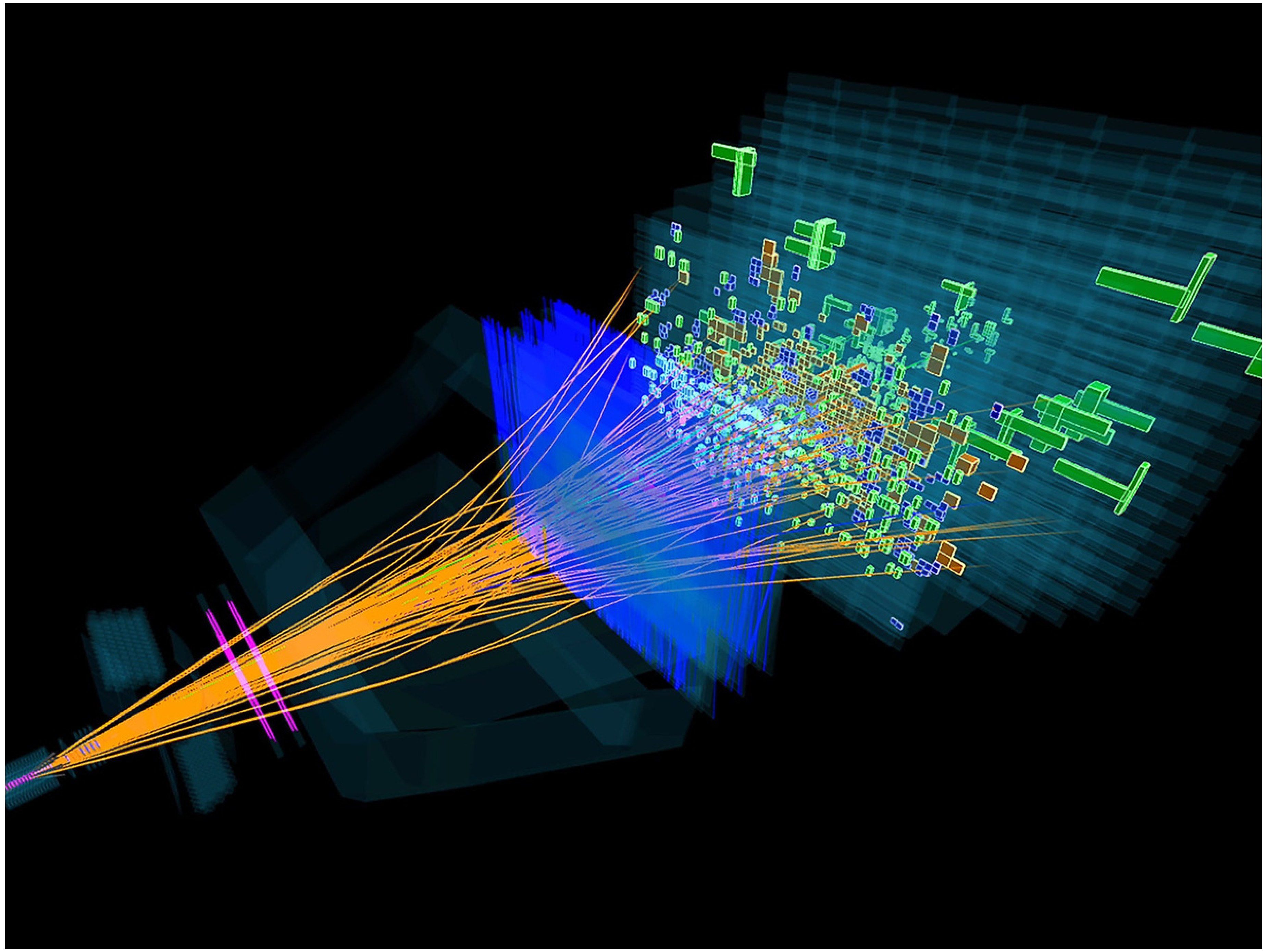




B-TAGGING: A REVIEW

Based on CMS and ATLAS Collaboration papers from Run 1 and 2 at LHC

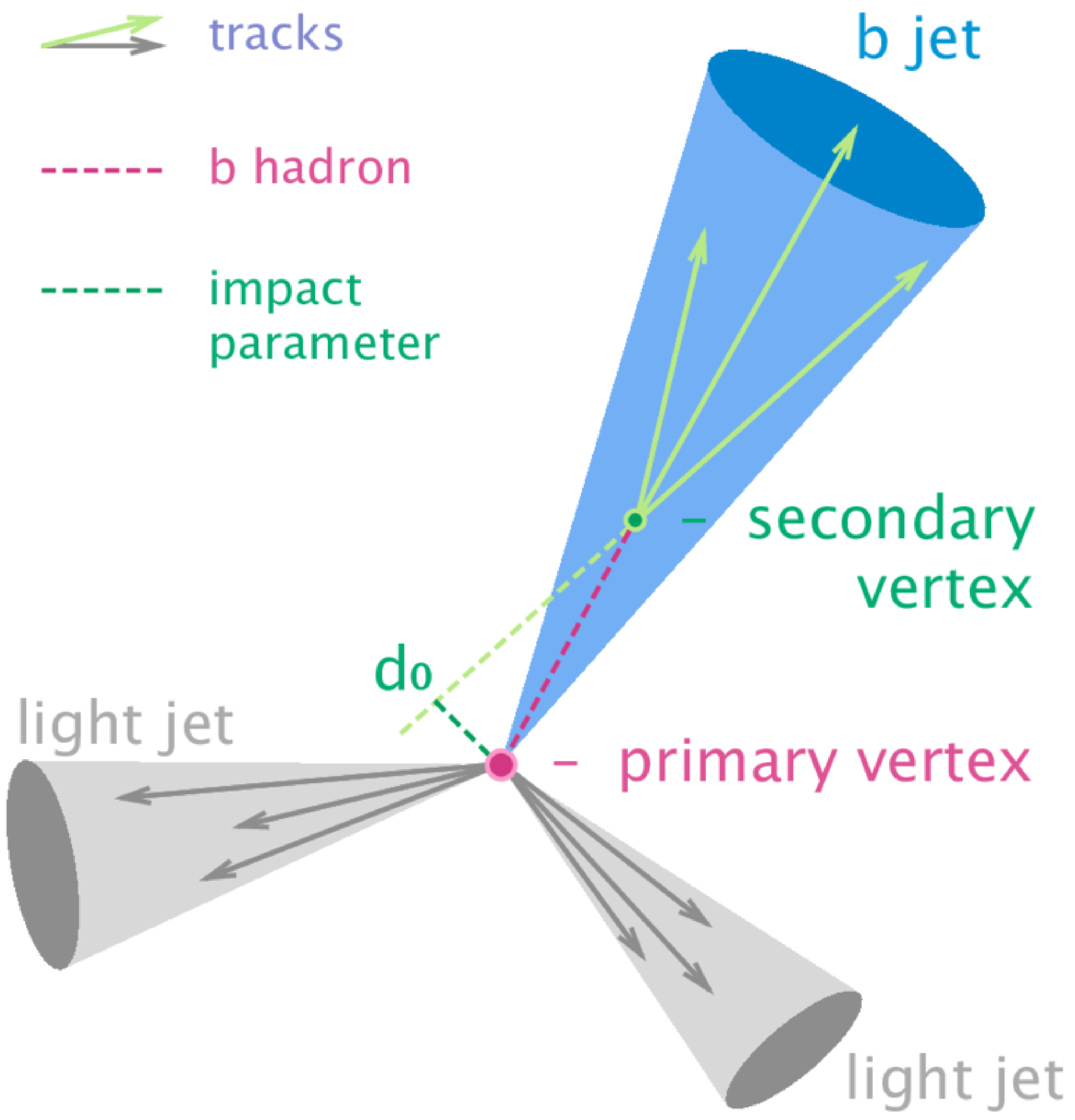
- Jet flavor tagging
- Identification of jet originated from bottom quarks (hadronization)



PROPERTIES?

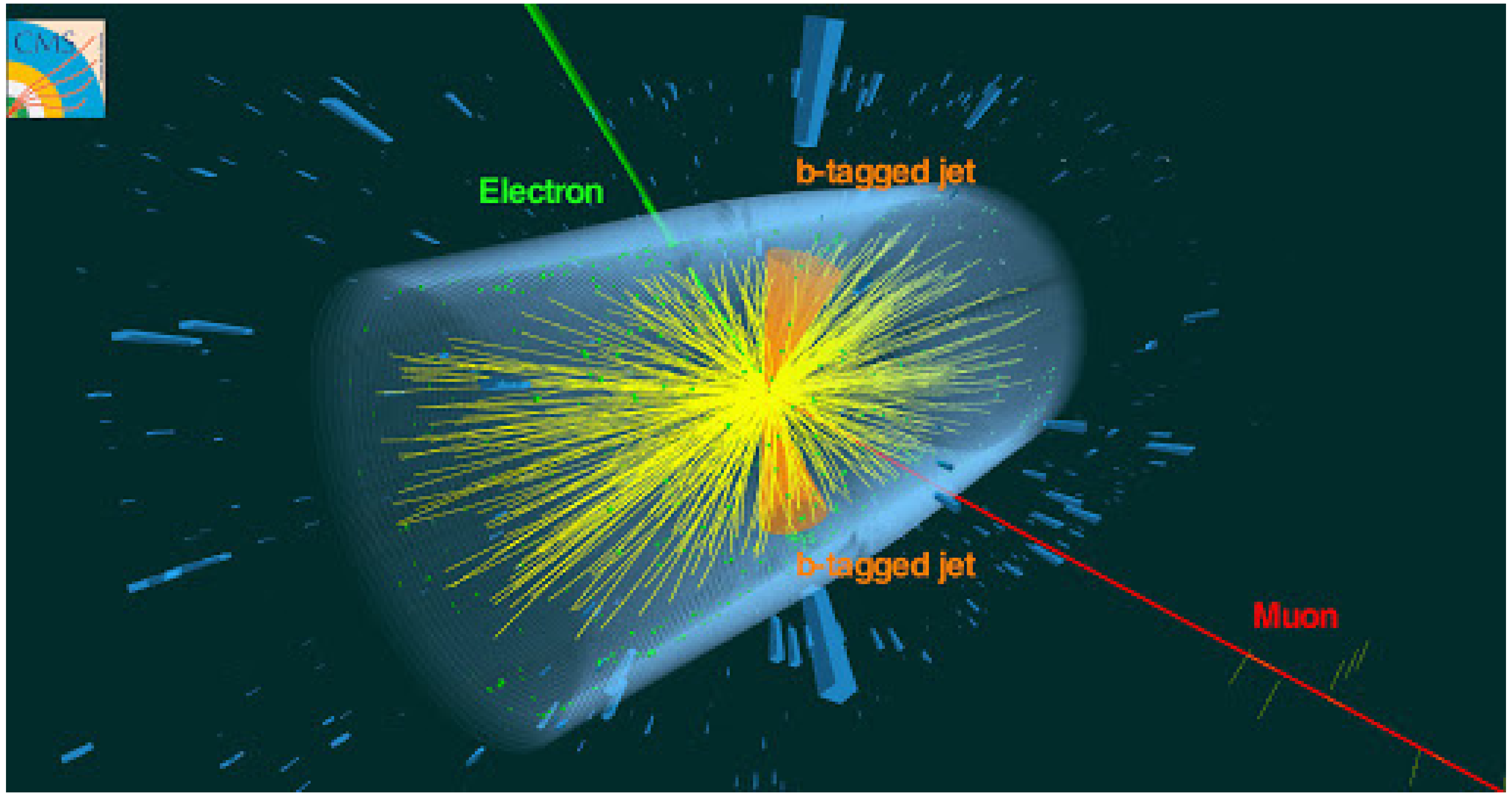
- b-quarks fragment into B-hadrons
- **Displaced vertex** (secondary) from primary vertex due to high life-time (**~1.5ps**)
- High mass of heavy hadrons (**few GeV**) also leads to decay products with a larger transverse momentum.
- presence of direct and indirect semileptonic decays, **b→μνX** (BR~**12%**), **b→c→μνX** (BR~**10%**) :

✓ Presence of soft leptons in jets



OVERVIEW OF B-TAGGING ALGORITHMS

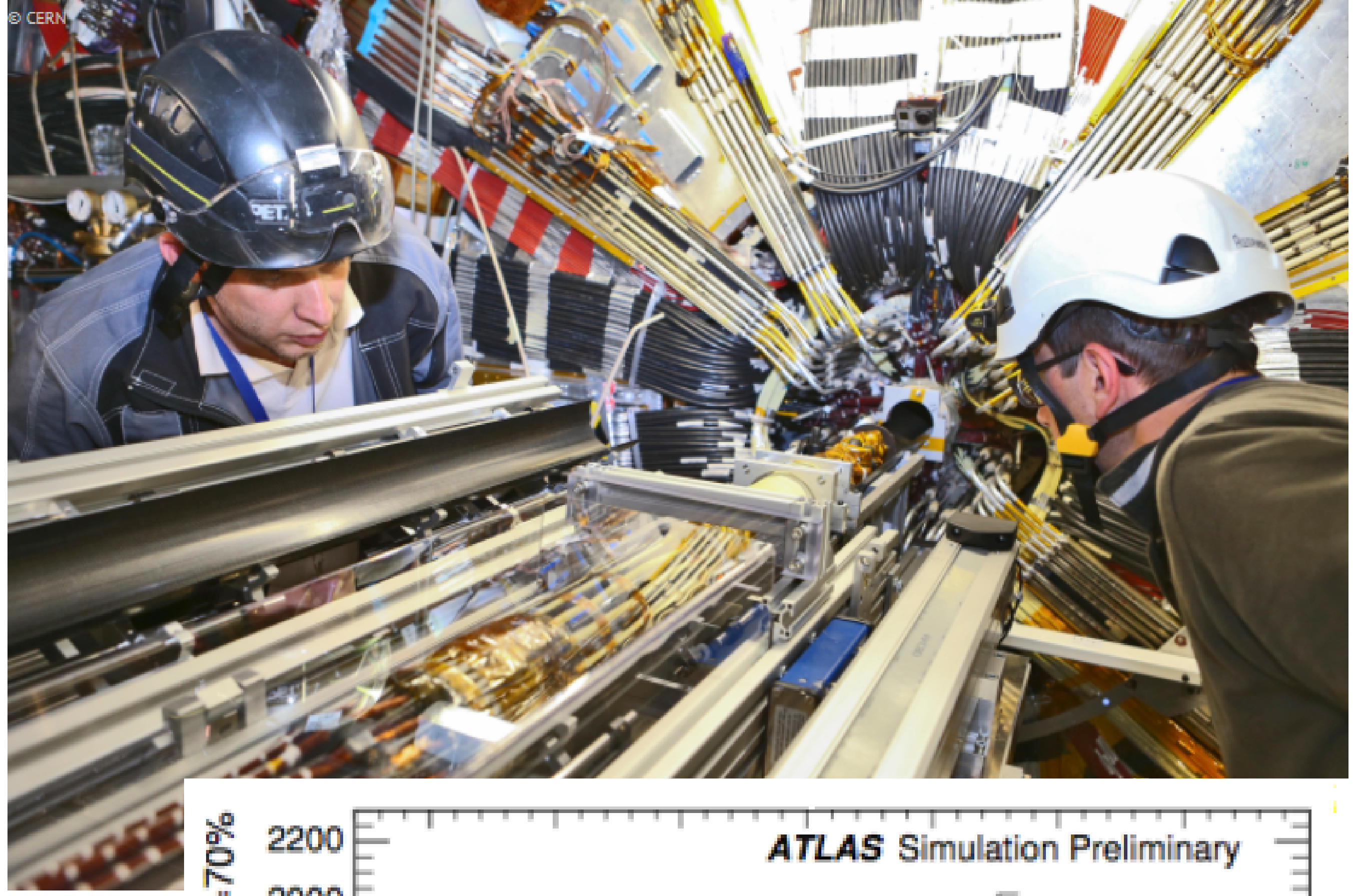
- **Different optimized WPs** in terms of b-efficiency vs the mistag.
- **b-jet efficiency** and purity is an important metric for tagger performance.



Properties	ATLAS	CMS
Tracks with large impact parameters (IP)	IP2D, IP3D, JetProb, Track-Counting	TCHP, TCHE, JetProb, JetBProb
Secondary vertices (SV)	SV0, SV1, SV2	SSVHP, SSVHE, IVF
Multiple vertices from $B \to C \to X$ decay chains	Jetfitter	
Soft leptons from semileptonic B decays	SMT, PtRel	Soft Lepton Taggers
Multivariate combinations	MV2MuRnn, MV2Mu, MV2c00, MV2c10, MV2c20	CSV, CSVv2, cMVAv2, DeepCSV

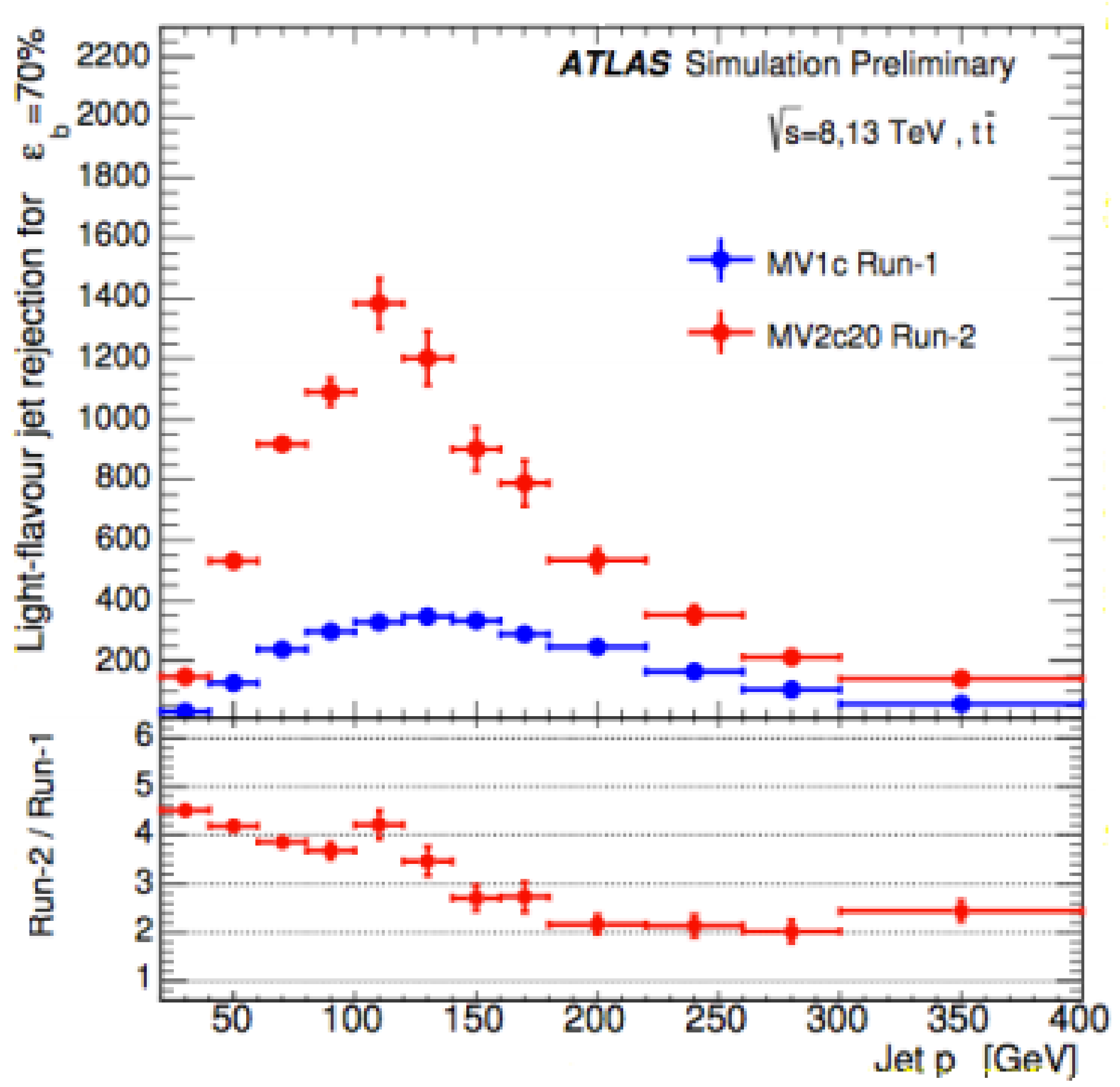
Table 1. Schematic of the algorithms used at Run 2 for different components of the b-tagging performed by the ATLAS and CMS collaborations

HIERARCHY AND ARCHITECTURE OF ALGORITHMS



Low level taggers:

- Account for features of the b-system (**impact parameter** of tracks associated to jets, **SV reconstruction, soft leptons**) and exploit discrimination wrt background jets (**light-flavor and c-jets**).



