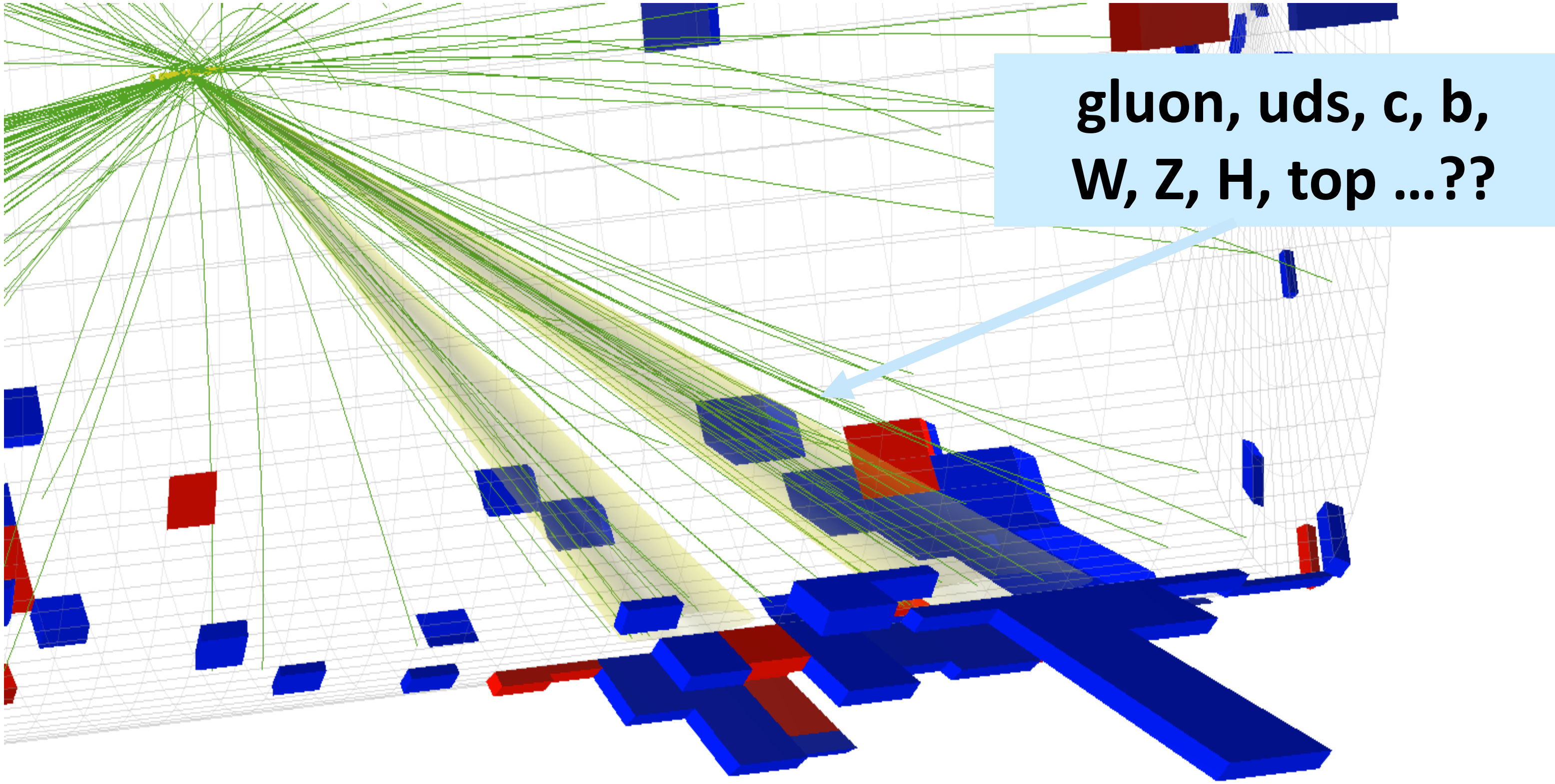
Review of flavor tagging algorithms at pp and ee and plans

Loukas Gouskos (CERN) Michele Selvaggi (CERN)

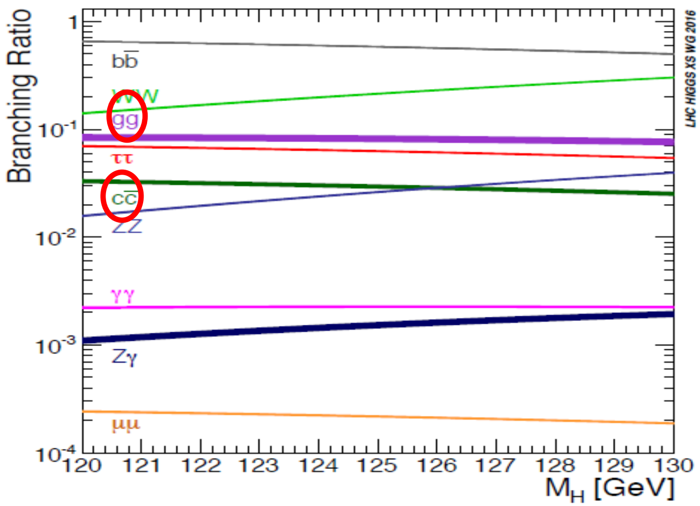
Introduction



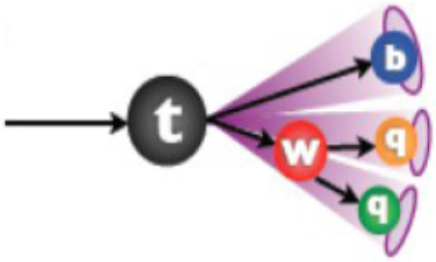
- People have been tagging jets for more than 30 years at colliders
  - starting with b jets at LEP and Tevatron, then top, W/Z and Higgs jets at the LHC.
- But it is only now that we have begun to develop powerful and multi-object tagging capabilities.
  - potential to open access to many new physics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

Physics motivation (ee collisions)

- Flavour tagging essential for the e^+e^- program, e.g.:
  - Higgs Sector:
    - (HL-)LHC can access 3rd gen. couplings and a few of 2nd generation
    - Future e^+e^-: Measure Higgs particle properties and interactions in challenging decay modes
      - E.g. cc, 1st gen quarks/fermions, gg [?]



  - Top quark physics [if E_{CM} sufficient]
    - Precise determination of top properties [mass, width, Yukawa]



Goal of this talk:
Discuss jet flavour tagging methods developed for the pp physics program at the LHC that could be explored at e^+e^- experiments as well and develop a tentative plan of action

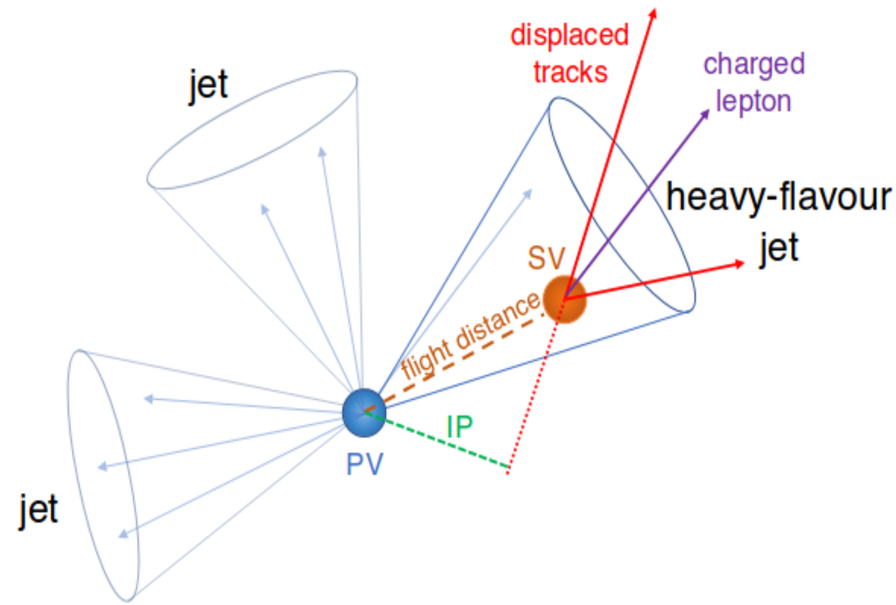
[Disclaimer: Focus on CMS results; similar methods developed by the other LHC experiments]

Flavour tagging in pp colliders

[Focus on the latest developments]

Heavy flavor (b/c) tagging basics

- Large lifetime:
  - b (c) ~ 1.5 (<1) ps
  - b (c) decay length: ~5 (2-3) mm for p_T~50 GeV
- Displaced vertices/tracks
  - Large impact parameters
  - Tertiary vertices when B decays to charm hadron
- Large track multiplicity
  - ~5 (~2) charged tracks/decay in b (c)
- Non-isolated leptons
  - from B/C decays
- Harder fragmentation wrt to light quarks
  - ~75 (50)% of the jet energy carried out by b (c) hadrons

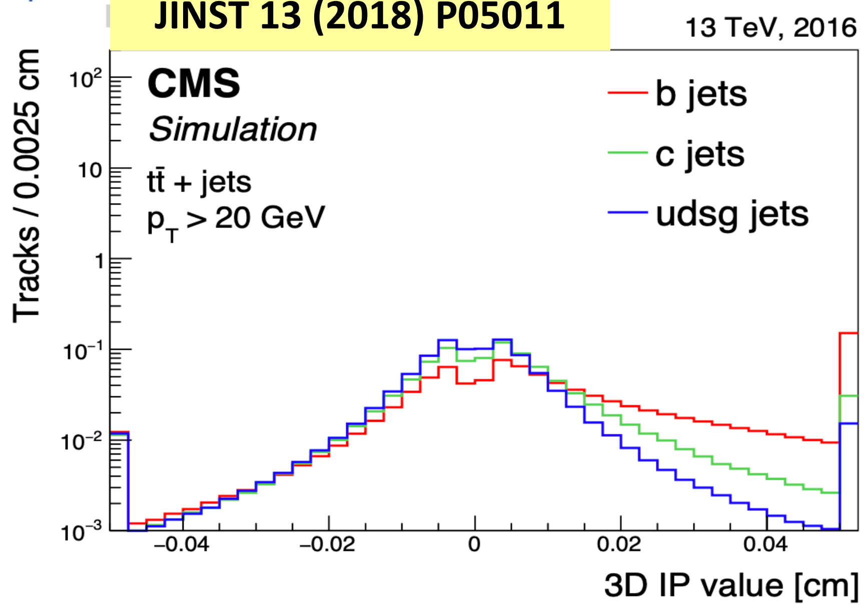


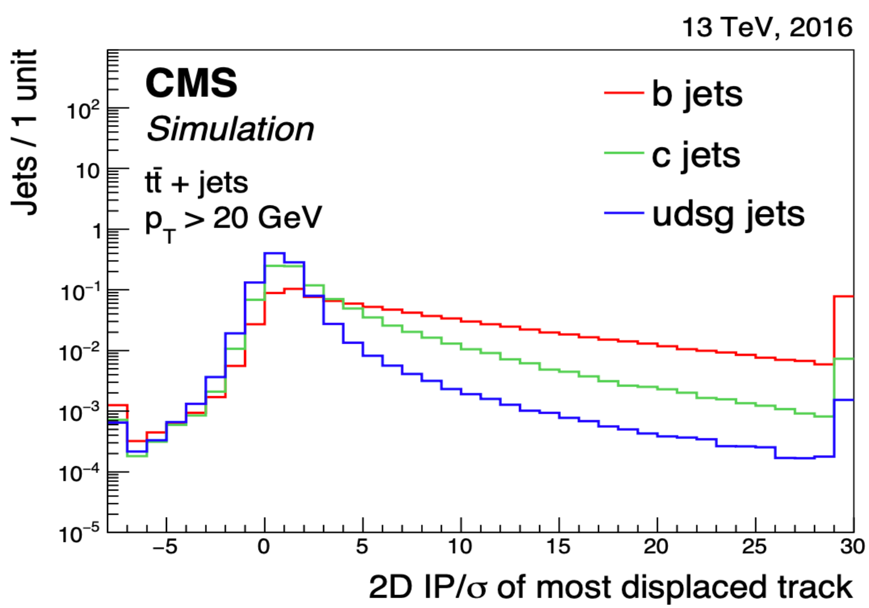
Charm has intermediate properties between **light** and **b-jets**

