Search for 4 top quarks production in final states containing hadronic tau decays

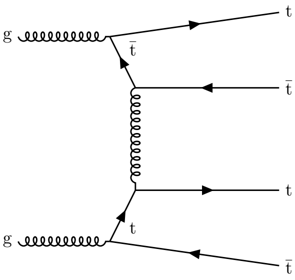
Huiling Hua [1] **Fabio Iemmi** [1] Duncan Leggat [2]
Hongbo Liao [1] Hideki Okawa [2] Yu Zhang [2]

[1]Institute of High Energy Physics (IHEP), Beijing

[2]Fudan University, Shanghai

November 24, 2021

4 top quarks production



- Process **yet to be observed** at the LHC
  - ATLAS close to observation: 4.7 σ obs (2.6 exp.)
- BSM contributions could enhance the cross section
- **Very rare** SM process: $\sigma^{\text{SM}}_{t\bar{t}t\bar{t}} \simeq 12$ **fb**
- **Challenging final states**: high jet multiplicity
- τ_h final states firstly explored in this analysis
- **Goal**: set UL on signal strength, give BSM interpretations using full Run2 data
- Following results based on **2016Legacy** data and MC

Object definitions

ELECTRONS

- $|\eta| < 2.5$; $p_T > 10$ GeV
- **MVA electron ID** developed by SUSY group
  - Tight WP
- **MultiIsolation** with LepAware JEC
  - Tight WP
- **2D IP** requirements
  - Tight WP
- **3D IP** requirements
  - Tight WP
- **Same cuts of 4tops SSDL**

MUONS

- $|\eta| < 2.4$; $p_T > 10$ GeV
- Muon ID developed by MUO POG
  - Medium WP
- **MultiIsolation** with LepAware JEC
  - Medium WP
- **2D IP** requirements
  - Tight WP
- **3D IP** requirements
  - Tight WP
- **Same cuts of 4tops SSDL**

Object definitions

JETS

- $|\eta| < 2.4$; $p_T > 25$ GeV
- AK4 jets
- CHS jets
- B tagging through **DeepJet**
  - Medium WP
- Cross-cleaned from lepton collection

TOPS

- Identified with resolved HOT
- 1% mistag rate WP

TAUS

- Reconstructed with HPS algorithm
- $|\eta| < 2.3$; $p_T > 20$ GeV
- **2D IP** requirements
- Identified with **DeepTau**
  - VVVLoose VsEle WP
  - VLoose VsMu WP
  - Medium VsJet WP
- Cross-cleaned from lepton collection

Simulated processes

- **$t\bar{t}t\bar{t}$ signal**
- **$t\bar{t}$ associated production**
  - Splitted in DL, SL and FH samples
- **$t\bar{t}$+X associated production**
  - $t\bar{t}W$, $t\bar{t}Z$, $t\bar{t}\gamma$, $t\bar{t}H$
- **QCD**
  - Splitted in H_T slices for $H_T > 200$ GeV
- **Single top**
  - tW
  - $\bar{t}W$
  - tZq($\ell\ell$)
  - tZq($\nu\nu$)
- **Single Higgs**
  - ggH($b\bar{b}$)
  - ggH(WW$\rightarrow \ell\nu$qq)
  - ggH(WW$\rightarrow \ell\nu\ell\nu$)
  - ggH($\gamma\gamma$)
  - ggH($\mu\mu$)
  - ggH($\tau\tau$)
  - VBF($b\bar{b}$)
  - VBF(WW$\rightarrow \ell\nu\ell\nu$)
  - VBF(4ℓ)
  - VBF($\tau\tau$)
  - VBF($\gamma\gamma$)
  - VBF($\mu\mu$)

Corrections to simulations

- **Implemented corrections** and scale factors to improve MC description of the data
  - JES and JER
  - MET filters
  - Pileup
  - Prefiring
  - Trigger
  - b tagging (BTagShapeCalibration)
  - Electron SFs (ID, ISO, IP)
  - Muon SFs (ID, ISO, IP)
- **To be implemented**:
  - DeepTau SFs
  - Reosolved HOT SFs
  - $t\bar{t}+b\bar{b}$ corrections

Baseline selection

- First, **apply loose preselection** to retain $t\bar{t}t\bar{t}(\tau_h)$-like events

- To do so, **use somehow looser object definitions** than the ones showed previously
- Three object definitions: **loose**, **fakeable** and **tight** objects
- Previous cuts define tight objects
- Analysis **baseline selection** is:
  1. $N_{\text{loose}\tau_h} > 0$
  2. $N_{\text{jets}} \geq 2$
  3. $N_{\text{loose b-jets}} \geq 2$

