Algorithms to identify high-energy B hadrons via their hit multiplicity increase through pixel detection layers

Manuel Sommerhalder

b tag meeting 19.11.18, CERN

Bachelor thesis supervised by
Prof. Dr. Ben Kilminster
Thea Årrestad
Dr. Yuta Takahashi

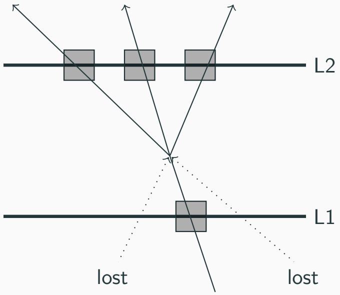
github.com/msommerh/bTag_HitCount

Motivation and method

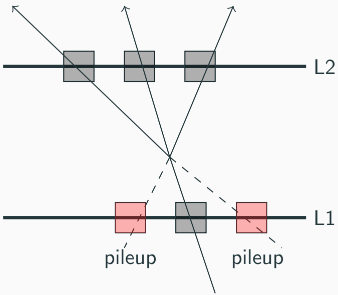
Motivation

- CSV(v2): track-based b tagging algorithm
- 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

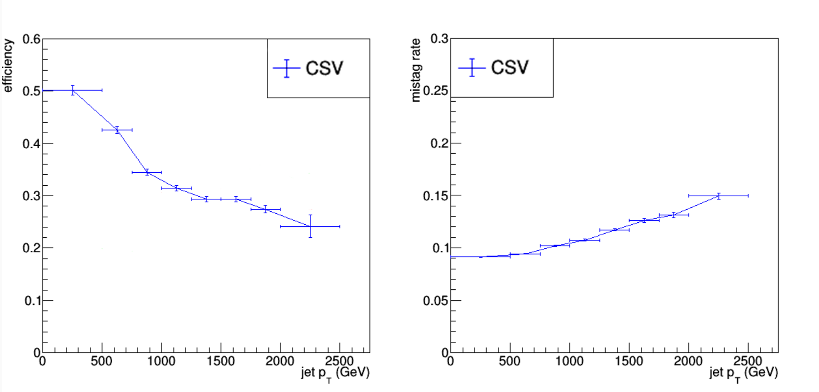


wrongly reconstructed tracks



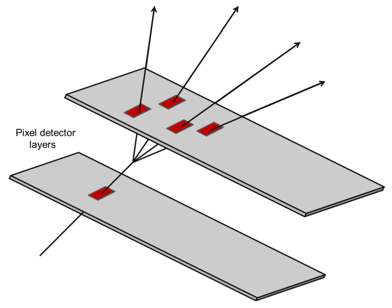
Motivation

- CSV(v2): track-based b tagging algorithm
- 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



Motivation

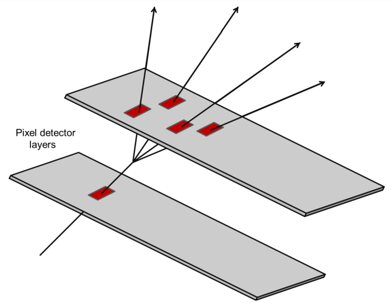
- alternative approach by B. Todd Huffman et al. (*arXiv:1604.05036* and *arXiv:1701.06832*)
- tagging high-p_T B hadrons based on an increase in hit multiplicity
- yields promising tagging efficiency on a DELPHES simulation



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 in addition to CSV leads to a gain in tagging efficiency
4. implement an MVA-based b tagger for a fair comparison to CSV



arXiv:1604.05036

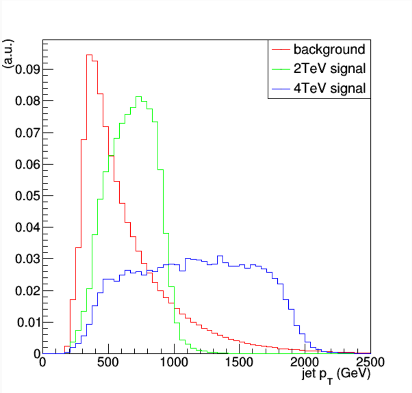
Method

1. generate high-p_T b jets ($p_T > 500$ GeV)
2. match hits in each layer to the jets
3. construct hit-based variables (ratios and differences between the number of hits in different layers)
4. compare the variables of b jets to those of generic QCD jets
5. develop discriminants to distinguish signal from background

