Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

New jet tagging techniques in Vector Boson Scattering (VBS) WV analysis in the semi-leptonic channel with the CMS experiment

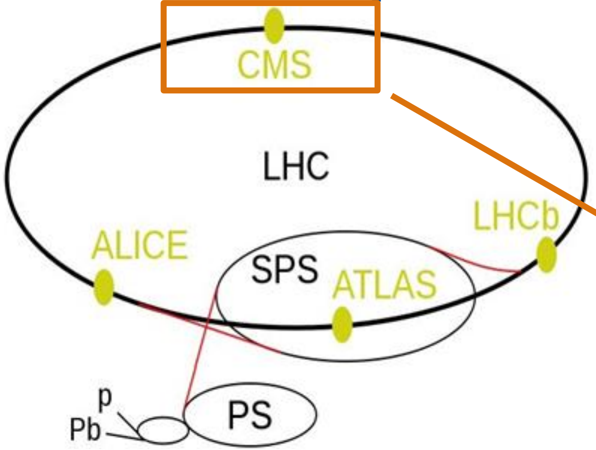
Raffaele Delli Gatti, Irene Zoi, Jennifer Ngadiuba

Final reports – Summer Trainings

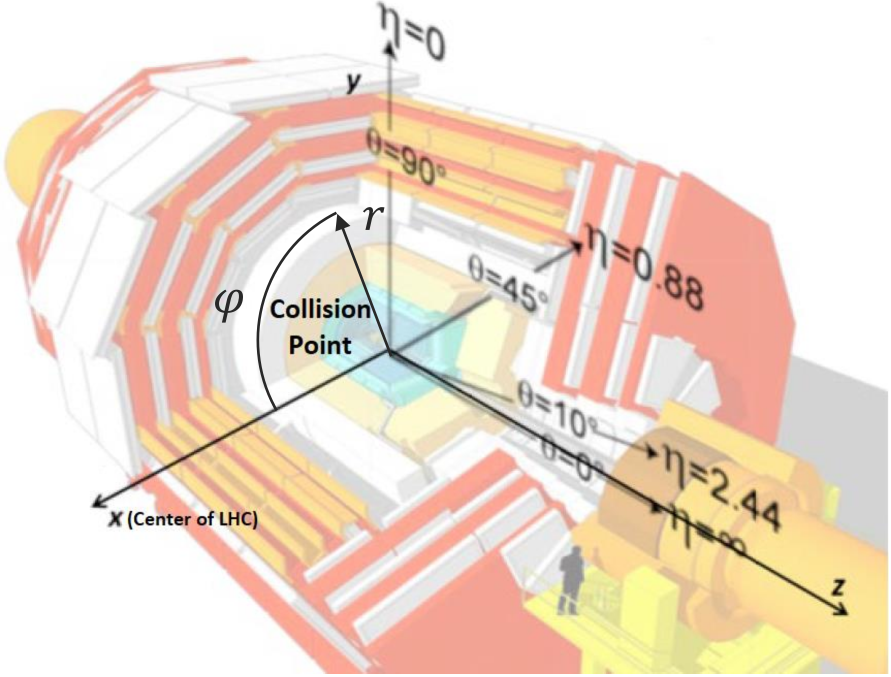
27 September 2023

In partnership with:

The Compact Muon Solenoid at the LHC



Carthesian (x, y, z) and cylindrical (r, η, φ) coordinates



CMS Detector

Run I (2010-2012) Run II (2015-2018) Run III (2022-)

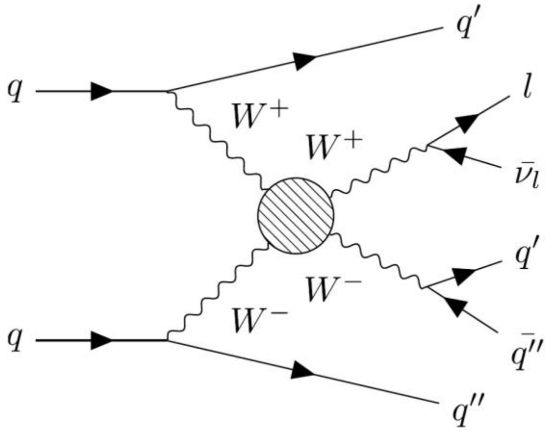
- **Pseudorapidity** $\eta = -\ln \tan\theta/2$
- **Angular distance** $\Delta R = \sqrt{\Delta\eta^2 + \Delta\varphi^2}$
- **Momentum, energy** measured in the **transverse** plane:

$$p_T = \sqrt{p_x^2 + p_y^2}, \qquad E_T = \sqrt{m^2 + p_T^2}$$

- **Missing Transverse Energy** (MET): $\not{E}_T = \left|-\sum_i \bar{p}_{T_i}\right|$

Introduction and motivation

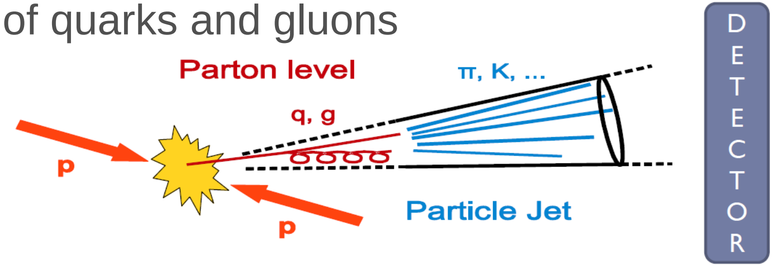
- **WV(V=Z,W) VBS scattering** in the **semileptonic channel**
- The W boson decaying leptonically and the other boson hadronically
- Two high energetic jets in the **forward regions** and reduced jet activity in the central region
- Purely **EWK** rare **process** at LO with 6 fermions in the final state and large background contamination



- First evidence of the SM process at the LHC in 2021 [**PLB 834 (2022) 137438**] using full Run II
- **Signal significance** of 4.4 standard deviations, to be increased for 5σ observation
- Also Important for BSM searches, such as anomalous Quartic Gauge Couplings

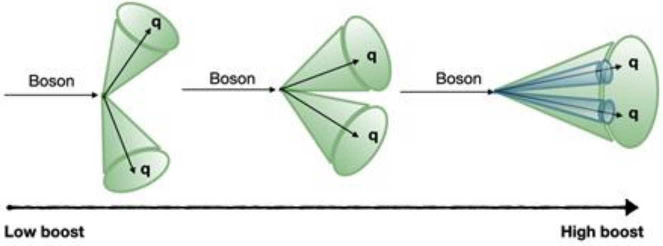
Jet clustering and tagging

- Jets are signatures of quarks and gluons

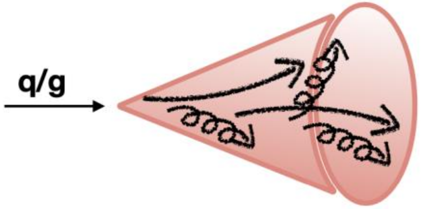


Parton Showering → Hadronization → Jets of colorless particles

- **Jet tagging:** identify the particle that initiated a jet
- **Boosted hadronic objects** have a different energy pattern than background jets of comparable invariant mass



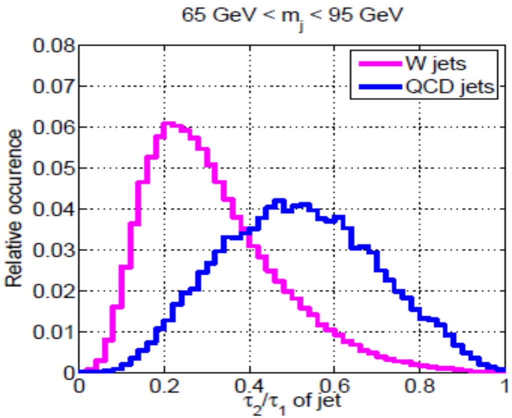
Boosted W boson jet



Boosted background jet

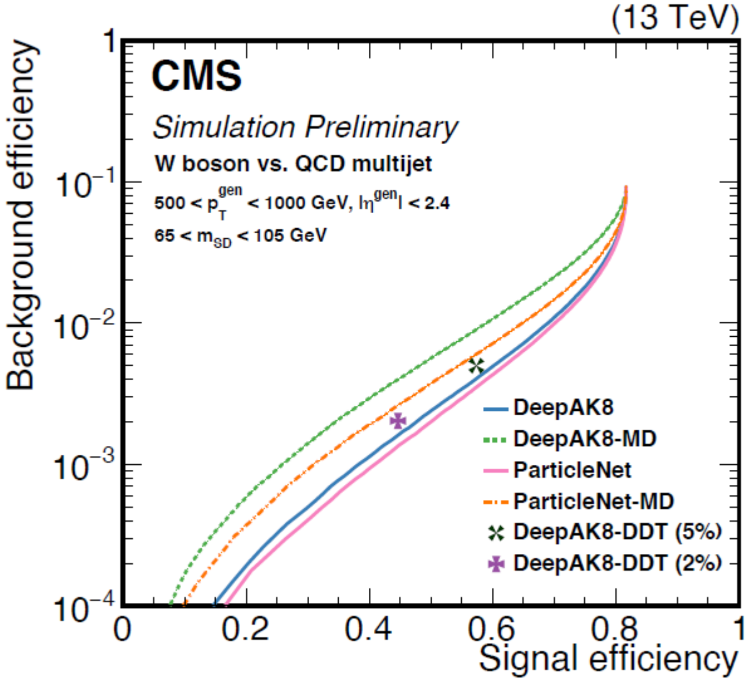
N-subjettiness

- Traditional tagging method **[JHEP 03 (2011) 015]**
- Effective discriminating variable for tagging boosted objects (at high Lorentz boost) and rejecting the background of QCD jets with large invariant mass
- For a large enough boost factor, the decay and fragmentation yields a collimated spray of hadrons which a standard jet algorithm would reconstruct as a single jet
- Jet shape variable (it tells how likely a jet has N subjets): τ_N
- τ_2/τ_1 is an effective discriminating variable to identify two-prong objects like boosted W, Z, and Higgs bosons



