Higgs to four leptons using nanoAODplus

Paula Martínez Suárez[1]
[1] *Supervised by Achim Geiser and Nur Zulaiha Jomhari*

Final report of the project developed during the DESY Summer Student Programme 2019

A validation of several nanoAODplus data sets from 2011 and 2012 (LHC Run 1) is carried out in order to reproduce the Higgs to four leptons plot available at the CERN Open Data Portal, created from AOD simplified data. All the variables are succesfully validated or at least partially validated, and the Higgs to four leptons plot can be reproduced with very slight differences.

WHAT IS nanoAODplus?

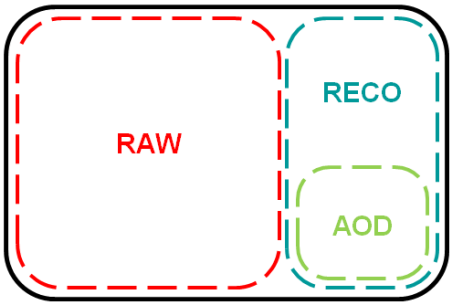


Figure 1: *RAW, RECO and AOD data tiers. While RECO contains different information from RAW about the event, AOD is a subset of RECO.*

The format of the data is nothing else than a C++ class which defines the structure of the data, and contains information about a particular attribute of an event. For each event, we can write the information contained in it in a data tier, which can be composed of more than one data format. Different data tiers will include different levels of information from the event. For example, the RAW data tier contains the detector output collected after the trigger selection (L1 and HLT). The RECO data tier goes one step further and uses the detector information to reconstruct physics objects (particles, jets, vertices, tracks) and hits/clusters, for the subsequent analysis. All this information is not necessary to perform most of the studies, so usually the working data tier is a subset of RECO, called AOD, which contains all the reconstructed objects, but not the detector information. For the sake of simplicity, from now on we will refer to the *data tiers* as *formats*, keeping in mind all the previous definitions.

The AOD format is the one used for analysis in Run 1 (2010-2012), but it is too slow for Run 2 (2015-2018) due to its large size per event ($\sim$ 400 kB), because the amount of produced data increases, and with it the resources that we need to process it. In this case a subset of the AOD, the miniAOD, is preferred, since its size per event is only a 10% of the AOD $\sim$ 40 kB).

Apart from AOD and miniAOD there is another format, the nanoAOD, only available from 2016 onwards. Its size per event is even smaller than in the miniAOD ($\sim$ 1 kB) because it only contains the information needed in the most generic analyses. The main advantage of this format is that it does not require a CMSSW environment, and can be read with plain ROOT. Since nanoAOD is only available for Run 2, a nanoAODplus data set format is being developed for Run 1. It is called 'plus' because it includes additional variables, that can help to debug or understand better the code. However, since the aim is to produce an exact replica of the official nanoAOD from Run 2, this variables should be removed in the end.

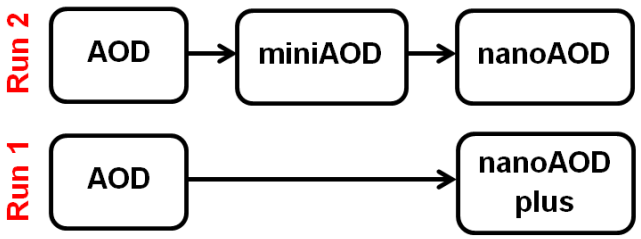


Figure 2: *Since the AOD format is the only available for Run 1, it is the starting point to produce the nanoAODplus.*

THE IMPORTANCE OF THE HIGGS BOSON

The Higgs boson is an elementary particle in the Standard Model, named after Peter Higgs, one of the physicists who proposed the Higgs mechanism in the 1960s to justify why some particles have mass.

The Higgs mechanism, when added to the Standard Model, establishes that at high enough energies all the electroweak force carriers are massles, but when a critical temperature is reached the W and Z bosons acquire a non-zero mass, due to what is known as electroweak symmetry breaking.

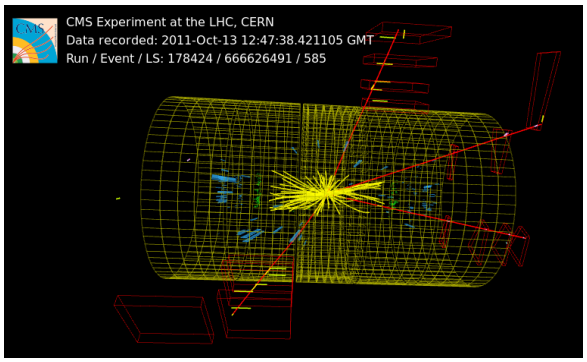


Figure 3: *Higgs decay to four muons in the CMS detector (Run 1, 2011).*

Besides this, when added to the Standard Model lagrangian, the Higgs field gives mass to the fermions (except for the neutrinos) through a Yukawa coupling. This coupling is proportional to the mass of the particle, which means that the Higgs boson will interact more strongly with heavy particles as, for example, the top or bottom quarks. This phenomenon is different from the Higgs mechanism; the coupling of the Higgs with the vector bosons is proportional to the mass squared.

This theory started to be developed in the 60s, but it was not confirmed until the Higgs boson discovery in 2012, by the CMS and ATLAS collaborations at the LHC.

The Higgs to four leptons decay channel

The Higgs discovery is the result of analysing five decay channels: $\gamma\gamma$, ZZ, WW, $\tau\tau$ and bb. The Higgs to ZZ to four leptons decay channel is one of the best options to reconstruct the mass peak of the Higgs boson. The reason is that leptons can be measured with high precision at CMS and ATLAS, and therefore the resolution of the peak is very good. Moreover, the background for this channel is not very significant, and it does not contain neutrinos that give rise to missing transverse energy.

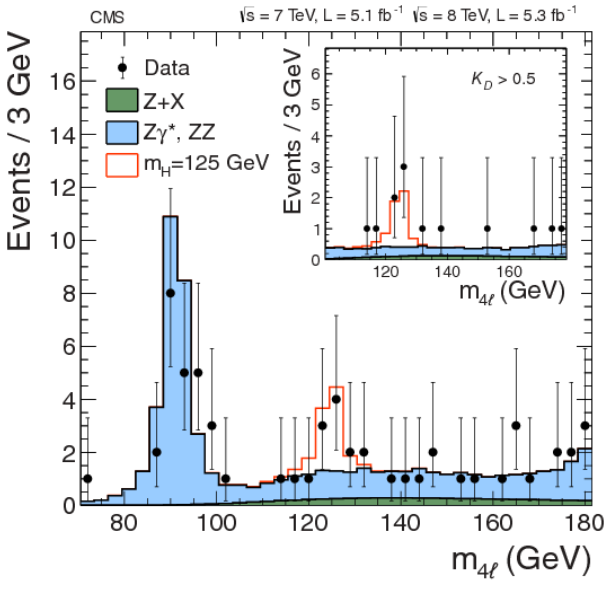


Figure 4: *Distribution of the four leptons invariant mass for the $ZZ \to 4\ell$ analysis. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass $m_H = 125$ GeV, added to the background expectation [4].*

