

## 1 Introduction

Four-top quark production ( $t\bar{t}t\bar{t}$ ) is a rare process within the Standard Model (SM) that serves as a crucial probe for both SM and beyond-the-Standard-Model (BSM) physics. Figure 1 illustrates representative leading-order Feynman diagrams for  $t\bar{t}t\bar{t}$  production. The latest predicted SM cross section for  $t\bar{t}t\bar{t}$  production at next-to-leading order (NLO) with next-to-leading logarithmic (NLL') correction is 13.37 fb at  $\sqrt{s} = 13$  TeV and 15.82 fb at  $\sqrt{s} = 13.6$  fb [3]. This process is sensitive to the top quark's Yukawa coupling with the Higgs boson while remaining unaffected by the Higgs boson decay, making it a prime candidate for identifying any deviations from SM predictions regarding the top quark Yukawa coupling. Furthermore, deviations in the four-top production cross section from SM predictions could also signal the presence of undiscovered heavy scalar or pseudo-scalar bosons decaying into top quarks, as proposed in several extensions of the SM, including top-philic resonances as shown by Figure 2.

13

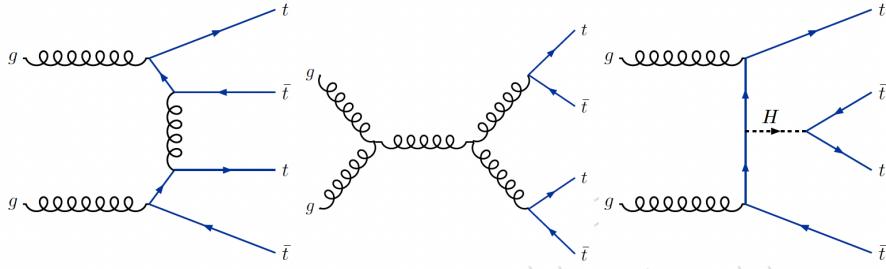


Figure 1: Representative LO Feynman diagrams for SM  $t\bar{t}t\bar{t}$  production.

Each top quark predominantly decays into a bottom quark and a W boson. The W boson then decays either hadronically, producing quarks, or leptonically, yielding a lepton and a neutrino. In this analysis, the fully hadronic final state, where all four top quarks decay entirely into hadrons, is being investigated for the first time, representing about 20% of all four-top quark events.

18

This analysis follows a previous CMS analysis on all-hadronic four-top production with Run2 dataset described in AN2020\_021. The observed significance of the  $t\bar{t}t\bar{t}$  signal in this channel was 2.5 standard deviations (with 0.4 expected). When combined [1] with other channels (Single Lepton, opposite-sign dilepton), the observed significance of the  $t\bar{t}t\bar{t}$  signal was 3.9 standard deviations (with 1.5 expected). In this analysis, we improve upon the previous all-hadronic channel analysis and expand it to include 2022, 2023 data. We will also include an interpretation against the top-philic resonance BSM model[2].

26

The search for four-top quark production in the fully hadronic final state faces significant challenges due to substantial backgrounds from top quark pair ( $t\bar{t}$ ) and QCD multijet production. To address these challenges, specialized machine learning tools for hadronic top quark tagging are utilized. In the "boosted" regime, where top quarks have high transverse momentum ( $p_T$ ) and their decay products produce collimated jets, the Particlenet boosted object tagger is employed for identification. For top quarks in the "resolved" regime, characterized by moderate  $p_T$  and distinct jets resulting from the hadronization of decay products, a dedicated BDT-based resolved top tagger has been developed specifically for this analysis.

The search strategy classifies events into distinct categories based on the number of reconstructed top quarks and the scalar sum of jet transverse momenta ( $H_T$ ). An event-level BDT, trained

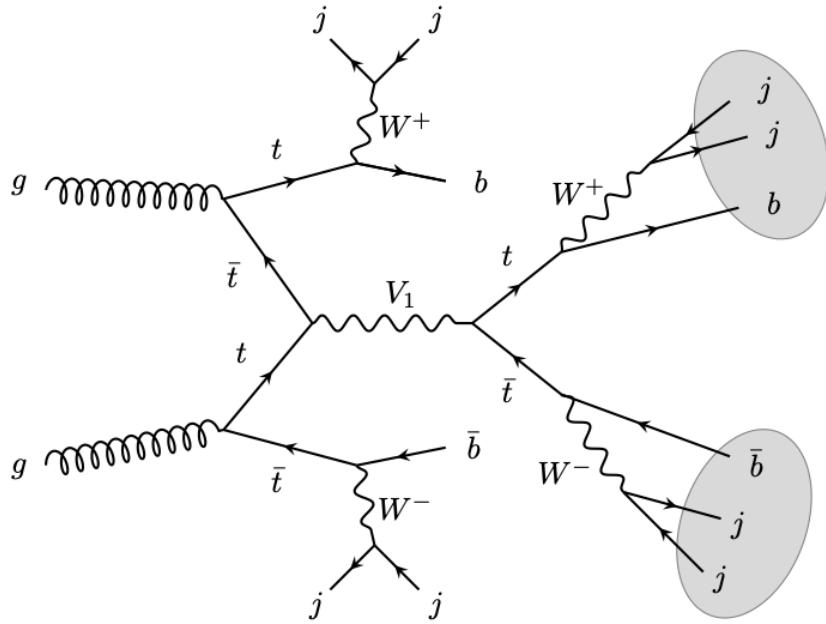


Figure 2: Representative LO Feynman diagrams for tree-level single production of the V1 decaying into the fully-hadronic channel in the BSM top-philic resonance model.

<sup>37</sup> on kinematic variables, is then employed to distinguish the  $t\bar{t}t\bar{t}$  signal from the background. To  
<sup>38</sup> estimate the dominant backgrounds, originating from  $t\bar{t}$  and QCD multijet production, data-driven  
<sup>39</sup> techniques are applied, ensuring accurate background modeling and robust signal extraction.

## 40 2 Triggers and Dataset

<sup>41</sup> The data for this search are recorded using a suite of cross-triggers requiring the presence of  $\geq 6$   
<sup>42</sup> jets,  $\geq 1$  or  $\geq 2$  b-tagged jets, and large  $H_T$ . For part of the 2017 run, a 4-jet, 3-b jet, high  
<sup>43</sup>  $H_T$  trigger is included as well in order to maximize the trigger efficiency. The HLT paths of the  
<sup>44</sup> triggers used for the 2016, 2017, and 2018 data-taking periods are listed in Table 1. The HLT  
<sup>45</sup> paths of the triggers used for the 2022 and 2023 data-taking periods are listed in Table 2. The  
<sup>46</sup> efficiency of these triggers is measured in an independent sample selected with a single muon trigger  
<sup>47</sup> (HLT\_IsoMu24 or HLT\_IsoMu27). The efficiency is measured as a function of  $N_{b-jet}$  and  $N_{jet}$ .  
<sup>48</sup> The efficiency is measured in the region where  $H_T > 900$  GeV, as the trigger efficiency is expected  
<sup>49</sup> to be independent of  $H_T$  in this region. In order to avoid overlapping to the analysis region, one  
<sup>50</sup> or more off-line leptons as defined in subsection 3.2 are required. The efficiency is computed as  
<sup>51</sup> follows:

$$\epsilon(N_j, N_b) = \frac{\text{Number of events passed OR of triggers and denominator selection}}{\text{Number of events that passed HLT_IsoMu24 or HLT_IsoMu27}}. \quad (1)$$

<sup>52</sup> The datasets used for measuring trigger efficiency, as listed in Tables 3 and 4, include events that  
<sup>53</sup> have passed the isolated muon trigger. Figure 1 and 2 show the measured efficiency in the  $N_{jet}$

## 2 Triggers and Dataset

Table 1: HLT paths corresponding to the triggers used for the search in 2016, 2017, and 2018.

| Year | Era                 | Trigger path  |
|------|---------------------|---|
| 2016 | B, C, D, E, F, G, H | HLT_PFHT400_SixJet30_DoubleBTagCSV_p056<br>HLT_PFHT450_SixJet40_BTagCSV_p056  |
| 2017 | B                   | HLT_PFHT380_SixJet32_DoubleBTagCSV_p075<br>HLT_PFHT430_SixJet40_BTagCSV_p080  |
|      | C, D, E, F          | HLT_PFHT380_SixPFJet32_DoublePFBTagCSV_2p2<br>HLT_PFHT430_SixPFJet40_PFBTagCSV_1p5<br>HLT_PFHT300PT30_QuadPFJet_75_60_45_40_TriplePFBTagCSV_3p0 |
| 2018 | A                   | HLT_PFHT380_SixPFJet32_DoublePFBTagDeepCSV_2p2<br>HLT_PFHT430_SixPFJet40_PFBTagDeepCSV_1p5  |
|      | B, C, D             | HLT_PFHT400_SixPFJet32_DoublePFBTagDeepCSV_2p94<br>HLT_PFHT450_SixPFJet36_PFBTagDeepCSV_1p59  |

Table 2: HLT paths corresponding to the triggers used for the search in 2022 and 2023

| Year | Era           | Trigger path   |
|------|---------------|--|
| 2022 | C, D, E, F, G | HLT_PFHT400_SixPFJet32_DoublePFBTagDeepJet_2p94<br>HLT_PFHT450_SixPFJet36_PFBTagDeepJet_1p59 |
| 2023 | C             | HLT_PFHT400_SixPFJet32_DoublePFBTagDeepJet_2p94<br>HLT_PFHT450_SixPFJet36_PFBTagDeepJet_1p59 |
|      | D             | HLT_PFHT400_SixPFJet32_PNet2BTagMean0p50<br>HLT_PFHT450_SixPFJet36_PNetBTag0p35              |

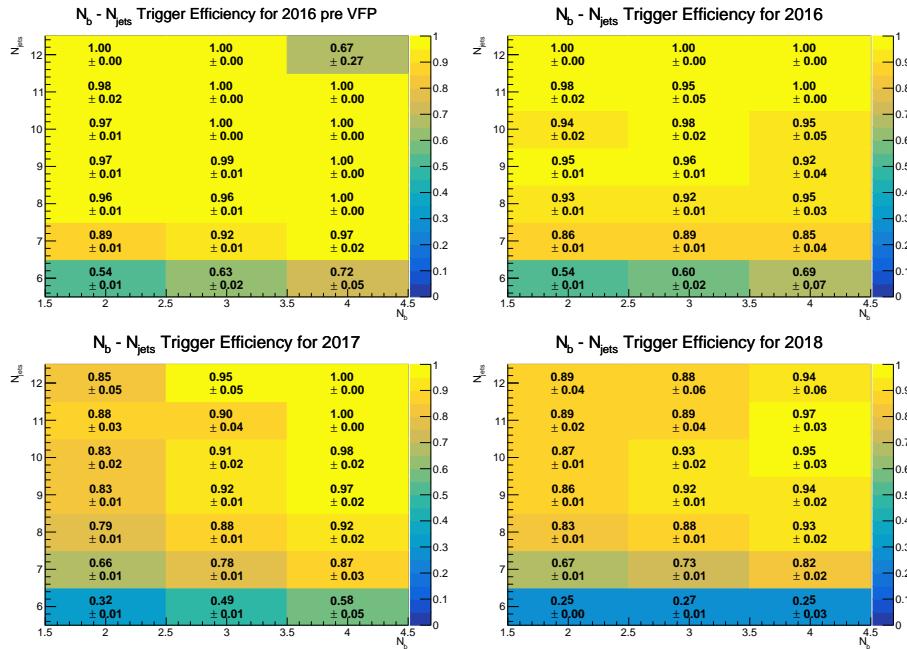


Figure 3: Efficiency measured for search triggers as a function of  $N_{jet}$  and  $N_b-jet$  for 2016 pre VFP, 2016 post VFP, 2017 and 2018 data. The  $N_{jet} = 6$  region used for the validation test is shown as well.