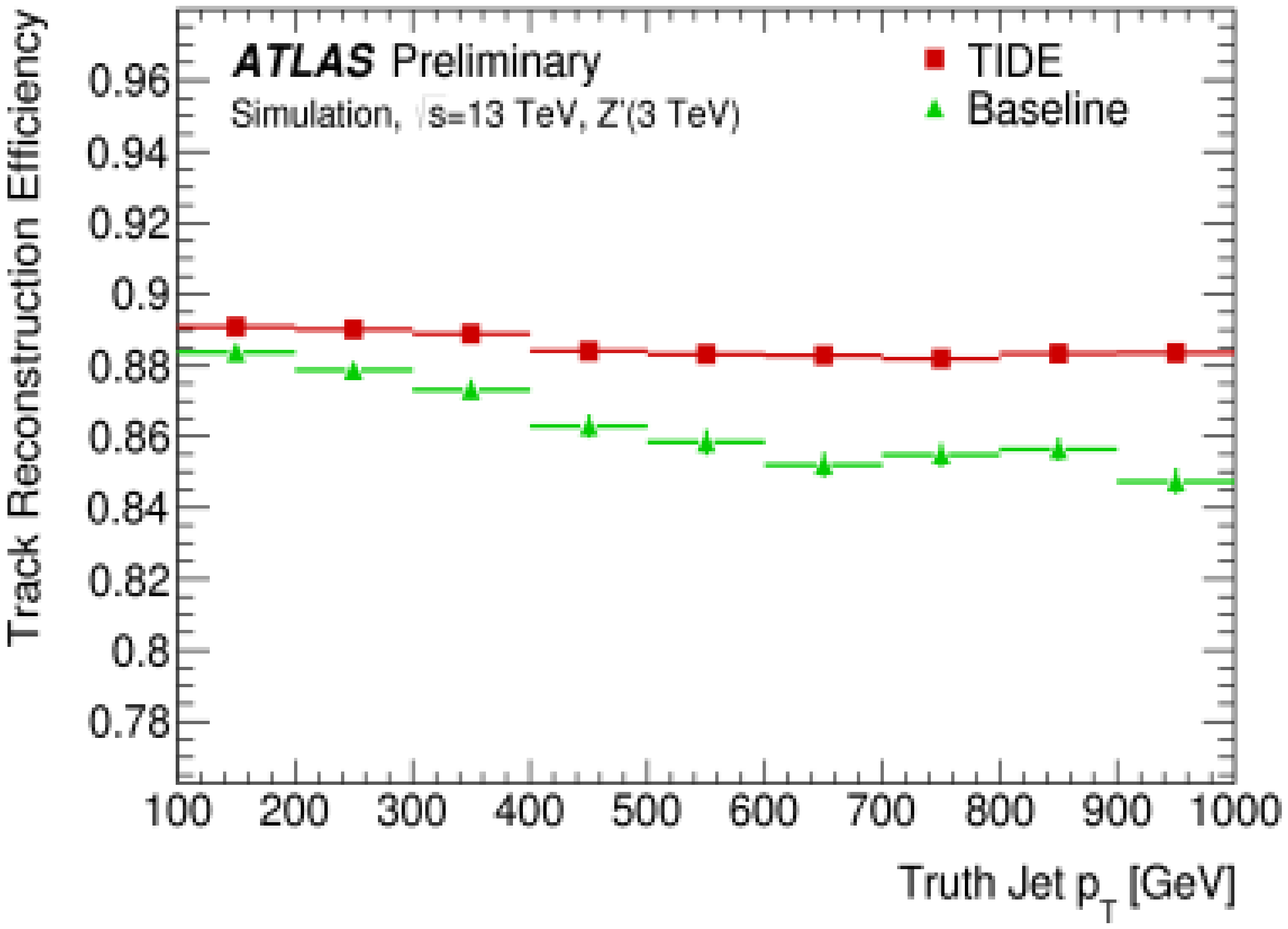
- Used as inputs of **high-level tagging discriminants.**
- Detector response crucial to achieving good discrimination (e.g IBL in ATLAS ensures better **track resolution and robust pattern recognition** for IP2D/ IP3D algorithm - important for IP and SV algorithms).

Tracking in ATLAS and CMS

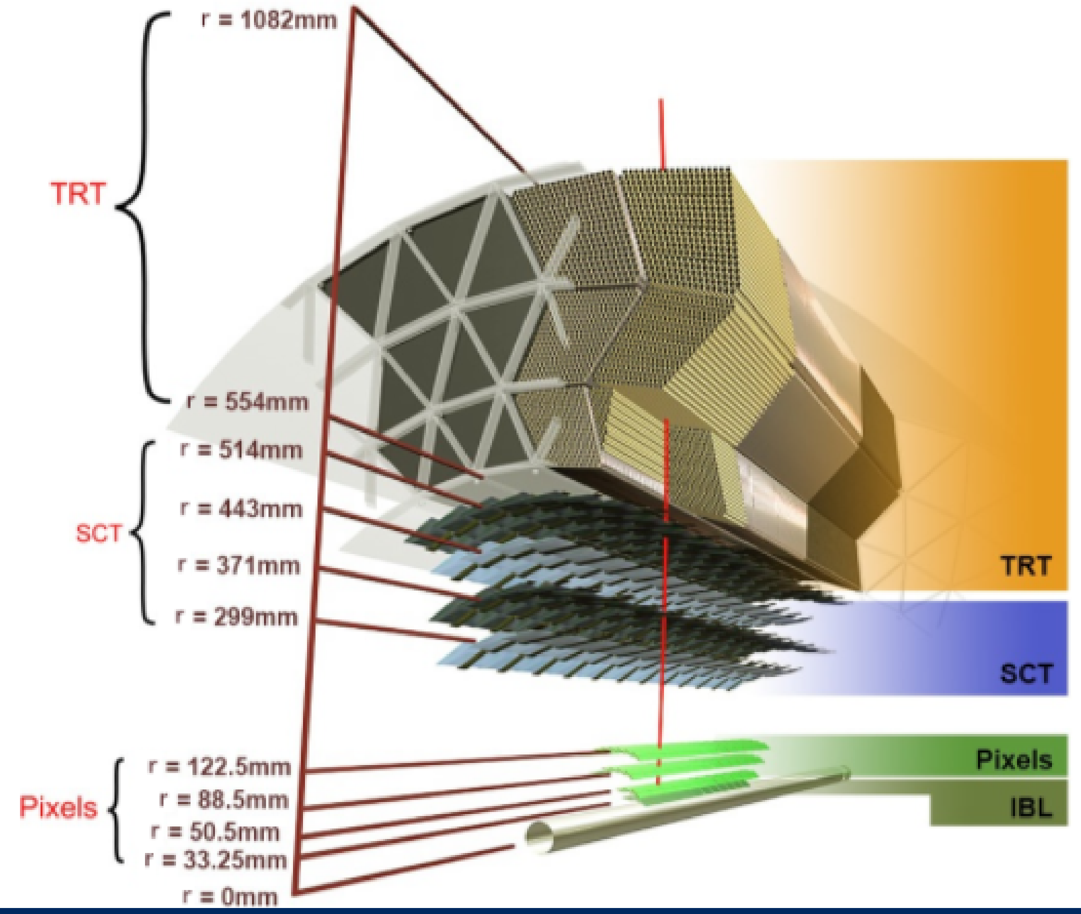
Both in **ATLAS** and **CMS**, iterative algorithms have been developed for **track reconstruction**:

- **Inside-out:** Silicon seed → extrapolate to TRT (primary tracking). **Outside-in:** TRT seed → extrapolate to silicon

- **Iterative combinatorial track finding** (Seed gen+track-fitting+track-selection)



1. Formation of space-points.
2. Track finding seeded by space-points (Combinatorial Kalman filter).
3. Ambiguity solving.
4. TRT extension.



Vertexing in:

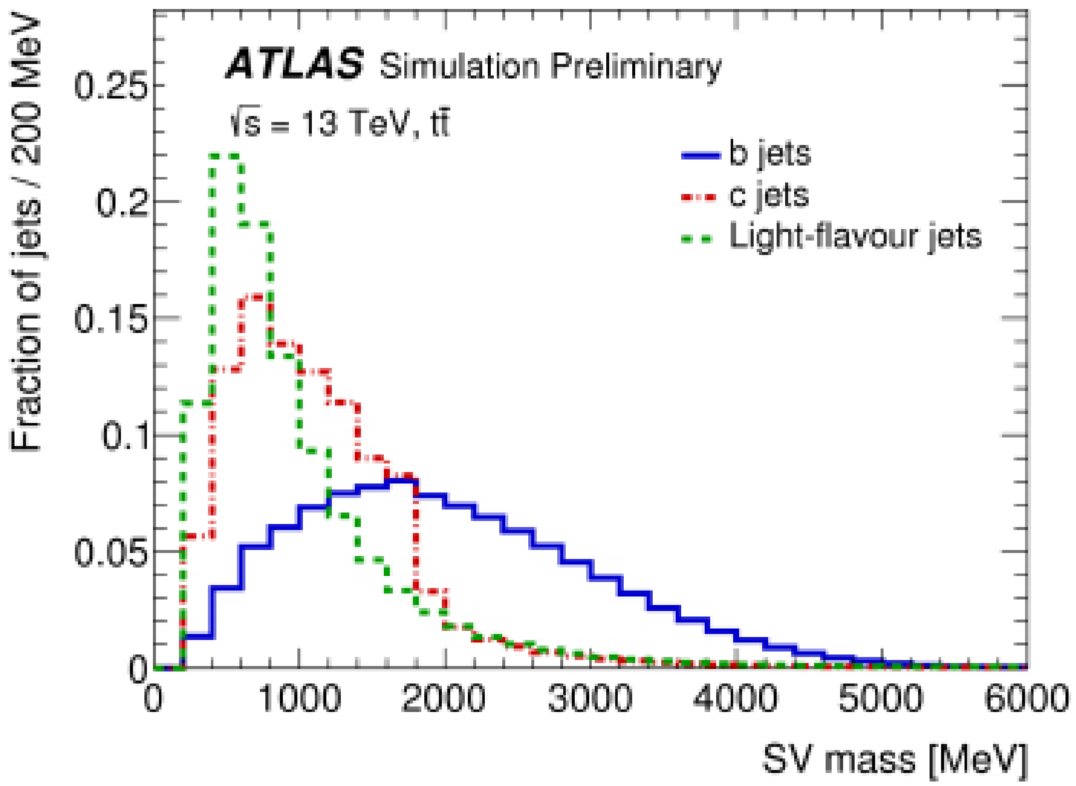
ATLAS

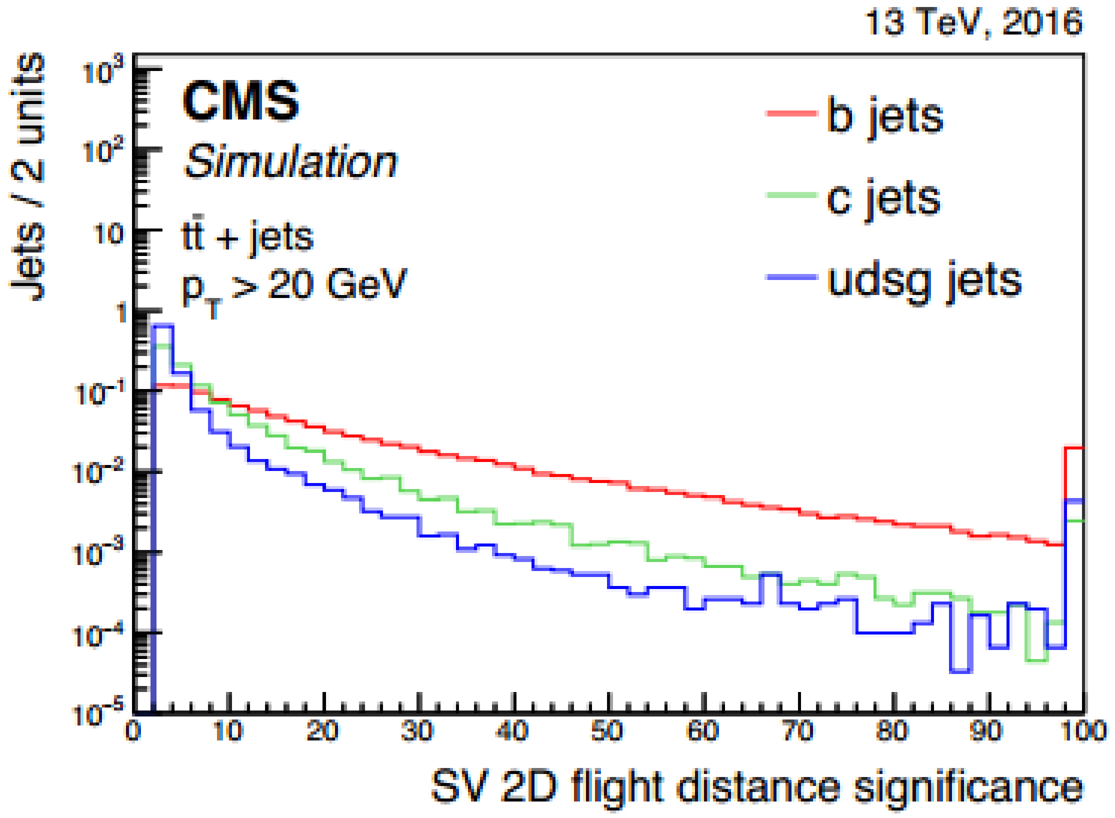
- **SVF**: consider tracks inside a jet cone f(pT), jet axis direction, and PV position.
- **SSVF**: based on SVF, tighter track selection + 3D-IP of a track with respect to the PV.

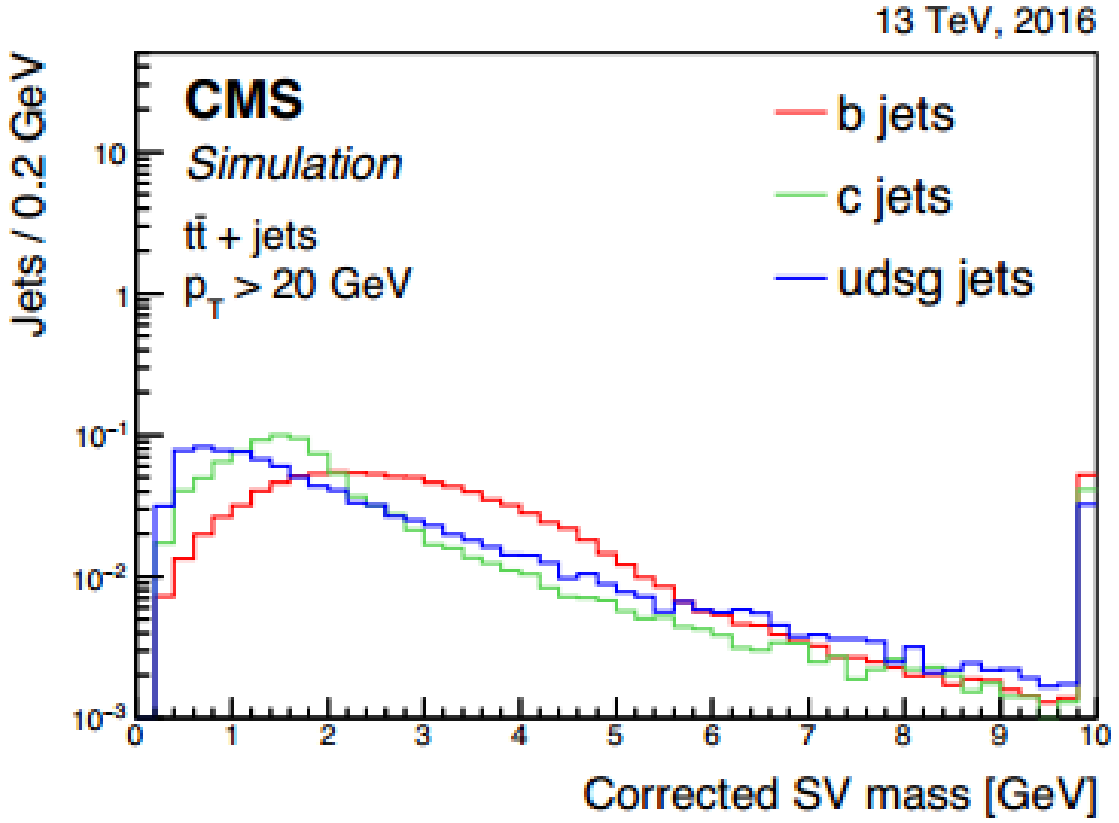
- Iteratively reconstruct a single secondary vertex per jet (b and c too close, or highest track multiplicity)

CMS

- **AVR**: Iteratively fit tracks clustered within a jet (dR<0.3) after passing basic selection.
- **IVR**: Starts from all recotracks with pt > 0.8 GeV: Track clustering + Secondary vertex fitting + track arbitration & cleaning + Secondary vertex refitting.

