ATLAS PUB Note
ATL-SOFT-PUB-2025-003
September 9, 2025

Photon showers in the ATLAS fast calorimeter simulation: A voxelized dataset with minimized information loss and improved ML models

The ATLAS Collaboration

For the upcoming high luminosity runs at ATLAS, it is necessary to replace the full calorimeter simulation with the fast calorimeter simulation in almost all cases. The current ATLAS fast calorimeter simulation, AtlFast3, uses parametric approaches to speed up the computing time significantly. Even though AtlFast3 generally achieves a high degree of accuracy, further quality improvements within the fast calorimeter simulation are necessary to enable this replacement. In this note, we present a new voxelization scheme for fast calorimeter simulation, with a specific focus on photon-induced showers. This updated discretization significantly reduces several artifacts previously observed in reconstructed photon shower shapes. While this work concentrates on photons, extensions to other particle types are foreseen in future developments. The resulting photon dataset is expected to be released for public usage in the near future. Furthermore, it is shown that the voxelized dataset can be learned by normalizing flows and diffusion models, which are able to generate highly similar datasets afterwards.

1 Introduction

The current fast simulation in ATLAS is done by AtlFast3 [1, 2], a fast simulation framework that parametrizes the calorimeter response and uses FastCaloSimV2, a mixture of classical histogram based parameterizations, and FastCaloGANV2, a generative adversarial network applied to showers at the event level. The classical, histogram based approach was very successful in the past. However, it is fundamentally not able to capture hit-correlations at the event level within individual calorimeter layers. Furthermore, it predicts the layer energies E_{layer} based on a complex but handcrafted model, which is unable to capture all correlation between these layer energies. These two potential disadvantages can be improved by employing generative machine learning, such as FastCaloGANV2, to generate individual calorimeter showers. To achieve this goal, two major strategies exist. The first is to train a generative model to predict individual hits. In this case the training dataset takes the form of a point cloud, which is the more general of the two approaches. Alternatively, hits can be binned into three dimensional voxels, where the third dimension corresponds to a layer index without further increase in granularity. This second approach has several benefits at the cost of requiring a suitable voxelization scheme:

1. The task becomes an image generation problem. This is well explored in the literature. In particular, this setup was used for the CaloChallenge [3].
2. The binning of the individual hits ensures a constant input dimensionality. This increases the number of applicable machine learning models.
3. Using a voxelized setup simplifies the exploitation of the local invariance wrt. shifts in η and ϕ [1]. Common machine learning techniques for voxelized datasets are able to implement this local symmetry efficiently.

These advantages are the reason why the voxelization approach was used in the past. FastCaloGANV2 [1] is a generative adversarial network that was trained on a voxelized dataset and used for Run 3. FastCaloSimV2 uses a voxelization to create the average showers that are used for its implementation, as well. Finally, a voxelization approach is also adopted in this work.

Even though FastCaloGANV2 is conceptually more expressive than the classical histogram based approach, its shower predictions were not accurate enough to be used for most of the photon simulation. The accuracy of the classical FastCaloSimV2 approach was generally better and used instead for electromagnetic showers in most cases. This is believed to be caused by two problems. The first fundamental problem is that the voxelization used for the training of the FastCaloGANV2 was required to be coarser than the underlying binning of the classical approach and never underwent a rigorous optimization. The problem arises because the parametrization of the calorimeter crucially relies on the actual voxelization scheme used, as it dictates the clear optimum which any model trained on it can reach. The second issue is that the model selection for FastCaloGANV2 relied solely on the total simulated energy, ignoring the actual shower shape. Since GANs depend heavily on proper model selection and lack an absolute loss, this limitation prevented the GAN from reaching its full potential.

[1] ATLAS uses a right-handed Cartesian coordinate system with its origin at the nominal interaction point (IP) in the center of the detector. The z-axis is along the beam pipe, and the x-axis points from the IP to the center of the LHC ring. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the beam pipe. The rapidity is defined as $y = \frac{1}{2} \ln \frac{E+p_z}{E-p_z}$, while the pseudorapidity η, equal to the rapidity in the relativistic limit, is defined in terms of the polar angle θ as $\eta = -\ln \tan\left(\frac{\theta}{2}\right)$.

In this work the first of these issues is addressed by directly investigating and re-optimizing the effects of voxelization, resulting in a voxelization that is able to surpass the quality of the classical approach while still being coarse enough to be learned by generative ML. The second problem is resolved by introducing a combination of normalizing flows and diffusion-like models that feature absolute losses, which result in a reduced model selection dependence. Efforts to improve the GAN-based approach are still ongoing, although these developments are not covered in this note.

2 Dataset creation and voxelization

2.1 GEANT4 sample generation

For the new dataset it was decided to strictly separate photons, electrons and pions into different datasets, as was done in the previously established approach. The main reason is that different calorimeter layers and binning strategies are needed to describe different incident particles to a sufficient degree of accuracy. This is especially true when comparing hadronic and electromagnetic showers, which are fundamentally different. Training a generative machine learning model on the joint dataset would artificially inflate its number of input features and force it to predict very sparse matrices.

The GEANT4 [4] simulation was executed using version 10.6 and the FTFP_BERT_ATL physics list [5]. The full simulation was performed without pileup, detector noise and vertex smearing. The GEANT4 photons are created directly on the calorimeter boundary, with their momentum consistent with production at the detector origin.

The dataset was generated using photons with incident kinetic energies of 256 MeV $\leq E_{\text{kin}} \leq$ 4.2 TeV and shower center pseudorapidities of $|\eta_{\text{center}}| < 1.35$. Here, the shower center is the energy-weighted center of all hits in one event. More details on the exact distributions can be found in subsection 2.3. In this detector region only the PreSamplerB, the electromagnetic calorimeter (EMB1-EMB3) [6] and the beginning of the hadronic calorimeter (TileBar0) [7] receive a relevant energy contribution (cf. Figure 1). In this context, a layer is determined to be relevant if the reconstructed shower shapes are influenced by not including the corresponding layer in the simulation. This means that all non-relevant layers can be safely ignored during the simulation. This definition of relevance mostly agrees with the previous definition in [1], where relevant layers are those which carry more than 0.1% of an average 1 TeV photons' energy. Out of these five relevant layers, EMB1 (layer 1), which is a strip layer with very high pseudorapidity resolution, and EMB2 (layer 2) have the largest influence over the shower shapes.

Internally, the simulation framework used in this study merges neighboring GEANT4 hits together to save memory. This merging is applied to hits if their absolute distance is smaller than a chosen "merging radius". Therefore, this merging radius imposes a maximal resolution of the GEANT4 simulation itself. During the voxelization procedure it was found that a finer GEANT4 hit merging radius, compared to previously generated samples, is necessary to improve the general simulation quality. Therefore, a merging radius of 0.5 mm was used, instead of 1.0 mm for EMB1 (and EME1) and 5.0 mm for the remaining layers as was done in the previous approach. To prevent the full simulation from consuming too much memory this finer merging radius was only applied in the plane orthogonal to the individual GEANT4 hit-trajectories. Parallel to this trajectory a much coarser merging radius was applied. These radii were chosen to be $R_{\text{merge, longitudinal}} \approx d_{\text{calorimeter layer}}/8$ but not larger than 5 mm. $d_{\text{calorimeter layer}}$ describes the thickness of the corresponding calorimeter layer. The coarser longitudinal merging should not affect the

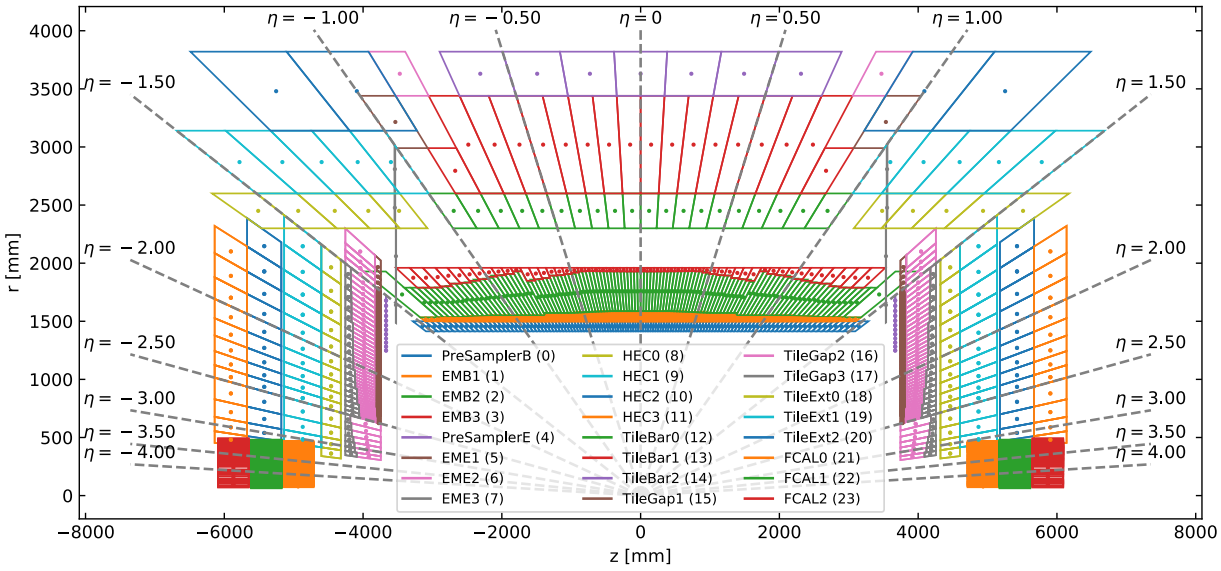


Figure 1: Schematic overview of the ATLAS calorimeter taken from [8]. It visualizes the position of the individual layers. In this note only the barrel is considered. Therefore, the PreSamplerB, the electromagnetic barrel and the beginning of the hadronic calorimeter are of special relevance. At $|\eta| = 0.8$, a transition region of the calorimeter geometry is visible: EMB2 is discontinuously growing while EMB3 is shrinking at the same time. This transition region is of special interest in the following sections as it is harder to simulate and thus a source for potential improvements.

simulation negatively, as the voxelization is blind towards this direction - besides for the boundaries of the calorimeter layers.

2.2 Voxelization procedure

For the voxelization, a polar coordinate frame was chosen. Calorimeter showers feature an approximate radial symmetry around the shower center, and this coordinate frame provides a means to best capture this symmetry. The transition from the cartesian detector coordinates η and ϕ to the local polar coordinates r and α is performed as follows: Let ϕ_i and η_i be the hit coordinates of the i_{th} hit and η_{center} and ϕ_{center} be the center position of the corresponding shower in the corresponding calorimeter layer. Furthermore, let z_{center} be the shower center position in the cartesian z-direction and $r_{\text{center}} = \frac{z_{\text{center}}}{\sinh(\eta_{\text{center}})} \equiv \sqrt{x^2_{\text{center}} + y^2_{\text{center}}}$. Then, the local polar hit coordinates r_i and α_i can be computed as follows:

$$\tilde{\phi}_i = \phi_i - \phi_{\text{center}}, \quad \tilde{\phi}_i \in [-\pi, \pi) \qquad \tilde{\eta}_i = \begin{cases} -(\eta_i - \eta_{\text{center}}) & \text{if } \eta_{\text{center}} < 0 \\ \eta_i - \eta_{\text{center}} & \text{else} \end{cases}$$

$$A = \sqrt{r^2_{\text{center}} + z^2_{\text{center}}} \qquad J = \left| \frac{2\exp(-\eta_{\text{center}})}{1 + \exp(-2\eta_{\text{center}})} \right|$$

$$r_i = \sqrt{(\tilde{\eta}_i \cdot J \cdot A)^2 + \left(\tilde{\phi} \cdot r_{\text{center}}\right)^2} \qquad \alpha_i = \arctan 2\left(\tilde{\phi}_i \cdot r_{\text{center}}, \tilde{\eta}_i \cdot J \cdot A\right), \quad \alpha \in [0, 2\pi)$$

Once these local polar coordinates have been calculated, a theoretically arbitrary voxelization can be applied. In this voxelization study, two "rules" were kept to reduce the degrees of freedom of the optimization:

1. An arbitrary number of "rings" $[R_1, R_2, ..., R_N]$ are defined.
2. $\forall$ hits (r_i, α_i) with $R_j \leq r_i < R_{j+1}$ the same uniform binning in r and α is used.

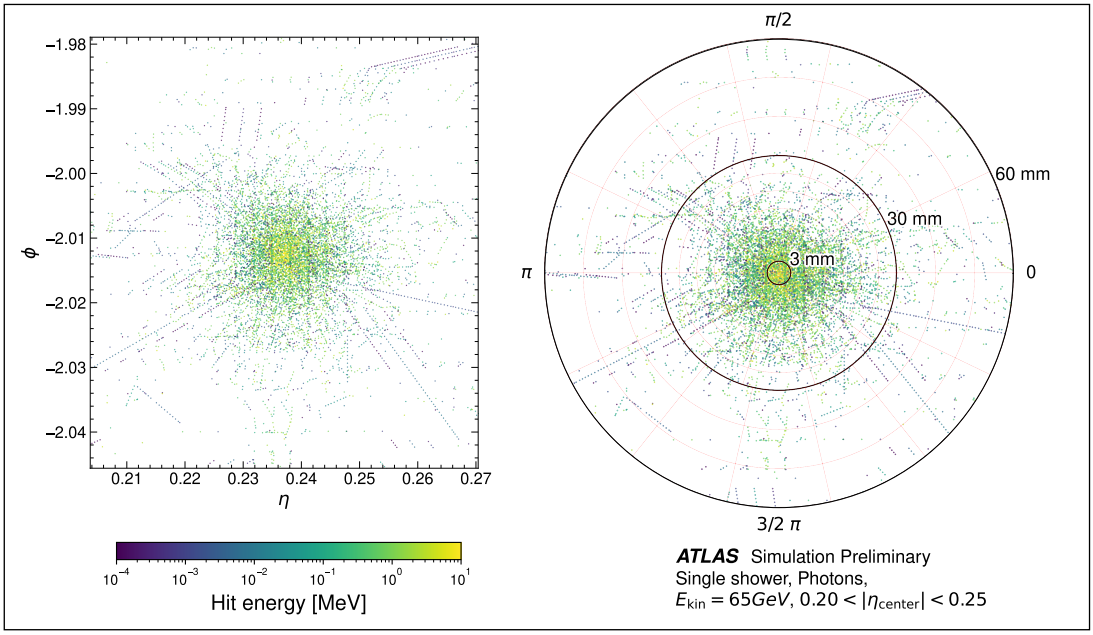


Figure 2: The transition from the cartesian hit coordinates η and ϕ to the local polar coordinates r and α. On the left side one can see a shower of a 65 GeV photon in the calorimeter EMB1. On the right side, the same shower is shown in the coordinates used for the voxelization. Additionally, the *optimal* voxelization (cf. Table 1) for this layer is superimposed using red lines. The change of the radial bin size at the "ring" boundaries at 3 mm and 30 mm is clearly visible.

