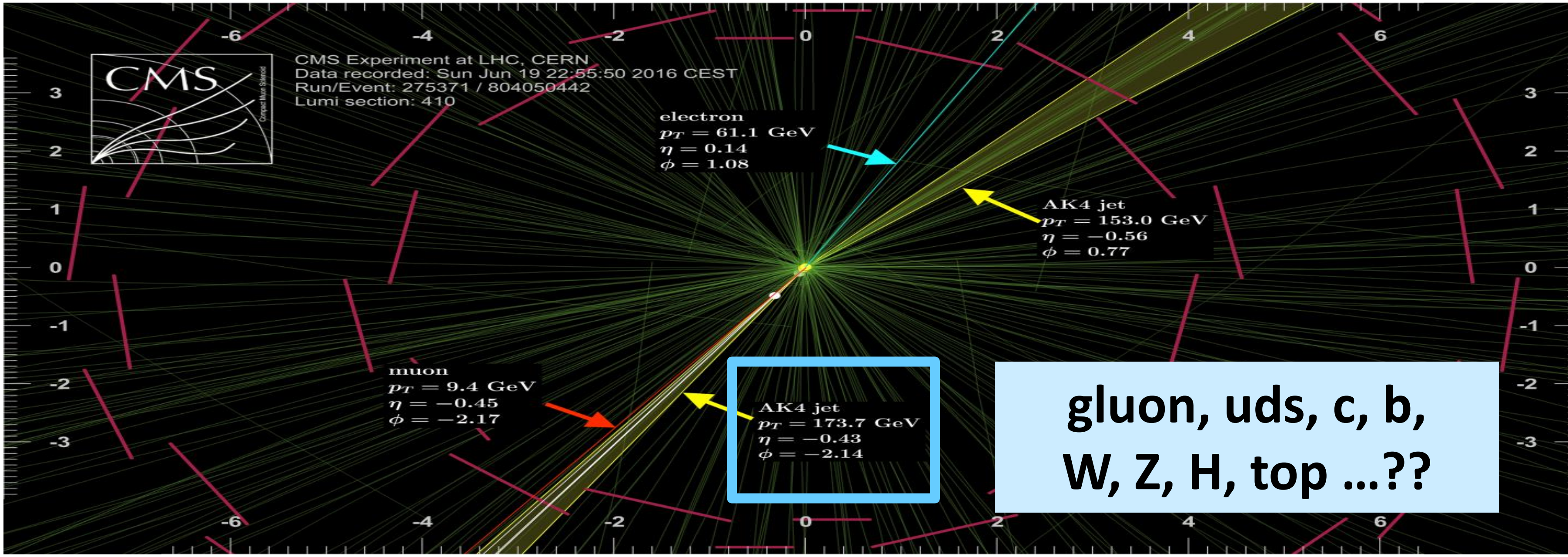
New Tagging approaches in CMS

Loukas Gouskos on behalf of the CMS Collaboration
University of California, Santa Barbara

Jet tagging

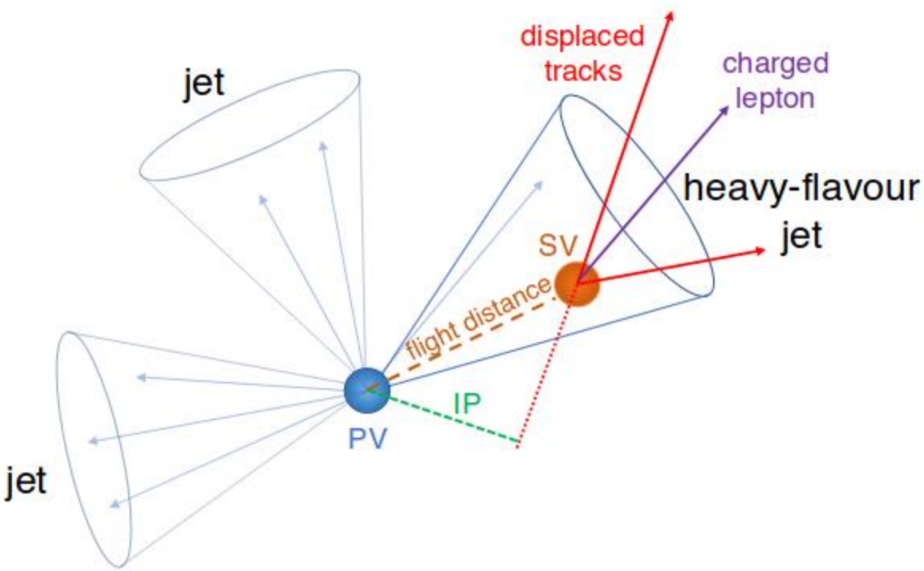


- People have been tagging jets for almost 30 years at Hadron colliders
  - starting with b jets at the Tevatron and LEP, then top, W/Z and Higgs jets at the LHC.
- But it is only now that we have begun to develop powerful and multi-Object tagging capabilities.
  - potential to open access to many new physics topics that had been written off previously

Tagging “slim (i.e.R=0.4)” jets

Jet tagging in CMS

- E.g. Heavy flavour tagging:

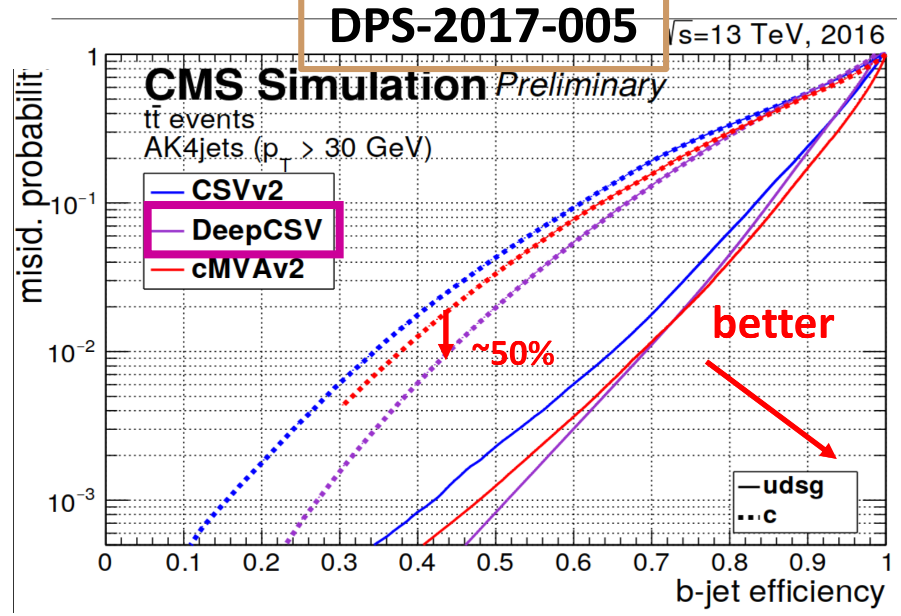


Standard tagging algorithms [CSV/cMVA]

- Large lifetime of hadrons with b
- Displaced vertices/tracks
- Large impact parameters
- Non-isolated leptons [in cMVA]
- Shallow ML (i.e. BDT) to improve performance

- From CSV/cMVA -> DeepCSV:
  - Use "standard" (high level) inputs with a few more tracks
  - BDT-> [simple] Dense Network (5 Hidden layers, 100 nodes)
  - Multiclass classifier: b, bb, c, udsg

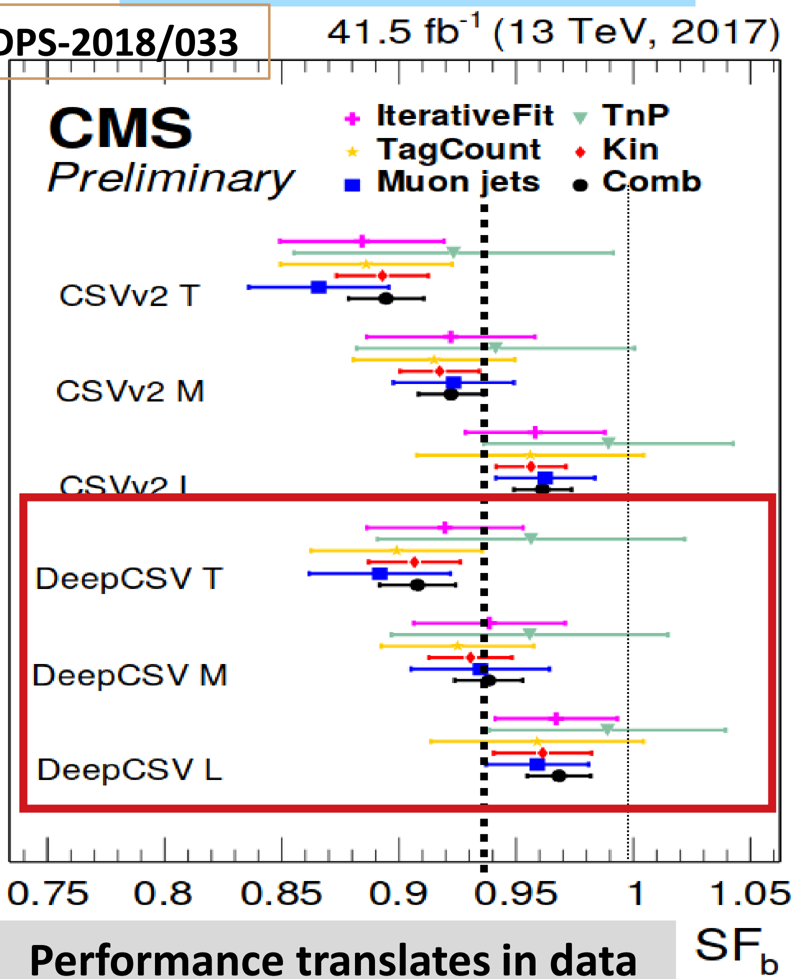
DPS-2017-005



Jet tagging in CMS (2)

Performance in data

DPS-2018/033



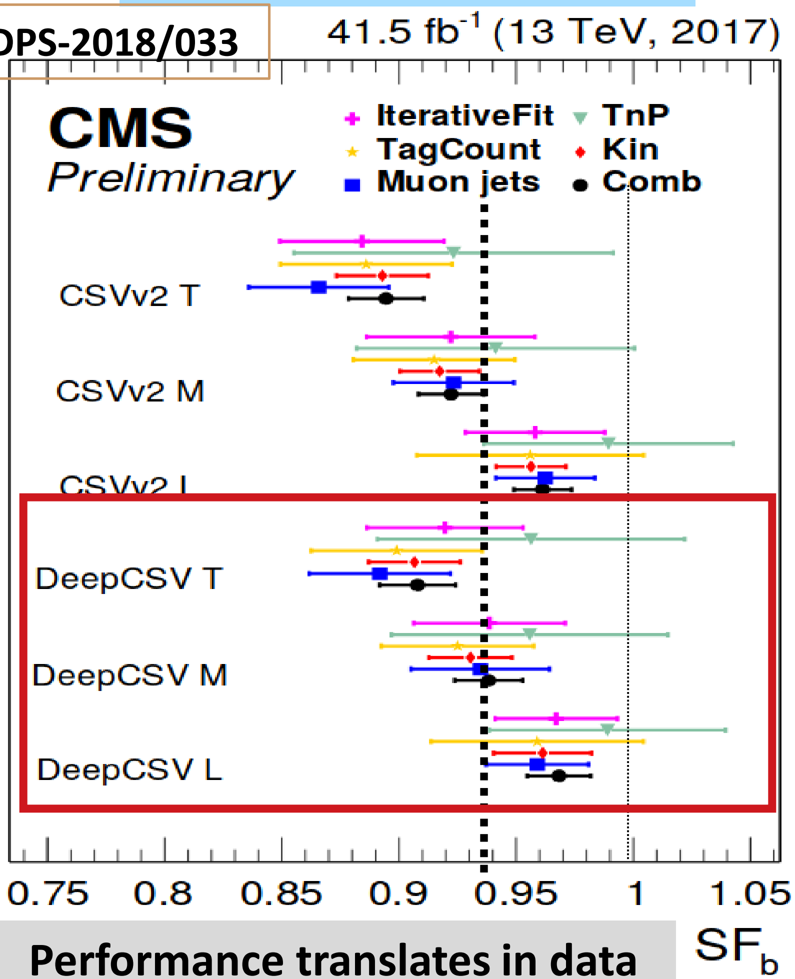
Performance translates in data

CMS default since 2017

Jet tagging in CMS (2)

