Karlsruhe Institute of Technology

ETP-KA/2022-18

Photon Reconstruction in the Belle II Calorimeter Using Graph Neural Networks

Florian Jochen Wemmer

Masterthesis

11th November 2022

Institute of Experimental Particle Physics (ETP)

Advisor:	Prof. Dr. Torben Ferber
Coadvisor:	Prof. Dr. Markus Klute

Editing time: 31st October 2021 – 11th November 2022

ETP-KA/2022-18

Photon Rekonstruktion im Belle II Kalorimeter mit Graph Neural Networks

Florian Jochen Wemmer

Masterarbeit

11. November 2022

Institut für Experimentelle Teilchenphysik (ETP)

Referent:	Prof. Dr. Torben Ferber
Korreferent:	Prof. Dr. Markus Klute

Bearbeitungszeit: 31. Oktober 2021 – 11. November 2022

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

Karlsruhe, den 11. November 2022

. . . .
Florian Jochen Wemmer

Abstract

This thesis presents the implementation and performance of the GravNet algorithm for the photon energy reconstruction in the Belle II electromagnetic calorimeter. GravNet is a machine learning algorithm based on the concept of graph neural networks. The Belle II Analysis Software Framework is the currently used reconstruction framework that serves as the baseline for comparison in several studies. GravNet solves many of the conceptual restrictions that limit the performance of the traditional reconstruction approach, especially in the presence of high levels of beam background. The studies in this thesis are considered a first validation and are exclusively based on Monte Carlo generated and simulated data. The GravNet implementation outperforms the baseline energy resolutions over a large range of photon energies from 0.01GeV to 3.0 GeV by up to 20 %. In addition, the studies demonstrate substantial improvements of up to 15 % in the reconstruction of neutral pions from the invariant mass of two-photon systems. GravNet proves to be a viable and versatile reconstruction algorithm with a promising outlook for a broad range of present and future applications.

Disclaimer

The Belle II Analysis Software Framework by the Belle II Framework Software Group [1] is used for the event generation of all events in this thesis and is responsible for the Monte Carlo generation and simulation. Beam background overlay file are provided centrally by the data production gorup of the Belle II experiment. I designed and implemented the custom module for crystal-wise Monte Carlo truth matching that enables the event selection of specific cluster signatures. Qasim et al. [2] proposed the GravNet architecture with an application to a toy model calorimeter. I reimplemented the GravNet architecture for the application in the Belle II calorimeter, designed the training pipeline and carried out the optimizations for all models. I developed the metrics used for the studies in this thesis with the exception of the Fuzzy Clustering Agreement Index by Rabbany and Zaïane [3]. My supervisor Prof. Torben Ferber proposed the studies. I produced all results and corresponding plots for the studies in this thesis with the `Matplotlib` package [4] unless the plots are labeled otherwise.

Contents

1. Introduction

A wise man by the name of Erwin Schrödinger, who made a great impact in the world of physics, once said something along the lines of: A purely rational world without mystery is absurd. Much in his spirit, the currently most successful theory of physics, the Standard Model, leaves nowadays particle physicists with many mysteries to understand and discover [5].

The Belle II experiment searches for new physics by measuring rare decays with the Belle II detector at the SuperKEKB electron-positron collider in Tsukuba, Japan. The SuperKEKB accelerator is tuned to the $\Upsilon(4S)$ resonance that provides a clean experimental setup in the decay of pairs of B-Mesons [6, 7]. In order to collect the amount of data required for high-precision measurements, the accelerator reached a world-record instantaneous luminosity as of June 2022. The development of SuperKEKB aims to further increase the luminosity by an order of magnitude within the next years.

Numerous physics analyses require excellent performance for the photon energy reconstruction. Increased beam background, induced by the extreme luminosity, poses a major challenge to the energy reconstruction processes at Belle II [6]. The detector received several upgrades to achieve a reconstruction performance in the new experimental setup that is on par or better in comparison to the preceding Belle detector. These upgrades consist of changes to the detector hardware, as well as to the algorithms used in the reconstruction software [7].

This thesis presents the implementation and performance of the GravNet algorithm for the photon energy reconstruction in the Belle II electromagnetic calorimeter. GravNet [2] is a machine learning algorithm based on the concept of graph neural networks. It is characterized by two representation spaces that allow for the end-to-end learning of complex detector geometries and input features. This machine learning approach mitigates conceptual limits of the currently used, traditional energy reconstruction which arise in the presence of high levels of background. By using the GravNet algorithm, this thesis improves the energy resolution for photons in present and future levels of background relative to the current reconstruction algorithm.

Chapter 2 starts with an overview of the Belle II experimental setup focusing on the electromagnetic calorimeter and beam background processes. Subsequently, the Belle II Analysis Software Framework, which serves as the baseline algorithm, is introduced. Chapter 3 describes event generation and selection that set the general conditions for the training of the machine learning algorithm. This thesis exclusively uses Monte Carlo generated and simulated data. Chapter 4 presents the GravNet algorithm and its implementation for the application to the Belle II electromagnetic calorimeter. Chapter 5 defines metrics that characterize and quantify the reconstruction performance for the following studies. Chapter 6 studies the behavior and performance of GravNet for the reconstruction of photon energies in several well-controlled scenarios. Chapter 6 brings GravNet to application for the reconstruction of the neutral pion mass from a two-photon system. Lastly, chapter 8 gives an outlook on possible directions for future work, before chapter 9 summarizes all presented results.

2. The Belle II Experiment

The Belle II experiment is a high-intensity electron-positron collision experiment hosted at the Japanese High-Energy Accelerator Research Organisation (KEK) in Tsukuba, Japan. The pursued physics program ranges from high-precision measurements of rare decays and the flavour sector to Dark Sector physics. The Belle II detector and SuperKEKB collider are the successors to Belle and KEKB respectively. The upgrade aims to collect 50 times the integrated luminosity for the search of new physics by the means of increased instantaneous luminosity and improved data-taking capabilities [6].

This chapter gives an introduction to the SuperKEKB collider and the Belle II detector in section 2.1. The goal of the GravNet algorithm proposed in this work is to improve the energy reconstruction in the Belle II electromagnetic calorimeter. Section 2.2 describes this sub-detector and the processes taking place in it in more detail. Finally, section 2.3 outlines the currently used reconstruction framework that serves as the baseline algorithm for the studies in this work.

2.1. SuperKEKB and the Belle II Detector

The following section is based on [6, 7]. The SuperKEKB collider is a so-called B-factory mainly operating at the $\Upsilon(4S)$ resonance of 10.58 GeV. The collision of 7 GeV electrons with 4 GeV positrons produces $B\overline{B}$ pairs in an exceptionally clean experimental environment. A preceding linear accelerator supplies electrons and positrons. The main accelerator has a circumference of about 3 km and consists of two rings, one for electrons and the other for positrons. The electron ring is referred to as high-energy ring (HER), and the positron ring as low-energy ring (LER). The two beams are designed to collide at the interaction point (IP) where the Belle II detector measures and records the particle reactions caused by the collision. Due to the asymmetric energies, the collision products receive significant boost in the laboratory frame. Figure 2.1 displays a schematic overview of the main accelerator.

The Belle II detector is a 4π general purpose detector comprised of seven sub-detectors. To achieve the largest possible angle coverage for the products of the asymmetric collisions, the sub-detectors are arranged around the IP in layers and are themselves asymmetric. The barrel, the forward endcap, and the backward endcap make up the full detector. Figure 2.2 shows a cut-through of the detector including concise sub-detector information. Vertex detectors and central drift chamber allow for the reconstruction of the tracks of charged particles, as well as precise decay vertex reconstruction. The two types of particle identification detectors discriminate particles in the barrel and forward endcap. The K_L^0 and muon detector identifies minimal-ionizing charged particles. The energy reconstruction of particles takes place in the electromagnetic calorimeter and is the main focus of this work. Section 2.2 presents the Belle II electromagnetic calorimeter.

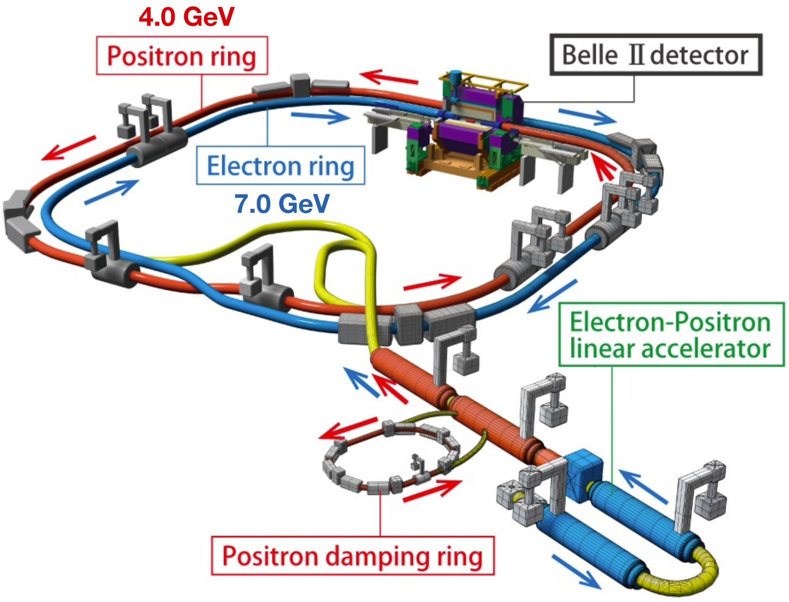


Figure 2.1.: Schematic overview of the SuperKEKB main accelerator ring and preceding linear accelerator. The main ring has a circumference of about 3 km. The interaction point is in the center of the Belle II detector. Figure adapted from [8].

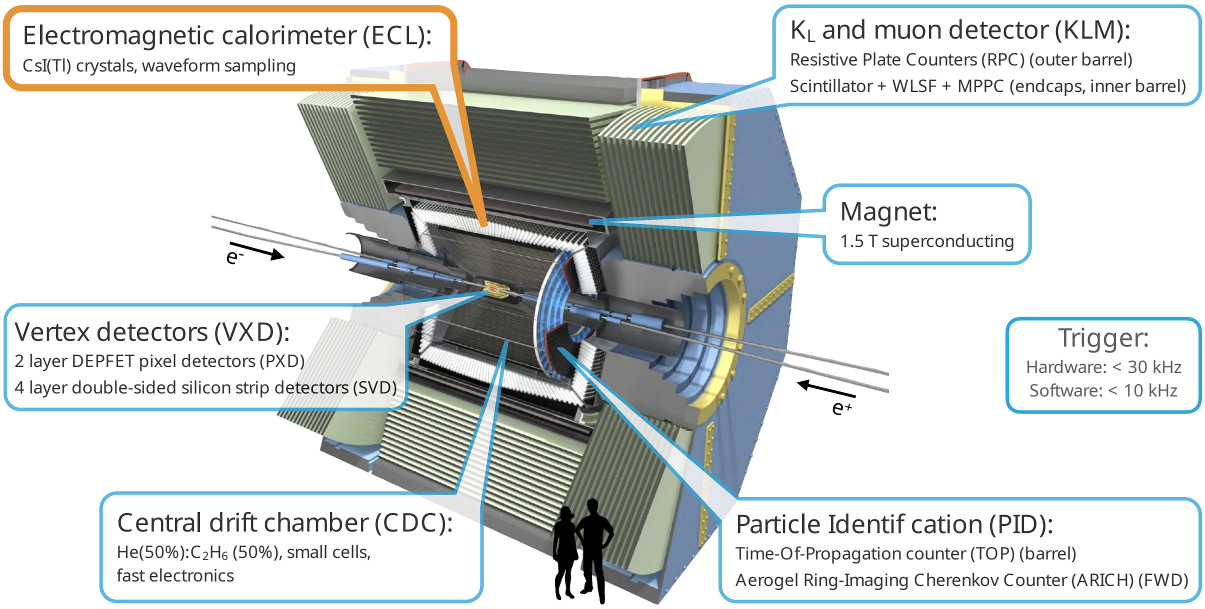

Electromagnetic calorimeter (ECL):
CsI(Tl) crystals, waveform sampling
K$_L$ and muon detector (KLM):
Resistive Plate Counters (RPC) (outer barrel)
Scintillator + WLSF + MPPC (endcaps, inner barrel)
Magnet:
1.5 T superconducting
e$^-$
e$^+$
Vertex detectors (VXD):
2 layer DEPFET pixel detectors (PXD)
4 layer double-sided silicon strip detectors (SVD)
Trigger:
Hardware: < 30 kHz
Software: < 10 kHz
Central drift chamber (CDC):
He(50%):C$_2$H$_6$ (50%), small cells,
fast electronics
Particle Identif cation (PID):
Time-Of-Propagation counter (TOP) (barrel)
Aerogel Ring-Imaging Cherenkov Counter (ARICH) (FWD)

Figure 2.2.: Schematic overview of the Belle II detector including brief information for the individual sub-detectors. The interaction point is in the center of the detector where the vertex detectors are located. The asymmetry of the detector is especially highlighted by the particle identification, which is only present in barrel and forward endcap. Figure adapted from [8].

The Belle II experiment constantly strives towards higher instantaneous luminosity in order to collect more data and improve the statistics for analyses. The experiment expects an integrated luminosity of about $50\,\mathrm{ab}^{-1}$ in ten years of data taking, in comparison to the $0.99\,\mathrm{ab}^{-1}$ collected by Belle over a span of eleven years. A higher instantaneous luminosity is mainly achieved by the means of two beam parameters:

- SuperKEKB is targeted to double the beam currents I_{beam} over KEKB. The beam current is asymmetric between LER and HER, therefore, I_{beam} is indicated with two values LER/HER.
- The beta-function characterizes the beam size at the IP vertical to the beam pipe in x-direction β_x^* and in y-direction β_y^*. Smaller beam sizes yield significantly more collisions per crossing, thereby increasing luminosity.

The beam parameters and resulting luminosity are associated with different phases of the experiment. The final stage of the SuperKEKB development aims to achieve a luminosity of $\mathrm{L} \approx 6 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, with $I_{\mathrm{beam}} \approx 2.8/2.0\,A$ and $\beta_y^* \approx 0.3\,\mathrm{mm}$ in nominal phase 3 operation by 2027. This is an increase in luminosity by a factor of 40 relative to the KEKB accelerator. As of June 2022, the collider operates in early phase 3 at a luminosity of $\mathrm{L} \approx 4.7 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, with $I_{\mathrm{beam}} \approx 1.5/1.2\,A$ and $\beta_y^* \approx 1.0\,\mathrm{mm}$.

Relative to Belle, considerably higher background levels arise at Belle II, caused by the measures for a higher luminosity, as well as the higher luminosity itself. Background greatly affects the energy reconstruction and is a major aspect throughout the studies in this work. Section 2.2.3 discusses background in more detail.

2.2. The Belle II Electromagnetic Calorimeter

The Belle II electromagnetic calorimeter (ECL) serves multiple purposes: First and foremost is the energy and position reconstruction of particles from energy depositions in ECL crystals. This primarily concerns photons, which are also the main focus of the studies in this work. Additionally, the ECL measurements are used to identify particles based on their shower shape and in combination with other sub-detector information. Lastly, the ECL generates trigger signals and assesses luminosity by measuring Bhabha scattering [6].

Section 2.2.1 describes the structure, geometry, and operation of the ECL. Section 2.2.2 introduces the concept of clusters and leakage. Section 2.2.3 gives an overview of the beam background processes at Belle II and how they affect the ECL.

2.2.1. Geometry and Operation

The following section is mostly based on [6, 7, 9]. The Belle II ECL consists of a total of 8736 individual scintillator crystals. The detector structure and desired energy resolution require a total-absorption calorimeter in limited space. Therefore, highly dense CsI(Tl) is chosen as the scintillator material. The incident of particles into the crystals produces particle showers that yield scintillation light. A thin, opaque foil made from Teflon and aluminum separates the crystals to contain the light emitted by scintillation within individual crystals. Finally, two photodiodes per crystal measure the light yield.

The crystals are arranged in a tightly packed crystal matrix that follows the cylindrical shape of the detector and is divided into three parts: A 3 m long barrel with an inner radius of 1.25 m, and forward and backward endcaps that enclose the barrel at the respective end. The IP in the center of the ECL is the origin of the spherical coordinates θ and ϕ, with θ being the polar angle to the beam pipe and ϕ being the azimuthal angle. Figure 2.3 shows a schematic representation of the ECL and the arrangement of crystals.

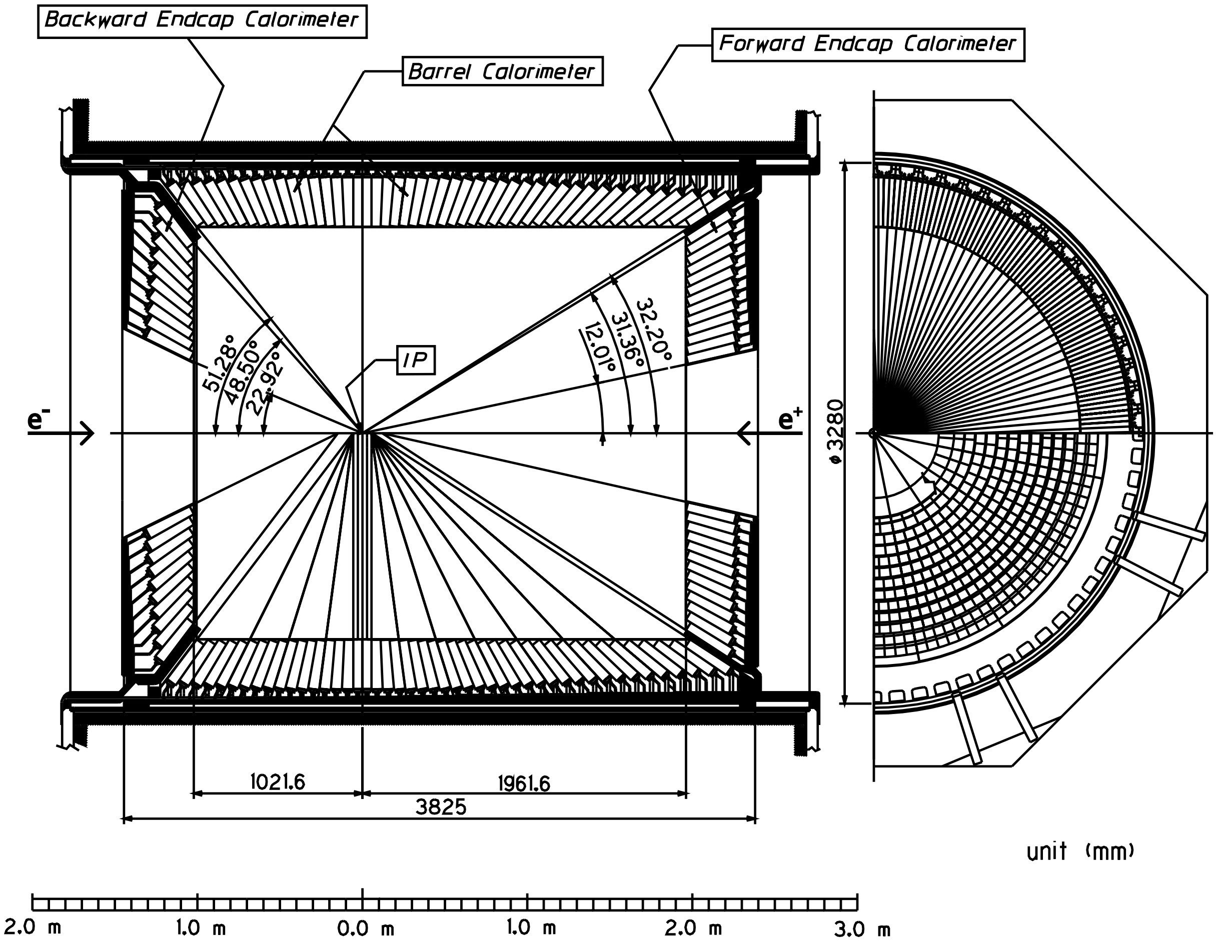


Figure 2.3.: Schematic representation of the electromagnetic calorimeter at the Belle experiment. The mechanical structure and crystal positions are identical for Belle and Belle II. The left side of the figure displays a cross-section of the overall calorimeter, with the interaction point (marked IP) in the center. The right side opens up the forward endcap and shows it from the perspective of the interaction point. The backward endcap has a similar, irregular crystal arrangement. Figure adapted from [9].

The barrel consists of 6624 crystals in 29 distinct shapes in order to achieve the smallest possible gaps between crystals. The crystals are narrow trapezoids with an average length of 30 cm and a face of about $5.5 \times 5.5\,\mathrm{cm}^2$ in the front and $6.5 \times 6.5\,\mathrm{cm}^2$ in the back. The crystal length corresponds to 16.1 radiation lengths. The front of each crystal is pointed towards the IP with a slight offset to avoid the escape of particles through the gaps.

2112 crystals in 69 shapes comprise the endcaps and are also pointed toward the IP with a slight offset. The crystal shapes in the endcaps are similar in length but more complicated than in the barrel. This is due to the irregular arrangement that is necessary to account for the decreasing number of crystals per θ-ring towards the beam pipe. On top of that, there are larger gaps between crystals that accommodate the support structure needed for the endcap geometry. The crystals in the backward endcap are less tightly packed than in the forward endcap.

The total calorimeter covers a region of approximately $12^\circ < \theta < 155^\circ$ which corresponds to about 90 % of total 4π acceptance in the laboratory frame. There are gaps around 1° wide between barrel and endcaps that are used for cables and pipes.

Crystal Measurements

Two photodiodes glued to the back of each crystal detect the scintillation light generated by particle incidents and subsequent showers. The signal of the diodes is pre-amplified and digitized by a field-programmable gate array. The digitized waveform provides time and energy measurements for the crystal which are then recorded. Additionally, pulse shape discrimination information is determined offline from the recorded data. These three quantities are the basic ECL measurements used for the energy reconstruction of particles.

The **recorded energy** of a crystal i is denoted by E_i^{rec}. The fit of a known response function to the digitized waveform yields the energy as the amplitude of the signal. The recorded energy is afflicted with electronic noise of around 0.35 MeV from the read-out process [10].

The **recorded time** of a crystal i is denoted by t_i^{rec}. Analogous to the recorded energy, it is determined in the fit of the response function to the waveform. The fit yields the starting time of the signal that is then related to the time of the interaction with the help of all detector measurements [10].

The **pulse shape discrimination (PSD)** information is strictly speaking not a crystal measurement but a property determined offline from the recorded waveform of the crystal. It consists of the hadron intensity HI_i in crystal i and its corresponding fit type and χ^2. The PSD information is based on the fact that different particles cause distinct scintillation responses in the ECL crystals. HI quantifies the amount of scintillation in a crystal that is caused by hadronic particles. The fit of several templates to the recorded waveform ultimately determines the amount of hadronic scintillation. Fit type and χ^2 value refer to that fit. The PSD information is only calculated for crystals with recorded energies above 30 MeV, which is a limit imposed by the data acquisition [11].

2.2.2. Clusters and Leakage

The following section is based on [6, 12]. Particles usually do not deposit their total energy in a single ECL crystal. Instead, the resulting shower spans over multiple crystals in the crystal matrix of the ECL. This means that energy depositions of the same particle end up in many crystals and are measured separately. The term cluster refers to a set of crystals with energy depositions that are corresponding to the same particle. Most times, the crystals that belong to a cluster form a connected region in the crystal matrix, however, this does not have to be the case. Additionally, a crystal can measure energy depositions of multiple particles at once. Therefore, the recorded energy in a crystal and the deposited energy by a cluster (more precisely by its associated particle) in that crystal are not necessarily identical. The last two factors are the major challenges for the clustering algorithms studied in this work whose task is to correctly allocate energy depositions to clusters.

Different types of particles in different settings result in distinct shower shapes in the ECL. The shower shape has a great influence on the arrangement of the set of crystals that belong to a cluster. The most simple shower shape is that of a photon, which deposits most of its energy in a central crystal and forms a radially symmetric shower around it. In comparison, electrons produce a less symmetric and more spread-out shower, thereby including more crystals with smaller energy depositions each in the cluster. The interactions of hadrons lead to strongly irregular clusters and in the case of charged hadrons oftentimes to an additional tilt in the shower shape. All of these showers can interact to form even more complicated clusters. This work presents the first validation of a new clustering algorithm. In order to gain insight into the basic functioning of the algorithm, the studies focus on the reconstruction of photon showers only. Additionally, chapter 3 introduces a specific set of criteria for the clusters, also referred to as cluster signatures, that ensure a well-controlled environment for the clustering.

Leakage is a crucial aspect of ECL clusters. The total deposited energy in all crystals of a cluster is less than the initial energy of the particle that created the cluster. The missing energy is not measured because some part of the particle shower leaked out of the crystals. For high-energy particles, leakage out of the back of crystals is dominant, for low-energy particles, side leakage becomes the prevalent loss. The small but unavoidable gaps between individual crystals enhance this issue. For this reason, the endcaps are at a significant disadvantage relative to the barrel. Furthermore, particles deposit energy in inactive material before reaching the ECL, for example in inner sub-detectors.

Leakage ultimately is a factor that limits the best possible energy resolution in the ECL. Typically, leakage amounts to a few percent of the initial particle energy, however, elaborated correction mechanisms allow pushing the resolution far beyond that limit (see section 2.3).

2.2.3. Beam Background

The following section is based on [6,13]. Beam background is the result of unstable beam particles colliding with the inner side of the beam pipe and creating mostly low-energetic showers that are then measured in the detector. This phenomenon affects all sub-detectors and in fact beam background accounts for a majority of the energy depositions in the Belle II ECL. The additional energy depositions pose a major challenge to the energy reconstruction of physics particles and underlying clustering processes. The most relevant sources for beam background in the ECL are:

- The **Touschek effect** refers to the Coulomb scattering of beam particles in the same bunch. The effect is proportional to the inverse beam size and the squared beam current. It accounts for up to 98 % of beam background in the ECL.
- **Beam gas scattering** refers to the Coulomb scattering of beam particles with residual gas atoms. The scattering rate is proportional to the beam current and the vacuum pressure. Beam gas scattering almost makes up the remaining 2 % of beam background.
- **Radiative Bhabha** are $e^-e^+ \rightarrow e^-e^+\gamma$ processes where the additional photon hits the beam pipe. This process is proportional to the luminosity. It is especially likely to produce high-energy beam background.

Usually, the resulting beam background ranges from 0.5 MeV to 1.0 MeV in the barrel up to 2.0 MeV in the endcaps. Most particles in this energy regime deposited their total energy in a single crystal. Seldom, background particles with much larger energy hit the ECL and create a cluster that is practically indistinguishable from clusters originating from the e^- e^+ collision and subsequent decays. These clusters are treated identically to every other cluster up to the physics analyses. It is then left to the individual requirements of the further analysis to identify and reject these types of clusters.

In general, the occurrence rate of beam background generating processes increases with smaller beam size, higher beam currents, and higher luminosity. These factors are major goals in the development of the experiment and were already improved by several factors throughout past runs (see section 2.1). The accompanying increase in background is mostly driven by the Touschek effect. Figure 2.4 visualizes the difference between the amount of background for two simulated phases of the experiment. First, higher background levels manifest in higher recorded energies per crystal because of pile-up. Second, the higher background causes more hits per event that go hand in hand with more cluster candidates in a more complex arrangement.

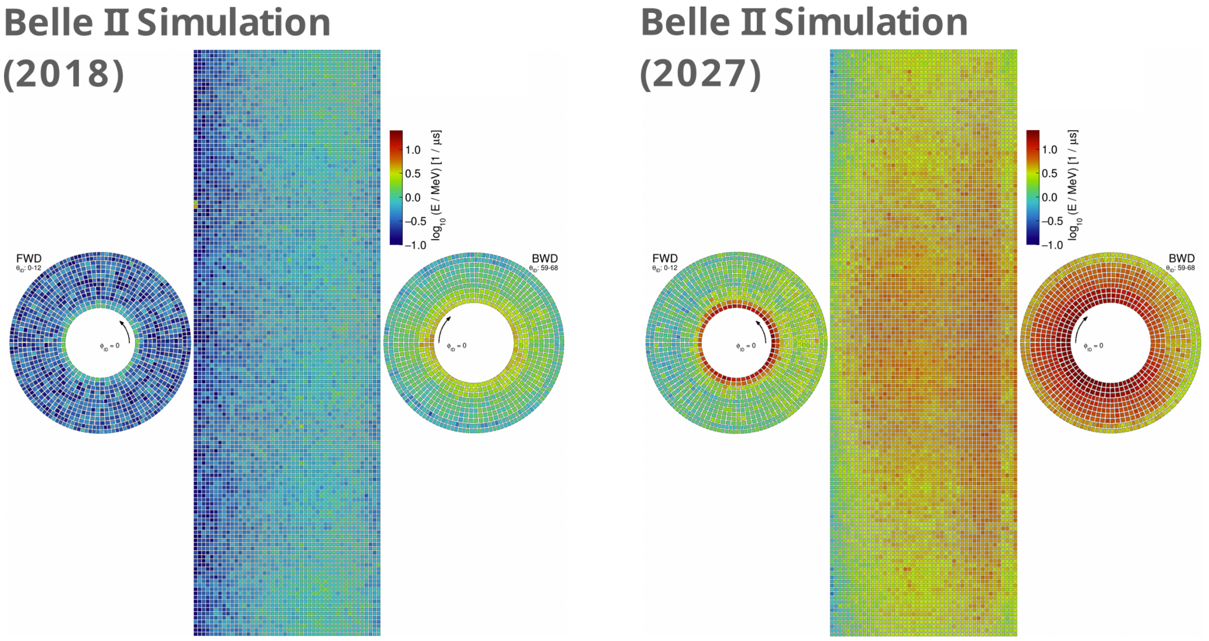


Figure 2.4.: Comparison of background energy depositions in the electromagnetic calorimeter for past phase 2 background as of 2018 and future nominal phase 3 background. Both plots show a projection of the calorimeter with the barrel rolled out and the endcaps from the perspective of the interaction point. The color of the crystals represents recorded energies. Figure adapted from [8].

Figure 2.4 highlights yet another important characteristic of beam background: The level of background is strongly dependent on the detector region. The backward endcap receives the largest doses of background, while the depositions decrease throughout the barrel down to the lowest level around the gap between the barrel and the forward endcap. Only in the very central rings of the forward endcap close to the beam pipe, the background level rises again. Figure 2.5 further illustrates the location dependence of the background as a function of the polar angle θ. Because the energy reconstruction greatly depends on the background, the upcoming studies for the most part analyze barrel, forward endcap, and backward endcap separately.

This work studies two types of backgrounds associated with different phases of the experiment:

- **Early Phase 3 Background:** The background simulation for the current experiment status as of November 2022 is referred to as early phase 3 background. The beta-function is set to $\beta_y^* = 1\,\mathrm{mm}$ at a luminosity of $\mathrm{L} = 0.3 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. This results in around 1500 crystals with energy measurements per event. Early phase 3 background is abbreviated as early background throughout the thesis.

- **Nominal Phase 3 Background:** The background simulation for the final status of the experiment in phase 3, as expected in 2027, is called nominal phase 3 background. For this background applies $\beta_y^* = 0.3\,\mathrm{mm}$ and at $\mathrm{L} = 6 \times 10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. This results in around 3500 crystals with energy measurements per event. Nominal phase 3 background is abbreviated as nominal background throughout the thesis.

The event generation for the studies described in chapter 3 uses the default early phase 3 and nominal phase 3 overlay files from simulation for release-05.

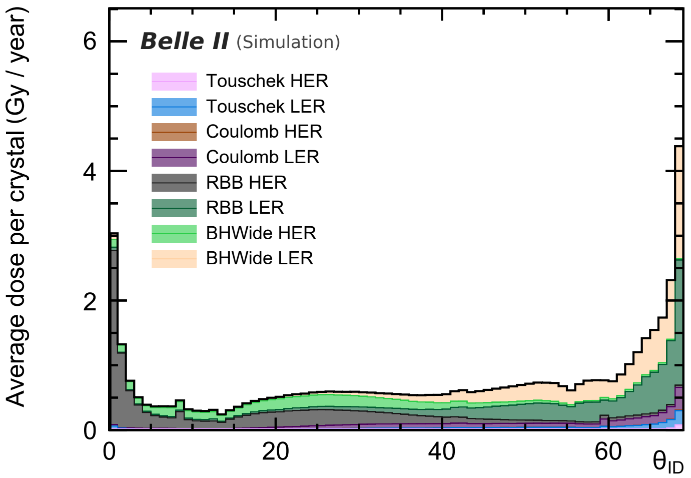


Figure 2.5.: Average energy dose per crystal in the calorimeter rings θ_{ID} for nominal phase 3 background. The left side at $\theta_{\mathrm{ID}} = 0$ is the most inner ring of the forward endcap, $\theta_{\mathrm{ID}} = 68$ corresponds to the most inner ring in the backward endcap. A different simulation than the one in this work creates the background for this plot, however, it still gives a good idea of the expected location dependence. Figure adapted from [6].

2.3. The Belle II Analysis Software Framework

The Belle II Analysis Software Framework (basf2) [1,14] is the currently used software framework that serves as the performance baseline in all upcoming studies. The functional scope of basf2 includes almost all software-related activities at Belle II and goes far beyond energy reconstruction and the ECL in general. In the scope of this work, the term basf2 specifically refers to the ECL reconstruction, namely the energy reconstruction including the clustering algorithm and the position reconstruction.

Energy Reconstruction

Per event, around 1500 crystals measure energy in early background events, and around 3500 in nominal background events. The recorded energies yield approximately between 40 and 100 cluster candidates respectively. The initial step of the energy reconstruction is the clustering of recorded energies. The objective of clustering is to assign the recorded energies E_i^{rec} in crystals i to clusters u that in turn are associated with particles. Figure 2.6 shows a representation of ECL crystals in an event that illustrates the clustering objective at Belle II. Basf2 employs a topological approach that consists of the following steps:

1. The algorithm finds seed crystals, that are defined by $E_{\mathrm{seed}}^{\mathrm{rec}} > 10\,\mathrm{MeV}$.
2. Starting from the seed crystals, the eight neighboring crystals are examined and added to a connected region (CR) if they fulfill $E_{\mathrm{i}}^{\mathrm{rec}} > 0.5\,\mathrm{MeV}$.
3. If a crystal in the eight immediate neighbors exceeds $E_{\mathrm{i}}^{\mathrm{rec}} > 10\,\mathrm{MeV}$, its neighbors are also considered according to step two.
4. Within the CRs, the algorithm finds crystals with a local maximum (LM) in the recorded energies. Each LM becomes a cluster candidate and is the origin for that cluster in basf2.
5. A 5 × 5 area around the LMs and excluding the four corner crystals is taken into account for the further clustering process. Figure 2.7 illustrates that selection of crystals.

If there is only one LM in the CR, all crystals of the CR that are within the 5 × 5 selection are included in that cluster. The membership of crystal i in cluster u is denoted by $p_i^{(u)} \in \{0, 1\}$.

If there are more LMs in the CR, the recorded energies must be divided among several clusters. For each cluster, the crystals of the CR within the respective 5 × 5 selection are assigned a partial membership of crystal i in cluster u denoted by $p_i^{(u)} \in [0, 1]$ and defined as

$$p_i^{(u)} = \frac{E_i^{\text{rec}} \cdot \exp\left(-Cd_i\right)}{\sum_k E_k^{\text{rec}} \cdot \exp\left(-Cd_k\right)}. \tag{2.1}$$

Here, simulation determines the constant $C = 0.7$ and d_i are the distances between cluster center and crystals i. The cluster center is defined by the position reconstruction described in the next paragraph. The process of splitting up the recorded energies starts out with the coordinates of the LM crystal as the cluster center. The center is updated iteratively and the procedure stops as soon as the centers of all clusters in the CR are stable within 1 mm.

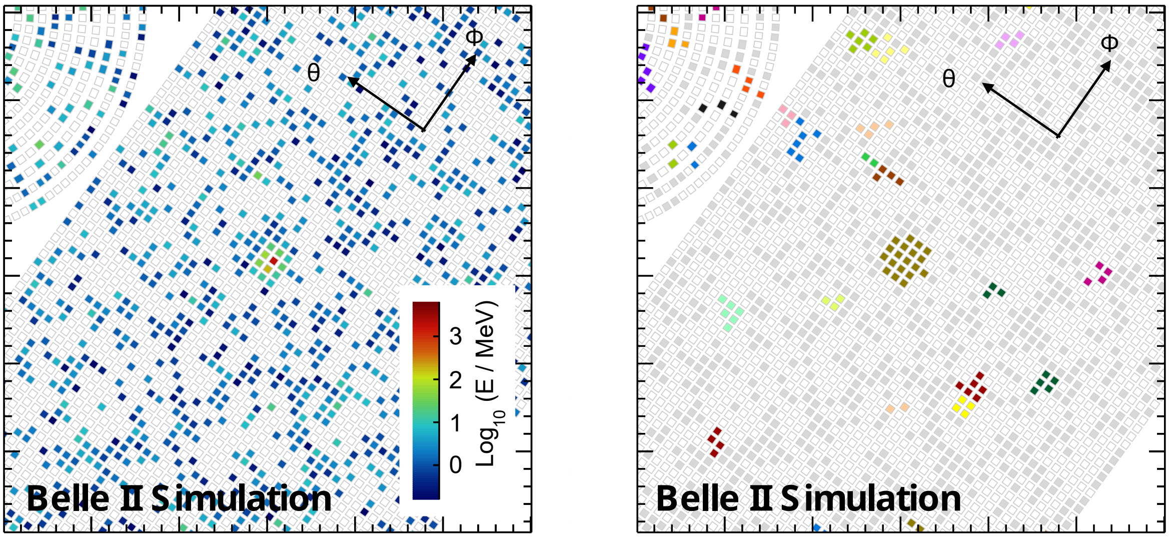


Figure 2.6.: The figure shows a flat projection of the crystals in the electromagnetic calorimeter in the $\theta - \phi$ plane. The barrel part is depicted to the right in both plots and rolled out from the viewpoint of the interaction point. The top left corner displays the forward endcap. The left plot by color visualizes recorded energies that are the input for the basf2 clustering algorithm. The right plot displays the clustering result with several clusters that are now each associated with a particle. Figure adapted from [8].

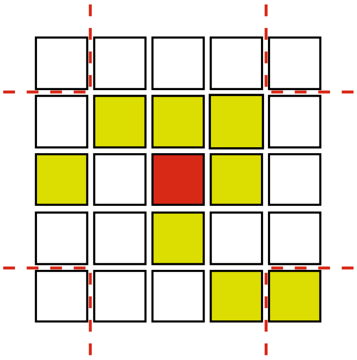


Figure 2.7.: Visualization of the basf2 clustering selection. The figure shows the crystal matrix from the perspective of the interaction point. The crystals considered in clustering are the 5 × 5 crystals surrounding the center with a local maximum in the recorded energies and excluding the four corners. Figure adapted from [15].

The sum of all energy depositions that belong to a cluster according to the initial clustering is later referred to as $E^{(u)}_{\text{pred}}$ and an important quantity for the comparison and analysis of basf2 (see section 5.1). However, it is not the cluster energy that basf2 yields for any further physics analyses. After the initial clustering, the crystal selection is optimized depending on several factors to correct for leakage and background. Depending on the expected background in the event and the energy of the LM and its eight immediate neighbors, between 2 (low-energy cluster in high background) and 21 crystals (high-energy cluster in low background) with the highest recorded energies are included in the final cluster. Additionally, the sum of the selected energy depositions is corrected as a function of reconstructed position, energy, and expected background level. The internal variable for the final reconstructed cluster energy by basf2 is called clusterE and is later referred to as $E^{\text{cor}}_{\text{pred}}$ (see section 5.3.1).

Ultimately, three characteristics of the basf2 algorithm are primarily notable for the upcoming studies: First, basf2 exclusively uses the recorded energies as input for the clustering algorithm. Second, it is only able to reconstruct connected regions of crystals surrounding the LM. Third, the algorithm does assign partial membership to energy depositions of particles associated with an LM, but it can not partially assign membership to background.

Position Reconstruction

The ECL position reconstruction determines the directions of the four-vectors of neutral particles whose reconstruction is exclusively based on the ECL. In this work, the basf2 position comes to application in the physics studies in chapter 7 for the reconstruction of photon four-vectors. Initially, the position reconstruction yields the coordinates $\vec{x}$ of the center of a cluster. It does so for CRs with just one LM and with multiple LMs alike by calculating

$$\vec{x} = \frac{\sum_i v_i \vec{x}_i}{\sum_i v_i}, \tag{2.2}$$

where $v_i = 4+\log(E_i/E_{\text{cluster}})$. Here, the cluster energy E_{cluster} is the sum of all energy depositions in that cluster. For CRs with multiple LMs, the process is carried out iteratively in alternation with the partial membership assignment until the cluster centers are stable. Finally, the directions of the four-vectors are the vectors from the well-known interaction point of the experiment to the cluster centers.

3. Event Generation and Selection

This work uses Monte Carlo (MC) generated and simulated events throughout all studies. The underlying truth information provided by MC generated and simulated data (or short MC data) is the foundation for the training of the GravNet models proposed in chapter 4. This chapter introduces the generation and selection of the events that the models train on and evaluate.

Section 3.1 describes the MC generation and simulation of events for various configurations. The configurations are set on a particle level, yet the reconstruction algorithms operate on the resulting detector response, more precisely in a crystal-level environment. This discrepancy requires a subsequent selection of events based on the characteristics of the clusters originating from the initial particles. Section 3.2 presents the criteria for the selection of events and describes the characteristics of the clusters that are referred to as cluster signatures.

3.1. Monte Carlo Generation and Simulation

The MC generation yields the initial positions and four-vectors for a desired particle configuration according to a physics model, which here is the Standard Model. For the studies in this work, basf2 release-06-00-03, namely its `ParticleGun` module (or single particle gun), performs the MC generation. The single particle gun takes a particle type, a vertex position, angles θ and ϕ in the laboratory frame, and a particle momentum p as input. It generates the four-vector of the desired particle, moving from the vertex in direction (θ, ϕ) with momentum p. Throughout all studies, the vertex position is set to the IP of the experiment. When generating multiple events, the single particle gun can automatically vary input values according to a given distribution. Most of the studies use this feature to create particles that are distributed uniformly in directions and momenta.

The particle gun is suitable for events with one particle. A single photon normally results in one cluster in the ECL. For the generation of overlapping photon clusters, the custom module `CloseByParticleGenerator` (or dual particle gun) is used. The input for this extended version of the single particle gun are two particle types, two momenta $p_{1,2}$, the angles (θ, ϕ), and an interval of opening angles $[\delta_{\text{min}}, \delta_{\text{max}}]$. It generates the two desired particles with the respective momenta moving from the IP towards (θ, ϕ) at an angular separation $\in [\delta_{\text{min}}, \delta_{\text{max}}]$.

The settings for all variables in the MC generation are stated with the respective studies in chapters 6 and 7.

After generating the initial particles, the simulation of the detector response takes into account all types of interactions of the particles with the detector like ionisation, scintillation, bremsstrahlung, pair production, Cherenkov radiation, and so on. The outputs of the simulation are detector measurements, like for the ECL crystals i the recorded energies E_i^{rec}, the recorded times t_i^{rec}, and the hadron intensity HI_i with its corresponding fit type and χ^2.

Adding background to MC generated physics processes is achieved by applying overlays to the detector response as part of the simulation. This work uses the default early phase 3 and nominal phase 3 overlay files from simulation for release-05. The type of background overlay (either early or nominal) is indicated in the settings for the respective study. The simulation is carried out with `GEANT4` [16] which is integrated into basf2.

In contrast to real detector data, MC data has the advantage of providing MC truth information. This is the information that is put into the event generation, as well as information about the true detector response obtained in the simulation. For the ECL, the true fractions of recorded energies in crystals i by particle u are provided as $t_i^{(u)}$.

Relying on MC information is key in the first validation of a reconstruction algorithm before the deployment on real detector data. However, the simulation is not perfect for a complex experiment like Belle II and is also limited by computational resources. The differences between simulated data and real data are subject to constant investigation by the Data Production group. For the simple scenarios studied in this work, the simulation yields results that are accurate enough for an outlook on the performance on real data [6].

3.2. Event Selection and Cluster Signatures

Both particle guns introduced in the previous section 3.1 operate on a particle level and set the kinematics of an event. Thereby they create a specific scenario of underlying physics. However, because the algorithms operate on a crystal level, the scenario is yet incomplete. Particles create distinct shower shapes in the ECL depending on the type of particle and the exact settings (see section 2.2.2). The resulting characteristics of the ECL crystals containing the shower are referred to as cluster signatures. A set of criteria defines the specific cluster signatures that GravNet trains on and is evaluated on. This is necessary for several reasons: A well-controlled environment is crucial for the interpretation and understanding of the performance of a new algorithm. It ensures comparability to the baseline algorithm. Also, in the particular case of machine learning, a set of specific samples without outliers often enables stable training in the first place. Most importantly, the criteria guarantee that one cluster is always associated with exactly one photon and vice versa. Clusters originating from a different number of particles are possible but excluded in the upcoming studies and part of future work.

A custom module in basf2 checks the event, as defined by the physics scenario and produced by the event generation, on a crystal level. Only events that fulfill all criteria for the cluster signatures are used in further studies. The procedure reduces the number of events after the event generation by a factor of three to ten depending on the scenario. Oftentimes machine learning algorithms are able to generalize to slightly different scenarios than the ones used in training. Nonetheless, a mechanism to not only identify signatures without MC information but also to deal with all occurring signatures instead of specific ones is an important addition in future work.

The studies in this work are associated with either one of two cluster signatures: The toy studies in section 6.2 and the physics studies in section 7.3 have to fulfill the criteria for one-cluster events described in section 3.2.1. The toy studies in section 6.3 and the physics studies in section 7.4 have to fulfill the criteria for two-cluster events described in section 3.2.1.

3.2.1. One Cluster

The cluster signatures are defined for a region of interest (ROI), 9×9 crystals around the center of an event. The global coordinates of a crystal with an LM in E_i^{rec} define the center for one-cluster events. Basf2 provides the identification of LMs and the 9×9 neighbors to that crystal, as well as the global crystal coordinates. The ROI extends the 5×5 area that is considered for a cluster by the baseline algorithm by two crystals to each side.

After the event center is set, the crystals in the resulting ROI have to fulfill the following criteria:

- The LM must have $E_{\mathrm{LM}}^{\mathrm{rec}} > 10\,\mathrm{MeV}$.
- The center of the event must be the only LM in the ROI.
- The particle associated with the energy deposition in the LM must be responsible for at least 20 % of recorded energy in that crystal. That is $t_{\mathrm{LM}}^{(1)} \geq 0.2$.
- The particle must be a primary photon according to basf2. That is, it has to be the photon that was generated by the particle gun and not a decay product of it.
- The photon must be the only particle that deposited energy in the ROI.