ParticleNET

- Improving boosted top, W, Z or Higgs tagging techniques by using machine learning, most notably deep neural networks (DNNs) [**PRD 101 (2020) 056019**]



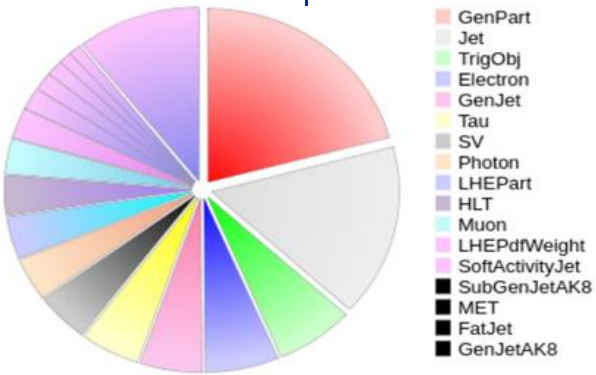
$$\text{Signal efficiency} = \frac{\#\ selected\ signal\ events}{\#\ signal\ events}$$

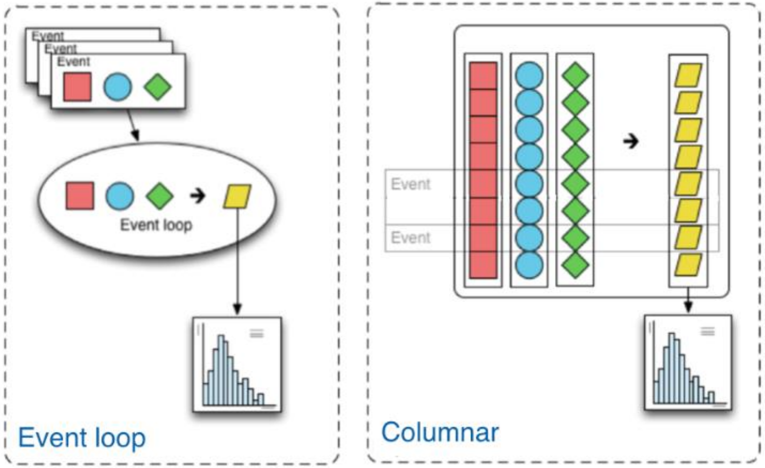
$$\text{Background efficiency} = \frac{\#\ selected\ bkg\ events}{\#\ bkg\ events}$$

NANOAOD and Coffea

- NANOAOD: a **new event data format** by CMS
- **ntuples** with per event information (~1kB per event)
- A factor of 20 smaller than the MINIAOD
- Only **top-level information** and physics objects used in the last steps of the analysis

NANOAOD composition



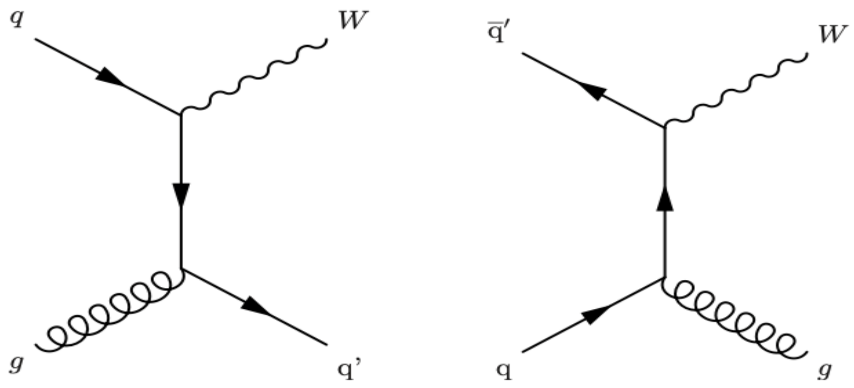


- **Columnar Analysis** with Coffea Framework
- The columnar approach has no explicit event loops: **100 times faster**
- The fields of data are treated as **awkward arrays:** array of subarrays of arbitrary length
- [[Muon, Muon], [Muon], [], [Muon, ... [Muon]]

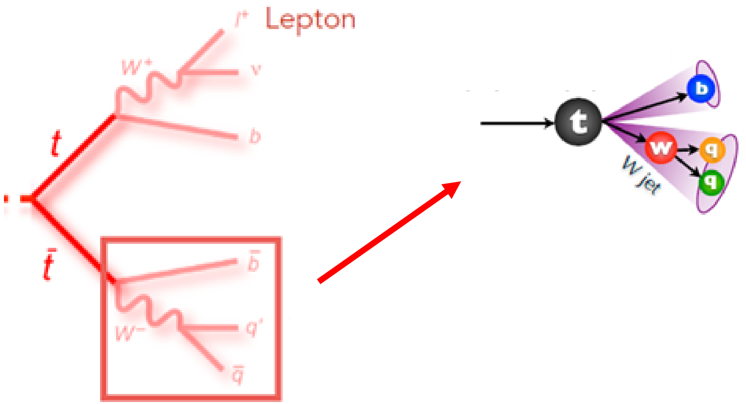
Backgrounds contributions

- Main sources

  - W + jets

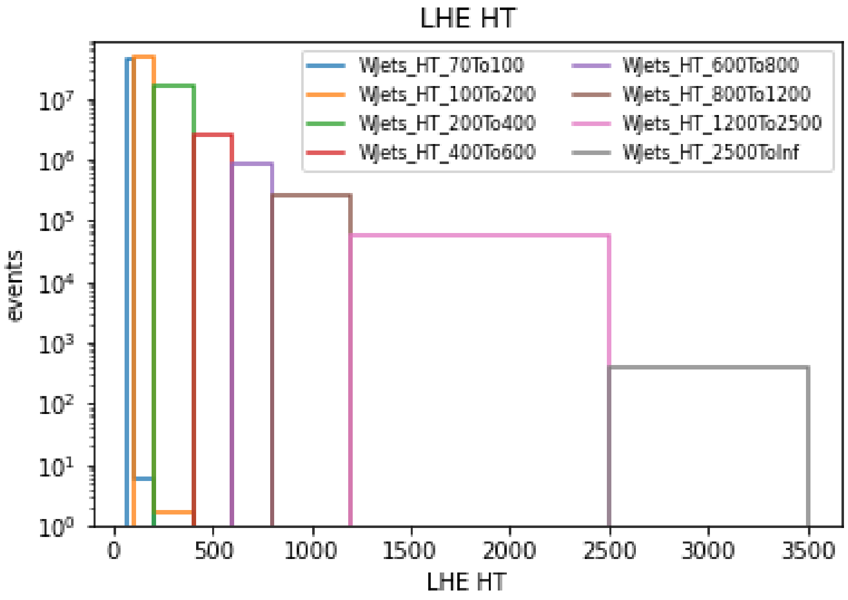


  - Top: $t\bar{t}$, single t, tW, tZ



W + jets background

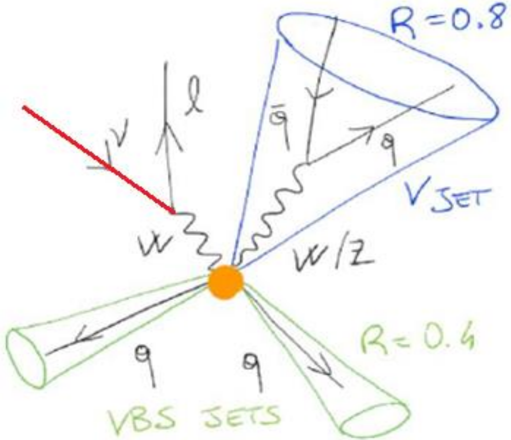
- W+jets HT-binned samples



- $HT = \sum_{i=1}^{N_{jet}} E_T$
- Variable characterizing the **visible energy** in the transverse plane
- **Increased statistic** in ensured at different scales of energy with respect to inclusive LO generation

