Response to Reviewers

Title: "BROOD: Bilevel and Robust Optimization and Outlier Detection for Efficient Tuning of High-Energy Physics Event Generators"

Authors: Wenjing Wang, Mohan Krishnamoorthy, Juliane Müller, Stephen Mrenna, Holger Schulz, Xiangyang Ju, Sven Leyffer, Zachary Marshall

We thank Greg Landsberg, Editor in charge, and the three reviewers for their thoughtful comments that have helped us improve the manuscript significantly. Below please find our responses to each of the issues raised in the review reports. We have numbered all detailed reviewer comments. The changes made in the manuscript are indicated in blue font and the number of the comment corresponding to the changes is indicated where applicable, e.g., "[Reviewer comment 10:] changed text".

Response to the Editor in charge

Dear Authors, Please consider referees' suggestions on your interesting paper. While a major revision is requested, it mainly affects the presentational aspects of the manuscript. It's clear that the overall consensus is that the paper is novel and important, so we are looking forward to publish it, once you address the referee comments. Best, Greg Landsberg Editor in charge

Response: Dear Editor, in the following we have addressed each reviewer's comments.

Response to Reviewer 1

My apologies for the delay in providing this report ... the paper is very long (!) Here is my report: This paper reports a study of multiple methods to automate the selection of bins, observables, and weights for parameter tuning for parton shower simulations in high energy physics. The paper is a serious study, it should certainly be published, and SciPost Physics is a reasonable venue for this. Before I can recommend publication, please see my comments and suggestions below.

Response: Thank you for your positive review. We are addressing your comments below and in the paper.

Here are two overall impressions: (1) Parameter tuning seems to be a bit of an art and this paper feels like it is adding a lot of mathematical rigor to a problem which lacks mathematical rigor behind the scenes (this comes out in some of my specific comments below). I know this is how tuning is done now, but maybe this could be stated somewhere near the beginning and/or end?

Response: Thank you for your suggestion. We explicitly call out the lack of mathematical rigor in the introduction (see blue text in the first paragraph on page 5).

(2) There is quite a mix of rigor and non-rigor (for lack of a better word). It would be useful to harmonize this across the draft. There are some concrete suggestions in my comments below.

Response: Thank you for the detailed comments and suggestions. We are addressing them below.

Detailed comments:

A *p5: "The uncertainty on the MC simulation comes from the numerical methods used to calculate the predictions, and it typically scales as the inverse of the square root of the number of simulated events in a particular bin." – Perhaps worth commenting that this is excluding non-statistical theory uncertainty? Some aspects of MCs are under theoretical control and thus have theory uncertainties (that I understand are usually ignored).*

Response: We should have used the word "precision" of the simulation instead of uncertainty. The fact that the model is incomplete is addressed elsewhere in the text (see page 5 at the bottom).

B *p5: Since you are being clear with your terminology, it would be good to say that $\Delta\mathcal{R}_b$ is the one-sigma uncertainty from the measurement and will be interpreted as the standard deviation of a Gaussian random variable (often, systematic uncertainties without any* *statistical origin dominate measurements so I think the word "interpreted" (or similar) is important to state).*

Response: We have updated the text (see page 5, first paragraph in Section 1.1)

C *p5: "A 'good' tune is one where the red line falls within the yellow band." – If the yellow band really is interpreted as the 68% CI, then shouldn't a good tune be one that contains the red line 68% of the time (so 1/3 of the time, it does not)? People like to look at plots and see all the points within error, but this is a sign of overfitting!*

Response: A good tune will have a χ^2 per bin near 1 on average. The yellow band in the plot is showing the denominator of the χ^2, and thus the simulation should, on average, fall within the yellow band. Please note that in order to shorten the paper, we removed a few pedagogical items that the SciPost community is likely familiar with, and this sentence has therefore been deleted (see Page 5 to top of page 7).

D *p5: Fig 1: Are these real data? I know you don't want to confuse the reader at this point, but if the data and simulations are real, please say what they are (feel free to forward-reference to a later section).*

Response: The data comes from analysis L3_2004_I652683 and the data is from a poor set of Pythia parameters. Please note that in order to shorten the paper, we removed the figure, assuming the SciPost community is familiar with the histogram concepts (see page 6).