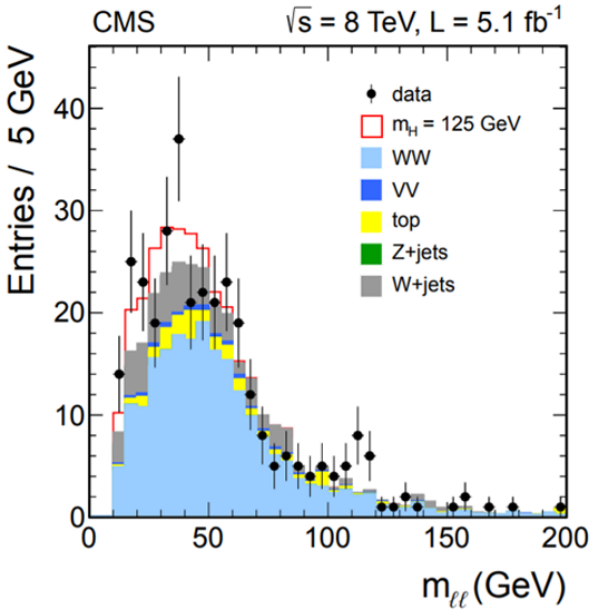


Figure 5: *The Higgs to WW to $2\ell 2\nu$ channel is not as clear as the ZZ to 4ℓ channel due to the neutrinos, which can not be detected [4].*

THE CERN OPEN DATA PORTAL

The CERN Open Data Portal (`http://opendata.cern.ch/`) is a platform which allows public access to the data produced by the different research activities at CERN. For CMS it contains simplified data sets, reconstructed data and simulations and the necessary analysis software. In this webpage we can find the Higgs to four leptons plot that we mentioned earlier, produced from 2011 and 2012 AOD data. As we will see, it is slightly different because:

- The Open Data example uses legacy versions of the original CMS data sets as well as the Monte Carlo simulations, while the original publication contains improved data.
- Only 50% of the of the Run 1 data is public, so the statistic significance is not the same.

The aim of this project is to reproduce the same plot using the nanoAODplus format. This is a non trivial task, since the nanoAOD format type can contain internal cuts to reduce the size per event. In figure 7 we can see one of them, which is applied to the transverse momentum of the electrons.

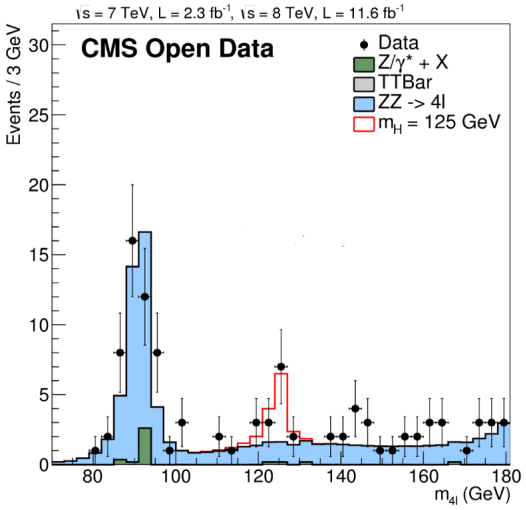


Figure 6: *Distribution of the four leptons invariant mass for the $ZZ \rightarrow 4\ell$ analysis in the Open Data example [2].*

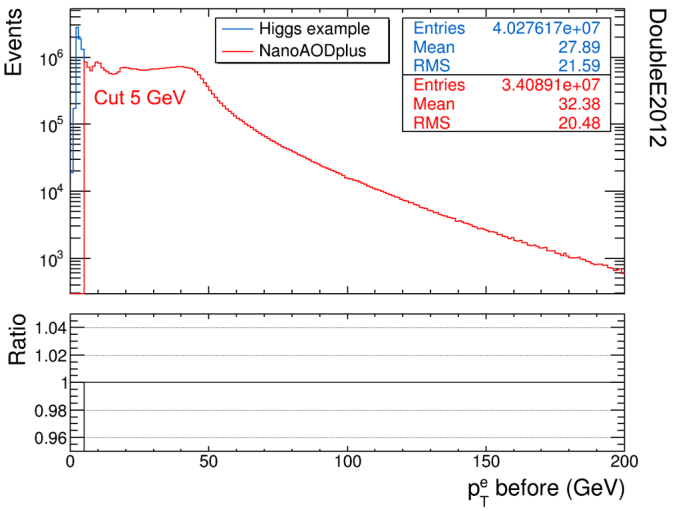


Figure 7: *p_T of electrons before applying any cut (2012 double electron data set). This picture shows a cut for $p_T < 5$ GeV in the nanoAODplus which is not present in the AOD data (Higgs example).*

DESCRIPTION OF THE PROJECT

As we mentioned before, our purpose is to reproduce the four leptons invariant mass plot that shows the Higgs boson's mass peak using nanoAODplus data. To do so, we need to perform the following steps:

- Match the variables from the Higgs example code and the nanoAODplus code. Most of them have different names, but contain the same information. This procedure also helps us to find missing variables, defined in the Higgs example but not in the nanoAODplus.
- Verify that they are correctly defined and identify internal cuts in the nanoAODplus. A way of doing this is to plot the variables from both sides before having applied any cuts.
- Find all the cuts used to create the Higgs to four leptons histogram in the Higgs example and reproduce them for the nanoAODplus.

Selection cuts

We will now see the necessary cuts to reproduce the Higgs to four leptons plot. The purpose of these cuts is to discard as many background events as possible, although there will be some irreducible background processes as well, as for example the direct ZZ production. First of all, we require some quality cuts to ensure that our muons and electrons are properly identified: the muons have to be global muons, and both muons and electrons must be particle flow

candidates. Next, we apply the kinematic cuts, which refer to variables related to the energy, direction and position of the particles. We want muons with $p_T > 5$ GeV and $|\eta| < 2.4$ and electrons with $p_T > 7$ GeV and supercluster $|\eta| < 2.5$, within the barrel or endcap acceptance and with a number of misshits less than or equal to one. For both muons and electrons we also require the impact parameter significance (SIP_{3D}) < 4, a distance to the vertex < 0.5 in the transverse plane and < 1 in the z direction and a relative isolation $\Delta R < 1$. Once selected the 'good' muons and electrons (the ones that survive all the cuts) we can define the final state. For the Higgs to four leptons plot we require four leptons (two pairs of same-flavour leptons) in the final state, with null total charge. As each pair of leptons should come from a Z boson, we also need that the charge of each pair of leptons equals zero.

RESULTS

The results from this project can be divided in two subsections. Firstly, we will use some control plots to validate the most relevant variables. Then, we will produce the Higgs to four leptons plot and compare it with the previous ones.

Control plots

The variables that we need to validate come from twelve different data sets. Four of them correspond to measured data, and eight of them to Monte Carlo simulations. Apart from these ones, we have six additional data sets for Monte Carlo simulations, corresponding to the Drell-Yan and $t\bar{t}$ backgrounds that are the same as in the Open Data example, but will need to be reproduced in the future.

