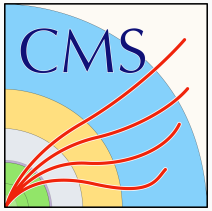


University of Zurich[UZH]

A new algorithm for b tagging very high p_T jets via pixel hit multiplicities

Manuel Sommerhalder

BTV meeting 26.06.19

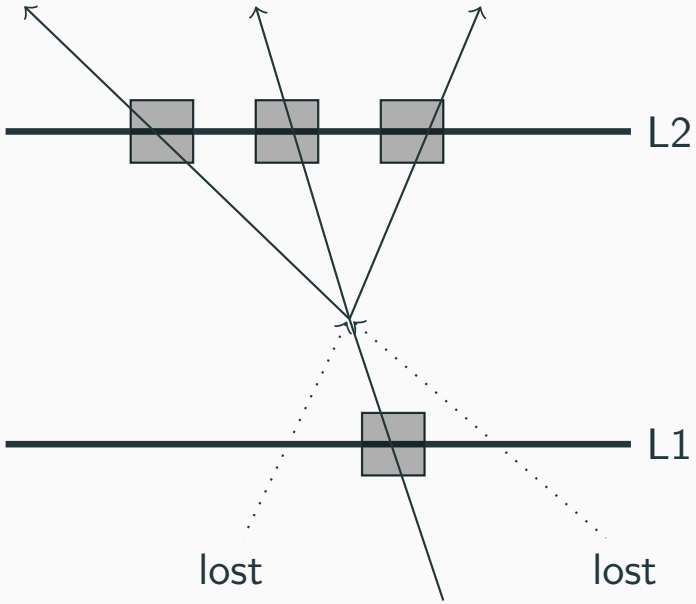
Bachelor/Master thesis supervised by
Prof. Dr. Ben Kilminster
Dr. Alberto Zucchetta
Dr. Thea Årrestad
Dr. Yuta Takahashi

github.com/msommerh/bTag_HitCount

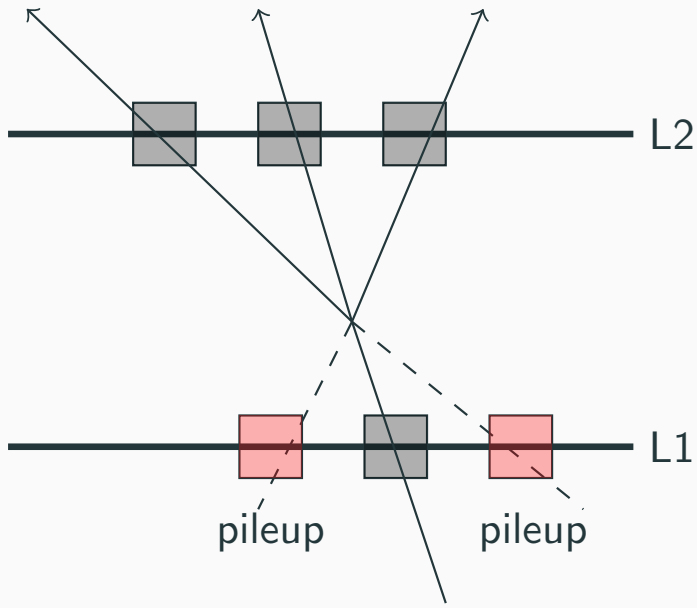
Motivation

- decay of highly boosted B hadrons between pixel detection layers causes a lack of hits in the earlier layer
- efficiency loss in track reconstruction at extreme p_T due to missing hits