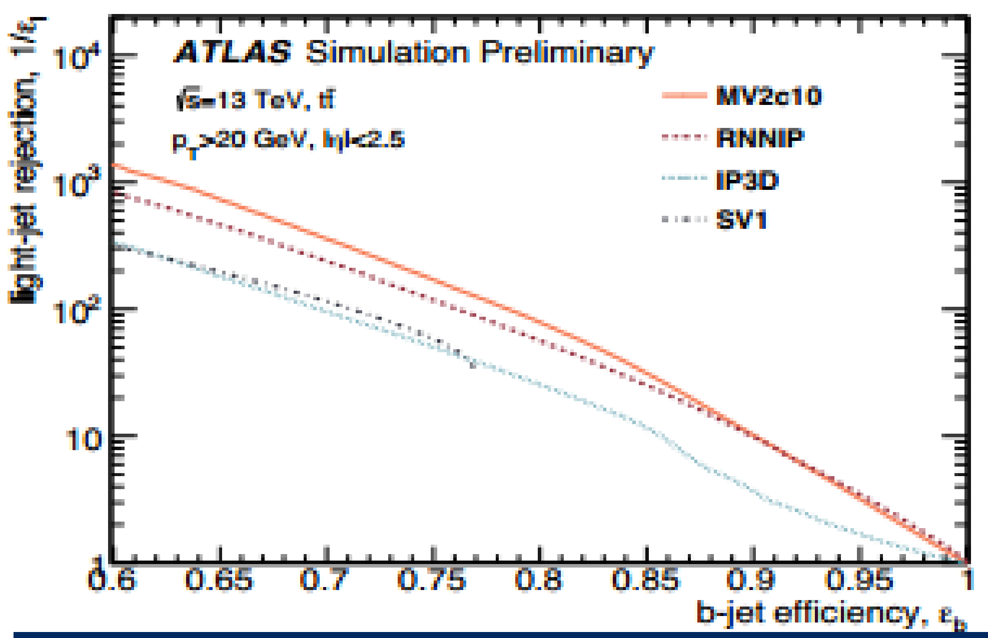


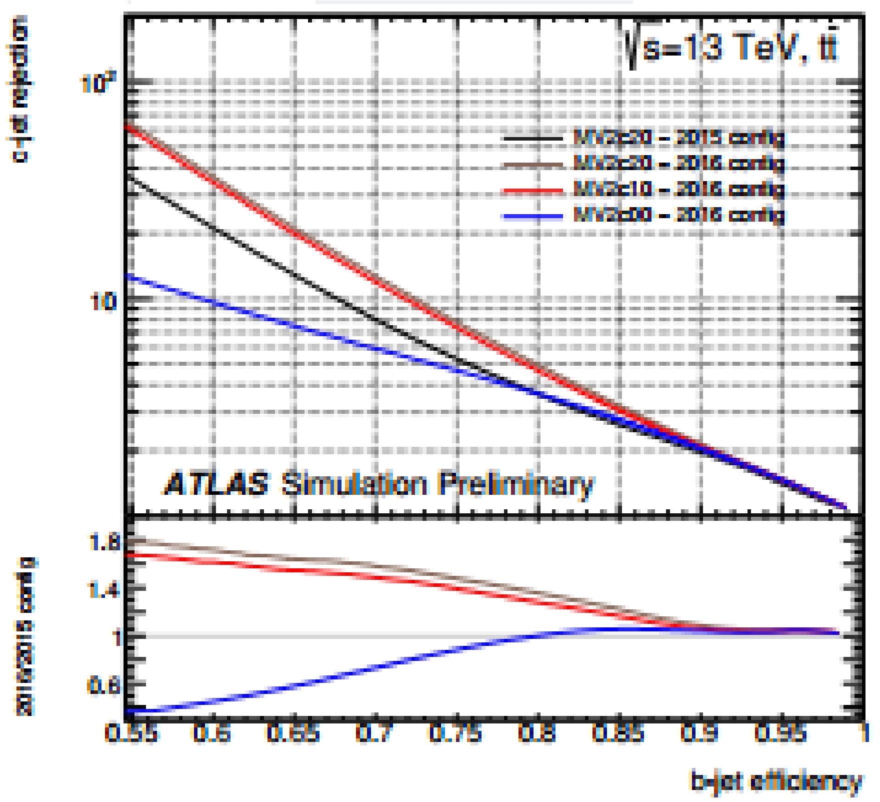


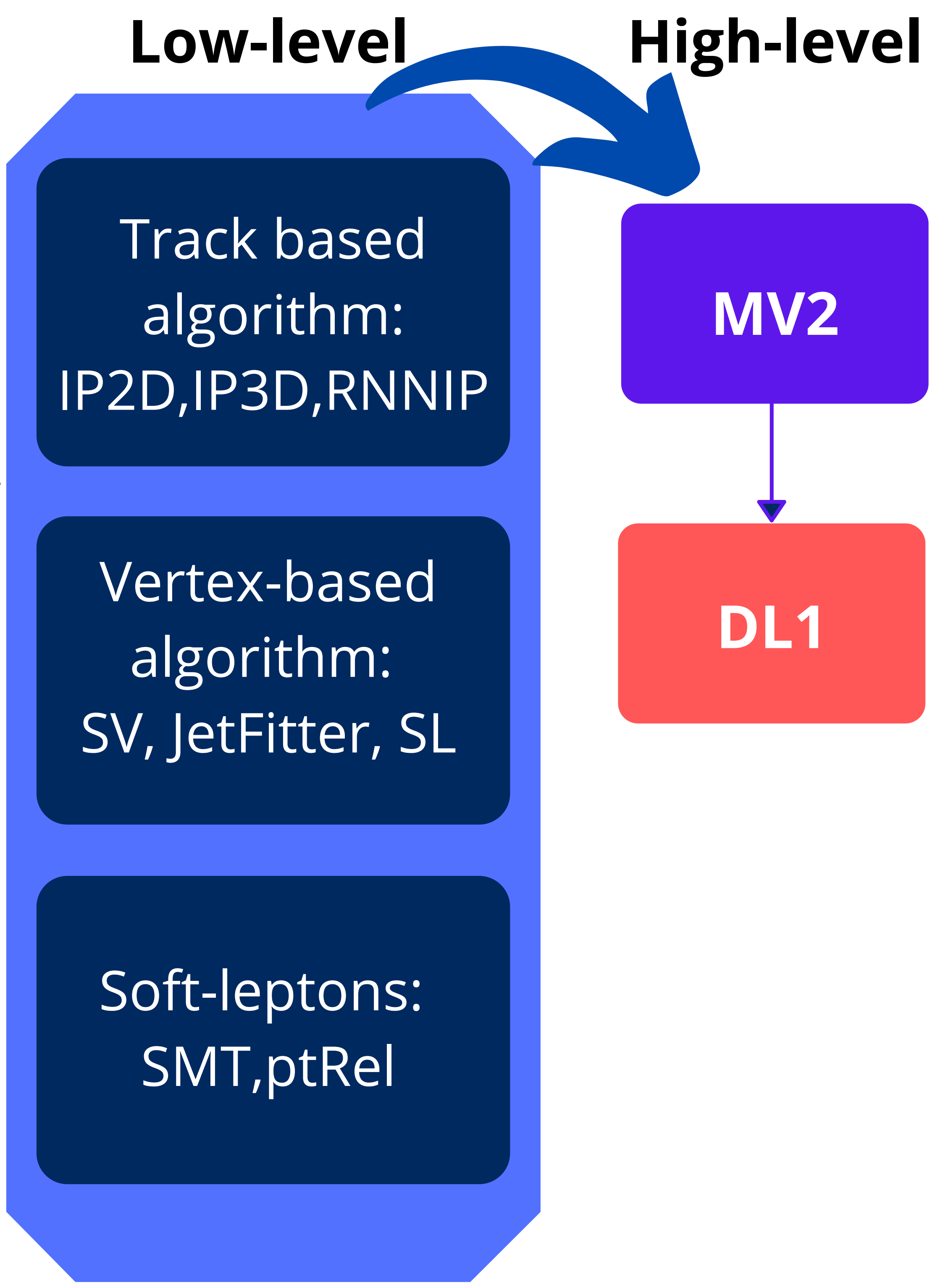


b-tagging strategies in ATLAS

- **IP2D, IP3D:** lifetime-based tagging algorithms.
- **RNNIP:** Recurrent NN tagger: better exploits correlations between tracks associated to jets.
- **SV, JetFitter: SV:** reconstruction of inclusive displaced SV within the jet. **JetFitter:** Reconstruct the decay chain exploiting decay topology.
- **MVA discriminators** (MV2* and MV2RNN): MVA based discriminator family exploiting jet kinematics, **IPs, SVs, and JetFitter** information. MV2* are based on BDT.
  DL1: DNN Multiclassifier.

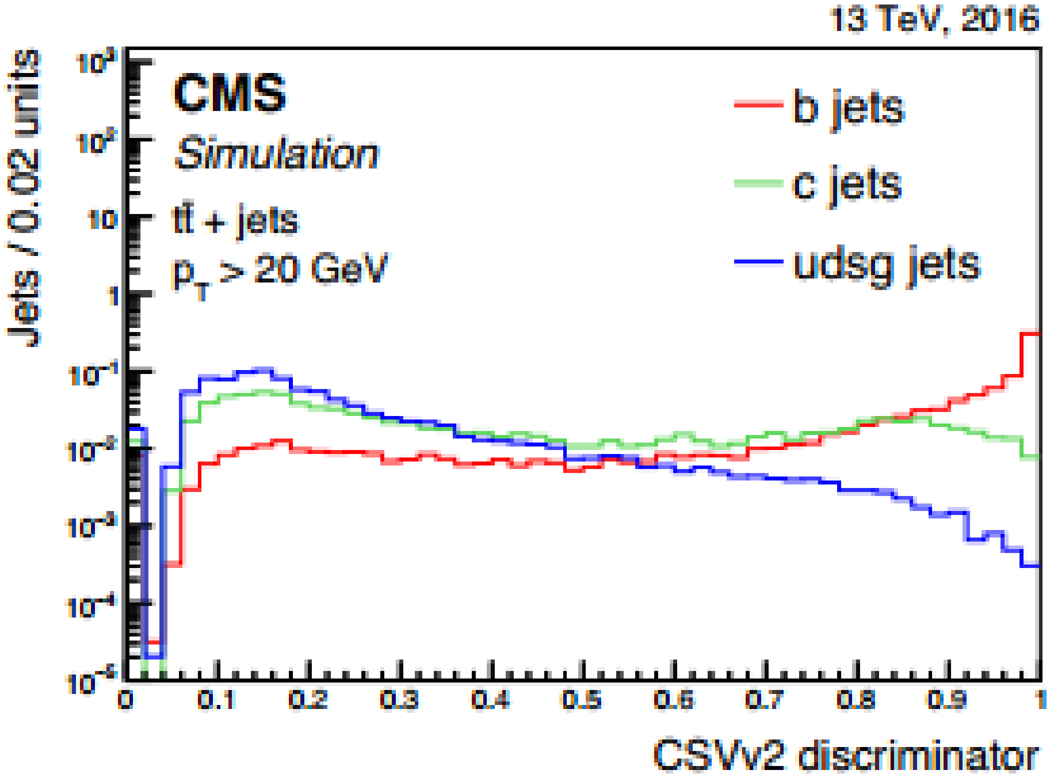


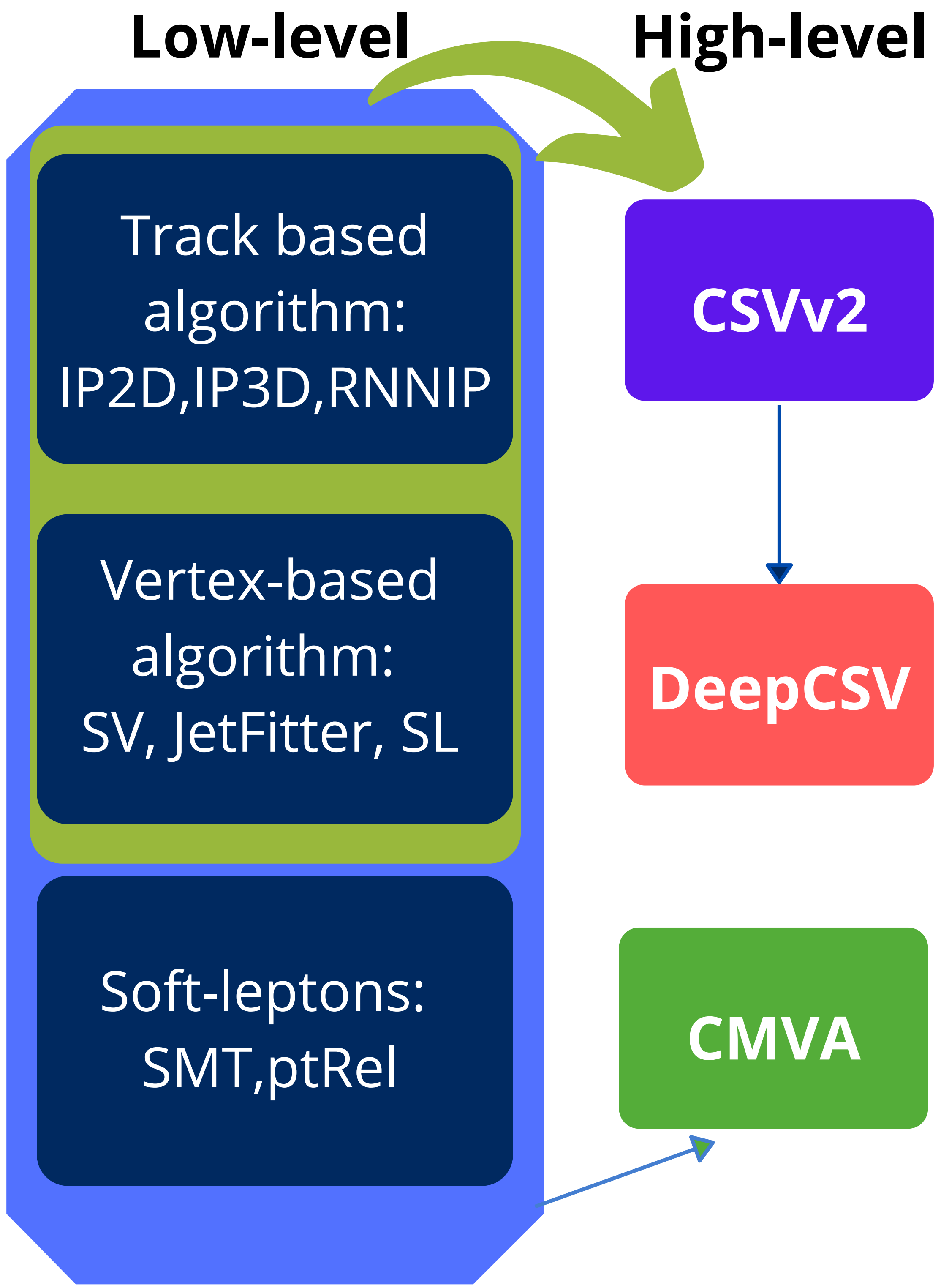




b-tagging strategies in CMS

- **JP and JBP:** Likelihood-based on the track properties (displacement). Returns p(b-jet).
- **CSV, CSVv2:** Combine displaced tracks with secondary vertices in BDTs (CSV) and in multilayer perceptrons (CSVv2).
- **CMVA:** combined multivariate analysis tagger, combines the discriminator values of the low-level tagger.
- **DeepCSV:** DNN Multiclassifier: same inputs as CSVv2 with a simple extension to use more charged particle tracks.
- **DeepFlavour:** DNN Multiclassifier.





Higher hierarchy b-tagging in ATLAS

MV2 tagger

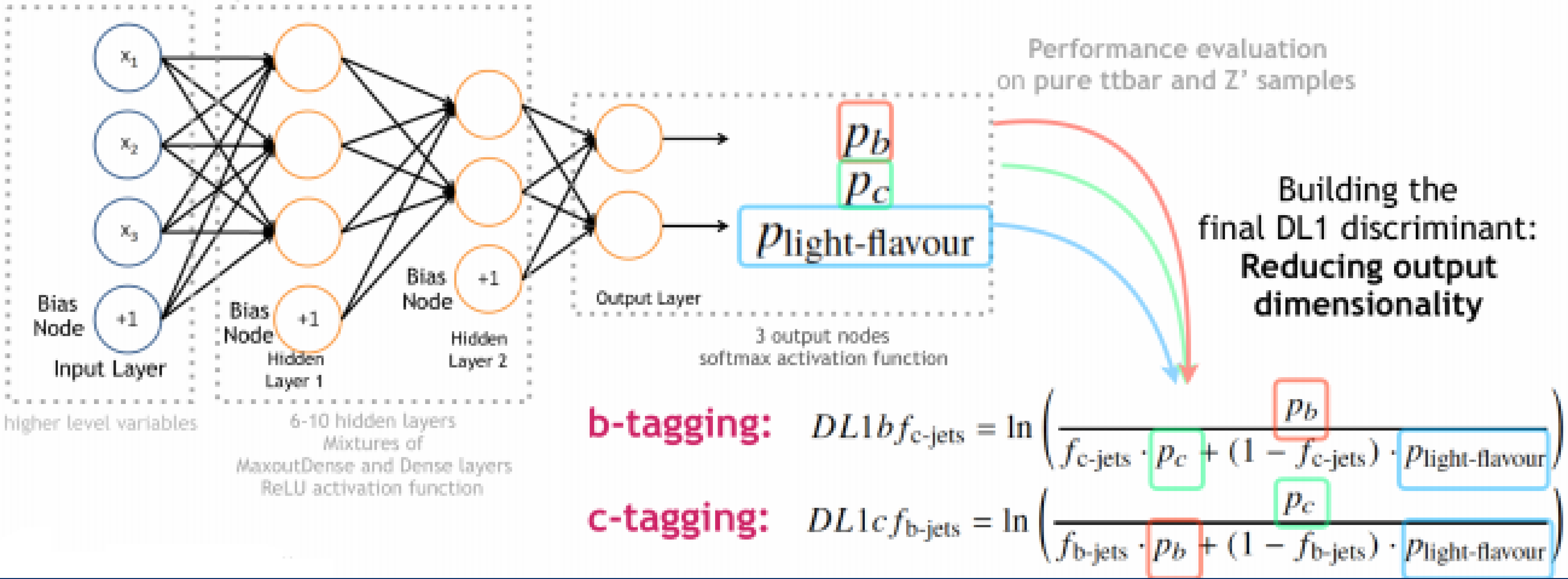
High level tagger that combine inputs from IP2D, IP3D, RNNIP, JetFitterand SL taggers:

- Based on the training of a BDT algorithm, targetting b-jet as signal and c-(light-)jets as background.
- Different versions of MV2 have been developed by ATLAS, relying on improvement in low-level taggers
- c-jet fraction in the training is set to 7% and that of light-flavoured jets to 93%.

