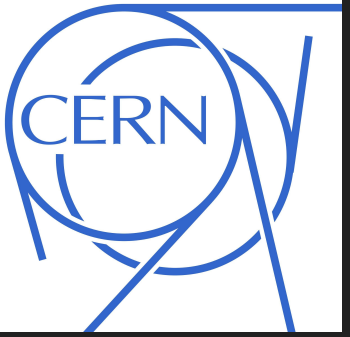


HF Tagging in CMS

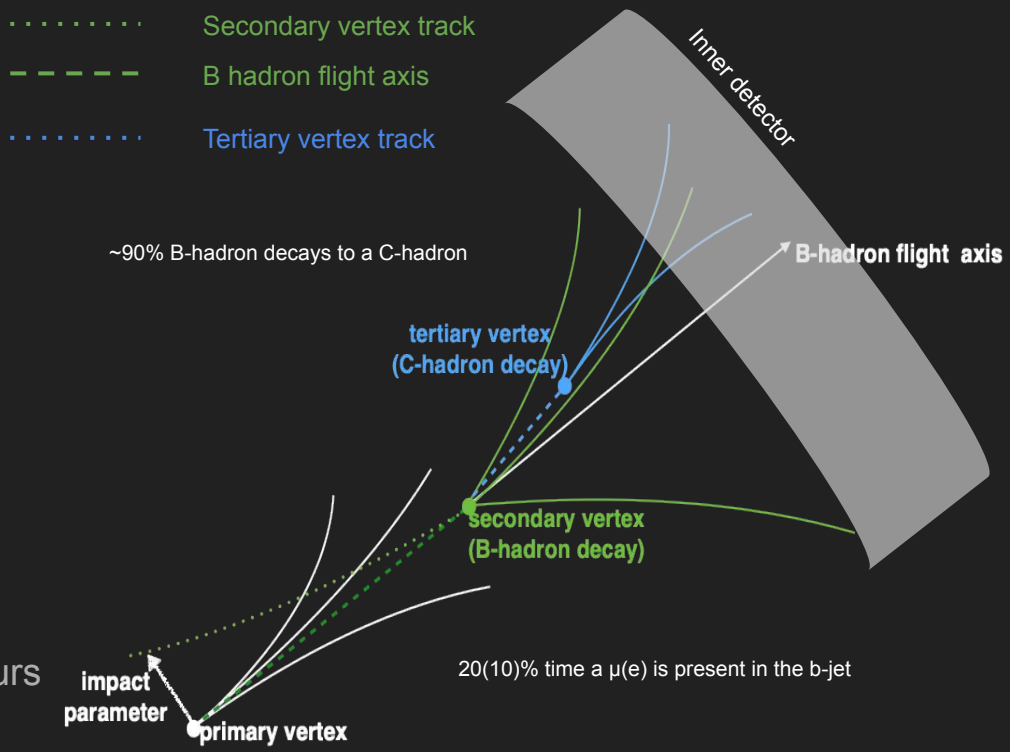
Joshuha Thomas-Wilsker, IHEP

What is b-tagging?

B-tagging is the identification of jets that originated from the hadronisation of b-quarks (b-jets).

Relies on unique properties of b-hadrons:

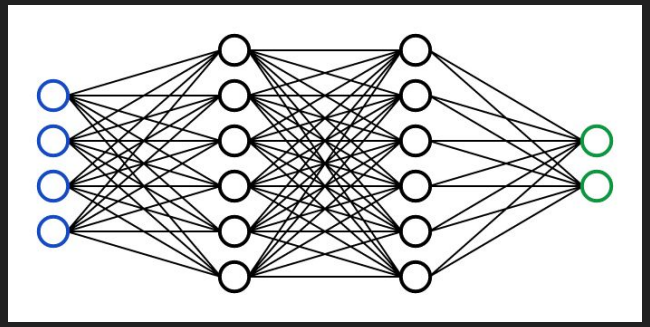
- Lifetime: large decay length(~5mm).
- Largest mass of any hadron: large decay product multiplicity (~5 charged particles per decay).
- Fragmentation: harder than other quark flavours (B hadrons carry ~75% of jet energy).



Important when searching/measuring many interesting physics processes as it is a distinguishable and frequent decay product of several large particle e.g. Higgs boson, top quarks etc.

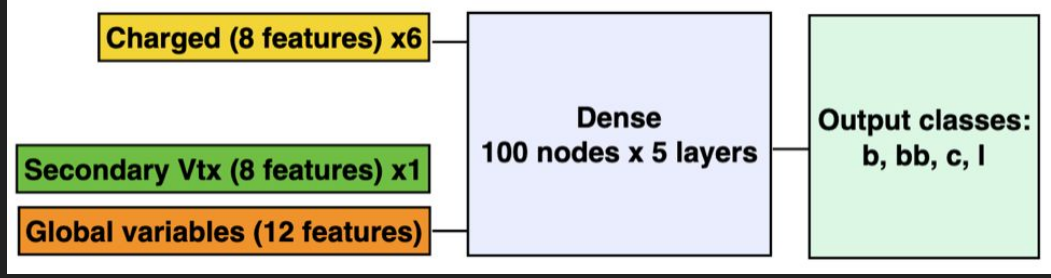
The AK4 b-jet ID Algorithms

Multi-classification deep neural network taggers to ‘tag’ b-jets.

