We would like to thank the two referees for the appreciation of our work and for their comments, which we address in turn.

Referee 1

- *This article is a valuable contribution to the ongoing study of the best way to utilize the SMEFT as a tool to constrain generic new physics using precision measurements at the LHC and other experiments; it explores the possibility of using reweighting techniques to produce a fast approximation to the results of a complete fit in the presence of a new data source.*

  We thank the referee for her/his kind words and for the positive feedback on our work.

- *I note that there is a notational error occurring in equation (2.9) and following; one imagines that $F_{(\mathrm{rw}}(\langle c_i \rangle))$ does not need the additional open-parens. On a similar note, there are typos on page 10 (ne for new), in the caption to Figure 4.3 (fo for of), and on page 14 (NNDPF for NNPDF) that jumped out at me.*

  We have corrected all these typos. Note that Eq. (2.9) is now Eq. (2.13).

- *Moving to address the physics and statistics of the contribution, I do have a couple of questions that aren't fully explained (at least at my level of comprehension) in the article. The most troublesome point is the inclusion of quadratic EFT effects, that are $\mathcal{O}(\Lambda^{-4})$ in the signal function. The effects at this order in the EFT expansion are not fully given by the terms kept by the authors, and following the usual rules of perturbation theory calculations these ought to be dropped, with a theoretical error introduced to parameterize our ignorance of the size of the effect at this order. With the current treatment, which is sadly common in the field, "constraints" are regularly produced on the EFT parameter space which do not hold in more-complete models, admitting instead model-building workarounds, which would not be the case with a robust theoretical treatment of the EFT expansion. It would at least be very beneficial to understand, in every article produced on this topic, how impactful the "quadratic" terms are on the fit itself (which the authors have tersely explored in their previous work, citation [23]).*

  In this respect, we shall emphasise that the validity of Bayesian reweighting is independent of the theory assumptions used. To demonstrate this statement, in particular when only the $\mathcal{O}(\Lambda^{-2})$ corrections are included in the theory calculation, we have redone the analysis leading to Figure 3.3 of the original manuscript; now, we have included only $\mathcal{O}(\Lambda^{-2})$ corrections in the prior fit, in the calculations necessary for reweighting, and in the new fit. This exercise has required to produce a new prior, made of $N_{\mathrm{rep}} = 10^4$ replicas, and a new fit with consistent theory.

  The results of this analysis are summarised, in the case of theory calculations truncated at $\mathcal{O}(\Lambda^{-2})$, in Fig. 1 below. We observe that, for those operators that are affected by the new data the most (for instance those which show an uncertainty reduction of 40% or larger, denoted by the horizontal dashed line), the bounds obtained with reweighting are reasonably similar to those obtained with a new fit. This test illustrates that Bayesian reweighting can be used to constrain the parameter space irrespective of the theory assumptions used in the prior, in particular whether (or not) $\mathcal{O}(\Lambda^{-4})$ terms are kept in the SMEFT computations.

  The list of operators that are affected by the new data the most differs between the $\mathcal{O}(\Lambda^{-4})$ and the $\mathcal{O}(\Lambda^{-2})$ fits. In the former case, they are: O13qq, OpQM, and OtZ; in the latter case, they are: O13qq, OpQM, and OtW and Opt. This is unsurprising, since in general each calculation probes different regions of the parameter space (as discussed in Ref. [23]).

