QUESTION 1. Please make sure that relevant works on ML-tagging at the LHC are cited. References is what gives our young people jobs;

Answer: Please note that our paper is on event classification. We agree with your suggestion and have added several more citations relevant to event classification and the use of graph and transformer models, including applications in jet tagging. The references have been updated in the text accordingly.

QUESTION 2. Please explain the attention layer a little more in detail, especially in view of the set transformer mentioned later. We should not assume that all readers know, for instance, the differences between a graph network and a transformer;

Answer: Thank you for pointing this out. We have added detailed explanations to the text, particularly in Sections 3.5 and 3.6.

QUESTION 3. Eq.(4) is a very sloppy version of a formula. It would be nice to first write the BCE loss and then the focal loss, including a definition of all symbols in a way that a student can just implement it;

Answer: We have revised Section 3.7 to provide a more precise formulation of the Binary Cross-Entropy (BCE) loss and the focal loss. We have included definitions for all symbols to ensure clarity.

QUESTION 4. In Sec.3 I am missing a discussion of network size and training data efficiency. I would expect the methods to be very different here, will come back to this for Fig.1;

Answer: To address this, we have expanded Appendix B to include a detailed explanation of the networks' optimization process for the BDT, CNN, and FCN. The hyperparameter scans for ParticleNet are discussed in the last paragraph of Section 3.4, and for the Particle Transformer in the last paragraph of Section 3.5.

QUESTION 5. In 4.1 the natural question arises - what is physics information and what is just a covariant representation under a known symmetry. Please comment and separate this carefully;

Answer: We agree and added that in section 4.1: "In our analysis, we distinguish between **covariant representations** and **physics information**. Covariant representations consist of features that respect the fundamental symmetries of the physical system, such as Lorentz invariance, ensuring that quantities like invariant mass remain unchanged under Lorentz transformations. In contrast, physics information encompasses additional insights into particle interactions, including coupling constants and interaction strengths derived from the Standard Model. By incorporating a Standard Model Interaction Matrix, we embed detailed physics governing particle interactions into our model. "

QUESTION 6. In Fig.1, please find a way to label the curves such that the reader does not have to spend significant time going back and forth between curves and labels. Line styles combined with colors might work.

Answer: Thank you for your suggestion regarding Fig. 1. We have updated the plot to include clearer labels for the curves.

QUESTION 7. Content-wise, I have a hard time understanding Fig.1. Why are PN and ParT without physics information so bad, comparable to the BDT? Or is the

BDT good because the problem is simple. And if that is the case, why does the physics information help? I am confused, some of the results are really counter-intuitive...

Answer: The situation may appear confusing, possibly due to our choice of colors (we hope the new mapping will clarify this). However, when only models with no or paired features are compared, the situation is as expected.

As also seen in other papers, fully connected networks (FCNs) and CNNs are not the best models for event classification with 4-vectors. The BDT, while slightly better (ranking third worst), is still a relatively poor model. PN and ParT outperform the BDT (if they all have no pairwise features). Therefore, the BDT does not outperform PN or ParT if no pairwise features are included.

QUESTION 8. Also in Fig.1, is the increase with more training data really the same for the graph and the transformer? Is that not against our general expectations, and also against the experience with the pretrained ParT?

Answer: Yes, both the graph and transformer models appear to improve similarly with more training data. Generally, with more data, the transformers typically tend to improve on data-rich problems with more and more training data (even compared to Graph-NN). However, this effect may not be evident in our case due to either a) we do not have enough training data, or b) the dataset is not rich enough in features. In any case, our results highlight the importance of including physical interactions and pairwise features. We have added a sentence to the paper to clarify this point.

QUESTION 9. I do not understand the comment at the very end of 5.2, that the PN is similar to the ParT;

Answer: In ParticleNet, the graph is constructed by connecting each particle to its k-nearest neighbors. When k is increased to n, each particle becomes connected to all other particles in the event, resulting in a fully connected graph. This means that during the graph convolution operations, information from every particle can be directly aggregated and passed to every other particle. Similarly, the ParT architecture employs a self-attention mechanism, where each particle attends to all other particles in the event. This global interaction allows the model to capture relationships between any pair of particles. Therefore, when $k = n$ in ParticleNet, both PN and ParT enable global interactions among all particles in the event. We have revised the manuscript to clarify this point.

QUESTION 10. Finally, the obvious question is if this improvement can be translated to the high-performance results for jet tagging. And since the highest-performing taggers to date are covariant, what do the authors expect to happen for those?

Answer: The improvements observed through the integration of physics information via the Standard Model Interaction Matrix suggest similar gains could apply to jet tagging.

Current high-performance taggers (e.g., ParticleNet, Particle Transformer) are covariant under symmetries like Lorentz invariance, ensuring outputs align with the underlying physics. Adding physics-specific information, such as particle-type correlations or interaction strengths, would complement these models, refining their decision-making process beyond symmetry properties alone.

Additionally, most jet taggers lack explicit information on how to relate different particle types, such as electrons and photons. Feynman rule-based features, like the ones we've introduced, could improve performance. Even in highly optimized models, this added information could significantly enhance background rejection, especially for rare signals.

We have addressed this in the Conclusion section.