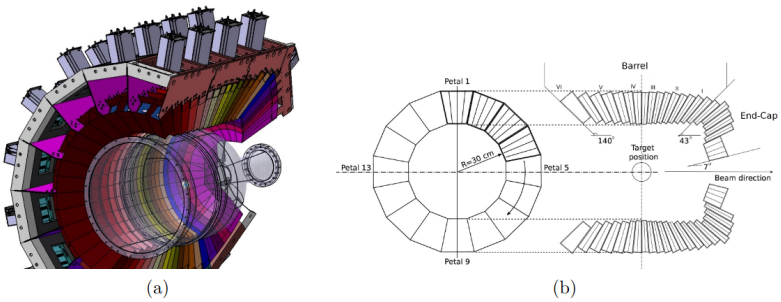
Dear reviewer,
Thank you for your thorough review of our manuscript. The detailed comments were useful and have allowed us to clarify our methodology to enhance our discussion of the results. We have prepared a point-by-point response to each comment below and have revised the manuscript to reflect these changes.

Note: blue font means that this is text and figures which were replaced/added to the paper

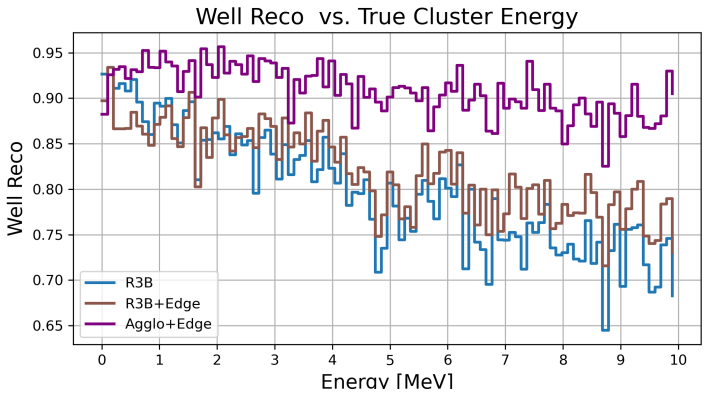
Q1:In general I think the paper would profit from a few more figures, especially
- a drawing of the CALIFA layout
- a more direct comparison of the reconstructed (clustered) energy to the true particle energy, and the dependence on the distance between particles (I think there must be such a dependence, but it is not really discussed in the paper)

This picture with description was added to the paper:



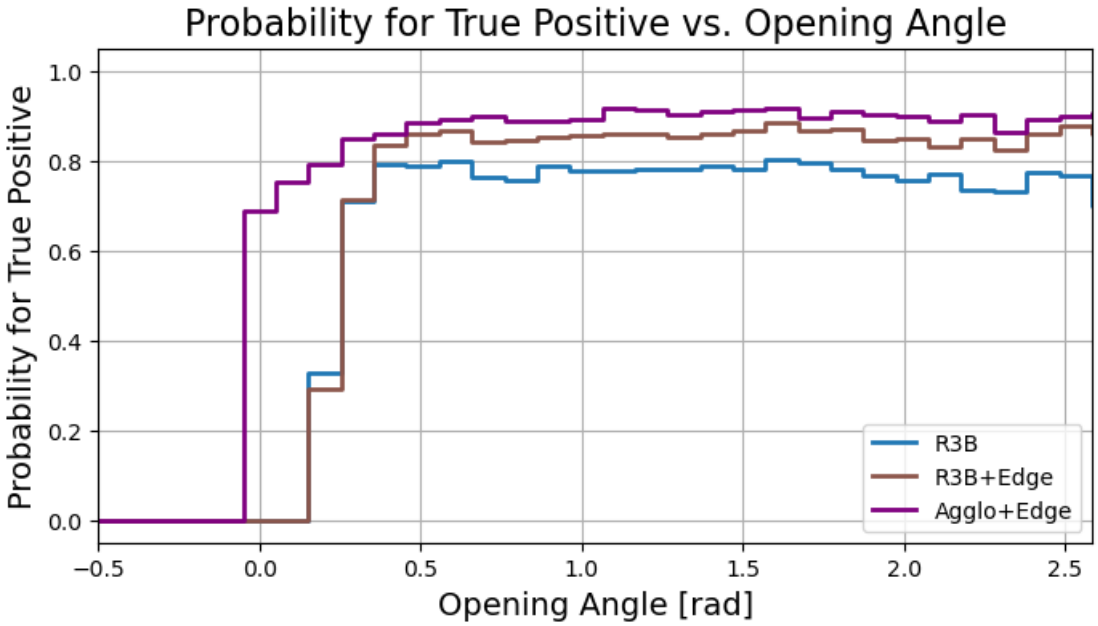
Caption: (a)Graphical representation of the CALIFA detector. Carbon fiber alveoli and aluminum holders fix the 15 to 22 cm long CsI(Tl) crystals. The gray boxes surrounding the holding structure represent the preamplifiers. (b) Cross profile and longitudinal section of the detector. The azimuthal angular coverage of the crystals vary between 1.5° (End-Cap) to 3° (Barrel). Figures taken from Ref.[X]