Mote Carlo samples

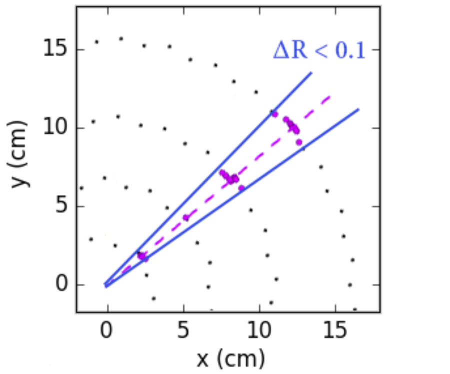
Monte Carlo samples

- signal: $Z' \to b\bar{b}$ with $M_{Z'} = 2$ TeV and $M_{Z'} = 4$ TeV, generated in PYTHIA 8
- background: *QCD_Pt-15to7000_TuneCUETP8M1_Flat_13TeV_pythia8 (92X)*
- momentum threshold of 350 GeV imposed on outgoing particles



Pixel cluster matching

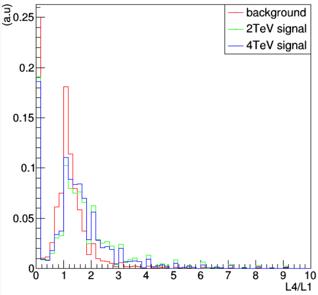
- hits given by clusters: adjacent pixels in the silicon pixel detector whose charge exceeds a pre-defned readout threshold
- clusters are counted for each jet and layer if they lie insiside a cone of fixed $\Delta R \equiv \sqrt{\Delta\eta^2 + \Delta\phi^2}$ around the jet axis
- different values for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

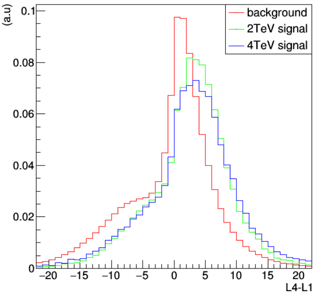


Cut-based discrimination using hit multiplicity variables

Discriminant and cone optimizations

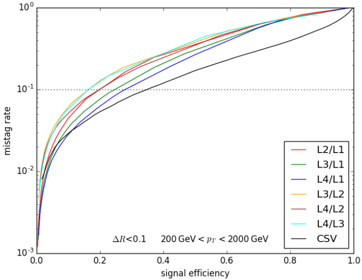
- Li: number of hits in layer $i \in (1, 2, 3, 4)$
- histograms for every combination of Li/Lj and $Li - Lj$ with $i > j$
- evaluated on every cone size $\Delta R =$ 0.04, 0.06, 0.08, 0.10, 0.16

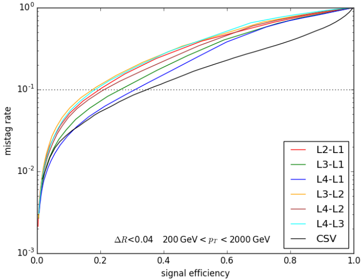




Discriminant and cone optimizations

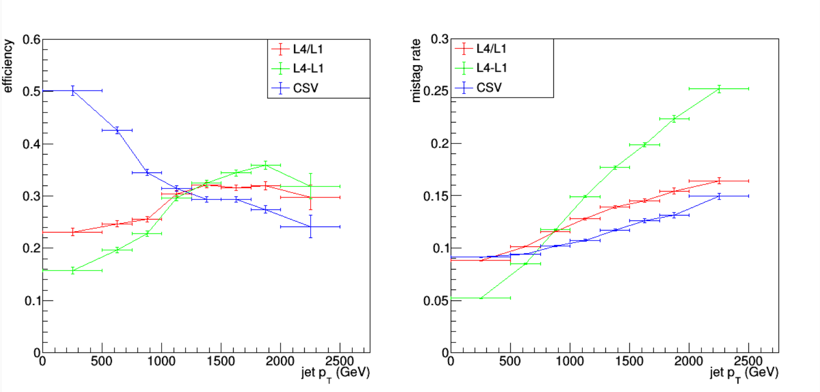
- compare performance of each discriminant at each ΔR
- mistag rate: fraction of generic QCD jets misidentified as b jets
- highest performance for $L4/L1$ with $\Delta R < 0.1$ and $L4 - L1$ with $\Delta R < 0.04$
- $L4/L1$ and $L4 - L1$ comparable to CSV at a 10% mistag rate





