Reply to Report 3 for 2308.00027v1

Title: "Returning CP-observables to the Frames They Belong"

**Authors: Jona Ackerschott, Rahool Kumar Barman,
Dorival Gonçalves, Theo Heimel, Tilman Plehn**

We would like to thank the referee for carefully reading our manuscript and providing detailed and valuable comments. The version of the manuscript that we resubmit addresses the aspects that the report brought to our consideration. Please find below the comments from the referee and our answers.

1. **Referee's comment**: *Page 3: It is not clear what Eqs. 1 and 2 mean, or what the variable r stands for. I assume the intended meaning of eq. 1 is that x_{reco} is sampled from $p(x_{reco}|x_{part})$, and operationally x_{reco} is computed using xpart and some standard-normally distributed random variables r. If this is the case, it might just be easier to explain this in words instead of equations.*

   Author response: We agree with the referee's suggestions and have accordingly updated the relevant discussion on Page 3 (first paragraph of Sec. 2.1).

2. **Referee's comment**: *Page 3: "guarantee a statistically correct and calibrated output" It is unclear what this means.*

   Author response: Statistically correct and calibrated output refers to a predicted posterior with a calibration curve [1] that matches the identity. Such a calibration curve measures the representation of the true posterior within the predicted posterior, averaged over the condition. For example a good calibration curve indicates that across all conditions we average over, 70% of truth samples lie in the 70% quantile of the predicted posterior for the respective condition. While this does not ensure agreement between the true and predicted posterior, a calibration curve that is equal to the identity is a necessary condition for the true posterior and thus a helpful sanity check.

   Now, we spell out the meaning of "statistically correct and well calibrated" in more detail on page 4, while adding a clear pointer to the reference for a detailed explanation: "[The loss] guarantees a reconstructed posterior whose quantiles are well calibrated to agree with the truth statistics, as demonstrated in Ref. [60]"

3. **Referee's comment**: *Page 6: "In our case, we will guarantee the correct descriptions of the intermediate top-mass peaks through the network architecture" It is unclear what this means. Is this about using MMD loss for the intermiate top-mass distributions?*

Author response: We do not employ MMD loss in our analysis. Instead, what we mean by the above statement is that the top quark's mass is integrated into the network architecture by explicitly choosing it as a degree of freedom in the event parametrization, as shown in Eq. (10) of the manuscript. This approach allows the network to learn it directly without requiring additional MMD terms in the loss function. We have updated the discussion in the first paragraph of Section 2.3 to clarify this.

4. **Referee's comment**: *A claim is made that iterative cINN in Ref. [62] can be used to reduce model dependence in the unfolding networks. I don't think this is true (explained below). However, one doesn't need a model independent unfolder for the purposes of this paper, so maybe this claim can be taken out.*

   On model independence: My understanding of Ref. [62] is as follows. Let's say we have a simulation model A for $p(x_{part})$ and another simulation model B for $p(x_{reco}|x_{part})$, and let's say model B could be wrong/mismodeled. Then assuming that model A is correct (which is fair to do in control regions only), iterative cINNs can learn an unfolder, which corrects for the mismodeling in B using the experimental data. In this paper, model dependence is used to mean dependence on model A. I don't think cINNs have been demonstrated to reduce dependence on A. In fact, an unfolder can be completely independent of model A only if the map from x_{part} to x_{reco} is invertible (i.e., for any given x_{reco} there exists only one possible x_{part}). This is typically not true in collider physics. Also, even if we're considering dependence on model B, I'd say the iterative technique corrects for simulation-model errors, and doesn't induce model independence, although that's arguably just semantics.

   Author response: As stated in Ref. [2], all unfolding methods require a trustable detector simulation, i.e. they cannot correct for mismodeling in model B. Contemporary unfolding methods, such as Omnifold [3], cINN [1] and iterative cINN unfolding [4], are Bayesian methods, thus rely on the accuracy of the likelihood $p(x_{\text{part}}|x_{\text{reco}})$, i.e. of the result of the detector simulation ("model B"). We would like to emphasize here that what can be corrected for in a Bayesian method, however, is prior dependence. This works by iteratively re-applying Bayes theorem on some initial, potentially inaccurate, prior [5]. Both iterative cINN unfolding and Omnifold use this idea to reduce the dependence on the SM prior usually assumed for unfolding. The plain cINN unfolding used in our paper applies Bayes theorem only once, which leads to a larger prior dependence that could be further reduced by the iterative method introduced in Ref. [4]. We agree with the referee that iterative cINN unfolding cannot render the unfolding results fully model-independent for a typical non-invertible detector simulation. Our intention behind the statement was only to emphasize that an iterative method might be able to reduce model dependence to an extent but not eliminate it.

