MACHINE LEARNING TECHNIQUES FOR MUON IDENTIFICATION AND ISOLATION AT CMS

Andrea Trapote Fernández

CPAN 2023

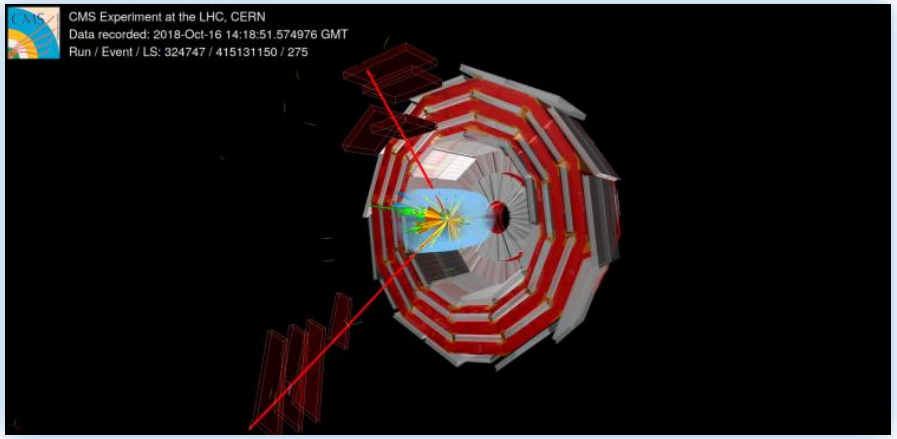
Contents

Two independent multivariate (MVA) techniques are developed to improve **muon identification.**

- **GENERAL MUON MVA ID:** to improve the current cut-based ID Working Points (WPs). The aim is to improve the **general muon identification** efficiency with a low misidentification rate.

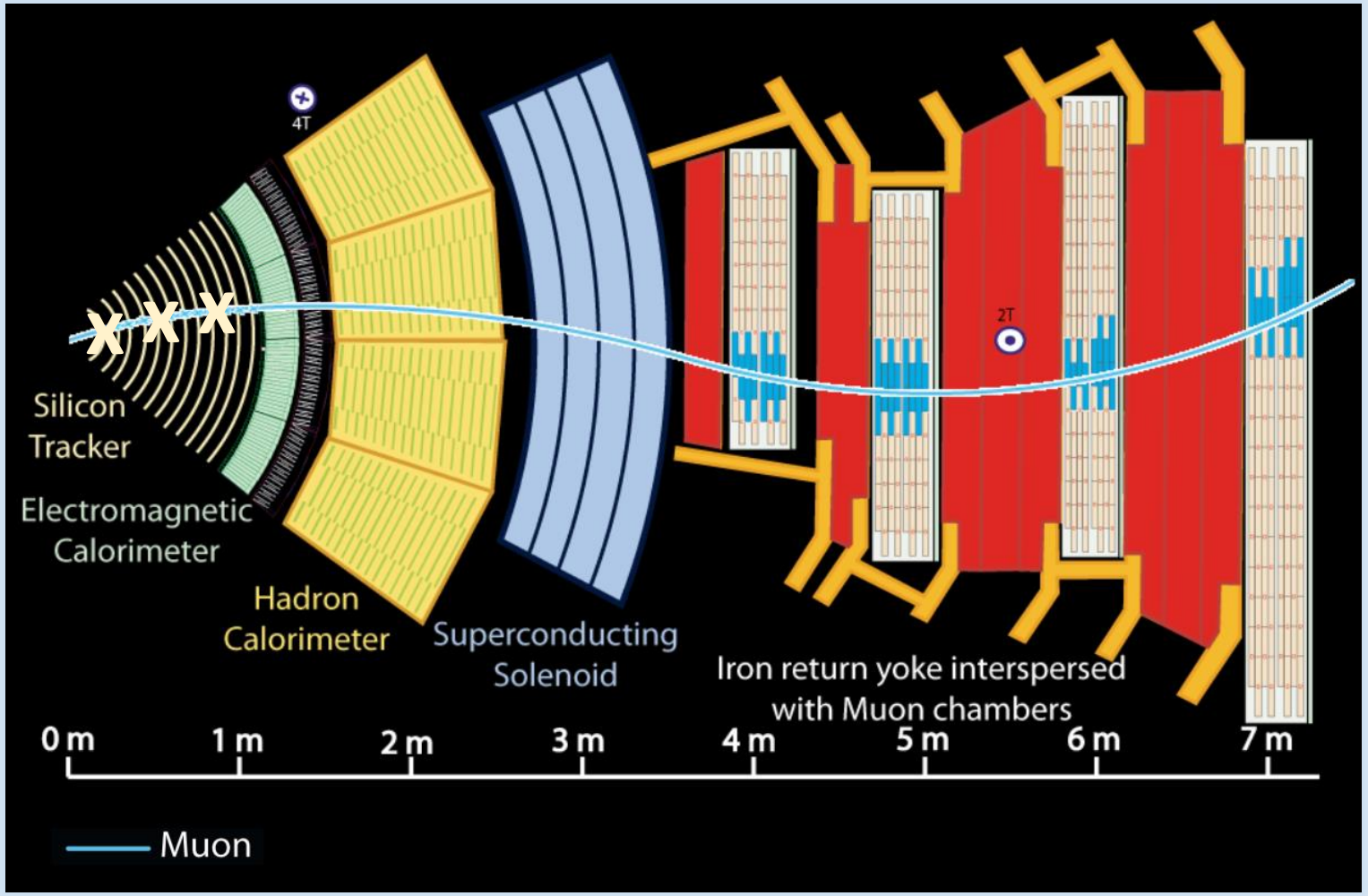
- **PROMPT MUON MVA:** to select **prompt muons** at analysis level. This MVA aims to improve the selection of prompt muons, arising from the decay of a W, Z, H bosons or τ leptons, and to reduce the contamination of muons from other sources.

These MVAs have been documented in a recent publication: CMS-PAS-MUO-22-001, (soon to be sent to JINST).



An event candidate for the ttH production, recorded in 2018 and identified using Machine Learning techniques [Physics briefing].

MUON RECONSTRUCTION



Standalone muons:
only use information from the muon system.

Global muons:
standalone muon propagated to the tracker.

Tracker muons:
tracker track propagated to the muon system.

1) General muon MVA ID

INTRODUCTION

Goal: Optimize the **standard muon identification** (cut-based ID) using a **Machine Learning** algorithm to discriminate spurious muons and instrumental backgrounds.

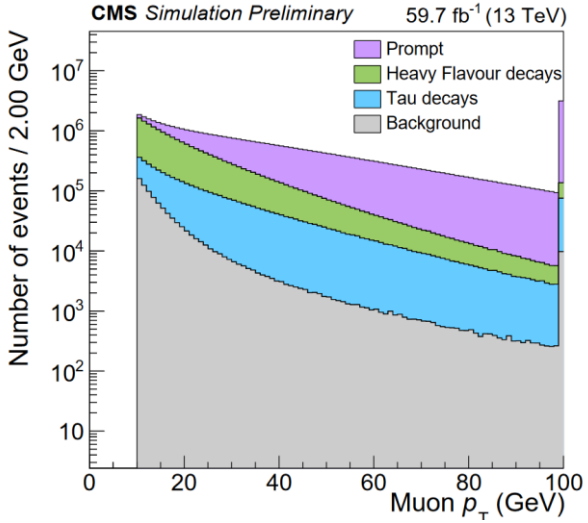
- **Selection**: muons with $p_T > 10$ GeV are used as initial selection from a **2018 $t\bar{t}$ sample**.
- **Classification**: muons are then classified in two categories according to its origin in 'GOOD' (signal) and 'BAD' (background) muons.
- **Inputs**: the ones used to define the Medium and Tight selections in the **cut-based ID** with the exception of the impact parameters (next slide).
- **Output**: probability of a muon to be a signal muon.
- **Model**: random forest trained with Scikit-learn.