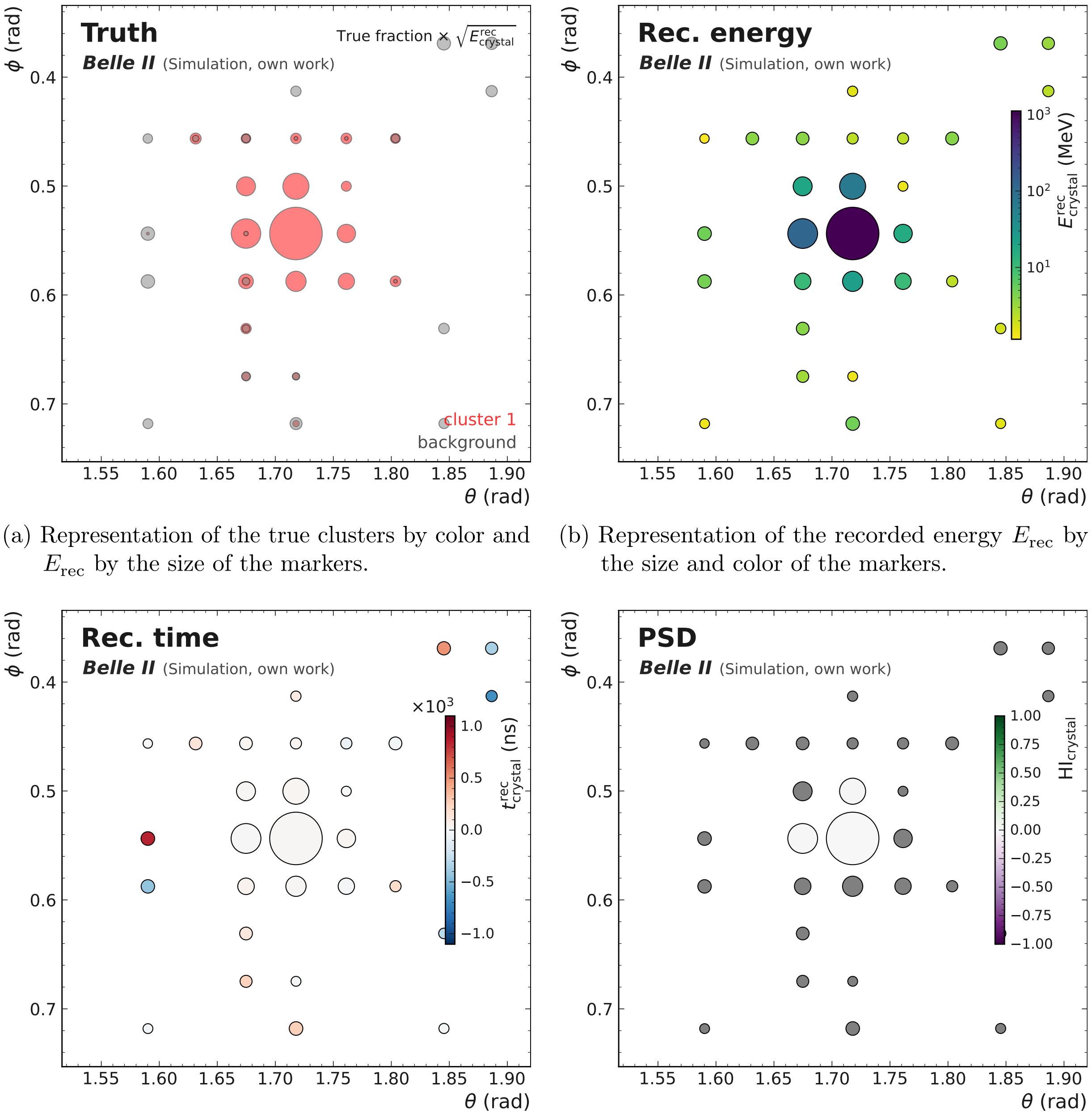
A so-called event display visualizes the ROI containing the cluster. The event display maps a flat projection of the ECL crystals from the view of the IP, similar to figure 2.6 but limited to a 9×9 area. Different crystal measurements like E_i^{rec} give a visual representation of the cluster close to the actual shape in the detector. Figure 3.1 shows an event display for an event that fulfills the criteria for one-cluster events with early background. Figure 3.2 shows an example of an event with nominal background.

3.2.2. Two Overlapping Clusters

The determination of the center of an event for two-cluster events again starts out with the selection of two LMs by basf2. The global coordinates (θ, ϕ) of the crystals containing the LMs are interpreted as latitude and longitude. The midpoint of the two LMs is found according to the haversine metric. The crystal closest to the midpoint becomes the center of the event. The 9×9 neighbors of that crystal are added in the same manner as for one cluster. Again, specific criteria are defined for the crystals within the ROI:

- Both LMs must have $E_{\mathrm{LM}}^{\mathrm{rec}} > 10\,\mathrm{MeV}$.
- The two LMs must be the only ones in the ROI.
- For either one LM applies $t_{\mathrm{LM}}^{(1)} \geq 0.2$, and for the other $t_{\mathrm{LM}}^{2)} \geq 0.2$.
- If $t_{\mathrm{LM}}^{(1)} \geq 0.2$ applies to an LM, then $t_{\mathrm{LM}}^{(1)} > t_{\mathrm{LM}}^{(2)}$ must also be true and vice versa.
- The largest deposited energy of a particle must be within a 5×5 area around its associated LM. In combination with the last criterion, this ensures that the particle is correctly associated with the LM. Otherwise, energy depositions from several particles or background adding up can cause a fake LM.
- Both particles associated with the LMs must be primary photons.
- The two photons must be the only particles with deposited energy in the ROI.
- Each of the photons has to deposit at least 10 MeV in the same (shared) crystals as the other photon in a 5×5 area around their respective LM. Because the ROI is extended to a 9×9 area, two clusters with an overlap in their 5×5 vicinity are still fully contained according to this selection.
- The LM with smaller θ, or larger ϕ if θ are identical, and all energy depositions corresponding to its particle are then labeled as cluster 1.

Figure 3.3 displays an event that is typical for two overlapping cluster events with early background. Figure 3.4 does the same for an event with nominal background.



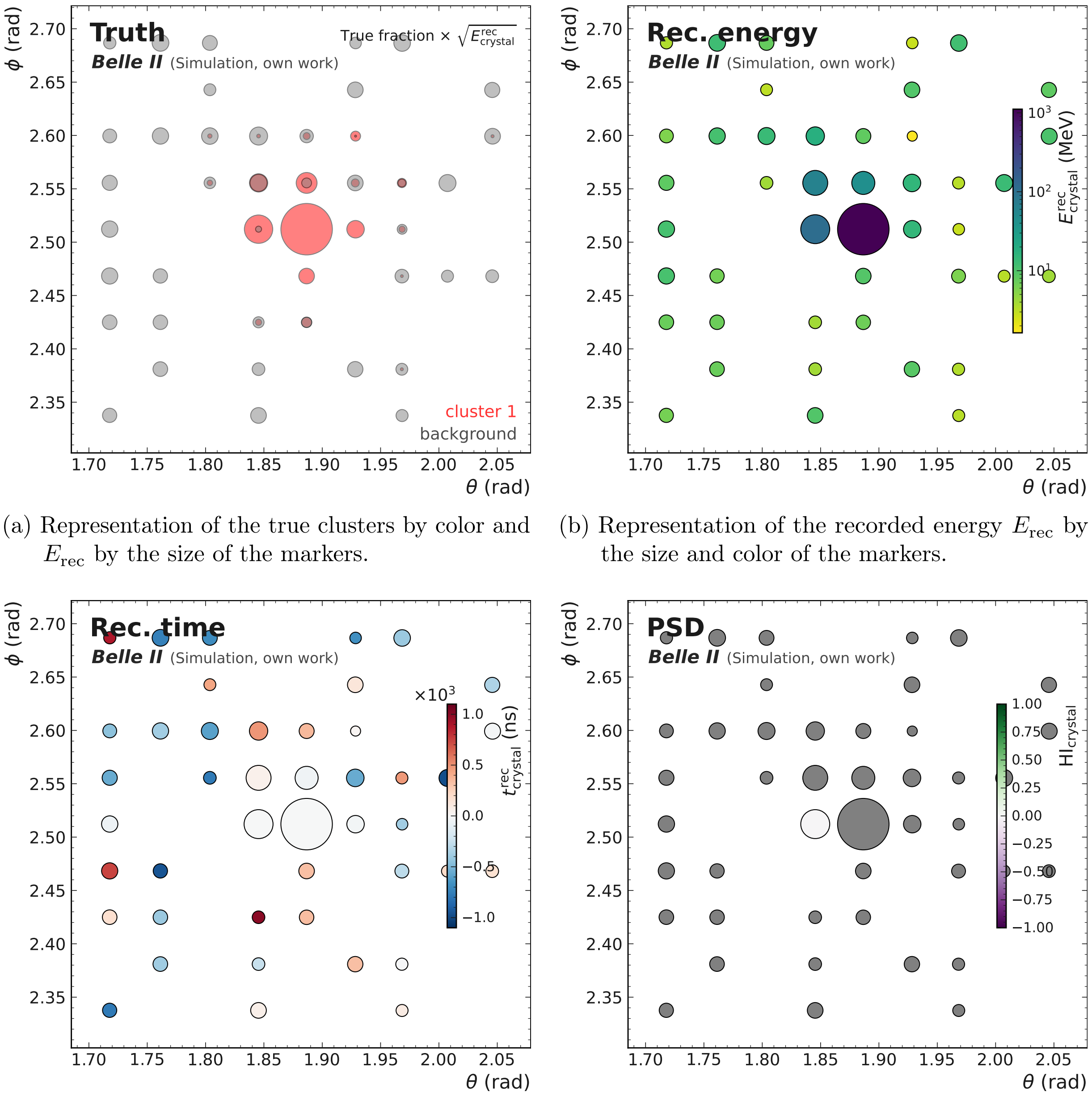
(a) Representation of the true clusters by color and E_{rec} by the size of the markers.

(b) Representation of the recorded energy E_{rec} by the size and color of the markers.

(c) Representation of the recorded time t_{rec} by color and E_{rec} by the size of the markers.

(d) Representation of the hadron intensity HI by color and E_{rec} by the size of the markers.

Figure 3.1.: Event display of a typical event with one cluster in the presence of early phase 3 background. θ and ϕ are the detector coordinates. Scaled and colored marks represent crystals and depict different measurements. The scaling of the recorded energy with $\sqrt{E_{\mathrm{rec}}}$ improves the visibility of low-energy crystals. Because only crystals with $E_{\mathrm{rec}} > 30\,\mathrm{MeV}$ contain pulse shape discrimination information (PSD) or the template fit fails, crystals with fit type -1 are masked in the corresponding PSD plot. The event is from the one-cluster toy study with early phase 3 background in section 6.2.



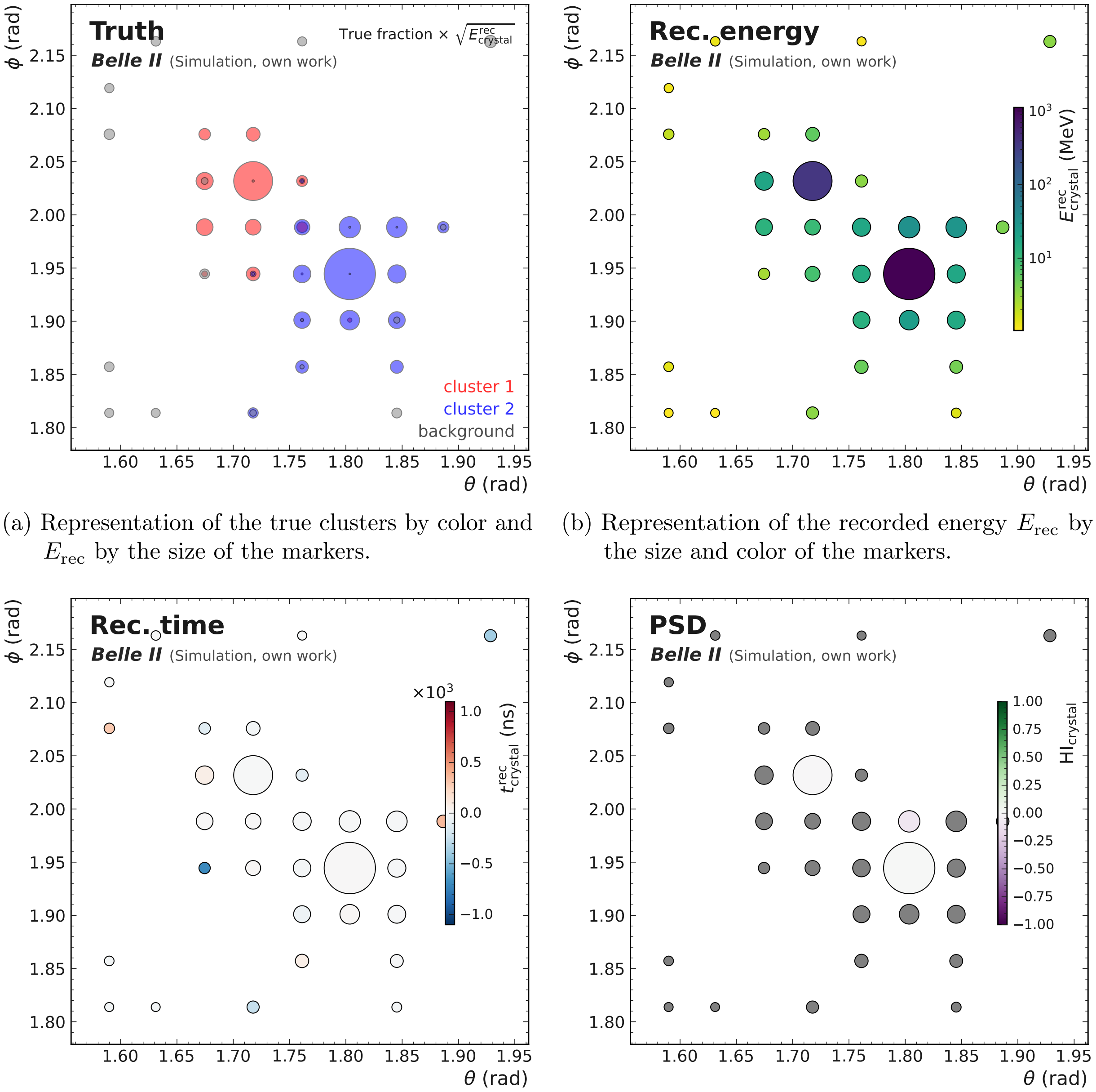
(a) Representation of the true clusters by color and E_{rec} by the size of the markers.

(b) Representation of the recorded energy E_{rec} by the size and color of the markers.

(c) Representation of the recorded time t_{rec} by color and E_{rec} by the size of the markers.

(d) Representation of the hadron intensity HI by color and E_{rec} by the size of the markers.

Figure 3.2.: Event display of a typical event with one cluster in the presence of nominal phase 3 background. θ and ϕ are the detector coordinates. Scaled and colored marks represent crystals and depict different measurements. The scaling of the recorded energy with $\sqrt{E_{\text{rec}}}$ improves the visibility of low-energy crystals. Because only crystals with $E_{\text{rec}} > 30\,\text{MeV}$ contain pulse shape discrimination information (PSD) or the template fit fails, crystals with fit type -1 are masked in the corresponding PSD plot. The event is from the one-cluster toy study with nominal phase 3 background in section 6.2.



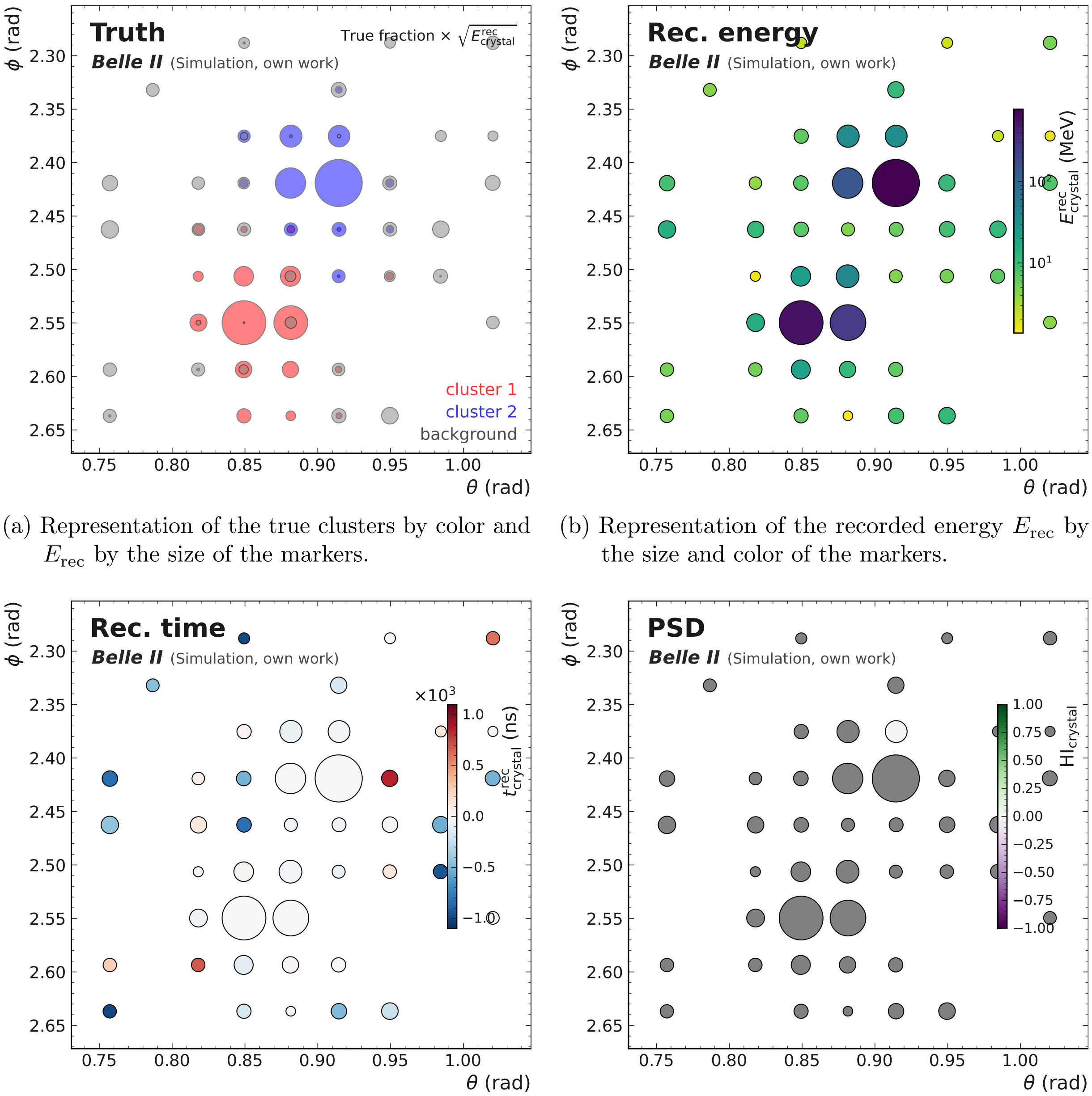
(a) Representation of the true clusters by color and E_{rec} by the size of the markers.

(b) Representation of the recorded energy E_{rec} by the size and color of the markers.

(c) Representation of the recorded time t_{rec} by color and E_{rec} by the size of the markers.

(d) Representation of the hadron intensity HI by color and E_{rec} by the size of the markers.

Figure 3.3.: Event display of a typical event with two overlapping clusters in the presence of early phase 3 background. θ and ϕ are the detector coordinates. Scaled and colored marks represent crystals and depict different measurements. The scaling of the recorded energy with $\sqrt{E_{\mathrm{rec}}}$ improves the visibility of low-energy crystals. Because only crystals with $E_{\mathrm{rec}} > 30\,\mathrm{MeV}$ contain pulse shape discrimination information (PSD) or the template fit fails, crystals with fit type -1 are masked in the corresponding PSD plot. The event is from the two-cluster toy study with early phase 3 background in section 6.3.



(a) Representation of the true clusters by color and E_{rec} by the size of the markers.

(b) Representation of the recorded energy E_{rec} by the size and color of the markers.

(c) Representation of the recorded time t_{rec} by color and E_{rec} by the size of the markers.

(d) Representation of the hadron intensity HI by color and E_{rec} by the size of the markers.

Figure 3.4.: Event display of a typical event with two overlapping clusters in the presence of nominal phase 3 background. θ and ϕ are the detector coordinates. Scaled and colored marks represent crystals and depict different measurements. The scaling of the recorded energy with $\sqrt{E_{\mathrm{rec}}}$ improves the visibility of low-energy crystals. Because only crystals with $E_{\mathrm{rec}} > 30\,\mathrm{MeV}$ contain pulse shape discrimination information (PSD) or the template fit fails, crystals with fit type -1 are masked in the corresponding PSD plot. The event is from the two-cluster toy study with nominal phase 3 background in section 6.3.

4. GravNet

GravNet is a machine learning (ML) algorithm based on the concept of graph neural networks (GNNs). Qasim et al. [2] initially propose the algorithm with an application to a toy model of a highly granular calorimeter. Section 4.1 introduces the fundamentals of the application of GravNet to the clustering of energy depositions in the Belle II ECL. Section 4.2 describes the details of the GravNet architecture for this application. Section 4.3 presents the implementation of the algorithm for the studies in this work. This chapter assumes basic knowledge of the functioning of GNNs and their training. For example, Liu and Zhou [17] give an introduction to the subject.

4.1. Fundamentals

This section gives an overview of relevant concepts for the application of the GravNet algorithm. Section 4.1.1 introduces the concept of fuzzy clustering, section 4.1.2 the objective for the training of GravNet, and section 4.1.3 the representation of events by graphs.

4.1.1. Introduction to Fuzzy Clustering

The term fuzzy clustering refers to the partial assignment of individual crystals to several clustering classes. This is contrary to the term hard clustering, which stands for the exclusive assignment of individual crystals to a single class. The GravNet implementation in present work strictly follows the concept of fuzzy clustering. Consequently, GravNet predicts fractions $p_i^{(u)}$ of recorded energies E_i^{rec}. These fractions denote how much of E_i^{rec} in crystal i belongs to cluster u. Background is its own clustering class, equivalent to the class of a cluster that is associated with a photon. Accordingly, for one-cluster signatures applies $u \in$ {background, cluster 1}, and for two-cluster signatures $u \in$ {background, cluster 1, cluster 2}. The predicted fractions $p_i^{(u)}$ from GravNet and basf2 alike, as well as the true fractions $t_i^{(u)}$ from the MC information, are also referred to as weights $w_i^{(u)}$. The following is true for $w_i^{(u)}$ in the context of fuzzy clustering:

$$\sum_u w_i^{(u)} = 1 \quad \forall i, \tag{4.1}$$

with $w_i^{(u)} \in [0, 1]$.

GravNet fuzzy clusters for all available classes including background. Basf2 is able to partially assign energy depositions to particles that are associated with the LM of a cluster. However, it is not able to partially assign energy depositions to background and as a result, is forced to either fully exclude or include crystals in a clustering. This conceptual difference stands out the most in the comparison between basf2 and GravNet predictions. The capability to accurately depict the underlying clustering can be an advantage, especially in future high background scenarios that are part of the nominal background studies in this work.

4.1.2. Training Objective

GravNet is a supervised ML algorithm. This means that the learnable parameters of the model are adjusted by the means of gradient descent in order to minimize a loss function. The loss function compares predicted fractions $p_i^{(u)}$ from the model to true fractions $t_i^{(u)}$ from the MC information for the samples of a training data set. The deployed $L2$-loss function is defined as

$$L2 = \sum_u \left(p_i^{(u)} - t_i^{(u)} \right)^2 , \tag{4.2}$$

and adds up the differences between truth and prediction for all classes u and crystals i. The $L2$-loss function goes hand in hand with GravNet yielding fractional rather than discrete outputs, making it a regression instead of a classification algorithm. It is worth noting, that the original paper proposes a loss that scales with the recorded energies in the crystals. However, for this application, the basic $L2$-loss is found to be more stable and delivers better results.

4.1.3. Representation of Events by Graphs

GravNet belongs to the class of GNNs and processes graphs consisting of nodes and edges. Therefore, the events for training and evaluation generated according to chapter 3 must be represented by graphs. GravNet does not operate on all crystals in the ECL simultaneously but evaluates the ROI of an event. Section 3.2 outlines the determination of the ROI that spans 9×9 crystals around the center of the event. Only these 81 crystals are taken into account for further processing of the event. Each crystal in the ROI is represented by a node in a graph of the event. Crystal measurements and properties of the respective crystals become the node features and ultimately the input features for GravNet. Section 4.3.1 specifies the input features.

In comparison to many other common ML algorithms like fully connected neural networks (FCCs) and convolutional neural networks (CNNs), GNNs do not require a fixed input size [17]. Hence GravNet does not depend on 81 nodes as input size but rather handles a variable number of nodes. This is used as an advantage by removing irrelevant nodes, namely crystals without recorded energies. A threshold of $E_{\text{rec}} > 1\,\text{MeV}$ on the recorded energy per crystal additionally eliminates crystals with energy measurements due to electronic noise. Roughly 20 to 45 nodes per event, depending on the type of background and the number of clusters, remain from the initial fixed number of nodes in a graph. The reduction of the input size not only comes with the advantage of GravNet not having to ignore nodes without information but also results in a speed-up for the inference times.

4.2. Architecture

This section describes the processing of the graph of an event by the GravNet algorithm. It starts with the introduction of the GravNet layer in section 4.2.1 and subsequently gives an overview of the overall structure of the architecture in section 4.2.2.

4.2.1. GravNet Layer

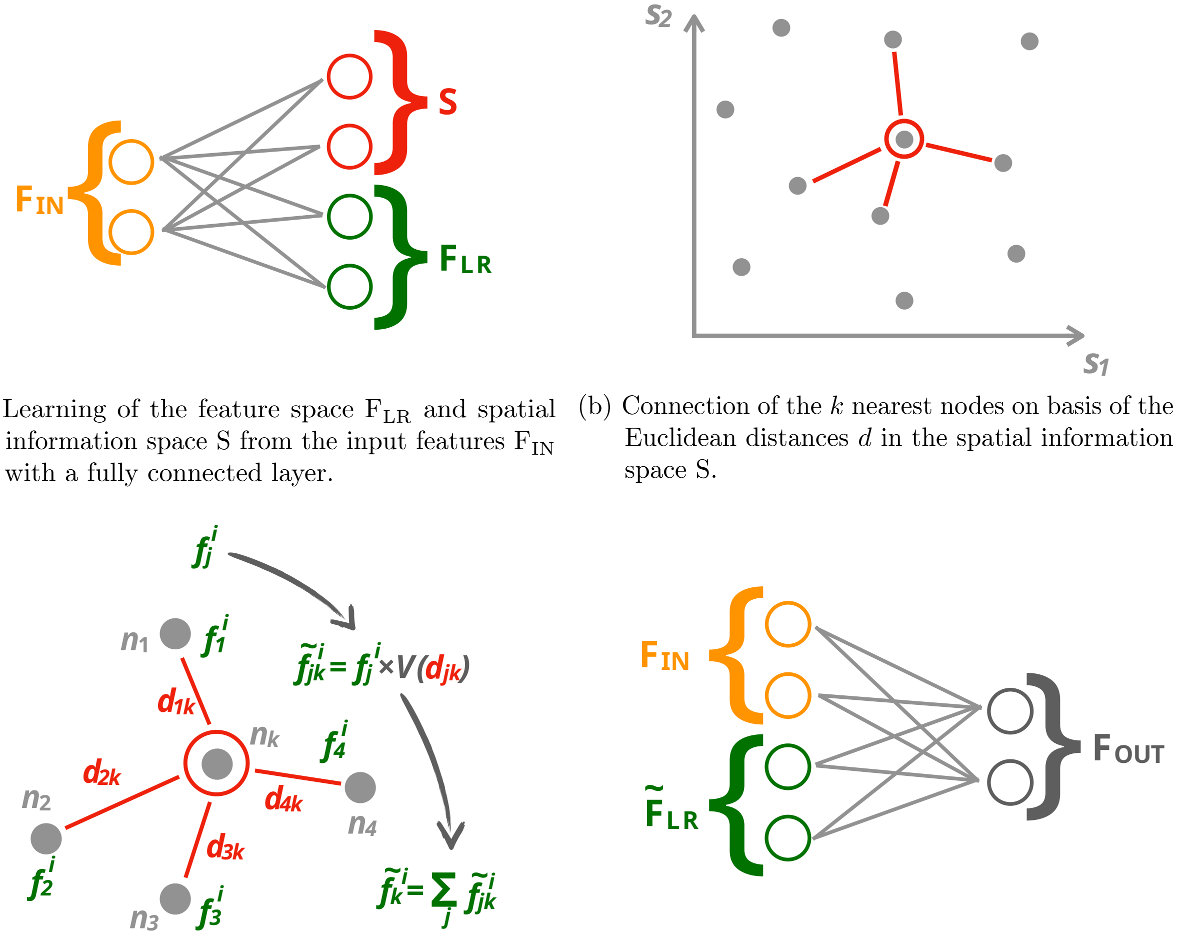
The GravNet layer stands at the core of the overall GravNet architecture and carries out the message passing between the nodes of a graph. Message passing means that nodes gather information, namely the features, of other connected nodes. However, so far the graph of an event is unconnected, making it a point cloud of nodes. For this reason, the GravNet layer first has to connect nodes which it does in an end-to-end learned representation space. Figure 4.1 visualizes the full operation of the GravNet layer that consists of four steps:

(a) A fully connected layer transforms the initial features F_{IN} into two learned representation spaces. The first space S represents spatial information, and the other F_{LR} transformed features used for message passing.

(b) Each node in the graph is connected to its k nearest neighbors according to the Euclidean distances d in S. This step turns the unconnected graph into an undirected graph.

(c) The incoming features f for each node are individually weighted by a Gaussian potential $V(d) = \exp\left(-10d^2\right)$ that depends on the distance between the connected nodes in S. The scaled features $\widetilde{f}$ of connected nodes are then aggregated to new features $\widetilde{\mathrm{F}}_{\mathrm{LR}}$ by summation. The scaling according to a decreasing potential in distance gives GravNet its *gravitational* name.

(d) All gathered features are concatenated with the initial features and processed to the output features $\mathrm{F}_{\mathrm{out}}$ using another fully connected layer.


F_IN
S
F_LR
S2
S1
f_j^i
n_1
f_1^i
$\widetilde{f}_{jk}^i = f_j^i \times V(d_{jk})$
d_{1k}
n_k
f_4^i
d_{2k}
n_2
d_{4k}
n_4
f_2^i
d_{3k}
n_3
f_3^i
$\widetilde{f}_k^i = \sum_j \widetilde{f}_{jk}^i$
F_IN
$\widetilde{\mathrm{F}}_{\mathrm{LR}}$
F_OUT

(a) Learning of the feature space F_{LR} and spatial information space S from the input features F_{IN} with a fully connected layer.

(b) Connection of the k nearest nodes on basis of the Euclidean distances d in the spatial information space S.

(c) The features f of connected nodes are scaled to $\widetilde{f}$ according to the Gaussian potential $V(d)$. Subsequently, $\widetilde{f}$ are aggregated by summation.

(d) The initial input features F_{IN} and the aggregated features $\widetilde{\mathrm{F}}_{\mathrm{LR}}$ are processed to the output features $\mathrm{F}_{\mathrm{OUT}}$ via a fully connected layer.

Figure 4.1.: Visualization of the GravNet layer. The figures display the learning of the representation spaces, the connection of nodes, the message passing, and the concatenation of the output for one node. In reality, the operation takes place for all nodes in the graph of an event simultaneously. Figure adapted from [2].

The number of initial features, the dimensions of the two representation spaces, and the number of nearest neighbors k are identified as hyperparameters and determined in a hyperparameter optimization in section 4.3.2. The dimension of the output features $\mathrm{F}_{\mathrm{OUT}}$ is set to the dimension of the initial features F_{IN}.

The GravNet layer makes the GravNet architecture stand out from other ML algorithms. Many network architectures like CNNs operate on data that contains ample relational information between elements. This approach assumes a regular or at least repeating grid for the input data [17]. Graphs do not assume any spatial structure and hence adapt to arbitrary detector geometry. In the ECL, this applies in particular to the endcaps which have irregular structures (see section 2.2.1). In addition, the end-to-end learning of the spatial information space allows GravNet to decide which nodes in a graph are the most relevant, unbiased by the actual detector geometry, and purely based on the features. However, this special nature of graphs is not indisputably a benefit. GravNet first has to confirm that it is actually able to learn a suitable representation of the geometry from the features.

4.2.2. Overall Structure

The GravNet layer alone is a fully operational ML architecture and sufficient for simple applications. For the studies in this work, the architecture is extended to reach the realm of deep learning. Three fully connected layers, a GravNet layer, and a batch norm layer are combined into a so-called GravNet block. Subsequently, a number of GravNet blocks are stacked to reach the desired depth. Figure 4.2 shows the resulting structure for three stacked GravNet blocks.

As the initial step, the input features are extended by appending the average of each feature per graph, effectively doubling the input size. This concept is called global exchange in the original paper since it acts as a global information gathering across all nodes of the graph. Additionally, because nodes now collect messages in successive message passing, information from nodes outside of the k connected neighbors is collected in the stacked GravNet blocks.

The full output of each GravNet block is used as an input for the next GravNet block. Additionally, the full output of each block is part of the input for the final fully connected layers by using skip connections as shown in figure 4.2. The final three fully connected layers then process the concatenated outputs of all blocks into the desired number of output classes. The number of clusters plus one class for background determines the number of output classes between either two or three (see section 4.1.1).

The architecture uses the exponential linear unit (ELU) [18] as an activation function for the first two fully connected layers of a GravNet block, and tanh-activation for the last fully connected layer of a block. A softmax function normalizes the final outputs of the overall architecture in order to guarantee the interpretation as fractions and the validity of equation (4.1).

The number of stacked GravNet blocks, the width of the fully connected layers, and the batch norm momenta are considered hyperparameters. A hyperparameter optimization presented in section 4.3.2 determines and lists the optimal parameters for the models. Ultimately, the architecture results in $\approx$ 16 000 learnable parameters of which $\approx$ 6000 are in the GravNet layers (more details in table 4.3). The low number of learnable parameters and computations distinguishes GravNet from other ML approaches. In comparison to FCCs and CNNs of usual complexity and comparable depth, the inference of GravNet is fast and memory-saving [17]. Furthermore, the input size for GravNet is smaller and variable, as only crystals with recorded energies are processed and crystals without information do not appear in the input. In combination with optimized implementations on specialized hardware like field-programmable gate arrays, this could open up real-time applications in future work.

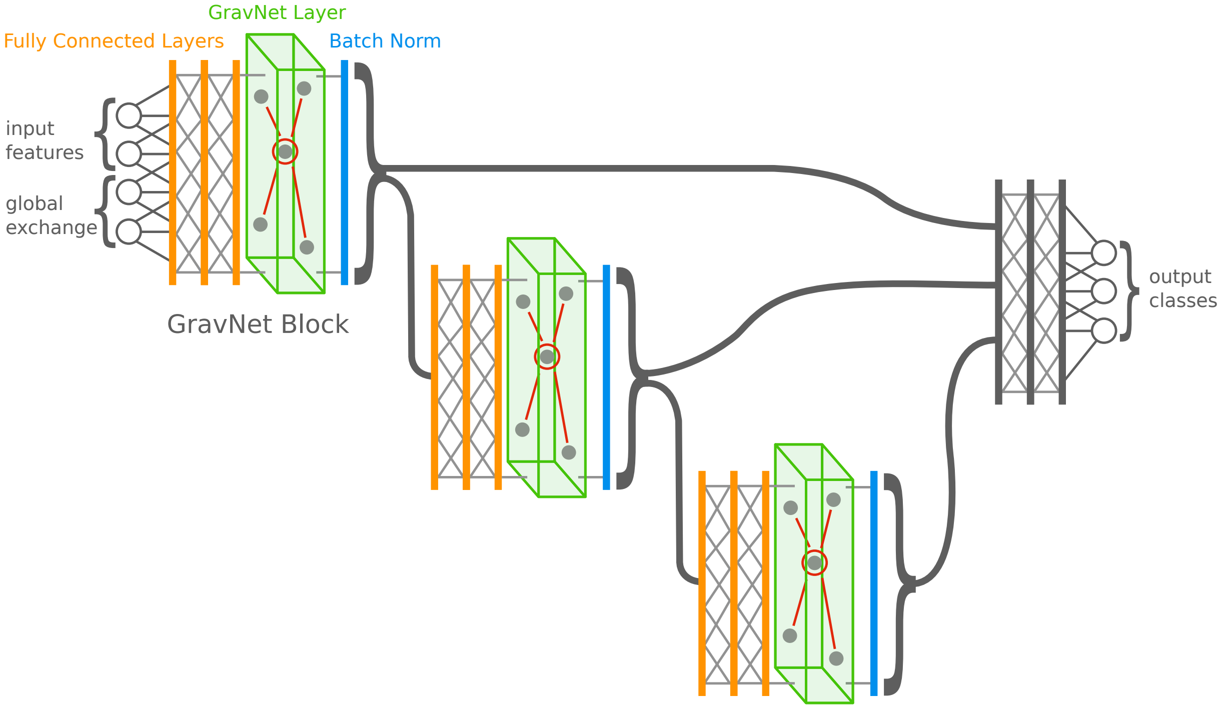


Figure 4.2.: Visualization of the overall GravNet architecture. The combination of three fully connected layers, GravNet layer, and batch norm layer forms a GravNet block. The exemplary architecture consists of three stacked GravNet blocks. A final fully connected network collects all GravNet block outputs and transforms them into the number of desired output classes.

4.3. Implementation

The overall architecture is implemented and trained using the ML library `PyTorch` [19]. The GravNet layers are implemented using the `GravNetConv` class from `PyTorch Geometric` [20]. A total of four distinct GravNet models are trained for the respective studies:

- **One cluster**: The architecture has two output features representing cluster 1 and background. One model is trained and evaluated on events with early background and one on events with nominal background. A hyperparameter optimization is carried out for the nominal background model and serves as the basis for the hyperparameters of the early background model. The events for training and evaluation fulfill the one-cluster criteria from section 3.2.1 and are associated with the one-cluster toy studies in section 6.2. In addition to the toy studies, the models evaluate the physics study events of weakly boosted pions in section 7.3.
- **Two clusters**: The architecture has three output features representing cluster 1, cluster 2, and background. One model is trained and evaluated on events with early background and one on events with nominal background. A hyperparameter optimization is carried out for the nominal background model and serves as the basis for the hyperparameters of the early background model. The events for training and evaluation fulfill the two-cluster criteria from section 3.2.2 and are associated with the two-cluster toy studies in section 6.2. In addition to the toy studies, the models evaluate the physics study events of highly boosted pions in section 7.4.

The focus is put on the optimization of the nominal background models, as basf2 reaches its limits in this future scenario and GravNet is expected to show the largest improvements. Section 4.3.1

gives an overview of the input features that all models have in common. Section 4.3.2 goes into the details of the training including a feature analysis and the hyperparameter optimization.

4.3.1. Input Features

The crystal measurements presented in section 2.2.1 and several crystal properties are the input features for GravNet. This section introduces these features and outlines the pre-processing that takes place to make these quantities suitable for ML. The general goal of the pre-processing is to normalize features to a range of [0,1] in order to value all features the same.

Crystal Measurements

The **recorded energy** E_{rec} per crystal in GeV is not subject to any pre-processing as the feature naturally ranges from 0.0 GeV to approximately 1.0 GeV in the studies presented here. High-energy photons usually deposit around 80 % of their energy in one central crystal of a cluster [9]. Given the energy range of photons in present work of up to 1.5 GeV, only a few outliers to the desired interval are expected and within tolerance.

The **recorded time** t_{rec} per crystal in μs is also not part of the pre-processing. The readout process caps the timing information at -1 μs to 1 μs. Albeit the interval is different than usual, it is similar enough and not found to cause issues. However, it may be an interesting aspect for investigations in future work.

The **PSD information** consists of three features: The hadron intensity HI, and its corresponding fit type and χ^2. HI ranges from HI ≤ 0.0 for fully electromagnetic showers to HI > 0.0 for hadronic showers. Since the exact hadron intensity does not benefit the present studies and only the discrimination between electromagnetic and hadronic showers is relevant, HI is capped at [-1, 1]. Once again the argument holds true, that the slightly different interval does not cause issues. The fit type indicates either the fitted template or a failed fit by four distinct values $\{-1, 0, 1, 2\}$ that are mapped to $\{0, 1/3, 2/3, 1\}$. The χ^2 of the corresponding template fit is divided by 200 and capped at 1 in order to contain 90 % of the distribution of χ^2 without taking extreme outliers into account.

Crystal Properties

The two **global coordinates** θ and ϕ are sufficient to locate each crystal in the ECL unambiguously, thus r is omitted. In contrast to Cartesian coordinates, spherical coordinates naturally represent the cyclical geometry of the detector. Min-max normalization via `scikit-learn` [21] globally scales $\theta \in [0, 2\pi]$ to the conventional interval [0, 1]. For $\phi \in [0, 2\pi]$ to actually represent a cyclical property including the edges of the interval, it is encoded in two features $\phi_{\text{sin}} = \sin\phi$ and $\phi_{\text{cos}} = \cos\phi$. After the transformation, min-max normalization globally scales both features to [0, 1]. The cyclical encoding of the coordinates can improve the performance for clusters that span over the interval gap in ϕ, however, this yet has to be fully confirmed in future work. The upcoming feature analysis includes a comparison between models with cyclical and non-cyclical coordinates that hints towards cyclical encoding being at an advantage.

The center of the event as defined in section 3.2 is the origin of the **local coordinates** θ' and ϕ' of the crystals in the ROI. θ' and ϕ' are the respective angular separations to the center in rad. Both local coordinates are explicitly not normalized globally but per event. This means that within the ROI, θ' and ϕ' extend over the entire interval [0,1].

One-hot features mark crystals that contain **LMs** in the recorded energy. One-cluster events have one LM feature, while two-cluster events have two LM features, each marking one LM.

The **masses** of the ECL crystals vary between 4.03 kg and 5.94 kg depending on their geometry and thereby indirectly on the location in the ECL. This range is globally normalized to [0,1].

4.3.2. Training

For each of the four GravNet models, two million events are generated and selected according to section 3.1. Out of the total events, 200 000 are chosen randomly as validation data set, leaving 1.8 million events for the actual training. Later on, for each study, additional test data sets are created independently but according to the exact same configuration and criteria.

The training is carried out in mini-batches of 1024 events for the one-cluster model, and 512 events for the two-cluster model. These batch sizes are the result of individual hyperparameter optimizations. The Adam algorithm [22] is used for the gradient-based optimization. The learning rate starts at $5 \times 10^{-}3$ and decays by a factor of 0.25 after every five epochs of plateauing validation loss. The training stops variably with conditional stopping depending on an additional objective. For early background models, the objective is the validation loss. The nominal background models employ the more profound energy resolution for the validation data set, denoted by $\mathrm{FWHM}_{\mathrm{dep}}$, as the additional objective. Smaller $\mathrm{FWHM}_{\mathrm{dep}}$ are better, the full definition of the measurement follows in section 5.3.2. Note that loss and $\mathrm{FWHM}_{\mathrm{dep}}$ are two distinct measurements that are not directly comparable. For both types of additional objectives, the training is stopped and the best configuration is saved after 15 epochs without improvements. In this configuration, the training takes approximately 3 1/2 hours on a single NVIDIA Titan X 12GB GPU, paired with an Intel Xeon E5-2630 CPU.

Figure 4.3 displays the loss progress for the nominal background models of GravNet with one cluster and two clusters. For these models, the conditional stopping depends on the $\mathrm{FWHM}_{\mathrm{dep}}$. The decaying learning rate manifests in the plateaus of the training loss. While the training loss is decaying steadily, both validation loss and $\mathrm{FWHM}_{\mathrm{dep}}$ fluctuate significantly before approaching a limit. Figure 4.4 visualizes the training progress with an event display of the GravNet prediction for an exemplary event from the validation data set at various epochs throughout the training. The initialization leads to a roughly equal prediction of each class before the first epoch. The visual differences to the prediction after the first epoch are large but expected given the model already trained on 1.8 million events at this point. The advances to the final epoch are far more subtle, yet GravNet is meant to improve the already excellent clustering of basf2 and the differences are not necessarily expected to be seen by the eye.

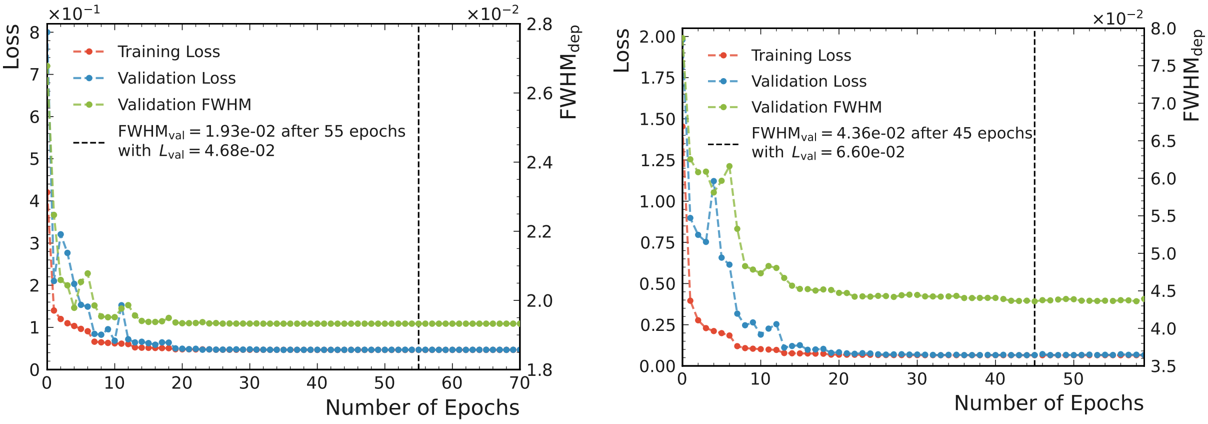


Figure 4.3.: Loss over progressing training epochs with conditional stopping at the indicated resolution $\mathrm{FWHM}_{\mathrm{val}}$ and at validation loss L_{val}. The left plot shows the loss for the one-cluster model of GravNet with nominal phase 3 background, and the right plot for the two-cluster model with nominal phase 3 background.