Characterization of the phase space

- Kinematic cuts to define the **electron signal region**



Boosted category

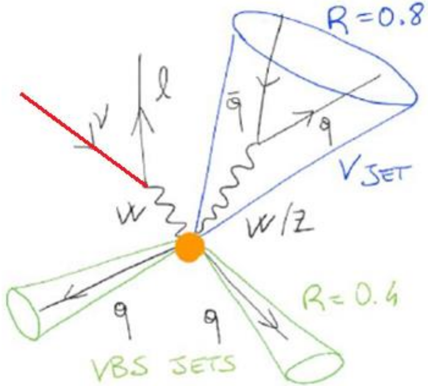
- **One isolated lepton** ($electron$) in the final state, moderate **Missing Transverse Energy**
- **1 FatJet** (anti-kt $R = 0.8$ jet) from hadronic decay of W boson

$$\not{E}_T = \left| -\sum_i \bar{p}_{T_i} \right|$$

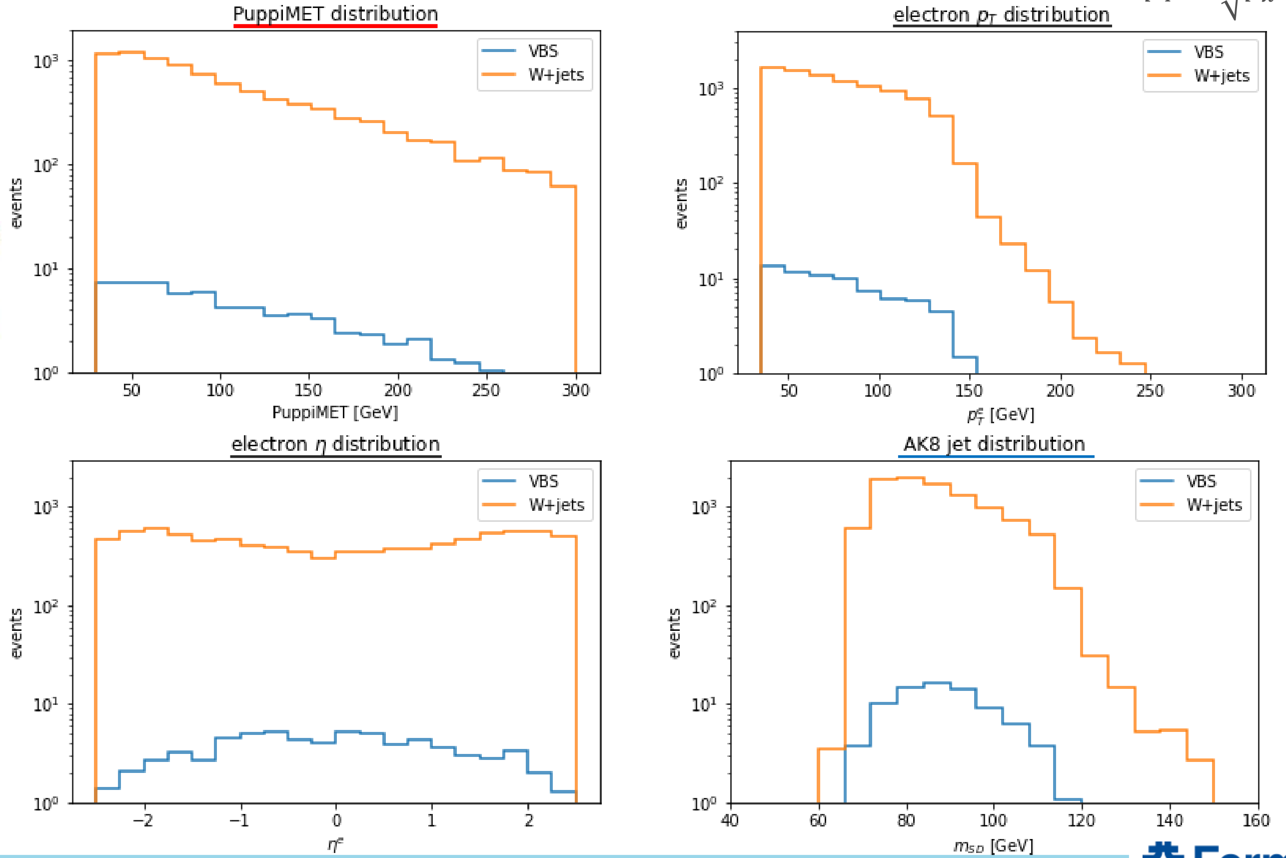
- **Invariant mass cut** on the hadronically decaying W: $70\ GeV < m_{SD} < 115\ GeV$
- At least **2 jets** (anti-kt R=0.4) tagged as VBS jets (max invariant mass pair)