DL1 tagger

Based on artificial DNN, trained with **KERAS** and **THEANO** backend using **ADAM** optimizer

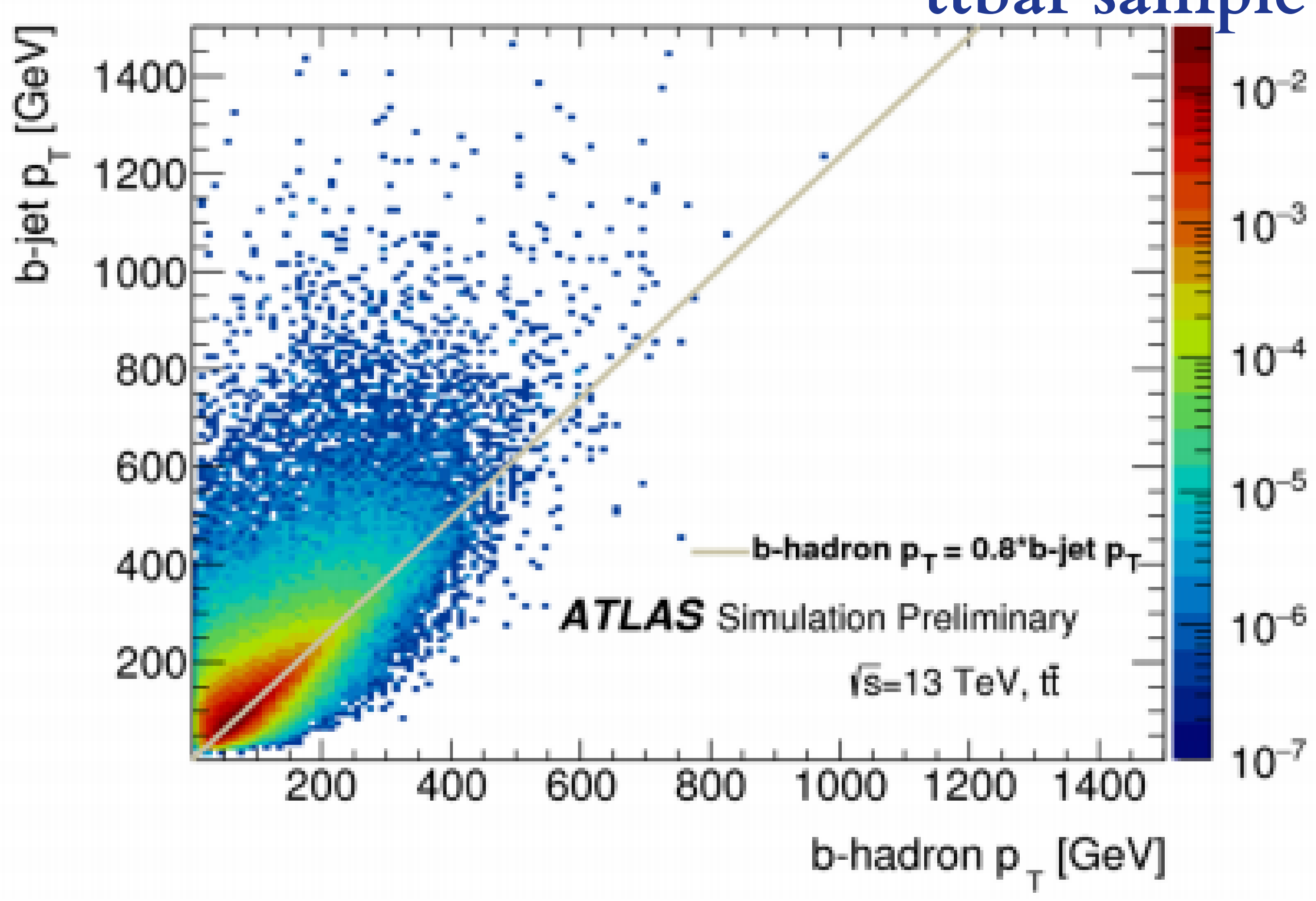
- Multiclassifier: returns the probability for a jet to be a b-,c- or light-flavor jet.
- It makes use of the same input features used in MV2 + full set of SMT and JetFitter c-taggiing variables



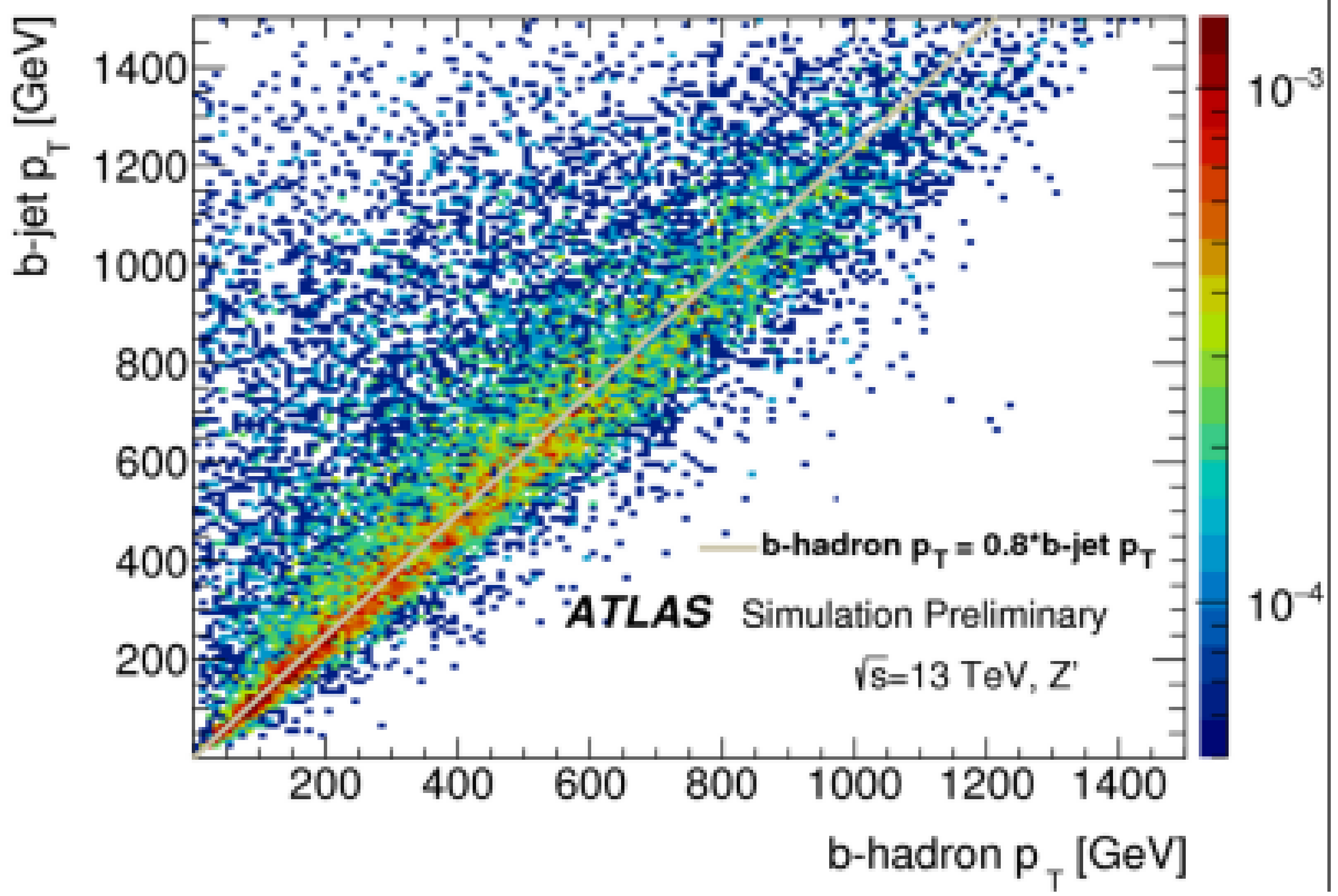
Higher hierarchy b-tagging in ATLAS

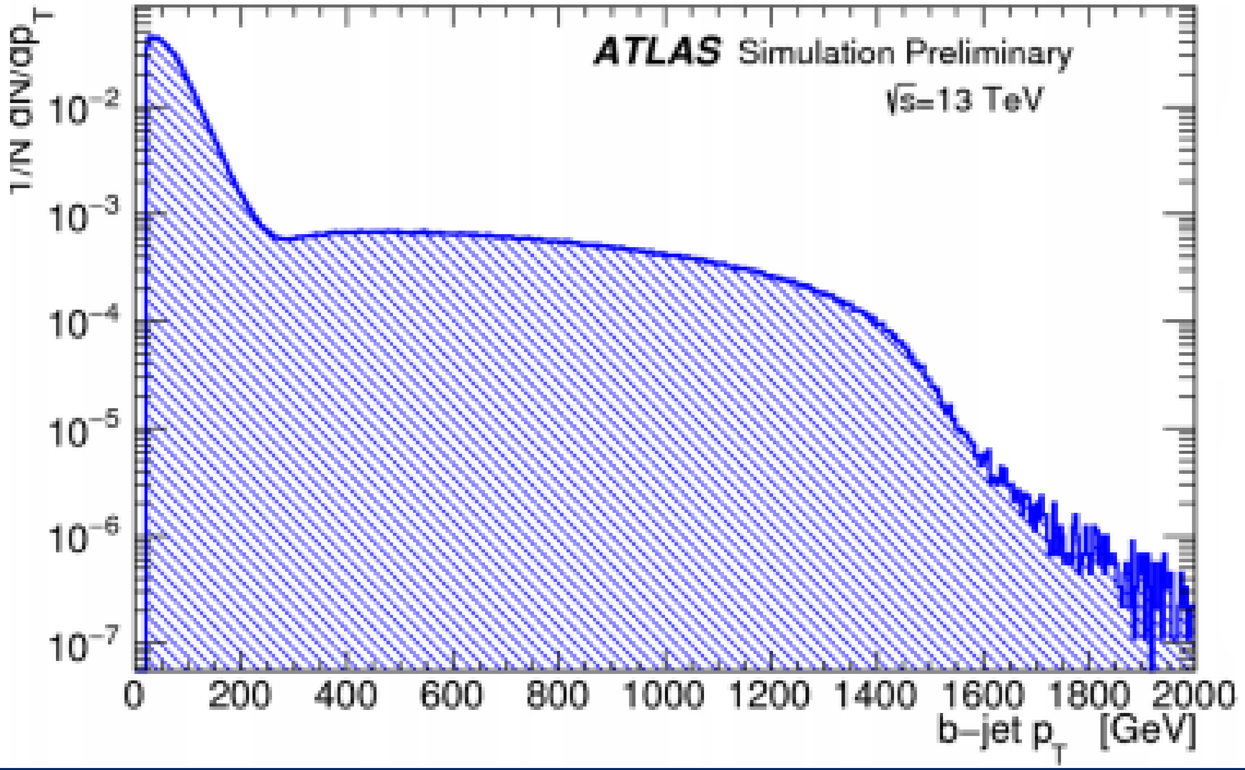
Training of DL1

ttbar sample



Broad Z sample





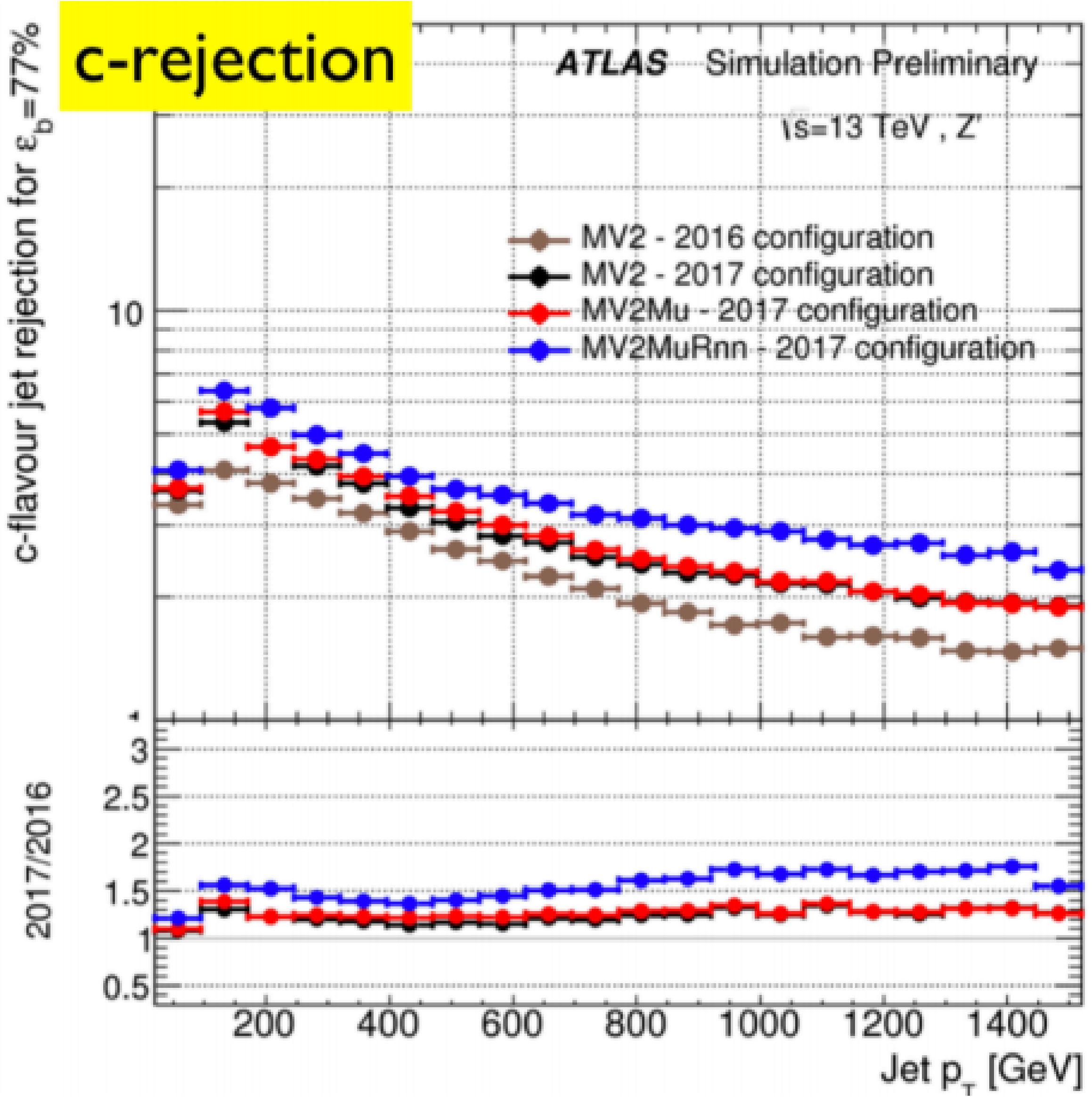
- ttbar sample loses correlation at high pT (merging of jets), while Z' fully characterizes the high pT phase space → Hybrid sample

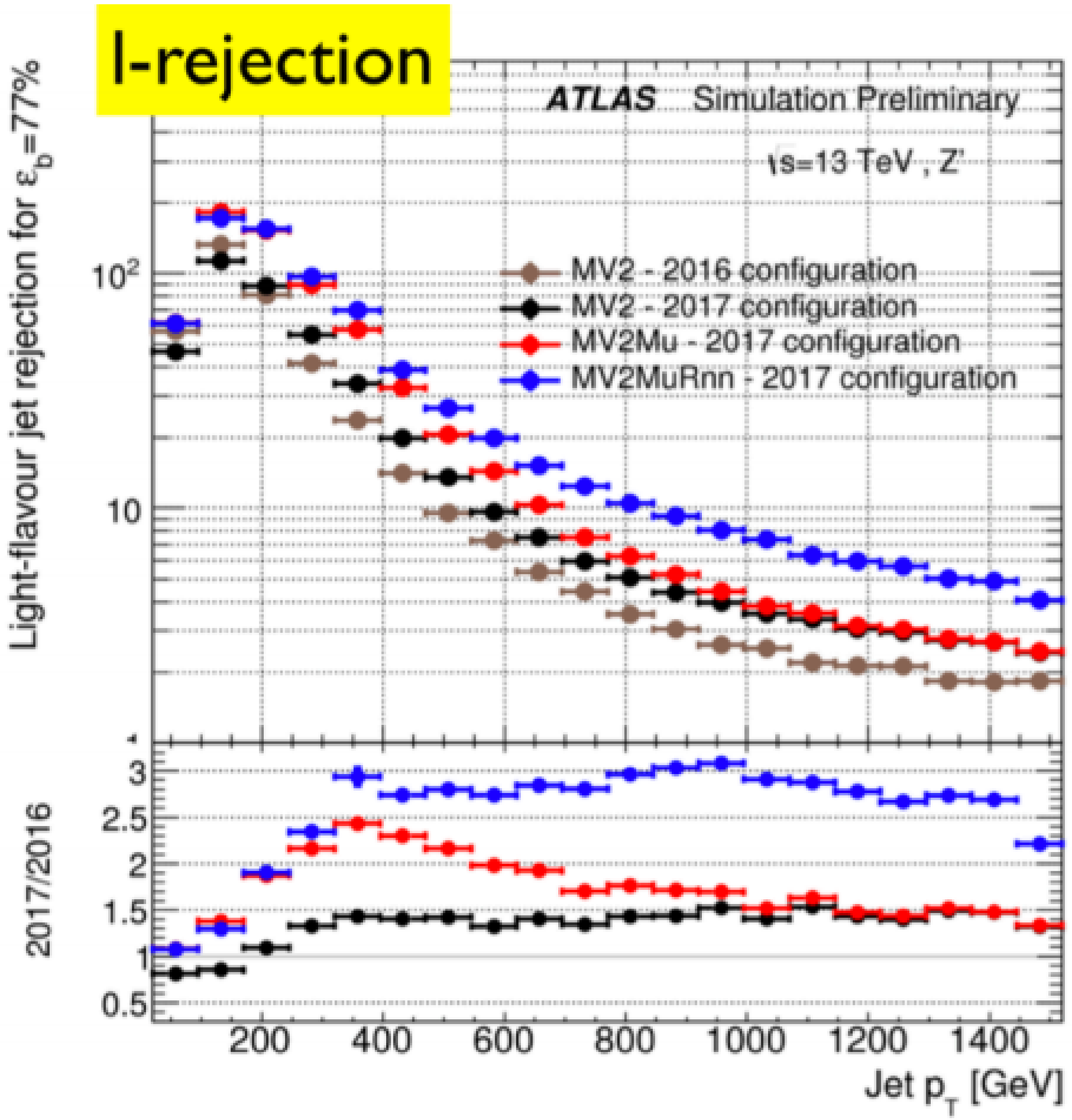
- Similar performance at low pt but significantly larger rejections at high pt