5. **Referee's comment**: *Eq. 3 shows the loss function for mapping to a Gaussian latent space, instead of not a generic latent space. However, this work uses uniform latent spaces as well. The loss function for that case could be provided.*

   Author response: The referee's comment is well-taken, and we appreciate your bringing this to our attention. We have updated Eq. (3) and the associated discussion on Page 4 (Para 1) to describe the loss function for a generic latent space.

6. **Referee's comment**: *In Eq. 5, does k take on different values for different x? If so, it is not obvious from the equation or the text.*

   Author response: We thank the referee for bringing this to our attention. Indeed, k takes on different values for different x. We have replaced $k \to k_x$ in the updated manuscript to clarify it.

7. **Referee's comment**: *Section 3.2 is confusing in a number of ways. Given that the approach is to only use the unfolding network as an analysis variable, the first part of section 3.2 feels unnecessary. In Eq. 26, $p(\alpha|x_{reco})$ and $p(\alpha|x_{part})$ only make sense if there is a prior on α. $p(x_{part}|x_{reco}, B)$ is not defined, since x_{part} (in the chosen phase space parameterization) doesn't exist for background events. This renders Eqs. 27, 28 and "$p(x_{part}|x_{reco}, B) = p(x_{part})$" meaningless in my opinion.*

   Author response: First, similarly to unfolding, we do Bayesian analysis here, so we assume an implicit prior on α, from which we want to infer the posterior $p(\alpha|x_{\text{reco}})$. So the second point is true, but as far as we can see it does not raise a problem for our argument.

   Second, our argument considers $p(x_{\text{part}}|x_{\text{reco}}, B)$ to be defined as the probability density of any parton-level signal event x_{part} occuring under the condition that a background event x_{reco} was measured on reco-level. Note that, since the detector-level measurement of a background event gives us no information about the probability of a signal event on parton-level, we trivially have $p(x_{\text{part}}|x_{\text{reco}}, B) = p(x_{\text{part}})$, so we only write down $p(x_{\text{part}}|x_{\text{reco}}, B)$ for clarity. In this sense it is unclear what the referee means with "x_{part} doesn't exist for background events", since background events on parton-level are not part of our argument.

   To address the referee's concerns, we clarified our argument concerning $p(x_{\text{part}}|x_{\text{reco}}, B)$ in the text below Eq. 30 of the new draft: "Let us consider $p(x_{\text{part}}|x_{\text{reco}}, B)$ for a moment. This is defined as the probability density of any parton-level signal event x_{part} occuring under the condition that a background event x_{reco} was measured on reco-level. However, the detector-level measurement of a background event gives us no information about the probability of a signal event on parton-level. For this reason, we can drop x_{reco} and write $p(x_{\text{part}}|x_{\text{reco}}, B) = p(x_{\text{part}})$, where $p(x_{\text{part}})$ is only constrained through prior knowledge. This includes our model assumptions as well as phase-space constraints due to a finite center-of-mass energy."

We also made the prior on α explicit in the second paragraph of Sec. 3.2 of the new draft: "Since we do not know if a particular reco-level event x_{reco} is signal or background, we only care about the full probability $p(\alpha|x_{\text{reco}})$ of our model parameter, given some reco-level event x_{reco} which is either signal or background and some prior $p(\alpha)$."

8. **Referee's comment**: *It is unclear how a variable number of jets is handled in section 3.3. The details weren't immediately obvious to me even after looking through Ref. [60].*

   Author response: A variable number of jets is handled by zero-padding the missing jets, so that the networks input vector maintains a fixed length. In contrast to Ref. [1] (Ref. [60] in the paper), we do not use the number of jets as an additional dimension of the input vector. We clarify this in the second paragraph of Sec. 3.3 of the revised manuscript: "In the latter case, we ensure that the input vector to the network has a fixed length by zero-padding the missing jets. We note that our approach to tackle a variable number of jets differs from that in Ref. [60] where the numbers of jets is incorporated as an observable in the training dataset."

9. **Referee's comment**: *The input and output shapes of the various neural networks could be provided.*

   Author response:

   We agree that this could serve as a sanity check when reproducing the paper, in the sense that one can check if the shapes of the network input match the provided shapes. However, the shapes can be easily inferred by the number of final states in the provided parameterization as well as the number of reco-level objects, which we provide. Hence, we do not think that this would provide much benefit.