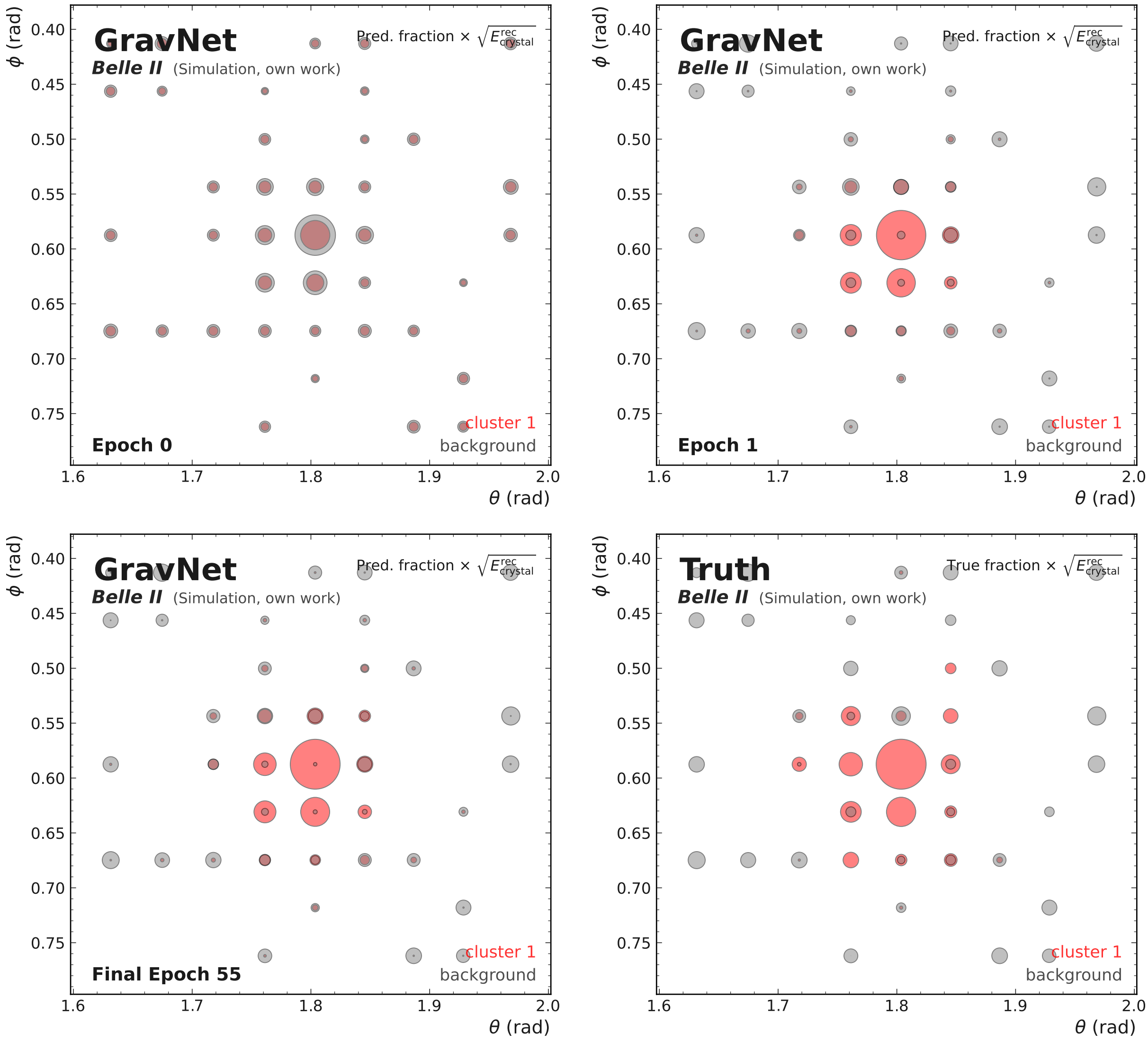
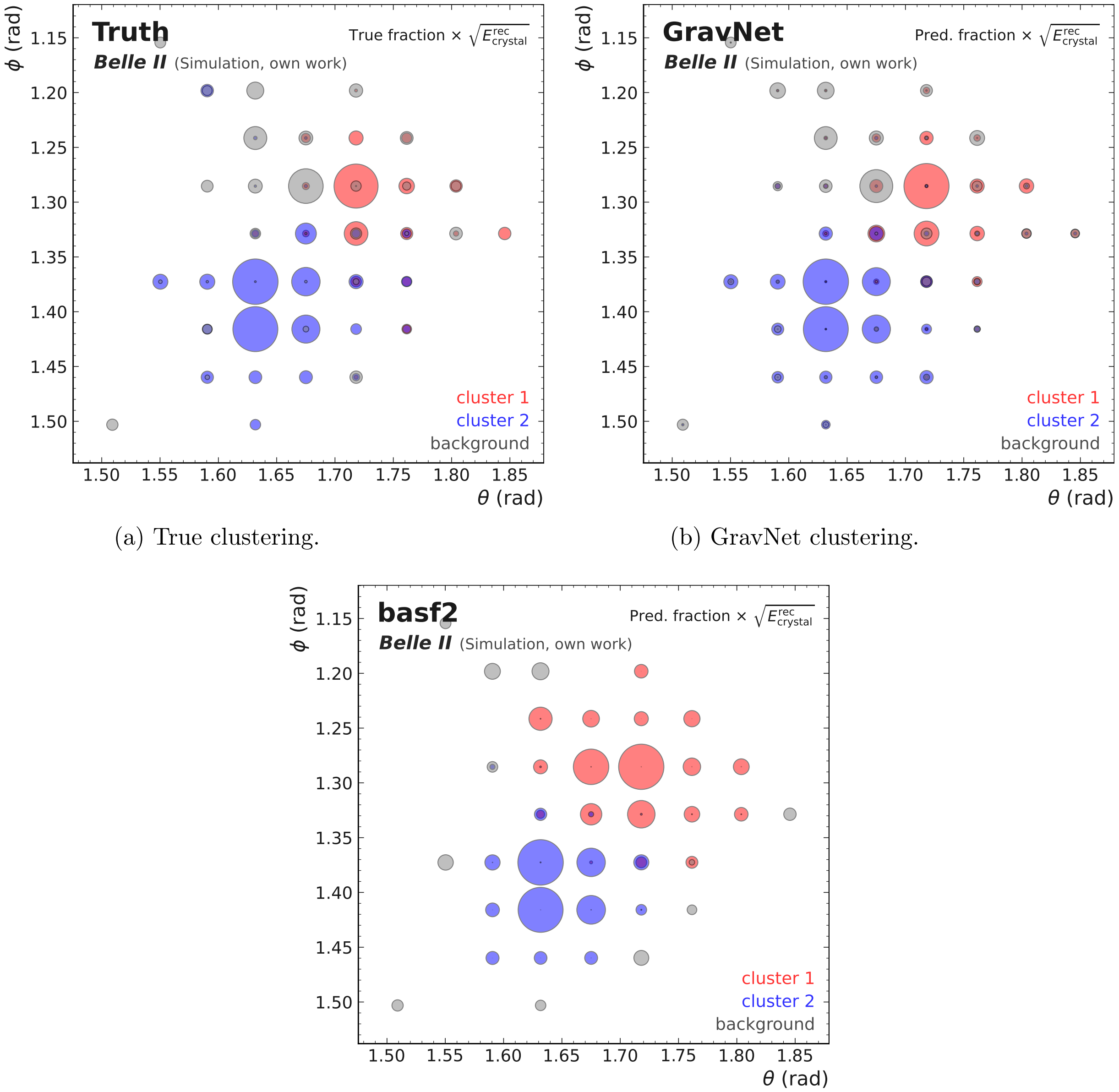


Figure 4.4.: Development of the clustering prediction throughout the training of GravNet. The four event displays show an event from the validation data set at different epochs. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\text{rec}}}$. The event fulfills the same criteria as the events in the one-cluster toy study with nominal phase 3 background in section 6.2.

On the other hand, a characteristic pattern of all GravNet models catches the eye: Even after the final epoch, the model continues to assign tiny fractions to all available classes in all crystals. This numerical behavior is the result of GravNet not being designed as a classification algorithm, but rather as a regression algorithm that is able to partially assign crystals to classes. A possible approach to mitigate that behavior would be to push GravNet more in the direction of hard clustering by setting thresholds to fully assign to, or remove crystals of a class. However, it needs clarification in future work whether it is necessary to revise this and if doing so indeed increases performance.

At the same time, the ability to partially assign crystals to background is the most prominent conceptual strength of GravNet in comparison to the basf2 baseline. Figure 4.5 highlights this ability in direct comparison to the basf2 prediction. In each crystal, basf2 assigns fractions of recorded energy to different particles that are associated with LMs. However, it does not partially assign any background energy depositions. Thus, in crystals with large overlap, GravNet depicts the underlying clustering more accurately and separates background and true energy clusters, whereas basf2 falsely includes the crystals in the clusters.



(a) True clustering.

(b) GravNet clustering.

(c) Basf2 clustering.

Figure 4.5.: Event display of the true clustering and the predicted clusterings of GravNet and basf2 for an exemplary event with large background overlap. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\text{rec}}}$. The event is from the highly boosted pion study with nominal phase 3 background in section 7.4.

Feature Analysis

The basic feature analysis carried out in this section aims to uncover the additional input features that drive the performance in comparison to the basf2 baseline. It can not substitute a full feature analysis that considers the correlations between features as well. Two-cluster nominal background events serve as the basis for the comparison of the performances of several models that only differ in their features. This scenario is associated with the two-cluster nominal background study in section 6.3. It is regarded the most challenging scenario for the algorithms and makes use of the full potential of the features. The performance is measured on a validation data set by the final validation loss, and the resolutions on the deposited energy $\mathrm{FWHM}_{\mathrm{dep}}$ and the generated energy $\mathrm{FWHM}_{\mathrm{gen}}$ that are explained in full detail in section 5.3.2. Table 4.1 lists the resulting performance measurements of six distinct models in comparison to the basf2 baseline. The corresponding plots of all models are in appendix A.2.

All models substantially outperform the basf2 baseline according to all measurements. This means that even the most basic model - which is comparable to basf2 in its feature set using only global coordinates, LMs, and the recorded energies - benefits from the conceptual differences of the ML approach. The analysis manifests that global coordinates in cyclical encoding slightly outperform non-cyclical encoding, leading to the decision to exclusively employ cyclical encoding. Local coordinates improve the performance even more and are therefore also part of the final feature set. The PSD information does not provide any improvements, however, this is expected due to the purely electromagnetic character of the events. The PSD information can be of high value in future scenarios that incorporate the clustering of hadronic energy depositions. The addition of the recorded time t_{rec} results in the biggest jump in performance, leading to the conclusion that in purely electromagnetic scenarios, timing information allows for better distinction of background energy depositions.

In order to achieve the best possible generalizability to any scenario, the final models combine all mentioned features as presented in section 4.3.1. This results in the best performance in the analysis, although the improvements over the model with just added t_{rec} are marginal.

Table 4.1.: The table compares the performances of GravNet models with different features and the performance of the basf2 baseline on the two-cluster nominal background validation data set. Listed are the resolutions on the deposited and on the generated energy $\mathrm{FWHM}_{\mathrm{dep,gen}}$, as well as the final losses. LM denotes the local maxima feature(s), E_{rec} is the recorded energy, PSD denotes the pulse shape discrimination information consisting of HI, χ^2-value and fit type, and t_{rec} is the recorded time.

Features	$\mathrm{FWHM}_{\mathrm{dep}}$ $\times 10^{-2}$	$\mathrm{FWHM}_{\mathrm{gen}}$ $\times 10^{-2}$	Final Loss $\times 10^{-1}$
Non-Cyclical Global Coord., LM, E_{rec}	4.54	7.13	7.66
Global Coord., LM, E_{rec}	4.50	7.07	7.61
Local Coord., LM, E_{rec}	4.47	7.07	7.71
Global Coord., LM, E_{rec}, PSD	4.52	7.09	7.61
Global Coord., LM, E_{rec}, t_{rec}	4.07	6.72	6.74
Global & Local Coord., Mass, LM, E_{rec}, t_{rec}, PSD	4.05	6.69	6.60
basf2	5.99	7.68	9.25

Hyperparameter Optimization

Subsequently to the selection of features, individual hyperparameter optimizations are carried out for the two GravNet models with nominal background. The hyperparameters found in this optimization are applied to the corresponding models with early background as well. The resolution $\mathrm{FWHM}_{\mathrm{dep}}$ on the respective validation data set is the objective according to which `Optuna` [23] determines the model architecture by varying:

- Width of the fully connected layers $\mathrm{F}_{\mathrm{IN}}, \mathrm{F}_{\mathrm{OUT}} \in [8, 32]$
- Dimension of the learned feature space $\mathrm{F}_{\mathrm{LR}} \in [8, 24]$
- Dimension of the learned spatial information space $\mathrm{S} \in [2, 8]$
- Number of nearest neighbors to connect in the spatial information space $k \in [10, 32]$
- Batch norm momentum $\in [0.0, 0.4]$
- Number of stacked GravNet blocks $\in \{3, 4, 5\}$
- Batch size for mini-batch optimization $\in [128, 4096]$

Table 4.2 presents the resulting hyperparameters for one-cluster and two-cluster models. Table 4.3 completes the overview over the resulting models by listing the parameters of specific layers.

Table 4.2.: The table lists the final hyperparameters of the one-cluster and two-cluster GravNet models. The hyperparameters are the result of an optimization of the resolution $\mathrm{FWHM}_{\mathrm{dep}}$ on the respective nominal background validation data set using `Optuna`.

Hyperparameter	One-Cluster Models	Two-Cluster Models
Width of the Fully Connected Layers, $\mathrm{F}_{\mathrm{IN}}, \mathrm{F}_{\mathrm{OUT}}$	22	24
Feature Space Dimension F_{LR}	16	16
Spatial Information Space Dimension S	6	6
Connected Nearest Neighbors k	14	16
Batch Norm Momentum	0.01	0.4
Stacked GravNet Blocks	4	4
Batch Size	1024	512

Table 4.3.: The table lists the number of parameters of specific layers in the GravNet architecture for the one-cluster and two-cluster GravNet models.

Layer	One-Cluster Models	Two-Cluster Models
Initial Fully Connected Layers	1606	1896
Initial GravNet Layer	1386	1462
Initial Batch Norm Layer	32	32
3× Stacked Fully Connected Layers	1386	1608
3× Stacked GravNet Layers	1386	1462
3× Stacked Batch Norm Layers	32	32
Final Fully Connected Layers	4363	4434
Total Number of Parameters	15 799	17 131

5. Metrics

This chapter proposes the metrics used to evaluate GravNet and the basf2 baseline in different scenarios. On the introduction of fundamental reconstruction quantities in section 5.1 follow two different types of metrics: The first set of metrics in section 5.2 defines properties for the characterization of events and clusterings. The second set of metrics in section 5.3 examines various aspects of the clustering performance by comparing predicted data as given by an algorithm and true data as given by the MC information.

A number of 200 000 test events for each study in chapters 6 and 7 provides statistical relevance and allows for the identification of trends. Consequently, all metrics are obtained per event and studied in histograms across all test events of a scenario. Shared energy (sec. 5.2.8), cluster center distance (sec. 5.2.9), cluster radius difference (sec. 5.2.6), and cluster energy difference (sec. 5.2.7) are designed specifically for events with two clusters and not used in events with just one cluster. On the other hand cluster, radius (sec. 5.2.5), reconstruction errors (sec. 5.3.1), and sensitivity and precision (sec. 5.3.3) are defined per cluster but used for events with two clusters as well. This is done by calculating the metrics for both clusters and taking the average. For this reason, they are not suitable for a direct comparison between one-cluster and two-cluster events.

For each metric, an example is given from the two-cluster toy study with early phase 3 background and clustering results from the baseline basf2 algorithm. Section 6.3 describes this scenario in detail. The plots are not interpreted yet, as this is done for the toy study in comparison with GravNet.

5.1. Fundamental Reconstruction Quantities

The weights $w_i^{(u)}$ define the clustering of an event, yet they do not entirely describe the underlying physics. In order to complete the concept of fuzzy clustering for measurements of performance, the weights have to be combined with the recorded energies per crystal E_i^{rec}. This yields the following fundamental quantities:

$$E_{i(u)}^{\text{dep}} = E_i^{\text{rec}} t_i^{(u)} \qquad \text{and} \qquad E_{i(u)}^{\text{pred}} = E_i^{\text{rec}} p_i^{(u)}. \tag{5.1}$$

Here $E_{i(u)}^{\text{dep}}$ corresponds to the true (deposited) amount of energy, and $E_{i(u)}^{\text{pred}}$ to the predicted amount of energy of cluster u in crystal i. As defined in section 3.2, a cluster is always associated with a photon. The terms are not interchangeable, but in the context of energies, it holds true that deposited and predicted energies also belong to the photon. Consequently,

$$E_{\text{dep}}^{(u)} = \sum_i E_{i(u)}^{\text{dep}} \qquad \text{and} \qquad E_{\text{pred}}^{(u)} = \sum_i E_{i(u)}^{\text{pred}} \tag{5.2}$$

define the total true (deposited) energy and the total predicted energy for a photon respectively. For better readability and clarity, the (u)-notation is, if possible, omitted in upcoming definitions. This is the case whenever metrics are defined per cluster, in these cases the quantities implicitly refer to that cluster.

5.2. Event Properties

The criteria for the event selection in section 3.2 lead to specific cluster signatures. The following metrics define properties that give further insight into those signatures, as well as into the resulting events as a whole. Their main purpose is to characterize, analyze and compare the events of different scenarios. Despite not directly measuring performance, most of them are also used to compare properties in different clustering results.

5.2.1. Leakage

Section 2.2.2 discusses leakage ΔE_{leak} and its causes in the ECL. Leakage has a special status because it is not comparing clusterings, but rather is an intrinsic property of clusters and events. It is defined per cluster using E_{dep} from equation (5.2) and the generated particle energy E_{gen}:

$$\Delta E_{\text{leak}} = E_{\text{gen}} - E_{\text{dep}}. \tag{5.3}$$

In contrast to other metrics, ΔE_{leak} is not averaged but added for two clusters. Figure 5.1 shows the leakage profile of the example data set. These distributions do not aim to uncover and explain details of leakage as a physics process, but rather give an estimation of the expected differences between the reconstruction errors on the deposited energy and on the generated energy that are described in section 5.3.1.

5.2.2. Sum of Weights

The sum of weights Σ_w reflects the total amount of weights that are assigned to a certain cluster:

$$\Sigma_w = \sum_i w_i. \tag{5.4}$$

For events with two clusters, Σ_w is calculated for both clusters and added up. Figure 5.2 shows the distribution of Σ_{w} for the true clustering of the example data set.

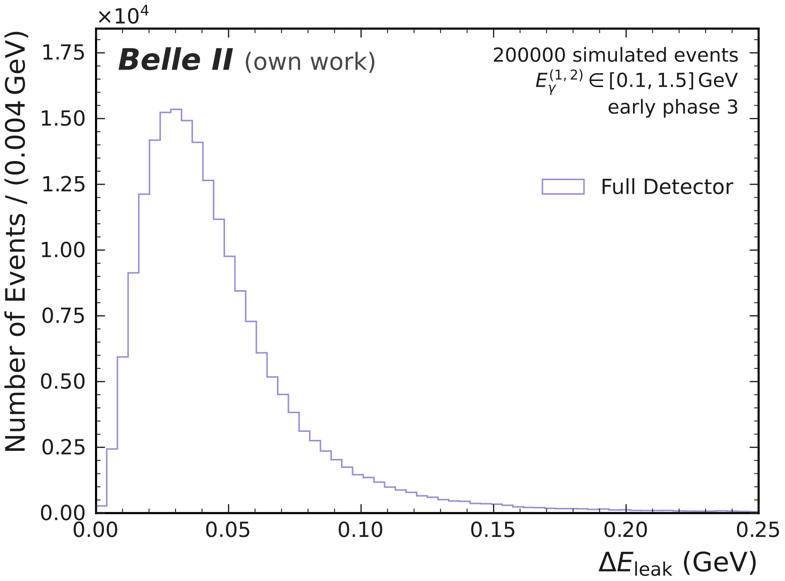


Figure 5.1.: Distribution of the leakages ΔE_{leak} for the example events from the two-cluster toy study with early phase 3 background in section 6.3.

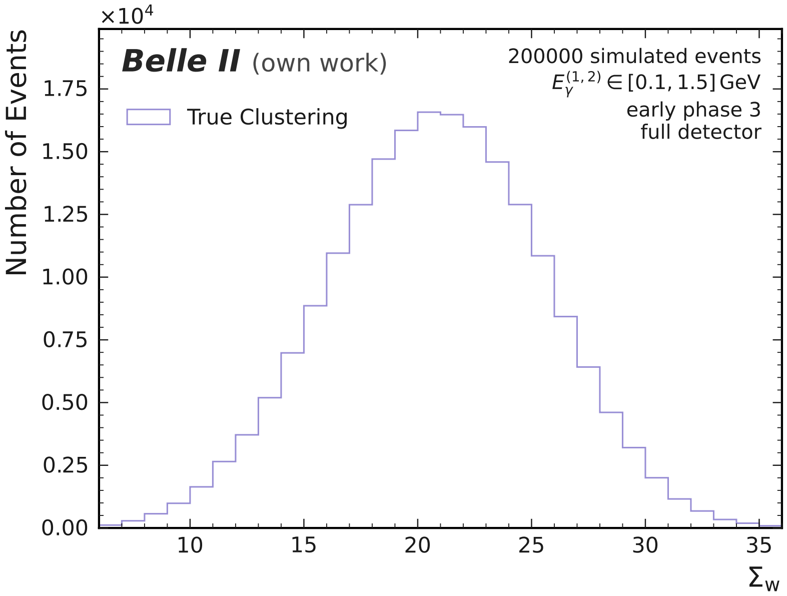


Figure 5.2.: Distribution of the sum of weights Σ_w for the true clustering in the example events from the two-cluster toy study with early phase 3 background in section 6.3.

5.2.3. Cluster Energy

The total cluster energy E_cluster according to equation (5.2) serves two purposes: On one hand, the distribution of E_dep is a sanity check for the event generation. A uniform distribution is expected for photons within a chosen interval. On the other hand, deviations in E_pred from the true uniformity reveal whether an algorithm misrepresents photons of a certain energy. For two-cluster events, the cluster energies are examined individually. This is relevant for GravNet which considers each cluster as a distinct class (see section 4.1.1) and could yield varying results.

Figure 5.3 shows the two distributions of E_cluster for cluster 1 and cluster 2 in the example data set. The event generation for this scenario produces two photons, each in a uniform interval $E_\mathrm{gen} \in [0.1, 1.5]$ GeV. Taking leakage into account, both distributions fulfill the expectations for a uniform distribution in this interval.

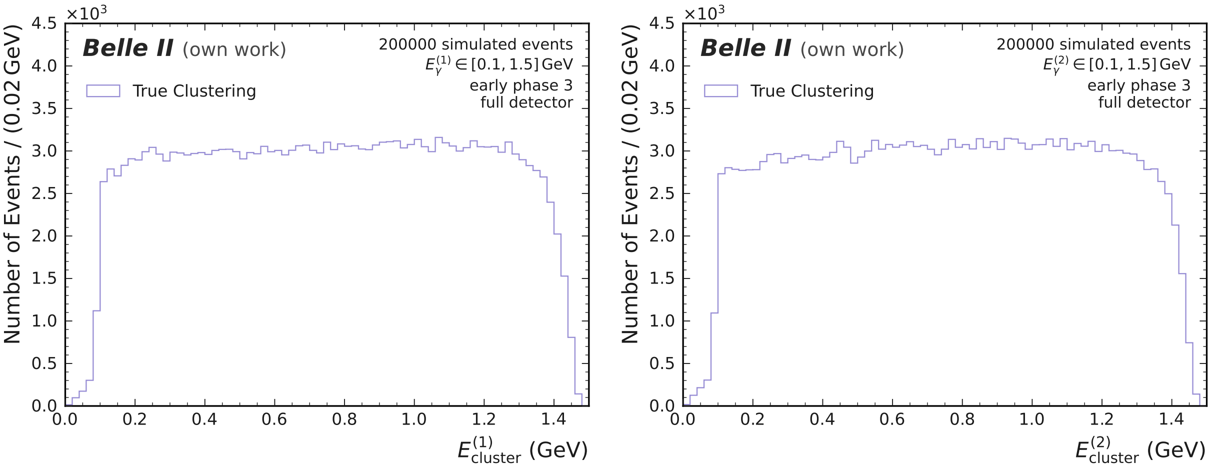


Figure 5.3.: Distribution of true cluster energies E_cluster for the example events from the two-cluster toy study with early phase 3 background in section 6.3. The left side depicts cluster 1, the right side cluster 2.

5.2.4. Cluster Center

The location $\vec{x}$ of the center of a cluster is calculated with the weighted average

$$\vec{x} = \frac{\sum_i \sqrt{E_i}\,\vec{x_i}}{\sum_i \sqrt{E_i}}. \tag{5.5}$$

Herein, $\vec{x_i}$ are the coordinates of crystal i, weighted by the energies E_i which can denote either predicted or deposited energies according to equation (5.1). Crystals with small energies gain more relevance due to the square root. Otherwise, the coordinates of the crystal with the LM dominate the cluster center. The scaling shifts the center in the direction of the likely, true point of impact of the particle.

Basf2 provides the coordinates of the centers of the crystals. The center is located around 15 cm within the crystal and not directly at the surface that faces the IP. In addition to that potential inaccuracy, ignoring the detector geometry in the calculation comes with another disadvantage. Metrics using cluster centers, defined in that manner, implicitly assume that the detector is flat within the local environment of an event. This works well for events that are located totally within the barrel or the endcaps, whereas the approximation is less accurate for the few events that are located right at the gaps between the barrel and the endcaps.

5.2.5. Cluster Radius

The cluster radius R describes the dimension of a cluster around the cluster center. Starting with the center as defined in equation (5.5), these steps follow:

1. Calculate the Euclidean distances d_i of all crystals in the cluster to the center.
2. In ascending order in distance, the energies in the crystals are summed up to the first crystal n where $\left(\sum_i^n E_i\right)/E_\text{cluster} \geq 0.95$. Here, E_i are either true or predicted energies according to equation (5.1), and E_cluster is the corresponding cluster energy as in equation (5.2).
3. This ratio and the ratio of the last crystal n' with $\left(\sum_i^{n'} E_i\right)/E_\text{cluster} < 0.95$ are the basis for a linear interpolation that is described in appendix B.1.
4. The interpolation yields a distance in between d_n and $d_{n'}$ that is set as the cluster radius R.

Looking at various cluster signatures, a cut-off at 95 % is a good compromise between full coverage of the cluster and the exclusion of extreme outliers. Figure 5.4 shows R drawn into an example event display. Figure 5.5 then shows the distribution of all radii R in the example data set.

5.2.6. Cluster Radius Difference

The cluster radius difference ΔR is explicitly defined for events with two clusters. It is an indication of asymmetry from a geometric point of view. The absolute difference between the cluster radii R_1 and R_2 as defined in the previous section 5.2.5 is obtained by:

$$\Delta R = |R_1 - R_2|\,. \tag{5.6}$$

Figure 5.6 shows the distribution of ΔR for the example data set.

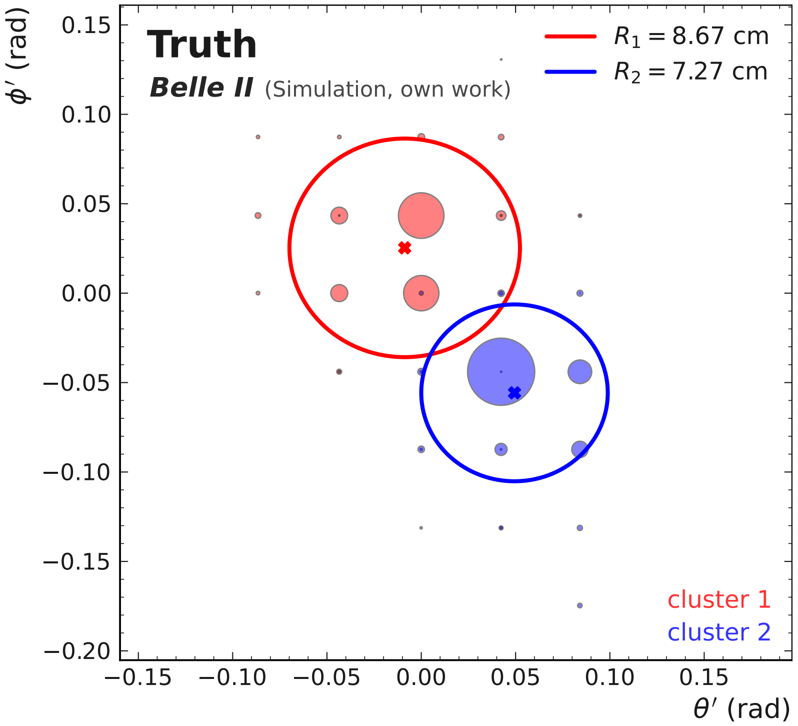


Figure 5.4.: Event display including the cluster radii R_1 and R_2. The local coordinates of the event are θ' and ϕ'. Background is excluded for better visibility of the true, radius defining energy depositions. The event is from the two-cluster toy study with early phase 3 background in section 6.3.

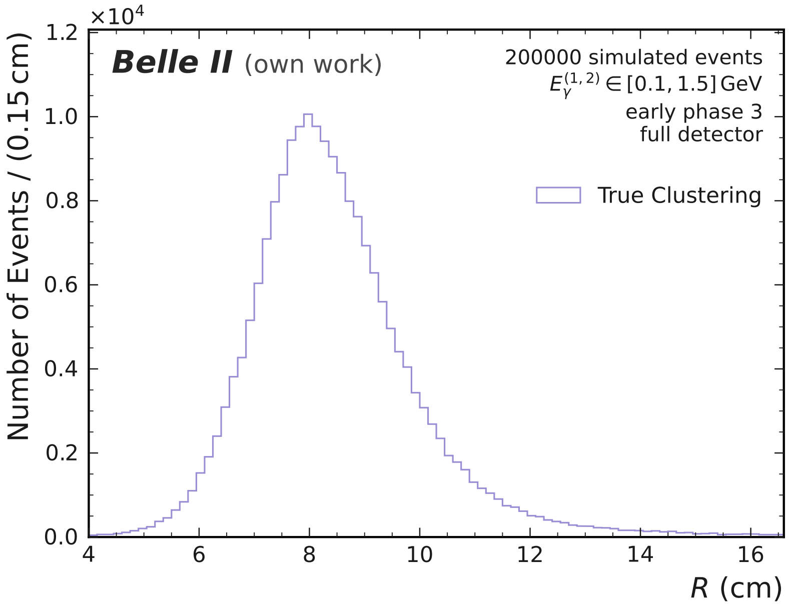


Figure 5.5.: Distribution of cluster radii R for the example events from the two-cluster toy study with early phase 3 background in section 6.3.

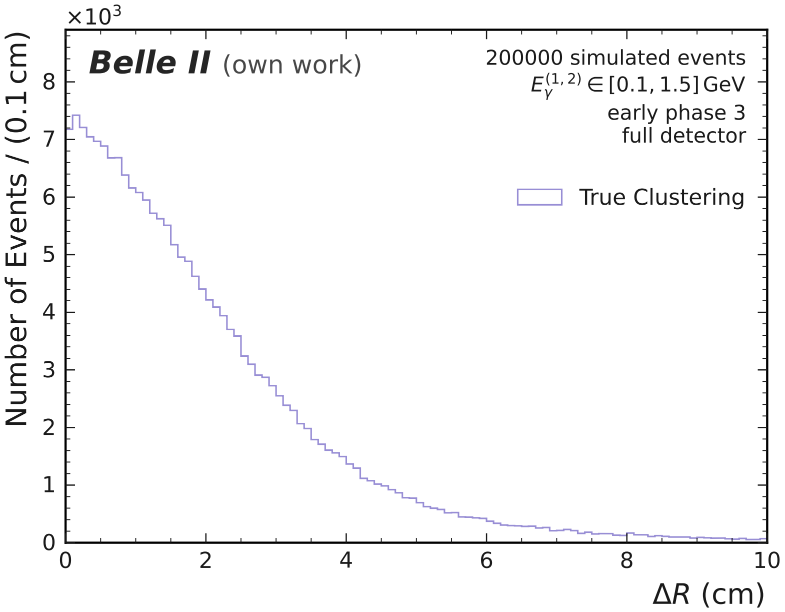


Figure 5.6.: Distribution of cluster radius differences ΔR for the example events from the two-cluster toy study with early phase 3 background in section 6.3.

5.2.7. Cluster Energy Difference

The cluster energy difference $\Delta E_{\text{cluster}}$ measures the asymmetry of two clusters from the perspective of energy:

$$\Delta E_{\text{cluster}} = \left| E_{cluster}^{(2)} - E_{cluster}^{(1)} \right| , \tag{5.7}$$

where $E_{cluster}^{(1,2)}$ are either true or predicted cluster energies as in equation (5.2). Figure 5.7 shows the distribution of $\Delta E_{\text{cluster}}$ for the example data set.

Together, $\Delta E_{\text{cluster}}$ and ΔR identify whether an algorithm is able to correctly separate strongly asymmetric clusters or if it neglects one cluster (intuitively the lower energetic / smaller one) in favor of the other.

5.2.8. Shared Energy

The shared energy E_{shared} characterizes overlap for events with two clusters. It measures the energy that the two clusters claim in common crystals above a threshold $E_{\text{th}} = 1\,\text{MeV}$. Using $E_i^{(1,2)}$ from equation (5.1), the shared energy is defined as

$$E_{\text{shared}} = \sum_i \left(E_i^{(1)} + E_i^{(2)} \right) \Theta \left(E_i^{(1)} - E_{\text{th}} \right) \Theta \left(E_i^{(2)} - E_{\text{th}} \right) , \tag{5.8}$$

with the Heaviside step function Θ.

Figure 5.8 shows the distribution of E_{shared} for the example data set. Most events fall within the peaking side of the distribution towards zero. Large E_{shared} means that one photon, despite having an individual LM, deposited most of its energy in the same crystals as the other.

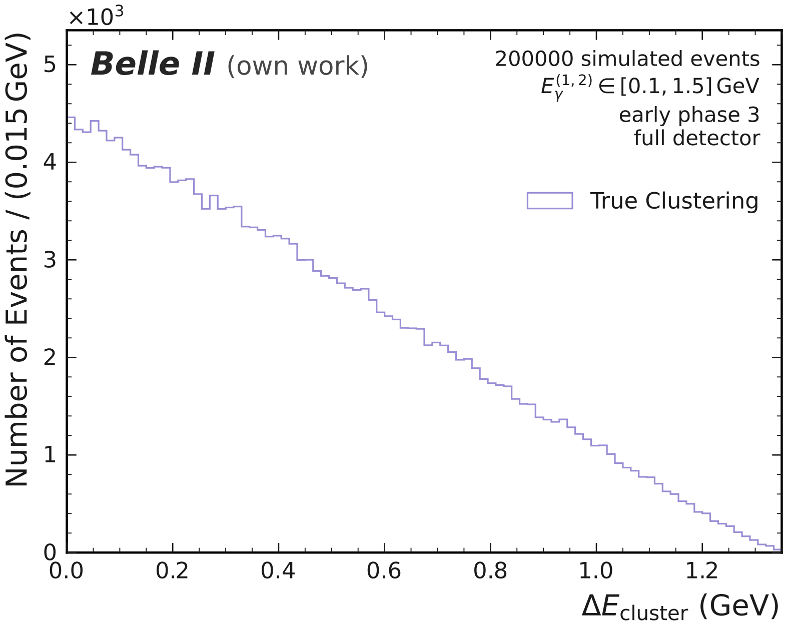


Figure 5.7.: Distribution of cluster energy differences $\Delta E_{\text{cluster}}$ for the example events from the two-cluster toy study with early phase 3 background in section 6.3.

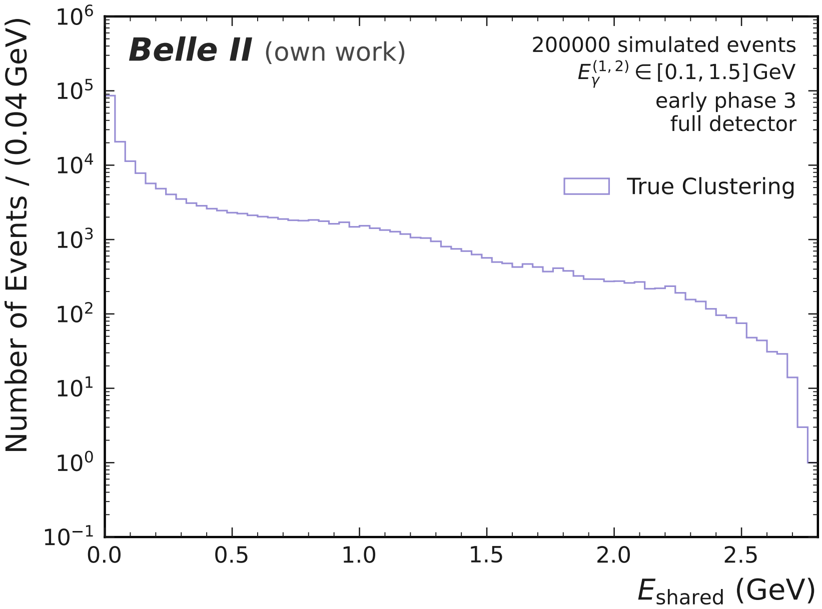


Figure 5.8.: Distribution of shared energies E_{shared} for the example events from the two-cluster toy study with early phase 3 background in section 6.3. The logarithmic scaling on the y-axis puts focus on the long tail of the distribution.

5.2.9. Cluster Center Distance

The cluster center distance Δx is another estimation of overlap specifically for events with two clusters. It is defined as the Euclidean distance between the cluster centers $\vec{x_{1,2}}$ according to section 5.2.4:

$$\Delta x = \sqrt{(\vec{x_1} - \vec{x_2})^2}. \tag{5.9}$$

Figure 5.9 shows the distribution of Δx for the example data set. Because Δx lacks information about the dimensions of the clusters (see cluster radius in section 5.2.5) it is not necessarily the case that close-by clusters also deposited a large amount of energy in identical crystals. The shared energy E_{shared} complements Δx accordingly.

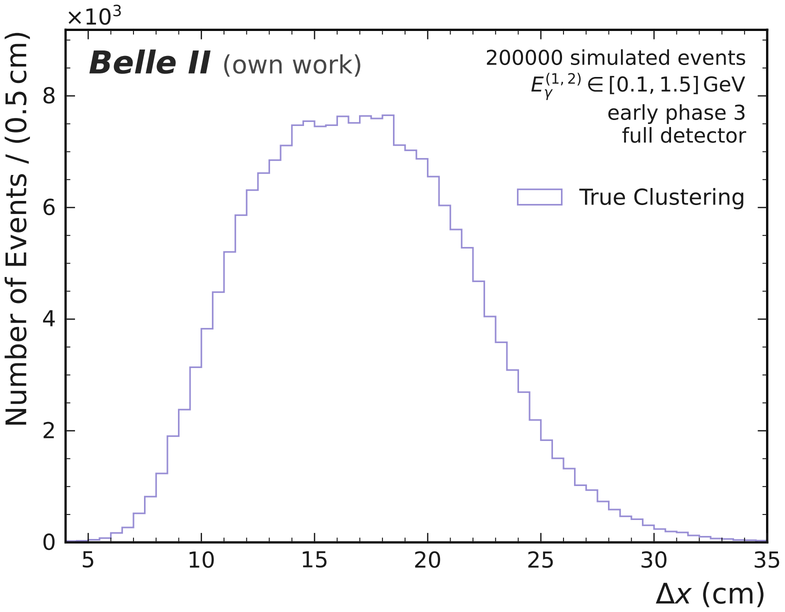


Figure 5.9.: Distribution of cluster center distances Δx for the example events from the two-cluster toy study with early phase 3 background in section 6.3.

5.3. Performance Evaluation

Using MC information, it is possible to compare predicted energy fractions to true energy fractions, which is the basis for the training of GravNet. However, analyzing and comparing just the weights does not fully assess the performance of the algorithms. For this reason, the metrics proposed in this section outline various aspects of the clustering of entire events and deliver a more intuitive representation of the performance of the algorithms. The reconstruction errors in section 5.3.1 directly evaluate the primary objective, which is to improve the photon energy reconstruction. Section 5.3.2 describes the determination of the energy resolution from the reconstruction errors. Sensitivity and precision in section 5.3.3 and the fuzzy clustering agreement index in section 5.3.4 each study aspects of the underlying clustering in more detail that are not accessible considering the photon energy resolution only.

5.3.1. Reconstruction Errors

The most relevant measurement associated with the primary objective is the relative reconstruction error between the true and predicted energy for a photon. This work studies two versions of relative reconstruction errors (or short reconstruction errors) depending on which quantities one wants to compare: The error on the deposited energy without detector effects and the error on the generated energy including detector effects. Looking at the distribution of reconstruction errors for all events in a test data set reveals a peak. The peak is characterized by the two properties full width half maximum and tail lengths. In the present work, these are the most important quantities for comparing algorithms and are described in detail in section 5.3.2.

Reconstruction Error on the Deposited Energy

The reconstruction error on the deposited energy η_{dep} is given by

$$\eta_{\mathrm{dep}} = \frac{E_{\mathrm{pred}} - E_{\mathrm{dep}}}{E_{\mathrm{dep}}}, \tag{5.10}$$

with the comparison between the predicted energy of a photon E_{pred} and the total energy the photon deposited E_{dep} as defined in equation (5.2).

η_{dep} gives access to the photon energy resolution solely based on the clustering and leaving out detector effects (primarily leakage). In contrast to other common definitions, the reconstruction errors in this work do not consist of absolute values of errors. This way the reconstruction errors distinguish whether an algorithm underestimates or overestimates energy. Figure 5.10 shows the full distribution of η_{dep} for the basf2 algorithm evaluated on the example data set.

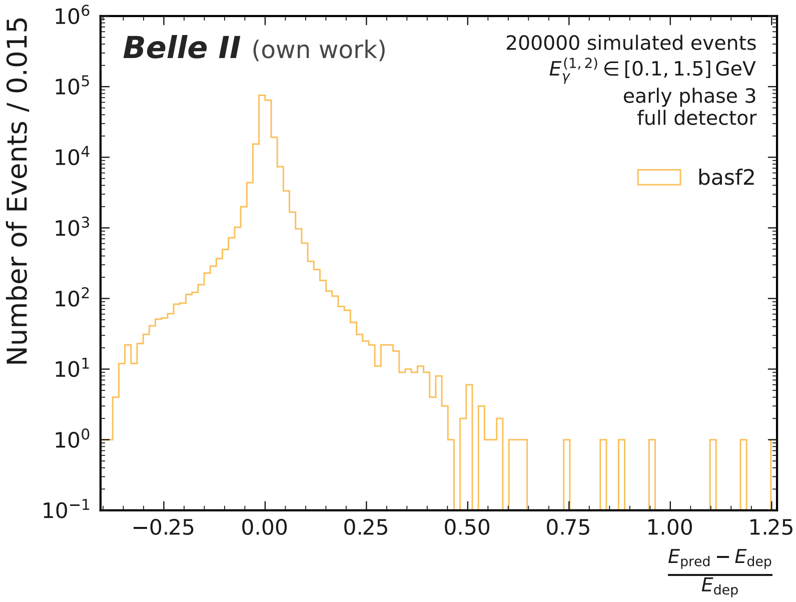


Figure 5.10.: Example of a distribution of reconstruction errors on the deposited energy $f\eta_{\mathrm{dep}}$. The logarithmic scale on the y-axis puts focus on the outliers. The results are from the two-cluster toy study with early phase 3 background and clustering of the baseline basf2 algorithm in section 6.3.

Reconstruction Error on the Generated Energy

The reconstruction error on the generated energy η_{gen} is defined as

$$\eta_{\text{gen}} = \frac{E_{\text{pred}}^{\text{cor}} - E_{\text{gen}}}{E_{\text{gen}}}, \tag{5.11}$$

and compares the generated photon energy $E_\gamma = E_{\text{gen}}$ to the leakage-corrected, predicted energy

$$E_{\text{pred}}^{\text{cor}} = \begin{cases} \text{clusterE} & \text{basf2} \\ E_{\text{pred}} & \text{GravNet.} \end{cases} \tag{5.12}$$

For basf2, the advanced and leakage-corrected version of the total photon energy, namely clusterE introduced in section 2.3, is the basis for comparison. GravNet trains with E_{dep} as ground truth. It is not aware of any detector effects and does not correct them, hence the $E_{\text{pred}}^{\text{cor}}$ is identical to E_{pred} in equation (5.2). The addition of a leakage correction mechanism, either as a separate step like in basf2 or within the GravNet training, is considered an important direction for future work.

The reconstruction error on the generated energy η_{gen} factors in detector effects and quantifies how much of the improvements to the underlying clustering carry over to further physics reconstruction. While it is theoretically possible to achieve perfect (zero) error on the deposited energy with perfect clustering, limitations of the hardware affect η_{gen}. For this reason, η_{gen} on average is significantly larger than η_{dep}. The effect appears in form of a wider, smeared-out peak when comparing the full distribution of η_{gen} in figure 5.11 with the peak of the η_{dep} distribution in figure 5.10. This is especially noticeable in the tail to the left side of the η_{gen} distribution.

Because it reflects the results that truly remain in the reconstruction process, arguably η_{gen} can be considered the more important measurement. However, η_{dep} demonstrates the full potential of an algorithm and thus is just as relevant in light of different (less) hardware limitations.

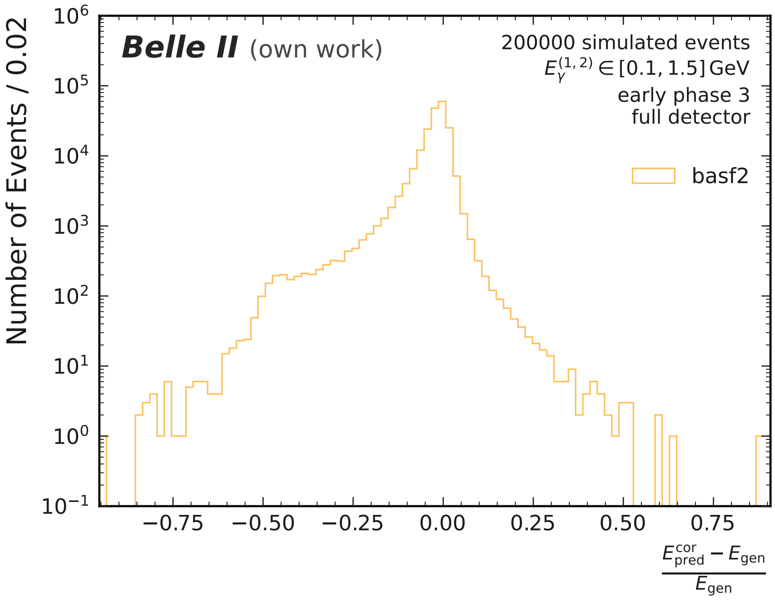


Figure 5.11.: Example of a distribution of reconstruction errors on the generated energy η_{gen}. The logarithmic scale on the y-axis puts focus on the outliers. The results are from the two-cluster toy study with early phase 3 background and clustering of the baseline basf2 algorithm in section 6.3.

5.3.2. Energy Resolution

The full width half maximum (FWHM) characterizes the distributions of η_{dep} and η_{gen} shown in the previous section 5.3.1. Within the distributions, the majority of evaluated events are found in the peak. The FWHM outlines the width of the peak and thereby represents an algorithm's performance on a test data set in a single measurement. The FWHM is also referred to as resolution. A more smeared-out peak leads to a larger FWHM which in turn is interpreted as a worse resolution. Besides the peak itself, the tails of the distributions are relevant for the characterization of outliers and are treated in the next paragraph.