	Loose	Fakeable	Tight
Electron	$\lvert\eta\rvert < 2.5$; $p_T > 10$ GeV Loose MVA ID Loose ISO (SSDL) 2D IP cut ExpMissInnerHits ≤ 1	$\lvert\eta\rvert < 2.5$; $p_T > 10$ GeV Loose MVA ID Loose ISO (SSDL) 2D IP cut 3D IP cut ExpMissInnerHits $= 0$	$\lvert\eta\rvert < 2.5$; $p_T > 10$ GeV Tight MVA ID Tight ISO (SSDL) 2D IP cut 3D IP cut ExpMissInnerHits $= 0$
Muon	$\lvert\eta\rvert < 2.4$; $p_T > 10$ GeV Loose MUO POG ID Loose ISO (SSDL) 2D IP cut	$\lvert\eta\rvert < 2.4$; $p_T > 10$ GeV Medium MUO POG ID Loose ISO (SSDL) 2D IP cut 3D IP cut	$\lvert\eta\rvert < 2.4$; $p_T > 10$ GeV Medium MUO POG ID Tight ISO (SSDL) 2D IP cut 3D IP cut
Tau	$\lvert\eta\rvert < 2.3$; $p_T > 20$ GeV 2D IP cut VVLooseVsJet	$\lvert\eta\rvert < 2.3$; $p_T > 20$ GeV 2D IP cut Exclude DM=5,6 VVLooseVsJet VLooseVsMu VVVLooseVsEle	$\lvert\eta\rvert < 2.3$; $p_T > 20$ GeV 2D IP cut Exclude DM=5,6 MediumVsJet VLooseVsMu VVVLooseVsEle

Trigger strategy

- Tried different trigger strategies
- **First**: try to inherit trigger setup of $t\bar{t}H$ multilepton (ML) analysis
  - Complicated combination of single-, double-, triple-lepton triggers, lepton+tau triggers, double tau triggers
- **Second**: use a simpler combination of multijet triggers
  - `HLT_PFHT450_SixJet40_BTagCSV_p056`
  - `HLT_PFHT400_SixJet30_DoubleBTagCSV_p056`
  - Same choice of $t\bar{t}H(b\bar{b})$ analysis
- **Tau triggers** are found to be **inefficient** for our signal
- **Multijet triggers** provide **good signal efficiency** (more in backup)
- Decided to **use the multijet triggers setup**

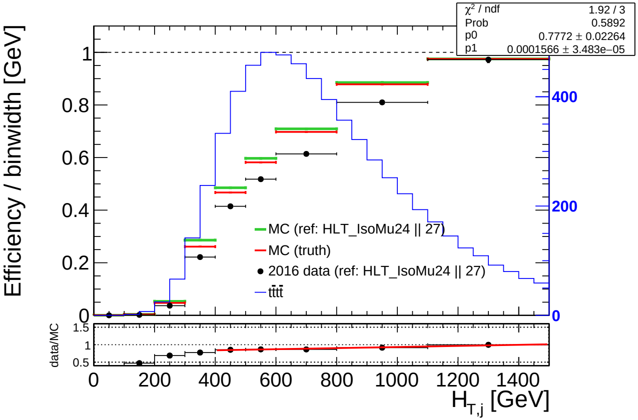
Some remarks about multijet triggers

- Well known issue with 2016 data: **efficiency in data is lower than in MC** by a non-negligible amount **at high H_T**
- The **issue is understood** (see, e.g., $t\bar{t}H(b\bar{b})$ AN):

> Initially a drop in efficiency in data at high HT was observed, which is attributed to the last run period of the LHC in 2016 (Run H) which had very high instantaneous luminosity. The L1 HT triggers suffered a problem in which saturated (high p_T) jets were excluded from the HT calculation [62]. A partial mitigation strategy of including an OR of a single jet trigger `HLT_PFJet450_v*` has been implemented, which recovers most of the lost efficiency at high HT.

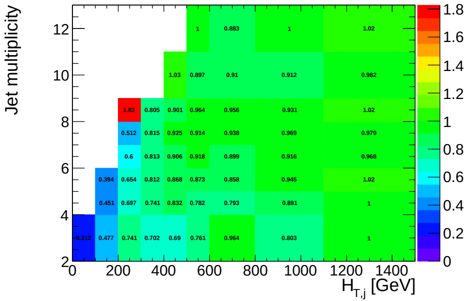
- **Include `HLT_PFJet450`** in the signal triggers
  - Recover some efficiency at high H_T

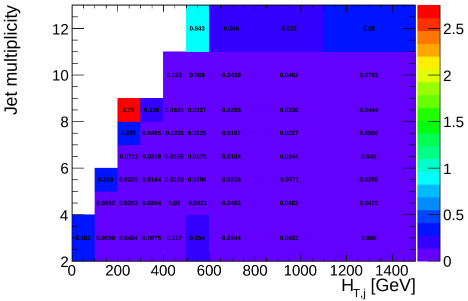
One-dimensional trigger efficiency



- Preselection + 1μ
- Plotted data, MC wrt reference, MC truth
- **Also plotted H_T distribution for signal**
- A $H_T > 400$ GeV cut saves enough signal and makes trigger efficient