Kinematic distributions

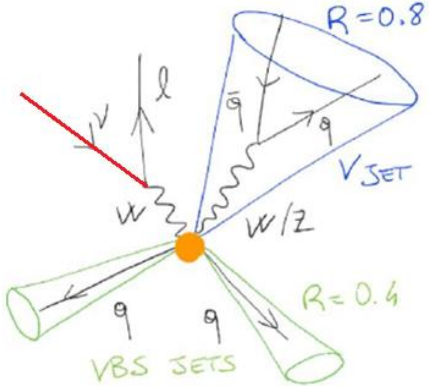
$$p_T = \sqrt{p_x^2 + p_y^2}$$

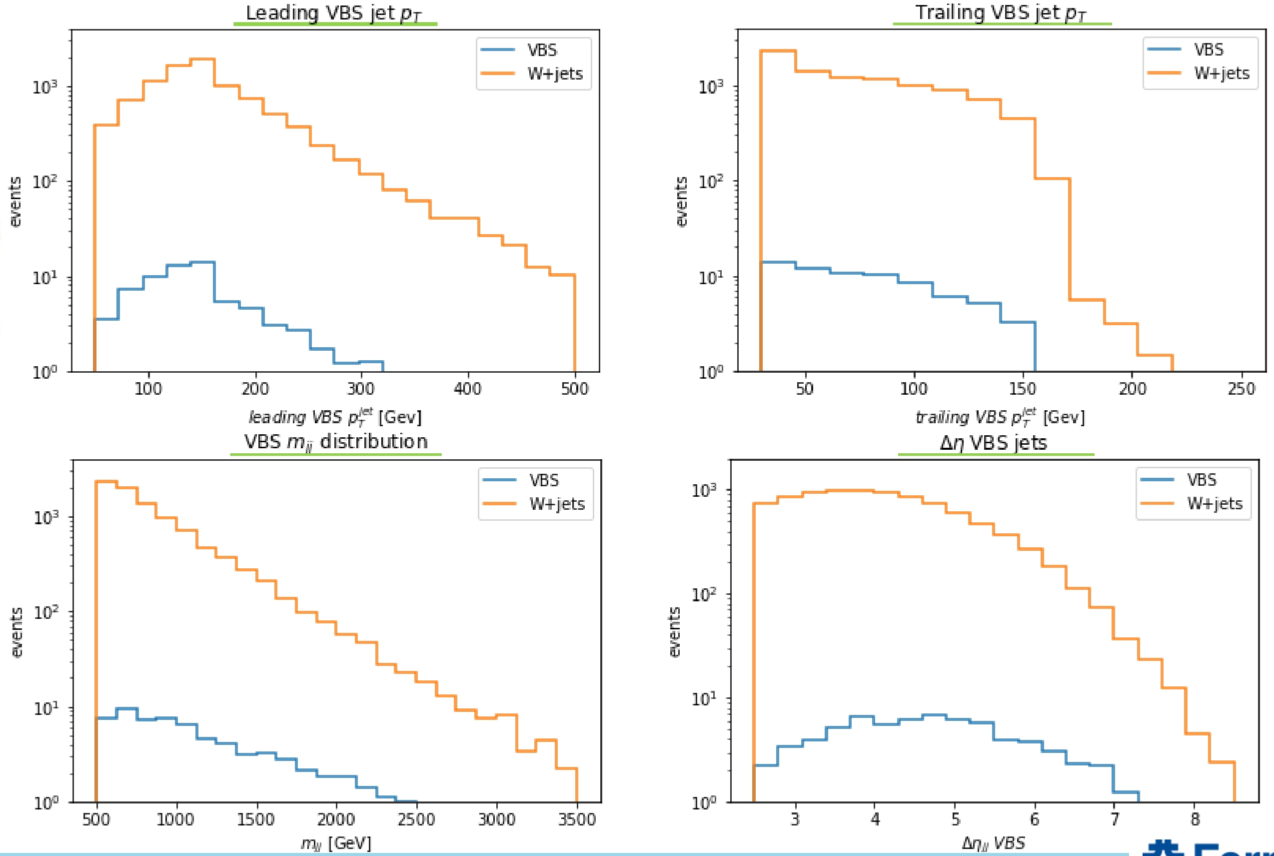


Pseudorapidity $\eta = -\ln \tan\theta/2$



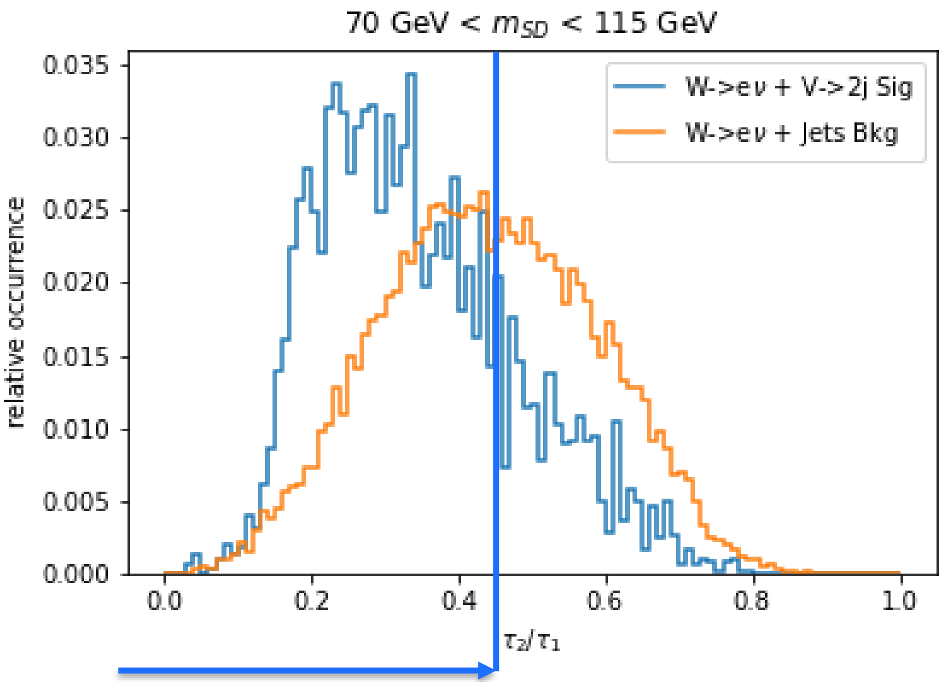
Kinematic distributions





Working point of the official analysis

N-subjettiness



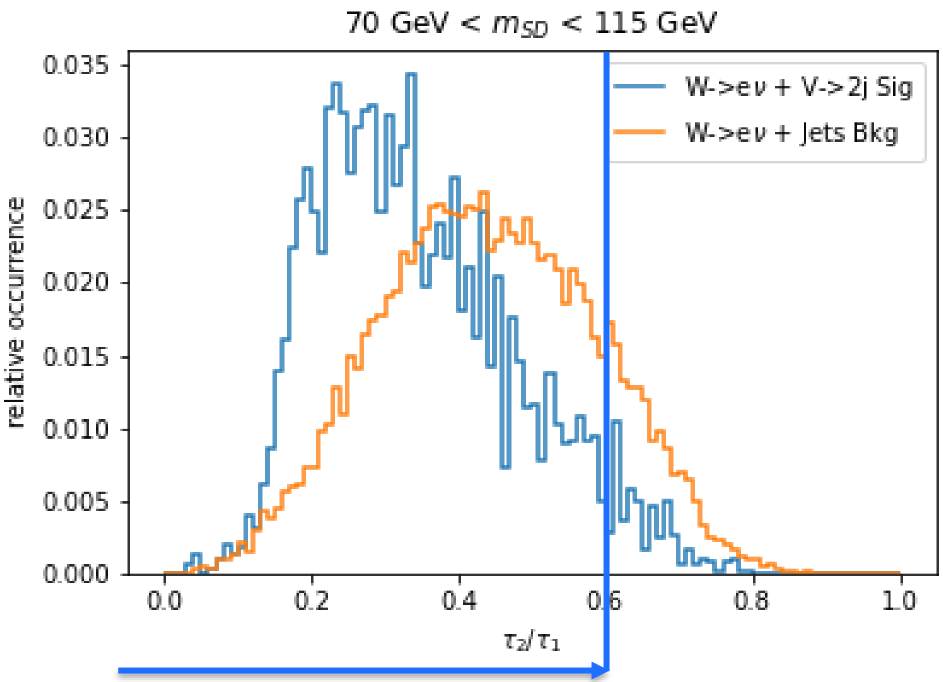
$$\tau_2/\ \tau_1 < 0.45$$

Signal efficiency = 78%

Background efficiency = 53%

Efficiency studies for different working points

N-subjettiness



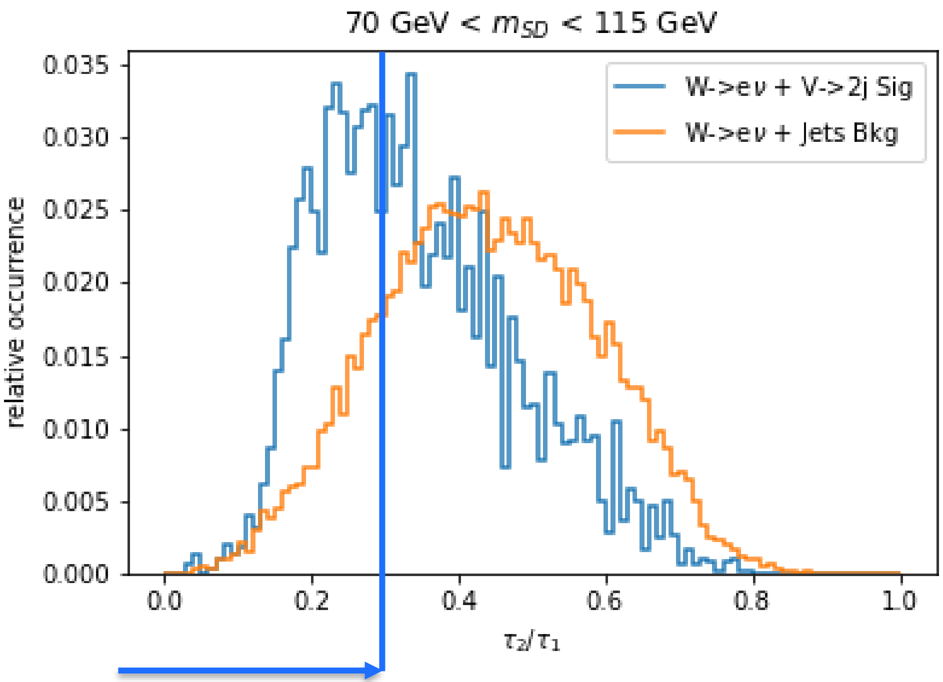
$\tau_2/\,\tau_1 < 0.6$

Signal efficiency = 95%

Background efficiency = 85%

Efficiency studies for different working points

N-subjettiness