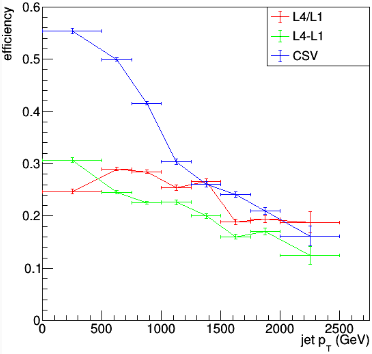
Single cut discriminants

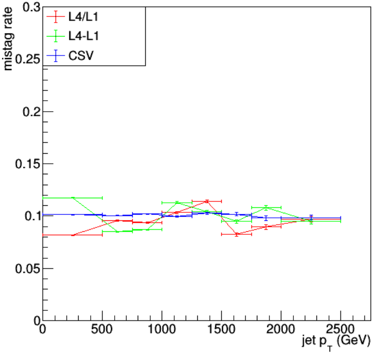
- cuts corresponding to a 10% mistag rate
- efficiency and mistag rate yield a high dependence on p_T
- no direct comparison can be made for specific points on the p_T spectrum



pt dependent cuts

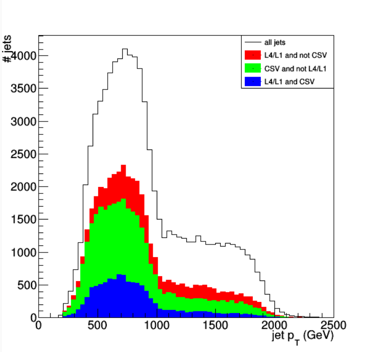
- vary cuts with p_T to achieve a flat 10% mistag rate
- comparable efficiency of CSV and $L4/L1$ at $p_T > 1200$ GeV

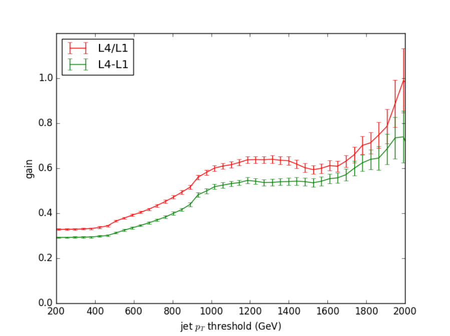




Comparison to CSV

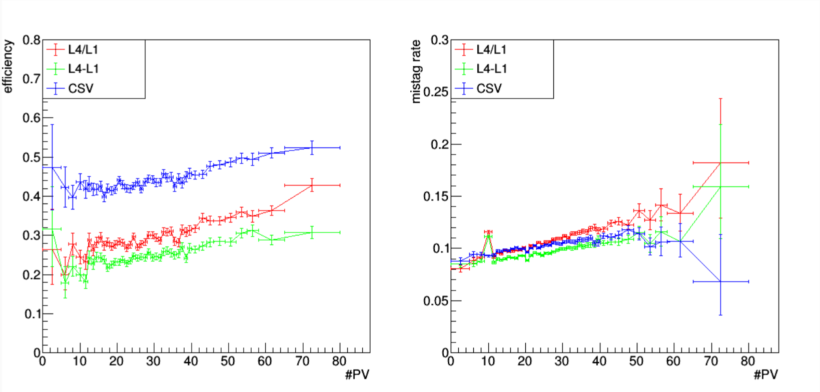
- LHS: tagged jets as a function of p_T at a flat 10% mistag rate
- RHS: relative gain when using each tagger in addition to CSV
- relative gain: $\frac{r}{g+b}$
- $L4/L1$ yields a 32% gain on the full spectrum, 63% above 1200 GeV