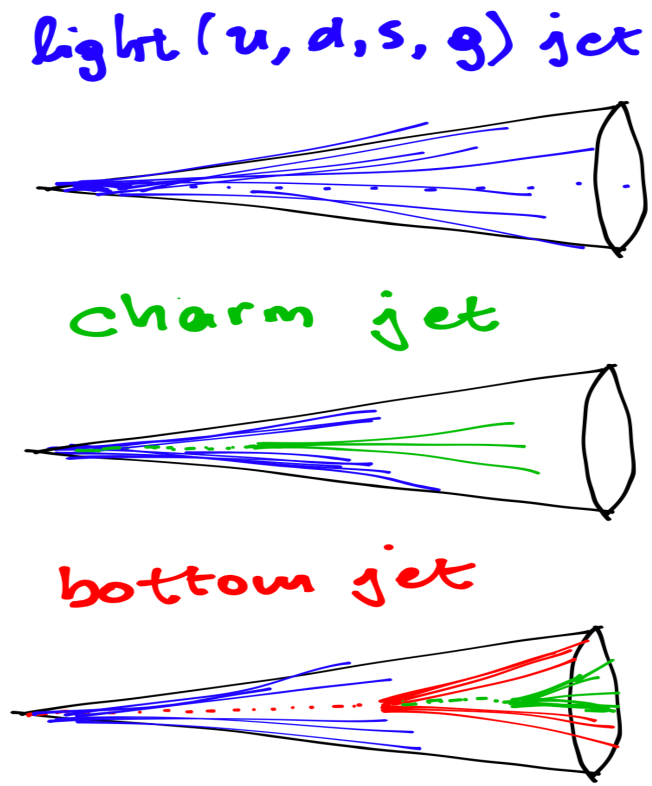
Detector side:
Need powerful pixel/tracking detectors
→ good spatial resolution
→ as little material as possible
→ precise track alignment

Heavy flavor (b/c) tagging basics (II)

JINST 13 (2018) P05011







→ Based on these properties develop "human-inspired" (high-level) variables

→ Usually used as inputs to a Machine Learning (ML) algorithm

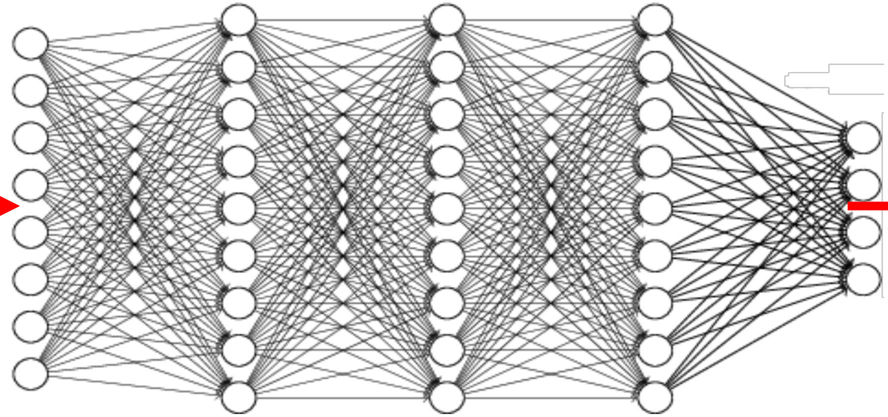
b/c tagging

- Exploit Deep Neural Networks (DNN) to improve b/c tagging
  - From BDT -> [simple] Dense Network (5 Hidden layers, 100 nodes)

Inputs

→ Use "human-made" (high-level) inputs

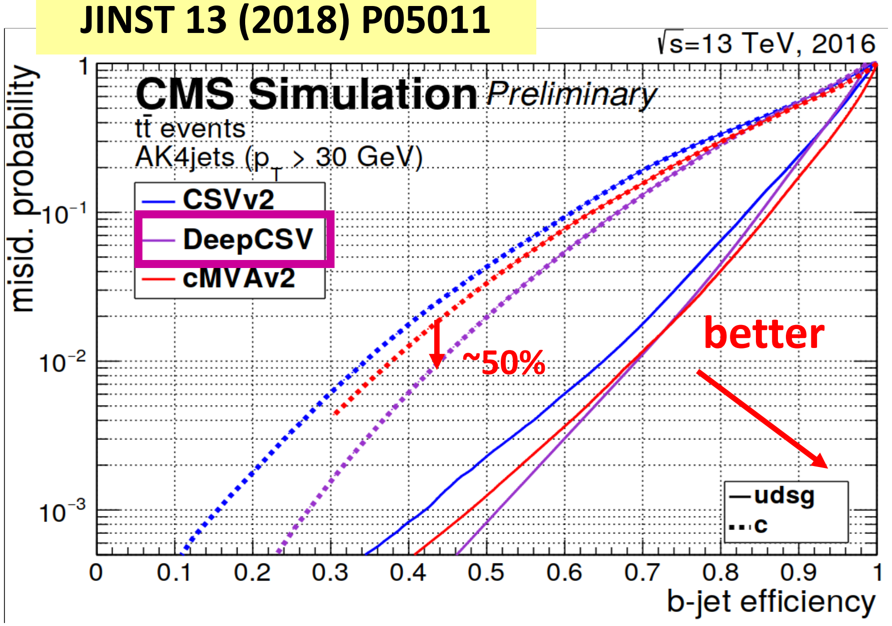
→ few more tacks (up to 6) wrt traditional approaches



Output

Multiclass classifier: b, bb, c, udsg

JINST 13 (2018) P05011

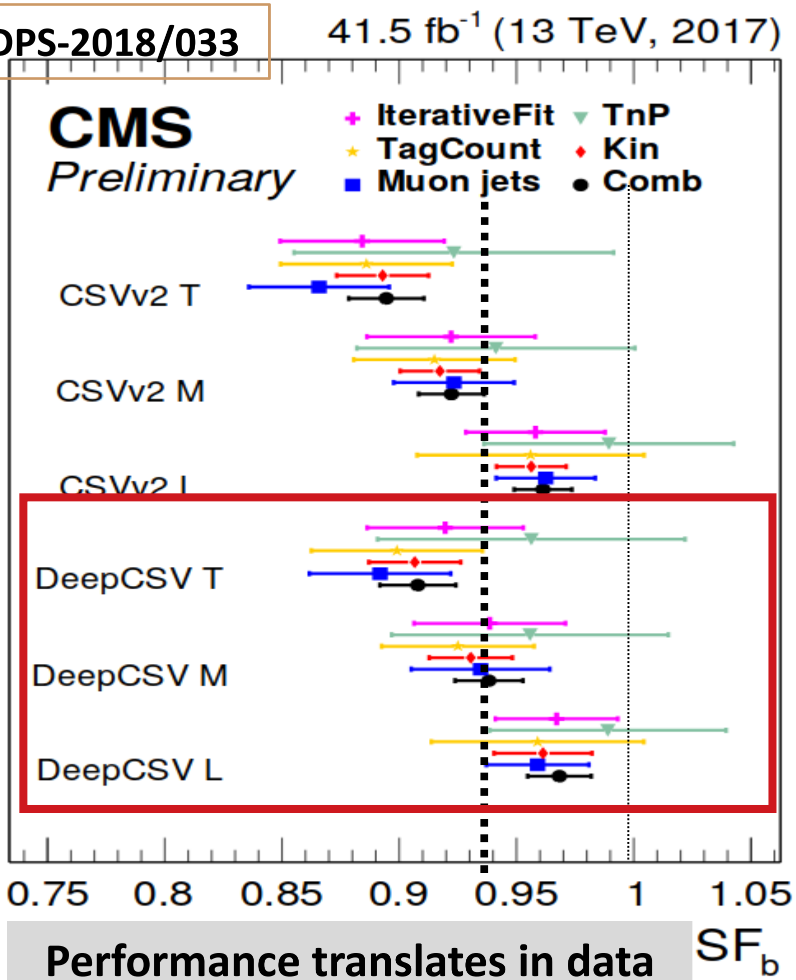


Significant gain in performance with respect to traditional methods [i.e. CSVv2, cMVAv2]

b/c tagging

Performance in data

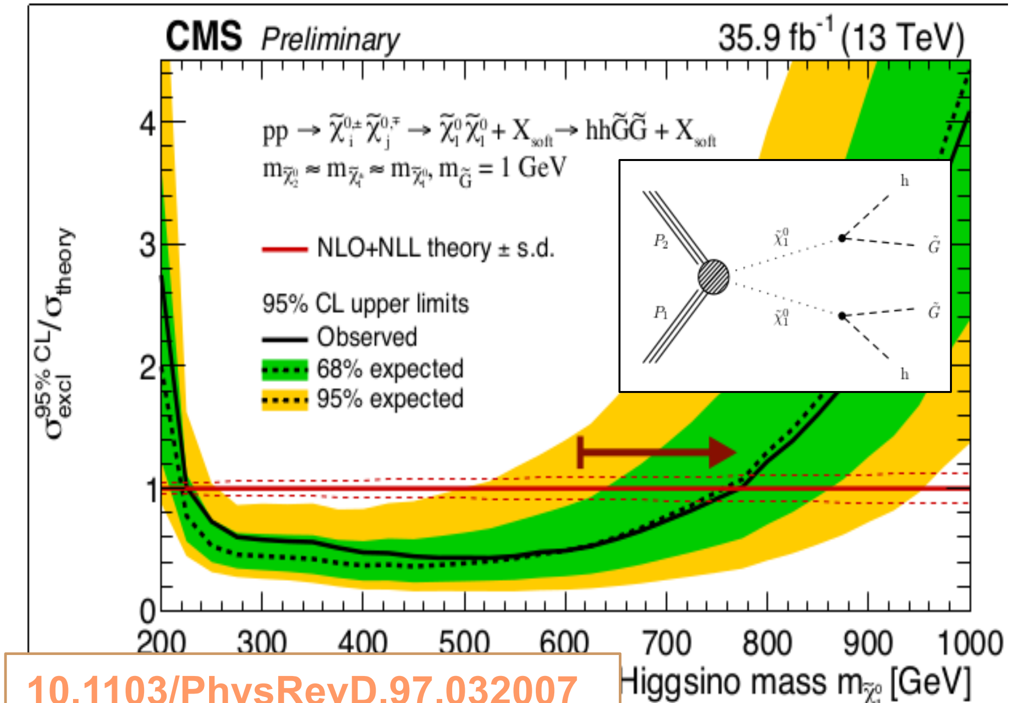
DPS-2018/033



Performance translates in data

Default in CMS since 2017

Impact on physics analyses



10.1103/PhysRevD.97.032007

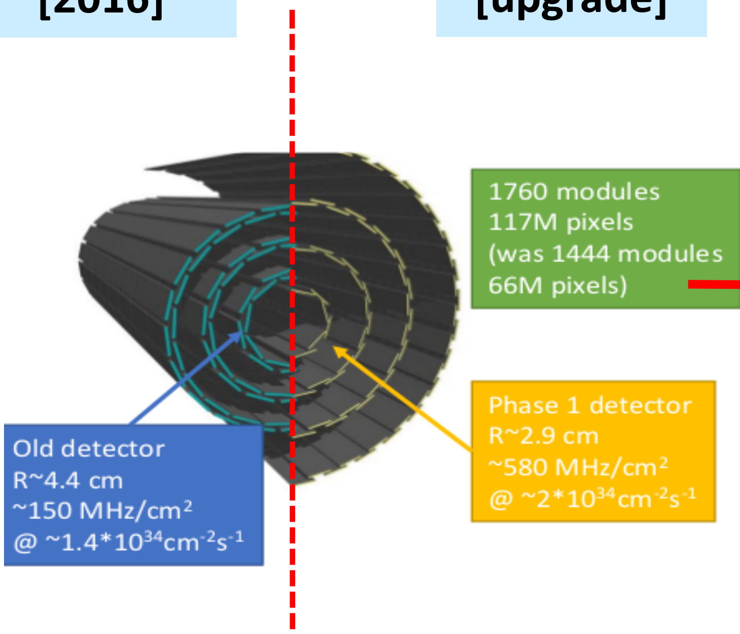
→ **Significant gain in sensitivity in final states with multiple b quarks**