lost tracks

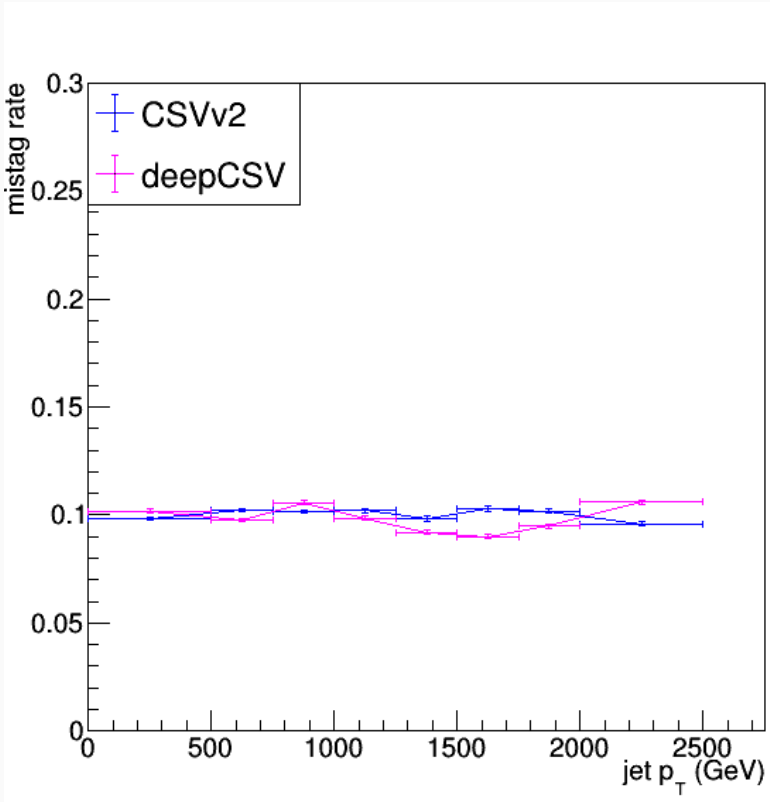


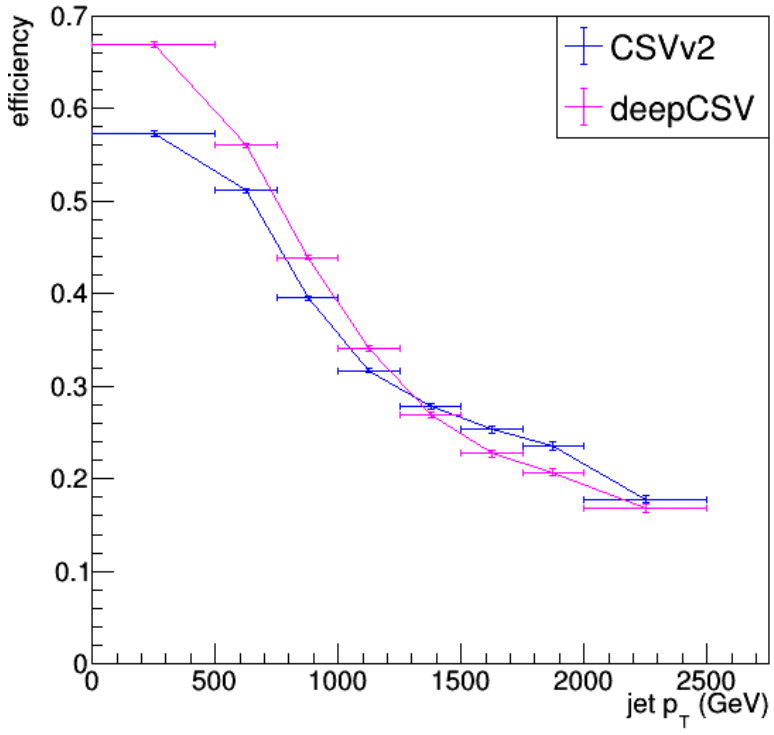
wrongly reconstructed tracks



Motivation

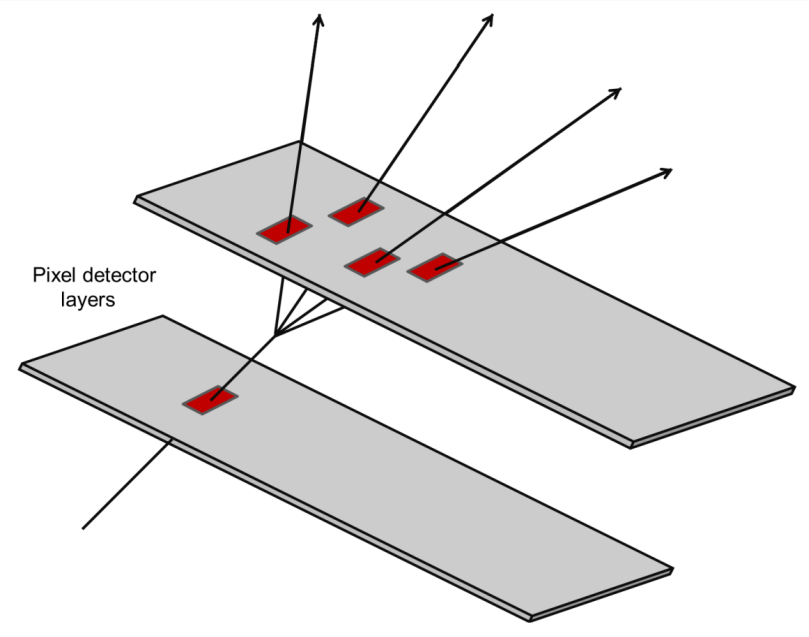
- all state-of-the-art taggers in CMS like CSV are based on PF candidates which in turn are based on tracks
- degrading b tagging performance at very high p_T
- what if we want to tag 6 or 8 TeV resonances in Full Run-II searches?