Data 2011	DoubleMu
	DoubleE
Monte Carlo 2011	HZZ
	ZZto4mu
	ZZto2mu2e
	ZZto4e
	DY10l1
	DY50TuneZ11
	TTBar11
Data 2012	DoubleMuParked
	DoubleEParked
Monte Carlo 2012	HZZ
	ZZto4mu
	ZZto2mu2e
	ZZto4e
	DY10l2
	DY50TuneZ12
	TTBar12

Figure 8: *Data sets used to produce the Higgs to four leptons plot. The ones written in gray are the same as in the Open Data example.*

These data sets describe the following information:

- **DoubleMu**, **DoubleE**, **DoubleMuParked**, **DoubleEParked**. Data collected in the CMS detector in 2011 and 2012 during Run 1. The muon data sets take muons as signal and electrons as background, and vice versa.
- **HZZ**. This data sets refer to the Monte Carlo simulation for the Higgs to ZZ to four leptons, for the case of a Higgs boson with a mass of 125 GeV.
- **ZZto4mu**, **ZZto2mu2e**, **ZZto4e**. These Monte Carlo simulations describe the most relevant background for this process, the direct ZZ production and its decay to four muons, two muons and two electrons or four electrons.
- **DY10**, **DY50TuneZ**. Drell-Yan background processes.
- **TTBar**. $t\bar{t}$ background processes.

While comparing the histograms, we will consider a variable validated if it is equal to the corresponding one in the Higgs AOD example and partially validated if it is similar but not exactly the same. Even though ideally we would want to have exactly the same variables, a partial validation can be enough for our purposes if we understand why the variable is different.

For the double muon 2011 data set, most of the variables before the cuts[1] match perfectly. There are small differences in the distances to the vertex (dxy and dz) and in the impact parameter significance (SIP_{3D}), that can come from a different definition of the primary vertex. These differences are not too significant, because when we plot the after variables we can see that they are similar enough to get reasonable results.

[1] We will use the name 'before the cuts' when only the quality cuts have been applied, and 'after the cuts' when we have all the cuts.

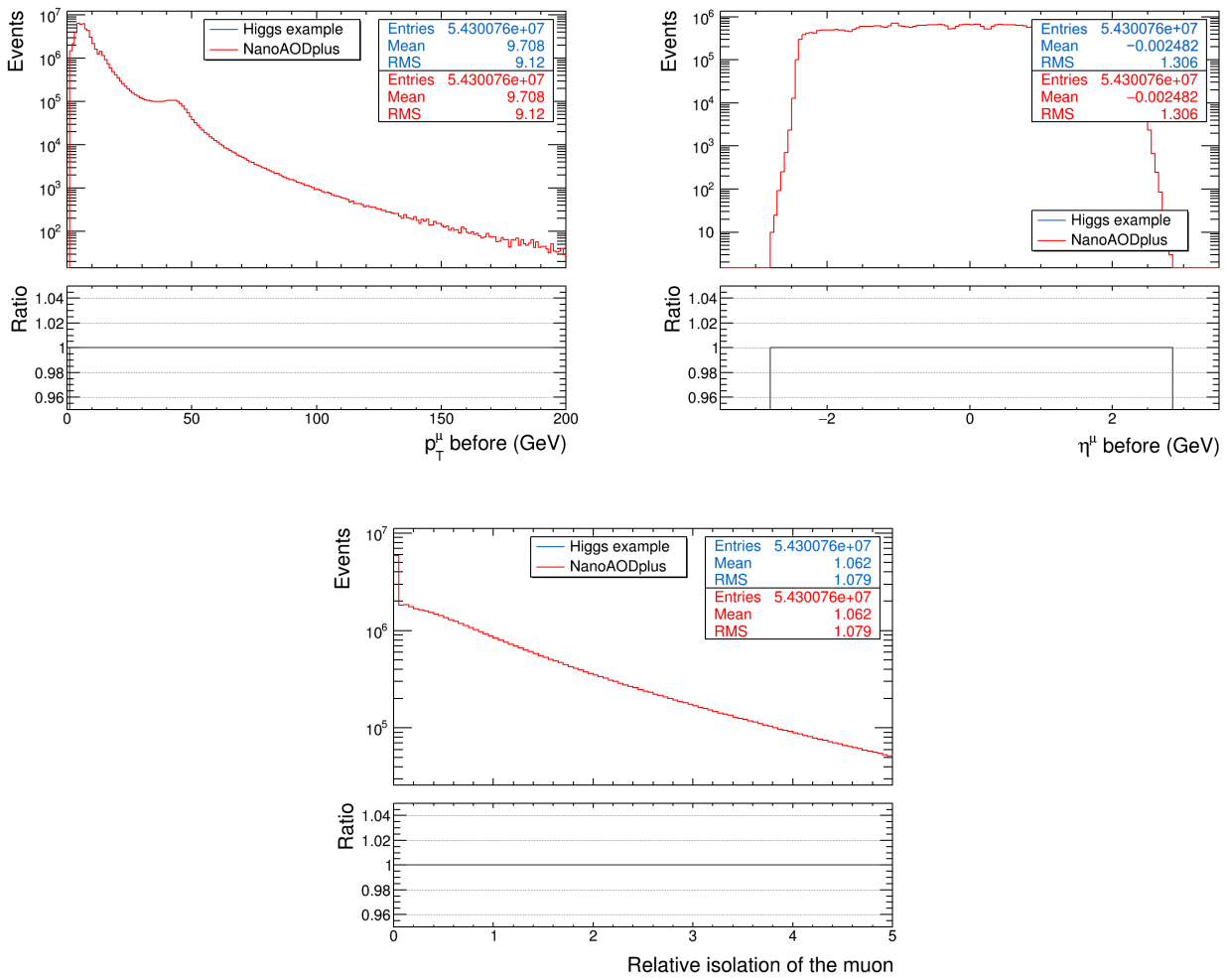


Figure 9: *Transverse momentum, pseudorapidity and relative isolation of the muons before the cuts. These variables match perfectly. 2011 double muon data.*