This figure with description was added to the Appendix of the paper:



Caption: Ratio of well-reconstructed clusters as a function of the true energy deposit in the cluster. For the geometrical R3B clustering algorithm, the well-reconstruction ratio decreases with increasing cluster energy. The R3B+Edge approach partially compensates this degradation, while the Agglo+Edge method maintains a consistently high well-reconstruction ratio over the full energy range.

Two-gamma analysis:
Here we reconstructed the clusters for events with exactly two gamma “true clusters” (the energy of the gammas is uniformly distributed 0.3 MeV < E < 10 MeV)
In the attached figure you see the probability for true positive – meaning that both true clusters in this case are correctly reconstructed – vs the opening angle of the true clusters. The angle is computed by using the position of the highest energy hit for each true cluster. Both models R3B+Edge and Agglo+Edge were trained on three-gamma events, as presented in the paper.
This figure with description was added to the Appendix of the paper:



Caption: True positive reconstruction probability as a function of the opening angle for the Geometrical R3B Clustering, R3B+Edge, and Agglo+Edge approaches. The true cluster angle is determined from the positions of the highest-energy hits in each true cluster. The analysis is restricted to events with exactly two gamma-ray true clusters, with energies uniformly distributed in the range 0.3 MeV < E < 10 MeV.

As expected the geometrical R3B Clustering fails for low opening angles between the clusters, since if the hits of both true clusters are smaller than 0.25 rad all hits will be merged to one cluster. For larger opening angles the distribution is flat, no dependence between the cluster reconstruction efficiency and position is observable.

Q2: l. 21 ff: some more details on the CALIFA layout would be useful, e.g. a drawing (see above), the (typical) size of the crystals (in cm and in Moliere radius and radiation length)

See figure with caption in Q1”Graphical representation of the CALIFA detector” which was added to the paper. This gives most geometrical details for the reader and gives a rough overview of the experimental setup.

Q3: l. 28: what is "high rate" in this context?

Misleading. The statement was removed.

- l. 98 ff: do I understand correctly that the cluster position is not updated while adding hits to the seed hit? and also that you always first finish the cluster seeded from the highest energy hit, even if there is a close-by second cluster where the seed hit is outside the cone of the first cluster, but there are some hits that sit in the cones of both seeds? has this procedure been optimised?

This is correct, there is no cluster center update. The main idea behind is that the hit with the highest energy is assumed to be the first interaction of the gamma (from a reaction of the projectile in the target) with the detector material. Moving the center position changes the polar angle which is crucial for the Doppler correction. Hence updating the center position may help to catch the outer hits but on the other side shift the "true" polar angle and therefore distorting the Doppler correction. The cluster reconstruction works step by step, meaning that one cluster after another is reconstructed in the geometrical R3B clustering and the Agglomerative clustering.
For all reconstruction methods involving the Edge model (Edge, Agglo+Edge, R3B+Edge) instead all pairwise combinations are evaluated in one step.

