Improving Zmm selection in ATLAS using LightGBM

Sara Dahl Andersen

with Malte Algren and Mads Storr-Hansen

Under supervision of Troels Petersen

Niels Bohr Institute, University of Copenhagen

Long term goal

- Improve Higgs channels: $H \rightarrow Z\gamma$, $H \rightarrow \gamma^*\gamma$, $H \rightarrow ZZ^*$

Long term goal

- Improve Higgs channels: $H \rightarrow Z\gamma$, $H \rightarrow \gamma^*\gamma$, $H \rightarrow ZZ^*$

Current goal

- Improve selection in $Z \rightarrow \mu\mu$ for both MC and Data
- Soon move on to $Z \rightarrow \mu\mu\gamma$

Outline

- Models on Monte Carlo data to get Zmm model for muon pairs: Similar approach as Troels just talked about
  - ML PID and Isolation models for single muons
  - ML Zmm model (ML PID and ML ISO serves as input)

- Moving on to Data
  - (Try to create) ML PID and Isolation models on Data
  - Using the Monte Carlo based Zmm model on Data with ML PID and ML ISO as input
  - Evaluating the performance on Data

Data

MC:

Signal:

mc16_13TeV.361107.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zmumu.deriv.DAOD_MUON1.e3601_e5984_s3126_r10201_r10210_p3629

Background:

mc16_13TeV.361101.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Wplusmunu.deriv.DAOD_MUON1.e3601_e5984_s3126_r10724_r10726_p3629

mc16_13TeV.361250.Pythia8B_A14_NNPDF23LO_bbTomu15.deriv.DAOD_MUON1.e3878_e5984_s3126_r10724_r10726_p3629

Eventually we would include Zmmgam files to get low pt muons

Data:

data18_13TeV.00349693.physics_Main.deriv.DAOD_MUON1.f933_m1960_p3553

LightGBM

- All models are created using the framework LightGBM
  - Tree-based models
  - Boosing: gradient boosting decision tree
  - Parameters are not hyperparameter optimized (yet!!)

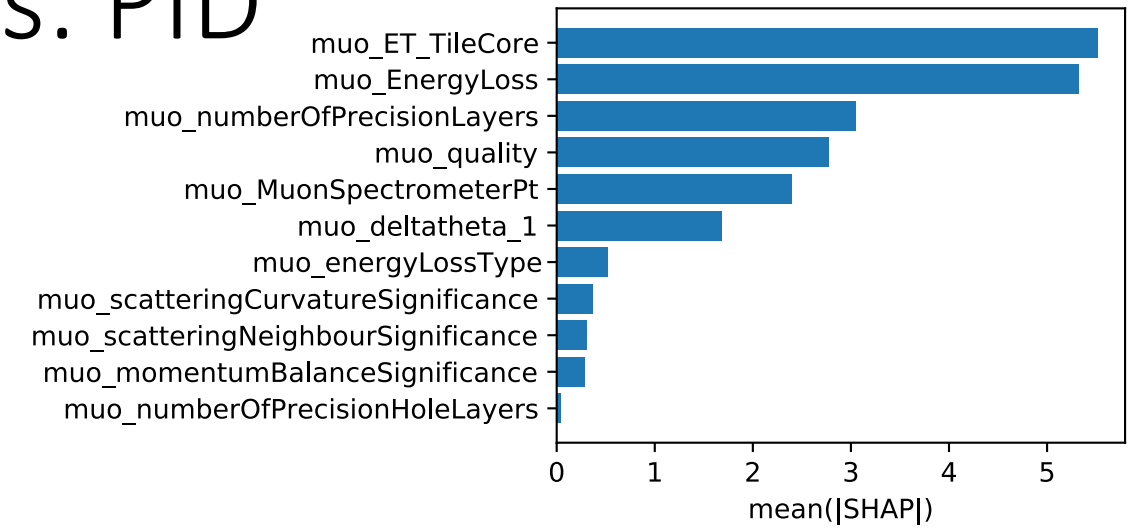
MC model on muons

We create two ML models on single muons:

Particle identification (PID)

Isolation (ISO)

MC model on single muons: PID



Particle identification (PID) model:

Signal selection: truthPgdId = 13 (muons)

Variables used in the model:

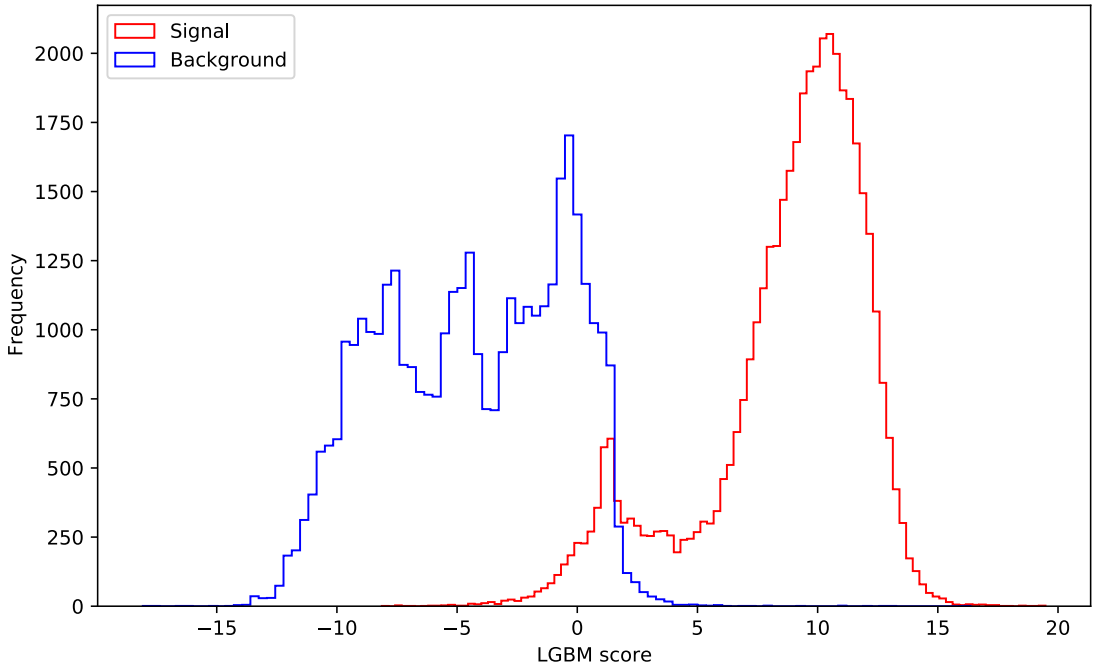
We have tested all PID variables and found the most important according to SHAP ranking.

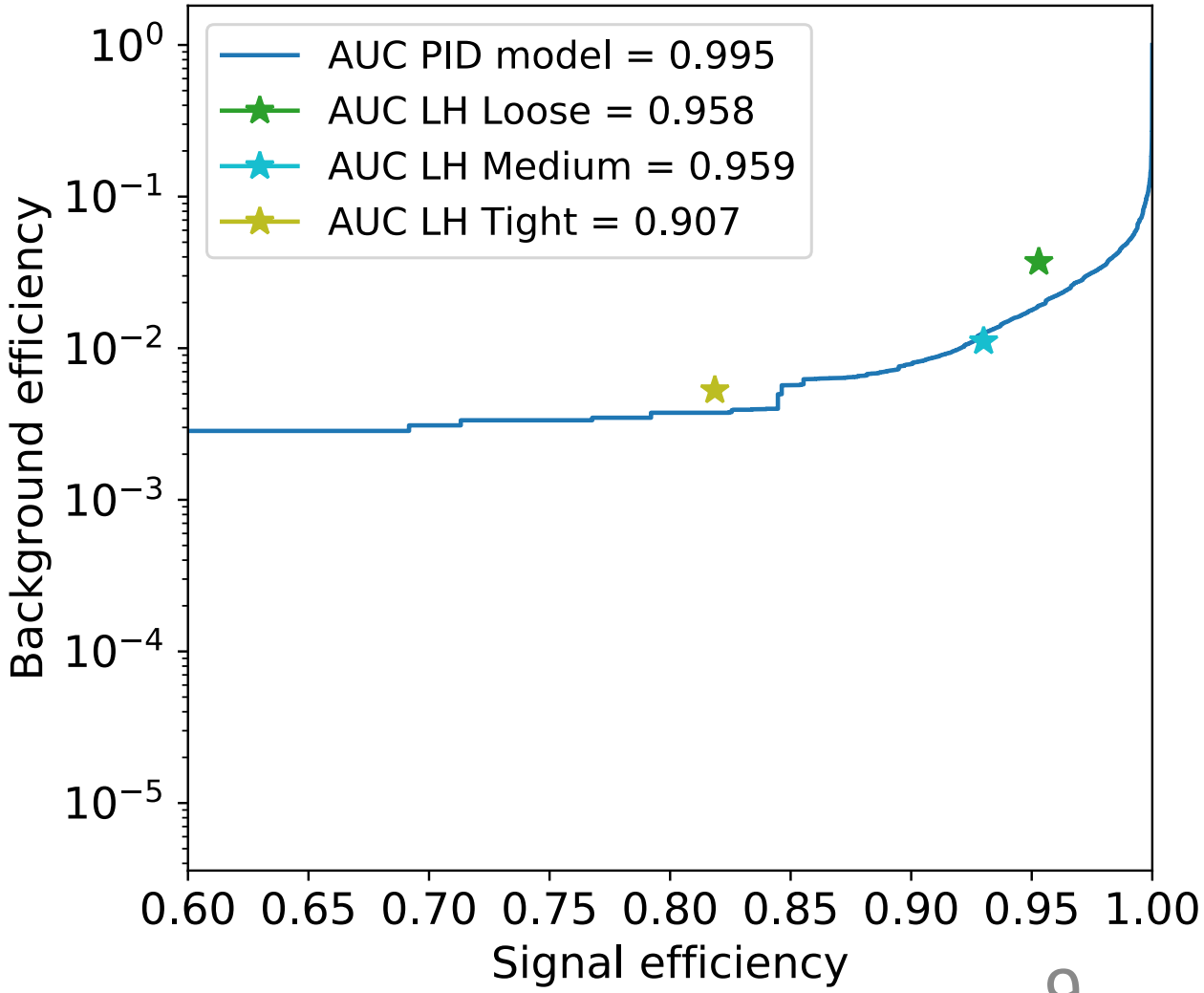
muo_numberOfPrecisionLayers	muo_numberOfPrecisionHoleLayers
muo_MuonSpectrometerPt	muo_ET_TileCore
muo_scatteringCurvatureSignificance	muo_scatteringNeighbourSignificance
muo_momentumBalanceSignificance	muo_energyLossType
muo_EnergyLoss	muo_deltatheta_1
muo_quality	

MC model on single muons: PID

Particle identification (PID) model:

Needs hyperparameter optimization

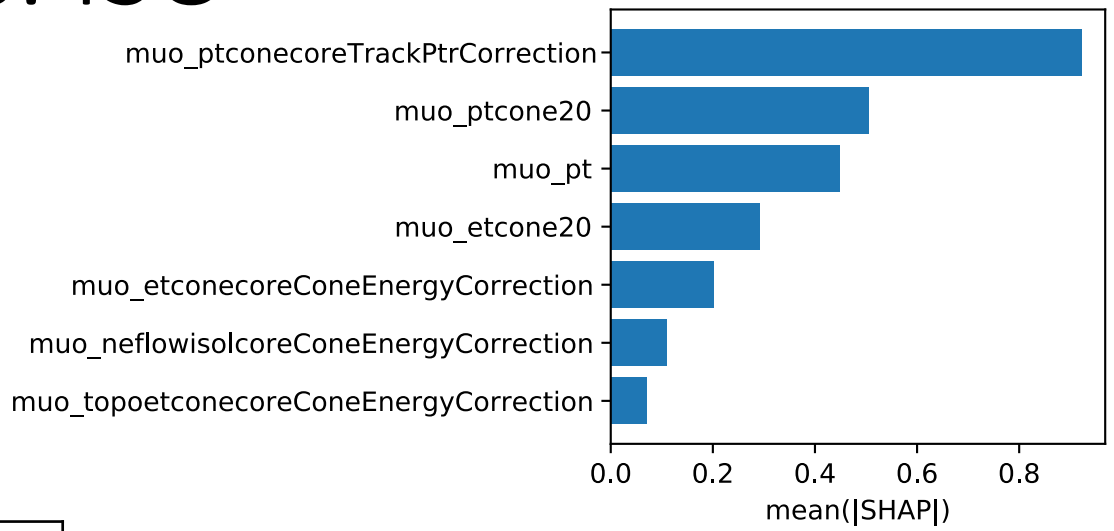




MC model on single muons: ISO

Isolation (Iso) model:

Signal selection: truthOrigin = 13 (Z)



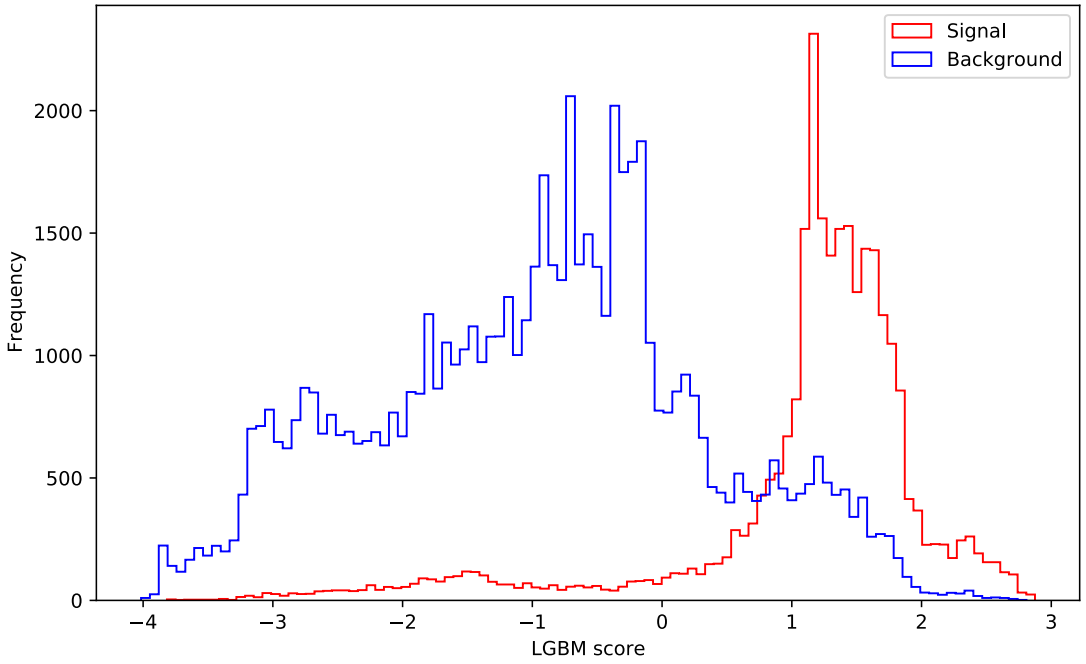
Variables used in the model:

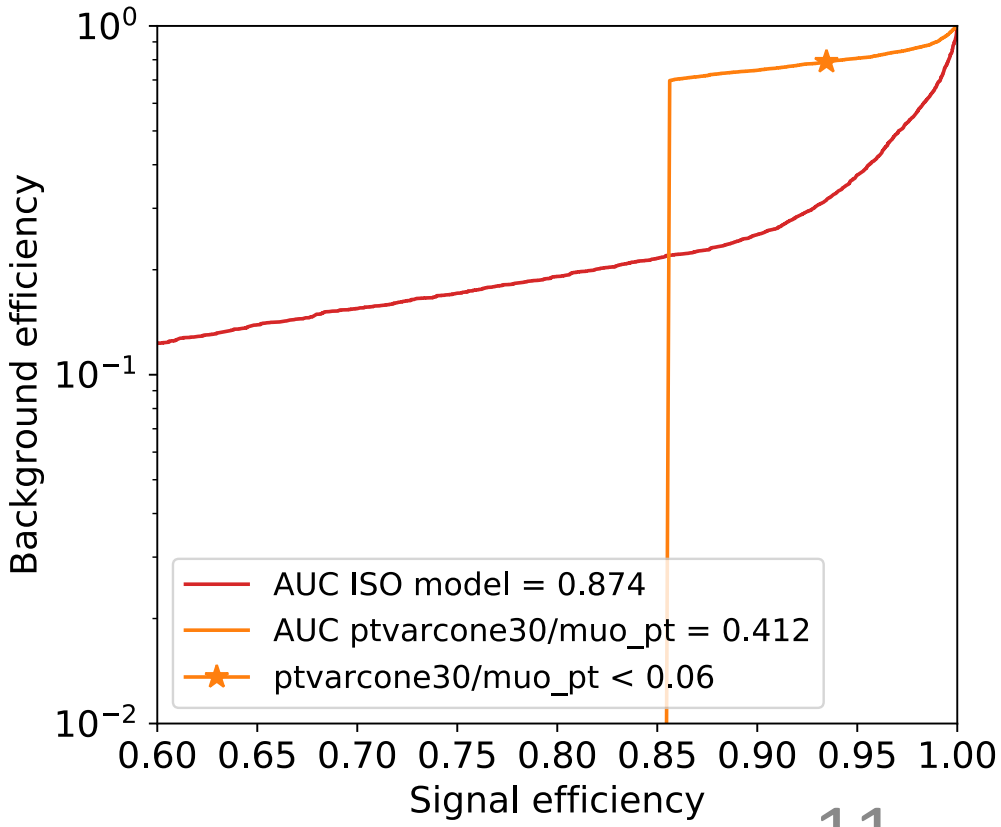
muo_ptconecoreTrackPtrCorrection	muo_ptcone20
muo_etconecoreConeEnergyCorrection	muo_etcone20
muo_neflowisolcoreConeEnergyCorrection	muo_pt
muo_topoetconecoreConeEnergyCorrection	

MC model on single muons: ISO

Isolation (ISO) model:

Compared with Isolation WP from ATLAS-CONF-2020-030:
$p_T^{varcone30} < 0.06 \cdot p_T^{\mu}$