Performance in data

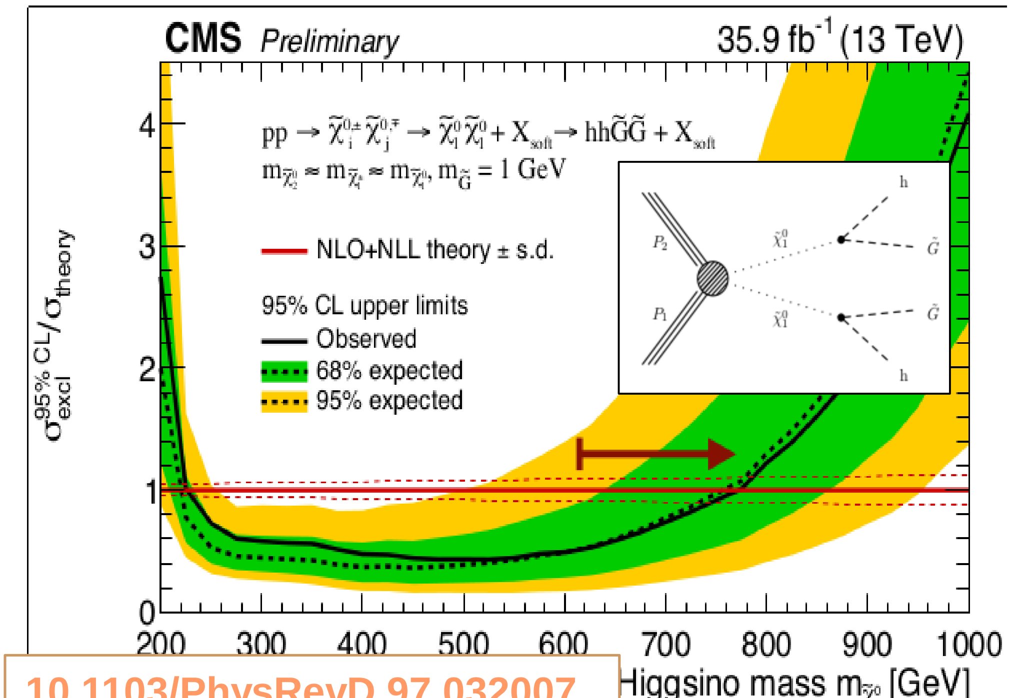
DPS-2018/033



Performance translates in data

CMS default since 2017

Impact on physics analyses



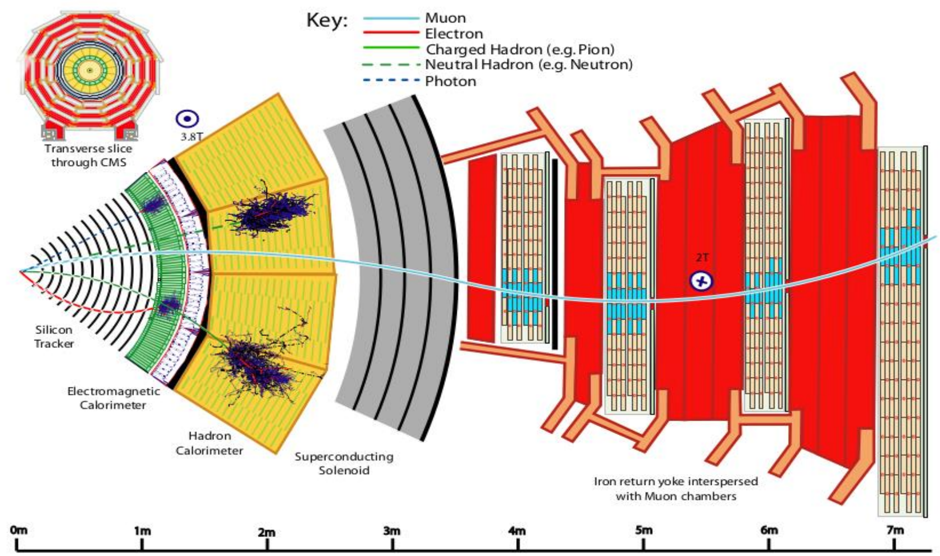
10.1103/PhysRevD.97.032007

→ **Significant gain in sensitivity in final states with multiple b quarks**

→ **~50% more signal for ~15% increase in the background**

Exploring [more] CMS potential

- CMS Event Reconstruction using Particle Flow (PF) algorithm
  - Combines information from all subdetectors
  - Mutually exclusive list of particles



- Rich information for each particle
  - Energy/momentum → **Inputs to jet substructure**
  - Particle category
  - Displacement from the PV
  - Reconstruction quality
  - ….

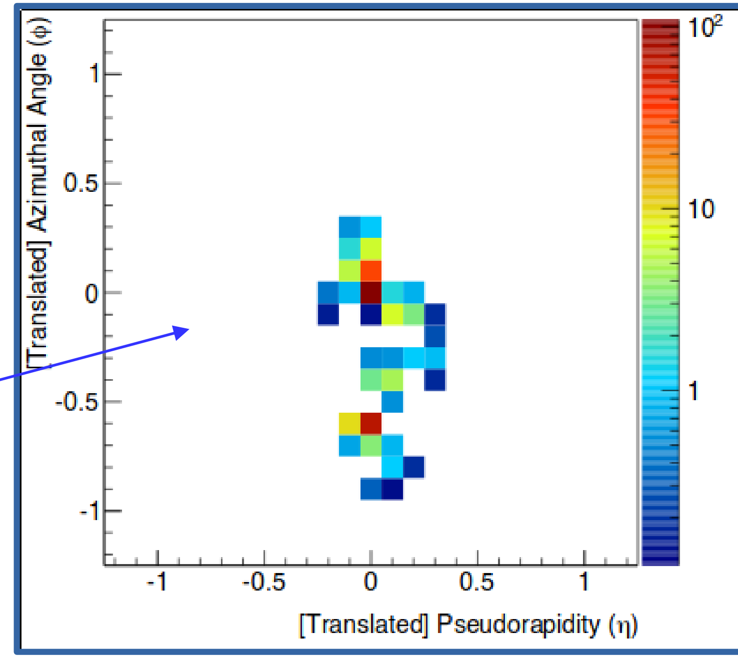
  → **Inputs for flavour tagging**

- [O(50) properties/particle] x [~50-100 particles/jet] ~O(1000) inputs/jet
- Begs for Deep Neural Networks (DNN) with "complex" architecture

Deep learning approaches for jet tagging

Boosted W

- Based on jet image:
  - Treat detector (i.e. calorimeters) as a camera & the jet as an image
  - Apply techniques used for image recognition (i.e. Convolutional Neural Networks – CNN)
  - <u>But:</u> jet images are very sparse
  - <u>And:</u> difficult to include information from other subdetectors (e.g. tracking)

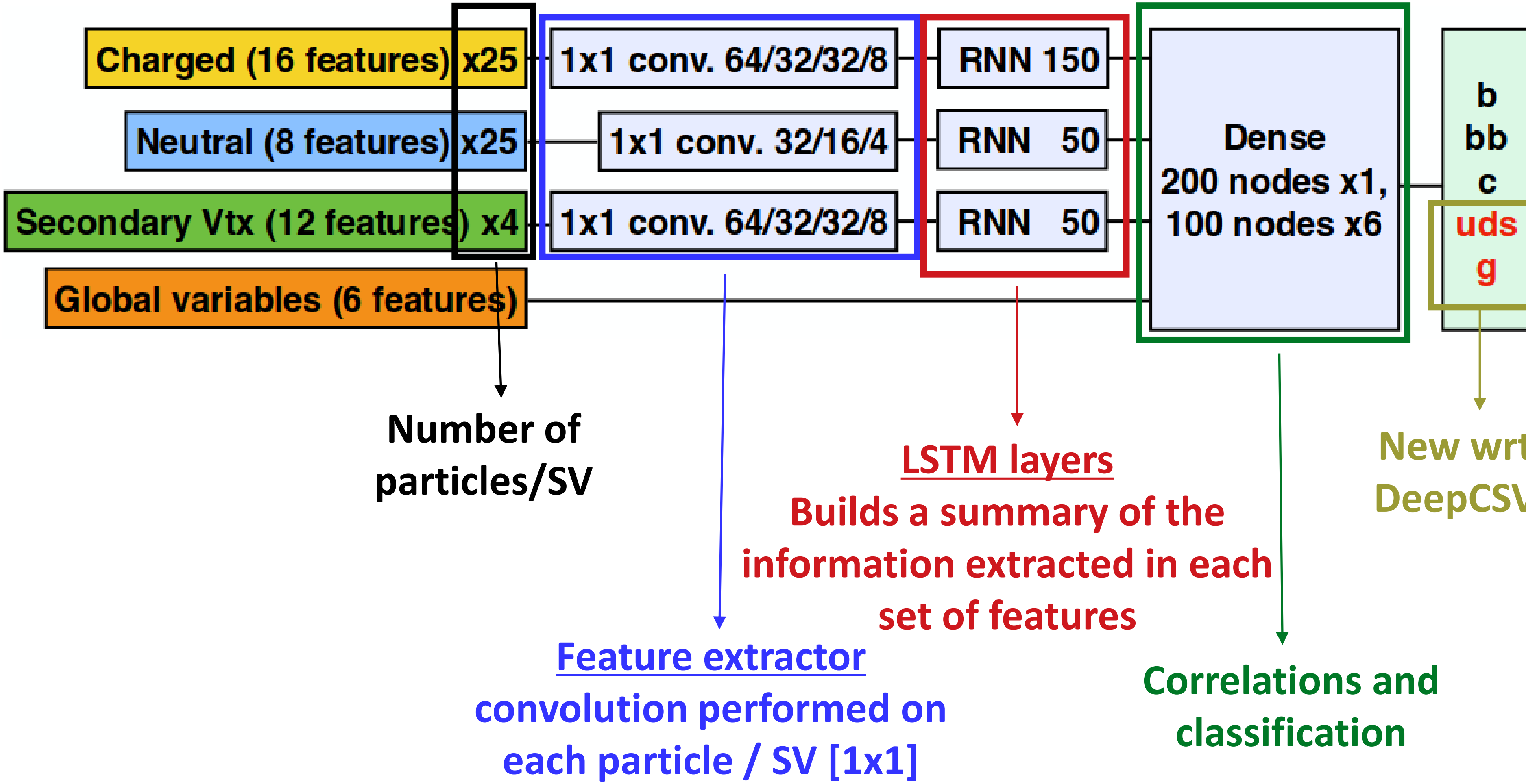


Deep learning approaches for jet tagging

- Based on particle sequence:
  - Jet as a sequence of constituent particles [i.e. PF Candidates]
  - Apply techniques used for natural language processing [e.g. RNN, CNN-1D, etc..]
  - Inclusion of more information [e.g. tracking] straight forward
  - Explore full potential of the CMS detector & event reconstruction

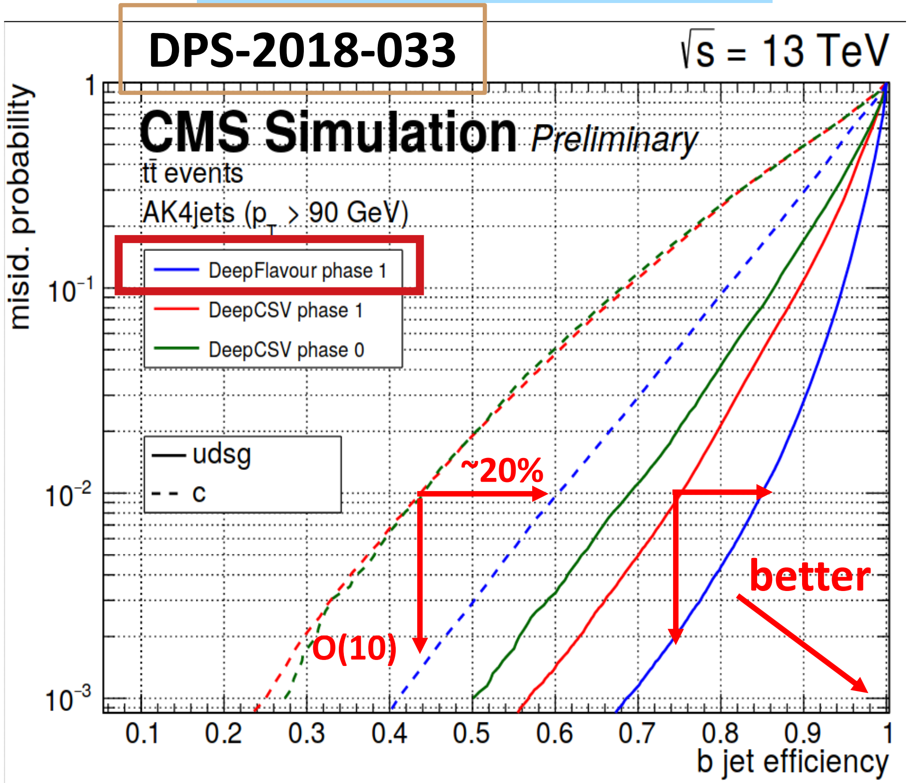
Particle-based "slim" jet tagging: DeepJet

- A multiclass classifier for: b, bb, c, uds, gluons
- Highlights from the architecture:



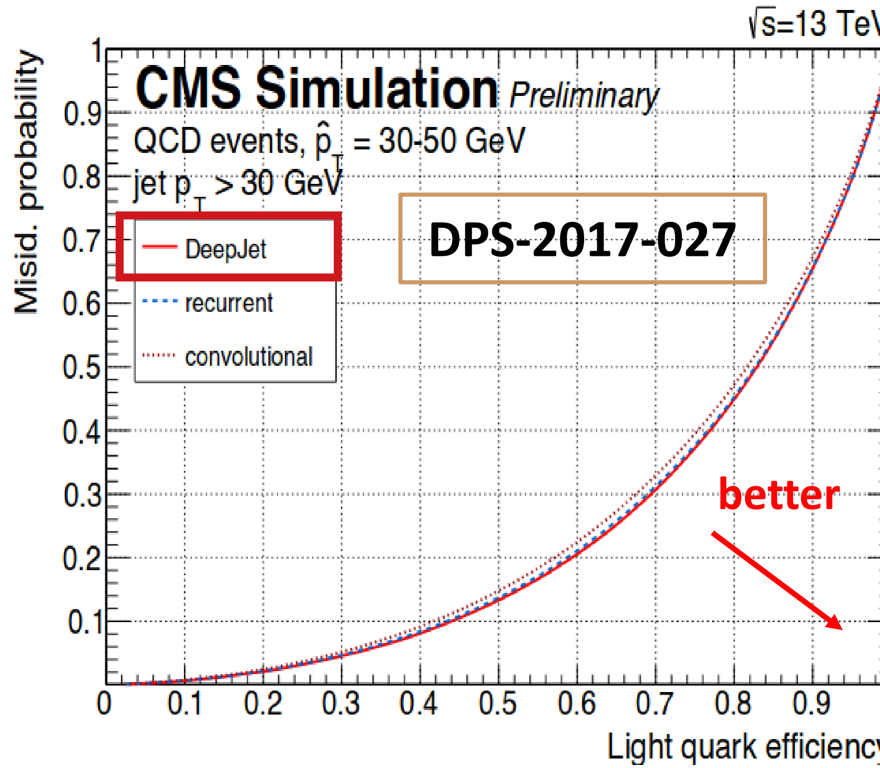
Performance of slim jet tagging

b vs. udsg / b vs. c



- → Significant gain in performance even more significant at higher pT
- → Large part of the performance loss of previous [non particle-based] taggers was due to track preselection

Quark – gluon separation



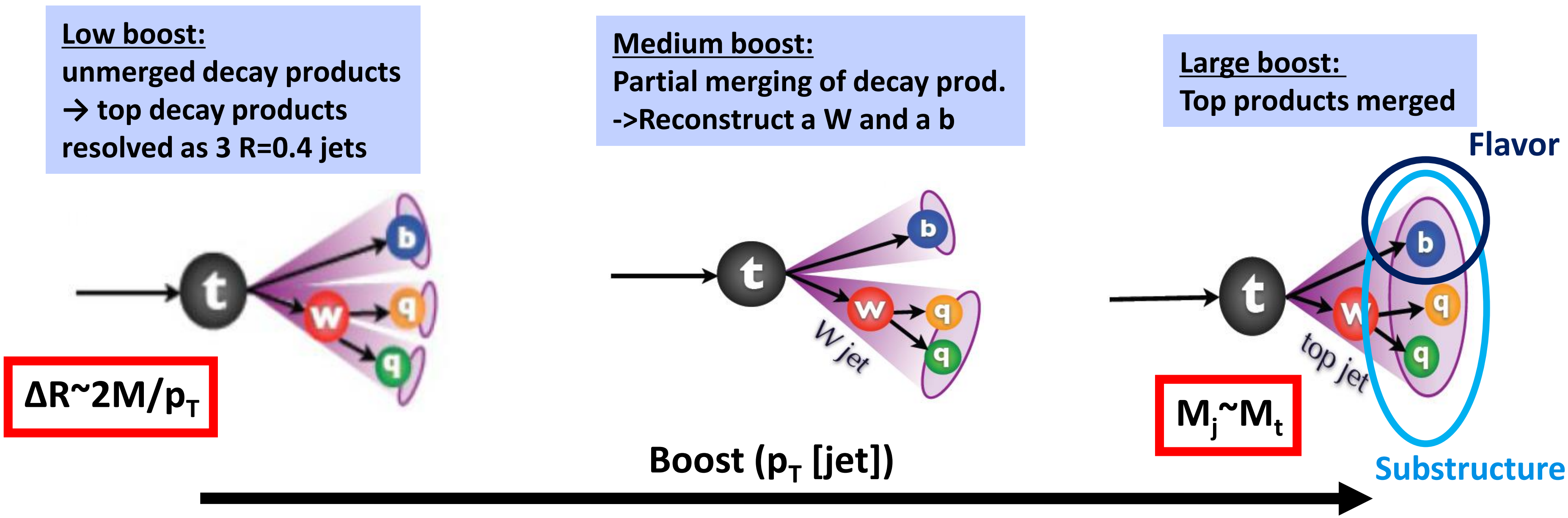
- → Generator level light quarks/gluons that did not split to heavy flavour
- → Similar performance to simpler & dedicated architectures

Boosted Jet tagging

i.e. large cone (R=0.8) jets

Boosted jet identification

- Use boosted top quark identification as an example:

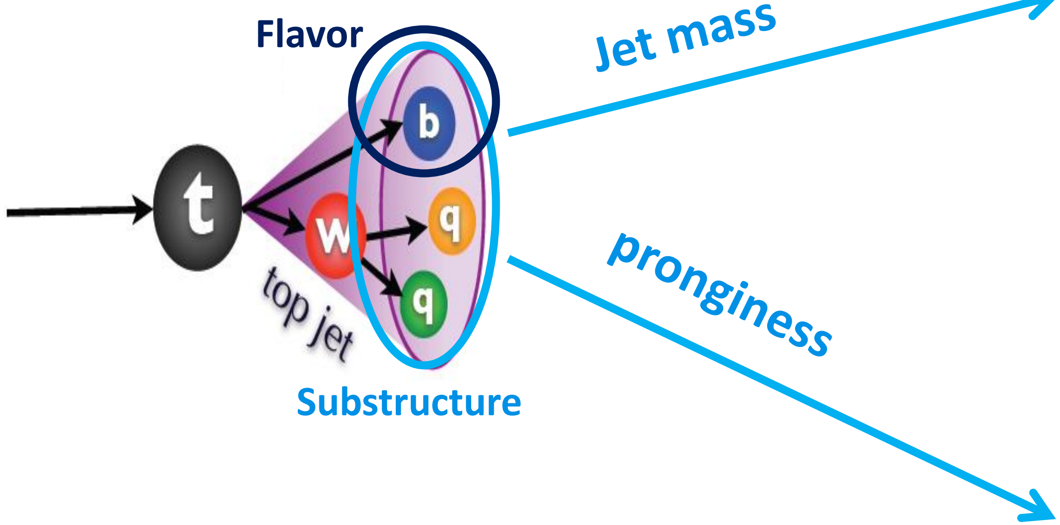


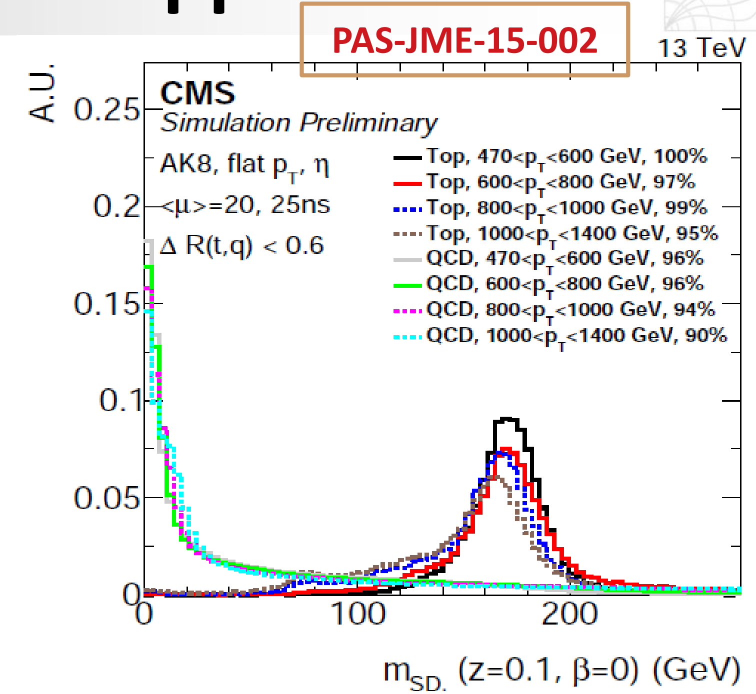
- In theory: A hadronic top quark decay to a W boson and a b quark → 3 quarks in total
  - **Substructure:** identify the 3-prong structure in a single “wide” jet
  - **Flavour:** Identify the b quark [or even W->cX]
  - **Pile-up**, etc..

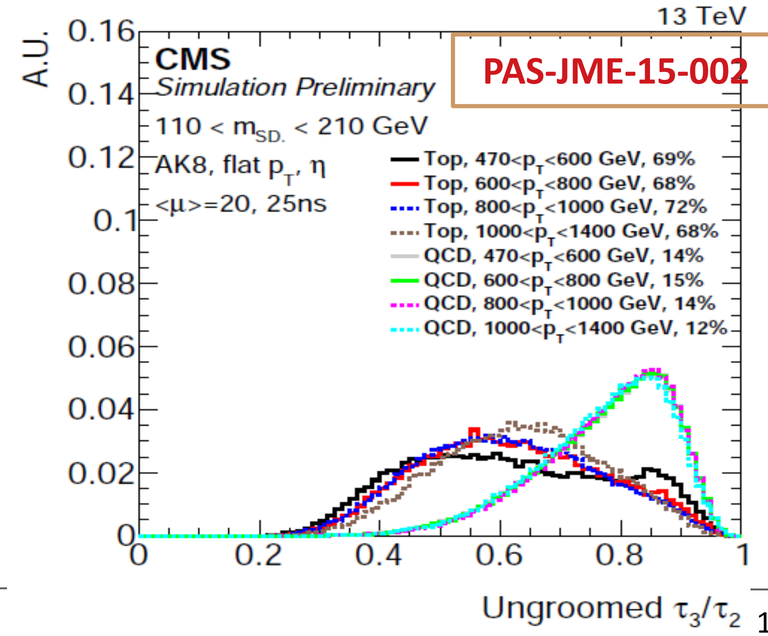
Not clear if these effects factorize
[need to exploited them simultaneously ?]

Boosted jet tagging: First approaches

- Simple "cut-based" approaches:
  - Extensively studied -> robust variables
  - Easy interpretation of results by the theory community



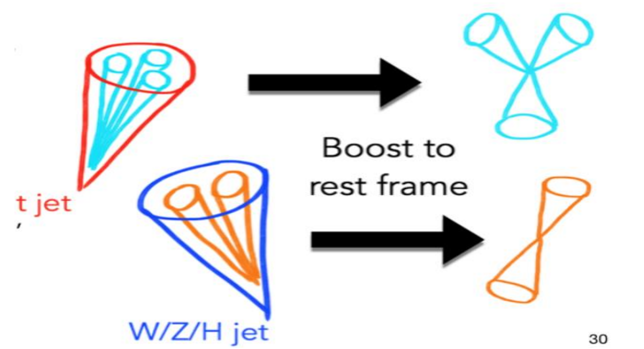


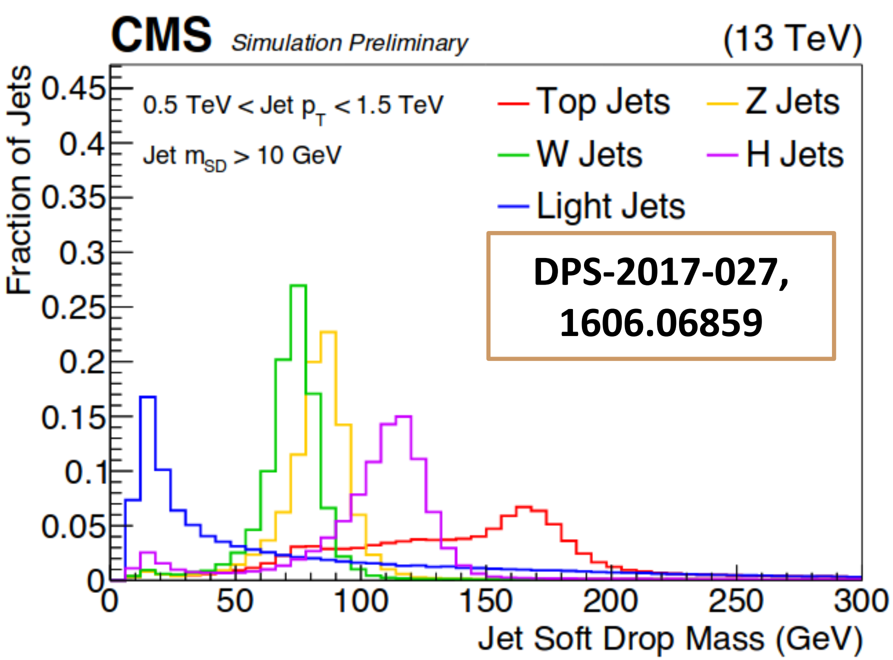


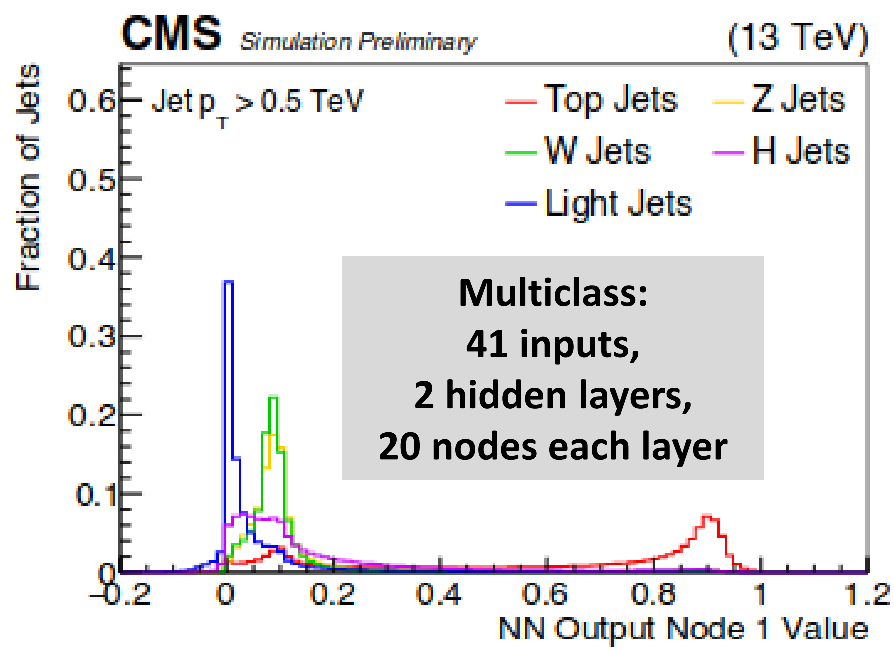
Flavor tagging:
b-tagging on softdrop subjets