Stability with respect to pileup

- analogous analysis done on samples with pileup
- expected PU in 2017 (20-30 PV)
- absolute performance of both taggers stable and similar to CSV



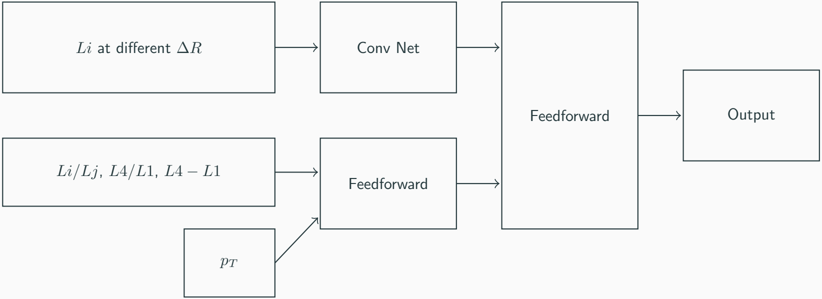
Multiplicity-based Artificial Neural network tagger: MANtag

MANtag structure

input variables:

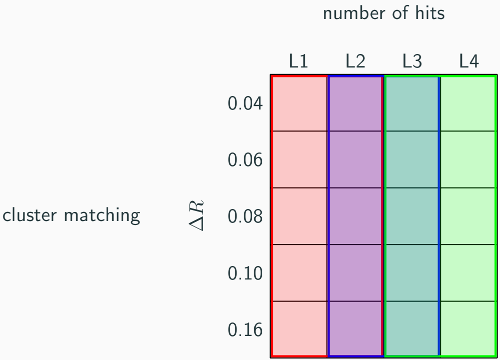
- number of hits in each Layer Li matched at 5 different ΔR
- ratio of hits for consecutive layer: $L2/L1$, $L3/L2$, $L4/L3$
- variables from previous discussion: $L4/L1$, $L4 - L1$
- p_T as input variable $\rightarrow$ need to reweight p_T profiles