Motivation

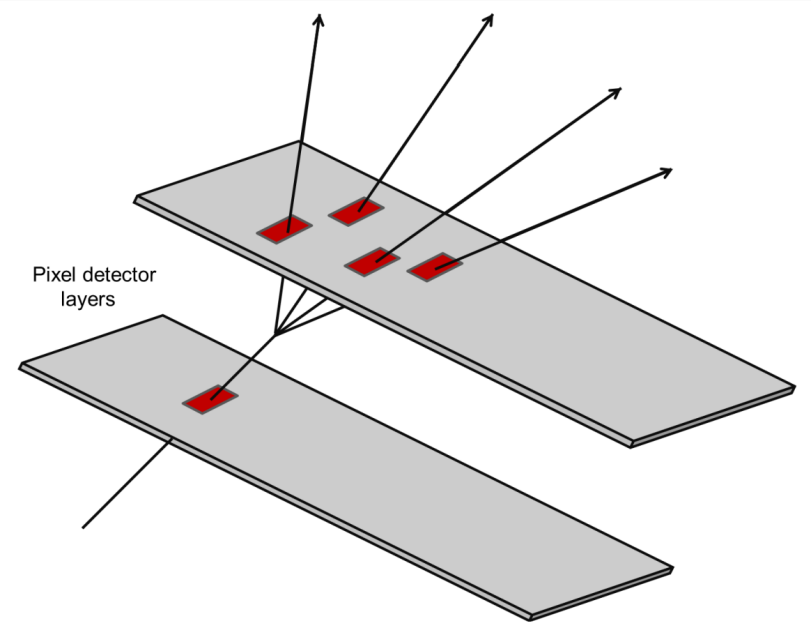
- alternative approach by B. Todd Huffman et al. (*arXiv:1604.05036* and *arXiv:1701.06832*)
- count pixel hits inside a small cone around the jet axis in each layer
- build tagging variables based on the increase in hit multiplicity in subsequent detector layers



arXiv:1604.05036

Aim of the study

1. check if a hit multiplicity increase is found for high p_T B hadrons on a CMS detector simulation
2. develop a simple cut-based b tagger from this
3. check if such a tagger is competitive to the CSV at very high p_T
4. further develop the tagger with MVA techniques to maximize the performance



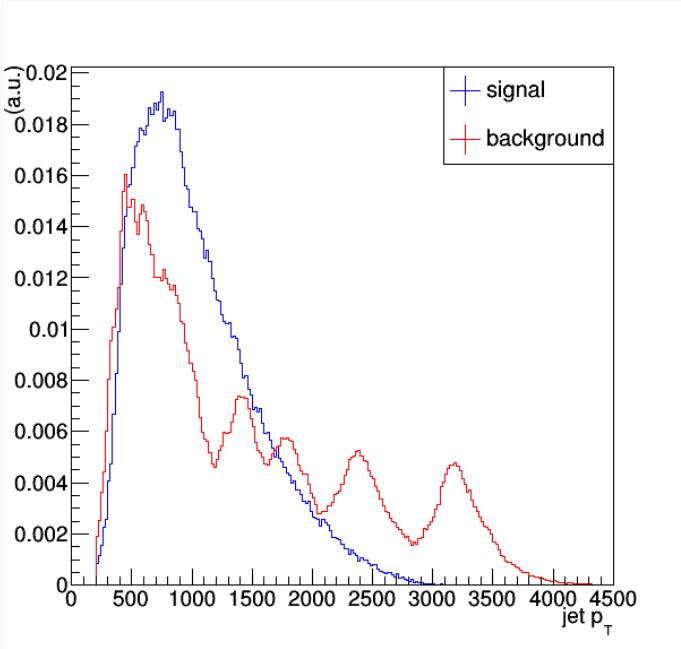
arXiv:1604.05036

Method

1. generate high-p_T b jets ($p_T > 500$ GeV)
2. geometrically match hits in each layer to the jets
3. construct hit-based variables
4. compare the variables of b jets (signal) to those of light-flavor jets (background)
5. develop discriminants to distinguish signal from background