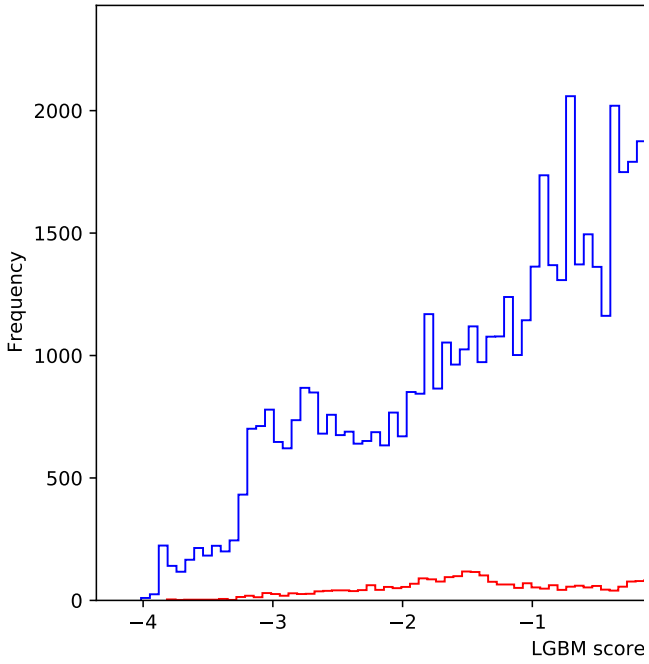


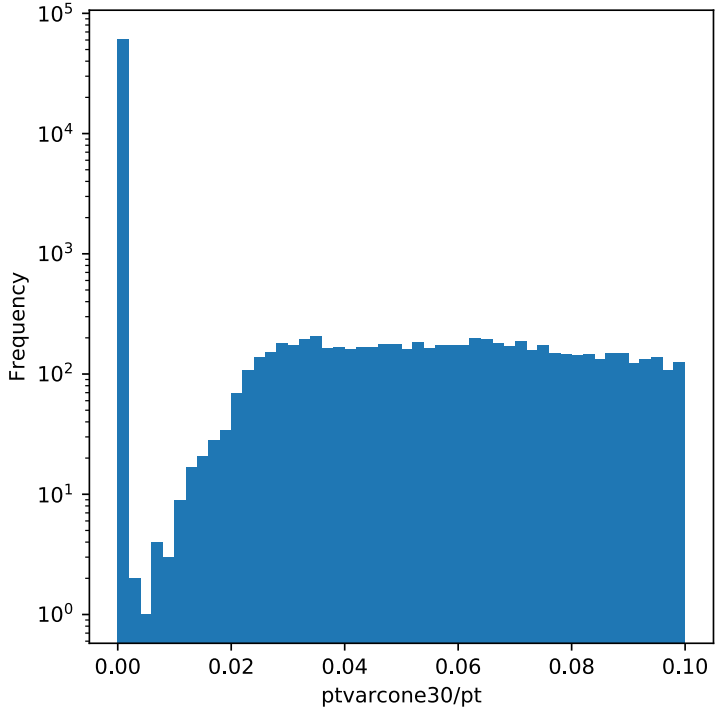
MC model on single muons: ISO

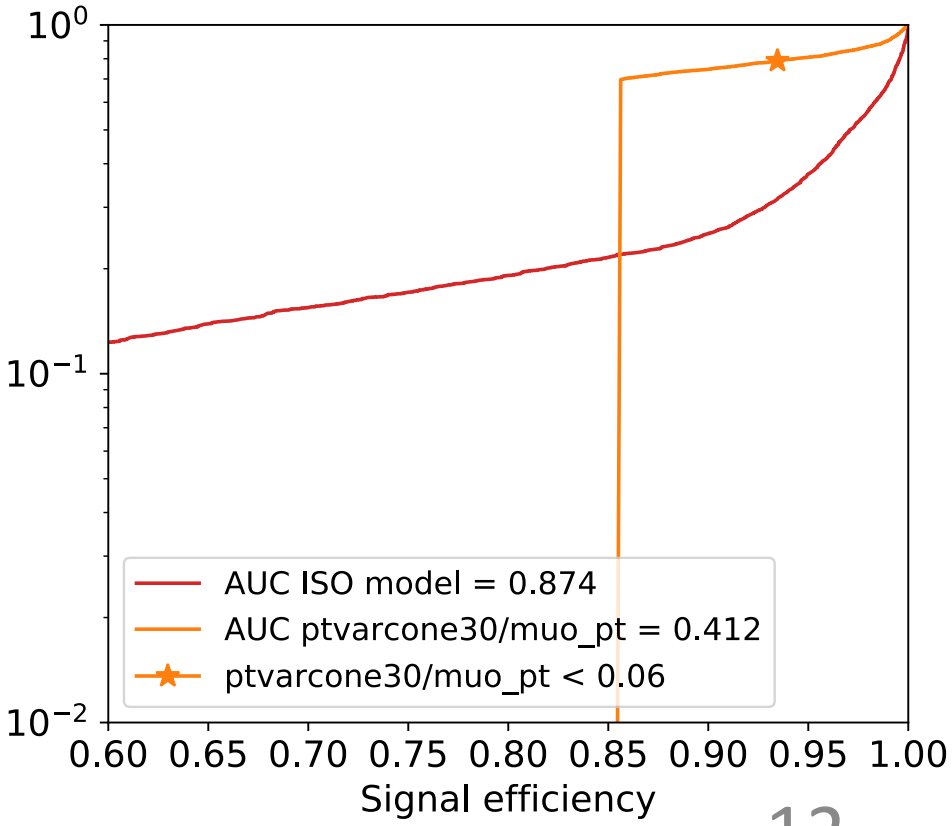
Isolation (ISO) model:

Compared with Isolation WP from ATLAS-CONF-2020-030:

$p_T^{varcone30} < 0.06 \cdot p_T^{\mu}$







MC model on Zmm
Model for muon pairs originating from the Z boson

- The two models for single muons will be used as input to the Zmm model
  - The two muons in the Z -> mm decay will each recieve a score from the ML models
  - The scores are used as variables in the LGBM model

MC model on Zmm: variables

Signal: Muon pairs originating from Z w. opposite sign
Background: Muon pairs not originating from Z

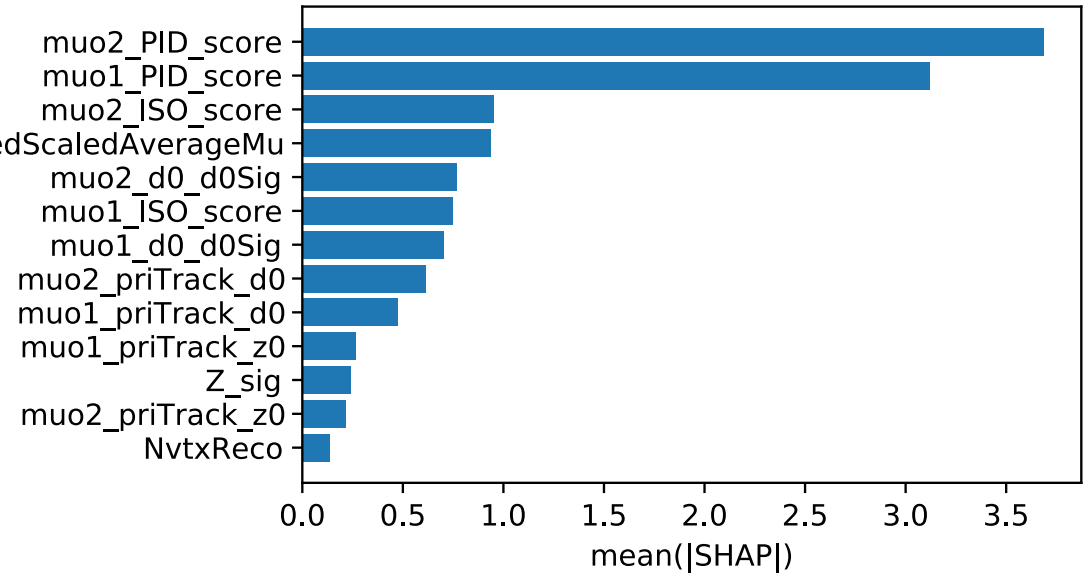
$$Z_{sig} = \frac{\mu_1^{z_0} - \mu_2^{z_0}}{\sqrt{\left(\mu_1^{z_0}\right)^2 - \left(\mu_2^{z_0}\right)^2}}$$

$$d_0^{\sigma_{d0}} = \frac{d_0^{pri\_track}}{\sigma_{d0}^{pri\_track}}$$

Training set:
- nSig: 48021, nBkg: 275969

Validation set:
- nSig: 11983, nBkg: 69015

Event variables		
correctedScaledAverageMu	NvtxReco	Z_sig
Muon variables (exists for both muons)		
muo_ISO_score	muo_PID_score	muo_d0_d0Sig
muo_priTrack_d0	muo_priTrack_z0	



MC model on Zmm: comparing with ATLAS

Using ATLAS selection from arXiv: 2005.05382

Two opposite-sign muons
pt > 10 GeV
$\|\eta\| < 2.7$
Medium ID
$d_0/\sigma_{d0} < 3$
~~High-quality track in ID or MS~~
~~$\|\Delta z_0 \cdot \sin\theta\| < 0.5mm$~~

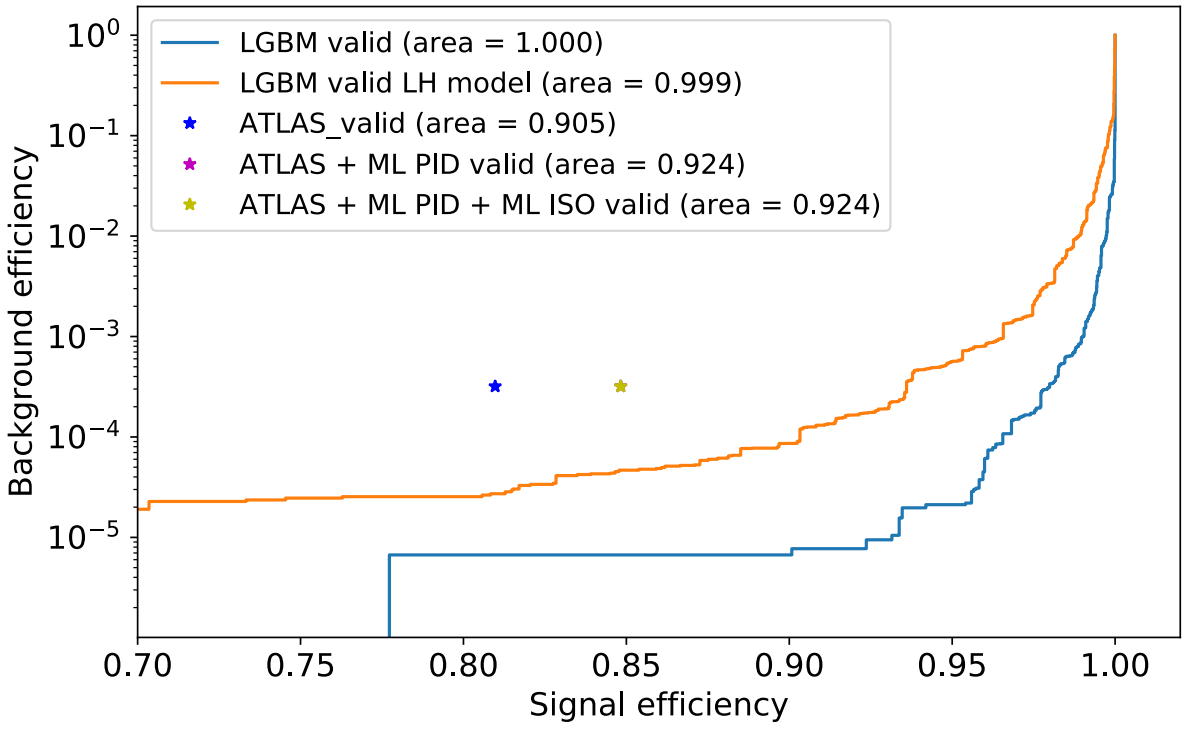
Modified ATLAS selection from arXiv: 2005.05382

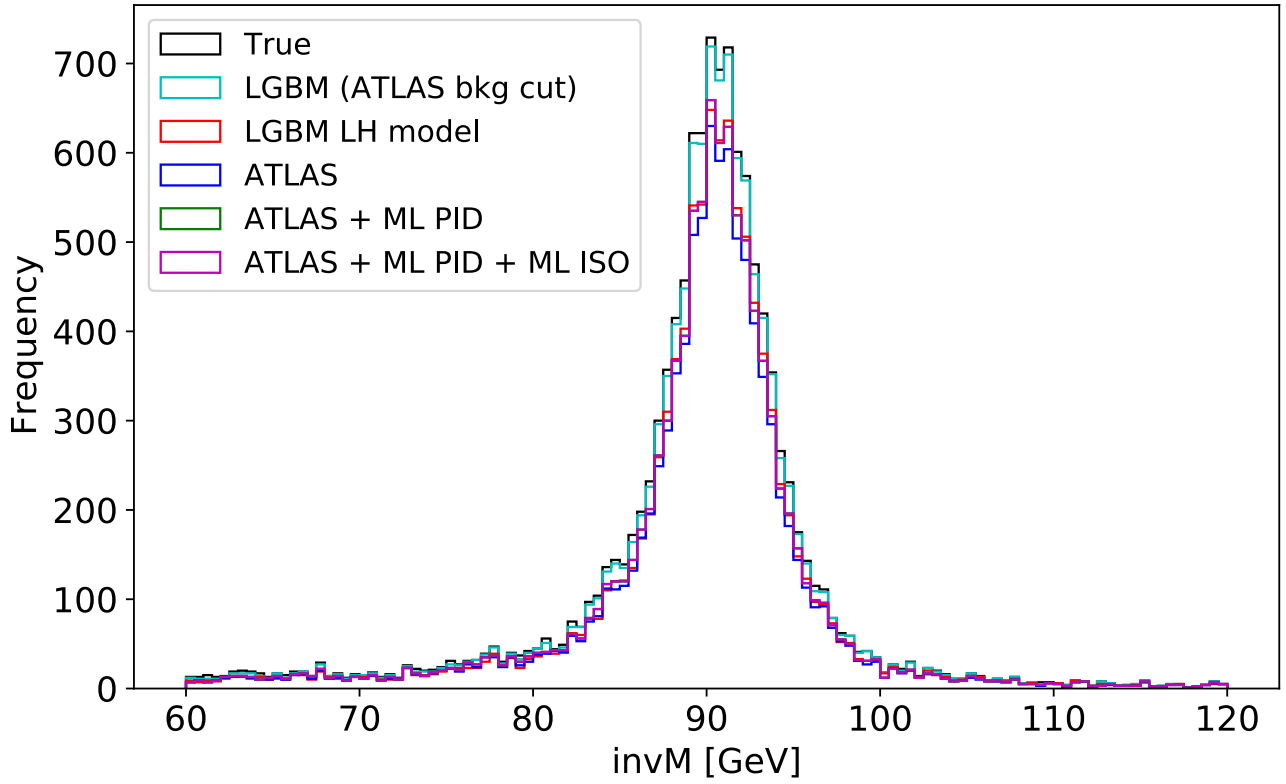
Two opposite-sign muons
pt > 10 GeV
$\|\eta\| < 2.7$
ML Pid / ML Iso
$d_0/\sigma_{d0} < 3$
~~High-quality track in ID or MS~~
~~$\|\Delta z_0 \cdot \sin\theta\| < 0.5mm$~~

