Major comments

COMMENT 1. I suggest to improve the clarity concerning the use of physics information in HEP classifiers in the abstract, introduction and conclusions to better highlight the novelty of the proposed with respect to well-established and other studied approaches. Here are a few concrete suggestions but I also invite the authors to read through the relevant paragraphs again and improve them where they see fit:

Answer: As suggested, we have revised the summaries and made changes to the introduction as described below.

- lines 43-45: "These methods are mainly used in the context of the classification of jet data. On the other hand, their application to event-level data has not yet been explored to the same degree, and BDTs are still the most commonly used method." - It is true that there are analyses that use BDTs as event level classifier, but there are numerous examples where deep learning architectures are used for this purpose. I suggest to rephrase this sentence and provide references to examples of HEP publications that use neural networks. In this context, I also suggest to discuss in the introduction (and in Section 3.1) that the use of high-level features is a standard technique in BDTs in HEP. In the same spirit for line 269: As it is common practice to use designed features as the input to BDTs, please provide citations to example publications of HEP Collaborations, such as ATLAS, CMS, ALICE and/or LHCb, instead of or in addition to Ref. [32].

Answer: We appreciate the feedback and agree that deep learning architectures have become increasingly common for event-level classification in HEP, alongside traditional BDTs. We have revised the sentence to reflect this broader usage and included citations to recent HEP publications that use neural networks for event-level classification. Additionally, we have expanded the discussion in the introduction and Section 3.1 to emphasize that the use of high-level features is a standard practice in BDTs for HEP, and provided appropriate citations to relevant publications from LHC experiments to support this point.

- lines 46-53: The use of physical information in the context of deep learning has been discussed in several papers before. Actually, such papers are cited in lines 268-272. In my opinion, the discussion of this previous work should be moved to a central part in the introduction of the paper to better lay the ground for the presented work.

Answer: Thank you for the suggestion. We have added a paragraph in the introduction section.

- lines 465-468: I suggest that the authors consider rephrasing the last sentence of the conclusions to highlight the potential gain of their novel SM interaction matrices as opposed to a general statement about the inclusion of physics information in networks, which had already been discussed in previous literature (such as Refs. [33-36]).

Answer: Thank you for your comment. We have added a paragraph in the conclusions section.

COMMENT 2) line 138: Isn't it a strong limitation to remove information about possible additional same-charge, same-flavor leptons from the BDT training, in particular for same-sign, same-flavor, 3-lepton and 4-lepton events? Same comment for the FCN (lines 158-159).

Answer: We appreciate the referee's observation regarding the exclusion of additional same-charge, same-flavor leptons from the BDT and FCN training. However, our analysis intentionally follows the same cuts and selection criteria as those employed by the ATLAS collaboration in the 4-top search analysis. These cuts, which define a signal region with at least six jets (including two b-tagged), $H_T >$ 500 GeV, and specific lepton configurations (two same-sign leptons or at least three leptons per event), are optimized to maximize the signal-to-background ratio.

COMMENT 3) Table 5: I am surprised that the focal loss ParticleTransformer performs very well in terms of significance. How is this related to the comparatively bad discrimination of this classifier in Figure 3, where it seems to separate the background and the signal much worse than for example the BDTs?

Answer: Thank you for your observation regarding the Particle Transformer with focal loss. While Figure 3 may suggest less clear separation between signal and background, this does not fully reflect the model's performance. As shown in Figure 5 and Table 4, the ROC curve and AUC metrics confirm that $\mathrm{ParT}_{\text{int. SM (FL)}}$ performs competitively. The higher significance values in Table 6 (in the new version) highlight the model's ability to focus on critical regions of the decision space, which is likely due to the focal loss emphasizing harder-to-classify events.

COMMENT 4) lines 418-420: I do not understand this part of the discussion, as the significances in the case of 20% systematic uncertainty on the background are compared, but the corresponding increase in statistics seems to be calculated neglecting systematic uncertainties, if I am not mistaken. Please clarify. Same comment for lines 447-449 for tth and lines 463-464 in the conclusions.