Table 3: Dataset used for run2 trigger efficiency measurement

| Year          | Dataset  |
|---------------|--|
| 2016 pre VFP  | /SingleMuon/Run2016B-ver1_HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD |
|               | /SingleMuon/Run2016C-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD      |
|               | /SingleMuon/Run2016D-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD      |
|               | /SingleMuon/Run2016E-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD      |
|               | /SingleMuon/Run2016F-HIPM_UL2016_MiniAODv2_NanoAODv9-v2/NANOAOD      |
| 2016 post VFP | /SingleMuon/Run2016F-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2016G-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2016H-UL2016_MiniAODv2_NanoAODv9-v1/NANOAOD           |
| 2017          | /SingleMuon/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD           |
|               | /SingleMuon/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD           |
| 2018          | /SingleMuon/Run2018A-UL2018_MiniAODv2_NanoAODv9-v2/NANOAOD           |
|               | /SingleMuon/Run2018B-UL2018_MiniAODv2_NanoAODv9-v2/NANOAOD           |
|               | /SingleMuon/Run2018C-UL2018_MiniAODv2_NanoAODv9-v2/NANOAOD           |
|               | /SingleMuon/Run2018D-UL2018_MiniAODv2_NanoAODv9-v1/NANOAOD           |

Table 4: Dataset used for early run3 trigger efficiency measurement

| Year     | Trigger path                            |
|----------|---|
| 2022     | /Muon/Run2022C-22Sep2023-v1/NANOAOD     |
|          | /Muon/Run2022D-22Sep2023-v1/NANOAOD     |
| 2022EE   | /Muon/Run2022E-22Sep2023-v1/NANOAOD     |
|          | /Muon/Run2022F-22Sep2023-v2/NANOAOD     |
|          | /Muon/Run2022G-22Sep2023-v1/NANOAOD     |
| 2023     | /Muon0/Run2023C-22Sep2023_v1-v1/NANOAOD |
|          | /Muon0/Run2023C-22Sep2023_v2-v1/NANOAOD |
|          | /Muon0/Run2023C-22Sep2023_v3-v1/NANOAOD |
|          | /Muon0/Run2023C-22Sep2023_v4-v1/NANOAOD |
|          | /Muon1/Run2023C-22Sep2023_v1-v1/NANOAOD |
|          | /Muon1/Run2023C-22Sep2023_v2-v1/NANOAOD |
|          | /Muon1/Run2023C-22Sep2023_v3-v1/NANOAOD |
|          | /Muon1/Run2023C-22Sep2023_v4-v1/NANOAOD |
| 2023BPix | /Muon0/Run2023D-22Sep2023_v1-v1/NANOAOD |
|          | /Muon0/Run2023D-22Sep2023_v2-v1/NANOAOD |
|          | /Muon1/Run2023D-22Sep2023_v1-v1/NANOAOD |
|          | /Muon1/Run2023D-22Sep2023_v2-v1/NANOAOD |

## 2 Triggers and Dataset

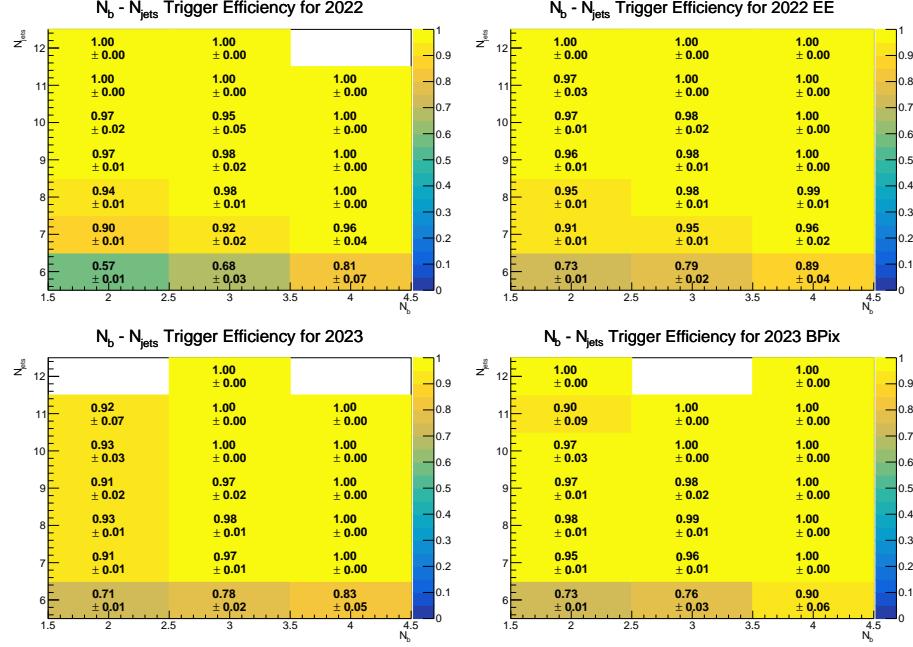


Figure 4: Efficiency measured for search triggers as a function of  $N_{jet}$  and  $N_{b-jet}$  for 2016 pre VFP, 2016 post VFP, 2017 and 2018 data. The  $N_{jet} = 6$  region used for the validation test is shown as well.

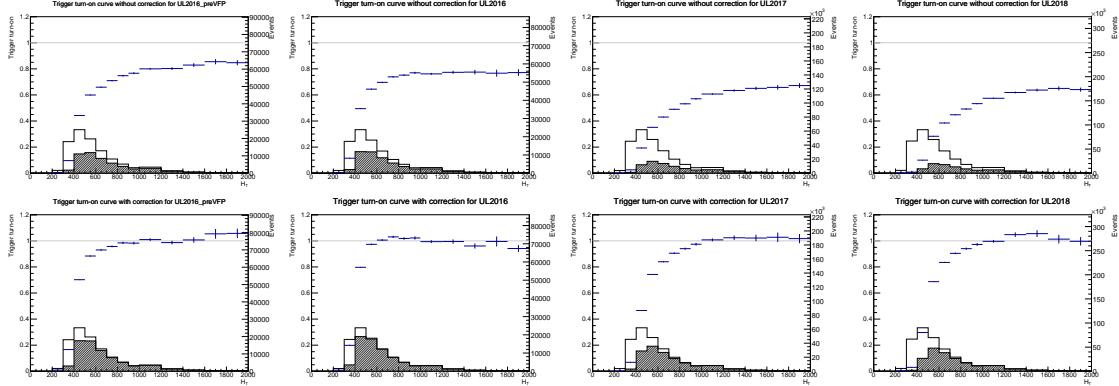


Figure 5: Trigger turn-on vs  $H_T$  before (top) and after (bottom) correcting for efficiencies measured as a function of  $N_{jet}$  and  $N_{b-jet}$  for 2016 pre VFP, 2016 post VFP, 2017 and 2018 data.

and  $N_{b-jet}$  plane for each year. There is a clear dependence on  $N_{jet}$  and  $N_{b-jet}$ , which justifies the necessity of corrections depending on  $N_{jet}$  and  $N_{b-jet}$ . The bottom row corresponds to  $N_{jet} = 6$  and shows low efficiencies due to the requirement of  $N_{jet} \geq 6$  in trigger paths. This region is used only for the validation of the background estimation methods. Details will be discussed in Section 5.

The trigger paths include selection criteria based on  $N_{jet}$ ,  $N_{b-jet}$ , and  $H_T$ . Ideally, the trigger efficiency should be measured as a function of all three variables. However, due to limited statistical precision, the efficiency is measured as a function of  $N_{jet}$  and  $N_{b-jet}$ . To validate that this approach

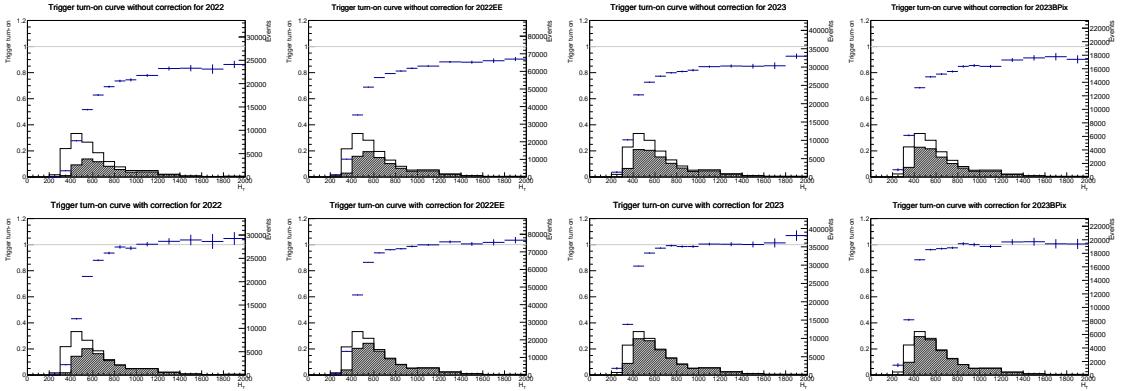


Figure 6: Trigger turn-on vs  $H_T$  for 2022 and 2023 data. Upper plot shows

is sufficient, the trigger turn-on curve is used. Figure 3 and 4 show the  $H_T$  turn-on for each year before and after applying the  $N_{jet}$  and  $N_{b-jet}$  dependent corrections. Before the corrections, the efficiency in the plateau region ( $H_T \geq 900$  GeV) deviates from unity. After the corrections are applied, the plateau region is close to 1. This behavior shows that the chosen approach is sufficient for measuring the trigger efficiency. The measured trigger efficiency is directly used as trigger SF to MC because the trigger efficiency is not applied at MC in this analysis.

## 68 3 Object Selection

### 69 3.1 Vertex Selection

70 Standard CMS vertex selection criteria corresponding to the “goodVertices” flag in NanoAOD is  
71 applied. Specifically, each event is required to have at least one primary vertex that satisfies the  
72 following criteria:

- 73     • The vertex must contain trajectories of reconstructed particle tracks with positive  $\chi^2$  values.
- 74     • There are at least 5 degrees of freedom in the vertex fit.
- 75     • The distance,  $|z|$ , along the beam line from the nominal center of the detector is less than  
76       24 cm.
- 77     • The transverse displacement,  $r$ , from the beam line is less than 2 cm.

### 78 3.2 Leptons

79 Events with electron or muon candidates identified with the following criteria are vetoed from the  
80 analysis. For the selection of electrons, the EGamma POG “Fall17-noIso-v2” MVA with a “loose”  
81 working point is used for Run 2. The EGamma “Winter22V1” MVA noIso score is used for Run  
82 3, with a working point that gives similar signal and background efficiencies as Run 2. We further  
83 require that the electrons have  $p_T > 15\text{GeV}$  and  $|\eta| < 2.5$ , excluding those in the barrel-endcap  
84 transition region. The “loose” muon ID recommended by the Muon POG is used to select muon  
85 candidates with  $p_T > 15\text{GeV}$  and  $|\eta| < 2.5$ . To isolate lepton candidates from hadronic activity,  
86 we require the “mini-isolation” of electrons to be less than 0.4 and “miniIsoId” for muons to be  
87 “loose”.

### 88 3.3 Jets

89 Particle Flow jets clustered with the anti- $k_T$  algorithm with distance parameter  $R = 0.4$ . To correct  
90 for pileup, charged hadron subtraction procedure is used for Run 2, where as PUPPI jets are used  
91 for Run 3. Latest jet energy corrections as recommended by the JetMET POG is reapplied to  
92 recalibrate from the NanoAoD values. The jets are further required to have  $p_T > 35\text{GeV}$  and  
93  $|\eta| < 2.4$ . Additionally, selected jets are required to pass “tight” working point of Jet ID as  
94 recommended by the JetMET POG.

#### 95 3.3.1 B-tagging

96 B jets are identified with the DeepJet algorithm with the “M” working point as recommended by  
97 the BTV POG, which corresponds to a b jet efficiency of about 80%. B tag SF corrections as  
98 recommended by the BTV POG is applied.

#### 99 3.3.2 W and Boosted Top tagging

100 For boosted top quarks with  $p_T > 400\text{GeV}$ , or W bosons with  $p_T > 200\text{GeV}$ , the decay products  
101 are expected to be contained within a  $\Delta R$  of 0.8. We apply Particlenet algorithm to jets clustered  
102 with anti- $k_T$  algorithm with a distance parameter of 0.8 to identify these W and boosted top  
103 quarks. Working point corresponding to a 1% mistag rate as recommended by the JetMET POG  
104 is used.

**105 3.3.3 Resolved Top tagging**

**106** For moderately boosted top quarks, the decay products can be resolved into separate AK4 jets. A  
**107** dedicated XGBoost BDT trigger is developed for tagging these resolved tops, following the same  
**108** methodology as the Run2 all-hadronic TTTT analysis [?]. From the AK4 jets collection, up to four  
**109** jets with the highest Deepjet b-tag scores are selected. For each b jet candidate, we identify all the  
**110** unique two-jet combinations in the AK4 jets collection (excluding the b jet candidate) as W subjet  
**111** candidates, with the condition that the combined mass of the two W subjet candidate is within 40  
**112** GeV of the true W boson mass, and that the mass of the combined three-jet candidate is within  
**113** 80 GeV of the true top quark mass. These selected three-jet combinations are the candidates for  
**114** resolved tops.

**115** The following variables form the inputs to the BDT:

- 116** • Mass of the b jet candidate
- 117** • Pairwise invariant mass of the b jet candidate with the W subjet candidate
- 118** • Mass of the resolved top candidate, from combining the four-vectors of the constituent jets
- 119** • Combined mass of the two W subjet candidates, from combining the four-vectors of the two
- 120** W subjet candidates
- 121** • The product of the top candidate  $p_T$  and  $\Delta R$  between the b jet and the W candidate.
- 122** • The product of the W candidate  $p_T$  and  $\Delta R$  between the W subjet candidates.
- 123** • The “soft-drop condition” from the soft-drop declustering algorithm reinterpreted as a vari-
- 124** able over the two W subjets:  $\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} \Delta R_{j_1, j_2}^{-2}$ , which tends to reject relatively soft collinear
- 125** jets.
- 126** • The DeepJet b-tag scores of all three constituent jets.
- 127** • The DeepJet c-tag scores of the W subjets.
- 128** • The quark-gluon likelihood scores of the W subjets. (Run2 Only)
- 129** • The DeepJet g vs uds discriminator scores of the W subjets. (Run3 Only)
- 130** • The jet constituent multiplicities of the W subjets.

