Dear Referee,

We would like to thank you for reviewing this paper and furnishing this report.

We have carefully considered the comments, and we have applied the corresponding changes to the original version of the paper to address the issues raised. A detailed response to the comments can be found below.

We are at your disposal for any further clarifications and/or additional information.

Sincerely,

Robert Schöfbeck, Matteo Presilla, Charlotte Knight, Saptaparna Bhattacharya

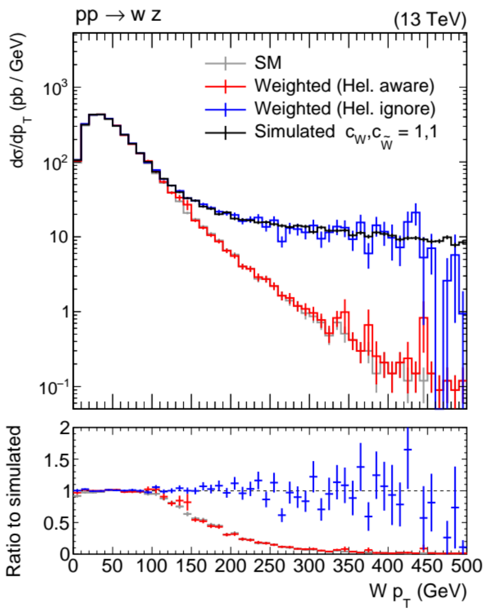
Dear authors,

many thanks for the well-researched and instructive paper! It was an interesting read and I learnt a lot from it. Overall, it is well written and the content is very much appropriate for this journal. However, before publishing it I would like to ask for a few clarifications, a bit more context at times, and improvements regarding formal aspects. These requests are summarised below. I hope I am not asking for unreasonably time-consuming additions, as I would be happy to see this paper in the journal very soon.

Comments on the content:

Fig 3: It is said that the helicity-ignorant reweighting is a bit off at high pT because "statistics is small" (more proper to refer to it as "small sample sizes"), but the error bars in the plot suggest that this is a significant effect. Does this comment refer to statistical fluctuations in the weight calculation? If so, would it be possible to quantify the size of the corresponding uncertainty? This information could e.g. be added as an additional set of error bars in the plot, or by giving the value of the uncertainty in the final bins in the text. If the source of the discrepancy is something else, of course that would be very good to clarify (and quantify).

We have changed the text to "small sample sizes". We have investigated that fluctuation in the tail and the discrepancy indeed appears to be a statistical fluctuation. The discrepancy at very high p_T for the helicity-ignorant reweighting was affected by an erroneous veto on large weights. Removing the safety condition, the agreement between helicity-ignorant reweighting and simulated predictions improved at high p_T. Fig. 3 (in particular the top left panel) is updated in the paper. As can be seen in the finer binned version of Fig. 3 (top left) a small number of events that receive a large weight in the helicity-ignorant scenario are the likely cause of the discrepancy. In a coarser binning the agreement is now within uncertainties.



Sec 5.2:

- Is there a reason why you mention in the text that pT(H) is an important probe to study BSM effects, but then use pT(Z) in the plots? Of course at LO these may be very strongly correlated, but is there a benefit to showing pT(Z) over pT(H)? (I do not suggest to change any plots, just asking for clarification.)

It is correct that $p_T(H)$ and $p_T(Z)$ are strongly correlated at the LO, so their distributions exhibit similar SMEFT sensitivity. There is no particular reason for this choice. We change in text "at high values of Higgs boson p_T" to "at high values of the Higgs or vector boson p_T".

- The description of helicities of reflected Z-bosons and angular distributions is not easy to follow. Would it be possible to e.g. add the plane of reflection into the sketch or show a distribution where the mirrored contribution is apparent?

We believe the word "reflected" causes confusion here. It is not used in the geometrical sense. We modify the text to
"The HVV operators, on the other hand, also modify the interference of scattering amplitudes with transverse Z boson helicities. This affects the distribution of angular observables Θ, hat-**θ**, hat-**ϕ**, which are measured in the ZH rest frame [37] as depicted in Fig. 4"

Sec 5.3:

- Unlike the previous sections, no PDF set is given here. For completeness, can it be added?

It is NNPDF 3.1 NNLO, and the information has been added to the text.

- Judging from the bottom left plot of Fig 8, the dedicated points used for ctGRe are not only -0.4, -0.2, 0.2, 0.4 as stated, but also -0.7 and 0.7. Is this correct?

Yes, that is correct. We modified the text accordingly.

- It is not clear to me whether helicity-aware or -ignorant reweighting is used in this section. As this is a focus of the paper, could it be mentioned? (No need to go into comparisons, pros and cons again)

In this section, the MG defaults are used. They correspond to helicity-aware reweighting at the LO and to helicity-ignorant at NLO. We added the following sentences for clarification:

“The event generator's defaults are used for the treatment of helicity in reweighting: At the LO, helicity-aware reweighting is used, while the NLO samples are reweighted with helicity-ignorant settings.”

- Fig 9:
- Would a ratio panel not be useful here? Was it omitted in order to be able to fit the figure into one page? If so, splitting the figure across two pages could be an option.

We added ratio pad to our plots, improved the binning, and the style.

- Having the SM distribution in addition to the sample1 and sample2 may be useful. It can serve as a validation for the “Reweighted to SM” distributions (as in principle, starting from either sample may be off). Moreover, it could help motivate the choice of a BSM starting point for the reweighting by showing that using the SM sample does not give as good agreement/statistical power when reweighting to Pt2. As an SM sample is already available, is this something that can be added?

The SM distributions are added to the plots.