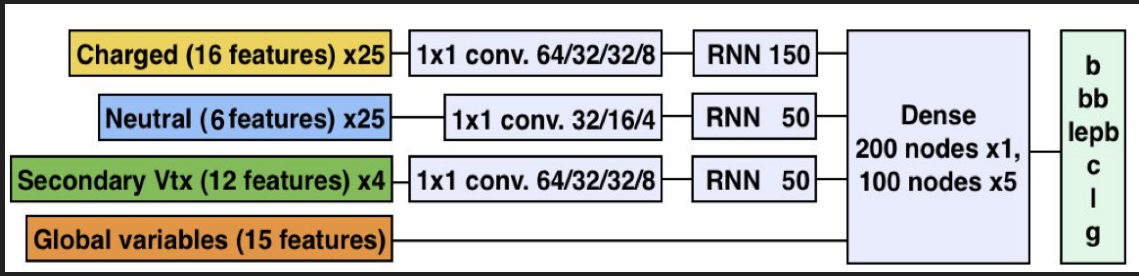


Discriminator values defined according to (combination of) probabilities on output nodes.

DeepCSV



DeepJet



	DeepCSV	DeepJet
b-jets	$\mathrm{P(b)} + \mathrm{P(bb)}$	$\mathrm{P(b)} = \mathrm{score(b)} + \mathrm{score(bb)} + \mathrm{score(lepb)}$
c-jets	$\mathrm{CvsB} = \frac{\mathrm{P(C)}}{\mathrm{P(C)} + \mathrm{P(B)}}$ $\mathrm{CvsL} = \frac{\mathrm{P(C)}}{\mathrm{P(C)} + \mathrm{P(L)}}$	$\mathrm{CvsB} = \frac{\mathrm{P(C)}}{\mathrm{P(C)} + \mathrm{P(B)}}$ $\mathrm{CvsL} = \frac{\mathrm{P(C)}}{\mathrm{P(C)} + \mathrm{P(L)}}$
light jets	$\mathrm{P(udsg)}$	$\mathrm{P(uds)} + \mathrm{P(g)}$

Workflow

- Very fast check of new production.
- Data/MC comparison of important variables in ~scale-factor phase space (see later).
- Variables checked include track parameters, secondary vertex information, lepton momentum.

Workflow

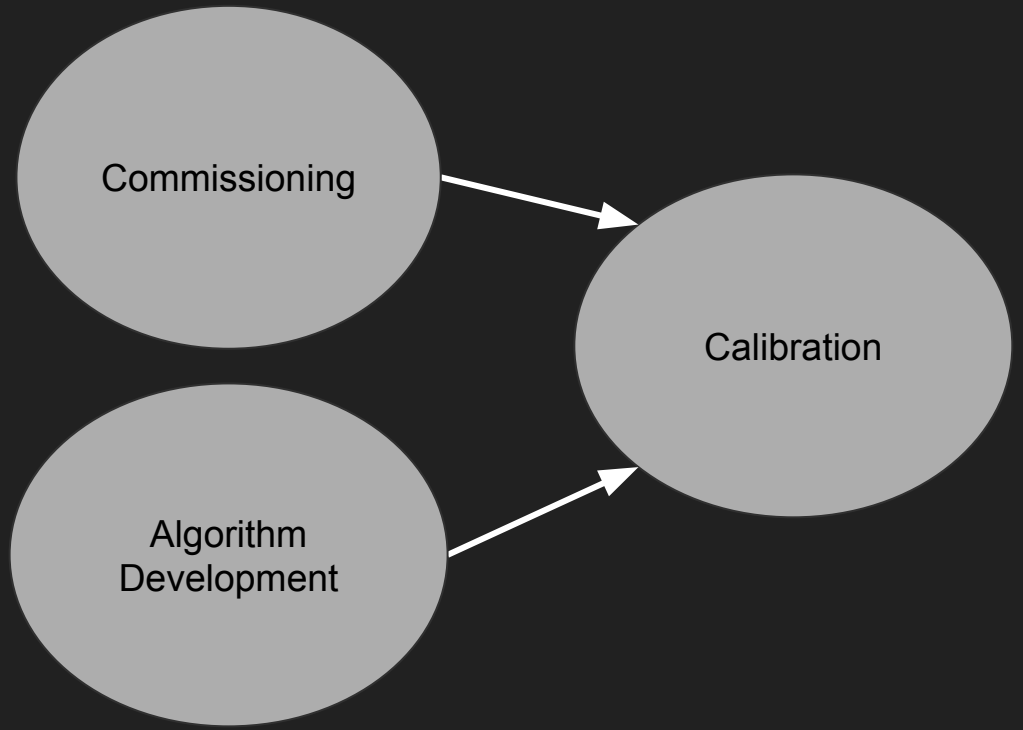
Commissioning

- Very fast check of new production.
- Data/MC comparison of important variables in ~scale-factor phase space (see later).
- Variables checked include track parameters, secondary vertex information, lepton momentum.