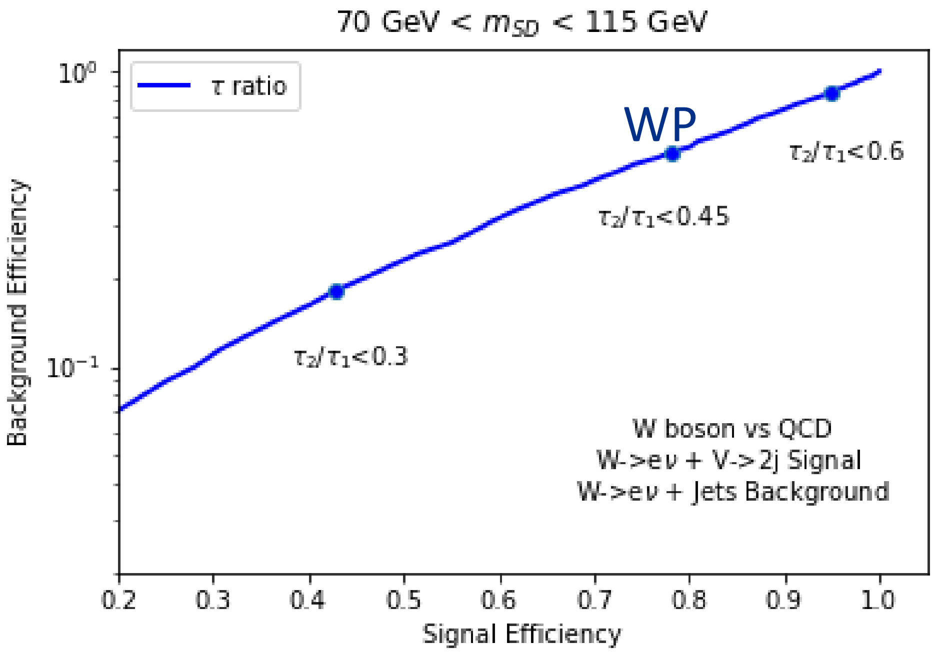
$\tau_2/\tau_1 < 0.3$

Signal efficiency = 43%

Background efficiency = 18%

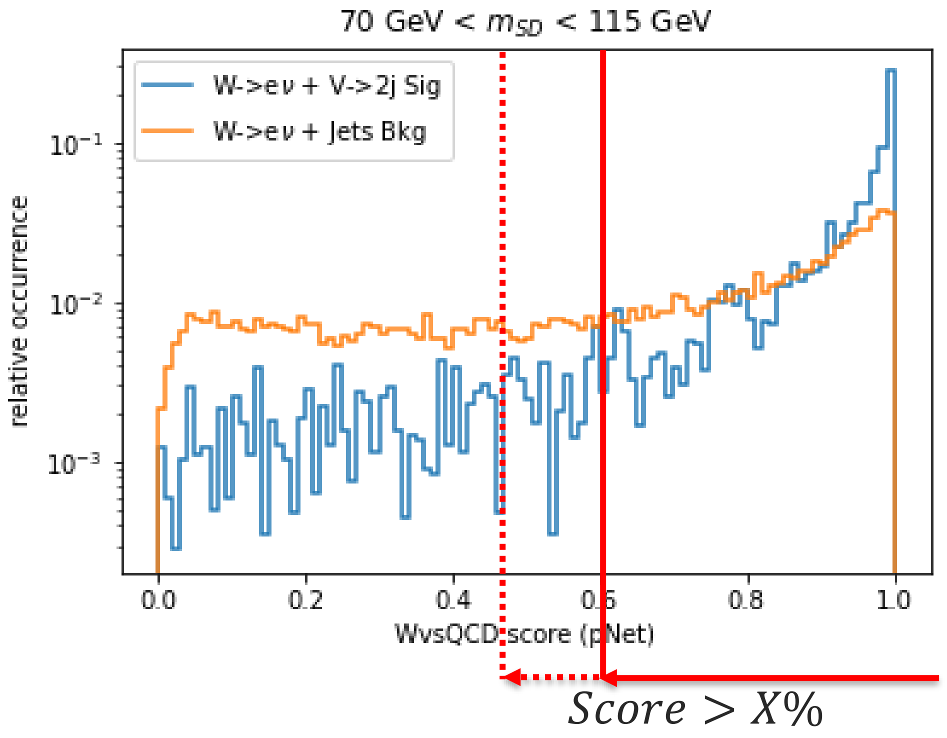
Tagging Efficiency

N-subjettiness



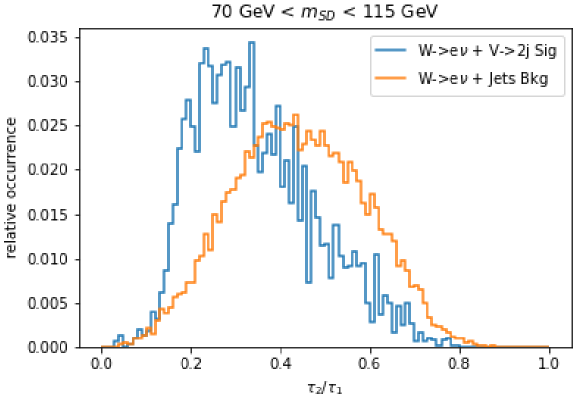
Efficiency studies for different working points

ParticleNET

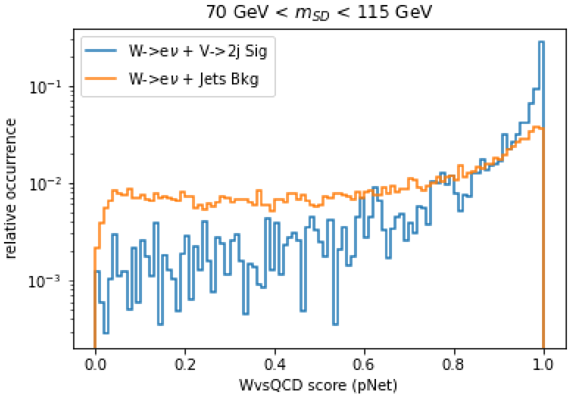


Jet tagging variables

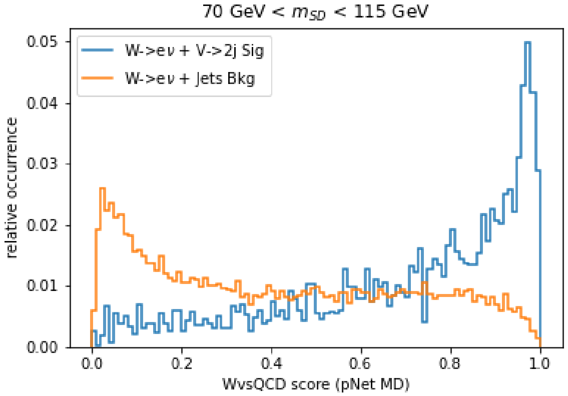
N-subjettiness



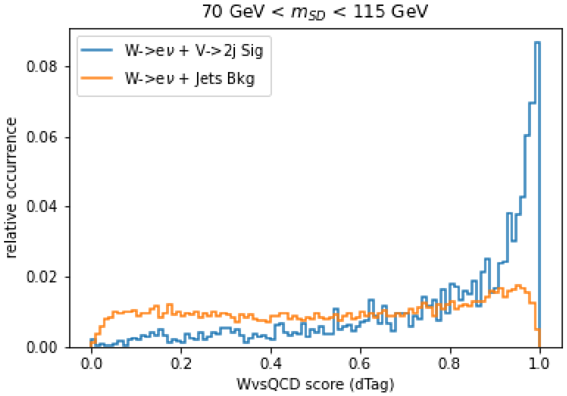
ParticleNET



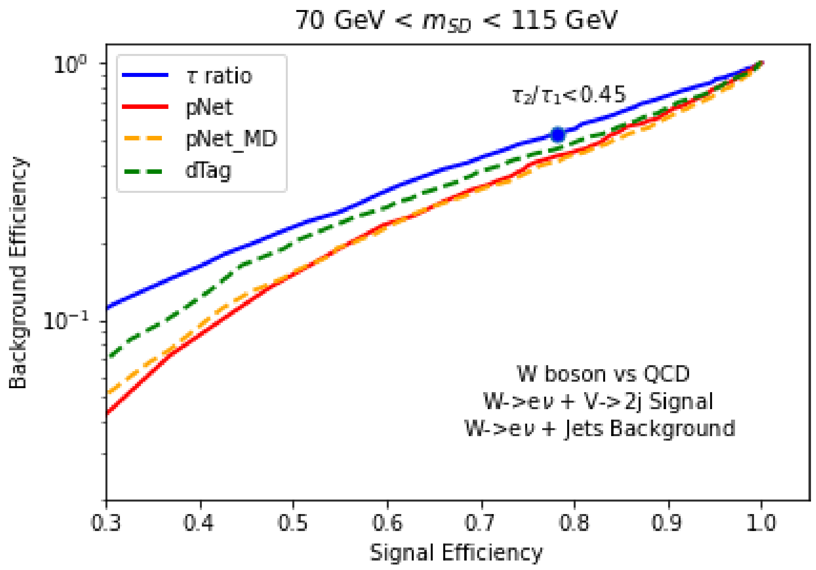
pNET MD

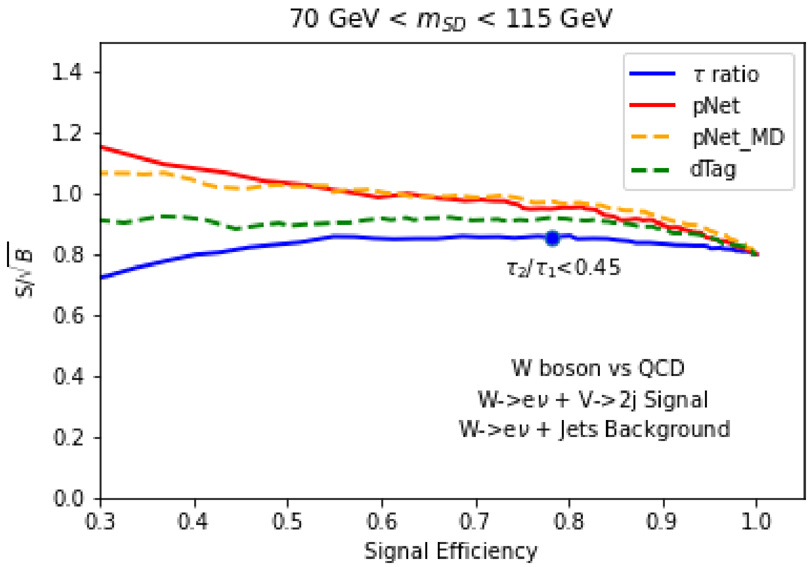


DeepTag



Efficiency and sensitivity studies





- Background efficiency reduction ~20%, sensitivity gain ~11% for a signal efficiency of 78%