Modified ML Zmm model

correctedScaledAverageMu
NvtxReco
Z_sig
Medium ID
muo_d0_d0Sig
muo_priTrack_z0
muo_priTrack_d0

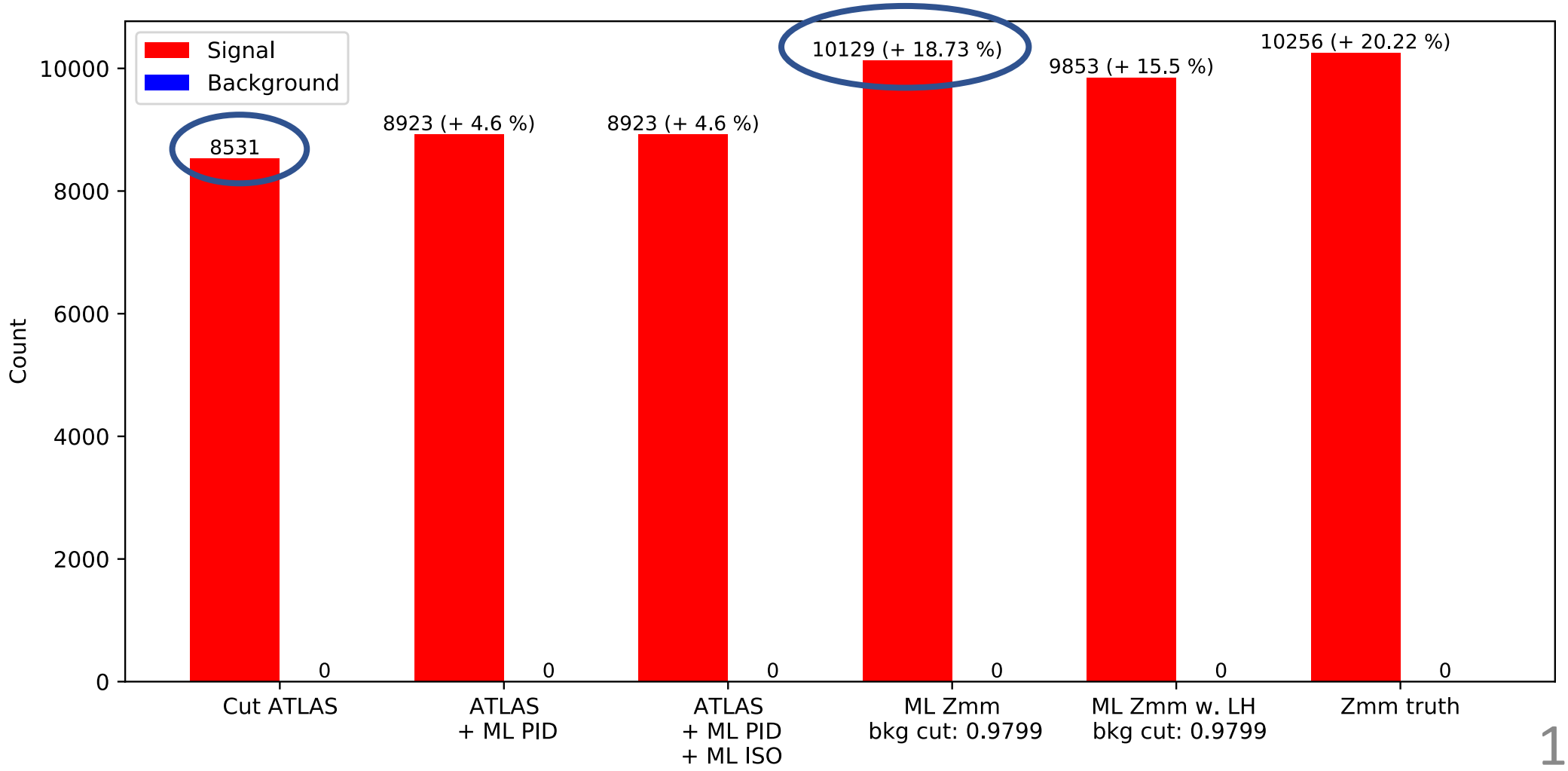
MC model on Zmm: prediction





MC model on Zmm: prediction

Range of invariant mass:
80 < invM < 100



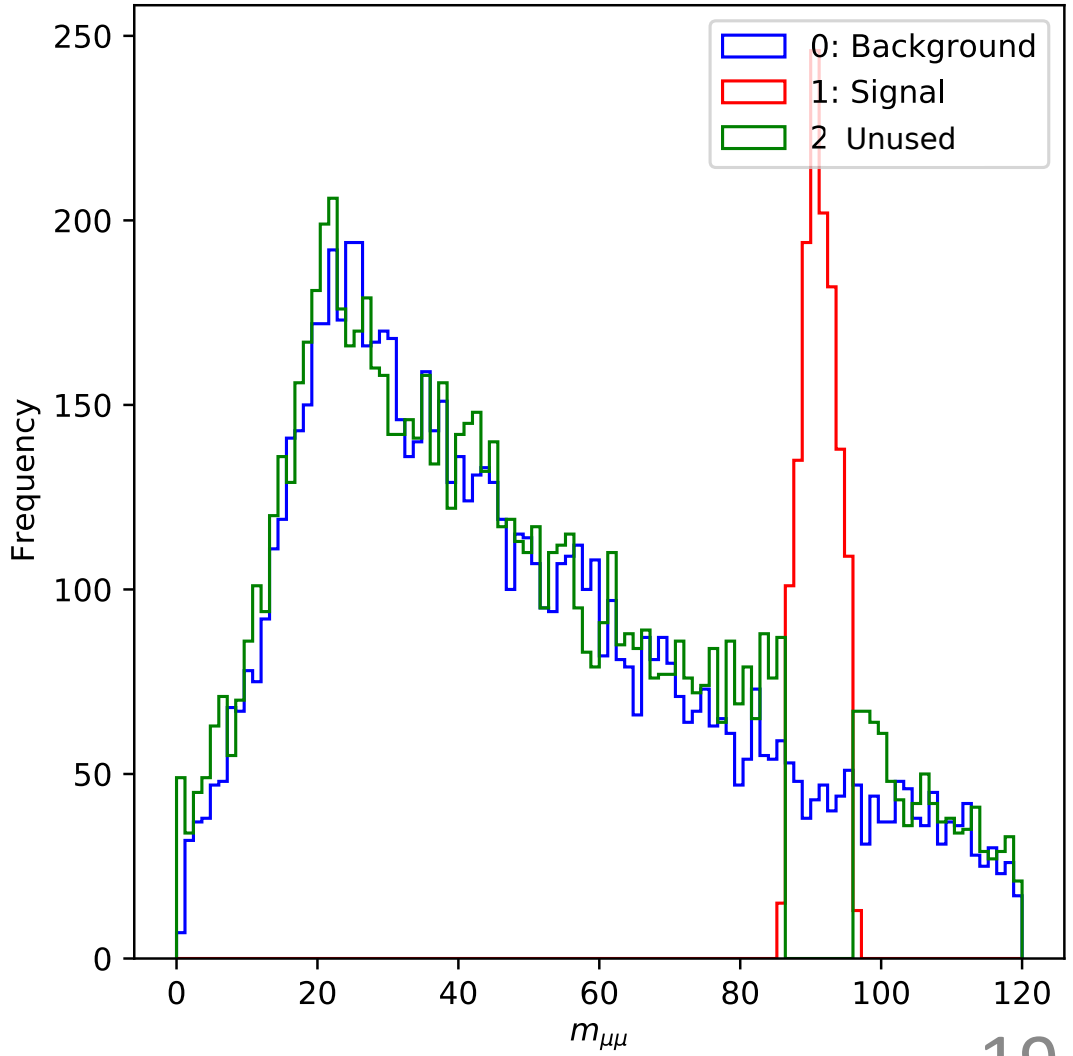
ML models on Data
PID and ISO models

Machine Learning on Data: Event selection

Tag and probe selection:

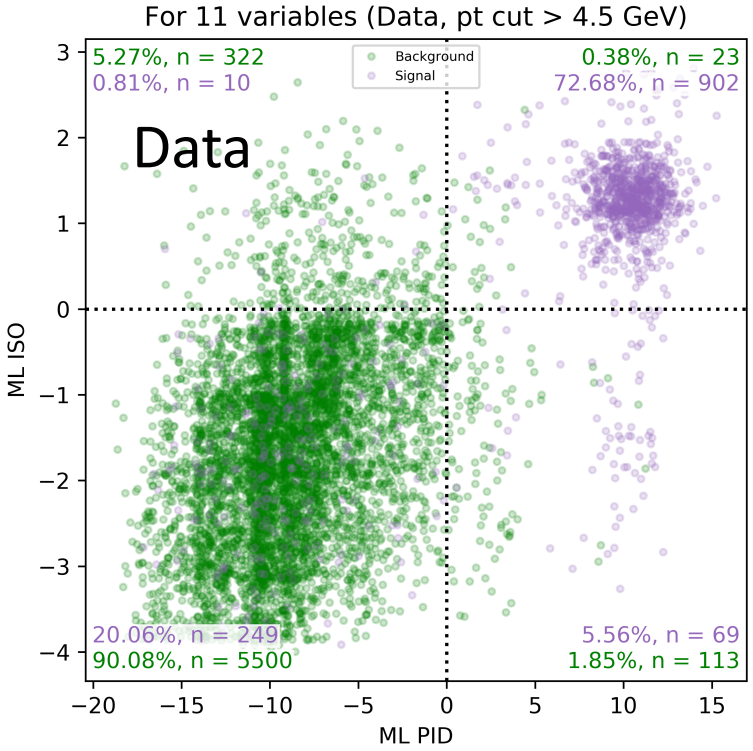
- Tag: Trigger muon + Tight ID
- Tag and probe pair
- Check their sign and Z mass range
  - accept if +/- 5 GeV from Z
- Save only the probe muon