**131** For training, genuine hadronic top candidates are taken from single-lepton  $t\bar{t}$  simulation samples,  
**132** and fake hadronic top candidates from di-lepton  $t\bar{t}$  simulation samples, with the requirement of  
**133**  $MET > 100$ ,  $N_{jets} \geq 5$  and  $N_{bjets} \geq 1$ . 2018 samples are used for Run2 model and 2022  
**134** samples for Run3. The BDT is trained with the XGBoost library on 500000 candidates for genuine  
**135** and fake samples each. Before training, to decorrelate from  $p_T$  of the top candidate, the candidates  
**136** are reweighted according to the  $p_T$  of the top candidate in bins of 25 GeV width from 0 GeV to 900  
**137** GeV, 50 GeV width from 900 GeV to 1000 GeV, and 100 GeV width from 1000 GeV to 5000 GeV,  
**138** such that the sum of weights of candidates in each bin is equal. If the resolved top candidates  
**139** overlap with one another within a  $\Delta R$  of 0.4, the lower scoring candidate is rejected.

**140** The working point for this resolved top tagger is selected at 2% False Positive Rate and given in  
**141** Table 5.

**142** Scales factors for tagging efficiencies and misidentification rates are derived for this resolved top  
**143** tagger. These scale factors are parameterized in the resolved top candidate  $p_T$ , and the uncertain-

**144** ties thereof are propagated as part of the systematic uncertainties. For scale factor derivations,  
**145** b-tag SF corrections, pileup weight and latest JERC corrections is applied to MC samples in the  
**146** following procedure.

| Year       | topWP |
|------------|-------|
| 2016preVFP | 0.961 |
| 2016       | 0.954 |
| 2017       | 0.961 |
| 2018       | 0.961 |
| 2022       | 0.966 |
| 2022EE     | 0.968 |
| 2023       | 0.963 |
| 2023BPix   | 0.969 |

Table 5: Resolved TopWP values for different years

147 For the derivation of the misidentification rates, we use a 0-lepton,  $N_b = 1$  region, selected with  
 148 pure  $H_T$  triggers and requiring  $H_T > 1200\text{GeV}$ . This region was selected to be enriched in  
 149 QCD multi-jet events and is orthogonal to the signal region via the  $N_b = 1$  requirement. In this  
 150 region, simulation is normalized to data. To correct the contamination of genuine tops in the  
 151 data, we estimate the number of genuine tops in data by the number of candidates matched to top  
 152 quarks at the generator level in simulation, and then subtract it from the data. In simulation, the  
 153 misidentification rates are measured to be the fraction of fake resolved top candidates that pass the  
 154 working point. In data, the misidentification rates are measured to be the fraction of candidates  
 155 (after subtraction of the estimated genuine tops) that pass the working point. In the end, the scale  
 156 factor is calculated as the ratio of the misidentification rates in data to those in simulation.

157 For the derivation of the tagging efficiency scale factor, we used single muon triggers to define  
 158 a control region selected to be enriched in semi-leptonic  $t\bar{t}$  events with similar kinematics to the  
 159 Signal Region. This region is orthogonal to the Signal Region via the requirement on the muon.  
 160 We further require that the events to be  $N_b \geq 2$ ,  $N_j \geq 4$ ,  $p_T^{\text{miss}} > 75\text{ GeV}$ , and muon  $p_T > 50\text{ GeV}$ .  
 161 In each event, jets are first cleaned against leptons, and resolved top candidates cleaned against  
 162 muons with a  $\Delta R$  of 0.4. We first normalize simulation to data. The misidentification scale factor is  
 163 applied to candidates in simulation that pass the working point but fail to be matched to top quarks  
 164 at generator level. Candidates are then split into three  $p_T$  categories: low ( $100 - 300\text{GeV}$ ), medium  
 165 ( $300 - 500\text{GeV}$ ) and high ( $\geq 500\text{GeV}$ ). We then perform a simultaneous template fit, where two  
 166 templates of candidates in simulation matched and unmatched to top quarks at generator level  
 167 respectively are used. The templates are binned in top candidate mass, since it is independent of  
 168 candidate  $p_T$ . The fit from simulation to data is performed simultaneously for candidates passing  
 169 and failing the working point. In each  $p_T$  category, the tagging efficiency scale factor is calculated  
 170 as  $\frac{\epsilon_{\text{post-fit}}}{\epsilon_{\text{pre-fit}}}$ , where tagging efficiency  $\epsilon$  is defined as the ratio of number of matched (genuine) top  
 171 candidates passing the working point to the total number of matched top candidates, and  $\epsilon_{\text{post-fit}}$   
 172 is the tagging efficiency in data estimated from fitting simulation to data.

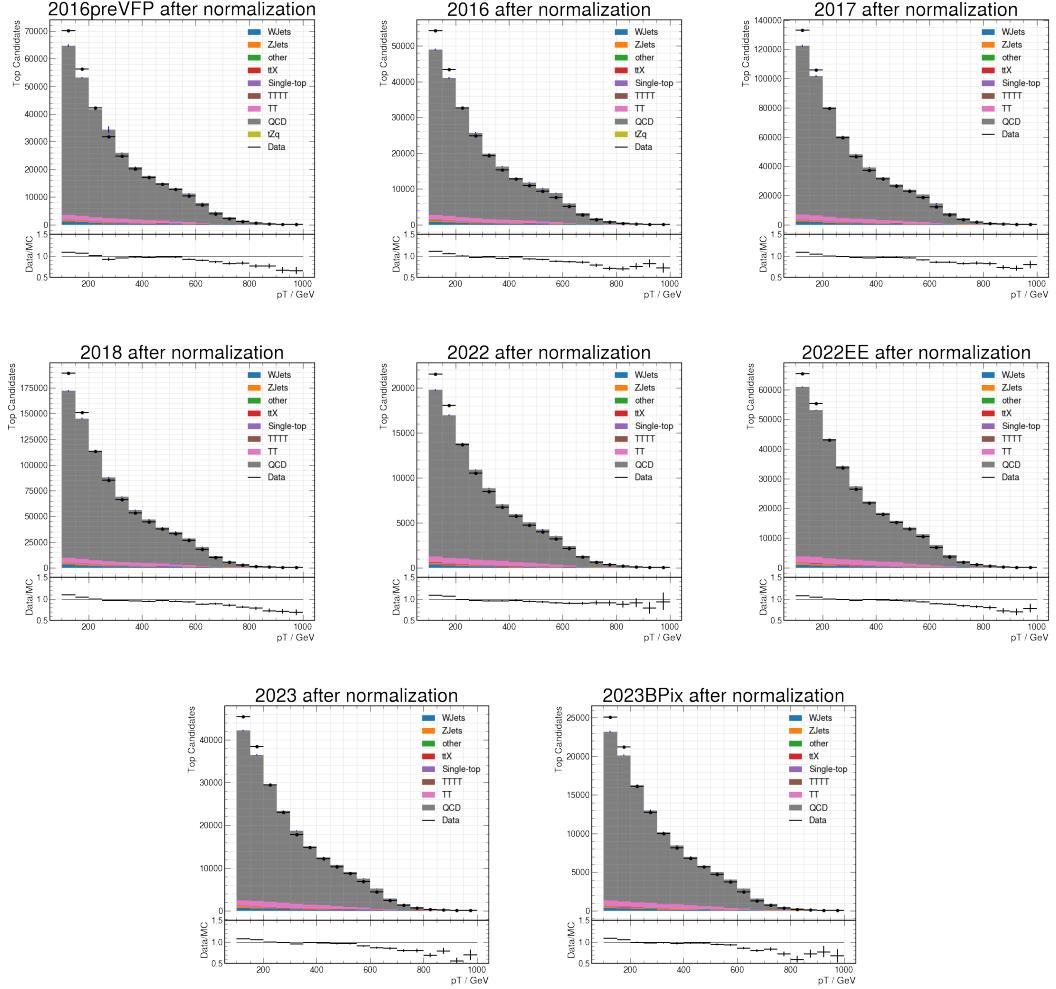


Figure 7: Distributions of resolved top candidate  $p_T$  in data and simulation for all candidates in the region used to derive the misidentification rate scale factor. Contributions from different processes as estimated from simulation are shown in the stacked histograms. The event yield in simulation is scaled to match data inclusively in this region, prior to the application of the top tagger working point.

### 3 Object Selection

#### 3.3 Jets

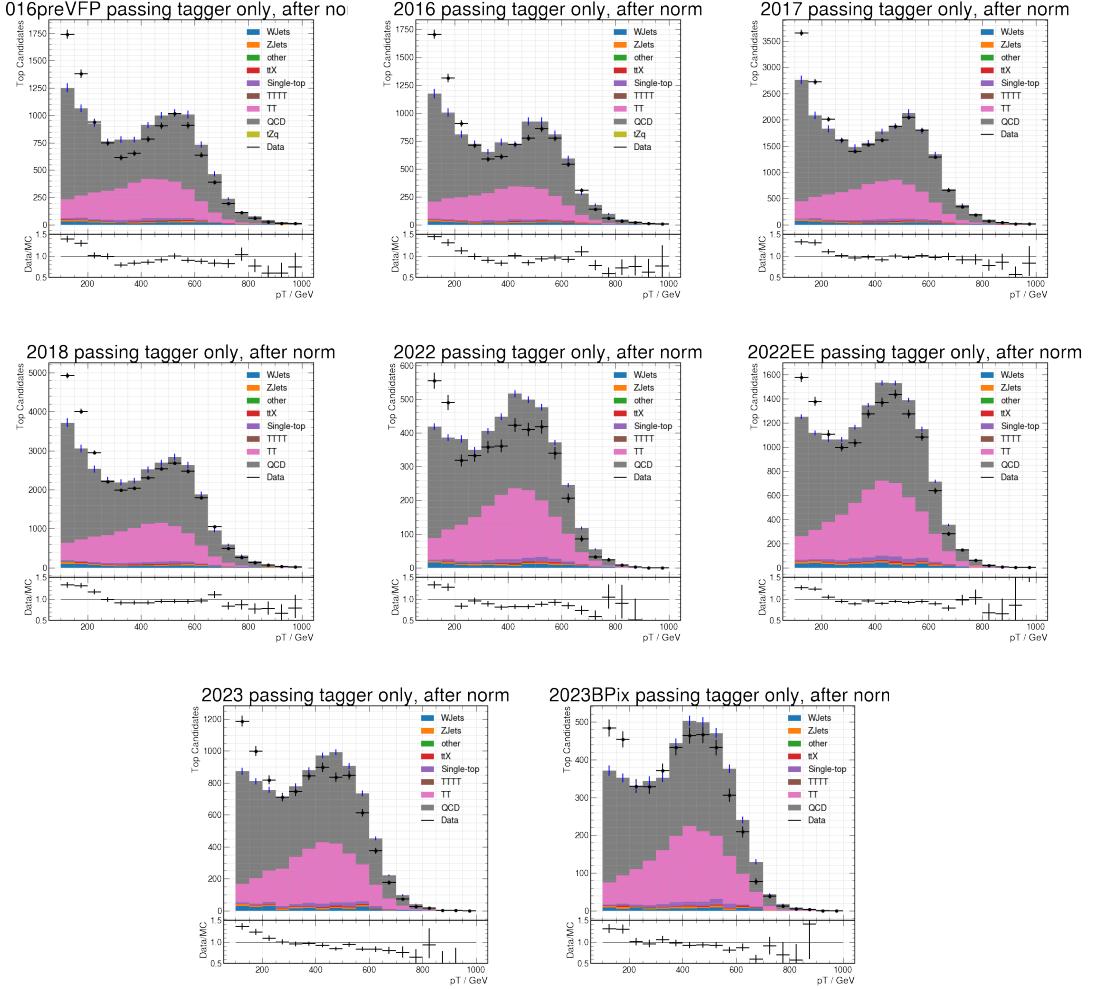


Figure 8: Distributions of resolved top candidate  $p_T$  in data and simulation for candidates passing the working point in the region used to derive the misidentification rate scale factor. Contributions from different processes as estimated from simulation are shown in the stacked histograms. The event yield in simulation is normalized to data by the scale obtained from the same distributions for all candidates in Figure 7

