# Search for the Higgs boson decaying to charm quarks using large-radius jets with the CMS experiment

Huilin Qu

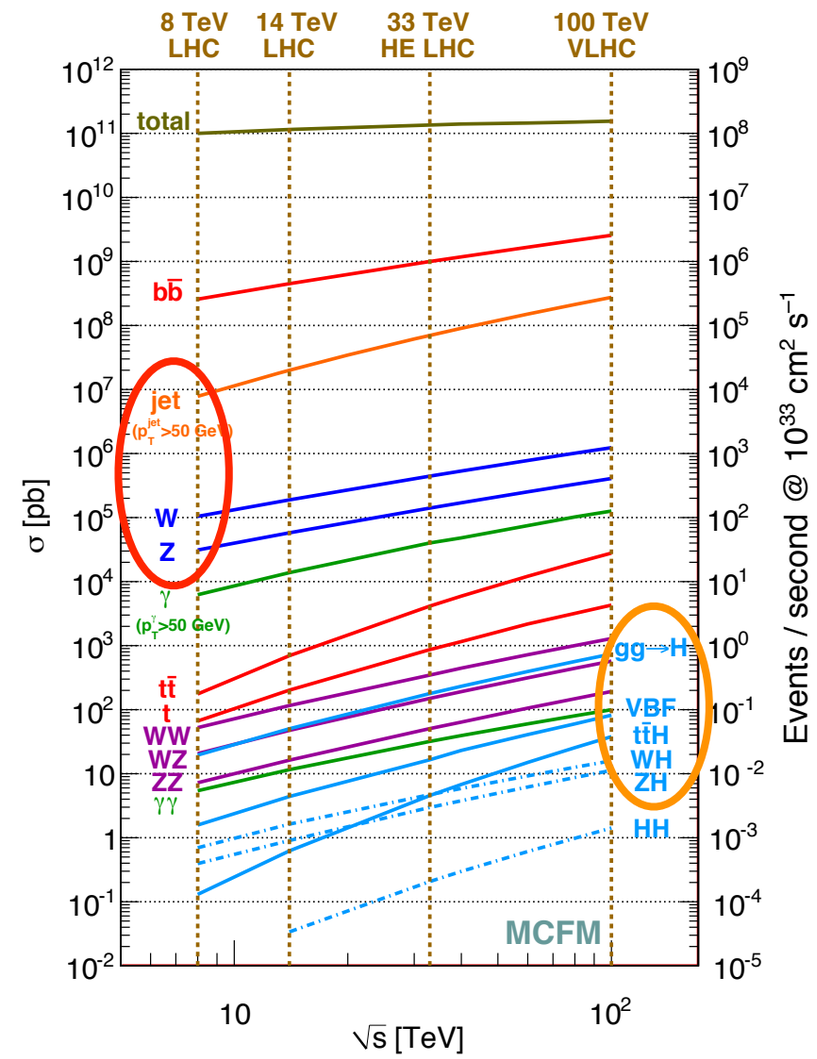
*on behalf of the CMS collaboration*

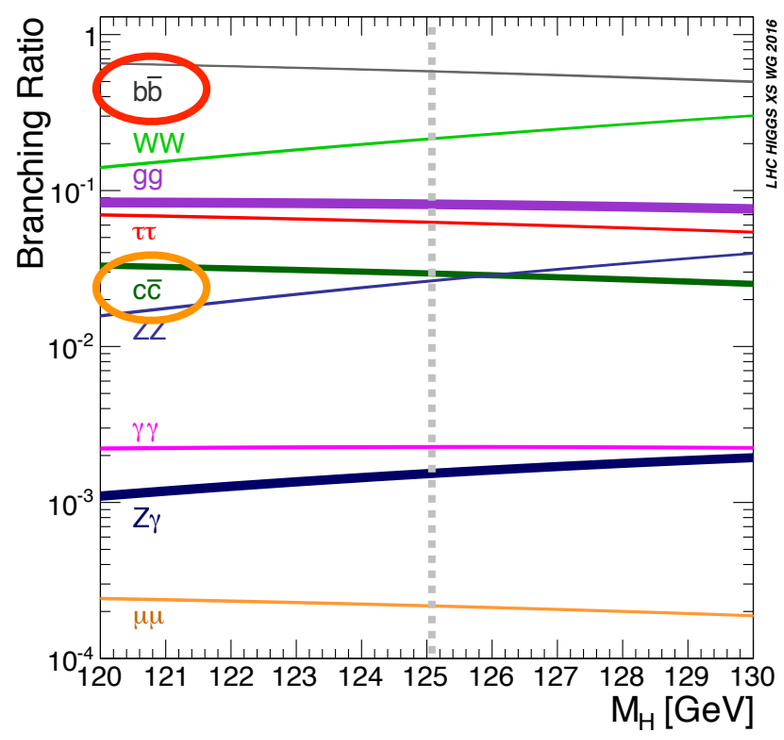
*Higgs Couplings 2019, Oxford*

October 1, 2019

# Introduction

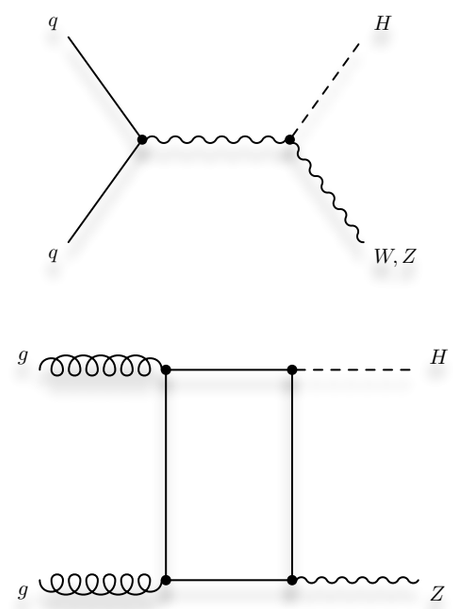
- Search for H→cc:
  - directly probes the Yukawa coupling to 2nd-generation quarks
  - next milestone in Higgs coupling measurements
- H→cc: very challenging to hunt at a hadron collider
  - small branching ratio in SM: ~2.9%
    - H→bb (BR=58%): a background in this search
  - very large (hadronic) backgrounds
  - charm quark identification is the key
- Direct searches for H→cc
  - ATLAS: Z(→ll)H, 36.1 $fb^{-1}$ data
    - upper limits on $\mu := (\sigma \times BR) / (\sigma_{SM} \times BR_{SM})$
    - $\mu < 110$ (150) obs. (exp.) [PRL 120 (2018) 211802]
  - CMS: VH, 35.9 $fb^{-1}$ data [*CMS-PAS-HIG-18-031*]
    - new results!
    - see Luca's talk for an overview