Answer: The referee is correct in noting that the increase in statistics has been calculated based on statistical uncertainties only, without accounting for systematic uncertainties. The reason for this is that including systematic uncertainties in the significance calculation introduces complexities that make it challenging to directly estimate the required increase in statistics. However, we believe this estimation still serves as a useful indication of potential gains.

COMMENT 5) Conclusions: "10% of the improvement directly attributable to the SM interaction matrix." - If I compare all background rejections of "int." and "int. SM" in Table 6, I do not see in any of the numbers an improvement near 10%. The same is true for the background rejections in Table 4. Or do you mean that it is 10% of the overall 10-40% improvement, i.e. only 1-4%? If this is the case, please rephrase the conclusions and the abstract for more clarity.

Answer: The number 10 percent comes from the results of table 4 and 5. Our baseline was the signal efficincy of 70 percent. Overall (Table 5) the background is reduced e.g. for the 4-top case from 0.18 to 0.17 for overall (i.e. 6 percent absolute) and e.g. from 0.13 to 0.119 for the Z+jets sample (i.e. 9 percent absolute (Table 4). We changed the conclusion to "...approximately up to 10% of this improvement..".

Minor comments

1) lines 113-114: "The dataset includes 302 072 events, half of which correspond to the four tops signal and half of which are background processes." - If I multiply the numbers from Table 1, I get for 4 tops: 32,463,742 * 0.007 = 230k, which is much more than half of the 300k events in the dataset. Please clarify.

Answer: We did not use the entire generated dataset. Our goal was to create a balanced dataset between signal and background events. Since the ttbar + Higgs process has a very low acceptance rate (making it challenging to generate), we produced approximately 30,000 events for this process. For the other background processes, we selected around 40,000 events each, resulting in a total of approximately 150,000 background events. This was matched with an equal number of 150,000 signal events to maintain balance in the dataset. This clarification has been added to the text.

2) line 115: "All background processes have an equal number of events." - Please clarify this statement, as it seems to be not consistent with the numbers in Table 1.

Answer: The statement refers to our approach for creating a balanced dataset between the signal and background processes. To achieve this, we randomly selected events from each background process to ensure an equal number of events across all backgrounds. While this may not reflect the original event yields listed in Table 1, it was necessary for balanced training.

3) line 153: I suggest to cite the Dropout paper.

Answer: We have added the reference to the Dropout paper.

4) line 154: I suggest to provide the values of the default parameters.

Answer: We appreciate the referee's suggestion. We have included the values of the default parameters for the Adam optimizer in the revised manuscript. These are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-7}$. Additionally, we have clarified that the learning rate was treated as a hyperparameter and optimized separately.

5) line 161: In my view, Ref. [21] is not appropriate as a general reference for the idea of convolutional neural networks. I suggest to cite original work on CNNs here.

Answer: We have replaced the reference with a citation to the original CNN work by Yann LeCun et al.

6) line 169: Please clarify why the maximum number of particles per event is 18, as each top decays to three particles, resulting in a maximum of 12 particles.

Answer: The maximum number of particles per event is 18 because we generate extra jets in our MC event generation, as outlined in Table 1.

7) Table 2: Please explain the meaning of $\gamma_{\mathbf{tag}}$, as it was not introduced before.

Answer: The variable γ_{tag}, refers to the identification tag assigned to photons in the event. This has been clarified in the text.

8) line 184: I would expected that also the particle type is part of the node features. Please clarify.

Answer: The particle types are indeed part of the node features. This has been clarified in Section 3.4 of the text.

9) line 192: Please provide more details about the "attention-weighted procedure". How do you define the attention heads?

Answer: Thank you for your comment. We have now included additional details about the attention-weighted procedure and how attention heads are defined in the context of ParticleNet. Specifically, the attention mechanism replaces the simple mean aggregation with a weighted sum, where the weights are learned through trainable linear transformations applied to the node features. This allows the model to assign different levels of importance to messages from neighboring particles, based on their relevance.