Two-dimensional trigger efficiency





- nJets vs H_T trigger efficiency
- Left: data/MC efficiency ratio; right: corresponding errors
- **Add $H_T > 400$ GeV cut to analysis selection** to make trigger efficient
- Use these histograms as **trigger efficiency scale factors and uncertainties**

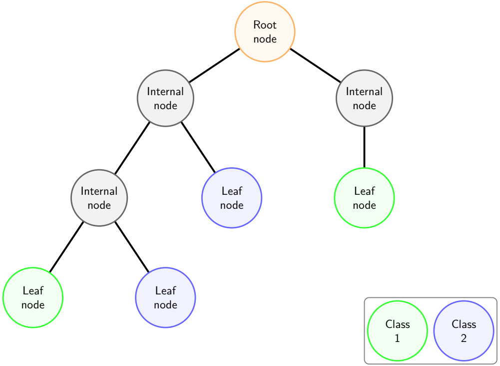
Analysis categories

Category	τ_h	ℓ	$\mathrm{N_{jets}}$	$\mathrm{N_{b\text{-}jets}}$
1tau0L	1	0	≥ 8	≥ 2
1tau1L	1	1	≥ 6	≥ 2
1tau2L	1	2	≥ 4	≥ 2
1tau3L	1	3	≥ 2	≥ 2
2tau0L	2	0	≥ 6	≥ 2
2tau1L	2	1	≥ 4	≥ 2
2tau2L	2	2	≥ 2	≥ 2

- **Phase space splitting** based on τ_h, ℓ, jets, b-jets multiplicities
- Two hadronic categories + six leptonic categories
- **Strategy**:
  - For categories with BDT, **fit BDT shape** (see the following)
  - For remaining categories, **fit $\mathbf{H_T}$ shape**

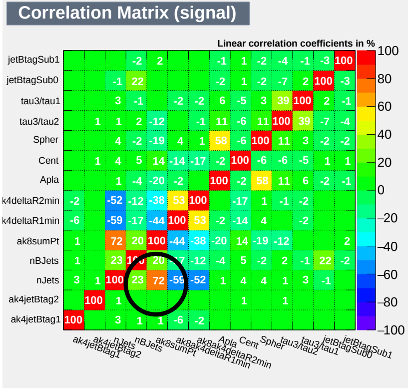
Multivariate analysis

- Train a **BDT** to better **separate signal from background**
- Use TMVA package
- Sufficient stats for training in
  - 1tau1L
  - 1tau2L
  - 2tauXL = 2tau1L + 2tau2L
- Input set of variables optimized by **correlation-based** removal
- **Goal**: achieve optimal performance while keeping the input set small



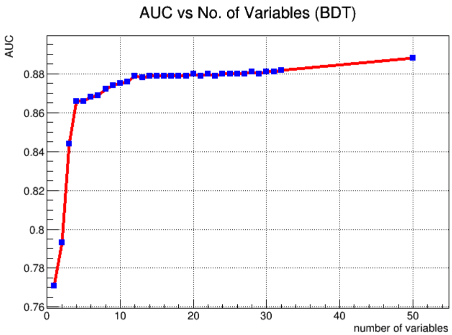
Correlation-based variable removal

- **Correlation-based variable removal**:
  - Start with set of 50 variables showing best separation power
  - **Find** pair of variables with **highest correlation**
  - **Remove** the one with lower separation power, **retrain**
  - **Repeat** until 1 variable is left in the set
  - **Plot AUC** as a function of number of variables
  - Choose smallest number of variables before performance drops

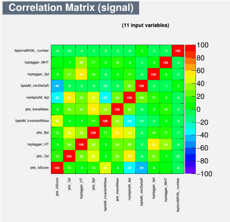


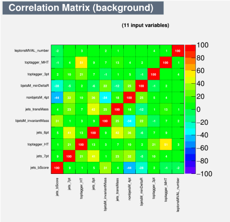
Correlation-based variable removal (1tau1L)

- AUC shows a **plateau above 11 input features**
- Knee below 11 input features
- Drop below 5 input features
- We **use 11 input features**
  - Sum of jets b tag scores
  - 7th jet p_T
  - Resolved tops H_T
  - 6th jet p_T
  - Invariant mass of b tagged jets
  - Transverse mass of jets
  - 4th non b tagged jet p_T
  - Minimum ΔR between b tagged jets
  - 3rd resolved top p_T
  - Vector sum of resolved top p_T
  - Number of loose leptons