This implies that the r bin size and the α bin size are allowed to change at the ring boundaries. Once the voxelization is performed all negative voxels are set to zero. The other voxels are normalized such that the total energy does not change. This allows the usage of logarithmic preprocessing steps to the data. These negative energies are a result of the treatment of the induced energy fluctuations caused by a hit in the neighboring cells. To handle this effect "mirror charges" are created, that are shifted by exactly one cell size. Depending on the configuration, this mirror charge can be negative. Nevertheless, the energy of a full cell has to be positive and the negative energies have to be highly localized. For the voxelization used, almost no negative voxel energies were present and their removal does not affect the reconstructed shower shapes. In a last step, all the voxel energies of each event are divided by the kinetic energy of the corresponding incident particle. The transition from the η- and ϕ-coordinates to the voxelized setup is shown for the second barrel layer in Figure 2. It should be noted that the energies of the individual detailed hits do not perfectly sum up to the expected corresponding GEANT4 cell energies. The sum of the energies of these detailed hits is lower than that of the standard GEANT4 hits [1]. This energy shift is resulting in visible differences of the reconstructed photon shower shapes. To reproduce the GEANT4 behavior, the hits corresponding the each cell were renormalized to sum up to the expected GEANT4 cell energy.

The smallest voxelization scheme found that was not observed to introduce artifacts in the reconstructed

shower shapes is shown in Table 1. It will be referred to as the *optimal* voxelization and it consists of 382 voxels. A visualization of this voxelization is given in Figure 3, which depicts an average shower of 65 GeV for events with $0.2 < |\eta_{\text{center}}| < 0.25$. The voxelization boundaries are indicated by semi transparent red lines. It uses finer voxelizations for EMB1, which needs a better resolution as it is a high granularity strip layer, and for EMB2, which carries most of the shower energy and needs to be modeled very accurately. The remaining layers, the PreSamplerB, the EMB3, and the TileBar0 are less relevant, as their spatial resolution and energy contribution are comparably lower.

Apart from this highly non-regular *optimal* voxelization, another voxelization was performed on the same GEANT4 simulation. It was constructed such that it was finer or equal to the *optimal* voxelization everywhere with the additional constraints of requiring the same binning in each layer and having a fixed number of α bins per layer. This "regular" voxelization consists of 1680 voxels, and is therefore about 4.5 times larger but enables the usage of convolutional or patchified [9] generative models. The regular voxelization is noted in Table 2 and visualized in Figure 4

Table 1: Table showing the *optimal* voxelization scheme that was used to create the smaller shower dataset for photons in the barrel. Δ_r is the size of the bins in the radial direction and n_α denotes the number of uniformly distributed α-bins. These two different quantities are given for the following reasons. n_α is given because the α bin size is not constant due to the polar nature of the coordinates. Δ_r is shown as the number of r bins depends on the size of the corresponding ring, meaning the r bin size has greater physical relevance. Unless otherwise stated, the "Default" binning is used for all relevant layers.

Ring [mm]	**EMB1**		**Ring [mm]**	**EMB2**		**Ring [mm]**	**Default**	
	n_α	Δr[mm]		n_α	Δr[mm]		n_α	Δr[mm]
0 – 3	4	1.0	0 – 30	10	6.0	0 – 20	4	10.0
3 – 30	10	4.5						
30 – 70	14	10.0	30 – 100	10	17.5	20 – 100	4	20.0
70 – 100	10	15.0						
100 – 220	4	60.0	100 – 500	4	100.0	100 – 500	4	200.0
220 – 1020	4	400.0						
			500 – 1000	4	500.0	500 – 1000	4	500.0

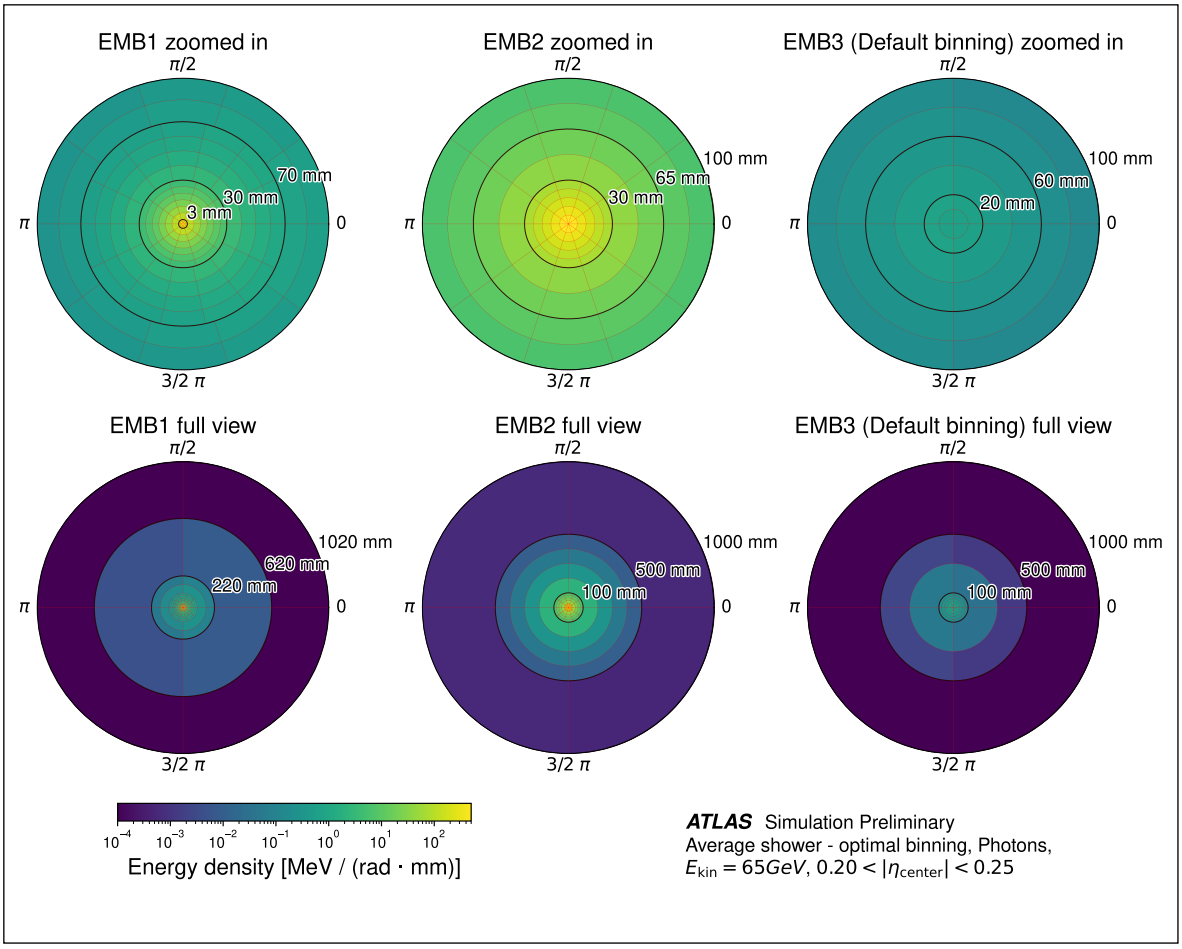


Figure 3: Visualization of the *optimal* voxelization found for photons in the barrel. An average shower is shown for 65 GeV photon samples with $0.2 \leq |\eta_{\text{center}}| \leq 0.25$. Two different radial ranges are depicted. In the upper row, the core of the shower can be seen. The voxelization is indicated by red lines. The voxel size is generally finer for the inner voxels, which tend to contain the highest energies. In the bottom row, the complete average shower is shown up to the outermost edges of the cylindrical voxelization region. After ~ 100 mm, the voxelization gets much coarser. These large voxels are needed to contain a sufficiently high fraction of the total energy of the shower, while not causing a drastic increase in the number of voxels. Any layers which are not shown have the same voxelization as the EMB3.

Table 2: The regular voxelization scheme that was used to create the larger, regular shower dataset for the photons in the barrel. The variable names are the same as those in Table 1.

Ring [mm]	**Default**	
	n_α	Δr[mm]
0 – 3	14	1.0
3 – 30	14	4.5
30 – 70	14	10.0
70 – 100	14	15.0
100 – 520	14	60.0
520 – 1320	14	400.0

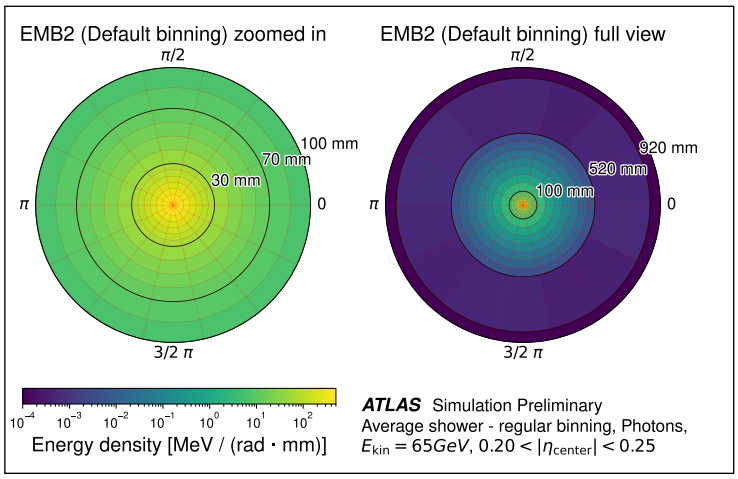


Figure 4: Average shower with the regular binning applied for 65 GeV photons, $0.2 \leq |\eta_{\text{center}}| \leq 0.25$. As was the case for the "optimal" binning, two radial ranges are depicted due to the strong bin size changes in the radius. A zoomed in view is shown in the left inset, while the full voxelization up to the edge of the cylindrical region is shown in the right inset. While EMB2 is shown, all layers use the same regular binning.

2.3 Dataset description

Both the *optimal* and regular datasets are expected to be released for public usage as HDF5-files in the near future. Both consist of a training dataset and a validation dataset. The training dataset consists of 27 η_{center} bins with 150k showers each. The integrated η_{center}-range of the shower centers spans from $|\eta_{\text{center}}| = 0$ to $|\eta_{\text{center}}| = 1.35$, within each η_{center} bin the η_{center} position of the individual events was sampled uniformly between η_{slice} and $\eta_{\text{slice}} + 0.05$. For the validation dataset, the simulation was performed by sampling $|\eta_{\text{center}}|$ uniformly within $0 \leq |\eta_{\text{center}}| < 1.35$, resulting in 4.05M events in total. The remaining generation was performed in an identical way for both datasets. The ϕ_{center}-position was sampled uniformly within $[0, 2\pi)$ and the logarithmic incident energies ($\log(E_{\text{kin}})$) were sampled according to the following probability-density function

$$p(\log(E_{\text{kin}})) = \frac{1}{A} \begin{cases} 1, & 2^8 \text{ MeV} < E_{\text{kin}} < E_{\text{transition}}, \\ \left(\frac{E_{\text{kin}}}{E_{\text{transition}}}\right)^{-k}, & 2^{22} \text{ MeV} > E_{\text{kin}} \geq E_{\text{transition}}, \end{cases}$$

with A being the normalization, $E_{\text{transition}} = 2^{19}$ MeV being the end of the log-uniform region and $k = \frac{1}{\log(2^{22}) - \log(E_{\text{transition}})} = \frac{1}{\log(8)}$ being a decay factor. This decay was mandatory as it was not possible to continue the log-uniform sampling for $E_{\text{kin}} < E_{\text{transition}}$ due to runtime reasons. The form of k was chosen to ensure a final decay of $\frac{1}{e}$, independent of $E_{\text{transition}}$. The energy dependence of this probability density function is visualized in Figure 5. The keys, needed to access the individual values in the HDF5-files, are listed in Table A.1 in the appendix subsection.

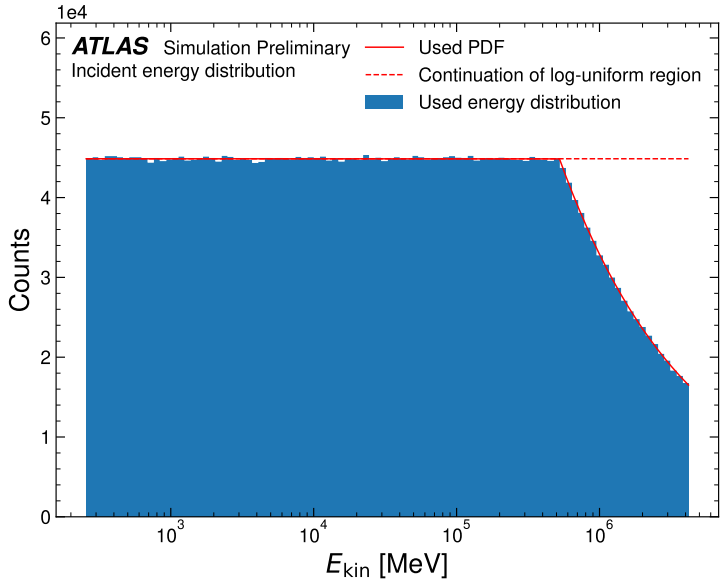


Figure 5: Distribution of the incident energies used for the GEANT4 simulation. The red line visualizes the analytic form of the probability-density function, while the blue histogram consists of 4.05 million events drawn from this distribution.

2.4 Voxelization verification

The voxelization scheme from Table 1 was found by utilizing the *BinnedShower* framework that is shown schematically in Figure 6. It directly uses individual GEANT4 events with the voxelization applied and pipes them to the end of the AtlFast3 simulation. Since no fast simulation algorithms are used this allows the performance of the voxelization to be verified in isolation, as the only possible source of differences from the full simulation is the voxelization scheme. After this step, the same procedure typically applied to fast simulation hits, including the assignment of hits to cells, is used. The exact procedure is as follows:

The BinnedShower internally stores a library with multiple voxelized events. In a first step a *best* event has to be selected. This can for example be done by minimizing the squared distance of the stored GEANT4 shower center from the shower center that is needed for the requested event. The incident energy is considered in addition, if more than one incident energy is part of the shower library. In the second step, the voxel that corresponds to the required hit index I_{hit} is found. Afterwards a hit is deposited inside the voxel's boundaries with an energy computed from the voxel energy. In the last step, the coordinate transformation to the polar coordinates, explained in subsection 2.2, is reverted.

