山东大学
SHANDONG UNIVERSITY

Department of Physics, Shandong University

# Compressed EWK study(ISRC1N2)

**Chengxin Liao**
liaocx@ihep.ac.cn

**Apr, Thu 17, 2025**

# Tasklist

- FF method var distribution check
- Multiclass result(failed)
- BSc thesis: https://www.overleaf.com/project/674e7119837a2580151a0868 (need to submit before the end of Apr)

# Introduction

Reference point: C1 mass = 100GeV, N2 mass = 70 GeV

SR optimization:  BDT method, optuna to auto-optimize hps

For FF method constriant

HH Pre-Selection

Had-Had channel: nBaseTau=2; nMediumTau=2; nLeps=0
MET >= 200,  MET trigger
Opposite-sign hadronic tau pair
bVeto
Mtt_reco < 40GeV && Mtt_reco > 130GeV
nBaseJet >= 1
Jet pt > 100GeV

HH Pre-Selection

Had-Had channel: nBaseTau=1; nMediumTau=1; nLeps>=1
MET >= 200,  MET trigger
Opposite-sign hadronic tau pair
bVeto
Mtl_reco < 40GeV && Mtl_reco > 130GeV
nBaseJet >= 1
Jet pt > 100GeV

# Multiclass(HH)

Hyperparameters: Ntrees = 200, MaxDepth = 6, MinNodeSize = 2%, Learning rate = 0.03(initial setting)

Feature engineering:

Select a simple model and put all features into model, choose Top 30 vars based on importance list, drop high correlated vars

Final feature list:

| Rank | Variable | Variable Importance |
|---|---|---|
| 1 | fb_dRtt | 8.238e-02 |
| 2 | fb_dRMax_xt | 7.068e-02 |
| 3 | fb_METsig | 6.205e-02 |
| 4 | fb_frac_MET_tt | 6.050e-02 |
| 5 | fb_dPhit1x | 5.751e-02 |
| 6 | fb_MIA | 5.460e-02 |
| 7 | fb_mt_taumin | 5.411e-02 |
| 8 | fb_Asy_tt | 5.363e-02 |
| 9 | fb_dEtat2j | 4.903e-02 |
| 10 | fb_MET_Soft | 4.737e-02 |
| 11 | fb_Asy_EH | 4.625e-02 |
| 12 | fb_Mll | 4.447e-02 |
| 13 | fb_frac_MET_MeffInc_40 | 4.317e-02 |
| 14 | fb_eta_jet2 | 4.282e-02 |
| 15 | fb_eta_jet1 | 4.229e-02 |
| 16 | fb_transSphericity | 4.140e-02 |
| 17 | fb_dRMax_jets | 4.086e-02 |
| 18 | fb_frac_MET_Meff | 4.007e-02 |
| 19 | fb_m_jet1 | 3.421e-02 |
| 20 | fb_nJets30 | 3.260e-02 |

Weight choose: no weight, abs(weight)

No weight have better performance
but abs(weight) fit our anaysis requirment

Split strategy: Separate entries by using mod 5, for Fake bkg, if separate follow sequence, all weighted entry will split into first fold

# Multiclass(HH)

Class: C1N2, VV, Other_bkg

Test_auc = \sum{Test_auc_class}

Train_auc = \sum{Train_auc_class}

Hyperparameter tune:
use optuna to auto-optimze

constraint:

average of AUC need to >= 0.6

penalty function: score = test_auc – 0.3*auc_gap (auc_gap = abs(train_auc- test_auc))

maximum(score)

After check some models, find C1N2 result is great, so the constraint and AUC calculation only in VV and Other bkg

**Grid Search**
**Ntrees: [200, 300, 400]**
**MaxDepth: [4, 6, 8, 10]**
**MinNode: [1, 3, 5, 7]**
**Learning rate: [0.001, 0.005, 0.01, 0.05, 0.1]**

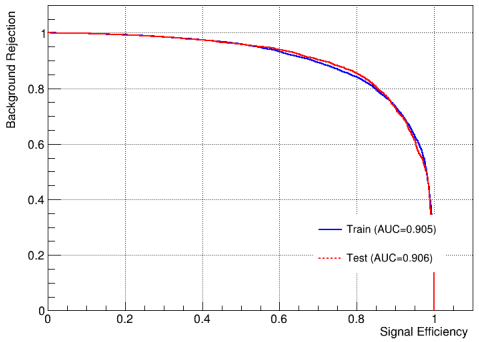
Best one: Ntree=300, MaxDepth=6, MinNode=1%, Learning Rate=0.05

There still have rooms to optimize for lr

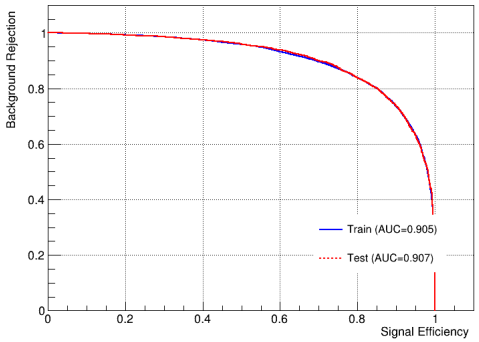
# Multiclass(HH)

## OverFit Check

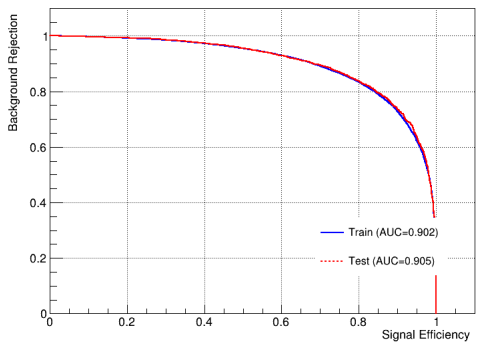
C1N2 (Fold 0: BDT0_300_6_1_005)



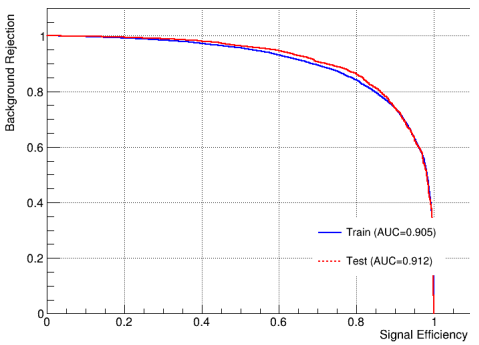

C1N2 (Fold 1: BDT1_300_6_1_005)



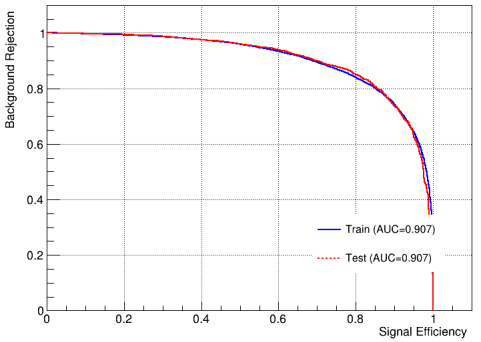

C1N2 (Fold 2: BDT2_300_6_1_005)



