To: Jose Monroy

From: Regina Werum

Re: Dissertation draft

Date 5/10/18

The substance of this dissertation is of course well beyond my area of expertise. That said, my contributions to your committee focus on the following aspects: I am reading it as a generalist with an eye towards science communication, and I am focusing on theories, hypotheses, and the experimental design.

As a generalist reader (imagine someone on an NSF panel from a different discipline to whom you have to communicate the scientific merit and broader impacts of the project), I recommend this:

* Clearly state in each chapter, not just the intro, that the two main goals of the dissertation are to answer the following questions (using my own words here):

1. “What causes the creation of a Higgs Boson with a single top quark/lepton-charged muon?” and
2. “How can we detect this particle more effectively or efficiently – how good is the device (FPIX modular detector) developed?”

The theoretical prediction of the existence of the process is already stated and what we want to do here in the thesis is to establish if the process is present in the data and that theory prediction is consistent with data. Also we explore an hypothetical case beyond the standard theory.

The second part of the thesis reports our contribution to the development of the Full CMS detector. FPix is just a part of the detector and not the part that we use to detect the Higgs boson or the top quark. We use the full detector.

As a social scientist, here are some additional observations and requests to clarify:

* In the intro and in the methods/data section you should make it clear how many times you ran the actual experiment,

I do not run a particular experiment, I use the full sample and the look for events that match some requirements. Collisions happen 40 million times per second but not all produce tHq processes and not all are stored. Depending on the cross section and the luminosity we can calculate the expected number of signal events and depending on the selection criteria we count the number of events found in the sample.

* how many events are part of the Monte Carlo component. If you did state it someplace and I missed it, it means you need to state it more prominently. This may be clear to you because there are norms in your field that circumscribe these practices. But, for instance, in our field, demonstrating that asymptotic tendencies apply is important. For that you need a certain sample size, whether experimental or MC based. By extension, talking about confidence intervals only makes sense if there is a reasonable scope to the project.

Normally, the statistics of the sample is represented in terms of the cross section, that’s why I did not mention the number of generated events but the cross section of the sample. Knowing the cross section and the luminosity one can find the number of generated events.

* You also need to describe the 2016 full data sample more explicitly. Outside of your committee no one knows what’s in it – or whether you use the full sample (n=?) or a subsample of it with particular properties.
* Please explain WHY you have chosen to compare the results of the MC with the 2016 sample – this is implicit, but worthwhile spelling out: If I understand correctly, you want to know whether the device you developed is more effective at spotting those special particles than we would get if we randomly picked them out??? In other words: the comparison is intended to help demonstrate the quality of the device. (Or is it designed to demonstrate the quality of the data??? I got confused along the way).

Actually not. 2016 data sample is used because it is the one available and not analyzed. Also, I am not using the data to evaluate the performance of the detector because it is made when commissioning the detector using other data samples. Both parts of the thesis are actually independent because the data sample was taken with a different detector. I am only reporting two parts of my work.

* P. 151ff please clarify the process used to identify the muon IDs – is this done automatically by the detector you developed, or manually by the researcher? What if any intercoder reliability checks are needed or recommended in this circumstance?

It is done in section 4.5.1, reconstruction is made by the Particle flow algorithm. The object definition in p151 describe how restrictive is the definition of the reconstructed muon, thus, a loose muon ID has less requirements than meduim moun ID.

* P. 155ff discussion of the jet-related variables: Please confirm and/or clarify: What are these variables? Are we talking about values of predictors of the outcome that are known to have been randomly miscoded or are missing? Or are these miscoded predictors associated with specific cases/events? Data errors? Do you treat them as random errors/white noise?

Jet related variables are additional variables used to improve the differentiation of reconstructed true jets from reconstructed jets that looks like a lepton. Thus we look at the lepton and the surroundings to see if there is a jet close the lepton, and if they are too close (delta R <0.5) then we consider the lepton as part of the jet.

* P. 157 Table 6.4 – the cutoffs for the fakeable object and the tight object appear identical. Purpose? Can you explain more plainly why muons have to be in 3 types of selection efficiencies (loose, fake, tight) and how they get assigned to those groups?

They are not, tight muons have more requirements than feakable muons. The selection is defined by the

reseracher and depends on the purpose of the definition, for instance, feakable and thigh selections are used to evaluate the fake rate. Loose muons are used in preseection. The assignment depends on the requirements fulfilled by the objects.

* P. 164: preselection process: seems you are constraining the 2016 sample. From what to what? Reason? Consequence? (where are you eliminating variation and why)?

No constrains, just select events using a loose set of selection criteria. Later, the selected events are used in a more deeper selection.

* One more question for you – found this note tucked into the folder on your diss:

p. 181: Input variables are correlated  -- how strongly?  Relates to Fig 6.18.  Could this affect multivariate modelling? Not sure what the norms are in physics regarding correlation and collinearity among input variables.

Numbers in the squares are the percentage of correlation. Boosted decision trees are not affected by correlations. Other methods like Fisher discriminator perform a correlation removal in the first stage.

* Reading the graphs vs. Multivariate output: You put the graphs in the text and the multivariate analyses in the appendix. That is fine, provided the graphs are self-explanatory. I did not find them to be self-explanatory and thus turned to the multivariate results to get a better sense. I recommend that you discuss the graphs and what they portray and mean more explicitly. Tie to hypotheses and to theory. Simply referencing them and restating the empirical results is not sufficient.

The graphs show essentially that the input variables have discrimination power, in 6.16 the individual contribution from several backgrounds and the signal are shown while the multivariate outputs are only for the dominant backgrounds. I tried to describe what each graph is about in text and the same trends are observed in BDT outputs. I might be not understanding this comment in full.

* The conclusion could easily be expanded. It would be useful to pull up a level – let go of the technical specs and instead focus on the big picture. This is about the God Particle, after all.  Think like a science communicator here. Tell us all – not just experts in particle physics – what are the findings, why are they valuable, and to whom? How do your findings push forward what we know about the universe as scientists? How might your findings influence future research? Perhaps even research policy and funding? What refinements to the analyses do you plan do to (hinted at on p. 274-5)? Are those refinements intended to improve the predictive model/reduce uncertainties? Do you see potential for that detector you developed to have commercial value – if so, for whom?

Although this analysis is very important for the Higgs community, it is just a small part of the big picture and what we can do here is to set up upper limits on the cross section and exclude some values from a range; in few words, we are testing if the H -t coupling could be negative and the results show that it is not. In a few words, what we found is that the H-t coupling behave as predicted by SM, constraining a bit more the existing exclusion on the coupling.

* Finally, how can we extrapolate from your experiment to others in the field – link to particle physics more generally. How can insights derived from your work inform work in other areas of physics? Future research in your and other areas of physics?

Minor suggestions:

* Make sure to run a spell check and do check manually. For example “Chapter4” should be “chapter 4” on p. 150.
* Make sure to annotate endnotes consistently, for instance p. 150 you have a sentence “blablabla [endnote].” It should be “blablabla. [endnote]”

I did this consistently in the full document, all references to chapters, sections, figure, table, appendices, start with uppercase. Also the references [] are always before the period