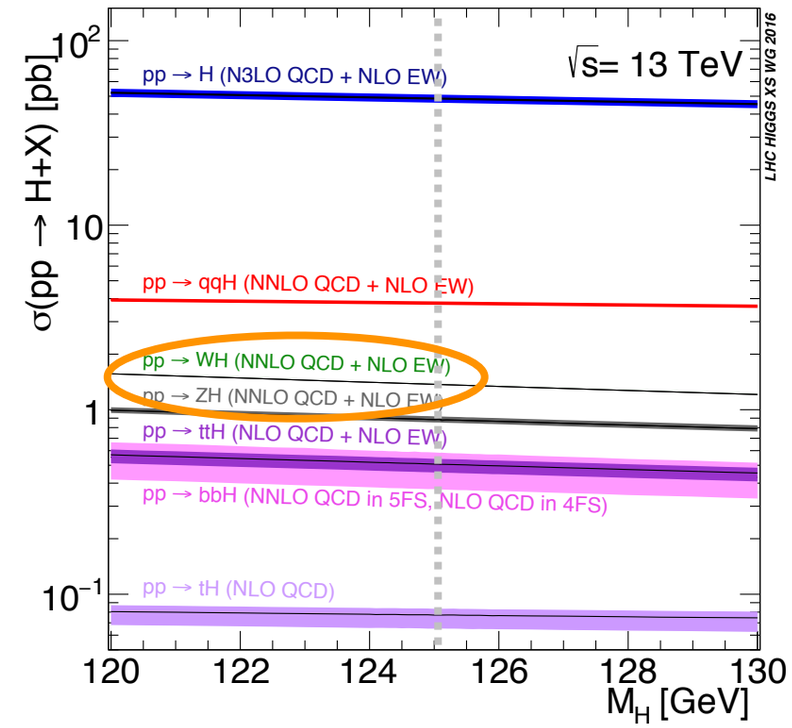


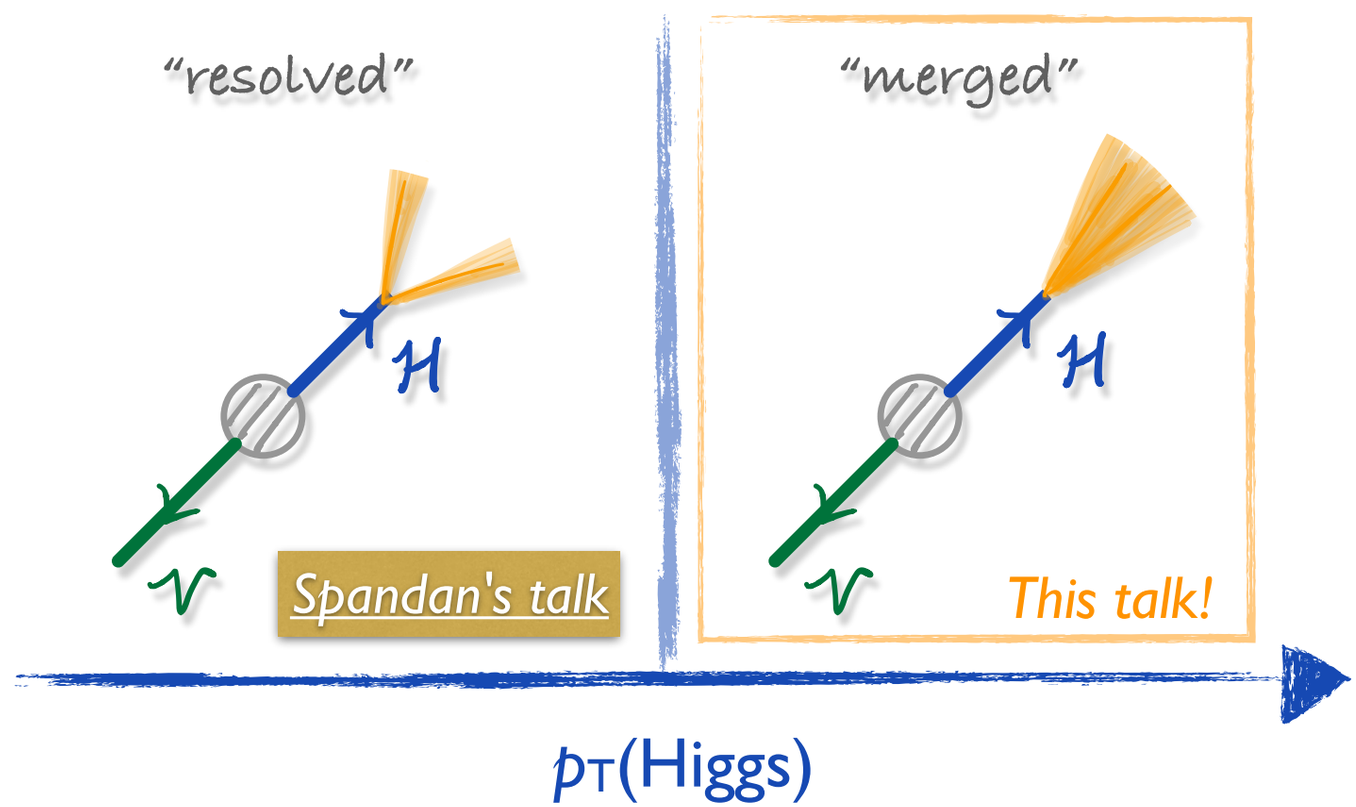



# FIRST DIRECT H→CC SEARCH IN CMS

- Exploits the VH production
  - leptonic V decay: Z→νν, W→lν, Z→ll
    - 3 mutually exclusive channels: 0L, 1L, and 2L (L = e, μ)
  - provides handles for event triggering and QCD background suppression
  - main backgrounds
    - W/Z + jets, ttbar, diboson
- Two complimentary approaches to fully explore the H→cc decay topology
  - resolved-jet topology: reconstruct H→cc decay with two resolved jets (R=0.4)
  - merged-jet topology: reconstruct H→cc decay with one large-R jets (R=1.5)
  - advanced charm-tagging techniques exploited

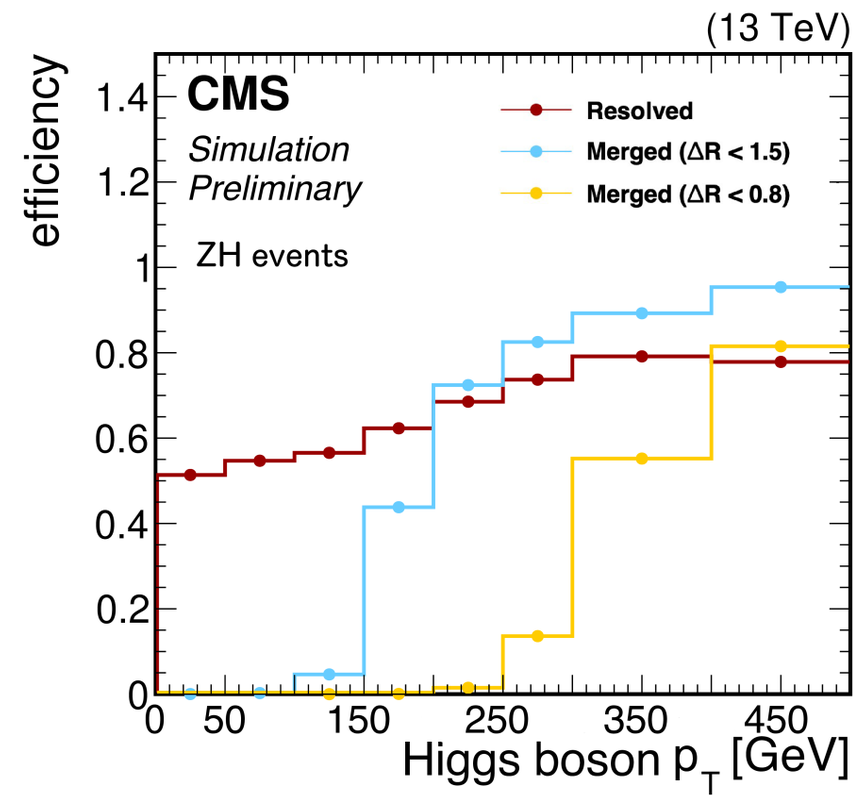


# Higgs Boson Reconstruction

- The cornerstone of the merged-jet analysis is the reconstruction of the H→cc decay with a single large-R jet
  - focus on the boosted regime
    - better signal purity: the $p_T$ spectrum in VH signals is harder than that in V+jets backgrounds
    - but lower signal acceptance: falling $p_T$ spectrum in both signal and backgrounds
  - choosing a suitable jet size
    - angular separation of the decay products $\Delta R \sim 2m_H / p_T$
    - R = 1.5 jets:
      - good efficiency to capture both quarks from Higgs with $p_T > \sim 150$ GeV
      - balance between signal purity and acceptance
  - capturing the showers of the two charm quarks in one jet can potentially lead to a better exploitation of the correlation between the two quarks from the Higgs decay

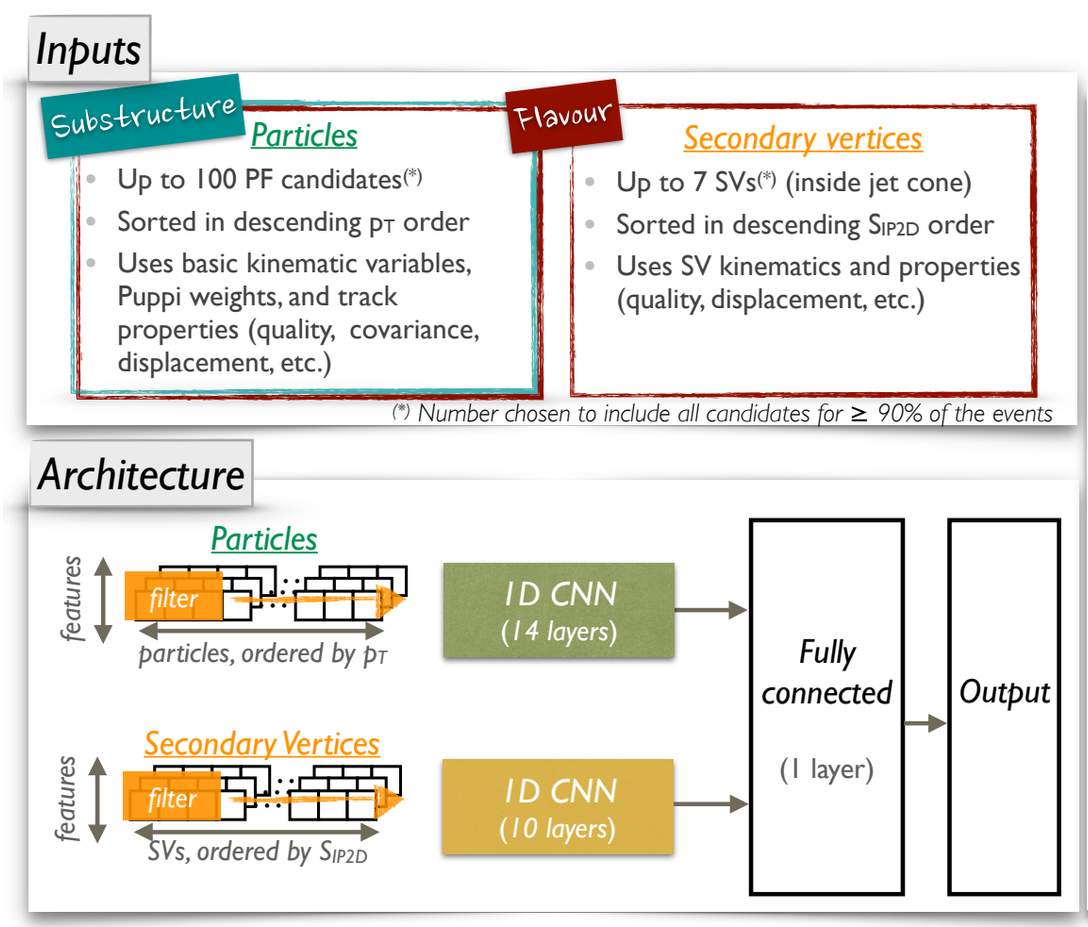



*Reconstruction efficiency:*

- *Merged (R=0.8 / R=1.5): both quarks contained in an AK8 / AK15 jet (with ΔR(jet, c-quark) < 0.8 / 1.5)*
- *Resolved: each quark is reconstructed as a resolved R=0.4 jet with $p_T$>25 GeV and |η|<2.4*

