Introduction to Subatomic Physics

Final Project

Dec 15 2025

For the final project, you will “discover” the Higgs to 4 lepton decay signal in LHC data from the 2011 and 2012 LHC runs. I will discuss additional details and examples related to the project in Lecture 10.

Instructions:

1. Your final submission will be a short report (as PDF file) containing several plots described below, and a short explanation of how you obtained the plots. Each figure/step will be worth between 20 and 40 points.
2. The first step of the project is to download the file FinalProject.zip from the Lecture 10 materials on the course website and unpack the file. This will create a directory called “FinalProject” containing the following files:

    - 1207.7235v2.pdf
    - data
    - FinalProject.ipynb
    - FinalProject.py
    - MC

    1207.7235v2.pdf is a copy of the CMS paper on the observation of the Higgs boson published in 2012. For this project it is useful to read section 5.2 and study Fig.4, which shows the invariant mass of Higgs boson candidates in the four lepton channel.
    “data” is a directory containing data from p+p collision events that were preselected to have four lepton candidates, as required for the Higgs to 4l decay channel. For each of the four leptons in each event, the Particle ID (electron or muon), charge, three momentum components px,py,pz and eta,phi are stored in the file.
    “MC” contains files from event simulations of Higgs decays and of various other processes that also produce events with 4 leptons. The format of the files is the same as for the data.
    FinalProject.ipynb and FinalProject.py contain a Jupyter notebook and plain Python code that can be used to read the input files, select events based on selection cuts that you will define, and plot the resulting invariant mass distribution of Higgs boson candidates.
3. (20 points) Run the Python script (either Jupyter notebook or plain Python) without further event selection, leading to a plot of the 4-lepton invariant mass distribution for data and simulation as shown in Lecture 9. Change the title of the last figure to include your name.

4. (20 points) In the next step you will define the basic object selection for electrons and muons) and event selections (based on Z candidates) to reduce the number of background events without losing too much of the Higgs decays. This will be done using cuts on the electron pT and eta and muon pT and eta. The values for these cuts are described in the second paragraph of section 5.2 of the CMS paper. The selection should be implemented in the objsel() function in the Python script. For all events that pass the object selection (after you have implemented the correct cuts), make a 2-panel figure that shows the invariant mass distribution of Z to e+e- and Z to mu+mu- candidates and another 2-panel figure that shows for each event the invariant mass of the highest mass Z candidate and of the lowest mass Z candidate.
5. (40 points) The next step requires some thinking and experimenting. You will now define an event selection based on kinematic variables that can be calculated for each event. This event selection should be implemented in the evtsel() function in the Python script. The event selection cuts should be chosen based on the simulated collision events, not based on the data. The most important input variables for the selection are the invariant masses of the Z candidates in each event. The cuts should NOT use the final 4-lepton invariant mass, but only the object cuts in objsel() and cuts based on the Z boson candidates. For this exercise we will assume that we know the expect Higgs mass mH = 125GeV. We will try to optimize the significance of the Higgs signal (in the simulation!) which we will calculate as z = s/(sqrt(s+b)), where s is the number of Higgs boson events with 119 GeV < m < 131 GeV and b is the number of background events in the same mass window. Note that you will need to retain the normalization of the various background processes that is performed in the Python script. For this step, report the significance you obtain for MC, the number of Higgs events after selection in MC and the number of events in the mass window in data, after subtracting the expected MC background. Also repeat the plot from step 3, but this time with the object and event selection.
6. (20 points) As the final step, you will see if the event selection can be further improved through a machine learning approach. You will split the simulated data into a training and test data set, and train a classifier (e.g., a BDT or MLP in scikit-learn) to classify simulated events as background or Higgs events, using the same input variables you used for step 5. For this step, you should report which ML model you used, the hyperparameters, solver and other ML setup information, the significance on the training and test data set and the number of Higgs events in data, as in step 5. Finally, make a plot showing the invariant mass distribution based on the ML selection.