aim: construct an MVA-based discriminant for a fair comparison to CSV

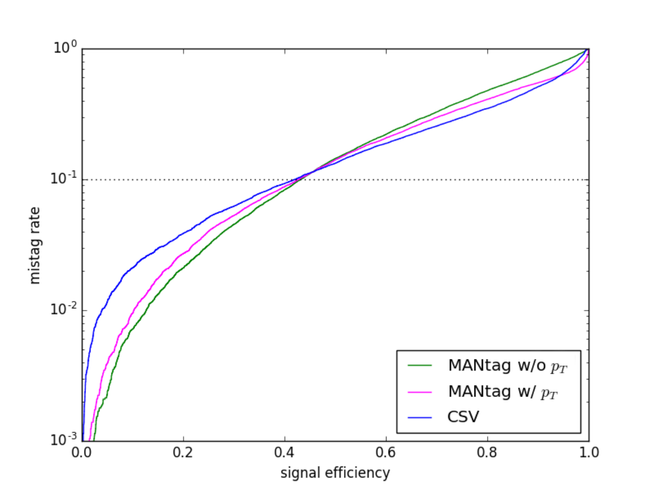


Convolutional layer

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

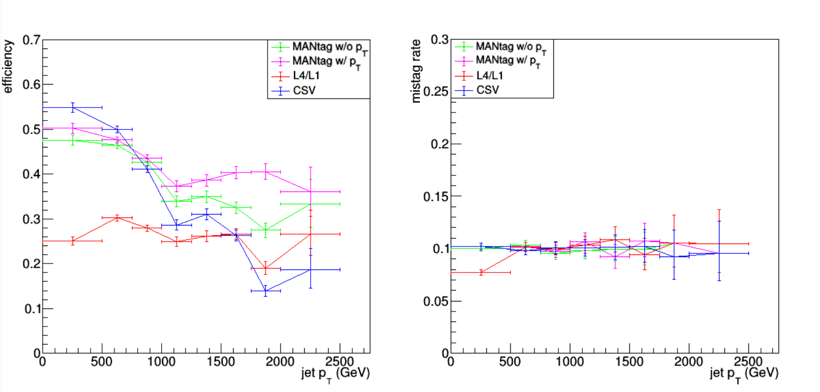


MANtag performance



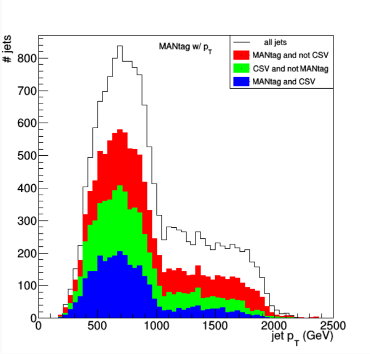
Efficiency and mistag rate

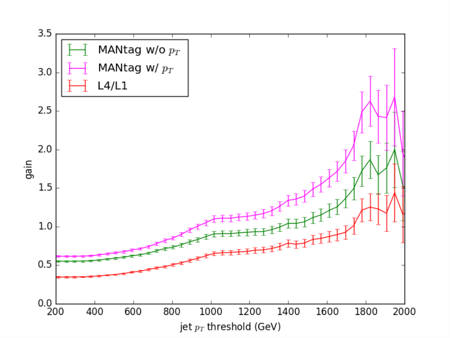
- superior performance of both neural networks at $p_T > 1200$ GeV
- higher efficiency of MANtag using p_T as input variable



Comparison to CSV

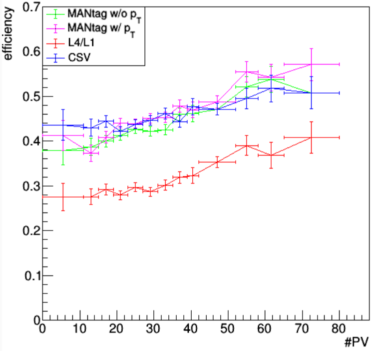
- relative gain to CSV at a 10% mistag rate
- MANtag with p_T variable yields a 60% gain on the full spectrum, 112% above 1200 GeV

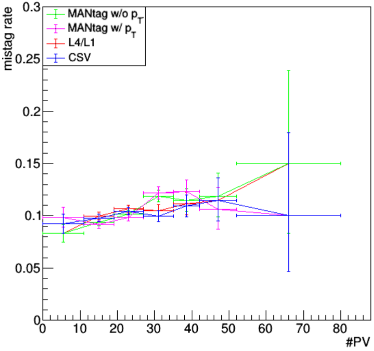




Stability with respect to pileup

- efficiency and mistag rate increasing as a function of PV for all taggers
- absolute performance of all taggers stable and similar to CSV





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:

- Implement a hit-based standalone tagger?
- Integrate hit-based variables into CSV?

read more at: *github.com/msommerh/bTag_HitCount*

Appendix

CSV

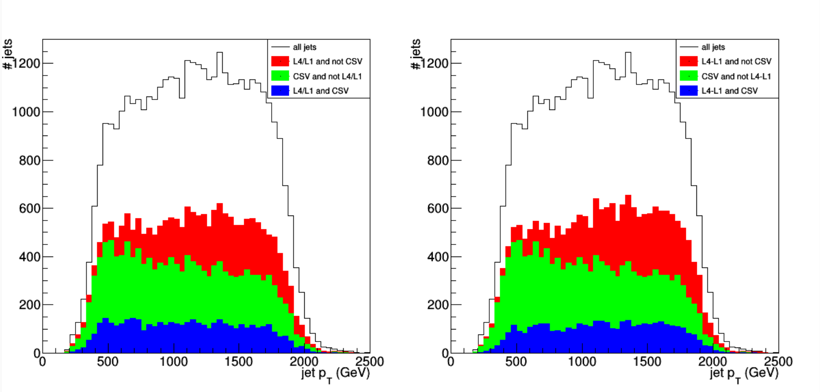
- Combined Secondary Vertex Tagger
- tracks reconstructed using an iterative procedure with a Kalman filter
- algorithms AVR and IVF reconstruct secondary vertices