Algorithm Development

- Inception of tagging algorithm + proof of concept.
- Research & development: prove it’s realistic, produce datasets etc.
- Implementation e.g. CMSSW using Light Weight Trained Neural Network (LWTNN).
- Dissemination: paper, fast release mechanism (e.g. CMS DP note).

Workflow



- Once commissioning distributions are understood we calibrate approved algorithms.
- Necessary as performance of algorithms on data and MC differs.
- Correct the efficiency on MC to match data as function of p_T & η.
- Fixed working point SF = combination of several calibration methods in different topologies.
- Can also correct as function of tagger output = discriminator reshaping

Calibration

Fixed WP
b-tag calibration:
tt, μ-enriched

c-tag calibration:
tt, W+C

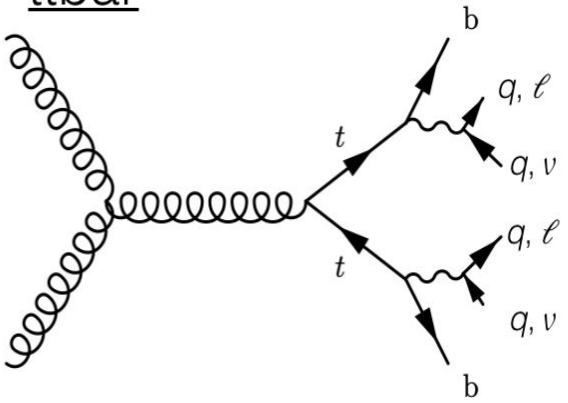
Mis-tag calibration:
QCD

Reshaping
b-tag calibration:
b-enriched=tt dilepton
light-enriched=Z+jets

c-tag calibration:
c-enriched=W+c
b-enriched=tt(DL/SL)
Z+jets=light-enriched

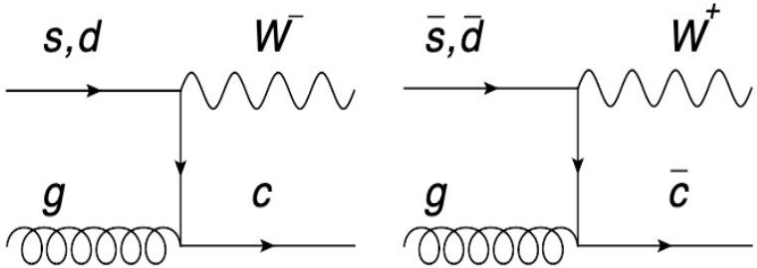
ttbar

t->bW branching ratio > 95%.

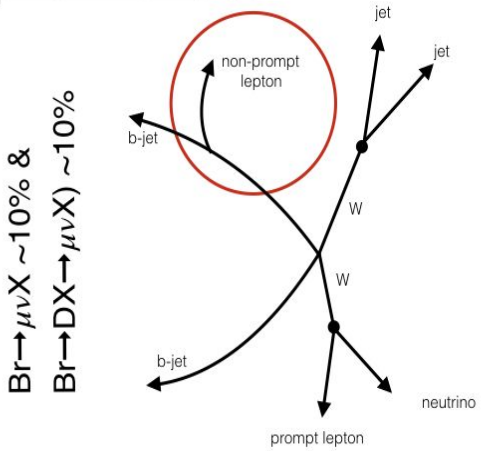


W+c

W+c always opposite sign (unlike backgrounds)

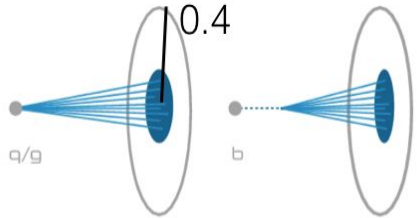


μ-enriched



QCD multijet

Dominated by light jets.



SF Derivation Example: Kinematic Fit Method

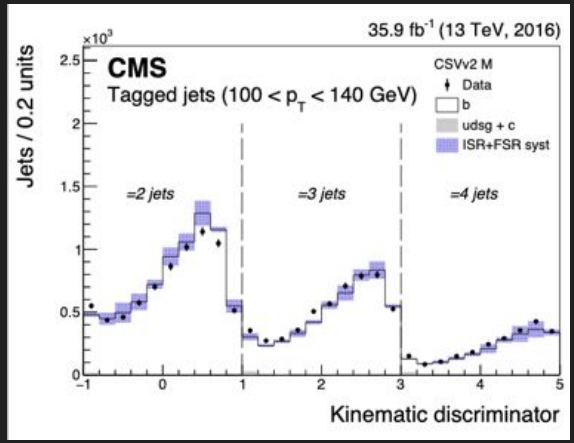
ttbar dilepton selection.