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

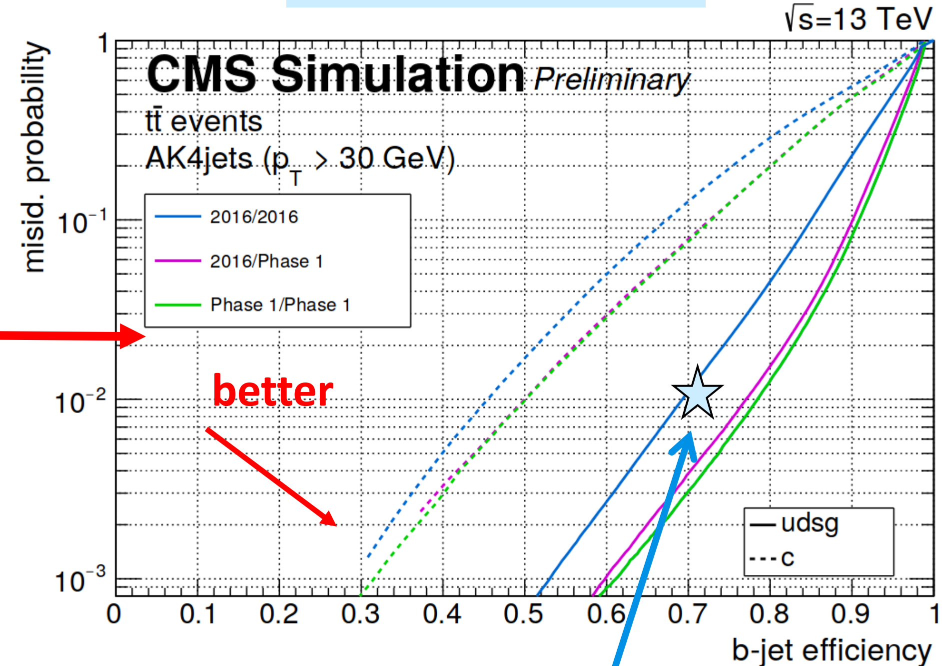
b/c tagging: detector upgrades

Phase 0 [2016]

Phase 1 [upgrade]



Phase 1 [upgrade]



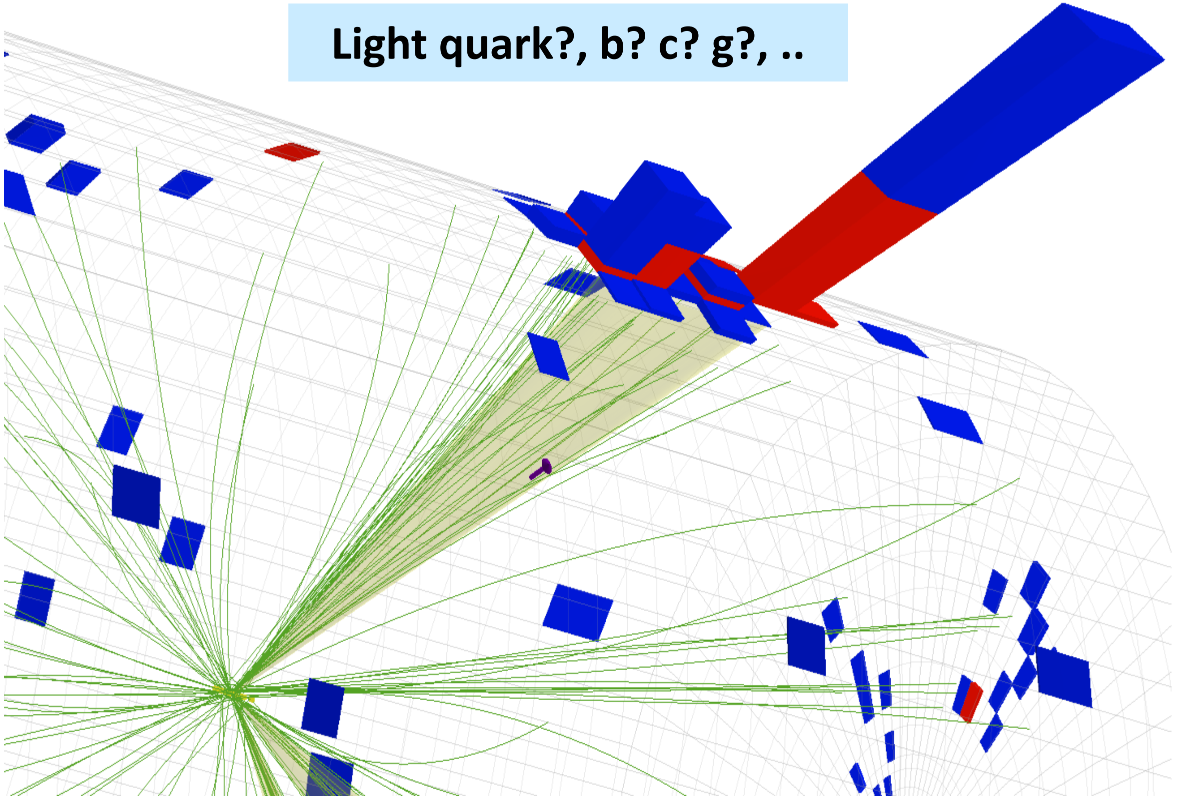
Performance of traditional algorithms [i.e. BDT-based] with upgraded detector

→ DNN-based flavour tagging [with old detector] ~similar performance with traditional flavour tagging algos using the upgraded detector

→ Significant gain in performance with the upgraded pixel detector

Exploring more of the detector's potential

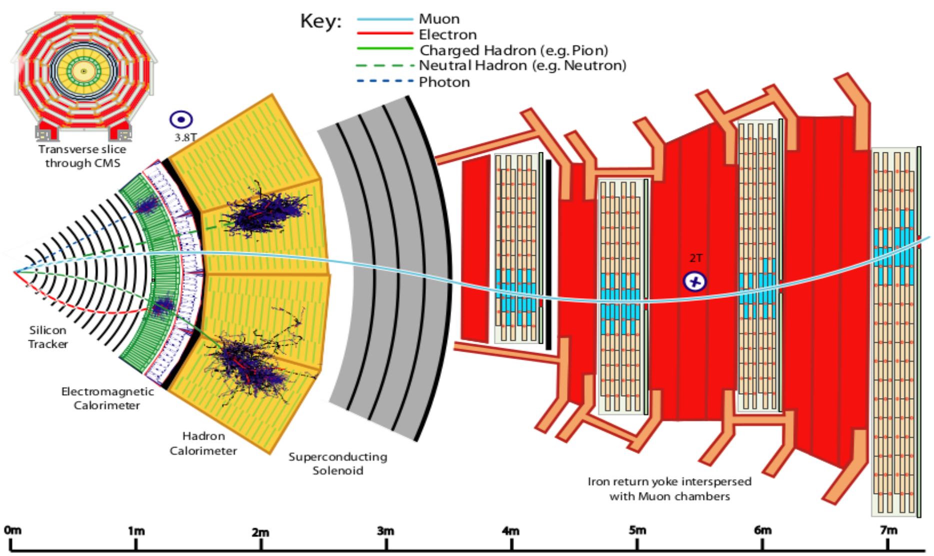
- **A Jet in theory:** Spray of particles produced by the hadronization of quarks and gluons
- **Experimentally:** A cone of reconstructed particles in the detector



- Can we gain by moving to **particle based jet tagging** with DNN?

Exploring more of the detector's potential (II)

- Event reconstruction at LHC & future experiments (will) have some flavour of PF event reconstruction:
  - Combines information from all subdetectors
  - Mutually exclusive list of particles



- Rich information for each particle
  - Energy/momentum
  - Position
  - Particle category
  - Displacement from the PV
  - Reconstruction quality
  - ….

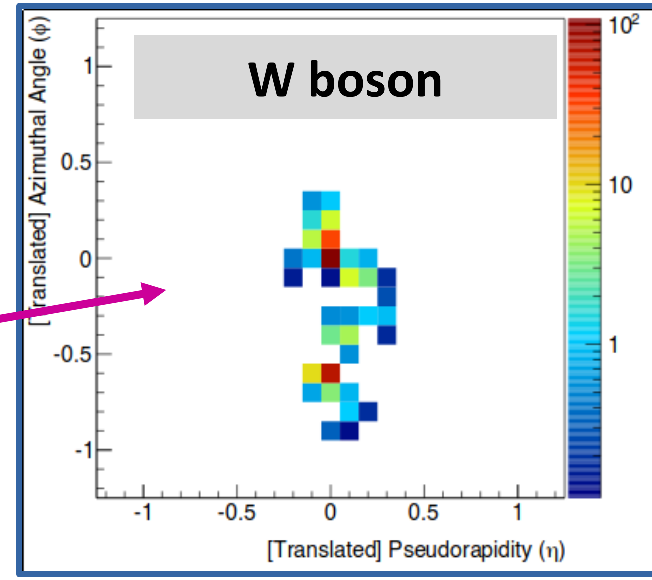
Energy/momentum, Position → **Inputs to jet substructure**

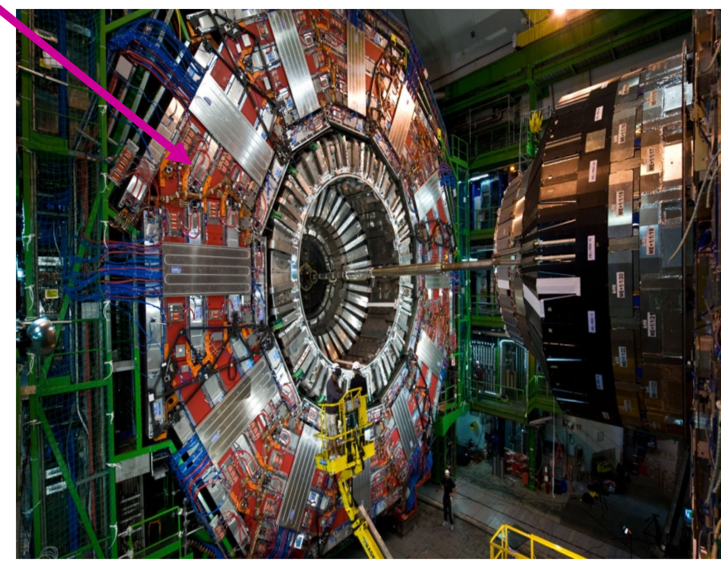
Particle category, Displacement from the PV, Reconstruction quality, …. → **Inputs for flavour tagging**

- [O(50) properties/particle] x [~50-100 particles/jet] ~O(1000) inputs/jet
- **Perfect case** for DNN with "complex" architecture

