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

Please edit your bibliography to limit the number of names per reference for articles with more than 5 authors to a maximum of three names, followed by et al. This journal prefers a style where only one author name, followed by et al., is cited when more than 5 authors are on the author list.

Done, thanks for the tip.

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 will be added to the paper:

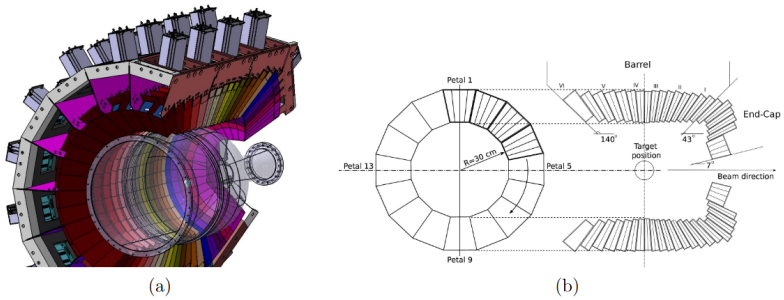


Fig. X. (a)Graphical representation of the CALIFA detector. Carbon fiber alveoli and aluminum holders fix the 15 to 22 cm long CsI(Tl) crystals. The gray boxes surrounding the holding structure represent the preamplifiers. (b) Cross profile and longitudinal section of the detector. The azimuthal angular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

This picture with description will be added in the “supplementary material” part of the paper:

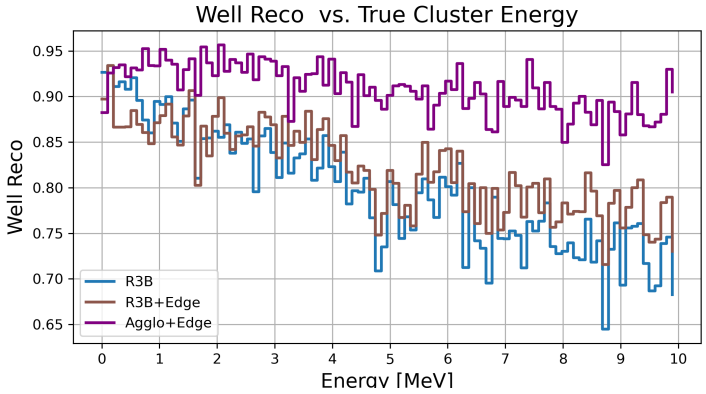
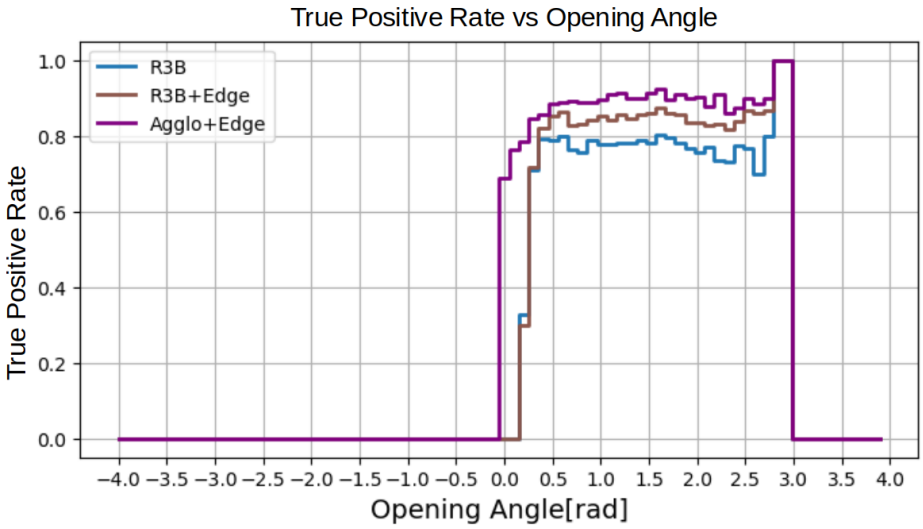


Fig. X: Ratio of well-reconstructed clusters as a function of the true cluster energy. For the geometrical R3B clustering algorithm, the well-reconstruction ratio decreases with increasing cluster energy. The R3B+Edge approach partially compensates this degradation, while the

Agglo+Edge method maintains a consistently high well-reconstruction ratio over the full energy range.