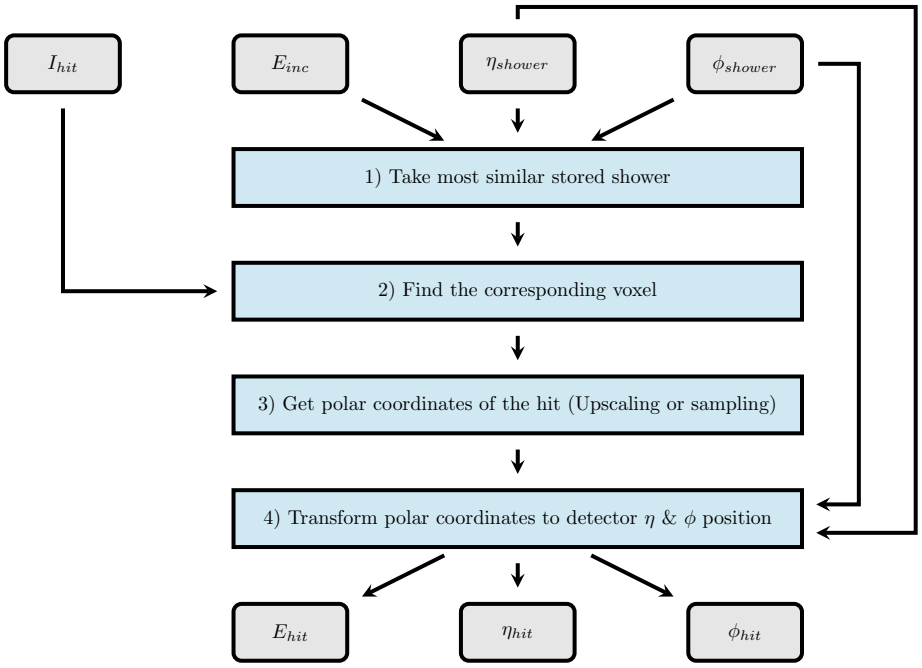


Figure 6: Visualization of the BinnedShower framework. The four steps, visualized in the flowchart, are executed for each hit that has to be simulated until the BinnedShower hits are received. The individual steps are explained in the text in more detail.

For the application of the BinnedShower, an assignment of voxel-energies to cells is necessary. Treating the voxels as single hits is not a good approach. Some observables, like e.g. ΔE (c.f. Table 4) are sensitive to minima in the cell energy distribution. To model these observables accurately, the correct number of hits must be generated to preserve the right statistics. If there are too few hits per cell, for instance if only one hit is created for each voxel, too many minima in the cell energy distribution are created, biasing the physical observables. The used procedure generates n hits per voxel, each carrying the same number of energy E_{hit}. The amount of hits n is only dependent on the energy of the corresponding voxel:

$$n\left(E_{\text{voxel}}\right) = \max\left(1, \begin{cases} \lfloor \frac{E_{\text{voxel}}}{\bar{E}_{\text{hit}}} \rfloor & \text{if } E_{\text{voxel}} \leq 100 \cdot \bar{E}_{\text{hit}} \\ 100 & \text{else} \end{cases}\right)$$

$$E_{\text{hit}} = \frac{E_{\text{voxel}}}{n}.$$

This formula is motivated by the fact that the expected uncertainty of the ATLAS calorimeter is proportional to the square root of the cell energy $\sigma(E_{\text{cell}}) \propto \sqrt{E_{\text{cell}}}$ [6]. Using that the Poissonian error scales in the same way with n, it follows $E_{\text{cell}} \propto \sigma(E_{\text{cell}})^2 \propto n$. The proportionality factor, in this case the expected hit energy $\bar{E}_{\text{hit}}$, is not obvious as the hit assignment has to happen relative to the voxels and not to the cells. Therefore, $\bar{E}_{\text{hit}}$ is treated as tuning parameter of the simulation. The value $\bar{E}_{\text{hit}} = 4$ MeV resulted in good agreement of the reconstructed shower shapes with the GEANT4 prediction for almost the whole barrel. Only for the pseudorapidity region $0.75 \leq |\eta_{\text{center}}| \leq 0.8$, a different expected hit energy of $\bar{E}_{\text{hit}} = 6$ MeV was used. This is a transition region in the calorimeter geometry and the changing cell sizes can affect the proportionality factor. The cutoff $n \leq 100$ was chosen to prevent unnecessary long simulations as 100 hits resulted in already no remaining visible fluctuations.

For the assignment of the hit positions, an average shower with a voxelization which is four times more granular is used. Both the radial and angular resolution of the shower is increased by a factor of two. This average shower is stored for fifteen different incident energies $E = 2^i$ with $i \in \mathbb{N} \wedge 8 \leq i \leq 22$. The discrete energies were obtained by averaging all showers around the individual discrete energy. This implies that, during the BinnedShower simulation, each voxel has four corresponding positive sub-voxels that are interpreted as the probability to map the hit to this sub-voxel. For all layers except EMB2, these discrete probabilities are directly used by sampling the sub voxel according to the implicit multinomial distribution. A linear distribution was only used for the radial interpolation of EMB2, in order to smooth the discontinuity at the sub-bin boundary in the radial direction. EMB2's angular interpolation is still using a binomial distribution. A visualization of the default interpolation and the radial interpolation for EMB2 can be seen in Figure 7.

For the voxelization study, the same events were used for binning and for binned shower simulation. This means that step 1) was trivial, as each event in the shower library had exactly one event during the simulation to which it corresponded. This eliminated all effects of global η and ϕ shifts from the binning optimization. In the following this procedure is referred to as *perfect event matching*.

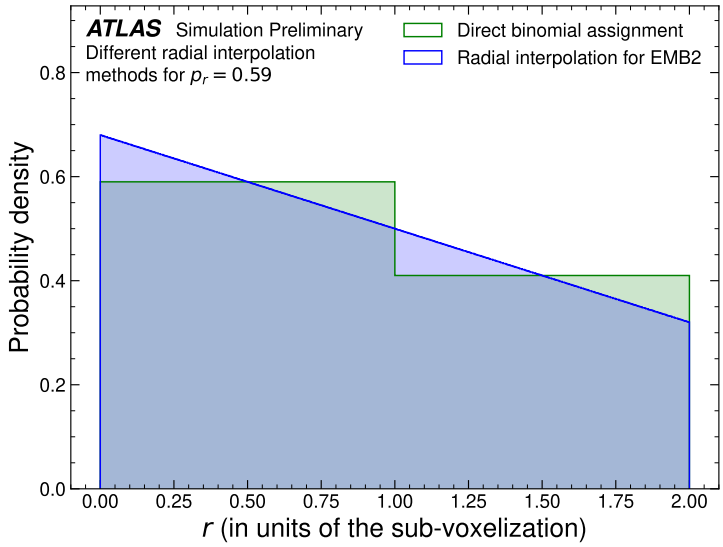


Figure 7: The two interpolation distributions used for radial sub-bin probabilities of $p_r = 59\%$ and $1 - p_r = 41\%$, respectively, which were common for EMB2. The first method assigns the hits to the sub bin using a binomial distribution with p_r underlying probability. The second method uses a linear PDF intersecting the two fixed points of p_r and $1 - p_r$. In this figure $r = 0$ corresponds to the left border of a voxel, $r = 1$ to the voxel-center and $r = 2$ to the right border.

2.5 ϕ modulation

The perfect event matching, as introduced at the end of subsection 2.4, cannot be used when the voxelized samples are generated using machine learning. However, when dropping the exact matching of physics events and binned events, artifacts due to the displacement in η and ϕ can occur.

The artifacts from the pseudorapidity displacement are relatively weak. Furthermore, the pseudorapidity displacement effects increase with the absolute difference in pseudorapidity of the event to be simulated and the event from the shower library. Therefore, the simplest solution is to condition machine learning models on η_{center}, solving the pseudorapidity displacement problems.

However, the effects from the ϕ-displacement are not related to the absolute ϕ difference. Instead, the accordion structure of the electromagnetic barrel of the ATLAS calorimeter creates a periodic modulation of the hit energies, depending on the hits' relative position to the center of its respective cell [6]. The accordion structure is visualized in Figure 8. The modulation has an amplitude of $\approx 10\%$ on the hit level. However, because of its periodic nature, most of this modulation is integrated out on the voxel level. At this resolution, the modulation is almost invisible. The sole remaining effect is the modulation of the total energy, which persists after integration due to the pronounced radial energy decay at the shower core and amounts to no more than $\lesssim 1\%$. The periodic and subtle nature of this effect makes it inherently hard to learn for machine learning techniques after the voxelization. As it is more strongly visible before the voxelization, it was decided to remove the modulation explicitly at this point, eliminating this effect from the dataset.

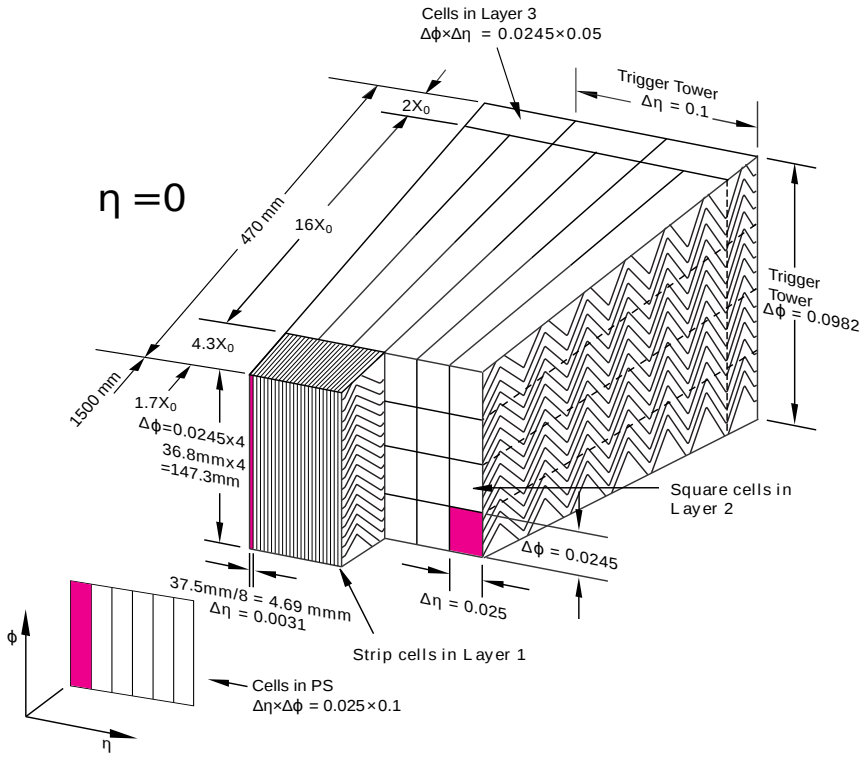


Figure 8: Sketch of a module of the ATLAS electromagnetic barrel. On the right side, the accordion structure is indicated. It is introducing non-uniform electromagnetic fields, that create modulations in the energy response within a cell. The figure was taken from [10].

The actual removal of the phi modulation is based on the sub-cell position of the individual hits $Y = \phi_{\text{hit}} - \phi_{\text{cell center}}(\phi_{\text{hit}})$.

Let $A(x)$ denote the expected hit energy of the average shower's hits at the position $x = \phi_{\text{hit}} - \phi_{\text{shower center}}$. This modulated shower can be factorized into two functions, the unmodulated shower $f(x)$, truly $\phi_{\text{shower center}}$ independent, and the pure phi modulation $g(Y)$.

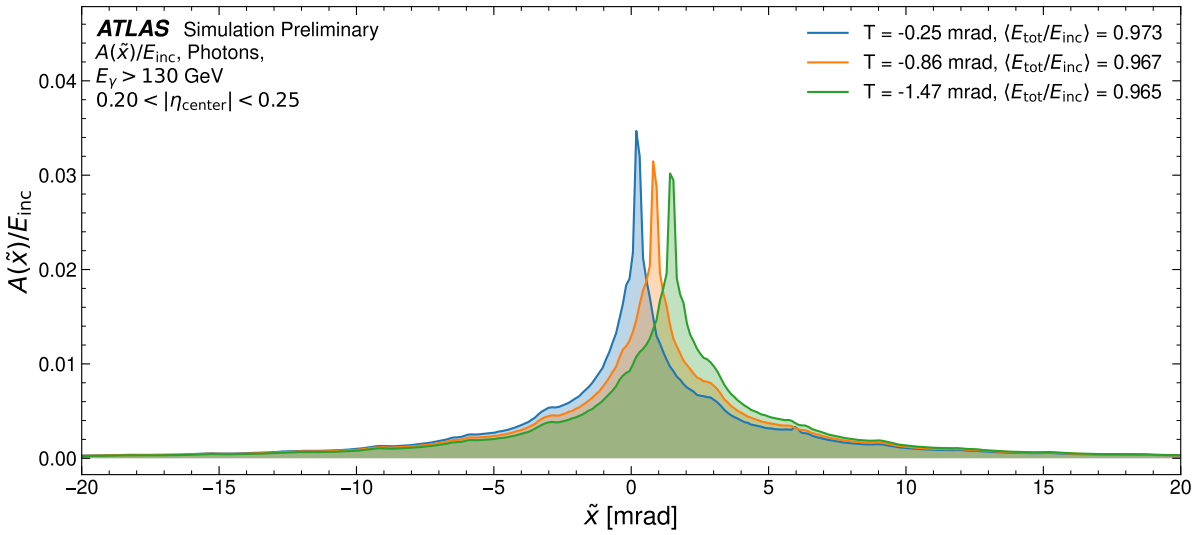
$$A(x, T) = f(x) \cdot g(Y) = f(x) \cdot g(x - T) = f(\tilde{x} + T) \cdot g(\tilde{x})$$

The second equation introduces $T = \phi_{\text{shower center}} - \phi_{\text{cell center}}(\phi_{\text{shower center}})$, the relative shower center position to its corresponding cell center, and uses the fact that g has to be periodic with the cell size in ϕ. The last equation introduces $\tilde{x} = x - T$, which aligns the g-modulations for different values of T. This essentially means that each shower is simply modulated by a periodic function g. Consequently, the whole ϕ-modulation problem boils down to the determination of g. The average shower, dependent on $\tilde{x}$, is visualized for different T bins in Figure 9(a).

