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 were replaced/added to the paper

Q1:What is the typical energy, spatial and timing distributions of the gamma rays emitting from the target? If this question is too general, maybe you can use the case of the most dominant type of events encountered at R3B, or the most interesting one.

Indeed this is not easy to answer due to the broad physics program envisioned by R3B. Typically energies from few 100keV for typical single particle excitations in odd nuclei up to more than 10 MeV in case of collective Dipole excitations get occupied in the reaction process. However, when forward boosted Doppler shifted energies up to 20 and more MeV are possible.

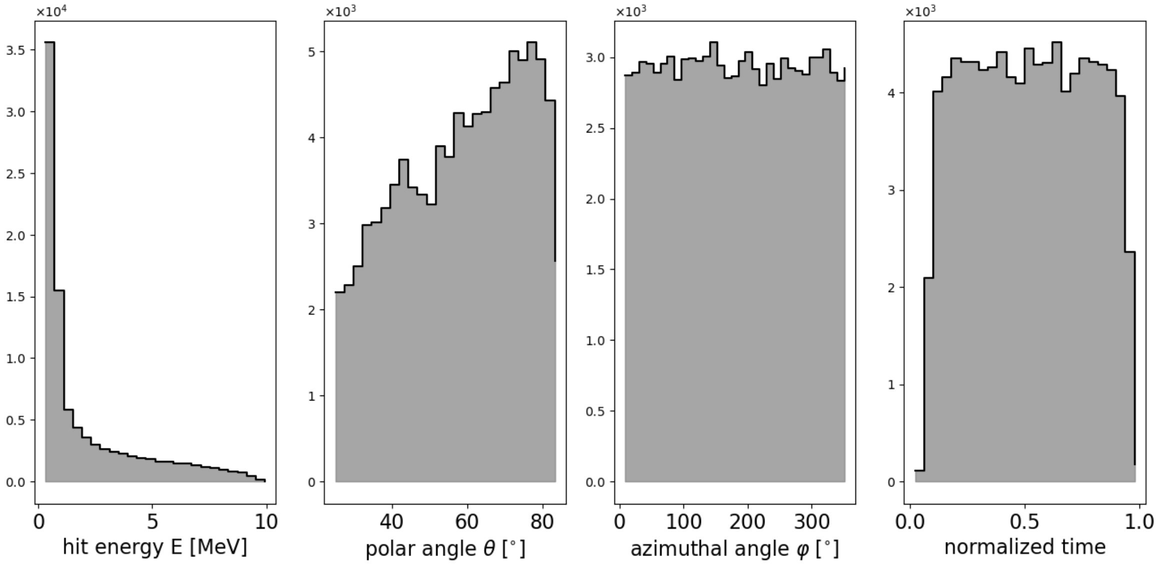
We used for the studies as example the commissioning experiment S444 in 2020 where we had a ^{12}C beam at 400/550/650 and 800 MeV per nucleon impinging on a carbon (or plastic) target. In the quasi-free-scattering reaction, where a proton from the ^{12}C projectile gets knocked out by a proton from the target and you end up with a ^{11}B fragment. This can be in the ground state or excited. If excited it immediately de-excites and emits a (boosted) gamma (2.1MeV in the ^{11}B frame) which we then detect in CALIFA.

Q2: What does the 4µs window correspond to, in term of the experiment? Is it around a beam signal, or a trigger signal?

The 4 µs window is defined relative to the trigger signal, which is provided by an upstream detector at the entrance of the experimental hall. Its width is chosen to ensure that all prompt and delayed signals correlated with the reaction are included. This covers environmental and δ-electrons, as well as weaker channels involving γ-ray cascades or delayed particle emission, as typically observed in heavy-ion fission experiments at comparable energies.

Q3: The input gamma rays are sampled from a uniform energy distribution between 0.3 MeV and 10 MeV, what is the angular distribution sampled, in particular the polar angle, also uniform?