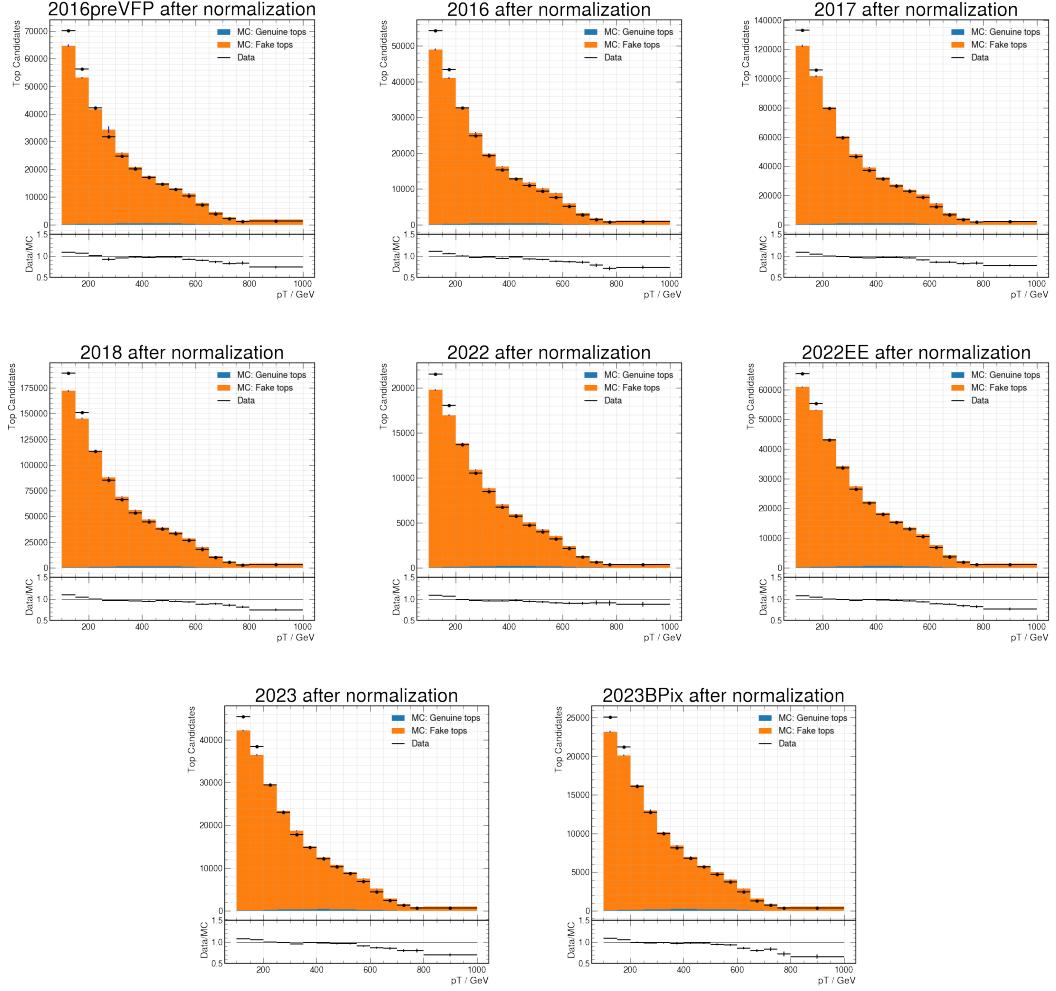


Figure 9: Distributions of resolved top candidate  $p_T$  in data and simulation for all candidates in the region used to derive the misidentification rate scale factor. Contributions from candidates matched and not matched to top quarks at the generator level in simulation are shown in the stacked histograms. The event yield in simulation is scaled to match data inclusively in this region, prior to the application of the top tagger working point.

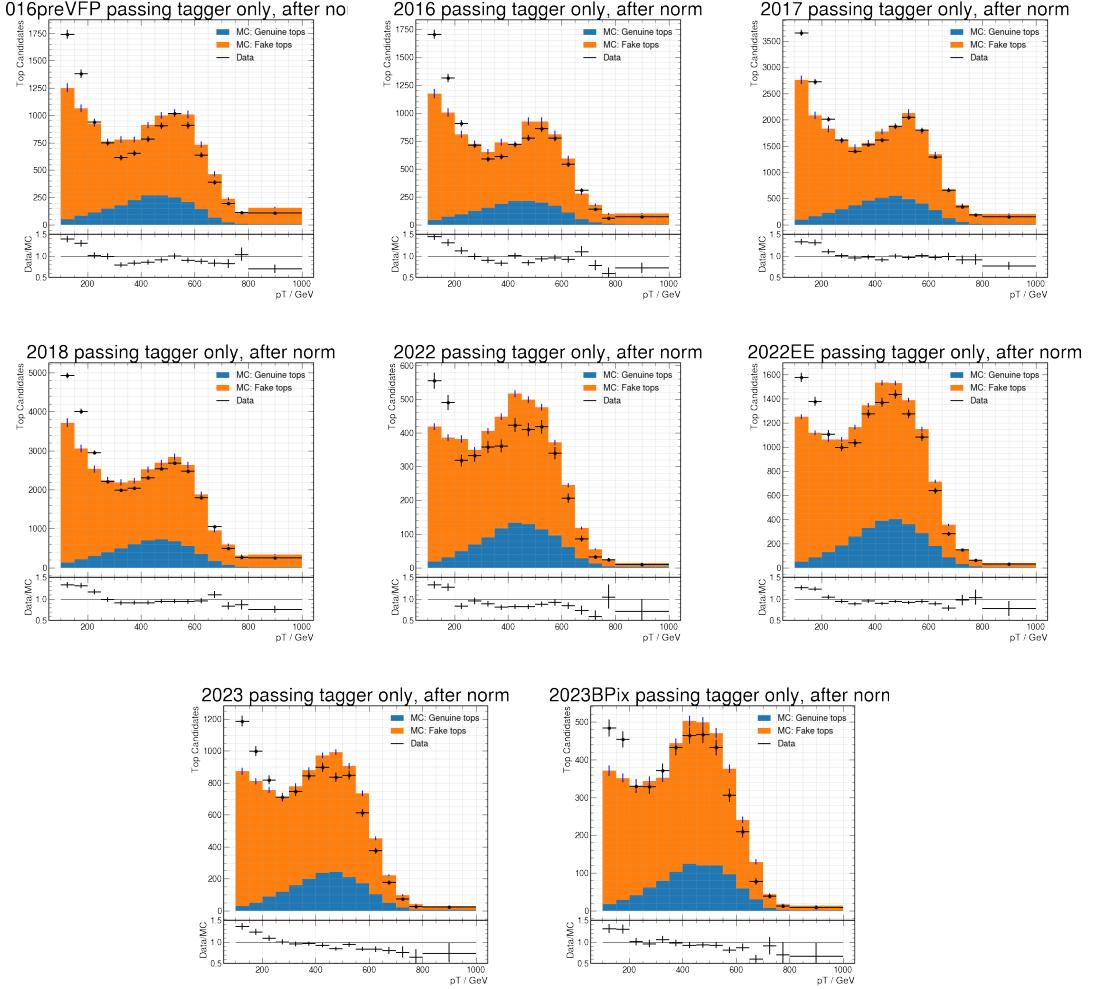


Figure 10: Distributions of resolved top candidate  $p_T$  in data and simulation for candidates passing the working point in the region used to derive the misidentification rate scale factor. Contributions from candidates matched and not matched to top quarks at the generator level in simulation are shown in the stacked histograms. The event yield in simulation is normalized to data by the scale obtained from the same distributions for all candidates in Figure 7

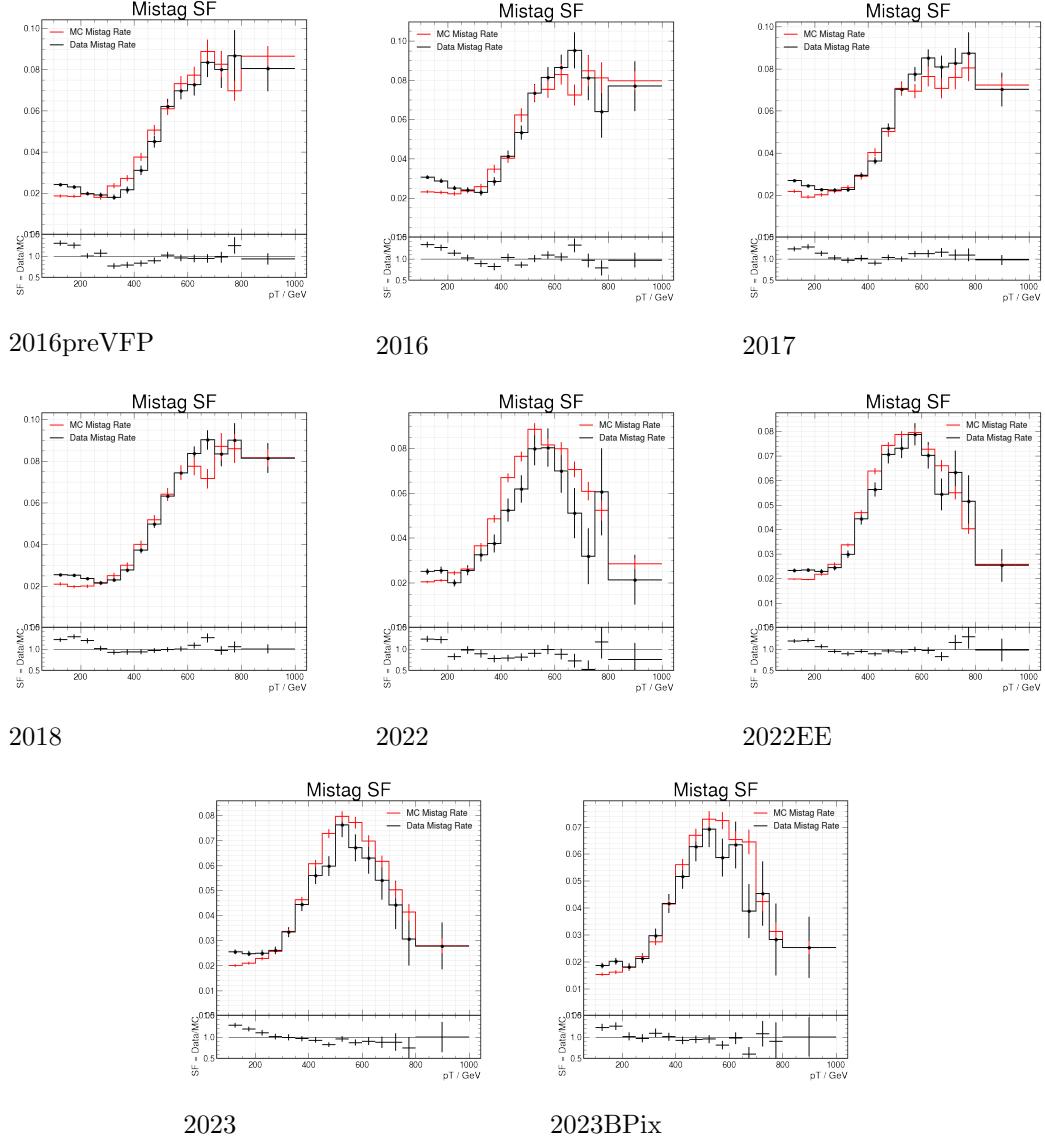


Figure 11: Misidentification rates measured in data and simulation after subtracting the estimated contribution from genuine tops. Misidentification rate scale factors are defined as the ratio of misidentification rate of data over simulation.

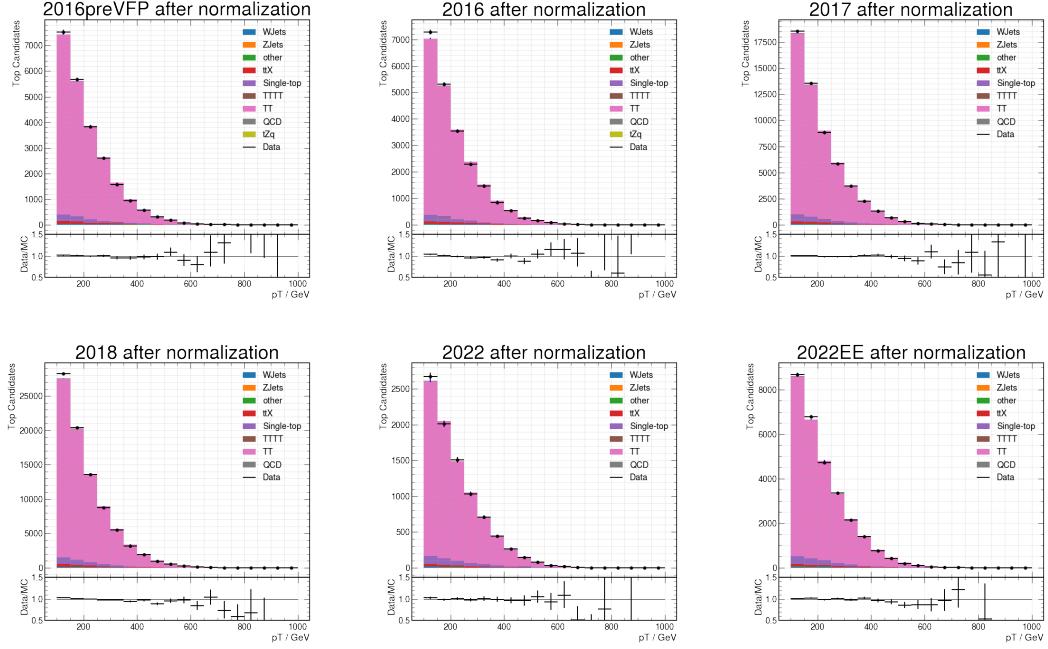


Figure 12: Distributions of resolved top candidate  $p_T$  in data and simulation for all candidates in the region used to derive the tagging efficiency scale factor. Contributions from different processes as estimated from simulation are shown in the stacked histograms. The event yield in simulation is scaled to match data inclusively in this region, prior to the application of the top tagger working point.