10. **Referee's comment**: *MMD is discussed only briefly in page 6. It is unclear whether (and for which distributions) MMD is used in the paper. "Its main advantage is that it only affects the target distribution and avoids an unnecessarily large model dependence." The meaning of this statement is unclear.*

    Author response: As previously stated in our response to Comment (c), we do not employ MMD loss in our analysis. Rather, the important sharp kinematic features, such as the invariant mass of the top quarks, are explicitly included in the event parametrization (see Eq. (10)). We have also updated the discussion in Section 2.3 to clarify this: 'For full high dimensional unfolding, as considered here, the simplest way of encoding LHC events at the parton-level is through the components of the final-state 4-momenta. However, it was observed that intermediate particle mass-peaks are poorly reconstructed in this parameterization. A problem that was already encountered in Ref. [81]. One way to improve the reconstruction quality of these peaks is to add a maximum mean discrepancy (MMD) between a given set of generated and truth distributions in the loss function [81,82]. As the name suggests, the MMD is a measure for the discrepancy between two distributions. It is reasonably efficient to compute from samples only, although it admits quadratic scaling with batch size, making it a potentially useful tool for generative network training. The disadvantage is that the additional loss term complicates the training and consequently limits the precision of the network. For our INN architecture, the computation of an MMD loss requires samples generated from the latent distribution, making the training twice as computationally expensive, while the usual INN loss works on latent-space samples. The MMD additionally adds a sizable amount of hyperparameters to the training, making it significantly more difficult to optimize. In this paper we use, instead of the MMD, a different phase-space parameterization for improving the reconstruction quality of intermediate particle mass-peaks.'

11. **Referee's comment**: *In figure 6 top 2 rows, it is confusing to have three truth distributions. The non-SM truth curves could be labelled differently in the legend. Also, in those plots, it cINN/truth should probably be computed with the SM truth for all three networks, instead of using different truth curves for each network.*

    Author response: We agree with the referee's comments and have updated the plots accordingly.

12. **Referee's comment**: *A description of how the cINN histograms are created could be provided. For instance, to get the central values of the bin counts in a histogram, is only one parton level event computed for each reco-level event? Or is the central bin-count value computed as an average over many samplings?*

    Author response: The histograms are created by unfolding each reco-level event only once and then compiling them from the resulting samples. We mentioned this in the first paragraph of Sec. 3.3 of the new draft: "Note that, here and in the following, we sample only one parton-level prediction of the cINN for each sampled reco-event in our plots." Note that we would expect the specific method, at least in the infinite statistics limit, to not impact the resulting histograms. If we assume that our predictions are continuous with respect to the reco-level input of the cINN, we would expect a sufficiently small neighborhood of x_{reco} to yield sufficiently similar results as just sampling multiple times from x_reco.

Some minor suggestions/comments from referee 3:

1. **Referee's comment**: *Page 14: "The conventional approach to complex kinematic correlations..." Is this supposed to be "compute kinematic correlations"?*

   Author response: Indeed, we have corrected the typo here.

2. **Referee's comment**: *Page 1: Maybe change "The, arguably, most interesting symmetry in the SM is CP" to "Arguably, the most interesting symmetry in the SM is CP"?*

   Author response: We have corrected the above statement.

3. **Referee's comment**: *Page 1: "CP ... potentially realized in an extended Higgs sector". It is not obvious that the authors are referring to CP **violation** here.*

   Author response: We have modified the statement to make this aspect clearer.

We have fully addressed all the comments and minor suggestions. We hope that with these clarifications and associated changes made to the manuscript, the paper can be accepted for publication in SciPost.

References

[1] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe, *Invertible Networks or Partons to Detector and Back Again*, SciPost Phys. **9** (2020) 074, arXiv:2006.06685 [hep-ph].

[2] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler, *OmniFold: A Method to Simultaneously Unfold All Observables*, Phys. Rev. Lett. **124** (2020) 18, 182001, arXiv:1911.09107 [hep-ph].

[3] A. Andreassen and B. Nachman, *Neural Networks for Full Phase-space Reweighting and Parameter Tuning*, Phys. Rev. D **101** (2020) 9, 091901, arXiv:1907.08209 [hep-ph].

[4] M. Backes, A. Butter, M. Dunford, and B. Malaescu, *An unfolding method based on conditional Invertible Neural Networks (cINN) using iterative training*, arXiv:2212.08674 [hep-ph].

[5] G. D'Agostini, *A multidimensional unfolding method based on bayes' theorem*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **362** (1995) 2, 487.