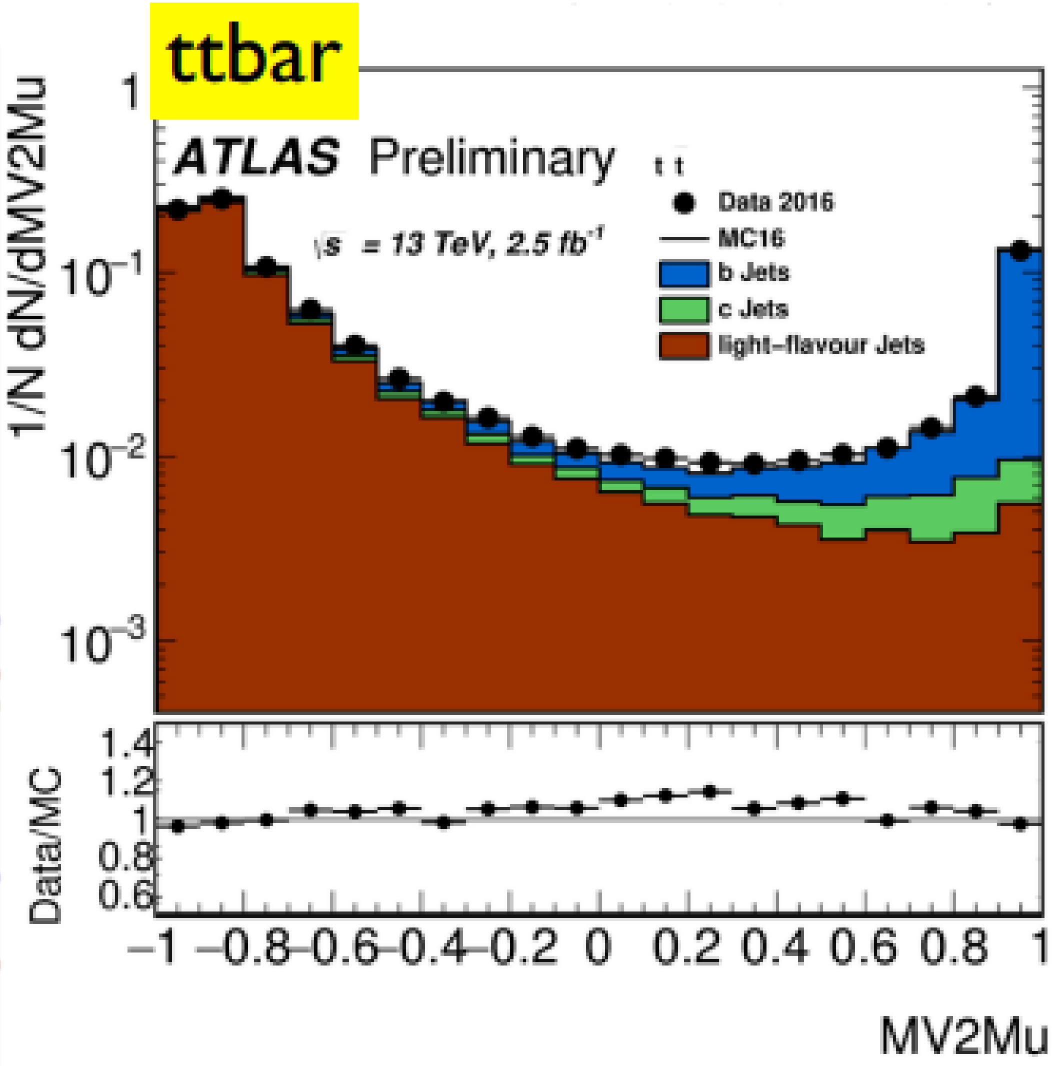
Higher hierarchy b-tagging in ATLAS

Three variants of the MV2 (and **DL1**) algorithms have been deployed

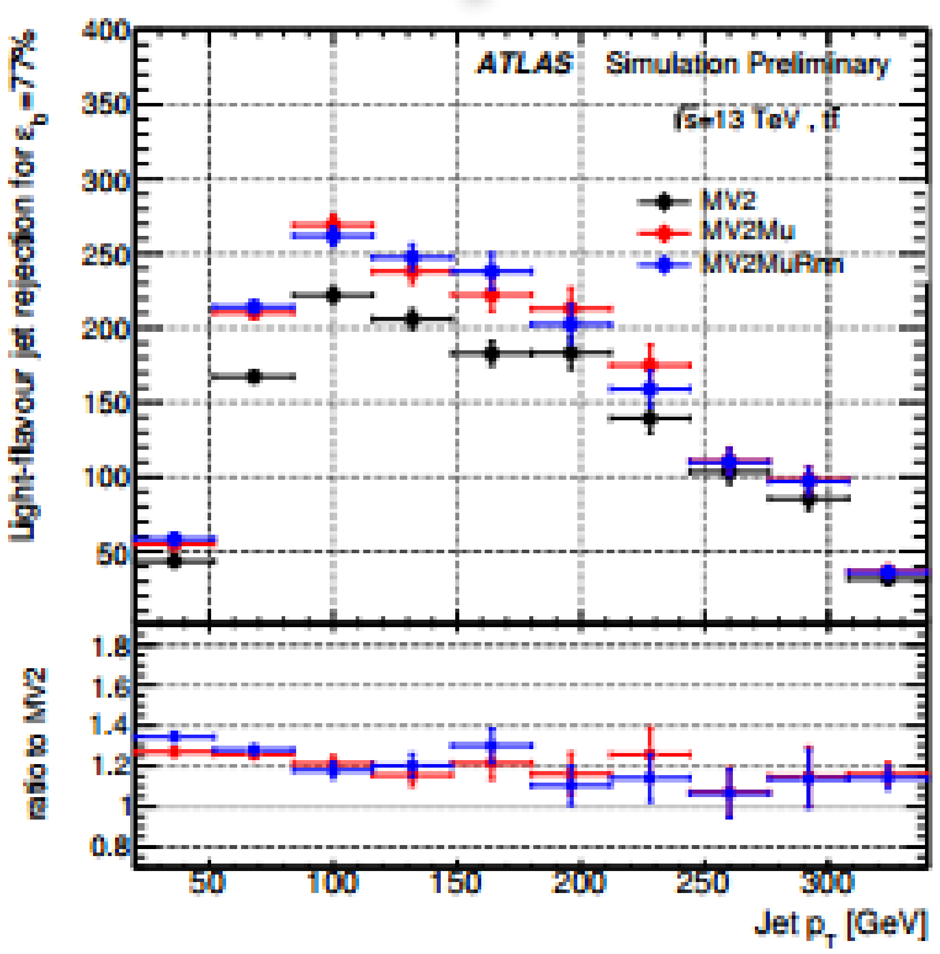
- standard impact parameter and secondary vertex-based inputs + kinematics of the jet (**MV2/DL1**) .
- standard inputs + soft muon tagger (**MV2Mu/DL1Mu**) .
- standard inputs + soft muon tagger+ RNNIP (**MV2MuRnn/ DL1MuRnn**).

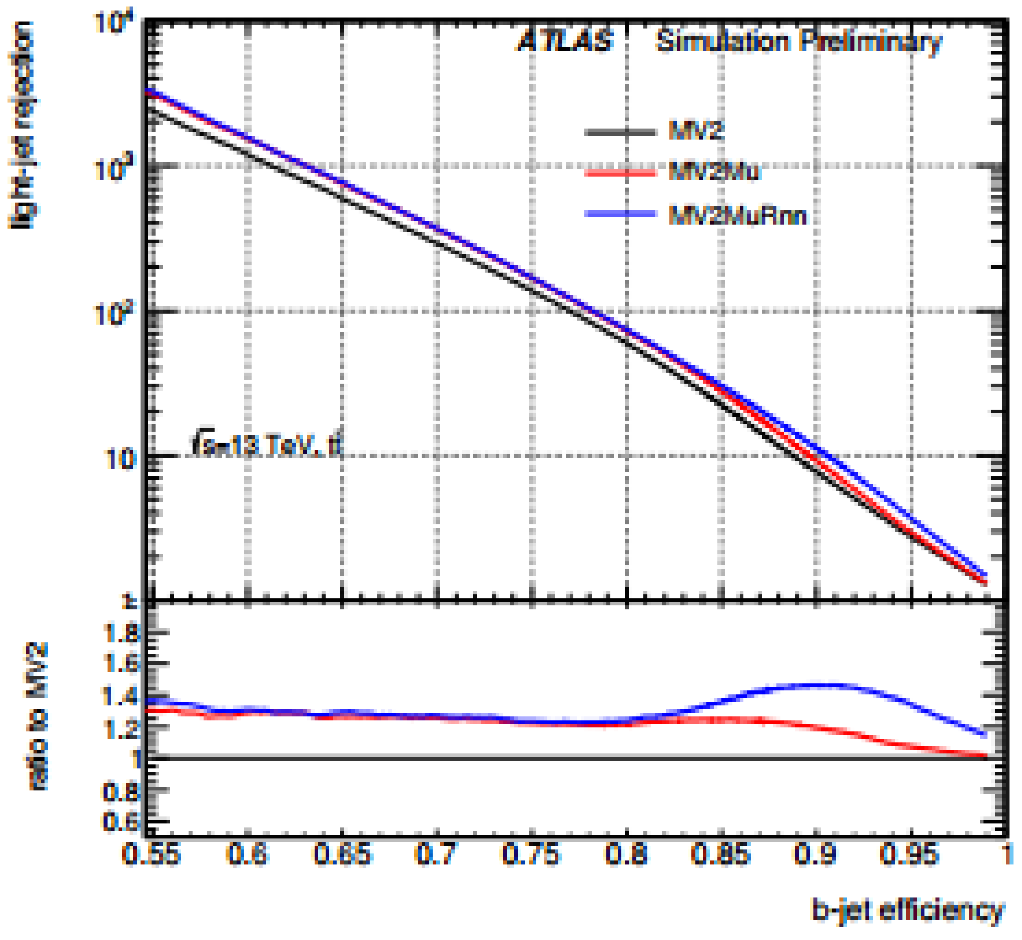




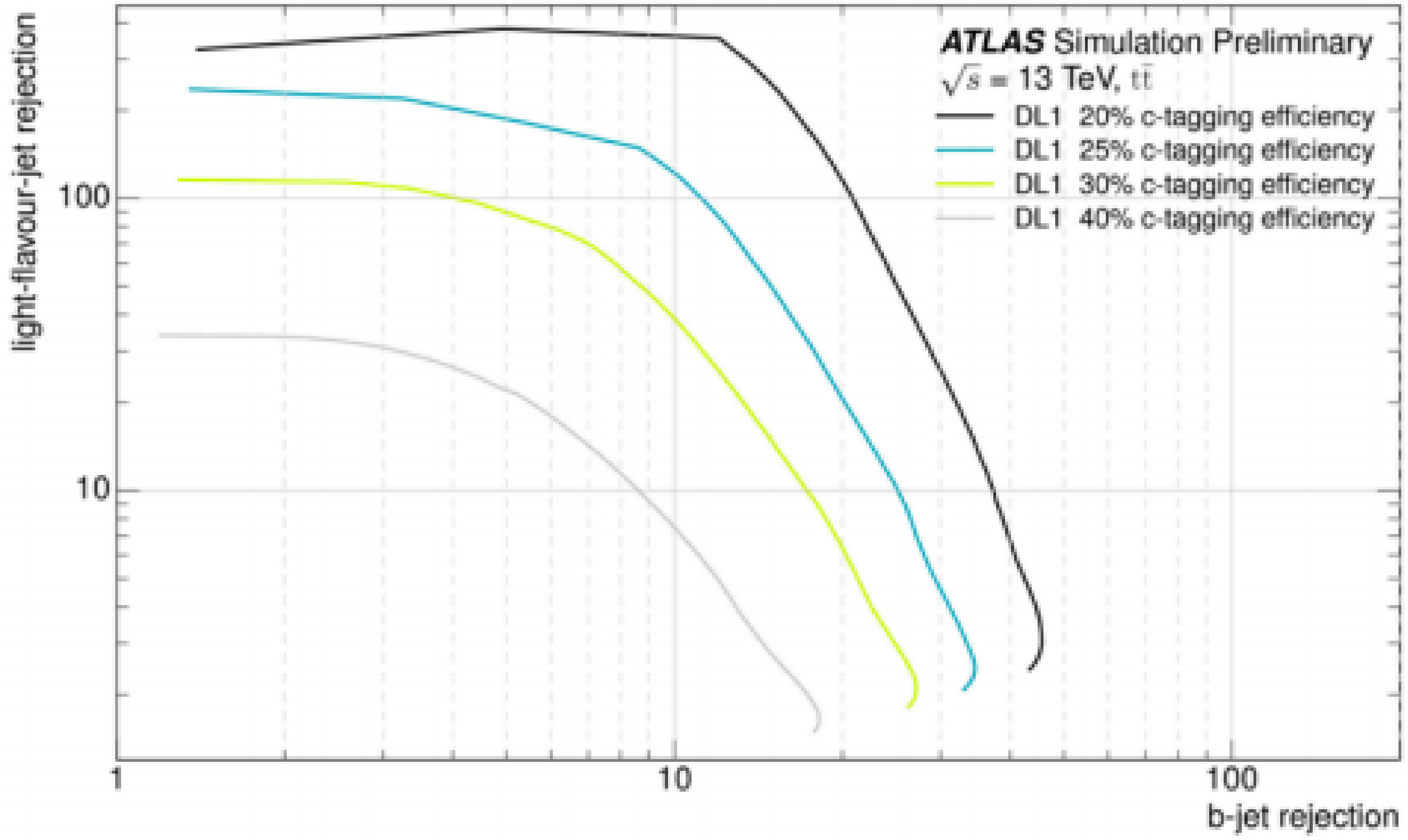


Performance of higher hierarchy b-tagging in ATLAS





- Light-jet rejection as a function of b-jet efficiency for **MV2 (black), MV2Mu (red), MV2MuRnn (blue)**
- Light-flavour rejection as a function of the jet pTfor **MV2 (black), MV2Mu (red), MV2MuRnn (blue).**
- The algorithm evaluation is performed on tt events for a flat **b-jet efficiency of 77% for each pTbin**

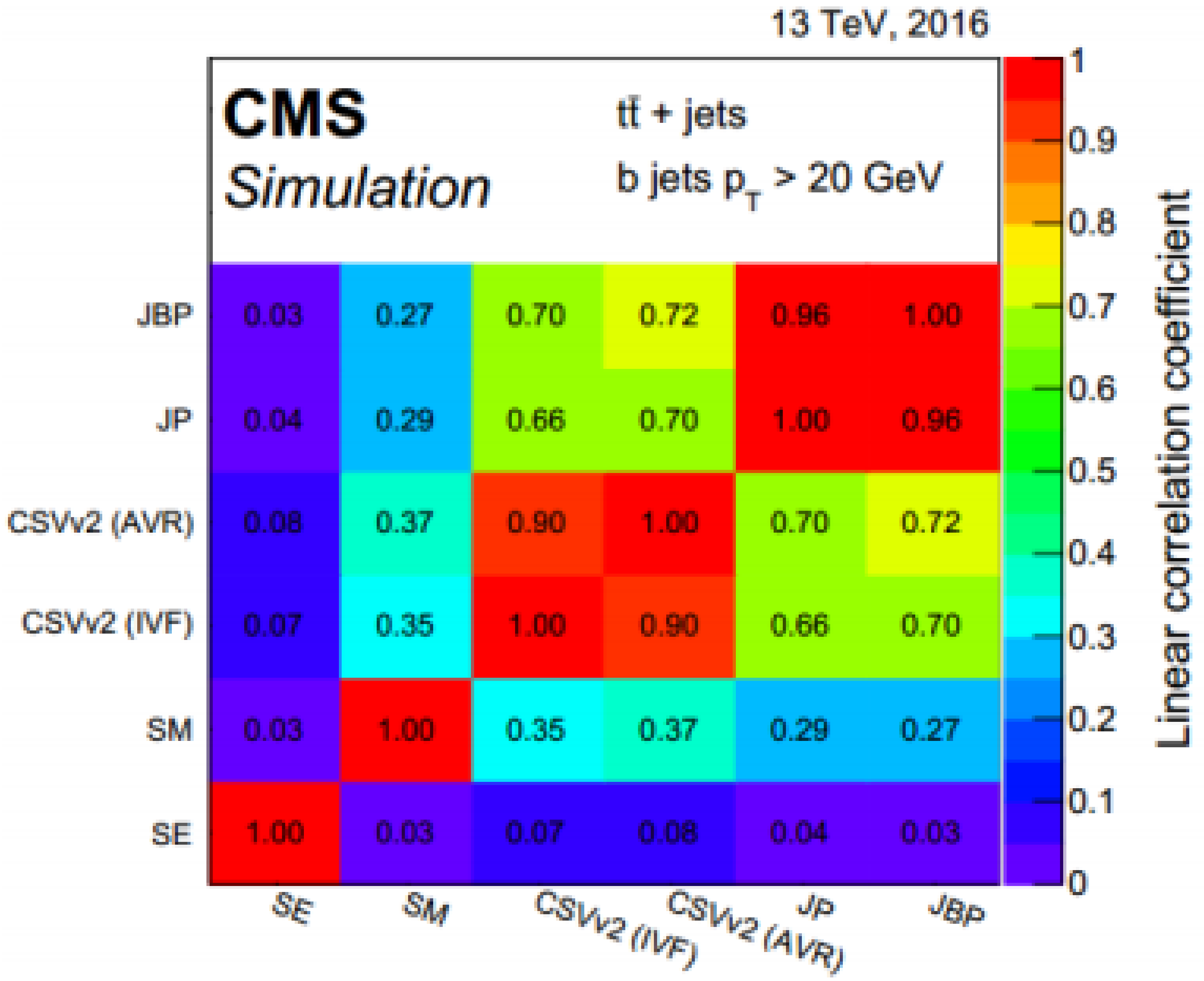


- Light flavor vs jet rejection
- **Performance comparable to MV2**, by having the same input.