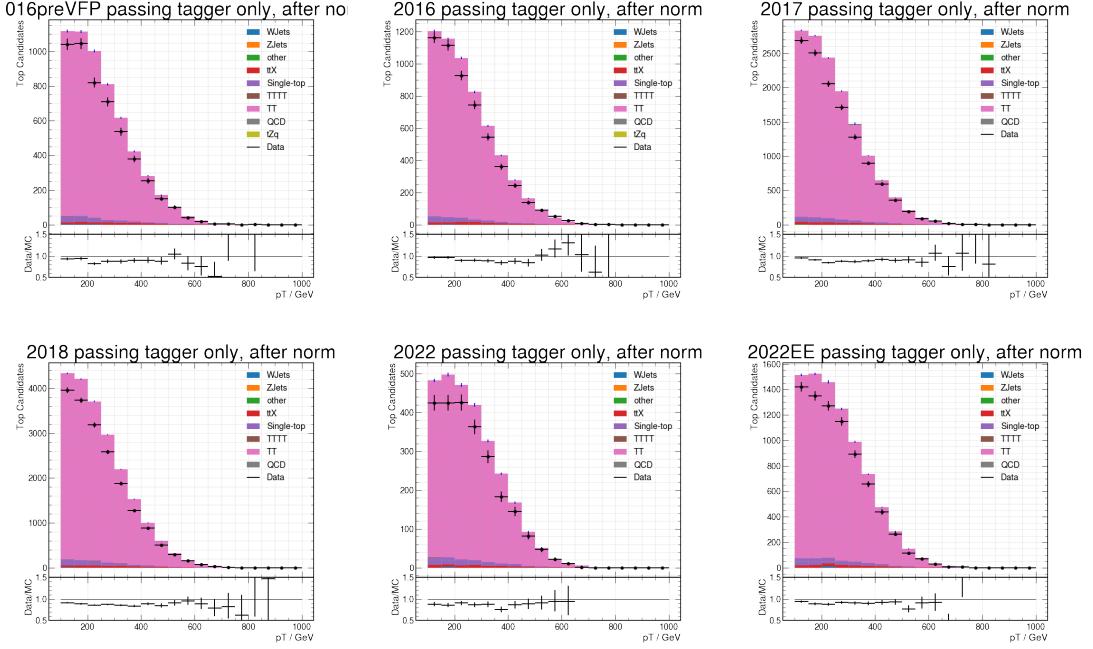


Figure 13: Distributions of resolved top candidate  $p_T$  in data and simulation for candidates passing the working point in the region used to derive the tagging efficiency scale factor. Contributions from different processes as estimated from simulation are shown in the stacked histograms. The event yield in simulation is normalized to data by the scale obtained from the same distributions for all candidates in Figure 12

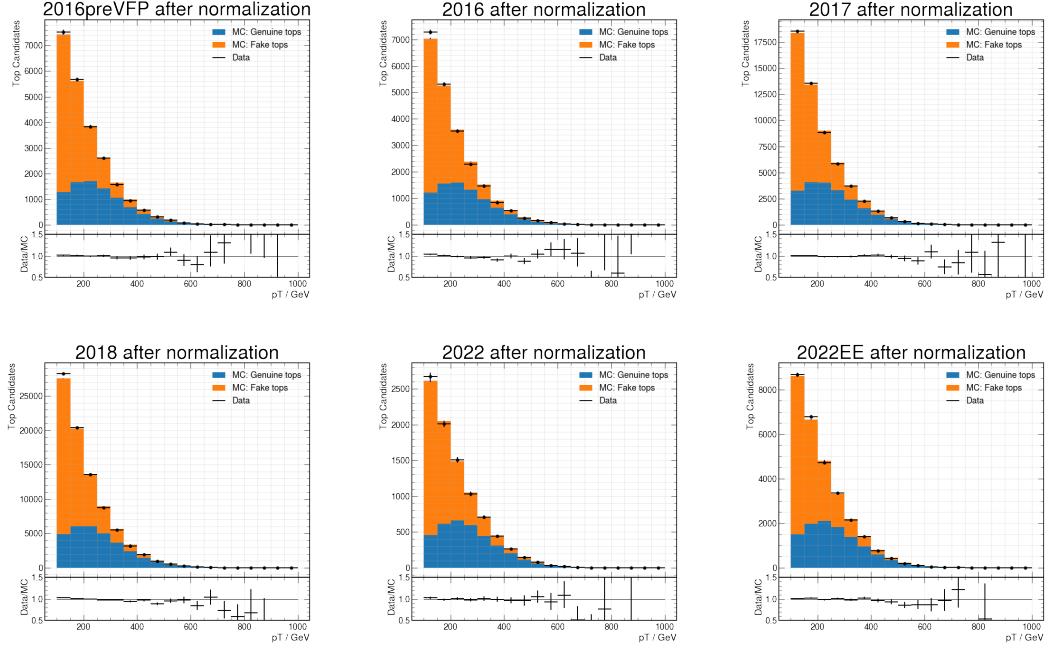


Figure 14: Distributions of resolved top candidate  $p_T$  in data and simulation for all candidates in the region used to derive the tagging efficiency scale factor. Contributions from candidates matched and not matched to top quarks at the generator level in simulation are shown in the stacked histograms. The event yield in simulation is scaled to match data inclusively in this region, prior to the application of the top tagger working point.

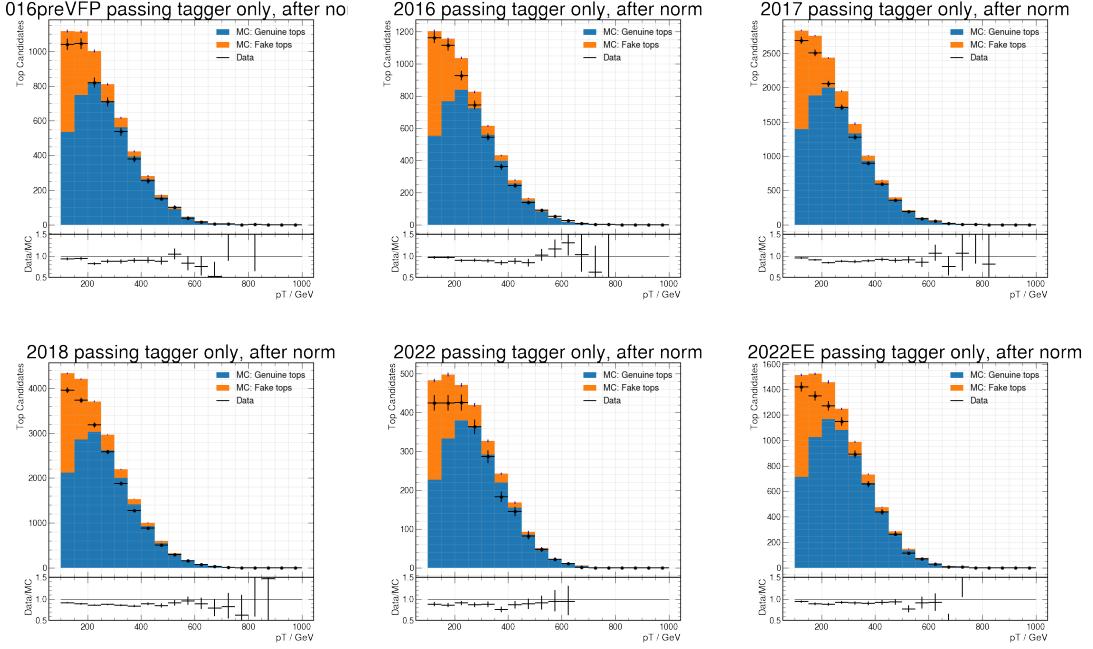
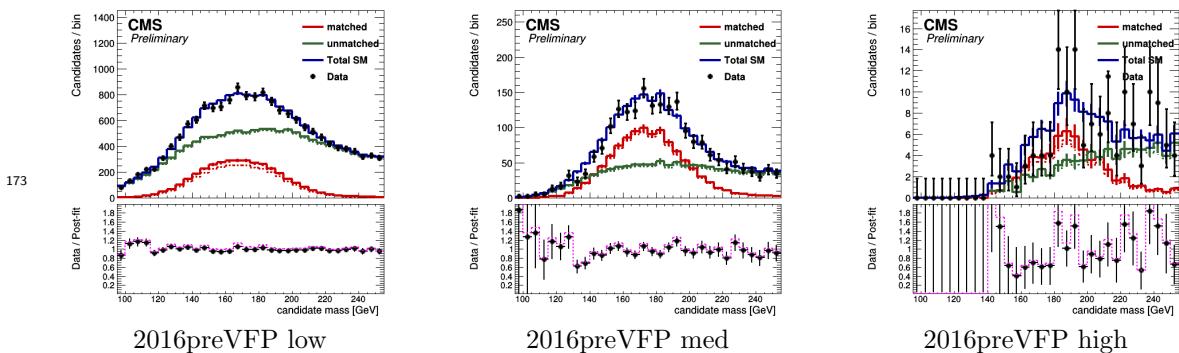
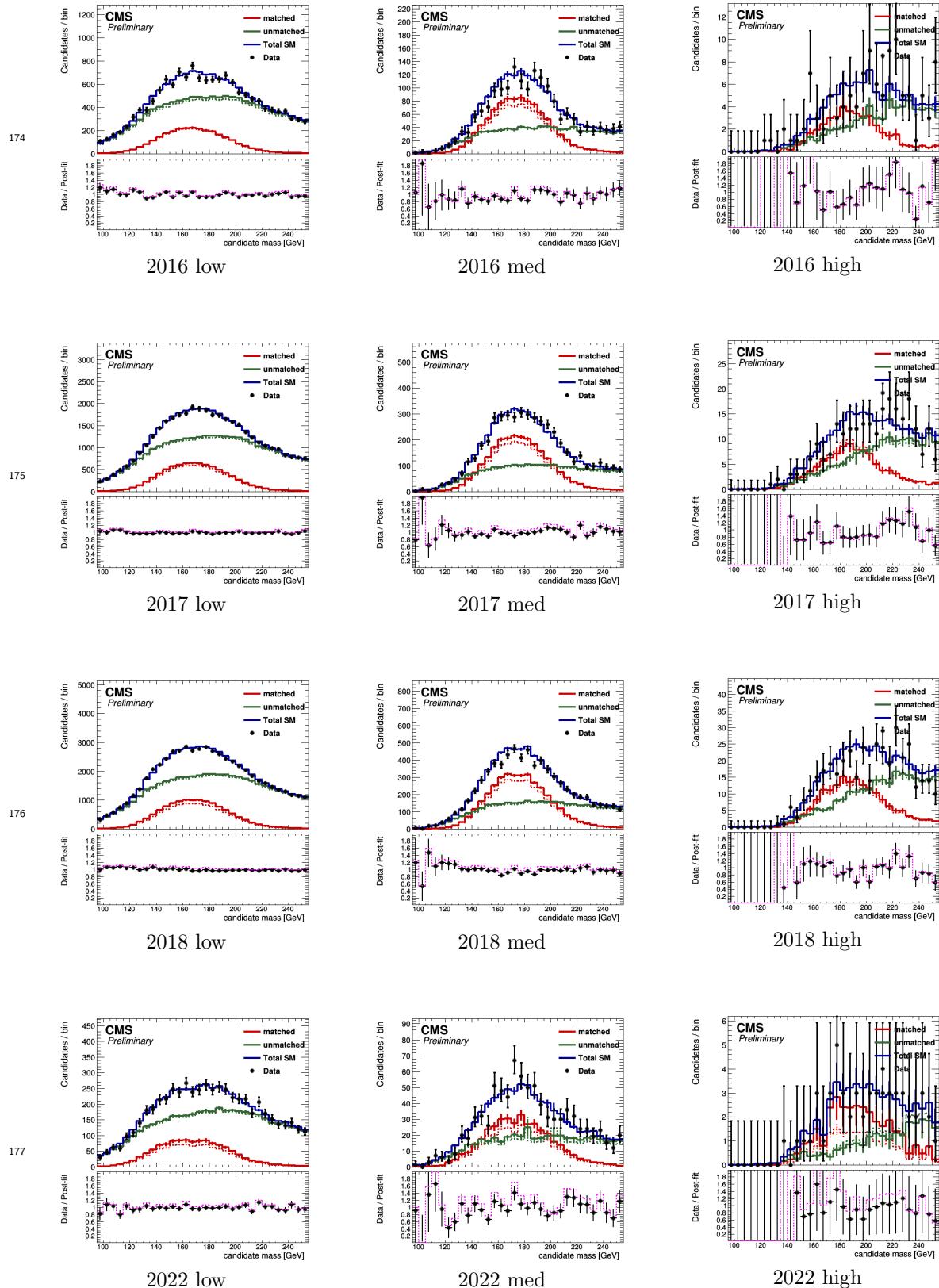


Figure 15: Distributions of resolved top candidate  $p_T$  in data and simulation for candidates passing the working point in the region used to derive the tagging efficiency scale factor. Contributions from candidates matched and not matched to top quarks at the generator level in simulation are shown in the stacked histograms. The event yield in simulation is normalized to data by the scale obtained from the same distributions for all candidates in Figure 12