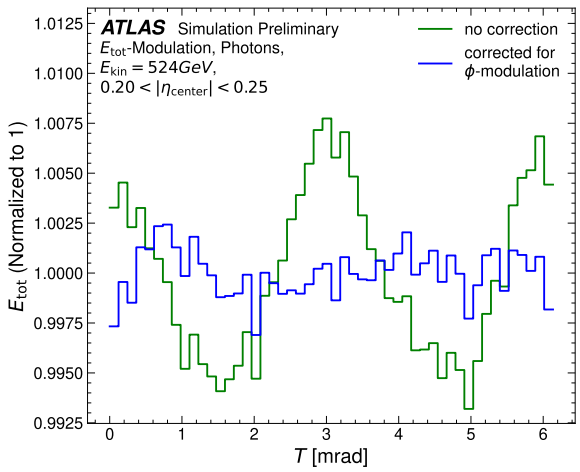
To prevent any transition biases it is desirable to extract the correction curve g directly from the GEANT4 photon simulation. However, the unmodulated photon hit energies are unknown. Therefore, it is impossible to extract the curve g from the data in a straightforward manner. Despite of this, the total energy E_{tot} and the relative shower center position T are enough the *learn* the curve g. Plotting the total energy E_{tot} over T will reveal a small but measurable modulation (c.f. Figure 9(b)). While this curve is obviously not the requested curve g, it is related to it. The amplitude of the total energy modulation $E_{\text{tot}}(T)$ should be zero for a perfect modulation correction, since it is purely introduced by the ϕ-modulation. Consequently, the amplitude of $E_{\text{tot}}(T)$ can be used as a measure that qualifies any applied correction curve $\tilde{g}$. This turns the search for the true g into an optimization problem of a parametric $\tilde{g}$. For this specific case, $\tilde{g}$ was parameterized as a free histogram of 50 bins. The correction curves were trained on a dataset containing all incident energies $E > 130$ GeV featuring a ϕ-binning fine enough to resolve the actual modulation. The lower energies were omitted since the ϕ-modulation only becomes relevant for high incident energies.

Using the ADAM optimizer [11] to minimize the standard deviation $\text{std}(E_{\text{tot}}(T))$ together with a denoising and a symmetrizing loss term, it is straightforward to obtain the correction curves that are depicted Figure 9(c). The corrections affect mainly the first two layers in the electromagnetic barrel. The PreSamplerB (layer 0) has no accordion structure but is apparently affected due to effects like backscattering from the first EMB layer. The correction for the final EMB layer (layer 3) becomes negligible, as the shower progressively broadens while traversing the calorimeter. For layer 3, the modulation becomes therefore weaker, as it is integrated over more oscillations. Additionally, layer 3 receives comparatively minimal energy. As a result, the loss function — and most physical observables — exhibit reduced sensitivity to the ϕ-modulation in layer 3.

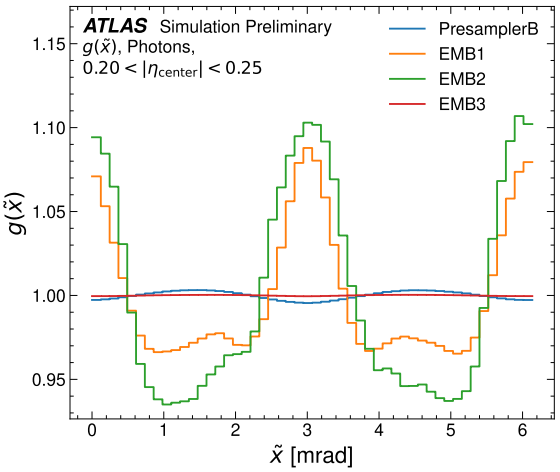
For the voxelized datasets discussed, the ϕ-modulation effects were removed using these extracted correction curves. The ϕ-modulation can therefore be ignored when training generative models on the voxelized datasets. The effects of the ϕ-modulation correction on the reconstructed shower shapes are shown in Figure 10. It should be noted that the correction for the ϕ-modulation is not tied exclusively to the simulation using ML models, but can be used to improve the fast simulation in general.



(a) Average shower $A(\tilde{x})$ for different T bins.

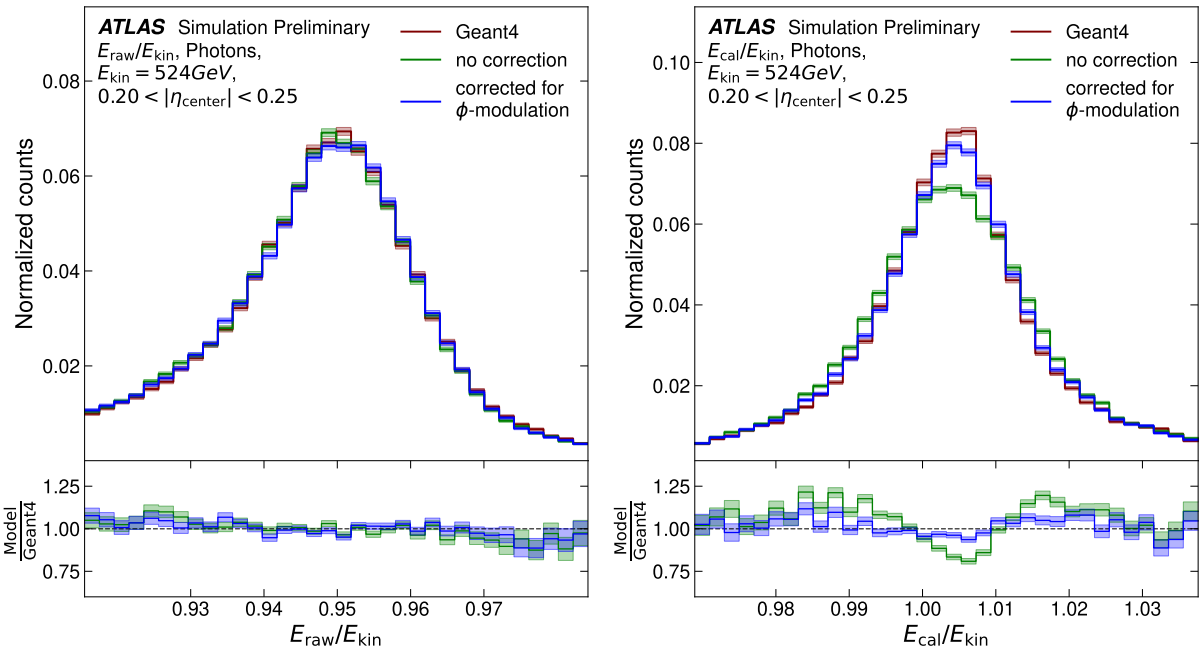


(b) Modulation of the total energy E_{tot} for different T bins.



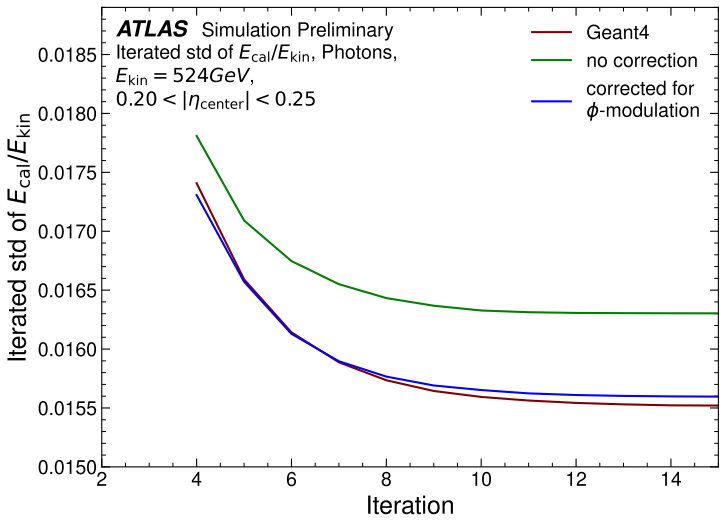
(c) Learned correction curves g as function of $\tilde{x}$.

Figure 9: T is the relative shower center position in ϕ to its corresponding cell center in ϕ. At the top (a), the dependence of the shower energy (A) is plotted for different T bins. As the plot uses $\tilde{x}$ as coordinate, the unmodulated shower f is shifted by T. The peak position therefore indicates the underlying value of T. It is clearly visible that the peak-height decreases with T. Due to the non-linear shape of the distribution, the integral E_{tot} is also affected.
On the bottom left (b), this E_{tot} modulation is shown for all T values (green line). The curves in the bottom right plot (c) show the actual ϕ-modulation $g(\tilde{x})$ on the hit level, as extracted by minimizing the fluctuation of E_{tot}. The remaining modulation of E_{tot} after the correction by the hit level ϕ-modulation is visualized by the blue line in (b).



(a) The effect of the ϕ-modulation correction on the raw energy of the calorimeter E_{raw}.

(b) The effect of the ϕ-modulation correction on the calibrated energy of the calorimeter E_{cal}.



(c) The effect of the ϕ-modulation correction on the width of the calibrated energy of the calorimeter $\sigma(E_{\text{cal}})$.

Figure 10: In the upper row, the raw (a) and calibrated (b) energy for photon showers of 524 GeV are depicted. The lines correspond to the GEANT4 reference curve, the binned shower application without any ϕ-modulation treatment and the ϕ-modulation correction extracted from the photon showers. It can be seen that the learned modulation is able to correct the differences almost perfectly.
In the lower plot (c), the iterated standard deviation of the calibrated energy is visible. This observable is computed by estimating the standard deviation within the 3σ interval of the standard deviation of the previous iteration, starting without any confinement, until convergence. This procedure was used to remove the low statistics in the tails of the distribution. For the iterated standard deviation, one can see that the learned correction improves the width of the calibrated energy significantly.

2.6 New generative models

Currently, two new generative model architectures for calorimeter simulation are being developed in ATLAS – a diffusion transformer [12], based on CaloDiT-2 [3], and a continuous normalizing flow [13], CaloCFM. Both models employ a 2-stage training as originally introduced for CaloFlow [14]. The energy model, in this case a downscaled CaloINN [15], predicts the energies of the relevant calorimeter layers E_{layer} and thus also the total deposited energy in all layers $E_{\text{tot}} = \sum E_{\text{layer}}$. The shape models mentioned above are conditioned on these layer energies and predict the energy distribution within the individual layers. This approach is beneficial as it reduces the total number of parameters, while it increases the accuracy of the layer and total energy at the same time. A visualization of this 2-stage approach is given in Figure 11. All three models (CFM, DiT and INN) were trained on multiple η_{center} bins at once. For the full pseudorapidity range, two different models were trained, respectively. One with $0 < |\eta_{\text{center}}| < 0.7$ and another with $0.7 < |\eta_{\text{center}}| < 1.3$. This split was chosen as the η_{center} dependence spikes around $|\eta_{\text{center}}| \approx 0.8$. At this point, one electrode ends and another begins. This effect results, for example, in a deficit in the measured energy and thus in a strong pseudorapidity correlation (cf. Figure 12). To fully confine this effect within one model, the shared boundary of the model domains was chosen as $|\eta_{\text{center}}| = 0.7$, slightly shifted relative to the electrode gap. All models were trained directly on logarithmic voxel or layer energies, respectively, not including any high-level shower shape observables in the model optimization algorithm. Specifically, the CaloINN uses the typical normalizing flow likelihood-loss [15], while the CaloCFM and the CaloDiT-2 use MSE-based losses [16, 17]. In the following sections, these three model architectures are explained in more detail.

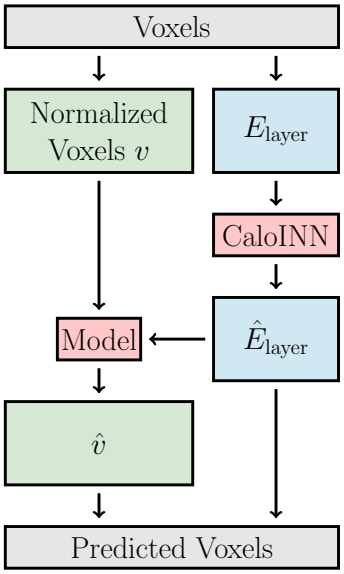


Figure 11: Visualization of the 2-stage setup. The CaloINN is used to predict the integrated energy content of the individual calorimeter layers. This simplifies the task of the shape model, which is only predicting the normalized energy distribution within each layer, not the absolute voxel energies. The variables indicated with the ˆ-symbol are predicted by the individual models.

2.6.1 CaloINN

For the modeling of the layer energies, it was decided to use a normalizing flow, more precisely, the CaloINN as presented in [15]. Normalizing flows are invertible mappings between two probability distributions. In the case at hand, the mapping was performed between a standard Gaussian - the latent space - and the distribution of the individual layer energies. The main CaloINN architecture, its training setup and the layer energy preprocessing was taken from [15]. Table 3 shows the adapted hyper parameters, using the same structure as table 5 in [15].

The center of the shower in the pseudorapidity direction was added as a condition to the model, in addition to the incident energy E_{kin}. It was given by $\eta_{\text{cond}} = 10 \cdot \eta_{\text{center}}$. The factor of 10 was chosen to have $\eta_{\text{cond}} \simeq 1$ as input values near to 1 are numerically better suited for the ML training.

Table 3: Network and training parameters for the CaloINN. Most parameters were taken from [15]. The main differences are the longer training, the larger batch size and the reduced number of neurons per layer. The increase in training time and batch size was needed to improve the pseudorapidity conditioning, while the smaller layer sizes are a result of the small input dimensionality.

Parameter	CaloINN
coupling blocks	RQS
# layers	3
hidden dimension	60
# of bins	10
# of blocks	14
# of epochs	2000
batch size	1024
lr scheduler	one cycle
max. lr	$1 \cdot 10^{-4}$
$\beta_{1,2}$ (ADAM)	$(0.9, 0.999)$
b	$1 \cdot 10^{-6}$
α	$1 \cdot 10^{-6}$

2.6.2 CaloCFM

The first shape model is a continuous normalizing flow (CNF) [13] trained via conditional flow matching (CFM) [18]. The training was inspired by the shape model of CaloDREAM [17], which was a model which performed well during the CaloChallenge [3]. However, for the dataset at hand, it was decided to use a fully connected backbone for the CFN instead of a vision transformer. The primary motivation for this was that this model was applied to the small, irregular dataset, which prevented the original vision transformer approach. The loss function and the linear optimal transport trajectory for the CFM were not changed. The voxel preprocessing was based on a regularized logit, in a similar fashion to the approach used in CaloINN. As was the case for the energy model, a factor of 10^{-6} was chosen for noise and logit regularization.