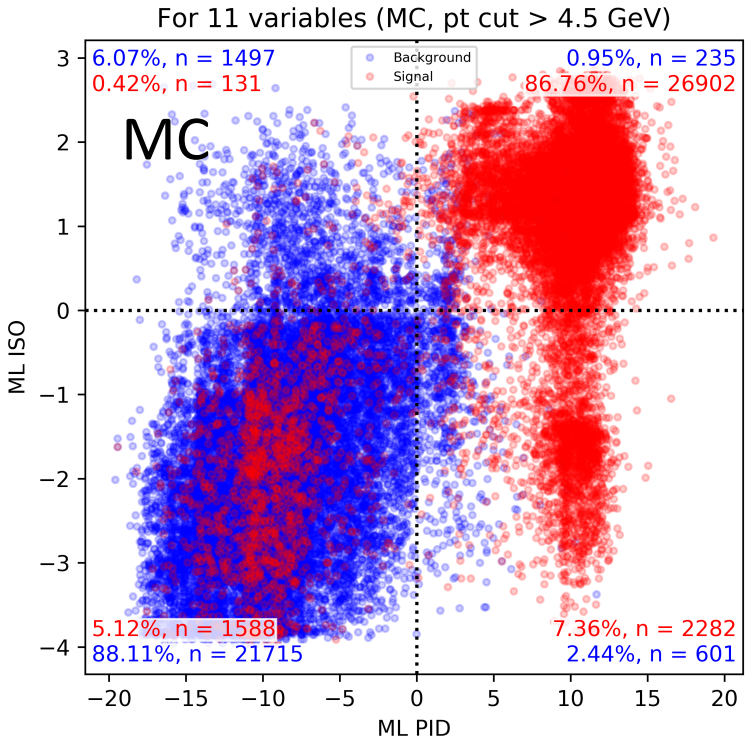
- Background is same-sign muons
- Unused is opposite-sign muons outside Z mass range

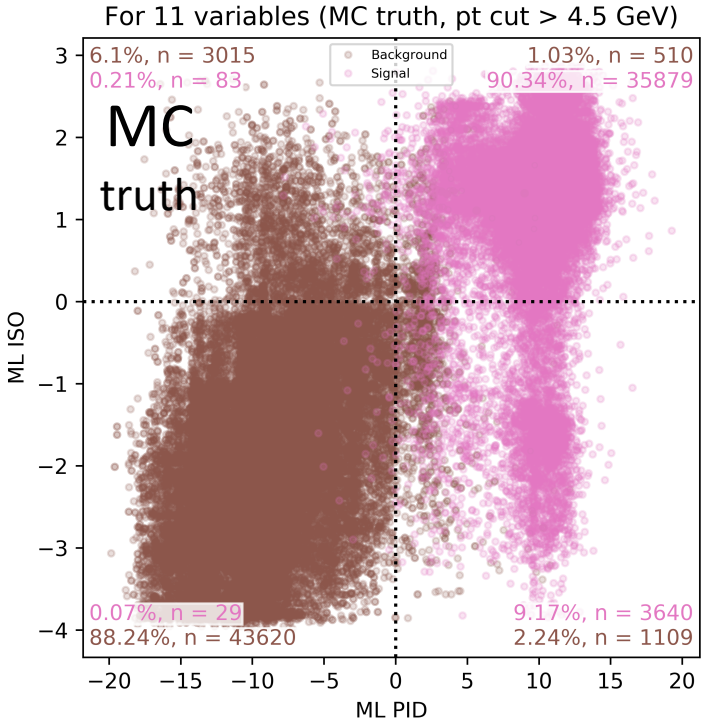


Machine Learning on Data: scoring using MC models

We use the T&P selection on both Data and MC and compare with MC truth



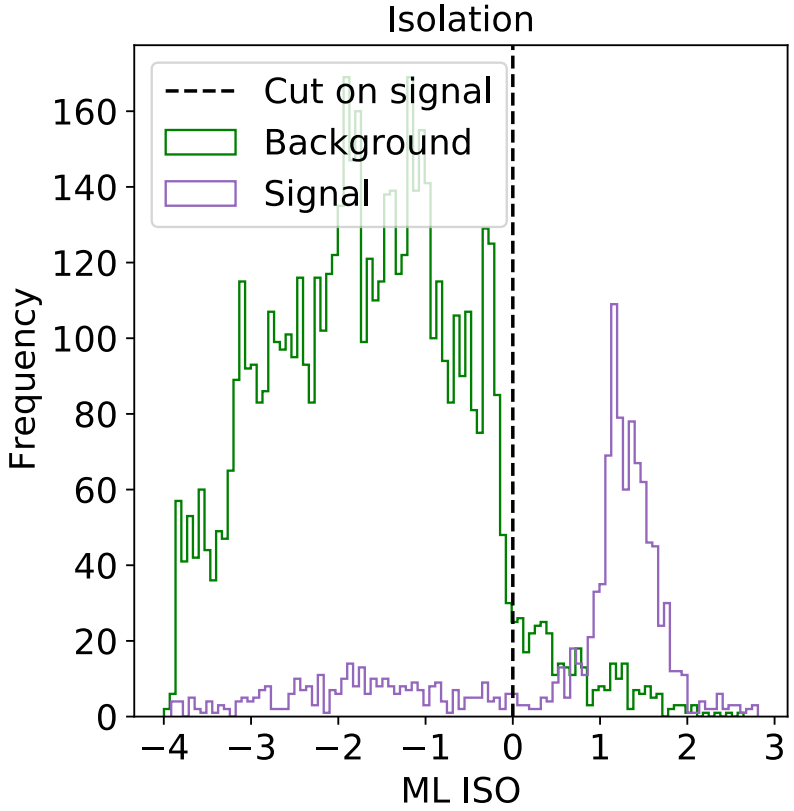




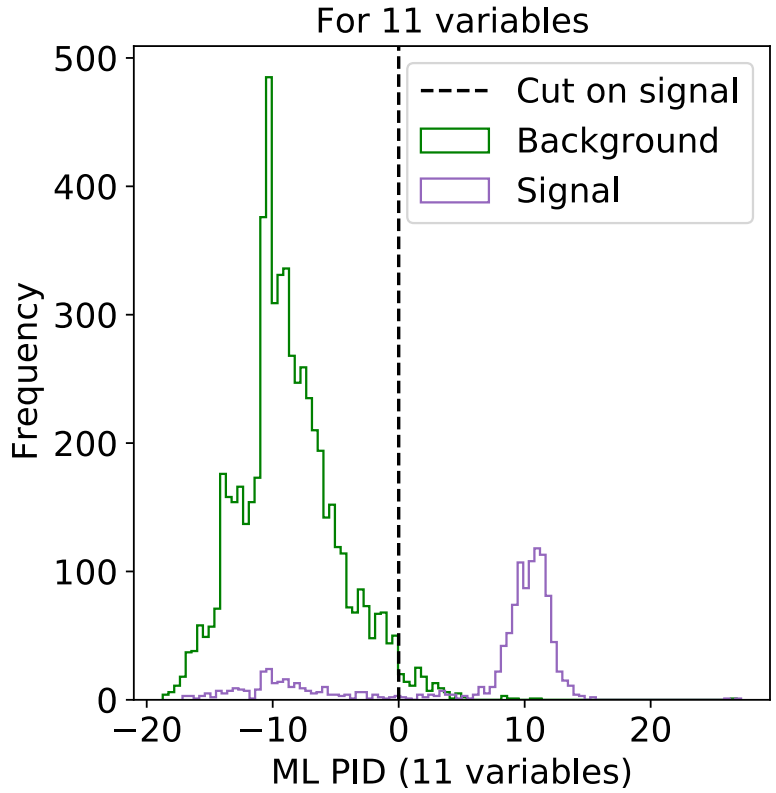
From the MC truth plot, we can conclude that the data can be cleaned further before it is used in PID and ISO models

Machine Learning on Data: ML models

PID dataset: We cut in the ML Iso variable



ISO dataset: We cut in the ML Pid variable



Zmm model on Data
Model for muon pairs originating from the Z boson

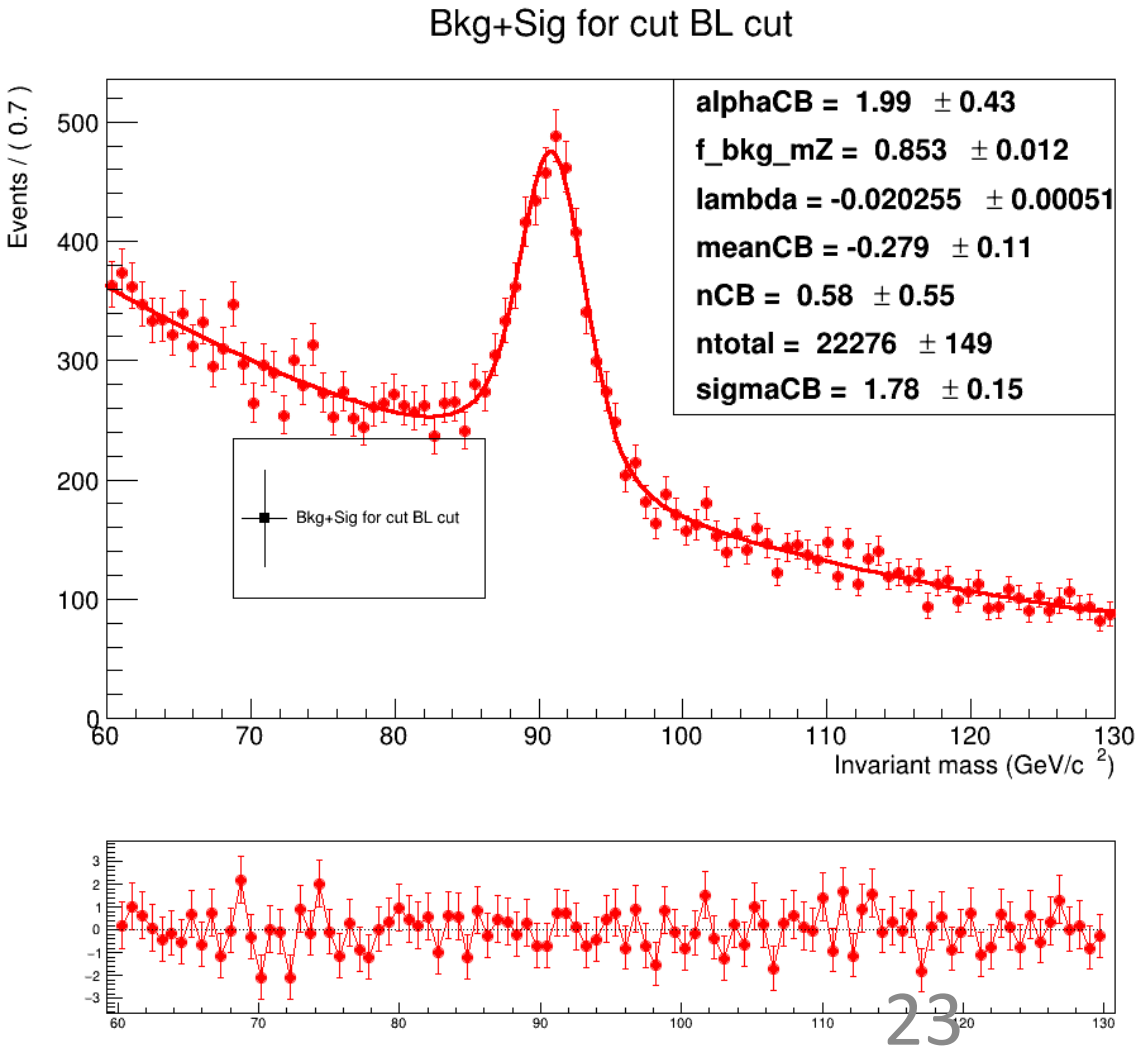
- We will use the ML Pid and ML Iso as input to the model as before
- Evaluating the model on Data