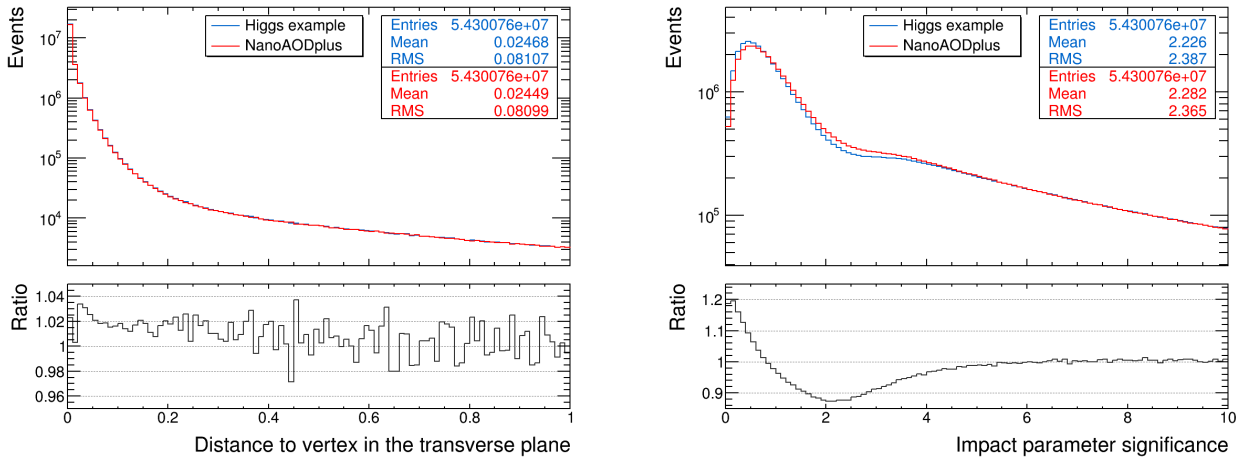


Figure 10: *Distace to the primary vertex in the transverse plane and impact parameter significance before the cuts. These variables are slightly different. They are partially (but not fully) validated. 2011 double muon data.*

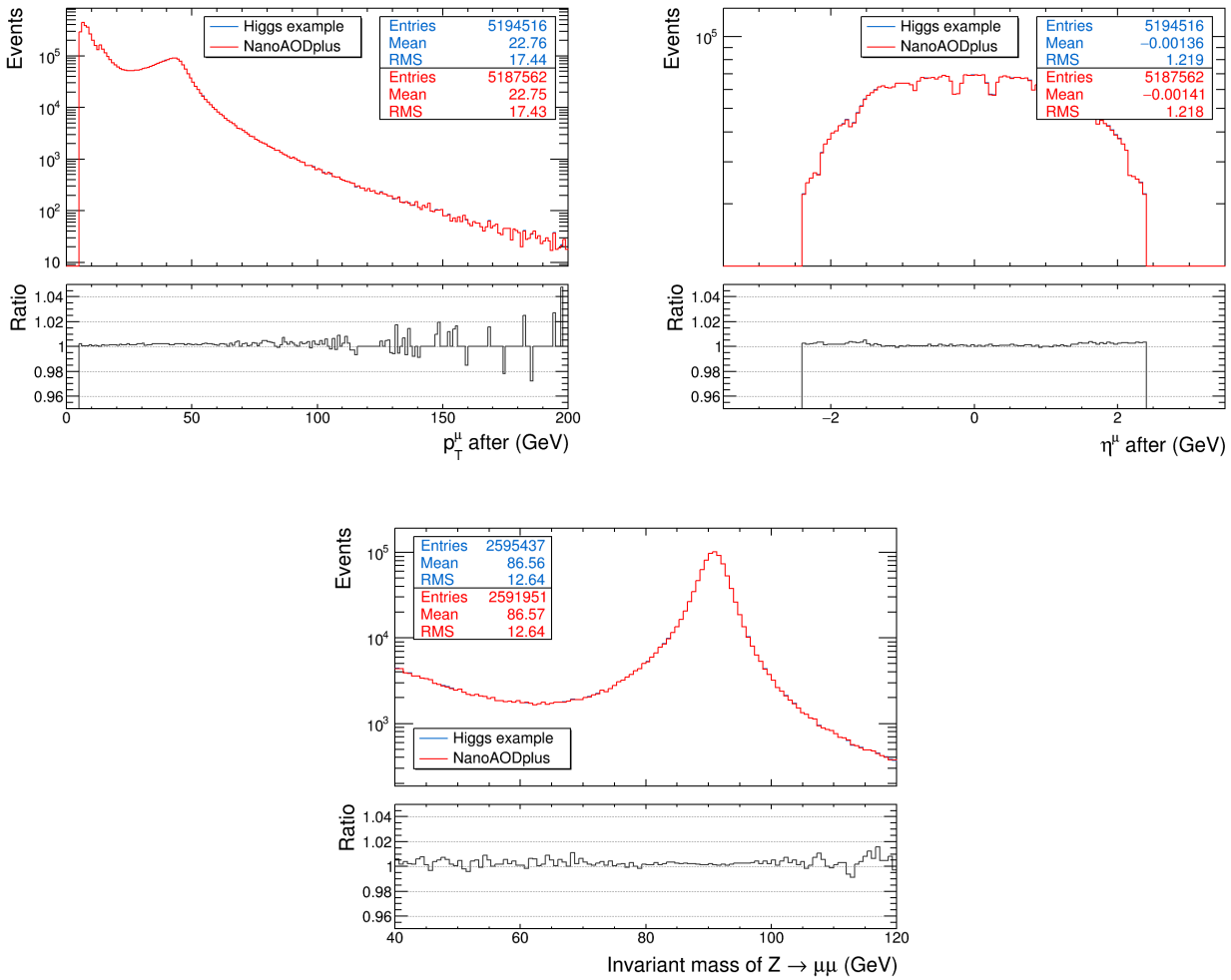
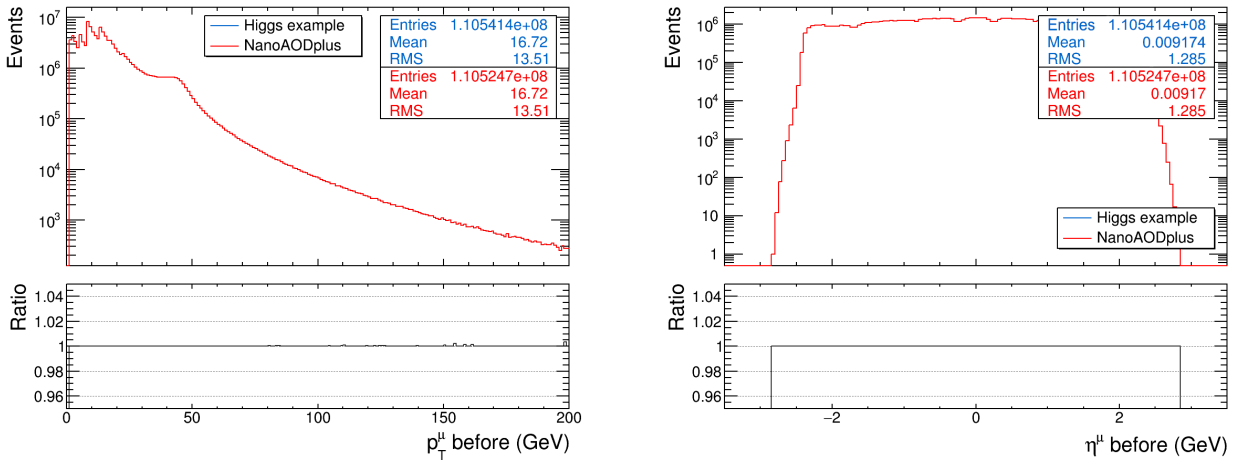


Figure 11: *Transverse momentum and pseudorapidity after the cuts, and invariant mass of dimuon. The differences are not too large. 2011 double muon data.*

When we move to the 2012 data set, we see that the before variables are slighly different. This happens due to the way in which we define the arrays containing the data for this set, but we can check that this becomes irrelevant once all cuts are applied.