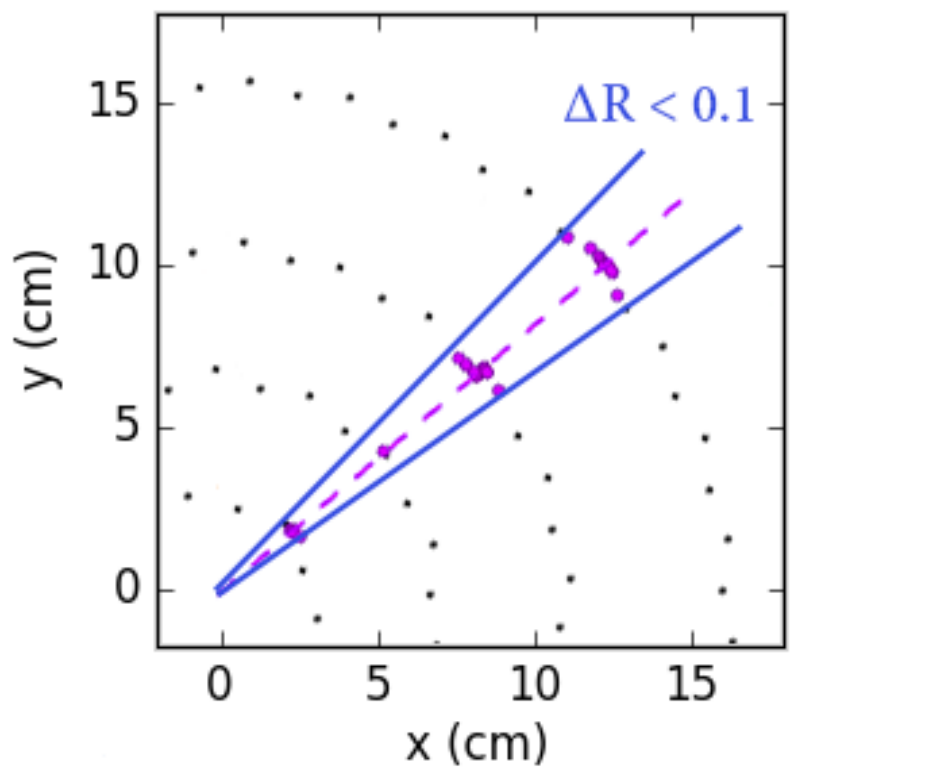
Monte Carlo samples

- signal: $Z' \to b\bar{b}$ at 9 mass points from 2 to 6 TeV, 2017 and 2018
- background: QCD samples produced in pT-bins, *TuneCUETP8M1 (102X)*, 2018
- momentum threshold of 350 GeV imposed on outgoing particles
- jets within 200 GeV$< p_T < 2500$ GeV selected
- with respect to the previous presentation, the samples have been updated to more realistic conditions
- p_T distributions before reweighting:



Pixel cluster matching

- hits given by clusters: adjacent pixels in the silicon pixel detector whose charge exceeds a pre-defined readout threshold
- clusters are counted for each jet and layer if they lie insiside a cone of fixed $\Delta R \equiv \sqrt{\Delta\eta^2 + \Delta\phi^2}$ around the jet axis
- different values for ΔR were tested: 0.04, 0.06, 0.08, 0.10, 0.16 (see later)