Machine Learning on Data: Event selection

We will not use tag & probe, as we need muon pairs

Instead we only look at sign of the muons:

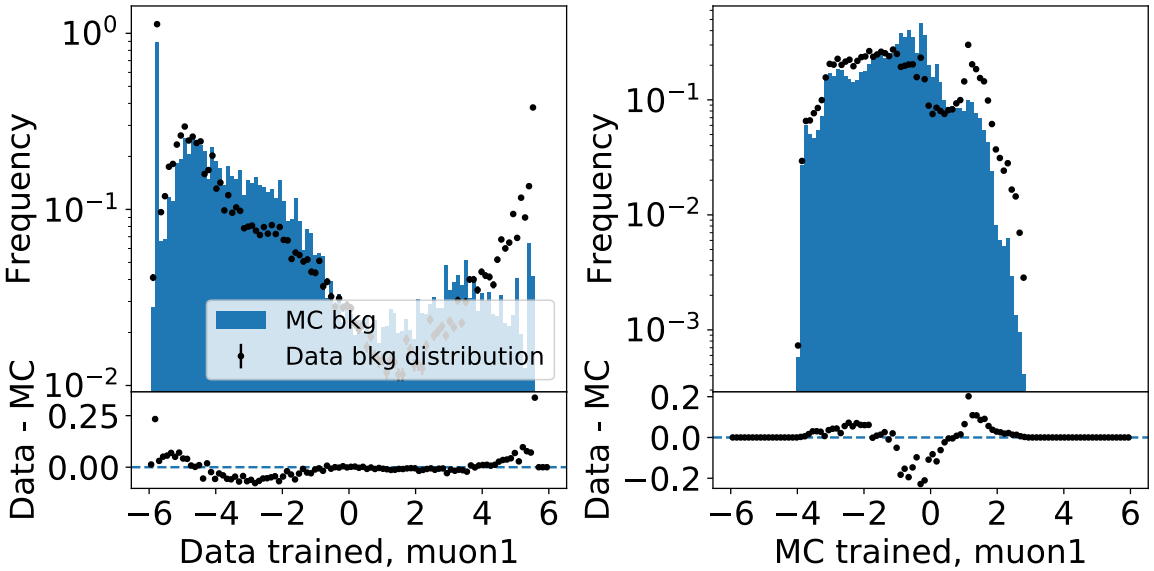
- Same-sign muons: Definetly background
- Opposite-sign muons: Some are signal while some are background
  - Numbers extracted from fit:
    - nSig: 3.274
    - nBkg: 19.001



Plot: Fit of opposite-sign muons

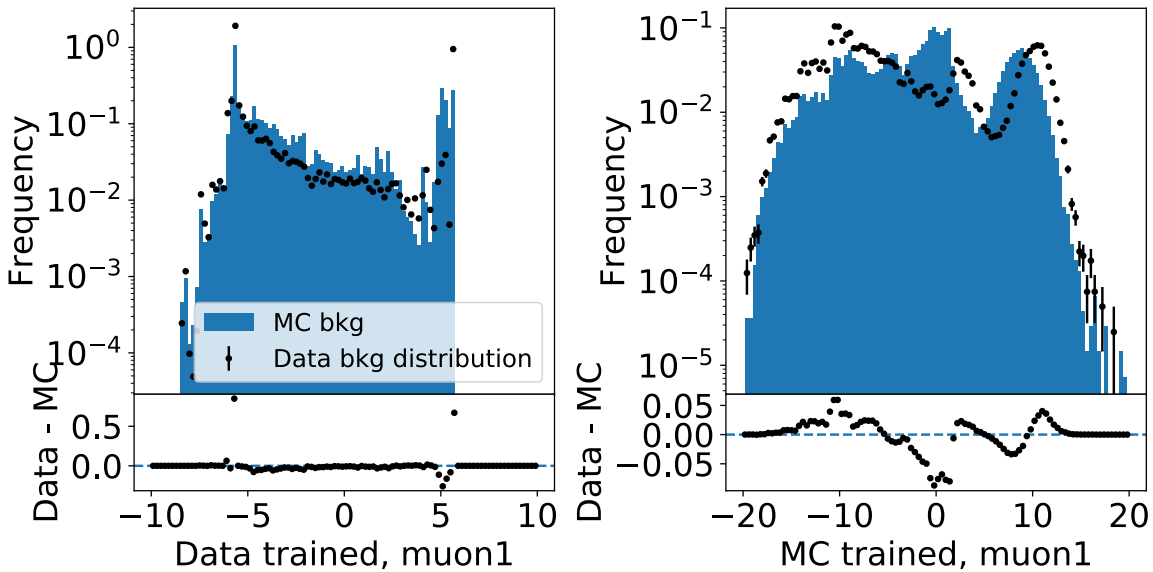
Machine Learning on Data same-sign: Prediction compared with MC background

Isolation



Score of Isolation ML model for Data compared with score of Isolation ML model for MC

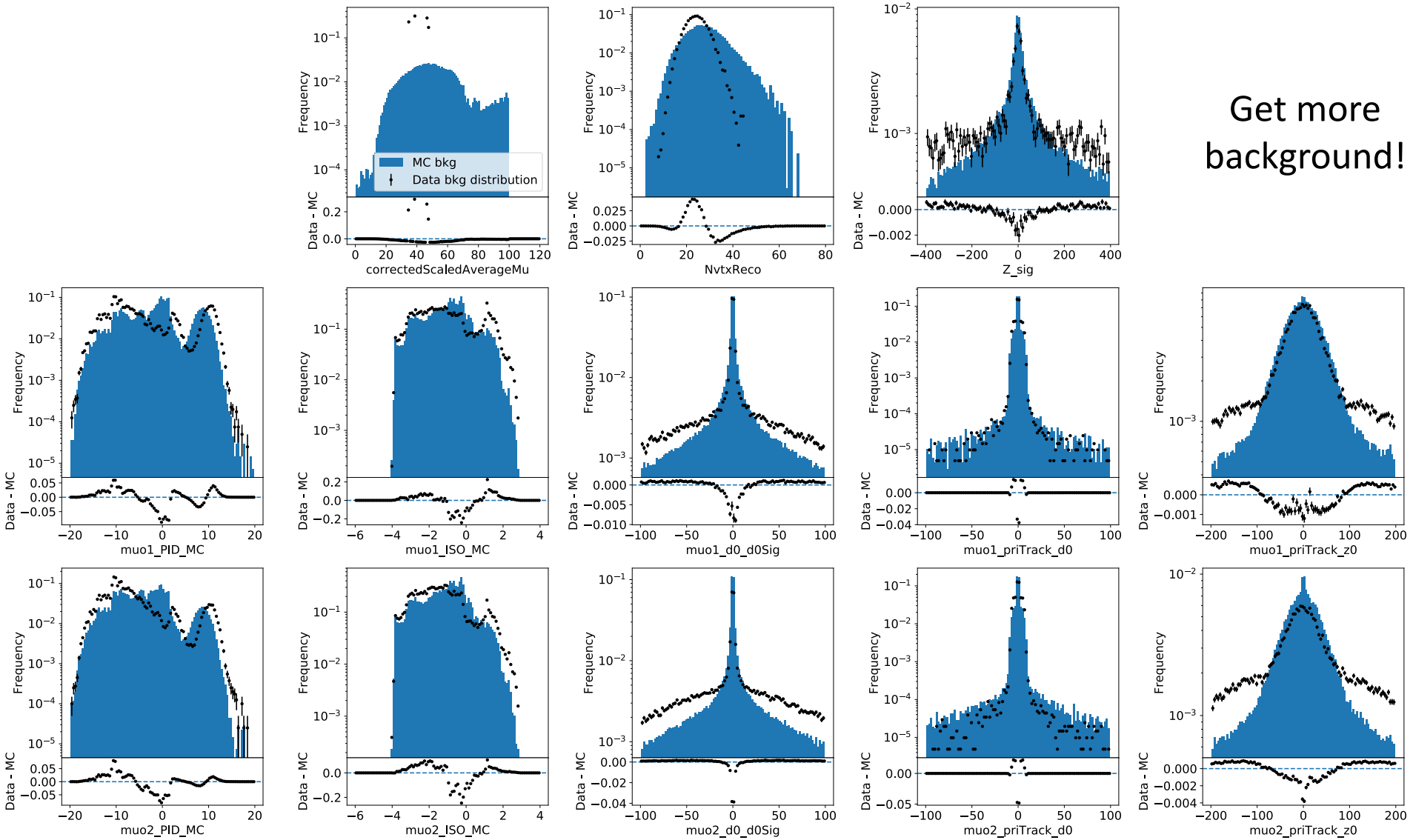
PID



Score of PID ML model for Data compared with score of PID ML model for MC

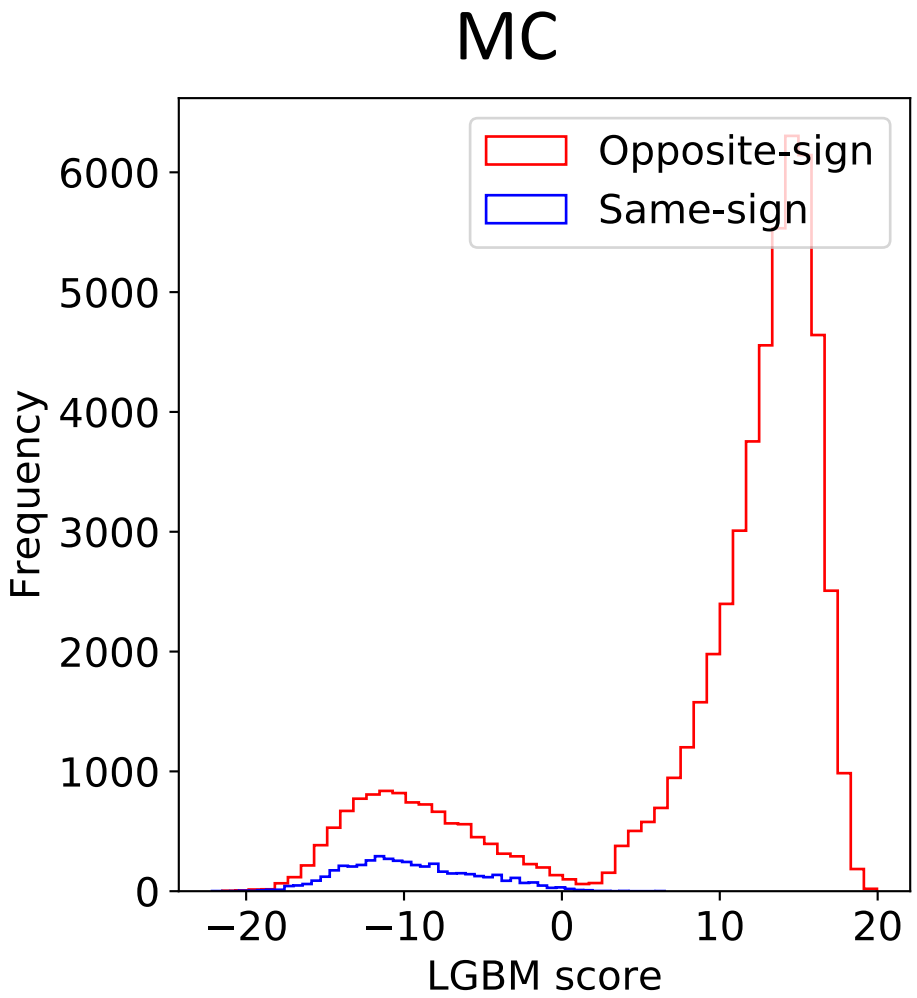
We will use MC model for data as well for now, we might have a solution for this: transforming the MC distribution