Given a distribution of reconstruction errors $f(\eta)$, the FWHM is defined as

$$\text{FWHM} = |\eta_1 - \eta_2|. \tag{5.13}$$

η_1 and η_2 satisfy the following relation to a maximum $f(\eta_{\text{max}})$ of $f(\eta)$

$$f(\eta_1) = f(\eta_2) = \frac{1}{2} f(\eta_{\text{max}}). \tag{5.14}$$

In order to find the FWHM, an unbinned χ^2-fit is performed using the `zfit` package [24]. First, a rough estimation of the peak dimension results in the fit range

$$r_{\text{fit}} = \left[r_{\text{L}}^{\text{fit}}, r_{\text{R}}^{\text{fit}} \right], \tag{5.15}$$

defined by the left and right limits $r_{\text{L,R}}^{\text{fit}}$. Looking at the histogram of the full distribution as pictured in figures 5.10 and 5.11, the bin containing the maximum number of events n_{max} is identified. Moving to the left from this bin, the position of the first bin i with a number of events $n_i \leq 0.05 \cdot n_{\text{max}}$ sets the left limit $r_{\text{L}}^{\text{fit}}$. Identically, $r_{\text{R}}^{\text{fit}}$ is obtained by moving from the maximum bin to the right. Limiting r_{fit} ensures a good fit in the region of the peak instead of focusing on the tails.

The fitted double-sided crystal ball function (DCB) is defined as

$$f(x; N, \mu, \sigma, \alpha_{\text{L}}, n_{\text{L}}, \alpha_{\text{R}}, n_{\text{R}}) = N \cdot \begin{cases} A_{\text{L}} \cdot (B_{\text{L}} - \frac{x-\mu}{\sigma})^{-n}, & \frac{x-\mu}{\sigma} < -\alpha_{\text{L}} \\ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & -\alpha_{\text{L}} \leqslant \frac{x-\mu}{\sigma} \leqslant \alpha_{\text{R}} \\ A_{\text{R}} \cdot (B_{\text{R}} - \frac{x-\mu}{\sigma})^{-n}, & \frac{x-\mu}{\sigma} > \alpha_{\text{R}} \end{cases} \tag{5.16}$$

with

$$A_{\text{L/R}} = \left(\frac{n_{\text{L/R}}}{\left|\alpha_{\text{L/R}}\right|} \right)^{n_{\text{L/R}}} \cdot \exp\left(-\frac{\left|\alpha_{\text{L/R}}\right|^2}{2} \right), \qquad B_{\text{L/R}} = \frac{n_{\text{L/R}}}{\left|\alpha_{\text{L/R}}\right|} - \left|\alpha_{\text{L/R}}\right|. \tag{5.17}$$

The DCB consists of a central Gaussian defined by the mean μ and the standard deviation σ. The tails adjoining to the left and right are defined by the transition factors $\alpha_{\text{L,R}}$ and the exponent of the tail $n_{\text{L,R}}$. The height is given by N.

After performing the fit, the FWHM of the fit function is obtained analytically. Because the DCB, as well as the underlying distribution, are asymmetric, left and right $\text{FWHM}_{\text{L,R}}$ are obtained separately. Appendix B.2 describes the full process. Correlated uncertainties on the fit parameters lead to the uncertainty δ_{FWHM} on the FWHM. The `uncertainties` package [25] propagates the uncertainties throughout the calculation.

To ensure a fair comparison between different algorithms, as the final step a correction of FWHM and δ_{FWHM} leads to:

$$\text{FWHM}_{\text{cor}} \pm \delta^{\text{cor}}_{\text{FWHM}} = (1-\mu)\cdot\Big(\text{FWHM} \pm \delta_{\text{FWHM}}\Big). \tag{5.18}$$

The correction is equivalent to shifting the peak and centering the underlying distribution around zero. In the corrected form, FWHM_{cor} accounts for the undesired bias of a distribution towards negative or positive reconstruction errors and is interpreted as a simple, virtual leakage correction. Without it, the FWHM puts algorithms with a sharp peak far off the center at an unrealistic advantage that gets lost in any correction process. In this form, the performance of GravNet is comparable to the baseline and a proper leakage correction can be added in future work. For the sake of readability, the corrected FWHM_{cor} are referred to as $\text{FWHM}_{\text{dep,gen}}$ depending on the type of reconstruction error.

Tail Lengths

After focusing solely on the peak of the distributions of η_{dep} and η_{gen}, the tail lengths characterize the rest of the distribution, especially outliers. Because of the potentially large asymmetry, the left tail length r_{L} and the right tail length r_{R} are studied separately. The same steps are carried out for the calculation of r_{L} and r_{R} individually:

1. Split the distribution in half at the peak position μ.
2. Determine the total number of events on the respective side of the distribution.
3. Moving away from μ, count the number of events up to 90 % of the total number of events.
4. The absolute distance between μ and the reconstruction error of the last event within the 90 % sets $r_{\text{L/R}}$.

The tail lengths are similar to an asymmetric 90th percentile of a zero-centered distribution. The process is repeated twice for a split at $\mu - \delta_\mu$ and at $\mu + \delta_\mu$ to find the uncertainties $\delta_r^{\text{L,R}}$ on $r_{\text{L,R}}$ that are induced by the fit.

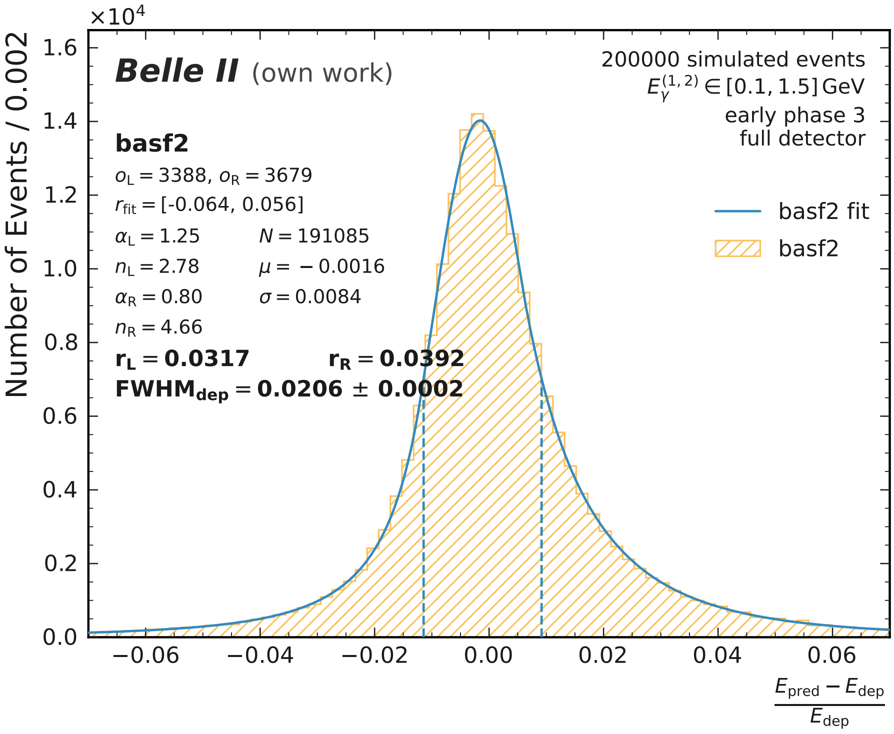
Subsequently, $r_{\text{L,R}}$ and the corresponding $\delta_r^{\text{L,R}}$ are corrected in the same way as the FWHM in equation (5.18) to account for a bias in the distribution:

$$r^{\text{cor}}_{\text{L/R}} \pm \delta^{\text{L/R}}_{r,\text{cor}} = (1-\mu)\cdot\Big(r_{\text{L/R}} \pm \delta_r^{\text{L/R}}\Big)\,. \tag{5.19}$$

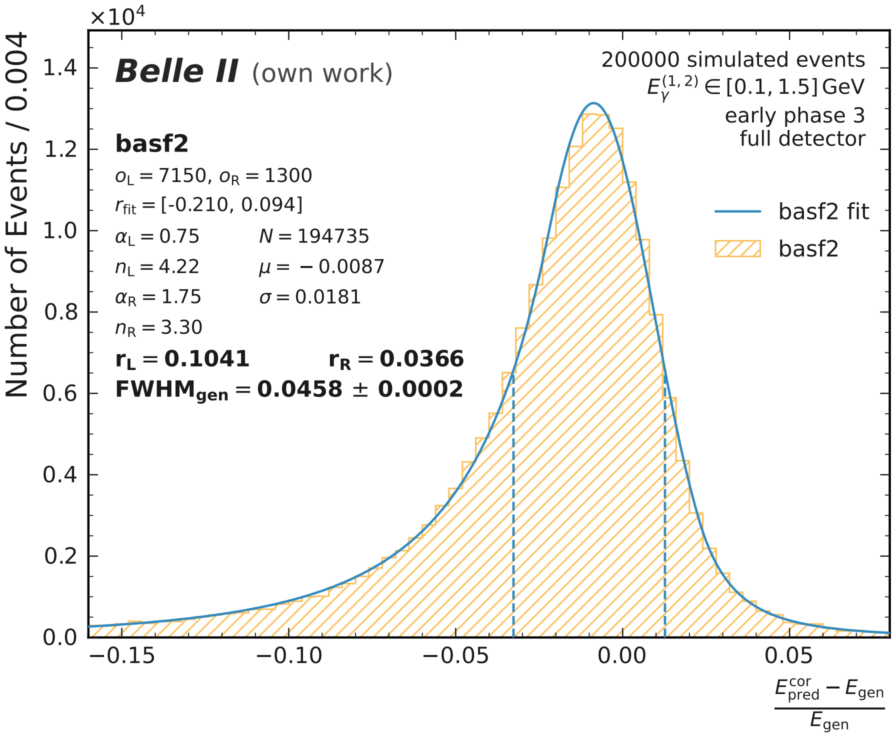
Analogous to the FWHM, the label for the corrected version $r^{\text{cor}}_{\text{L/R}}$ is omitted and henceforth $r_{\text{L/R}}$ always refers to the corrected lengths.

Figure 5.12a shows the distribution of η_{dep} for the example data set with the focus on the peak. Because of the focus on the peak, the overflow values $o_{\text{L,R}}$ specify how many of the total events are not displayed on each side of the distribution. All quantities $o_{\text{L,R}}$, r_{fit}, the fit parameters, as well as the resulting FWHM and $r_{\text{L,R}}$ are indicated within the figure. Given $\mathcal{O}(\delta_r^{\text{L,R}}/r_{\text{L,R}}) \approx 0.1\,\%$, the uncertainties on the tail lengths turn out to be minuscule and are therefore omitted in the limited space of the plot. Since the fits are carried out on unbinned data, the histogram depicted in the plot is not the basis for the fit and serves only a visual purpose. Figure 5.12b shows the example plot for the distribution of η_{dep} including the same information.

η_{dep} as well as η_{gen} are averaged over both clusters for two-cluster events. The advantage of this approach is that FWHM and $r^{\text{L,R}}_{\text{tail}}$ of the resulting distributions still hold true to the interpretation per event.



(a) Reconstruction errors on the deposited energy $f(\eta_{dep})$.



(b) Reconstruction errors on the generated energy $f(\eta_{gen})$.

Figure 5.12.: Example of a double-sided crystal ball fit for the distributions of the two variants of reconstruction errors. The overflow $o_{L,R}$, the fit range r_{fit}, the fit parameters, as well as the resulting $FWHM_{dep}$ and the tail lengths $r_{L,R}$ are indicated in the figure. The results are from the two-cluster toy study with early phase 3 background and clustering results of the baseline basf2 algorithm in section 6.3.

Energy Dependence

So far, the distributions of $\eta_{\text{dep,gen}}$ are only studied with a test data set that covers a broad spectrum of photon energies. For the analysis of the energy dependence of the resolution, photons are generated at various fixed energies. The resolution for each set of fixed-energy photons is then determined individually according to section 6.2.2. Plotting the resulting FWHMs over the generated photon energies E_γ reveals a relationship that is modeled by the inverse square function:

$$\text{FWHM}_{\text{dep,gen}}(E_\gamma) = \frac{a}{\sqrt{E_\gamma(\text{GeV})}} - b. \tag{5.20}$$

$\text{FWHM}_{\text{dep,gen}}$ respectively characterize the peaks of the distributions $\eta_{\text{dep,gen}}$, yielding the same relation but different parameters. The energy resolution of the ECL crystals, as well as the subsequent energy reconstruction, are highly dependent on the energy of the incident particle. The study of the energy dependence gives access to the full performance spectrum of the algorithms.

In this form, the analysis is only carried out for events with one photon cluster. Consequently, figure 5.13 shows an example of the energy dependence of FWHM_{gen} for one-cluster toy study events with early phase 3 background and clustering of the baseline basf2 algorithm in section 6.2. The resolution for one-cluster events is not directly comparable to that of two-cluster events. The energy dependence for two-cluster events is approached in another way in section 6.3.

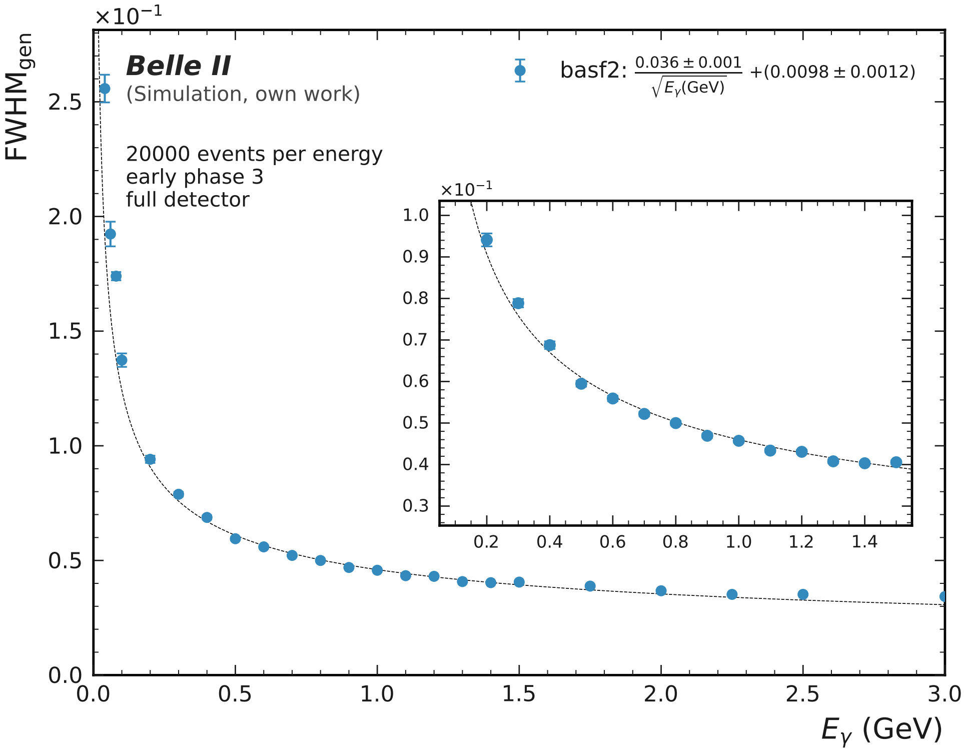


Figure 5.13.: Example of the resolution dependence $\text{FWHM}_{\text{gen}}(E_\gamma)$ for the reconstruction error on the generated energy η_{gen}. Each data point represents the FWHM_{gen} of 20 000 photons at the specific energy. The results are from the one-cluster toy study with early phase 3 background and clustering of the baseline basf2 algorithm in section 6.2.

5.3.3. Sensitivity and Precision

Sensitivity S and precision P are two common metrics in classification tasks and therein defined as

$$S = \frac{TP}{P} \quad \text{and} \quad P = \frac{TP}{TP + FP}, \tag{5.21}$$

with TP being the number of correctly detected cases by an algorithm, P being the total number of positive cases in a data set, and FP being the number of incorrectly detected cases [26]. Motivated by this definition it is viable to apply the concept to amounts of energy:

$$\begin{aligned} S_{\text{avg}} &= \frac{1}{n}\sum_{i}^{n} \frac{\min\left(E_i^{\text{dep}}, E_i^{\text{pred}}\right)}{E_i^{\text{true}}}, \\ P_{\text{avg}} &= \frac{1}{n}\sum_{i}^{n} \frac{\min\left(E_i^{\text{dep}}, E_i^{\text{pred}}\right)}{E_i^{\text{pred}}}. \end{aligned} \tag{5.22}$$

E_i^{pred} denotes the predicted energy, independent of it being true or false, that is equivalent to $P = TP + FP$. In correspondence to T, E_i^{dep} is the true (deposited) energy. The minimum of the two quantities is identified as the amount of correctly predicted energy. Sensitivity and precision are calculated separately for each crystal i and averaged over all crystals n that belong to the cluster. For numerical stability of the division, in this context, a crystal belongs to a cluster if its weight fulfills $w_i^{(u)} \geq 5\,\%$.

As a result of the averaging, S_{avg} and P_{avg} do not represent the performance for the event as a whole. Looking at the small reconstruction errors in figure 5.12a, it is evident that for the total energies in an event, sensitivity and precision would spike at one. However, in this crystal-wise formulation, they are an interesting measurement as well as a sanity check for algorithms that make decisions per crystal. As usual, for events with two clusters, the average of both clusters is taken to determine the event-wise value.

Sensitivity and precision are also referred to as true positive rate and positive predictive value. Analogous to the application of these metrics to classification tasks, there is a constant trade-off between them. High sensitivity is only worth something if the algorithm does accurately detect true energy only - that is to have high precision - and vice versa. Thus S_{avg} and P_{avg} are depicted in one plot revealing their correlation as well as marginal distributions. Figure 5.14 presents the plot for the example data set.

5.3.4. Fuzzy Clustering Agreement Index

The two previous metrics are focused on performance in regard to the energy reconstruction. The fuzzy clustering agreement index (FCAI) proposed by Rabbany and Zaïane [3] complements the performance evaluation by providing intuitive access to the quality of the underlying clustering. The quality of a clustering is defined by the agreement or similarity between the true clustering and the predicted clustering. The FCAI is an extension to the normalized mutual information that generalizes to fuzzy clustering and rates the agreement between two clusterings W and V on a scale from 0 (completely disjoint) to 1 (identical). Analogous to the mutual information, the FCAI determines the accordance at each data point (crystal) and is subsequently normalized to

$$\text{FCAI} = \frac{\mathcal{O}_{WV} - \mathcal{E}_{WV}}{\frac{1}{2}\left(\mathcal{O}_{WW} + \mathcal{O}_{VV}\right) - \mathcal{E}_{WV}}. \tag{5.23}$$

$\mathcal{O}_{WV}$ is the accordance of all pairwise comparisons between W and V, $\mathcal{O}_{WW}$ is the self-accordance of W, $\mathcal{O}_{VV}$ is the self-accordance of V, and $\mathcal{E}_{WV}$ is the expected accordance of the two clusterings

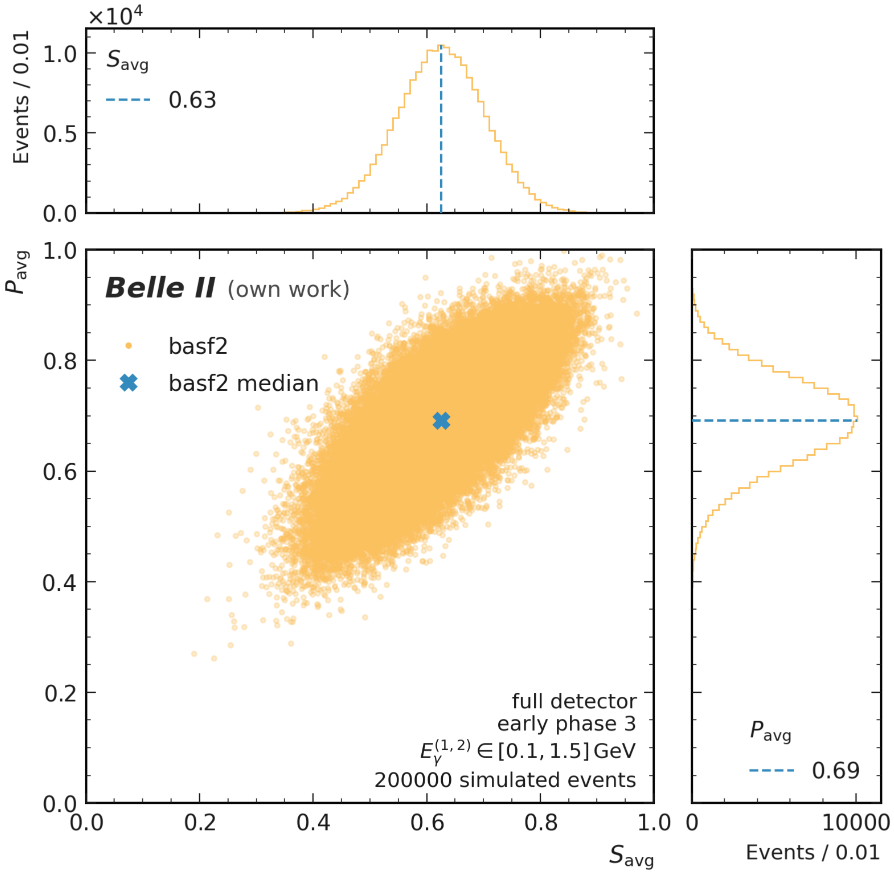


Figure 5.14.: Example of the distributions and correlations of the average sensitivity S_{avg} and precision P_{avg}. Each point in the 2D plot represents one event, in spite of that, sensitivity and precision are metrics regarding a single crystal and are averaged over all crystals in a cluster. The results are from the two-cluster toy study with early phase 3 background and clustering of the baseline basf2 algorithm in section 6.3.

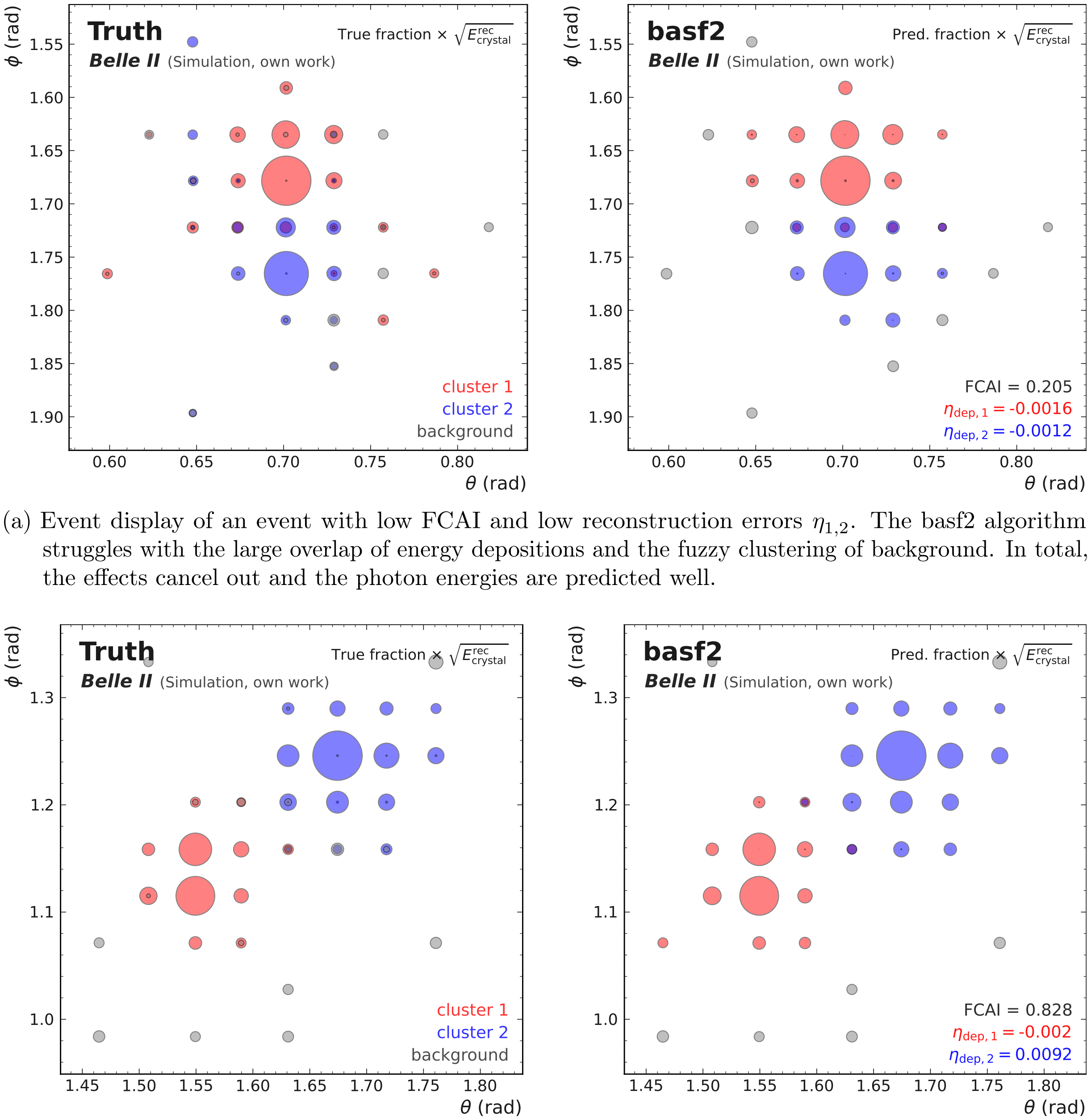
by chance. A detailed definition of these quantities and the pairwise comparison is part of appendix B.3.

Figure 5.15 shows a comparison of two events that have visibly different clustering qualities. Although the effect on photon energy resolution is negligible, the effects of the different clusterings on the FCAI are clear. For this reason, the FCAI can help identify problematic events that lead to the failure of an algorithm and are not as obvious in derived metrics like η.

Figure 5.16 shows the distribution of FCAIs of the basf2 baseline on the example data set including the median, which is the basis for comparison of different algorithms and is also referred to as the FCAI score for a given data set.

The definition of the FCAI is applied to one-cluster, as well as two-cluster events. Unlike previous metrics, the FCAI takes only the weights as the basis for comparison and is unique in treating background as an independent class besides the photon cluster(s). This approach comes with the caveat that no conclusions for the energy resolution should be drawn entirely based on the FCAI. To the FCAI it is irrelevant how much energy a correctly or incorrectly assigned crystal contains. All crystals are weighted the same and low-energy crystals have the same impact as more important high-energy ones. However, a high FCAI > 0.5 indicates that the algorithm works correctly and is not achieving good energy resolution by chance. The case in which two clusterings in a given setting are identical by chance would result in FCAI = 0. This work primarily uses the FCAI to verify that the energy resolution is backed up by the clustering quality.

While not the primary goal for now, the ability to accurately depict the true clustering can become more relevant in future scenarios. One outlook with noteworthy impact is an application to hardware with less leakage where the causality between clustering and resolution is more distinct.



(a) Event display of an event with low FCAI and low reconstruction errors $\eta_{1,2}$. The basf2 algorithm struggles with the large overlap of energy depositions and the fuzzy clustering of background. In total, the effects cancel out and the photon energies are predicted well.

(b) Event display of an event with higher FCAI and larger reconstruction errors $\eta_{1,2}$. Little overlap in the event leads to an accurate clustering result including in the shared crystals. Yet the reconstruction errors deteriorate due to the wrong assignment of background.

Figure 5.15.: Shown are two events with distinct FCAI. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\mathrm{rec}}}$. The plots to the left present the underlying true clustering. The right side depicts the basf2 clustering result including the FCAI and the two reconstruction errors $\eta^{(1,2)}_{\mathrm{dep}}$. The events are from the two-cluster toy study with early phase 3 background in section 6.3.

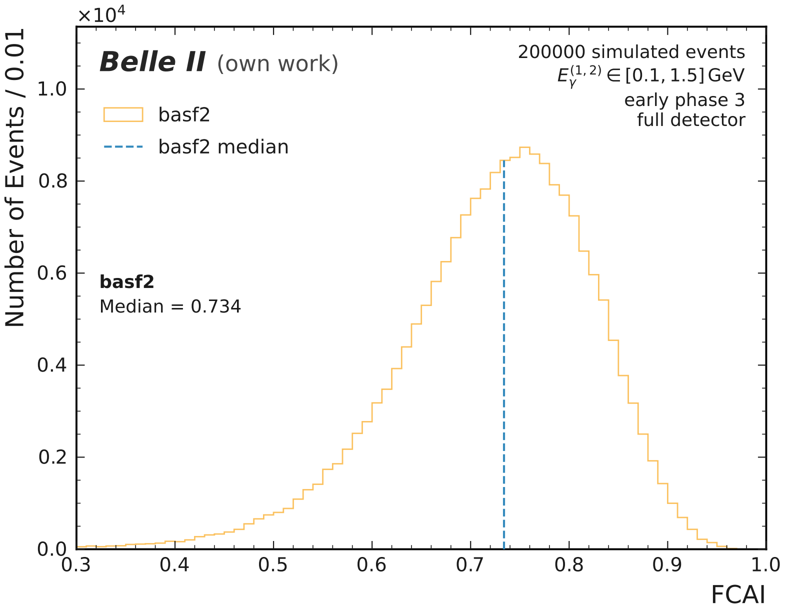


Figure 5.16.: Example of a distribution of FCAIs including the median from the two-cluster toy study with early phase 3 background and clustering results of the baseline basf2 algorithm in section 6.3.

6. Toy Studies

It is paramount for any physics application to gain a deep understanding of the performance of an algorithm and its underlying functioning. The toy studies in this chapter aim to evaluate and comprehend the behavior of the algorithms in a simple and well-controlled environment. This is especially important for neural networks like GravNet, where the high number of calculations relative to classical algorithms makes it nearly impossible to trace back results to certain operations. The four GravNet models introduced in chapter 4 are trained and optimized using the toy studies and only then brought to application in further physics reconstruction in the next chapter 7.

Two scenarios are of interest: One cluster originating from one photon, and two overlapping clusters originating from two photons. Both scenarios are studied with early background as well as nominal background. For each scenario, the characteristics and performances of the two types of backgrounds are analyzed together and compared. In addition, several metrics point out different and interesting behavior of the algorithms depending on the detector region. Therefore, whenever relevant, the studies separately examine the barrel, forward endcap, and backward endcap instead of the full detector.

The first section 6.1 motivates and presents the settings used in the event generation. The following studies of the two scenarios assess the same aspects and hence are structured identically: In sections 6.2.1 and 6.3.1, the events and the algorithms are examined using the event properties as defined in section 5.2. The next parts in sections 6.2.2 and 6.3.2 evaluate the performance according to the metrics in section 5.3. This includes the most important part of the studies, namely the analysis of the energy dependence of the resolution. Lastly, sections 6.2.3 and 6.3.3 concisely summarize and review the results of each study.

6.1. Motivation and Settings

One photon clusters are the most basic and abundant signature in the ECL. This scenario gives first insights into the general behavior and performance of the algorithms. The two corresponding studies lay the foundation for the physics studies with weakly boosted neutral pions in section 7.3. Two overlapping photon clusters are another common and more challenging signature occurring in the case of highly boosted particles. The physics studies in section 7.4 deal with this scenario for highly boosted neutral pions. Another possible application are the new physics decays from light dark photons and light axion like particles. Because of their numerous occurrences, improvements to these scenarios affect a wide range of physics analyses [6].

To get an idea of the energy range of photons in realistic experiment conditions, the decays of $\mathrm{B}^{+}\,\mathrm{B}^{-}$ and $\mathrm{B}^{0}\,\overline{\mathrm{B}}^{0}$ at Belle II are looked at in simulation. Figure 6.1 shows the energy spectrum of photons originating from these decays, where photons up to the fourth generation in the decay chain are taken into account. Many photons fall within an interval of $[0.1, 1.5]$ GeV that

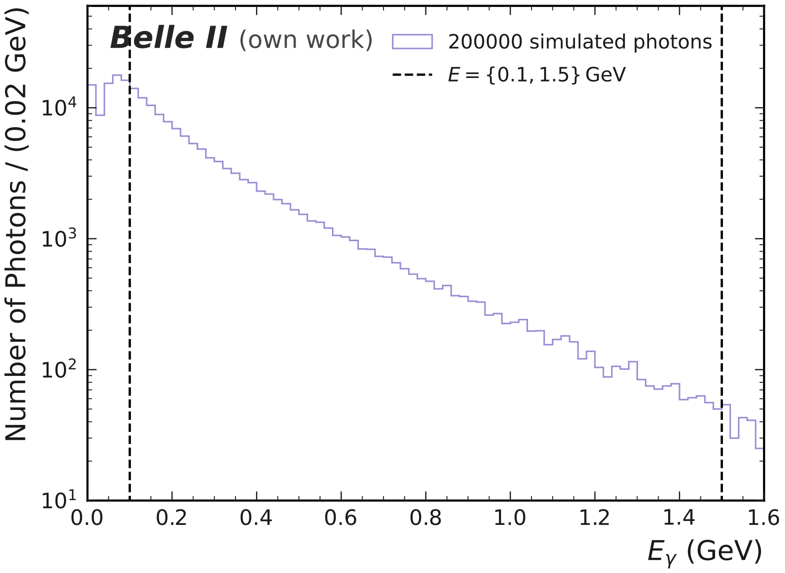


Figure 6.1.: Photon energy E_γ spectrum for simulated photons, originating from $\mathrm{B}^+\,\mathrm{B}^-$ and $\mathrm{B}^0\,\overline{\mathrm{B}}^0$ decays at Belle II. Photons up to the fourth generation in the decay chain of the initial particles are taken into account.

is highlighted in the plot. This interval is chosen for the generation of photons for both toy study scenarios. Lower energies are considered too challenging for the initial studies. In addition, regardless of the high number of low-energy photons, these are mostly due to general radiation processes and not relevant to physics analyses. Extending the interval to even higher energies even though there are only a few photons, potentially limits the ability of GravNet to specialize to photon clusters with more frequently occurring energies. However, GravNet is still able to generalize to energies outside of its training range to some extent. The energy dependence analysis in section 6.2.2 demonstrates the ability to generalize to lower, as well as to higher energetic photons.

The detector is separated into barrel and endcaps with significantly dissimilar characteristics, as is described in section 2.2. For all studies, the detector regions are defined by intervals in azimuthal angle θ that are stated in table 6.1. Some buffer is left to the edges of the individual components to ensure that the cluster is fully contained within the detector part. Thus the geometry of the studied detector regions is not exactly identical to the actual detector parts. The different regions have in common that polar angle ϕ is set to cover the whole circumference.

Table 6.1.: Intervals in θ for the event generation in different detector regions.

Detector Region	θ (deg)
Full Detector	$[17.0, 150.0]$
Barrel	$[37.2, 123.7]$
Forward Endcap	$[17.0, 31.36]$
Backward Endcap	$[131.5, 150.0]$

6.2. Single Photon Cluster

The single particle gun, introduced in section 3.1, creates events with a single photon. The momentum is set to $p_\gamma = E_\gamma \in [0.1, 1.5]\,\mathrm{GeV/c}$, the angles are set to $\phi \in [0.0, 360.0]\,^\circ$, and θ depending on the detector region according to table 6.1. Over several events, this yields photons that are distributed uniformly in space and momenta within the given ranges. Subsequently, either early or nominal background is added. The ROI then has to fulfill the specifications for one-cluster events given in section 3.2.1. After the selection, for each study, a total of two million events are left for the training of GravNet, and 200 000 events for the testing of the algorithms. Events for training and for testing are generated independently but use the same settings and fulfill the same criteria. While keeping comparability, this ensures that GravNet is not evaluated on the same events it is trained on.

One model of GravNet is trained and analyzed using events that cover the full detector range and are combined with early background. A second model, including a hyperparameter optimization (see section 4.3.2), trains on and evaluates exclusively events with nominal background. The basf2 baseline algorithm is the same for both types of backgrounds.

The studies compare the two GravNet models with their respective backgrounds to the basf2 baseline, as well as among each other. Whenever noteworthy, separate plots are shown for the different detector regions: Barrel, forward endcap, and backward endcap. The same full-coverage GravNet is used for the evaluation of all detector regions. For all metrics in all detector regions, extended plots including fit parameters are in appendix C.

6.2.1. Event Properties

The first part of the analysis focuses on the characterization of the test events using the properties defined in section 5.2. This section contains only metrics that present interesting behavior and outline differences between the algorithms. Less distinct properties and metrics that are used as a sanity check are in appendix C.1.

Leakage

Leakage ΔE_leak is examined as intrinsic propriety of the events. Figure 6.2 shows the distribution of ΔE_leak for both early and nominal background in different detector regions. The distributions roughly peak at 0.02 GeV for early background events and at 0.035 GeV for nominal background events. The peaks of the endcaps, especially in the backward endcap, are smeared out. The leakage will not show up in the reconstruction error on the deposited energy η_dep, however, it does have an effect on the reconstruction error on the generated energy η_gen. Therefore, a larger difference between the reconstruction error on the deposited energy η_dep and η_gen is expected in the endcaps.

Sum of Weights

The sum of weights Σ_w displays only minor variations in different detector regions. One conspicuity is the backward endcap for nominal background events. Figure 6.3 shows the distribution for the full detector for both types of backgrounds, as well as the backward endcap for nominal background. Larger Σ_w means that on average more percentage of the total energy of an event (independent of any detector effects) belongs to the photon cluster. For this reason, the peaks of the true clustering are shifted towards smaller Σ_w in the presence of nominal background. GravNet specifically trains with nominal background and models that characteristic well, whereas basf2 is not able to adapt to the smaller weights.

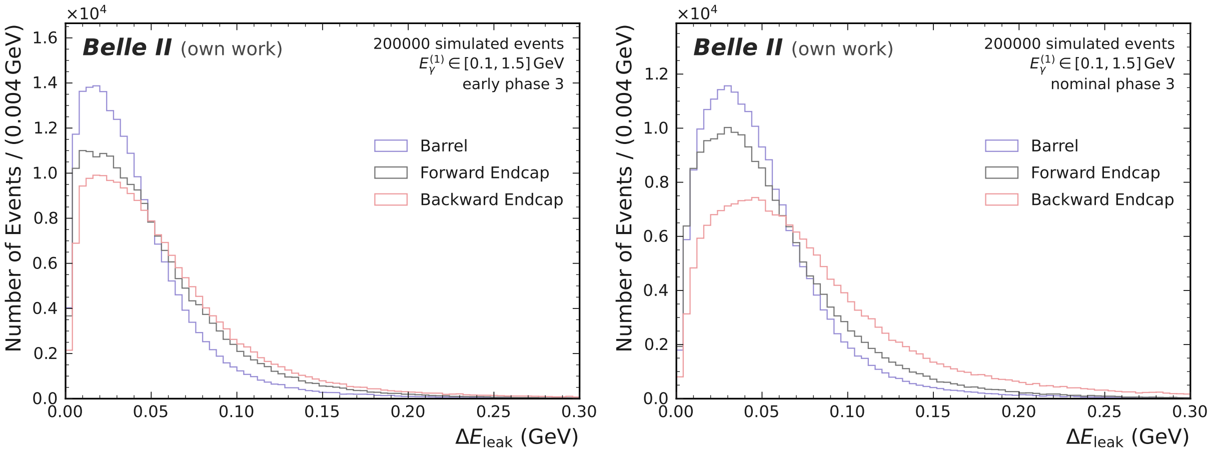
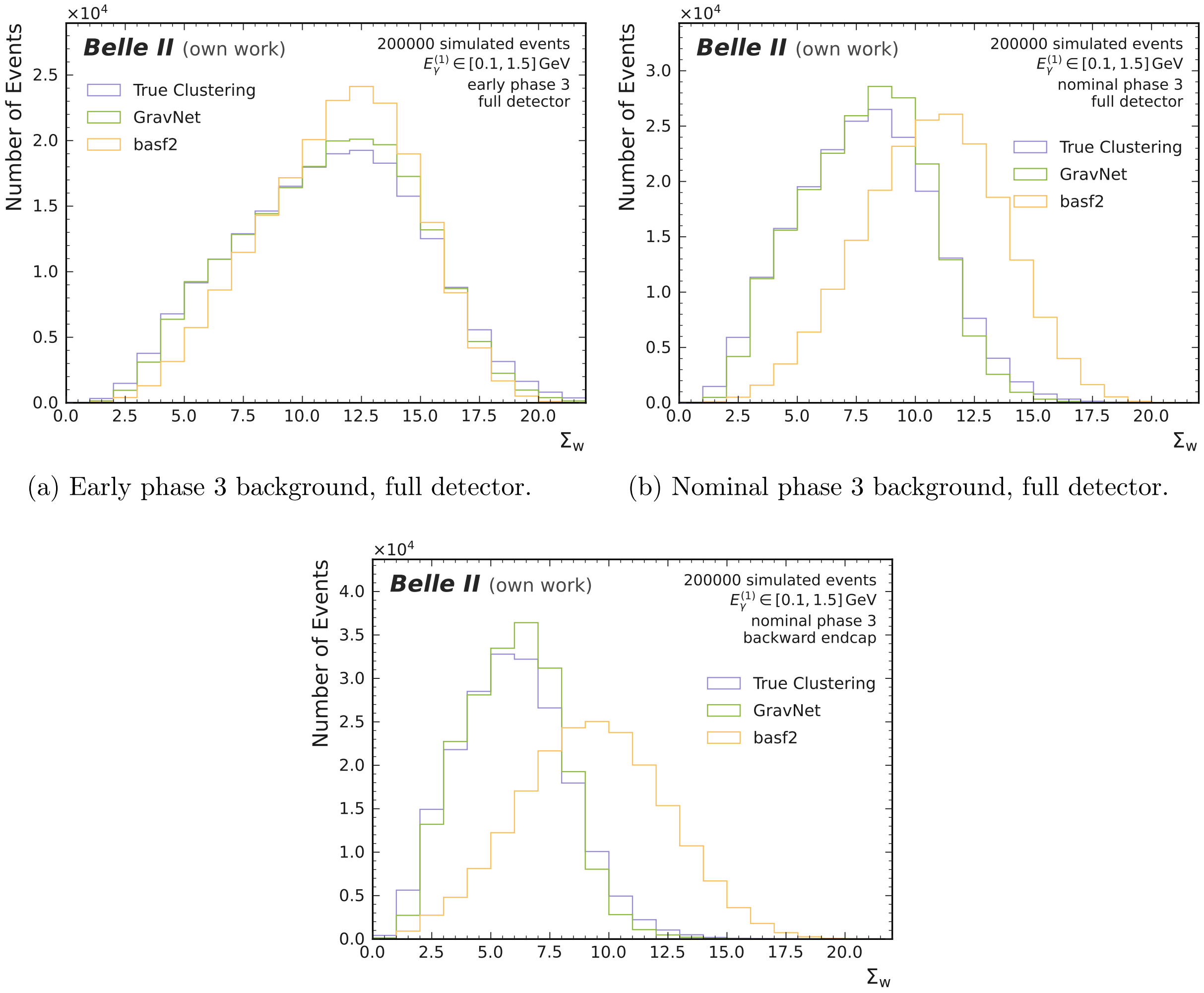


Figure 6.2.: Distribution of the leakages ΔE_{leak} for the one-cluster toy studies with early and nominal phase 3 background. Barrel, forward, and backward endcaps are shown in comparison.



(a) Early phase 3 background, full detector.

(b) Nominal phase 3 background, full detector.

(c) Nominal phase 3 background, backward endcap.

Figure 6.3.: Distribution of the the sum of weights Σ_w for the one-cluster toy studies. Each plot compares the true clustering, GravNet, and the basf2 baseline.

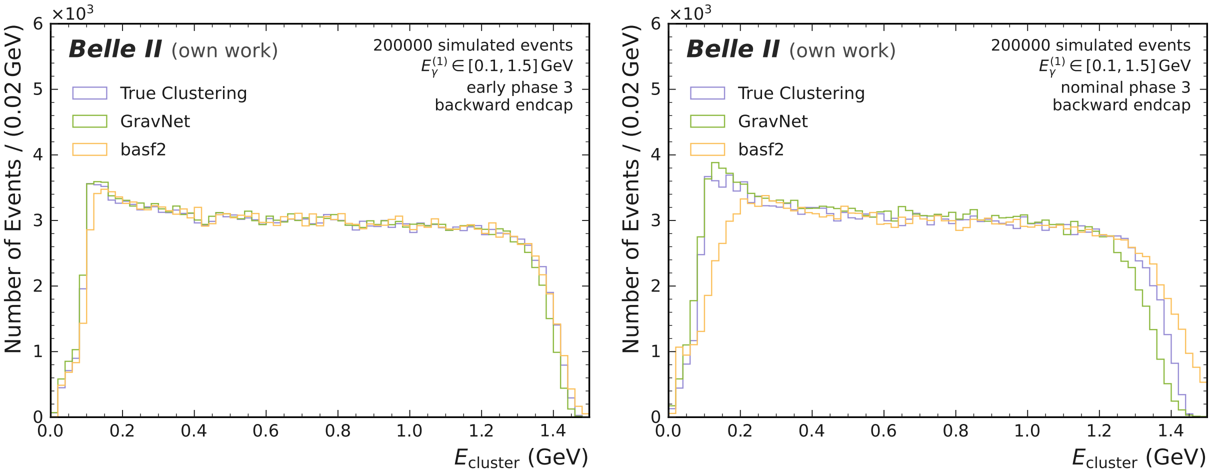


Figure 6.4.: Distribution of cluster energies E_{cluster} for the one-cluster toy studies. Both plots compare the true clustering, GravNet, and the basf2 baseline in the backward endcap. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Cluster Energy

Figure 6.4 shows the distribution of cluster energies E_{cluster} in the backward endcaps for both types of backgrounds. A peak towards lower energies stands out for both true deposited energies. This accumulation of lower energies is likely caused by a combination of higher leakage and the event selection criteria. For early background, both algorithms find proportionally more low-energy photons. For nominal background, GravNet still identifies low-energy photons, whereas basf2 smoothens the peak which is in agreement with the results for the sum of weights.

Cluster Radius

Figure 6.5 shows the distribution of cluster radii R for early and nominal background in full detector coverage. The graphs for the true clustering are practically identical for the two types of backgrounds. This is expected considering R depends solely on the particles true energy depositions which themselves are independent of the background. GravNet and basf2 model the property well for early background but have distinctive reactions to nominal background. While the peak of Gravnet is slightly more pronounced, basf2 predicts clusters with significantly larger radii thereby smearing out the peak. On the bottom line GravNet likely overspecializes to include a certain number of crystals in a cluster. Basf2 assigns overly many crystals to a cluster or too much weight to crystals at the edge of the cluster. A possible cause is the inability to fuzzy cluster background energy depositions.

6.2.2. Performance Evaluation

This section uses the metrics introduced in section 5.3 to quantify the performances of the algorithms. The FCAI is tailored towards describing the accuracy of the underlying clustering. Sensitivity and precision aim to assess the performance on a crystal level. The energy resolution directly evaluates the primary objective, which is to reduce the reconstruction errors for the energy reconstruction. The energy dependence of the resolution reveals the full performance spectrum of the algorithms. Again, the most relevant metrics are presented in this section with the rest of the metrics, including fit parameters, being in appendix C.2.