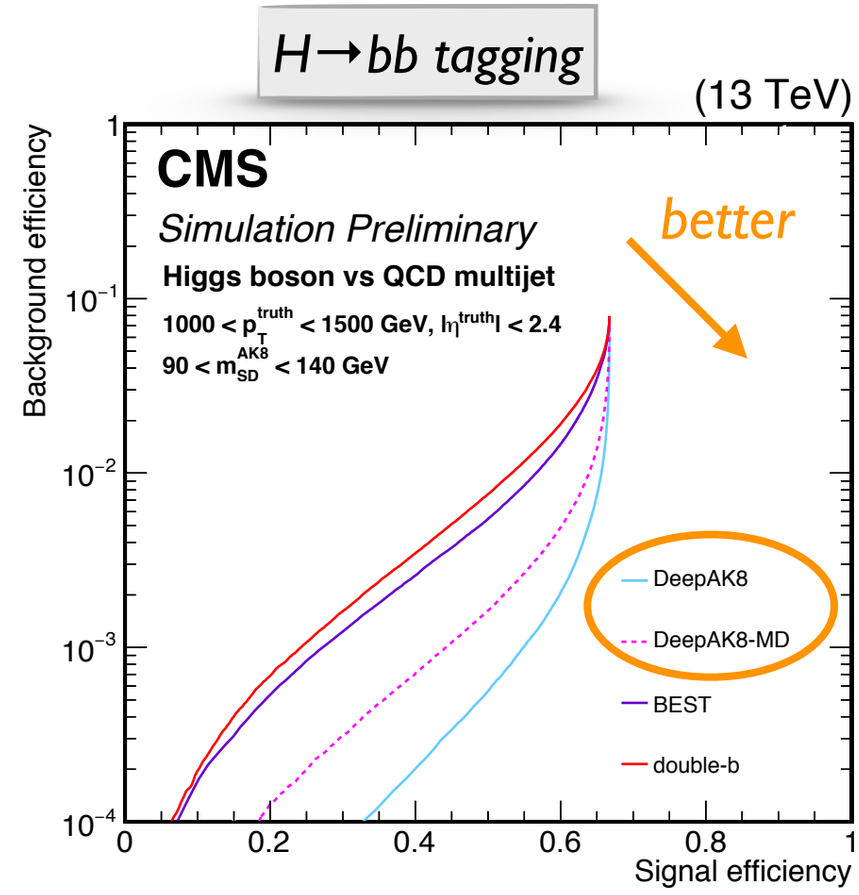
# H→cc Identification

- Advanced machine learning-based algorithm to identify the H→cc decay: "DeepAK8"
  - multi-class classifier for top quark and W, Z, Higgs boson tagging
    - sub-classes based on decay modes (e.g., Z→bb, Z→cc, Z→qq)
    - output scores can be aggregated/transformed for different tasks -> highly versatile tagger
  - uses deep neural networks to directly process jet constituents (PF candidates / secondary vertices)
    - architecture: ResNet inspired 1D convolutional neural networks
  - significant performance improvement




Output

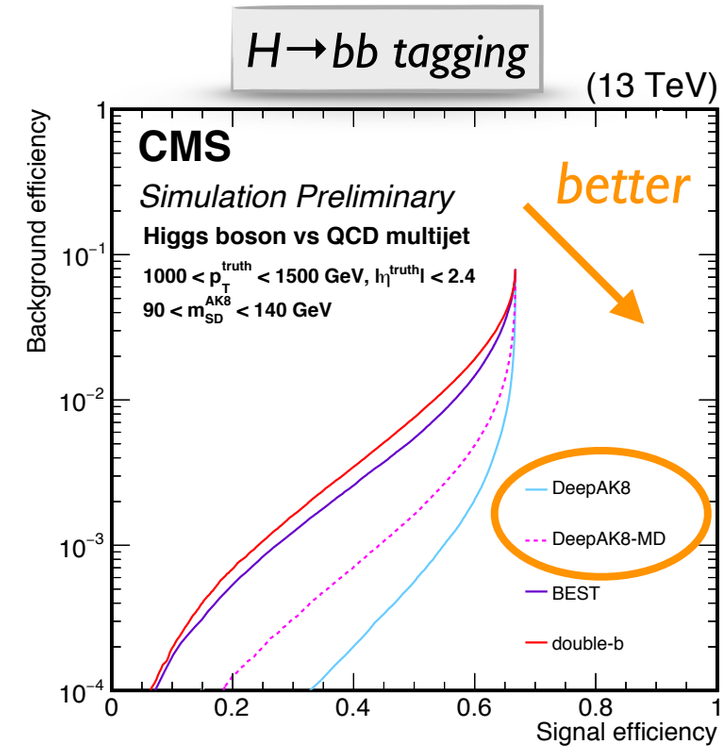
| Category | Label |
|---|---|
| Higgs | H (bb) |
| | H (cc) |
| | H (VV*→qqqq) |
| Top | top (bcq) |
| | top (bqq) |
| | top (bc) |
| | top (bq) |
| W | W (cq) |
| | W (qq) |
| Z | Z (bb) |
| | Z (cc) |
| | Z (qq) |
| QCD | QCD (bb) |
| | QCD (cc) |
| | QCD (b) |
| | QCD (c) |
| | QCD (others) |

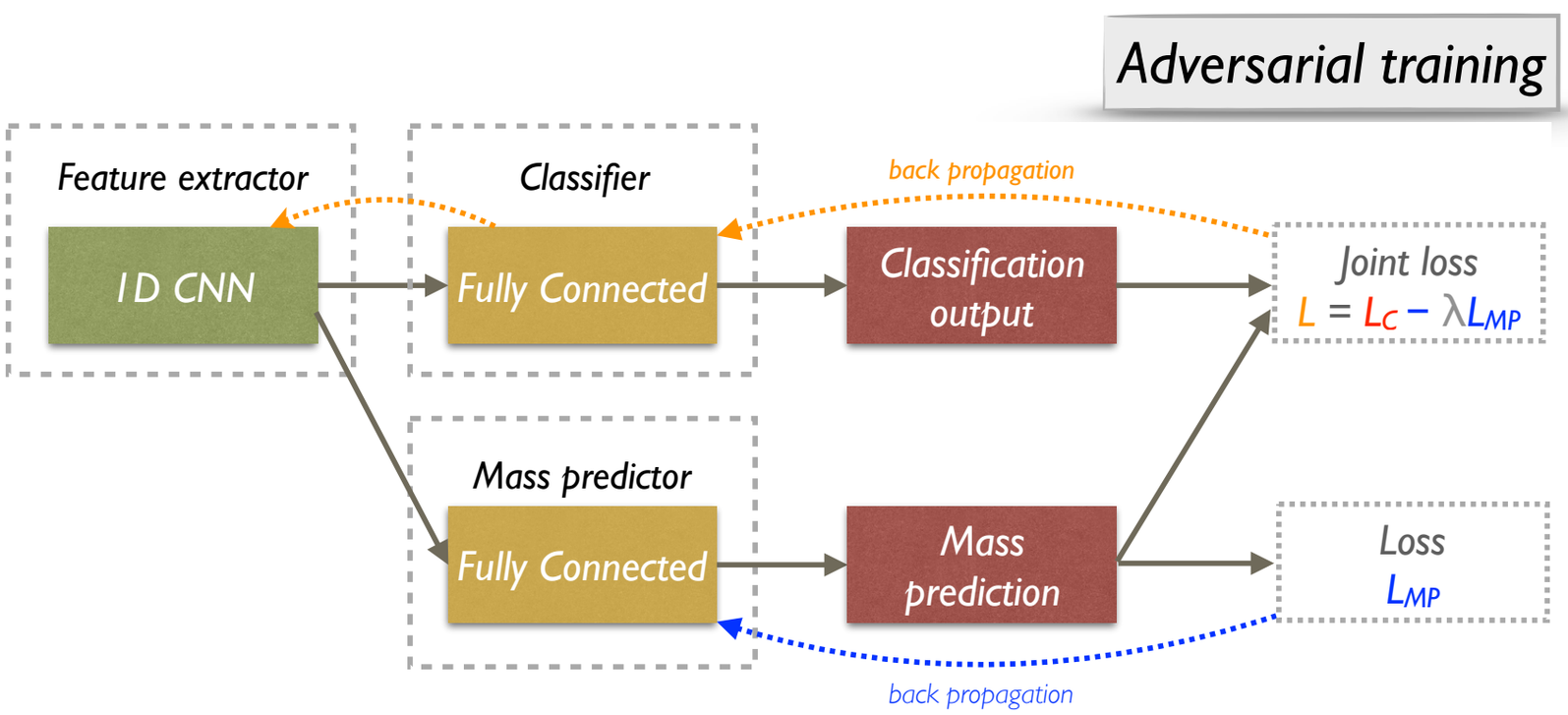


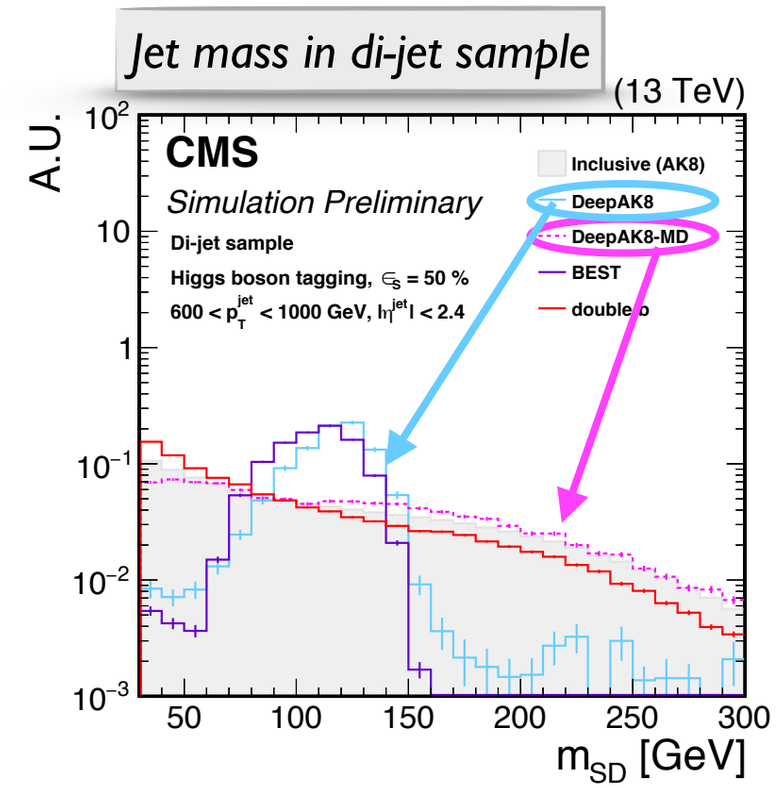

CMS-PAS-JME-18-002

# H→cc Identification (II)

- Mass-decorrelated tagger: "DeepAK8-MD"
  - the nominal version of DeepAK8 shows significantly improved performance, but also features strong "mass sculpting"
    - i.e., modification of the jet mass shape in background samples after tagging requirements
  - dedicated version designed to minimize mass sculpting
    - using "adversarial training" technique
  - significantly reduced mass sculpting yet still strong performance
    - allows us to fit the mass distribution for signal extraction