Boosted jet tagging: LHC Run 2

- Move to approaches that exploit more advanced methods [i.e shallow ML]
- Include additional handles by constructing higher level observables
  - Quark-gluon separation, ECF, Event shapes, etc..
- Highlights from BOOST 2017;
  - E.g. Boosted Event Shape Tagger (BEST)

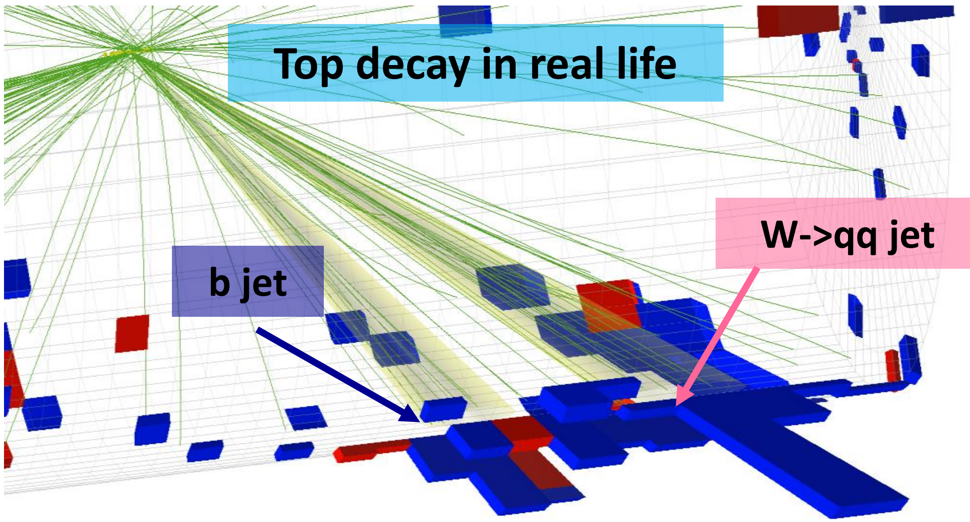


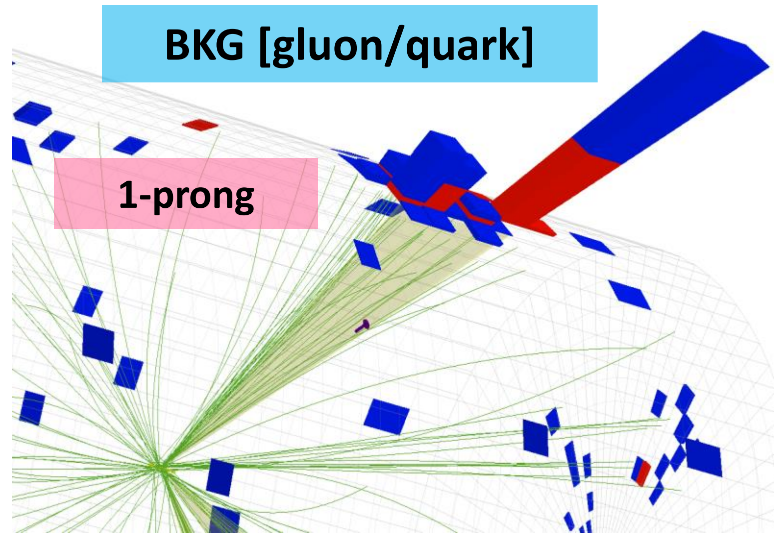




Particle based boosted jet tagging

- In practice: A jet is a cone full of reconstructed particles in the detector
  - With mass and kinematics consistent with the top decay

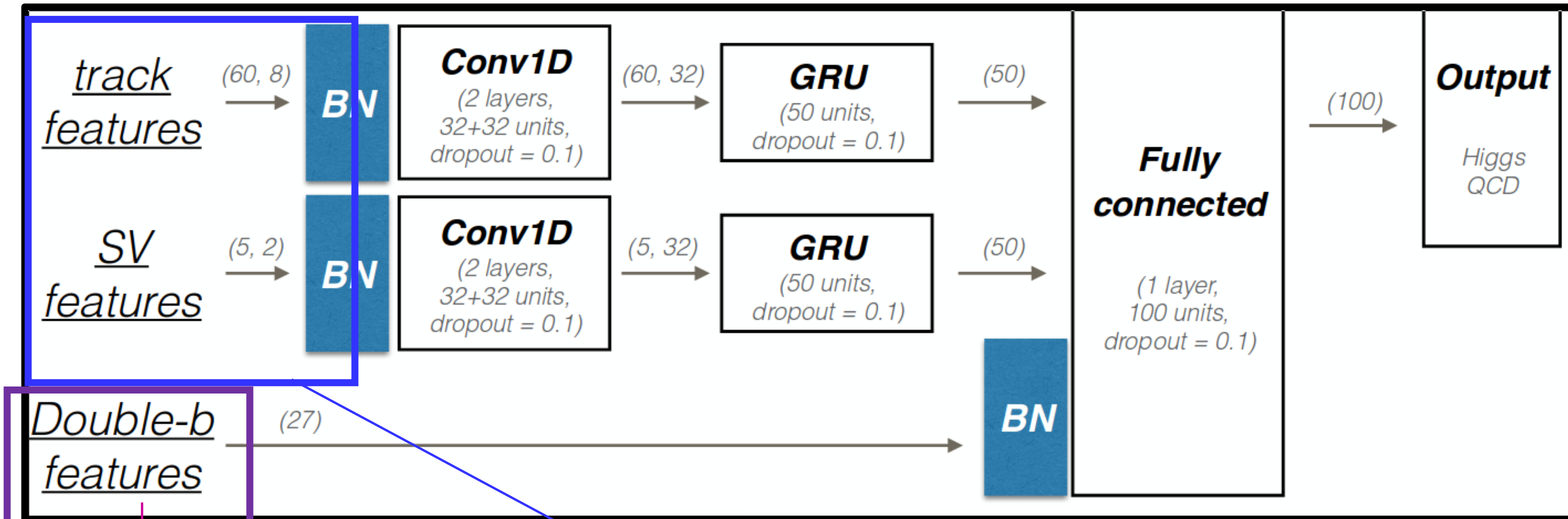




- Can we gain by moving to particle based jet tagging with DNN?
- Different challenges in larger cone jet tagging:
  - Much larger number of constituents
  - More complicated problem to solve [e.g. overlapping decay products]
  - More complex algorithms could/should be necessary

Particle-based Heavy Flavor (HF) Tagging [on boosted jets]

- Similar-ish approach as the one developed for "slim" jet tagging
  - Design motivated by the need to improve performance **while** keeping BKG jet mass sculpting under control [more later]

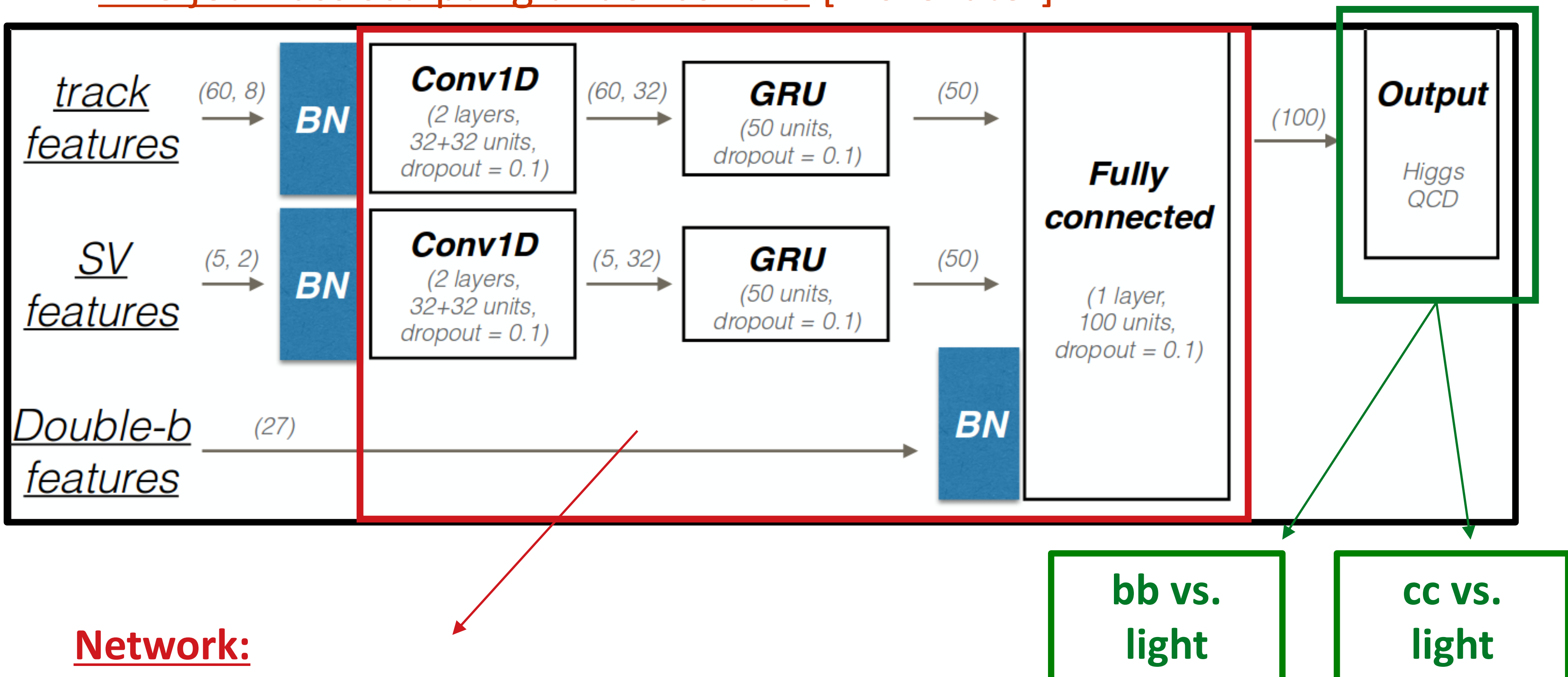


Double-b features:
-> vertexing, tracking & substructure
-> inputs largely mass & p_T independent

Inputs (differences wrt "slim" jet case):
→ Increased num of particles/SV (larger R)
→ Reduced num of features (less mass sculpting)
→ Drop Neutral PF candidates

Particle-based Heavy Flavor (HF) Tagging [on boosted jets]

- Similar-ish approach as the one developed for "slim" jet tagging
  - Design motivated by the need to improve performance **while** keeping BKG jet mass sculpting under control [more later]