Deep learning approaches for jet tagging

- 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)
  - **But:** jet images are very sparse
  - **Also:** LHC & future detectors are very heterogeneous/complex **not "image-like"**
    - difficult to include information from other subdetectors (e.g. tracking)

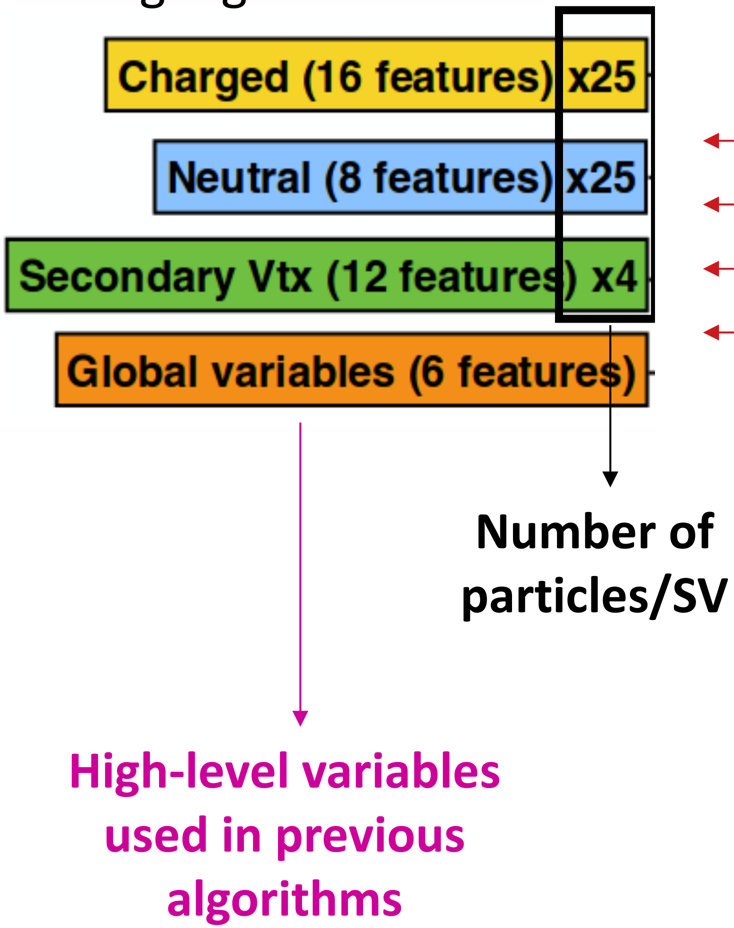




- Based on **particle sequence**:
  - **Jet** as a **sequence of constituent particles**
  - Apply techniques used for **natural language processing** [e.g. CNN-1D,..]
  - Inclusion of more information straight forward
  - Explore **more** of the detector & **event reconstruction potential**

ParticleBased jet tagging: Inputs

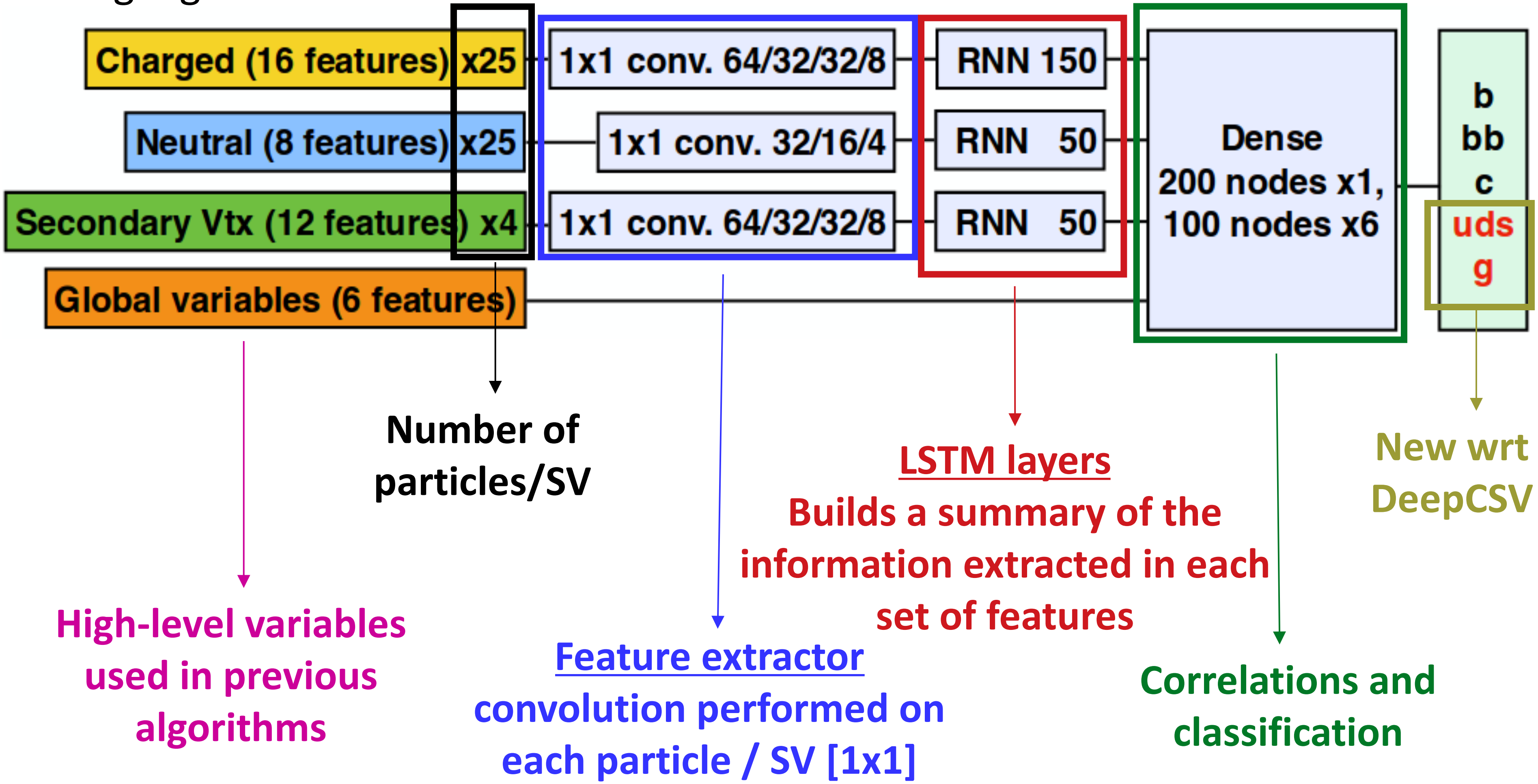
- Treat the jet as a particle sequence and develop a multiclass classifier for: b, bb, c, uds, gluon tagging
- Highlights from the archi



- "Low level" inputs:
  - p_T, η, ϕ of PF candidates
  - Particle ID
  - Impact parameters & significance of charged tracks - PV
  - Various track parameters, etc…
  - p_T, η, ϕ of secondary vertices within jet cone

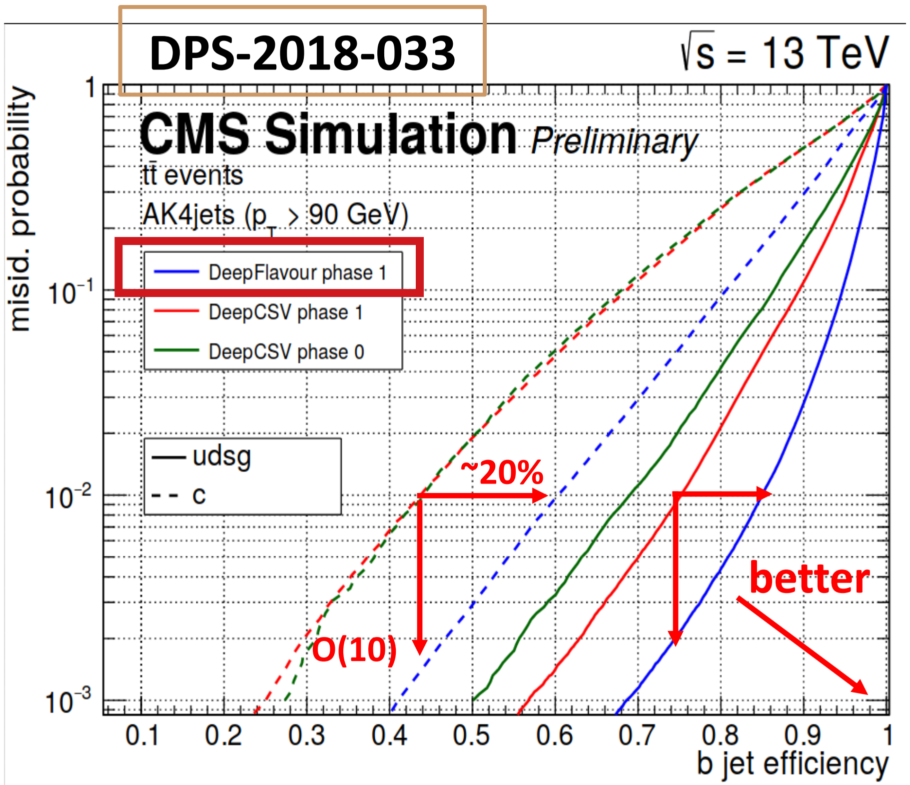
ParticleBased jet tagging

- Treat the jet as a particle sequence and develop a multiclass classifier for: b, bb, c, uds, gluon tagging
- Highlights from the architecture:



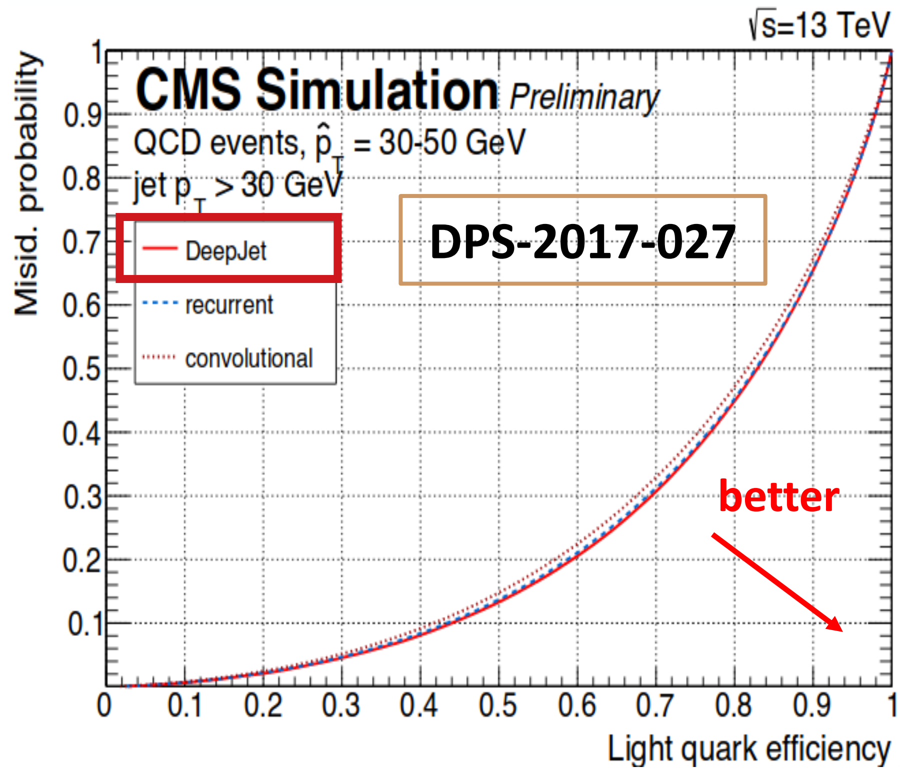
Performance

b tagging



- → 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

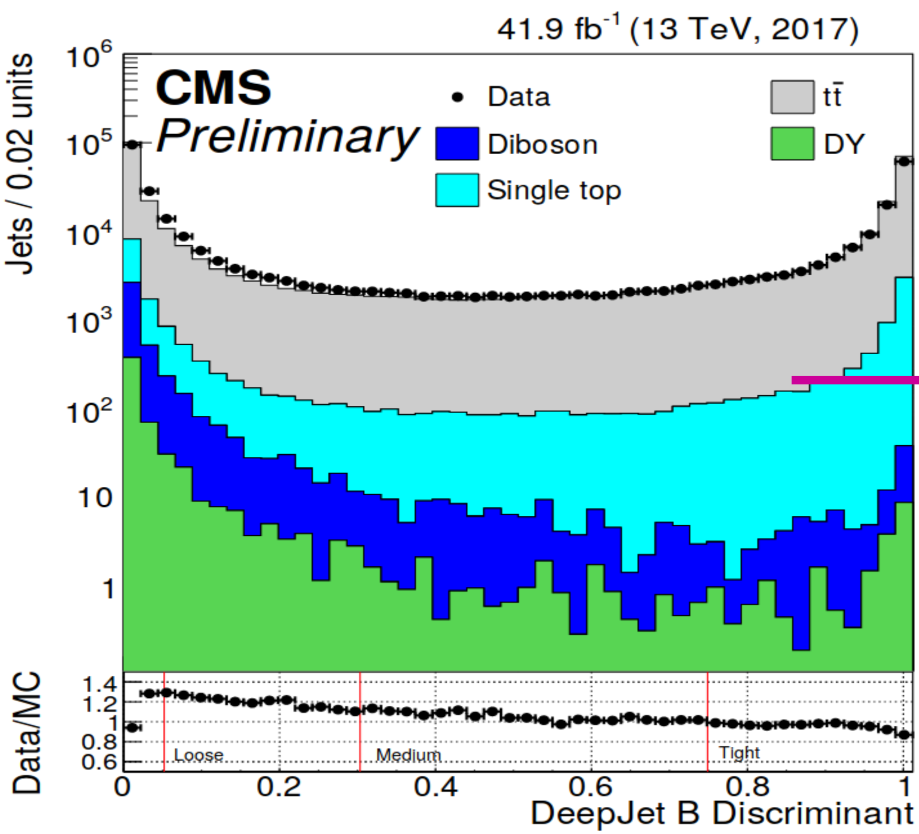
quark/gluon tagger

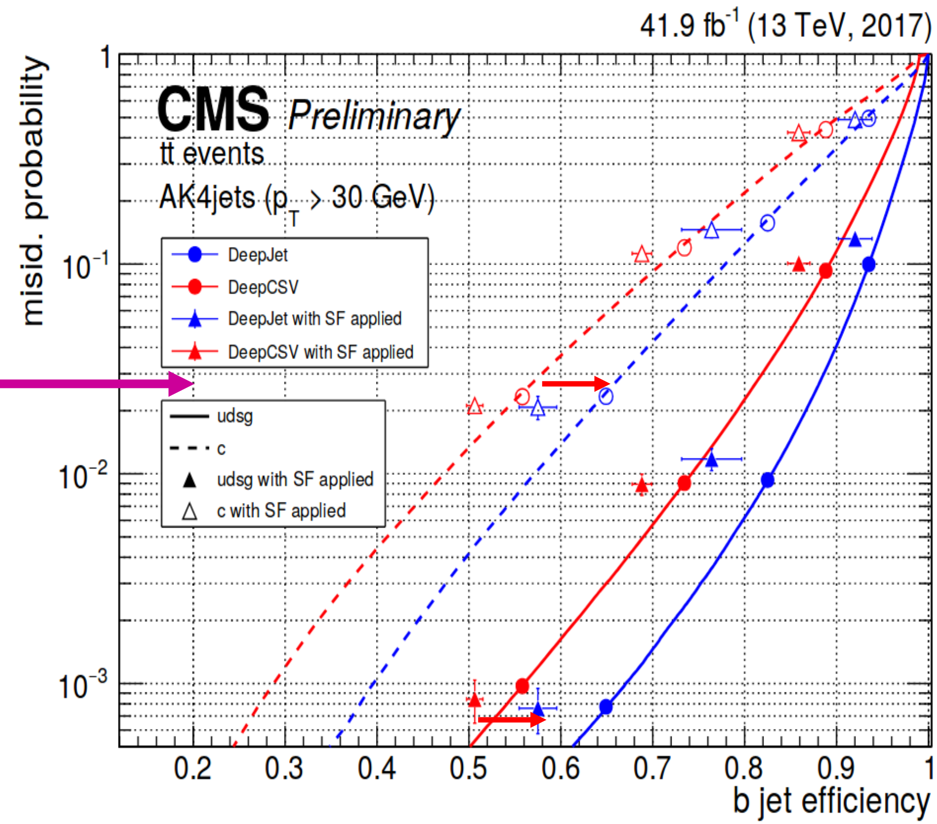


- → Generator level light quarks/gluons that did not split to heavy flavour
- → Similar performance to dedicated implementations

Calibration in data

- Three main data samples with different flavour composition
  - ttbar [b-enhanced], W+c [c-enhanced] and multijet [uds/g enhanced]





Good overall Data/MC aggreement

Gain remains after accounting for the efficiency/mistagging correction factors

Teaser: "ParticleNet" for jet tagging

- How to represent a jet is one of the key aspects of ML algos for jet physics
  - Improve performance → extend physics reach
  - Lead to fresh insight into jets → deepen our understanding of jet physics
- Inspired by the "point cloud" approach introduced **"Particle cloud"** for jets

Point cloud

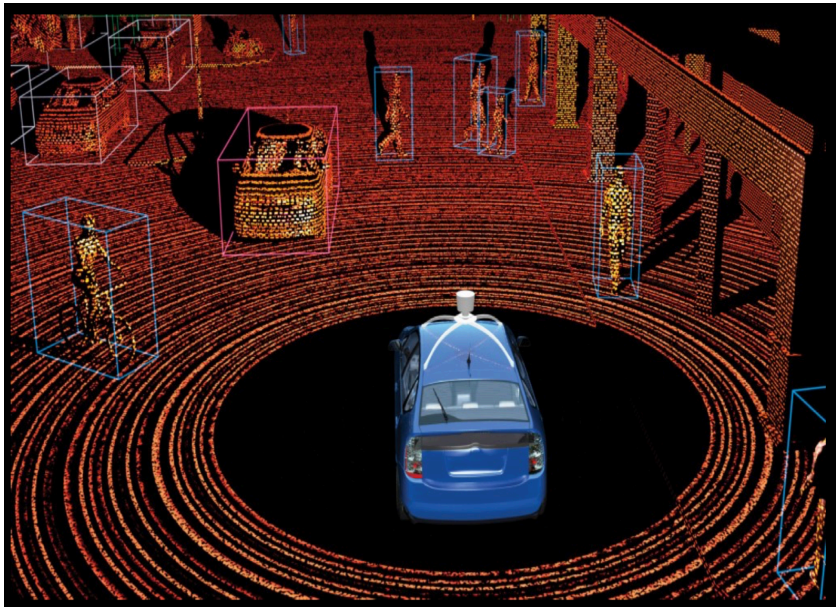
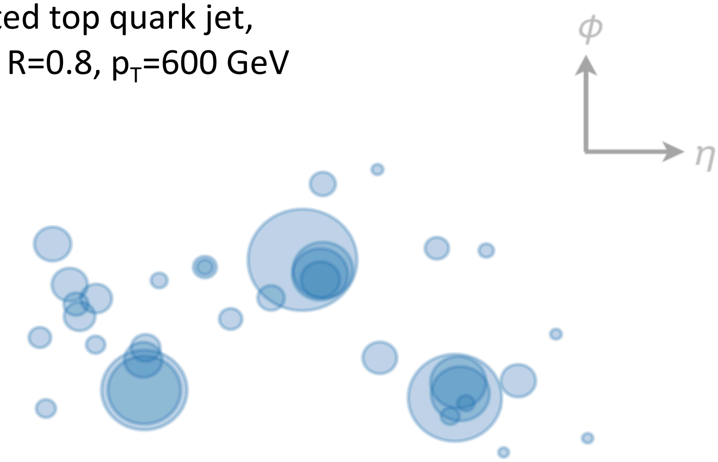


Image from: https://news.voyage.auto/an-introduction-to-lidar-the-key-self-driving-car-sensor-a7e405590cff

Particle cloud

Simulated top quark jet,
Anti-k_T, R=0.8, p_T=600 GeV

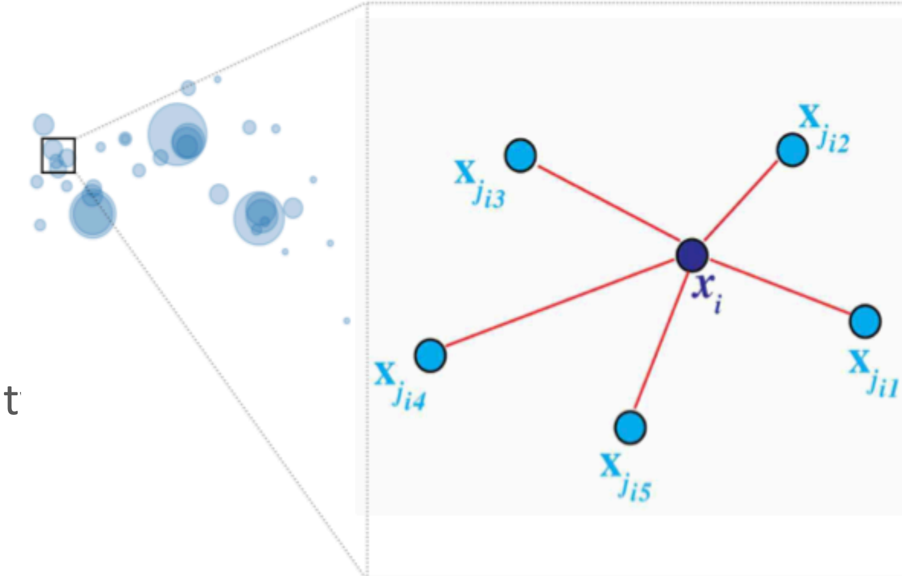


ParticleNet [Huilin Qu, LG]

PRD 101 (2020) 5, 056019
CMS-DP-2020-002

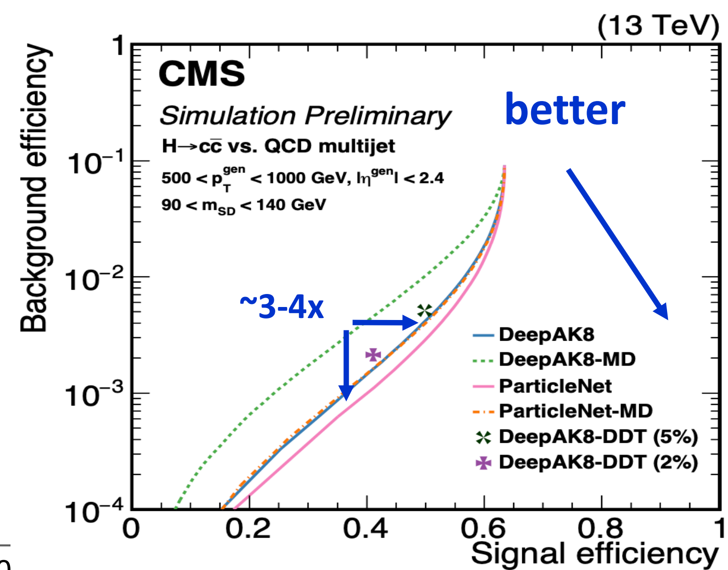
Teaser: "ParticleNet" for jet tagging

- **Jet representation:** particle cloud
  - Particles are intrinsically unordered
    - Primary info: 2D coordinates in the η-ϕ space
  - But also exploit additional info
    - Energy, momentum, charge, particle t
    - Track quality, displacements, ..