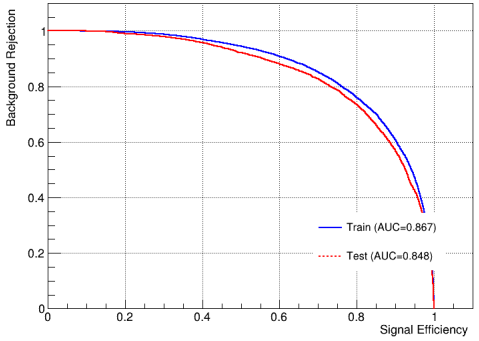

C1N2 (Fold 3: BDT3_300_6_1_005)



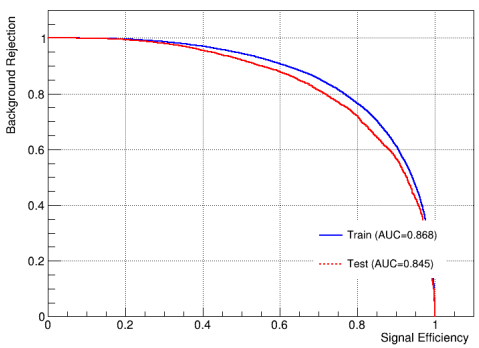

C1N2 (Fold 4: BDT4_300_6_1_005)



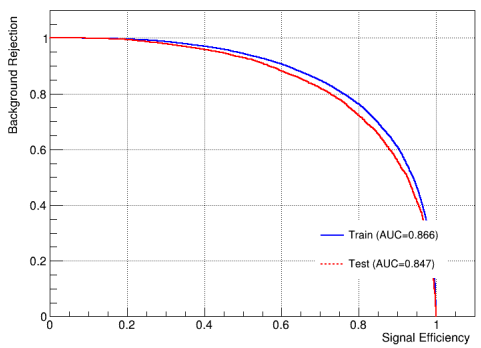

Other_bkg (Fold 0: BDT0_300_6_1_005)



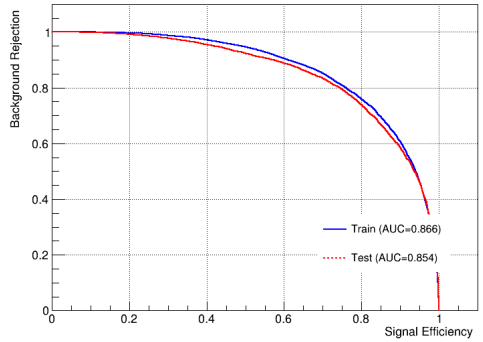

Other_bkg (Fold 1: BDT1_300_6_1_005)



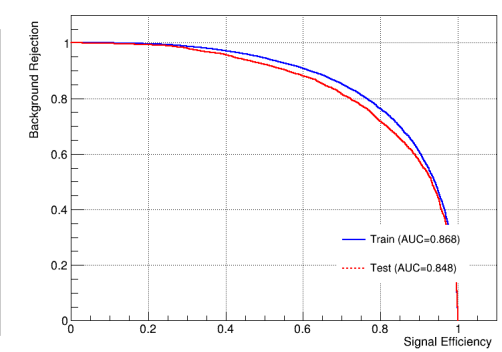

Other_bkg (Fold 2: BDT2_300_6_1_005)



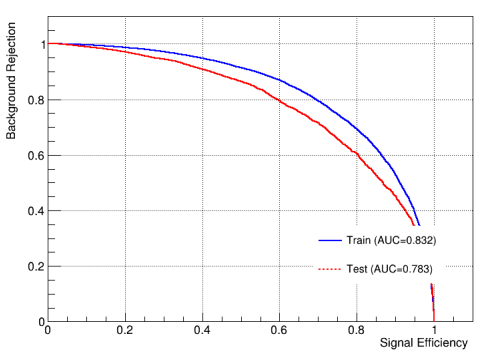

Other_bkg (Fold 3: BDT3_300_6_1_005)



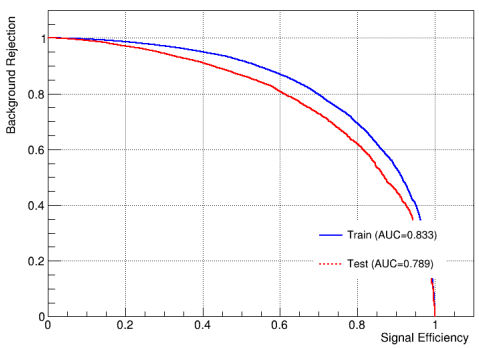

Other_bkg (Fold 4: BDT4_300_6_1_005)



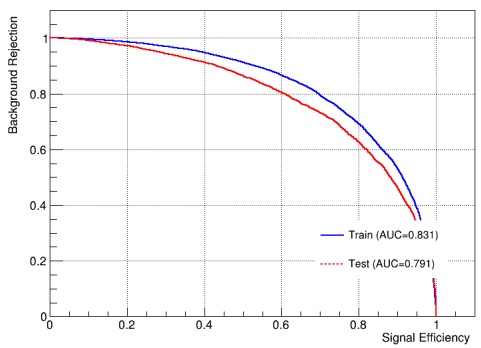

VV (Fold 0: BDT0_300_6_1_005)



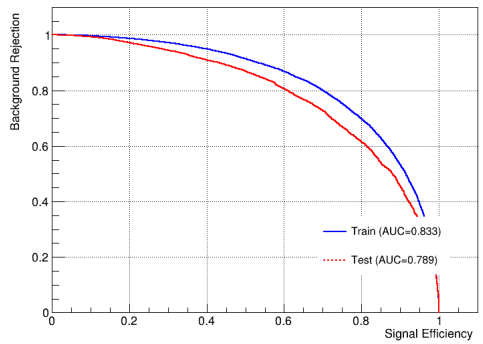

VV (Fold 1: BDT1_300_6_1_005)



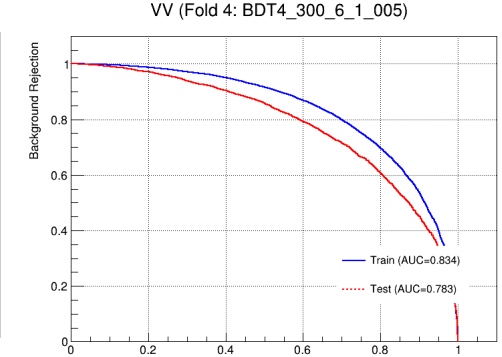

VV (Fold 2: BDT2_300_6_1_005)




VV (Fold 3: BDT3_300_6_1_005)