### 3 Object Selection

#### 3.3 Jets



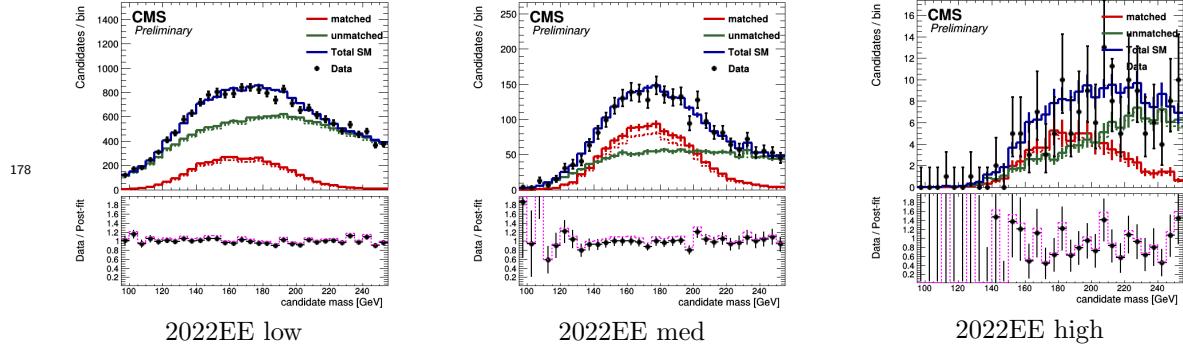
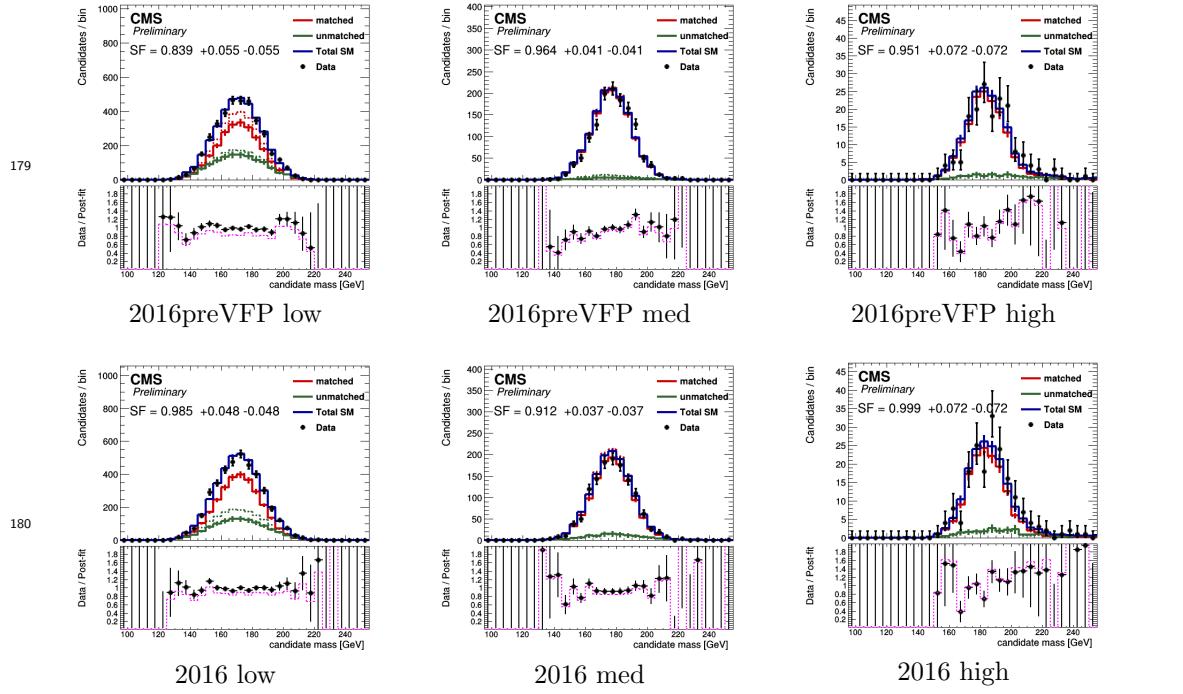


Figure 16: Data vs pre- and post-fit distributions in candidate mass for top candidates failing the working point in years for low (left column), medium (middle column), and high (right column)  $p_T$  categories. Solid lines correspond to post-fit distributions and dashed lines to pre-fit distributions. Candidates matched to generator-level tops are shown in red. Candidates unmatched to generator-level tops and total candidates are shown in green and blue respectively. The ratio of data to the total pre- and post-fit simulation is shown in the ratio panels.



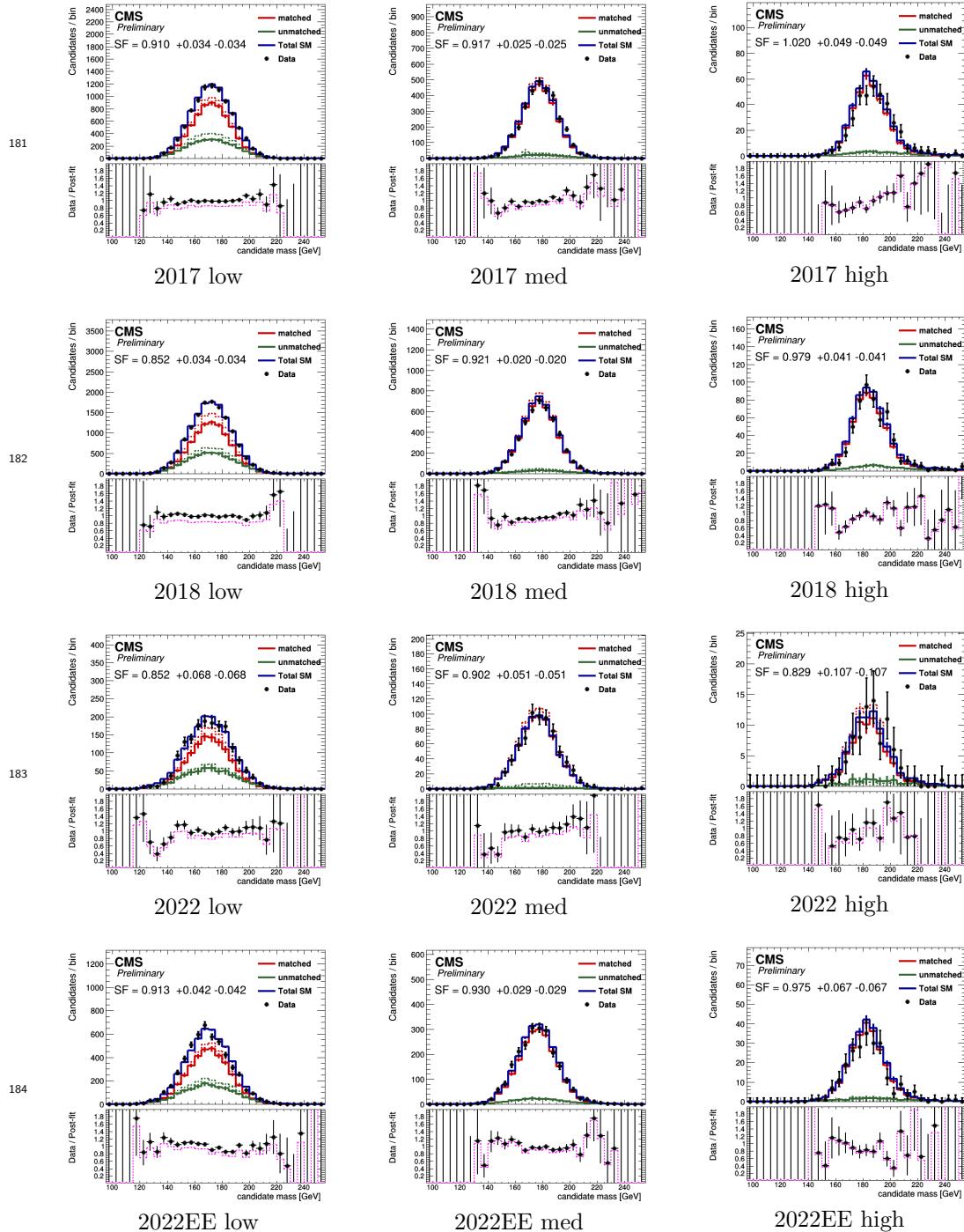


Figure 17: Data vs pre- and post-fit distributions in candidate mass for top candidates passing the working point in years for low (left column), medium (middle column), and high (right column)  $p_T$  categories. Solid lines correspond to post-fit distributions and dashed lines to pre-fit distributions. Candidates matched to generator-level tops are shown in red. Candidates unmatched to generator-level tops and total candidates are shown in green and blue respectively. The ratio of data to the total pre- and post-fit simulation is shown in the ratio panels. The scale factor extracted from the fit as shown by this figure and figure 16 is shown.

## 4 Event Selection

### 4.1 Baseline Selection and Signal Region

We apply to the events passing the triggers described in section 2 the following baseline selection:  $N_{jet} \geq 9$ ,  $N_{bjet} \geq 3$ ,  $H_T > 700$  GeV, and no leptons. The objects are defined according to the criteria in section 3

The signal region (SR) is defined with the additional requirement of at least one tagged resolved top. Events in the SR are then subdivided into 12 categories based on the number of tagged resolved tops ( $N_{RT}$ ), the number of tagged boosted tops ( $N_{BT}$ ), and  $H_T$ . Table 6 defines these categories.

| Top tags                    | $H_T$ [GeV] |         |          |           |           |           |       |
|-----------------------------|-------------|---------|----------|-----------|-----------|-----------|-------|
| $N_{RT} = 1, N_{BT} = 0$    | 700–800     | 800–900 | 900–1000 | 1000–1100 | 1100–1200 | 1200–1300 | >1500 |
| $N_{RT} = 1, N_{BT} \geq 1$ | 700–1400    |         |          |           | 1200–1300 | >1400     |       |
| $N_{RT} \geq 2$             | 700–1100    |         |          |           |           | >1100     |       |

Table 6: Definitions of the SR categories based on the number of resolved tops ( $N_{RT}$ ), number of boosted tops ( $N_{BT}$ ), and  $H_T$ .

### 4.2 Event-level BDT

To further discriminate between signal and background, we implement an event-level BDT with the "CatBoost" library. This library is selected because it is found to perform better as compared to XGBoost BDT and neural networks. We use simulated  $t\bar{t}t\bar{t}$  events as signal and a mixture of simulated  $t\bar{t}$  and QCD multijet events as background. Training against of mixture of background processes as such are found to perform better than training against  $t\bar{t}$  alone. The BDT is trained and applied to each year separately.

Following the pre legacy iteration of this analysis, we choose the kinematics of jets, b-jets and the associated variables thereof as inputs for the even-level BDT. The number of boosted W candidates, the  $p_T$  of the leading resolved top candidate and b-jet are also used. Event shape variables that reflect event topology information are also included as inputs to the BDT. On the other hand, the number of resolved tops and boosted tops are deliberately excluded from the inputs to the BDT so that that can be used as independent variables for binning in SR. Resolved and boosted top discriminants are also excluded to prevent the potential dependence on the shape of these discriminants. The following is the optimized set of variables input to the event-level BDT:

- The number of jets present in the event,  $N_j$
- The number of b-tagged jets present in the event,  $N_b$
- The number of boosted W candidates
- The sum of the masses of  $R = 0.8$  jets
- The missing transverse energy,  $P_T^{\text{miss}}$

## 5 Background Estimation

---

- The scalar sum of  $p_T$  of jets,  $H_T$
- The scalar sum of  $p_T$  of b-tagged jets
- The  $P_T^{\text{miss}}$  divided by square root of  $H_T$
- The  $p_T$  of the leading b jet
- The  $p_T$  of the leading resolved top candidate
- The  $\eta$  difference between the leading and sub-leading jets
- The  $\eta$  difference between the leading and sub-leading b-tagged jets
- The absolute  $\phi$  difference between the leading and sub-leading jets
- The absolute  $\phi$  difference between the leading and sub-leading b-tagged jets
- The mean of the DeepJet b-tag scores of the b jets in the event
- The  $H_T$  of the six highest- $p_T$  jets divided by the total  $H_T$  in the event
- The transverse momenta of the jet with the seventh-largest  $p_T$  in the event
- Hadronic centrality ( $C$ ), defined as the value of  $H_T$  divided by the sum of the energies of all jets in the event
- Event sphericity ( $S$ ), calculated from all of the jets in the event in terms of the tensor

$$S^{\alpha\beta} = \sum_i p_i^\alpha p_i^\beta / \sum_i |\vec{p}_i|^2,$$

where  $\alpha$  and  $\beta$  refer to the three-components of the momentum of the  $i$ -th jet. The sphericity is then

$$S = (3/2)(\lambda_2 + \lambda_3),$$

where  $\lambda_2$  and  $\lambda_3$  are the two smallest eigenvalues of  $S^{\alpha\beta}$ .

- Event aplanarity ( $A$ ), defined as

$$A = (3/2)(\lambda_3)$$

## 5 Background Estimation

QCD multijet and hadronic  $t\bar{t}$  processes are the dominant background of this analysis. Next-to-leading order (NLO) calculations in strong interactions and limited statistics lead to large uncertainties in the simulation of the very high jet and b-jet multiplicity regions of these processes used in this analysis. Therefore, we cannot rely on the simulation-based approach to predict the QCD multijet and  $t\bar{t}$  + jets background. To overcome this, we adopt two data-driven methods: the “extended ABCD” method for data-driven estimation of the absolute rate of the background, and the “ABCDnn” method for data-driven estimation of the shape of the background.

### 5.1 Extended ABCD method

We divide the phase space into two dimensions using  $N_j$ , the number of jets, and  $N_b$ , the number of b-jets. In the vanilla ABCD method, the phase space is divided into four regions along the two dimensions, one of which is the Signal Region (SR) D and the other three are Control Regions (CR) A, B and C. The yield in the Signal region D is estimated as  $\hat{F}_D = \frac{F_B}{F_A} \times F_C$ , where  $\hat{F}_D$  is the estimated yield in Signal Region D and  $F_A$ ,  $F_B$  and  $F_C$  are the observed yield in Control Regions A, B and C respectively. This method assumes joint distributions in  $N_j$  and  $N_b$  are mostly factorizable.