Revised Section on the geometrical R3B clustering:

For the cluster reconstruction all detected hits are sorted in a list by descending energy. A user-defined cluster shape, typically a cone with an aperture of 0.25 rad, is chosen. This value represents an optimal balance between the compact, high-energy clusters characteristic of light charged particles and the more diffuse showers produced by gamma rays.
The clustering process begins by assigning the hit with the highest energy as the center of the first cluster. The algorithm iterates through the remaining hits in the sorted list. A hit is added to the current cluster if its angle relative to the cluster's central hit is within the defined aperture. After all hits in the list have been processed, the assigned hits are removed. The hit with the highest remaining energy is then selected as the center of a new cluster, and the process is repeated until no unassigned hits remain.

***Q4**: l. 109: please mention the physics list that has been used in Geant4, and the Geant4 version number*
geant4-11-02, Physics List:QGSP_BERT_HP . This was added to the text.

***Q5**: l. 122: I cannot really judge how realistic these event topologies are, but I find it at least surprising that the number of gammas is always 3.*

The choice of the three-gamma event topology was made as a showcase scenario, motivated by the fact that in realistic experimental conditions the average number of reconstructed clusters from both signal and background reactions with comparable energies is close to three. The specific example shown in Fig. 2 corresponds to the emission of three 2.1 MeV photons and can be regarded as a "worst-case" scenario, since at this energy the event topology is strongly dominated by Compton scattering, leading to a comparatively broad spatial distribution of the energy deposits, as depicted in Fig. 1 of the paper.

Using mono-energetic photons in this way provides the reader with a direct and intuitive comparison of the reconstruction methods, as the differences in performance become clearly visible

in the reconstructed cluster energy spectra shown in Fig. 2. Furthermore, the improvement achieved by the ML models trained on the three-gamma dataset has also been verified for two-gamma events, as shown in the response to Q1.

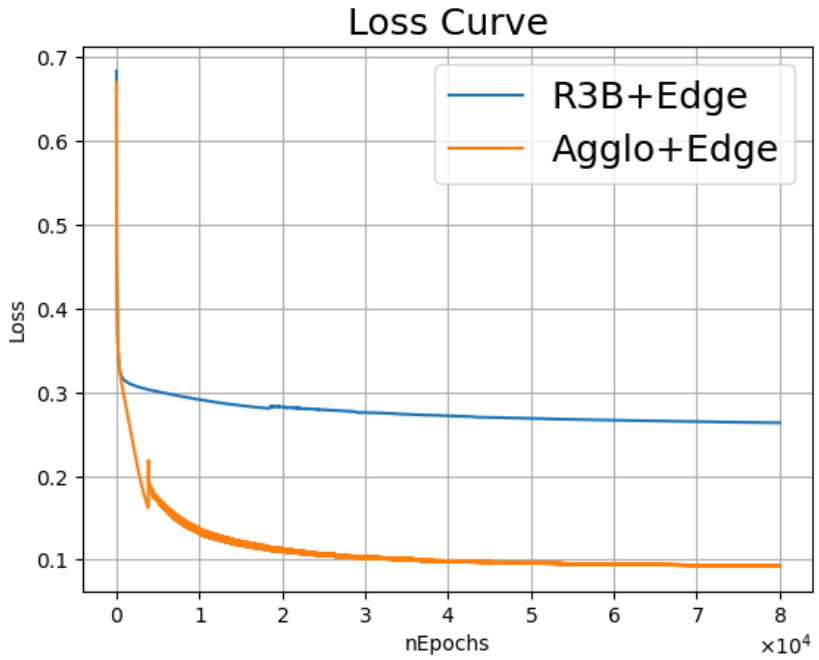
***Q6**: l. 130: I find it a bit odd that the true energy is corrected for gammas that deposit only a fraction of their energy in the detector volume.*
Would it not make more sense to restrict the angles in the generation such that the showers of all gammas are fully contained in the
detector?