Conclusions

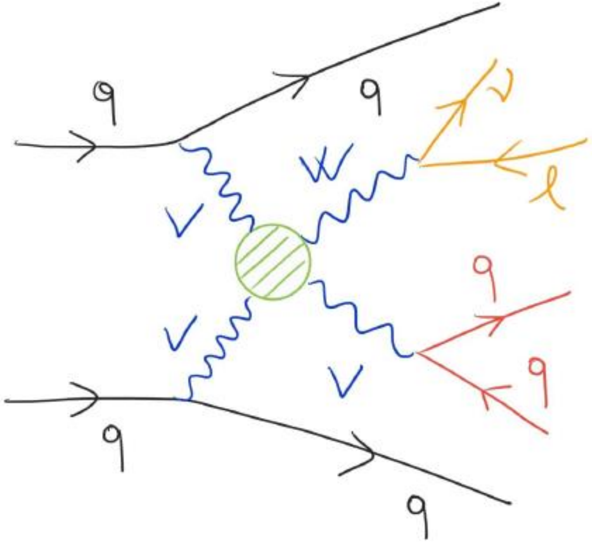
- $\mathrm{WV}(\mathrm{V}=\mathrm{Z},\mathrm{W})$ Vector Boson Scattering in the semi-leptonic $\mathrm{l}\nu q\bar{q}$ channel
- Coffea Framework to analyze NANOAOD data version 9
- Electron signal region, W + jets background
- Latest developments in ML-based identification algorithms of highly Lorentz boosted heavy particles in CMS
- Compared to N-subjettiness, ParticleNet shows background efficiency reduction $\sim 20\%$ and sensitivity gain $\sim 11\%$, for a signal efficiency of 78%, in the context of the VBS WV analysis
- Further studies:
  - Tagging efficiency and sensitivity also in the muon signal region
  - Z vs QCD score
  - Study of the top background
  - Include full Run II statistics

Thank you for your attention

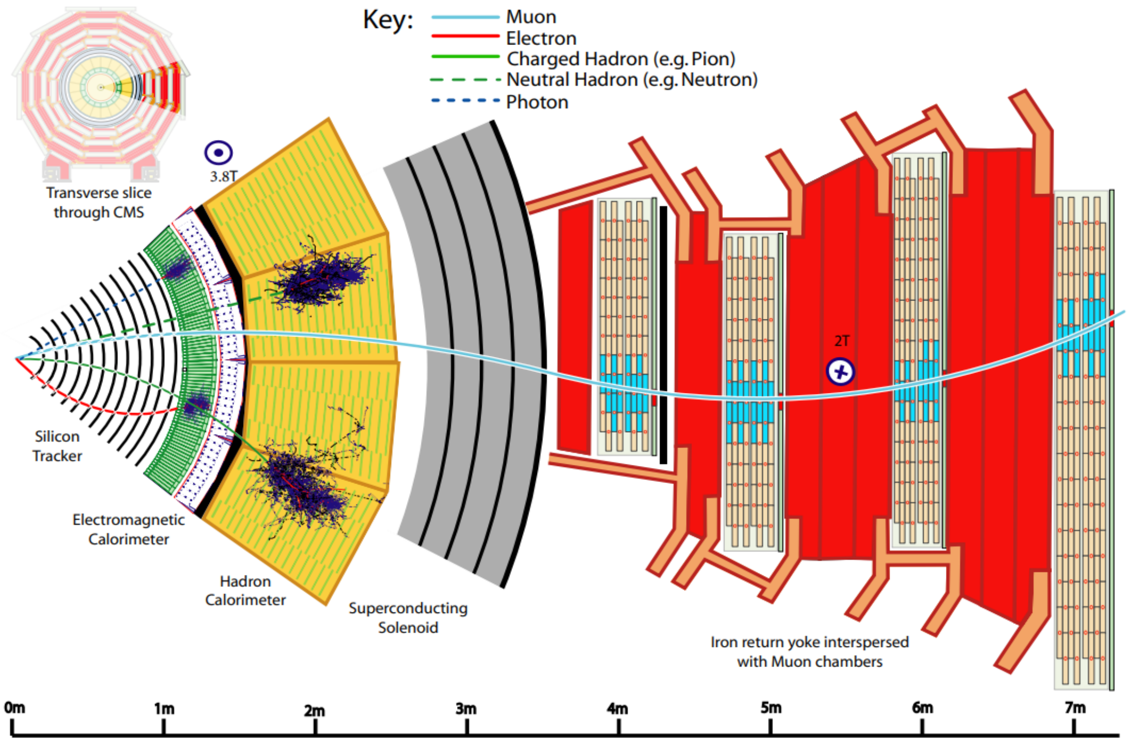
Backup

Outline

1. The Compact Muon Solenoid at the Large Hadron Collider
2. $\mathrm{WV}(\mathrm{V} = \mathrm{Z}, \mathrm{W})$ Vector Boson Scattering in the semi-leptonic $\mathrm{l}\nu q\bar{q}$ channel
3. Jet tagging techniques
4. N-subjettiness and ParticleNet
5. NANOAOD and Coffea Framework
6. Phase space and distributions
7. Efficiency and sensitivity gain

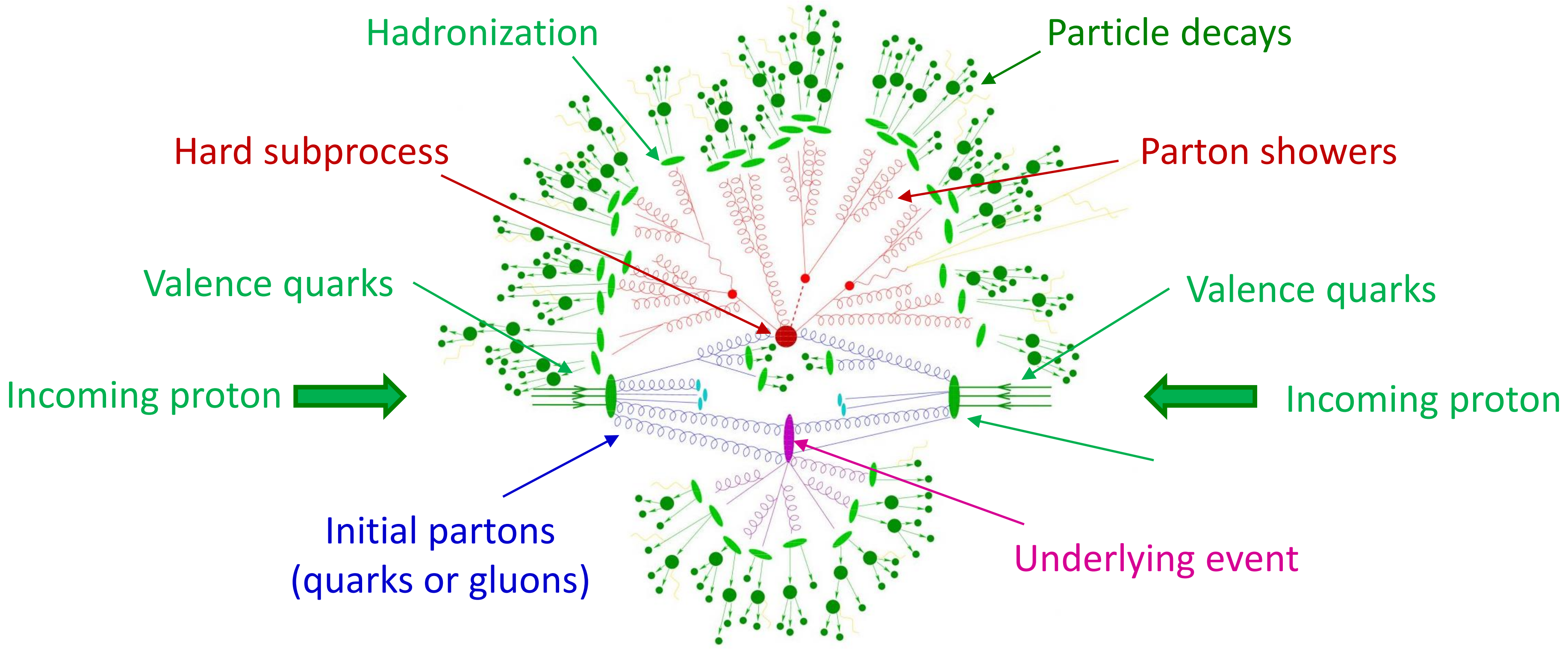


The Compact Muon Solenoid



- **Multi-purpose detector** (Cessy, France)
- **Cylindrical structure**, around the interaction point
- **Sub-detectors**
- **Superconducting solenoid** magnet, 3.8 T
- **Barrel** and **endcap** regions
- **Particle flow** (PF) algorithm to reconstruct final state particles