- **Network architecture:** Dynamic Graph CNN (DGCNN)
  - Treat the particle cloud as a graph: each point is a vertex
  - for each point, a local patch is defined by finding its k-nearest neighbors

H→cc



Top tagging:

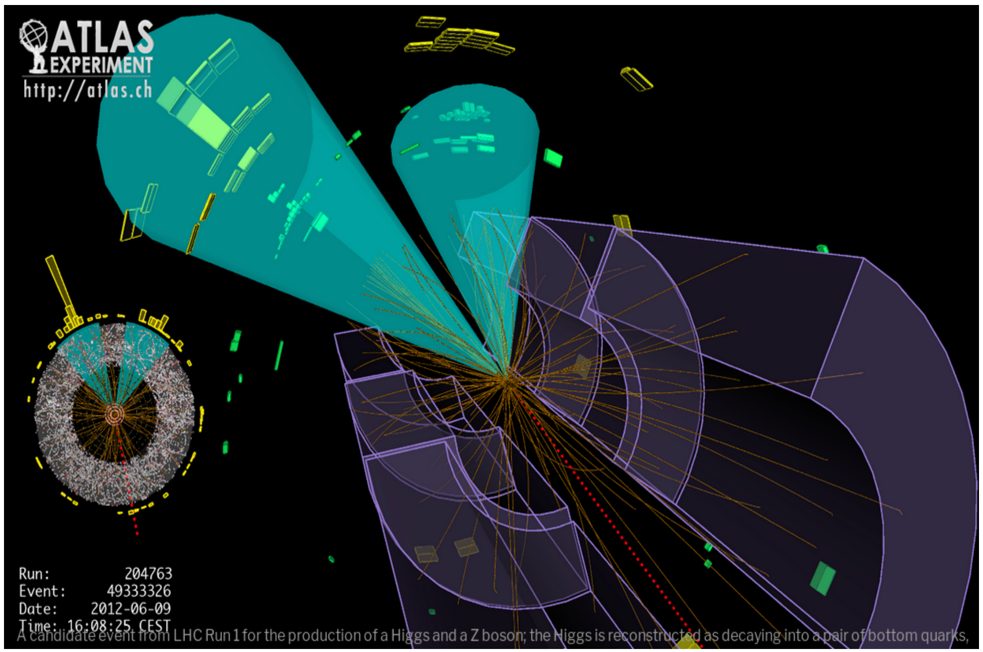
	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$	$1/\varepsilon_b$ at $\varepsilon_s = 30\%$
ResNeXt-50	0.936	0.9837	302 ± 5	1147 ± 58
P-CNN	0.930	0.9803	201 ± 4	759 ± 24
PFN	-	0.9819	247 ± 3	888 ± 17
ParticleNet-Lite	0.937	0.9844	325 ± 5	1262 ± 49
ParticleNet	**0.940**	**0.9858**	$\mathbf{397 \pm 7}$	$\mathbf{1615 \pm 93}$

Existing flavour tagging tools in e^+e^- colliders

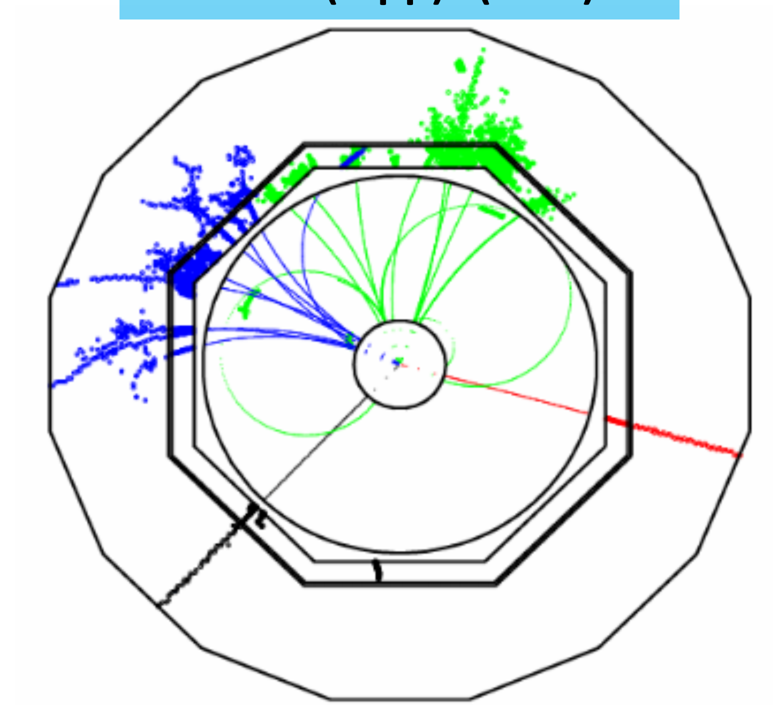
From pp to e^+e^-

- e^+e^- colliders provide a very clean environment
  - Lower occupancy, no pileup

LHC: Z(->νν)H(->bb)



e^+e^-: Z(->μμ)H(->bb)



From pp to e^+e^-

- e^+e^- colliders provide a very clean environment
  - Lower occupancy, no pileup

- Future e^+e^- pixel/tracking detectors tailored for b/c tagging
  - Higher granularity wrt to LHC detectors
    - ATLAS/CMS pixel size: O(~100x100 μm^2)
  - Less tracking material
    - ~0.4% X_0/layer CMS/ATLAS Pixel, ~0.15-0.2% X_0/layer in e^+e^- detectors
      - better impact parameter resolution/ less multiple scattering
      - CMS/ATLAS Pixel resolution: O(10) μm; ~2-5 μm in e^+e^-
  - Smaller fluence in e^+e^-: allows PIXEL detectors with layers closer to beam pipe
  - More precise track reconstruction in e^+e^- due to more hits/layer in the pixel/tracker detectors

Jet flavour tagging in e^+e^-

- Algorithms under development follow similar-ish methods as the previous generation of algorithms used in pp and past e^+e^-

- Use as an example the "LCFIPlus" flavour tagging tool developed for ILC/CLIC

NIM A 808 (2016) 109–116

Main steps in a nutshell:

- Vertex finding:
  - Identify primary and secondary vertices (PV and SV)
- Jet clustering: optimized for flavour tagging [also in multijet final states]
  - SV and leptons [from B/C decays] used as seed for jet clustering
  - Different clustering algorithms explored
- Jet vertex refiner:
  - Uses jet information to improve the b/c reconstruction
    - single tracks pass some selection are also considered as "pseudo-vertices"
  - Combine vertices to reduce #SV down to at most two
- Flavour tagging: exploit shallow-ML [BDT]
  - Inputs: high-level variables based on PV, SV, track and jet properties
  - Output: "b", "c" and "uds" [no gluon-tagging]

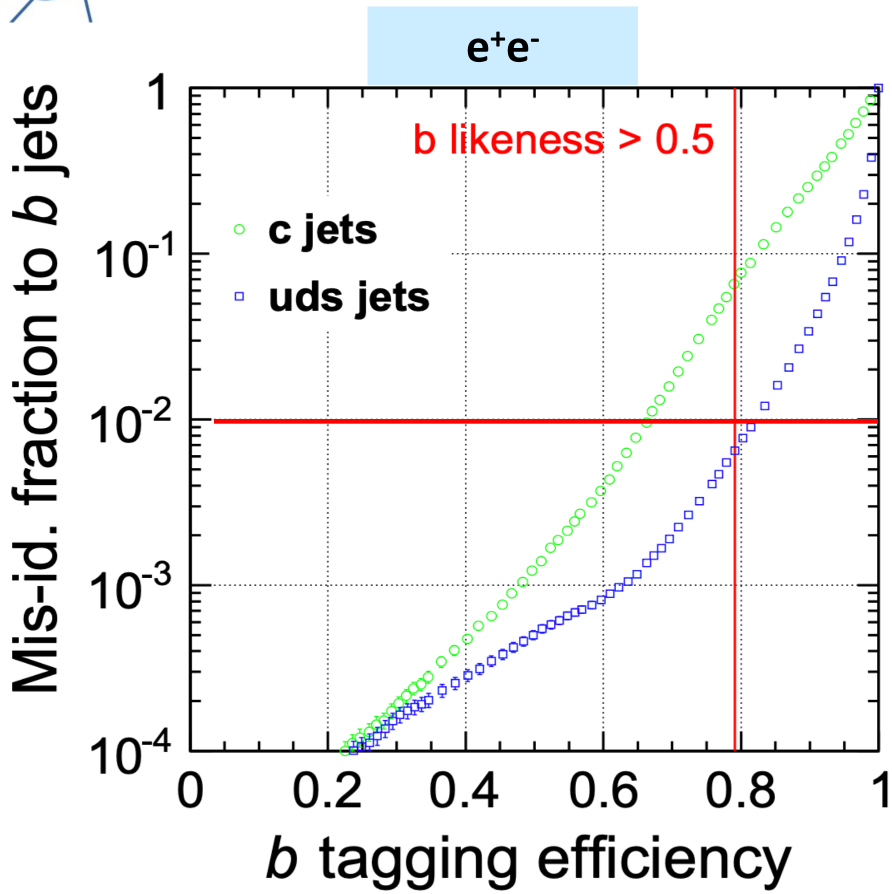
jet flavour tagging in e^+e^-: Inputs

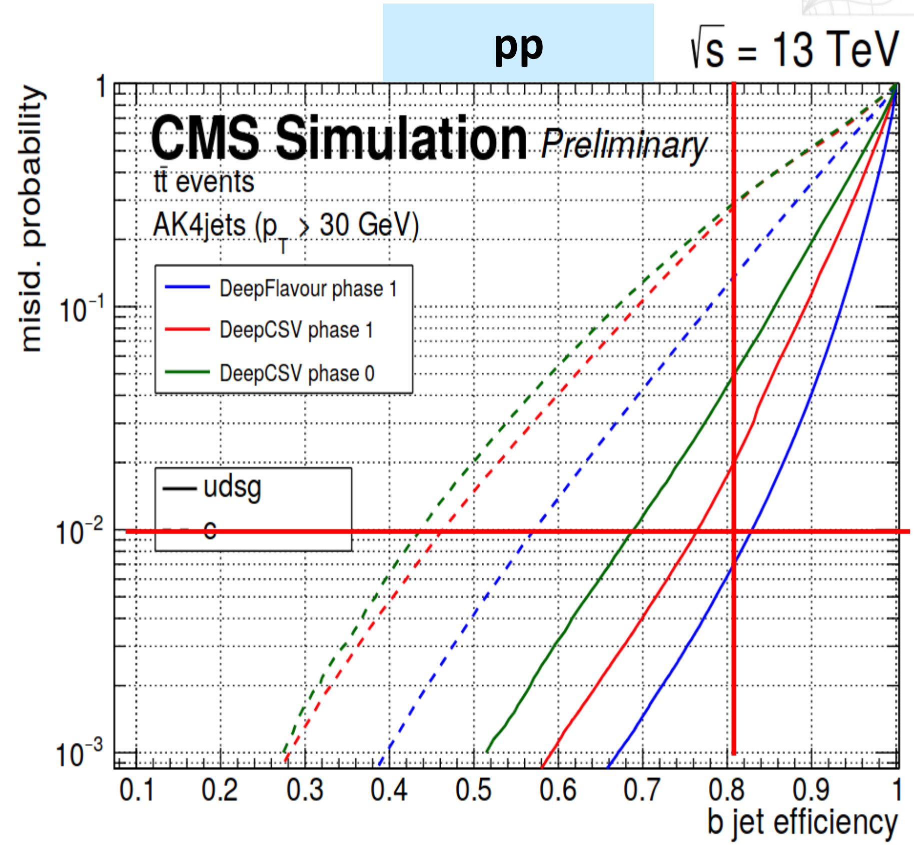
- Inputs categorized based on the number of SV (N_{SV})
  - $N_{SV} = 0$:
    - pT and displacement of the two tracks with highest sig(d0)
    - high-level variables based on tracks associated to jet/SV
    - muon and electron multiplicity
  - $N_{SV} = 1, =2$:
    - Additional variables related to the SV properties
    - Correlations between the SVs [when $N_{SV} = 2$]
    - etc…