The fully connected CNF backbone consisted of 4 layers with a hidden size of 1024 neurons. The E_{kin} and $\eta_{\text{cond}} = 10 \cdot \eta_{\text{center}}$ conditions were embedded using two fully connected subnetworks with three layers of 30 internal neurons, that increased their dimensionality to 128 each. For training, layer norm was used to enhance the training stability. The CNF shape model was trained for 160,000 epochs with a one cycle learning rate between 10^{-4} and $3 \cdot 10^{-3}$, and a batch size of 4096. For the CNF model with $|\eta_{\text{center}}| > 0.7$,

the number of neurons was increased to 1500 to capture the difficult pseudorapidity dependence more accurately. With this larger expressivity it also converged faster, and so was only trained for 40,000 epochs.

Within the general CFM MSE-loss, it was decided to weight each voxel with

$$w_i = \frac{\sum_{j \in \text{layer}(i)} 1}{n_{\text{voxels}}}.$$

layer(i) denotes the set of all voxel indices that are in the same layer as voxel i. The reason for this scaling was to prevent the loss from over-prioritizing the layers with more voxels. Otherwise, the non-linear logit-preprocessing would result in exactly this bias.

For the evaluation, the midpoint solver with 5 steps (10 model evaluations) was found to produce good results after reconstruction. The reconstructed shower shapes did not improve by using smaller step sizes.

2.6.3 CaloDiT-2

The second shape model is a diffusion transformer (DiT) model [12], which is trained via a continuous-time diffusion process [16], and subsequently consistency distillation [19]. The CaloDiT-2 model presented here is tailored for use with the regular dataset, as a result of the use of vision-transformer-style patching in the architecture [9]. CaloDiT-2 features significant improvements with respect to the original CaloDiT model from the CaloChallenge [3]. The foremost is the change from a discrete-time (DDPM) diffusion process [20] to a continuous-time (EDM) diffusion process [16, 21]. This allowed a reduction in the number of diffusion steps used during training from 400 to 32, while achieving similar generative fidelity. It also allowed consistency distillation to be used to subsequently reduce the model to a single diffusion step.

The voxel preprocessing consisted of a simple log transformation followed by normalization. The incident energy E_{kin} and pseudorapidity η_{cond} conditions were scaled with the inverse of their corresponding maximum values. As the regular dataset described in subsection 2.2 was used for training CaloDiT-2, each shower image consisted of a regular tensor of size $[r, \alpha, \text{layer}] = [24, 14, 5]$. ViT-like patching [9] was used together with sinusoidal position embeddings, with both being modified to operate in three dimensions. A patch size of $[P_r, P_\alpha, P_{\text{layer}}] = [3, 2, 1]$ was used, resulting in a sequence length of 280. These patches then underwent a linear projection and summation with the positional embeddings. This, together with independent embeddings for the diffusion timestep and the conditions (E_{kin} and η_{cond}), produced the input for the CaloDiT-2 backbone.

The CaloDiT-2 backbone consists of a stack of DiT (adaLN-zero) [12] blocks. The model consists of 4 such blocks, with an embedding dimension of 144 and 8 attention heads. A series of linear projections are then applied to recover patches of the correct size. The patches are then rearranged to recover the original $[24, 14, 5]$ shower image.

The model was trained using an EDM diffusion process [16] with 32 diffusion steps for 800 epochs, with a batch size of 256. The AdamW [22] optimizer was used during training, with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4}.

Consistency distillation [19] was then further used to produce a target model, with weights calculated as an exponential moving average (EMA) of an "online" (i.e trained via minimization of the consistency loss) student model. 32 discretization steps were used for the distillation, with the $l_2(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2^2$ distance

used in the consistency loss computation. The EMA decay was set to 0.95. This distillation procedure was run for 300 epochs.

For evaluation, the distilled target model was sampled with a single diffusion step using the 2^{nd} order Heun ODE solver [16]. By aggressively reducing the number of diffusion steps in this fashion, the conscious choice was made to trade-off some model performance for inference speed.

3 Results

In this subsection, the accuracy of the fast simulation using the binned shower or the two ML models is evaluated. The accuracy of the fast simulation using the binned shower or the two ML models was evaluated by comparing the reconstructed shower shapes to the GEANT4 prediction. For all observables originating from the new voxelization, the ϕ-modulation was removed before the voxelization and reapplied during the simulation. Specifically, this means, that AtlFast3 did not receive the new ϕ-modulation correction. GEANT4 did not need a correction, as it is not introducing shifts in the ϕ-direction. In Figure 12, the predicted total energy of the CaloINN, conditioned on η_{center} is shown. It is able to predict the total energy of the individual events with a high degree of accuracy. In particular, the difficult detector-gap regions are reproduced faithfully. A χ^2 analysis of the individual total energies, as in [1, 2] was performed. The result is $\chi^2 = 1.2$ within the central barrel ($0.2 < |\eta_{\text{center}}| < 0.25$) and $\chi^2 = 2.0$ for the difficult transition region ($0.75 < |\eta_{\text{center}}| < 0.85$). The full χ^2-study can be found in the Appendix in Appendix C.

To assess the quality of the complete fast simulation, several post reconstruction shower shapes [23] are investigated. The considered shower shapes are explained in Table 4.

Table 4: Definition of the investigated shower shapes. E_{ijk} denotes the energy sum of the $j \times k$ cells within layer i, within a $j \times k$ rectangle around the cell with the highest energy. j corresponds to the η-direction and k to the ϕ-direction.

Observable	Description
E_{raw}	The raw sum over all cell energies
E_{cal}	The calibrated cluster energy
f_1	Energy reconstructed in the first layer, divided by the full reconstructed energy
ΔE	Difference between the second 3-cell-wide energy maximum in EMB1 and the cell with the lowest energy that corresponds to either the first or second maximum
E_{ratio}	Ratio between A, the difference between the brightest cell in EMB1 and the second 3-cell-wide energy maximum in EMB1, and B, the sum of these two values
$\omega_{\eta 1}$	η-width of the shower in EMB1, calculated by using the brightest cell in EMB1 and its 3 nearest cells within positive and negative η direction, respectively
fracs_1	$\frac{E_{117}-E_{113}}{E_{113}}$ - Ratio of the outer shower to the inner shower in EMB1
R_ϕ	$\frac{E_{237}}{E_{277}}$ - Ratio, describing the energy decay in ϕ-direction in EMB2
R_η	$\frac{E_{233}}{E_{237}}$ - Ratio, describing the energy decay in η-direction in EMB2

In Figure 13, Figure 14, and Figure 15 the pure binned shower, the CaloCFM, the CaloDiT-2 and the currently used AtlFast3 are compared to the GEANT4 reference in terms of these shower shapes for different pseudorapidity regions. In the investigated regions AtlFast3 used FastCaloSimV2, the classical parametrization. The error bands visualize the expected statistical poissonian error. It should be noted, that AtlFast3 is not using the new voxelization and its input files were created using Run 3 simulation and reconstruction. However, all showers for these figures were created using the current Run 4 configuration. While the corresponding GEANT4 version of the Run 3 and Run 4 setup is identical (c.f. subsection 2.1), small changes in reconstruction can slightly deform some distributions. For the pseudorapidity region of $|\eta_{\text{center}}| < 0.05$, the self-consistent Run 3 AtlFast3 simulation and the corresponding Run 3 GEANT4

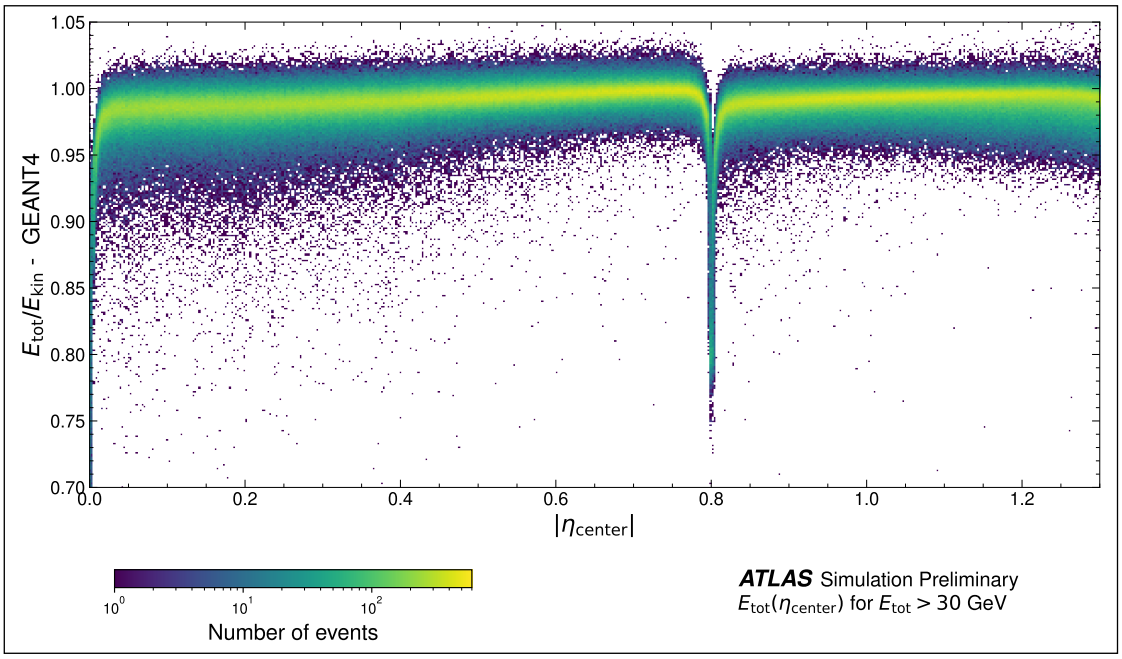
samples were recreated. Within this configuration, only the shift related to the E_{raw} observable vanished. The other effects do not seem to be correlated to this version difference. Therefore, the remaining AtlFast3 artifacts are either related to the imperfect voxelization or the missing hit-correlations. Currently, it is believed that updating AtlFast3 with the new voxelization removes most of the remaining artifacts. Nevertheless, the observed artifacts can be used as a measure of improvement. Furthermore, it should be considered, that the shower shapes were extracted from a simulation without noise and pileup, implying that some of the artifacts might not be visible anymore once these distortions are active.

In Figure 14 the shower shapes for a "simple" mid-barrel region are visualized. The corresponding pseudorapidity region does not feature a strong pseudorapidity dependence and can be seen as representative for the entire dataset except $|\eta_{\text{center}}| < 0.05$ and $0.75 < |\eta_{\text{center}}| < 0.85$. These two regions feature an electrode gap and thus a strong local pseudorapidity dependence. These difficult regions are shown explicitly within Figure 13 and Figure 15, respectively.

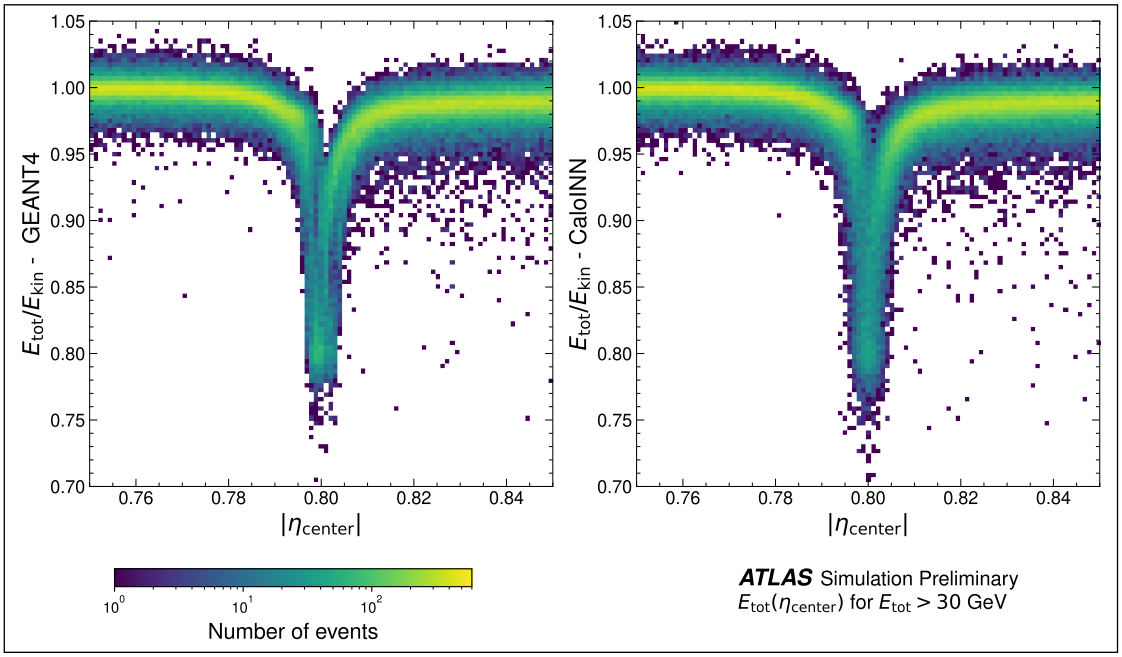
Within the simple detector regions, neither AtlFast3 nor any of the new models show significant problems. AtlFast3 seems to show some problems with ΔE and E_{ratio}, but these effects cannot be observed with detector noise, as low energy tails get washed out. For the difficult pseudorapidity regions (Figure 13 and Figure 15), the benefit of the pseudorapidity-conditioning becomes clearly visible. The calibrated energy E_{cal} of the binned shower and the ML models is clearly sharper than the AtlFast3 counterpart for which a dedicated calibration of this observable is needed. This is expected, as AtlFast3 cannot model this strong pseudorapidity dependence. The other shower shapes in these regions are generally improved compared to AtlFast3 with the new models.

In Figure 16 the shower shapes with $0.75 < |\eta| < 0.8$ are visualized for a significantly higher incident energy of $E_{\text{kin}} = 1$ TeV. For the higher energies where the training statistics fall, CaloDiT-2 appears to introduce some differences, e.g. within the observables R_η or fracs_1. These differences seem to originate from the aggressive single-step distillation process, but are mostly smaller than the AtlFast3 differences. Only for R_ϕ, CaloDiT-2 deviates further from GEANT4. This means, that for the shown version of the CaloDiT-2 simulation accuracy was traded for reduced evaluation time. Improved or less aggressive distillation would be a solution to reduce these differences. Apart from this, it can be seen that the calibrated energy is not reproduced perfectly for high energies in this pseudorapidity region. This is a joint effect of the strong pseudorapitidy dependence together with ϕ-modulation, which is seen here in its worst form. Lastly, the E_{ratio} observable for the *optimal* binning related simulations seems to be slightly worse. The finer regular binning improves the optimal result moderately for this observable. However, it could probably optimized for the *optimal* binning by finetuning the average hit energy E_{hit} and the sub-voxel interpolation procedure, if necessary.