Structure of a proton-proton collision

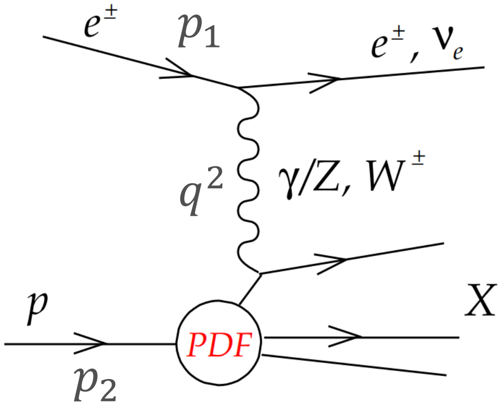


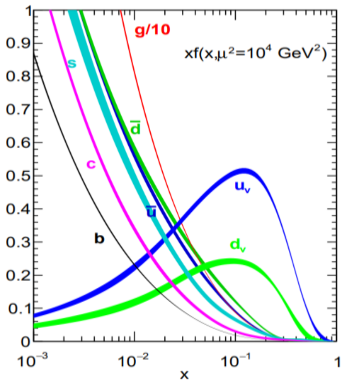
Monte Carlo generation

PARTON DISTRIBUTION FUNCTIONS (PDFs)

- In hadron collisions at the LHC, the colliding protons presents an internal structure: **valence quarks** interact via **gluons**, from which virtual quark-antiquark pairs arise (**sea quarks**)

- Dynamics of the systems resulting in a distribution of the parton momenta, to be determined by experiments (non-perturbative QCD), such as electron-proton **deep inelastic scattering**

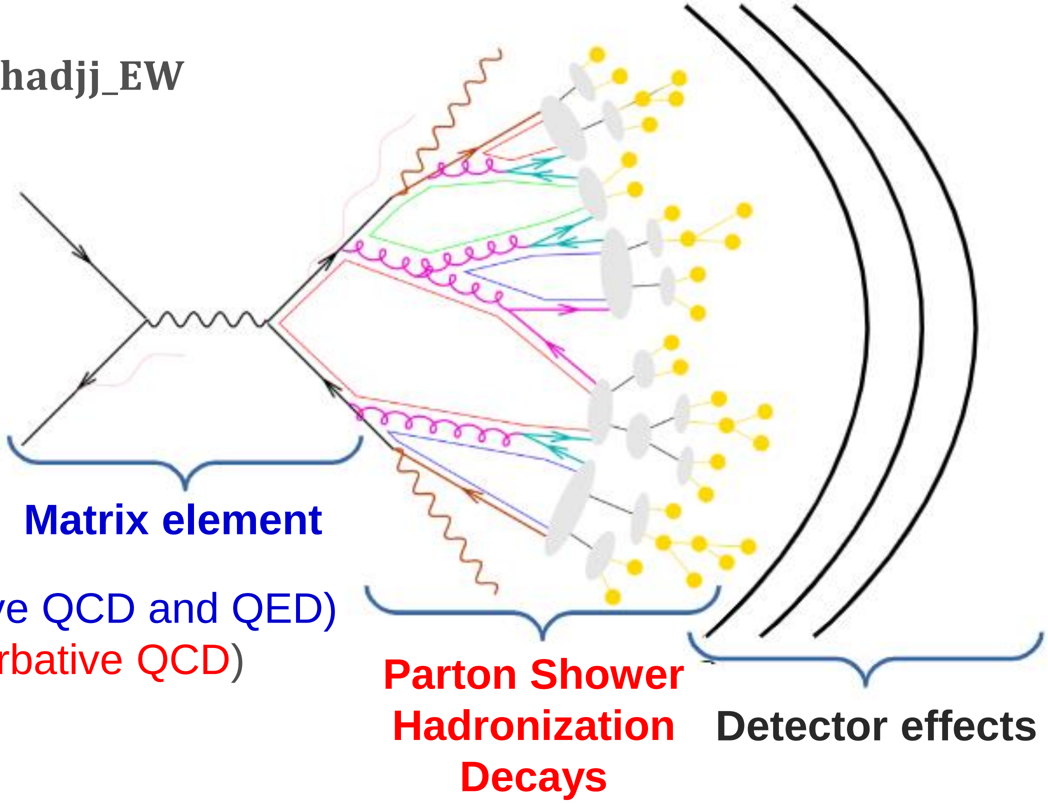




- Each parton carries an unknown fraction ξ of the proton momentum: statistical distribution f(ξ). There are two independents variables:
  - **Bjorken scaling** $x \equiv \frac{Q^2}{2p_2 \cdot q}$ with $Q^2 = -q^2, \quad 0 \leq x \leq 1$
  - **Inelasticity** $y \equiv \frac{p_2 \cdot q}{p_2 \cdot p_1} \approx \frac{Q^2}{x \cdot s}, \quad s = (p_1 + p_2)^2 \approx 2p_1 \cdot p_2, \quad 0 \leq y \leq 1$
- At high energies $E \gg m_p, \quad x \equiv \xi \rightarrow$ **PDF**: $\mathrm{f}(x, Q^2)$.

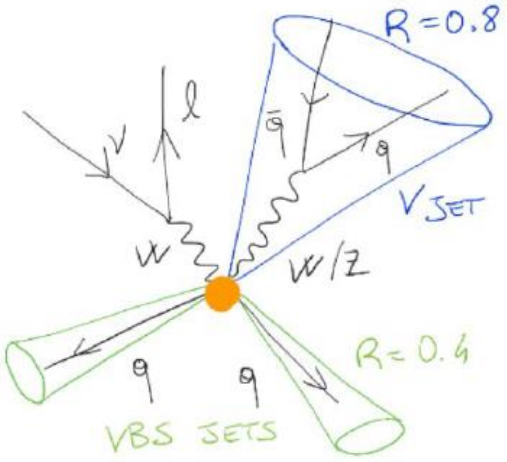
Simulated Monte Carlo sample of the signal

- Process: **WlepWhadjj_EW**, **WlepZhadjj_EW**
- Production campaign: **Run II 2018**
- Collisions: **proton-proton**
- Center-of-mass energy: **13 TeV**
- PDF: **NNPDF3.1**
- Tuning: **CP5 MC Tune**
- Format: **NANOAODSIMv9**
- Generators: **Madgraph** (perturbative QCD and QED) interfaced with **Pythia8** (non-perturbative QCD)



Characterization of the phase space

- Kinematic cuts to define the **signal region**
- Only **one isolated tight lepton** (e, μ) in the final state: $p_T^e > 35\ GeV$, $p_T^\mu > 30\ GeV$
- Events containing a second loosely identified lepton with $p_T > 10\ GeV$ are vetoed
- **Lepton pseudorapidity** : $|\eta^e| < 2.5$, $|\eta^\mu| < 2.4$
- **Moderate Missing Transverse Energy** : $\not{E}_T > 30\ GeV$
- **1 FatJet** (anti-kt $R = 0.8$ jet) from hadronic decay of W boson: $p_T > 200\ GeV$, $\eta < 4.7$
- **At least 2 jets** (anti-kt R=0.4) tagged as VBS jets (max invariant mass pair): leading $p_T > 50\ GeV$, trailing $p_T > 30\ GeV$, $\eta < 2.4$, $\Delta\eta_{VBS} > 2.5$, $m_{jj}^{VBS} > 500\ GeV$



Boosted category

Characterization of the phase space

- Reconstructed jet no overlapping with isolated leptons:
  $\Delta R(j_{AK4}, l) > 0.4$, $\Delta R(j_{AK8}, l) > 0.8$

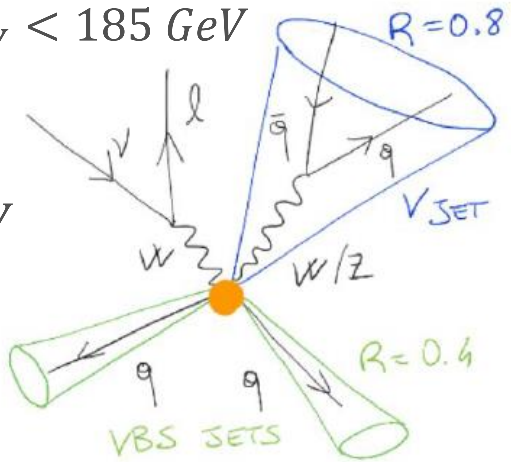
- AK4 jets no overlapping with AK8 jets: $\Delta R(j_{AK4}, j_{AK8}) > 0.8$