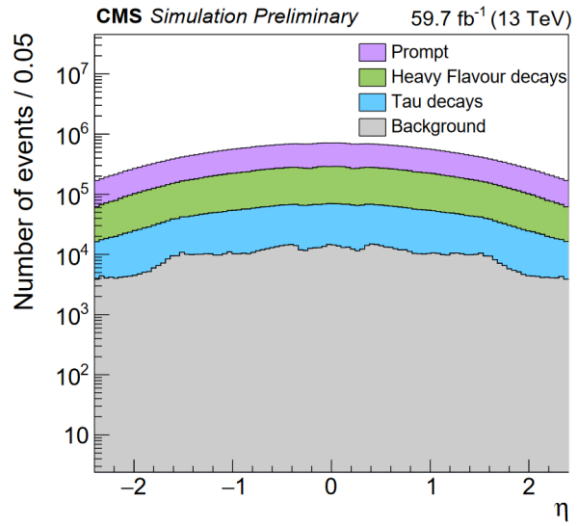
INPUT VARIABLES

12 input variables taken from the Medium and Tight WPs definition of the cut-based ID:

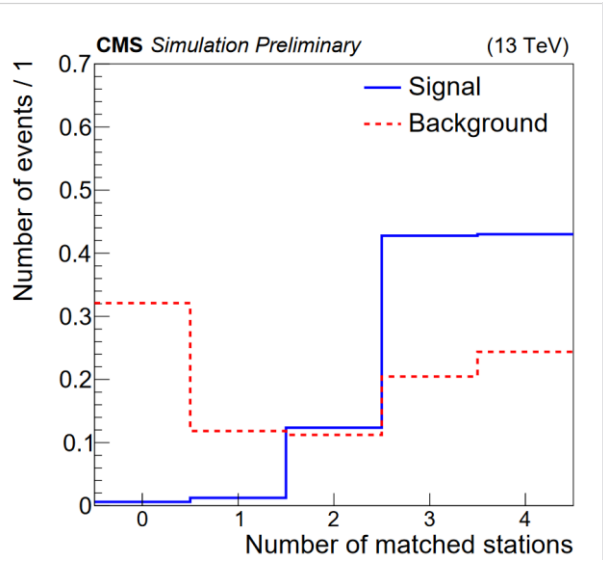
- Muon p_T and η: we are reweighing these variables to not introduce correlation.
- Segment compatibility.
- χ^2 from the kink-finder algorithm on the inner track.
- ~~Transverse impact parameter with respect to the primary vertex (dz).~~
- ~~The longitudinal distance of the tracker track wrt. the primary vertex (dxy).~~
- Tracker-standalone position match (local χ^2).
- Normalized χ^2 of the muon track fit.
- Fraction of valid tracker hits.
- Number of valid pixel hits.
- Number of tracker layers with hits.
- Number of muon stations with muon segments.
- Number of muon-chambers hits included in the global-muon track fit.
- Flag: is global muon.

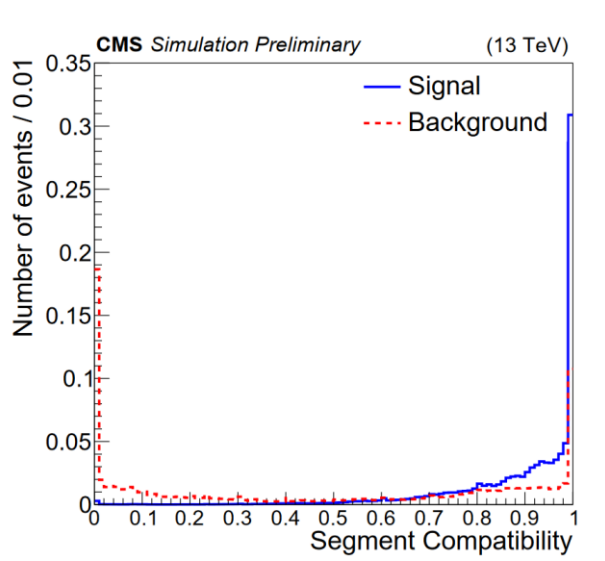
Impact parameters (IP) are **not** included to have a more general training.