  In the revised version of the paper, we have emphasised that Bayesian reweighting can be applicable irrespective of the theory assumptions adopted, provided that they are consistently adopted in the prior fit and in the computations needed for reweighting. We have mentioned that we have explicitly validated the method by using fits based on SMEFT calculations truncated at
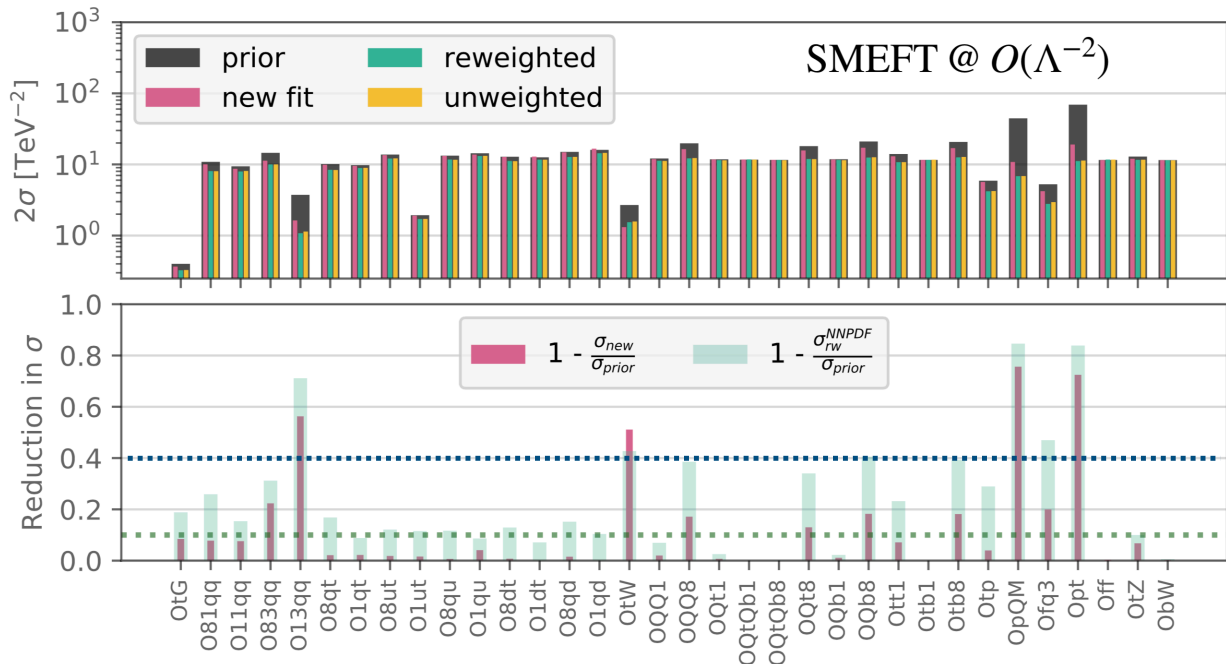
Figure 1. Same as Figure 3.3 in the original manuscript but now with the SMEFT calculations truncated at $\mathcal{O}(\Lambda^{-2})$ instead of $\mathcal{O}(\Lambda^{-4})$ (which is the default choice used in the results shown in the paper) both in the prior fit, in the posterior fit, and in the reweighted results, The dotted horizontal line in the bottom panel indicates an uncertainty reduction by 40% (for visualization purposes).

$\mathcal{O}(\Lambda^{-2})$ or not. To this purpose, we have added a couple of new paragraphs at the end of Sect. 3.3. A more extended discussion on whether or not it is advisable to include $\mathcal{O}(\Lambda^{-4})$ corrections in a SMEFT analysis is beyond the scope of our study, and thus not further investigated here.

- *More generally, it is very important to acknowledge that the EFT is a series expansion in something like s/Λ^2, and as such it doesn't make sense to consider scales $\Lambda \simeq 1$ TeV at the LHC; it's perfectly clear that this will not converge. I would recommend then that the authors re-benchmark their scale of new physics to $\Lambda \simeq 5$ TeV instead.*

As the referee knows very well, in a given SMEFT analysis one can only be sensitive to the ratio c_k/Λ^2, therefore it is not possible to access Λ directly from the data. While we have assumed $\Lambda = 1$ TeV for illustrative purposes here, it would also be possible to present our results for any other value of Λ. For instance, interpreting the results in terms of $\Lambda = 5$ TeV would lead to an increase in the 95% CL bounds reported in the paper by a factor $(5\text{ TeV})^2/(1\text{ TeV})^2 = 25$, and likewise for any other value of Λ. However, the validity of the Bayesian reweighting method does not depend on the interpretation of the results in terms of a specific value of Λ.