E *p5: "optimal set of physics parameters" → Perhaps it would be good to be clear what you mean by "optimal". Since the title of this section is "mathematical formulation", it would be sensible to state mathematically what you mean by "optimal". Along these lines, it would be good to explicitly state somewhere around Eq. 1 that you are ignoring correlations between measurements.*

Response: Thank you for the comment. Here, we define optimal as either a locally or globally optimal solution. Since we are dealing with a multimodal problem, our goal is to find at least a locally optimal solution. Please see the blue text below Eq. (1) for the change we made in the manuscript.

F *p6: More measurements are starting to provide proper covariance matrices, so you can at least get correlations between bins so going from Eq. 1 to Eq. 2 is a non-trivial approximation. You say this "implicitly assumes that each bin b is completely independent of all other bins." but I would have expected some statement about the impact on the results.*

Response: Please see our response to Referee 2, Comment 2. This has been addressed in the text (below Eq. (1)). We acknowledge that more measurements start to provide proper covariance matrices. It is interesting to study the impact of including them in the tuning procedure. However, we would like to defer such a study to future research work when we have a sufficient amount of such measurements available for conducting a systematic study. Our proposed approach follows the ideas introduced in the PROFESSOR paper (Buckley et al., 2010, `https://doi.org/10.1140/epjc/s10052-009-1196-7`) and extends it to rational approximation models.

G *Eq. 3c: why is it $\hat{p}_w \in (...)$ and not $\hat{p}_w = (...)$?*

Response: In general, multiple (local) minimizers may exist, and thus there is a set of solutions to the problem and we write $\hat{p}_w$ as an element of this set of minimizers. See below Eq. (3c).

H *p9: Isn't it redundant to write $e(\hat{p}_w|w)$ since the w is already part of the symbol $\hat{p}_w$?*

Response: Thank you for pointing this out. We realize that our notation is a bit confusing, so we updated it. Please see page 10.

I *Eq. 5a, 5b, and 6: Something seems strange here; in optimal portfolio theory, the goal is to identify weights of each component asset. However, your function 6 only depends implicitly on these weights - they do not enter in the "expected return" (Eq. 5a) or the "return variance" (Eq. 5b). Am I missing something?*

Response: We realize that our notation may have been confusing here. Please note the notation change in the previous comment H, and the updated notation in Eq's 5a, 5b, and 6. This notation change should make the dependence on $\mathbf{w}$ explicit (besides simplifying our notation). Please see page 11.

J *Sec. 2.1.2: can you please provide some intuition here instead of making the reader dig through [13] to find Eq. 27?*

Response: We added the following sentence in the manuscript:"The intuition behind this is that it takes both the model fit and data uncertainty into consideration." (page 11).

K *Sec. 2.2: I don't understand the notion of "uncertainty set" - can you please expand? The interval does not represent the 1σ interval or the maximum possible variation (which is infinite). The text after Eq. 10 suggests it has some meaning and is not just a definition for the symbol $\mathcal{U}_b$ in Eq. 10.*

Response: In Section 2.2, we added an explanation for uncertainty set $\mathcal{U}_b$ and clarified how it is different from a probability distribution.

L *p15: "It would be non-physical to adjust the model parameters to explain these extremes." → I agree, but then why don't you drop these bins from all histograms? If you don't, then you will tune away these effects in some cases because by chance the* *simulator happens to have region of parameter space that can explain it (physical or otherwise).*

Response :We have added a better explanation to the text. Since we don't know the correlation between bins of different histograms, we cannot perform such a global removal (see page 16, bottom).

M *I believe the precise wording for the null hypothesis is that the mean of R_b is $f_b(b)$ (?) ("appropriately described by" and "no significant difference between" are not precise). Same for the alternative.*

Response: We fixed the definition of the null and alternate hypothesis in Section 3.2.

N *What is the level that you actually pick?*

Response: We clarified the value of α used in the experiments in Section 3.2.