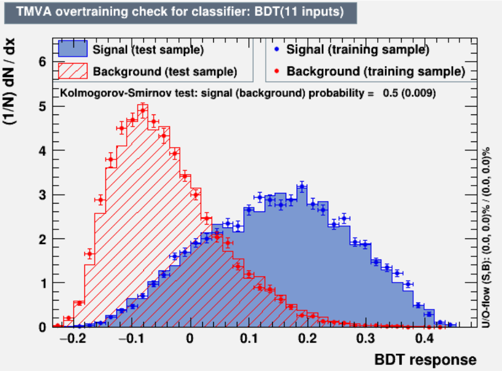


Training results (1tau1L)

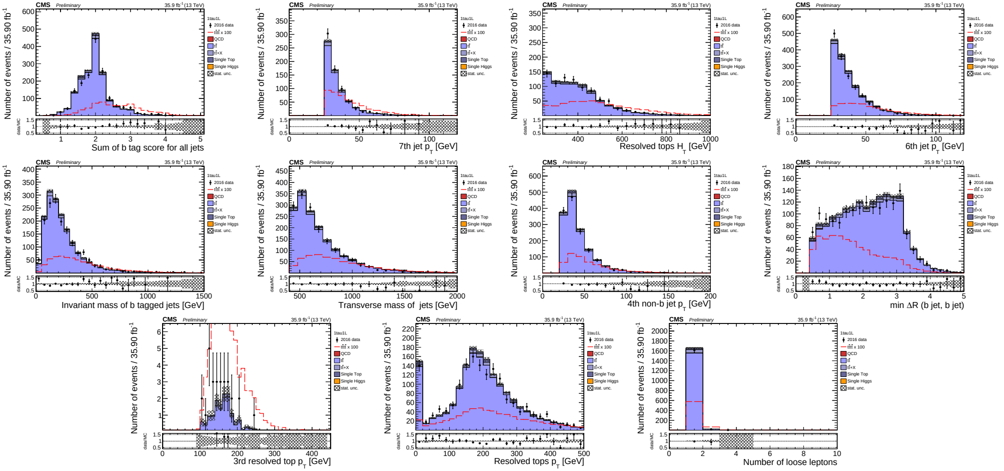




- Correlation-based removal keeps variables with low-medium correlation
- Results obtained with standard hyperparameters $\implies$ **room for improvement after hyperparameter tuning**



Validation of input variables (1tau1L)



QCD estimation in 1tau0L

- We have an **hadronic category** in our analysis, **1tau0L**
  - 1 τ_{h}, no leptons, $\geq$ 8 jets, $\geq$ 2 b jets
- **Dominant background** in 1tau0L is **QCD** multijet production
- **MC predictions for QCD cannot be safely used**
  - Big theoretical uncertainties on cross sections and NLO corrections
  - Usually very low selection efficiency $\implies$ poor statistics
- Look for a **data-driven estimation of the QCD background**
- **Both yield and shape** are estimated **from data**
  - **Yield**: from fake rate method
  - **Shape**: from control region in data

QCD regions

- **Three regions** are involved in our QCD studies:
  - **Signal region** (SR): where analysis is performed
  - **Control region** (CR): where fake rates and shapes are extracted
  - **Validation region** (VR): where QCD estimation is validated

	N_{τ_h}	N_ℓ	N_{jets}	N_{bjets}	$t\bar{t}t\bar{t}$	$t\bar{t}$	QCD	$t\bar{t}+X$
SR	1	0	≥ 8	≥ 2	10	6371	7461	192
VR	1	0	≥ 8	1	1	2321	7792	79
CR	1	0	≥ 8	0	0	294	8979	8

- The **large QCD simulated yield** that we get **in CR comes from fake taus**
- Use **fake rate method** to estimate this yield from data

Fake rate method

- Estimate the background completely from data by doing

$$N_{\text{fake-}\tau} = \sum_{p_T,\eta} N_{\text{fake-}\tau}(p_T,\eta) = \sum_{p_T,\eta} \left[N_{F,\overline{T}}(p_T,\eta) \times \frac{FR(p_T,\eta)}{1 - FR(p_T,\eta)} \right]$$

- $N_{F,\overline{T}}(p_T,\eta)$, number of fakeable-non-tight taus in SR
- $FR(p_T,\eta)$, probability for a fakeable tau to be a tight tau, computed in the CR
- Parametrize as a function of p_T, η of fakeable tau
- Binning in (p_T,η): $p_T \in [20, 30, 75, 150, 300, \text{Inf}]$; $\eta \in [0, 1.5, 2.3]$
- Performed several **sanity checks** before applying FR method, see backup