Compared to taggers developed for pp
→ Less input variables and fewer features per track/SV
→ Simpler network architecture

Room for improvement:
→ Exploit advanced network architectures and lower-level features
→ Inclusion of features relevant for gluon-tagging

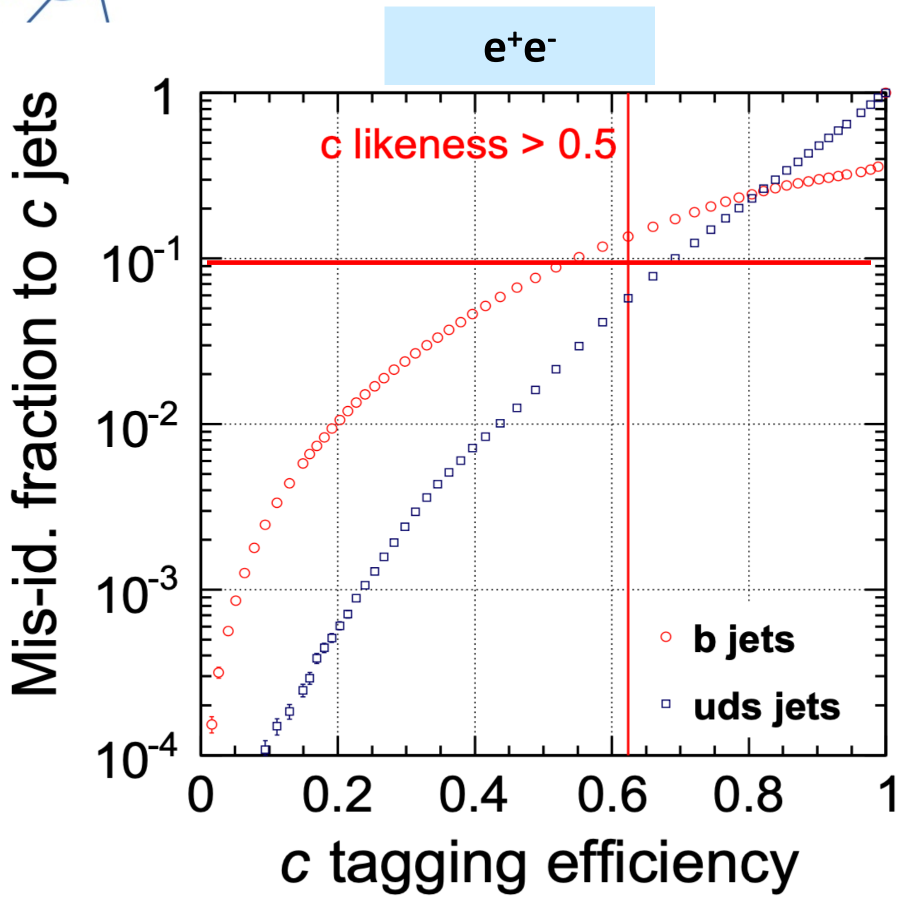
b-tagging performance

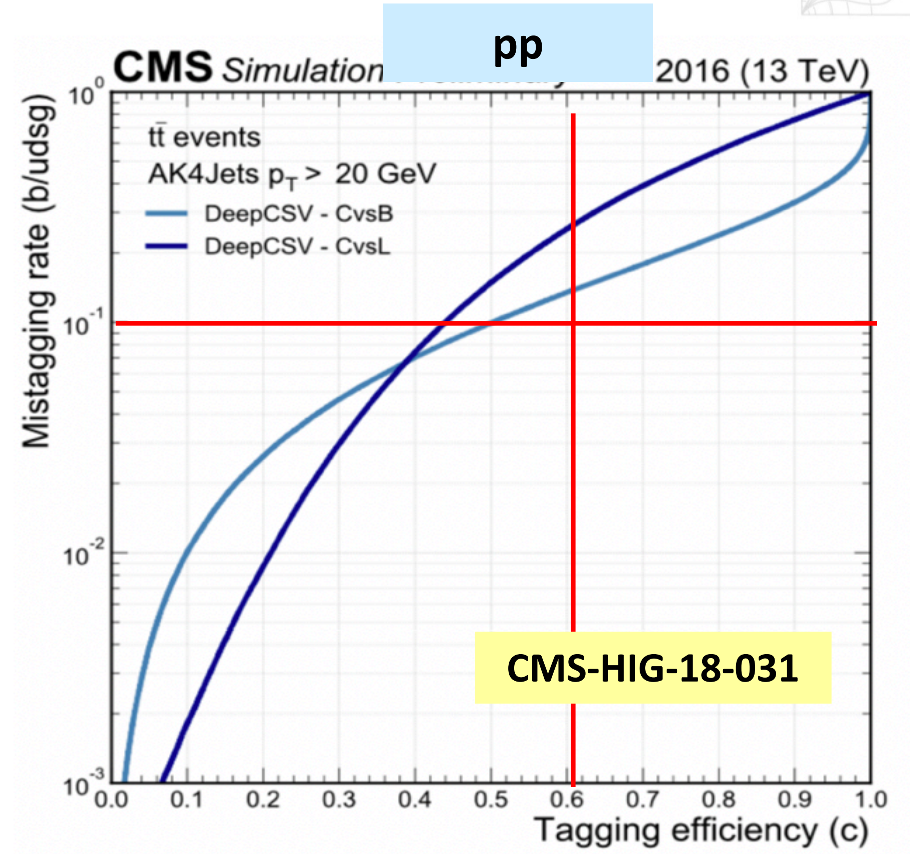




- Similar performance [take it with a lot of grain of salt]
  - yet conditions and detector potential very different [favoring the e^+e^- case]
- Definitely worth exloring the recent developments in pp colliders
  - (a) Improve performance and/or achieve necessary performance with less complex (cheaper) detector solutions

c-tagging performance

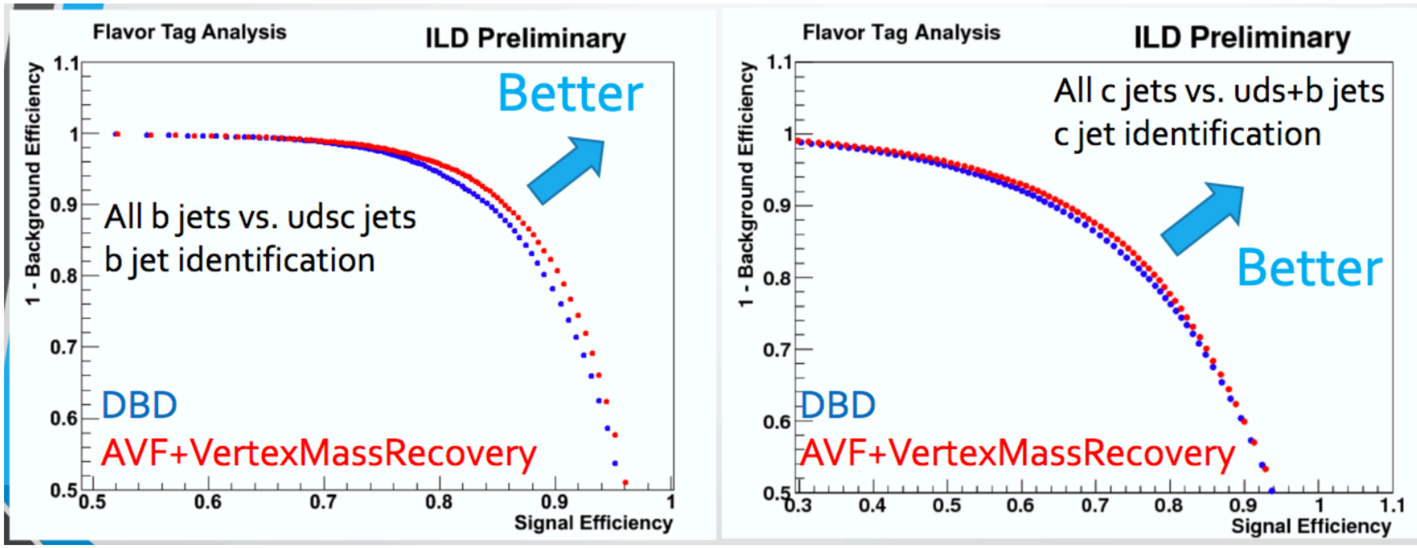




- **Charm-bottom separation:** similar performance
- **Charm-light separation:** e^+e^- shows better performance, but for the pp case:
  - results derived <u>before</u> the upgraded PIXEL detector
  - Algorithm does not explore the latest tagging developments
    - i.e. low-level features and advanced ML architecture

Possible improvements

- Based on a dedicated talk at the CLIC workshop in 2019 [slides]
  - Adaptive Vertex Fitting
    - currently strict track selection applied to prevent fake tracks



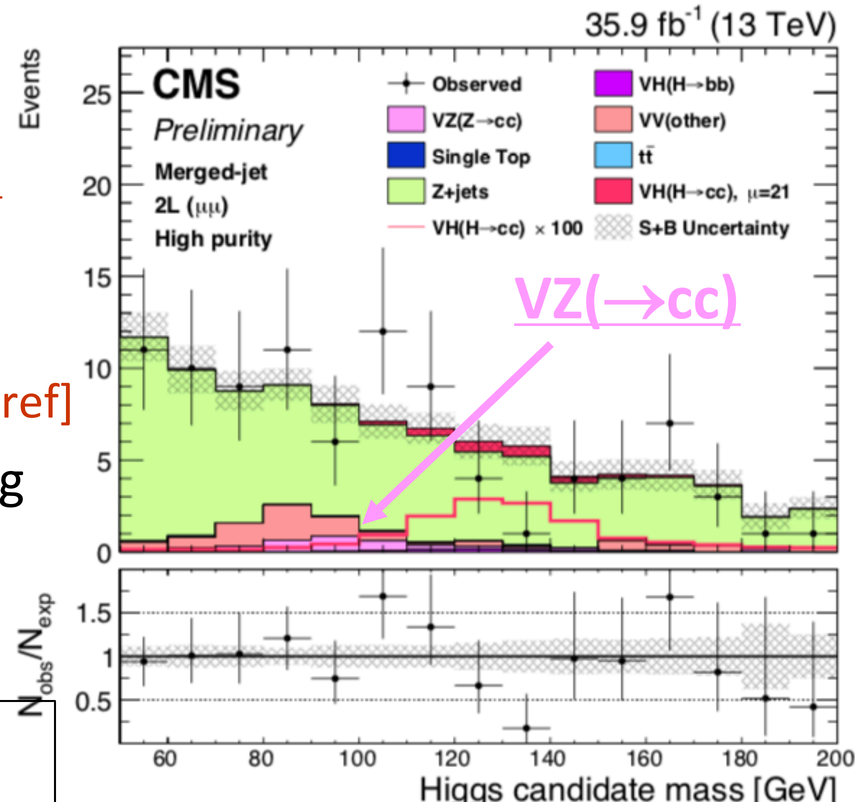
  - Improve tracks from B-hadron using ML [e.g. BDT]
  - Introduce Vertex Mass Recovery for better B/C separation
    - computed by charged tracks; typically smaller than its original mass
    - If π^0 is reconstructed as a part of vertex, adding the mass helps to recover the mass. Find best assignment to a vertex using ML
  - Additional inputs to the flavour tagging MVA + move from BDT → NN