We agree that the text was not explaining the procedure very clearly. This was handled as such for simplicity. It makes it from the technical point easier to do simulation and consequent analysis. These effects by partial contained clusters cannot be avoided in the real experiment while additional contributions from cluster reconstruction should be avoided.
Moreover this truncation should have no effect in the model training stage, since all the models only do hit-wise comparisons. See more about this in the answer to Q11.

Here the revised version of the sentence to clarify:
Event selection is limited to cases in which all three gamma rays are emitted within the geometrical acceptance of the CALIFA detector, which only partially encloses the target region. For gamma rays that deposit only a fraction of their energy in the detector volume – such as in cases where the incident gamma ray undergoes Compton scattering, deposits part of its energy in the calorimeter, and subsequently escapes the active volume - the corresponding true energy is adjusted to reflect only the energy actually deposited in CALIFA.

***Q7**: l. 133: the total sample size of 20000 events looks rather small to me*

Since the feature size of 12 is small, the complexity of reconstruction task is not that high. As shown by the loss curve the loss converges. And from the metrics results for training and validation no overfitting was observed.



This figure of the loss curve with description was added to the Appendix of the paper:

Caption: Loss curves for R3B+Edge (blue) and Agglo+Edge (orange) clustering models as a function of training epochs, demonstrating convergence performance. Both models were trained using the binary cross-entropy loss with a learning rate of 5×10^{-3} , demonstrating stable convergence behavior.

***Q8**: section 2.4: I think in addition to the performance metrics defined here, that are purely based on hit-counting, it would be interesting to also look at quantities related to the energy*

This was discussed in the answer to Q1. The plot “well reco vs true_energy”, plotted in Q1, was added to the Appendix of the paper

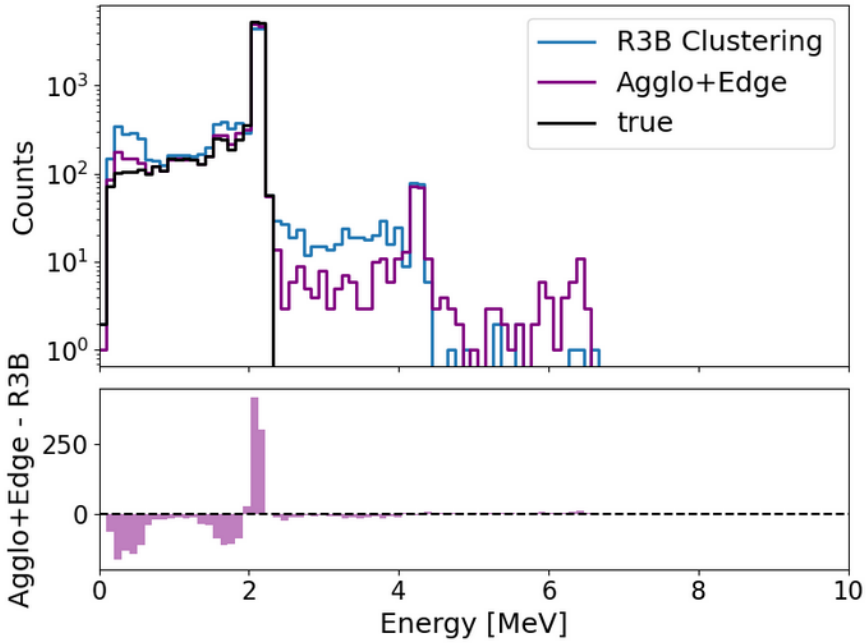
***Q9**: l. 207: how did you decide to stop after 8 * 10^4 epochs? How do the losses look like for the training and the test data samples?*
- discussion of figure 2: I'm not sure what is plotted here: individual cluster energy, or sum of all clustered energies? And are all the clusters fully contained in the detector volume, or do you expect reduced true energies due to the procedure described in l. 130? By how much deviates the number of reconstructed clusters from the true number of clusters?