Fake rate method in data

- Compute FR in CR, **apply** the method **using previous formula**
- **Important**: take care of subtracting $t\bar{t}$ and $t\bar{t}$+X from $N_{\mathrm{F},\overline{\mathrm{T}}}(p_{\mathrm{T}}, \eta)$

	MC QCD	FR method
Yield	7461 ± 1681	7679 ± 273

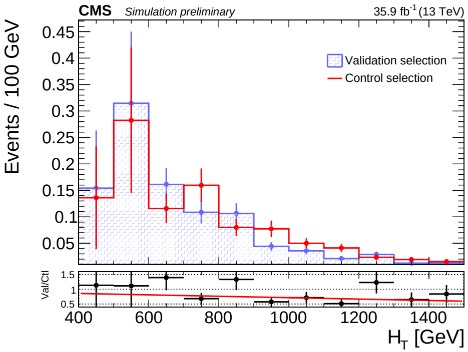
- The **estimated yield** from FR method is **in agreement with QCD MC predictions**
- But it comes with **way lower uncertainty** (4% vs 23%)

QCD shape estimation: general idea

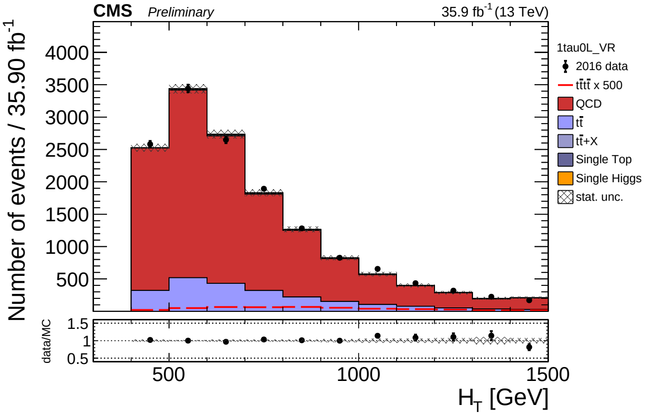
- Take the **QCD shape from** the **CR in data**
- Need to extrapolate from shape in CR to shape in different QCD regions
- **Correct for kinematic differences** between CR and region of interest **using the simulation**
- Take the ratio of H_T shapes in CR and region of interest, fit it and **get a transition function**
- **Apply the transition function** to the data distribution in CR **to get the final shape**

Transition function

- To validate method, compute QCD shape in the VR
- **Just compare shapes**: normalize areas to 1
- Smoothen the ratio by **fitting with a straight line**
- This straight **transition function is applied to the H_T distribution of data in the CR** to obtain the final shape



QCD estimation: validation



- Good data/MC closure in VR $\Longrightarrow$ validates the QCD estimation
- Estimate systematic unc. on yield by the level of disagreement in ratio plot

Expected yields

	$t\bar{t}t\bar{t}$	$t\bar{t}$	$t\bar{t}$+X	QCD	Single top	Single Higgs
1tau0L	8.79	5389.6	171.0	7679	111.1	-0.29
1tau1L	6.47	1570.4	73.3	2.2	31.7	0.029
1tau2L	1.25	24.5	10.0	0	0.22	0
1tau3L	0.07	0	0.57	0	0	0
2tau0L	0.44	168.2	13.0	1.70	6.3	0.015
2tau1L	0.17	8.6	3.7	0	0.08	0
2tau2L	0.014	0.08	0.20	0	0	0

Preliminary results

- We wrote a first, **stat-only datacard**
- Fit simultaneously shapes in 1tau0L, 1tau1L, 1tau2L, 1tau3L, 2tau0L, 2tauXL
- Get expected upper limit on signal strength:
  `combine -M AsymptoticLimits datacardname.root --run blind`
  - **Expected upper limit** on $\mu_{t\bar{t}t\bar{t}}$ at 95% CL:

$$\boxed{\mu_{t\bar{t}t\bar{t}} < 2.0156}$$

- Get expected significance of the measurement:
  `combine -M Significance datacardname.root -t -1 --expectSignal=1`
  - **Expected significance**:

$$\boxed{1.0389\ \sigma}$$

Summary

Prospects and publication strategy

- Looking for conveners' advice about long-term strategy
- Concerning a **possible combination**, main point is **hadronic tau veto**
  - Other channels currently do not apply such a veto
- Checked **DeepTau** documentation looking for **efficiencies and mistag rates**
- Found some information in AN Study of the misidentification of jets, electrons and muons as hadronically decaying tau leptons with the DeepTau ID for the full Run II data.
- Document provides mistag rate for VsJet discriminant only
  - Our working point: 70% eff., 0.2% mistag rate (tables 1 and 19)
- Numbers seem encouraging, but further studies should be carried on