Higher hierarchy b-tagging in CMS

cMVAv2

Correlation between the different input variables for the cMVAv2 tagger for b-jets in tt events



DeepCSV: DNN architecture

- Input variables go through **4 fully connected layers**, each layer has **100 nodes.**
- ReLu activation function used in each of the hidden nodes
- Output layer → soft max activation function → multiclassification

Training:

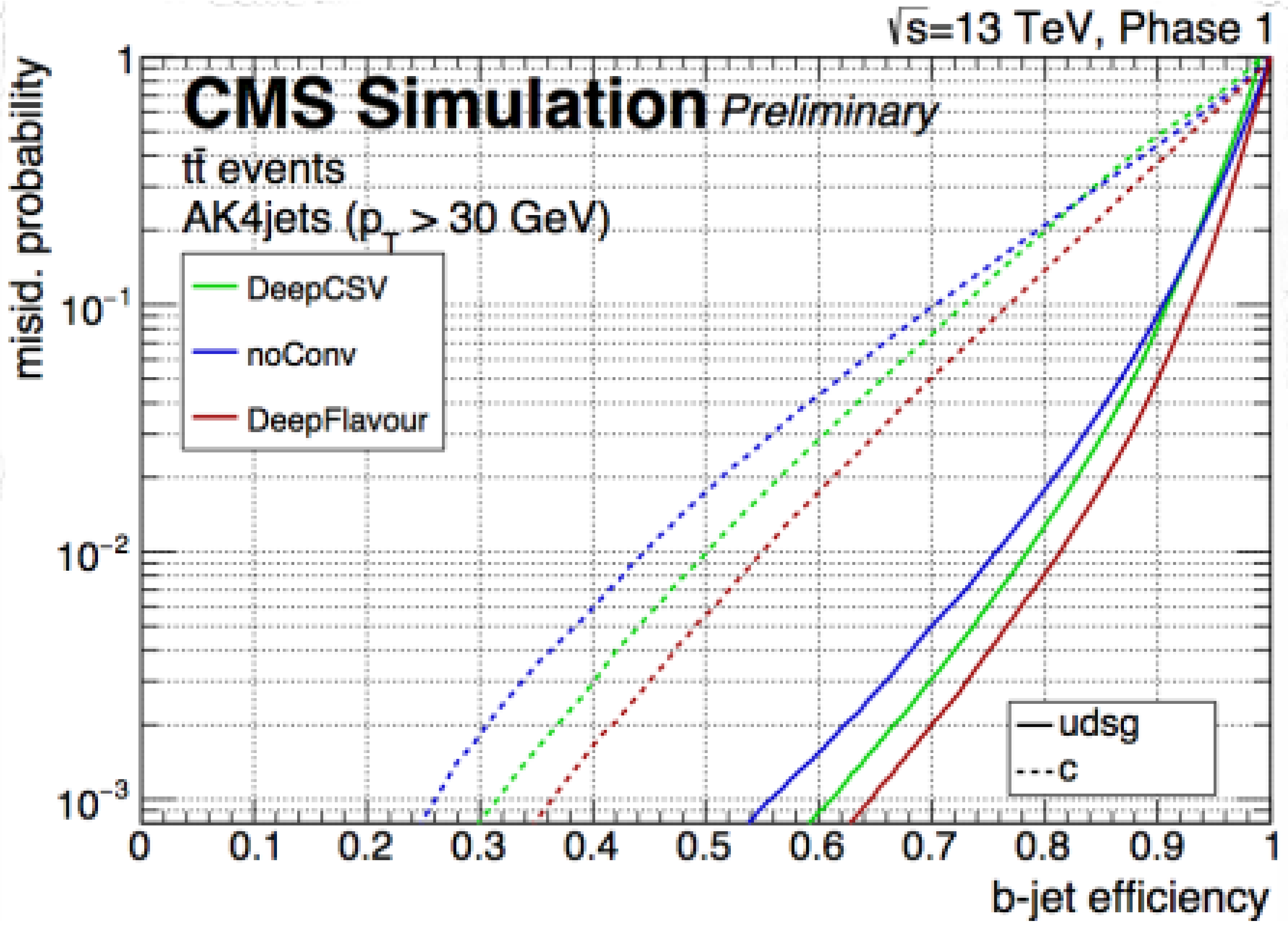
- Jets with **pTin [20,1000] GeV** and flavour ratio fixed to **2 : 1 : 4 for b : c : light**
- Mixture of **tt and multijets events** reduce dependence on heavy-flavour quarks production process.

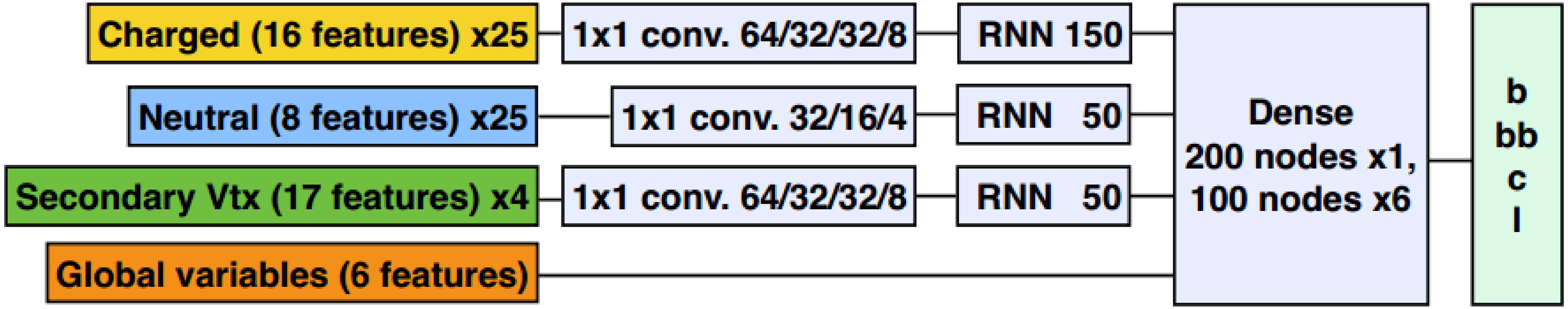
Higher hierarchy b-tagging in CMS

DeepFlavour

- Going further in exploiting DNN techniques
→ DeepFlavour Tagger

- Using properties of jet constituents and topological features of the reconstructed SV
- Four output nodes for b, bb, c and light
- Added a convolutional layer



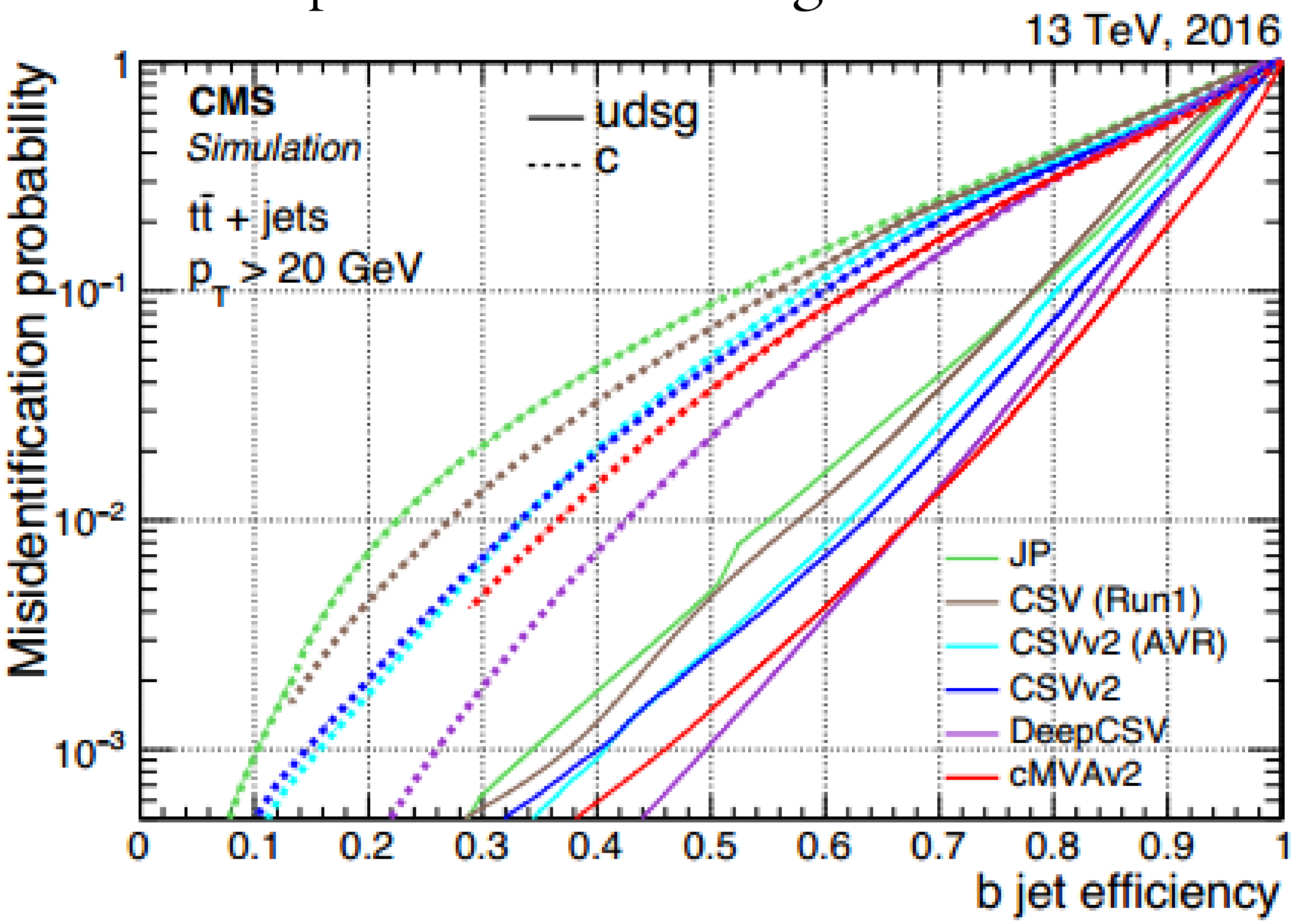


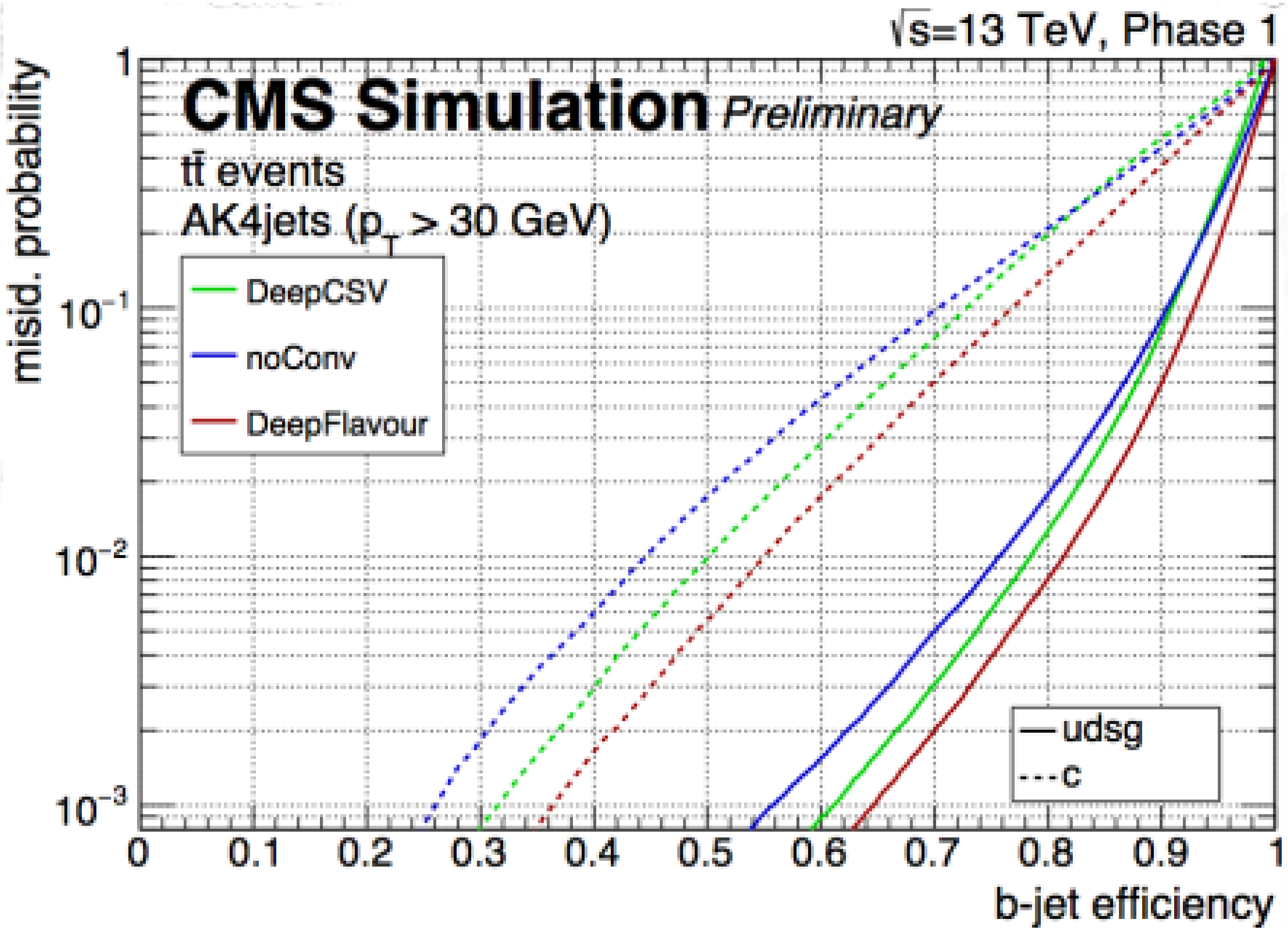
Performance of higher hierarchy b-tagging in CMS

Tagger	Working point	ε_b (%)	ε_c (%)	ε_{udsg} (%)
	DeepCSV L	84	41	11
Deep combined secondary vertex (DeepCSV) $P(b) + P(bb)$	DeepCSV M	68	12	1.1
	DeepCSV T	50	2.4	0.1

DeepCSV WPs are defined as **values of the discriminator cut for which the light mistag-rate is 10%, 1%, and 0.1%**

- Large improvement in the performance of DeepCSV wrt CSVv2 algorithm





- Overall **non-negligible improvement** wrt **DeepCSV of 5% at 0.1% mistag rate!**

Performance in data: CMS

Efficiency for b-jets (Measurement in ttbar events, 4 methods):

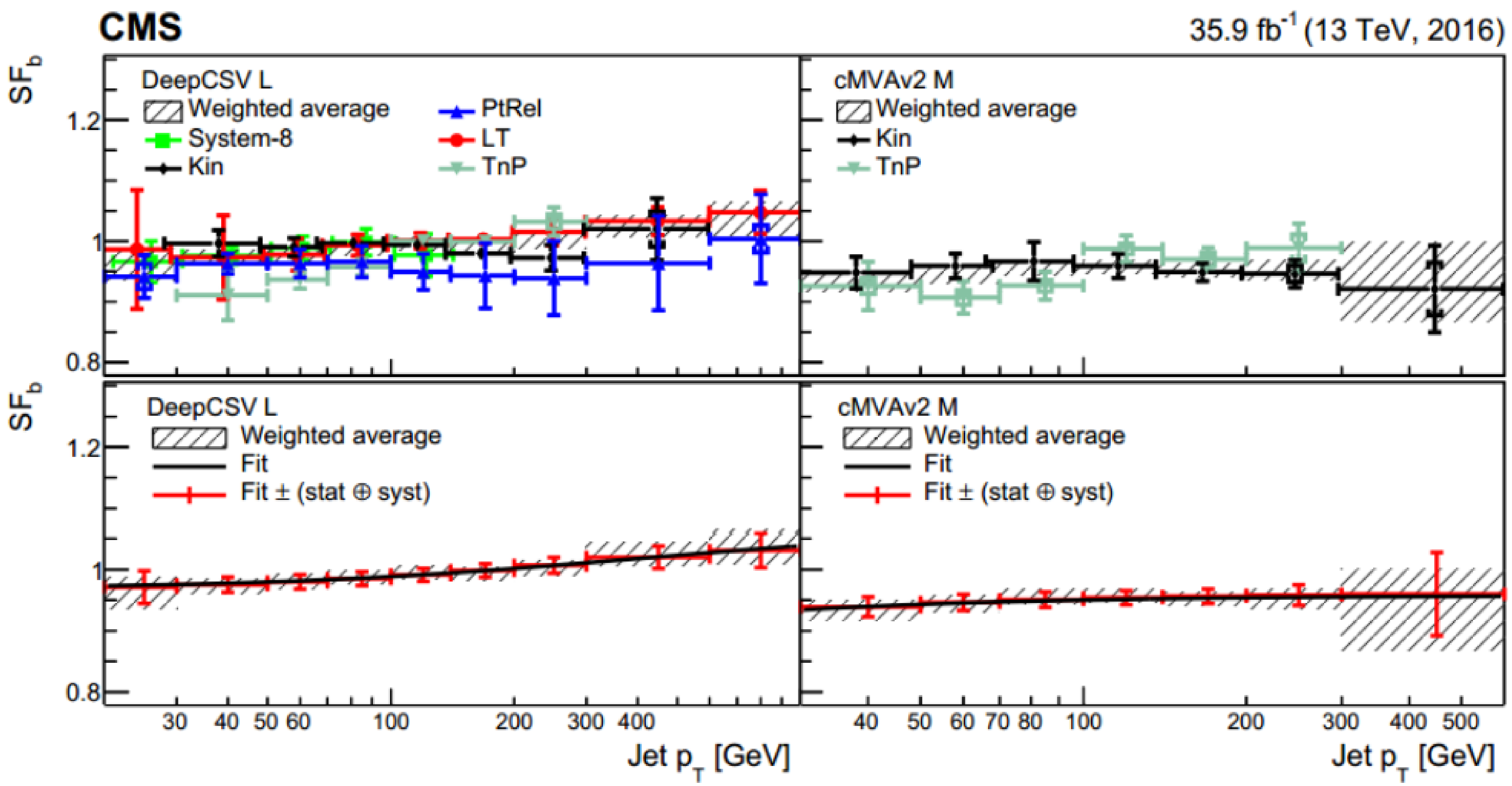
- **Kin:** di-leptonic channel, based on template fit.
- **TnP:** semileptonic channel. Tag: CSV medium requirement to either b-jets. Probe: the other.
- **TagCount:** di-leptonic channel, based on counting events with two b-tagged jets in selected sample.
- **Iterative-Fit:** di-leptonic channel, based on calibration of the full b-tagging shape.