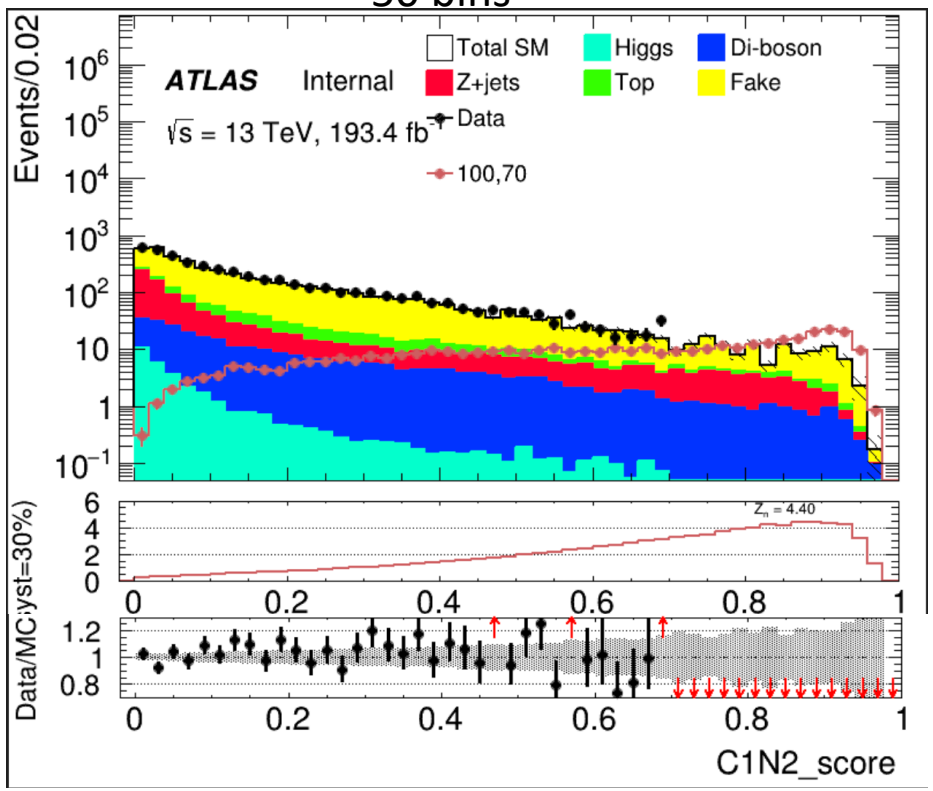



VV (Fold 4: BDT4_300_6_1_005)



# SR(HH) Multi

50 bins




Each index stand for a region, and index 10 stands for sum

# Zjets(HH)

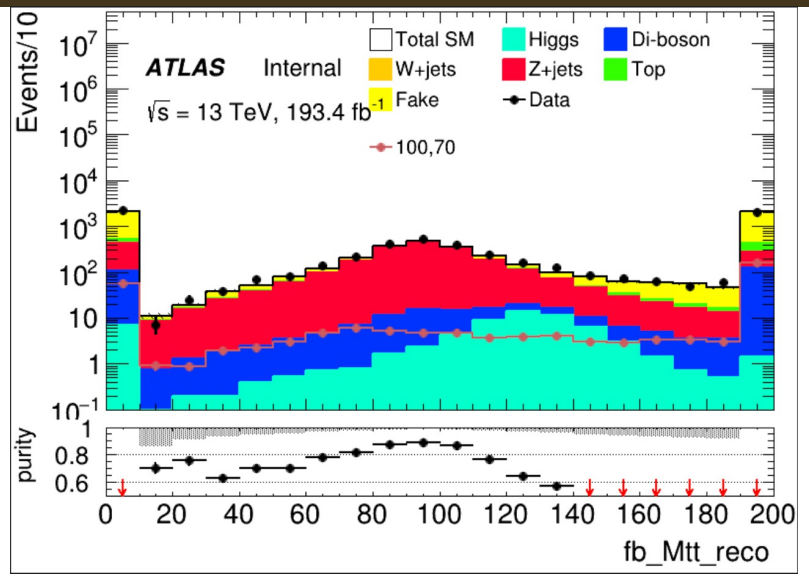
== 2 medium tau
== 0 lepton
METtrig && MET >= 200
OS
bVeto
nBaseJet >= 1
Jet pt > 100GeV

Same with pre-selection

C1N2 score <= 0.7(Othrogonal with SR)

CR: Mtt reco >= 80 && Mtt reco  <= 110
VR: (Mtt reco >= 40 && Mtt reco < 80) || (Mtt reco > 110 && Mtt reco < 130)

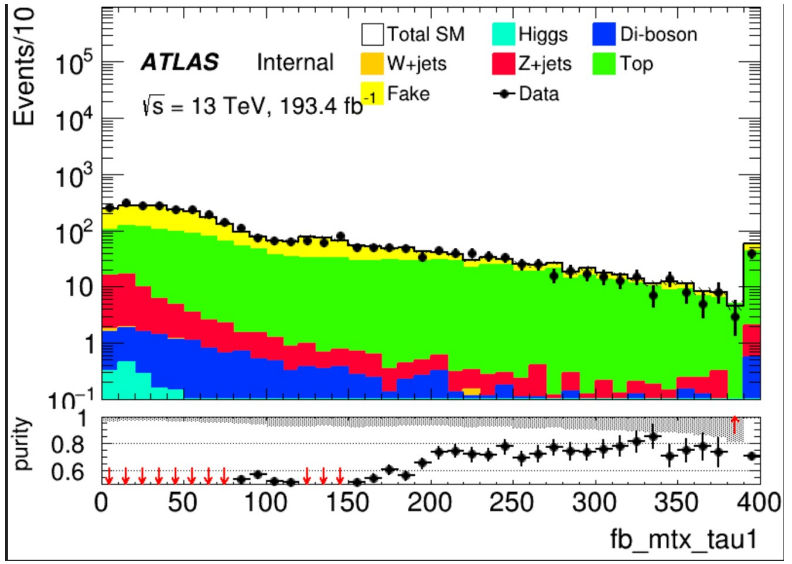
| Region | TotalBkg | Zjets | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 1433.93+-11.003 | 1232.36+-5.476 | 0.859 | 1571 | 1.096 |
| VR1 | 824.938+-8.468 | 674.307+-4.082 | 0.817 | 904 | 1.097 |
| VR2 | 469.252+-7.834 | 321.877+-2.913 | 0.685 | 524 | 1.117 |



# Top(HH)

== 2 medium tau
== 0 lepton
METtrig && MET >= 200
OS
nBaseJet >= 1
Jet pt > 100GeV
Mtt_reco < 40GeV && Mtt_reco > 130GeV

Same with pre-selection




>= 1 bTag(improve top events)
C1N2 score <= 0.7(Othrogonal with SR)

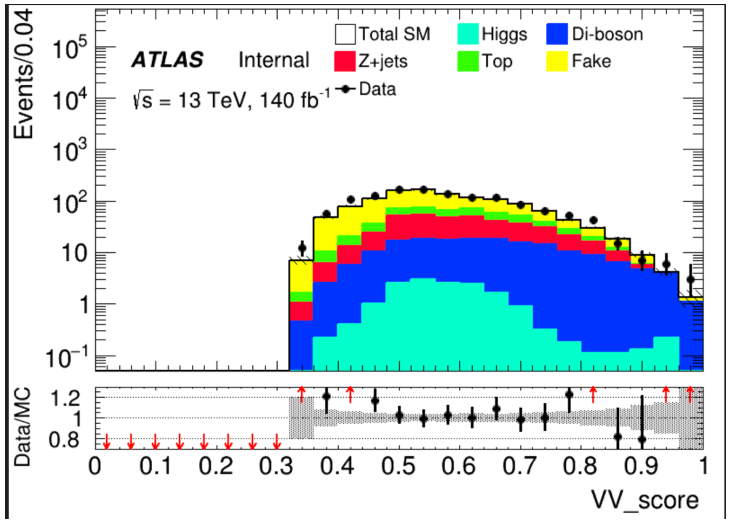
CR: $M_T(\tau, MET) > 250$
VR: $200 < M_T(\tau, MET) < 250$