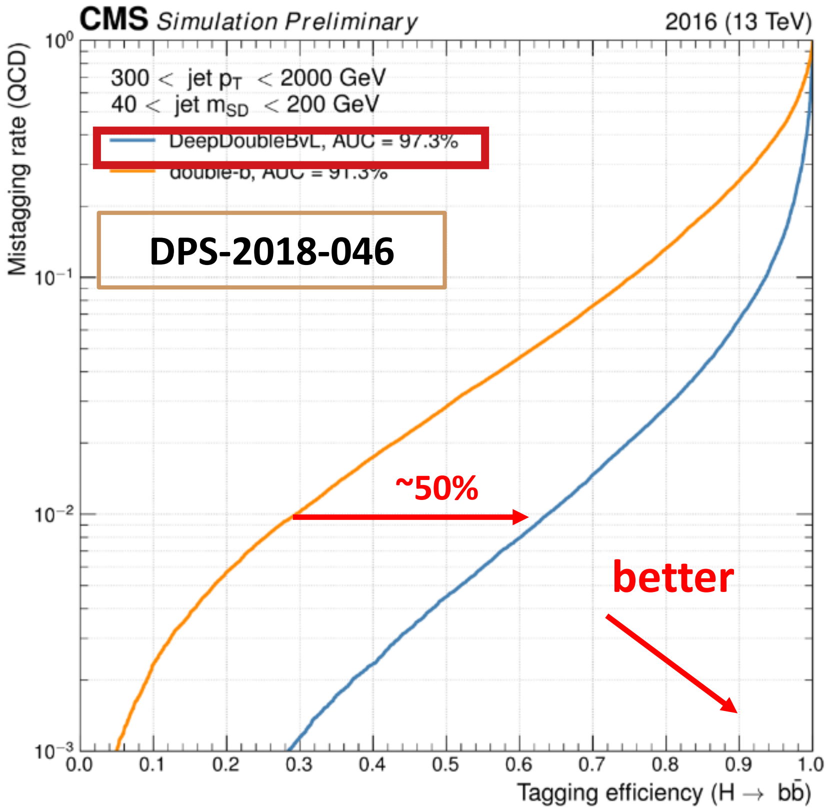


Network:
Simpler version of "slim" jet network
Main motivation: reduction of mass sculpting

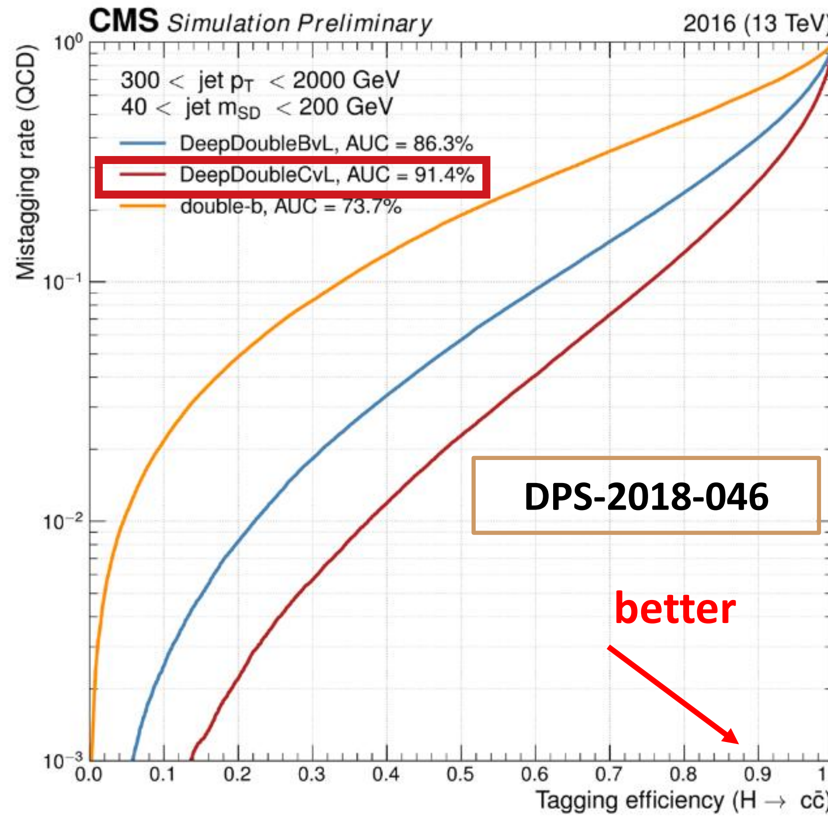
[binary classifier]

Performance of Boosted HF Tagging

H→bb vs. QCD



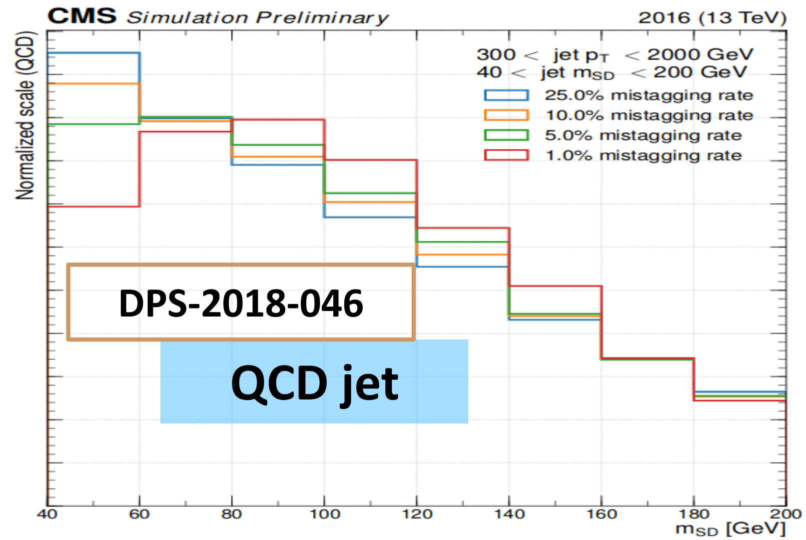
H→cc vs. QCD

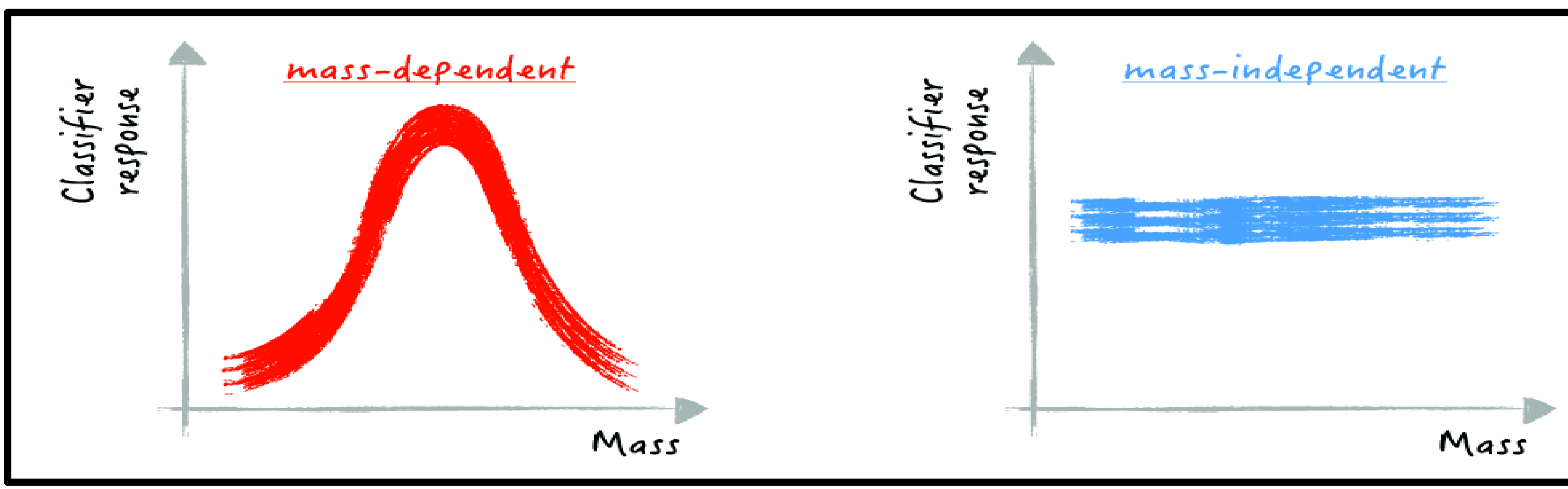


→ **Significant gain in performance wrt to existing BDT-based approaches**
→ **bb vs. QCD: ~2x higher signal efficiency for ~1% BKG rate**

Mass sculpting of BKG Jets

- **But:** it is correlated with the jet mass
- Is this a show-stopper?
  - Depends on the physics analysis
- What does "mass independence" means for a tagger?





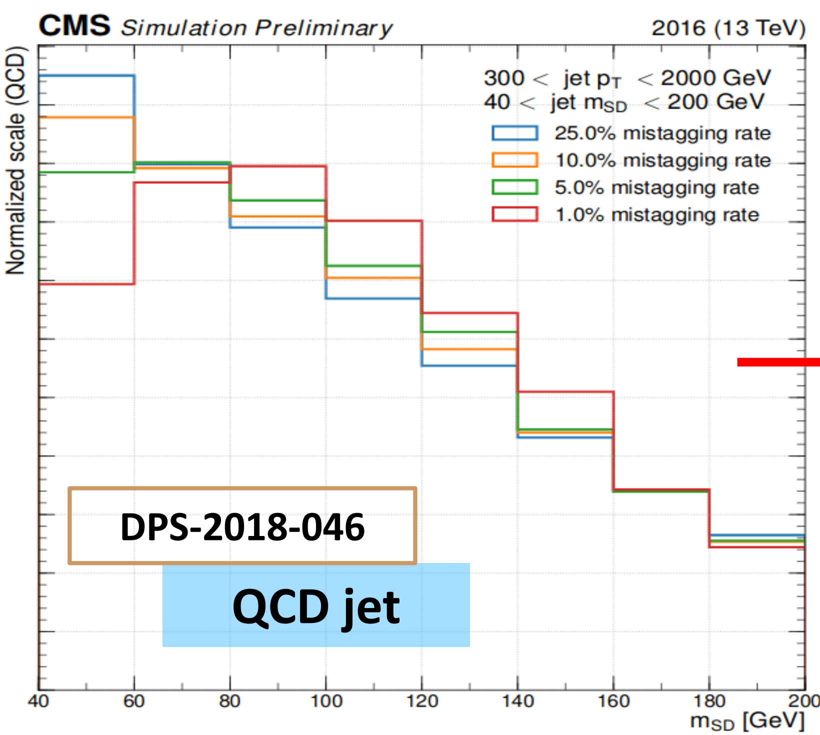
Classifier's response is non-uniform in mass

Classifier's response is uniform in mass

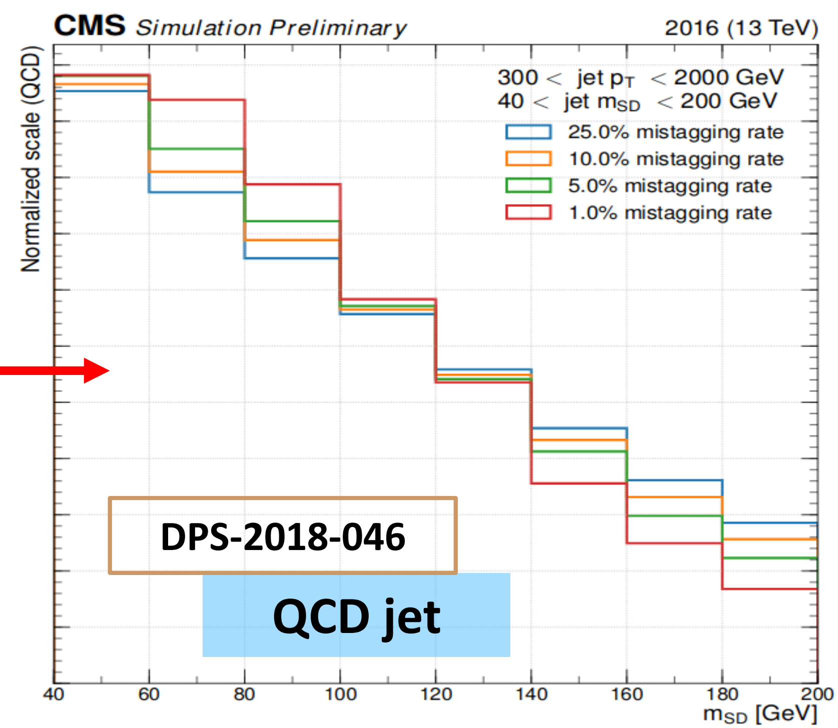
Addressing mass sculpting

- One method to decorrelate tagger's response with jet mass:
  - Add a penalty term in the loss function proportional to the Kullback-Liebler divergence
    - i.e. how much the bkg jet mass shape changes after classification