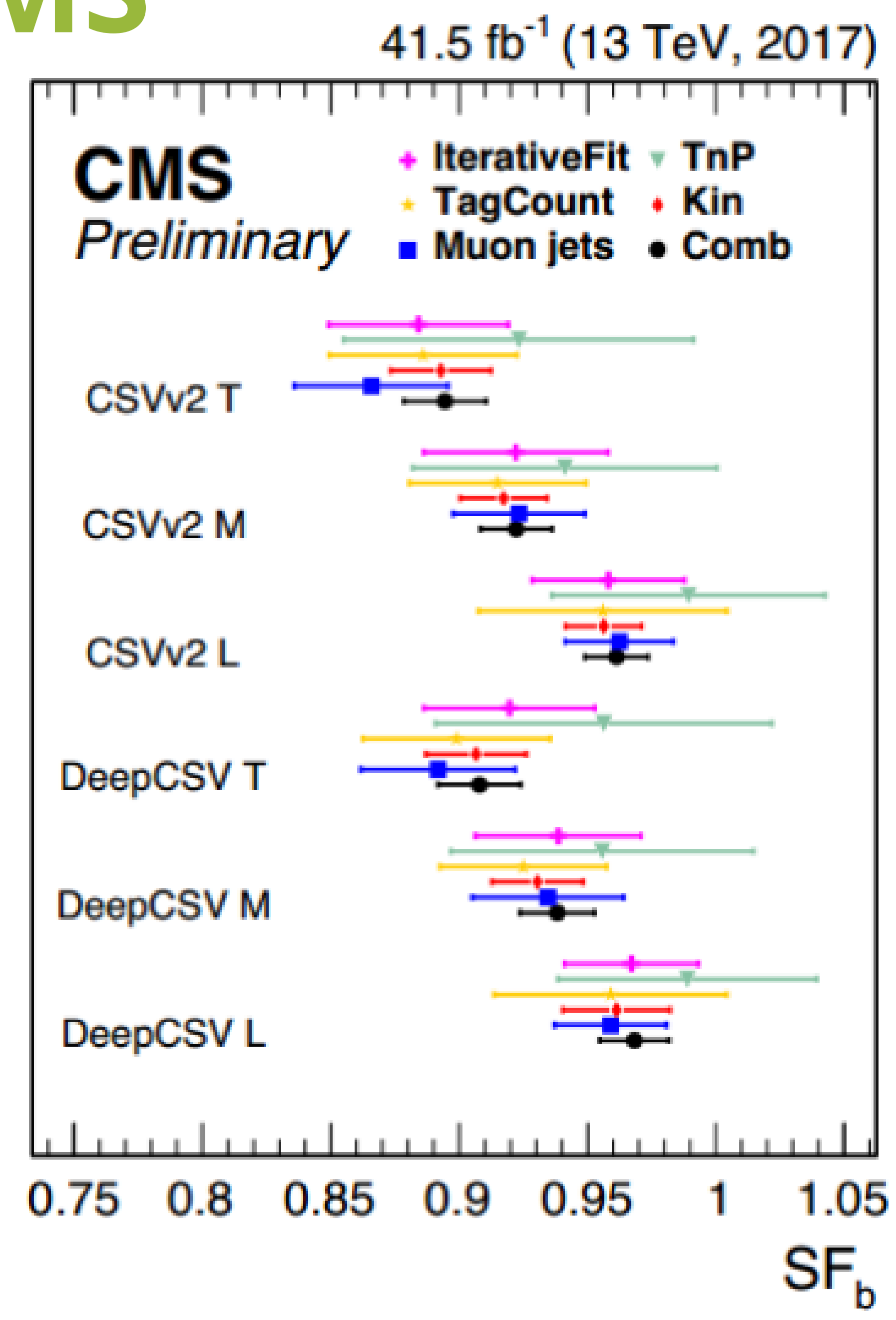
Scale factors are produced

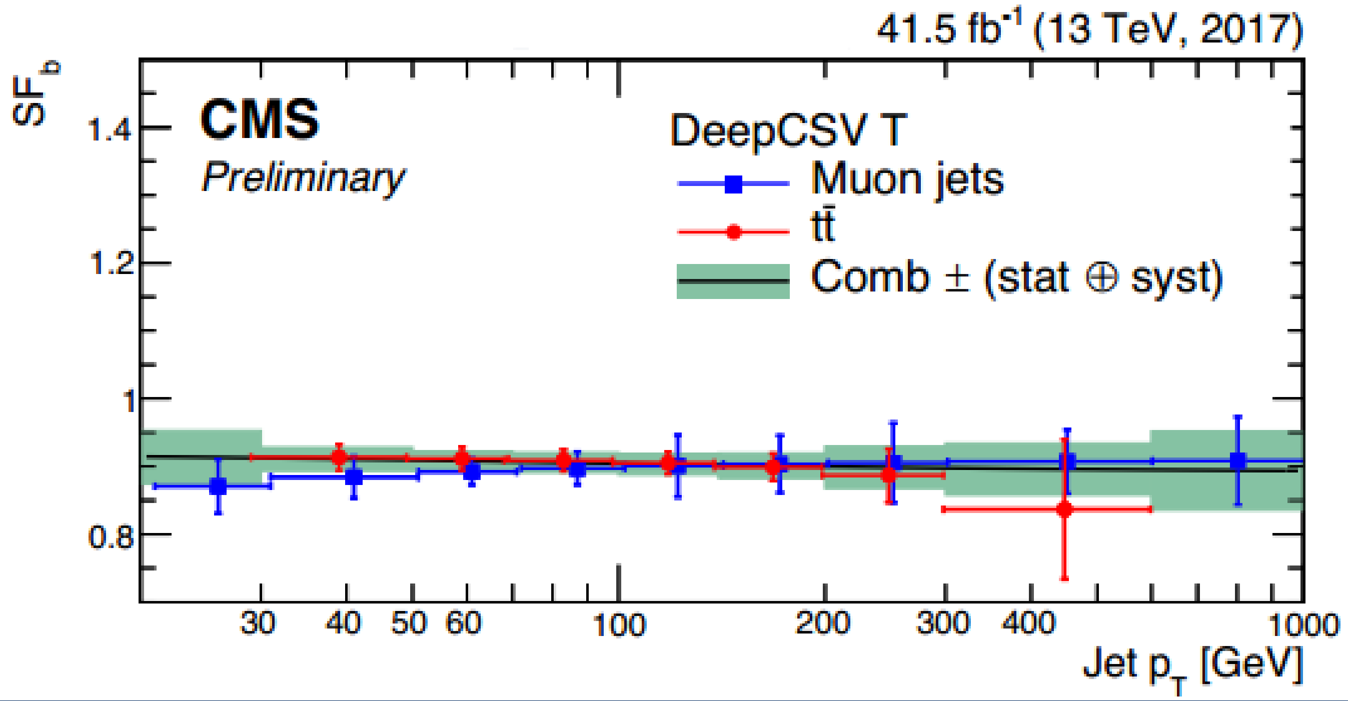
Mu-enriched sample

- Configuration of **two AK4 jets** with **$50 < p_T < 250$ GeV**, with at least one containing a **muon with $p_T > 5$ GeV**.
- **PtRel**, the **LT** and the **System8** methods

Performance in data: CMS

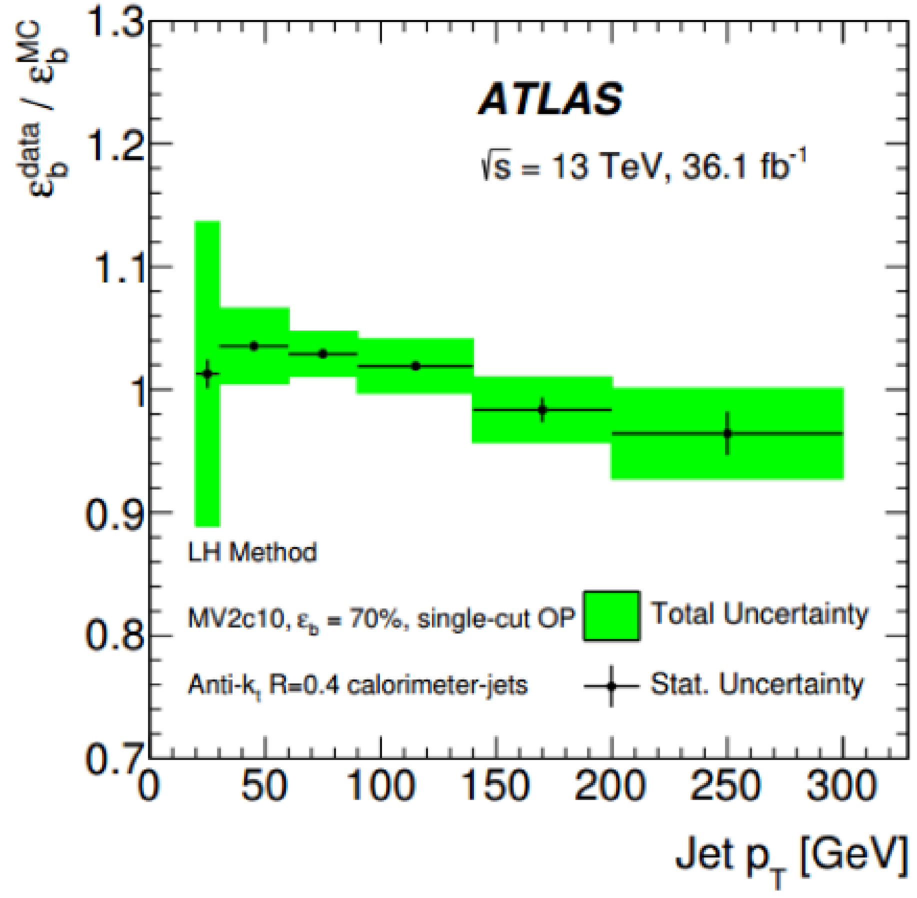


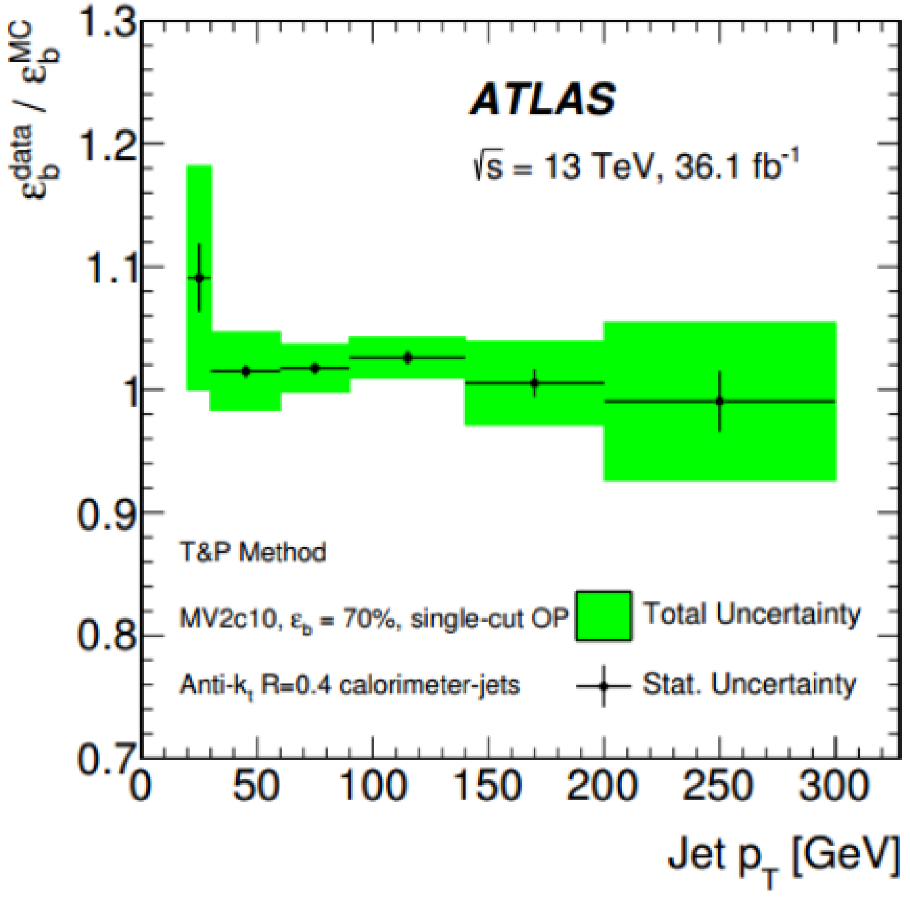




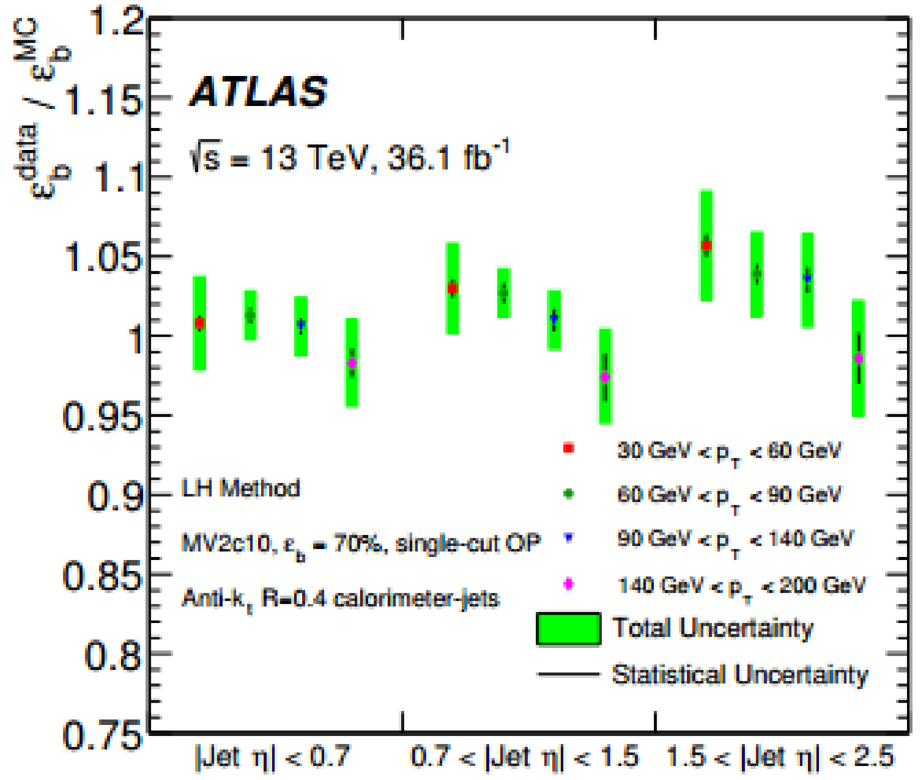
- The values of the scale factors are **averaged over the spectrum of b-jets** from ttbar events.
- The **scale factors** measured with the different methods **agree within their uncertainties.**

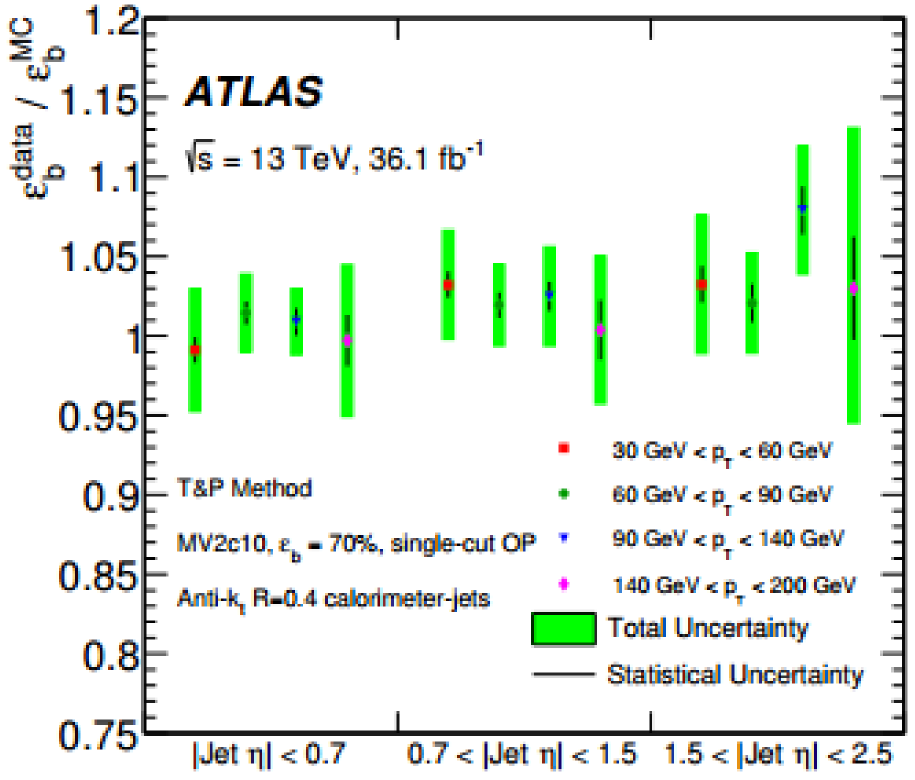
Performance in data: ATLAS





- **70% b-jet tagging efficiency** single-cut OP, for **R = 0.4** calorimeter-jets
- **Combinatorial likelihood method** or **TnP**





- The SFs found via the two different methods **agree within the experimental uncertainty.**
- Scale factors have been derived in different pseudorapidity bins.

Uncertainties:

ATLAS

- **MC generator modeling** → Kinematic distribution and flavor composition.
  - Using alternative generators alter the scale factors.
  - Changing settings in the nominal generator→ changes parton distribution function.
- **Normalization uncertainties** → 6% in from the predicted cross-section of the single top sample.
  - 20% uncertainty for the extrapolation of Z + jets normalization.
- **Experimental uncertainties** → reconstruction of electrons, muons, jet mismatching, mismodelling pile-up.

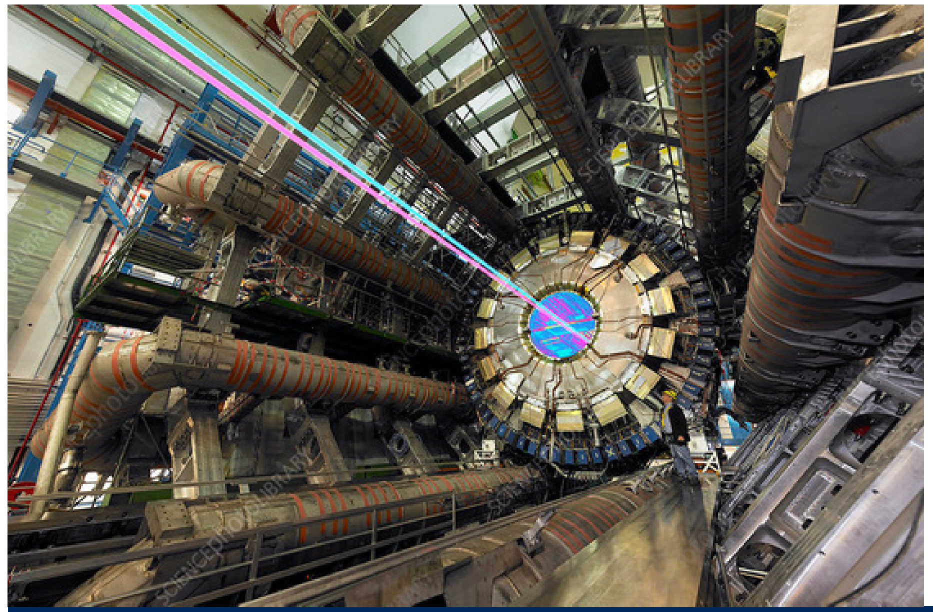
Uncertainties highly dependent on the energy.