From one of the above plot of the loss curves in answer to Q8, the loss flattens out and remains almost constant. Which means that the model does not learn much more after 60k epochs. But there is no strict condition to finish at 80k or 100k or 150k epochs. This was just limited by the computing time.

In Fig. 2 of the paper the individual cluster energy is plotted. We updated the caption to make this clear. And we have added the true cluster energy distribution. As pointed out we expect reduced true energies due to the procedure described in l.130 (i.e. partial contained clusters in the calorimeter), which is visible in the long black tail in the updated Fig. 2.
Short summary to the number of reconstructed clusters in Fig. 2:
True cluster number: 13500
Geometrical R3B Clustering: 14028 (due to the false negative rate)
Agglo+Edge: 13353

Reviewed caption of Fig. 2:
Reconstructed gamma cluster energy spectrum from simulated events, each consisting of three 2.1 MeV photons emitted from the target point. This showcase can be regarded as a “worst-case” scenario, since at this energy the event topology is strongly dominated by Compton scattering, leading to a comparatively broad spatial distribution of the energy deposits, as depicted in Fig. 1. The upper histogram shows the reconstructed cluster energy distribution using the geometrical R3B clustering (blue),the Agglo+Edge(pink) method accordingly, and in black the true energy cluster distribution. The lower panel displays the bin-by-bin count difference between the two approaches. The Agglo+Edge model demonstrates a significant improvement by successfully reattaching escaped hits, notably in cases where sparse energy deposits around 1.6 MeV and 0.5 MeV result from pair production and subsequent annihilation processes of the original gamma photons. This clean-up step leads to a marked reduction in false negatives (i.e. reduction of bin counts at 0.5 and 1.6 MeV) compared to the geometrical R3B clustering and an enhancement of 2.1 MeV peak.

Reviewed
Fig. 2:



***Q10**: discussion of the results: I think there are some rather easy additional studies and cross checks that might provide additional insights:*
- what happens if you apply the algorithms to events with 2 or 4 gammas?

The two gamma analysis and the figure “Probability for True Positive vs. Opening Angle” was already presented in answer to Q1.

***Q11**: discussion of the results: I find it really surprising that the edge clustering with the (relatively simple) pre-clustering with the geometric algorithm leads to a better result than the edge clustering alone. Is there any way to get an idea why, e.g. by looking into events that are correctly clustered in one case, and incorrectly in the other?*

At first glance, the observation indeed seems surprising. However, several aspects may help explain this behavior:

→ Across all clustering models studied, the weakest metric is the false negative (FN) rate, i.e. the models more frequently fail to merge a hit into its correct cluster than mistakenly assign a hit to a wrong cluster. The pre-clustering step with the agglomerative algorithm (Agglo) provides a useful prior grouping and allows the subsequent Edge model to specialize more effectively in the decision of whether a hit (or cluster) should be merged or not.

→ In the Agglo+Edge configuration, the Edge model is trained only on pre-clustered events corresponding to true positives (TP) or false positives (FP), but not on false mixed (FM) events. This again enhances the learning process.

→ All “Edge” models used in this work operate exclusively on pairwise hit (or pairwise cluster, in the pre-clustered case) comparisons. No snapshot of the full event is done. This would involve some sophisticated transformer models (which did actually not succeed for our reconstruction

problem). In principle, such models should be more powerful as they could learn full cluster shapes. As pointed out in the outlook, it would be a natural next step to