10) line 247: Please check the grammar in "i.e. pt were is large" at the end of the bullet.

Answer: The sentence "i.e. pt were is large" was a typo. We have corrected the text.

11) line 327: Do you reduce nf to values smaller than 6 if the scale is below the top quark mass? Please clarify.

Answer: In our model, we choose to keep $n_f = 6$ constant throughout the calculations, even for energy scales below the top quark mass. This simplification was made to avoid additional complexity in the implementation of the model and to maintain consistency across different energy scales. The effect of varying n_f would primarily be significant at very low energy scales, where heavy quarks decouple. Since our analysis focuses on energy ranges where the influence of this variation is relatively minor, we chose this approach. However, we acknowledge that varying n_f based on the energy scale is a more precise treatment, and we may consider implementing this refinement in future work to improve accuracy at lower energy scales.

12) Figure 1: I suggest to add the focal loss ParticleTransformer and the SetTransformer curves to this figure, as you include these in other relevant parts of the results section, such as Table 4, Figure 3 and Table 5.

Answer: We appreciate the suggestion to add the focal loss Particle Transformer and Set Transformer curves to Figure 1. However, retraining these models and incorporating their results at this stage would require significant computational resources and time. We believe that the inclusion of these models in other key parts of the results section, such as Table 4, Figure 3, and Table 5, already provides a thorough comparison of their performance. These tables and figures offer a detailed and representative evaluation of the models' effectiveness, ensuring that the reader can fully understand their impact without further additions to Figure 1.

13) Figure 1: Typo in "FNC" $\rightarrow$ "FCN". I also suggest to remove the label "no pair int." from the legend for consistency with Table 3.

Answer: The typo "FNC" has been corrected to "FCN". We also removed the label"no pair int" from the legend to ensure consistency with Table 3. Thank you for pointing it out.

14) Figure 2: It is curious that the "int. SM" models clearly outperform the benchmark ParticleNet for tth, ttWW and ttZ but for ttW the improved is not that pronounced, in particular not for lower signal efficiencies. It would be instructive to discuss

the origin of this behavior in the body of the text. Is this for example connected to the share of ttW in the total background or to the similarity of ttW to the 4 top signal?

Answer: The referee correctly observes that the 'int. SM' models outperform the baseline ParticleNet (PN) for processes like $t\bar{t}H$, $t\bar{t}WW$, and $t\bar{t}Z$, but show a less pronounced improvement for $t\bar{t}W$, particularly at lower signal efficiencies. This behavior can be attributed to the kinematic similarities between the 4-top signal and the $t\bar{t}W$ background, particularly in terms of jet and lepton multiplicities. Both processes produce high jet multiplicities and can have multiple isolated leptons in the final state, making it challenging for the models to distinguish between the two. This is especially true at lower signal efficiencies, where models tend to prioritize the rejection of more distinguishable backgrounds. Although $t\bar{t}WW$ is also kinematically similar to the 4-top signal, it involves a more complex final state, with additional jets and W bosons. This increased complexity provides more features for the models to exploit, allowing the 'int. SM' models to outperform the baseline more significantly for $t\bar{t}WW$ than for $t\bar{t}W$.

15) Figure 3: What is the "total cross section"? Is it sum the of the SM signal and background cross sections? Please clarify.

Answer: Thank you for pointing this out. In Figure 3, the "total cross-section" (σ_{tot}) is defined as the sum of the cross-sections for the SM signal and all background processes considered in this study. We have clarified this in the figure caption to avoid any confusion.

16) Figure 3: Typo in "FNC" $\rightarrow$ "FCN".

Answer: The typo "FNC" has been corrected to "FCN". Thank you for pointing it out.

17) line 431: Should "ongoing couplings constants" rather be "running coupling constants"?

Answer: The referee is correct. We have revised the text to use "running coupling constants" instead of "ongoing coupling constants".

18) I suggest to reference Table 8 somewhere in text of the Appendix.

Answer: We have added a corresponding reference in the text of the Appendix.