So far, only preliminary memory and timing measurements were performed. Nevertheless, it seems that the required memory of the presented models is slightly smaller than the memory needed to evaluate the FastCaloGANV2. This stems from the fact that at least one GAN was trained for each η-bin of size 0.05. The new models are conditioned on $13 - 14$ of these bins, reducing their effective memory by this factor. The execution time of the FastCaloGANV2 seems to be approximately $2\times$ to $4\times$ faster. This is not expected to be problematic, as the runtime of the GAN was, so far, not limiting the fast simulation. Instead, the slowest component was the GEANT4 based simulation of the inner detector. Some preliminary benchmark measurements are stated in Appendix B.

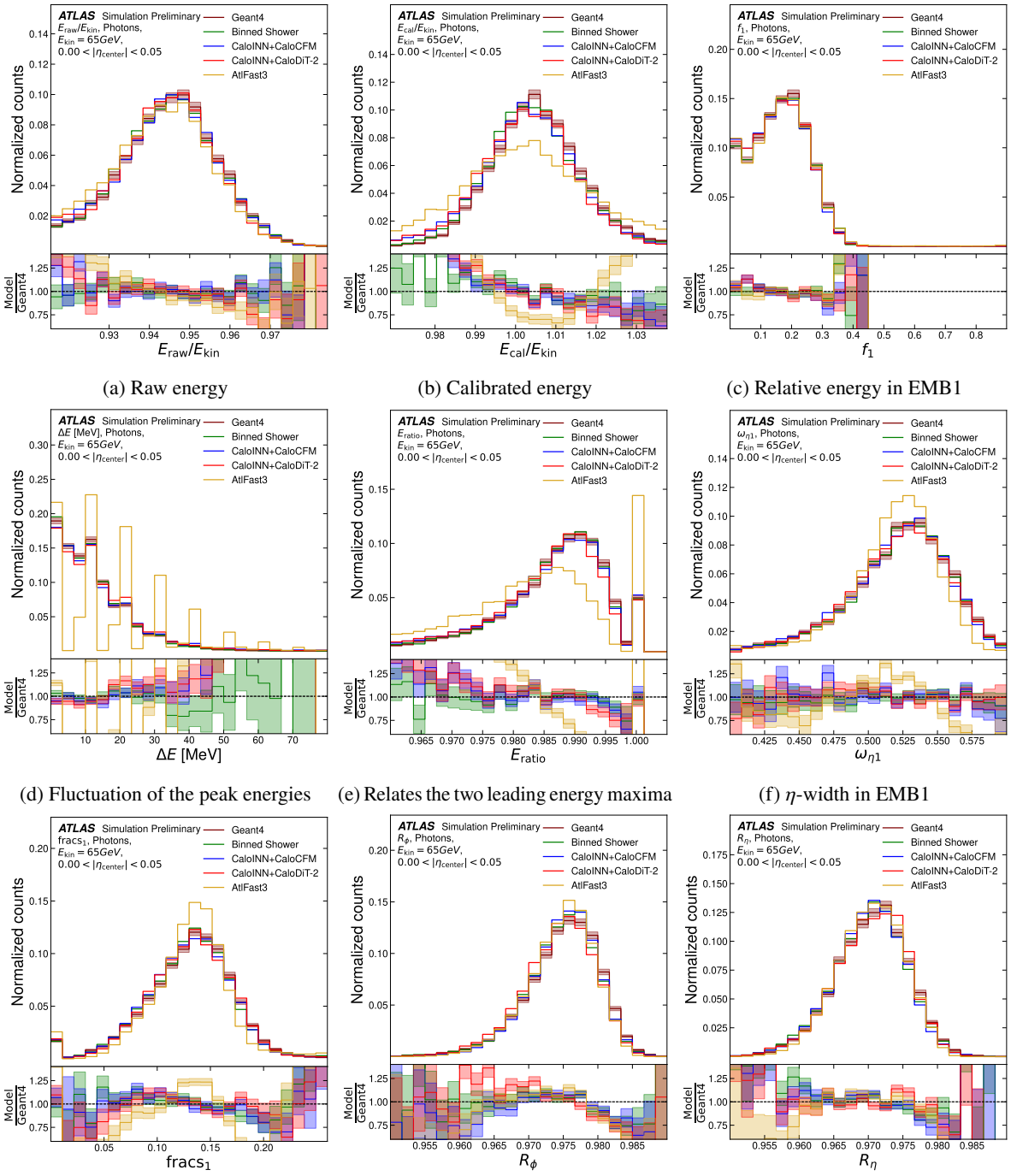


(a) Total energy E_{tot} for $|\eta_{\text{center}}| < 1.3$ as predicted by GEANT4.



(b) Total energy E_{tot} for $0.75 < |\eta_{\text{center}}| < 0.85$ as predicted by GEANT4 compared to the CaloINN modeling.

Figure 12: The distribution of $\frac{E_{\text{tot}}}{E_{\text{kin}}}$ as a function of $|\eta_{\text{center}}|$. The strong η-dependence at the electrode gaps at $|\eta_{\text{center}}| \approx 0.0$ and $|\eta_{\text{center}}| \approx 0.8$ can directly be seen (a). In the bottom row (b), the CaloINN prediction is compared to the GEANT4 reference, within the region $|\eta_{\text{center}}| \approx 0.8$. The CaloINN is able to model the energy dip and the width change of the distribution at $|\eta_{\text{center}}| \approx 0.8$ faithfully - down to the spread of single events. The accuracy for the dip at $|\eta_{\text{center}}| \approx 0.0$ is similar.

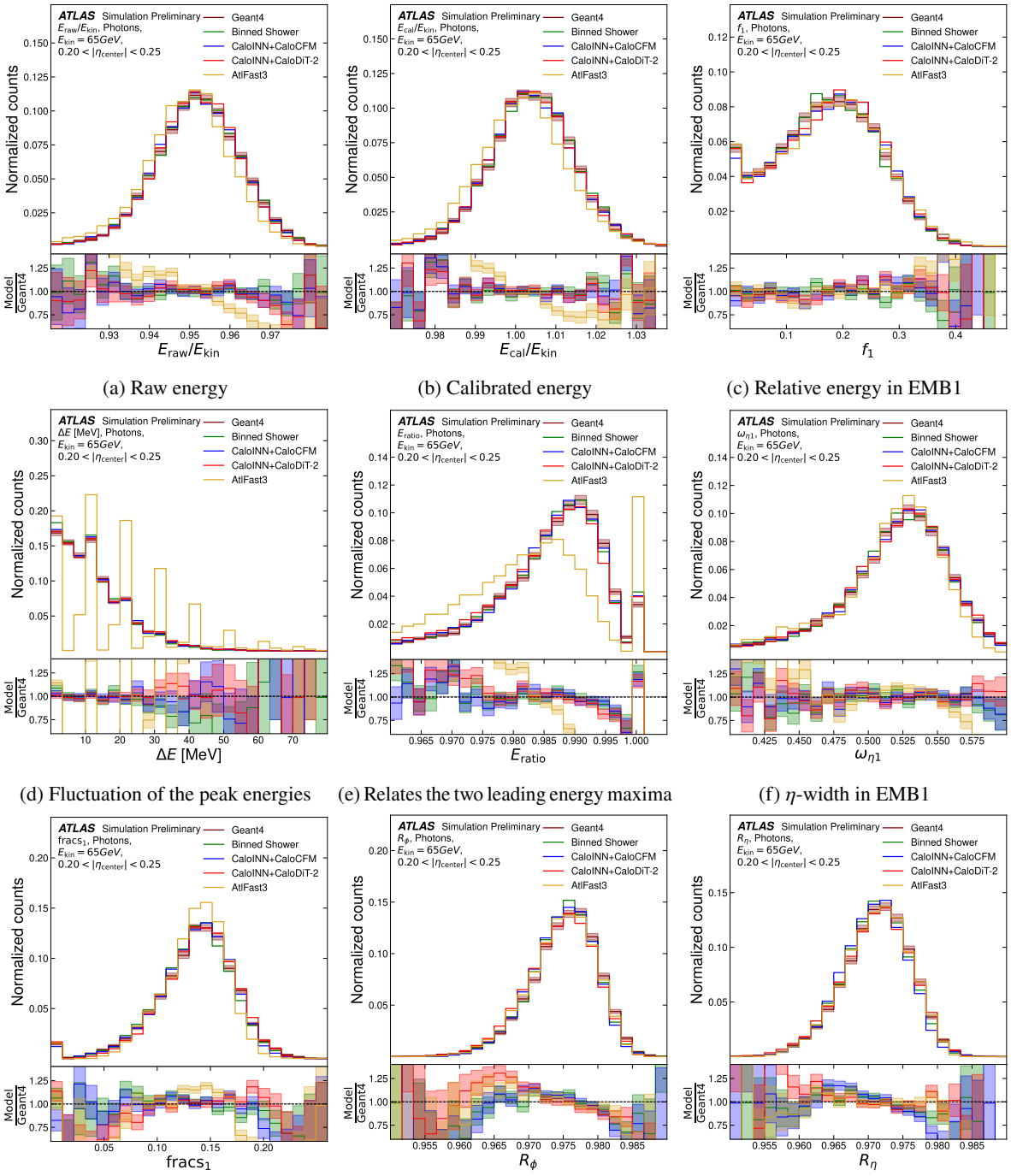


(a) Raw energy (b) Calibrated energy (c) Relative energy in EMB1

(d) Fluctuation of the peak energies (e) Relates the two leading energy maxima (f) η-width in EMB1

(g) Shower decay measure (h) ϕ-width in EMB2 (i) η-width in EMB2

Figure 13: Visualization of the shower shapes of showers with $|\eta_{\text{center}}| < 0.05$. The definition of these shower shapes is given in Table 4. The shower shapes (a), (b) and (c) describe general energy distributions, the shower shapes (d) to (g) are sensitive to the first electromagnetic barrel layer (EMB1) and the shower shapes (h) and (i) focus on the second layer (EMB2). The incident energies of the used events satisfy $\frac{3}{4} \cdot 65$ GeV $< E_{\text{kin}} < \frac{3}{2} \cdot 65$ GeV, where the energy related variables were rescaled to $E_{\text{kin}} = 65$ GeV. The error bands are indicating the expected Poissonian error. As all distribution feature the same number of events, the errors are identical for all displayed lines. To improve readability, only the error bands of GEANT4 are shown.

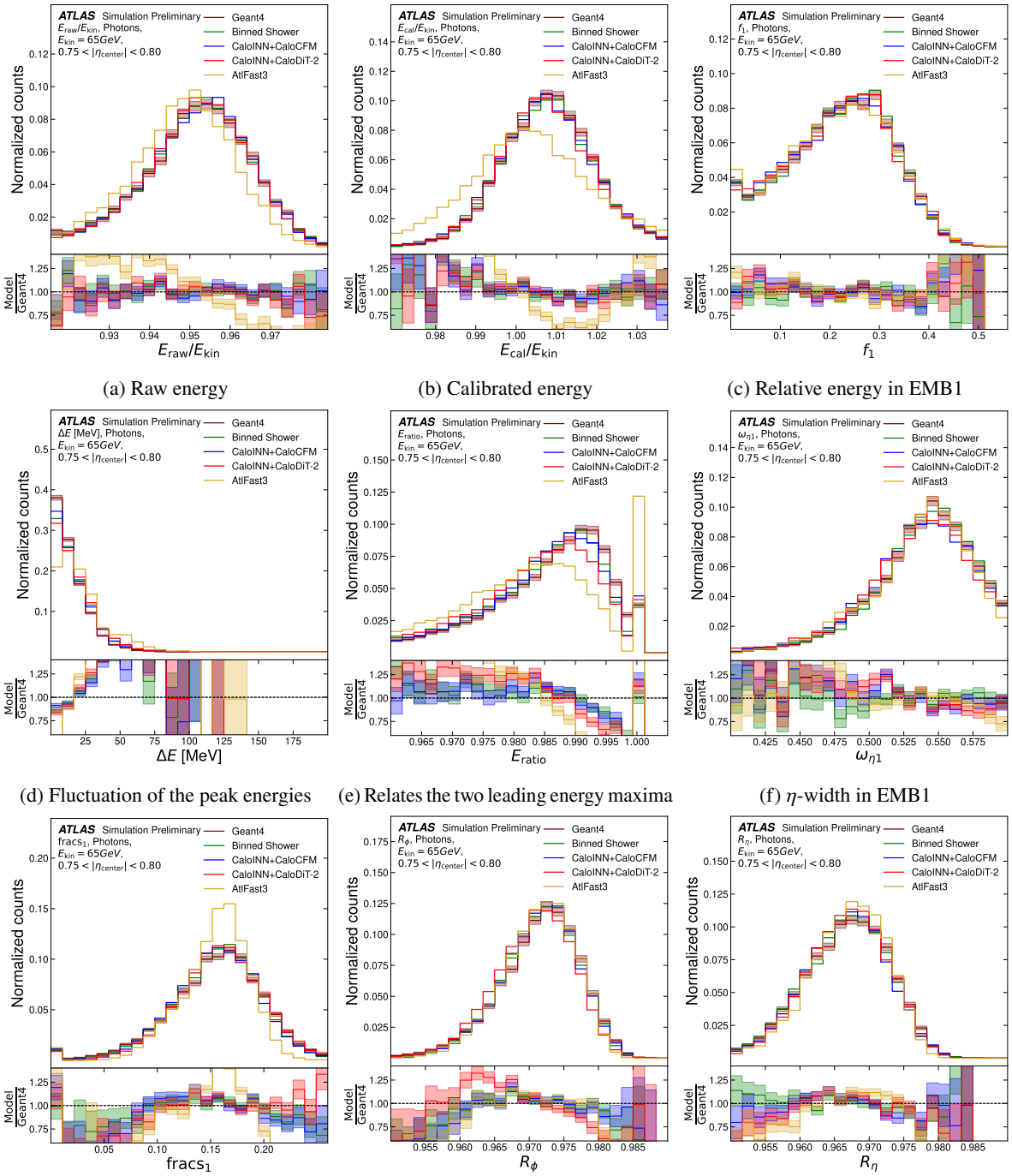


(a) Raw energy (b) Calibrated energy (c) Relative energy in EMB1

(d) Fluctuation of the peak energies (e) Relates the two leading energy maxima (f) η-width in EMB1

(g) Shower decay measure (h) ϕ-width in EMB2 (i) η-width in EMB2

Figure 14: Visualization of the shower shapes of showers with $0.2 < |\eta_{\text{center}}| < 0.25$. The definition of these shower shapes is given in Table 4. The shower shapes (a), (b) and (c) describe general energy distributions, the shower shapes (d) to (g) are sensitive to the first electromagnetic barrel layer (EMB1) and the shower shapes (h) and (i) focus on the second layer (EMB2). The incident energies of the used events satisfy $\frac{3}{4} \cdot 65$ GeV $< E_{\text{kin}} < \frac{3}{2} \cdot 65$ GeV, where the energy related variables were rescaled to $E_{\text{kin}} = 65$ GeV. The error bands are indicating the expected Poissonian error. As all distribution feature the same number of events, the errors are identical for all displayed lines. To improve readability, only the error bands of GEANT4 are shown.