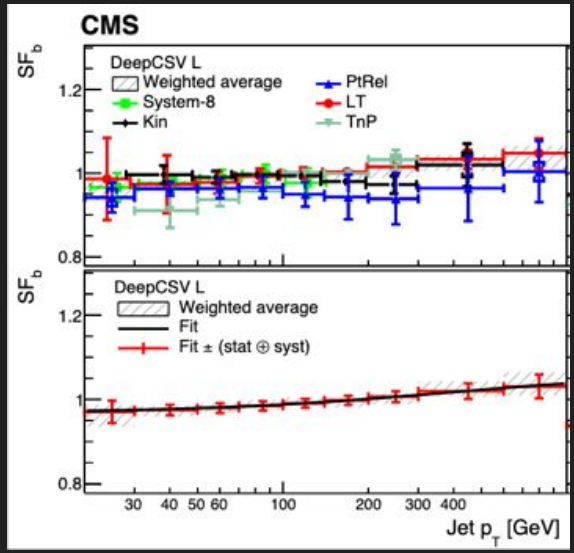
Train BDT using kinematic variables only, to discriminate jets ttbar system from ISR/FSR jets.

A binned likelihood fit to data of the kinematic discriminator distribution.

SFb extracted from fit - only free parameter.

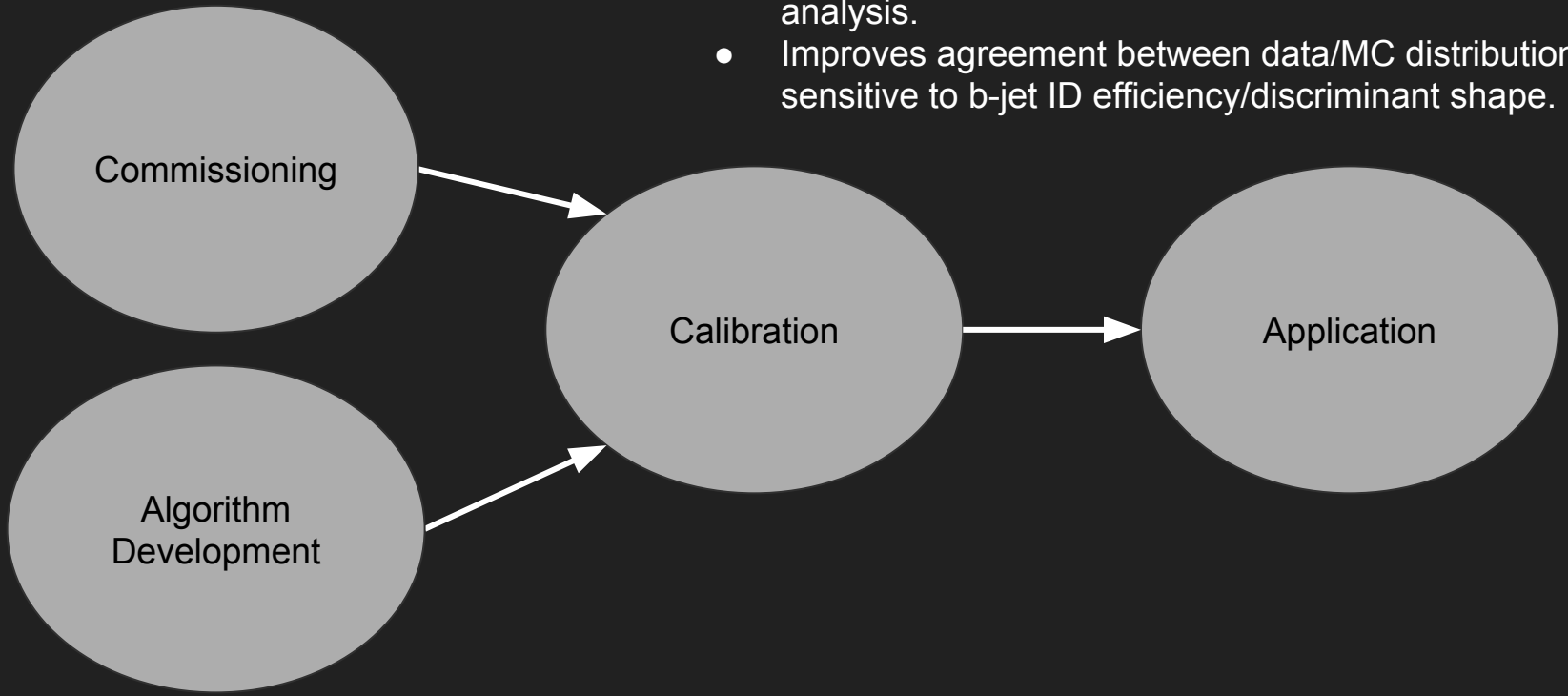
Dominant systematic = ISR/FSR uncertainty in ttbar simulation.





Workflow

- Analyses use output distribution of b-tagging algorithms for selection and/or fitting shape.
- Calculate event weights using SFs which are applied in analysis.
- Improves agreement between data/MC distributions sensitive to b-jet ID efficiency/discriminant shape.