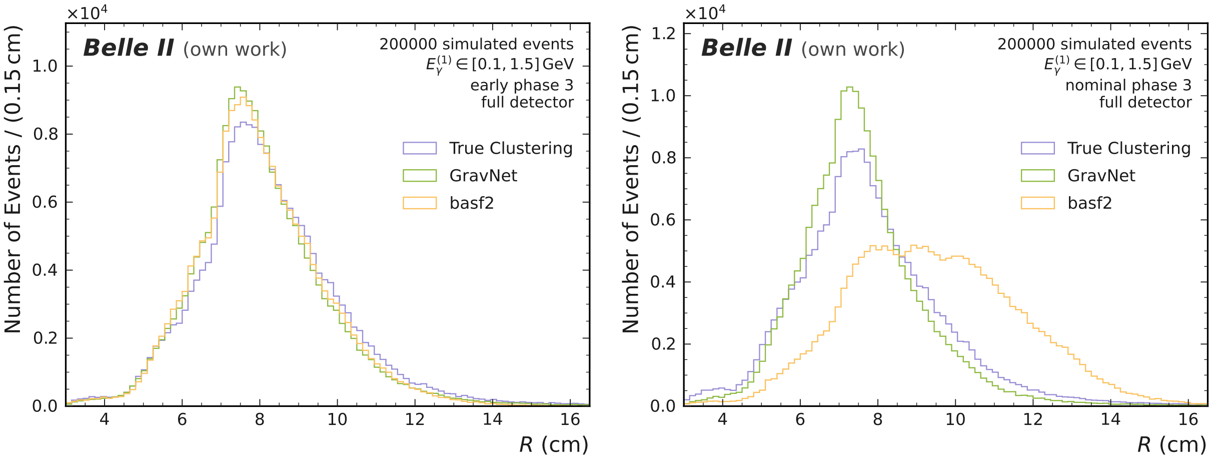


Figure 6.5.: Distribution cluster radii R for the one-cluster toy studies. Both plots compare the true clustering, GravNet, and the basf2 baseline in full detector coverage. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Fuzzy Clustering Agreement Index

The FCAI quantifies the quality of a predicted clustering in comparison with the true clustering. Figure 6.6 displays the distribution of FCAIs for the full detector coverage. According to the median of the distribution, GravNet depicts the underlying clustering more accurately for both early background and nominal background. The performance remains almost stable with nominal background which is in line with the results seen in the event properties section 6.2.1. A good clustering result has to be interpreted carefully though and is only an addition to the performance evaluation from a distinct aspect. Due to the fact that the FCAI treats crystals independently of their energy, it does not specifically evaluate if an algorithm correctly assigns the most important crystals. However, a good FCAI $>$0.5 does signal that the algorithm works correctly and is not getting a good energy resolution by chance (which would be implied by FCAI$\,=\,$0). This is the case for both algorithms, with the basf2 baseline even achieving some early background events with perfect FCAI$\,=\,$1.0. These are events where some isolated crystals purely contain true energy depositions and are correctly identified by basf2. Figure 6.7 displays an exemplary event display of such an event. However, this cluster signature is an exception and the FCAI, in general, is expected to be worse for basf2 due to the algorithm not fuzzy clustering background energy depositions.

Sensitivity and Precision

Average sensitivity S_{avg} and average precision P_{avg} give insight into the ability of an algorithm to correctly identify energy depositions per crystal. Figures 6.8a and 6.8b show the correlation plots and marginal distributions for early and nominal background in full detector coverage. For the most part S_{avg} and P_{avg} are similar in different detector regions. The forward endcap for early background is an exception to that and is additionally displayed in figure 6.8c.

Generally, nominal background events are more spread out and in the median have lower S_{avg} and P_{avg}, as is intuitively expected. In all cases, basf2 demonstrates a significantly higher S_{avg} than GravNet. Given the approach of basf2 to rather assign a crystal to a cluster and remove it later in the optimization (see section 2.3), this is unsurprising. GravNet has roughly the same advancements in P_{avg} that are then seen in the resolution. In the forward endcap with early background, GravNet and basf2 have almost identical median in S_{avg} but a large discrepancy in P_{avg}. The next paragraph uses these observations to draw a connection to the resolution.

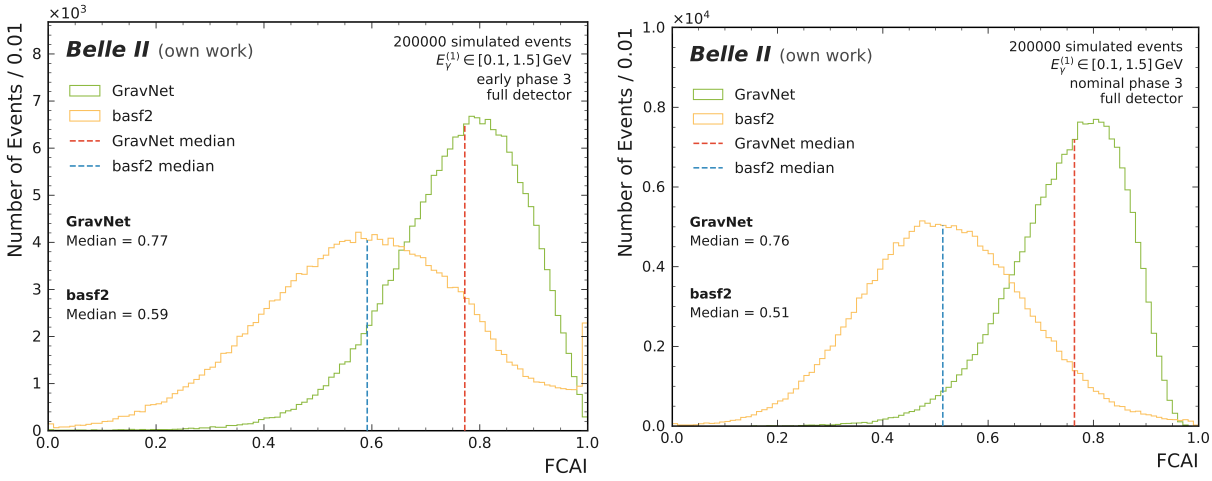


Figure 6.6.: Distribution of FCAIs for the one-cluster toy studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

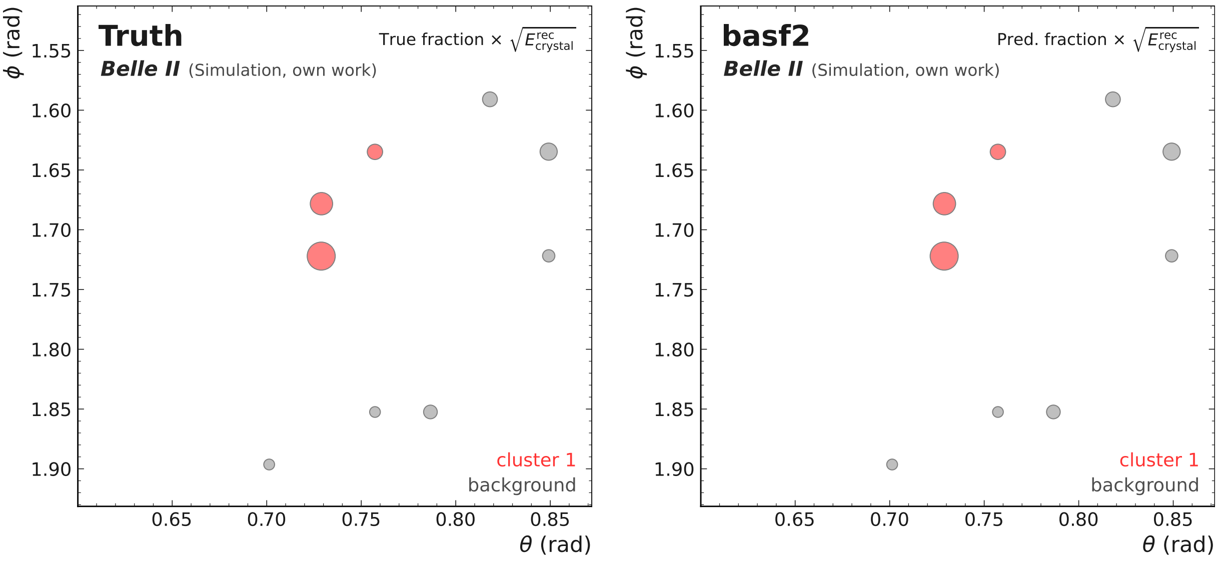
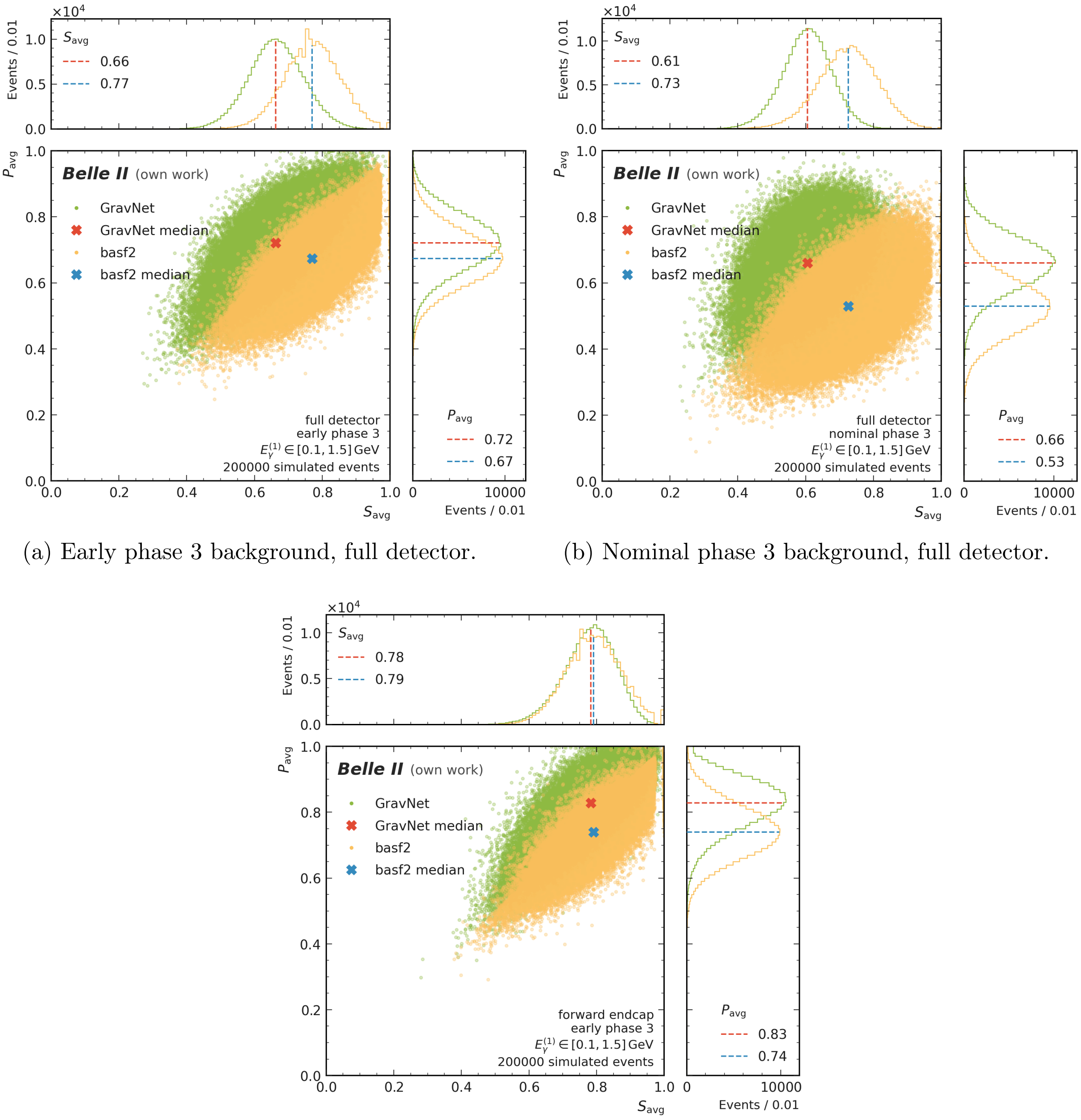


Figure 6.7.: Example of an event with perfect FCAI = 1.0 from the one-cluster toy study with early phase 3 background. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\mathrm{rec}}}$. The plot to the left presents the true clustering, the right plot side depicts the basf2 clustering result.



(a) Early phase 3 background, full detector.

(b) Nominal phase 3 background, full detector.

(c) Early phase 3 background, forward endcap.

Figure 6.8.: Correlation and marginal distributions for average sensitivity S_{avg} and average precision P_{avg} for the one-cluster toy studies. Each plot compares GravNet and the basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated in the marginal distributions.

Energy Resolution

The analysis of the energy resolution starts with a look at the total distributions of reconstruction errors on the deposited energy η_{dep}. Figure 6.9 displays the distribution of η_{dep} for GravNet and basf2 with early and nominal background. The shape of the distributions in all detector parts only differs in width, hence, the full detector coverage is presented here. The peaks are approximately symmetrical around zero. This is especially true for GravNet with both types of backgrounds. Basf2 overestimates energies for nominal background events which manifests in a smeared-out peak towards positive η_{dep}. Even more noticeable is the secondary peak for negative η_{dep}. In these few events, basf2 only assigns the LM to the cluster and is unable to recognize crystals with large true deposited energies at a greater distance to the LM. Figure 6.10 shows the event display for such an event. GravNet is at an advantage in this situation with the ability to cluster disjoint energy depositions.

Figure 6.11 shows the distributions of reconstruction errors on the generated energy η_{gen}. The peaks of the distributions are much wider and biased towards negative values, which is caused by leakage. GravNet shows a similar overall shape for early and nominal backgrounds. Basf2 again has the tendency to overestimate the energy for some nominal background events, reflected by a smeared-out peak. This is despite basf2 now employing its leakage correction which is optimized to correct for detector effects. The absence of such a mechanism for GravNet results in a shifted peak. This is illustrated in more detail in the upcoming plots with a focus on the peaks.

In the next step, the FWHMs of the distributions of reconstruction errors are determined to quantify the performances. Figure 6.12 shows the fits for η_{dep}, the resulting resolutions as well as the tails for all detector regions and the two types of backgrounds. Both algorithms and types of backgrounds have in common that the resolution is best for the forward endcap and the worst for the backward endcap. The resolution in the backward endcap can be explained by higher levels of background and the irregular structure of crystals. Most true event properties shown in the previous section are practically identical for different detector regions and leakage is no factor in η_{dep}. Only the small accumulation of lower energetic events, seen in figure 6.4, is likely to have a small impact on the resolution in the backward endcap. On the other hand, the excellent resolution in the equally irregular forward endcap comes arises unexpectedly. This leads to the conclusion that the resolution is mainly dominated by the amount of background and

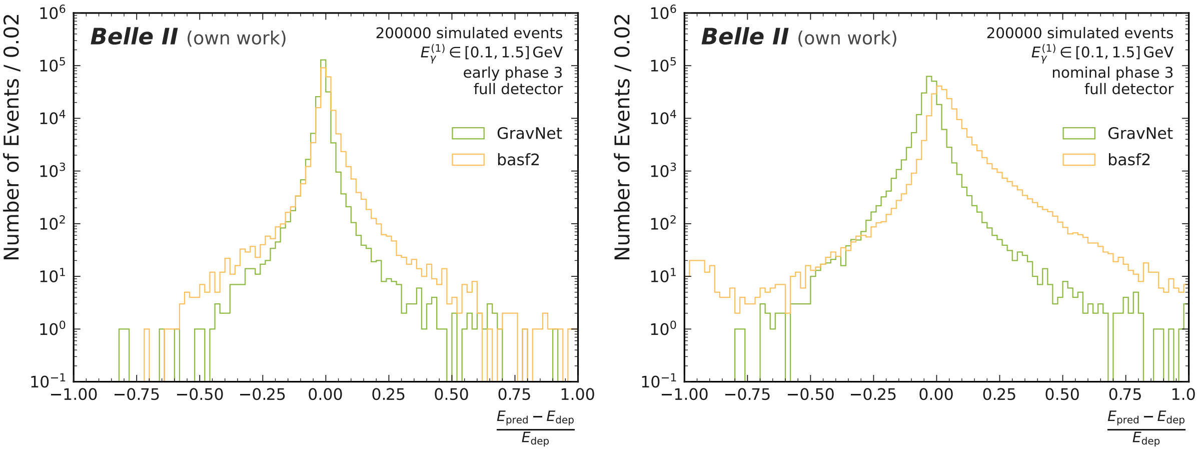
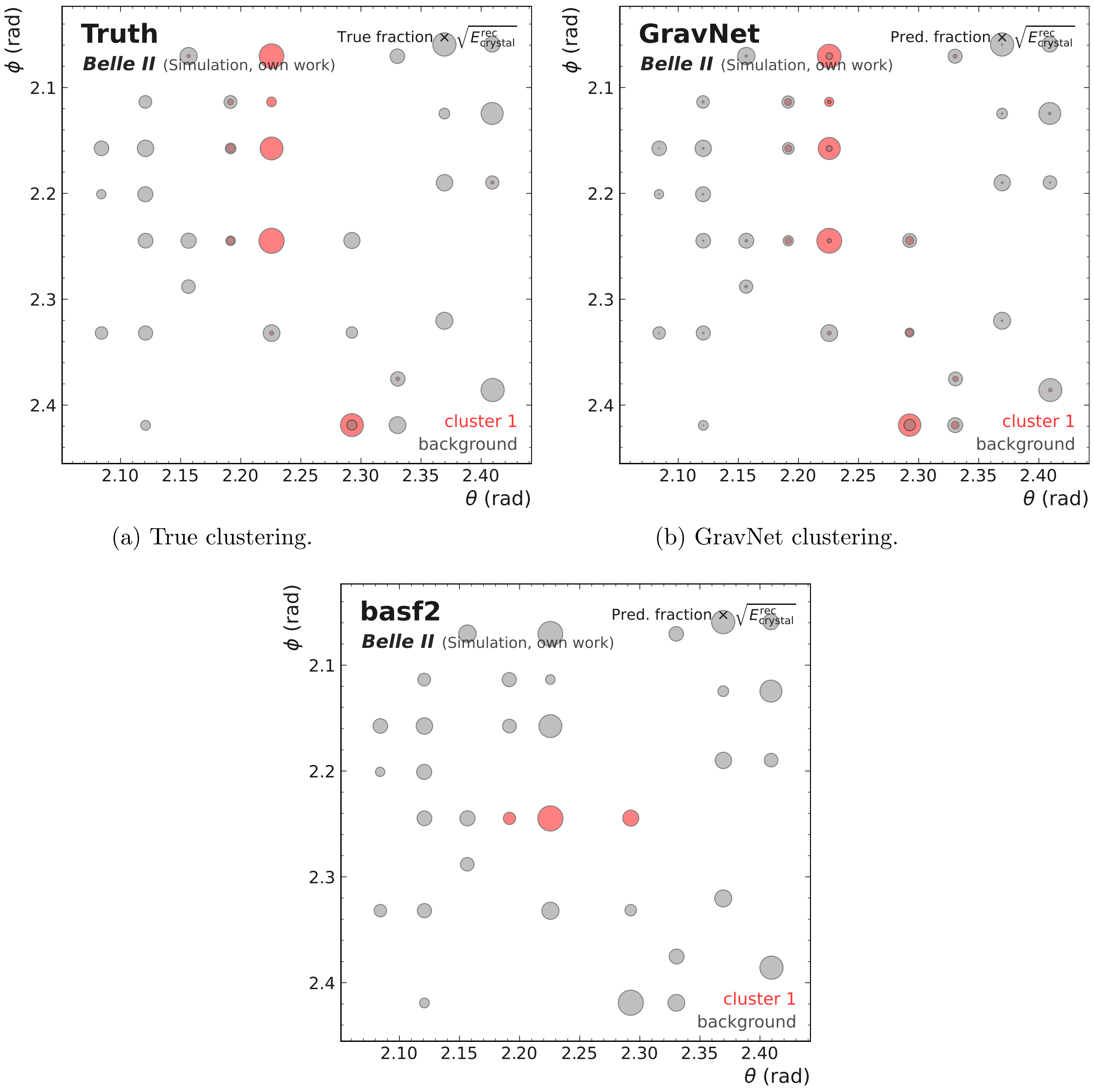


Figure 6.9.: Distribution of reconstruction errors on the deposited energy η_{dep} for the one-cluster toy studies. Both plots compare GravNet and the basf2 baseline in full detector coverage. The left side depicts early phase 3 events, and the right side nominal phase 3 events.



(a) True clustering.

(b) GravNet clustering.

(c) Basf2 clustering.

Figure 6.10.: Example of an event with a large negative reconstruction error on the deposited energy η_{dep} from the one-cluster toy study with nominal phase 3 background. $\eta_{\text{dep}} = -0.883$ for the basf2 clustering, and $\eta_{\text{dep}} = -0.139$ for the GravNet clustering. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\text{rec}}}$.

both algorithms have no problem dealing with the different detector geometry. That conclusion is backed up by the comparison between early and nominal backgrounds. The resolution in any scenario with nominal background is worse than its counterpart with early background by a factor of about 2.5. The left tails of GravNet and basf2 measure roughly the same independent of the background. However, the right tails of the distributions are significantly shorter for GravNet, meaning it is less prone to overestimate the energy. This is particularly noticeable with nominal background where GravNet at least reduces the tails by half in every detector region.

A notable feature of all GravNet peaks is the shift toward negative values. Due to leakage and the lack of a correction mechanism, this shift is expected for the distribution of η_{gen} but not of η_{dep}.

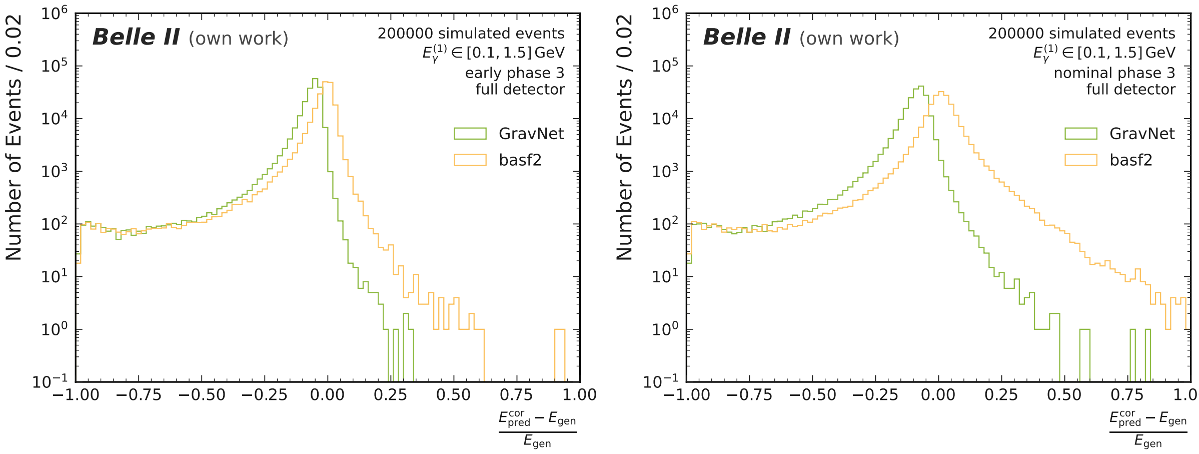


Figure 6.11.: Distribution of reconstruction errors on the generated energy η_gen for the one-cluster toy studies. Both plots compare GravNet and the basf2 baseline in full detector coverage. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

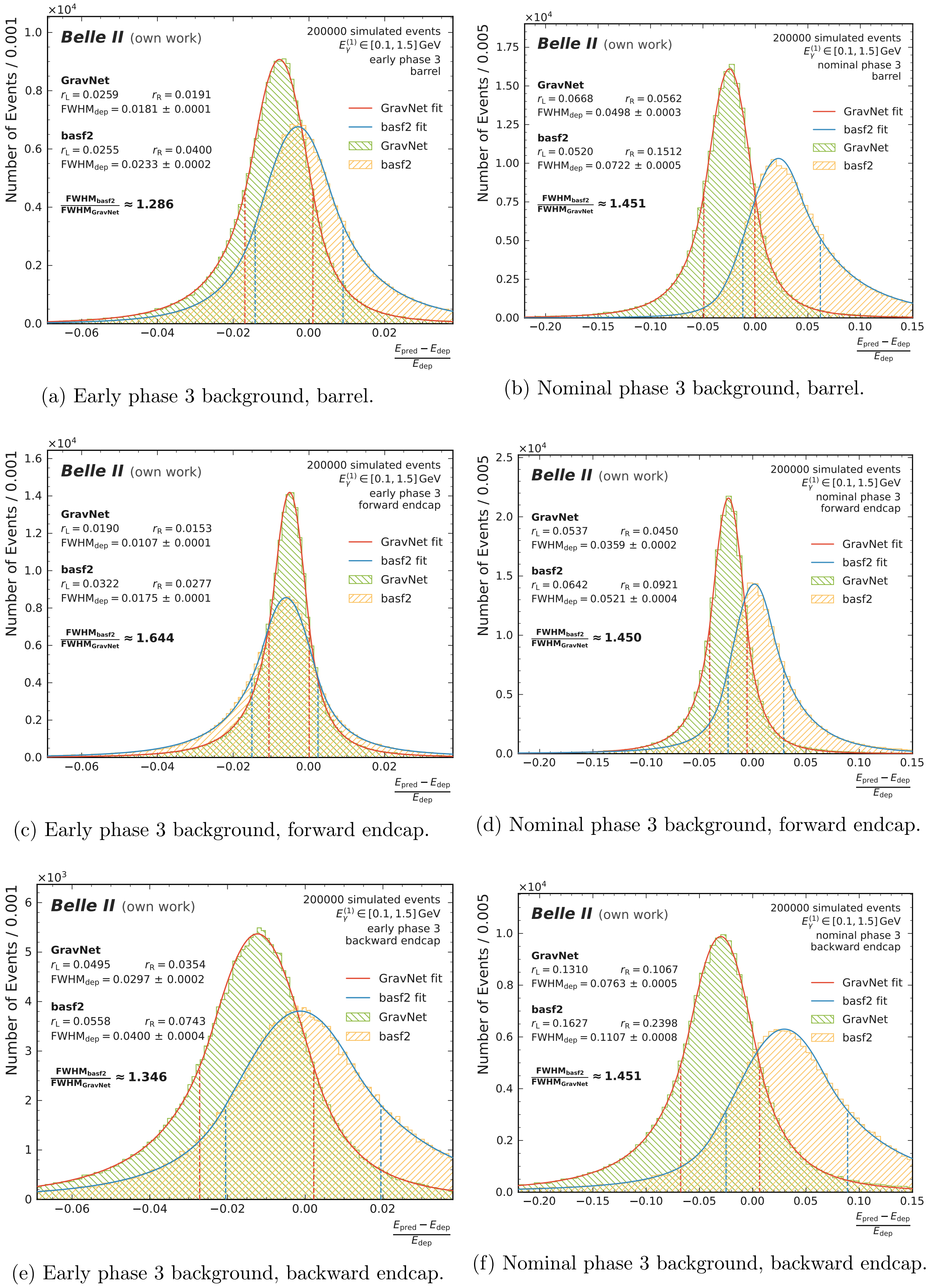
Two assumptions could explain the bias: First, a difference between the mean of the $L2$-loss from equation (4.2) and the mean of the η_dep distribution with GravNet naturally striving towards a zero mean for its loss. Second, the numerical behavior of the regression architecture, namely to always assign (although tiny) fractions to all classes in every crystal, which leads to more energy loss in high-energy crystals than is compensated for in low-energy ones. The bias is corrected in the FWHMs, however, further investigations are necessary for a satisfactory explanation and understanding of GravNet in future work.

The largest relative improvement in resolution is found in the forward endcap and coincides with the exceptional result for S_avg and P_avg. In this scenario (see figure 6.8), GravNet and basf2 have almost an identical median in S_avg but a large discrepancy in P_avg. These factors lead to the conclusion that the resolution is indeed dominated by the precision of an algorithm. The forward endcap also highlights the association of lower sensitivity to the bias in the peaks of GravNet. The otherwise constant tendency towards lower energies is significantly reduced in the forward endcap, where S_avg of GravNet catches up to that of basf2.

Overall, GravNet consistently improves the resolution on the deposited energy in nominal background events and in any detector region by about 45 % over basf2. The improvements for early background are within a range of 28 % in the barrel and up to 64 % in the forward endcap. Additionally, while the left tails are roughly the same, GravNet significantly reduces the right tails of the distributions in both studies.

Figure 6.13 shows the fits for η_gen, the resulting resolutions as well as the tails for all detector regions and the two types of backgrounds. The reconstruction errors on the generated energy ultimately determine the resolution that is left for further physics reconstruction from an improvement to the clustering. For early background, basf2 now displays a peak close to zero for all detector regions, indicating a well-functioning leakage correction. In the case of nominal background, the correction is not accurate anymore and the constant overestimation of energies that is already present in the distribution of η_dep remains.

From the initially large improvements to η_dep, only some are left after detector effects. GravNet delivers around 9 % improvement on η_gen for early background. The relative improvements to the right tails decrease, with basf2 even reducing the tails by the means of leakage correction. For nominal background events, the improvements range from 15 % to 21 % depending on the detector region. GravNet additionally maintains a reduction of the right tails by a factor of two.



(a) Early phase 3 background, barrel.

(b) Nominal phase 3 background, barrel.

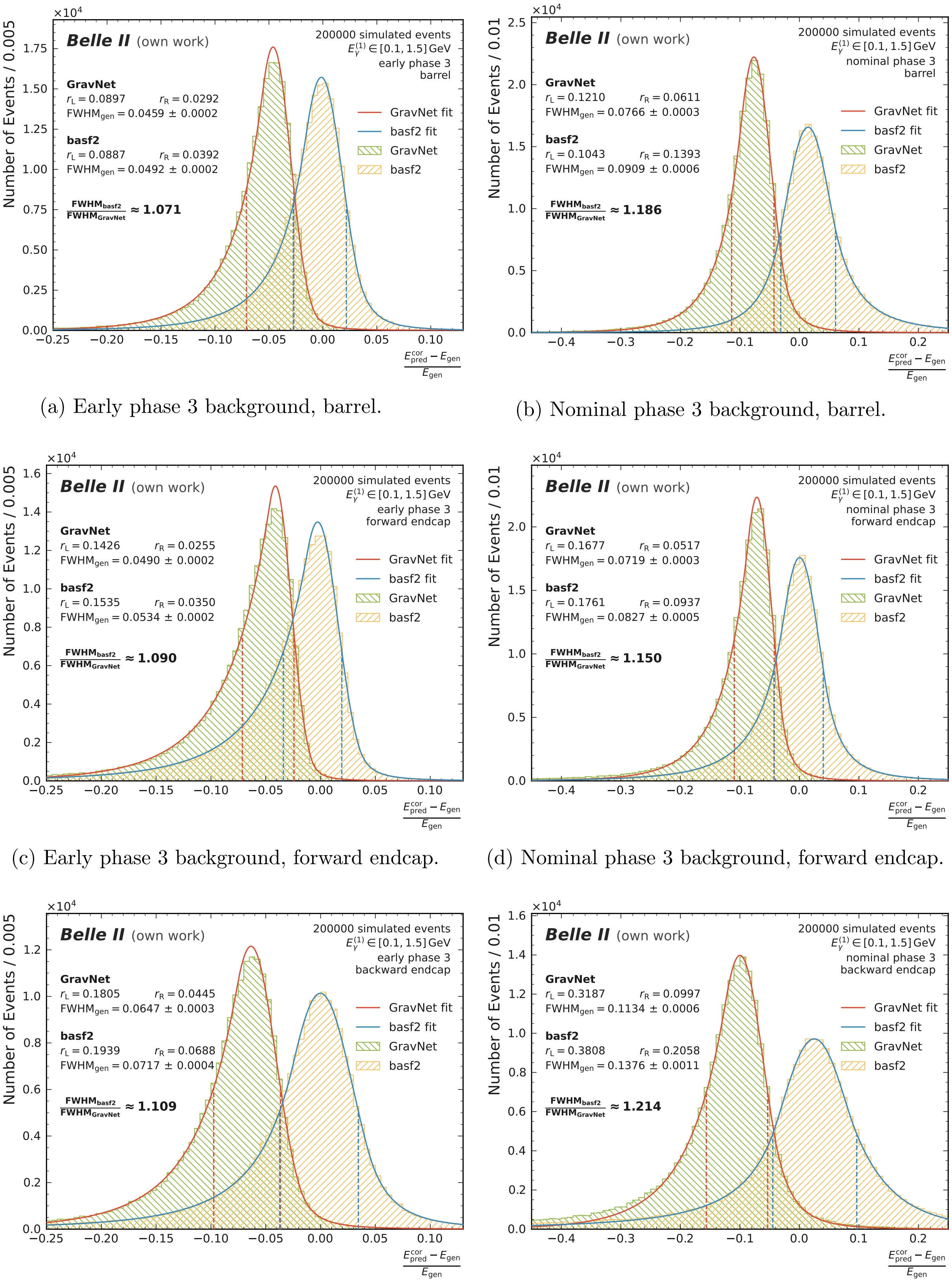
(c) Early phase 3 background, forward endcap.

(d) Nominal phase 3 background, forward endcap.

(e) Early phase 3 background, backward endcap.

(f) Nominal phase 3 background, backward endcap.

Figure 6.12.: Fits for the distributions in reconstruction errors on the deposited energy η_dep for the one-cluster toy studies. Each plot compares GravNet to the basf2 baseline. The left plots depict early phase 3 events, the right plots nominal phase 3 events. Barrel, forward, and backward endcaps are displayed separately in this order.



(a) Early phase 3 background, barrel.

(b) Nominal phase 3 background, barrel.

(c) Early phase 3 background, forward endcap.

(d) Nominal phase 3 background, forward endcap.

(e) Early phase 3 background, backward endcap.

(f) Nominal phase 3 background, backward endcap.

Figure 6.13.: Fits for the distributions in reconstruction errors on the generated energy η_{gen} for the one-cluster toy studies. Each plot compares GravNet to the basf2 baseline. The left plots depict early phase 3 events, and the right plots nominal phase 3 events. Barrel, forward, and backward endcaps are displayed separately in this order.

Energy Dependence

The energy dependence is considered the most important part of the study as it reveals the full performance spectrum of an algorithm. The analysis is carried out according to section 5.3.2 with 20 000 events per fixed energy ranging from 0.01 GeV up to 3.0 GeV. The energy range exceeds the energies that GravNet is trained on, thus the analysis also evaluates the ability of GravNet to generalize to other photon energies.

The endcaps behave as expected from the previous analysis, therefore, focus is put on the full detector coverage and plots of all detector regions are in appendix C.3. Figure 6.14 displays early background events, figure 6.15 nominal background events. Both figures compare the plots for $\mathrm{FWHM}_{\mathrm{dep}}$ and $\mathrm{FWHM}_{\mathrm{gen}}$, revealing the potential resolution based only on the clustering, and the remaining resolution in a physics reconstruction given the current detector hardware.

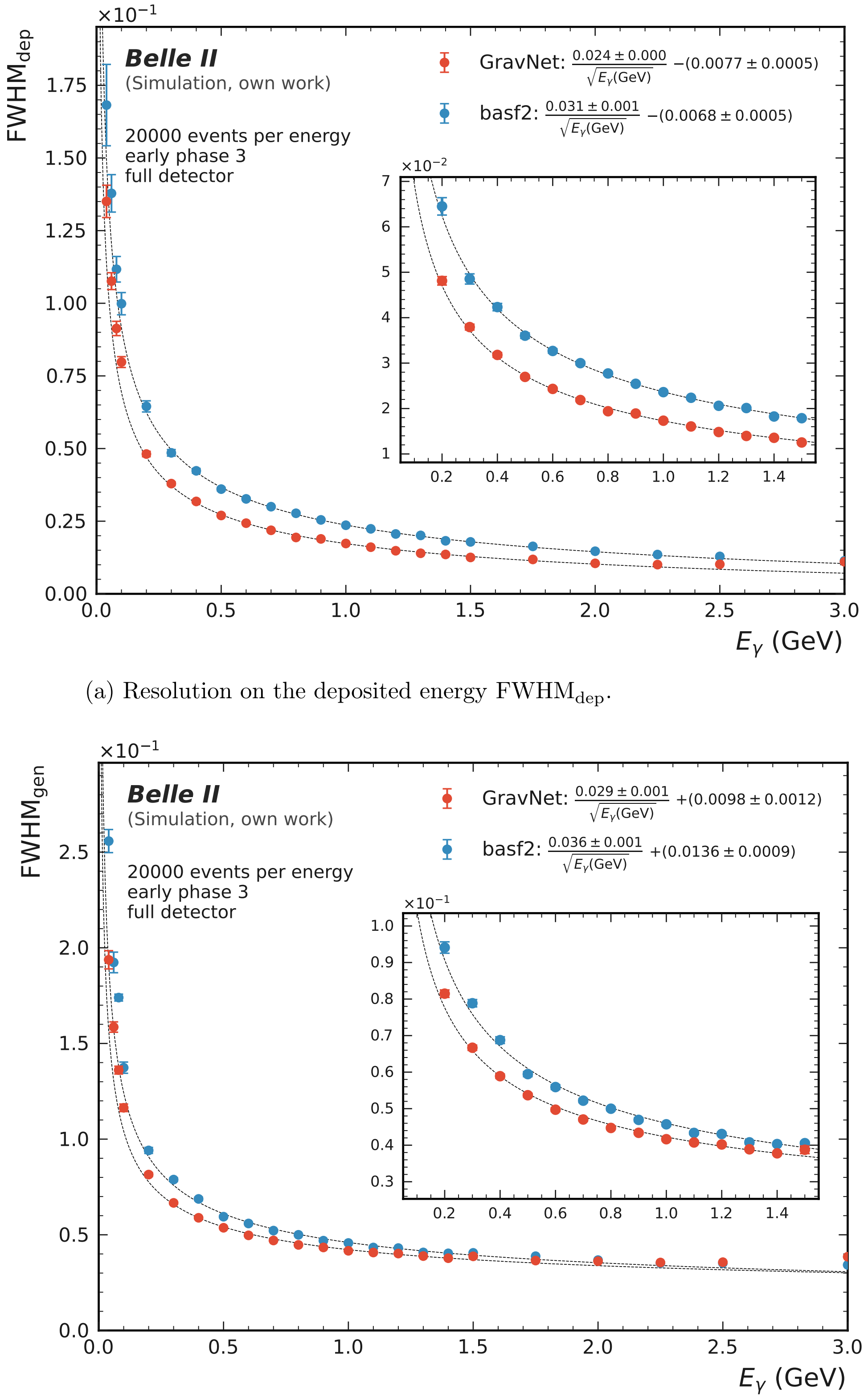
For early and nominal backgrounds alike, GravNet outperforms the basf2 baseline significantly in the low-energy regime. The distributions converge for higher energies as a combination of GravNet not being trained at such energies and basf2 already providing excellent resolution, especially with leakage correction. Again the improvements diminish moving from $\mathrm{FWHM}_{\mathrm{dep}}$ to $\mathrm{FWHM}_{\mathrm{gen}}$ in light of detector effects, however, the decrease is less pronounced with nominal background. For photons with energies $E_{\mathrm{gen}} < 0.1\,\mathrm{GeV}$, outside of the training range, GravNet outperforms the basf2 baseline by up to 100 %. Overall, GravNet generalizes better to lower-energy photons than to higher-energy ones. Low-energy photons often deposit all their energy in just a few crystals, which is a signature GravNet already learns with 0.1 GeV photons. Higher energy photons can lead to a larger number of crystals in a cluster than the algorithm is trained for.

6.2.3. Overview and Conclusions

This section sums up the performance analysis and draws final conclusions combining the results. Table 6.2 summarizes essential performance metrics of the early background results for different detector regions and calculates the improvement to the basf2 baseline in percent. Table 6.3 presents the same for the nominal background events.

GravNet proves to be a stable and reliable algorithm for the clustering of simple one-cluster events. The ability to fuzzy cluster background depositions stands out in several metrics, like the sensitivity and precision, FCAI, or the sum of weights. GravNet holds up its performance in edge cases like disjoint clusters or with tiny energy depositions. In general, GravNet is able to model underlying clusters more accurately than basf2, which is especially true in the presence of nominal background. The large improvements with nominal background are most likely the result of the hyperparameter optimization and feature analysis specifically for this scenario. When it comes to different detector regions, both algorithms are able to deal with the irregular geometries in the endcaps, which is one of the main motivations for GravNet. While the clustering quality is similar in different detector regions, it does not equally carry over to the resolution in these regions.

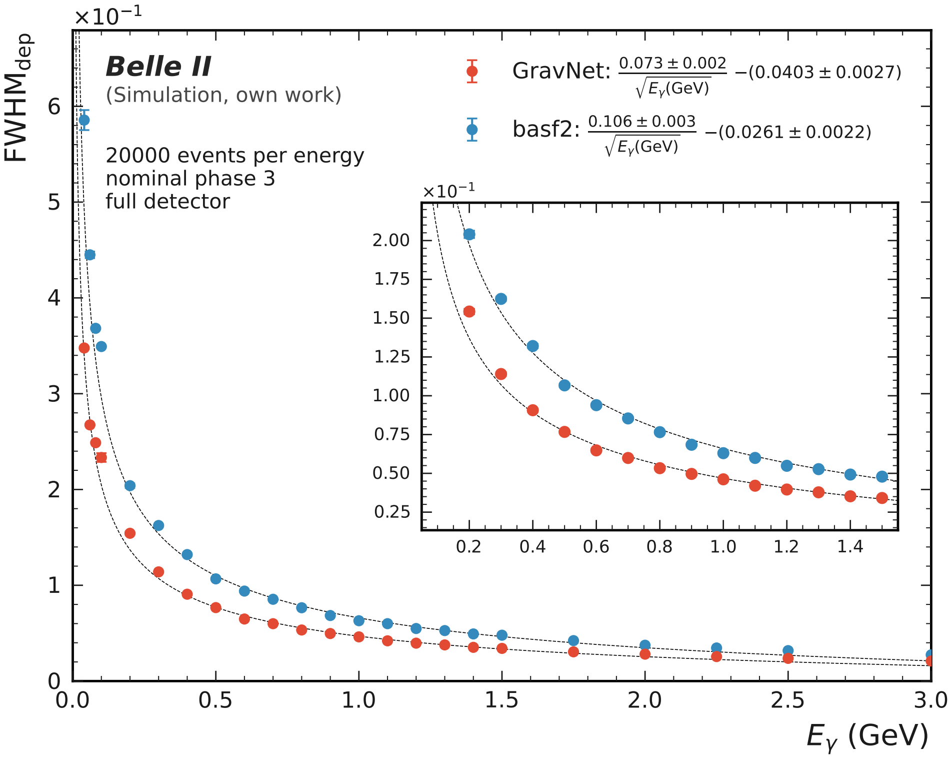
The effects of the improved clustering on the energy resolution are most noticeable in low-energy photons and with high levels of background. In this regime, GravNet improves the resolution on the generated energy up to 17.2 % in full detector coverage. The resolution on the deposited energy promises a large potential for many applications in future scenarios. GravNet comes with a noteworthy bias towards lower energies. Even though this bias is explicable by the absence of leakage correction for η_{gen}, its presence in η_{dep} needs further investigation. Basf2 provides comparable resolution for early background. Despite having the advantage of a profound correction, it is not optimized for nominal background and lacks behind in the performance for this future scenario.



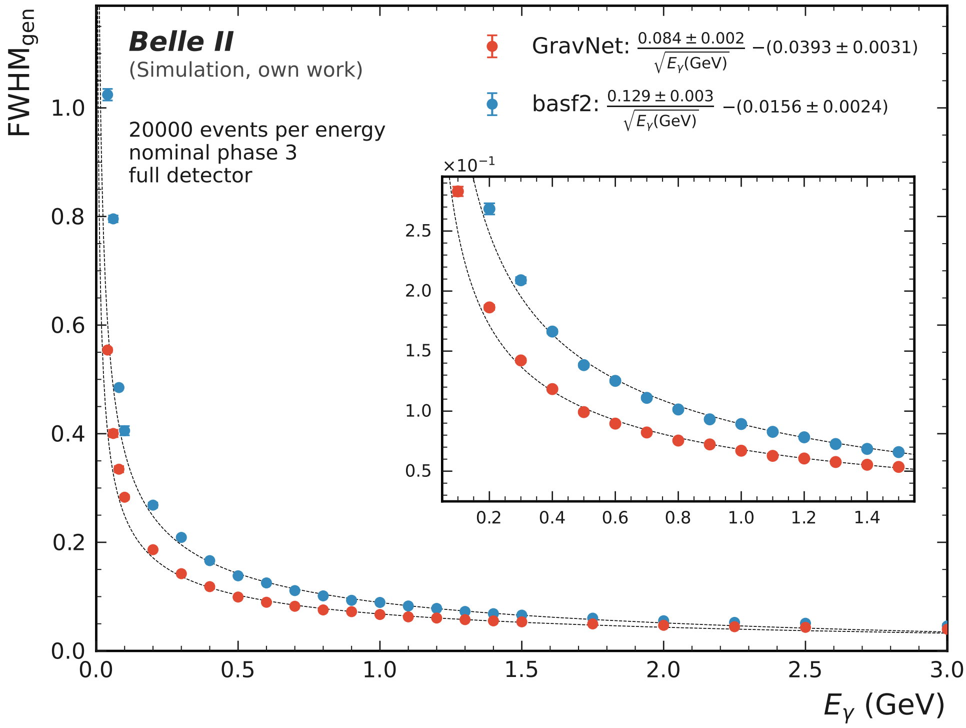
(a) Resolution on the deposited energy FWHM$_{\mathrm{dep}}$.

(b) Resolution on the generated energy FWHM$_{\mathrm{gen}}$.

Figure 6.14.: Shown are the resolutions FWHM$_{\mathrm{dep,gen}}$ in dependence of the generated photon energy E_γ for the one-cluster toy studies with early phase 3 background. Both plots compare GravNet and basf2 in full detector coverage. Each data point marks the FWHM$_{\mathrm{dep/gen}}$ of 20 000 events at a fixed energy $E_\gamma \in [0.01, 3.0]$ GeV. A fit models the relation with an inverse square root and the one sigma band is highlighted.



(a) Resolution on the deposited energy $\mathrm{FWHM}_{\mathrm{dep}}$.



(b) Resolution on the generated energy $\mathrm{FWHM}_{\mathrm{gen}}$.

Figure 6.15.: Shown are the resolutions $\mathrm{FWHM}_{\mathrm{dep,gen}}$ in dependence of the generated photon energy E_γ for the one-cluster toy studies with nominal phase 3 background. Both plots compare GravNet and basf2 in full detector coverage. Each data point marks the $\mathrm{FWHM}_{\mathrm{dep/gen}}$ of 20 000 events at a fixed energy $E_\gamma \in [0.01, 3.0]$ GeV. A fit models the relation with an inverse square root and the one sigma band is highlighted.

Table 6.2.: Summary and comparison of the performance of the algorithms for one-cluster toy study events with early phase 3 background. The metrics $\mathrm{FWHM}_{\mathrm{dep,gen}}$, average sensitivity S_{avg} and precision P_{avg}, and FCAI are listed for different detector regions. The improvement to the basf2 baseline is stated in percent for each region.