Backup slides

Samples

- $t\bar{t}t\bar{t}$
  - TTTT_TuneCP5_PSweights_13TeV-amcatnlo-pythia8_correctnPartonsInBorn
- $t\bar{t}$
  - TTTo2L2Nu_TuneCP5_PSweights_13TeV-powheg-pythia8
  - TTToSemiLeptonic_TuneCP5_PSweights_13TeV-powheg-pythia8
  - TTToHadronic_TuneCP5_PSweights_13TeV-powheg-pythia8
- $t\bar{t}+X$
  - ttWJets_13TeV_madgraphMLM
  - ttZJets_13TeV_madgraphMLM-pythia8
  - TTGJets_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8
  - ttH_4f_ctcvcp_TuneCP5_13TeV_madgraph_pythia8

Samples

- QCD
  - QCD_HT200to300_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT300to500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT500to700_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT700to1000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT1000to1500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT1500to2000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
  - QCD_HT2000toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8
- Single top
  - ST_tW_top_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M2T4
  - ST_tW_antitop_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M2T4
  - tZq_ll_4f_ckm_NLO_TuneCP5_PSweights_13TeV-amcatnlo-pythia8
  - tZq_nunu_4f_13TeV-amcatnlo-pythia8_TuneCUETP8M1

Trigger efficiency for $t\bar{t}t\bar{t}$ signal

$$\varepsilon^i = \frac{N^i_{\text{trig}}}{N^i},$$

N^i = number of events falling in category i; N^i_{trig} = number of events also passing trigger

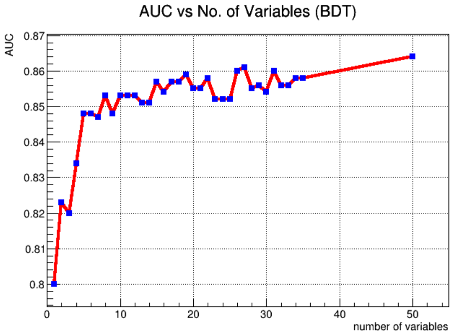
	1tau0L	1tau1L	1tau2L	1tau3L	2tau0L	2tau1L	2tau2L
N^i_{trig}	498.17	413.49	110.48	7.70	25.97	12.74	1.94
N^i	505.17	425.05	118.31	8.14	26.89	14.03	2.23
ε^i	0.986	0.973	0.933	0.946	0.966	0.908	0.870

- With $H_T > 400$ GeV cut:

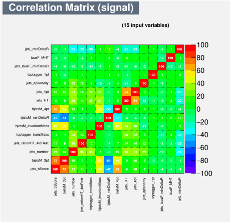
	1tau0L	1tau1L	1tau2L	1tau3L	2tau0L	2tau1L	2tau2L
N^i_{trig}	484.31	347.75	70.43	3.58	24.60	8.58	1.03
N^i	488.43	352.27	71.94	3.62	25.01	8.88	1.03
ε^i	0.991	0.986	0.979	0.989	0.984	0.966	1

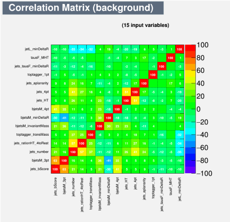
Correlation-based variable removal (1tau2L)

- AUC shows a **plateau above 15 input features**
- Drop below 5 input features
- We **use 15 input features**
  - to be filled

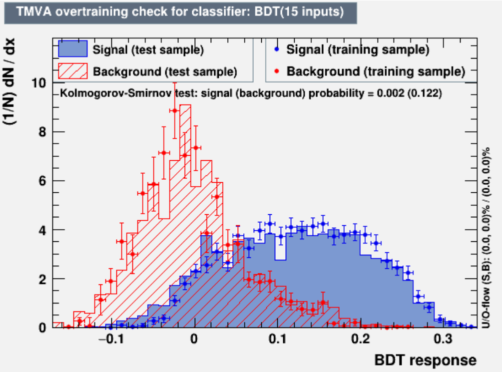


Training results (1tau2L)