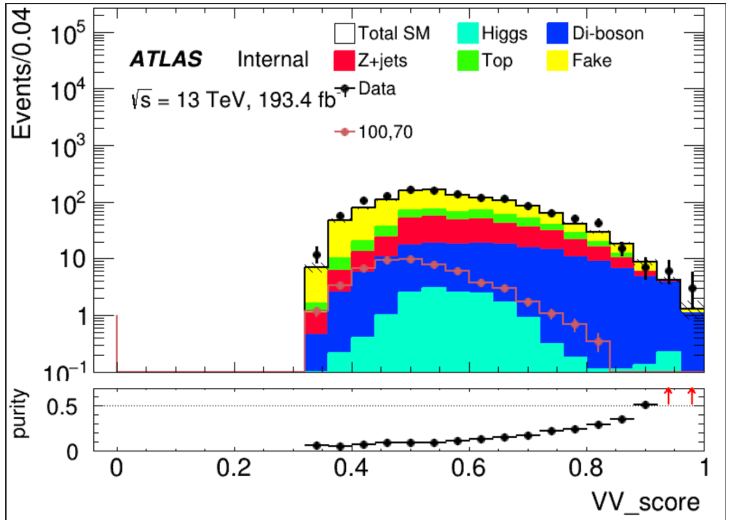
Huge overestimation

| Region | TotalBkg | Top | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 330.84+-7.206 | 244.867+-4.875 | 0.740 | 261 | 0.7889 |
| VR | 198.391+-6.284 | 145.391+-4.064 | 0.733 | 217 | 1.096 |

# VV(HH)

Pre-selection && C1N2_score <= 0.7 && Max(score) == VV_score && VV_score >= 0.85




| Region | TotalBkg | VV | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| VR | 32.4944+-2.09 | 15.4902 | 0.476 | 31 | 0.954 |

# Binary class(HH)

Hyperparameters: Ntrees = 200, MaxDepth = 6, MinNodeSize = 2%, Learning rate = 0.03(initial setting)

Feature engineering:
Select a simple model and put all features into model, choose Top 30 vars based on importance list, drop high correlated vars

Final feature list:

| Rank | Variable | Variable Importance |
|---|---|---|
| 1 | fb_dRtt | 8.238e-02 |
| 2 | fb_dRMax_xt | 7.068e-02 |
| 3 | fb_METsig | 6.205e-02 |
| 4 | fb_frac_MET_tt | 6.050e-02 |
| 5 | fb_dPhit1x | 5.751e-02 |
| 6 | fb_MIA | 5.460e-02 |
| 7 | fb_mt_taumin | 5.411e-02 |
| 8 | fb_Asy_tt | 5.363e-02 |
| 9 | fb_dEtat2j | 4.903e-02 |
| 10 | fb_MET_Soft | 4.737e-02 |
| 11 | fb_Asy_EH | 4.625e-02 |
| 12 | fb_Mll | 4.447e-02 |
| 13 | fb_frac_MET_MeffInc_40 | 4.317e-02 |
| 14 | fb_eta_jet2 | 4.282e-02 |
| 15 | fb_eta_jet1 | 4.229e-02 |
| 16 | fb_transSphericity | 4.140e-02 |
| 17 | fb_dRMax_jets | 4.086e-02 |
| 18 | fb_frac_MET_Meff | 4.007e-02 |
| 19 | fb_m_jet1 | 3.421e-02 |
| 20 | fb_nJets30 | 3.260e-02 |

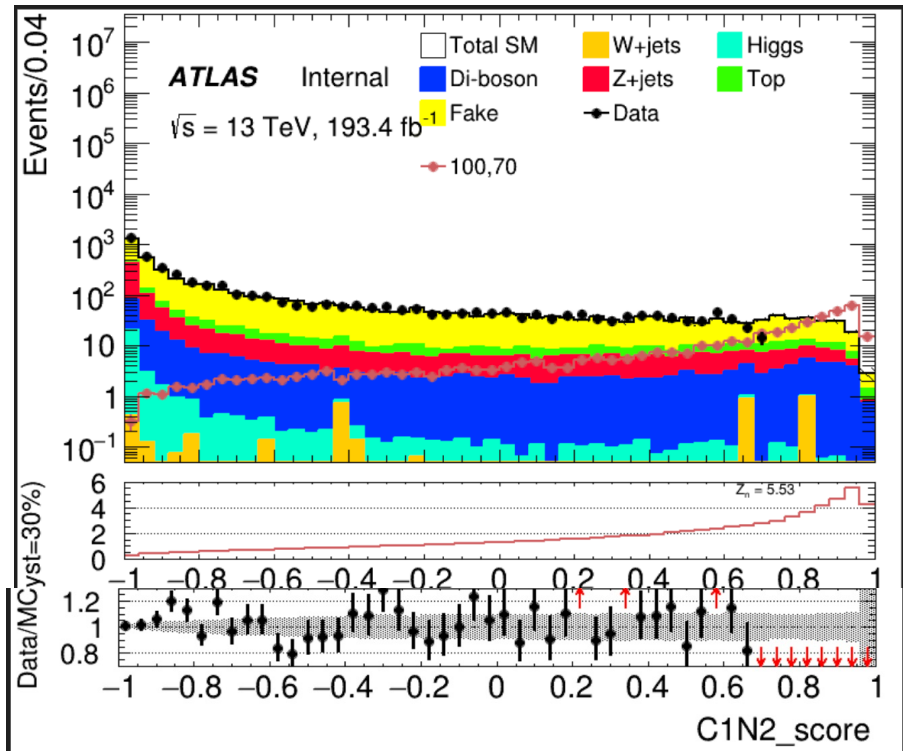
Weight choose: no weight, abs(weight)

No weight have better performance
but abs(weight) fit our anaysis requirment

Split strategy: Separate entries by using mod 5, for Fake bkg, if separate follow sequence, all weighted entry will split into first fold

# SR(HH) Binary

50 bins




Each index stand for a region, and index 10 stands for sum

# Zjets(HH) Binary

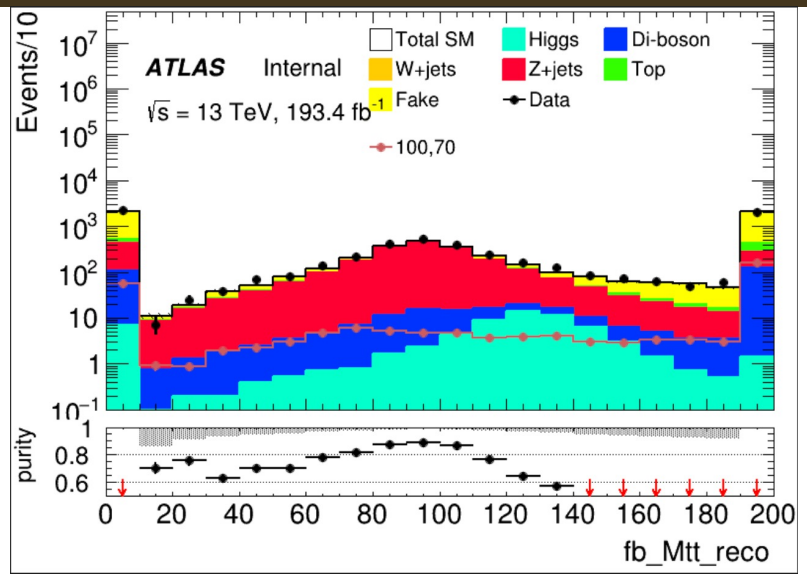
== 2 medium tau
== 0 lepton
METtrig && MET >= 200
OS
bVeto
nBaseJet >= 1
Jet pt > 100GeV

Same with pre-selection

C1N2 score <= 0.7(Othrogonal with SR)