- **Transverse mass** of the **leptonically** decaying W: $m_W^T < 185\ GeV$

- **Invariant mass** of the **hadronically** decaying W:

$$70\ GeV < m_W < 115\ GeV$$

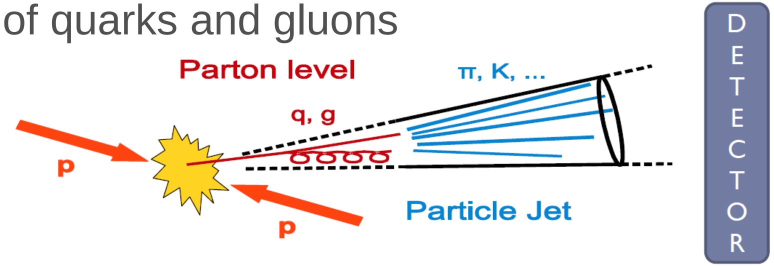
- **bVeto** (no b jets)



Boosted category

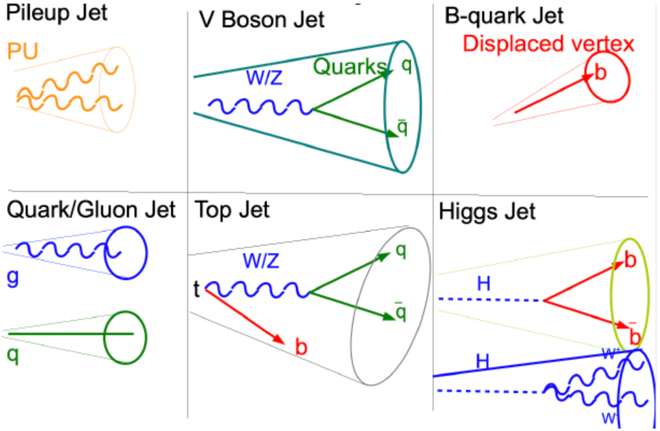
Jet clustering and tagging

- Jets are signatures of quarks and gluons



Parton Showering → Hadronization → Jets of colorless particles

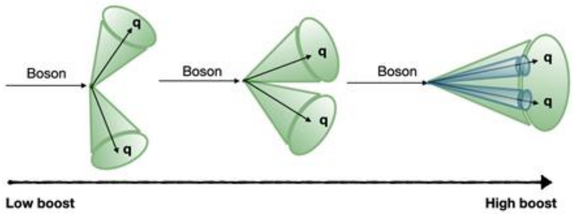
- Jet algorithm allows to collect iteratively the particles belonging to a jet

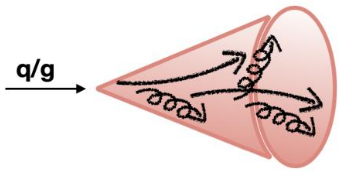


- Different types of jets
  - Larger cone: **top jet**
  - Smaller cone: **W jet**
  - Much smaller: **b jet**
- **Cone radius:** $\boxed{\Delta R = \frac{2m}{p_T}}$
  - **AK8:** $R = 0.8$
  - **AK4:** $R = 0.4$
- **Jet tagging:** identify the particle that initiated a jet

Boosted objects

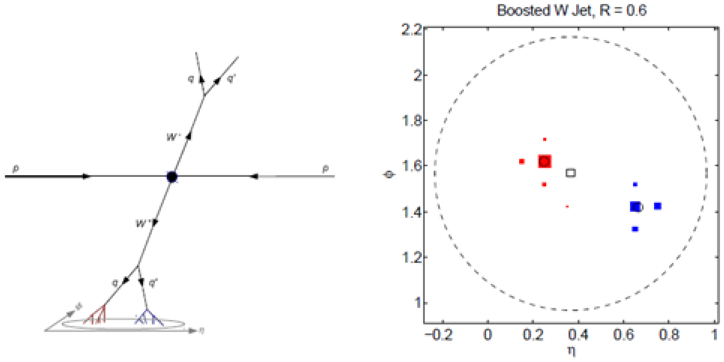
- Boosted hadronic objects have a different energy pattern than QCD jets of comparable invariant mass



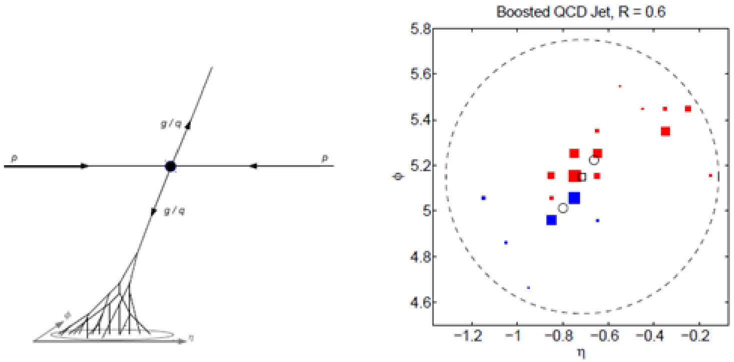


A jet containing a **boosted W boson** should be composed of **two distinct hard subjets**, with invariant mass of 80 GeV

A **Boosted QCD jet** of 80 GeV originates from a hard parton, it gains mass through **large angle soft splittings**



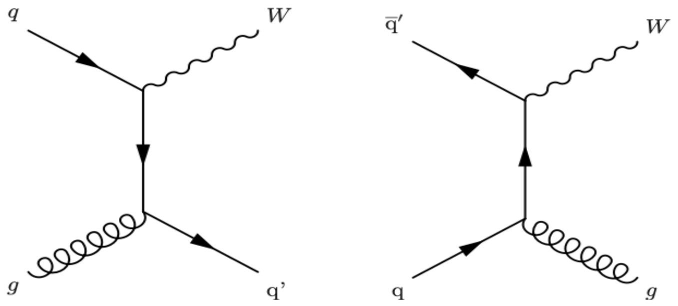
Boosted W jet, $R = 0.6$



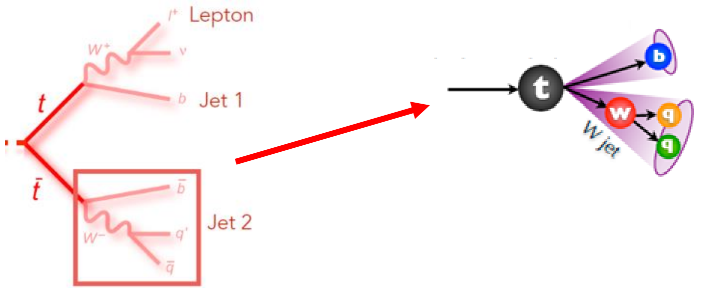
Boosted QCD jet, $R = 0.6$

Background contributions

- **Main sources**
  - W + jets



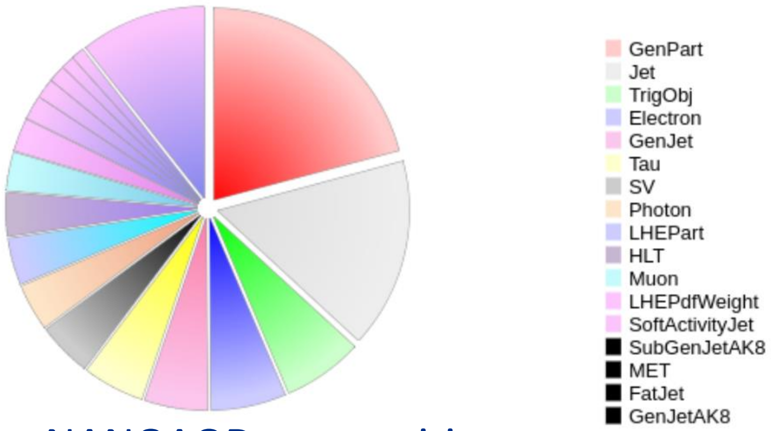
  - Top: ttbar, single top, tW, tZ



- **Other contributions**
  - QCD-VV
  - Non-prompt: data-driven estimation with fakable object technique
  - VBF-V: single V boson EWK production
  - Drell-Yan
  - ggWW, VVV, Vgamma: very small contribution

NANOAOD data format

- A new event data format designed by the CMS collaboration
- Satisfy the needs of a large fraction of physics analyses (at least 50%) with a per event size of order 1 kB.
- More than a factor of 20 smaller than the MINIAOD format
- Only top level information and physics objects typically used in the last steps of the analysis
- Typical format of user ntuples, containing per event information



NANOAOD composition

Coffea Framework

- **Columnar Analysis** with Coffea Framework
- While the traditional way of analyzing data in HEP involves the event loop, the columnar approach has no explicit loops: **100 times faster**
- The fields of data are treated as arrays and analysis is done by way of numpy-like array operations.
- Coffea builds upon **awkward arrays** with a variety of features that better enable us to do analyses
- Array of subarrays, which have arbitrary length (they can even be empty)!
- [[Muon, Muon], [], [Muon], [], [Muon, ... [Muon]]