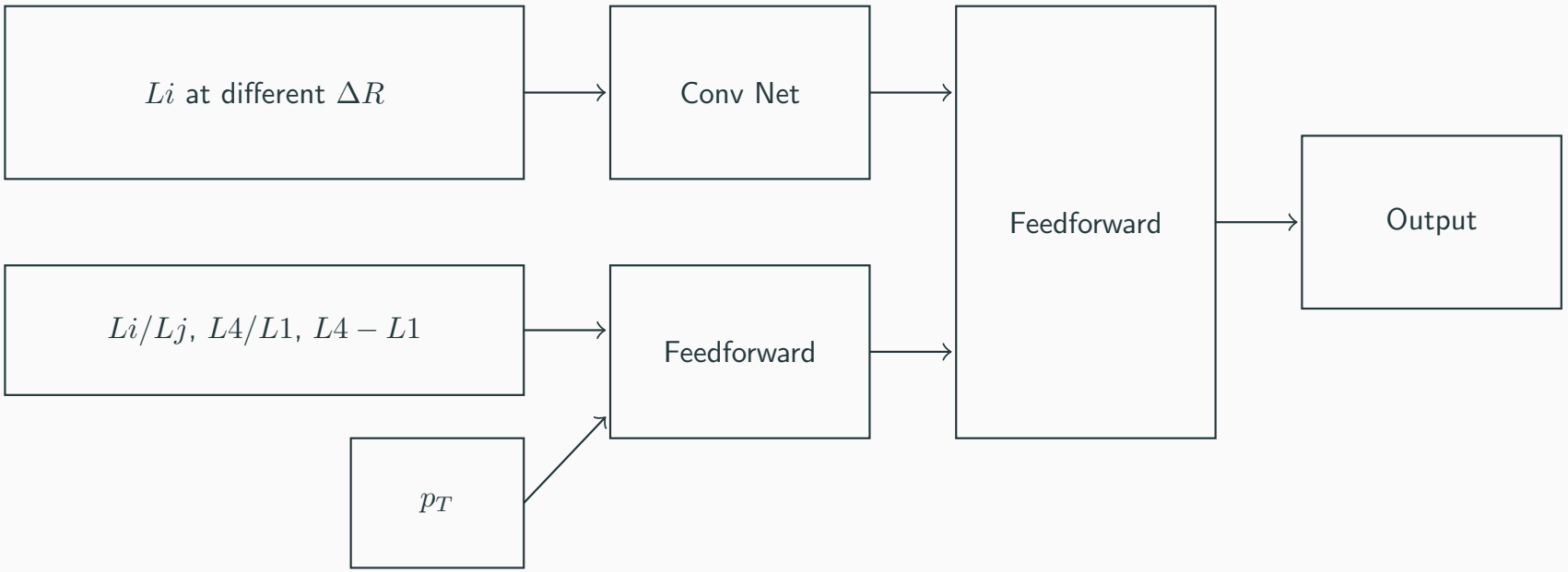


MANtag structure

Multiplicity-based Artificial Neural network tagger (MANtag)

input variables:

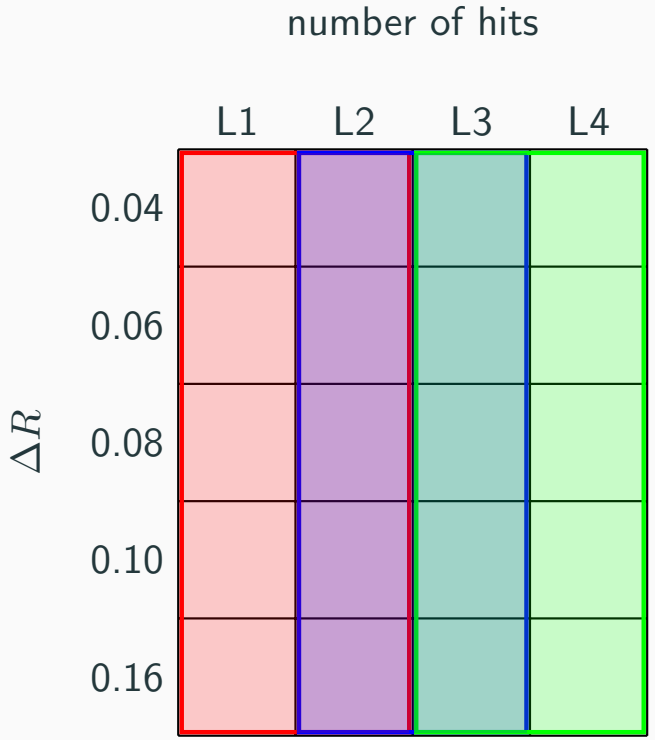
- number of hits in each Layer Li matched at 5 different ΔR
- ratio of hits for consecutive layer: $L2/L1$, $L3/L2$, $L4/L3$
- variables from previous discussion: $L4/L1$, $L4 - L1$
- p_T as input variable $\rightarrow$ need to reweight p_T profiles