We have added a new paragraph at the end of Sect. (3.3) which emphasises that the value of Λ itself is not directly accessible at the fit level, and that the results that we quote for $\Lambda = 1$ TeV can be rescaled to correspond to any other value of Λ. Specifically, the upper and lower bounds of the Wilson coefficient c_k need to be rescaled as

$$\delta\widetilde{c}_k^{\pm} = \delta c_k^{\pm} \times \left(\frac{\widetilde{\Lambda}}{\Lambda}\right)^2. \tag{1}$$

We have also added an explicit calrification that any SMEFT analysis should validate their assumptions to fulfil the EFT regime, as already discussed by some of us in [23].

- *I'm also confused by the comments regarding double-counting and whether or not it is problematic in this context on page 6; if the goal is to explicitly reproduce the results of the fit of [23] through sub-fitting and then reweighting why should I be including additional data which was excluded there? Does the reweighting procedure meaningfully depend on that data to reproduce the fit accurately? More generally, it isn't clear to me why double-counting would be less of a concern in the reweighting, fast-fit production context than it is in the context of a full-blown fit.*

When extracting information from experimental data one should be careful not to count the same measurement twice. This is true both for PDF and SMEFT analyses. In the former case the issue of double counting was discussed in some detail in [JHEP 1704 (2017) 044]. If one includes absolute differential distributions for top quark pair production, then one cannot include the corresponding total cross-section, since the two measurements are actually extracted from the same data set. If one includes the normalised distributions, then including the total cross-sections is fine since there is no double counting.

In the SMEFiT analysis of Ref. [23] we proceeded in a similar way: we included the absolute differential distributions of top quark pair production, and we excluded the corresponding total cross-sections. In the present analysis, however, we inadvertently included also these total cross-sections. This has been done everywhere: in the generation of the prior fit, in the computation of the SMEFT theory used for reweighting, and in the generation of the new fit. Since we are interested only in checking the validity of reweighting, this is therefore not a problem. However, because of double counting, we acknowledge that the bounds on some operators could be slightly more stringent than what they should be. Hence our comment concerning the double counting, that anyways does not affect the validation of Bayesian reweighting presented in our work.

- *The testing for reliability of reweighted results based on the KS statistic seems to have been employed in a thoroughly ad-hoc way here as well; is there any mathematical/statistical reason why we should expect these thresholds of 0.3 or 0.2 to be dispositive as to the reliability of the reweighting procedure? Given the differences in adopted thresholds for different sub-analyses, is there some meaningful interpolating formula, perhaps one that takes into account the amount of data points being added, which could suggest a reasonable threshold for reliability in future reweighting exercises? Also, and more worryingly, given that e.g. the result for* `O13qq` *is officially reliable but the result for* `Ofq3`, *which is very strongly correlated with* `O13qq` *in the full fit of [23], is not, how are we to think about correlations and flat directions in the reweighting framework?*

This is indeed an important point. We agree with the referee that the choice of the thresholds to determine when the reweighting of a given SMEFT degree of freedom is reliable or not is currently mostly based on a case-to-case phenomenological evidence. In this respect, the difficulty in defining objective criteria to establish when reweighting is trustable reflects the fact that the SMEFT parameter space is very different from the PDF one. In a PDF fit all directions are somehow correlated, *e.g.* by DGLAP evolution or sum rules, while in the SMEFT case each direction is in general (theoretically) uncorrelated from the others. Therefore, while reweighting is conceptually the same in the PDF and the SMEFT cases, the interpretation of the results is more delicate in the latter.

We aim at defining more objective criteria to identify for which SMEFT operators the reweighting method can be trusted in future work. While we recognise the value of the referee's comment, we prefer to resort to the current ad-hoc criteria for the time being, which anyways seem to work reasonably well in the cases that we have explicitly studied.

We have added a comment stating that the current criteria to select the operator are derived mostly on phenomenological evidence, and that the search for more formal and robust criteria will be addressed in future work, see Sect. 3.3 towards the end of the paragraph after the bullet points on page 11.