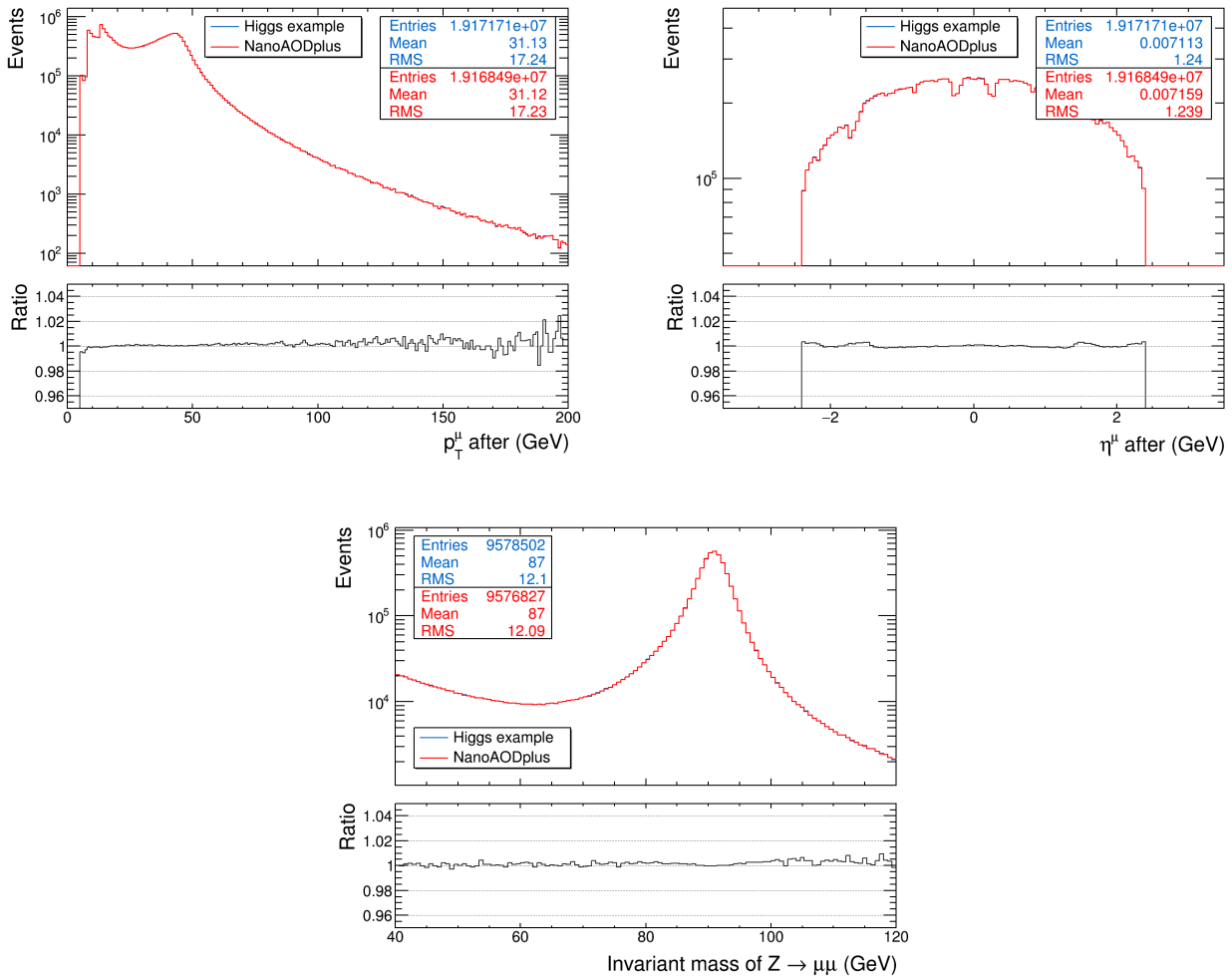
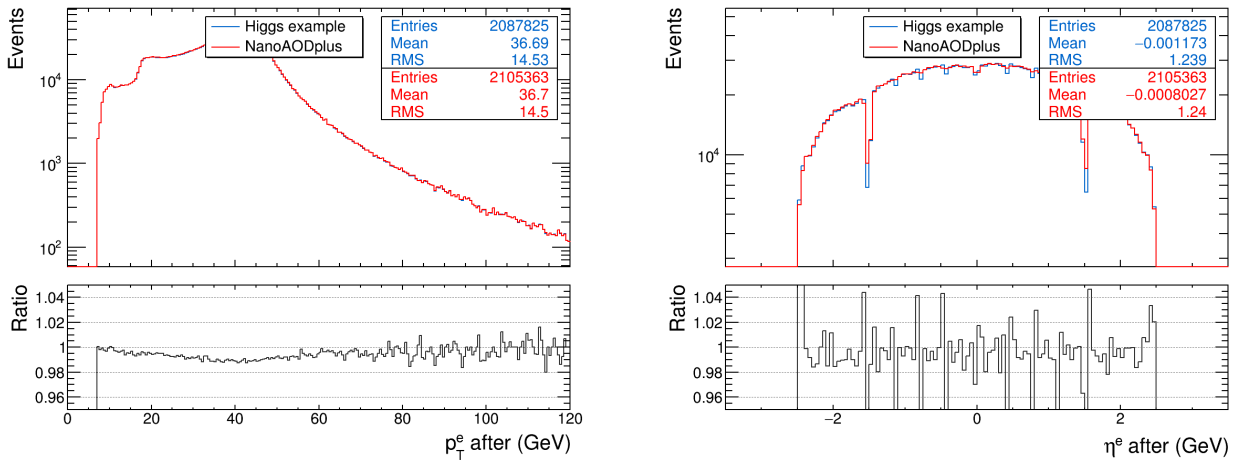


Figure 12: *Transverse momentum and pseudorapidity before the cuts (previous page) and after the cuts (above), and invariant mass of the dimuon. Again the order of the differences is small. 2012 double muon data.*

In the case of the electron data sets we find an additional problem, which is that the nanoAODplus variables have an internal cut in the electron p_T that is not present in the AOD, as we see in the figure 7. For this data sets we can produce the histograms without any cuts from the program that creates the nanoAODplus and check that the variables are effectively the same, except for the distances to the vertex (dxy and dz) and in the impact parameter significance (SIP_{3D}). Here we will only show the variables after the cuts, to prove that they are very similar to the Higgs example.