Yes, all distributions are flat as shown in the following figure with description which was added to the Appendix of the paper:



Caption: Distributions of the single hit features in CALIFA, obtained from Geant4 simulations with incident photons. While the primary photons were generated with uniform distributions in energy, polar and azimuthal angle, as well as normalized time, the reconstructed hit features exhibit detector- and physics-driven effects. The hit energy spectrum is dominated by low-energy deposits, reflecting the enhanced probability of pair production and Compton scattering at higher photon energies, while the contribution of the photoelectric effect decreases (cf. Fig. 1). The reduced statistics at small polar angles arise from the lower solid-angle coverage of the crystals in this region ($d\Omega = \sin\theta \, d\theta \, d\varphi$).

Q4: Does the MC simulation include the detector response and electronics readout effect?

The MC simulation includes the detector response thanks to the Geant4 library which simulates the interaction of the emitted gammas with the detector material CsI(Tl).
The electronics readout effect/response is not simulated with Geant4. By smearing out the times of the hits belonging to one cluster (true information) with a Gaussian with sigma of 200us we try to approximate the electronic response with respect to the time.
Geant4 version and used physics list was added to the paper:
geant4-11-02, Physics List:QGSP_BERT_HP

Q5: Can you show the distributions of 12 inputs for the Edge model training? Is it possible to compare these variables between MC simulation and real data with some kind of calibration runs?

The four input feature distributions, as shown in answer to Q3 will be added to the text.
In principle it is possible to compare these variables between MC simulation and real data - more precisely with experimental runs (with a source we have no time information).

Q6: Can you compare the output of Edge model output between MC and real data from Calibration runs?

This is not possible since in calibration runs with source no reference time is available. In principle an artificial method of summing up several events could be applied.
This sentence was added to the text:
To further validate the cluster reconstruction models presented here, conceptual methods are envisioned for application to source calibration data.

**Q7: Figure 2: why do you select the case of 3 mono-energetic gammas at 2.1 MeV? If it is to represent a mixture of signal and
background, shouldn't the gamma rays be of different energy?**

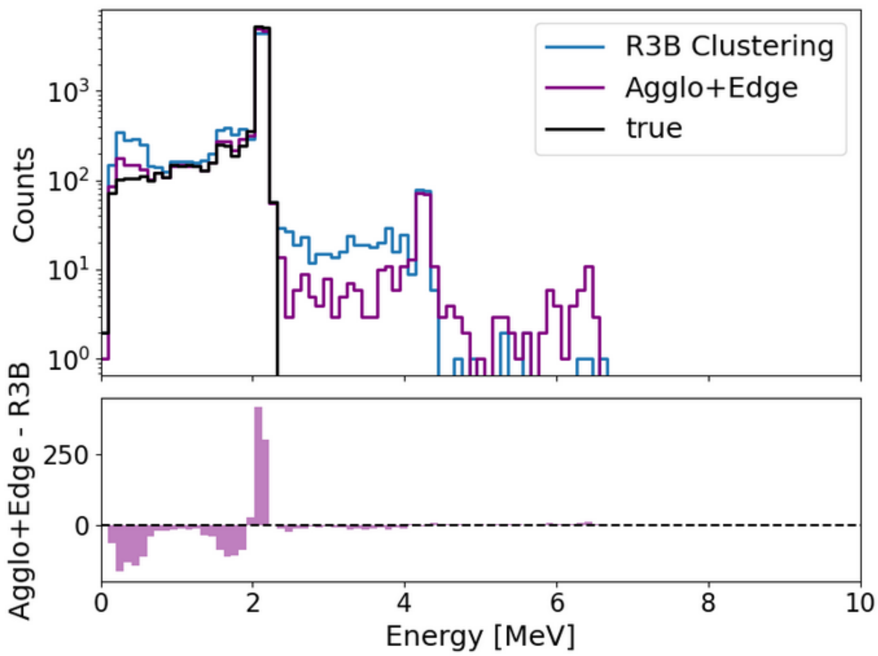
The specific "show case" in Fig. 2 with the emission of three 2.1 MeV photons, motivated by the fact that in realistic experimental conditions the average number of reconstructed clusters from both signal and background reactions is close to three, and can be regarded as a "worst-case" situation, since 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.