CMS-PAS-JME-18-002

# H→cc Identification (III)

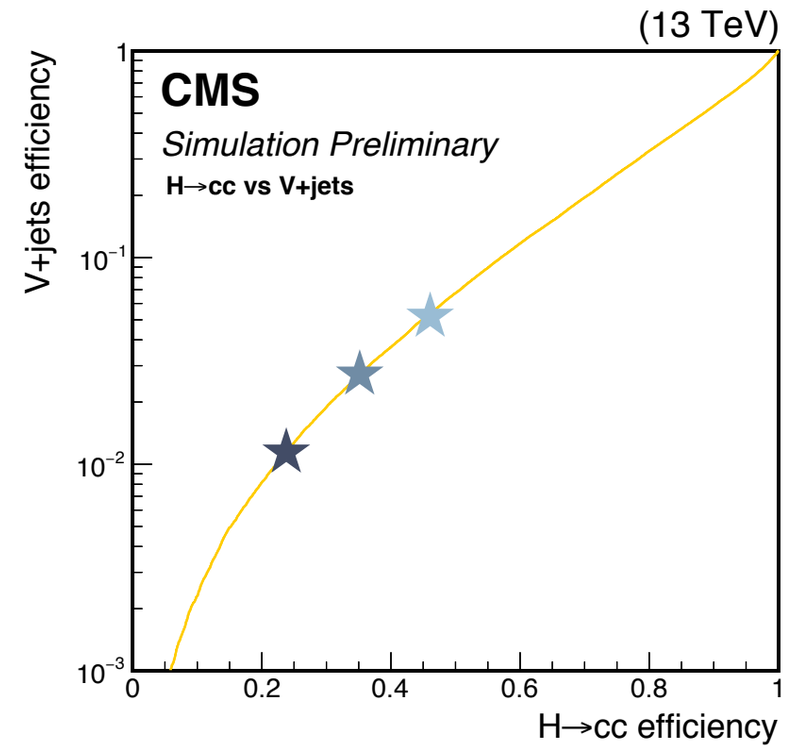
- The DeepAK8-MD algorithm has been adapted to R=1.5 jets for the H→cc analysis with a dedicated training
  - cc-tagging discriminant defined as:

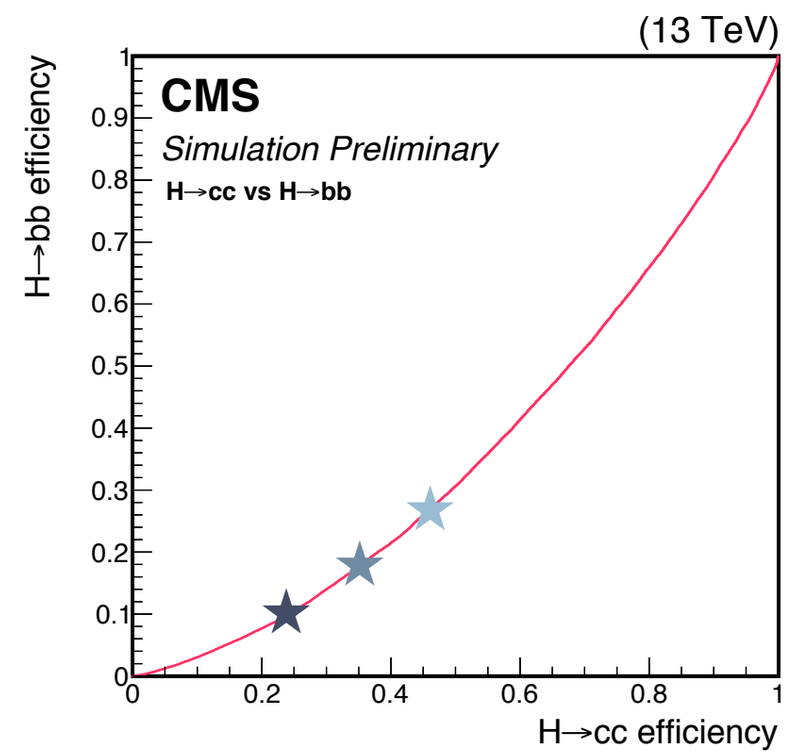
$$\frac{\text{score}(Z \to c\bar{c}) + \text{score}(H \to c\bar{c})}{\text{score}(Z \to c\bar{c}) + \text{score}(H \to c\bar{c}) + \text{score}(QCD)}$$

  - right: performance in MC
- Three working points defined:

| | Loose | Medium | Tight |
|---|---|---|---|
| cc-discriminant | >0.72 | >0.83 | >0.91 |
| ε(V+jets) | 5% | 2.5% | 1% |
| ε(H→cc) | 46% | 35% | 23% |
| ε(H→bb) | 27% | 17% | 9% |

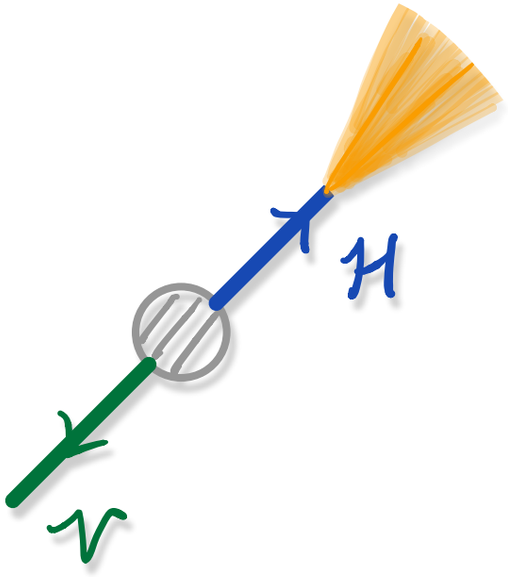
- Events are categorized into three mutually exclusive categories, based on the 3 WPs, to improve sensitivity
  - high/medium/low purity (HP/MP/LP) categories
- cc-tagging discriminant calibrated in data
  - using "proxy" jets from g(gluon)→cc
    - similar characteristics as signal jets
  - scale factors applied to H→cc / Z→cc jets



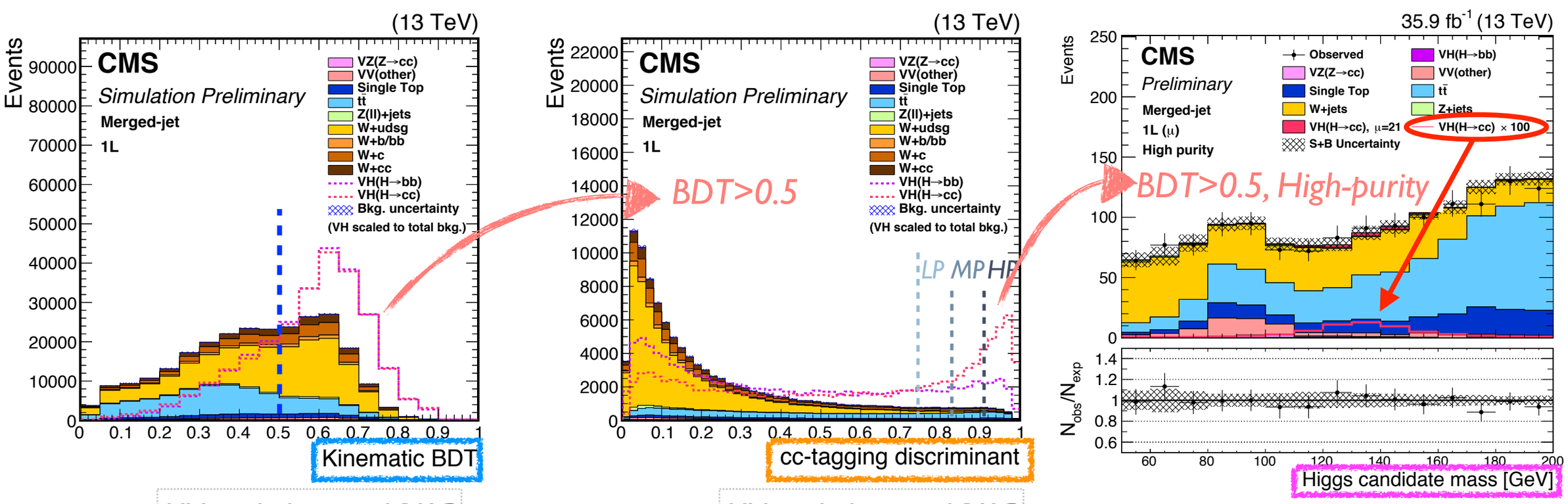
# Baseline Event Selection

- VH events have a clear signature
  - vector boson recoiling against the Higgs boson
  - little additional activity in the event
- Vector boson reconstructed with lepton and/or missing transverse momentum (MET)
  - 2L: V:=opposite-sign same-flavor lepton pair; 75 < m(LL) < 105 GeV [compatible w/ Z mass]
  - 1L: V:=lepton + MET; $\Delta\varphi$(lep, MET)<2.0 [compatible w/ W decay]
  - 0L: V:=MET; MET>170 GeV [due to trigger requirement], $\Delta\varphi$(MET, j)>0.5, $\Delta\varphi$(pfMET, tkMET)<0.5 [suppress QCD]
- Baseline selection
  - high $p_T$ (>200GeV) vector boson and $H_{cand}$, back-to-back ($\Delta\varphi(V, H_{cand})>2.5$)
  - the large-R jet leading in $p_T$ selected as the Higgs candidate ($H_{cand}$)
    - requires $p_T$ > 200 GeV, soft-drop (SD) groomed jet mass $m_{SD}(H_{cand}) \in [50, 200]$ GeV
  - veto events with additional R=0.4 jets ($\Delta R(j, H_{cand})>1.5$) to suppress ttbar contribution