Before



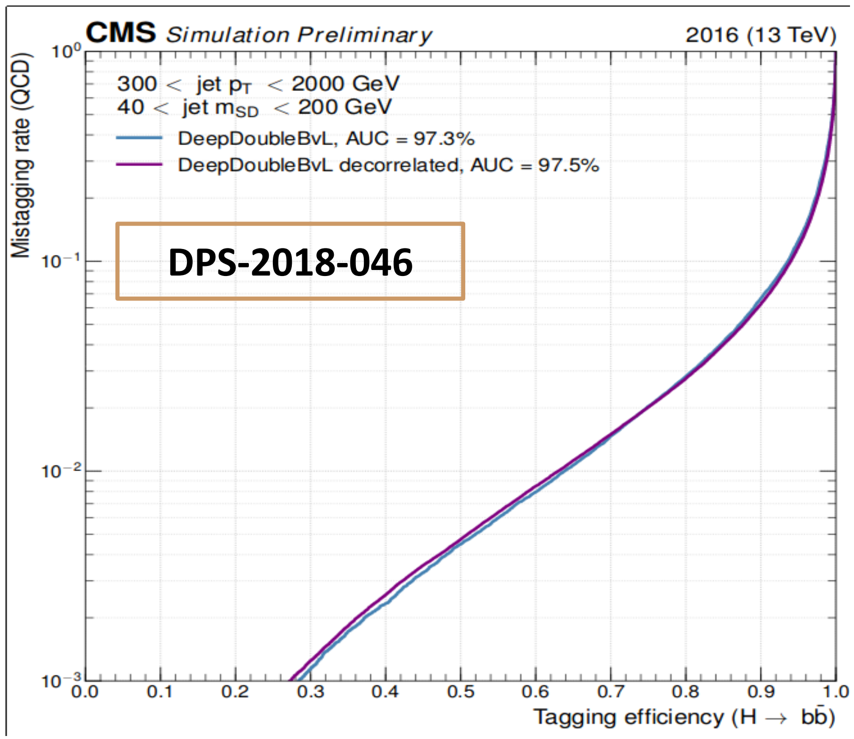
After



Mass sculpting significantly reduced

Addressing mass sculpting (2)

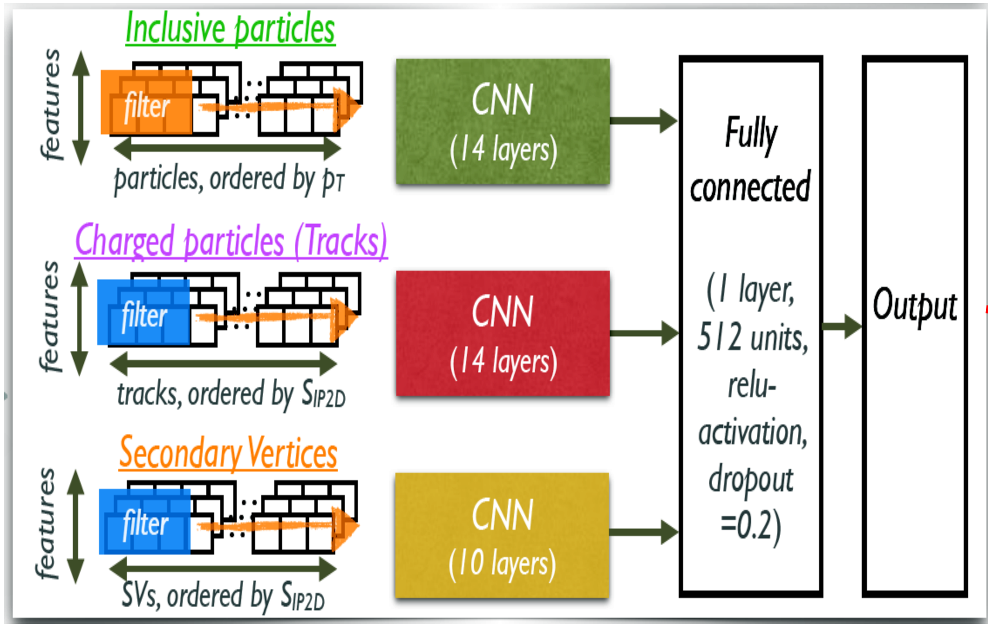
- One method to decorrelate tagger's response with jet mass:
  - For each training batch add a penalty term proportional to the Kullback-Liebler divergence



Negligible impact in performance
[NB. Tagger was already designed for small mass sculpting]

Particle-based Boosted Jet Tagger

- Deep (boosted) Jet: Multiclass classifier for W, Z, H and top tagging
  - Designed for maximal performance



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)

Inputs:

→ **Up to 100 inclusive particles**
10 features/particle

→ **Up to 60 charged particles**
60 features/particle

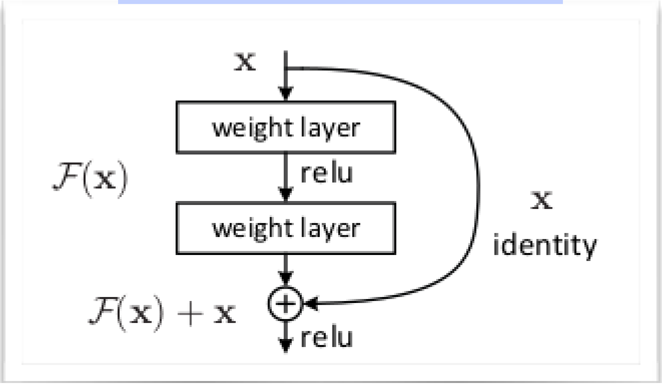
→ **Up to 5 SV (14 features/SV)**

A very versatile boosted jet tagger
-> various decay modes with different flavor content

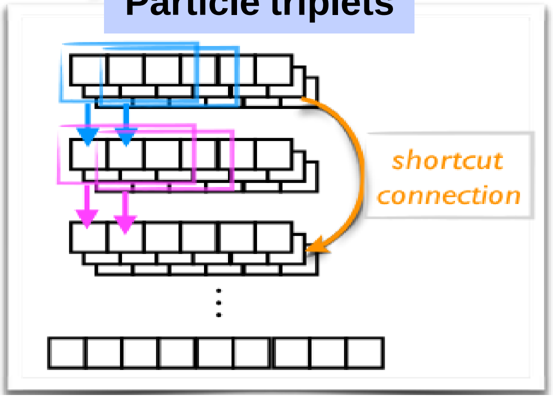
DeepBoostedJet: Network architecture

- Advanced/Complex network architecture is necessary to achieve maximum performance
  - Architecture based solely on 1D-CNN
    - Less computationally expensive
  - Fairly deep network to better exploit correlations between particles
  - CNN architecture inspired by the ResNet model for image recognition
    - Improves performance in deep networks and makes training easier
  - “Move” in particle triplets
    - Exploit correlations between nearby particles faster

1512.03385,1603.05027



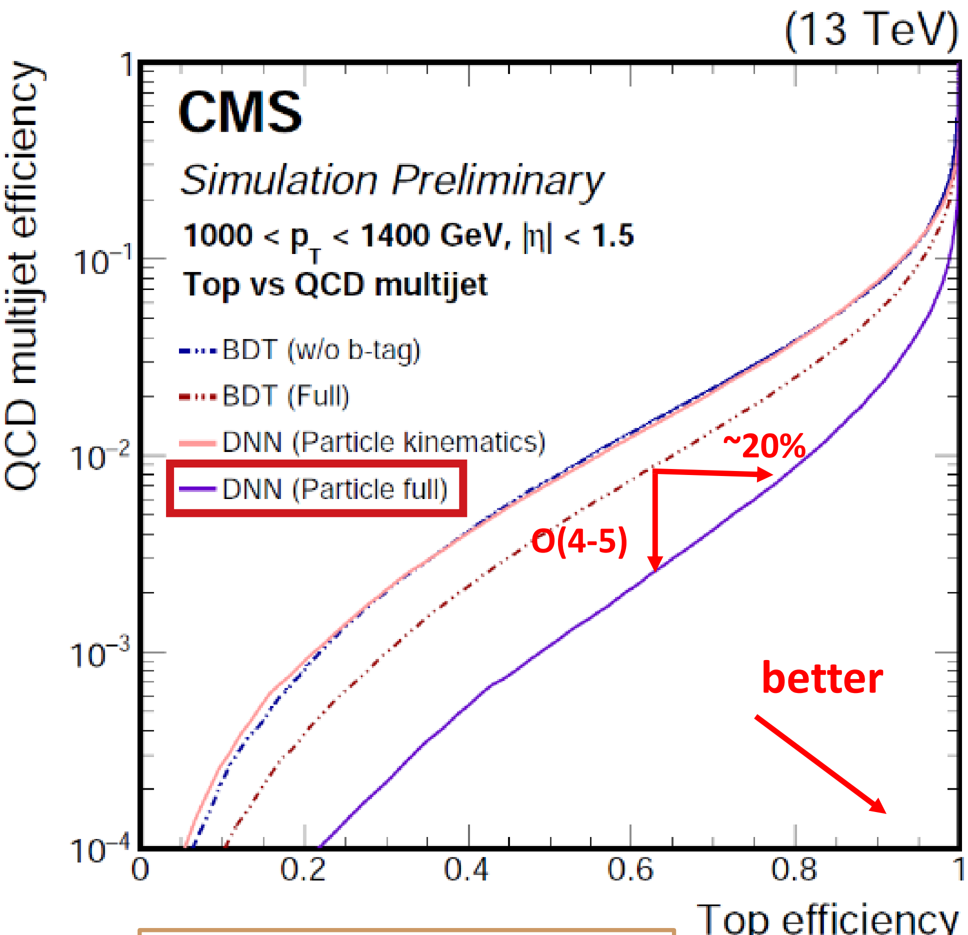
Particle triplets



- Also: A version decorrelated with the jet mass
  - Same architecture and inputs as nominal version
  - Use of adversarial networks to predict the jet mass

Performance in MC

e.g. Top vs. QCD



DPS-2017-049,
NIPS 2017 paper

- Compare performance vs. a BDT-based approach which uses state of the art high-level features [ref. 10.1007/JHEP10(2017)005]
  - inputs: jet substructure, b/c flavour tagging, quark-gluon tagging
- Investigate impact of flavour tagging in performance
- Similar performance between the DNN and the BDT approach without flavour info
- Very significant gain [~4-5x] for the Particle-based DNN approach after including flavour

Summary

- Jet tagging is essential for the success of the LHC Physics program
  - Large effort at the LHC to improve existing / develop new jet tagging methods
- Key player in these developments: Advanced machine learning algorithms
  - Allows us see much more of the true potential of the CMS apparatus
  - Still room for improvement / other ideas to try
    - Strong interest by the theory and experiment communities
- Effort pays off: Large gain in performance wrt previous approaches
  - First results in data look encouraging; work on-going -> **stay tuned!**
  - Lots of effort to better understand what the DNN learns
  - Understand complementarity between the various approaches in CMS

    .. But also between ATLAS & Theory:
    - Use the same training/testing sample & compare the various top-tagging archs
    - More details in: https://mmm.cern.ch/owa/redir.aspx?C=8GGPliT9bPPgi3hzi5byLjurHZHnRw1z7XmkgL-w4UKKwRmprezVCA..&URL=https%3a%2f%2fgoo.gl%2fXGYju3

Please send your results!

Extra slides

CSV, cMVA, DeepCSV inputs

CSVv2 Inputs

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

-> DeepCSV: CSVv2 + more tracks [up to 6 tracks]

-> cMVA inputs: discriminators of other b-tagging variables & uses a BDT

DeepJet Inputs

Charged particles

feature	offset	lower bound	upper bound	comment
trackEtaRel	-	-5	15	BTV
trackPtRel	-	-	4	BTV
trackPPar	-	-10^5	10^5	BTV
trackDeltaR	-	-5	5	BTV
trackPParRatio	-10	100	-	BTV
trackSip2dVal	-	-	70	BTV
trackSip2dSig	-	-	$4 \cdot 10^4$	BTV
trackSip3dVal	-	-	10^5	BTV
trackSip3dSig	-	-	$4 \cdot 10^4$	BTV
trackJetDistVal	-	-20	1	BTV
trackJetDistSig	-	-1	10^5	BTV
$p_T(cPF)/p_T(j)$	-1	-1	0	
$\Delta R_m(cPF, SV)$	-5	-5	0	
fromPV	-	-	-	
VTXass	-	-	-	
$w_p(cPF)$	-	-	-	
χ^2	-	-	-	
Npixel hits	-	-	-	

Neutral particles

feature	offset	lower bound	upper bound
$p_T(nPF)/p_T(j)$	-1	-1	0
$\Delta R_m(nPF, SV)$	-5	-5	0
isGamma	-	-	-
hadFrac	-	-	-
$\Delta R(nPF)$	-0.6	-0.6	0
$w_p(cPF)$	-	-	-

Secondary vertices

feature	offset	lower bound	upper bound
$p_T(SV)$			
$\Delta R(SV)$	-0.5	-2	0
$m_{SV})$	-	-	-
$N_{tracks}(SV)$	-	-	-
$\chi^2(SV)$			
$\chi^2_n(SV)$	0	-1000	1000
$d_{xy}(SV)$	-	-	-
$S_{xy}(SV)$	-	-	800
$d_{3D}(SV)$	-	-	-
$S_{3D}(SV)$	-2	-2	0
$\cos\theta(SV)$	-	-	-
$E_{rel}(SV)$	-	-	-

Global features

feature	comment
$p_T(j)$	
$\eta(j)$	
N_{cPF}	
N_{nPF}	
N_{SV}	
N_{PV}	
trackSumJetEtRatio	BTV
trackSumJetDeltaR	BTV
vertexCategory	BTV
trackSip2dValAboveCharm	BTV
trackSip2dSigAboveCharm	BTV
trackSip3dValAboveCharm	BTV
trackSip3dSigAboveCharm	BTV
jetNSelectedTracks	BTV
jetNTracksEtaRel	BTV

DeepDoubleBB/CC Inputs

double-b features (27)