Detector Region	Algorithm	$\mathrm{FWHM}_{\mathrm{dep}}$ $\times 10^{-2}$	$\mathrm{FWHM}_{\mathrm{gen}}$ $\times 10^{-2}$	S_{avg}	P_{avg}	FCAI
Barrel	GravNet	1.81	4.59	0.65	0.71	0.77
	basf2	2.33	4.92	0.77	0.67	0.60
	Improvement	**28.6 %**	**7.1 %**	**-15.6 %**	**6.0 %**	**28.3 %**
Forward Endcap	GravNet	1.07	4.90	0.78	0.83	0.77
	basf2	1.75	5.34	0.79	0.74	0.49
	Improvement	**64.4 %**	**9.0 %**	**-1.3 %**	**12.2 %**	**57.1 %**
Backward Endcap	GravNet	2.97	6.47	0.61	0.67	0.80
	basf2	4.00	7.17	0.75	0.61	0.61
	Improvement	**34.6 %**	**10.9 %**	**-18.7 %**	**9.8 %**	**31.1 %**
Full Detector	GravNet	1.72	4.88	0.66	0.72	0.77
	basf2	2.34	5.13	0.77	0.67	0.59
	Improvement	**36.2 %**	**5.2 %**	**-14.2 %**	**7.4 %**	**30.5 %**

Table 6.3.: Summary and comparison of the performance of the algorithms for one-cluster toy study events with nominal phase 3 background. The metrics $\mathrm{FWHM}_{\mathrm{dep,gen}}$, average sensitivity S_{avg} and precision P_{avg}, and FCAI are listed for different detector regions. The improvement to the basf2 baseline is stated in percent for each region.

Detector Region	Algorithm	$\mathrm{FWHM}_{\mathrm{dep}}$ $\times 10^{-2}$	$\mathrm{FWHM}_{\mathrm{gen}}$ $\times 10^{-2}$	S_{avg}	P_{avg}	FCAI
Barrel	GravNet	4.98	7.66	0.61	0.66	0.76
	basf2	7.22	9.09	0.73	0.52	0.51
	Improvement	**45.1 %**	**18.6 %**	**-16.4 %**	**26.9 %**	**49.0 %**
Forward Endcap	GravNet	3.59	7.19	0.60	0.66	0.76
	basf2	5.21	8.27	0.73	0.57	0.54
	Improvement	**45.0 %**	**15.0 %**	**-17.8 %**	**15.8 %**	**40.7 %**
Backward Endcap	GravNet	7.63	11.34	0.62	0.66	0.77
	basf2	11.07	13.76	0.72	0.47	0.45
	Improvement	**45.1 %**	**21.4 %**	**-13.9 %**	**40.4 %**	**71.1 %**
Full Detector	GravNet	4.50	7.66	0.61	0.66	0.76
	basf2	6.87	8.97	0.73	0.53	0.51
	Improvement	**52.7 %**	**17.2 %**	**-16.4 %**	**24.5 %**	**49.0 %**

6.3. Two Overlapping Photon Clusters

The dual particle gun introduced in section 3.1 is used for the creation of events with two photons. The momentum is set to $p_\gamma = E_\gamma \in [0.1, 1.5]\,\text{GeV}$, the angles are set to $\phi \in [0.0, 360.0]\,^\circ$, and θ depending on the detector region according to table 6.1. The generated events contain two photons that are distributed uniformly in space and momenta within the given ranges. The variation in energy provides asymmetry associated with the energy difference ΔE_{leak}. The angular separation between the photons is set to a uniform $\delta \in [2.8, 9.7]\,^\circ$. This leads to various levels of overlap associated with the cluster center distance Δx. After the photon generation, either early or nominal background is added to the event. The resulting events have to fulfill the criteria for two overlapping clusters in section 3.2.2. Two million events are used for the training of GravNet, and 200 000 independent events for the testing of the algorithms.

Again, one model of GravNet specializes in early background events. Another, hyperparameter-optimized model of GravNet deals with nominal background. The performances for the two types of backgrounds are analyzed together and compared. Particularly interesting metrics are shown for the three detector regions barrel, forward, and backward endcaps. Extended plots for all metrics in all detector regions are located in appendix D.

6.3.1. Event Properties

As with the one-cluster toy study, the metrics from section 5.2 give a first description of the events. The section starts with a brief summary of the metrics that behave identically to the one-cluster data set, namely leakage, sum of weights, cluster energy, and cluster radius. Focus is then put on metrics that describe overlap and asymmetry in order to see how the algorithms handle this challenging scenario. Again the full set of metrics including less distinct properties and sanity checks for all detector regions is in appendix D.1.

One Cluster Metrics

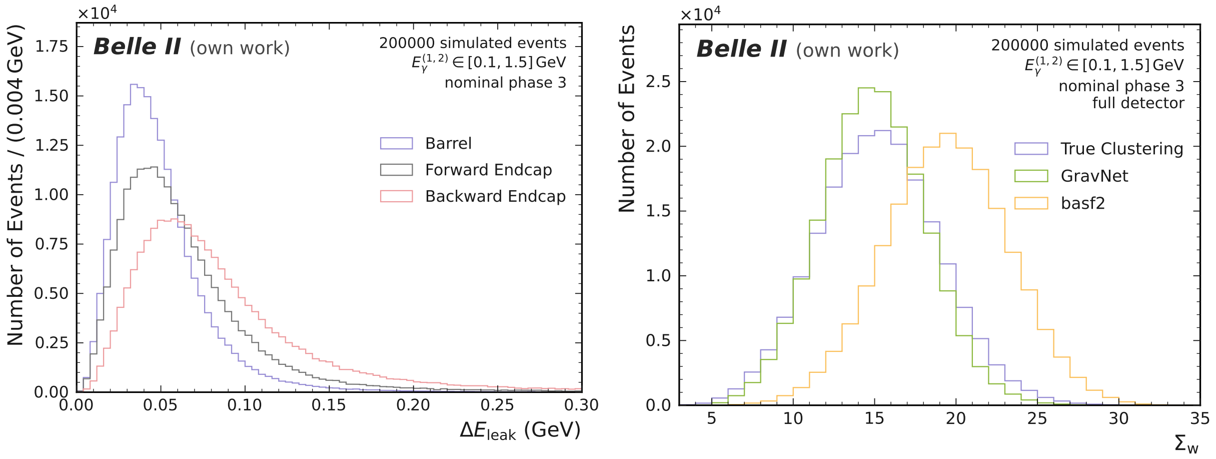
Leakage and sum of weights behave identically to the one-cluster study. Figure 6.16 shows the plots for both properties with nominal background in full detector coverage. In comparison to one-cluster events, the peaks of both quantities are shifted towards higher values. The sum of weights approximately doubles because of the second cluster. Once more basf2 has problems with assigning less weight in nominal background levels. On a relative scale, leakage is shifted less due to the overlap of the two clusters.

Figure 6.17 shows the distribution of cluster energies and radii in the backward endcap with nominal background. It is noteworthy that in contrast to the one-cluster studies, the cluster energies are uniformly distributed in all detector regions, including the backward endcap. For the true clustering, the distribution of cluster radii is identical to that of one-cluster events. GravNet follows the trend of basf2 towards larger radii, which was not the case with one-cluster events. This means that GravNet now also tends to include more crystals in a cluster, however, this behavior is only observed in the backward endcap.

Cluster Radius Difference

The cluster radius difference ΔR describes the asymmetry between the two cluster dimensions. The previous paragraph shows that the clusters in two-cluster events have the same radii as one-cluster events. However, for the algorithms, it is much more challenging to correctly estimate the dimension of a cluster in the presence of overlap. The cluster asymmetry reveals if an algorithm does prefer certain radii in this process. The distribution of ΔR is shown for early and nominal background events in full detector coverage in figures 6.18a and 6.18b.

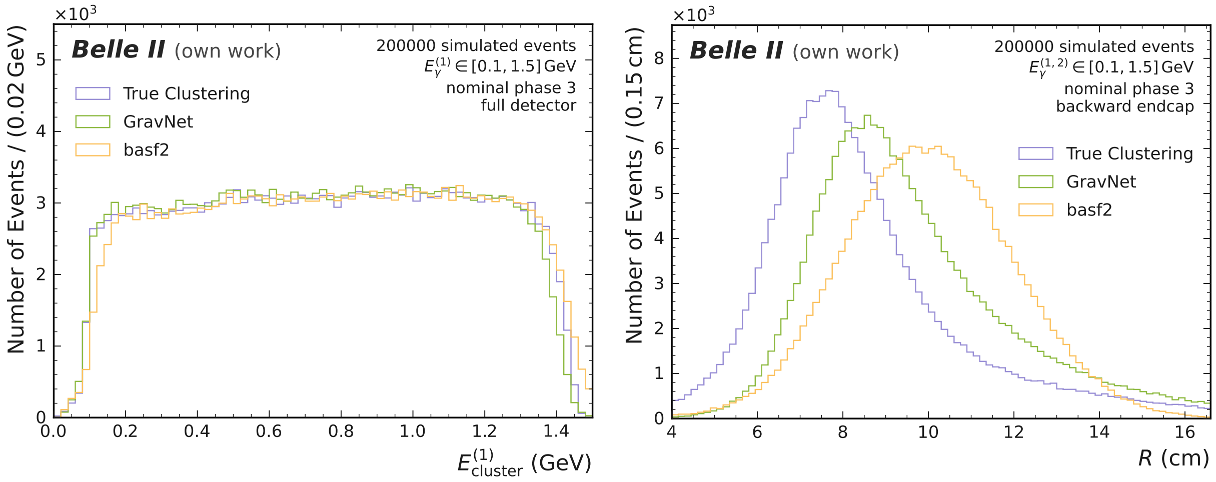
As expected, the true asymmetry does not change between the types of backgrounds. However, distinct trends emerge for the two algorithms: For early background, both GravNet and basf2 significantly prefer symmetric clusters. This is plausible for basf2 which is optimized to a certain cluster shape. However, GravNet should theoretically be able to learn strongly asymmetric events. It is likely that these kinds of events are too rare for the model to pick up in training. The trend towards symmetrical clusters continues for nominal background where GravNet draws a distribution comparable to that of early background. Basf2 is able to map more of the asymmetries correctly. An exception is the backward endcap, shown in figure 6.18c. GravNet performs virtually perfectly in this scenario, whereas basf2 exhibits a modest inclination for larger radii.



(a) ΔE_{leak} in comparison for barrel, forward, and backward endcaps.

(b) Σ_w in full detector coverage. True clustering, GravNet, and basf2 baseline are compared.

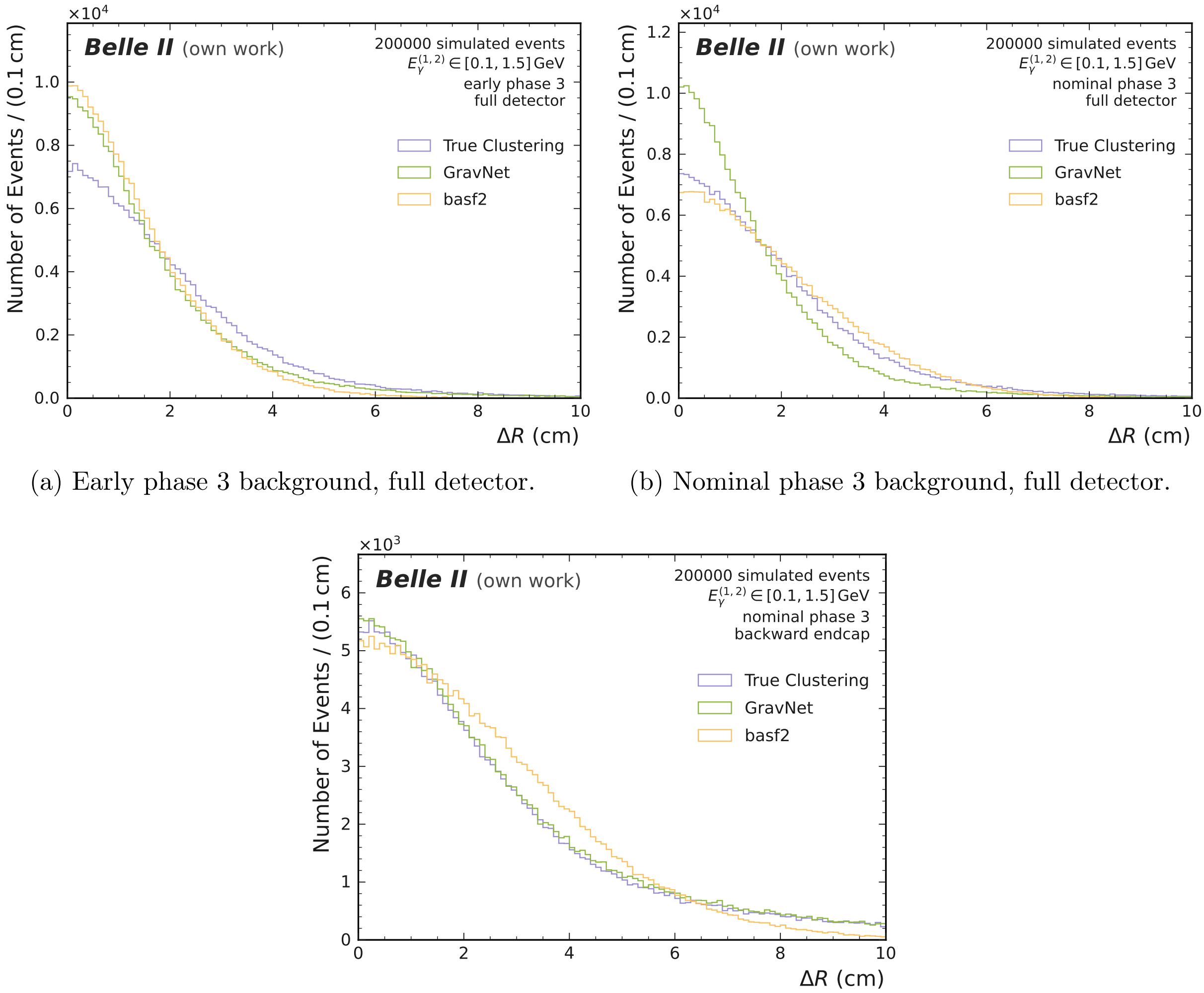
Figure 6.16.: Distribution of leakages ΔE_{leak} and sum of weights Σ_w for the two-cluster toy studies with nominal phase 3 background.



(a) E_{cluster} of photon 1 according to true clustering, GravNet, and basf2 baseline.

(b) R in comparison for true clustering, GravNet, and basf2 baseline.

Figure 6.17.: Distribution of cluster energies E_{cluster} and cluster radii R for the two-cluster toy studies with nominal phase 3 background in the backward endcap.



(a) Early phase 3 background, full detector.

(b) Nominal phase 3 background, full detector.

(c) Nominal phase 3 background, backward endcap.

Figure 6.18.: Distribution of cluster radius differences ΔR for the two-cluster toy studies. Each plot compares the true clustering, GravNet, and the basf2 baseline.

Cluster Energy Difference

The cluster energy difference $\Delta E_{\text{cluster}}$ is another measurement of asymmetry between the two clusters from the perspective of cluster energy. In general, both algorithms are able to separate strongly asymmetric cluster energies in all detector regions. Figure 6.19 displays a minuscule trend that is notable in the backward endcap with nominal background. Here, basf2 prefers more asymmetric cluster energies and neglects events with energetically symmetrical clusters by a small amount. This behavior is again apparent in the energy dependence analysis.

Shared Energy

The distribution of shared energies E_{shared} between the two clusters provides insight into whether an algorithm is able to correctly fuzzy cluster mutual crystals. The behavior of E_{shared} is alike in different detector regions. Figure 6.20 shows the distribution for the full detector with early and nominal background with a focus on the long tails at high E_{shared}. The tails of the distribution only account for a limited amount of the total events, therefore, the figure is complemented by detailed plots of the peaks.

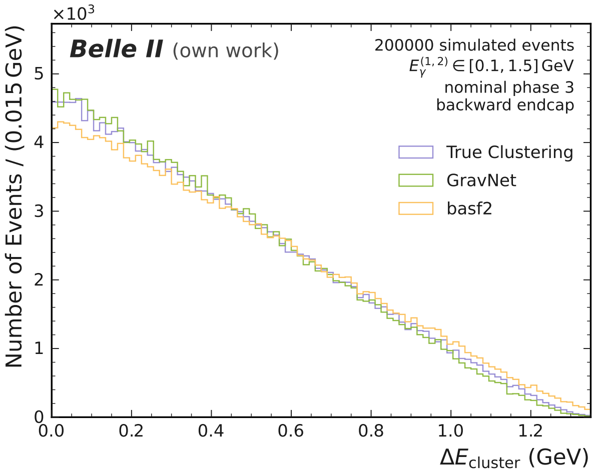


Figure 6.19.: Distribution of cluster energy differences $\Delta E_{\text{cluster}}$ for the two-cluster toy study with nominal phase 3 background in the backward endcap. The plot compares the true clustering, GravNet, and the basf2 baseline.

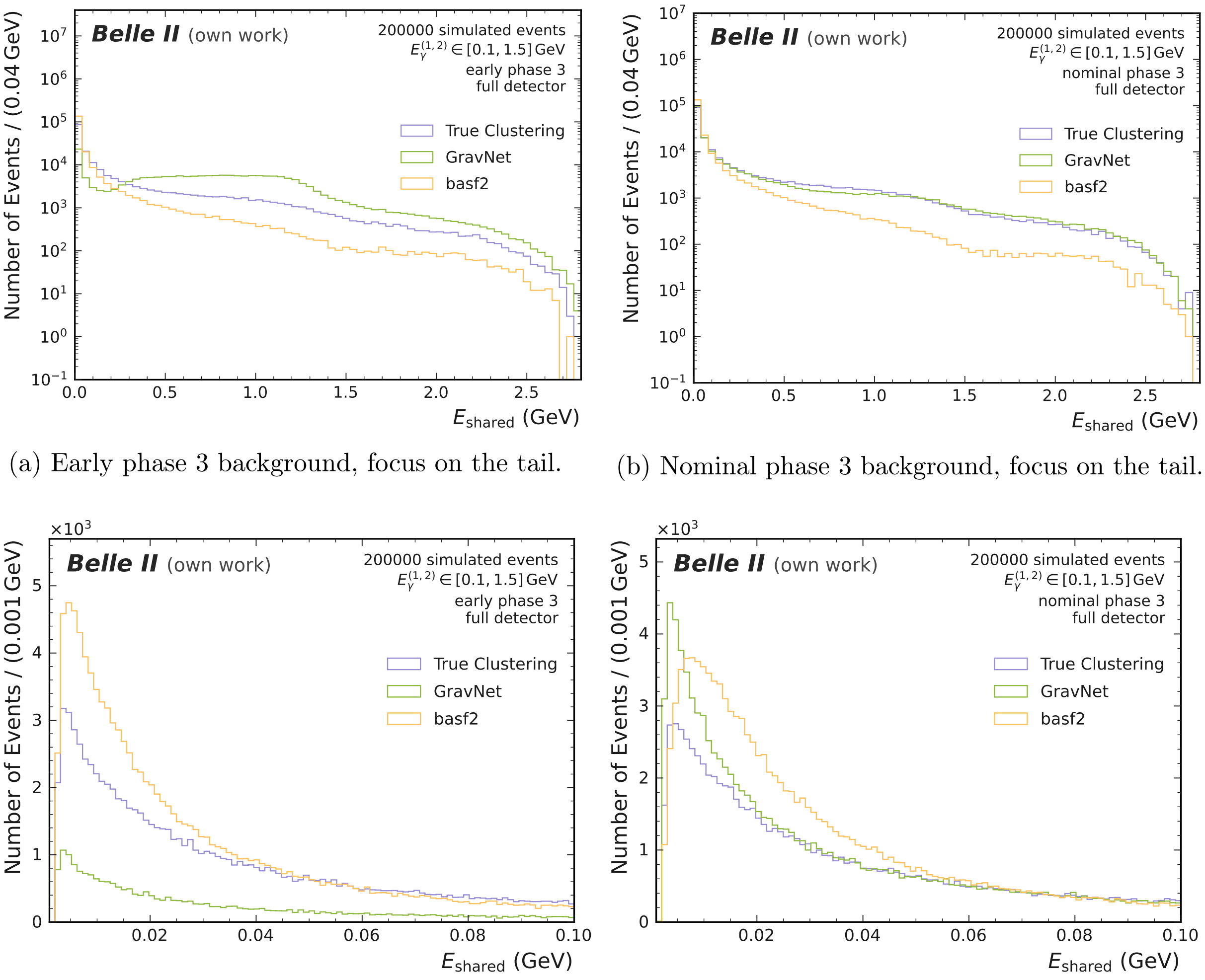
Once again, the distributions of the true clustering are almost identical for early and nominal backgrounds, thereby emphasizing that the photon generation is interchangeable for the two types of backgrounds. For both types of backgrounds, basf2 on average underestimates E_{shared}. This is explained by basf2 fully assigning crystals to either cluster instead of making use of its ability to fuzzy cluster (that is only for energy depositions associated with particles, not background).

GravNet constantly overestimates the shared energy in early background events. The dip at low E_{shared} and subsequent wide peak are likely caused by a specific pattern of events occurring in training but not in the test data set. The true distribution and GravNet match well in the high-energy tails for nominal background. This is attributed to the optimization of GravNet for nominal background (see section 4.3.2), which likely extends the capability of GravNet to recognize rare high-energy events. Contrary to that finding, in the peak of the distribution, GravNet displays a substantial preference to separate clusters instead of fuzzy clustering.

In conclusion, neither basf2 nor GravNet accurately model E_{shared}. Both algorithms present issues with the fuzzy clustering of shared crystals which is the most challenging part of the two-cluster scenario. GravNet takes a slight edge in the clustering of nominal background events. However, as already seen in the one-cluster toy studies, the effects of these clustering issues do not necessarily carry over to the performance in terms of resolution.

Cluster Center Distance

The cluster center distance Δx is another measurement of overlap. The distributions of Δx reveal identical trends and properties across all types of backgrounds and detector regions. Figure 6.21 shows the exemplary plot for early background events in the full detector. Basf2 significantly outperforms GravNet in reconstructing the correct distances between clusters. Part of the reason is that basf2 is inclined to fully assign the two LMs to each cluster. This decision is based on physics knowledge that GravNet first has to find in the training samples and even then does not automatically learn it. The bias towards smaller Δx also goes hand in hand with the finding that GravNet yields higher E_{shared}.



(a) Early phase 3 background, focus on the tail.

(b) Nominal phase 3 background, focus on the tail.

(c) Early phase 3 background, focus on the peak.

(d) Nominal phase 3 background, focus on the peak.

Figure 6.20.: Distribution of shared energies E_{shared} for the two-cluster toy studies. Each plot compares true clustering, GravNet, and basf2 baseline. The left side displays early phase 3 events, and the right side nominal phase 3 events. The top plots show the full distributions, the bottom plots zoom in on the peaks towards low E_{shared}.

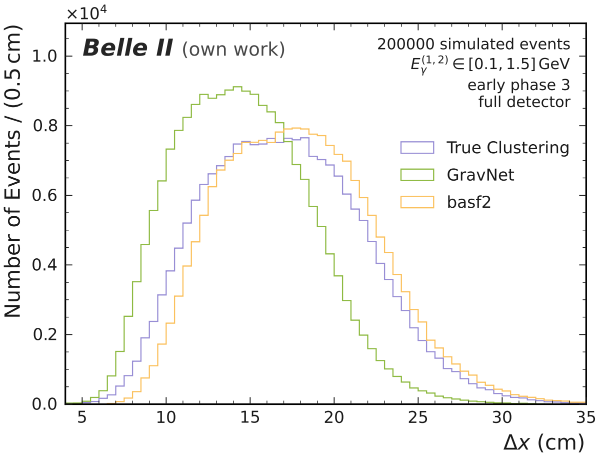


Figure 6.21.: Distribution of cluster center distances Δx for the two-cluster toy study with early phase 3 background in full detector coverage. The plot compares the true clustering, GravNet, and the basf2 baseline.

6.3.2. Performance Evaluation

The performance evaluation starts with a look at the clustering quality using the FCAI. Sensitivity and precision aim to uncover the mechanics of the underlying clustering. Afterward, the energy resolution studies the overall results of the energy reconstruction. Lastly, the energy dependence of the resolution is examined in more detail. The most relevant metrics are highlighted, and all metrics, including fit parameters, for all detector regions are located in appendix D.2.

Fuzzy Clustering Agreement Index

Figure 6.22 shows the distribution of FCAIs for early and nominal backgrounds. The medians confirm that the clustering results are far from random and both algorithms produce excellent clustering results in the presence of challenging overlap. The FCAI score backs up that GravNet is able to adapt exceptionally well to nominal background. This is most likely justified by the hyperparameter and feature optimization for this exact scenario. While basf2 significantly loses performance when confronted with nominal background, GravNet has a more stable median. It is noteworthy that despite the otherwise excellent performance, no events with perfect FCAI exist for two-cluster events contrary to the basf2 results with one-cluster events. Nevertheless, basf2 reduces the gap in performance, partly because it is able to use its ability to fuzzy cluster energy depositions of two particles now.

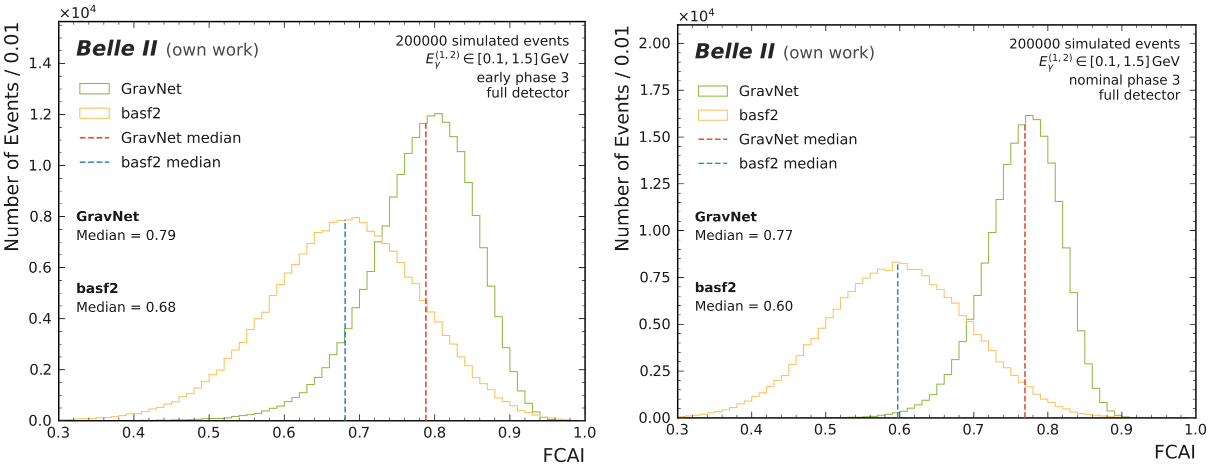


Figure 6.22.: Distribution of FCAIs for the two-cluster toy studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Sensitivity and Precision

The differences in averaged sensitivity S_{avg} and precision P_{avg} in the three detector parts are minor for two-cluster events. Figure 6.23 displays the correlation and marginal distributions for the full detector in early and nominal background. For early background, GravNet is significantly less sensitive and only marginally improves on the precision. In contrast to the one-cluster toy study, this result does not vary substantially in different detector regions. The improvements with nominal background events are more pronounced and expected considering the one-cluster toy study results and the optimization of GravNet for nominal background. In general, S_{avg} and P_{avg} are in the median a bit lower in the two-cluster studies than in the one-cluster studies. Nonetheless, both algorithms are clearly able to correctly identify energy depositions also on a crystal level.

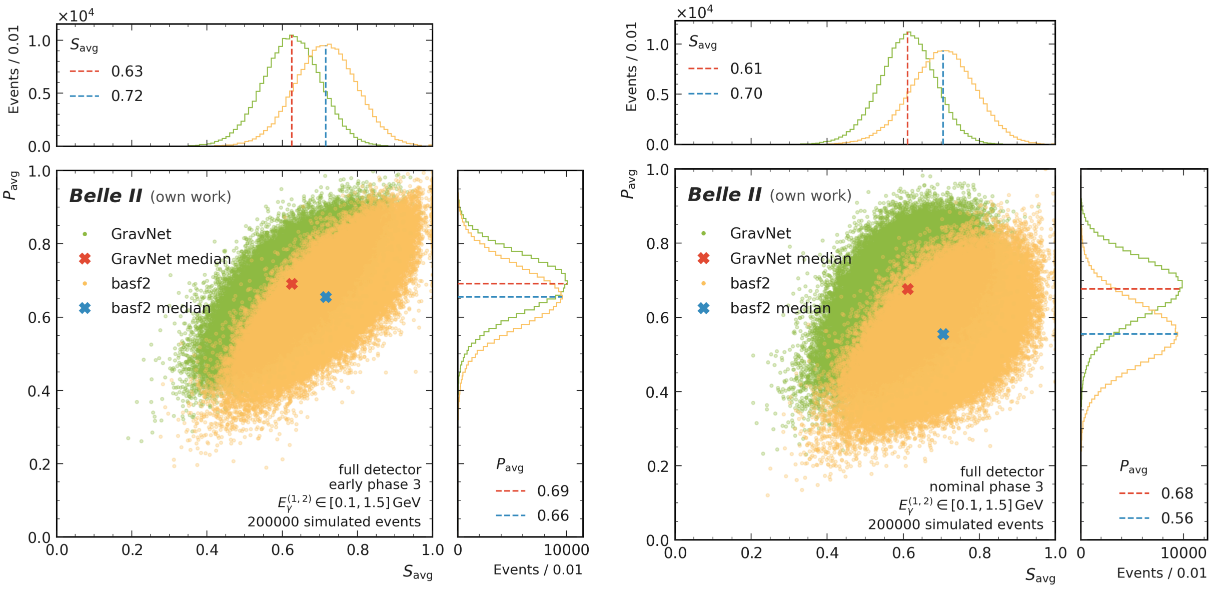


Figure 6.23.: Correlation and marginal distributions for average sensitivity S_{avg} and average precision P_{avg} for the two-cluster toy studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated in the marginal distributions. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

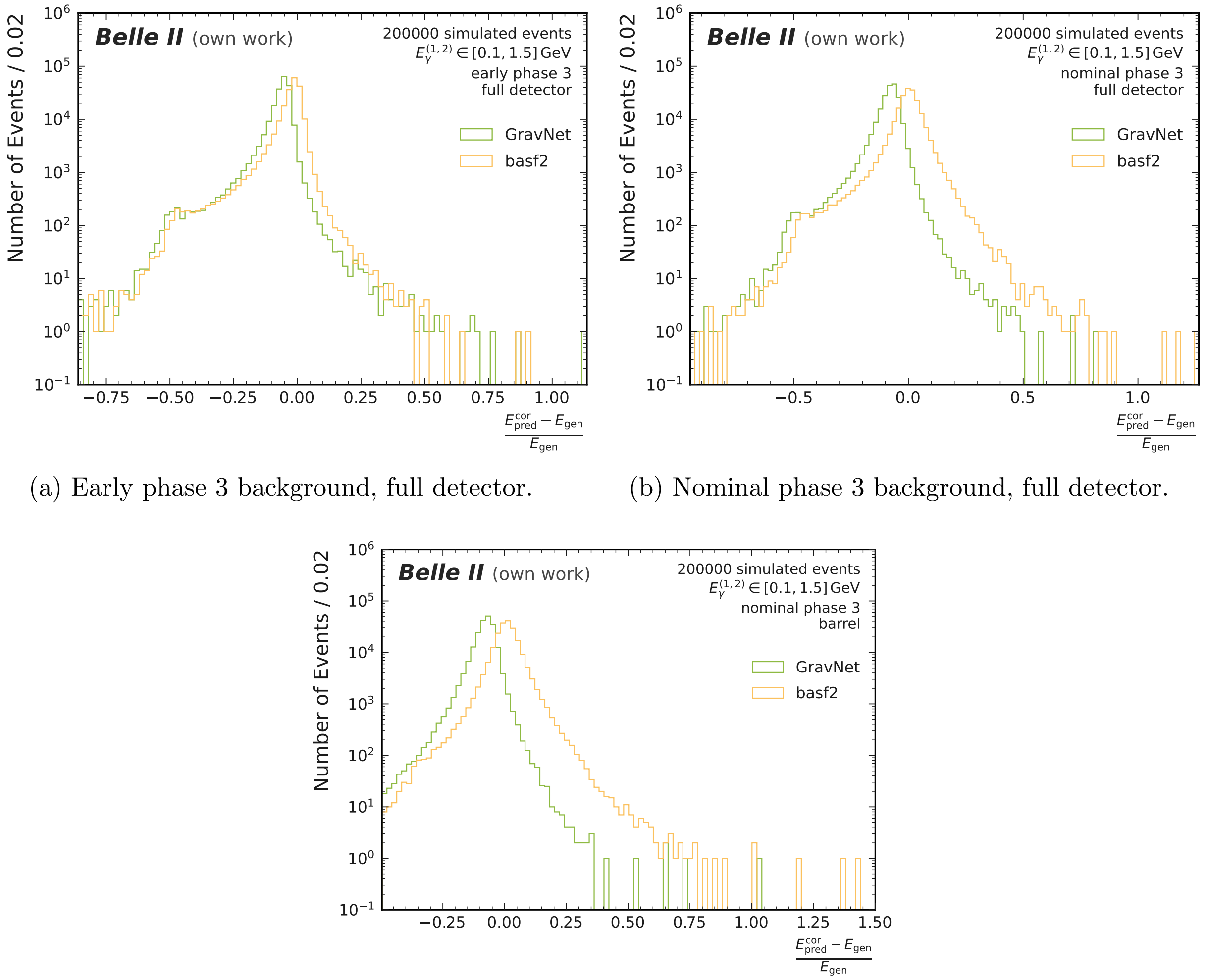
Energy Resolution

Before looking at the FWHM of the reconstruction errors η_{dep} and η_{gen}, it is worth paying attention to the full distributions of η_{gen} including outliers. Figure 6.24 shows the distribution of η_{gen} for early and nominal background in the full detector. The figure is complemented by η_{gen} in the barrel for nominal background. In both early and nominal background there is a notable bump in the tails towards negative errors. This bump is caused by events with overlapping photons, such that energy depositions of one cluster are located on the opposite side of the second cluster. This does not necessarily lead to large E_{shared} but is rather pointed out by a small cluster center distance Δx. Figure 6.27 shows an event display that demonstrates that both algorithms have difficulties with the correct identification. The bump does not exist in the barrel, implying that these most challenging signatures are less frequent in the regular structure of the barrel.

After examining the full distributions, the resulting resolutions for all detector parts are determined. Figure 6.25 displays the fits for η_{dep} for both types of backgrounds in barrel, forward, and backward endcap. The FWHM does not quantify the resolution of one photon anymore, but rather the resolution of the whole event. As a result, resolutions are not comparable to the one-cluster studies (see section 5.3.2). However, identical trends emerge with the best resolution in the forward endcap, then the barrel, and the worst in the backward endcap. These trends are equal for early and nominal backgrounds. The relative improvements are similar to the one-cluster toy studies with approximately 40 % in nominal background and a range of 32 % to 54 % in early background. With early background, GravNet reduces the left tails by a factor of two whereas the right tails remain roughly the same. For nominal background, GravNet reduces left and right tails alike, although not as drastically. This indicates that GravNet is able to better learn the edge cases that were presented throughout. The biases for the peaks towards negative and positive values alike are discussed in the one-cluster toy studies section 6.2.2 and traced back to the same causes here.

Figure 6.26 displays the resolution on η_{gen} for both types of backgrounds and all detector regions. In contrast to previous findings, the best resolution is found in the barrel, then the forward endcap, then the backward endcap for both types of backgrounds. Overall, the advances are less pronounced than in the one-cluster toy studies in early background events. The background endcap receives 9.4 % improvement, while especially in the forward endcap, the improvements diminish to the point where they are within the uncertainties. This is unexpected, given the large improvements to the underlying clustering and η_{dep}. Uncorrected leakage which now affects two photon energies, causes the even larger discrepancy between the errors on deposited and generated energies. The tails in early background receive only little change in comparison to the basf2 baseline.

Even though the improvements in η_{dep} are about equally large in early and nominal background events, much more carries over to η_{gen} for nominal background events. Over 10 % improvements are found in every detector part, with the backward endcap being the most advanced at 17 %. Additionally, GravNet again is less prone to overestimate energies and significantly reduces the right tails.

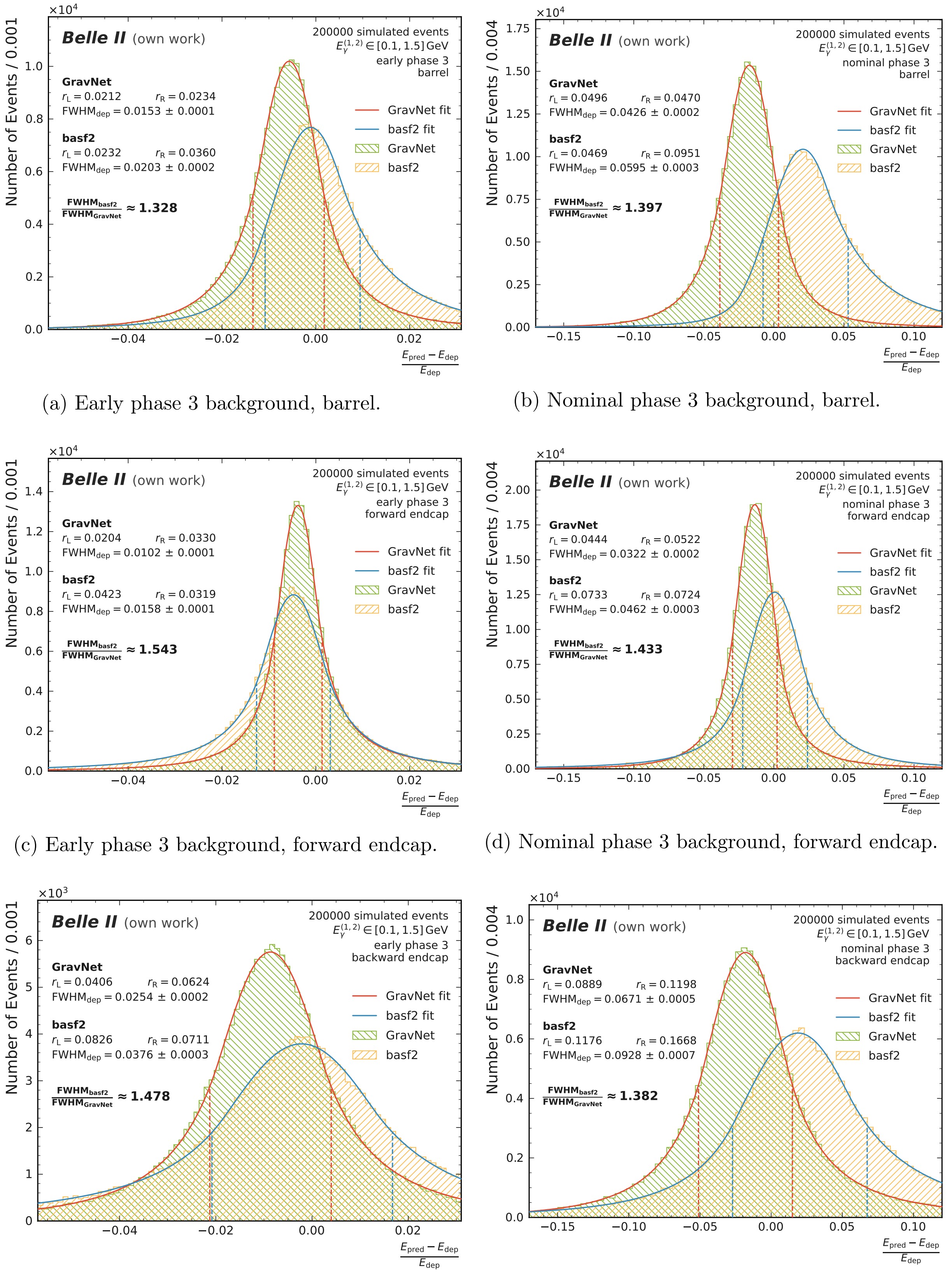


(a) Early phase 3 background, full detector.

(b) Nominal phase 3 background, full detector.

(c) Nominal phase 3 background, barrel.

Figure 6.24.: Distributions of reconstruction errors on the generated energy η_{gen} for the two-cluster toy studies. Each plot compares GravNet and basf2 baseline.



(a) Early phase 3 background, barrel.

(b) Nominal phase 3 background, barrel.

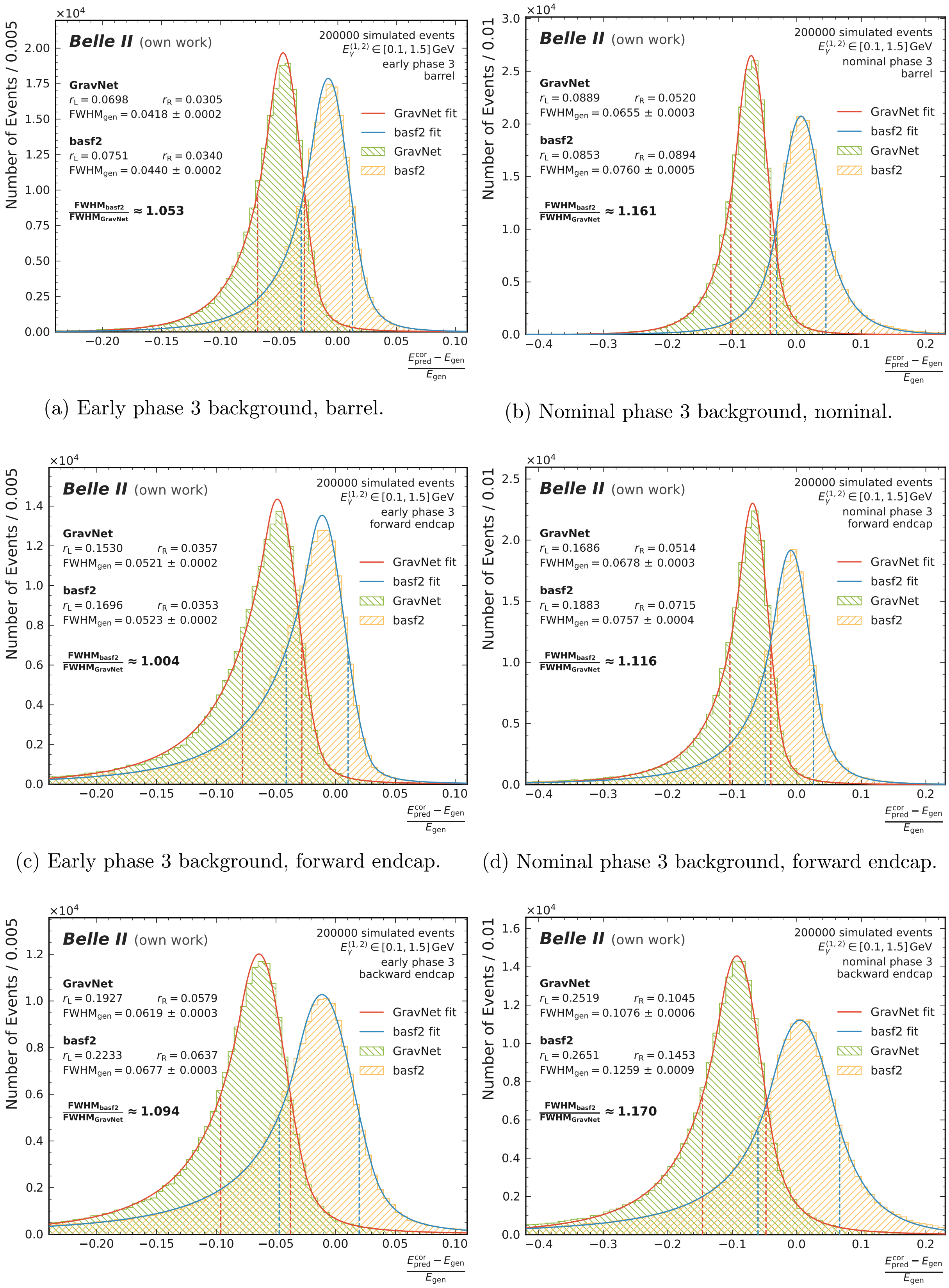
(c) Early phase 3 background, forward endcap.

(d) Nominal phase 3 background, forward endcap.

(e) Early phase 3 background, backward endcap.

(f) Nominal phase 3 background, backward endcap.

Figure 6.25.: Fits for the distributions of reconstruction errors on the deposited energy η_{dep} for the two-cluster toy studies. Each plot compares GravNet to the basf2 baseline. The left plots depict early phase 3 events, and the right plots nominal phase 3 events. Barrel, forward, and backward endcaps are displayed separately in this order.



(a) Early phase 3 background, barrel.

(b) Nominal phase 3 background, nominal.

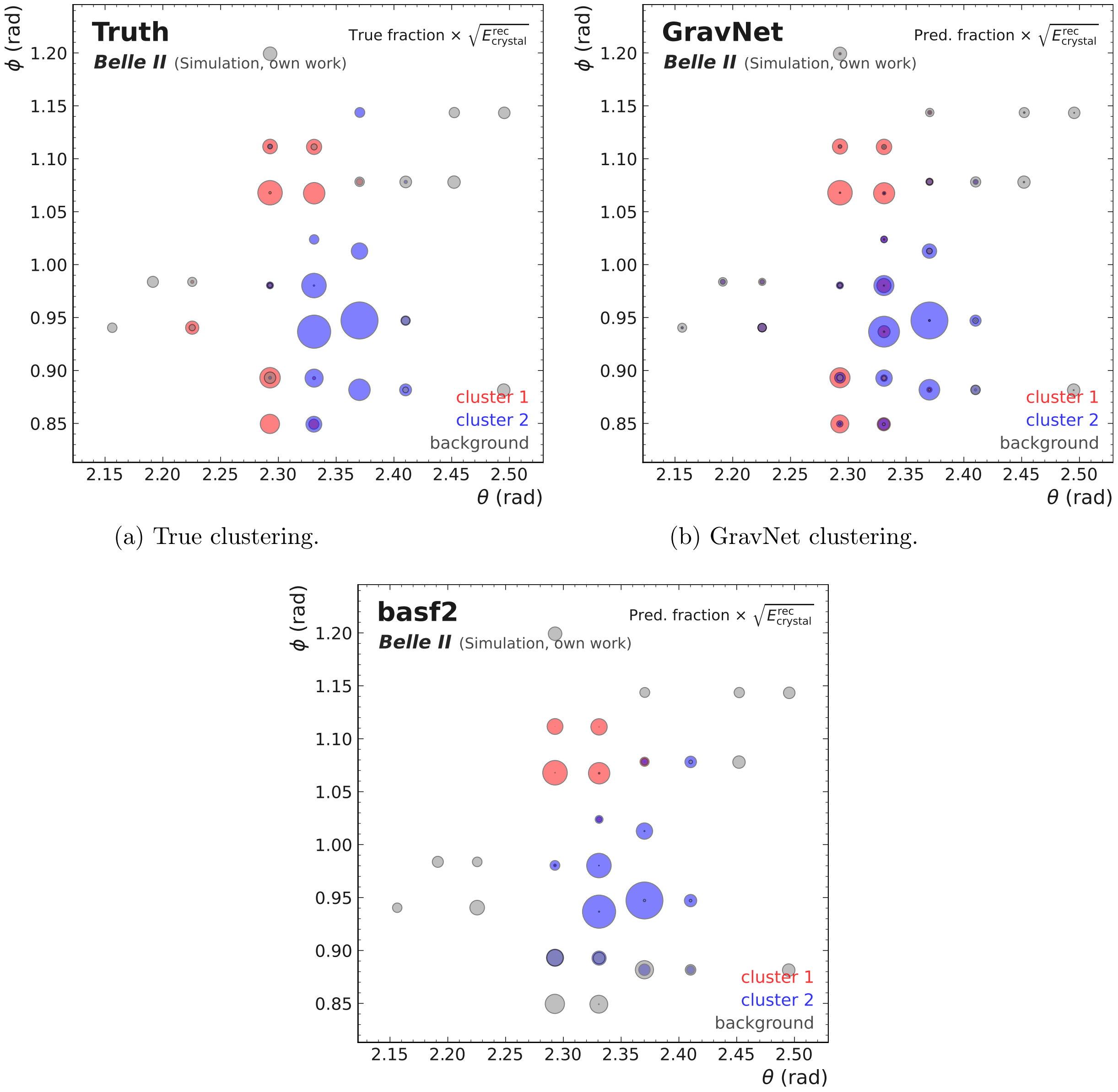
(c) Early phase 3 background, forward endcap.