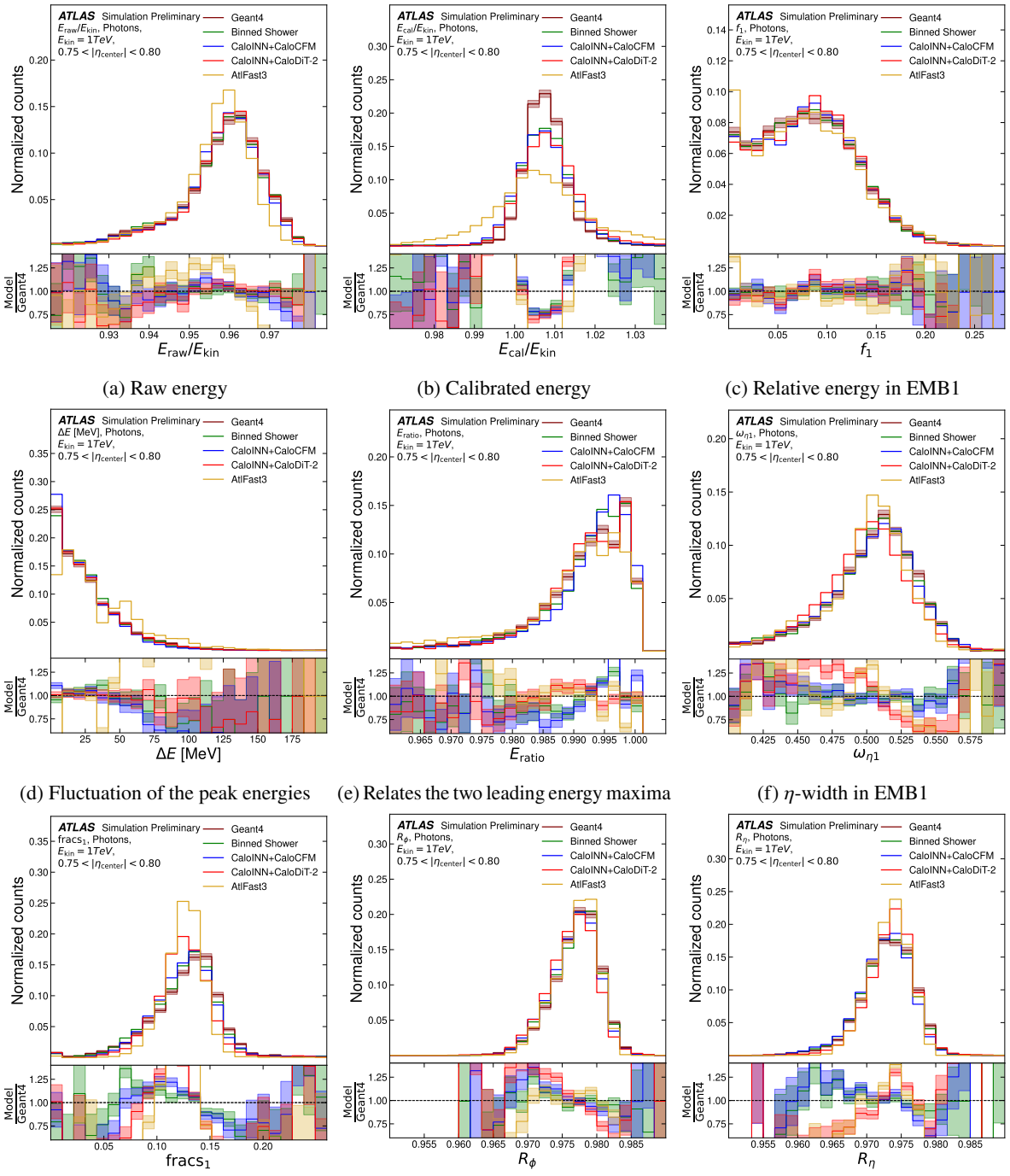


(a) Raw energy (b) Calibrated energy (c) Relative energy in EMB1

(d) Fluctuation of the peak energies (e) Relates the two leading energy maxima (f) η-width in EMB1

(g) Shower decay measure (h) ϕ-width in EMB2 (i) η-width in EMB2

Figure 15: Visualization of the shower shapes of showers with $0.75 < |\eta_{\text{center}}| < 0.8$. The definition of these shower shapes is given in Table 4. The shower shapes (a), (b) and (c) describe general energy distributions, the shower shapes (d) to (g) are sensitive to the first electromagnetic barrel layer (EMB1) and the shower shapes (h) and (i) focus on the second layer (EMB2). The incident energies of the used events satisfy $\frac{3}{4} \cdot 65$ GeV $< E_{\text{kin}} < \frac{3}{2} \cdot 65$ GeV, where the energy related variables were rescaled to $E_{\text{kin}} = 65$ GeV. The error bands are indicating the expected Poissonian error. As all distribution feature the same number of events, the errors are identical for all displayed lines. To improve readability, only the error bands of GEANT4 are shown.



(a) Raw energy (b) Calibrated energy (c) Relative energy in EMB1

(d) Fluctuation of the peak energies (e) Relates the two leading energy maxima (f) η-width in EMB1

(g) Shower decay measure (h) ϕ-width in EMB2 (i) η-width in EMB2

Figure 16: Visualization of the shower shapes of showers with $0.75 < |\eta_{\text{center}}| < 0.8$. The definition of these shower shapes is given in Table 4. The shower shapes (a), (b) and (c) describe general energy distributions, the shower shapes (d) to (g) are sensitive to the first electromagnetic barrel layer (EMB1) and the shower shapes (h) and (i) focus on the second layer (EMB2). The incident energies of the used events satisfy $\frac{3}{4} \cdot 1\text{ TeV} < E_{\text{kin}} < \frac{3}{2} \cdot 1\text{ TeV}$, where the energy related variables were rescaled to $E_{\text{kin}} = 1$ TeV. These histograms feature the biggest differences that were observed after investigating all reconstructed shower shapes in the barrel. The error bands are indicating the expected Poissonian error. As all distribution feature the same number of events, the errors are identical for all displayed lines. To improve readability, only the error bands of GEANT4 are shown.

4 Conclusion

An approach has been developed to test what information loss is present for different voxelizations of the calorimeter shower. It has been shown that showers in the electromagnetic barrel can be simulated well with 382 voxels that are appropriately arranged through the calorimeter layers. This is an important step towards a highly accurate fast calorimeter simulation for Run 4, as it prevents generative machine learning models from being limited in their accuracy by the voxelization of the training dataset.

Additionally, it has been shown that current state-of-the-art ML architectures like (continuous) normalizing flows and diffusion transformers can reproduce the voxelized dataset with great precision, solving this memory problem. For the observed shower shapes, the reconstructed shower shapes obtained by using new generative machine learning models agree broadly within their expected statistical errors. This means that calorimeter simulations using current state-of-the-art generative machine learning are not just comparable to the AtlFast3 baseline, but clearly surpass it for photons in the barrel.

This represents a new milestone in the usage of ML for the ATLAS calorimeter simulation. The models are also comparable to the AtlFast3 simulation in terms of speed and memory. Future work could focus on improving the distillation applied to CaloDiT-2, and potentially explore a version without patching applied to the smaller, irregular binning.

The next step is to extend the reoptimized training dataset to the full ATLAS calorimeter, $0.0 < |\eta_{\text{center}}| < 5.0$, training further ML models on the extended dataset. This work should focus on combining as many pseudorapidity bins per model as possible. Afterwards, the binning reoptimization will have to be performed for pions and electrons. The resulting datasets are expected to be public as well.

References

[1] ATLAS collaboration, *AtlFast3: The Next Generation of Fast Simulation in ATLAS*, Computing and Software for Big Science **6** (2022), ISSN: 2510-2044, URL: http://dx.doi.org/10.1007/s41781-021-00079-7 (cit. on pp. 2, 3, 5, 21, 30).

[2] ATLAS collaboration, *Software and computing for Run 3 of the ATLAS experiment at the LHC*, Eur. Phys. J. C **85** (2025) 234, arXiv: 2404.06335 [hep-ex] (cit. on pp. 2, 21, 30, 32).

[3] O. Amram et al., *CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation*, (2024), ed. by C. Krause et al., arXiv: 2410.21611 [physics.ins-det] (cit. on pp. 2, 17–19).

[4] S. Agostinelli et al., *GEANT4 - A Simulation Toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250 (cit. on p. 3).

[5] J. Allison et al., *Recent developments in Geant4*, Nucl. Instrum. Meth. A **835** (2016) 186 (cit. on p. 3).

[6] ATLAS collaboration, *ATLAS liquid-argon calorimeter*, Technical design report. (1996), URL: https://cds.cern.ch/record/331061 (cit. on pp. 3, 11, 13).

[7] ATLAS collaboration, *ATLAS tile calorimeter*, Technical design report. (1996), URL: https://cds.cern.ch/record/331062 (cit. on p. 3).

[8] P. Gessinger et al., *The Acts project: track reconstruction software for HL-LHC and beyond*, EPJ Web Conf. **245** (2020) 10003, URL: https://cds.cern.ch/record/2752944 (cit. on p. 4).

[9] A. Dosovitskiy et al.,
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021,
arXiv: 2010.11929 [cs.CV], URL: https://arxiv.org/abs/2010.11929 (cit. on pp. 6, 19).

[10] N. Nikiforou, *Performance of the ATLAS Liquid Argon Calorimeter after three years of LHC operation and plans for a future upgrade*, 2013, arXiv: 1306.6756 [physics.ins-det],
URL: https://arxiv.org/abs/1306.6756 (cit. on p. 13).

[11] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017,
arXiv: 1412.6980 [cs.LG], URL: https://arxiv.org/abs/1412.6980 (cit. on p. 14).

[12] W. Peebles and S. Xie, *Scalable Diffusion Models with Transformers*, 2023,
arXiv: 2212.09748 [cs.CV], URL: https://arxiv.org/abs/2212.09748 (cit. on pp. 17, 19).

[13] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud,
Neural Ordinary Differential Equations, 2019, arXiv: 1806.07366 [cs.LG],
URL: https://arxiv.org/abs/1806.07366 (cit. on pp. 17, 18).

[14] C. Krause and D. Shih,
Fast and accurate simulations of calorimeter showers with normalizing flows,
Phys. Rev. D **107** (2023) 113003, arXiv: 2106.05285 [physics.ins-det] (cit. on p. 17).

[15] F. Ernst, L. Favaro, C. Krause, T. Plehn, and D. Shih,
Normalizing Flows for High-Dimensional Detector Simulations, SciPost Phys. **18** (2025) 081,
arXiv: 2312.09290 [hep-ph] (cit. on pp. 17, 18).

[16] T. Karras, M. Aittala, T. Aila, and S. Laine,
Elucidating the Design Space of Diffusion-Based Generative Models, 2022,
arXiv: 2206.00364 [cs.CV] (cit. on pp. 17, 19, 20).

[17] L. Favaro, A. Ore, S. P. Schweitzer, and T. Plehn,
CaloDREAM – Detector Response Emulation via Attentive flow Matching,
SciPost Phys. **18** (2025) 088, arXiv: 2405.09629 [hep-ph] (cit. on pp. 17, 18).

[18] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le,
Flow Matching for Generative Modeling, 2023, arXiv: 2210.02747 [cs.LG],
URL: https://arxiv.org/abs/2210.02747 (cit. on p. 18).

[19] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, *Consistency Models*, 2023,
arXiv: 2303.01469 [cs.LG], URL: https://arxiv.org/abs/2303.01469 (cit. on p. 19).

[20] J. Ho, A. Jain, and P. Abbeel, *Denoising Diffusion Probabilistic Models*, 2020,
arXiv: 2006.11239 [cs.LG], URL: https://arxiv.org/abs/2006.11239 (cit. on p. 19).

[21] Y. Song et al., *Score-Based Generative Modeling through Stochastic Differential Equations*, 2021,
arXiv: 2011.13456 [cs.LG], URL: https://arxiv.org/abs/2011.13456 (cit. on p. 19).

[22] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, 2019,
arXiv: 1711.05101 [cs.LG], URL: https://arxiv.org/abs/1711.05101 (cit. on p. 19).

[23] ATLAS collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data*,
Journal of Instrumentation **14** (2019) P12006, ISSN: 1748-0221,
URL: http://dx.doi.org/10.1088/1748-0221/14/12/P12006 (cit. on p. 21).

Appendices

A HDF5 Keys

Table A.1: The dataset keys of the HDF5 files, corresponding to the variable names used in the text. $\vec{v}_j$ denotes the list of the voxels in layer j. j represents the numerical layer index according to Figure 1. $\vec{\alpha}_{0,j}$ and $\vec{r}_{0,j}$ describe the starting points in α and r direction, respectively, of the voxels $\vec{v}_j$, while $\vec{\Delta\alpha}_j$ and $\vec{\Delta r}_j$ describe the corresponding bin sizes.

Variable	HDF5 dataset-name
E_{kin}	incident_energy
η_{center}	eta_center
$\vec{v}_j$	energy_layer_j
$\vec{\alpha}_{0,j}$	binstart_alpha_layer_j
$\vec{r}_{0,j}$	binstart_radius_layer_j
$\vec{\Delta\alpha}_j$	binsize_alpha_layer_j
$\vec{\Delta r}_j$	binsize_radius_layer_j

B Inference Benchmarks

Preliminary inference benchmarks for each of the models investigated were performed. The execution times were measured on a single core of an AMD EPYC 9654 CPU for a batch size of one. The time shown is the average taken over 10 repeats of 100 model passes, with the error estimated from the standard deviation. Model inference was performed with ONNX Runtime in a pythonic environment. Memory consumption was measured after the model was loaded with ONNX Runtime in a C++ environment.

Table B.1: Inference benchmarking results for each of the models investigated, including number of parameters, memory consumption and time per event.