This we now pointed out in the revised caption of Fig.2, see response to Q9.

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.

Q8: Figure 2: can you add the truth distribution: i.e. where the fake rates are 0 and TP=1?

This was implemented in Fig. 2.



The observed tail on the true-positive distribution is described in the revised version in the paper: 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.

Q9: Figure 2: It would also be interesting to have the same plot but comparing cluster energy, rather than event energy, between different models. I guess this will show up as peaks centered at 2.1 MeV with different height (efficiency) and width (resolution).

Sorry for the misleading caption. Fig. 2 caption was changed to:
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.

Q10:What is the significance of various fake rates and WR with respect to physics measurements of experiment? This study focuses only on the correct association of hits to clusters. Is this the most important requirement of the calorimeter reconstruction? For example, I can imagine in some applications where noise hits are non-negatable missing a few low energy hits may be less important than a precise localization or energy measurement of the cluster. Is there other "Performance Metrics" worth exploring for the CALIFA calorimeter reconstruction?

A reduction in the energy reconstruction is not negligible, even for low energy hits missing (crystal energy threshold ~ 100keV) since for experimental runs the Doppler correction scales this missing energy. Every non correlated cluster generates an additional gamma that spoils the spectrum at low energies which as consequence could impede the reconstruction of typical gamma cascades in nuclear excitations.

Furthermore, we agree that precise spatial localization is of great importance. Given that gamma rays cannot be tracked by typical pixel detectors, our position reconstruction is inherently limited by the crystal size. This is why a precise Doppler correction is so highly dependent on accurate position reconstruction.
While this work primarily focuses on the energy cluster reconstruction, we recognize the broader significance of localization and will address its impact more explicitly in future work.

Q11: The numbers in table 1 are measured with the uniform energy distribution between 0.3 MeV and 10 MeV. Is there any energy dependence of these numbers?

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

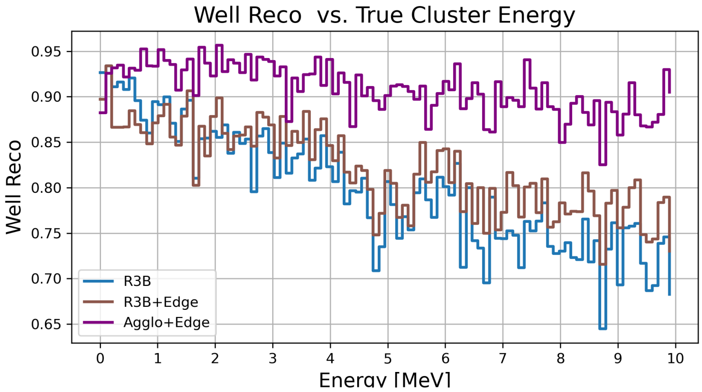


Fig. X: Ratio of well-reconstructed clusters as a function of the true cluster energy. For the geometrical R3B clustering algorithm, the well-reconstruction ratio decreases with increasing cluster energy. The R3B+Edge approach partially compensates this degradation, while the Agglo+Edge method maintains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

Q12: Can you discuss in the "outlook" section how this model will be integrated into the R3B data analysis to produce improved measurements, and cite some examples where the improvements will bring significant impact on the science output of R3B?

This text was added to the paper:
The methods developed in this work can be directly integrated into the R3B data analysis chain as analysis tasks within the R3BRoot framework [6]. Their application is of particular relevance for heavy-ion experiments where the signal reconstruction is challenged by large background contributions, the production of δ-electrons, and the simultaneous emission of a large number of neutrons and γ-rays. In such scenarios, the improved reconstruction performance is expected to enhance the precision of invariant-mass spectroscopy and kinematical reconstruction, thereby contributing directly to the scientific output of the R3B program.

Q13: Publication details are missing for Reference [5].

Done, thanks for the hint!

Q14: It would be useful to define "A GeV".

We decided to use GeV per nucleon instead of the ambiguous and undefined term AGeV.