Convolutional layer

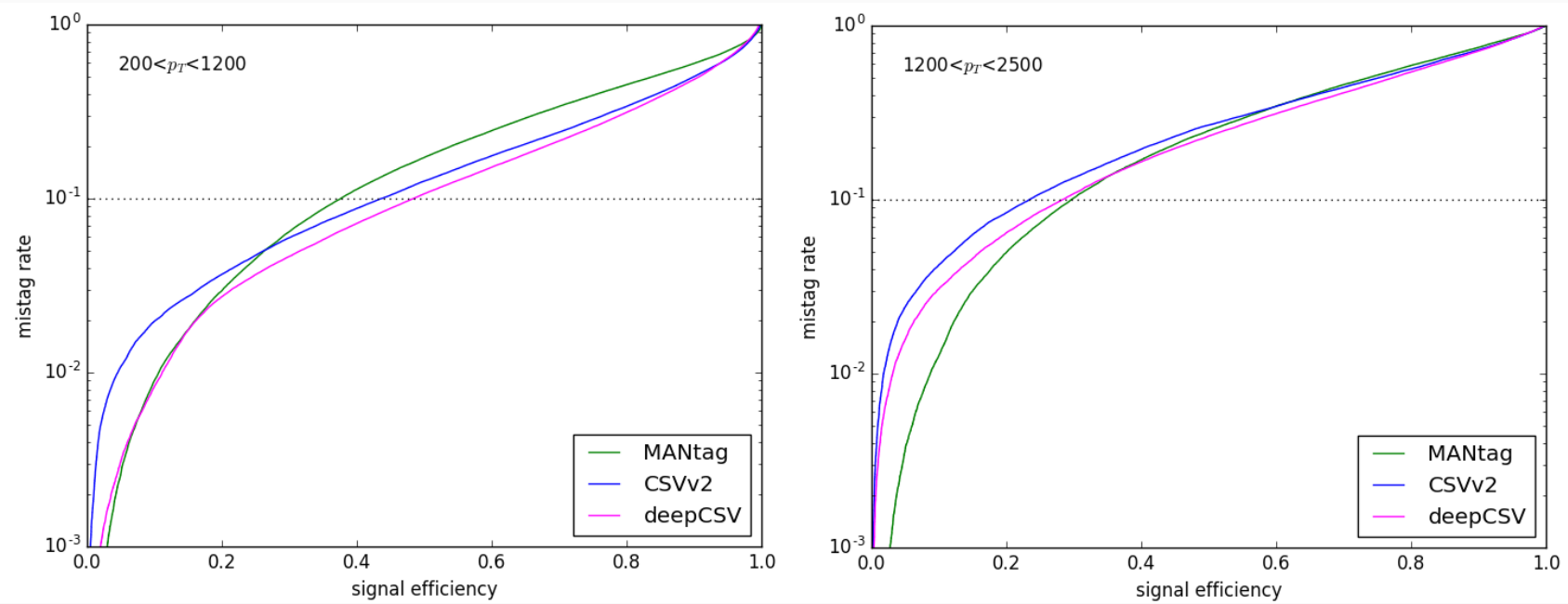
- Li and cone size arranged in a 5x4 matrix
- convolutional 5x2 filter sliding over input matrix
- take advantage of spatial structure