Sec 5.4:
- The 5-flavour scheme is mentioned only here and was not mentioned earlier, e.g. in Sec 5.3. Is this worth mentioning? Also, a comment on whether 5FS vs 4FS has any effect on the expected accuracy of EFT reweighting (I assume there is none) may be helpful for readers.

The 5FS scheme was used in both, Sec. 5.3 and Sec. 5.4. It has no relevance for the reweighting. We have added this information to the text.

- It is mentioned that a comparison between SMEFTsim and SMEFT@NLO is made, using the LO implementation of both packages. However, no results are shown or described. I assume things looked consistent, but can this agreement be quantified? By giving a max deviation in obtained inclusive cross-sections in the text, for example, or a similar measure of consistency.

We change the text in the paper to avoid the misunderstanding of a quantitative comparison. Indeed, we only translate the coefficients. A comparison of inclusive cross-sections is available [here] by the author of the SMEFTSim paper. It provides at least some checks, although not in a polished format. We have added the following sentence:

“ For the SMEFTsim prediction, the conversion of Wilson coefficients provided in Ref. [28] is employed.”

- Fig 11: Is it worth mentioning which helicity configuration(s) are suppressed in the SM sample that lead to the disagreement at high pT when reweighting to cTZ = 1?

We have added the following sentence to the caption of Fig.12.

“The distribution shows that specific helicity configurations, for example, those with $h_{g1} = h_{g2}$, are suppressed in the SM and, therefore, helicity-aware reweighting cannot populate those regions of the phase space.”

Sec 5.6:
- Fig 16 left: Similar comment as on Fig 3, is it understood why both the reweighting and separate simulation don’t agree so well with the direct simulation at very high pT(H)? Given the error bars, this seems to be a real effect. A comment on how sensitive LHC experiments currently are to a potential bias of this size would also be helpful.
- Fig 17-20 (and also 16): If the disagreement in the reweighted distributions at very high pT(H) is due to fluctuations in the computation of weights, it would be good to quantify the corresponding uncertainty. But I find it odd that all the plots in this section have the same rather significant trend at highest pT(H). If it was a downward fluctuation in the SM sample, it would be very unfortunately pronounced, since this appears to be at least a 2sigma effect in every pT(H) plot. And from my understanding, the weights for different Wilson coefficients are calculated independently, so it is also hard to believe that all of them would fluctuate downwards by similar amounts. Could you add a brief summary of your investigations on this?

We have investigated the discrepancy and identified its source as an incorrect choice of scale settings in the simulation. As described in Sec. 4, the scale $H_T/2$ should have been used. The default scale choice had incorrectly been used. We have corrected the simulation and the closure is now within uncertainties.

Sec 5.7:
- Judging from the mathematical expression, it looks like options 1 and 4 are the same. What is the difference? And could the max sum index N be defined in the text?

The formula in Option 1 is not appropriate. The quantity N is the number of decay products. The list is updated accordingly.

- Is it correct to use T0=1 but T0^2==1, i.e. both "=" and "=="?
This syntax is taken from [JHEP 04 (2021) 073] and has been extensively validated in the course of this work.

- The purpose of showing p p > v v j j in Fig 22 is not clear to me. How does it connect to the point being made about scale choices?

The figure shows the impact of various parameters in the simulation namely, scale choices, PDFs and the variation in the top width for a key multiboson process (vector boson scattering). The point is that all these variations have a collective impact on the simulation and are therefore choices that should be made carefully when producing MC samples for EFT analyses.

Sec 5.8:
- In the text, the green histogram in Fig 23 is labeled the SM reweighted histogram, while the plot gives this label to the blue histogram. The green one has much smaller uncertainties, so I assume the label within the plot is correct?

The text correctly states that the green line is obtained from reweighting a SM sample that was combined with predictions from simulation with non-zero Wilson coefficients. We clarify the text in the following way.

"The green line in Fig. 23 shows an example where the reweighted prediction is obtained from a simulation that combines the SM tt sample with a sample with Wilson coefficient values of c_{tG} = 1 and c_{tG} = 3. This choice improves the population of tails in the simulation, resulting in the lowest statistical uncertainty."

Comments on formalities:
- Harmonisation between British and American English, e.g. "parametrization" (Section 1) vs "parameterised" (Section 5.3)
Fixed.
- arXiv links in references are broken
- "not implemented in [the] MG5aMC re-weighting tool" Fixed.
- Table 1 caption: "thee" fixed
- Table 2 caption: remove "multiboson" as these operators are all covered in Table 1? fixed
- Clarifying the differences in what plots represent in the plot itself would make things easier to parse. For example in Fig 3, the difference in left and right plots is not evident from the plots themselves.
We have added the relevant information to these figures.
- Fig 4: Should the angle \theta have a \hat? In the sketch there is one, but not in the text. Are these not the same quantity? These are the same quantity, and there should be a hat. Fixed.
- At the top of p19 it says "p_t" while otherwise it is denoted "p_T" Fixed.
- MadSpin in Sec 5.5 may want a reference? https://arxiv.org/pdf/1212.3460 Fixed.
- 5.6:

- As I did not quite understand the term "transverse mass of the 2->2 system" at first (and naively I would call VBF a 2->3 process), I looked into it and found Eq. (3.1) in https://arxiv.org/pdf/1507.00020 which states that all the final state particles are taken into account and the initial state particles are omitted. If my understanding is correct, could it be rephrased?
Fixed, but we kept the “2->2 system” so that it is clearly just a reproduction of the table in the Madgraph paper.
- 5.7:
- Starts with subsubsection 5.7.1, but there is no 5.7.2 Fixed.
- Several instances of a Figure being shown before being mentioned in the text, which can throw a reader off.
Figures now appear on the same or the following page where they are mentioned.
- Fig 22 caption: "choice of scale choice" Fixed.