O *Eq. 15: I think if you do this, then the χ^2 hypothesis is not true. If you are comparing many subsets, then something like an F-test would be more appropriate, or maybe a sentence that says that this is motivated by statistics but does not have a strict type I error at the set point (and then also probably good to remove all of the ultra pedagogical and likely not applicable explanation at the top of p16)*

Response: We have clarified in the text below Eq. (13) that since the test statistic is being calculated per bin and then summed over a subset $\mathcal{B}$ of contiguous bins to get the total test statistic, we believe that the χ^2 hypothesis test is appropriate to use here. Additionally, we would prefer to keep the paragraph about the critical value in the manuscript since it is relevant to the discussion in section 3.2 and it describes how we obtain the critical value as a function of α and $\rho_{\mathcal{B}}$.

P *p19: What does "some of the simulation data were available to us" mean?*

Response: We have removed the sentence, since it adds no useful content. It means that we had available the exact output of the ATLAS generator calculations for their choices of parameters, but we did not have access to their exact software.

Q *p21: I found it strange that you did not cite the original data papers that go into the A14 tune (sorry if I missed it!) I also see that Fig. 9, 10, 11 do not provide a reference for the data - please add it!*

Response: Thank you for pointing this out. This was an oversight. We have added the data references from the original A14 paper by ATLAS (see references in the last paragraph of Section 4.3 (page 22) and in the captions for figures 9, 10, and 11).

R *Table 5-7; 13-15: What should I take away from that fact that there is a huge spread in performance and the ranking from the different metrics is quite different? (in some cases, the worse in one metric is the best in another!)*

Response: The three proposed metrics essentially compare different aspects of the tuning results for different employed methods. It is not surprising to see these methods are ranked differently for different metrics. In the end, which metric to use really depends on the objective of the tuning work. If the objective is to achieve the smallest differences between MC simulations and data, weighted χ^2 would be used. If the objective is to find parameters with small uncertainties, A/D-optimality would be used. As addressed in the following comment *T*, we added a few sentences to Section 7 to clarify this in the updated note.

S *Sec. 5: I was surprised that this comes after the results. It is a bit hard to compare your tunes to the "expert ones" if I don't have a sense for the "uncertainty". Can you maybe add the expert values to Table 21?*

Response: Since the primary objective of the manuscript was to introduce automated optimization methods for parameter tuning, we prefer to leave the eigentune results separate in Section 5 in order to first focus on the optimization outcomes and the performance of our proposed methods for finding optimal solutions and then analyze the uncertainty without creating too much confusion. We added a forward reference at the end of Section 4.6.3 to the eigentune results in Section 5. We also added the expert eigentune values to Table 13.

T *Sec. 7: You have compared many method variations - which one do you suggest as a baseline recommendation?*

Response: We added a paragraph in Section 7 to address this question.

Response to Reviewer 2

The paper is very interesting, extremely relevant, and has great potential. However, it is missing physics aspects here and there, so I am asking the authors to add more physics discussions for the typical LHC physicists. All my comments are included in the attached pdf file (in red). Please feel free to ignore some of them, if they do not make any sense, but you will get the idea. What I am missing most is the final step, namely an application of the eigentunes to something new...

Response: We thank the reviewer for your careful reading and thoughtful comments that have helped us improve the manuscript.

Detailed comments:

1. *p5 line 120: Say explicitly that theory/model uncertainties are not included*

   Response: Thank you. This is an important clarification. We have added this to the text. See Section 1.1.

2. *p6 line 145: Mention that observables are not some kind of basis for observed events, unlike phase space. So there will be correlations which are not considered here?*

   Response: We have added these caveats to the text (page 7).

3. *p7 line 172 "where weights are assigned based on how influential data is on constraining parameters" – Does that not come in through the errors?*

   Response: The text has been amended to more correctly reflect the method of that paper, which deals more with how to reduce a large dimensional tuning problem to a more tractable one (page 8).

4. *p9 Eq.(4) This is becoming a little formula-heavy. Fine if the authors want to be precise, but the average reader will skip the details*

   Response: We prefer to keep the equations and details here for reasons of completeness.

5. *p10 Eq.(6) Any chance this key formula of 2.1.1 could be explained in two sentences?*

   Response: We added a brief explanation below Eq. (6).

6. *p12, the end of section 2.1, Not sure if this is relevant here, but this algorithm sounds very stable once it sits in a local minimum, so how does it ensure that it finds a good global minimum? Any kind of mutation or annealing?*

   Response: We added a comment regarding the balance of local and global search on page 13. Please note however that because of the multimodality of the optimization problem, we cannot guarantee to find the globally optimal solution. There is more explanation of the local and global search procedures in the online supplement, Section 8.1, where the optimizer is explained in detail.