	CaloINN + CaloCFM (small/large)	CaloINN + CaloDiT-2
No. Params	106098 + 3151230 / 6043882	106098 + 2001958
Memory Consumption (MB)	75 / 176	59
Time / Event (ms)	4.87 ± 0.05 / 9.62 ± 0.05	12.75 ± 0.05

C χ^2 evaluation of the CaloINN

For the total energy, as predicted by the CaloINN, a χ^2 test, as in [1, 2] was performed. The full test, including the plots for several investigated energy slices, are shown in Figure C.1 and Figure C.2. For the central barrel region ($0.2 < |\eta_{\text{center}}| < 0.25$), the average χ^2 values scatter around 1.2. For for the difficult

transition region ($0.75 < |\eta_{\text{center}}| < 0.85$), the model prediction is slightly worse with a reduced $\chi^2 = 2.0$.

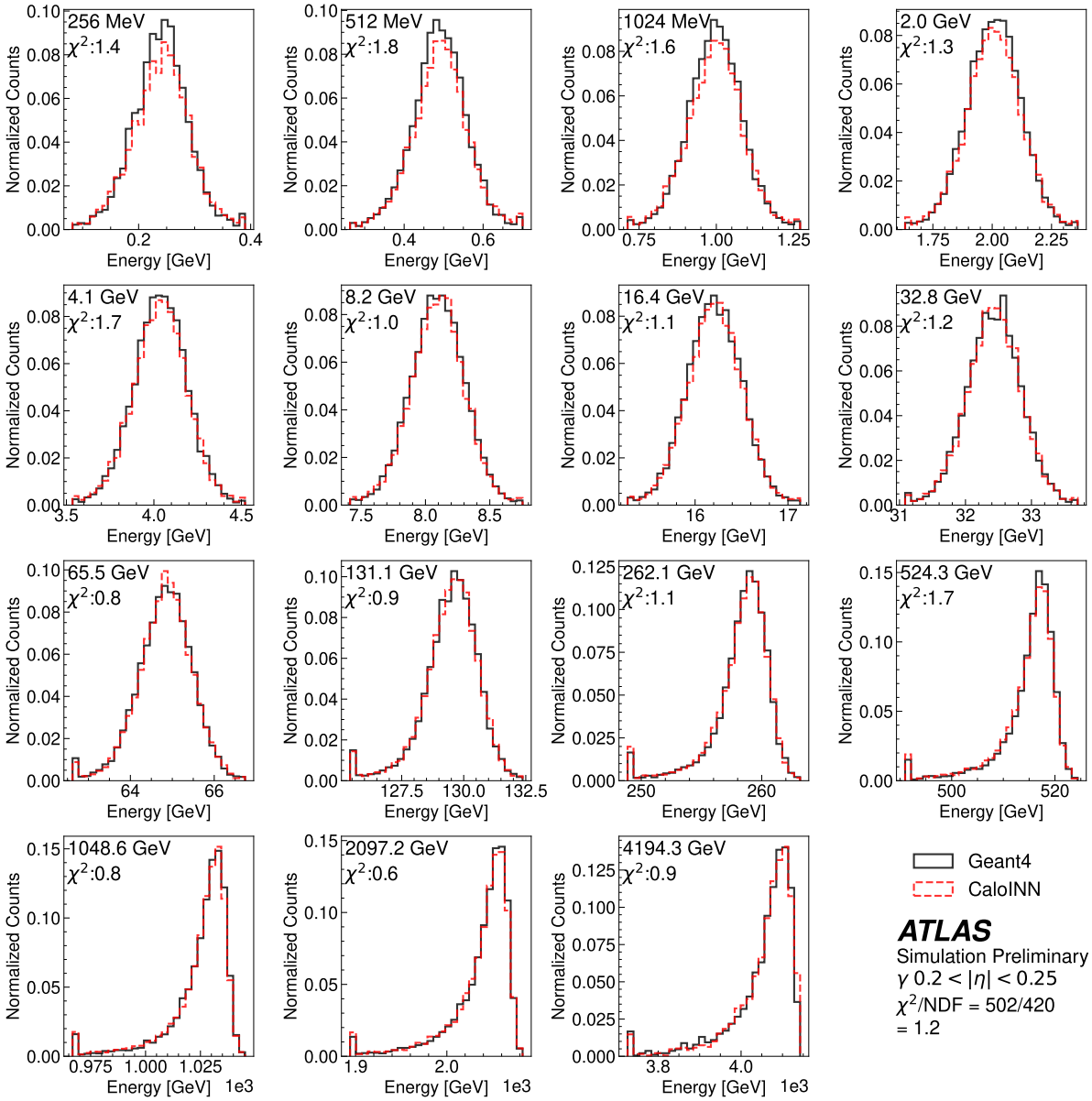


Figure C.1: χ^2-comparison of the CaloINN predictions of the total deposited photon energy (red line) to the GEANT4 reference (solid black line). All showers in this figure are from the simple central barrel region $0.2 < |\eta_{\text{center}}| < 0.25$, in which also the FastCaloGANV2 was evaluated in [2]. The energy prediction for low incident energies ($E_{\text{kin}} \leq 4.1$ GeV) feature slight deviations, while the higher energies are effectively perfectly modeled. Nevertheless, the energy modeling is a clear improvement over the FastCaloGANV2 in [2].

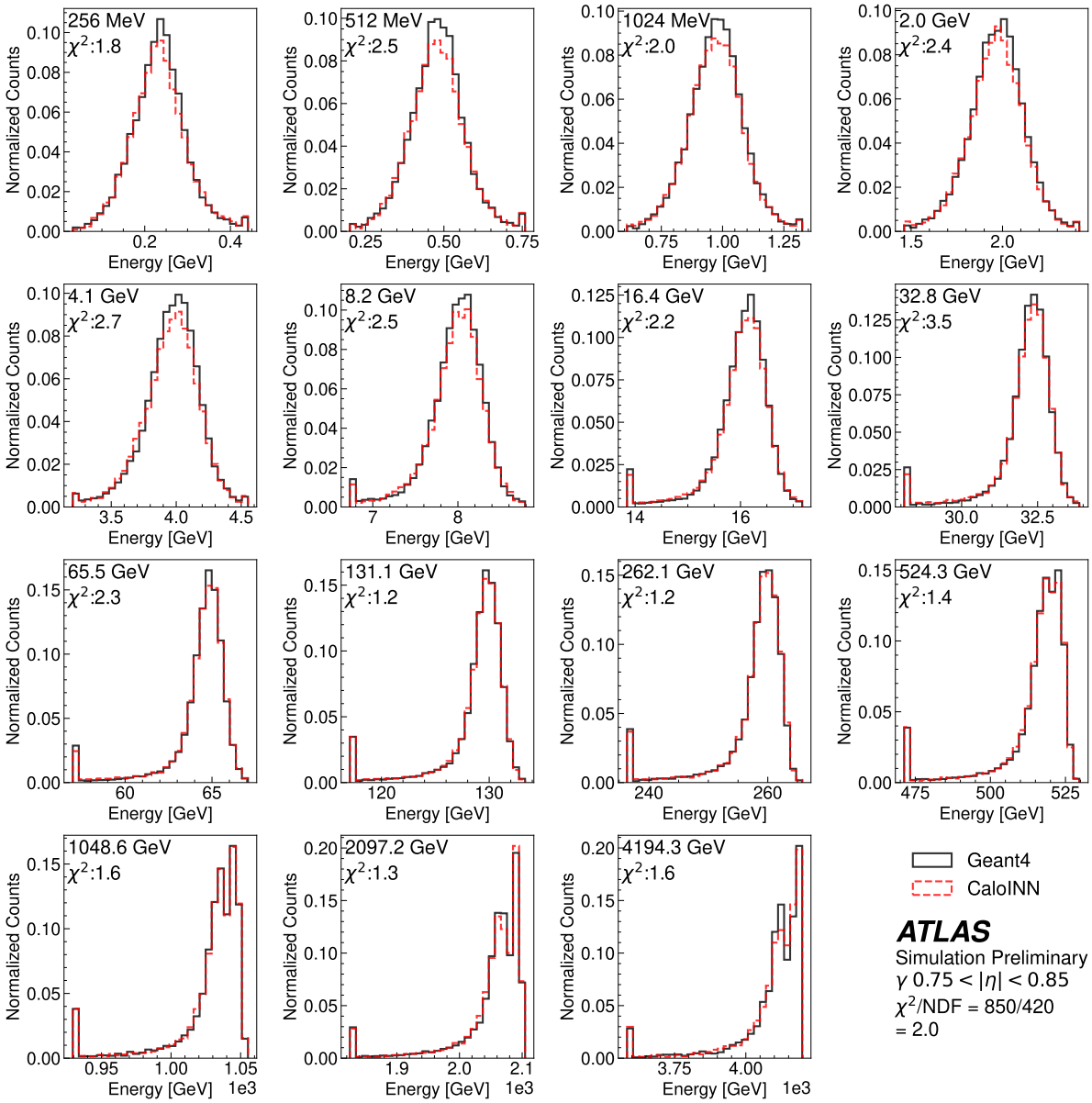
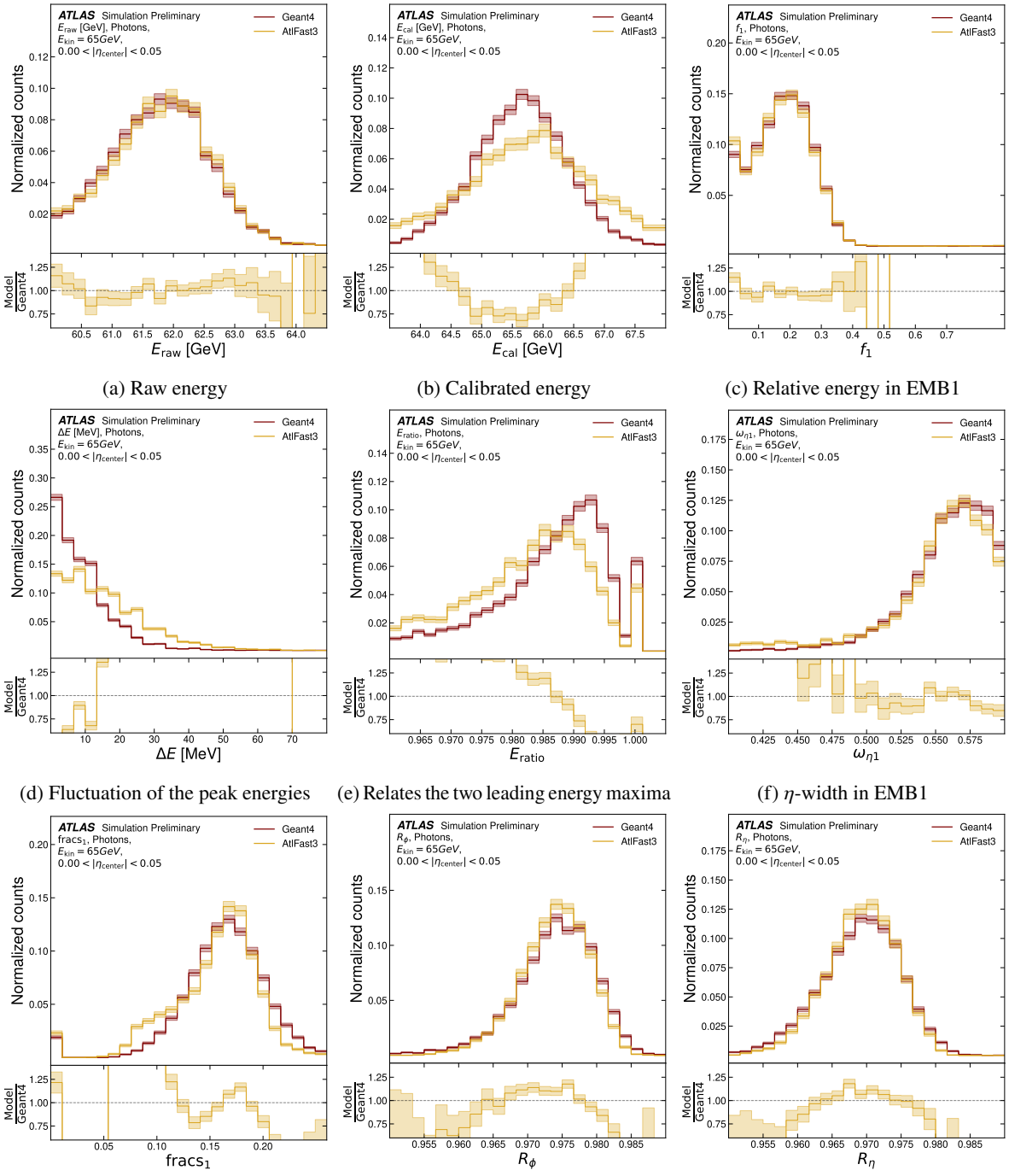


Figure C.2: χ^2-comparison of the CaloINN predictions of the total deposited photon energy (red line) to the GEANT4 reference (solid black line). All showers in this figure are from the difficult transition region $0.75 < |\eta_{\text{center}}| < 0.85$. For this region, no FastCaloGAN evaluations were performed. In this detector region, the energies are modeled generally slightly worse than in the central barrel, not featuring a clear trend anymore.

D Reproduction of Run 3 AtlFast3 plots

In Figure D.1, reproduced shower shapes are shown, that correspond to the same simulation and reconstruction release version for both, the full GEANT4 and the fast AtlFast3 simulation. It is visible, that the shift within E_{raw}, vanishes compared to Figure 13, while the artifacts related to the other observables persist. Therefore, it seems, that the observed differences in the reconstructed shower shapes are mainly unrelated to the different Run 3 reconstruction algorithm used to generate the AtlFast3 input samples.



(a) Raw energy (b) Calibrated energy (c) Relative energy in EMB1

(d) Fluctuation of the peak energies (e) Relates the two leading energy maxima (f) η-width in EMB1

(g) Shower decay measure (h) ϕ-width in EMB2 (i) η-width in EMB2

Figure D.1: Visualization of the shower shapes of showers with $|\eta_{\text{center}}| < 0.05$, when using the same simulation and reconstruction version for the GEANT4 samples and the AtlFast3 execution. The definition of these shower shapes is given in Table 4. The shower shapes (a), (b) and (c) describe general energy distributions, the shower shapes (d) to (g) are sensitive to the first electromagnetic barrel layer (EMB1) and the shower shapes (h) and (i) focus on the second layer (EMB2). These plots were created by using discrete incident energies. Therefore, no energy scaling was applied. The error bands are indicating the expected Poissonian error.