To improve the accuracy of the estimation, we extend the CRs to lower multiplicity in  $N_j$  as seen in fig 20. Then, yield in the SR can be estimated as  $\hat{F}_D = \left(\frac{F_B F_C}{F_A}\right)^2 \left(\frac{F_X}{F_B F_Y}\right)$ . Here,  $F$  is the estimated  $t\bar{t}$  + QCD multijet yield in each Control Region. This is estimated by

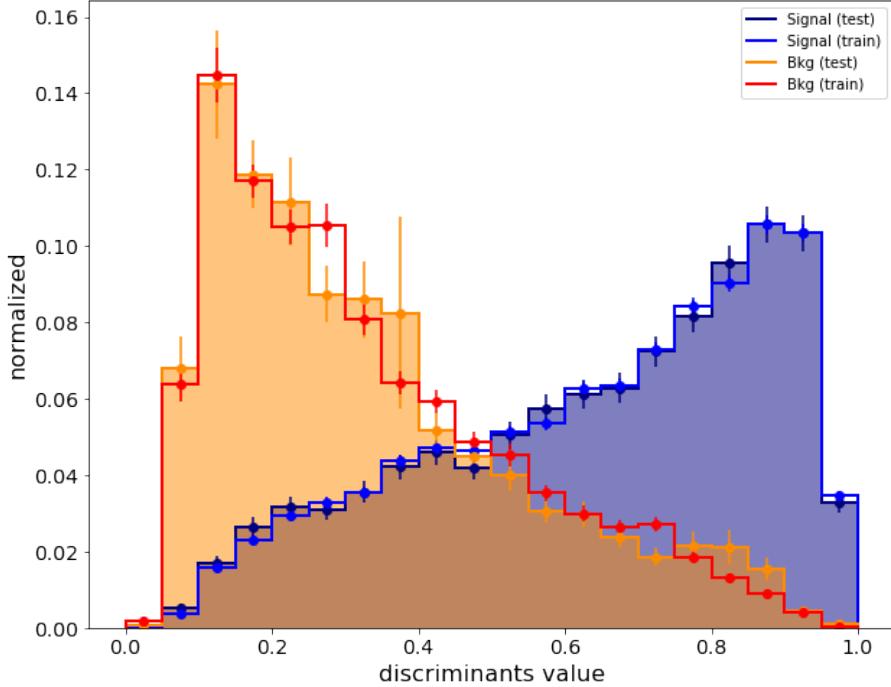


Figure 18: As an example, the discriminant distributions for the event-level BDT for signal and background in training and validation samples for 2018.

256 subtracting the  $t\bar{t}t\bar{t}$  MC yield and minor background processes MC yield from the observed yield  
 257 in data in each Control Region.

258  
 259 The extended ABCD method is applied to each  $h_T$  bin as specified in table 6 for each year  
 260 separately.

261  
 262 Closure test conducted on prelegacy 2016  $t\bar{t}$  events requiring 0 lepton and same regions defined on  
 263  $N_j$  and  $N_b$  shows that yield predicted by vanilla ABCD method if off by 18%. The disagreement  
 264 between the estimated and true yield improves to 7% for the extended ABCD method.

265  
 266 Appendix D includes yields for control regions and estimated yield for the signal region for each  
 267  $H_T$  bin and for each year.

## 268 5.2 ABCDnn method

269 We use the “ABCDnn” method to estimate the shape of the event-level BDT discriminant for the  
 270  $t\bar{t} + \text{QCD multijet}$  background in the Signal Region. Specifically, “ABCDnn” method utilizes a  
 271 Machine Learning technique named Neural Autoregressive Flow, which creates transformations  
 272 between distributions of the feature variables under different conditions. These transformations  
 273 are constructed as invertible bijective functions implemented as DNNs and learnt during training.  
 274 In addition, the method automatically take into account the complex correlations between feature  
 275 variables, so that multi-dimensional distributions can also be estimated.

276  
 277 Specifically, for this analysis, we use the same definition of Control Regions and Signal Regions  
 278 as the ones defined for extended ABCD. That is, the neural networks are trained on Control

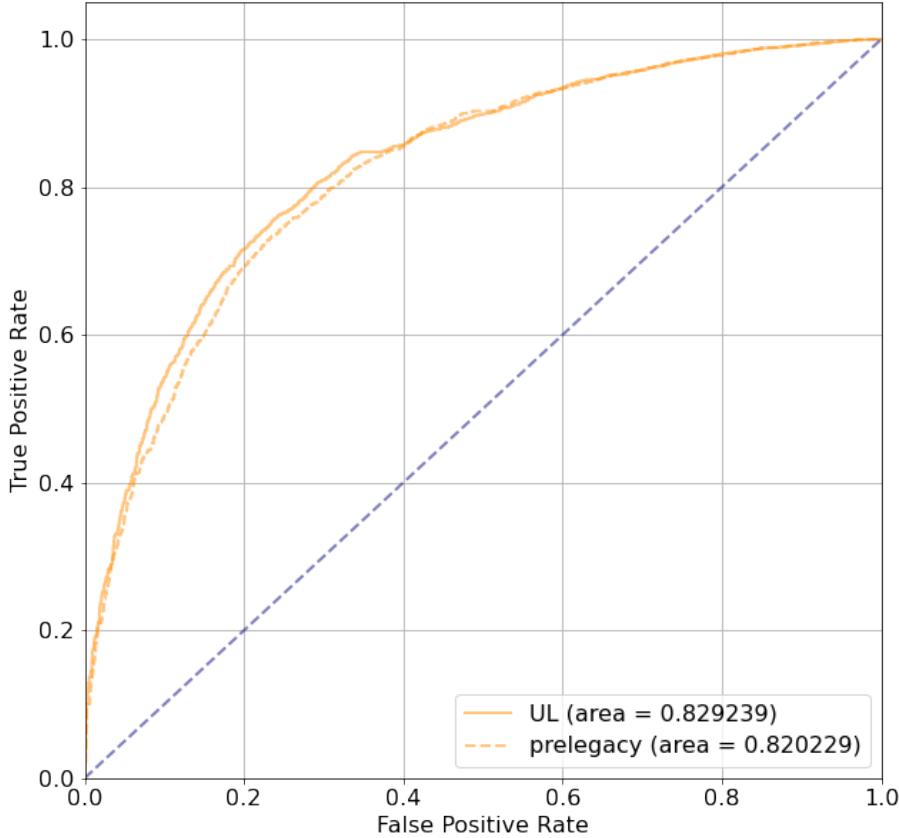


Figure 19: ROC curves of the event-level BDT for 2018, as compared against the prelegacy model.

Regions and applied to the Signal Region. For each Control Region, the neural network takes two distributions for training: the normalized distribution of  $t\bar{t}$  MC as the input distribution, and the normalized data-driven estimated  $t\bar{t} + \text{QCD multijet}$  distribution as the target distribution. The target distribution is created by taking the actual observed data distribution and subtract it by the  $t\bar{t}t\bar{t}$  MC distribution and other minor background MC distributions. The neural network learns the transformation from the input distribution to the target distribution in each Control Region, as well as the condition, i.e. the label of the Control Region in the phase space, so that it learns how the transformation should change across the phase space defined by  $N_j$  and  $N_b$ . It does this by minimizing the maximum-mean-discrepancy between the output predicted distribution and the target distribution. Finally, the neural network is applied to the input distribution of  $t\bar{t}$  MC in the Signal Region to obtain the desired shape of data-driven estimated  $t\bar{t} + \text{QCD multijet}$  distribution in the Signal Region.

The learned models in this analysis not only transforms from MC distribution to data-driven distribution to compensate for potential data-MC disagreement, but also transforms from a single process distribution ( $t\bar{t}$ ) to dual process ( $t\bar{t} + \text{QCD multijet distribution}$ ). We do not use the MC QCD multijet distribution as input due to limited statistics, especially in the Signal Region.

This training is done for each top-tag instead of each  $H_T$  bin because of the limited statistics in each  $H_T$  bin. Therefore, for this analysis the feature variables underlying the transformed distributions are event-level BDT discriminant and  $H_T$ . The predicted  $H_T$  of each event in the

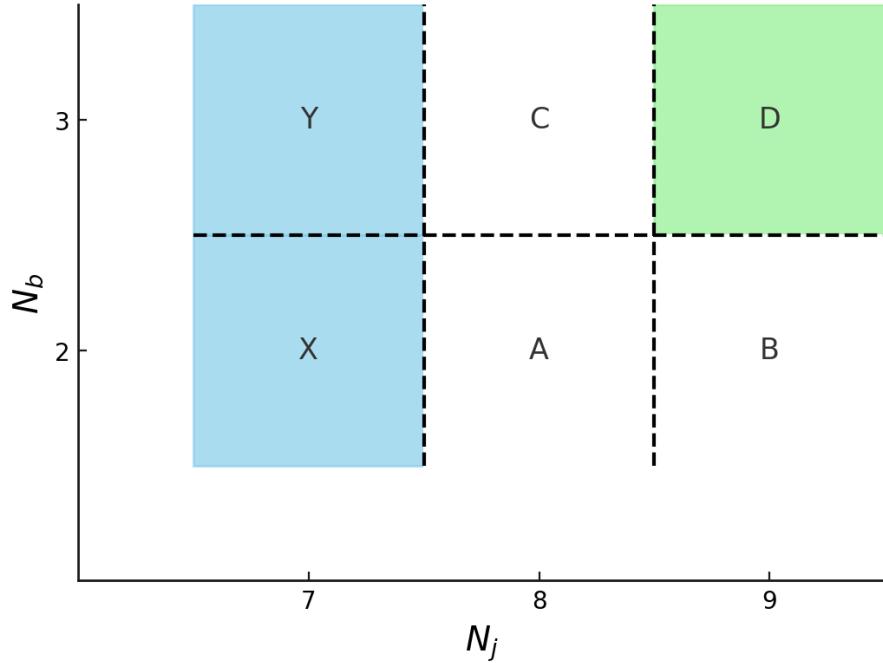


Figure 20: A, B and C are the Control Regions for the vanilla ABCD method and Y and X are the added adjacent Control Regions for the extended ABCD method. D is the Signal Region

300 input distribution is used to place them in the corresponding  $H_T$  bin, and the BDT discriminant  
 301 is used in each  $H_T$  bin to create the predicted shape. A separate “ABCDnn” model is trained  
 302 for each year.

303  
 304 A closure test for the method has been done by the previous iteration of this analysis morphing  
 305  $t\bar{t}$  MC distribution into  $t\bar{t}X$  MC distributions, with the same event selection and Control/Signal  
 306 Region definition as the main analysis. The method was found to predict the target shape well.  
 307

### 308 5.3 Validation of the background estimation methods

309 The two background estimation methods are validated in a validation region (VR) requiring lower  
 310  $N_j$  than the previously defined SR. Specifically, we use  $N_j = 8, N_b \geq 3$  as the VR, and define the  
 311 Control Regions for this VR correspondingly. We then perform the same procedures as defined  
 312 previously for “extendedABCD” and “ABCDnn”, and compare the predicted yield and shape  
 313 with the truth.

314  
 315 We find that the estimated yield in the Validation Region in each  $H_T$  bin agrees with the true  
 316 yield after accounting for statistical error from MC propagated through the extended ABCD  
 317 formula and statistical error in the true yield. The result is in Appendix E. We account for the  
 318 disagreement as a systematic uncertainty detailed in section 6.1.1.

319  
 320 We find that the predicted shape in the Validation Region largely agrees well with the true shape,  
 321 especially for  $H_T$  bins with higher statistics. The result is in Appendix G. We account for the  
 322 disagreement as a systematic uncertainty detailed in section 6.1.2.

## 323 6 Systematic uncertainties

### 324 6.1 Systematics from background estimation

325 We derive the systematic uncertainties for our background estimation methods from the Validation  
326 Region defined in section 5.3, where we compare the output (yield/shape) of the specific background  
327 estimation method against the truth. We then apply the derived systematics to the Signal Region.  
328 These uncertainties are calculated separately for each  $H_T$  bin defined in table 6.

#### 329 6.1.1 ExtendedABCD

330 We derive a normalization uncertainty from the Validation Region based on the disagreement  
331 between the data and predicted background in the VRs. This uncertainty is calculated in each VR  
332  $H_T$  bin as the sum in quadrature of two quantities: the deviation of the weighted mean (average)  
333 of events in that VR  $H_T$  bin from 1, and the weighted RMS of events in that VR  $H_T$  bin. The  
334 weighted mean  $\langle f \rangle$  is defined as

$$\langle f \rangle = \frac{\sum_i f_i w_i}{\sum_i w_i},$$

355 for each discriminant histogram bin  $i$  in the VR  $H_T$  bin of interest, where the weight  $w_i$  is the  
356 number of events in that histogram bin and  $f_i$  is the ratio of observed events to predicted events  
357 ( $N_{\text{data}}/N_{\text{pred}}$ ) in that histogram bin. The mean is weighted to reflect the distribution of events in  
358 the histogram, and the deviation of this weighted mean from 1, i.e.  $(1 - \langle f \rangle)$  reflects the overall  
359 offset in normalization between the prediction and the data. Similarly, the weighted RMS is defined  
360 as

$$\sqrt{\langle f^2 \rangle - \langle f \rangle^2},$$

341 where

$$\langle f^2 \rangle = \frac{\sum_i f_i^2 w_i}{\sum_i w_i},$$

342 and reflects the spread of the disagreement between the prediction and the data in the histogram  
343 bins. This method of quantifying the data-prediction discrepancy is based on recommendations  
344 from the CMS Statistics Committee. These uncertainties are taken to be uncorrelated across  $H_T$   
345 bins and years.

