24/05/2024

Possible detection of the Higgs decay into muons

Computing methods in High Energy Physics 2024

Shankar Bhandari

Jarno Vierros

Contents

1 Introduction

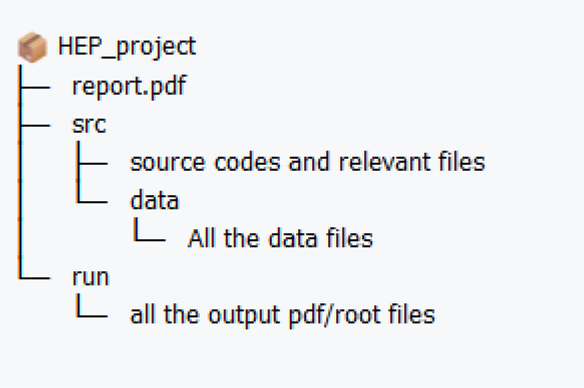
The topic of this final project for the course is investigating the possible detection of Higgs particles decaying into muon anti-muon pairs during run 3 of the LHC. To investigate this we used particle physics event generation simulations. The program we used for the simulation is PYTHIA 8.3 [1].

Since we wanted to study the $H \to \mu^-\mu^+$ channel, we also had to consider the main background processes. The main background processes for $H \to \mu^-\mu^+$ are the Drell-Yan process and $t\bar{t}$ production. This means we had to simulate not only the production of muons from Higgs decay but also the background processes.

After we were done with the simulations and we had the raw particle data, we organised the data for it to be ready for analysis. During the organising, we also simulated the measurement uncertainty by applying 1% Gaussian smearing to the momenta and 2 mrad Gaussian smearing to the angles θ and ϕ of the muons. After smearing we did the data analysis.

To do the simulations, smearing and analysis we built custom C++ programs. The instructions for their usage are on the README file in their directory. We also use the ROOT framework to store the data and to do the analysis. All of the programs, input file(s) and output file(s) are in the file structure as presented in the figure (1). In the GitHub, we have all of the files except for the data files and executables.

We simulated the events for the DY process in different stages of working on the project work with some changes to how the program works. This was because the initial amount of simulations (20 million events) that took 2 days did not give a smooth enough background for the analysis. We also had to change how the code works due to bugs in how the data was being stored, for both the DY and $t\bar{t}$ programs. Although there was a bug on how the data was being stored, it was possible to fix it so we created a program to fix the simulated data.



```
HEP_project
├── report.pdf
├── src
│   ├── source codes and relevant files
│   └── data
│       └── All the data files
└── run
    └── all the output pdf/root files
```

Figure 1: File structure of the project

2 Simulation and the custom programs

2.1 Simulating using PYTHIA

We made three separate custom programs to simulate the muon production from Drell-Yan, $t\bar{t}$ and Higgs decay using PYTHIA. To get the Drell-Yan muons, we only turned on the process described in the Feynman diagram in figure (2). In the diagram, f is a fermion and $\bar{f}$ is an anti-fermion and the process happens only via virtual photon/Z-boson. To get the muons from $t\bar{t}$-production we turned on all $t\bar{t}$-production. To get the muons from Higgs decay, we turned on all of the ways to produce the Higgs boson.

For the top production, at first we accidentally only had $f\bar{f} \to t\bar{t}$ process turned on, which didn't represent all top production. However, the shape of the distribution in the resulting data was close to other top quark background distributions we found on-

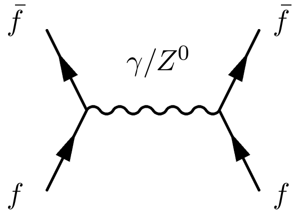


Figure 2: Feynman graphs for the process that was turned on to simulate the Drell-Yan production

line, such as (https://medium.com/@yasunsafak/top-quark-analysis-using-hep-tutorial-f709376d5fb3). Therefore, we considered it acceptable to use the $f\bar{f} \rightarrow t\bar{t}$ dataset to represent all top production processes. This choice was also motivated by the fact that once we realized our mistake, there was no time left to regenerate all our data. However, we did simulate $100,000$ general top production processes to determine the total top production cross section for our normalization.

From the simulations, we selected final state muons and stored the following data: components of the three momentum, energy, mass, charge and whether it came from Higgs production or not. The last part is not necessary for the simulation analysis, but it is there to make it easier to debug any problems that arise.

When we added smearing to the data, we also replaced the components of the three momentum and energy with total momentum, transverse momentum, pseudorapidity and the polar angle of the three momentum. This was done because pseudorapidity and transverse momentum are more useful variables for physics analysis than the three momentum and they match the trigger conditions.

2.2 custom programs

We built many custom programs to process and analyze the data. In total, we have made ten programs for various purposes, including generating, processing, reading, analysing and repairing data. The repairing program is for fixing data originating from simulations effected by bugs in the simulation programs (that have since been fixed!). The instruction on how to use the programs are explained in the README of the folder they are in.

3 Results and analysis

3.1 Results

At first, we simulated $10 \cdot 10^6$ Higgs decay events, $20 \cdot 10^6$ Drell-Yan events and $20 \cdot 10^6$ $t\bar{t}$ production events. We quickly realized that the signal was very weak compared to the Drell-Yan background, so we would need to have an extremely smooth Drell-Yan curve to see the contribution of the signal. In practice, this meant simulating as many Drell-Yan events as possible.

Only Drell-Yan events producing muons would pass the trigger and contribute to the smoothness of our background. Furthermore, events with viable pseudorapidity and transverse momentum seemed to originate predominantly form Drell-Yan processes mediated by a Z boson rather than a photon. Therefore, we decided to disable the photon-mediated Drell-Yan process as well as all decays of the Z boson except the muon decay. This meant that every Drell-Yan event simulated resulted in muons in the final state, which dramatically increased the rate at which we were able to generate Drell-Yan events that would pass the trigger.

We ran many additional simulations with this new setup, which generated a total of $46 \cdot 10^6$ events. All of these events involved muon production. Although this was far from a minimum bias simulation, this did not prevent us from normalizing the data, since we had gained all necessary information for normalization from the earlier simulations.

3.2 Trigger efficiency

Out of the $10 \cdot 10^6$ Higgs bosons simulated 2196 decayed into muons, which is consistent with the value predicted by the Particle Data Group's branching ratio: 2600 ± 1300 [2]. Out of these 1248 events contained at least two muons with $|\eta < 2.1|$ and $p_T > 20$ GeV, which gives the following trigger efficiency: 56.83%.

The number of events passing the selection for other datasets is shown in table (1). Passing event counts after normalization are also shown.

Process	selected events	normalized events
Drell-Yan	$21,123,005$	$159,973,659$
ttbar	$237,119$	$2,623,840$
Signal (H)	1238	1295

Table 1: The number of events passing the selection for each dataset both before and after normalization.

3.3 Normalization

The number of events expected to happen in a collider experiment can be calculated via the following formula:

$$N = \sigma \int \mathcal{L}(t)dt \quad , \tag{1}$$

where σ is the cross section of the process and the integral gives the integrated luminosity L which we assume to be 300fb^{-1}. We use the cross section given by PYTHIA after the simulation. The process cross section along with the expected number of events N and the coefficient used to normalize the Histograms is given in the table (2). The Normalization coefficient X is calculated with the following equation:

$$XN_{\text{pythia}} = N_{\text{LHC}} = N = \sigma \int \mathcal{L}(t) \quad , \tag{2}$$

where N_{pythia} is the number of events simulated in PYTHIA and the cross section is for the whole process.

The normalization coefficient found in equation (2) can be directly used to normalize the signal and $t\bar{t}$ events, since they are produced with the pythia process corresponding to the cross section with no additional modification. However, we have heavily tampered with the Drell-Yan process, which means the cross section is no longer accurate, so an additional step is needed. In our original unmodified simulation of $20 \cdot 10^6$ events, we found $38,613$ events passing the trigger. Based on the expected number of events happening in reality given in table (2), the number of events passing the trigger in reality

Process	$\sigma(mb)$	N (expected)	Normalisation coefficient X
Drell-Yan	$2.762 \cdot 10^{-4} \pm 1.471 \cdot 10^{-7}$	$8.286 \cdot 10^{10} \pm 4.413 \cdot 10^{7}$	7.5734
ttbar	$7.377 \cdot 10^{-7} \pm 1.219 \cdot 10^{-09}$	$2.2131 \cdot 10^{8} \pm 3.657 \cdot 10^{5}$	11.0655
Signal (H)	$3.489 \cdot 10^{-8} \pm 2.459 \cdot 10^{-10}$	$1.046 \cdot 10^{7} \pm 7.377 \cdot 10^{4}$	1.046

Table 2: The cross sections (σ), expected events (N) and the normalization coefficient. These cross section are from PYTHIA simulations. The normalization coefficients for ttbar and the signal are calculated using Equation (2), while for the Drell-Yan process the normalization coefficient is calculated using Equation (3) due to the biased nature of the data used.

would be $159,973,659$. Therefore, we normalized the Drell-Yan events in such a way that after the normalization the amount of simulated events passing the trigger would match this value. In mathematical terms:

$$X_{DY} N_{\text{trigger, pythia}} = N_{\text{trigger, LHC}} = \frac{n_{\text{trigger, pythia}}}{n_{\text{pythia}}} N_{\text{LHC}} \quad , \tag{3}$$

where X_{DY} is the normalization coefficient used for Drell-Yan events, $N_{\text{trigger, pythia}}$ is the number of simulated Drell-Yan events passing the trigger, $N_{\text{trigger, LHC}}$ is the number of Drell-Yan events passing the trigger in reality, n_{pythia} is the number of Drell-Yan events simulated without forcing muon decays, $n_{\text{trigger, pythia}}$ is the subset of n_{pythia} that pass the trigger and N_{LHC} is the total number of Drell-Yan events happening in reality.

3.4 Analysis

We selected events with exactly two muons with opposite charges that pass the trigger and reconstructed these two muons into a single particle using the conservation of energy and momentum. We neglected events with more than two eligible muons since it would be difficult to determine which muons should be selected for reconstruction. We then extracted the invariant mass of this reconstructed particle and plotted it in Figure (3).

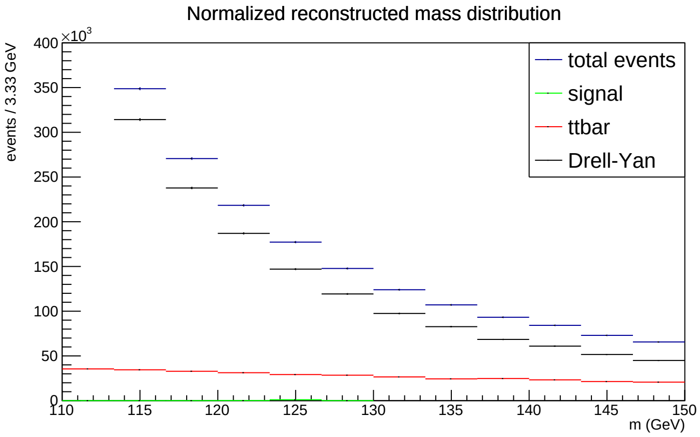


Figure 3: Normalized mass of the reconstructed particle. The colour means the following: black: Drell-Yan, red: $t\bar{t}$, green: signal, blue: background + signal

We then fitted the background with the sum of a Breit-Wigner distribution representing the Drell-Yan background and a first-degree polynomial representing the $t\bar{t}$ background as shown in Figure (4). The signal area 123.5 GeV – 125.5 GeV marked by the dotted line was ignored when fitting.

The signal is extremely weak compared to the background. Therefore, we attempted to remove the background using the background fit. Figure (5) shows the excess of events compared to what is expected based on the background fit. No significant peak is seen near the Higgs boson mass of 125 GeV, which indicates that the signal is weaker than the statistical variation in the background.

3.5 p_T constraint

To uncover the signal, the background must somehow be reduced. We extracted the pseudorapidity and transverse momentum of the reconstructed particle and studied their distributions for the background and signal. We noticed that on average the transverse momentum of the reconstructed particle was higher in signal events than in Drell-Yan events. Therefore, we decided to apply a constraint on the transverse momentum of the reconstructed particle. After some experimentation, we chose $p_T > 90$ GeV.

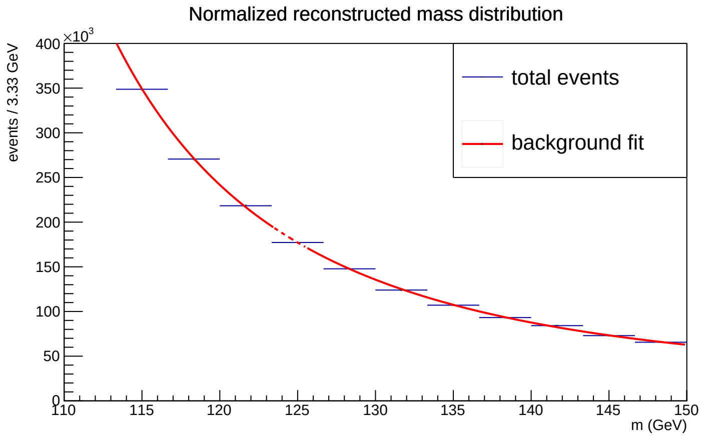


Figure 4: Fit of the background using a Breit-Wigner distribution and a first-order polynomial. The area with a dotted line was ignored during the fitting.

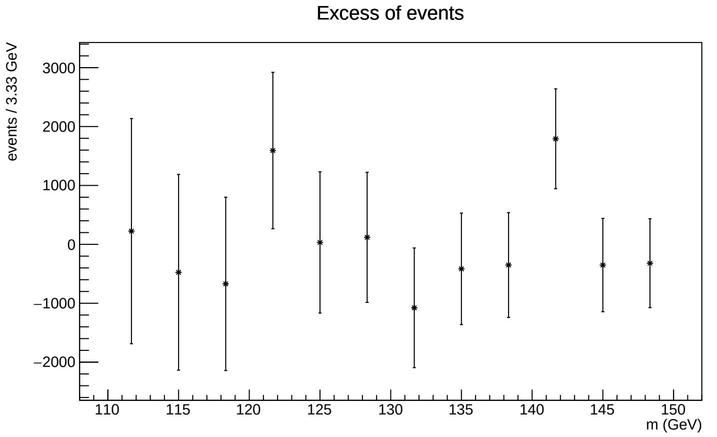


Figure 5: Excess of events compared to what is expected by the background fit.

Process	selected events	normalized events
Drell-Yan	$611,227$	$5,037,888$
ttbar	7002	$77,480$
Signal (H)	267	279

Table 3: The number of events passing both the trigger and the $p_T > 90$ GeV constraint for each dataset both before and after normalization.

Naturally, the constraint changes how many events pass the selection. The new values are displayed in Table (3). For the signal and $t\bar{t}$ events, the normalization coefficient remains the same. However, since the Drell-Yan normalization coefficient is determined based on the number of events passing the selection, it must be recalculated with equation (3) using the P_T constraint in addition to the trigger. The new Drell-Yan normalization coefficient is 8.2426.

Figures (6), (7) and (8) show the results of the analysis with the new transverse momentum constraint. The signal is still weak, but now it might be just barely visible. In Figure (8) the excess of events in the bin at 125 GeV is 561 ± 392.

We can calculate the statistical significance of this signal peak using a naive expression $\frac{N_S}{\sqrt{N_B}}$, where the number of signal events $N_S = 561$ is from Figure (8) and the number of background events $N_B = 15795$ at 125 GeV is from the background fit. Using the aforementioned values of N_S and N_B we get the statistical significance of $\frac{N_S}{\sqrt{N_B}} \approx 4.5$, i.e. 4.5σ. This is a surprisingly high statistical significance. However, it's not very accurate, since it doesn't take into consideration many sources of uncertainty in the analysis, e.g. the uncertainty in the background fit. For comparison, with this logic the statistical significance of the peak at 142 GeV in Figure (5) would be approximately 9σ despite the fact that it's almost certainly just a statistical variation in the background.

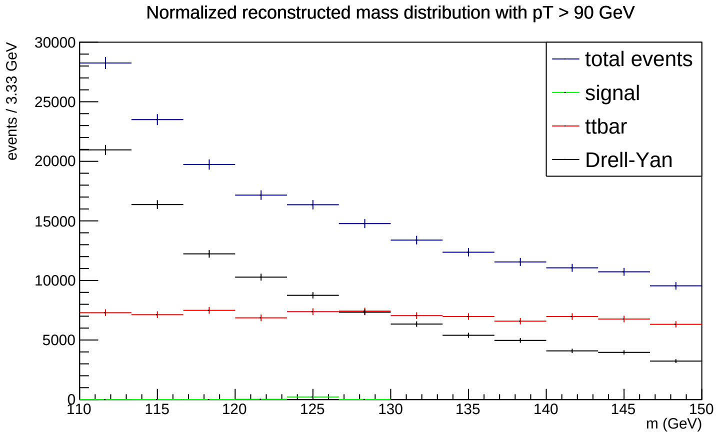


Figure 6: Normalized mass of the reconstructed particle with $p_T > 90$ GeV constraint. The colour means the following: black: Drell-Yan, red: $t\bar{t}$, green: signal, blue: background + signal

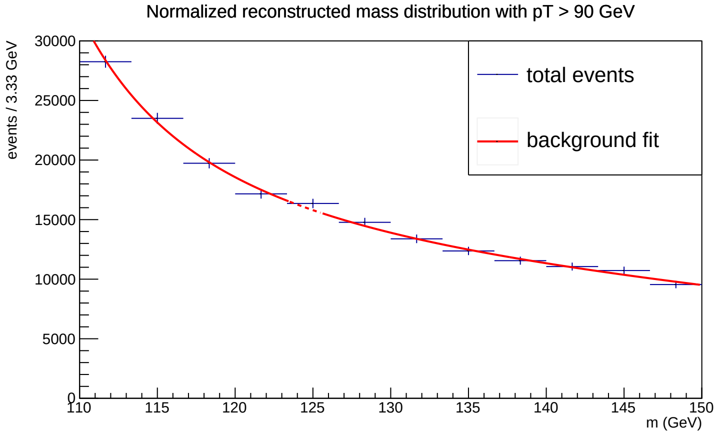


Figure 7: Fit of the background with $p_T > 90$ GeV constraint using a Breit-Wigner distribution and a first-order polynomial. The area with dotted lines was ignored during the fitting.

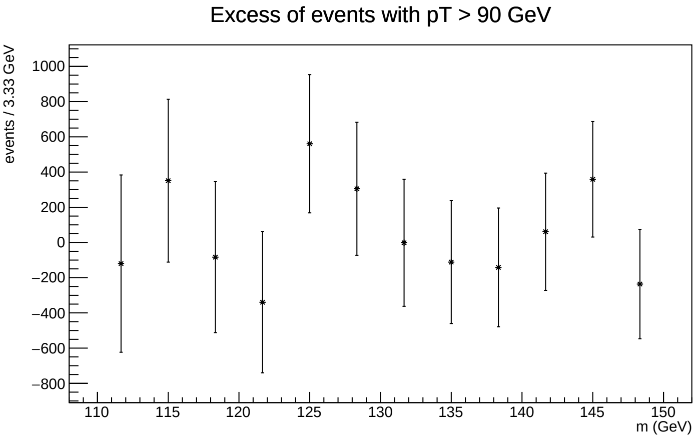


Figure 8: Excess of events with $p_T > 90$ GeV constraint compared to what is expected by the background fit.

4 Conclusions and discussion

If we refrain from using the transverse momentum constraint, we conclude that we cannot separate the signal from the background. The signal could potentially be uncovered by reducing the statistical variation in the background by simulating more background events. However, the number of events we have simulated is already relatively close to what would be seen at the LHC, so there isn't much room for improvement left in this regard.

If the transverse momentum of the reconstructed particle is constrained to $p_T > 90$ GeV, we can see a very weak signal at 125 GeV. The naive statistical significance of this signal is 4.5σ, but the true significance is likely much smaller. This result is not sufficient to confirm a discovery of the $H \to \mu\mu$ decay channel, and it's likely not even the most significant signal found so far. However, it does suggest that there is a signal that can be found.

The selection efficiency with the $P_T > 90$ GeV constraint is 12.16%. Therefore, if we naively assumed that the entire excess of events found at 125 GeV were signal events, we would conclude that a total of 4614 ± 3224 Higgs muon decays transpired. This is consistent with the number of muon decays that were actually simulated 2196, but the uncertainty is extremenly high. However, it could in principle be used to give upper and maybe even lower limits to the Higgs to muons branching ratio, assuming the Higgs production cross section is known.

4.1 Discussion

Our analysis was relatively surface level, since we ignored all other tracks than the muons and didn't attempt to optimize our constraints. A more thorough analysis done by experts would likely achieve a much better reduction in the background. The statistical variation of the background could also be suppressed further by using the full dataset. Therefore, based on this analysis, the Higgs muon decay channel should be seen in the run 3 data with acceptable statistical significance. However, although the scientists analysing the real data will have some advantages over our analysis, they will have many more disadvantages. There will be detector effects, pileup and many more sources of background to name a few. Consequently, the final verdict is that it is possible that the $H \to \mu\mu$ signal will be seen with good significance in the run 3 data, but it is not guaranteed.

If the data were real and our results were presented to the scientific community, the results should be believable, since we have completed the steps of the analysis carefully and documented our methods. It should be relatively easy for someone else to repeat this study to confirm our result. Furthermore, our result is not an extraordinary discovery, but a confirmation of what most scientists already expect to be the case, i.e. there probably is a signal but it's difficult to separate from the background. Consequently, the results themselves might not be challenged, but the validity of our methods could still be questioned.

Believeable or not, our study is very far from perfect. The most important improvement we could do is double check everything to make sure there are no mistakes in the programming or the analysis. Another way would be to generate more data to see if our results vanish into the background. Particularly fixing the error in our $t\bar{t}$-production data would be good, although it's unlikely to have a noticeable effect on the results. An obvious way to improve the simulation would be to use a proper detector simulation instead of a simplistic Gaussian smearing. Also, pileup should be simulated, which would make choosing which muons are paired in the reconstruction much more difficult. Currently we consider only the muons found in the final state, but by studying all detected tracks we might be able to better identify Higgs decays.

All of the above improvements would require a large amount of time to implement, so they will have to wait for another day. All things considered, if we assumed the data to be real, we would be able to convince ourselves with the results since we have put a considerable amount of time and care into this project.

References

[1] Christian Bierlich et al. *A comprehensive guide to the physics and usage of PYTHIA 8.3.* 2022. arXiv: `2203.11601 [hep-ph]`.

[2] R. L. Workman et al. “Review of Particle Physics”. In: *PTEP* 2022 (2022), p. 083C01. DOI: `10.1093/ptep/ptac097`.