How do we use it?

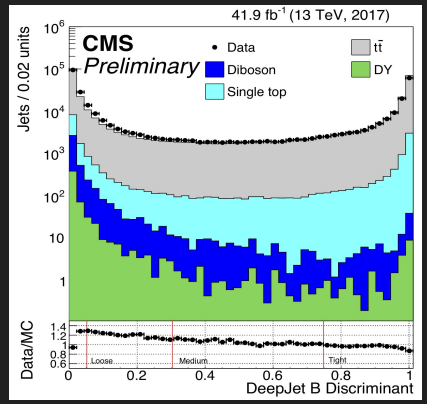
Fixed working point (based on mis-tag efficiencies of 0.1, 1 and 10%):

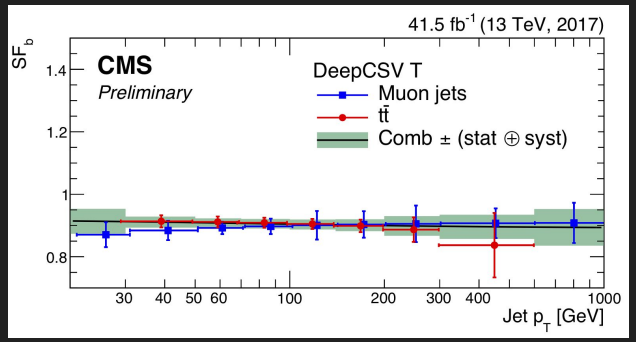
$$SF_{f,p_T,\eta} = \frac{\epsilon^{Data}_{f,p_T,\eta}}{\epsilon^{MC}_{f,p_T,\eta}}$$

$$P(MC) = \prod_{i=tagged} \epsilon_i \prod_{j=not-tagged} (1-\epsilon_j)$$

$$P(Data) = \prod_{i=tagged} SF_i\epsilon_i \prod_{j=not-tagged} (1-SF_j\epsilon_j)$$

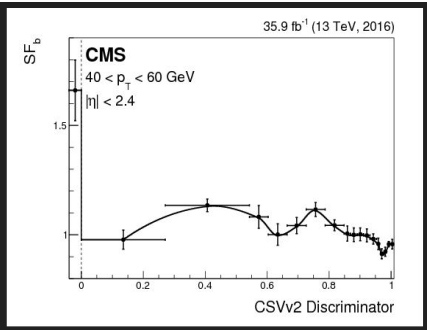
$$\rightarrow \quad w = \frac{P(Data)}{P(MC)}$$





Or reweight the entire distribution as a function of the discriminator:

$$w_{event} = \prod_{i}^{N_{jets}} SF(D_i, p_{Ti}, \eta_i)$$



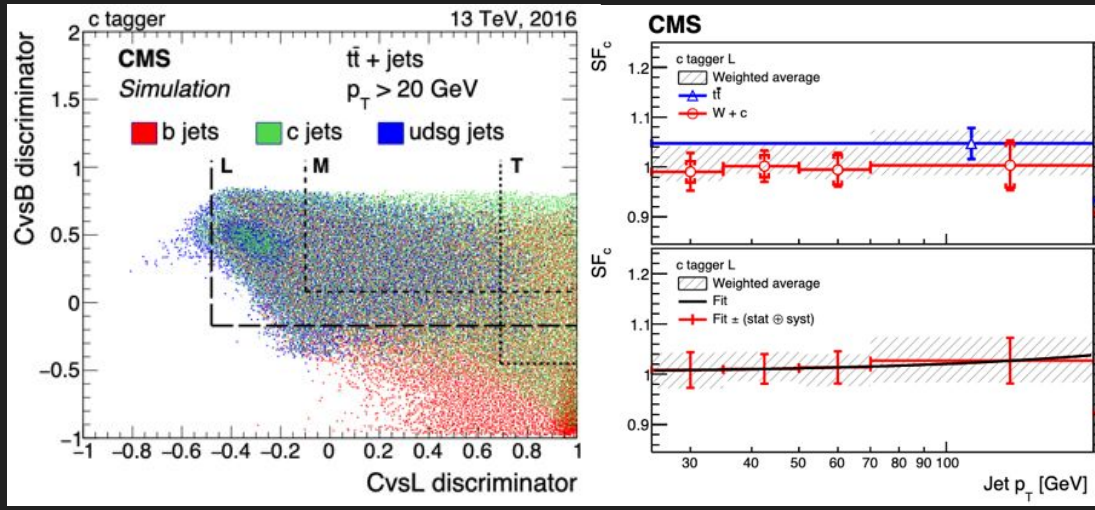
c-tagging

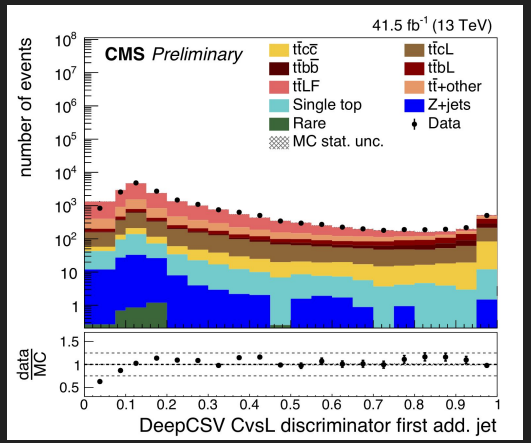
Using DNN model with the same training but different combination of probabilities.