Input variable	Run 1 CSV	CSVv2
SV 2D flight distance significance	x	x
Number of SV	—	x
Track η_{rel}	x	x
Corrected SV mass	x	x
Number of tracks from SV	x	x
SV energy ratio	x	x
ΔR(SV, jet)	—	x
3D IP significance of the first four tracks	x	x
Track $p_{\mathrm{T,rel}}$	—	x
ΔR(track, jet)	—	x
Track $p_{\mathrm{T,rel}}$ ratio	—	x
Track distance	—	x
Track decay length	—	x
Summed tracks E_{T} ratio	—	x
ΔR(summed tracks, jet)	—	x
First track 2D IP significance above c threshold	—	x
Number of selected tracks	—	x
Jet p_{T}	—	x
Jet η	—	x

arXiv:1712.07158

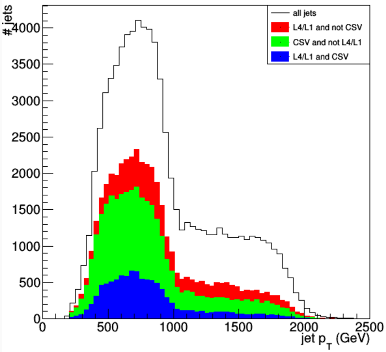
Single cut discriminants

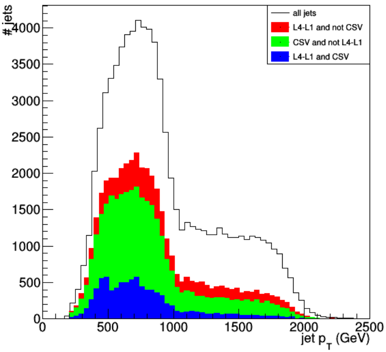
- number of jets correctly tagged by CSV and each tagger at a 10% mistag rate corresponding to a single cut
- red area corresponds to gain when using each tagger in addition to CSV



pT dependent cuts

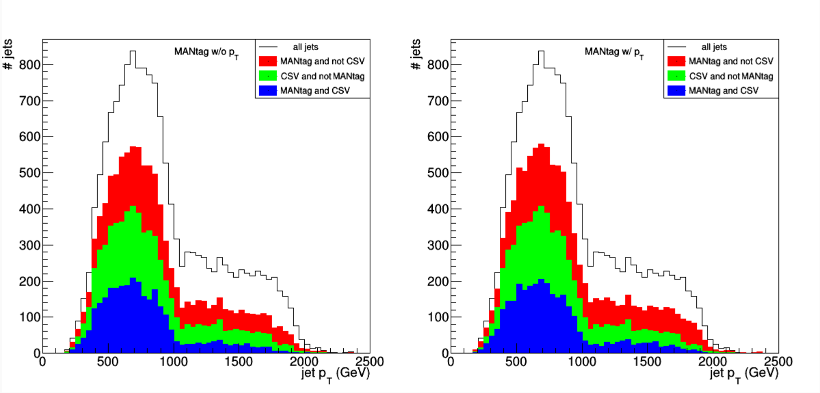
- number of jets correctly tagged by CSV and each tagger at a flat 10% mistag rate over the entire p_T spectrum
- red area corresponds to gain when using each tagger in addition to CSV





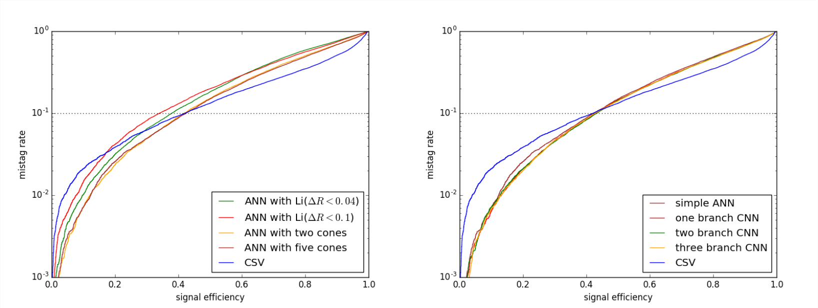
MANtag performance

- number of jets correctly tagged by CSV and MANtag without p_T (LHS) and with p_T (RHS) at a flat 10% mistag rate
- red area corresponds to gain when using each tagger in addition to CSV



Choice of artificial neural network model

- LHS: densely connected network with different cone size inputs
- RHS: 5 cone size inputs processed through ANNs of different complexity



Pileup profiles