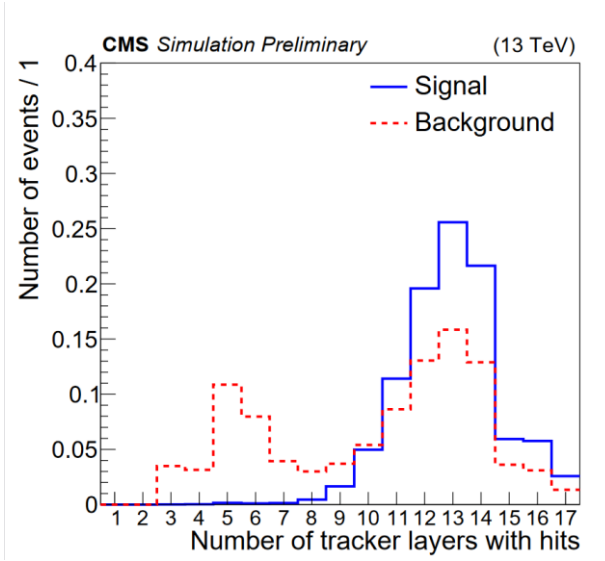


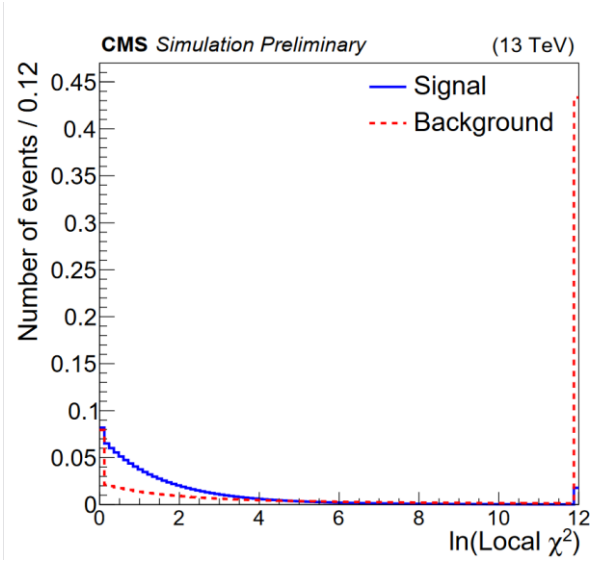


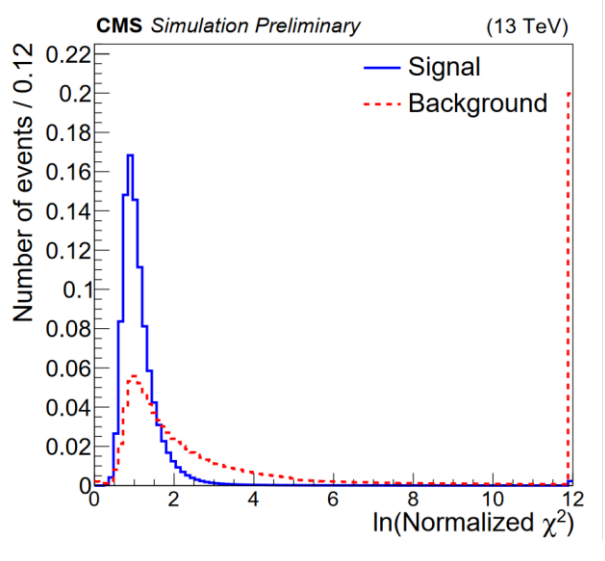
INPUT VARIABLES

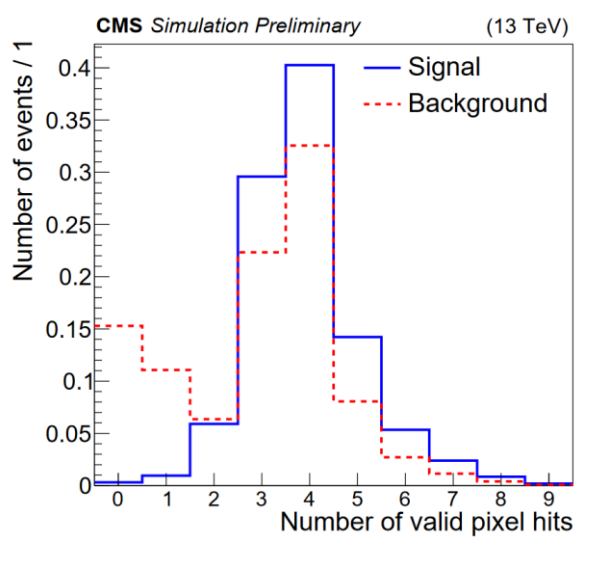


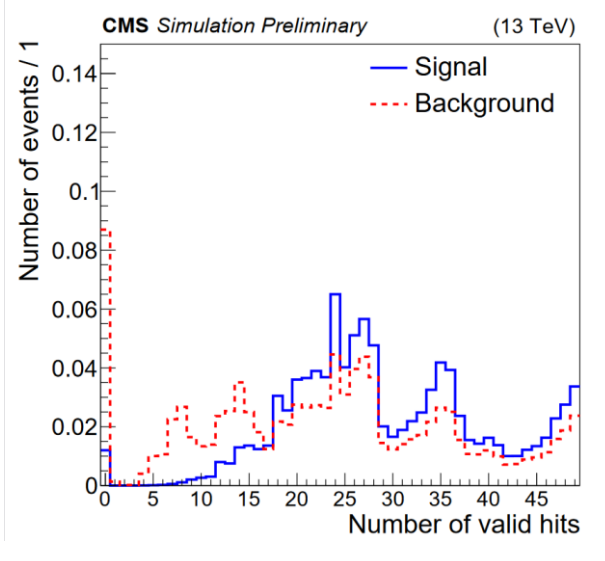


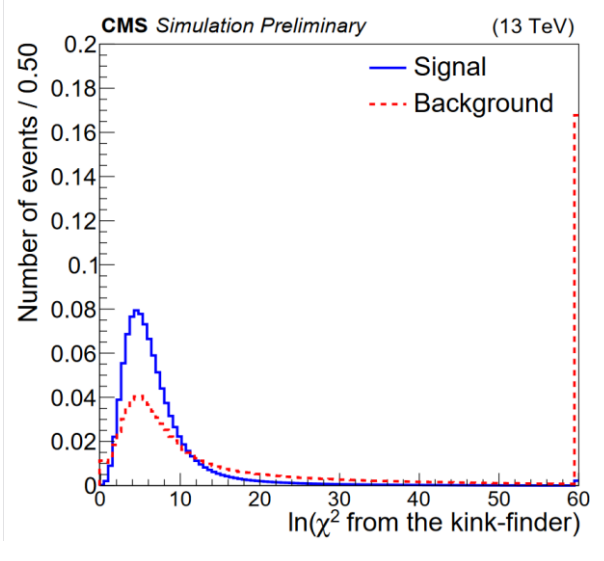








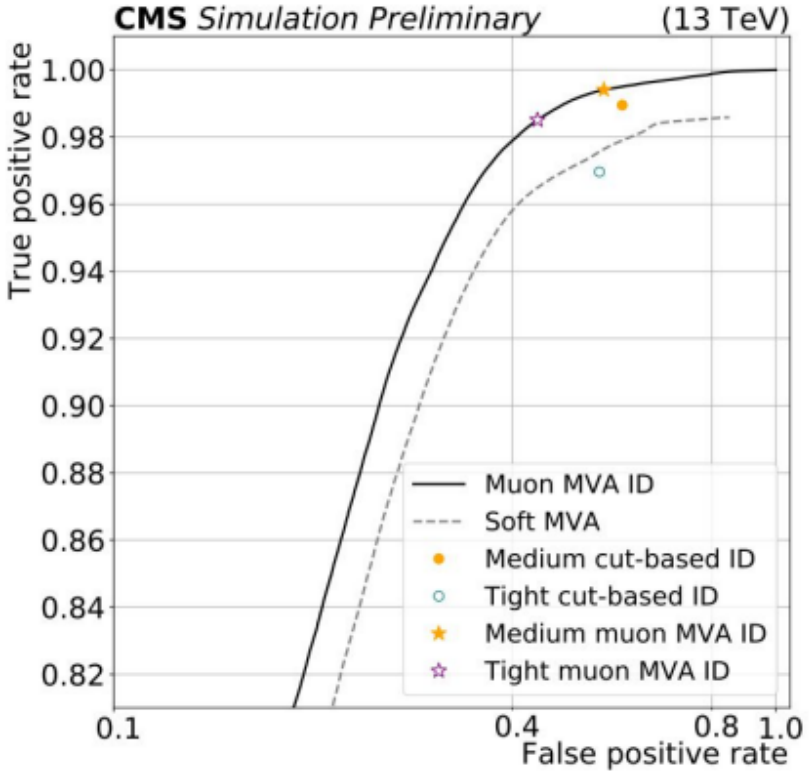




➢ Good discrimination between signal and background in some variables.

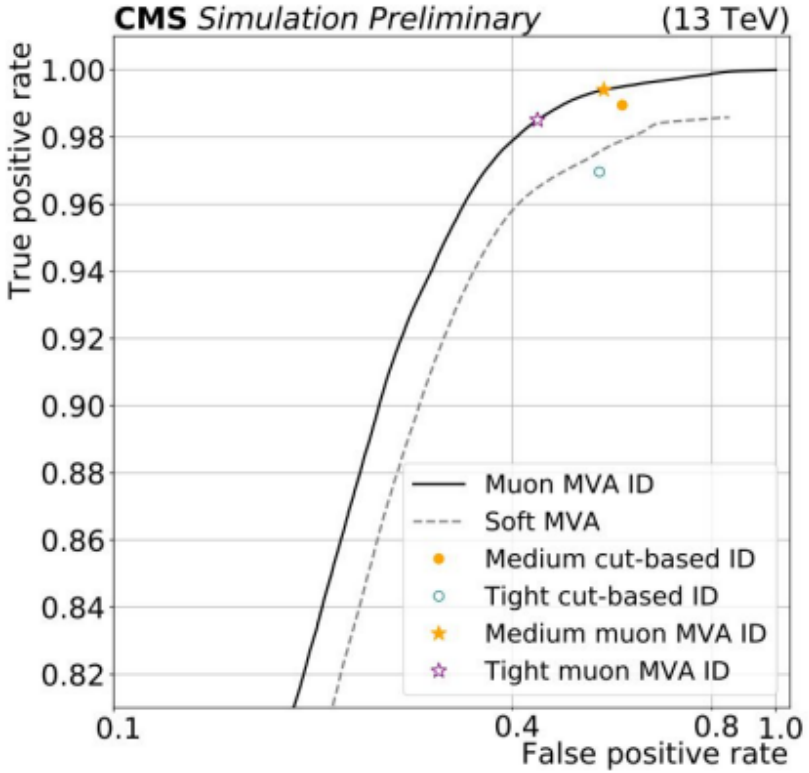
MUON MVA ID PERFORMANCE

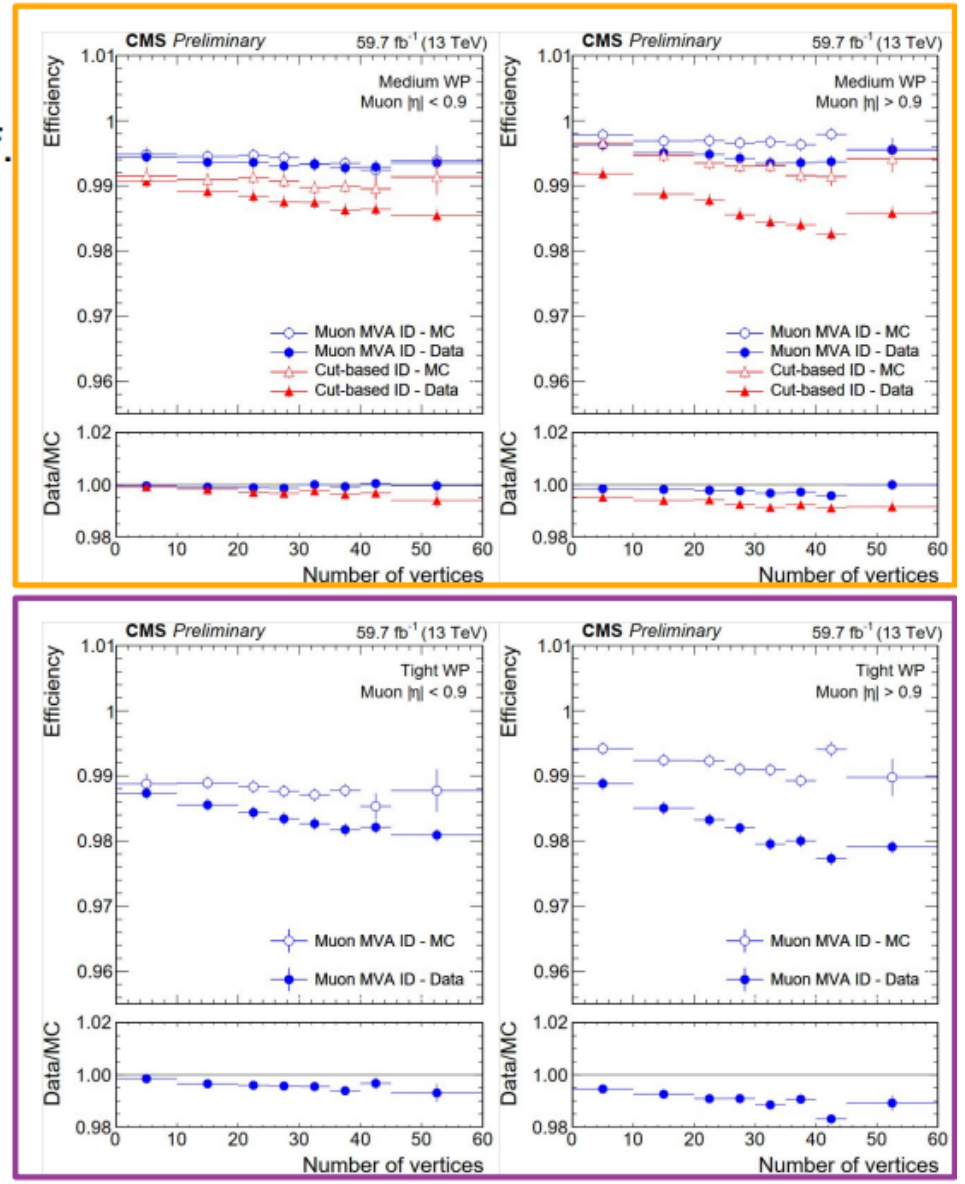
- **Good performance achieved with the MVA, promising for Run 3!!**