Two-gamma analysis:
Here I reconstruct 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 ratio of 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:



As expected the geometric 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.
*the entries of the two last bins are misleading. They only contain 5 and 1 entries accordingly due to the partly filled detector.

- 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 first picture with caption in this document. I hope this gives enough info for the reader and gives a rough overview of the experimental setup.

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

Up to O(10^4) ions/s. This is added in the text.

- 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?

Yes correctly, 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 azimuthal angle which is crucial for the Doppler correction. Hence updating the center position may help to catch the outmost hits but on the other side shift the “true” azimuthal angle and therefore distort the Doppler corrected energy.
The cluster reconstruction works step by step, meaning that one cluster after another is reconstructed. There is no simultaneous cluster building with hit sharing etc.

Revised Section:

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.

- 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 is added in the text.

- 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.

You are correct that from a pure physics standpoint, it is unexpected to have exactly three gammas with same energy for each event. This was a rudimentary observation based on specific experimental campaigns, such as (p2p) experiments where we observed around 10 low-energy hits, which should be assigned to 1-2 background clusters and around one cluster from particle de-excitation in the reaction.

For other physics cases, like fission, we might expect to see more clusters in CALIFA, while others may produce fewer.

The main goal of our algorithm is not to focus on the exact number of gammas. Instead, we want the algorithm to learn the correlation between hits so that it can accurately group hits together. The choice to use a fixed number of gammas per event with same energy was an initial step and makes also easier to interpret reconstructed cluster energy distributions, as in Fig2. We can see there the single escape peak at ~1.6MeV and the corresponding 511keV peak. Using uniformly distributed gamma energies would smear out the distributions and would it make also more difficult to see just by looking at the plot, if the applied model improves reconstruction.

- 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?

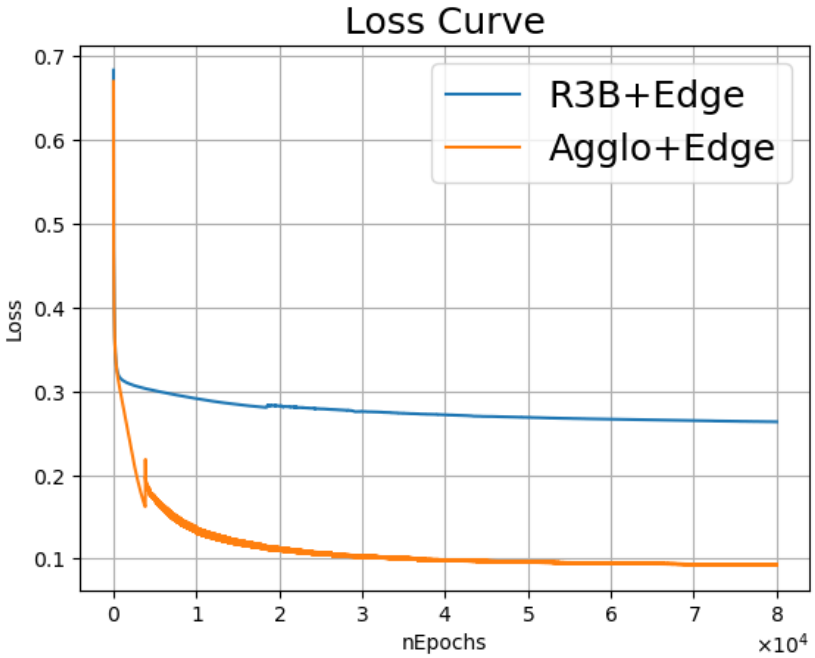
Here the revised version of the sentence to clarify:
Event selection is limited to cases in which all three gamma rays are emitted within the geometrical acceptance of the CALIFA detector, which only partially encloses the target region. For gamma rays that deposit only a fraction of their energy in 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.

I see your point. This was handled as such for simplicity. It makes it from the technical point easier to do simulation and consequent analysis. However this truncation should have no effect in the model training stage, since all the models only do hit-wise comparisons. See more about this discussion in one of the next answers.

- 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 for both training and validation the loss converges for both of them. And from the metrics results for training and validation no overfitting is observed.