7. *p13 line 288, Along the same line, what about observables where I know that I cannot go beyond a given level of precision?*

   Response: We have amended the text to make the issue of model uncertainties more clear (Section 3, page 15). The limitations of the models are not clearly known. An exception is the case (addressed later) where process-dependent corrections to the hard process might be known but not included for technical reasons.

8. *p14, line 322, Is the limit on z motivated by a number of observations included?*

   Response: The z-value (3) is chosen based on the rule of thumb for selecting a z-value for outlier detection, where almost all of the data (99.7%) should be within three standard deviations from the mean, assuming that the data follows a normal distribution (see Section 3.1).

9. *p15 line 335, At some level, the separation into bins and observables is also ad-hoc, because each bin constitutes a statistically independent measurement, no?*

   Response: Observables themselves are typically chosen to test theoretical or phenomenological models, and the binning is chosen so that it represents the detector resolution. In that sense, we do not feel either the selection of observables or binning is ad hoc.

10. *p17, line 391, Are the results for bi-level and robust optimization produces with a comparable computing time? Ah, found the table, please mention here that it exists.*

    Response: We added a forward reference to the computation times at the beginning of Section 4.

11. *p18 line 422, Would a Gaussianity test of the output be interesting for those metrics?*

    Response: Since the models, $f_i(\cdot), i \in \{1, \ldots, |\mathcal{O}|\}$ are nonlinear, we evaluate the tangent-linear model (TLM) $\mathbf{F}_{\mathcal{O}}(\hat{\mathbf{p}}_{w^*})$ for each observable $\mathcal{O}$. Using the TLMs, the weighted posterior is approximated as a Gaussian around the parameter estimate. Hence, a Gaussianity test should agree that the weighted posterior is Gaussian distributed by design. However, we are unsure about what the referee means by "output". Additionally, we believe that a Gaussianity test should in no way affect the A-optimality and log D-optimality comparison metric since these metrics quantitatively describe the shape and size of the confidence ellipsoid around the parameter. We added a sentence at the end of Section 4.2 to clarify this.

12. *p21 line 457, As a physics reader I would really want to know those tuning parameters and their impact. Can you add an appendix with them, I really think that would be cool to have.*

    Response: The parameters are listed and briefly described in Table 15. We have made a note of this in the text and have expanded the descriptions in the Table (see Sections 4.3 and 8.3.

13. *p22 line 491, Reading this part, the SHERPA data set seems to be much less interesting than the A14 set. What is the reason to include it? Some specific strength?*

    Response: The reason for including the SHERPA dataset was to try out our optimizers on something that had not been done before, thus there is also no expert tune available for this dataset. We want to show the general applicability of the methods to new datasets and tuning tasks. See Section 4.4. for the changes we made.

14. *p23 line 514 "all bins are removed from some observables" which? discuss!*

    Response: In section 4.5, we forward referenced to the table that enumerates the observables from which all bins are removed.

15. *p25 line 545, Please do not just state the comparison results, but explain them or at least put them into context.*

    Response: Thank you for your comment. We added a few explaining words in the document at the end of Section 4.6.1.

16. *p25, Maybe combine Tabs.5-7 into one, with a layout that makes them easy to compare?*

    Response: Thank you for the suggestion. We combine Tables 5-7 and changed the table references in the text accordingly (see Table 5).

17. *p27, Not sure I understand this right, how would this kind of distribution look for a Gaussian statistics sample or for some kind of systematics or anything like that? Maybe I could derive this somehow, but you are the authors :)*

    Response: We updated Figure 3 and added an explanation to Section 4.6.2 to address your comment.

18. *p28, Again, maybe there is a way to combine Tabs. 8-10 into something easier to read and compare and get the point?*

    Response: We concatenated the tables into one for better readability and easier comparison (please see Table 6).

19. *p33, Figure 6, binFiltered, observableFiltered.*

    Response: The figure has been updated (see Figure 5).

20. *p34 line 624, Again, I would argue that in a physics paper this would be where we enter some discussion of these observables and their challenges etc.*

    Response: The discussion of the observables and challenges is already addressed in the original A14 tune paper. A statement about this has been added to the text on page 21. This is a mixture of an applied math and phenomenology paper, and we prefer to keep the text organized as it is.