cluster matching



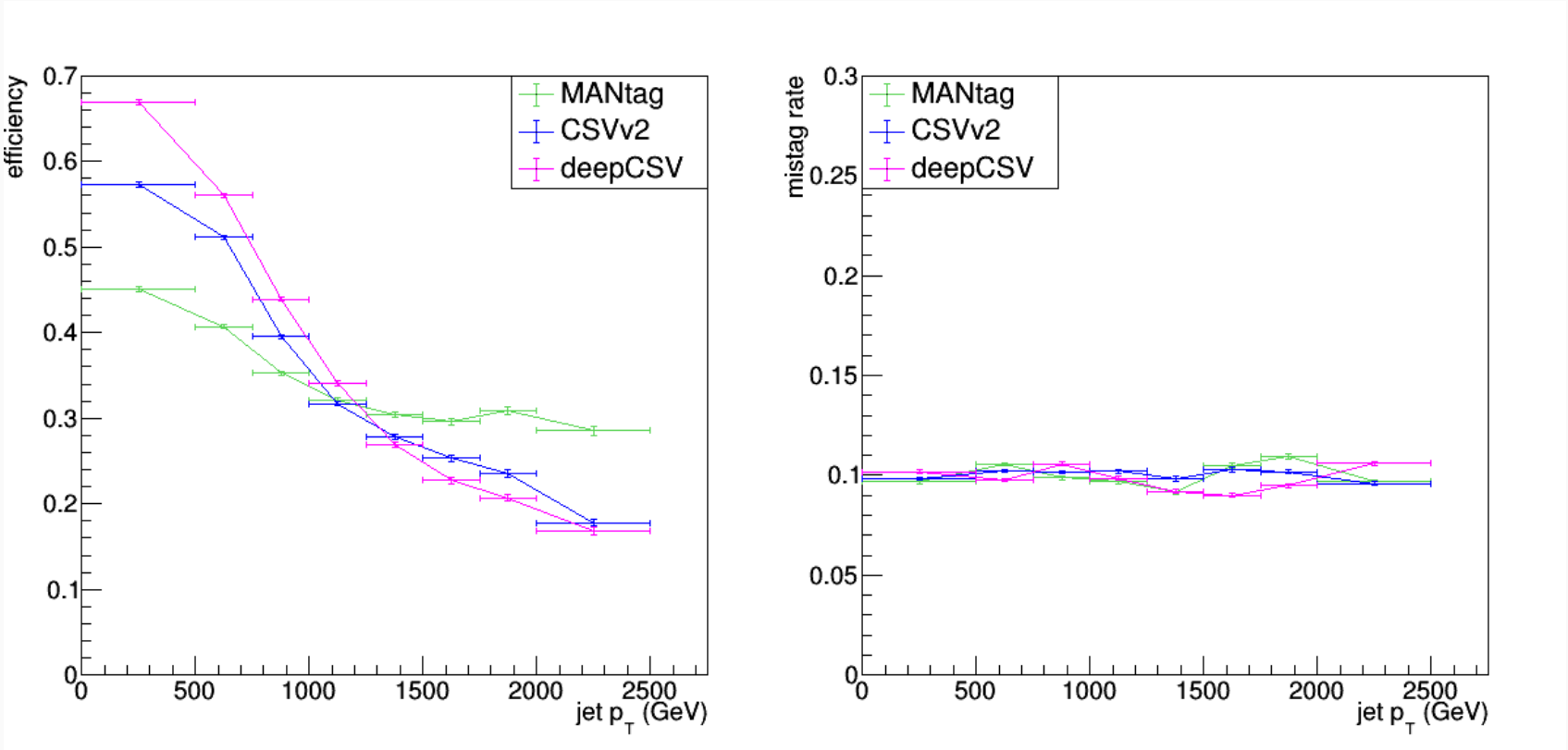
ROC curves

- hit-based approach more stable at higher p_T than the CSVs
- MANtag yields a higher efficiency than CSV at low mistag rates



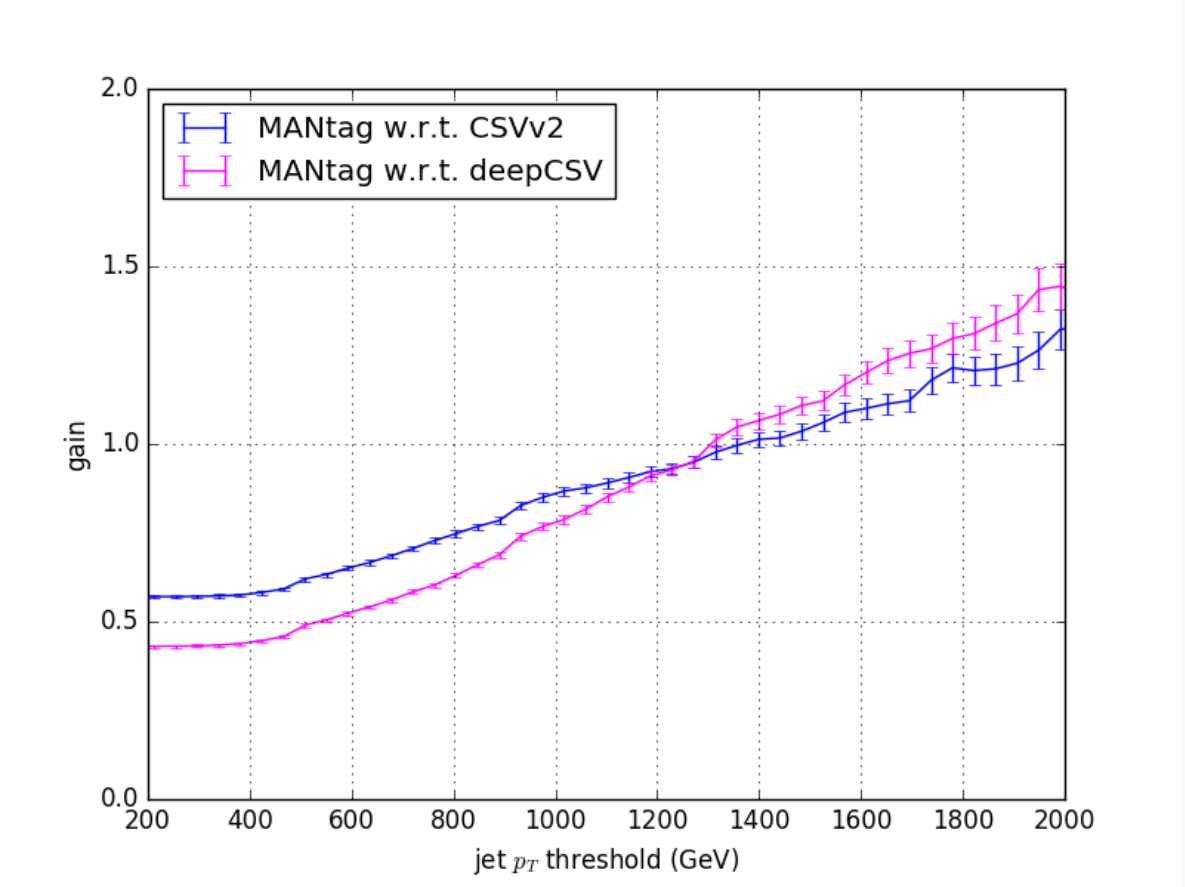
Efficiency vs pT

- At a flat 10% mistag rate working point, the MANtag efficiency exceeds the CSVs at $p_T > 1200$ GeV