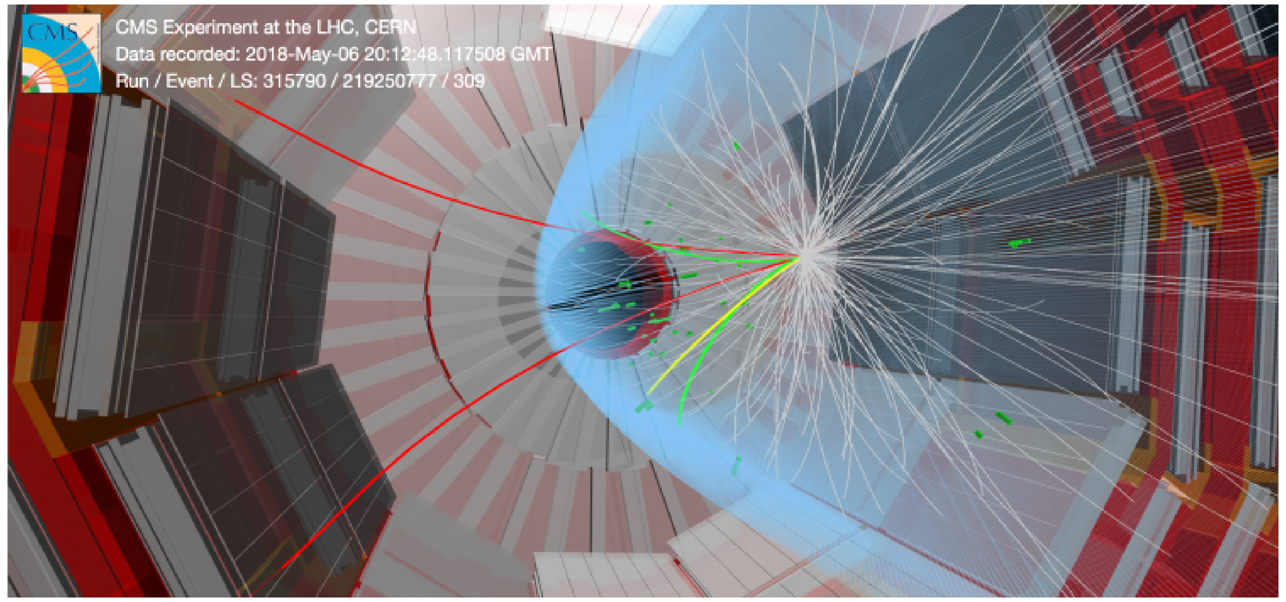
CMS

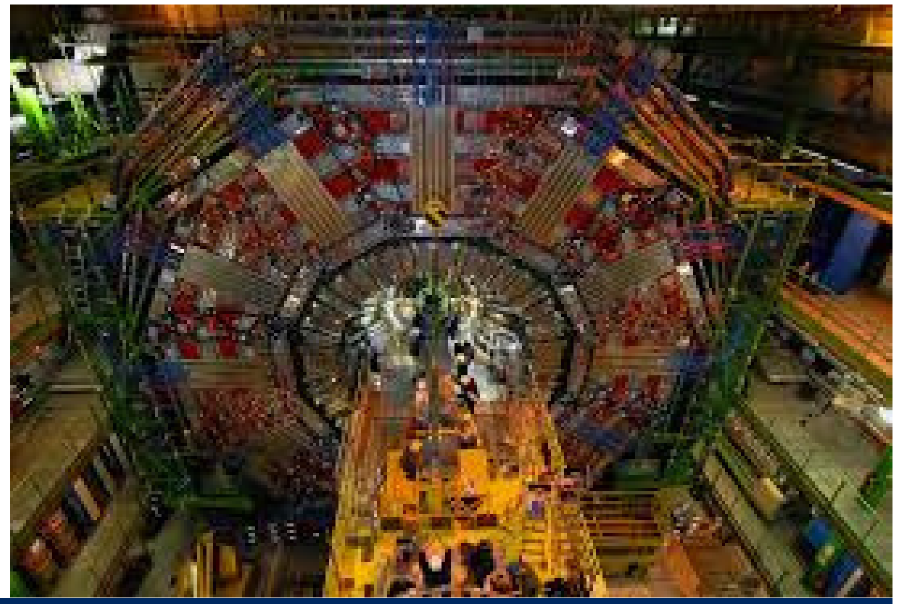
- **Scale factors influenced by gluon-splitting** → highest track multiplicity on heavy flavor jets → **25% variation estimation.**
- Modeling of b-quark fragmentation affects pT of b-sample (**5%**).
- Uncertainty by applying a calibration derived from simulation onto data (eg. **LT relying on JP calibration**).
- Muon methods (eg. **System-8**) take the longest deviation in the scale factor as systematic error.
- Jet energy corrections and pile-up interactions.

Conclusions:

- **b-Tagging is a fundamental tool in most physics analyses.**
- Both **ATLAS and CMS reached a significant improvement on their algorithms in Run-II**, in particular through the development of **Deep Neural Network-based algorithms.**
- **High-level taggers strongly rely on inputs from low-level algorithms** exploiting the kinematics of the b-jets to ensure separation against c-/light-flavour-jet backgrounds.
- **Similar approaches in ATLAS and CMS on how to tackle the extraction of the data/MC scale factors for b-, c- and light-flavor jets.**
- **Still some challenges ahead**, i.e. nature of light-flavor fake-rates (**resolution effects**), **high-pt extrapolation for b-jets,...**





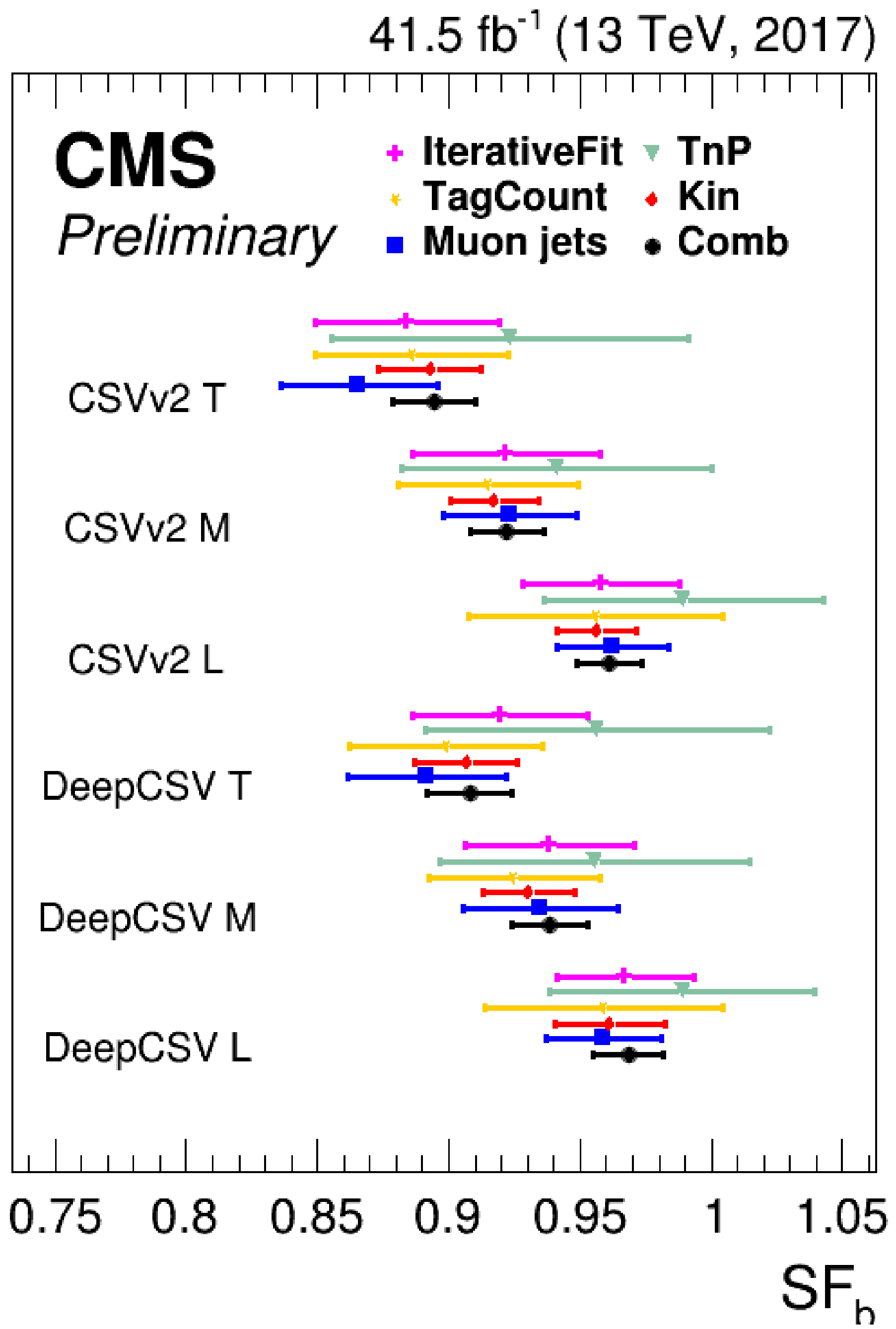


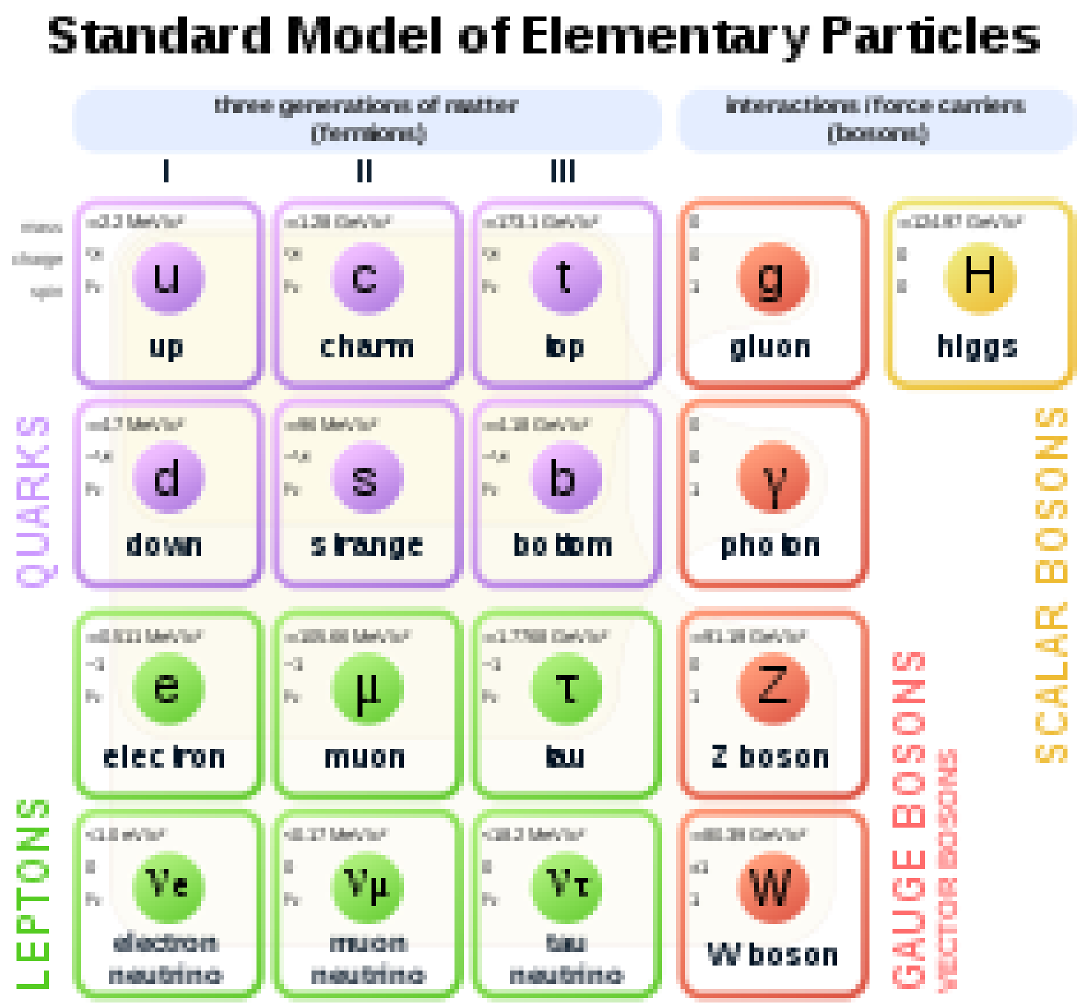
Additional slides: CMS calibration with muon

- **LT**: Lifetime tagging method is used for the measurement of the b-tagging efficiency in multijet events, based on template fits the JP or CSV distributions. The strategy behind it is similar to the PtRel method, but the fit is performed on the JP discriminator distribution having also the same sample.

- **System-8**: Method used for the measurement of the b-tagging efficiency in multijet events with
a muon, solving a system of equations vertices.

- **PtRel**: The pT of the muon relative to the jet axis, pT_rel, is a variable that is able to discriminate
between b-jets and non-b-jets.





Additional slides: CMS and ATLAS detectors

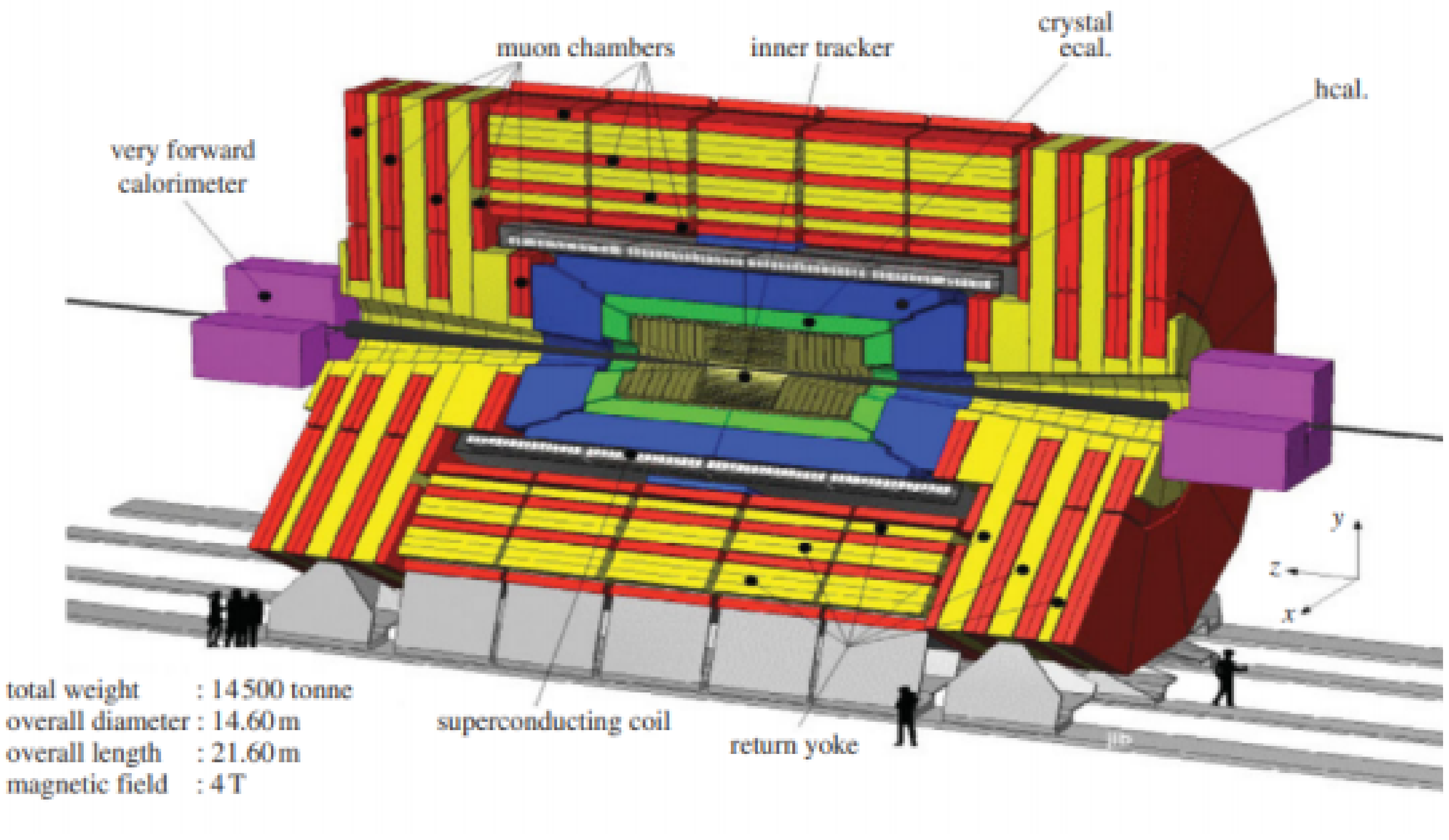


Figure 1. Schematic of the CMS detector.

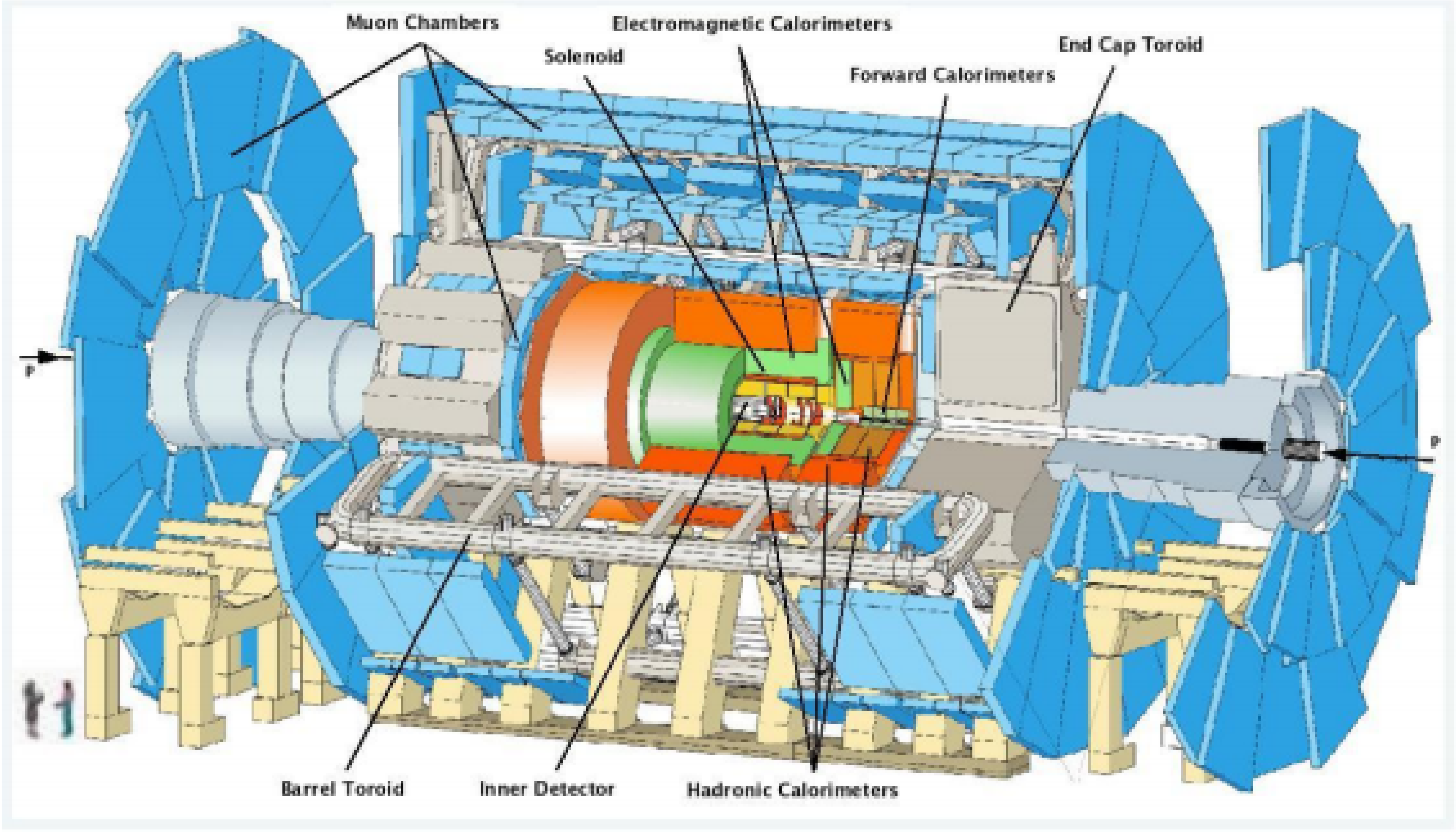


Figure 2. Schematic of ATLAS detector.