1) L60: Maybe I'm missing the reasoning but why background samples such as tt+ jets are not included in this study? I would expect the contribution to 4tops (and later to ttH) to be important (such as ttbb for example).

Answer: The $t\bar{t}$+jets backgrounds, including $t\bar{t}b\bar{b}$, are significantly suppressed in the selected signal regions (SSML: same-sign dilepton or trilepton), where irreducible backgrounds such as $t\bar{t}W$ and $t\bar{t}Z$ dominate. In the ATLAS 4-top analysis (Eur. Phys. J. C 83:496, 2023), the contribution from $t\bar{t}b\bar{b}$ is considered negligible in these regions but is still included as part of the reducible background. Its impact is modeled with systematic uncertainties of up to 50% for events with three or more b-jets. Additionally, since $t\bar{t}$+jets mainly affects the analysis through misidentified or non-prompt leptons and charge misassignment, which are treated using data-driven methods, its direct contribution is expected to be minimal.

2) L80: How do you define b-jets in this context? Based on reco-level tagging or based on gen objects matched to reco jets?

Answer: We used the ATLAS Delphes default card, which is based on truth-level (gen-level) matching to partons, as discussed in ATL-PHYS-PUB-2015-022. We have added this to the text.

3) L136: BDTs compared to NNs are knows for their resilience against irrelevant features (also investigated in the context of anomaly detection in arXiv:2309.13111 and arXiv:2310.13057), so it seems surprising that a zero-padding strategy would result in degradation. Do you have an idea why that would be the case?

Answer: We have checked this, and in our case, zero-padding noticeably decreased performance. We believe the reason is that zero padding increases the feature dimensionality of the data without adding information. The BDT then has to find the classification boundaries in a much larger feature space. In addition, the BDT could (incorrectly) use zeros as information for the split.

4) Additionally, both BDTs, MLPs, and CNNs would require some specific ordering of the inputs. How is that defined? Based on the pT of the objects? Sorry if I missed in the text.

Answer: Yes, the information is p_T ordered. We have added this to the caption of Table 2.

5) L200: In the case of ParticleNet, what is the number of k neighbors used? In the space of jets and leptons, I imagine the number of constituents to be much smaller compared to previous studies using particles clustered to jets (which can go above 100 objects per jet). If the original number k=16 from ParticleNet is used, I imagine the graph is almost fully connected, is that correct?

Answer: Yes, this is correct. The performance improves with increasing k, eventually saturating when k approaches the number of particles in the event, effectively making the graph fully connected. We have a plot on the performance as function of the number of neighbors; please see Fig. 4.

6) Eq5: Have you tried including the embedding U as a concatenation rather than an addition? That would increase the complexity of the model, but could also lead to better use of these features.

Answer: Yes, concatenating U with $Q * K$ increases the input dimensionality to the attention mechanism and increases model complexity. We did try this approach, but the performance did not

improve, and the training slowed down due to the increased complexity.

7) Eq6: The result matrix multiplication QK^T is a NxN matrix (N is the number of objects), however I would expect U to be of shape NxNxF with F being the number of pairwise features calculated for each of the pairs. How is that summed with the QK^T term?

Answer: The referee's observation is correct: the matrix U is initially computed with a dimensionality of $N \times N \times F$, where F represents the number of pairwise features. To match the dimensionality of the QK^T term (which is $N \times N$), the U matrix undergoes dimensionality reduction through a 1D convolution applied along the F-axis. This operation ensures that U and QK^T have compatible shapes for the subsequent operations.

8) L280: How is a 3-body invariant mass calculated from a pair ij?

Answer: We addressed this by calculating the two-body invariant mass m_{ij} and extended it to include three-body systems. Specifically, we added the masses of the pair ij combined with the hardest and second hardest particles in the event (based on their p_T), which we denote as $m_{ij,1}$ and $m_{ij,2}$, respectively. However, after testing these additional features (including the hardest, second hardest, and more), we found that they did not significantly improve performance in ParticleNet, Transformer, or BDT models.

9) Sec. 4.2: Some of the interaction terms introduced require the knowledge of a jet being a b or a light jet. If there is a mistag and a b-jet is wrongly identified (or a light jet is misidentified), does that affect the performance?

Answer: We do not expect mistagging (incorrect identification of b-jets or light jets) to affect performance differently than it does in other algorithms or cases of mistags.

10) L283: Isn't ΔR_{ij} already somewhat included in the edge information of particlenet in Eq.5? the $x_i - x_j$ term, besides the norm, should also carry the same information. Did the inclusion of ΔR_{ij} bring any benefit to the particlenet implementation?

Answer: While the term $x_j - x_i$ in Equation 4.1 inherently captures relative differences in angular and kinematic space, explicitly including ΔR_{ij} as a separate feature allows the model to leverage this information directly. This explicit inclusion was found to improve the performance slightly for models like ParticleNet, as shown in Table 4, consistent with prior observations in jet physics.