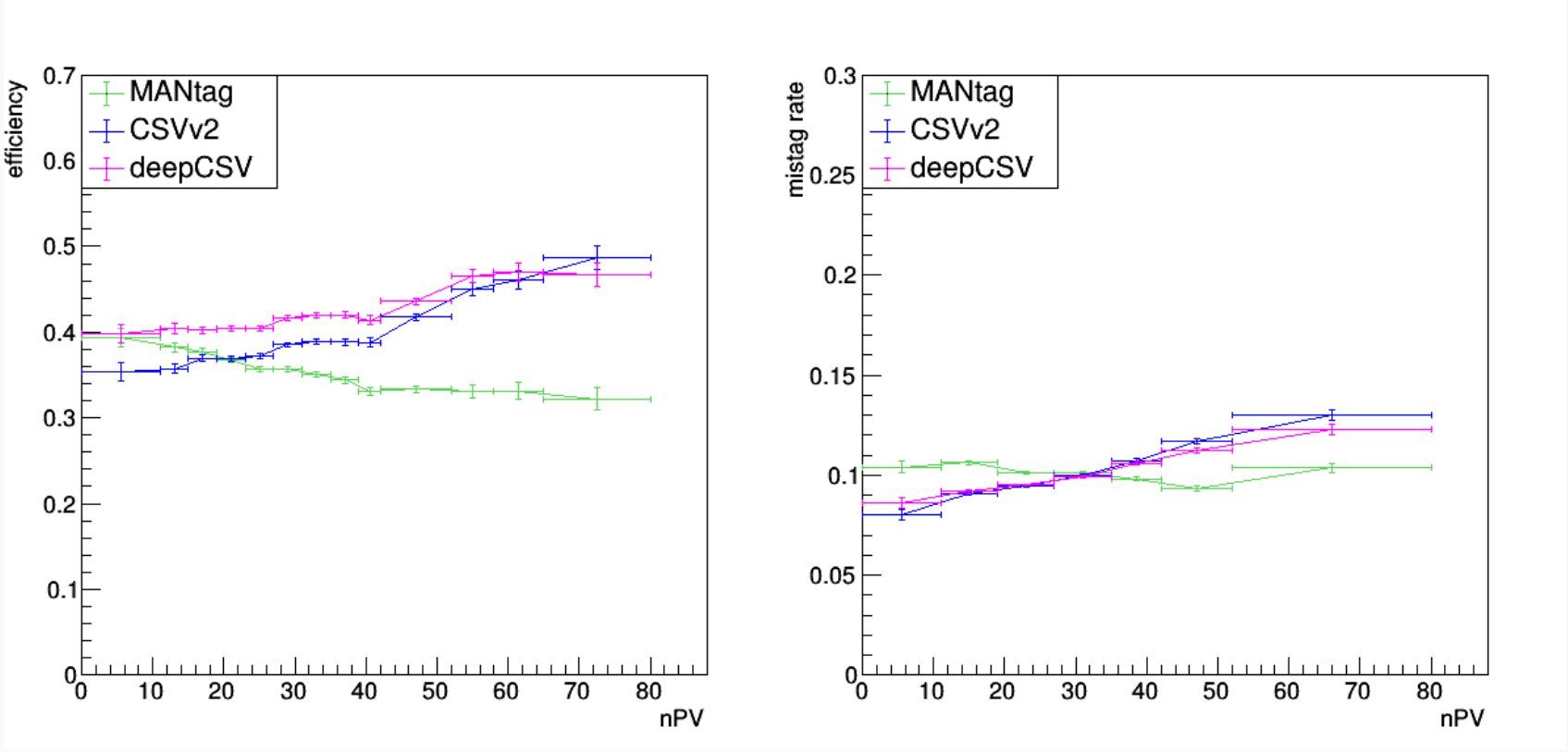
Relative gain

- relative gain = (additionally tagged jets) / (already tagged by CSV)
- at a flat 10% mistag rate, a significant relative gain could be reproduced
- gain with respect to deepCSV more strongly dependent on the p_T threshold



Stability with respect to pileup

- again considering the same flat 10% mistag rate working point
- efficiency and mistag rate of MANtag are relatively flat at an increasing number of primary vertices



Conclusion and outlook

Counting hits in a small angular region around the jet axis in each pixel detection layer...

- ...results in remarkably simple variables.
- ...yields a significant potential for improving b tagging at extreme p_T.
- ...has a stable absolute performance with respect to pileup.

Next steps:

- short-term: add the information to CMSSW for further studies with the Ultra-Legacy re-reco
- long-term: provide pixel hit multiplicity information to DeepJet if improvements are confirmed

read more at: *github.com/msommerh/bTag_HitCount*
or see: *github.com/cms-sw/cmssw/issues/27286*