c-jet WP defined simultaneously on CvsL & CvsB:

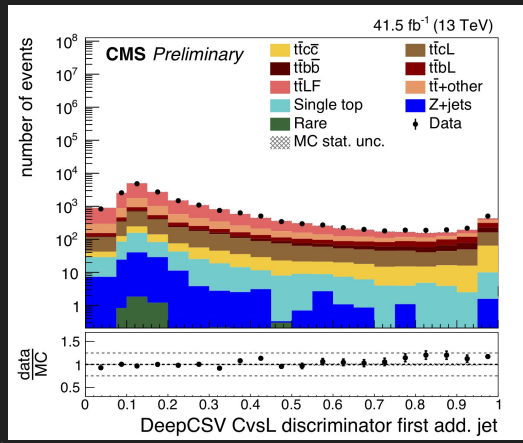
- Loose WP focus = discriminate C from B
- Tight WP focus = rejecting light
- Medium WP = trade off

Shape calibration also exists - successfully applied in recent tt+cc XS measurement! [link]





Apply SF



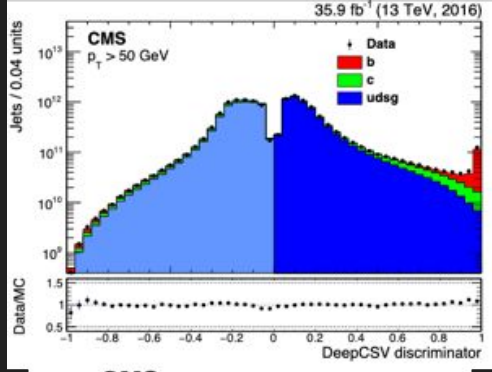
Mis-tag Scale Factors

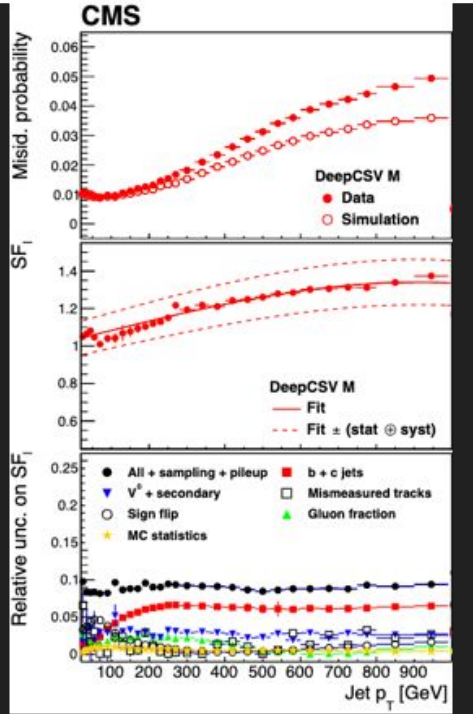
- Negative-tag Method.
- Inclusive multi-jet events.
- Uses same DNN model but only with +ve(-ve) impact parameter values and SV's with +ve(-ve) flight distance.
- Distributions should be approximately symmetric for light-jets because non-zero values of IP / SV flight distance mostly c.f. resolution effects.
- Mis-ID probability = $\epsilon_l = \epsilon^- R_{light}$

Fraction of -ve tagged jets passing WP in inclusive multi-jet sample

$$R_{light} = \frac{\epsilon_l^{MC}}{\epsilon^{-,MC}}$$

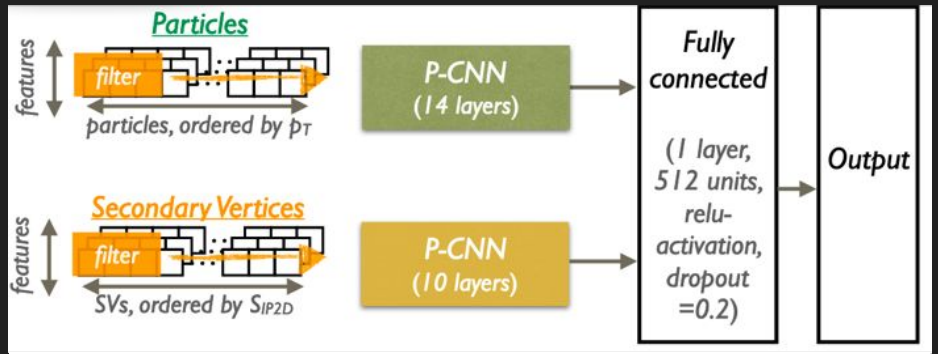
Ratio of mis-id probability for light-jets w.r.t. the negative tagging efficiency of all jets in MC





Boosted AK8 Taggers

- Radius 0.8 jets.
- Taggers: DeepDoubleX/DeepAK8.
- Focus on bb and cc tagging.
- Iterative fit method used to extract SF using 3 templates: bb, cc and udsg.