- The first four SIP values for selected tracks ordered in decreasing SIP;
- For each τ-axis we consider the first two SIP values for their respective associated tracks ordered in decreasing SIP, to further discriminate against single b quark and light flavor jets from QCD when one or both SV are not reconstructed due to IVF inefficiencies;
- The measured IP significance in the plane transverse to the beam axis, 2D SIP, of the first two tracks (first track) that raises the SV invariant mass above the bottom (charm) threshold of 5.2 (1.5) GeV;
- The number of SV associated to the jet;
- The significance of the 2D distance between the primary vertex and the secondary vertex, flight distance, for the SV with the smallest 3D flight distance uncertainty, for each of the two τ-axes;
- The ΔR between the SVs with the smallest 3D flight distance uncertainty and its τ-axis, for each of the two τ-axes;
- The relative pseudorapidity, η_{rel}, of the tracks from all SVs with respect to their τ-axis for the three leading tracks ordered in increasing η_{rel}, for each of the two τ-axes;
- The total SV mass, defined as the total mass of all SVs associated to a given τ-axis, for each of the two τ-axes;
- The ratio of the total SV energy, defined as the total energy of all SVs associated to a given τ-axis, and the total energy of all the tracks associated to the fat jet that are consistent with the primary vertex, for each of the two τ-axes;
- The information related to the two-SV system, the z variable, defined as:

$$z = \Delta R(\mathrm{SV}_0, \mathrm{SV}_1) \cdot \frac{p_{T,\mathrm{SV}_1}}{m(\mathrm{SV}_0, \mathrm{SV}_1)} \tag{2}$$

where SV_0 and SV_1 are SVs with the smallest 3D flight distance uncertainty. The z variable helps rejecting the $b\bar{b}$ background from gluon splitting relying on the different kinematic properties compared to the $b\bar{b}$ pair from the decay of a massive resonance.

track features (60×8)

Table 10: Full list of charged PF candidate features used as input to the DeepAK8 network

feature	comment
trackEtaRel	BTV
trackPtRatio	BTV
trackPParRatio	BTV
trackSip2dVal	BTV
trackSip2dSig	BTV
trackSip3dVal	BTV
trackSip3dSig	BTV
trackJetDistVal	BTV

SV features (5×2)

$d_{3D}(SV)$
$S_{3D}(SV)$

DeepBoostedJet Inputs

PF candidates

$\log p_{\mathrm{T}}$
$\log(p_{\mathrm{T}}/p_{\mathrm{T}}(\mathrm{jet}))$
$\log E$
$|\eta|$
$\Delta\phi(\mathrm{jet})$
$\Delta\eta(\mathrm{jet})$
Puppi weight
$\min \Delta R(\mathrm{SV})$
$\Delta R(\mathrm{jet})$
$\Delta R(\mathrm{subjet\ 1})$
$\Delta R(\mathrm{subjet\ 2})$
q (electric charge)
isMuon
isElectron
isPhoton
isChargedHadron
isNeutralHadron
hcalFraction
pvAssociationQuality
lostInnerHits
d_{xy}
d_z
$\sigma_{d_{xy}}$
σ_{d_z}
χ^2_n
qualityMask
dptdpt
detadeta
dphidphi
dxydxy
dzdz
dxydz
dphidxy
dlambdadz
trackEtaRel
trackPtRatio
trackPParRatio
trackSip2dVal
trackSip2dSig
trackSip3dVal
trackSip3dSig
trackJetDistVal

Secondary vertices

$\log p_{\mathrm{T}}$
$\log(p_{\mathrm{T}}/p_{\mathrm{T}}(\mathrm{jet}))$
$\log E$
$|\eta|$
$\Delta\phi(\mathrm{jet})$
$\Delta\eta(\mathrm{jet})$
$\Delta R(\mathrm{jet})$
m_{SV}
N_{tracks}
χ^2_n
d_{xy}
$\sigma_{d_{xy}}$
$d_{3\mathrm{D}}$
$\sigma_{d_{3\mathrm{D}}}$
$\cos(\overrightarrow{(\mathrm{PV},\mathrm{SV})}, \vec{p}_{\mathrm{SV}})$

DeepX technical details

- Training/validation/testing:
  - DeepJet: 80 / 10 / 10 M jets
  - Deep(Boosted)Jet: 40 / 5 / 5 M jets
  - DeepDoubleB/C : 3 / 0.5 / 0.5
- ReLu, Dropout (0.1), batch normalization, softmax
- Adam optimizer

DeepDoubleBB/CC: Addressing mass sculpting

- One method to decorrelate tagger's response with jet mass:
  - For each training batch add a penalty term proportional to the Kullback-Liebler divergence

$$D_{\mathrm{KL}}(\vec{w}, \vec{z}) = \sum_j w_j \log\left(\frac{w_j}{z_j}\right)$$

    - Measures how similar are two distributions/pdf

- (a) Mass histograms of each class (QCD or H->bb) weighted by the probability to be classified as QCD or H->bb then normalized to 1
- (b) minimize the KL divergence:

KL(p(QCD|QCD), p(H(bb), QCD))

KL(p(H(bb), H(bb)), p(QCD|H(bb)))

$$h_{(0|0)\ell} = \frac{\sum_{k=1}^{P} h_\ell^k p_0^k y_0^k}{\sum_{\ell=1}^{Q}\sum_{k=1}^{P} h_\ell^k p_0^k y_0^k}$$

$$h_{(1|0)\ell} = \frac{\sum_{k=1}^{P} h_\ell^k p_1^k y_0^k}{\sum_{\ell=1}^{Q}\sum_{k=1}^{P} h_\ell^k p_1^k y_0^k}$$

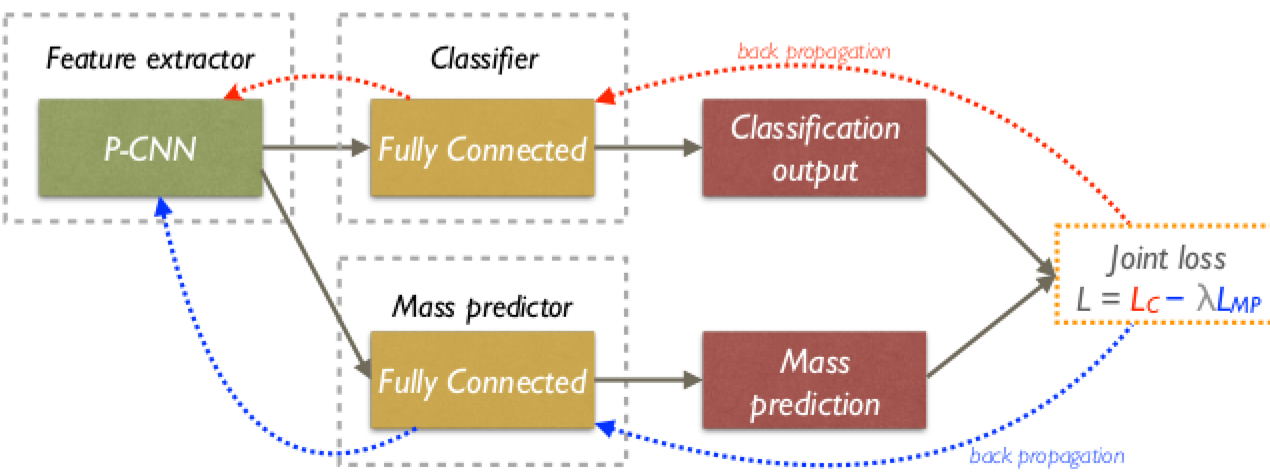
$$h_{(0|1)\ell} = \frac{\sum_{k=1}^{P} h_\ell^k p_0^k y_1^k}{\sum_{\ell=1}^{Q}\sum_{k=1}^{P} h_\ell^k p_0^k y_1^k}$$

$$h_{(1|1)\ell} = \frac{\sum_{k=1}^{P} h_\ell^k p_1^k y_1^k}{\sum_{\ell=1}^{Q}\sum_{k=1}^{P} h_\ell^k p_1^k y_1^k}$$

- (c) add to final loss

DeepBoostedJet: mass decorrelation

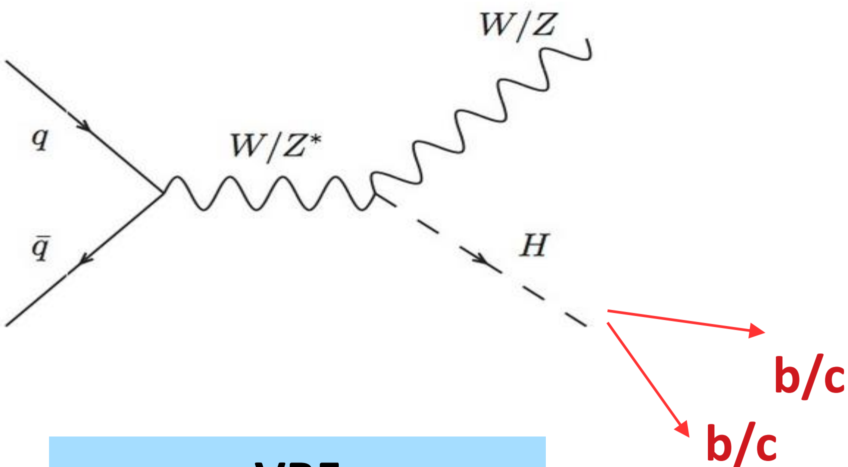
- Use adversarial training to regulate the behaviour of the network
  - Introduce a mass prediction network to predict the jet mass from the features extracted by the CNNs
  - It's loss (L_{MP}) is an indicator for mass correlation
    - Smaller L_{MP} more accurate mass prediction ; the features extracted by the CNNs have a higher correlation with jet mass
  - Introduce a joint loss: $L = L_C - \lambda L_{MP}$, second term a penalty on mass correlation
    - Minimizing L -> simultaneously improve classification & reduce mass correlation
    - λ: hyperparameter balancing between performance and mass independence



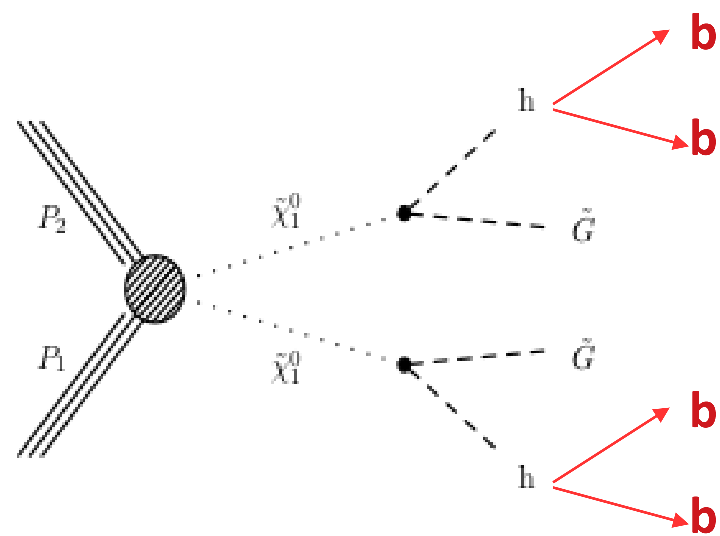
Physics with slim jet tagging

- Jet tagging on single parton jets (i.e. anti-kT R=0.4 “slim” jets):
  - Powerful tool for SM measurements and BSM searches, e.g.:

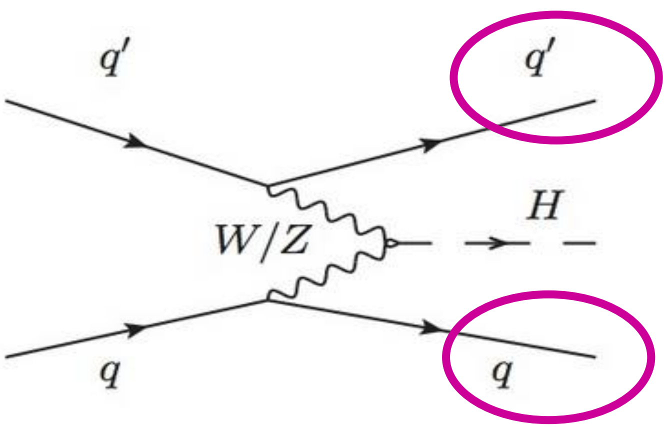
VH (→ bb/cc)



BSM (e.g. SUSY)



VBF

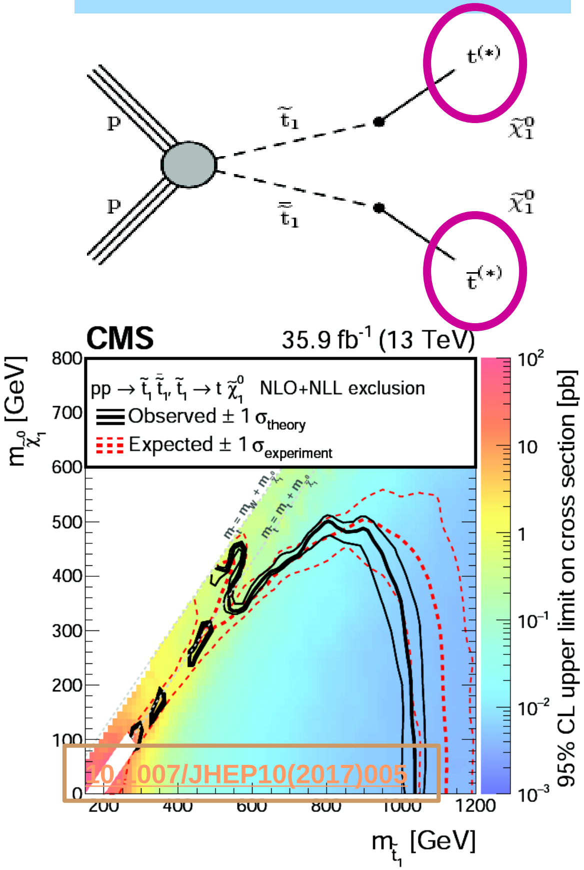


Heavy flavour tagging, q/g discrimination remain important tools that must constantly evolve as pileup rises.

