Referee 3: Report and responses

Strengths
1. It is always interesting to see a new idea to solve long-standing challenges in LHC physics;
2. Combining distributions and fractions for several jets at a time might be useful for well-separated jets;
3. The paper comes with code, so these methods can be tested.

Weaknesses
1. I am missing a well-defined case for a combined analysis of many jets, this is assumed, but the reality is much more complicated than even numbers of heavy flavors;

Report
I generally like new ideas for long-standing problems, and the paper presents a new, global approach to flavor tagging. Learning fractions and distributions at the same time is also not unrealistic, and the paper nicely illustrates how the problem can be tackled using a set of standard numerical approaches. In that sense, the paper is interesting and could eventually be published.

We are grateful to the Referee for reading in detail the manuscript and for the useful comments that have considerably improved the work.

However, I see a series of shortcomings:
1. It would be nice to have an actual use case, the reality of jet tagging in events is much more complicated than an even number of heavy flavors;
The Referee is correct about this point, it is possible that one of the jets lies beyond the |eta|=2.5 limit, and therefore we would lack the knowledge of its b-score. We have added a discussion in the Discussion section about this. We argue that if this happens, for instance in the hh > bbbb process, then one can still infer a probability of the event belonging to the sought signal by studying its invariant mass with the other jets in the event. Also other features could be included, such as DeltaR, which of the jets (ordered by p_T) pairs with which jet; etc. This point indicated by the Referee should be studied in a future investigation, as it counts for ~25% of the selected cases in searches such as in 2301.03212.

2. Following a proper use case, the results could also be presented in a physics context. For a journal like SciPost I do not consider pure toy applications without physics interpretation appropriate;

This is a fair point and we could agree to some extent. However, let us expand in some details the reason for our choice. The paper is very physical in the sense that it is about a physics problem, and it is addressing many details of a physics situation. The used statistical techniques are in some sense a novelty, but the problem we address is physical. Another aspect of our argument is that if we would have simulated (MadGraph+Pythia) a physical process, then we would have needed a b-tagging score for the b-jets, but including the usual detector level simulation through Delphes would have given an oversimplified b-tagging distribution, since the b-tagging Delphes cards are a simple draw of a random number. And in such a case we would have been in something very similar as in the current situation. Moreover, it would have had less complexity. The other options were running a full GN1 or using a b-tagger such as Secondary Vertex Displacement in Pythia, but this was beyond the purposes of this work and would have been misleading to the main idea that we are presenting in the manuscript.

Henceforth, we considered a suitable option to perform the presented calculations. A next work will include features as those suggested by the Referee in this and in other points of this report.

3. In Sec.2 the four different approaches are defined, but it is not clear how they are selected and why these ansatzes have been chosen, please clarify and maybe also specify expected strengths and weaknesses motivating this selection;

Perhaps the most intuitive way to summarize a stream of numbers is a histogram: to count the number of occurrences within each bin. This histogram is typically the “default” approximation to the density function, and can be viewed as a posterior inference with a Dirichelt prior with alpha=1, or a uniform prior (alpha being the Dirichlet distribution parameters). Starting from this purely “uninformative” prior, we designed three models to make use of the prior physical knowledge on the shape of the density. (1) The Dirichelt prior describes how concentrated the bin heights are, as alpha = infinity enforces only one bin to have mass and alpha = 0 enforces uniformity. (2) The self-normalized Gaussian process encodes the continuity of density function, as we believe that the neighboring bins should have similar heights. This can be understood as a local smooth. (3) The unimodality prior enforces the density function to be unimodal, or the bin height only has one local mode. The continuity and unimodality assumptions will probably be even more relevant when modeling physically meaningful quantities. From our experiments, we find these structured priors (2 and 3) are able to extract more information from noisy data, and hence are generally recommended.

3. The rest of the paper is very text-heavy and hard to read. It would help to maybe split up the figures and align them with the text, so one does not always have to flip to the back of the paper.

We thank the Referee for this very suitable suggestion. We have adopted it and brought the relevant figures closer to their corresponding Section. We have also created an Appendix for figures that intend to show the generalities of the method as well as a testing for the method with other seeds.

4. I am not sure if I am missing something, but it would be nice to discuss in more detail the impact of the prior and how this Bayesian approach works in Sec.3;

We are sorry for the confusion: here the term “prior” may refer to both (a) the structure of the model prior that encodes the shape of the density, which we have carefully designed in Section 2, and (b) the prior-guess of the density, on which we typically do not have too much knowledge. The prior structure of the density shape has a persistent effect on the posterior fitting: for example if we enforce unimodality, the inferred density needs to be unimodal regardless of the sample size. In the paper, we have extensive experiments on this comparison between different prior structures, and our proposed method exhibits a robust advantage over the non-informative benchmarks. On the other hand, our Figure 3 and 4 demonstrate that the effect of the prior-guess soon vanishes as the sample size goes up, as the posterior distribution concentrates to the truth exponentially fast with a big enough sample size regardless of the prior-guess.

5. Along the same lines, is the Bayesian approach guaranteed to converge or lead to a stable result? It is not clear to me if I should deduce this from the shown results;

This is a very fair concern. The answer to this question is subtle and, at least, two-folded. On one hand, the parameter $\hat R$ is an estimator that assesses whether the Bayesian method has reached a good convergence in the sense that the chains' sampling is unbiased. However, this is not enough to guarantee that the proposed model is correct. To this purpose, the Bayesian framework provides tools such as the Posterior Predictive Check (see e.g. 2011.01808), which can assess the probability of the data to have been sampled from the proposed model with the inferred parameters. In a real case scenario this should be performed, and we have added a sentence about this in the Discussion section.

6. So what is the bottom line? Does the method work? Better than what? Or does it need additional work? The paper leaves the reader a little in the dark.

To address this point, we have added Section 3.4 in which we perform a comparison of the proposed analysis with usual cut-based analyses. To perform such a comparison we study the prospects of each framework to recognize the class bbbb (which in many cases it would be the signal). We plot the ROC curves for both frameworks and find that the Bayesian framework performs a better classification. We also find that a bias in the prior belief on the individual c- and b-curves yields a biased ROC curve for the cut-based method, but it does not for the Bayesian framework, being this also an important result.

We have also included a new figure with the results for 10% and 1% bbbb signal. Interestingly we find that the 1% case still works quite well, being the reason that the bbcc background class helps to learn the b-distribution while simultaneously learning the bbbb fraction.

Altogether, I think the paper needs a more realistic use case/example/bottom line to be really helpful as a physics paper. The idea is very exciting, though, and I am looking forward to some more quantitative results.

We thank the Referee for all their comments. We are looking forward to connecting many building blocks and being able to implement these ideas in a real analysis. Improving sensitivity in LHC analyses could provide important progress for better understanding the physics at the LHC.