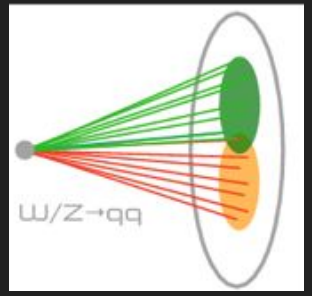


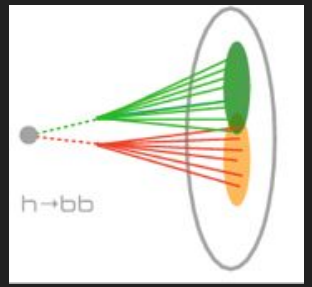
bb vs. QCD:

$$\text{DeepAK8-bb} = \frac{\text{score}(Z \to b\bar{b}) + \text{score}(H \to b\bar{b})}{\text{score}(Z \to b\bar{b}) + \text{score}(H \to b\bar{b}) + \text{score}(QCD)}$$

cc vs. QCD:

$$\text{DeepAK8-cc} = \frac{\text{score}(Z \to c\bar{c}) + \text{score}(H \to c\bar{c})}{\text{score}(Z \to c\bar{c}) + \text{score}(H \to c\bar{c}) + \text{score}(QCD)}$$





Output

Category	Label
Higgs	H (bb)
	H (cc)
	H (VV*→qqqq)
Top	top (bcq)
	top (bqq)
	top (bc)
	top (bq)
W	W (cq)
	W (qq)
Z	Z (bb)
	Z (cc)
	Z (qq)
QCD	QCD (bb)
	QCD (cc)
	QCD (b)
	QCD (c)
	QCD (others)

Recent Developments: ParticleNet

New fat-jet tagger being established.

Graph convolutional neural network for jet tagging.

Uses unordered set of jets constituent particles as input.

Improvement over currently used DeepAK8 (especially mass-decorrelated versions).

MD: Mass decorrelated = trained on artificial sample with flat mass spectrum for signal.

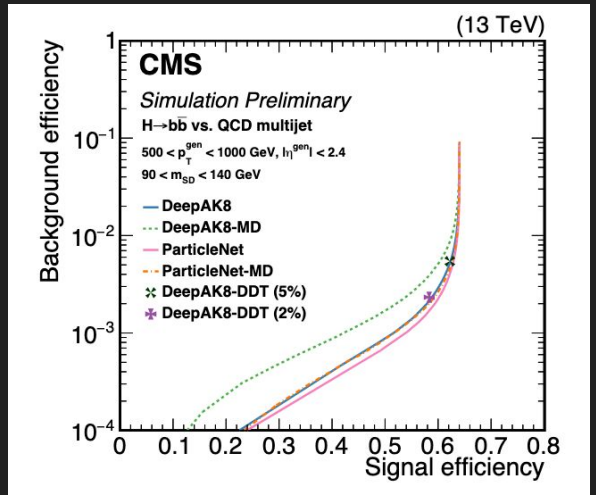
DDT: Designing Decorrelated Taggers [link]

Jet Tagging via Particle Clouds

Huilin Qu*
Department of Physics, University of California, Santa Barbara, California 93106, USA

Loukas Gouskos†
CERN, CH-1211 Geneva 23, Switzerland

How to represent a jet is at the core of machine learning on jet physics. Inspired by the notion of point clouds, we propose a new approach that considers a jet as an unordered set of its constituent particles, effectively a "particle cloud". Such a particle cloud representation of jets is efficient in incorporating raw information of jets and also explicitly respects the permutation symmetry. Based on the particle cloud representation, we propose ParticleNet, a customized neural network architecture using Dynamic Graph Convolutional Neural Network for jet tagging problems. The ParticleNet architecture achieves state-of-the-art performance on two representative jet tagging benchmarks and is improved significantly over existing methods.



Recent Developments: Deep Scale Factors

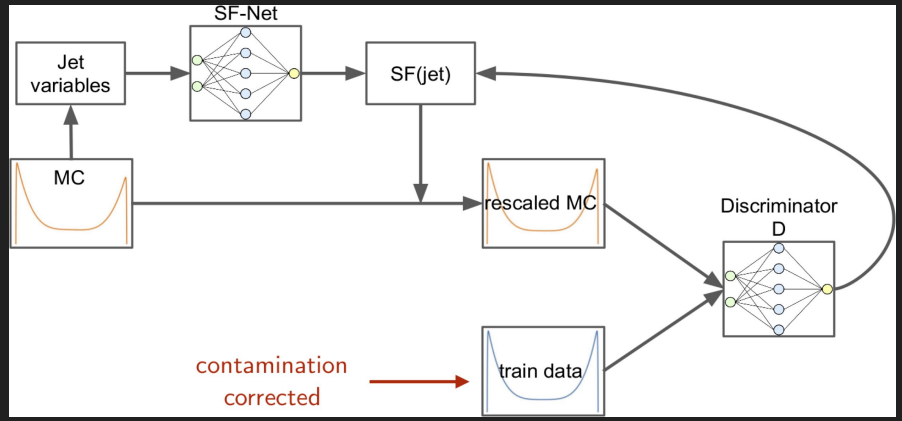
Variable dependent SF's to improve data/MC shape agreement of taggers.