21. *p35 line 637, "we are not certain about the optima found by the methods." Is this a problem with the global structure of the parameter space?*

    Response: We realize that this sentence is not conveying the message very well. We deleted that part of the sentence and explain what we mean better. The larger values for A- and D-optimality indicate that we are less certain about the validity of the optimal solutions found by the methods for SHERPA than we are for A14. The uncertainty is impacted by (1) the data we are using and (2) the rational approximation we construct. Note that we are not making a statement about the global optimality of the parameters because we cannot guarantee that we found the global minimum. Instead, we use a multistart approach with local optimization to obtain the best local minimum, which greatly mitigates the problem of finding the global minimum. See the end of Section 4.7.1.

22. *p41 line 732, Again a physics question, what does alphaSvalue mean here? It sounds a little weird...*

    Response: This is related to Comment 12. We have added a footnote to the table with a brief description of the physics content of these parameters. See Section 5, page 43.

23. *p43 line 752, The physics discussion is coming too late for the typical SciPost reader. Expand and integrate in the text?*

    Response: Please see comment 20 above, where we have addressed a similar concern. Due to the nature of the paper (mix of phenomenology and applied math), we prefer to keep the structure as is, thereby also minimizing the number of repetitions we have in the text.

24. *p46 line 821, "In fact, even at the time of the A14 tune, methods existed to better describe the Multijets category using Pythia, but it requires a process-dependent correction." Please say it, rather than talking around it...*

    Response: We have amended the text (Section 6.3) to make it clear that we are referring to matched or merged calculations

25. *p49, So here is one thing I am missing - we have these eigentunes, so the authors could take some kind of process and observable not in the tune and illustrate what we can*

learn from the new tunes in terms of uncertainties. Or not? That is what I would have expected as the outcome of the study...

Response: In Section 8.9, we obtain the parameters $\mathbf{p}_b$ and $\mathbf{p}_o$ when the optimization methods are run on data in which some bins or observables are filtered out, respectively. Then we obtain the parameter $\mathbf{p}_a$ when the optimization methods are run on all the data. In Figures 13-16, we show that except for when the bin variance levels $r_b(\mathbf{p}) \leq 10^{-1}$, the bins have very similar variance levels for both kinds of parameters i.e., $\mathbf{p}_a$ and $\mathbf{p}_b$ or $\mathbf{p}_o$. This shows that removing bins or entire observables from the tuning process does not reduce the information required to achieve a good tune as it performs very similarly to when all bins are used for tuning. As for when $r_b(\mathbf{p}) \leq 10^{-1}$, the disagreement in the cumulative distribution of bins is not significant since the number of bins is small with all of them having small levels of variance. We updated the text in Section 8.9 to reflect this.

Moreover, we updated our Eigentune section (Section 5) and show on two observables (Figure 8) the uncertainty bands obtained with the eigenvectors.

Response to Reviewer 3

i *Strengths: New procedure for computing demanding task, that is reduced to a few hours run with the proposed methodology. In addition, overall quality improvement demonstrated with specific examples. A clear improvement over the state of the art.*

Response: Thank you for your positive comments, we appreciate your support.

ii *Weaknesses: The procedure follows closely what is done normally (a chisquare fit over histograms). So, as much as the accepted procedure, it makes little sense to me. Correlations are systematically neglected and the N-dim distribution of the observed quantities taking as input is simplified to a set of 1D distributions. In my mind, this can introduce biases in the problem.*

Response: Please note that the innovation of our proposed approach does not lie in finding new ways to solve the χ^2 minimization problem (the inner optimization problem). The innovation lies in finding the best weights automatically and without bias. We do not have correlations for the data available, thus we are not systematically neglecting them. If we had the correlation information, it would have to be incorporated into the χ^2 formulation (the inner problem), which does not change the need for improved methods for finding the weights. We discuss the formulation and challenges in performing the χ^2 optimization with N-dimensional distribution at the end of Section 7.