- *I also am struck by the (admittedly not phenomenologically relevant) increasing feature in Figure 4.2 in going from 6 to 7 datasets included; I would naively have assumed that introducing additional data should only be further narrowing the range of replicas that were consistent with the data, but that doesn't seem to be the case here. Is this behavior understood by the authors? A short comment explaining it would be valuable to the reader I believe.*

  When adding new measurements, the value of $N_{\rm eff}$ should decrease up to statistical fluctuations (reweighting it is not a deterministic procedure). These statistical fluctuations are negligible when the number of starting replicas is large enough. However they can appear when we deal with a smaller sample of replicas and with new data that does not add any new information, as is the case here between 6 and 7. We have added a comment to clarify this state of affairs in a new footnote on page 15.

- *Finally, I find myself thoroughly confused by the discussion of NNPDF versus GK weights; it is clear that having accidentally found something that fits very well can be damaging in the case of GK weights, but it isn't obvious why that sample should be fully discarded as in the NNPDF formalism; fitting well doesn't generally get punished in statistics, but it shouldn't be overly rewarded. Is it plausible that some middle-ground treatment exists, which for instance treats any fit better than that which maximizes the NNPDF weight as equally-worthy with that maximizing fit?*

  The rationale for comparing results obtained with the NNPDF and GK weights arises from previous studies, based on different analytical (theoretical) arguments, that claimed in turn that either the former or the latter gives the correct answer. Here we want to revisit such claims from a purely phenomenological point of view (the discussion on the mathematical derivation of the weight expressions is beyond the scope of this paper). It should be clear that some form of reweighting will always work (namely by removing from the sample those replicas that disagree the most with the new data). The question we wanted to address is whether or not NNPDF and GK weights lead to the same results in the SMEFT case, and if so under which conditions. In our studies, we have found that both choices of weights lead to comparable results, provided that the initial number of replicas is sufficiently large; however, NNPDF weights are more efficient than GK weights, *i.e.* they reproduce the fit results with a larger number of effective replicas.

  All this said, in principle it might be possible that some alternative form of the weights exhibits an even superior performance, though we are not aware of any theoretical work that motivates such a choice. Nevertheless, to investigate this point from a phenomenological point of view, we have repeated the reweighting exercise that leads to Figure 3.3 (including the t-channel single top production measurements). Now, we have used a one-parameter family of weights

  $$\omega_k^{(p)} \propto \left[\left(\chi_k^2\right)^{(n_{\rm dat}-1)/2}\right]^{1/p} \exp\left(-\chi_k^2/2\right), \quad k=1,\ldots,N_{\rm rep}, \tag{2}$$

  where p is a parameter that interpolates between the NNPDF weights ($p=1$), Eq. (2.4), and the GK weights ($p \to \infty$), Eq. (4.1). We have re-evaluated Figure 3.3 for different choices of p, and we have estimated the corresponding effective number of replicas. This is what one finds for two intermediate values, $p=2$ and $p=3$:

p	1 (NNPDF)	2	3	∞ (GK)
$N_{\rm eff}$	306	115	68	22

  Therefore there is no benefit in using the intermediate weights option: all values of p decrease the efficiency of the reweighting procedure in comparison to the NNPDF case.

Furthermore, we have verified that, provided the effective number of replicas N_{eff} is large enough ($N_{\text{eff}} \sim 100$), the results obtained with Eq. (2) still reproduce the corresponding results of a new fit. This is shown explicitly in Fig. 2 below, the counterpart of Figure 3.3 but now using the hybrid weights in Eq. (2) with $p = 2$. As can be seen, for those operators for which the KS-statistic and the reduction of uncertainties lie above the given threshold, the fit results are well reproduced by the reweighting with these hybrid weights. However, for $p = 2$, we end up with $N_{\text{eff}} = 115$ effective replicas, a value significantly smaller than the one obtained in the case of NNPDF weights ($N_{\text{eff}} = 306$).