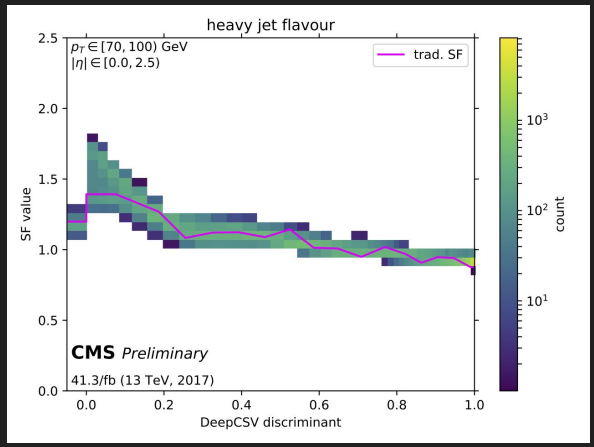
Underlying model often requires tuning e.g. binning, fit function.

Primary network uses jet variables to derive per-jet SF.

Adversary aims to discriminate between data and rescaled MC.

Comparison with iterative fit method which gives a single value for all jets in a given bin.





Summary

Seen an overview of the activities from the CMS HF tagging group.

Jet tagging algorithms utilise advanced in machine learning techniques.

Taggers have become incredibly versatile.

Commissioning of inputs and calibration of taggers is essential for usage in physics analyses - many well established methods mixed in with new and novel techniques.

Jet tagging is a great playground for machine learning.

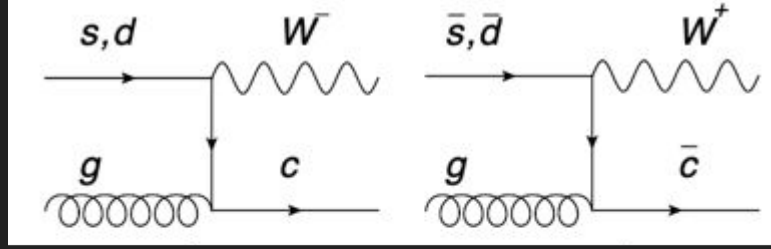
Several recent developments and many more to come.

Backup

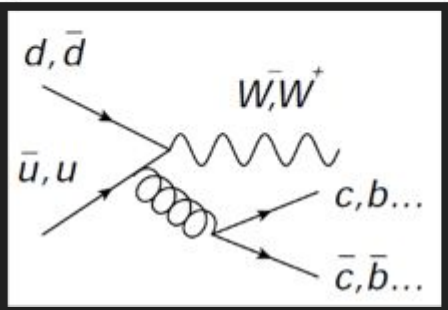
Iterative Fit Method (reshaping SF)

- Required by analyses using full distribution of discriminant.
- Tag and probe method using dilepton events (==2 jets, tag passes WP of discriminator).
- tt enriched (87% purity) region for HF SF, Z+jets enriched (99.9% purity) for light SF.
- SFb measured by subtracting light-jet contribution from HF-enriched region using MC with SF-light applied.
- Then performed in light-enriched region subtracting HF component with SFb applied.
- Repeated iteratively until SF’s converge. First iteration no SF applied.
- Large uncertainty on the combined effect of several uncertainties that could affect the sample purities.
- For c-tag reshaping SF same method but with additional W+c region.

Fixed WP c-tag SF's



- W+c Method
- At LO, W+c production due mainly to processes in which the W+c are opposite sign.
- Dominant bckg is W+qq where the OS/SS rate is balanced.
- OS-SS subtraction provides enriched W+c sample.
- Expected signal purity 60%(80%) for W→μν(W→eν) w. remaining bckg dominated by Z+j/ttbar(just ttbar).
- Large uncertainty c.f. background subtraction where fraction of W+c in MC and data assumed to be the same.



$$\epsilon_c = \frac{N(W+c)^{OS-SS}_{tagged}}{N(W+c)^{OS-SS}}$$

observed # OS-SS w. c-tagged jet x fraction of W+c events w. c-tagged jet.
observed # OS-SS events x fraction of W+c events.

What does the future hold?

Treatment of HF tagging nuisance parameter correlations in fit model needs coherent approach.

Correlation scheme for systematic uncertainties affecting scale factor measurement and physics analysis. Common tool to assess correlation scheme.

Need for advanced pileup mitigation techniques in tracking/b-tagging/jet selection e.g. PUPPI jets. Derivation of SFs for such jets.

Investigation into SF dependence as a function of the jet environment.