(d) Nominal phase 3 background, forward endcap.

(e) Early phase 3 background, backward endcap.

(f) Nominal phase 3 background, backward endcap.

Figure 6.26.: Fits for the distributions of reconstruction errors on the generated energy η_{gen} for the two-cluster toy studies. Each plot compares GravNet to the basf2 baseline. The left plots depict early phase 3 events, and the right plots nominal phase 3 events. Barrel, forward, and backward endcaps are displayed separately in this order.



(a) True clustering.

(b) GravNet clustering.

(c) Basf2 clustering.

Figure 6.27.: Example of an event with a negative reconstruction error on the generated energy η_{gen} from the two-cluster toy study with early phase 3 background. $\eta_{\text{gen}} = -0.609$ for the basf2 clustering, and $\eta_{\text{gen}} = -0.505$ for the GravNet clustering. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\text{rec}}}$.

Energy Dependence

The approach to analyzing the energy dependence for two-cluster events differs from that in the previous studies. In two-photon events, the energy resolution for one photon is on one hand dependent on its own energy as is analyzed in the one-cluster toy study in section 6.2.2. On the other hand, the presence of the second photon including its overlapping energy depositions affects the resolution as well. When combining two high-energy photons this effect is not as pronounced as in the combination of a high-energy photon with a low-energy one. Intuitively, low-energy photons are at a disadvantage and might get lost in the larger energy depositions of a high-energy photon. The paragraph about the cluster energy difference discusses the fundamentals of energetically asymmetric clusters and shows that both algorithms are able to separate strongly asymmetric clusters. The upcoming analysis extends this concept to combine the dependence of the resolution on the photon energy itself with the dependence on the asymmetry in energies.

Four intervals in generated photon energies are defined in order to categorize photons for the analysis: $E_\gamma^{(1,2)} \in \{[0.1, 0.2],\ [0.2, 0.5],\ [0.5, 1.0],\ [1.0, 1.5]\}$ GeV. These intervals are chosen to roughly represent photon energies that result in comparable cluster radii and thereby indirectly also a similar number of included crystals in a cluster. Next, all 16 possible combinations of two photons from these intervals are formed. Events from the existing data sets are selected according to these combinations, with one photon becoming the primary photon and the other the secondary. The resolution for the primary photons is determined for every combination of intervals. Note that this is different from the event-wise resolution considered otherwise and comparable to the one-cluster toy studies. Figure 6.28 shows an example of the resolutions for two combinations of photon energy intervals. As expected, the high-energy photons from the interval $E_\gamma \in [1.0, 1.5]$ GeV have a better energy resolution than the photons with $E_\gamma \in [0.1, 0.2]$ GeV. The rest of the plots are in appendix D.3.

Figure 6.29 presents the results for early and nominal background events in a heat map. The exact numbers are additionally summarized in appendix D.3. In comparison between early and nominal background events, the FWHMs differ in absolute values, however, identical trends are identified. Analogous to the one-cluster toy studies, the resolution is mainly driven by the photon energy and improves rapidly going from low-energy to high-energy photons. The dependence on the second photon is much more subtle, intuitively worsening with the second photon increasing in energy for both algorithms. Despite the absolute resolution following only a little pronounced trend, the implications on the relative improvements (indicated by the color) between GravNet and basf2 baseline are sizable. For events with a smaller difference in cluster energies, GravNet generally outperforms basf2 by a larger margin than it does for large cluster energy differences. The findings in the distributions of cluster energy differences and cluster radius differences are in line with this result.

Ultimately, in comparison to the photon resolution of one-cluster events shown in figures 6.14 and 6.15, the two-cluster photon resolution holds up to the one-cluster benchmark with early background. In nominal background events, the basf2 resolution deteriorates for low-energy photons, while GravNet manages to achieve very similar results as in the one-cluster toy study over the full photon energy range.

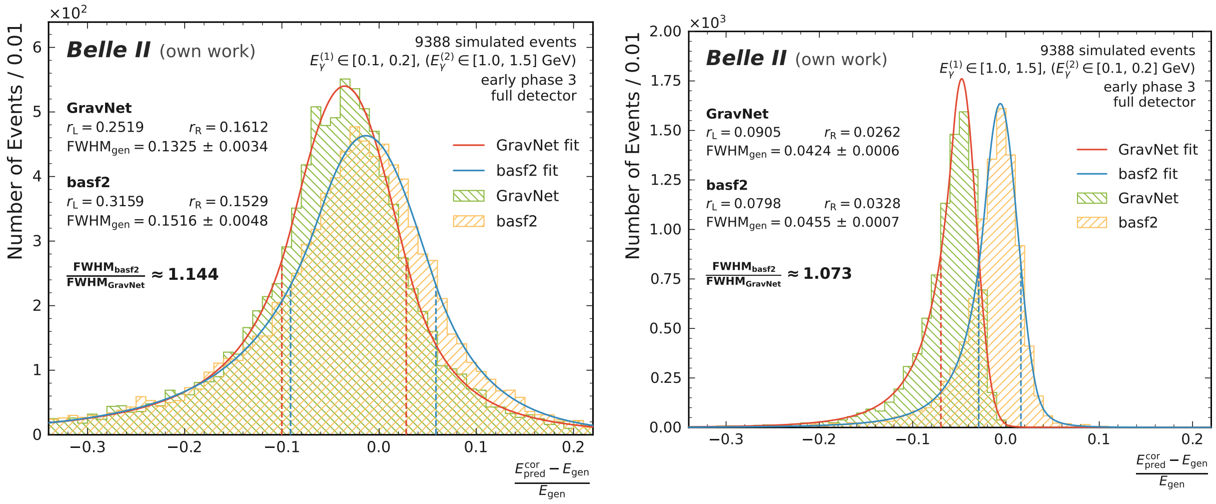


Figure 6.28.: Fits for the distributions of reconstruction errors on the generated energy η_{gen} for the two-cluster toy study events with early phase 3 background in full detector coverage. The left plot displays the resolution for photons with $E_\gamma^{(1)} \in [0.1, 0.2]$ GeV in the presence of a second photon with $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV. The plot to the right displays the resolution for photons with $E_\gamma^{(1)} \in [1.0, 1.5]$ GeV in the presence of a second photon with $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

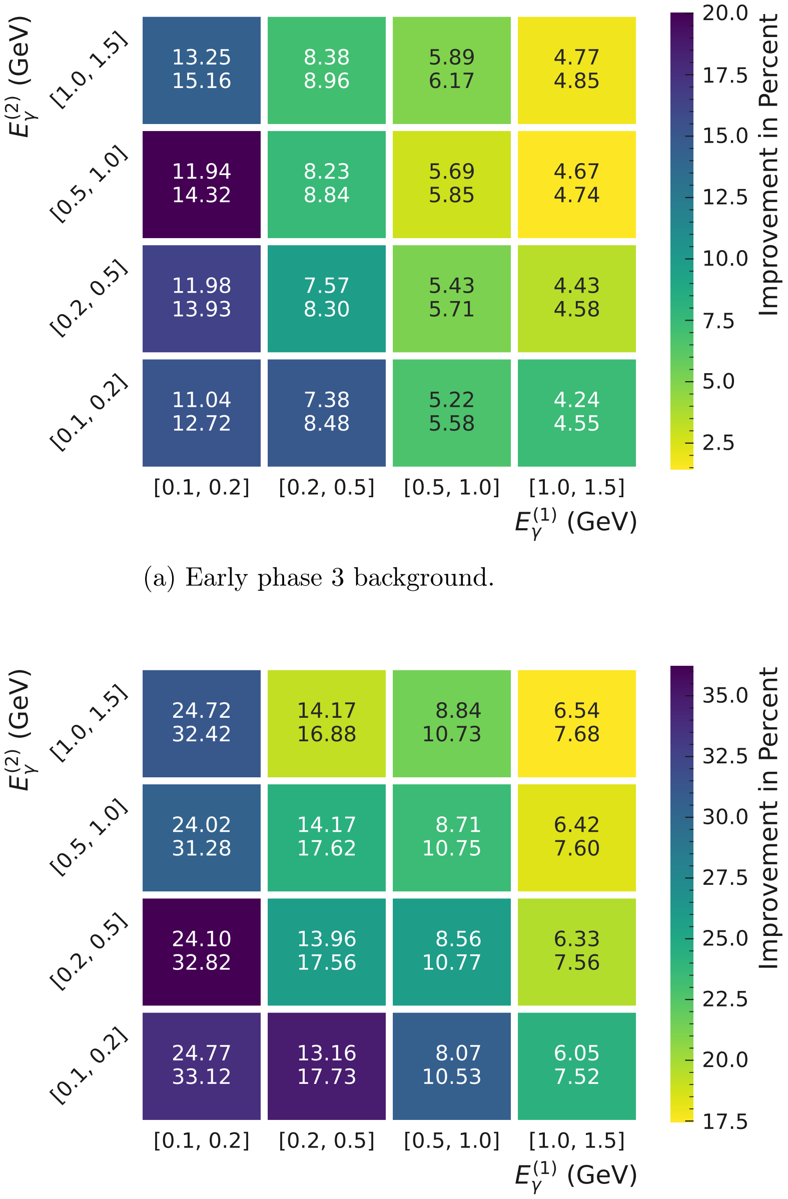
6.3.3. Overview and Conclusions

The most important results for $\mathrm{FWHM_{dep,gen}}$, S_{avg} and P_{avg} are summed up in table 6.4 for early background events. Table 6.5 presents the results of the nominal background study. Many of the results and correlations from the one-cluster toy studies reappear for two overlapping clusters.

GravNet proves to be the better clustering algorithm in the FCAI score and in the clustering of extreme cluster signatures. This is especially true in nominal background events where the excellent clustering is most likely the result of the hyperparameter optimization and feature analysis for this specific scenario. Basf2 has a small advantage in dealing with asymmetric clusters as seen in cluster energy difference, cluster radius difference, and the energy dependence analysis.

In comparison to the one-cluster toy studies, the resolution for two overlapping photons turns out to be no issue in early background for both algorithms. Even so, GravNet manages to outperform the basf2 baseline by a magnitude of a few percent. With nominal background, the results are remarkable given the challenging scenario, but basf2 takes a significant hit in performance relative to one-cluster events, while GravNet performs similarly. This finding is not only present on an event level but also on a crystal level, emphasized by sensitivity and precision.

Ultimately, the improvements in the event-wide resolution amount to approximately 5 % in early background and 15 % in nominal background. For the single photon resolution, the advancements span a wide range of 2 − 15 % in early background depending on the photon energy. In nominal background at least 17 % to 35 % improvement on the single photon resolution are found. Once more, GravNet brings the largest advancements to low-energy photons, especially in nominal background events. Edge cases like one disjoint cluster and high shared energies are another regime where GravNet substantially outperforms the basf2 baseline thanks to its better clustering capabilities.



(a) Early phase 3 background.

(b) Nominal phase 3 background.

Figure 6.29.: Heat maps of the relative improvements of GravNet over basf2 for the two-cluster toy studies in full detector coverage. Shown are the resolutions $\mathrm{FWHM}_{\mathrm{gen}} \times 10^{-2}$ of photons with generated energies $E_\gamma^{(1)}$ in the presence of a second photon with generated energy $E_\gamma^{(2)}$. The top number in each cell displays the absolute GravNet resolution, and the bottom number the basf2 resolution. The relative improvement between the two resolutions is indicated by the color.

Table 6.4.: Summary and comparison of the performance of the algorithms for two-cluster toy study events with early phase 3 background. The metrics $\mathrm{FWHM}_{\mathrm{dep,gen}}$, average sensitivity S_{avg} and precision P_{avg}, and FCAI are listed for different detector regions. The improvement to the basf2 baseline is stated in percent for each region.

Detector Region	Algorithm	$\mathrm{FWHM}_{\mathrm{dep}}$ $\times 10^{-2}$	$\mathrm{FWHM}_{\mathrm{gen}}$ $\times 10^{-2}$	S_{avg}	P_{avg}	FCAI
Barrel	GravNet	1.53	4.18	0.63	0.69	0.79
	basf2	2.03	4.40	0.72	0.66	0.69
	Improvement	**32.8 %**	**5.3 %**	**-12.5,%**	**4.5 %**	**14.5 %**
Forward Endcap	GravNet	1.02	5.21	0.67	0.73	0.78
	basf2	1.58	5.23	0.73	0.70	0.64
	Improvement	**54.3 %**	**0.4 %**	**-8.2 %**	**4.2 %**	**21.9 %**
Backward Endcap	GravNet	2.54	6.19	0.59	0.66	0.80
	basf2	3.76	6.77	0.70	0.61	0.67
	Improvement	**47.8 %**	**9.4 %**	**-15.7 %**	**8.2 %**	**19.4 %**
Full Detector	GravNet	1.53	4.44	0.63	0.69	0.79
	basf2	2.06	4.58	0.72	0.66	0.68
	Improvement	**34.6 %**	**3.2 %**	**-12.5 %**	**4.5 %**	**16.1 %**

Table 6.5.: Summary and comparison of the performance of the algorithms for two-cluster toy study events with nominal phase 3 background. The metrics $\mathrm{FWHM}_{\mathrm{dep,gen}}$, average sensitivity S_{avg} and precision P_{avg}, and FCAI are listed for different detector regions. The improvement to the basf2 baseline is stated in percent for each region.

Detector Region	Algorithm	$\mathrm{FWHM}_{\mathrm{dep}}$ $\times 10^{-2}$	$\mathrm{FWHM}_{\mathrm{gen}}$ $\times 10^{-2}$	S_{avg}	P_{avg}	FCAI
Barrel	GravNet	4.26	6.55	0.62	0.68	0.77
	basf2	5.95	7.60	0.71	0.55	0.59
	Improvement	**39.7 %**	**16.1 %**	**-12.7 %**	**23.6 %**	**30.5 %**
Forward Endcap	GravNet	3.22	6.78	0.60	0.67	0.77
	basf2	4.62	7.57	0.70	0.59	0.63
	Improvement	**43.3 %**	**11.6 %**	**-14.3 %**	**13.6 %**	**22.2 %**
Backward Endcap	GravNet	6.71	10.76	0.61	0.63	0.76
	basf2	9.28	12.59	0.70	0.51	0.55
	Improvement	**38.2 %**	**17.0 %**	**-12.9 %**	**23.5 %**	**38.2 %**
Full Detector	GravNet	4.05	6.69	0.61	0.68	0.77
	basf2	5.99	7.68	0.70	0.56	0.60
	Improvement	**47.9 %**	**14.8 %**	**-12.9 %**	**21.4 %**	**28.3 %**

7. Physics Studies

ML models often struggle to transfer the performance on selected toy data into real-life applications. This chapter brings the GravNet photon reconstruction from the toy studies to application in the reconstruction of neutral pions wehich primarily decay into two photons. More specifically, the pion mass is reconstructed as the invariant mass of a two-photon system. The results demonstrate how much of the improvements in photon energy resolution carry over to the reconstruction process and thereby determine the actual value of the algorithm to further physics analyses. Albeit still relying on MC data instead of actual detector data, checking the performance for possible applications in this setting including an MC truth is an essential first step in validation.

Each toy study from the previous chapter is related to one physics study in this chapter. The GravNet models trained for the respective toy study are now used to evaluate an independent physics data set. The models for early and nominal background from the one-cluster toy study in section 6.2 find application in the reconstruction of weakly boosted neutral pions in section 7.3. The following study of the decays of highly boosted pions is handled by the models from the two overlapping cluster toy study in section 6.3 for the two types of backgrounds.

The first section 7.1 motivates and introduces the different settings for the studies. Section 7.2 presents the reconstruction of neutral pions in these settings and the evaluation process used in the following studies. The photons in the toy study scenarios are intended to fit the kinematics and characteristics of neutral pion decays at Belle II. For this reason, the models from the toy studies are used in the physics studies without alterations or retraining. Nonetheless, the scenarios are not exactly identical and sections 7.3.1 and 7.4.1 give a brief overview of the photon performance of the algorithms in the new scenarios. The main parts in sections 7.3.2 and 7.4.2 then focus on the pion performance. Section 7.5 sums up the results of the studies and comes to final conclusions.

7.1. Motivation and Settings

The decay of neutral pions $\pi^0 \rightarrow \gamma\gamma$ is a prevalent source for two photons that are registered in the Belle II ECL. Neutral pions occur in direct production and as decay products in many physics processes, having a large impact on frequent physics analyses [6]. In the scope of the physics studies in this chapter, only neutral pions are considered, therefore, the term pion always refers to the neutral pion.

Once again the decays of $\mathrm{B}^+\ \mathrm{B}^-$ and $\mathrm{B}^0\ \overline{\mathrm{B}}^0$ at Belle II are used to get an idea of the relevant momentum range of pions. In accordance with the photon spectrum in section 6.1, pions up to the third generation in the decay chain of these initial states are considered. Figure 7.1 shows the resulting spectrum for the pion momenta.

Based on this spectrum, a momentum range of $p_{\pi^0} \in [0.2, 2.7]\,\mathrm{GeV/c}$, is chosen for the studies of pion decays. However, the full momentum range is not treated in one study but rather split up into weakly boosted pions and highly boosted pions, the reason being that different momenta result in distinct cluster signatures in the ECL.

Weakly boosted pions in a momentum range of $p_{\pi^0} \in [0.2, 2.0]$ GeV/c leave two isolated photon clusters in the ECL. The two clusters are treated individually, one after another, by the GravNet models from the one-cluster toy study in section 6.2. Section 7.3 studies the reconstruction of weakly boosted pions with early background and nominal background. Due to the large boost, pions with momenta $p_{\pi^0} \in [2.0, 2.7]$ GeV/c result in two overlapping clusters. Section 7.4 uses the GravNet models from the two overlapping cluster toy study in section 6.3 to evaluate these events and study the pion reconstruction.

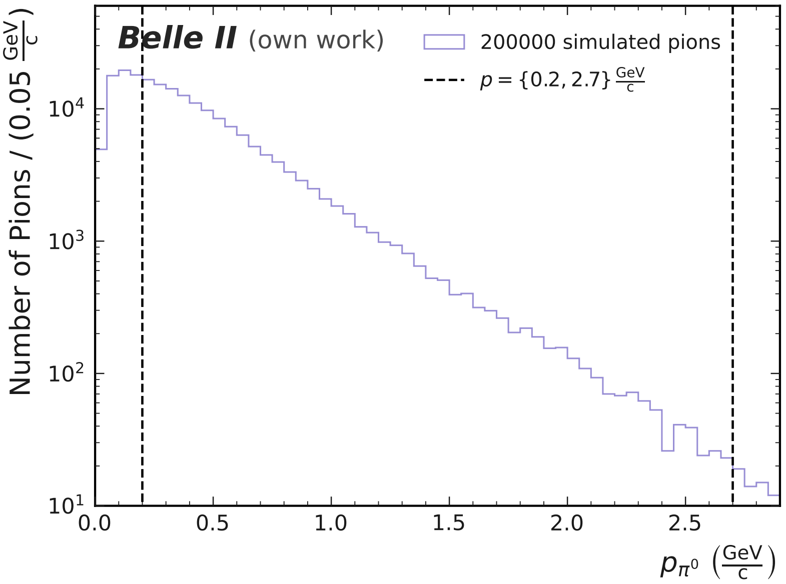


Figure 7.1.: Pion momentum p_{π^0} spectrum for simulated neutral pions, originating from $\mathrm{B}^+\ \mathrm{B}^-$ and $\mathrm{B}^0\ \overline{\mathrm{B}}^0$ decays at Belle II. Pions up to the third generation in the decay chain of these initial particles are taken into account.

7.2. Fundamentals of Neutral Pion Reconstruction

This section explains the basics of pion reconstruction and the methods used for the evaluation of the pion performance. This includes the introduction of a leakage correction to GravNet and the theoretical analysis of the energy and position dependence of the reconstruction.

The studies in this chapter reconstruct the pion mass from the photon four-vectors $P_\gamma^{(1,2)}$. The reconstructed pion mass is the invariant mass of the two-photon system $M_{\gamma\gamma}$, derived as

$$\begin{aligned} M_{\gamma\gamma} &= \sqrt{\left(P_\gamma^{(1)} + P_\gamma^{(2)}\right)^2} \\ &= \sqrt{2E_\gamma^{(1)} E_\gamma^{(2)}\,(1 - \cos\alpha)}, \end{aligned} \tag{7.1}$$

with the separation angle α between $P_\gamma^{(1)}$ and $P_\gamma^{(2)}$ in the laboratory frame. GravNet aims to improve the energy resolution for the photons. The other contributing factor is the position reconstruction that ultimately determines α. The four-vector directions are either provided by the basf2 position reconstruction (see section 2.3) or by the true MC information.

Several ways exist to obtain and combine energies and positions in order to analyze the pion reconstruction from various aspects. The present work will focus on three possible realizations:

- **Reconstructed energy + reconstructed position:** The reconstructed energy $E_{\text{pred}}^{\text{cor}}$ from either basf2 or GravNet is combined with the reconstructed position of basf2. This combination specifies the currently possible resolution using only reconstructed values from existing frameworks. It is the main point of comparison for the performance evaluation.
- **Reconstructed energy + MC true position:** The reconstructed energy from either basf2 or GravNet is combined with the true position from MC information. This combination reveals the best possible resolution with the respective energy reconstruction, independent of the position reconstruction. In contrast to the first variant, it emphasizes the full potential of the energy reconstruction algorithms.
- **MC true energy + reconstructed position:** The true MC energy $E_{\text{gen}} = E_\gamma$ is combined with the reconstructed position of basf2. This combination demonstrates the resolution that is theoretically possible with perfect energy reconstruction but is limited by the current position reconstruction. A confluence of perfect clustering and perfect leakage correction would result in this resolution.

Given the reconstructed invariant mass, the determination of the mass resolution is along the lines of the photon energy resolution introduced in section 5.3.2. Like the reconstruction errors on the photon energy before, the reconstructed masses are looked at for all events in a test data set. The resulting distribution is characterized by a peak around the true pion mass at $m_{\pi^0} = 134.97\,\text{MeV/c}^2$. The peak is fitted with the DCB function defined in equation (5.16) and described by the FWHM and the tail lengths $r_{\text{L,R}}$. These properties are corrected for potential biases to

$$\text{FWHM}_{\text{cor}} \pm \delta_{\text{FWHM}}^{\text{cor}} = \left(\frac{m_{\pi^0}}{\mu}\right) \cdot \Big(\text{FWHM} \pm \delta_{\text{FWHM}}\Big). \tag{7.2}$$

This is equivalent to shifting the peaks to the true pion mass and thereby yields the actual resolution of an otherwise biased algorithm. The correction of the tail lengths follows the same principle.

GravNet Leakage Correction

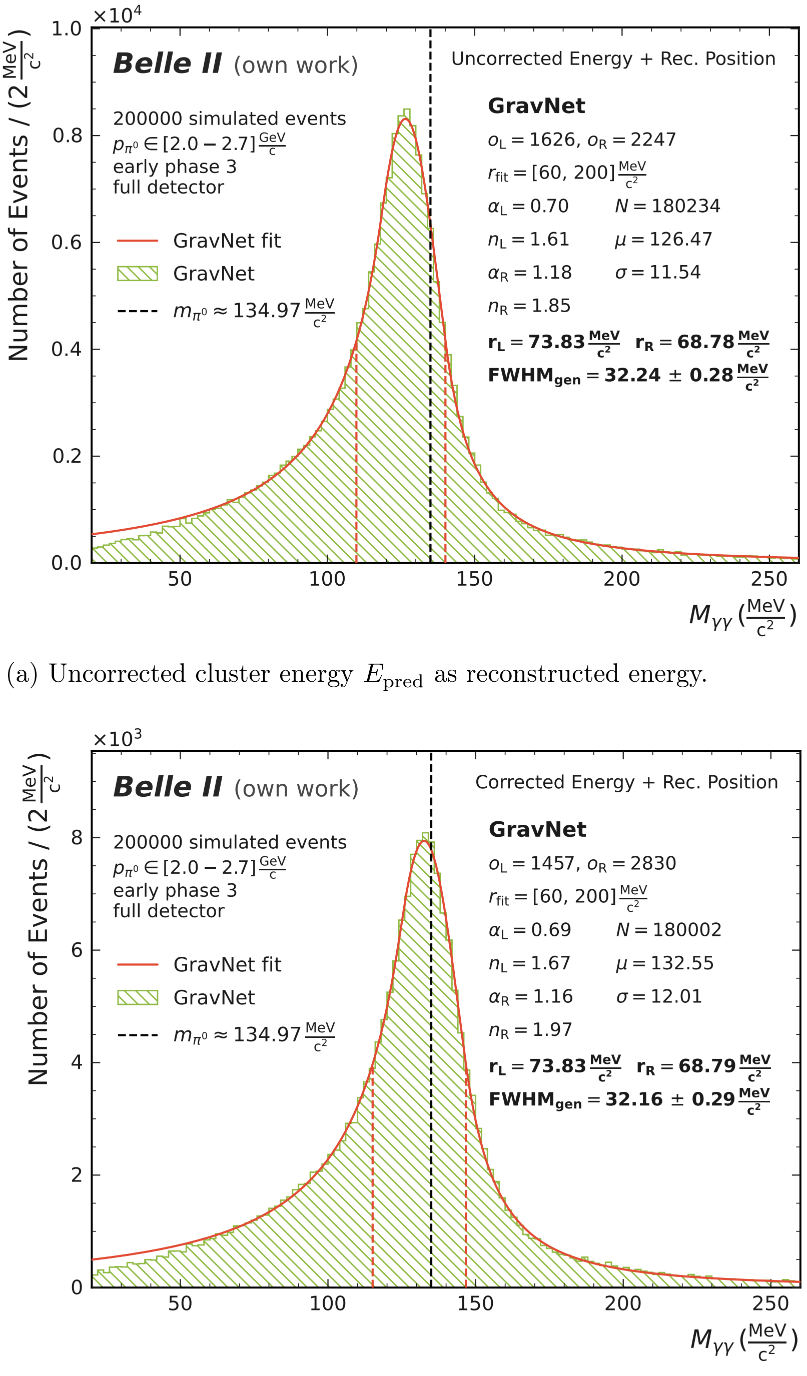
The toy studies in section 6.2.2 discuss the bias of GravNet towards lower energies. The bias in η_{gen} is mostly caused by the absence of a leakage correction for GravNet, while basf2 uses an advanced correction. So far, the simple predicted energy E_{pred} from equation (5.2) was the basis for any performance analysis of GravNet. The effects of a missing correction aggregate in the pion reconstruction from two photons and lead to an even larger offset to the peak. For this reason, the upcoming pion studies provide a fitting scenario for the investigation of the effects of a basic leakage correction for GravNet. This correction is based on the toy studies and yields an individual correction factor c_E for each of the four scenarios:

$$c_E = 1 - \mu. \tag{7.3}$$

Here, μ is the position of the peak of the distribution of η_{gen}, according to the fit parameter. The fit is taken from the corresponding toy study in full detector coverage. The correction factors are stated at the beginning of each study. Note that the corrections do not depend on any of the physics studies, but are exclusively determined by the generic toy studies. This concept is not far from actual leakage correction processes and yields a universal correction that is not tuned toward specific studies. The correction factor is used to calculate the corrected energy $E_{\text{pred}}^{\text{cor}}$ for GravNet as

$$E_{\text{pred}}^{\text{cor}} = c_E \cdot E_{\text{pred}}. \tag{7.4}$$

Figure 7.2 shows an example of the fit for the distribution of reconstructed invariant masses, highlighting the effects of the leakage correction on GravNet. The physics studies are entirely built around the corrected energies for both GravNet and basf2. Despite the leakage correction, the additional corrections to the FWHM, discussed in the previous paragraph and defined in equation (7.2), are applied throughout all studies and account for any remaining biases.



(a) Uncorrected cluster energy E_{pred} as reconstructed energy.

(b) Leakage-corrected cluster energy $E^{\text{cor}}_{\text{pred}}$ as reconstructed energy.

Figure 7.2.: Distribution of reconstructed invariant masses of the two-photon system $M_{\gamma\gamma}$. The plots compare two versions of GravNet energy reconstruction, combined with the basf2 position reconstruction. The results are from the highly boosted pion study with early background in section 7.3.

Energy and Position Dependence

Energy and position resolution have varying impacts on the mass resolution, depending on the energies and the separation of the photons. The contributions are estimated by the propagation of uncertainty for the squared invariant mass. In small-angle approximation for α and assuming independent values for both energies and angle, the uncertainty on the pion mass is given by

$$\sigma_{m^2} \approx m_{\pi^0}^2 \left(\frac{\sigma_{E_\gamma^{(1)}}}{E_\gamma^{(1)}} + \frac{\sigma_{E_\gamma^{(2)}}}{E_\gamma^{(2)}} + \frac{2\sigma_\alpha}{\alpha} \right). \tag{7.5}$$

Here, $\sigma_{E_\gamma^{(1,2)}}$ denotes the uncertainties on the photon energy, identified as the energy resolution. σ_α corresponds to the position resolution. Photons from the decay of highly boosted pions have smaller angular separation in the detector frame of reference and higher energies. The same is true the other way around for weakly boosted pions. Looking at σ_{m^2} immediately leads to the consequence that the mass resolution is dominated by the energy resolution for small pion momenta and by the position resolution for large momenta.

This behavior is confirmed by plotting the pion mass resolution over a range of pion momenta for the different reconstruction combinations presented at the beginning of the section. In this analysis, only the basf2 reconstruction and MC information is used to determine the pion masses. In contrast to the following studies, no restrictions are imposed on the photons and the analysis represents the full spectrum of π^0 decays at Belle II. Figure 7.3 displays the resolution of the three reconstruction variants for $p_{\pi^0} \in [0.0, 4.0]$ Gev/c. Approximately at 1.7 GeV/c and above, the mass resolution is dominated by the position reconstruction. Therefore, it is expected that the energy resolution delivers the largest improvements to pions with relatively small momenta and has decreasing impact, especially in the second study of highly boosted pions.

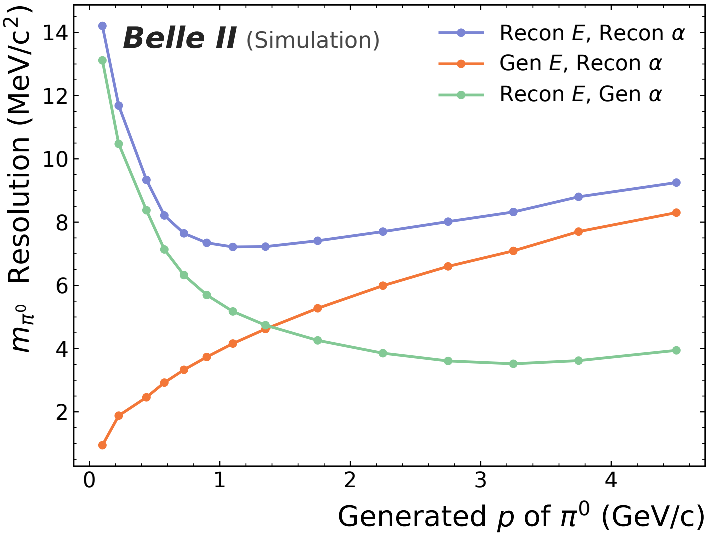


Figure 7.3.: Pion mass resolution in dependence of the generated pion momentum in a range of $p_{\pi^0} \in [0.1, 4.0]$. Three different reconstruction variants are compared: Reconstructed energy and reconstructed position, reconstructed energy and MC true position, and MC true energy and reconstructed position. The reconstruction is handled by basf2 with no additional restrictions imposed on the photons. Courtesy of Miho Wakai [27].

7.3. Weakly Boosted Neutral Pions

The single particle gun from section 3.1 is used to generate pions in a uniform momentum range of $p_{\pi^0} \in [0.2, 2.0]$ GeV/c. The studies take place in full detector coverage according to table 6.1 and generated pions of multiple events are distributed uniformly in space. The decaying pion leaves two isolated photon clusters in the ECL. Both clusters are identified and treated independently by creating two ROIs. Each of the two ROIs in the event then has to fulfill the criteria for one-cluster events stated in section 3.2.1. After the selection, 200 000 events are the foundation for the evaluation of the algorithms. There is no need for events for the training of GravNet, as only the already trained models from the toy studies are used.

The GravNet models from section 6.2 one by one evaluate the ROIs in an event and predict two photon energies. The photon energies are corrected by $c_E = 1.047$ for early background GravNet and events. The energy predictions of the nominal background GravNet are corrected by $c_E = 1.076$.

Early and nominal background events are studied together and compared. Section 7.3.1 gives an overview of the photon performance according to the metrics in section 5.3 and draws a connection to the performance in the corresponding toy studies. Subsequently, section 7.3.2 evaluates the performance of the pion mass reconstruction. A summary and conclusion are given for the weakly and highly boosted pions together in the final section 7.5.

7.3.1. Photon Performance

Photons in the toy study scenarios, in good approximation, fit the kinematics and characteristics of neutral pion decays at Belle II. The models were trained on the toy study events and are now brought to application in a scenario that is very alike but not identical. Before analyzing the pion reconstruction, it is crucial to review the photon performance on its own and identify potential deviations from the toy studies.

Fuzzy Clustering Agreement Index

Figure 7.4 shows the distribution of FCAIs for early and nominal background events. In both types of backgrounds, the algorithms achieve high FCAI scores, confirming non-random clustering results. GravNet remains the better clustering algorithm by a margin. However, basf2 catches up and the difference is significantly smaller than in the one-cluster toy study (see figure 6.6). GravNet once again shows the largest advancements with nominal background, which are probably due to the hyperparameter and feature optimization. Overall, the increments in the median indicate that on a clustering level the events of the weakly boosted pion studies are less complicated than the ones in the toy studies.

Sensitivity and Precision

Figure 7.5 plots the distributions of sensitivity and precision for early and nominal background events. Looking at the distribution with the experience from the toy studies, little to no improvements are expected to the photon resolution with early background. GravNet remains marginally more precise, but basf2 has a large lead in sensitivity. As acknowledged in the toy study section 6.2.2, the lack of sensitivity leads to a bias in the photon energy peak. This time, the bias is rectified through leakage correction though, and will not have a direct effect on the pion reconstruction. In nominal background, both algorithms take a hit in performance leaving results that are similar to the toy study. Here the discrepancy is much larger, predicting larger differences in the photon resolution as well.

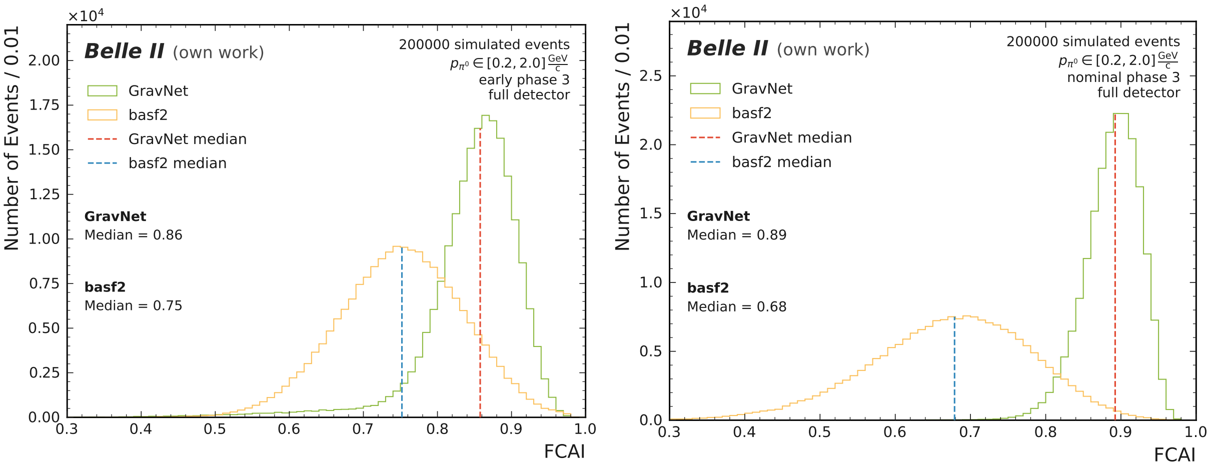


Figure 7.4.: Distribution of FCAIs for the weakly boosted pion physics studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

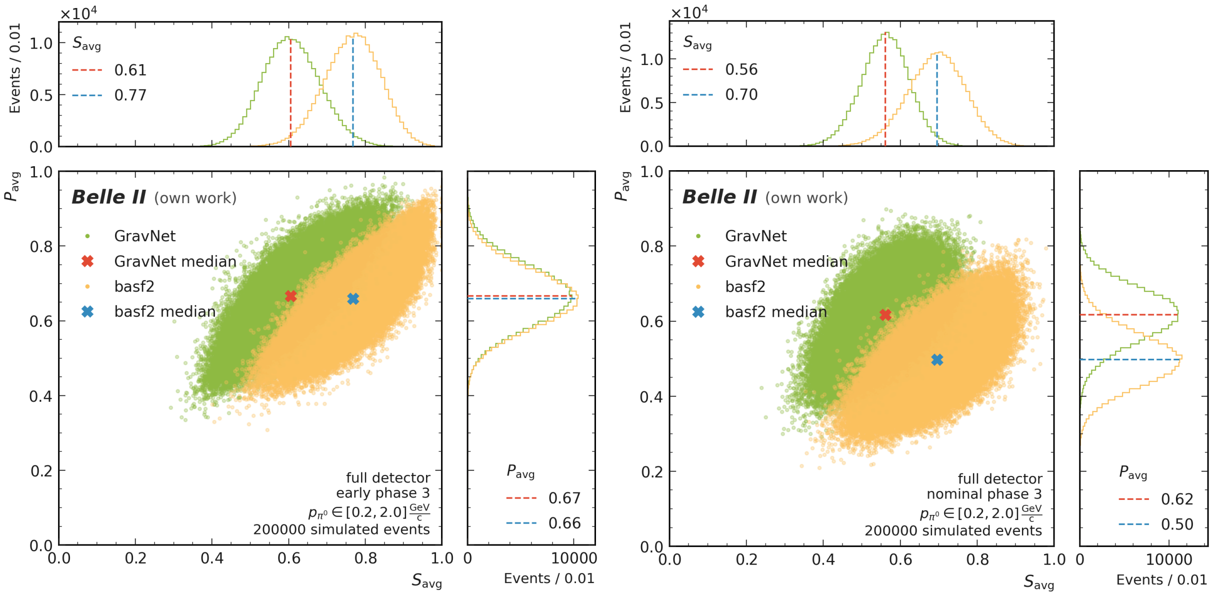


Figure 7.5.: Correlation and marginal distributions for average sensitivity S_{avg} and average precision P_{avg} for the weakly boosted pion physics studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated in the marginal distributions. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Energy Resolution

So far, in early background, GravNet loses some of its advancements over basf2 in comparison to the toy study. On one hand, from FCAI score and sensitivity and precision, a performance that is on par with the basf2 baseline is anticipated for early background events. On the other hand, the much higher precision in combination with the better clustering, promises an improvement to the photon resolution in nominal background. Figure 7.6 displays the photon resolution for early and nominal background. For early background, the improvement to the resolution is reduced from 5.2 % in the toy study to now 4.4 %. This result supports the theory that basf2 is able to adapt to the physics study better than GravNet. For nominal background, the improvements increased from 17.2 % to 22.2 %, giving a promising outlook for the pion reconstruction.

Note that GravNet is not leakage-corrected yet and despite the better resolution, the propagation of the offset in the peak to the pion reconstruction is undesired. Figure 7.7 presents the corrected photon resolution of GravNet, revealing the energies that are ultimately used in further pion reconstruction. The corrected FWHMs throughout all studies virtually shift the peak so that it is centered around zero (see section 5.3.2). This is equivalent to a leakage correction based on this exact scenario, however, the correction is deliberately based on the universal toy studies. With regard to this fact, the results of the crude correction are excellent and the true shift of the distribution only costs GravNet performance within the uncertainties.

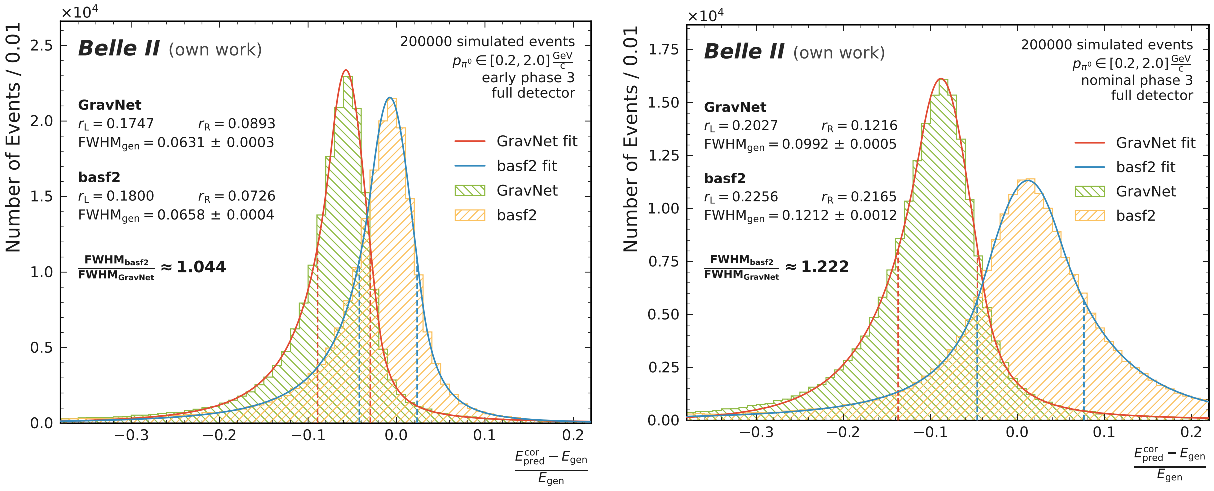


Figure 7.6.: Fits for the distributions in generated errors η_{gen} for the weakly boosted pion physics studies. Both plots compare GravNet to the basf2 baseline in full detector coverage. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

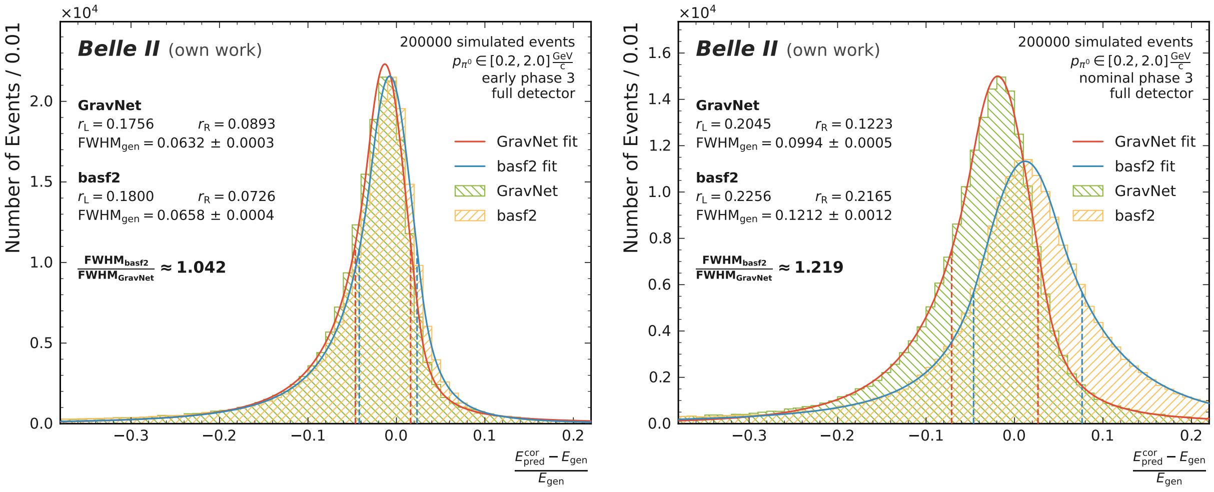


Figure 7.7.: Fits for the distributions in generated errors η_{gen} for the weakly boosted pion physics studies. Both plots compare GravNet to the basf2 baseline in full detector coverage. The GravNet prediction is corrected for leakage with a correction factor using the generic toy study results in section 6.2. The left side depicts early phase 3 events corrected with $c_E = 1.047$, and the right side nominal phase 3 events with $c_E = 1.076$.

7.3.2. Pion Performance

The photon resolution improvements are marginal for early background but extensive for nominal background. Nevertheless, the momentum range in this study suggests that even small improvements have a high impact on the pion mass resolution. The first part of the pion performance evaluation directly assesses the mass resolution. The three variants of reconstruction introduced in section 7.2 are compared to see the current status of the reconstruction, potential resolution with perfect position reconstruction, and potential resolution with perfect energy reconstruction. The second part then focuses on the dependence of the mass resolution on the pion momentum.

Mass Resolution

Figure 7.8 displays the three versions of pion reconstruction for both early and nominal background. Starting with early background, the 4.2 % improvement to the photon resolution does not carry over equally to the pion mass resolution. 0.6 % improvement are within the uncertainties of the fit and not considered significant in this context. Looking at the potential resolution with perfect position reconstruction in figure 7.8c reveals the same relation and confirms that GravNet is not held back by the position reconstruction. Comparing the FWHMs of the fully reconstruction-based pion mass with the MC energy reconstruction in figure 7.8e accentuates that there is still a lot of potential with perfect reconstruction. The MC energy reconstruction yields $\mathrm{FWHM} \approx 18\,\mathrm{MeV/c^2}$ in comparison to the $\mathrm{FWHM} \approx 31\,\mathrm{MeV/c^2}$ for both algorithms.