Figure 2. Same as Figure 3.3 in the original manuscript but now using the hybrid weights Eq. (2) with $p = 2$.

We have added a new subsection (4.2) to discuss the results obtained with hybrid weights, that contains in particular the new Fig. 2.

Requested changes

1. *Correct typographical errors in notation and text as identified.*

   Done

2. *Renormalize theory to a cutoff scale $\Lambda \sim 5$ TeV where the SMEFT approach is theoretically consistent.*

   As discussed above, we have added to the text (see the discussion around the new Eq. (3.1)) a clarification to remind the reader that the fit is sensitive only to c_k/Λ^2 and not to Λ alone, and that results for other values of Λ can be obtained using the rescaling defined in Eq. (1).

3. *Explore the importance of keeping versus dropping "quadratic" contributions to the reweighting procedure, as well as the impact of introducing new theoretically errors for missing higher orders in Λ^{-2}*

   See the discussion above. We have added a short summary of the studies presented in this reply by adding two new paragraphs at the end of Sect. 3.3. They demonstrate that reweighting works as well when the calculations are based on $\mathcal{O}(\Lambda^{-2})$ theory as when the calculations are based on $\mathcal{O}(\Lambda^{-4})$ theory, used as a default in the paper.

4. *Explore and discuss the impact of flat directions on reliability metrics for reweighting results.*

   See the discussion above. We believe this is an important point but beyond the scope of this paper and will be left for future work.

5. *Add explanation for rising feature to caption of Figure 4.2.*

   Done, see above and the new footnote on page 15.

6. *Explore and discuss potential intermediate-case weights for the reweighting procedure.*

   Done, see above and the new subsection 4.2.

Referee 2

In the manuscript, the authors investigate if Bayesian reweighting can be applicable to the SME-FiT framework (global analysis of SMEFT) using the data set used in previous literature. Performing a new fit on the newly updated data set might be computationally expensive when a large number of EFT coefficients and/or a large set of data are involved. The Bayesian reweighting method seems to allow us to estimate the impact of those new extended data on the EFT parameter space. I find that the approach proposed in the manuscript is an interesting idea and it will be eventually very useful if it works as was described. The importance of the global fit in the SMEFT is growing, and we see various examples where the global fit makes a significant difference. However, even at the level of the dimension-six operators it is difficult to carry out the global fit due to a large number of operators. This type of approach in the manuscript looks in the right direction to make the SMEFT more practical and to improve the mapping between data and Wilson Coefficients (WCs) in a more accurate way. I feel that the review in Section 2 is too short to fully understand how the method works. I would suggest to make it more self-contained. The remaining of the manuscript is clearly written and organized. I only have a few basic questions. I do not mean that they have to be implemented in the revision if they are too basic or not necessary. It is up to the authors.

We thank the referee for her/his kind words and for the positive feddback on our work.

1. *Section 2: It would be helpful if some explanation on how N_{rep} Monte Carlo replicas are generated from the initial data set (instead of just giving a ref) since it seems essential to understand it. Similarly for the chi-squared in eq. 2.4. When comparing the theory prediction from each replica with the new data set in chi-squared, is the uncertainty of the new data set also taken into account?*

   We have extended the discussion in Sect. 2 to make the paper self-consistent. In particular we have added a discussion on how Monte Carlo replicas are affected by the inclusion of additional data sets and which definition of the χ^2 function is used for both the fits an the reweighting.

   Note that when we compare the theory prediction for each replica to the new data by means of the χ^2 we take into account all the relevant sources of uncertainty in the same way as we would do in a fit. This is crucial, because otherwise the results of a new fit and of the reweighting will not coincide. This point should now be clearer in the extended version of Sect. 2.

2. *Section 2: Is it obvious why all N_{op} k-th replica has a universal weight? A theory prediction (for a new measurement) constructed from k-th replica might involve smaller set of WCs. Then, do not we need reweighting of those finite subset of WCs?*

   The weight of a given Monte Carlo replica is determined by the value of the χ^2 to the new data. Each replica corresponds to a given point in the SMEFT parameter space and is associated to a given weight that cannot be separated into the contributions from individual operators. However, in the exercise we presented in our paper, the theory prediction is always the same, only the values of the Wilson coefficients are changed, and the relevant basis of operators is always the same.