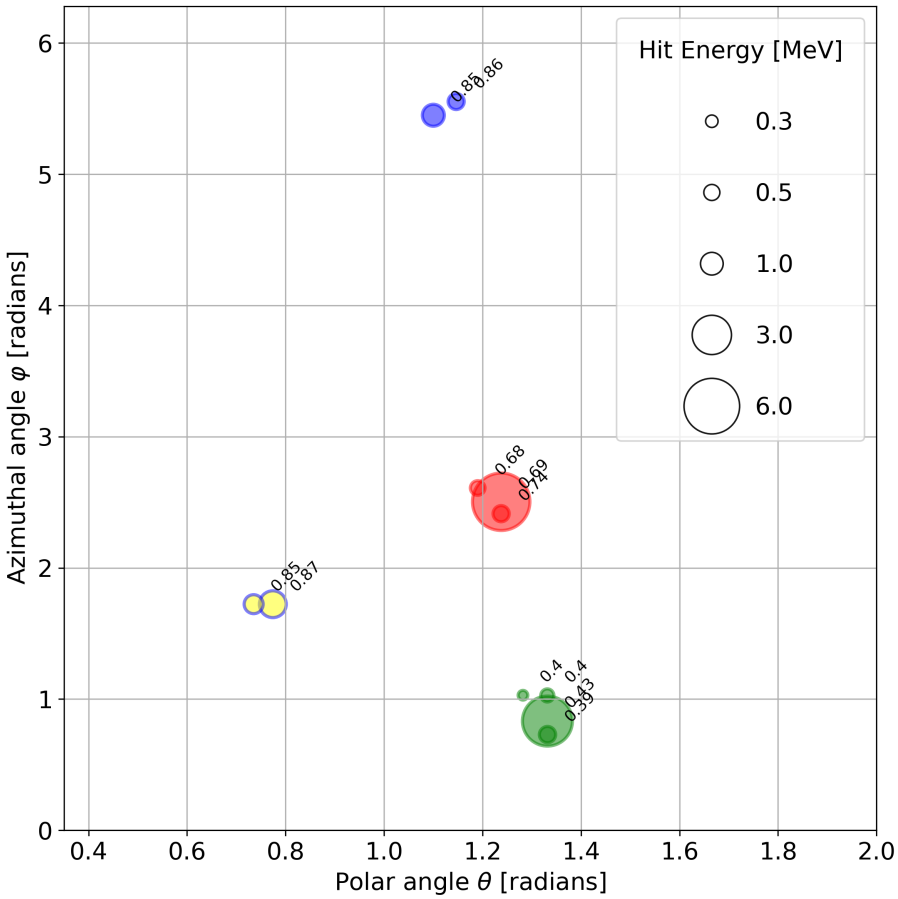
Taken together, this explains why the standalone Edge clustering, despite having access to more raw information, cannot fully exploit it due to its pairwise nature and consequently exhibits a relatively high FN rate (≈20%). By contrast, the Agglo+Edge approach benefits from the structured prior of the pre-clustering and is therefore able to achieve superior overall performance.

Following sentence was added to the “3. Discussion” part of the paper to motivate the results:

The incorporation of pre-clustering step in both Agglo+Edge and R3B+Edge modes acts as a clean-up stage, reducing false negatives and enabling the Edge model to specialize more effectively in merging decisions.

***Q12**: figure 3, example event: this seems to be a rather easy-to-reconstruct event, where a slightly increased cone radius would give the*
correct result for the geometrical clustering as well. Maybe a more complex event would be more interesting?

This figure of an example event with a more complex topology – the blue cluster with widely spread hits - and description was added to the paper:



Caption: Example of a simulated event involving three primary photons, illustrating the performance difference between the Agglo+Edge clustering method and the geometrical R3B clustering approach. Each marker represents a detected hit, plotted as a function of the polar angle θ and the azimuthal angle ϕ. The edge color of each circle indicates the true cluster assignment (ground truth), while the fill color denotes the cluster assignment according to the geometrical R3B clustering. The size of each circle reflects the energy deposited in the detector segment. Numbers adjacent to the hits represent the normalized hit times. In this event, the geometrical R3B clustering incorrectly assigns the two hits at (θ ≈ 0.8 rad, ϕ ≈ 1.8 rad), with a normalized times of 0.85 and 0.87 (blue edge, yellow fill) respectively, to a separate cluster, resulting in a False Negative (FN). In contrast, the Agglo+Edge method correctly assigns all hits to their respective clusters.

***Q13**: references: some references look incomplete (or not published), e.g. [1], [5]*
also, since LLMs have been used in the preparation, I would find it useful to have a link for each reference so I can easily check if it exists

Done, thanks for the hint!