Things That Go Bump in the Data: Analysis of Four-Lepton Decays in the CMS Open Dataset

Anna Barth, Andrés Cook, Brian Shuve

Abstract

The Compact Muon Solenoid (CMS), a particle collider experiment, recently publicly released a large dataset of collision events. We analyze four-lepton events in this dataset to search for an as-yet undiscovered dark sector particle that decays into four leptons. We develop a system that allowed us to deal with large quantities of data quickly, find and characterize interesting prospective signals in the data, and calculate numerical constraints on the production rate of any particle that decays into four leptons.

Background

The Standard Model, which describes all the laws of physics in terms of a handful of particles and their interactions, is one of the most successful theories modern physics has to offer. Even so, it fails to explain our universe in a major way – mysterious dark matter makes up 80% of all observed matter, and no Standard Model particle matches the profile. This problem has motivated physicists to search for as-yet-undiscovered particles that could explain the existence of dark matter. At the Large Hadron Collider (LHC), physicists observe millions of particle collisions per second in the hopes of seeing rare signatures of particles outside the Standard Model.

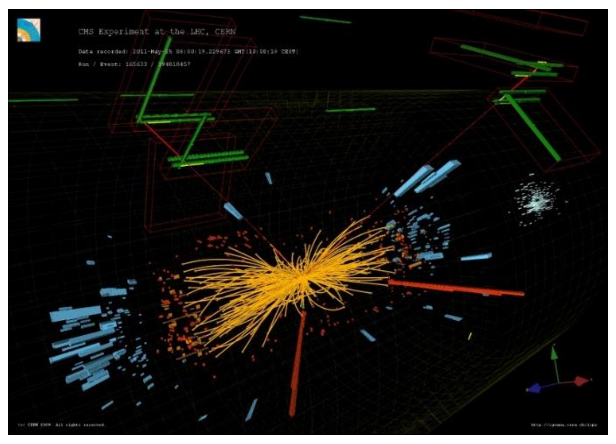

Fig 1: A collision event at CMS. The colored lines represent particles shooting out of the collision. [1]

Recently, one LHC experiment called the Compact Muon Solenoid (CMS) released some of its massive datasets to the public [2]. This summer, we analyzed the 2011 CMS open dataset, searching for the signature of a theoretical dark Higgs particle [3]. We modeled our search off the Standard Model Higgs, which can decay into four leptons (electrons and their heavier cousins). We searched for a dark Higgs that would also decay into four leptons.

Analysis of Decays

Each collision in the dataset is represented by an 'event' containing the leptons produced in that collision. We extract the leptons from each event and isolate leptons that likely originate from a decay event – those that have a high momentum and deviate significantly from the main beam. Then, we reconstruct particle candidates decaying into those leptons. If the leptons originate from the decay of a particle, their total energy in the center-of-mass frame should equal that particle's mass. So, given the four leptons' four-momenta, we can reconstruct the original particle's mass. The reconstruction process is described in Figure 2.

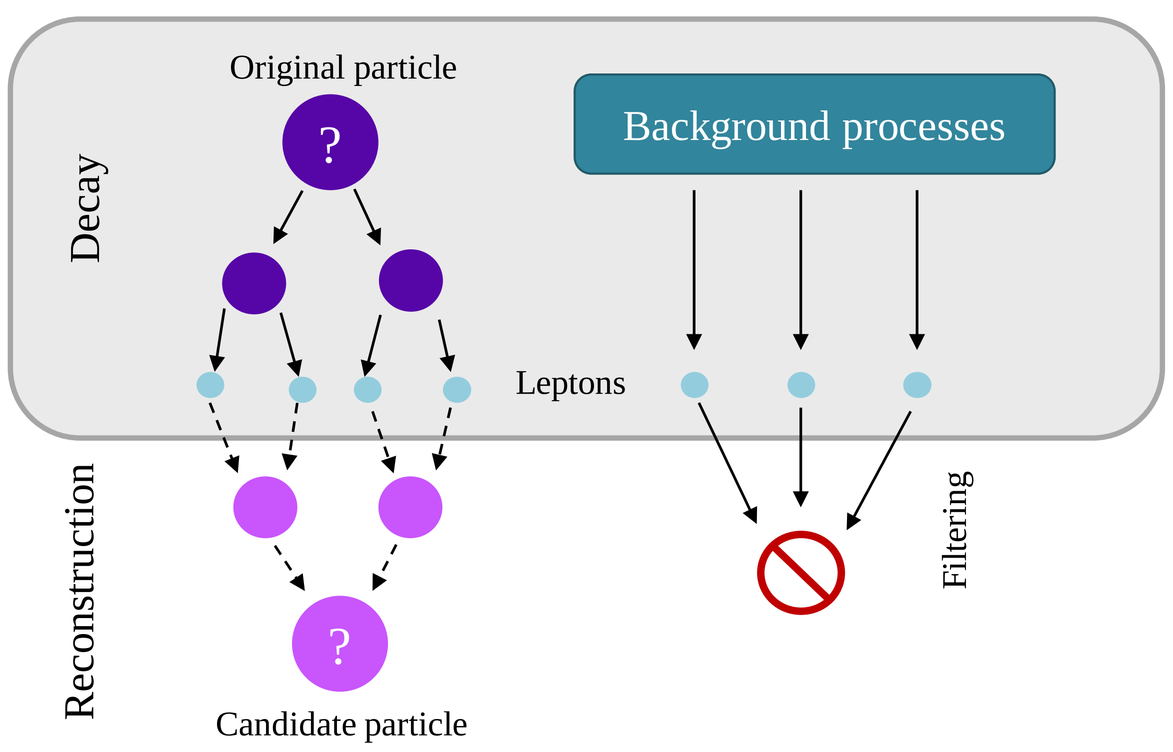

Fig 2: A collision event, from the target particle's decay in the detector to our reconstruction algorithm.

We plot the masses of hypothetical particle candidates. Real particles that decay into four leptons manifest as peaks on the resulting histogram. These peaks make up our *signal*, while the surrounding noise, generated by groups of four leptons that do not come from a real-particle decay product, is termed the *background*. We adjust our search selections (lepton momenta, directions, and identification criteria) depending on the mass region we were investigating or the particular lepton (electron or muon). We based our selections on existing CMS 4-lepton searches for the Z boson [4].

Next, we analyze the significance of the various peaks we observe based on the Poisson distribution. We calculate this by fitting a quadratic background to the region around the peak. A peak is considered significant if it deviates from the background by at least 4 standard deviations.

When we plot all reconstructed 4-lepton masses in the 2011 dataset, we find two particularly interesting statistically significant bumps, one at 225 GeV (Figure 2) and one at 268 GeV (Figure 3), with significances of 4.51 and 5.81, respectively. No known particle has either of these masses, so it is unclear what the sources of these bumps are.

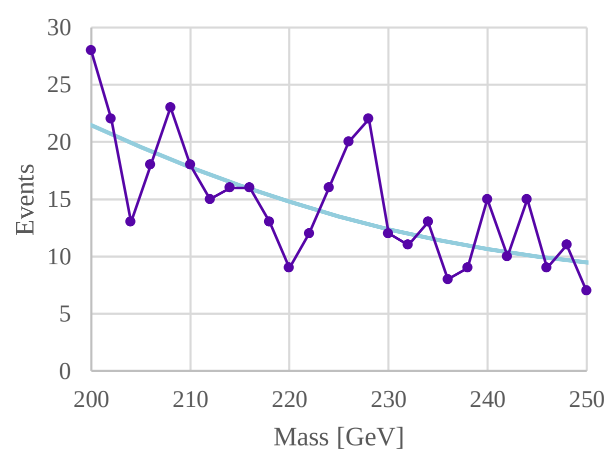

Fig 2: The peak at 225 GeV

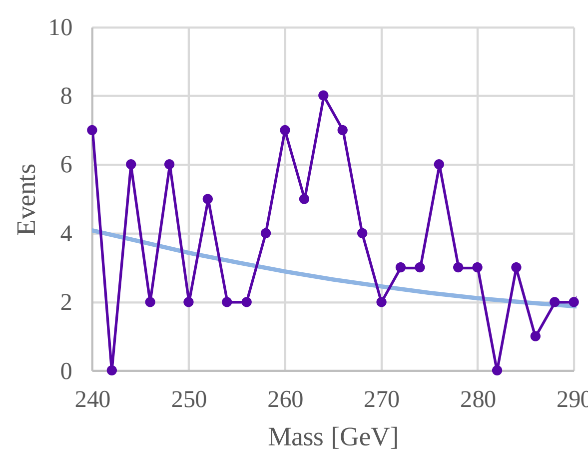

Fig 3: The peak at 268 GeV

These results are preliminary (see Future Work), and we expect that these bumps are statistical anomalies. We suggest these mass regions as areas for future exploration.

Upper Limits

In every mass range, we see some number of events – each of these numbers falls around some theoretical average. We can infer that the number of events we see is probably not too far away from the true average. We calculate an upper limit on the number of signal events for each mass region, the theoretical average number of events for a given mass range which is too high given the number of events we actually saw, at a 95% confidence level.

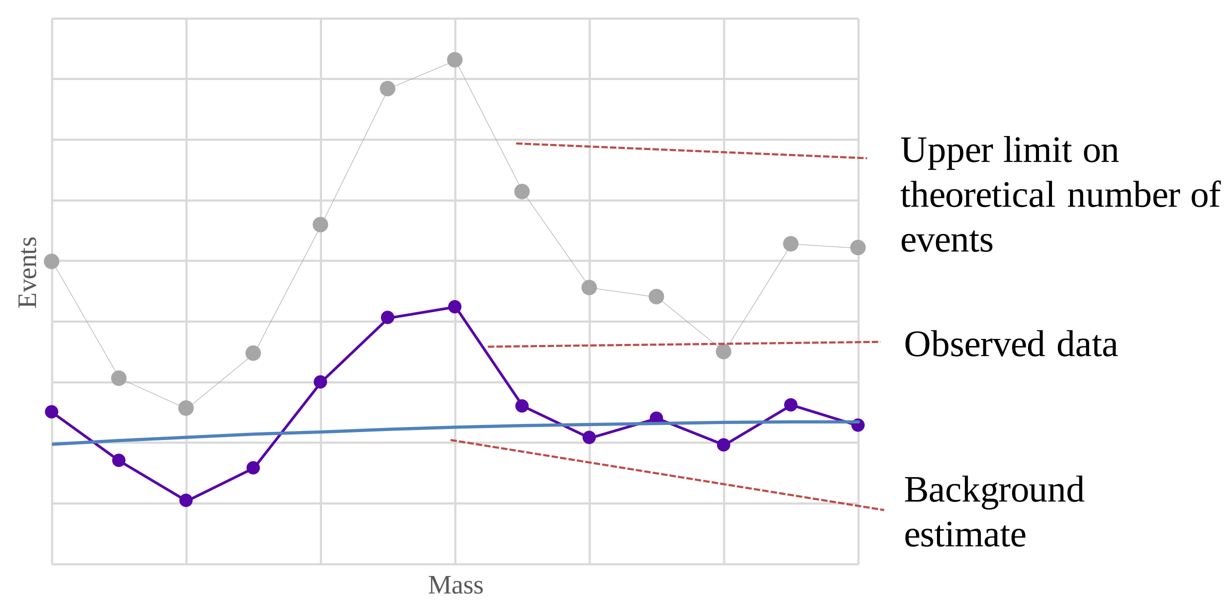

Fig 4: Schematic illustration of upper limits on hypothetical data

We make corrections to our calculated upper limits to compensate for the fact that the detector does not detect every lepton perfectly, and our filtering algorithm cuts out some real leptons. Using data from simulated particle collisions being run through a model of the detector, we rescale the number of events we detect to an estimation of how many events occurred including detector error. Specifically, we determined that only approximately 40% of events are detected, so we corrected our data accordingly.

Peak location	Signal events observed	Rescaled upper limit
225 GeV	37 ± 6.1	146 ± 12.1
268 GeV	25 ± 5.0	99 ± 10.0

Table 1: Upper limits at two masses of interest

Working with an Open Dataset

Our work is complicated by the fact that we are working with an open dataset. As non-collaborators in the CMS experiment, we do not have access to many of the collaboration resources, making it difficult to find documentation for aspects of analyzing the data. This lack of documentation leads to issues in accessing our data (*e.g.*, in file conversion and reading) and in understanding the contents of the data. Learning our way around the data, by piecing together public information from many online sources and experimental code snippets, was a major component of our work.

We also lack computational resources to deal with the massive volume of data we have. To increase our efficiency, we developed a system to convert the data into more streamlined file types. Ultimately, we increased our efficiency 100,000-fold (see Table 2).

File type	Events processed per minute
AOD	145
PATTuple	9,000
CSV	2,100,000
Distilled CSV (stores only 4-lepton events)	10,000,000

Table 2: File types used in our analysis, and how quickly we can process them.

Future Work

Going forward, we intend to continue our analysis on the 2012 CMS dataset, which is much larger and was collected at higher beam energies. The increased number of events will improve the statistical significance; in particular, we will be able to take a closer look at the peaks at 225 GeV and 268 GeV. We will also improve on our analysis of prospective signals by estimating a systematic uncertainty on our background prediction, or by modeling an expected signal. We plan to continue investigating the CMS trigger system, which will improve how we compensate for detector inefficiencies and false reconstruction of leptons.

Acknowledgements

We would like to thank Brian Shuve and the HMC physics department for supporting our work as well as the Class of 1985 Summer Experiential Learning Endowment, the Arnold and Mabel Beckman Foundation, and the Claire Stevens Summer Research Fund for their financial support. We are also grateful to Yangyang Cheng at CMS for her help in getting this project off the ground. Finally, thanks to the CERN CMS collaboration for their data, helpful publications, and online resources.

References

[1] CERN Document Server, https://cds.cern.ch/record/1374433

[2] CERN Open Data Portal, http://opendata.cern.ch/

[3] N. Blinov, E. Izaguirre, and B. Shuve, Rare Z Boson Decays to a Hidden Sector, Phys. Rev. D97 (2018) no. 1, 015009, arXiv:1710.07635.

[4] CMS Collaboration, "Observation of Z decays to four leptons with the CMS detector at the LHC", JHEP 12 (2012) 034, doi:10.1007/JHEP12(2012)034, arXiv:1210.3844.