MUON MVA ID PERFORMANCE

➢ **Good performance achieved with the MVA, promising for Run 3!!**





- **Medium MVA WP**: same background contamination as the medium cut-based WP with 0.5-1% higher efficiency.
- **Tight MVA WP**: achieves a 10% smaller background contamination than the medium MVA ID and the efficiency is about 99% .
- MVA ID is **more stable as a function of PU** than the cut-based ID .

2) Prompt muon MVA

INTRODUCTION

Goal: Select **prompt muons** araising from the decay of a W, Z, H bosons or τ leptons and to reduce the contamination of muons from other sources.

- **Prompt muon MVA broadly used** in CMS analyses: searches for supersymmetry, standard model precision measurements, studies of the top quark properties and measurements in the Higgs boson sector.

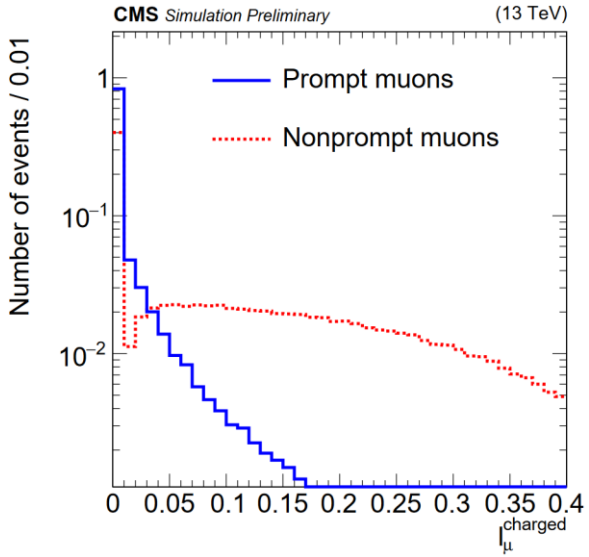
- **Selection**: muons with $p_T > 5$ GeV are used as initial selection from **2018 $t\bar{t}/t\bar{t}$H samples**.

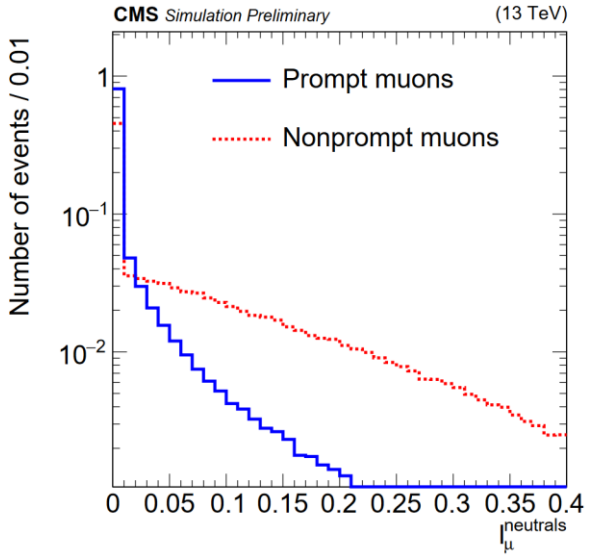
- **Inputs**: muon identification and isolation variables.

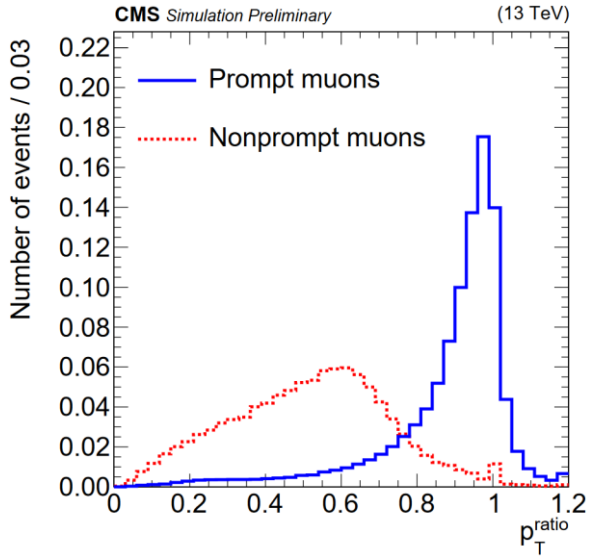
- **Classification**:
  - **Signal**: muons form $t\bar{t}$**H** matched to a generator-level muon produced in the prompt decay of a H, W, Z or τ.
  - **Background**: muons form $t\bar{t}$ not matched to a generator-level muon produced in the prompt decay of a H, W, Z or τ.

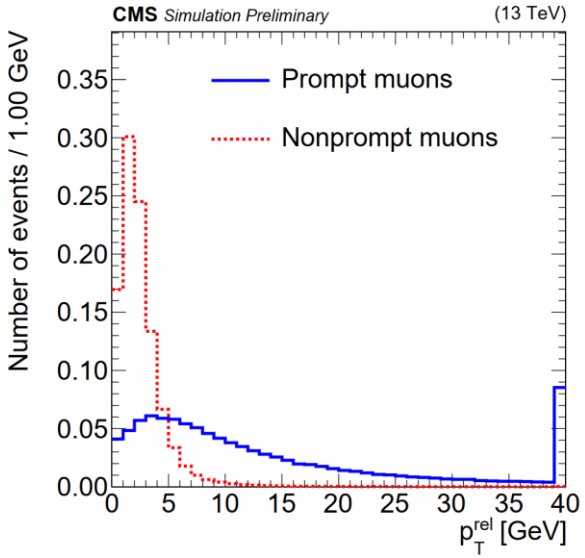
- **Model**: BDT trained with TMVA.