CR: Mtt reco >= 80 && Mtt reco  <= 110
VR: (Mtt reco >= 40 && Mtt reco < 80) || (Mtt reco > 110 && Mtt reco < 130)

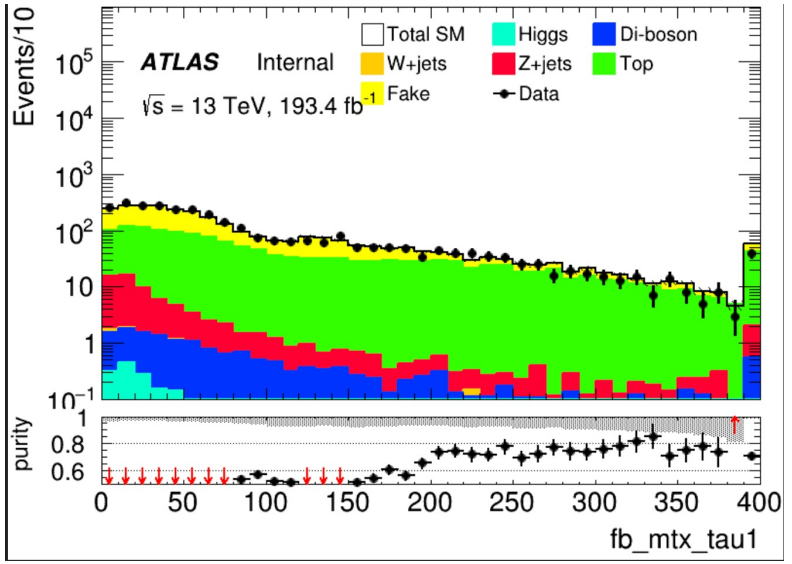
| Region | TotalBkg | Zjets | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 1420.2+-10.9304 | 1221.51+-5.434 | 0.860 | 1559 | |
| VR1 | 821.998+-9.435 | 673.277+-4.073 | 0.819 | 904 | |
| VR2 | 465.375+-7.800 | 320.194+-2.908 | 0.688 | 523 | |



# Top(HH) Binary

== 2 medium tau
== 0 lepton
METtrig && MET >= 200
OS
nBaseJet >= 1
Jet pt > 100GeV
Mtt_reco < 40GeV && Mtt_reco > 130GeV

Same with pre-selection




>= 1 bTag(improve top events)
C1N2 score <= 0.7(Othrogonal with SR)

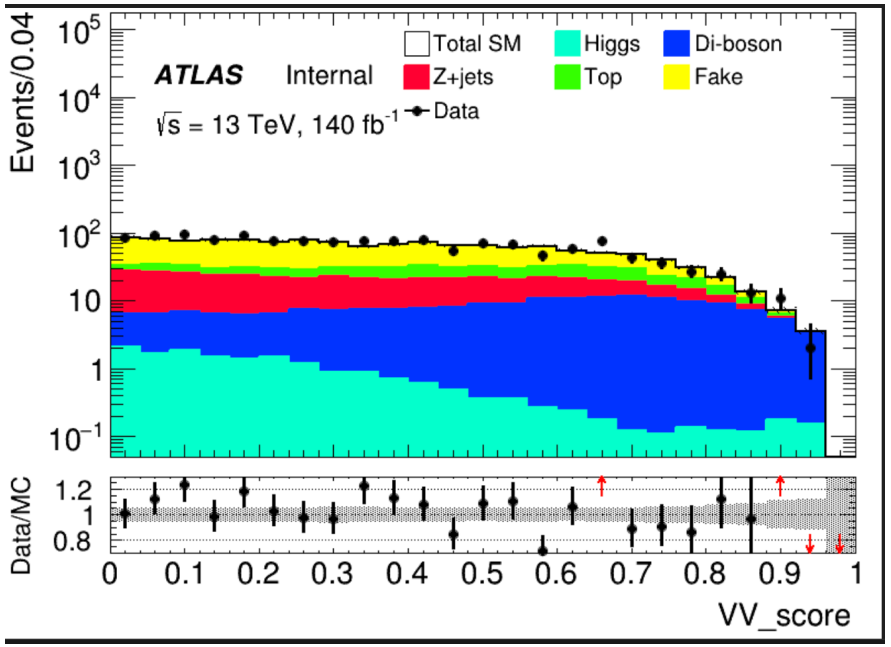
CR: $M_T(\tau, MET) > 250$
VR: $200 < M_T(\tau, MET) < 250$

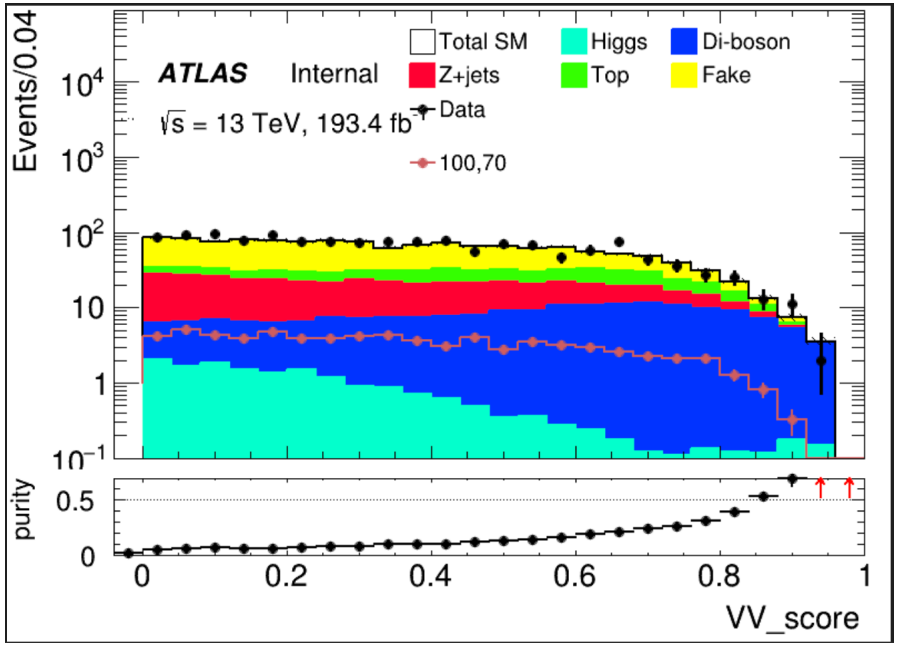
Huge overestimation

| Region | TotalBkg | Top | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 330.534+-7.2017 | 244.813+-4.875 | 0.740 | 261 | 0.7889 |
| VR | 195.634+-6.232 | 143.499+-4.036 | 0.733 | 216 | 1.096 |

# VV(HH) Binary

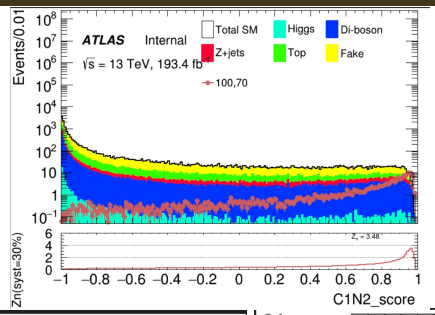
Pre-selection && C1N2_score <= 0.7 && VV_score >= 0.80