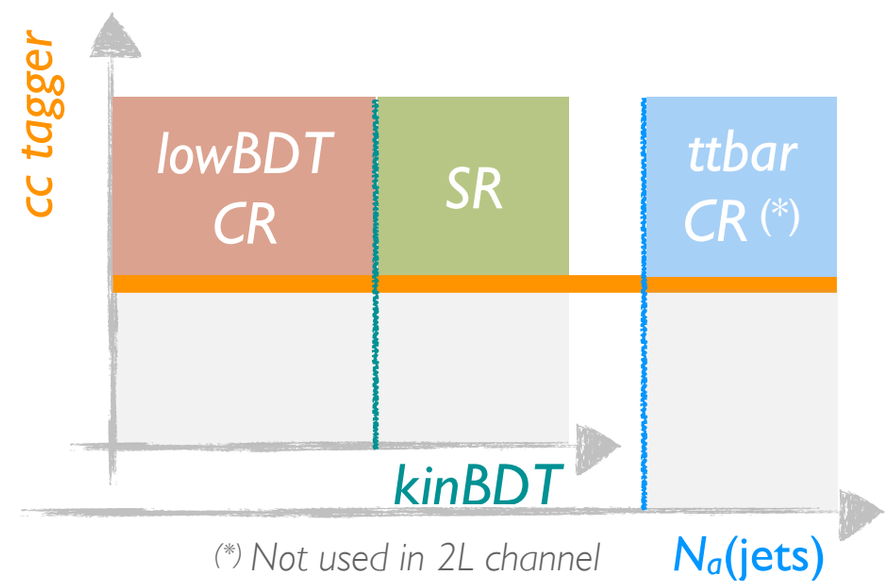
# Analysis Strategy

- Analysis strategy overview
  - event-level kinematic BDT developed in each channel to better suppress the dominant backgrounds (V+jets, ttbar)
    - using only event kinematics, NOT the intrinsic properties (e.g., flavor/mass) of $H_{cand}$
  - cc-tagging discriminant used to select cc-flavor jets and reject light/bb-flavor jets
  - distinct $m(H_{cand})$ shapes between signal and V+jets/ttbar background: fit the $m(H_{cand})$ shape to extract the $H \rightarrow cc$ signal
- Kinematic BDT, cc-tagging discriminant and $m(H_{cand})$ largely independent of each other
  - allowing for a simple and robust strategy for background estimation and signal extraction



# Signal Extraction

- The VH(cc) signal is extracted via a binned fit to the mass of the Higgs candidate [$m(H_{cand})$]
  - $m(H_{cand})$ shapes taken directly from MC
    - validated in control regions: very good data/MC agreement
- Dedicated control regions (CRs) are set up to constrain the normalizations of major backgrounds
  - V+jets: use low BDT region (i.e., BDT<0.5)
  - ttbar: invert the cut on N(additional R=0.4 jets) (i.e, $N_{aj}$>=2)
    - only for 0L and 1L; ttbar contribution is negligible for 2L
  - CRs designed to have similar flavour composition as SRs
    - by applying the same cc-tagging requirement as the corresponding SR
- Normalization of the major backgrounds (V+jets and ttbar) are obtained via a simultaneous fit of SR and the CRs
  - effects of the mistag SFs of the cc-tagging discriminant will be taken into account
    - because the same cc-tagging requirement is applied in CRs and the SR
    - therefore, cc-tagging SFs only needed for VH(cc)/VZ(cc) (not needed for BKGs)




*Full analysis validated in two data samples:*

- ✓ *low $p_T(V)$*
- ✓ *low values of the cc-discriminant*

# SYSTEMATICS

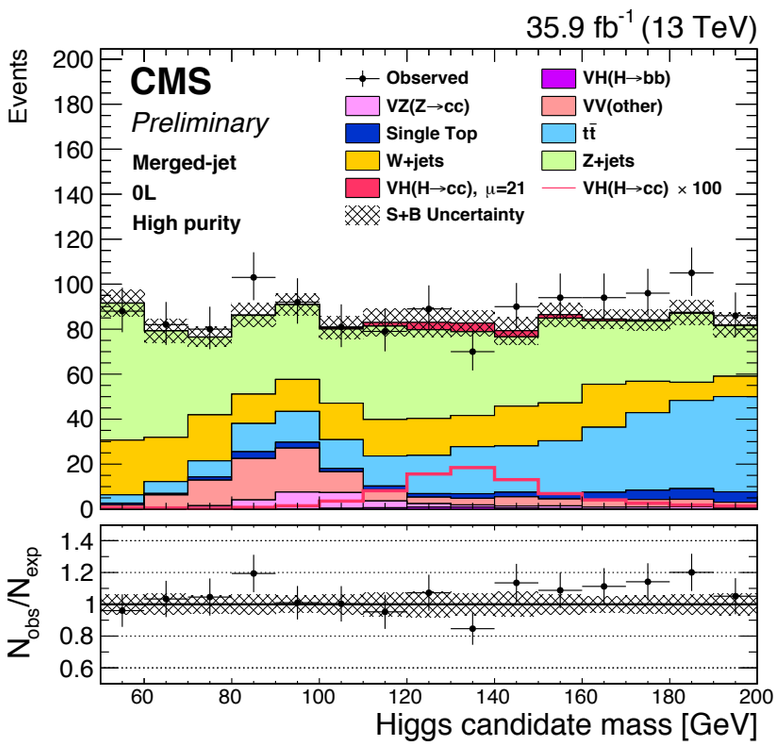
| Source | Type | 0-lepton | 1-lepton | 2-lepton |
|---|---|---|---|---|
| Size of simulated samples | shape | ✓ | ✓ | ✓ |
| Jet energy scale | shape | ✓ | ✓ | ✓ |
| Jet energy resolution | shape | ✓ | ✓ | ✓ |
| MET unclustered energy | shape | ✓ | ✓ | |
| c tagging efficiency | shape | ✓ | ✓ | ✓ |
| Lepton efficiency | shape (rate) | | ✓ | ✓ |
| Pileup reweighting | shape | ✓ | ✓ | ✓ |
| top $p_{\mathrm{T}}$ reweighting | shape | ✓ | ✓ | ✓ |
| $p_{\mathrm{T}}(\mathrm{V})$reweighting | shape | ✓ | ✓ | ✓ |
| PDF | shape | ✓ | ✓ | ✓ |
| Renormalization and factorization scales | shape | ✓ | ✓ | ✓ |
| VH: $p_{\mathrm{T}}(\mathrm{V})$NLO EWK correction | shape | ✓ | ✓ | ✓ |
| Luminosity | rate | 2.5% | 2.5% | 2.5% |
| MET trigger efficiency | rate | 2% | | |
| Single top cross section | rate | 15% | 15% | 15% |
| Diboson cross section | rate | 10% | 10% | 10% |
| VH: cross section (PDF) | rate | ✓ | ✓ | ✓ |
| VH: cross section (scale) | rate | ✓ | ✓ | ✓ |

- Dominant sources:
  - size of the MC simulation / data control samples, cc-tagging, simulation modeling