#### 346 6.1.2 ABCDnn

347 We derive an uncertainty for the shape of ABCDnn output from the Validation Region, based  
348 on observed disagreement between the shape of data and predicted background in the VRs. This  
349 uncertainty is calculated in each VR  $H_T$  bin as a linear shift, both up and down, of the BDT  
350 discriminant value of each event. For example, a shift up of 2% ( $s = +0.02$ ) for a particular VR  
351  $H_T$  bin means that the BDT discriminant value of every event in that VR  $H_T$  bin is multiplied  
352 by 1.02. Qualitatively, in both up and down directions, we shift the BDT discriminants such that  
353 the disagreement between the shape of the normalized BDT discriminant distribution of the data  
354 and the predicted background is as small as possible. Quantitatively, the extent of the shift is  
355 determined by minimizing the following metric:

$$s = \arg \min_s \sum_i \max \left( \frac{|f_i(s) - 1|}{\sigma_i(s)}, \frac{|f_i(0.0) - 1|}{\sigma_i(0.0)} \right)$$

356 where  $s$  is a particular shift up (positive) or down (negative),  $f_i = \frac{N_{\text{pred},i}}{N_{\text{truth},i}}$ , the ratio of predicted  
357 events over observed events in BDT discriminant bin  $i$ , and  $\sigma_i$  is the statistical uncertainty in  
358 bin  $i$ .  $\frac{|f_i(s)-1|}{\sigma_i(s)}$  measures the deviation of ratio in bin  $i$  from 1 under shift  $s$ , inversely scaled by

359 the statistical uncertainty in that bin. In most cases, the deviation increases for some bins and  
 360 decreases for the others under a particular shift, therefore a naive minimization of this term across  
 361 the bins results in bad coverage in the bins whose deviation decreases. These bins are the bins  
 362 that are supposed to be covered by the shift, since the bins whose deviation increases would be  
 363 covered by the shift in the opposite direction.

364 To overcome this, we include a maximization against the second term  $\frac{|f_i(0.0) - 1|}{\sigma_i(0.0)}$ , which is the same  
 365 scaled deviation of ratio before shifting. The maximization of the first term against the unshifted  
 366 scaled deviation of ratio removes the effect of bins whose deviation worsened under a particular  
 367 shift direction, so that by minimizing this metric, the shifting method can obtain good coverage  
 368 of disagreement in all bins.

370 Currently, we consider this shape uncertainty as uncorrelated across  $H_T$  bins and across  
 371 years.

### 373 6.1.3 Closure Test on systematic uncertainty estimation for background estima- 374 tion

375 We perform closure tests to make sure that the methods used to estimate the systematic  
 376 uncertainties of the two background estimation are reasonable. We use ultralegacy 2018 data,  
 377 using exactly the same requirement as our event selection and inputs to extended ABCD formula,  
 378 and using the same  $H_T$  bin definition as  $N_{RT}=1$ ,  $N_{BT}=0$  top-tag, but requiring 0 resolved top  
 379 instead of at least 1 so that the regions are orthogonal from the main analysis. In this region, we  
 380 estimate the systematic uncertainties of the two background estimation methods in the VR, and  
 381 apply them to the SR, and finally compare the predicted yield and shape with the truth in SR  
 382 while taking into account the estimated systematic uncertainties.

383 For “extendedABCD”, the result shows that the estimated yield in the Signal Region in each  $H_T$   
 384 bin agrees with the true yield after accounting for statistical error from MC propagated through  
 385 the extended ABCD formula and statistical error in the true yield, as well as the systematic  
 386 uncertainty as defined in sec 6.1.1. The result is in Appendix H.

388 For “ABCDnn”, the result shows that the systematic uncertainty as defined in sec 6.1.2 mostly  
 389 covers the disagreement between the predicted and the true shape in each  $H_T$  bin in the Closure  
 390 Test region. The result is in Appendix I.

## 393 6.2 Other sources of systematic uncertainties

- 394 • The statistical uncertainties of SM/top philippic  $t\bar{t}t\bar{t}$  signal samples and background processes  
 395 modeled with MC samples.
- 396 • The statistical uncertainties of the estimated QCD +  $t\bar{t}$  yields in each CR, propagated through  
 397 the extendedABCD formula.
- 398 • The statistical uncertainties of the simulated  $t\bar{t}$  events transformed by ABCDnn to form the  
 399 predicted discriminant distribution of BDT.
- 400 • The statistical uncertainties from the trigger efficiency corrections, parameterized in  $N_b$  and  
 401  $N_j$ .
- 402 • Uncertainties related to the boosted top and W tagging correction factors used to correct  
 403 the performance of the Particlenet algorithm in simulation to match data. Two uncertainties  
 404 are included: boosted top and boosted W uncertainties related to the non-mass-decorrelated  
 405 scale factors. These uncertainties are correlated between processes but uncorrelated between  
 406 years.

---

## 7 Results and interpretation

---

- Tagging and misidentification uncertainties related to resolved top tagging correction factors used to correct the performance of the resolved top tagging algorithm in simulation to match data, as described in Section 3.3.3. Two uncertainties for both tagging efficiency and mistag rate scale factors are considered, and uncertainties in these stem from statistical uncertainties that arise in calculating the scale factors. These are correlated between processes and uncorrelated between years.
- Uncertainties related to correction factors used to correct for the performance of the DeepJet b-tagging algorithm in simulation with respect to data. The DeepJet iterative fit/shape scale factors are used. Scale factors per event are calculated as the product of the scale factors for all jets in a given event passing pre-selection. The uncertainties are split into several sources, including those affecting HF (heavy-flavor) or LF (light-flavor) jets, uncertainties from charm jets (cferr1 and cferr2), and linear and quadratic statistical fluctuations (LFstats1, LFstats2, HFstats1, HFstats2). These are all considered as separate systematic variations, with statistical variations correlated between processes and uncorrelated between years (LFstats1, LFstats2, HFstats1, HFstats2) and others correlated between both processes and years (HF, LF, cferr1 and cferr2).
- Uncertainties associated to the pileup reweighting correction factor.
- Jet energy scale (JES) and resolution (JER) uncertainties impacting the reconstruction of jets, including  $R = 0.4$  and  $R = 0.8$  jets,  $p_T^{\text{miss}}$ , and tagged top, W, and b candidates and their correction factors. The JES systematic uncertainty is currently applied using the combined JES uncertainty. We vary the JES and JER up and down by their systematic uncertainties and propagate the effects through all analysis objects and selections, including in selecting jets and boosted objects and re-propagating the varied jets through our resolved top algorithm. JER and JES variations are correlated between processes but not between years.
- Uncertainties assigned to the integrated luminosity measured by the CMS experiment.
- Uncertainties in the theoretical cross sections used to normalize ttX backgrounds are considered. These impact the number of ttX events in each SR bin. For ttX processes ( $X=W,Z,H$ ), we assign an uncertainty of 26% based on the largest deviation of the signal strength observed by CMS for any of these processes (relative to the SM prediction).

## 7 Results and interpretation

A binned likelihood analysis is carried out in the SR categories split by resolved and boosted top multiplicities and  $H_T$ , as defined in Table 6. As input to each SR category, the shape and yields of the event-level BDT discriminant of  $t\bar{t}t\bar{t}$ ,  $t\bar{t}X$  and other minor background processes are provided to CMSCombine from Monte Carlo, and those of  $t\bar{t}$  and QCD processes are provided from data-driven background estimation methods extendedABCD (Section 6.1.1) and ABCDnn (Section 6.1.2).

The systematic uncertainties described in Section 6 are included in the analysis as nuisance parameters. Up and down variations of the BDT discriminant shapes are provided as inputs for each nuisance parameter corresponding to shape-based systematic variations, while uncertainties affecting only the normalizations of processes are implemented using log-normally distributed constraints on the simulation rates in each data-taking period.

Currently at the blinded stage, we have the following preliminary result:

|                   | <b>Expected significance</b> | <b>Median expected limit</b> |
|-------------------|------------------------------|------------------------------|
| 2016 + 2016preVFP | 0.753                        | 3.281                        |
| 2017              | 0.669                        | 4.813                        |
| 2018              | 0.493                        | 5.281                        |
| 2022 + 2022EE     | 0.753                        | 3.688                        |
| <b>Total</b>      | <b>1.441</b>                 | <b>1.500</b>                 |

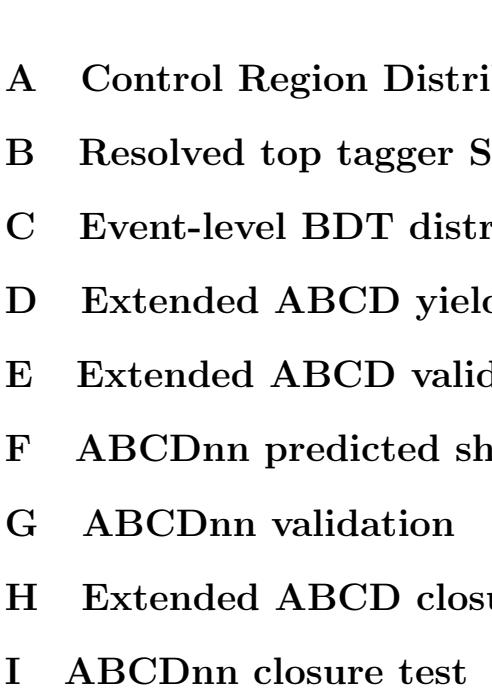
Table 7: Expected significance and median expected limits for SM  $T\bar{T}T\bar{T}$ .

| $M_{v_1} / \text{GeV}$ | $c_t$ | <b>Expected significance</b> | <b>Median expected limit</b> |
|------------------------|-------|------------------------------|------------------------------|
| 1400                   | 3.0   | 0.928                        | 2.289                        |
|                        | 2.5   | 0.700                        | 3.047                        |
|                        | 2.0   | 0.446                        | 4.813                        |
| 1200                   | 3.0   | 1.838                        | 1.129                        |
|                        | 2.5   | 1.330                        | 1.578                        |
|                        | 2.0   | 0.947                        | 2.242                        |
| 1000                   | 3.0   | 3.594                        | <b>0.561</b>                 |
|                        | 2.5   | 2.744                        | <b>0.746</b>                 |
|                        | 2.0   | 2.001                        | 1.039                        |
| 800                    | 2.5   | 5.444                        | <b>0.361</b>                 |
|                        | 2.0   | 3.965                        | <b>0.506</b>                 |
|                        | 1.5   | 2.586                        | <b>0.793</b>                 |
| 600                    | 2.5   | 10.062                       | <b>0.192</b>                 |
|                        | 2.0   | 7.471                        | <b>0.259</b>                 |
|                        | 1.5   | 5.013                        | <b>0.393</b>                 |

Table 8: Expected significance and median expected limits for various  $M_{v_1}$  and  $c_t$  values for the topophilic BSM model for 2018.

- 453   **A Control Region Distributions**
- 454   **B Resolved top tagger SF distributions**
- 455   **C Event-level BDT distributions**
- 456   **D Extended ABCD yields**
- 457   **E Extended ABCD validation**
- 458   **F ABCDnn predicted shape**
- 459   **G ABCDnn validation**
- 460   **H Extended ABCD closure test**
- 461   **I ABCDnn closure test**

462   **References**

- 463   [1] CMS Collaboration. Evidence for four-top quark production in proton-proton  
464   collisions at  $\sqrt{s} = 13\text{ TeV}$ . *Phys. Lett. B*, 844:138076, September 2023.  
465     
466    $\text{atmimg}=\text{"si1.svg"}$   $\text{atmml:msqrt}$   $\text{atmml:mrow}$   $\text{atmml:mi}$   $\text{atmml:msqrt}$   $\text{atmml:mrow}$   
467    $\text{atmml:mi}$   $\text{atmml:msqrt}$   $\text{atmml:mrow}$   $\text{atmml:mi}$   $\text{atmml:msqrt}$   $\text{atmml:mrow}$   
468    $\text{atmml:mi}$   $\text{atmml:mspace}$   $\text{atmml:width}$   $\text{atmml:mtex}$   $\text{atmml:mtex}$   $\text{atmml:math}$ . *Physics Letters B*, 844:138076,  
469   September 2023.
- 470   [2] Jeong Han Kim, Kyoungchul Kong, Seung J. Lee, and Gopolang Mohlabeng. Probing tev scale  
471   top-philic resonances with boosted top-tagging at the high luminosity lhc. *Physical Review D*,  
472   94(3), August 2016.
- 473   [3] Melissa van Beekveld, Anna Kulesza, and Laura Moreno Valero. Threshold resummation for  
474   the production of four top quarks at the lhc, 2022.