This loss curve plots will be added to the paper as additional material:



- 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 one of the answers above. The plot “well reco vs true_energy” will be inserted to the paper as “supplementary material”.

- 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?

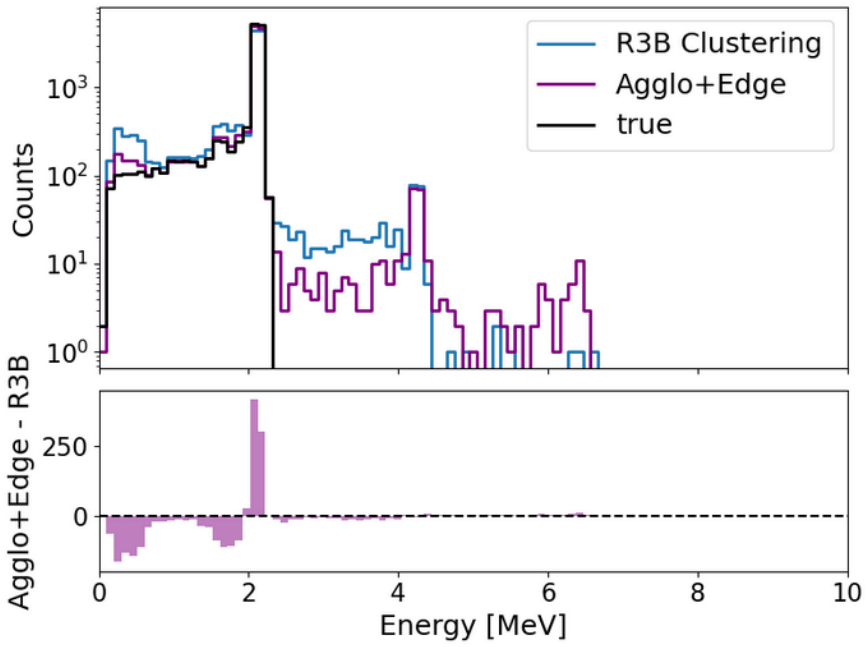
As you can see from one of the above plots of the loss curves, the loss flattens out and remains almost constant. Which means that the model does not learn much more after 60k epochs. But there is no strict constraint to finish at 80k or 100k or 150k epochs.

In Fig. 2 the individual cluster energy is plotted. I updated the caption to make this clearer. And I have added the true cluster energy distribution. As you point out I expect reduced true energies due to the procedure described in l.130, which is visible in the long black tail in the updated Fig. 2.
Just to give you the number of reconstructed clusters in Fig2.
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 gamma photons emitted from the target point. The upper histogram shows the reconstructed cluster energy distribution using the geometrical R3B clustering (blue) and the

Agglo+Edge(pink) method accordingly. 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:



- 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?

See two gamma analysis in a previous answer.

- 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?

Yes, at the first glance this seems to be surprising. However, there are several aspects which may offer an explanation:
→ no matter which clustering model we look at - the metric which perform worst is the FN, meaning that the model more often do not merge a hit to the correct cluster than wrongly incorporating a hit to another cluster. The preclustering step with the Agglo clustering allows the algorithm to specialize itself in decision making if a hit (or cluster) should be merged or not.
→ moreover to enforce this capability in the Agglo+Edge model, the Edge model here is only trained on preclustered events which are either correctly clustered (TP) or on FP events, no false mixed (FM) events. This enhances again the learning process.
→ important to understand is that all “Edge” models discussed in this paper only do pairwise hit (or in preclusterd data also pairwise cluster) 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). Models which take a snapshot of the event should (in theory) be more powerful, they should really learn whole cluster shapes. Building such a model would be the next step (as mentioned in the outlook). For such models the data should not be truncated as done here for clusters where only partial energy is deposited in the detector, as mentioned above and in the text, since it could disturb the learning process.
Since we do only pairwise hit comparisons the standalone edge clustering cannot profit from the higher amount of info that we have in the not-preclustered data-sample, on the contrary it does not aggressively enough merge hits (FN rate: 20%).

- 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?

I have all interesting events (meaning: events where the Agglo+Edge correctly assigns all hits, while the geometrical R3B Clustering fails) in :
https://syncandshare.lrz.de/getlink/fiCFrguk3hJdMMif4oGLq6/
TODO: an eligible will be selected.

- 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 tip!