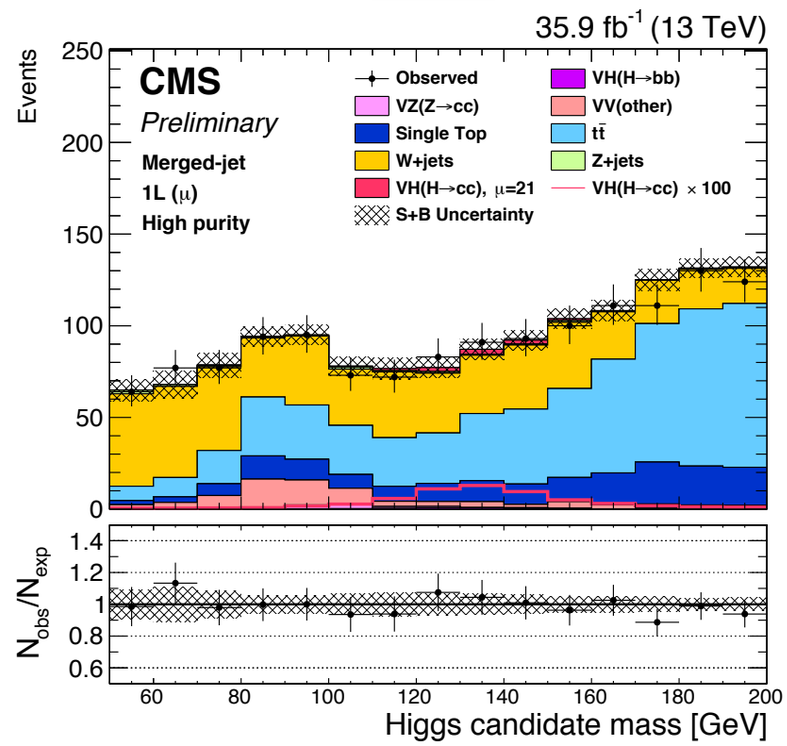
# RESULTS: POST-FIT DISTRIBUTIONS

- Good agreement between the predicted background and the observed data

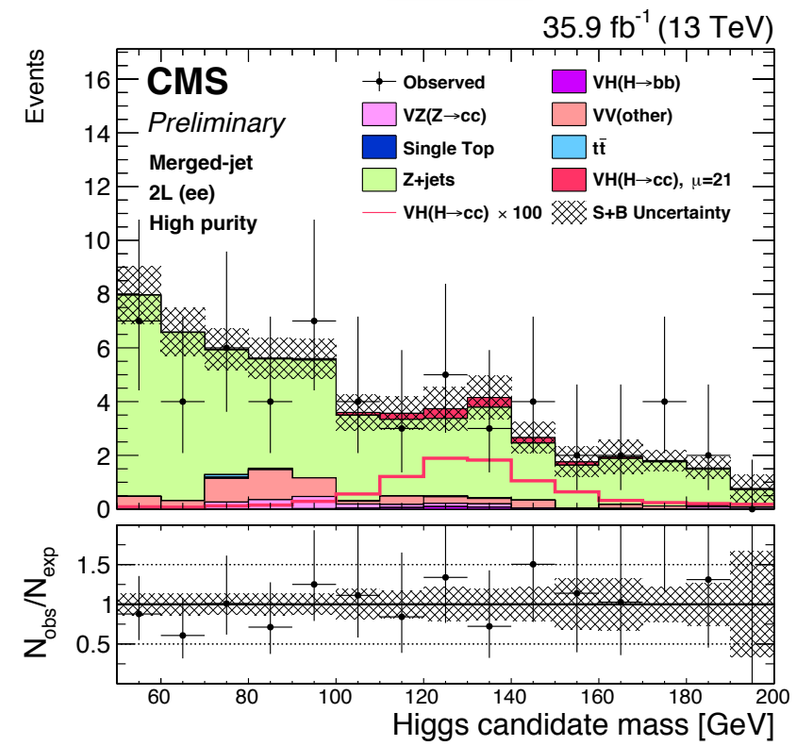
0L




1L (μ)




2L (ee)



# VZ(cc) Validation

- The full procedure of this analysis is validated by measuring the VZ(cc) process
  - following exactly the same procedure, but extract the VZ(cc) signal strength instead of VH(cc)
  - VH(cc) fixed to the SM expectation
- Results:
  - best-fit signal strength: $\mu_{VZ(cc)} = 0.69^{+0.89}_{-0.75}$
    - consistent with SM expectation ($\mu_{VZ(cc)}$=1) within uncertainty
  - observed (expected) significance: 0.9 (1.3) $\sigma$

# VH(cc) Results

- Upper limits on the signal strength $\mu_{VH(cc)}$ at 95% confidence level
  - $\mu_{VH(cc)}$ < 71 obs. ( $49^{+24}_{-15}$ exp.)

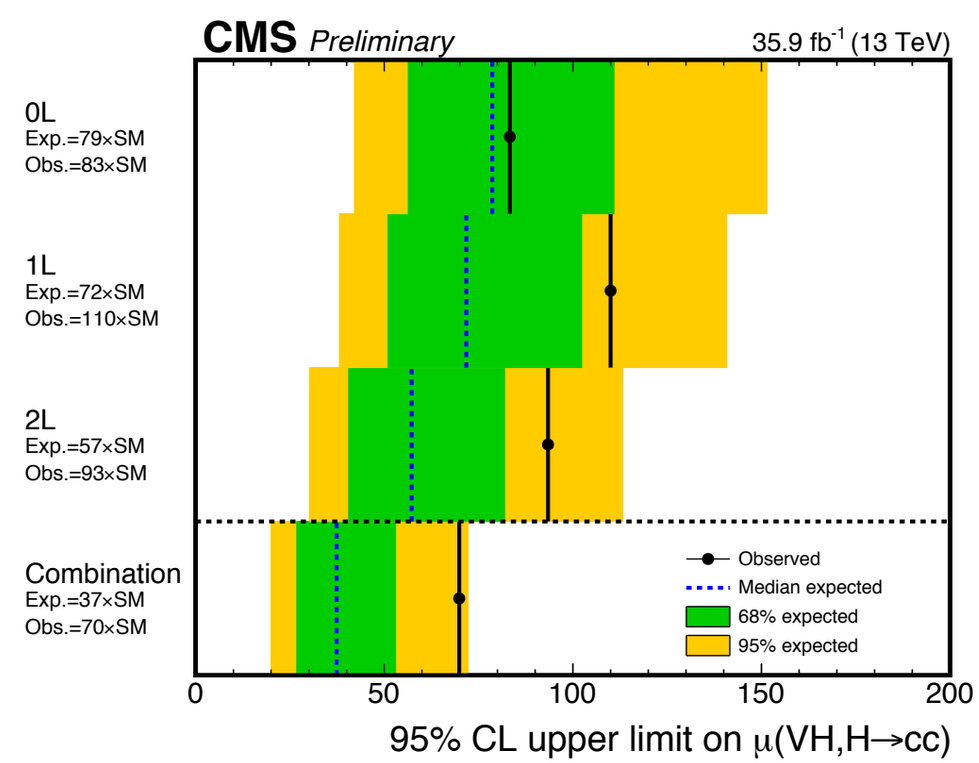
| | Merged-jet (inclusive) | | | |
|---|---|---|---|---|
| | 0L | 1L | 2L | All channels |
| Expected UL | $81^{+39}_{-24}$ | $88^{+43}_{-27}$ | $90^{+48}_{-29}$ | $49^{+24}_{-15}$ |
| Observed UL | 74 | 120 | 76 | 71 |

- Best-fit signal strength: $\mu_{VH(cc)} = 21^{+26}_{-24}$

- Results are combined with resolved-jet analysis
  - to remove overlap, requires:
    - $p_T(V)$ < 300 GeV for the resolved-jet topology
    - $p_T(V)$ >= 300 GeV for the merged-jet topology
      - "inclusive" merged-jet analysis requires $p_T(V)$ > 200 GeV

*Upper limits at 95% confidence level*

| | resolved-jet ($p_T(V) < 300$ GeV) | merged-jet ($p_T(V) \geq 300$ GeV) | combination |
|---|---|---|---|
| expected | $45^{+18}_{-13}$ | $73^{+34}_{-22}$ | $37^{+16}_{-11}$ |
| observed | 86 | 75 | 70 |



# SUMMARY

- A search for the Higgs boson decaying to charm quarks using large-radius jets with the CMS experiment is presented
  - a novel approach
    - reconstructs both quarks from the Higgs decay with a single large-R jet
    - utilizes an advanced ML-based algorithm to identify H→cc decays
  - very competitive results
    - an observed (expected) upper limit on the VH production cross section times the H→cc branching ratio of 71 (49) times the SM expectation
- Still, a long way ahead
  - so far we have explored only ~25% of the collected Run 2 data, and less than ~1% of the full expected dataset of the (HL-)LHC
  - needs breakthroughs in many areas:
    - better charm quark (pair) identification algorithm
    - more advanced signal extraction / background estimation methods
    - reduced systematics with improved event generators / simulation tools
    - upgrades of the detector (tracking / timing / etc.)
- The charming journey has just started!