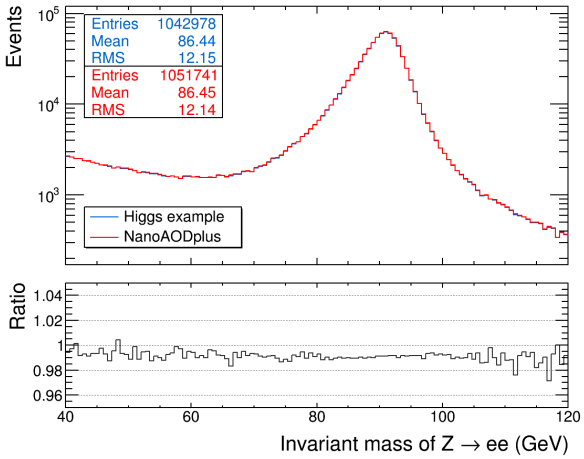
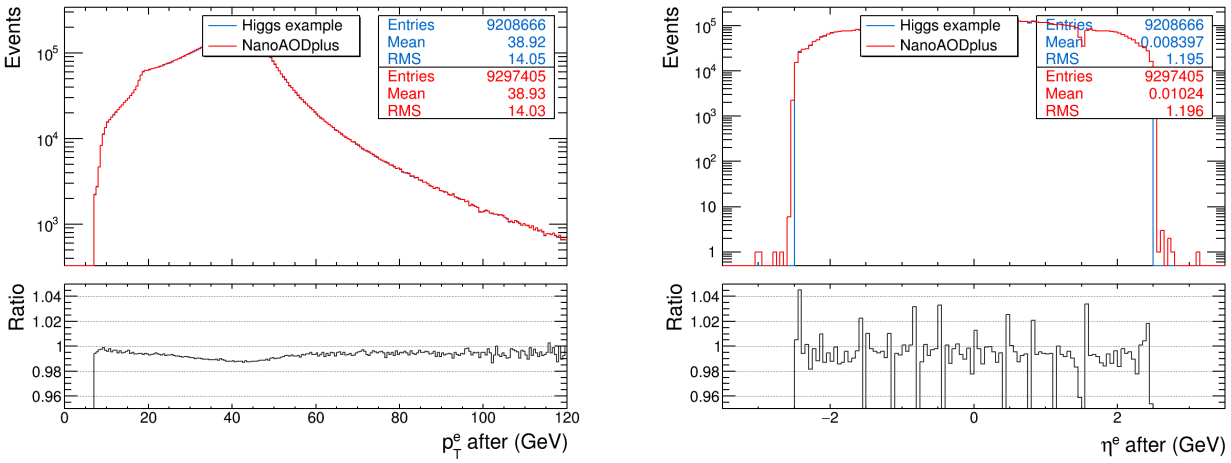


Figure 13: *Transverse momentum and pseudorapidity after the cuts (previous page) and invariant mass of the dielectron (above). 2011 double electron data.*

The same thing happens for the double electron 2012 data set.



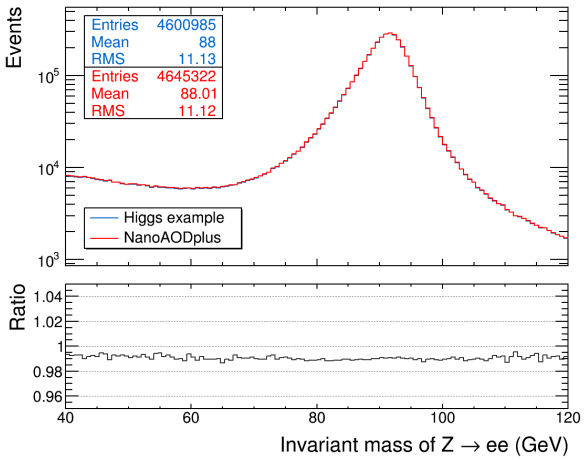


Figure 14: *Transverse momentum and pseudorapidity after the cuts and invariant mass of the dielectron. 2012 double electron data.*

Following the same procedure we can also validate or partially validate the variables in the Monte Carlo data sets. The agreement is good.

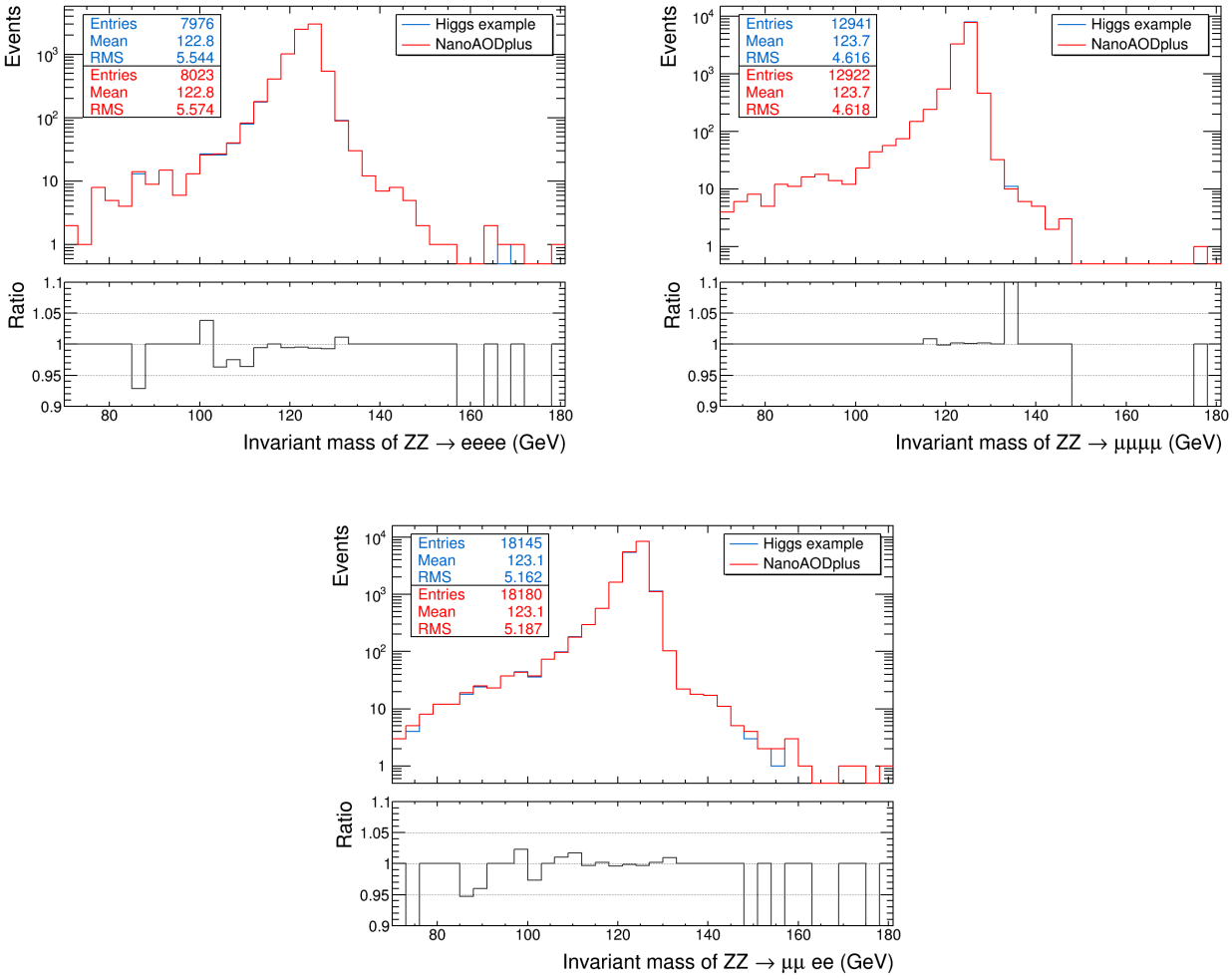
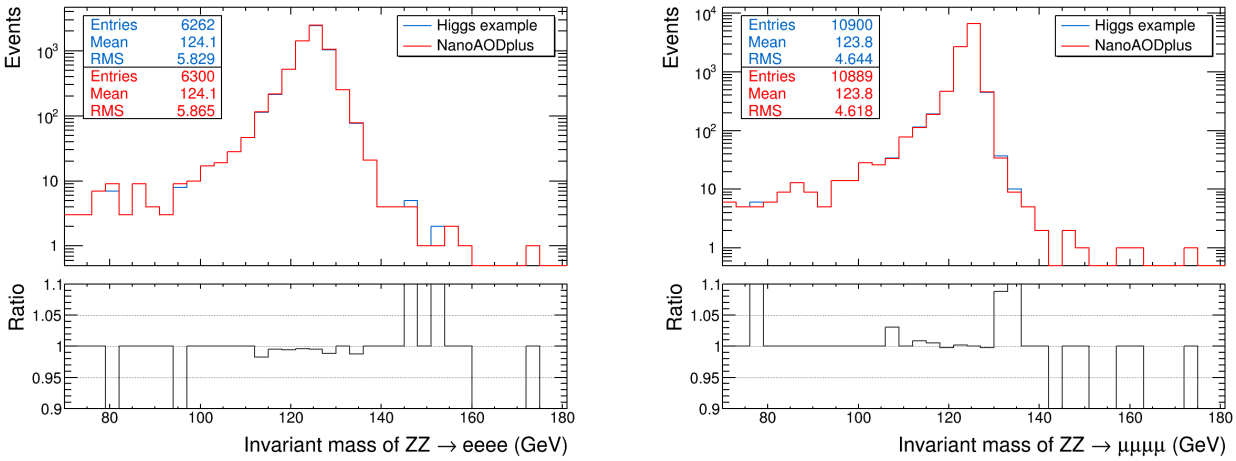