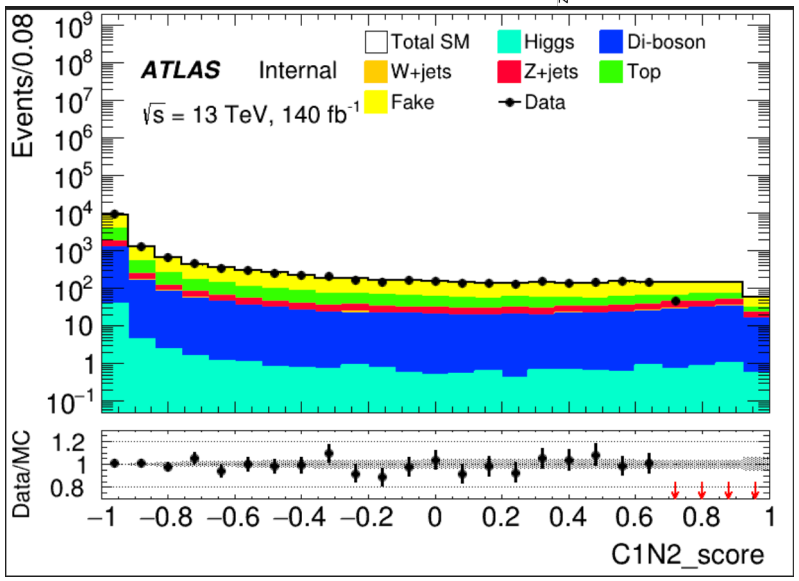




| Region | TotalBkg | VV | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| VR | 46.6004+-2.77 | 24.193+-0.816 | 0.519 | 51 | 1.108 |

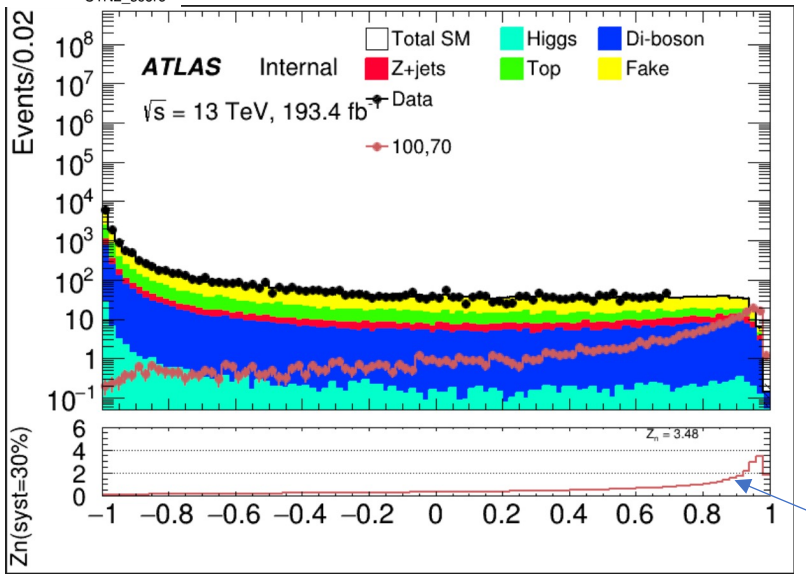
# SR(LH) Binary



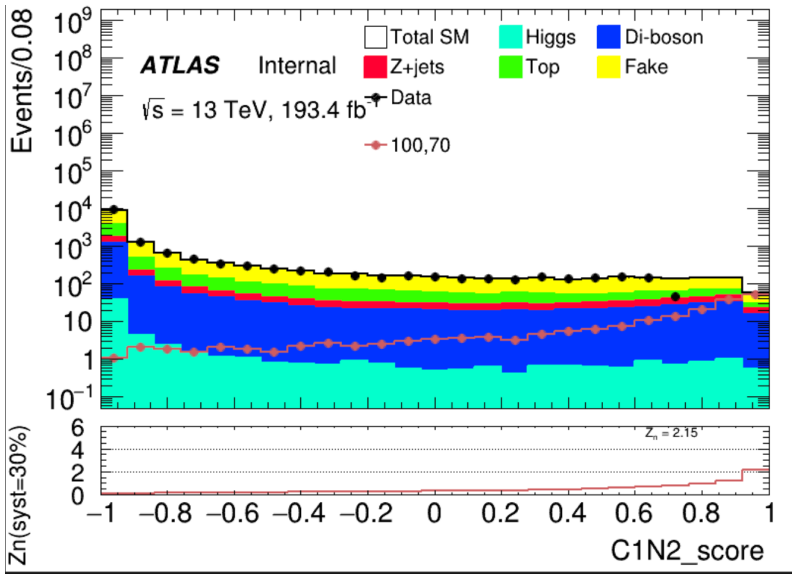

25 bins




100 bins




25 bins




Narrow peak

Each index stand for a region, and index 10 stands for sum

# Zjets(LH) Binay

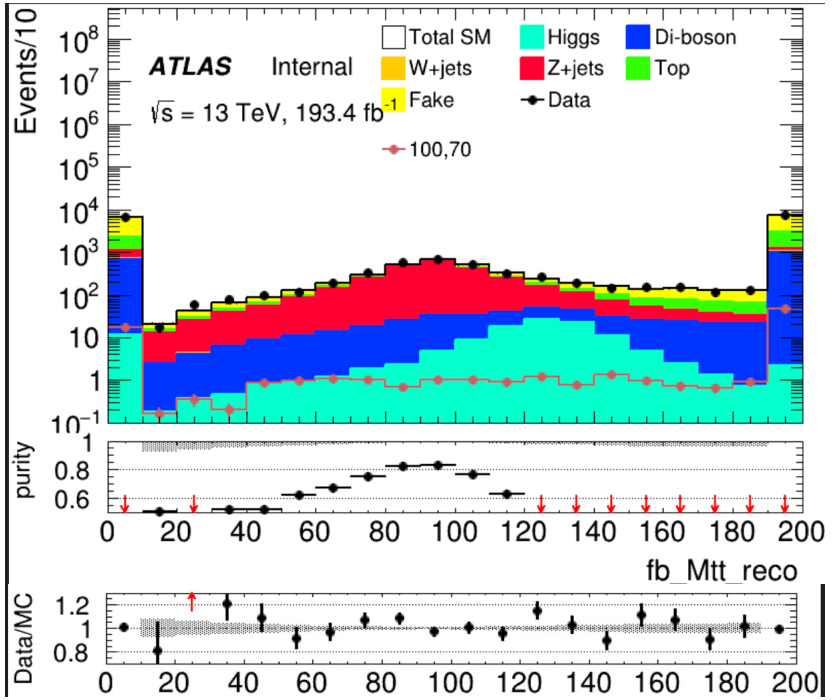
== 1 medium tau
>= 1 lepton
METtrig && MET >= 200
OS
bVeto
nBaseJet >= 1
Jet pt > 100GeV

Same with pre-selection

C1N2 score <= 0.7(Othrogonal with SR)

CR: Mtt reco >= 80 && Mtt reco  <= 110
VR: (Mtt reco >= 40 && Mtt reco < 80) || (Mtt reco > 110 && Mtt reco < 130)

| Region | TotalBkg | Zjets | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 1420.2+-10.9304 | 1221.51+-5.434 | 0.860 | 1559 | |
| VR1 | 821.998+-9.435 | 673.277+-4.073 | 0.819 | 904 | |
| VR2 | 465.375+-7.800 | 320.194+-2.908 | 0.688 | 523 | |