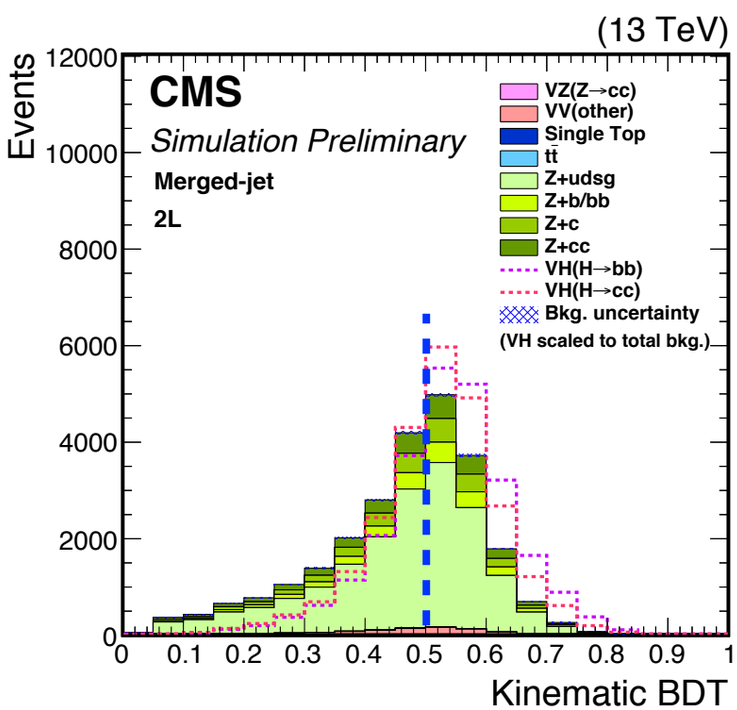
# Backups

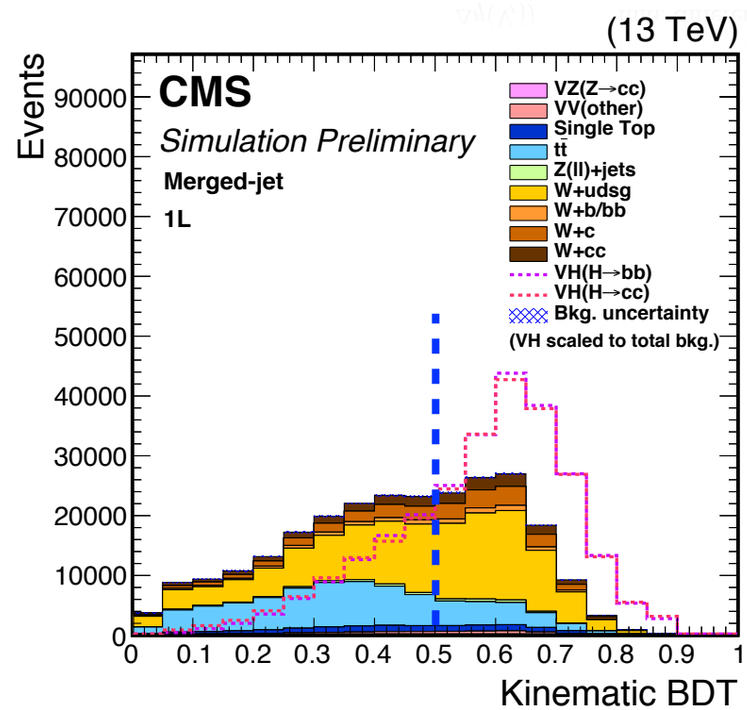
# Kinematic BDT

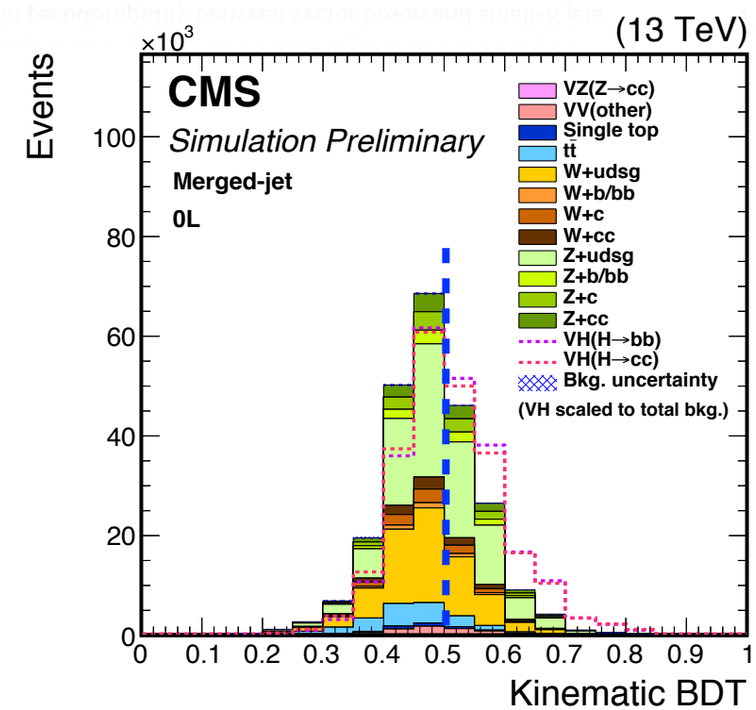
- Kinematic BDT developed to separate VH signals from major backgrounds (V+jets, ttbar)
  - using only event kinematics, NOT the intrinsic properties (e.g., flavor/mass) of $H_{cand}$
  - the resulting BDT is largely uncorrelated with mass and the cc-tagging discriminant of $H_{cand}$
- Two regions are defined based on the BDT
  - search region (SR): high BDT (>=0.5)
  - control region (CR): low BDT (<0.5)

*BDT Inputs*

| Variable | Description | 0L | 1L | 2L |
|---|---|---|---|---|
| $p_T(V)$ | vector boson transverse momentum | ✓ | ✓ | ✓ |
| $p_T(H_{cand})$ | $H_{cand}$ transverse momentum | ✓ | ✓ | ✓ |
| $\lvert\eta(H_{cand})\rvert$ | absolute value of the $H_{cand}$ pseudorapidity | ✓ | | |
| $\Delta\phi(V, H_{cand})$ | azimuthal angle between vector boson and $H_{cand}$ | ✓ | ✓ | ✓ |
| $p_T^{miss}$ | missing transverse momentum | | ✓ | |
| $\Delta\eta(H_{cand}, \ell)$ | difference in pseudorapidity between $H_{cand}$ and the lepton | | ✓ | |
| $\Delta\eta(H_{cand}, V)$ | difference in pseudorapidity between $H_{cand}$ and vector boson | | | ✓ |
| $\Delta\eta(H_{cand}, j)$ | min. difference in pseudorapidity between $H_{cand}$ and small-$R$ jets | ✓ | ✓ | ✓ |
| $\Delta\eta(\ell, j)$ | min. difference in pseudorapidity between the lepton and small-$R$ jets | | ✓ | |
| $\Delta\eta(V, j)$ | min. difference in pseudorapidity between vector boson and small-$R$ jets | | | ✓ |
| $\Delta\phi(\vec{p}_T^{miss}, j)$ | azimuthal angle between $\vec{p}_T^{miss}$ and closest small-$R$ jet | ✓ | | |
| $\Delta\phi(\vec{p}_T^{miss}, \ell)$ | azimuthal angle between $\vec{p}_T^{miss}$ and lepton | | ✓ | |
| $m_T$ | transverse mass of lepton $\vec{p}_T + \vec{p}_T^{miss}$ | | ✓ | |
| $N_{aj}$ | number of small-$R$ jets | ✓ | ✓ | ✓ |




*Signal scaled to total BKG*

# Calibration of cc-tagging Discriminant

- cc-tagging discriminant calibrated via proxy jets
  - impossible to isolate a pure Z/H→cc sample...
  - instead, uses proxy jets (gluon→cc) that share similar characteristics as signal jets
  - corrections are then transferred to signal jets
- Proxy jets obtained from a di-jet sample
  - requires the presence of at least one secondary vertex in each subjet
    - similar cc-tagging discriminant shapes between proxy and signal jets after this selection
    - further enhances g→bb/cc fraction
- Template fit method used to extract the data/MC scale factors (SFs)
  - define 3 MC templates: bb(+b), cc(+c) and udsg
  - fit variable: the CSVv2 b-tagging discriminant
- SFs typically between 0.9 to 1.4, with 10 - 30% uncertainty
  - also validated in γ+jets sample: consistent results
- SFs applied only on VH(cc) signal and VZ(cc)
  - and bb-mistag SF applied on VH(bb) and VZ(bb)
  - systematics uncertainties propagated
  - not applied on BKG (estimation is data-driven)

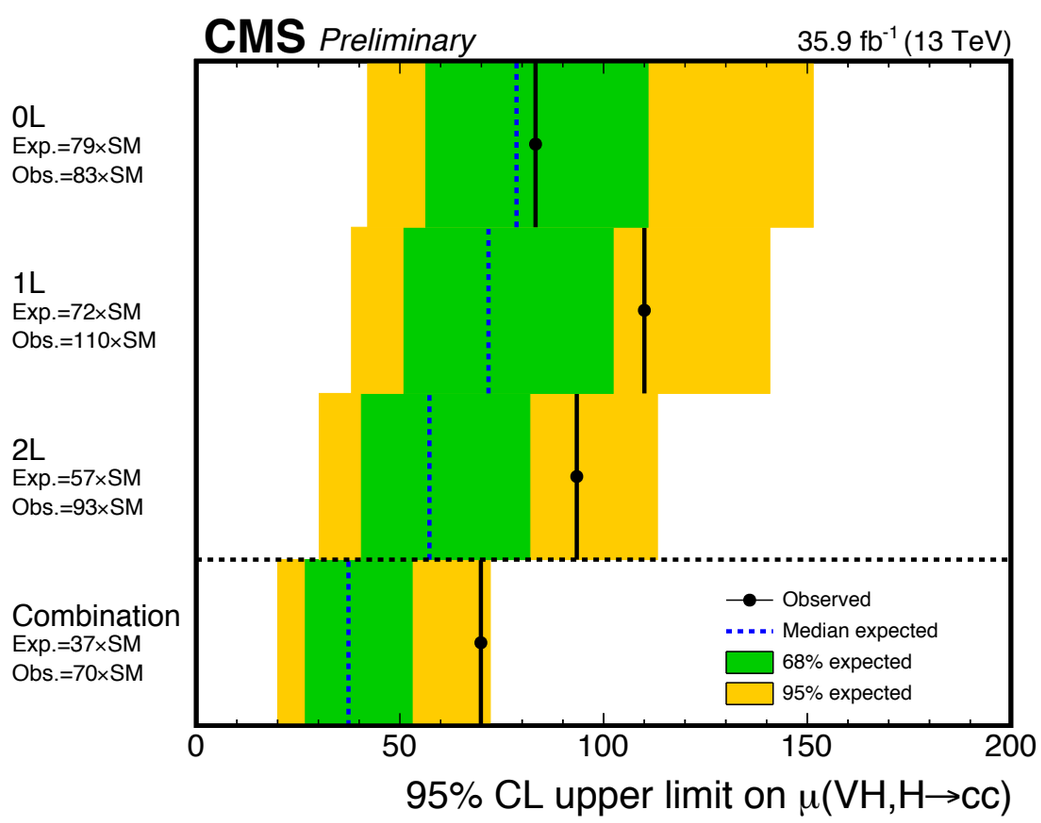
# Results (Resolved & Merged)

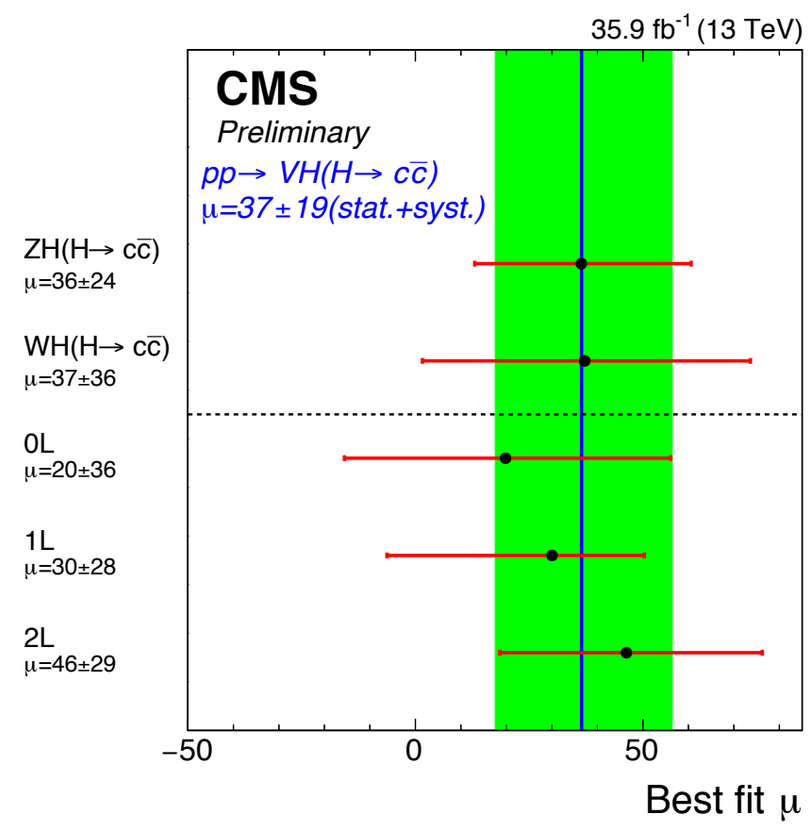
- Resolved & Merged: Inclusive

| | Resolved-jet (inclusive) | | | | Merged-jet (inclusive) | | | |
|---|---|---|---|---|---|---|---|---|
| | 0L | 1L | 2L | All channels | 0L | 1L | 2L | All channels |
| expected UL | 84 | 79 | 59 | 38 | 81 | 88 | 90 | 49 |
| observed UL | 66 | 120 | 116 | 75 | 74 | 120 | 76 | 71 |