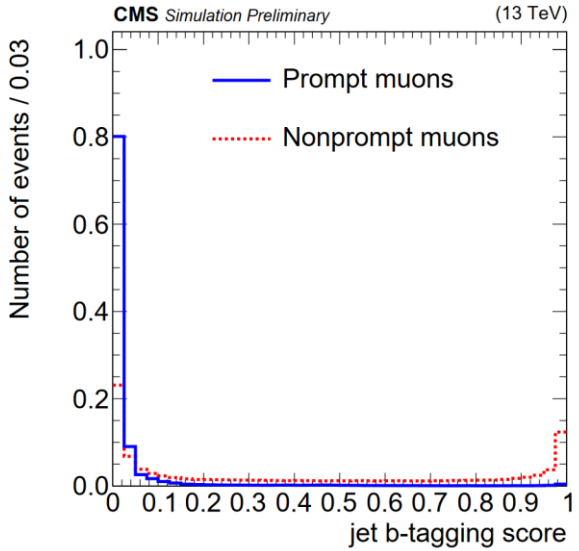
INPUT VARIABLES

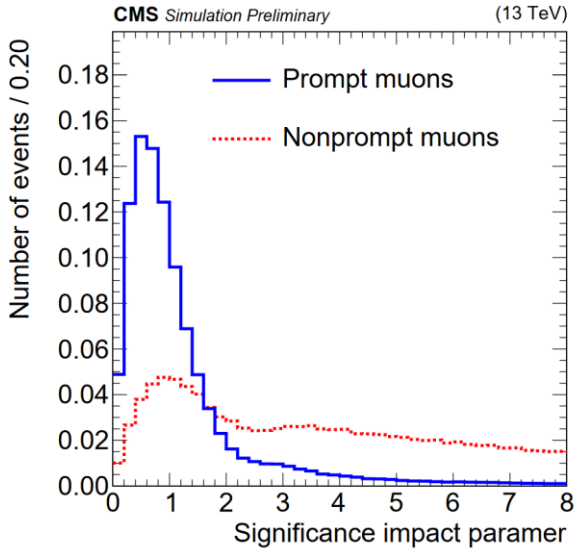


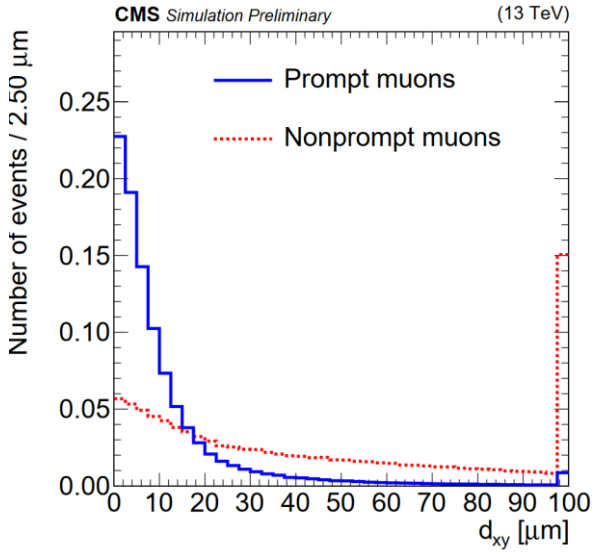


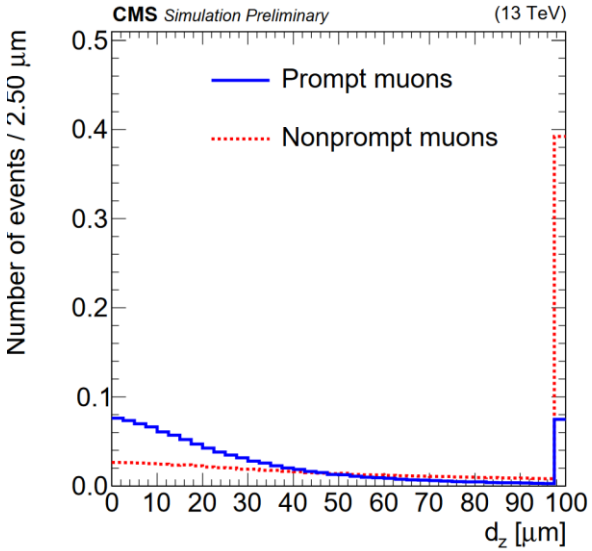






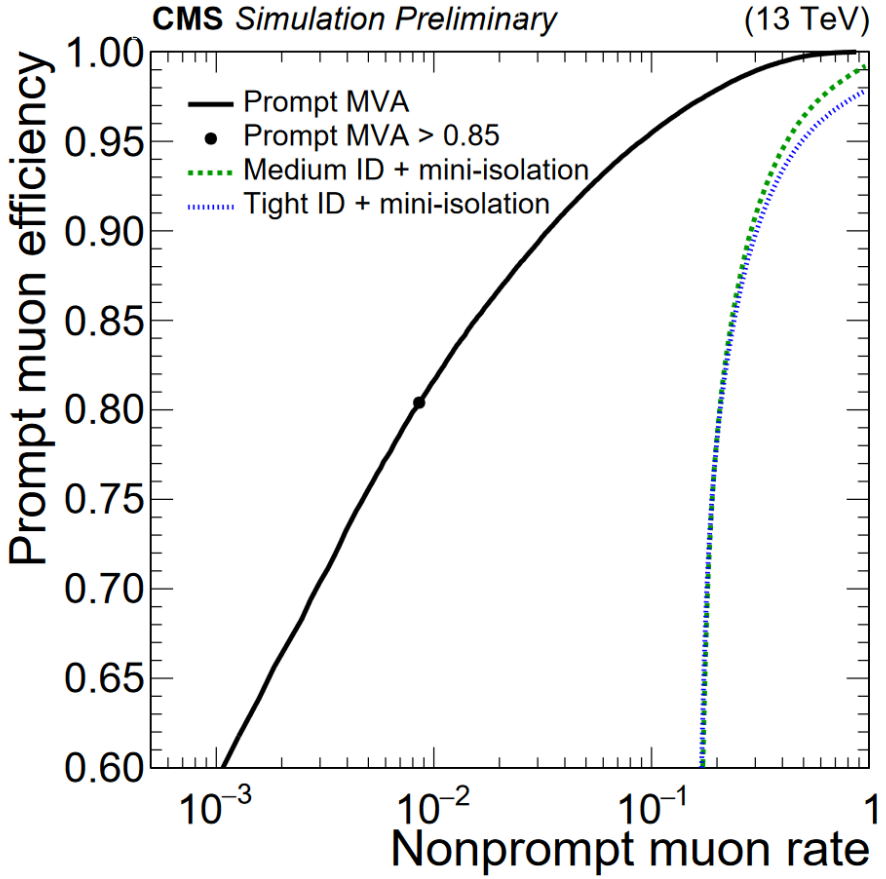






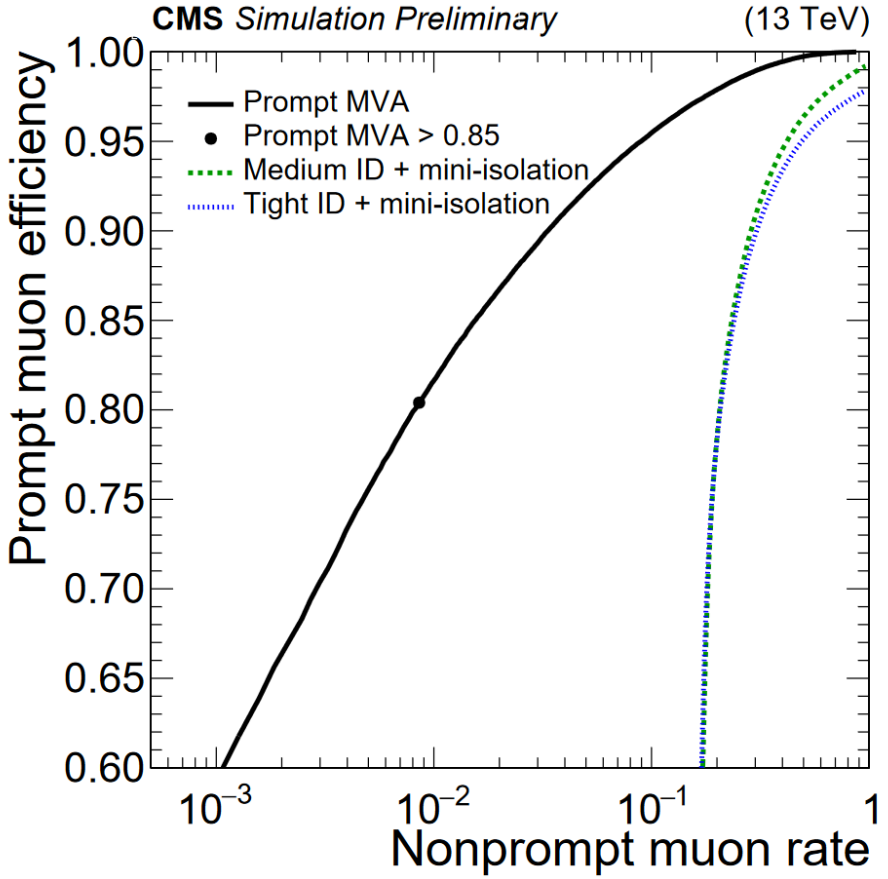
PROMPT MUON MVA EFFICIENCY

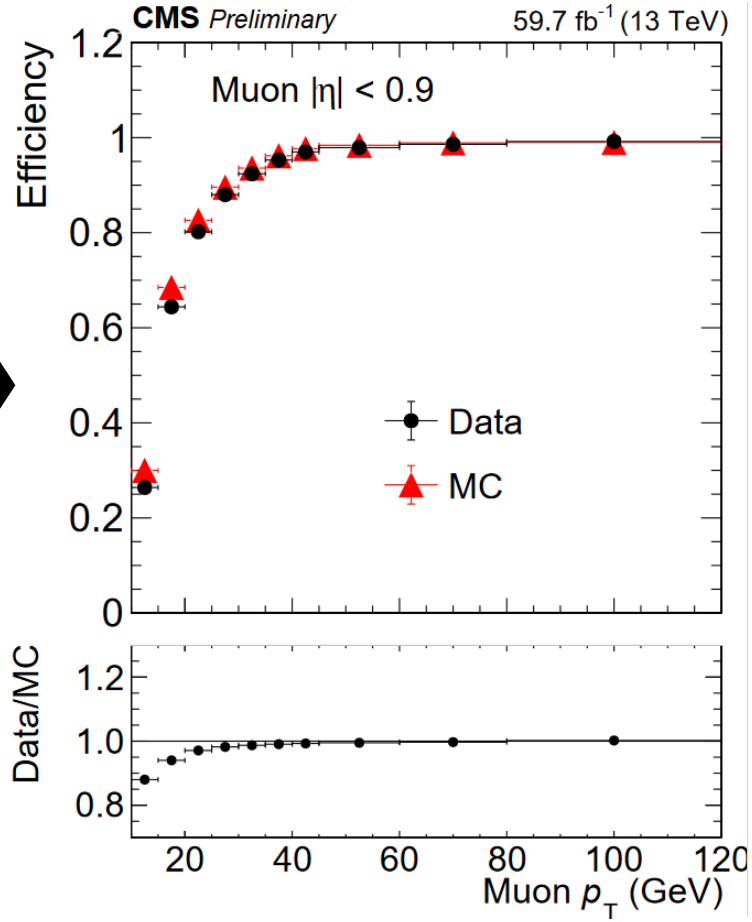
- **Working point defined as MVA > 0.85.** It was optimized for a $t\bar{t}$H measurement.

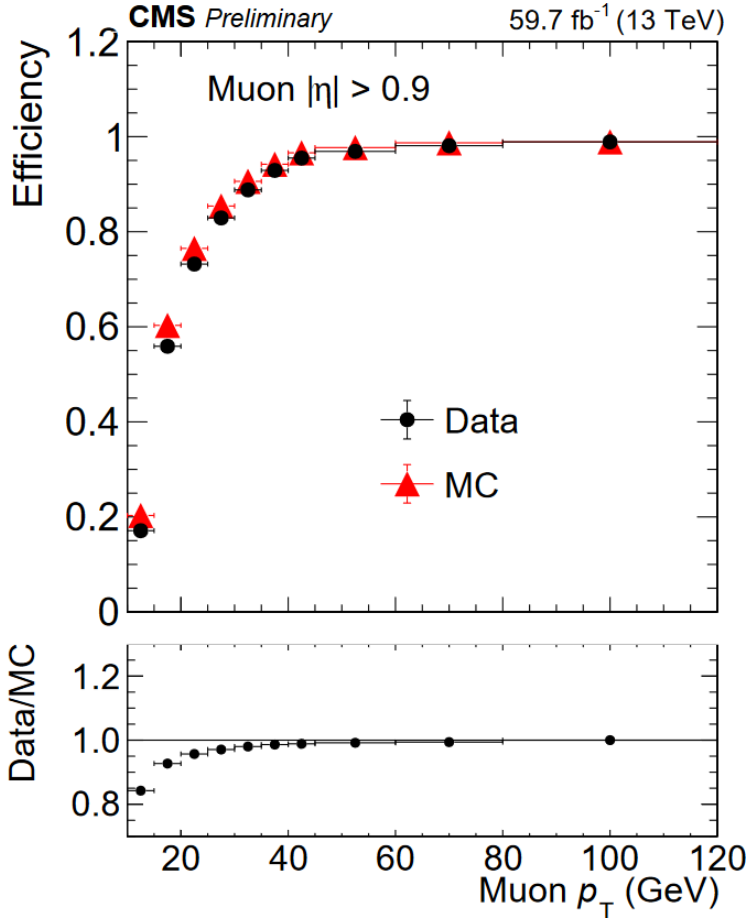


PROMPT MUON MVA EFFICIENCY