iii *The paper discusses the task of MC tuning and proposes a new procedure to improve over current state of the art, by speeding up the computation and reaching a better agreement on a real-life example. This is done utilising a few new elements, including a better minimisation strategy (a few are proposed) and an outlier removal procedure. I have a few questions regarding the big picture proposed: 1) To which extent the outlier removal is sound? Even in absence of systematic issues, outliers will occur. Removing them might help to established the expectation value of the parameter one is fitting, but any sense of statistical interpretation of the uncertainty range is lost (i.e., coverage is broken). The paper offers no discussion of this point and how crucial this is.*

Response: As explained in the response to reviewer comment 25, removing bins or entire observables from the tuning process does not reduce the information required to achieve a good tune. Additionally, the filtering approaches only eliminate parts of the MC model that are highly unlikely to be explained by data. Hence, it is a conservative approach since the range of the function within the domain is usually much larger than the range of the values that could be used to fit the data. The outlier removal is based on the intuition that the models that are highly unlikely to be explained by data could be removed to (a) get a better estimate of the tune, and (b) prevent the algorithms from going into regions of extrapolation. We added this explanation in Section 7.

iv *2) The paper is TOO LONG. The same content can be delivered in 1/2 the length. Authors explain established concepts (I doubt that one needs to explain what a histogram is in a physics paper). Often, the authors repeat sentences twice, assuming that this comes as an explanation (e.g., page 34). I think that an effort should be taken to reduce the paper length and make the paper more readable. The text can be reduced with no impact on the amount of transmitted information.*

Response: Please note that we have made an effort to reduce the length of the paper, for example by eliminating figures and descriptions that are not necessary for the audience and by concatenating tables where appropriate. These updates are throughout the manuscript and indicated by strike-out text and rephrased sentences.

v *3) I would expect that a paper of this kind would address what I consider the elephant in the room. MC tuning uses a set of correlated 1D quantities as uncorrelated quantities, instead of taking as input an N-dim distribution. To me, this is potentially dangerous. At least, I would expect a paper of this kind to discuss this as a potential problem and discuss the balance between what is right and what is doable with the existing information. It is true that authors discuss input weights to alleviate the issue. But this sounds to me as the survival of the "by hand" intervention that this paper aims to remove from the tuning procedure, since there is an arbitrariness in the weight setting strategy (it is not obvious to me that Eq.3 is the only choice).*

Response: The reviewer is correct, ideally, we would have an MC generator that can take as input an N-dimensional distribution. However, we do not have that available. We also do not have the data correlations available. As explained in comment (ii), our method does not aim at improving the χ^2 optimization, but rather to improve the current method of adjusting the weights. Had we correlation information available, our method would still be applicable. We discuss the formulation and challenges in performing the χ^2 optimization with N-dimensional distribution at the end of Section 7.

vi *4) The analysis of the accuracy benefit is left on a qualitative level. I would be interested to understand the coverage property of the fitting method, the validity of the quoted uncertainties, and the capability to converge to the correct minimum.*

Response: Coverage and validity of quoted uncertainties: As explained in the response to reviewer comment 25, removing bins or entire observables from the tuning process does not reduce the information required to achieve a good tune. If we had unlimited resources, we could potentially perform some sort of k-fold cross validation study to better understand this performance. Another idea would be to perturb the parameter values, run the MC at these parameter values, and perform the study with this data. But we would require a lot of resources to run the MC at these new parameters. Also, none of these approaches deal with the fact that the approximation model is imperfect due to the gap between the model and the MC event generator. To address this, we need to use the MC directly in a derivative-free optimization approach. The details of this approach and its challenges are briefly described in Section 7.

Convergence: Since the problem is a non-convex global optimization problem, there are likely many local optima. It is not possible to guarantee convergence to the globally optimal solution. For any fixed set of weights, the lower level optimizer (APPRENTICE) is able to locate a local optimum. In the bilevel optimization, however, this does not necessarily imply that one will also converge to a locally optimal solution in the weights space. The bilevel surrogate model optimizer is set up such that it never samples a weight vector more than once. Therefore, if we keep sampling, eventually we densely sample the whole weights space and thus we will sample at the globally optimal weights. However, since we cannot guarantee that the lower level problem gives us the globally optimal parameters for any weights, we are not able to guarantee that we recognize the globally optimal solution in the weights space. The robust optimization approach has nonlinear constraints that may result in locally optimal solutions for the parameters and the weights. See Sections 2.1.3 and 2.2 for an update of the manuscript.