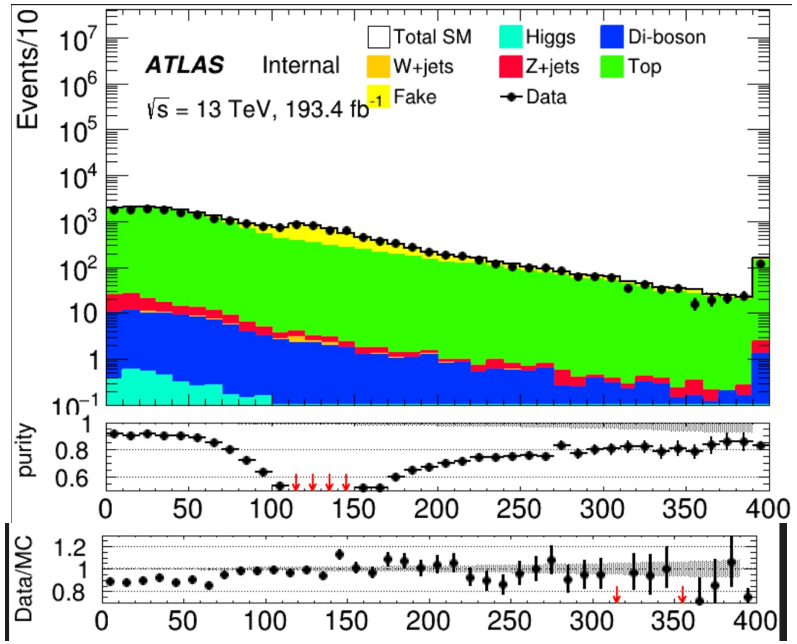


# Top(LH) Binary

== 1 medium tau
>= 0 lepton
METtrig && MET >= 200
OS
nBaseJet >= 1
Jet pt > 100GeV
Mtt_reco < 40GeV && Mtt_reco > 130GeV

Same with pre-selection

>= 1 bTag(improve top events)
C1N2 score <= 0.7(Othrogonal with SR)



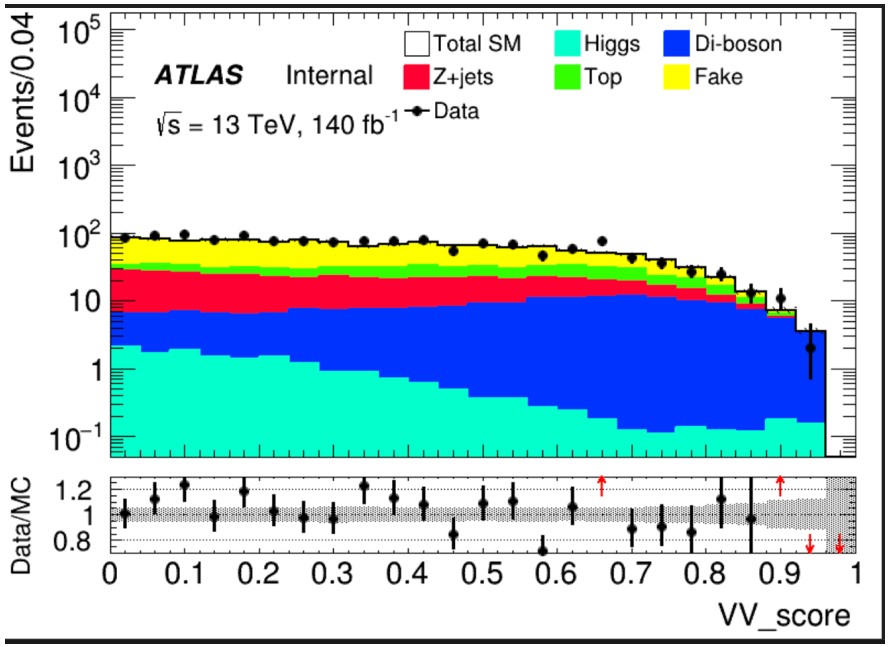

CR: $M_T(\tau, MET) < 100$
VR: $200 < M_T(\tau, MET) < 250$

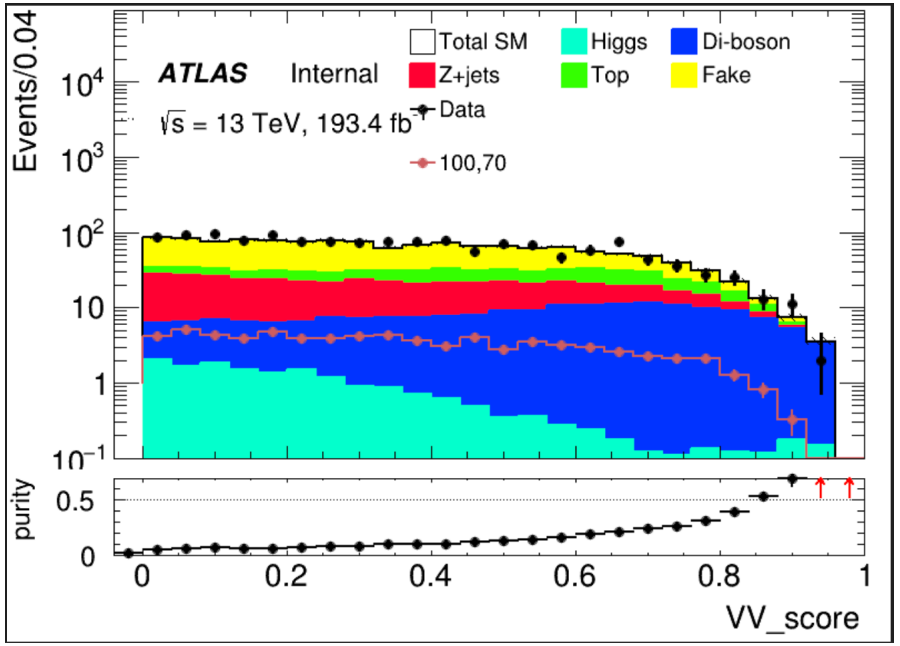
Huge overestimation

| Region | TotalBkg | Top | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| CR | 16211.7+-52.8677 | 13843.8+-39.5036 | 0.8539 | 14766 | |
| VR | 865.092+-11.2306 | 633.439+-8.38922 | 0.7332 | 836 | |