Figure 15: *Invariant masses for different decays of the ZZ pair in the 2011 HZZ Monte Carlo sample.*



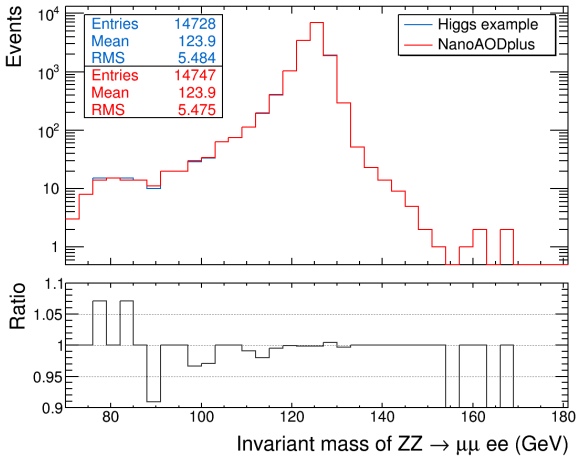


Figure 16: *(From previous page) Invariant masses for different decays of the ZZ pair in the 2012 HZZ Monte Carlo sample.*

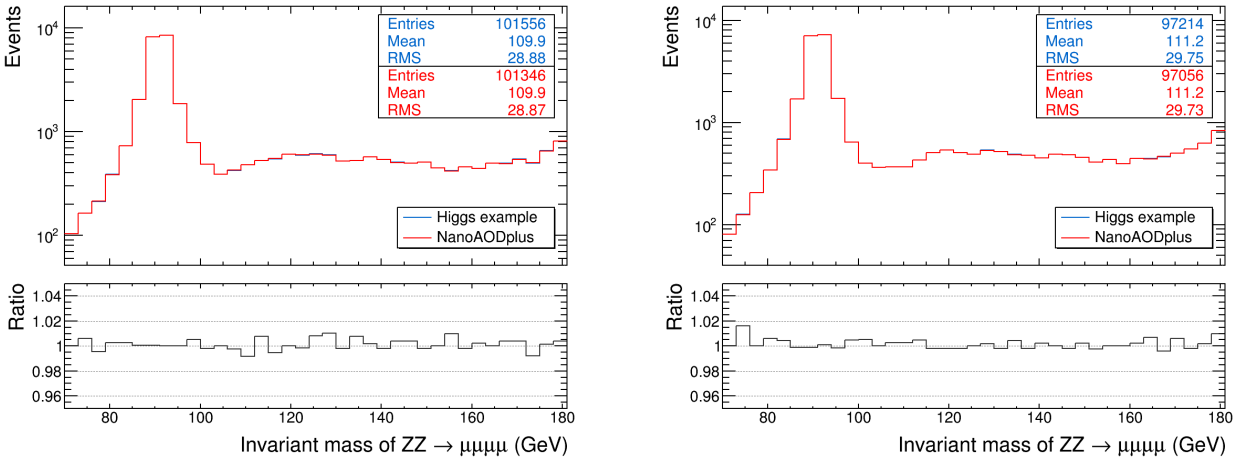


Figure 17: *Invariant mass of $ZZ \rightarrow \mu\mu\mu\mu$ for the 2011 (left) and 2012 (right) ZZto4mu Monte Carlo sample.*

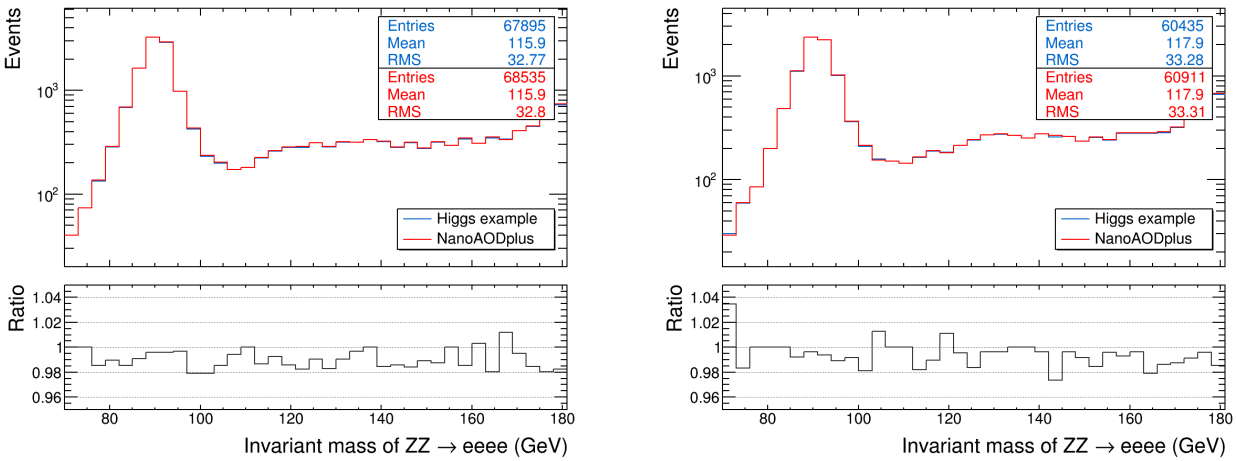


Figure 18: *Invariant mass of $ZZ \rightarrow eeee$ for the 2011 (left) and 2012 (right) ZZto4e Monte Carlo sample.*