vii *5) Related to the previous points, I have the impression that the paper would benefit from including a closure test on some toy dataset in which one knows the "right" answer and could demonstrate that the procedure would provide an unbiased result. In view of the previous comments, I think that a revision is needed. Considering the last few comments. I also ask the authors to consider the list of proposed changes, given below.*

Response: We thank the reviewer for the suggestion. For the physics datasets, we see that the best approach depends on the metric that the physicist wants to minimize. This happens because we propose multiple different algorithms for finding the optimal weights that are based on different optimization objectives with nonlinear approximations. Hence, with each method, we obtain different parameter (and weight) results that optimize the corresponding metric.

However, for a toy linear model (where the experimental data is made up of constant variance and mean values obtained by computing a linear function for known parameters), it is possible to recover the same known parameters with the proposed approaches. We describe the setup of the toy model and the results in Section 4.8.

Requested changes

a *Sec 1.1: I think one can live w/o the explanation of what a histogram is*

Response: Thank you for your comment. Please note that in the revision, we deleted the description of the histogram and the example histogram (Section 1.1) and forward referenced to a later histogram in the results section as an example for one.

b *Pag.6: Isn't this replacement of the histogram by a surrogate model subject to a systematic uncertainty related to the choice of the functional form? I would expect an* *analysis of performance benefits vs accuracy costs.*

Response: We agree with the reviewer that the specific choice of the surrogate model for the MC simulator introduces uncertainty. In the present study, we analyze the results obtained with a cubic polynomial (as done in [1]) and a rational approximation (see [2]). Our optimization methods are general enough to use different types of surrogate models. The type of surrogate model used will impact the uncertainty. However, quantifying the uncertainty due to model choice / misspecification will introduce another level of complexity in the problem and is outside the scope of this paper. We added a note regarding this issue in the manuscript below equation (2).

c *Eq. 3.b. s.t. for "subject to" is a very uncommon notation for the physics literature I am used to. You might want to consider to re-format the three equations*

Response: Thanks for pointing this out. We replaced "s.t." with "subject to" in Eq. (3b).

d *Pag 13: Are the optimized parameters varied in a range? If so, a discrepancy could also be induced by too strict boundaries. This possibility is not considered but it might be relevant.*

Response: We thank the reviewer for this observation. For both datasets, the parameters were varied in predefined intervals. For the A14 dataset, the parameter bounds were carefully chosen such that the polynomial parameterizations are valid within the bounds and to give a physically meaningful coverage in the sense that the experimentally observed data was "covered" by the range of predictions (see the update to Section 4.3). The optimal parameters found by the different methods all lie inside the intervals, i.e., not on the boundary. This indicates that the specified bounds were appropriately chosen and no extrapolation happened. For the Sherpa dataset, however, multiple parameters ended up on the boundary, meaning that the model extrapolated and better results can potentially be found by adjusting the boundaries. Additionally, we could solve this problem in an unbounded/domain bounded only by physics constraints using the MC directly in a derivative-free optimization approach as described in comment vi. The details of this approach and its challenges are briefly described in Section 7.

e *Pag 15: the offered procedure is effectively a single-hypothesis test. Only Fisher called this kind of test hypothesis test. He disagreed with the Neyman Pearson hypothesis test, which requires two hypotheses to be specified. My understanding is that Fisher's problem is ill-posed, as decades of frequentist literature exposed problems related to this. In particular, no any claim of optimization in this context is typically an overstatement. This point should be discussed.*

Response: We thank the reviewer for this comment. Fisher's approach considered the p-value could be interpreted as a continuous measure of evidence against the null hypothesis. On the other hand, Neyman Pearson's approach said one can use the p-value to either reject the null hypothesis, fail to reject the null hypothesis, or end up with type 1/2 errors. Given these definitions, the approach described in Section 3.2 is a Neyman Pearson approach since every subset $\mathcal{B}$ of observable $\mathcal{O}$ is tested to check if the null hypothesis is rejected or otherwise. We realize that the explanation of the optimization formulation may be misleading and we have added a few clarifying sentences below equation (15) to help the reader better understand our approach.