➢ **Working point defined as MVA > 0.85.** It was optimized for a $t\bar{t}$H measurement.







- Efficiency **higher than 80%** for muon with p_T > 20 GeV.
- Discrepancies between data and MC smaller than 3% above 20 GeV.

NONPROMPT RATE: MEASUREMENT PROCEDURE

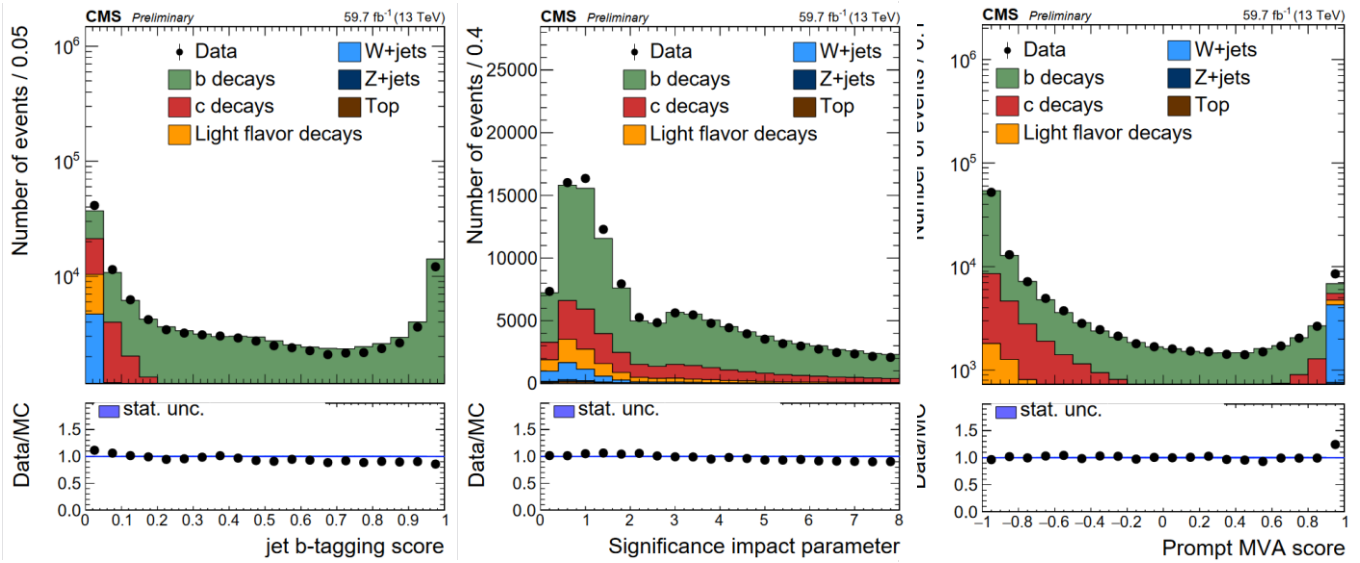
The nonprompt rate is measured in a sample enriched in **multijet events.**

- **Selection:**
  - Pass non-isolated single muon trigger.
  - Pass probe selection and medium cut-based ID.
  - Jet recoiling against the muon with a $\Delta R > 0.7$.