   Note that, for those operators that are not being constrained by the new data, the effect of reweighting will be negligible: in other words, they will have exactly the same probability distribution before and after reweighting. This can be seen in plots such as those in Figure 3.3: for most operators the value of the KS statistic is very small when adding single-top data. This is a nice feature of the method: it automatically identifies the subset of operators that are being constrained.

   A limitation of the method is that one cannot extend the operator basis probed in the prior fit. In other words, the method cannot be applied to a prior determined from top data only

to include measurements where new directions in the parameter space open up (such ass Higgs data). In this case, carrying out a new fit is required.

3. *Section 2: In this approach, is it essential to have an ensemble of all WCs (including those that will be covered in the new data set) in the prior irrespective of the status of the initial data set? For instance, suppose, we have an initial data set that is sensitive to a subset of WCs (say set I). Later, new data set is added, and new one is sensitive to mostly different types of WCs (say set II) compared to the initial ones. Is it correct that at least sets I $\bigcup$ II have to be included in the ensemble of prior for this approach to work? In this situation, if we decide to add a second new data set later which is sensitive to WCs set III. Then, we have to go back to the beginning and should start again with constructing an ensemble of MC replicas for WCs in sets I $\bigcup$ II $\bigcup$ III?*

   This is correct, and, as also mentioned above, it corresponds to an intrinsic limitation of the method. A prior fit based on data set I can only be used to describe data set II via reweighting, provided that no additional dimension-6 operators are required in II. In other words, the set of operators relevant to describe I should be the same as those required to describe I $\bigcup$ II. Note however that this does not imply that II will constrain *all* the WCs: in general *only a subset* will be constrained (as shown by the case of single-top measurements).

4. *Section 3.2: In Fig. 3.2, N_{eff} abruptly drops when the s-channel measurements are subsequently added. It is explained that it is due to a large amount of new information being added, specifically sensitivity to new combinations of SMEFT parameters that are unconstrained by the measurements previously considered. I feel that this goes against what one does in the SMEFT. In the New Physics search via the SMEFT, our goal would be to come up with as many new measurements that give constraints on new combinations of WCs as possible or constraints on as many new set of WCs as possible. It sounds like the better new types of processes we come up with, the less the approach become effective. I wonder if the authors have any comment on this.*

   The drop in the number of effective replicas N_{eff} actually shows that this is the case: once the s-channel data is added, we are including a lot more information into the fit, which is what we want. One can think about the amount of information that a new data set is adding as being measured by $N_{\mathrm{eff}}^{(1)} - N_{\mathrm{eff}}^{(2)}$, the difference between the values of N_{eff} before and after reweighting. Since $N_{\mathrm{eff}}^{(1)} - N_{\mathrm{eff}}^{(2)}$ is large when adding s-channel data, then we conclude that this data set is bringing in a lot of new information.

   Perhaps here there is some confusion since N_{eff} needs to be used in two different ways: one to quantify the amount of new information being added (measured by $N_{\mathrm{eff}}^{(1)} - N_{\mathrm{eff}}^{(2)}$); and the other to monitor the validity of the method: if N_{eff} drops below about 100 replicas or so, the method becomes unreliable. That is, the ensemble of left replicas will not have the required statistical power to faithfully describe the underlying probability distribution in the space of Wilson coefficients. Therefore, the approach does not become less efficient if the process is more constraining: it only indicates that we would need to start with a much larger sample of initial replicas to use reweighting. Provided that the initial number of replicas is large enough, then one can add as many new experiments as possible (and as constraining as possible).

Requested changes

- *Please improve section 2 so that it becomes more self-contained.*

  We have extended the discussion in Sect. 2 to ensure that the manuscript is self-contained. In particular, we have added information about the SMEFiT methodology and the explicit form of the χ^2 estimator that we use.