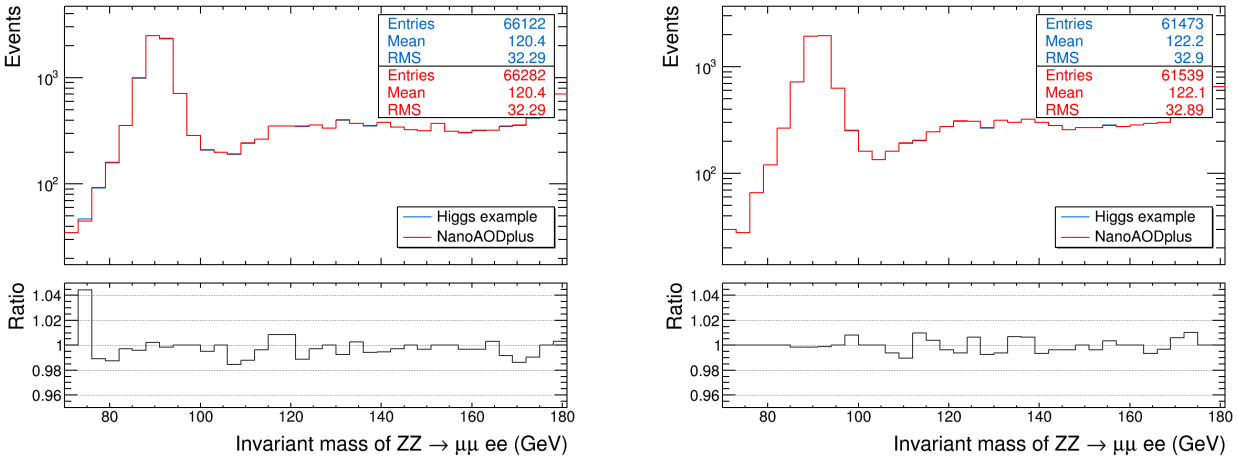


Figure 19: *Invariant mass of $ZZ \rightarrow \mu\mu ee$ for the 2011 (left) and 2012 (right) ZZto2mu2e Monte Carlo sample.*

The reason why we show these mass plots is that they (among others) are the ones that we will later use to produce the Higgs to four leptons mass plot.

The Higgs to four leptons plot

The reason for validating all this data sets is that now we can use them to produce the Higgs to four leptons plot. To do so, we need to add all the decay channels for each process. We have data, Higgs to ZZ, ZZ, Drell-Yan and $t\bar{t}$. For each of them we need to calculate a scale factor, as the Monte Carlo is usually produced with higher statistics to avoid fluctuations. The results are shown in figure 20.

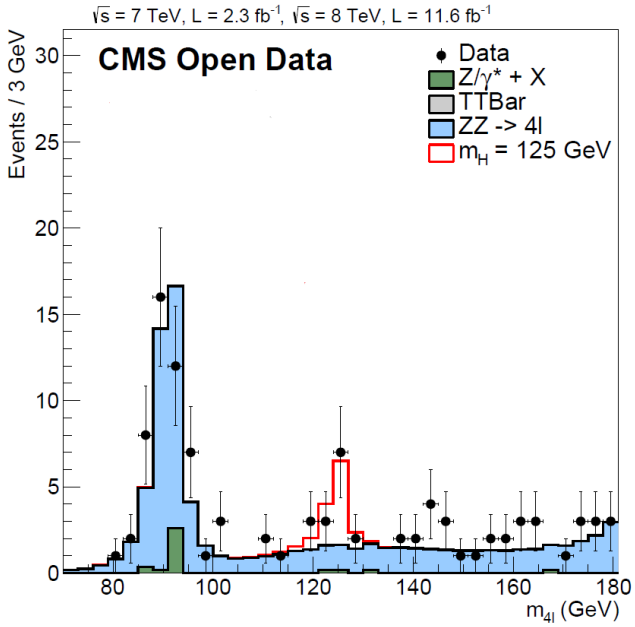


Figure 20: *Higgs to four leptons plot produced from nanoAODplus data. Even though there are some differences, the result reproduces very well the Open Data example.*

DoubleMu	$ZZ \rightarrow \mu\mu\mu\mu$
	$ZZ \rightarrow \mu\mu ee$
DoubleE	$ZZ \rightarrow eeee$
HZZ	$ZZ \rightarrow \mu\mu\mu\mu$
	$ZZ \rightarrow eeee$
	$ZZ \rightarrow \mu\mu ee$
ZZto4mu	$ZZ \rightarrow \mu\mu\mu\mu$
ZZto4e	$ZZ \rightarrow eeee$
ZZto2mu2e	$ZZ \rightarrow \mu\mu ee$

Figure 21: *Datasets used to fill the different decays of the ZZ.*

Except for Drell-Yan and $t\bar{t}$, the rest of the datasets contain information about different leptonic decays of Higgs to ZZ, but they are not valid in general to describe the three of them (4μ, $4e$ and $2\mu\,2e$). The double muon datasets are used for 4μ and $2\mu\,2e$ decays, and the double electron for $4e$. The HZZ Monte Carlo samples are valid for all the decays, whereas ZZto4mu is only used for 4μ, ZZto4e for $4e$ and ZZto2mu2e for $2\mu\,2e$. A summary of this can be seen in figure 20.

CONCLUSIONS AND OUTLOOK

The agreement between AOD and nanoAODplus is good, even perfect in some cases, but it can still be improved. The differences can come from slightly different definitions or small bugs. Some of them could be inevitable if the official nanoAOD contains different object definitions from the AOD, since our goal is not to reproduce exactly the AOD results, but reproduce them as well as possible using nanoAOD variables.

The nanoAODplus data type is currently under development, so it is still not the same as the official nanoAODplus. Our next step is to remove all the extra variables, so that it looks exactly as tue official nanoAOD. Once done this, it can be used to analyse Run 2 data and be directly compared with the official nanoAOD results for further validation.

REFERENCES

References

[1] CERN Open Data Portal (`http://opendata.cern.ch/`)

[2] CERN Open Data Portal, Higgs-to-four-lepton analysis example using 2011-2012 data (`http://opendata.cern.ch/record/5500`)

[3] CERN Open Data Portal, Root files for Higgs-to-four-lepton analysis example using 2011-2012 data (`http://opendata.cern.ch/record/5501`)

[4] The CMS Collaboration, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC (arXiv:1207.7235)

[5] G. Petrucciani, A. Rizzi and C. Vuosalo. Mini-AOD: A New Analysis Data Format for CMS (arXiv:1702.04685)

[6] WorkBookNanoAOD (2019-09-13, AndreyPopov) (`https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD#NanoAOD_format`)