- In order to subtract possible contributions from EW processes ($t\bar{t}$ and W-Jets), a fit to m_T^{fix} is used:

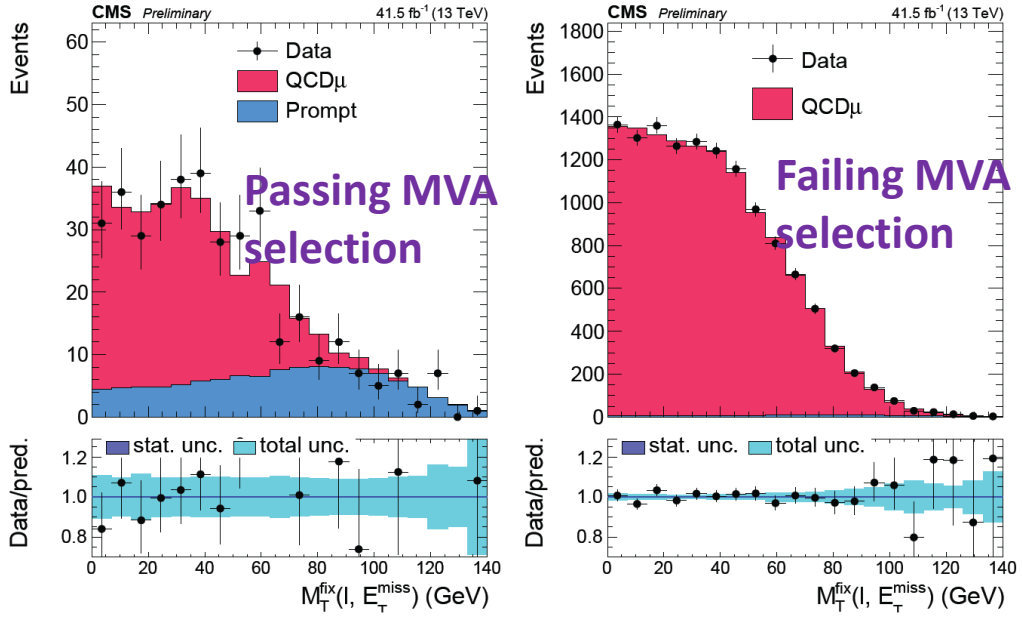
$$m_T^{fix} = \sqrt{2\, p_T^{fix} p_T^{miss} (1 - \cos\Delta\phi)}$$

$$p_T^{fix} = 35\text{GeV}$$



- **Multijet and EW contributions** to data are parametrized using templates that are derived from samples of simulated events of each.

- **Nuisance parameters** added to the fit:
  - Statistical uncertainties in the templates.
  - Systematic deformations of their shapes.



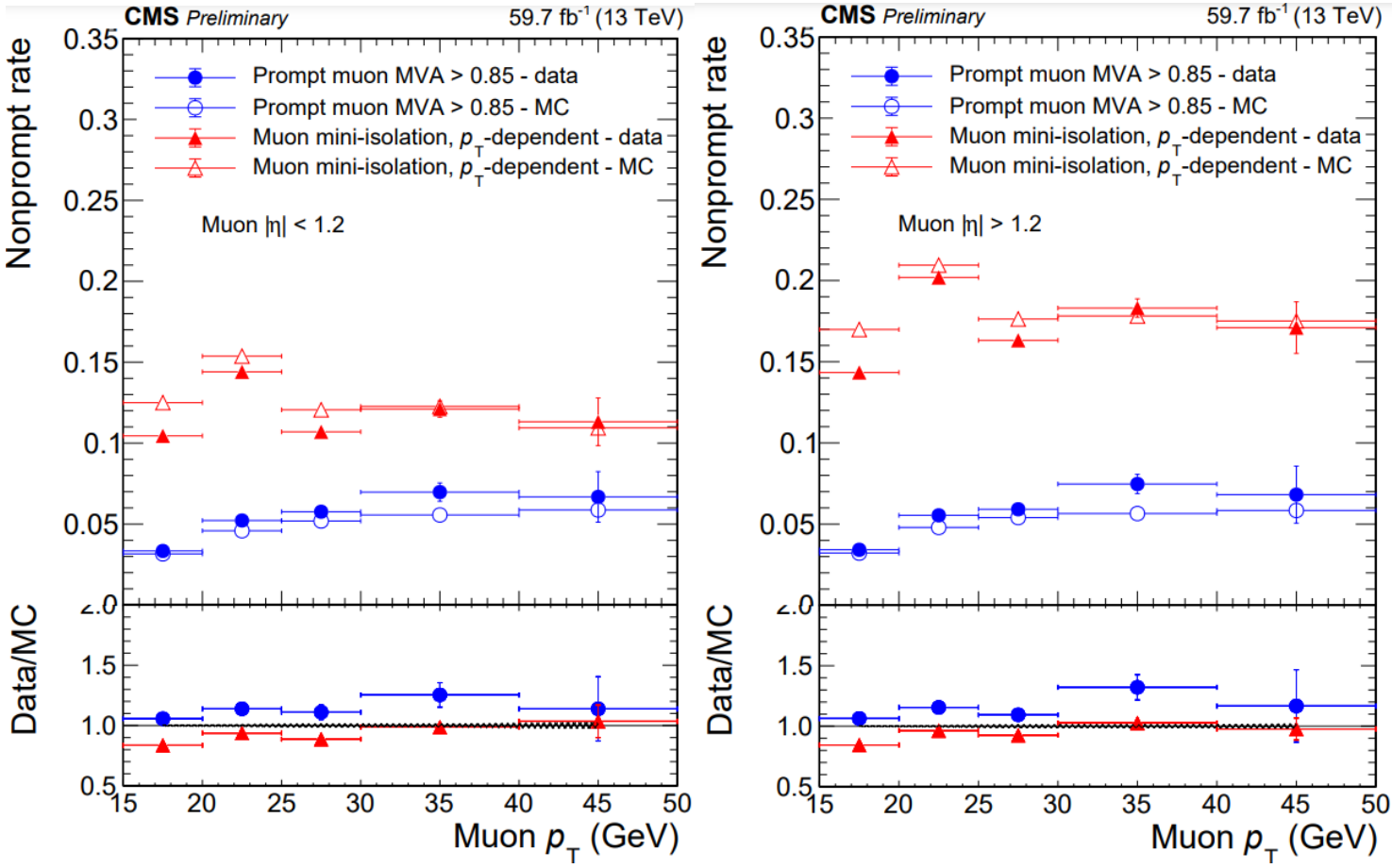
NONPROMPT RATE: RESULTS

- The fit is performed separately for cases in which the muon **passes or fails** a MVA > 0.85.
- **Nonprompt rate** is defined as:

$$f = \frac{N_{\text{pass}}}{N_{\text{pass}} + N_{\text{fail}}}$$

- The nonprompt rate with a selection of **mini-isolation** is also computed to compare.

- Since the prompt MVA and the mini-isolation have a different **efficiency dependence** as a function of the muon p_T , we consider a requirement on mini-isolation that gives the same efficiency as the prompt MVA.



- The MVA shows **a factor 2 (3) smaller nonprompt rate** compared to mini-isolation in $|\eta| < 1.2$ ($|\eta| > 1.2$)

Summary

➢ **Two different multivariate techniques** have been developed. The first one to improve general muon identification efficiency with a low misidentification rate, while the second one targets the selection of prompt muons from bosons.

GENERAL MUON MVA ID

- The signal efficiency has been measured showing **higher efficiencies** than those obtained with the standard cut-based ID.
- Results show extraordinary performance even at high PU and thus ensuring a **great potential for future** leptonic analyses with the CMS experiment with **Run 3 data.**
- Furthermore, it gives a continuous score which offers more flexibility for analyzers.

PROMPT MUON MVA

- The prompt muon MVA, already **used in many Run 2 analyses**, have been documented [CMS-PAS-MUO-22-001].
- The **nonprompt rate achieved with this MVA is a factor 2-3 times smaller** than the one obtained with the standard cut-based ID and isolation.
- **This model is the key to reject fakes in many analyses.**

Back up

Composition

Classification	Matching	$t\bar{t}$ 2018	DY 2018	$t\bar{t}$ Run 3
Background muons	Not matched	669267 (0.69%)	344557 (0.72%)	79266 (2.06%)
	Punchthrough	211541 (0.22%)	23468 (0.05%)	9295 (0.24%)
Signal muons	From boson	58360911 (60.41%)	43296756 (90.51%)	1923782 (49.91%)
	From tau	7411462 (7.67%)	4037901 (8.44%)	259719 (6.74%)
	From B	19375609 (20.06%)	51245 (0.11%)	989074 (25.66%)
	From B to C	7034538 (7.28%)	11519 (0.02%)	362189 (9.40%)
	From C	2537186 (2.63%)	30724 (0.06%)	171967 (4.46%)
	From other light	28158 (0.03%)	520 (0.00%)	1710 (0.04%)
	From PiKppMuX	970196 (1.00%)	36893 (0.08%)	56665 (1.47%)
	From PiKNotppMuX	12652 (0.01%)	462 (0.00%)	728 (0.01%)
Total		96611520	47834045	3854395

SELECTION

- GOOD muons (signal):
  - From boson, from tau, from b, from b to c, From C, From other light.
- BAD muons (background):
  - Not matched, punchthroug. from PiKppMuX, From PiKNotppMuX.