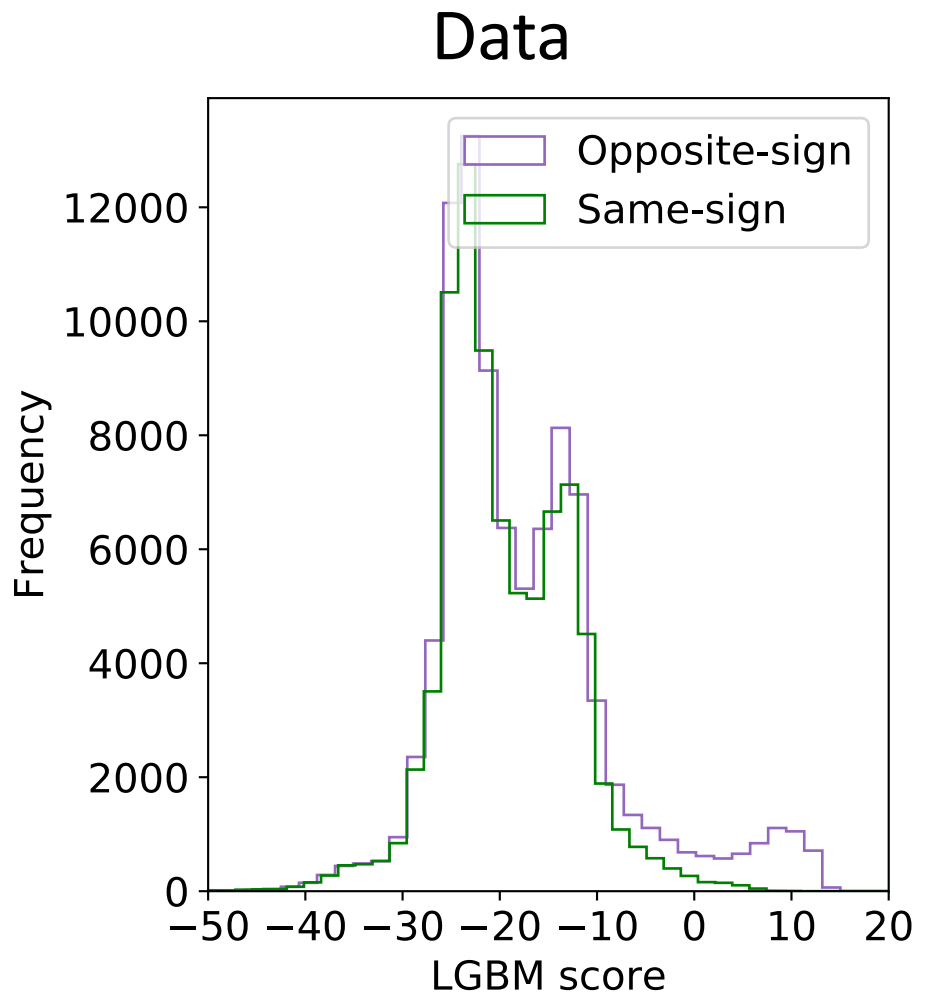
Machine Learning on Data: variables



Get more background!

Machine Learning on Data: LGBM model



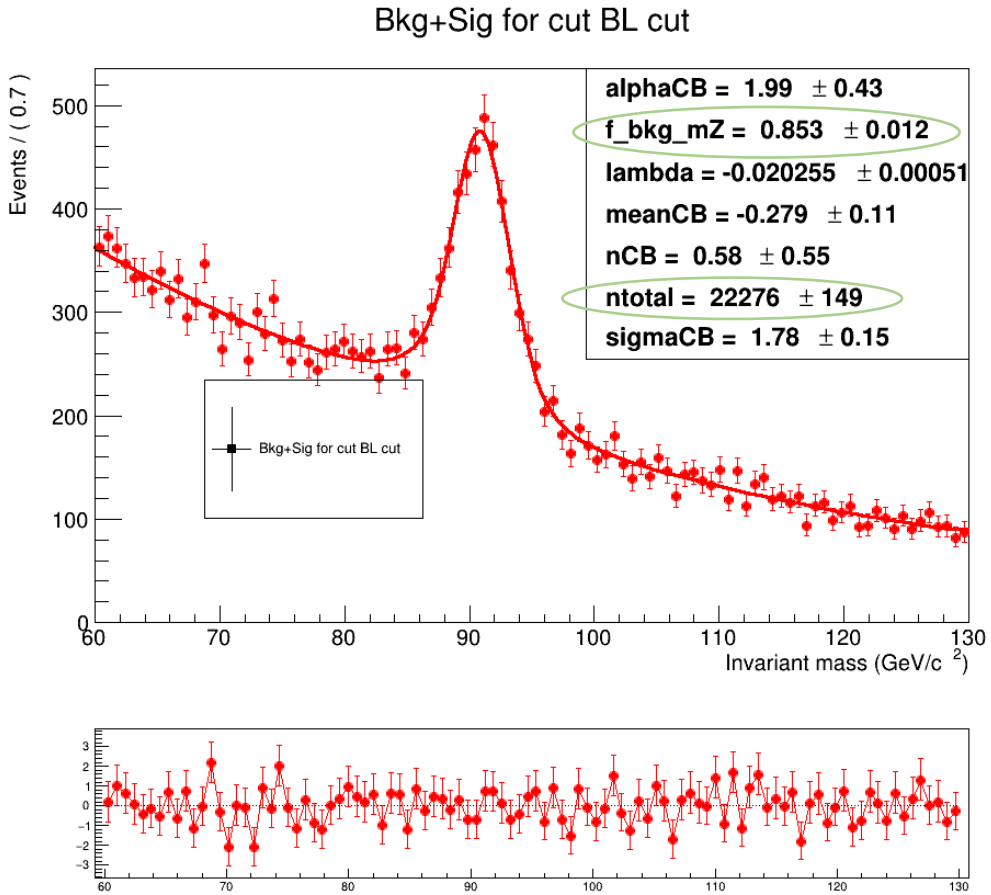


Machine Learning on Data: Fitting Z peak

Signal: Breit-Wigner x Crystal Ball

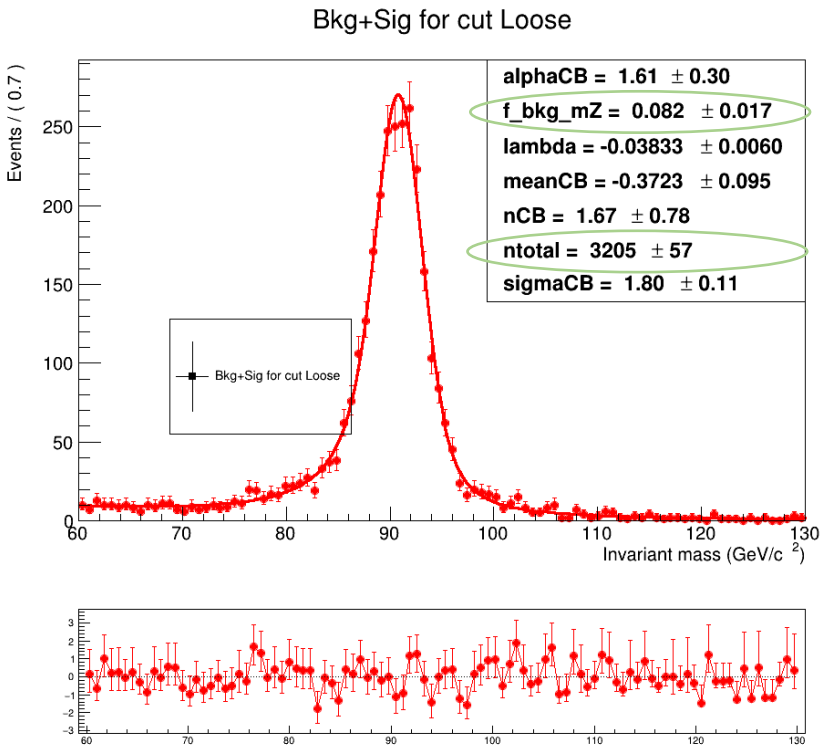
Background: $N \cdot \exp(\lambda \cdot m)$

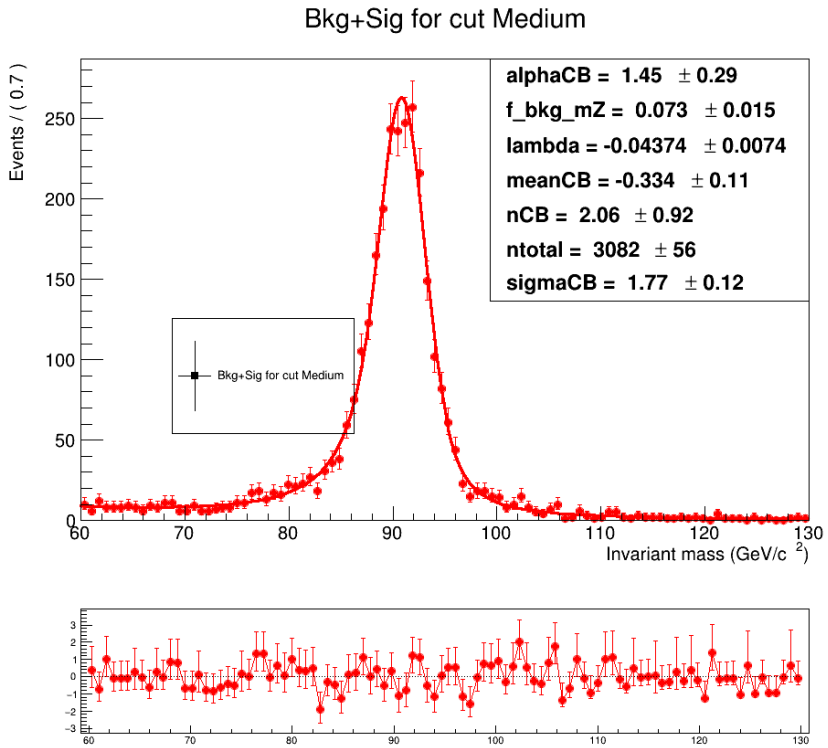
Data shown is opposite-sign. This is our baseline cut, we will find the signal efficiency from cutting in the LGBM score to compare with ATLAS' likelihood cuts.