Measure Higgs-charm coupling at the FCCee

The ZH(→cc) analysis

- ZH→cc can serve as a very useful benchmark measurement
  - Identify / motivate the detector requirements
  - Assess performance of event reconstruction and jet flavour tagging algorithms
  - Goal: measure Higgs-charm coupling O(~1%) at the FCCee
    - Can be significantly modified by the presence of BSM Physics

CMS-HIG-18-031



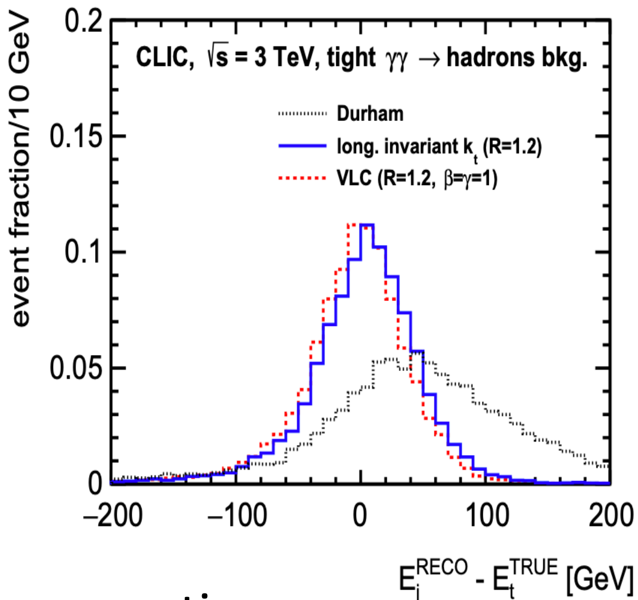
- Current limits at the LHC:
  - $\mu = \sigma/\sigma_{SM} < 70$ (37) obs (exp) using ~36 fb^{-1} of Run 2 data in the VH channel
    - clearly a long way ahead…
  - Projections for HL-LHC: UL on $\mu < 6.3$ [add ref]
- Bottom line: FCCee essential for measuring rare Higgs decays such as H→cc
  - lots of interest already;
    Case study team (CS1) formed:

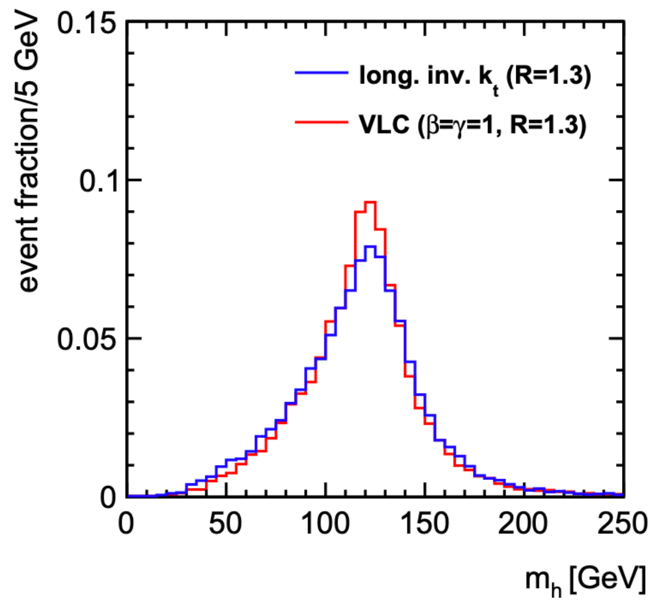
CS1: Aram Apyan, Matthew Baldwin, Franco Bedeschi, Gregorio Bernardi, Alain Blondel, Loukas Gouskos, Patrick Janot,, Markus Klute, Giovanni Marchiori, Michele Selvaggi

Jet clustering

- Exclusive (e.g., Durham at LEP) vs. inclusive (e.g., ak$_T$ at LHC) jets clustering?
  - Latest developments: Valencia algorithm
    - better beam BKG rejection → improved mass resolution
    - NB: CLIC-based studies – need to repeat for FCCee

Eur. Phys. J. C (2018) 78:144





- More questions:
  - No clustering (use full event content) vs large-*R* / small-*R* jets?
    - No clustering could be more appropriate for the ννcc final state?
- Currently open questions:
  - Final answer will come also from iterations with the analysis strategy

Jet flavour calibration strategy

- Initial plan: design a flavour tagger that does not include information related to the mother particle [e.g. Z→cc or H→cc]

- Advantage: Use Z→bb/cc events for calibration
  - e.g., at FCCee: ~10^{11} Z→bb at the Z pole; a tremendous number compared to ~millions at LEP
    - great opportunity to calibrate heavy flavour tagging algorithm with an unprecedented precision

- LHC: Z→bb events [possible also Z→cc] started very recently to be used for the calibration of last generation of H→bb taggers
  - with significantly reduced systematic uncertainties

Where do we stand:

- Finalizing infrastructure to develop a jet flavour tagging algorithm ala LHC
  - Samples (Delphes-based) with a first set of inputs already produced
    - Samples: Z(→vv)H(→bb, cc, qq)
    - Two jet clustering algorithms and detector configurations included
  - Working on final touches on the network architecture
    - Rough estimate: first results in ~2 week time
  - Access to the training package will be provided to all collaborators

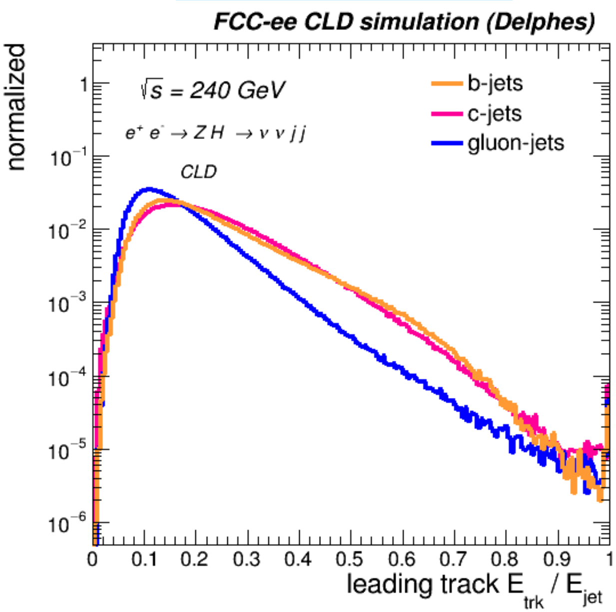
- Analysis front:
  - A first version targeting Z(→qq)H(→cc) final state in place [Delphes-based]
    - tagger's performance parametrized using performance from ILC/CLIC
    - Obviously, other channels will be explored in parallel

- Ultimate goal: incorporating all tagging and analysis developments in FCCSW

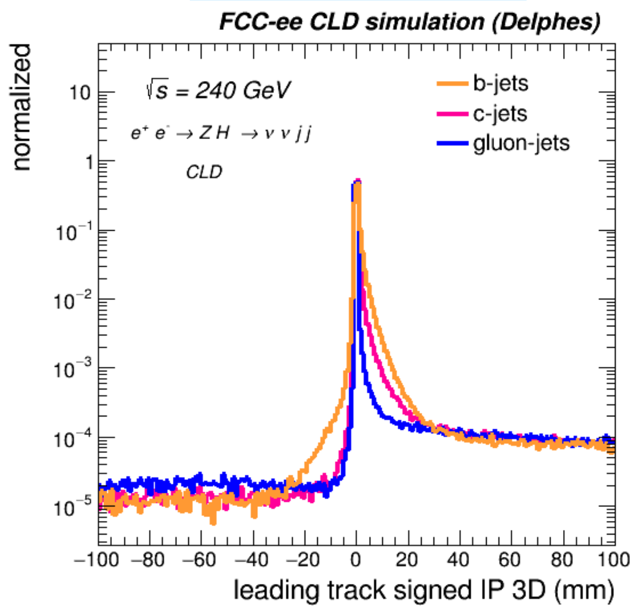
Tentative plan

- Two routes [to be followed in parallel]:
  - Short term plan: based on Delphes samples (~few months)
    - Most of the necessary ingredients are in place
      - list of all charged pf candidates with all kinematic info and PID (including leptons)
      - Full correlated track parameters (computed with TrackCovariance code from Franco)
      - work in progress: secondary vertex reconstruction
    - Relevant for providing a "qualitative" understanding of the impact of different setups (e.g., detector configuration, tagging algos, etc..) on the physics outcome
      - Precise optimization can only be done using Full Simulation

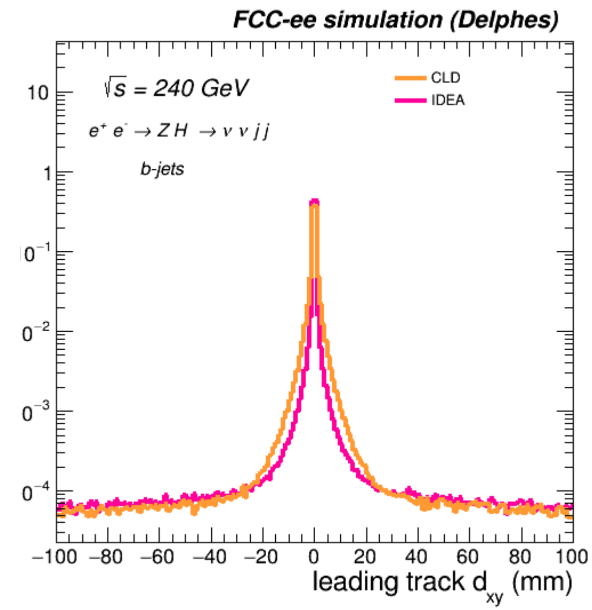
Fragmentation



sig3D



CLD vs. IDEA



Tentative plan (II)

- Two routes [to be followed in parallel]:
  - Short term plan: based on Delphes samples (~few months)

  - Longer term plans (>6 months - year): Implement necessary pieces in FCCSW
    - Detector simulation: Some subdetectors are included; no complete detector geometry in place yet
    - PF event reconstruction: Currently using Delphes PF candidates
      - possible solution: Use Gaudi-Marlin Processor (GMP) to include Pandora PF
    - Vertexing / tracking: possible solution use GMP to include ILC/CLIC algorithms
      - Tracking should be our main goal [PV and SV reconstruction can come later]
      - Assuming that we move to particle based tagging, dedicated b-tagging algorithm not urgent [but will be useful for better understanding the performance]
    - Jet clustering: results from Delphes-based simulation useful
      - Consider even reconstructing the full event: Achieve optimal performance
    - ML infrastructure: will benefit from the Delphes-based developments
      - Code for design/training/inference will be ported from the Delphes-based effort

Summary and Outlook

Summary & outlook

- Powerful jet flavour identification is essential for the success of the e^+e^- physics program

- Jet tagging methods developed at the LHC can be explored at FCCee and potentially enhance the e^+e^- physics program
  - And/or motivate the design of future detectors

- Large effort at the LHC to improve existing / develop new jet tagging methods
  - Key player in these developments: Advanced machine learning algorithms
    - Explore much more of the detector's and evt reconstruction true potential
  - Large gain in performance wrt traditional approaches; which translates in data

- A highly motivated group (i.e. CS#1) in place; finalizing plan of action
  - Close collaboration with the FCCee Physics and Software coordinators essential for the success of the effort.