- Resolved & Merged: Exclusive & Combination

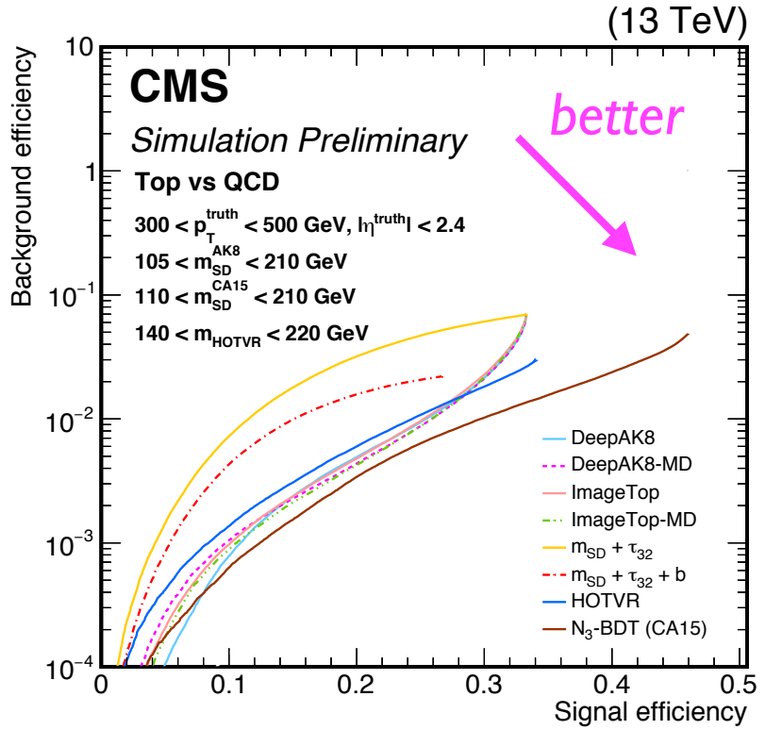
| 95% CL exclusion limit | | | | | | |
|---|---|---|---|---|---|---|
| | resolved-jet ($p_T(V) < 300$ GeV) | merged-jet ($p_T(V) \geq 300$ GeV) | combination | | | |
| | | | 0L | 1L | 2L | All channels |
| expected | $45^{+18}_{-13}$ | $73^{+34}_{-22}$ | $79^{+32}_{-22}$ | $72^{+31}_{-21}$ | $57^{+25}_{-17}$ | $37^{+16}_{-11}$ |
| observed | 86 | 75 | 83 | 110 | 93 | 70 |

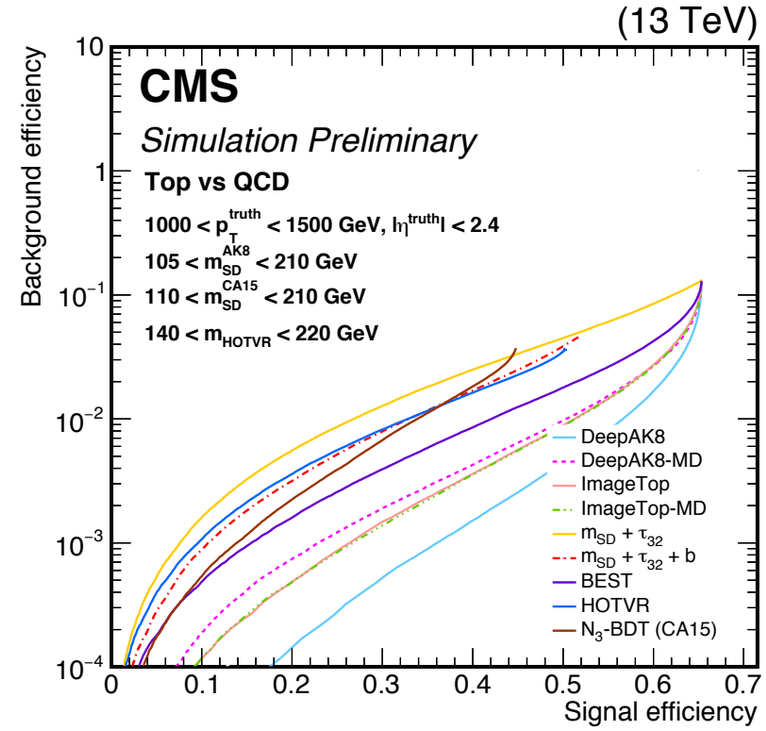


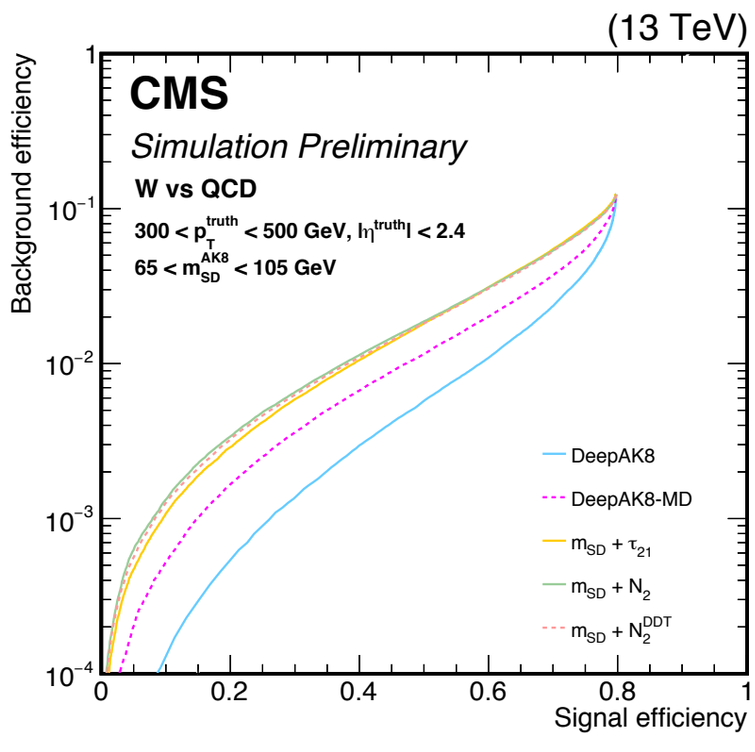
# DEEPAK8

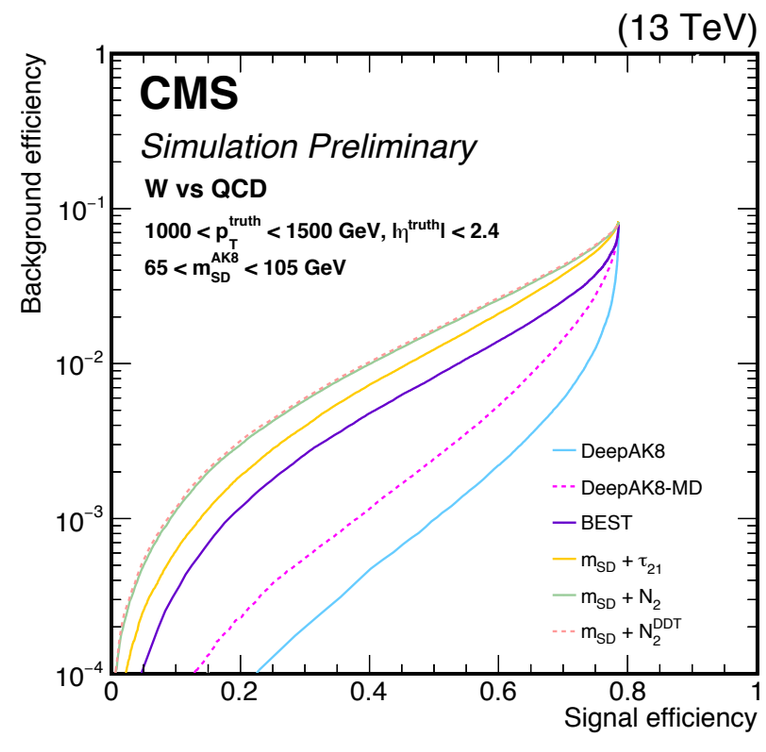
*top vs QCD*




*W vs QCD*



# Ablation Study of DeepAK8

- DeepAK8 shows substantial gain compared to traditional approaches *CMS-PAS-JME-18-002*
- To understand the main sources of the improvement, alternative versions of DeepAK8 were trained using a subset of the input features
  - Particle (kinematics): only kinematic info of PF candidates
    - four momenta, distances to the jet and subjet axes, etc.
  - Particle (w/o Flavour): adding experimental info
    - charge, particle identification, track quality, etc.
  - Particle Full + SV (the full DeepAK8): adding features related to heavy-flavour tagging
    - track displacement, track-vertex association, SV features, etc.