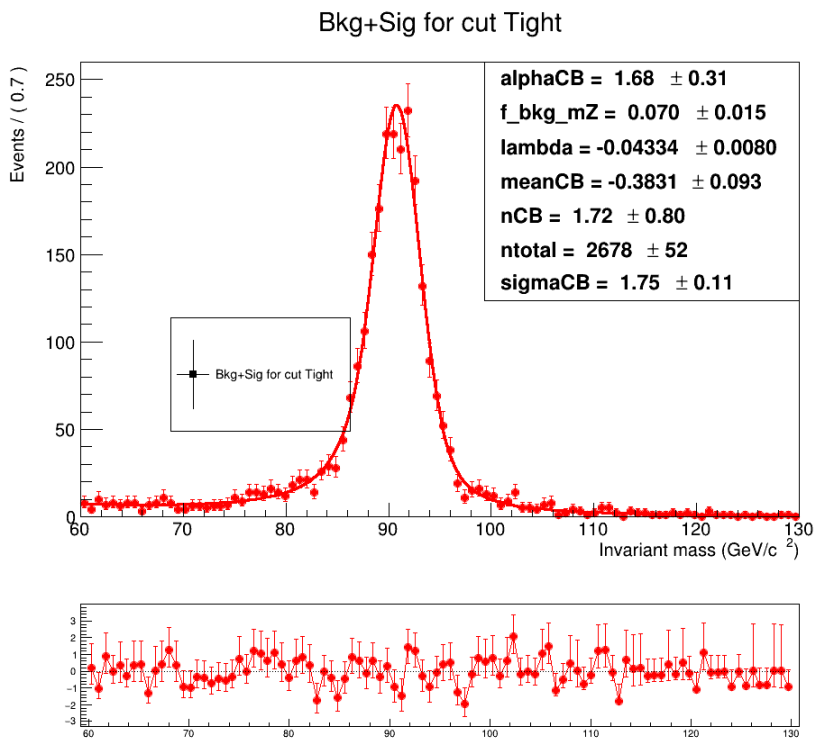


Machine Learning on Data: Fitting Z peak

Using Loose/Medium/Tight selection on opposite-sign







Machine Learning on Data: ”ROC” curve

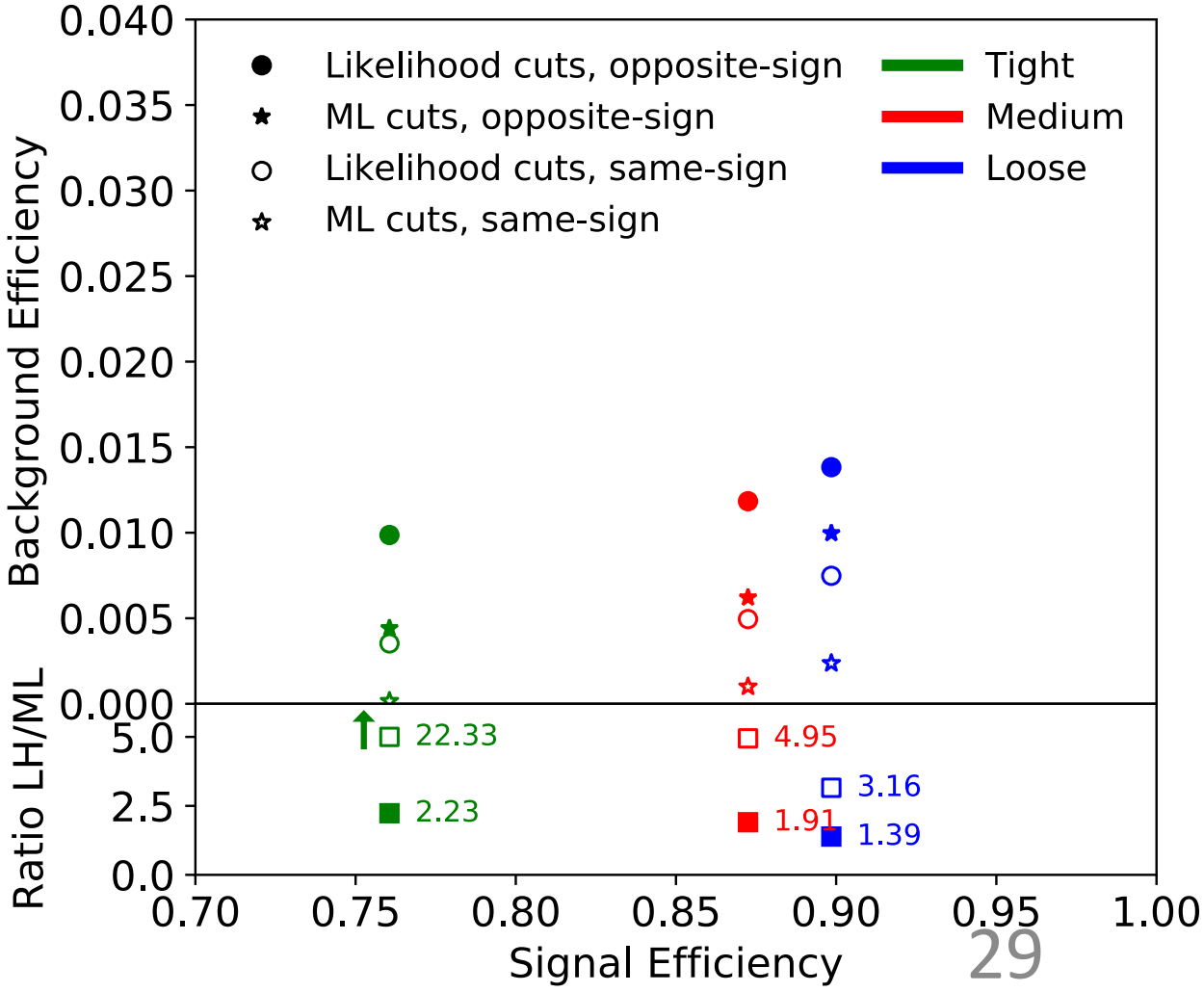
We look at the signal efficiency for LH cuts and cut in our LGBM score to get the same signal efficiency

We can compare the background efficiencies

For same-sign, we have no signal, but we make the same cut in LGBM as for opposite-sign and count how much background is left.

Ratio plot show the decrease in background at the same signal efficiency

The same plot should be made in MC to compare – we need to get a bigger background sample before this can be done.



Machine Learning on Data: ”ROC” curve

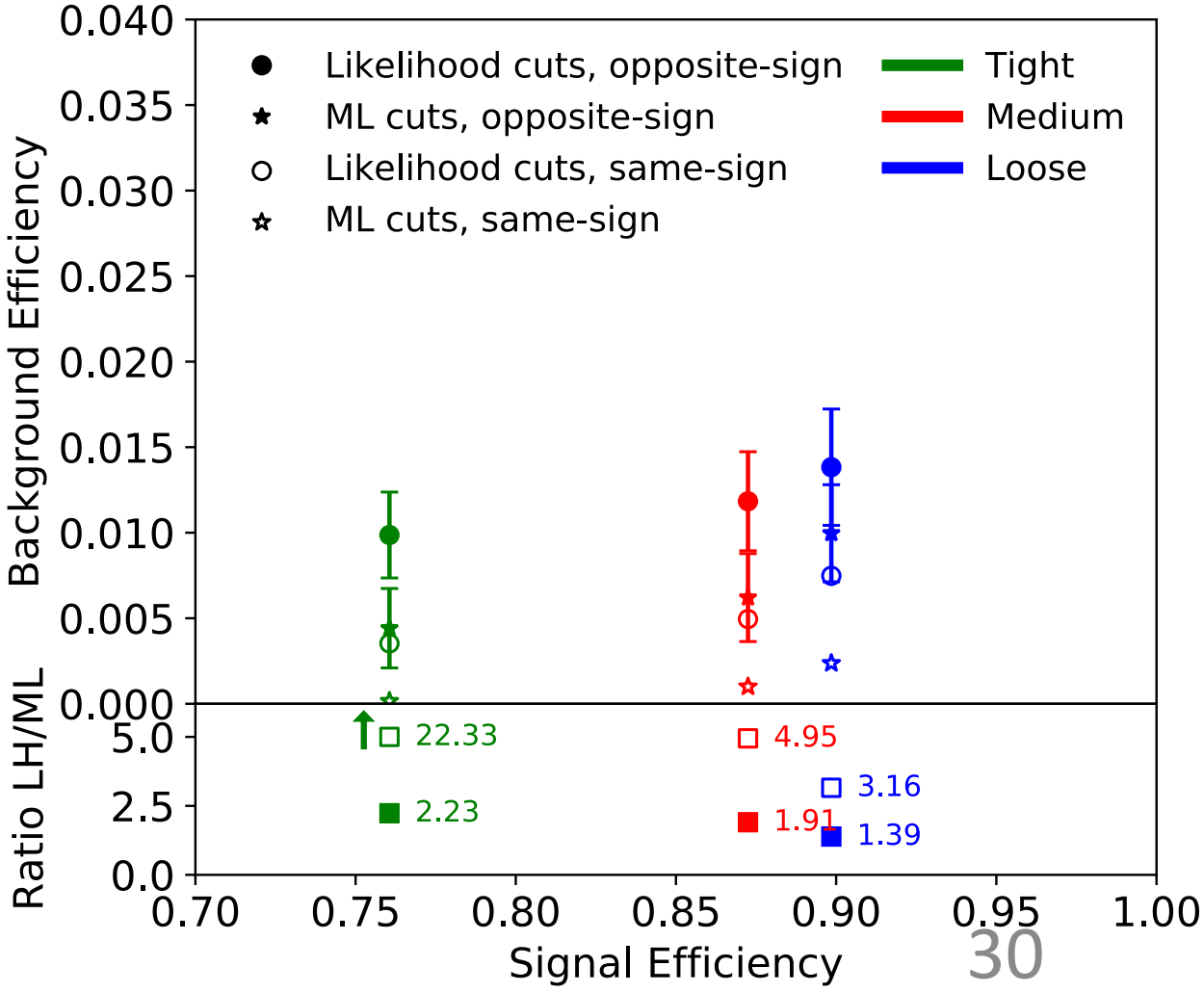
We look at the signal efficiency for LH cuts and cut in our LGBM score to get the same signal efficiency

We can compare the background efficiencies

For same-sign, we have no signal, but we make the same cut in LGBM as for opposite-sign and count how much background is left.

Ratio plot show the decrease in background at the same signal efficiency

The same plot should be made in MC to compare – we need to get a bigger background sample before this can be done.



Conclusion

- The LightGBM model improves the Z selection for both MC and Data
- We will try to get ML PID/ISO trained on Data as input to the Zmm Data model

Next step

- Create a ML model on $Z \rightarrow mm\gamma$ and $H \rightarrow Z(mm)\gamma$

Bonus slides

MC model on Zmm: signal selection and reweighing