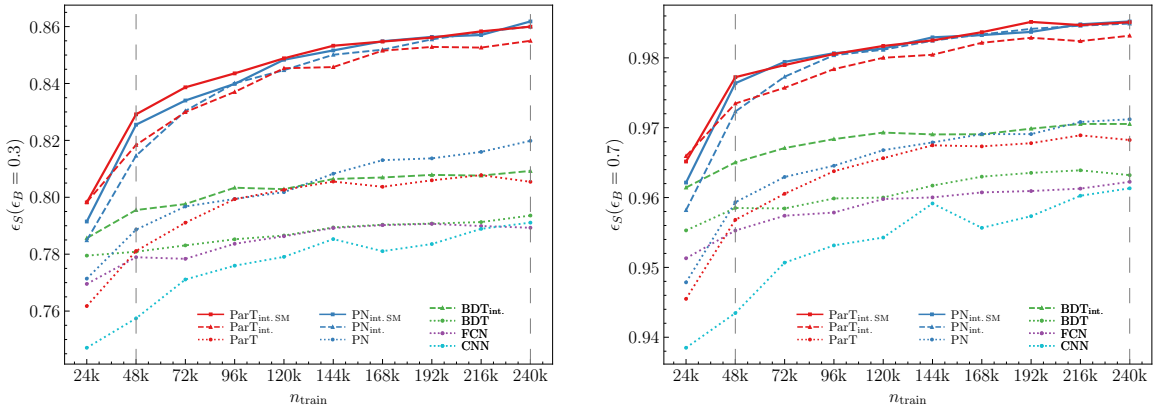
11) L338: The large negative number is understood in the PartT implementation using attention, but is that also used for the ParticleNet application? There there was no softmax operation applied, unless that was introduced as part of the interaction matrix formulation in this work?

Answer: This was not used in the ParticleNet implementation. We have updated the text to clarify this.

12) Fig1: This figure is really interesting and I partially agree with the authors conclusions. The AUC is often too simplistic to provide the complete picture and even if the AUC does not improve dramatically, the signal efficiency for fixed background efficiency can still change by a good margin. Would be great to have a similar plot but

instead of the AUC using the signal efficiency at fixed background efficiency (say 30% or 70%), which is also closer to how these classifiers would be used in a realistic analysis.

Answer: We have added both plots showing the signal efficiency (ϵ_S) as a function of training data size for each algorithm, at the fixed background efficiencies (ϵ_B) of 30% and 70%. Please refer to Appendix A.1: Signal efficiency at fixed Background efficiencies.



(a) Signal efficiency (ϵ_S) at a fixed background efficiency (ϵ_B) of 30%

(b) Signal efficiency (ϵ_S) at a fixed background efficiency (ϵ_B) of 70%

Figure 1: Signal efficiency as a function of training data size for each algorithm, evaluated at two fixed background efficiencies: (a) 30% and (b) 70%. These plots allow for comparison between different models' abilities to detect signal as the amount of training data increases

13) Fig2: The background rejection improvement with respect to a baseline model is a great distribution to show, however I do not understand why the 48k data sample is chosen to be the baseline comparison. Is there some specific restriction in a real analysis that would only allow around 50k simulated samples to be available for training? Given the focus of the paper on the improvements obtained by adding SM inspired quantities, I would use as a baseline the particlenet results with full data but without any interaction terms. That would also help the reader navigate the results to see the improvements brought by the different choices of interaction terms added.

Answer: The baseline model is indeed the ParticleNet results with full data. However, the graph was created for our statistical base scenario, which consists of 50k events. The reason we chose 50k as the base is: a) that this roughly corresponds to the number of available background events after all cuts in ATLAS, making it a realistic scenario, and b) that we initially evaluated all algorithms with this dataset and then produced the training data in a second attempt to see how the models scale with larger training sets.

14) L383: Similar to the previous comment, the numbers described as improvements wrt the baseline do not seem very meaningful unless there is a strong reason why a dataset with only 50k samples is needed as the baseline.

Answer: We have added the same illustration in the appendix for the full data scenario, but as mentioned earlier, the 50k scenario is more realistic. Additionally, with infinite training data, all

models are likely to perform the same according to the universal approximation theorems.

15) Tab: 5: I appreciate the comparison of the predicted significances and how they are modified in the presence of a systematic uncertainty affecting the background processes. In this context, using the signal efficiency value that maximizes the SIC curve (sig. eff/ sqrt(bkg. eff) vs sig. eff.) would perhaps be a better/more realistic choice. How the expected significance changes if instead of fixing the signal efficiency one used the maximum SIC point? That should also give you the best expected significance for each classifier, which is always good to know.

Answer: This is a different (maybe more complicated) but interesting metric that we have not analysed. However, we believe that such a comparison would not change the overall conclusions of the table. In fact, we examined two other signal yields and arrived at similar conclusions. Optimizing the significances also depends on the luminosity, which is beyond the scope of this paper. We only wanted to illustrate how small effects in the AUC can nevertheless lead to improvements in significance.

16) Fig.4 Again, these results would be more interesting when evaluated over the full dataset instead of only part of the data.

Answer: As mentioned previously, we have added this plot to the appendix for the full dataset scenario.