# VV(LH) Binary

Pre-selection && C1N2_score <= 0.7 && VV_score >= 0.80



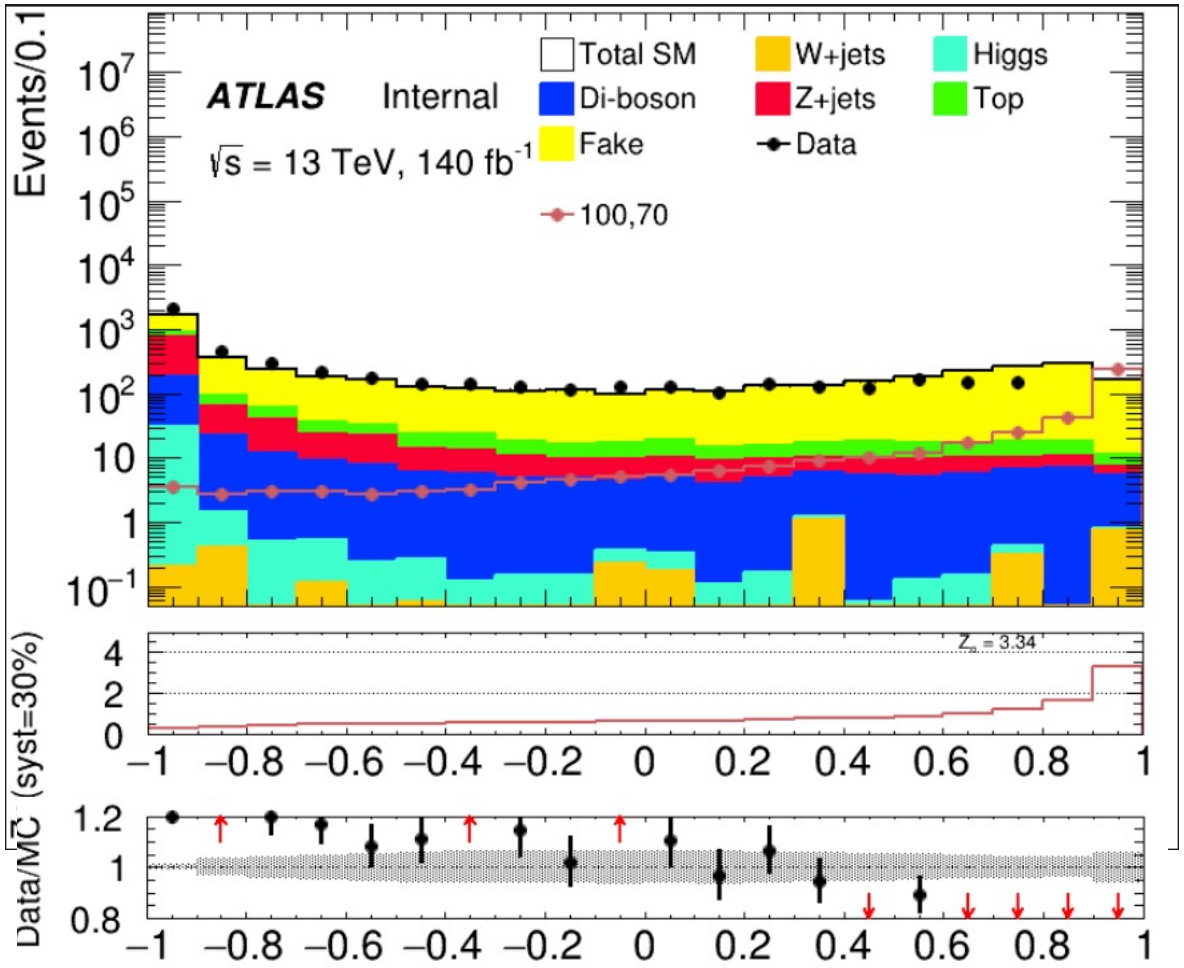




| Region | TotalBkg | VV | purity | Data | Data/Bkg |
|---|---|---|---|---|---|
| VR | 78.6819+-2.26 | 53.8201+-1.124 | 0.68402 | 72 | |

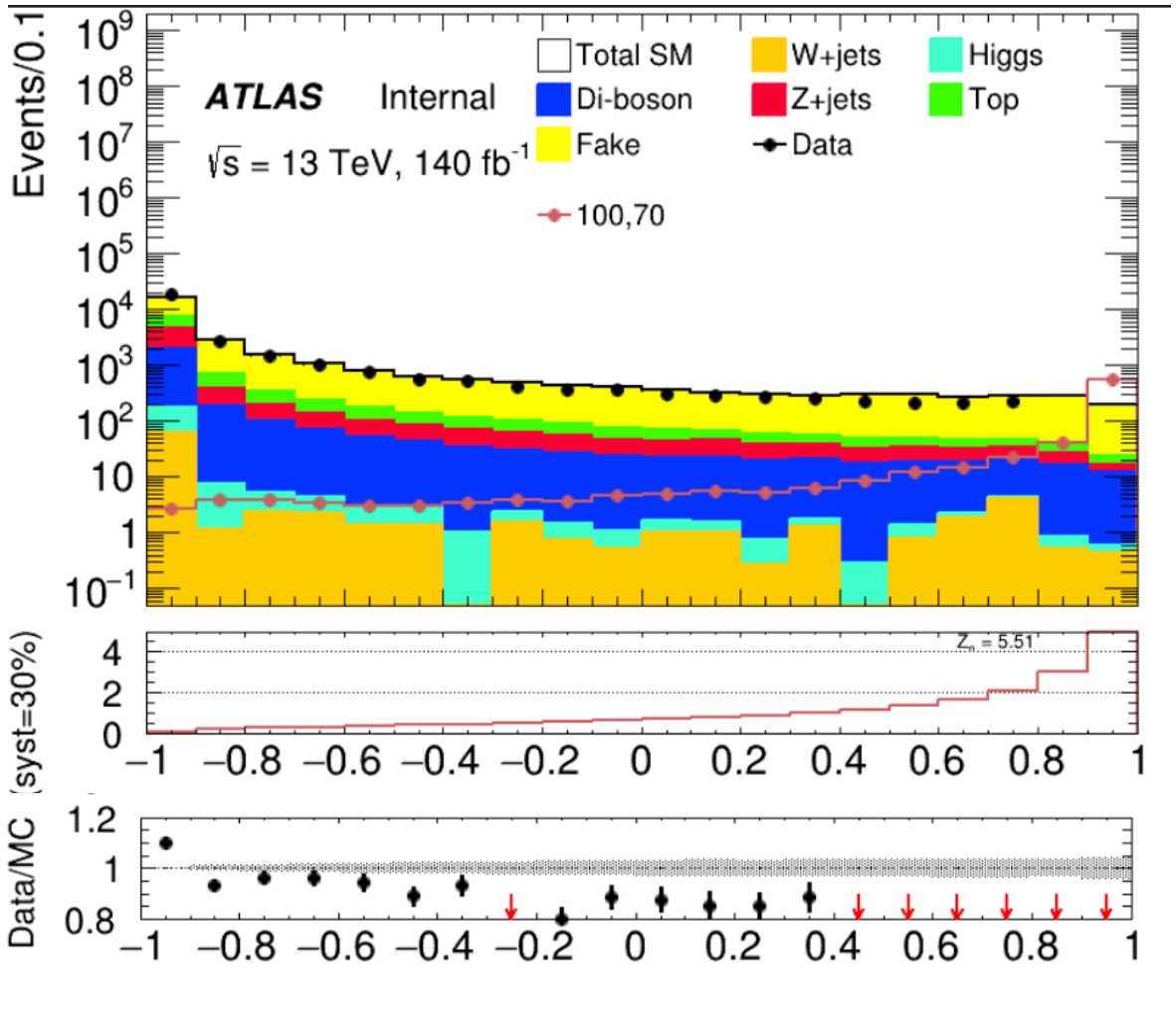
# Backup

# SR

HH channel




LH channel



# Bkg decay mode

Wjets: W->e/muon + nu
W->tau+nu(can contribute true tau_had)
jet misidentified to a fake tau

Zjets: Z->ll/tautau
jet misindentified to fake tau

Top: top->W+b, W can contribute a true tau_had
b-quark is a source of fake

VV: W/Z

LH channel: $\geq 1tau,\ \geq 1lep$
Wjets: W contribute lep, jets misidentified to fake
Zjets:
SingleTop: W contribute lep, b-quark misidentified to fake
VV:

HH channel: $\geq 2tau,\ == 0lep$
Wjets: W contribute tau_had, plus a fake tau
Zjets: Z->tautau(had) or 2 fake tau
SingleTop: W contribute a tau_had, plus a fake tau
VV: