For the reviewers of paper with id 154.

Thank you very much for your comments! We tried to address them in the new version of the paper. Please find hereafter our answers with some explanation or additional information when useful.

1. Introduction:
Please add a reference to ROOT and Minuit

A. done.

2. Pseudo-experiments....
I would add a numerated list (each in a separate line) for the description of each toy fit

A. done since it helps readibility. Itemization would me more elegant but would require to sacrifice further text.

Figure 2.
It is not clear why the speed-up of Goofit/MPS vs GooFit and RooFit/Proof-Lite vs RooFit is not perfect (30), since I presume each process runs a full toy cycle. Is it due to hyper-threading ?

A. No. This can be argumented in different ways. For instance for GooFit the MPS is limited to 16 concurrent processes per GPU; each process uses a shared GPU and one exclusively assigned CPU and the used physical servers have 16 (or 20) *physical* CPUs.
As shown in the slides (8, 9, backup) of the talk at the conference and also commented in the text of the paper, the speed-up as a function of the # of CPUs follows the Amdhal law.
The best estimate of the fit parameter for the Amdhal model is similar for the speed-up of both Goofit/MPS vs GooFit & RooFit/Proof-Lite vs RooFit. See also below.

It would be maybe more interesting seeing a graph of the speed-up as function of the number of CPU. I don't see the interest to show it as function of # MC Toys. This variable should have no influence on the speedup.

A. We agree (indeed in Fig.2 the speed-up performances look rather steady with respect to # MC Toys). Thus we substituted the old plot of Fig.2 with the speed-up plot as a function of the # of CPUs, following your suggestion.

Eventually one could show the speed-up as function of the
number of bins, which are responsible for the time spent in the ML evaluation.

A. This is not possible for the specific physical case (binning is fixed and comes from a background subtraction procedure). We have discussed in the talk given at the conference that the maximum advantage of using GPUs is for Unbinned Maximum Likelihood fits and that the advantage for Binned fits diminishes with smaller # of bins (the smaller the # of bins the smaller the time required to calculate the pdf) [slide 3].
Unfortunately the 5pages-constraint does not allow to add this (preliminary) study also in the paper.

It could be interesting to see also the speedup of a single GooFit vs a single RooFit, when using the GPU and when using only the CPU. In this last case one would see the direct contribution of using GPU and the possible advantage (or disadvantage) of using GooFit vs RooFit on the single CPU.

A. Done. We added in the text these numbers: the speed-up for single GooFit vs single RooFit is either 64 (TeslaK40) or 48 (TeslaK20).

Paragraph 3.: Exploring the applicability limits of Wilks theorem. In the case the mass and the width are fixed the condition of the Wilks theorem are satisfied if the correct test statistics is used (see paper on asymptotic formulae from Cranmer et al., ref 11).This should be mentioned.

A. We think this is addressed with the case in 3.1
Indeed in this case μ is not constrained to be positive and doesn't go to its limit when $H_1 \rightarrow H_0$

3.1 Is in this case mu allowed to be negative ?

A. Yes! We now make it explicit writing "….free (either positive or negative)"

Not clear what is the scope of this case, instead of presenting only 3.2.

A. 3.1 addresses the case in which Wilks theorem should apply. We find that this is correct. This is important above all for validation purposes (in the known case we get what expected).

3.1 What is the chi2 probability for the fit for the test statistic distribution ? If the chi2/dof =1.37 and dof is large (like in histogram of fig. 5), this probability is then rather small.
3.2 Same comment on the chi2 probability of the fit as before

A. The plot shown (fig.6) was replaced because there was a bug in the **after-fit** evaluation of the *normalized* chi-square. The latter is now 1.013 (probability ~3.5%) for the fit in 3.2 and 1.009 (probability ~11.8%) for the fit in 3.1. The fits are fairly good given the 9M entries.

What are the fit results in the case of keeping m and gamma free ? These cases are more interesting, because there the Wilks theorem does not apply.

A. We agree that these cases are interesting. Even if there is no reference model for them we tried to fit with a chi-squared function which unfortunately turns to be not good at low or very low values of the test statistic. Limiting the fit to the range where chi2 model seems to work we get - for the number of degrees of freedom – values of its best estimate that are not particularly significant.
Of course further study is needed here and we deserve it for a later step (@ later conference).

Some conclusions on the applicability of the Wilks theorem are missing in the paper.

A. We have shown that the statistic distributions we get for our application (in the two cases 3.1 & 3.2) follow enough well the predicted distribution. For case 3.2 this is really new.
The 5pages-constraint is strict for the amount of information & results that we convey in the paper and it is difficult to find space to stress further this point. Let us know if you don't agree.

- 4 Future developments
The RooFit based approach is not unreliable, but just unbearable for the long required processing time".

A. We agree that your proposal is much more correct and we changed accordingly.