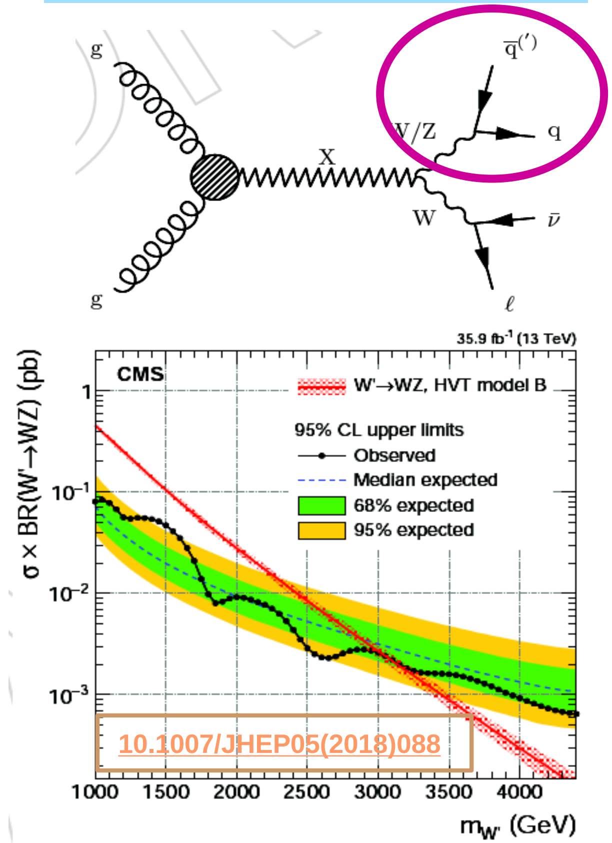
Physics with boosted jets

- Boosted jet tagging, as we all know, has real power and we are finding ever more applications but it too has to constantly evolve and improve.

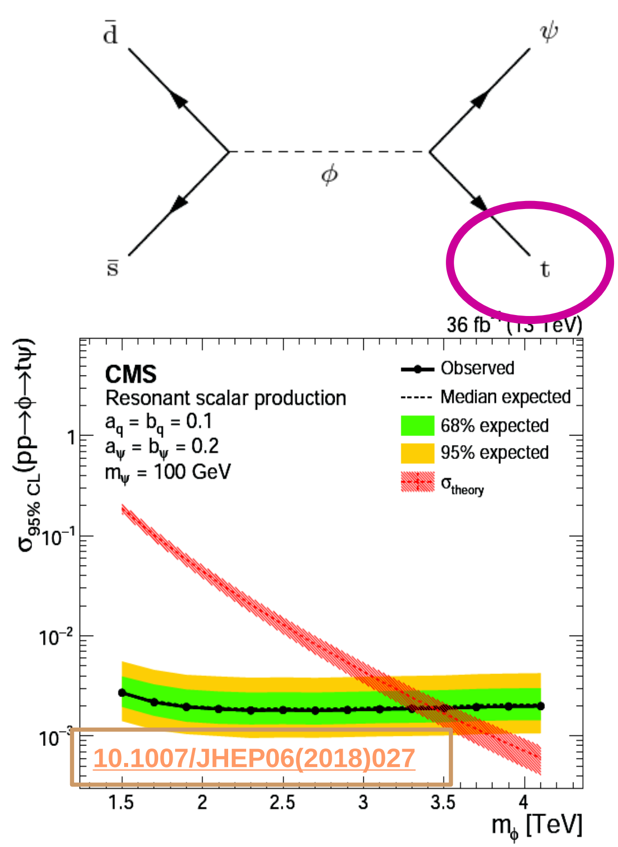
SUSY (top squarks)



Heavy Resonance (VV)



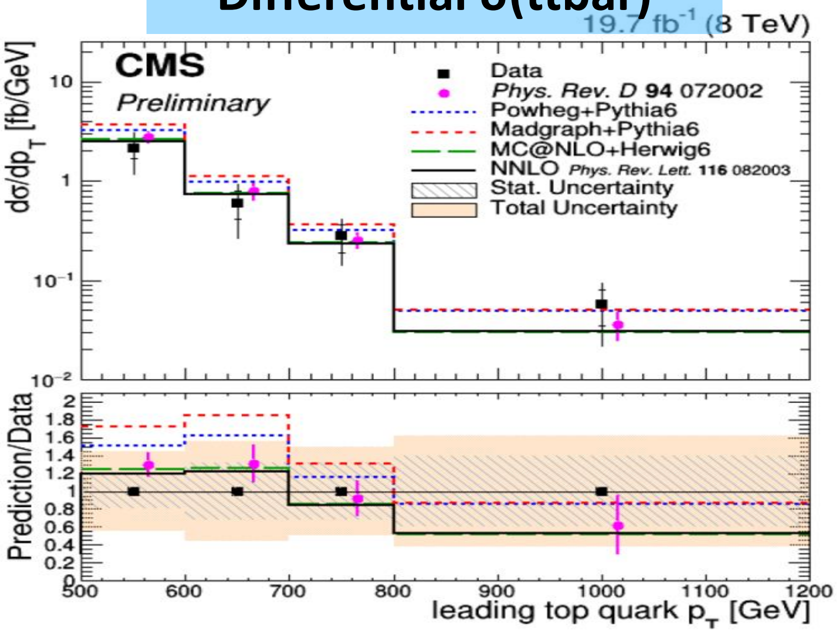
DM (mono-top)



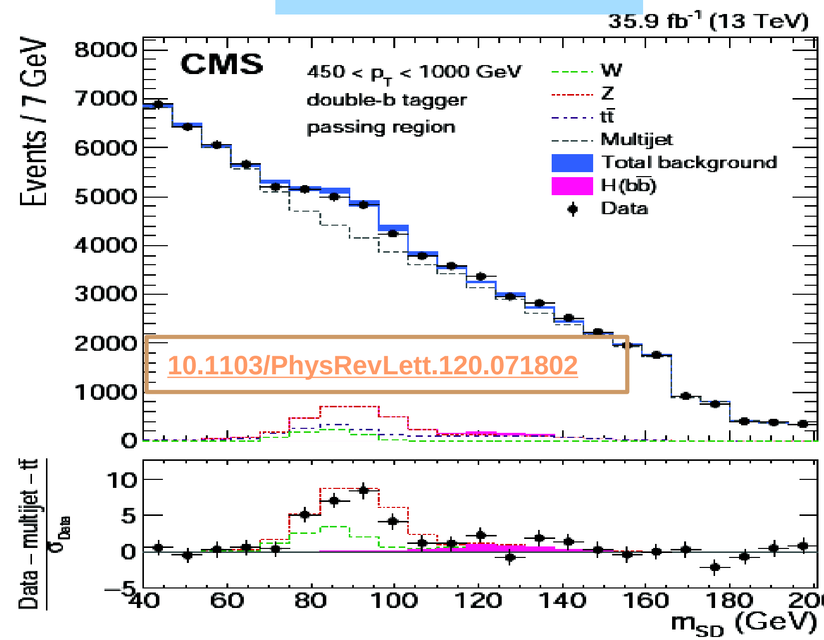
Physics case: "Boosted" jet tagging

- We have known for a long time that boosted jet tagging has great potential and we are seeing more and more new and important results because of it, such as these here....

Differential σ(ttbar)



H -> bb

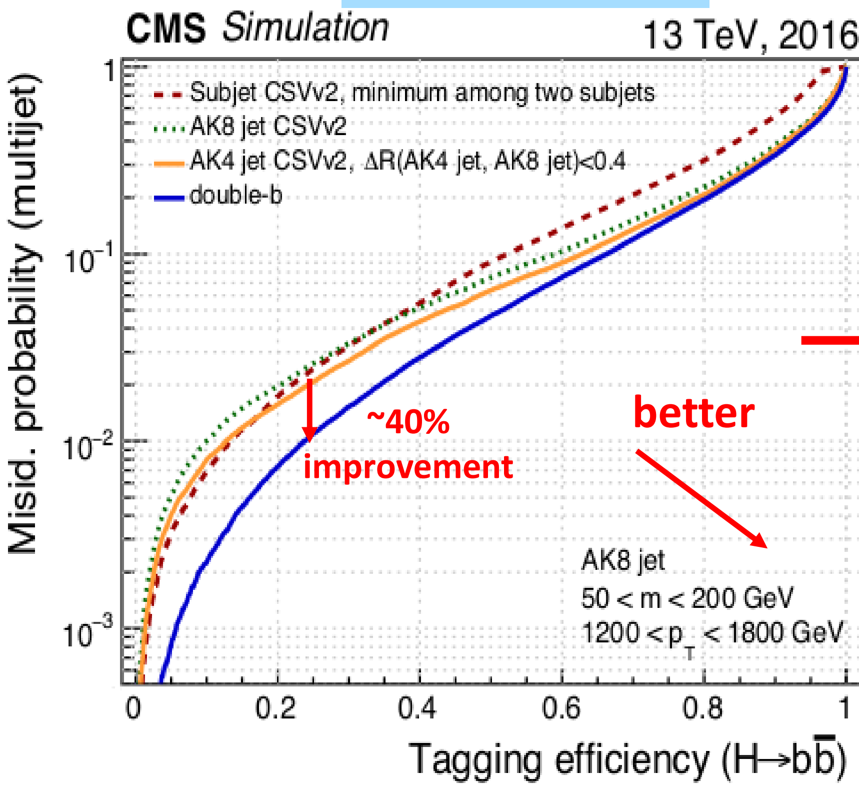


Precision measurements can benefit from improved boosted object reconstruction while keeping <u>systematic uncertainties under control</u>

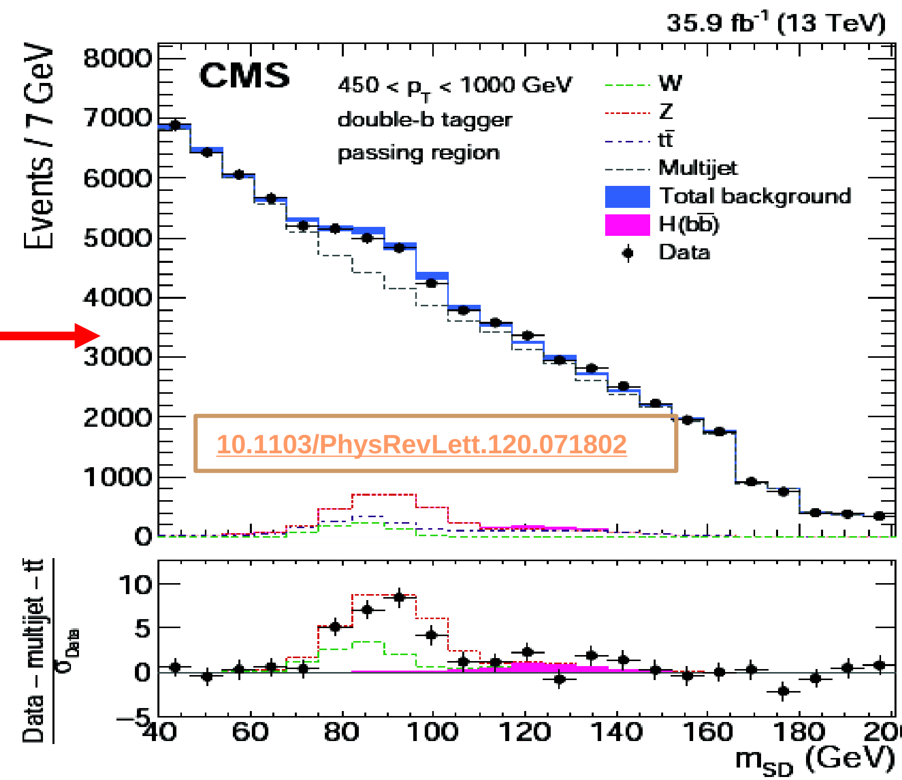
Boosted jet tagging: LHC Run 2 (2)

- The “Double-b” tagger: Identification of Z/H/X -> bb against QCD
  - Exploits tracking, vertexing & substructure observables as inputs to a BDT
  - Designed to be largely independent on jet mass & jet pT

Double-b



H -> bb



10.1103/PhysRevLett.120.071802