f *Line 451: "The main reason for this discrepancy is the fact that we use a better optimization routine" Certainly a different one. But better in which context? I think the paper should offer more evidence of this claim. My understanding is that most of the improvement comes from the surrogate model, which has pros and cons (see previous comment) that should be assessed.*

Response: We realize that our description may have not been precise enough and we updated it (see Section 4.3). In the original study, the optimizer Minuit was used to obtain the NNPDF parameters. In the present study, within Apprentice, we use the truncated Newton method as default, and thus we can take advantage of the exact gradient and Hessian information, which Minuit did not have. Moreover, Apprentice is significantly faster than Minuit, which allows for an efficient multistart optimization strategy and thus increases the probability of finding better optima of the χ^2 minimization. Multistart was not done in Minuit. Regarding the choice of the surrogate model, we included the cubic polynomial and a rational approximation in the study. For a comparison between the two, we refer to [2]. For updates in the manuscript, please see above Equation (2) and Section 4.3 (page 22).

g *Table 3: I assume the quoted uncertainties correspond to a "1-sigma" range because it's related to the eigenvectors of the linearised problem. If so, it should be clarified. I have doubts about the fact that your outlier removal procedure is not altering the statistical interpretability of these uncertainties.*

Response: We added "68%" confidence level in the caption of Table 3 to make it clear. It turns out the eigentunes presented in the last draft were intermediate results; we updated them in the new draft and they are now consistent with those in Section 5 (Eigentune).

The "outlier" identification was verified manually and carefully reviewing each observable by domain experts. The removal of certain observables was primarily driven by physics, i.e., removing them makes the interpretation of the tuned physics parameters sensible. It does not alter the statistical nature. However, it could possibly change the tuned physics parameters. The tuned parameters presented in Table 3 are obtained with the same weights and data as those used in the A14 paper and we found very similar results. That gave us confidence in proceeding with our study of automating the weight adjustment.

h *Line 544: Isn't this obvious (and not necessarily right)? Outliers were removed, so I would expect exactly this. Am I too naive? I would like to see this procedure repeated on a toy data in which you know that the method can describe the model, i.e., in which you should not remove anything in principle. What happens? (see general comment in the report). I think this deserves some comment.*

Response:

This sentence conveyed an incorrect message and we have added/updated the text at the end of Section 4.6.1 and Section 8.9 to correct it. Our response to this comment also includes our response from comment (25).

j *Line 559: I am concerned with the fact that 50% of the points come from an observable. Are your weights acting as a regularization of this? My understanding was that their main purpose was to limit the redundancy among correlated quantities. Or does the number of bins/observable enter?*

Response: We thank the reviewer for bringing this up and we have improved our description of the results (see page 27). In fact, our study uses the same data as used by ATLAS. Figures 3 and 4 show that the relative contribution from the track jet properties seems to be the largest. However, from Table 7, we can see that (1) Robust optimization is in fact able to detect the redundancy in the track jet properties observables and it gives only one subcategory a large weight while setting the weights of the other 3 subcategories to (near-)zero. In contrast, the bilevel optimization methods have a different goal, namely fitting each observable approximately equally well, thus mimicking the tuning performed by the expert. Thus, the outcomes of the bilevel optimization (in terms of weights) are similar to the expert tune (the weights for all track jet properties observables are about equal). Lastly, from Table 7 we can also see that the expert assigned large weights to the multijets and the fit (Figure 4) is accordingly better. See also Sec 4.6.4 for additional discussion.

k *Fig. 6: The label on the bottom-right is cut.*

Response: Thanks for pointing this out. The figure has been updated (now Figure 5).

l *Sec 4.8: why not running everything on one machine and make a one-to-one comparison?*

Response: Please note that we were not primarily concerned with the run time of the methods, but rather automation, and one goal was to ensure that the code runs on different architectures. See Section 4.9.

References

[1] A. Buckley, H. Hoeth, H. Lacker, H. Schulz, and J. von Seggern, "Systematic event generator tuning for the LHC," *The European Physical Journal C*, vol. 65, pp. 331–357, 2010.

[2] A. P. Austin, M. Krishnamoorthy, S. Leyffer, S. Mrenna, J. Müller, and H. Schulz, "Practical algorithms for multivariate rational approximation," *Computer Physics Communications*, vol. 261, p. 107663, 2021.