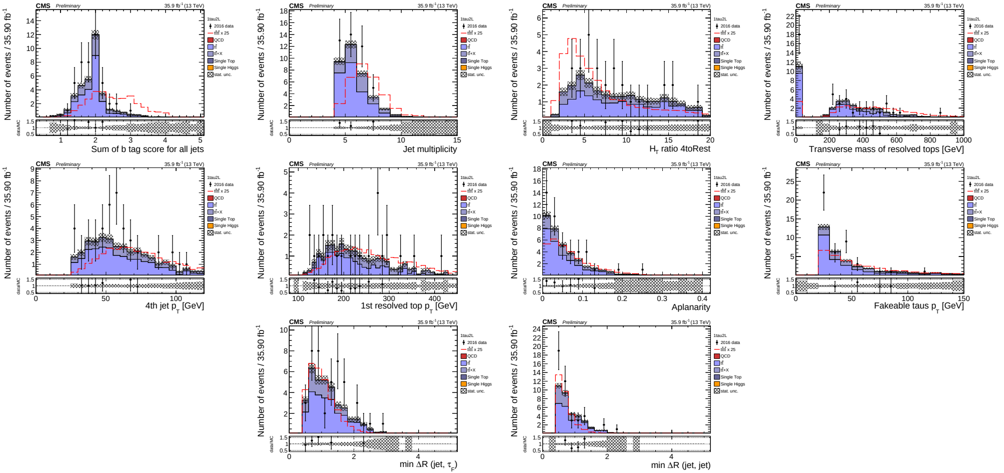




- Correlation-based removal keeps variables with low-medium correlation



Validation of input variables (1tau2L)

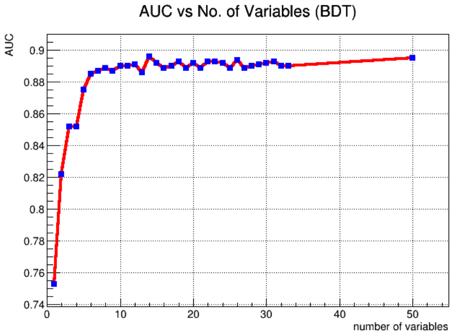


Status report

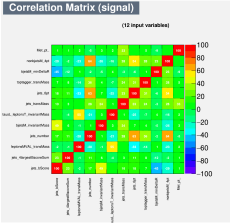
F. Iemmi

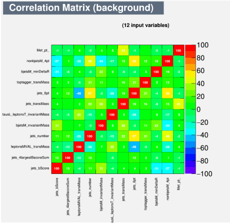
Correlation-based variable removal (2tauXL)

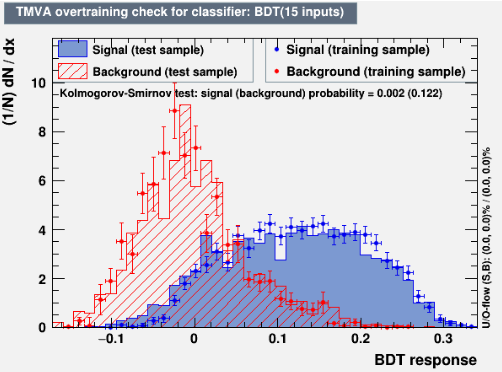
- AUC shows a **plateau above 12 input features**
- Drop below 5 input features
- We **use 12 input features**
  - to be filled



Training results (1tau2L)

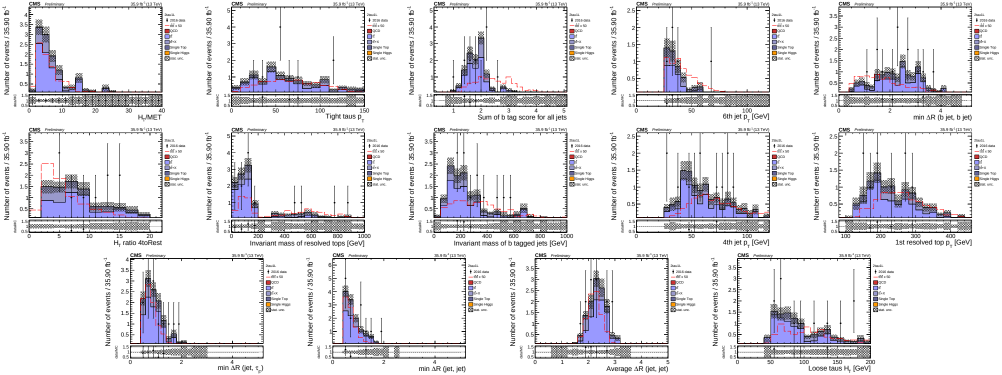






- Correlation-based removal keeps variables with low-medium correlation

Validation of input variables (2tau1L)



Closure test in MC QCD

- Compute FR in CR, **apply the method in** the same **CR**
- **Compare with number** of events in CR **you count from MC**
- These numbers should close

	Value	Raw entries
Counting	7979 ± 1350	547
Fake rate method	8636 ± 2321	–

- Values are in **agreement** within the uncertainties, **closure is not perfect** (8% discrepancy)
  - **Due to approximations** in weighting and summing `TEfficiency` objects
  - See my discussion with ROOT developer Lorenzo Moneta here

Fake rate method in MC QCD

- Compute FR in CR, **apply the method in** the **application region** (same as signal region, but use fakeable-not-tight taus)
- **Compare with number** of events in SR **you count from MC**

	Value	Raw entries
Counting	7461 ± 1681	315
Fake rate method	5887 ± 1782	–

- Values are in **agreement within the uncertainties**
- Uncertainties are big due to poor statistics in MC samples