➢ To compare the cut-based ID with the MVA we use the **efficiency** and the **fake rate (FR)**, which are defined as:

$$\text{Efficiency} = \frac{GOOD\ loose\ muons\ \ passing\ WP\ cut}{GOOD\ loose\ muons}$$

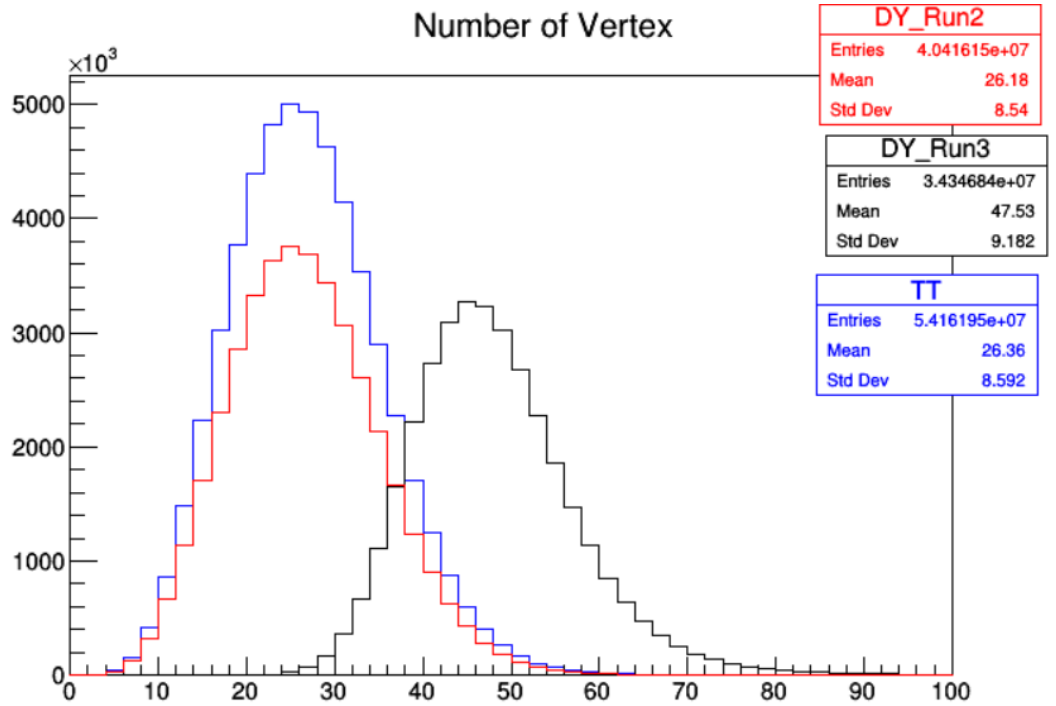
$$Fake\ rate = \frac{BAD\ loose\ muons\ \ passing\ WP\ cut}{BAD\ loose\ muons}$$

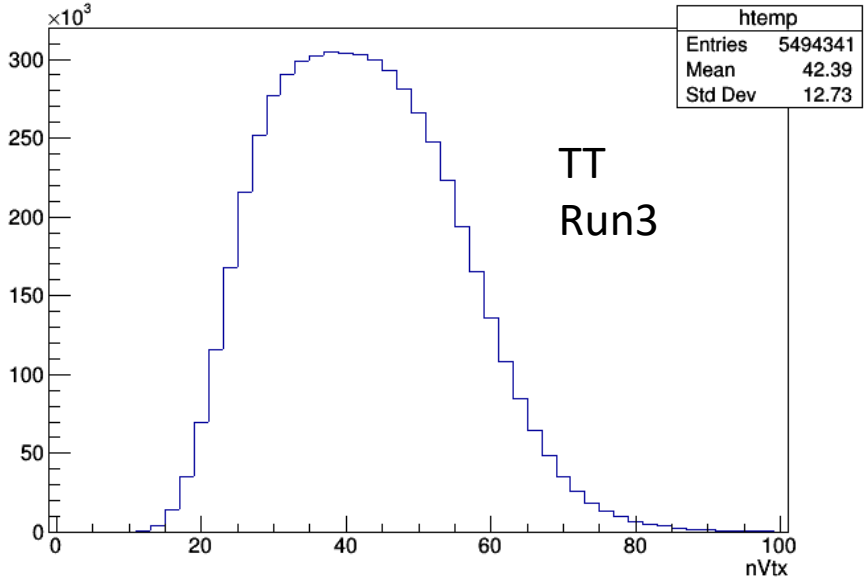
Uncertainties in plots and tables

```
def calcula_eff_error(m_total,m_pass):
    theEffCalculator = r.TEfficiency()
    dataEff = float(m_pass)/float(m_total)
    EffLow = theEffCalculator.ClopperPearson(m_total,m_pass,0.68,False)
    EffUp = theEffCalculator.ClopperPearson(m_total,m_pass,0.68,True)
    return (dataEff-EffLow, EffUp-dataEff)
    theEffCalculator = 0
```

Samples muon MVA ID

- DY Run2 : /DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/MINIAODSIM
- DY Run3 : /DYJetsToMuMu_M-50_TuneCP5_14TeV-madgraphMLM-pythia8/Run3Summer19MiniAOD-2023Scenario_106X_mcRun3_2023_realistic_v3_ext1-v1/MINIAODSIM
- Ttbar Run2: /TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15_ext3-v2/MINIAODSIM
- Ttbar Run3: /TT_TuneCP5_14TeV-powheg-pythia8/Run3Winter21DRMiniAOD-FlatPU30to80_112X_mcRun3_2021_realistic_v16-v2/MINIAODSIM





Samples prompt MVA

Process	Sample name
$t\bar{t}$+jets	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8[1]
Drell-Yan	/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/[3,4]
W+jets	/WJetsToLNu_TuneCP5_13TeV-madgraphMLM-pythia8/[2,5]
Single top quark	/ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8/[6]
	/ST_t-channel_top_4f_inclusiveDecays_TuneCP5_13TeV-powhegV2-madspin-pythia8/[3]
	/ST_t-channel_antitop_4f_inclusiveDecays_TuneCP5_13TeV-powhegV2-madspin-pythia8/[6]
	/ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[6]
	/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/[6]
Multijet	/QCD_Pt-20to30_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[1]
	/QCD_Pt-30to50_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[1]
	/QCD_Pt-50to80_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[1]
	/QCD_Pt-80to120_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[1,2]
	/QCD_Pt-120to170_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[1,3]
	/QCD_Pt-170to300_MuEnrichedPt5_TuneCUETP8M1_13TeV_pythia8/[2,3,4]

Dataset name	Run-range
/SingleMuon/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	297047–299329
/SingleMuon/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	299368–302029
/SingleMuon/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	302031–302663
/SingleMuon/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	303824–304797
/SingleMuon/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	305040–306462
/DoubleMuon/Run2017B-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	297047–299329
/DoubleMuon/Run2017C-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	299368–302029
/DoubleMuon/Run2017D-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	302031–302663
/DoubleMuon/Run2017E-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	303824–304797
/DoubleMuon/Run2017F-UL2017_MiniAODv2_NanoAODv9-v1/NANOAOD	305040–306462

TRAINING

- Main challenge: **class imbalance**. More than 98% of the muons are from signal.
- Several machine learning models, such as tree classifiers, BDTs, KNN and neural networks, were tested, but the best performance is achieved with the **random forest**.
- For the training we use the **Scikit-learn package**.
- We take the 60% of the muons to **train** the model and the rest for **testing**.
- All the **hyperparameters** were optimized, combining manual and grid-search strategies, with the aim of achieving the best performance while preventing over-fitting.
- The **memory usage** was also considered in the optimization.

Muon MVA ID efficiency vs pT