Closure test in data

- Compute FR in CR, **apply the method in** the same **CR**
- **Compare with number** of events in CR **you count from data**
- This should close (at least approximately)

	Value	Raw entries
Counting	11561 ± 108	11561
Fake rate method	11561 ± 384	–

- Values are in **agreement** within the uncertainties, **perfect closure**
  - No weighting of any kind of objects is needed for data

Uncertainties on the QCD shape

- QCD H_T shapes are taken from the CR in data and translated to VR or SR **using** the corresponding **transition functions (TFs)**
- TFs are the result of a fit: ROOT gives you the **fitted parameters and the correlation matrix** $\mathcal{V}$ of the fit
- In our case, we fitted with straight lines of the form

$$y = mx + q,$$

so the correlation matrix will look like

$$\mathcal{V} = \begin{bmatrix} \sigma_q^2 & \rho_{qm} \\ \rho_{mq} & \sigma_m^2 \end{bmatrix},$$

where $\sigma_{q/m}^2$ are the variances of the parameters and $\rho_{qm} = \rho_{mq}$ are the correlation coefficients between m and q

Uncertainties on the QCD shape

- In general, $\rho_{qm} = \rho_{mq} \neq 0$, i.e., some degree of correlation exists between the two parameters
- This means **one cannot shift** m **and** q **up and down independently**
- $\mathcal{V}$ is a real, symmetric matrix $\implies$ it can always been diagonalized by means of an orthogonal transformation
- This means **it exists some auxiliary parameter space in which** m **and** q **are fully decorrelated**
  - One can shift them up/down independently in this space
- Linear algebra theorem: the orthogonal diagonalizing matrix $\mathcal{O}$ has the eigenvectors of $\mathcal{V}$ as columns

$$\mathcal{D} = \mathcal{O}^{-1}\mathcal{V}\mathcal{O} = \mathcal{O}^{T}\mathcal{V}\mathcal{O}$$

Uncertainties on the QCD shape

- **Idea**: Starting from the "real" parameters, described by the vector $\mathbf{p}^T = (q, m)$, we first transform them to some auxiliary parameters $\tilde{\mathbf{p}}^T = (\tilde{q}, \tilde{m})$:

$$\tilde{\mathbf{p}} = \mathcal{O}\mathbf{p}$$

- In the auxiliary space, the correlation matrix is diagonal and its non-zero elements are the variances of $\tilde{m}, \tilde{q}$

$$\mathcal{D} = \begin{bmatrix} \tilde{\sigma}^2_{\tilde{q}} & 0 \\ 0 & \tilde{\sigma}^2_{\tilde{m}} \end{bmatrix}.$$

- Now the parameters can be shifted independently, so we define the shifted TFs in the auxiliary space to be described by

$$\tilde{\mathbf{p}}^T_{\text{up}} = (\tilde{q} + \tilde{\sigma}_{\tilde{q}}, \tilde{m} + \tilde{\sigma}_{\tilde{m}})$$
$$\tilde{\mathbf{p}}^T_{\text{down}} = (\tilde{q} - \tilde{\sigma}_{\tilde{q}}, \tilde{m} - \tilde{\sigma}_{\tilde{m}})$$

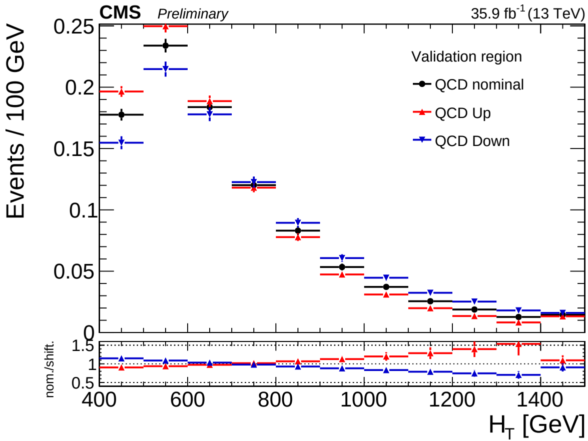
Uncertainties on the QCD shape

- Finally, we perform the **inverse transformation** to go back and get the parameters describing the **TFs in the original space**

$$\mathbf{p}_{\text{up}} = \mathcal{O}^{-1}\tilde{\mathbf{p}}_{\text{up}}$$
$$\mathbf{p}_{\text{down}} = \mathcal{O}^{-1}\tilde{\mathbf{p}}_{\text{down}}$$

- Now compare nominal shapes with the upwards/downwards shifted shapes
- Scale all areas to one: we are **interested in the shape differences**
  - The yield will be coming from FR for all of them

QCD shape uncertainty: VR



	q	m
Nominal	0.95	-0.00024
Up	1.31	-0.00064
Down	0.59	0.00016