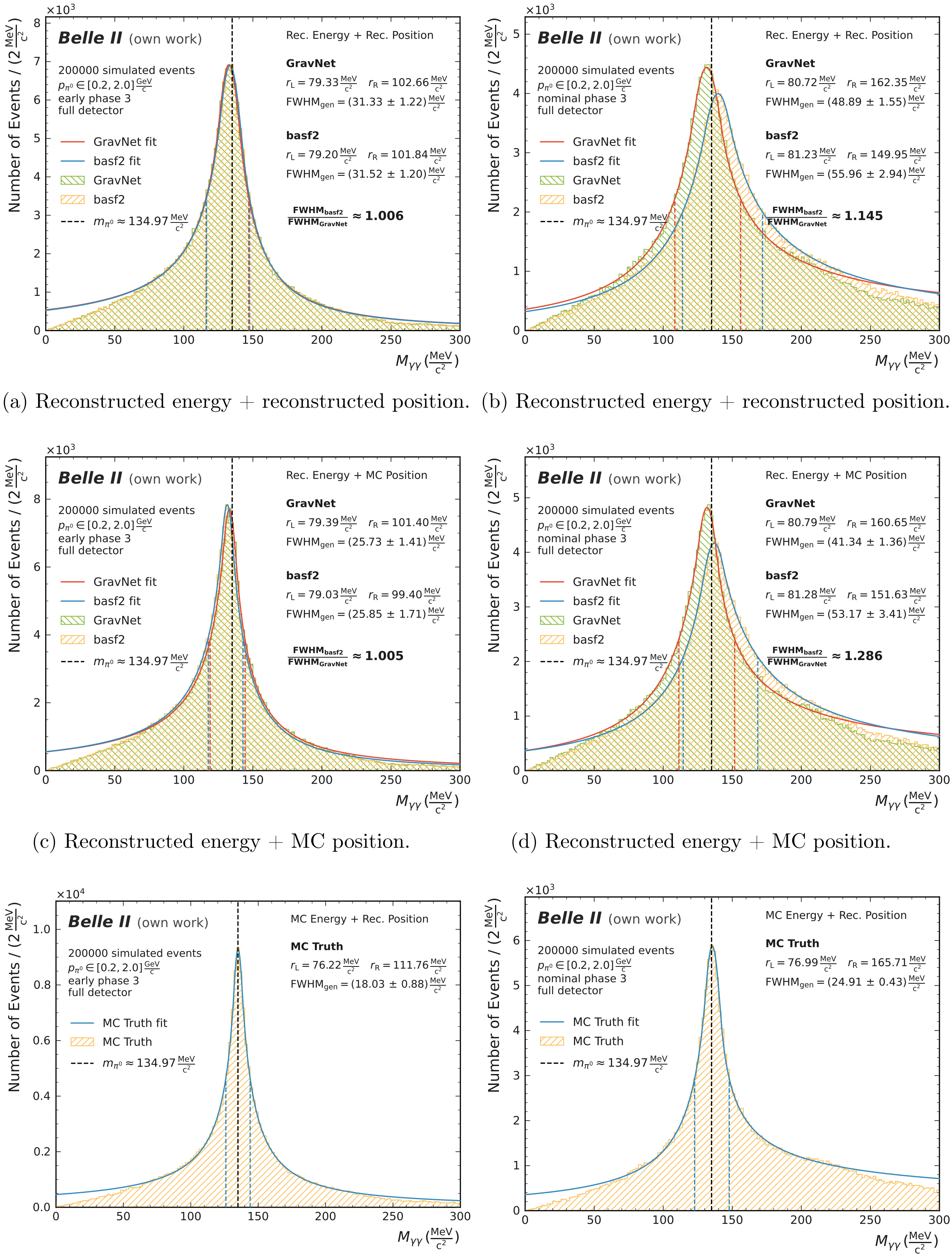
As seen in all toy studies and in the analysis of the photon resolution, GravNet performs to its full potential in nominal background. Figure 7.8b illustrates that even with the current position reconstruction, the improvements to the pion mass resolution amount to 14.5 %. While basf2 does not increase in resolution by much when going from reconstructed to MC position in figure 7.8d, GravNet demonstrates great potential by increasing the relative improvements to 28.5 %. This jump in performance is expected in the low momentum regime of this study and highlights that in this scenario the performance of GravNet is limited by the position reconstruction. Nevertheless, there are also a lot of improvements to be found in energy reconstruction. Figure 7.8f states the theoretically possible FWHM with the current position reconstruction at $\mathrm{FWHM} \approx 35\,\mathrm{MeV/c^2}$. This leaves plenty of room for improvements solely based on the energy reconstruction.

Both types of backgrounds have in common that the pion mass resolution is dominated by the energy resolution. This results in approximately 90 % of unrealized potential relative to the MC energy reconstruction for both types of backgrounds. GravNet unveils negligible improvement to the pion mass resolution in early background. The advancements in nominal background are far better at approximately 15 % and promise even larger differences to the basf2 baseline with better position reconstruction.

Momentum Dependence

Analogous to the energy dependence analysis in the one-cluster toy study in section 6.2.2, pions are generated at various fixed momenta. Per fixed momentum in a range from 0.2 GeV/c up to 2.0 GeV/c, 20 000 events are created for the evaluation. This range of momenta already includes decays that result in photons with energies $< 0.1\,\mathrm{GeV}$ and $> 1.5\,\mathrm{GeV}$. These are outside the range of energies on which GravNet is trained, therefore especially upper and lower bounds of this interval test the ability of GravNet to generalize to unknown scenarios.

Figure 7.9 displays the dependence of the resulting $\mathrm{FWHM}_{\mathrm{gen}}$ on the generated pion momenta. The plots for both types of backgrounds share a similar central dip at 1.1GeV/c. This dip is also found in the study in figure 7.3 and a result of the energy and position dependence of the pion reconstruction. Over the whole range, the resolution is approximately half as good with nominal



(a) Reconstructed energy + reconstructed position.

(b) Reconstructed energy + reconstructed position.

(c) Reconstructed energy + MC position.

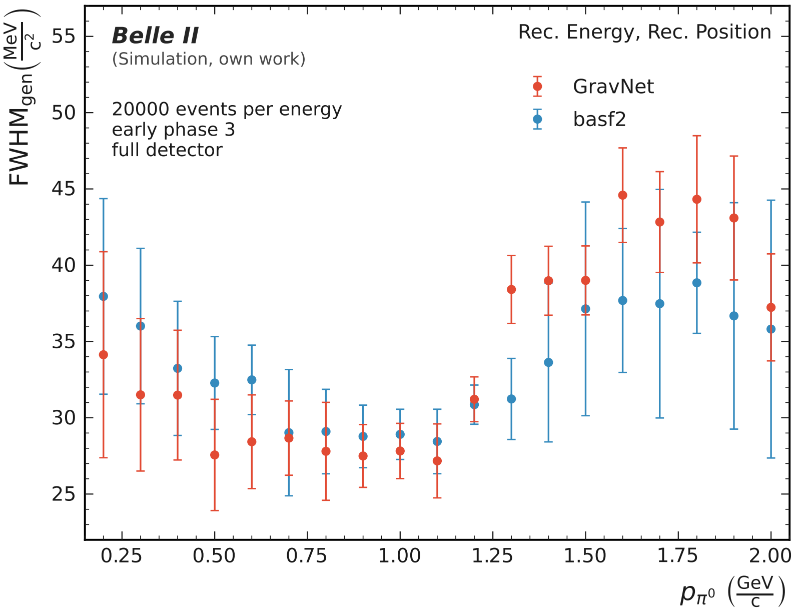
(d) Reconstructed energy + MC position.

(e) MC energy + reconstructed position.

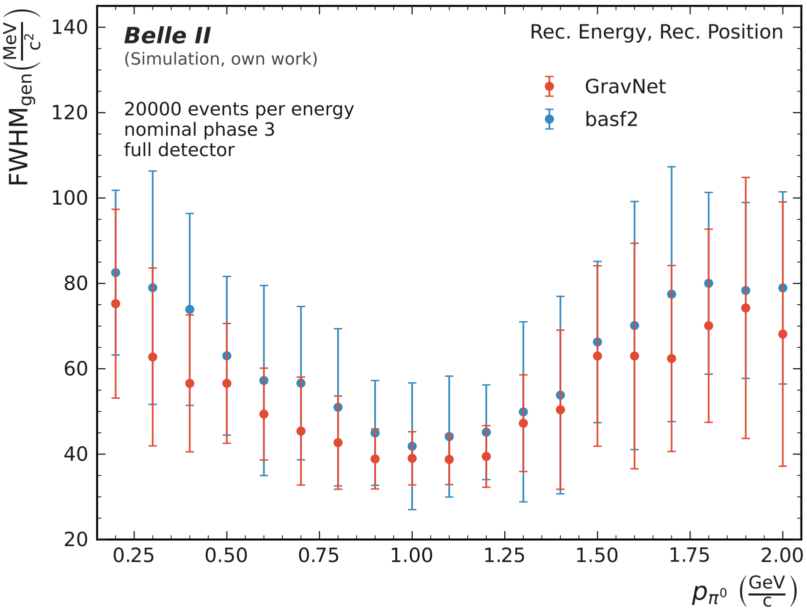
(f) MC energy + reconstructed position.

Figure 7.8.: Fits for the distributions in invariant masses of the two-photon system $M_{\gamma\gamma}$ for the weakly boosted pion studies. Each plot depicts different combinations of energy and positions used in the pion mass reconstruction. The plots on the left side depict early phase 3 events, and the plots on the right side nominal phase 3 events.

background as with early background. Even though GravNet outperforms basf2 in the low momenta regime for both types of backgrounds, the resolution is significantly worsening with higher momenta for early background events. Part of the reason for that peculiarity is likely that GravNet is not fully specialized to the high energy photons occurring with these momenta. In nominal background, GravNet manages to keep an advantage at all momenta. However, the trends should be treated with care in light of the pronounced uncertainties on the FWHM.



(a) Early phase 3 background.



(b) Nominal phase 3 background.

Figure 7.9.: Shown are the resolutions FWHM_{gen} on the invariant mass of the two-photon system, in dependence of the generated pion momentum p_{π^0} for the weakly boosted pion studies. Both plots compare GravNet and basf2 in full detector coverage. Each data point marks the FWHM_{gen} of 20 000 events at a fixed momentum $p_{\pi^0} \in [0.2, 2.0]\ \frac{\text{GeV}}{\text{c}}$.

7.4. Highly Boosted Neutral Pions

The single particle gun from section 3.1 is used to generate pions in a uniform momentum range of $p_{\pi^0} \in [2.0, 2.7]\,\mathrm{GeV/c}$. The studies take place in full detector coverage according to table 6.1. Once again, generated pions of multiple events are evenly distributed in the detector. Two overlapping clusters are the result of the decay of the highly boosted pion. The ROI of the event is identified and has to fulfill the criteria for two overlapping clusters stated in section 3.2.2. 200 000 events are used for testing with no additional training of the existing GravNet models.

The evaluation is based on the two GravNet models from the two-cluster toy studies in section 6.3 with early and nominal background. In contrast to the weakly boosted pion study, just one ROI with the two clusters is inferred. The two predicted photon energies are corrected by $c_E = 1.048$ for early background GravNet and events. The predictions of the nominal background GravNet are corrected by $c_E = 1.071$.

Section 7.4.1 first gives an overview of the photon performance according to the metrics in section 5.3. The comprehensive evaluation of the pion performance follows in section 7.4.2. The final summary and conclusions to the studies are found together with the weakly boosted pion results in section 7.5.

7.4.1. Photon Performance

Deviations between the toy studies and the upcoming physics studies are first analyzed by considering only the photon performance. The metrics used in this section are the two-cluster metrics already known from the evaluation of the corresponding toy studies. As usual, the focus is put on the clustering quality and the photon energy resolution.

Fuzzy Clustering Agreement Index

The distributions in FCAI are shown in figure 7.10 for early and nominal background. As with the weakly boosted pion studies, the overall increased FCAI scores propose that the events are slightly less complicated in the physics studies than in the corresponding toy studies. Nevertheless, it has to be seen how these excellent FCAI scores carry over to photon energy resolution and finally the pion mass resolution.

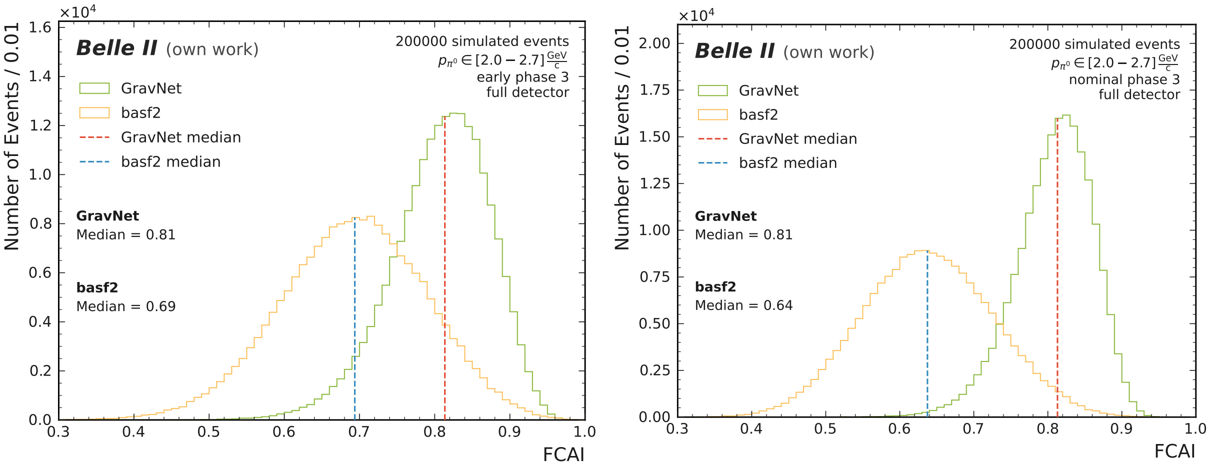


Figure 7.10.: Distribution of FCAIs for the highly boosted pion physics studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Sensitivity and Precision

Sensitivity and precision in figure 7.11 depict practically indistinguishable behavior in comparison to the toy studies for both types of backgrounds. Both algorithms have marginally increased medians in S_{avg} and P_{avg}, hinting towards the events being more simple in the toy studies. This is in line with the suggestion from the FCAI analysis.

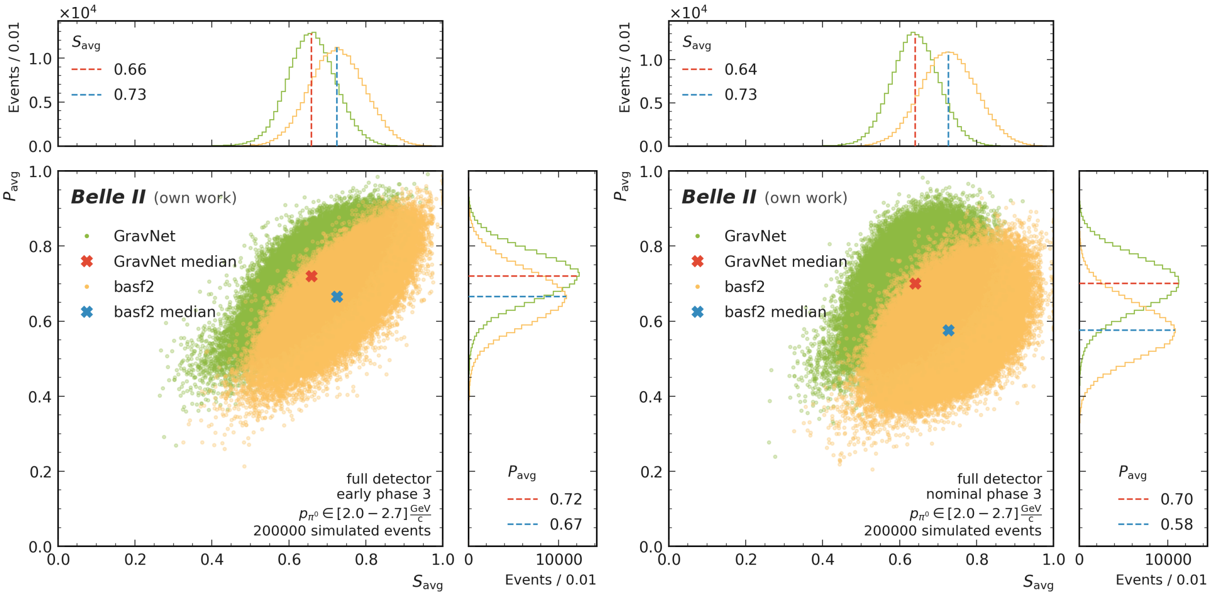


Figure 7.11.: Correlation and marginal distributions for average sensitivity S_{avg} and average precision P_{avg} for the highly boosted pion physics studies. Both plots compare GravNet and basf2 baseline in full detector coverage. The medians of the distributions are marked and indicated in the marginal distributions. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

Energy Resolution

According to the metrics so far, the kinematics between the two-cluster toy studies and the highly boosted pion studies are more alike than was the case for weakly boosted pions. Thus, for the photon energy resolution, the same improvements are presumed for toy studies and physics studies. Figure 7.12 displays the uncorrected photon energy resolution, figure 7.13 the corrected version of GravNet. The two-cluster toy study for early background presents 3.2 % improvement in photon resolution over the basf2 baseline. In the physics study, 1.4 % improvement remains, which is still true after leakage correction but nevertheless a larger loss in performance than expected. This emphasizes that the clustering results are not necessarily a good indicator of the photon energy resolution. An improvement of 20 % in comparison to the 14.5 % in the toy study is found for nominal background. 19.8 % are left after the leakage correction of GravNet, forecasting much potential for the pion mass reconstruction.

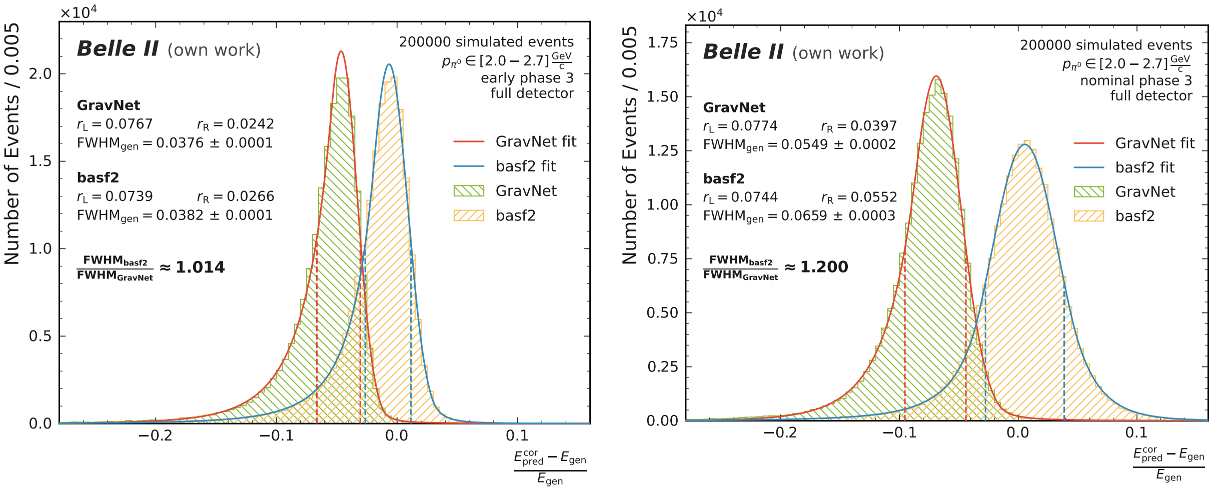


Figure 7.12.: Fits for the distributions in generated errors η_{gen} for the highly boosted pion physics studies. Both plots compare GravNet to the basf2 baseline in full detector coverage. The left side depicts early phase 3 events, and the right side nominal phase 3 events.

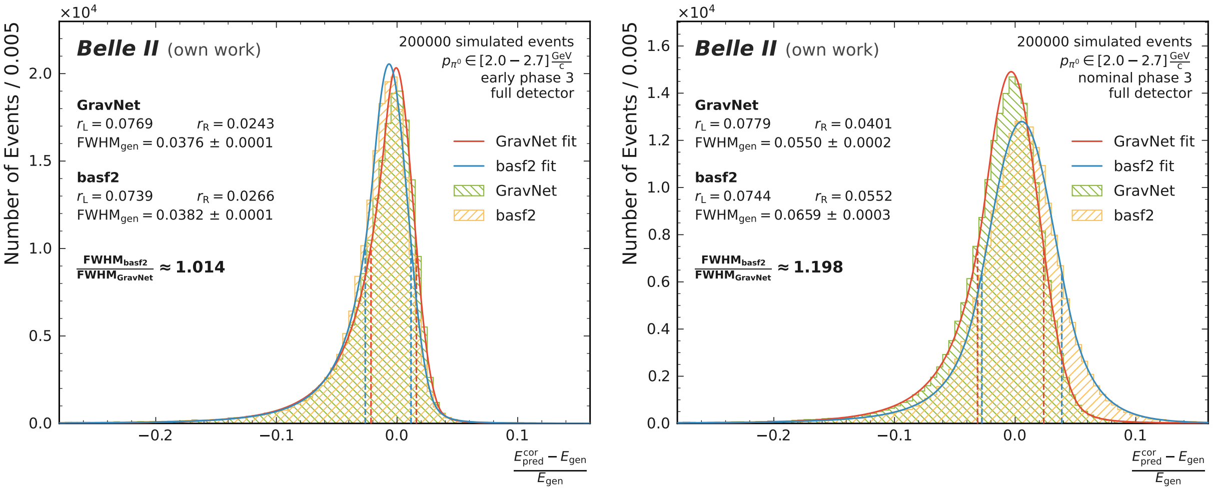


Figure 7.13.: Fits for the distributions in generated errors η_{gen} for the highly boosted pion physics studies. Both plots compare GravNet to the basf2 baseline in full detector coverage. The GravNet prediction is corrected for leakage with a correction factor using the generic toy study results in section 6.2. The left side depicts early phase 3 events corrected with $c_E = 1.048$, and the right side nominal phase 3 events corrected with $c_E = 1.071$.

7.4.2. Pion Performance

The photon resolution improvements are large for the highly boosted pion study with nominal background. However, now the momentum range suggests that even large improvements only have a diminishing impact on the pion mass resolution as seen in figure 7.3. For the same reason, it is likely that few to none of the already small improvements in early background are seen in the pion mass resolution. After looking at the mass resolution for the different reconstruction variants, the momentum dependence of the pion mass resolution is studied in the second part of the section.

Mass Resolution

This section compares the three variants of reconstruction introduced in section 7.2 in order to analyze the current status of the reconstruction, the resolution with perfect position reconstruction, and the resolution with perfect energy reconstruction. Figure 7.14 displays the resulting resolutions for all variants with early and nominal background events. The already small improvements of GravNet to the photon resolution in early background vanish completely in the pion mass reconstruction. In the scenario with current energy and position reconstruction, basf2 outperforms GravNet within the uncertainties. Figure 7.14c indicates an opposite trend, leading to the conclusion that the algorithms perform equally well (within the uncertainties). Regardless, the MC energy reconstruction in figure 7.14e points out that both algorithms achieve results close to the best possible resolution given the limits of the position reconstruction. That the position reconstruction limits the pion mass resolution in this momentum regime also stands out in the comparison between MC energy and MC position reconstruction.

Identical trends are found for the nominal background events. In general, the absolute resolutions for early and nominal backgrounds do not differ by much. Even in this more challenging scenario, both algorithms perform close to the theoretical limit stated in figure 7.14f. Nonetheless, GravNet claims a significant improvement of 2.1 % in the current scenario and suggests an improvement of up to 8.6 % given a perfect position reconstruction.

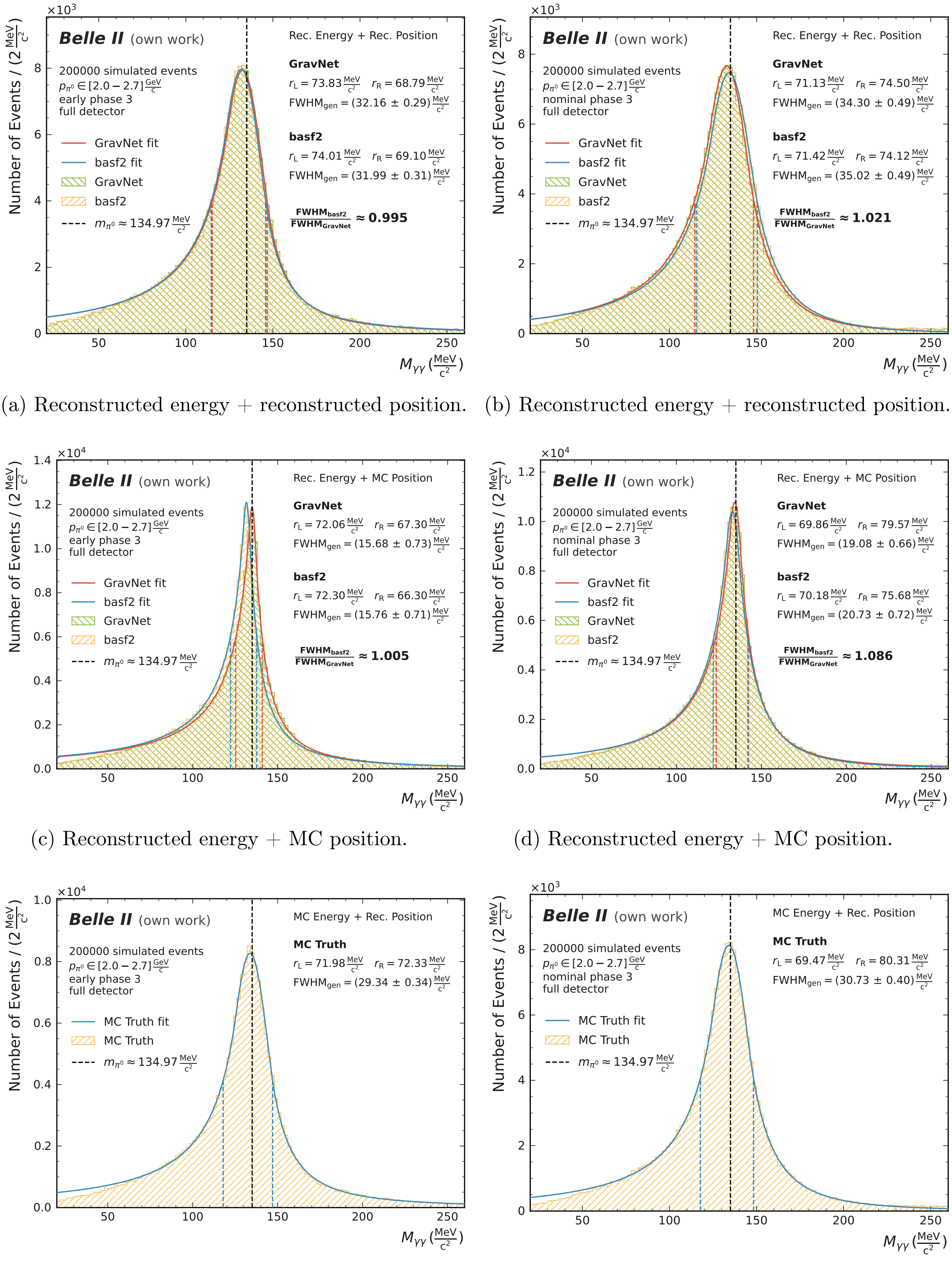
Momentum Dependence

Once again, pions are generated at various fixed momenta to examine the momentum dependence. This time the momenta range from 2.0 GeV/c up to 2.7 GeV/c with 20 000 events per energy. Figure 7.15 displays the dependence for both types of backgrounds. For early background events, GravNet and basf2 perform practically indistinguishable. The largest improvements are found at lower momenta for nominal background. Both early and nominal backgrounds display a trend towards lower resolution with higher momenta, which is also seen in the theoretical analysis in figure 7.3.

7.5. Overview and Conclusions

The physics studies in the two previous sections evaluate the effects of the GravNet energy reconstruction on the pion mass resolution by the means of photon energy resolution improvements. In early background, the improved photon resolution has a vanishing effect on the pion mass resolution. Even for low pion momenta where small improvements have a high impact, GravNet is not improving the energy resolution enough to achieve meaningful changes. This is likely traced back to the specialized GravNet having difficulty adapting to the new kinematics. The studies of nominal background events draw a different picture. The versatility of GravNet manifests in large improvements to the photon resolutions that equally amount to 20 % for weakly and highly boosted pions. In comparison to the basf2 baseline, GravNet improves the pion mass resolution by 14.5 % in the decay of weakly boosted pions and by 2.1 % for highly boosted pions. GravNet manages to carry over far more improvements associated with low pion momenta that allow for a large influence of the energy resolution in the reconstruction. However, in both cases GravNet loses some of its potential to the position reconstruction, leaving room for further improvements.

On one hand, in early background events the momentum dependence analysis reveals small improvements to the mass resolution only for low momenta. These do not have a relevant impact in the larger data set over the full range of momenta. On the other hand, the significant advancements for all pion momenta in nominal background leave a promising outlook for the application in future background scenarios, especially with further optimization towards the different event kinematics.



(a) Reconstructed energy + reconstructed position.

(b) Reconstructed energy + reconstructed position.

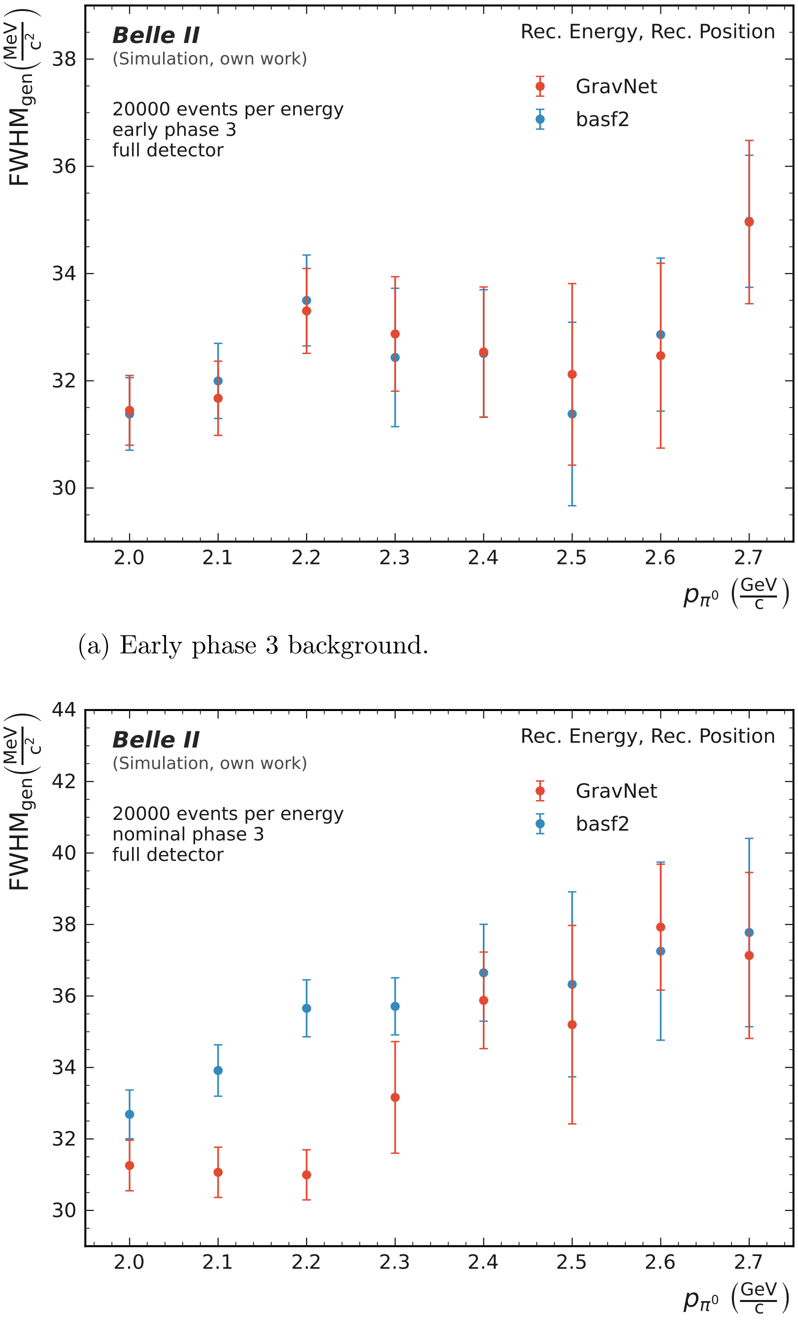
(c) Reconstructed energy + MC position.

(d) Reconstructed energy + MC position.

(e) MC energy + reconstructed position.

(f) MC energy + reconstructed position.

Figure 7.14.: Fits for the distributions in invariant masses of the two-photon system $M_{\gamma\gamma}$ for the highly boosted pion studies. Each plot depicts different combinations of energy and positions used in the pion mass reconstruction. The plots on the left side depict early phase 3 events, and the plots on the right side nominal phase 3 events.



(a) Early phase 3 background.

(b) Nominal phase 3 background.

Figure 7.15.: Shown are the resolutions $\mathrm{FWHM_{gen}}$ on the invariant mass of the two-photon system, in dependence of the generated pion momentum p_{π^0} for the highly boosted pion studies. Both plots compare GravNet and basf2 in full detector coverage. Each data point marks the $\mathrm{FWHM_{gen}}$ of 20 000 events at a fixed momentum $p_{\pi^0} \in [2.0, 2.7]\,\frac{\mathrm{GeV}}{\mathrm{c}}$.

8. Outlook

This chapter summarizes options for further expansion and exploration of the applications of the GravNet algorithm, but also existing issues that need additional investigation.

Metrics

The metrics presented in chapter 5 lay the foundation for the comparison of a wide range of characteristics in the events and of the performances of the algorithms. The toy studies in chapter 6 investigate correlations of metrics and underlying mechanics of the clustering algorithms. Additional analyses can be performed using these metrics: First to find further attributes that make events challenging for the algorithms. Second, to find which aspects of the clustering are dominating the contribution to the energy resolution. Accordingly, the dependence of the energy resolution with respect to different metrics can be studied in more detail.

Additionally, metrics can be fine-tuned to quantify the properties they represent more accurately. This concerns mostly metrics that depend on a certain threshold like the cluster radius or the tail length, but also the definition of properties like the cluster center.

Machine Learning

Chapter 4 discusses the numerical behavior of GravNet to assign small fractions to all available classes in each crystal. This behavior can for example be fixed by a mechanism that corrects the outputs. One possibility for such correction is to set thresholds on a full assignment or removal of crystals of a class. The latter would push GravNet more in the direction of hard clustering. However, further studies are needed to assess the necessity of such changes and to evaluate whether an actual increase in performance is achieved.

GravNet is not the only feasible machine learning approach to energy reconstruction in the Belle II electromagnetic calorimeter. The capability of GravNet to adapt to arbitrary geometries and input sizes, thereby being a universal algorithm and saving computing resources, stands out. Nevertheless, fully connected neural networks and convolutional neural networks theoretically also have many machine learning advantages and deserve a discussion on their own.

Another aspect for potential investigation is the particular GravNet architecture. On one hand, there are several other options opposing the GravNet layer to realize graph neural networks and utilize message passing. On the other hand, the existing GravNet structure can be optimized or extended using hyperparameters.

Oftentimes input features play a large role in the performance of machine learning algorithms. The input features for GravNet can be studied further and optimized thoroughly. An example is an in-depth study of their influence on the performance in different scenarios and a corresponding addition or removal of features.

Regardless of the architecture, the training process can be subject to further optimization. GravNet is trained and tested on a broad range of kinematics. One option is to train several models of GravNet to specific scenarios. Examples are the training on smaller energy intervals or on the particular kinematics of neutral pion decays. Looking at the application on real detector data, various scenarios with distinct kinematics occur which are mixed and not known beforehand. The application to unknown kinematics requires a mechanism that either decides which specific model to use or which model provides the best prediction. Also in light of an application on real detector data, the distribution of kinematics in the training should be chosen to represent the true distribution of kinematics occurring at Belle II.

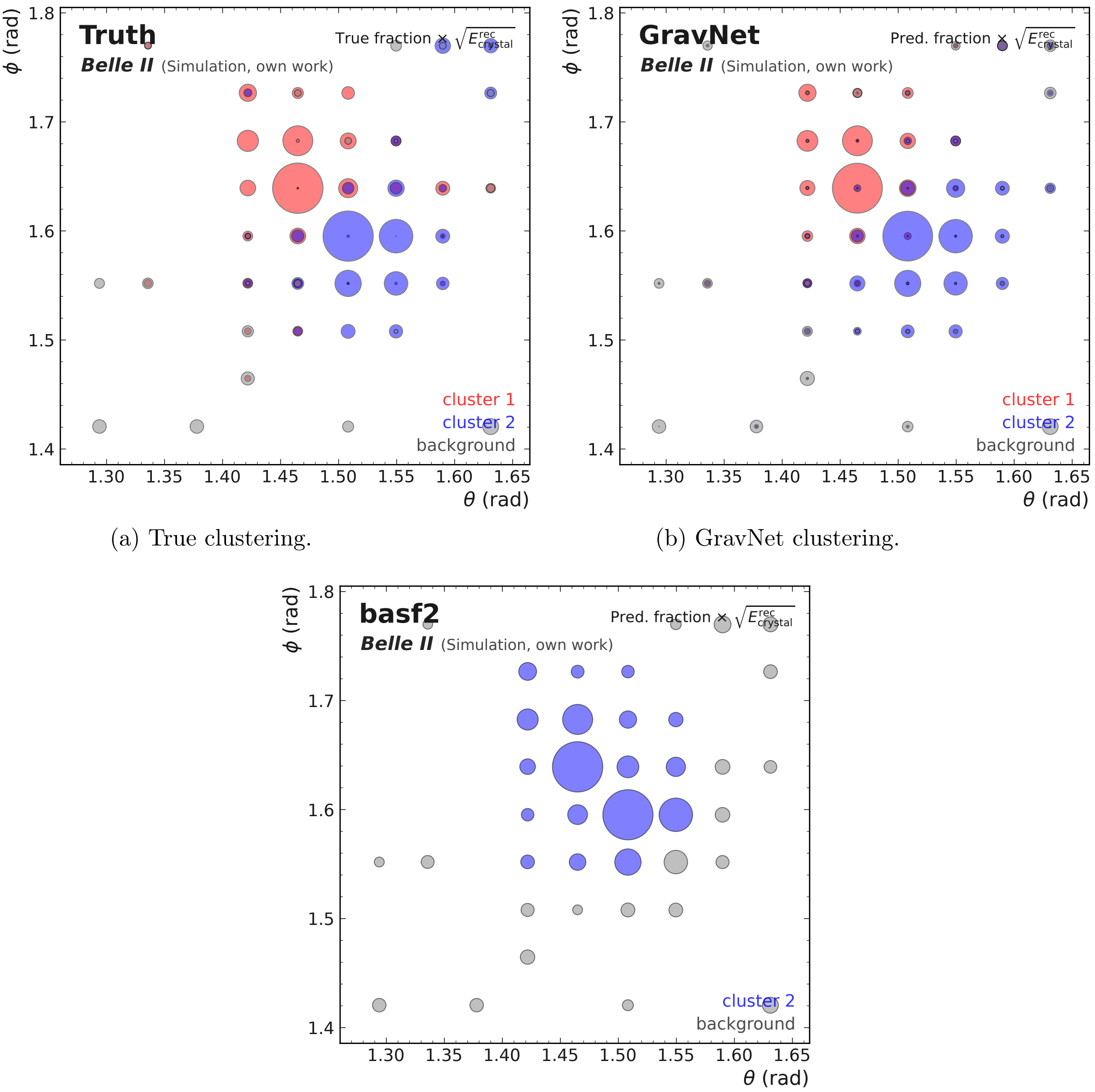
Physics Applications

In comparison to other machine learning approaches, the small number of computations and parameters in the GravNet algorithm, as well as the small and variable input size, make inference fast and memory-saving. In combination with an optimized implementation on specialized hardware like field-programmable gate arrays, the fast inference can open up real-time applications. An example is the usage of GravNet to generate trigger signals, which in addition does not necessarily require equally high energy resolution as in the studies in this work. Therefore, the architecture can possibly be reduced to achieve a further speed-up.

The deployment of the GravNet algorithm for real detector data applications opens up a number of challenges. First of all, a dedicated algorithm with information from the full detector data needs to determine a specific region of interest for the GravNet model to evaluate. Only if the preceding algorithm is able to select relevant regions of interest is it possible to take advantage of the energy resolution of GravNet. Another logical extension is the addition of an advanced leakage correction. The basic leakage correction used for the pion mass reconstruction in chapter 7 demonstrates the benefits of even crude improvements to the predicted energies. A more profound leakage correction for example with energy-dependent corrections is likely to substantially improve the overall performance of GravNet. In order to achieve the generalization necessary for the cluster signatures (or shower shapes) that appear in real detector data, a less stringent event selection than the one described in chapter 3 is needed. After the first validation of GravNet provided by this work, the selection does not have to consider comparability to the Belle II Analysis Software Framework baseline anymore. Even so, potentially new cluster signatures could lead to issues or even complete failure of the existing models and require retraining. It would be necessary to reevaluate the performances for specific energies as well as different detector regions as the selection criteria might have varying effects.

The application to a broader range of reconstruction tasks in the Belle II electromagnetic calorimeter is another compelling direction for future work. First, there is the reconstruction of other types of particle showers in the calorimeter. On one hand, the adaption to electromagnetic showers originating from electrons should not need many changes, and a performance similar to that for photon clusters is expected. On the other hand, hadronic showers result in dissimilar clusters to the ones studied in this work and the reconstruction needs not only retraining but a distinct study of the performance. However, the outlook of that type of study is especially promising in a combination of electromagnetic and hadronic clusters, where GravNet could make full use of its additional input features.

Second, the reconstruction of photon clusters could be extended. The present work shows that GravNet has great potential in accurately depicting the true clustering of an event. This includes edge cases like disjoint clusters or excessive overlap. Two edge cases that are explicitly excluded in this work and in which basf2 fails per definition are: Two local maxima that originate from just one particle and one local maximum that consists of two particles. Figure 8.1 shows the event display of an event with just one local maximum but associated with two particles. GravNet is theoretically independent of local maxima and therefore able to identify and label such events just as well, whereas basf2 does not recognize the second cluster. For this type of application, a decision mechanism has to be employed in order to decide beforehand or afterward which prediction fits the cluster(s) better.



(a) True clustering.

(b) GravNet clustering.

(c) Basf2 clustering.

Figure 8.1.: Event with a cluster with only one local maximum that originates from two particles. θ and ϕ are the detector coordinates. The recorded energy is scaled with $\sqrt{E_{\text{rec}}}$.

9. Summary

This work introduces and studies a machine learning approach based on graph neural networks for the clustering of energy depositions in the Belle II electromagnetic calorimeter. Chapter 4 proposes the GravNet architecture which consists of stacked GravNet layers that utilize message passing between nodes in an end-to-end learned representation space of the calorimeter. This concept comes with two main advantages in comparison to other machine learning methods: First, the algorithm does not rely on a regular arrangement of calorimeter crystals and therefore readily handles irregular calorimeter geometries like the Belle II endcaps. Second, contrary to convolutional and fully connected neural networks, not only the input size but also the number of computations are minimal and allow for a resource-saving implementation. In addition, the algorithm enables the use of virtually any input features that greatly enhance the performance depending on the scenario.

The Belle II Analysis Software Framework is the currently used reconstruction framework that serves as the baseline for comparison in all studies. Chapter 5 introduces various metrics for the evaluation and comparison of the performances from distinct perspectives. The toy studies in chapter 6 investigate the basic functioning and behavior of GravNet by evaluating the reconstruction of single and two overlapping photon clusters. Overall, GravNet depicts the underlying clustering more accurately than the baseline. This especially applies to edge cases like the reconstruction of disjoint clusters or clusters with vast overlap, where the topological approach of the Belle II Analysis Software Framework reaches its limits. While this does not necessarily transfer to great energy resolution due to detector hardware effects, it is especially useful in future scenarios with fewer hardware restrictions. Nevertheless, when it comes to the photon energy resolution, GravNet significantly outperforms the baseline over a large range of photon energies from 0.01GeV to 3.0 GeV. This applies to both, the current early phase 3 background, and the future nominal phase 3 background conditions. The improvements to the photon energy resolution range from approximately 5 % in early phase 3 background events, up to 20 % for low energy photons with high levels of background in nominal phase 3 events.

The physics studies in chapter 7 apply the GravNet algorithm to the reconstruction of the neutral pion mass from the invariant mass of a two-photon system. Of the large advancements to the photon energy resolution, no significant improvements carry over to the neutral pion mass resolution in early phase 3 background. Considering nominal phase 3 background conditions, the reconstruction of both weakly and highly boosted neutral pions yields substantial improvements of up to 15 %. As a matter of fact, the pion mass reconstruction is limited by the current position reconstruction for the photons and the improvements of GravNet suggest great potential in light of better position resolution.

All studies are considered the first validation of GravNet and use Monte Carlo generated and simulated data including Monte Carlo truth information. A lot of work is required in order to bring GravNet to application on real data: First, the algorithm has to learn more universal cluster signatures like the ones originating from charged electromagnetic and hadronic showers. Second, the algorithm has to be extended by an elaborate leakage correction to account for detector hardware effects. However, the application on real data is also expected to deliver considerable benefits as GravNet could make full use of its extended set of input features to distinguish many different cluster signatures.

Overall, GravNet proves to be a viable and versatile algorithm for the clustering of energy depositions, leaving a promising outlook for a broad range of present and future applications.

Appendix

A. GravNet Model Plots

A.1. Loss Progression

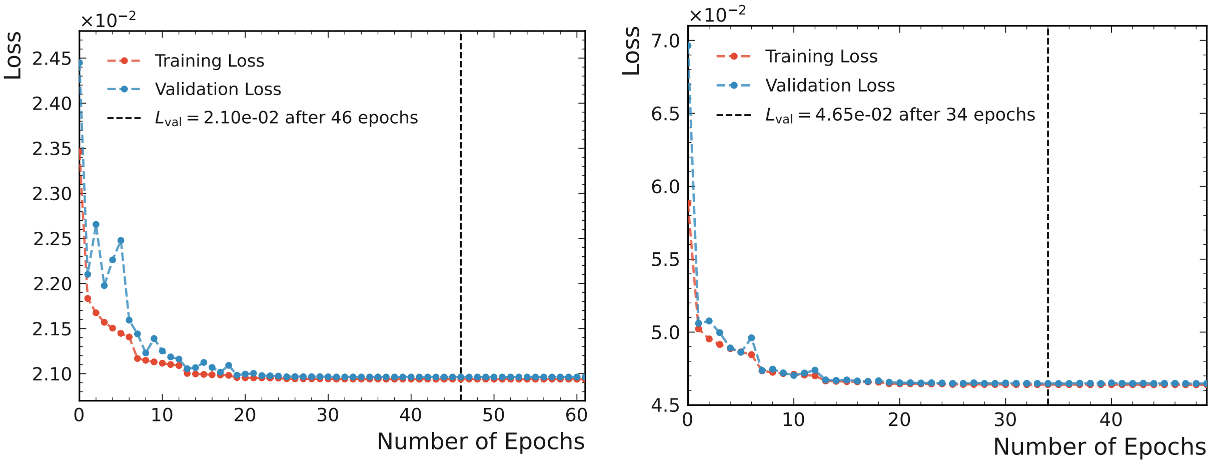
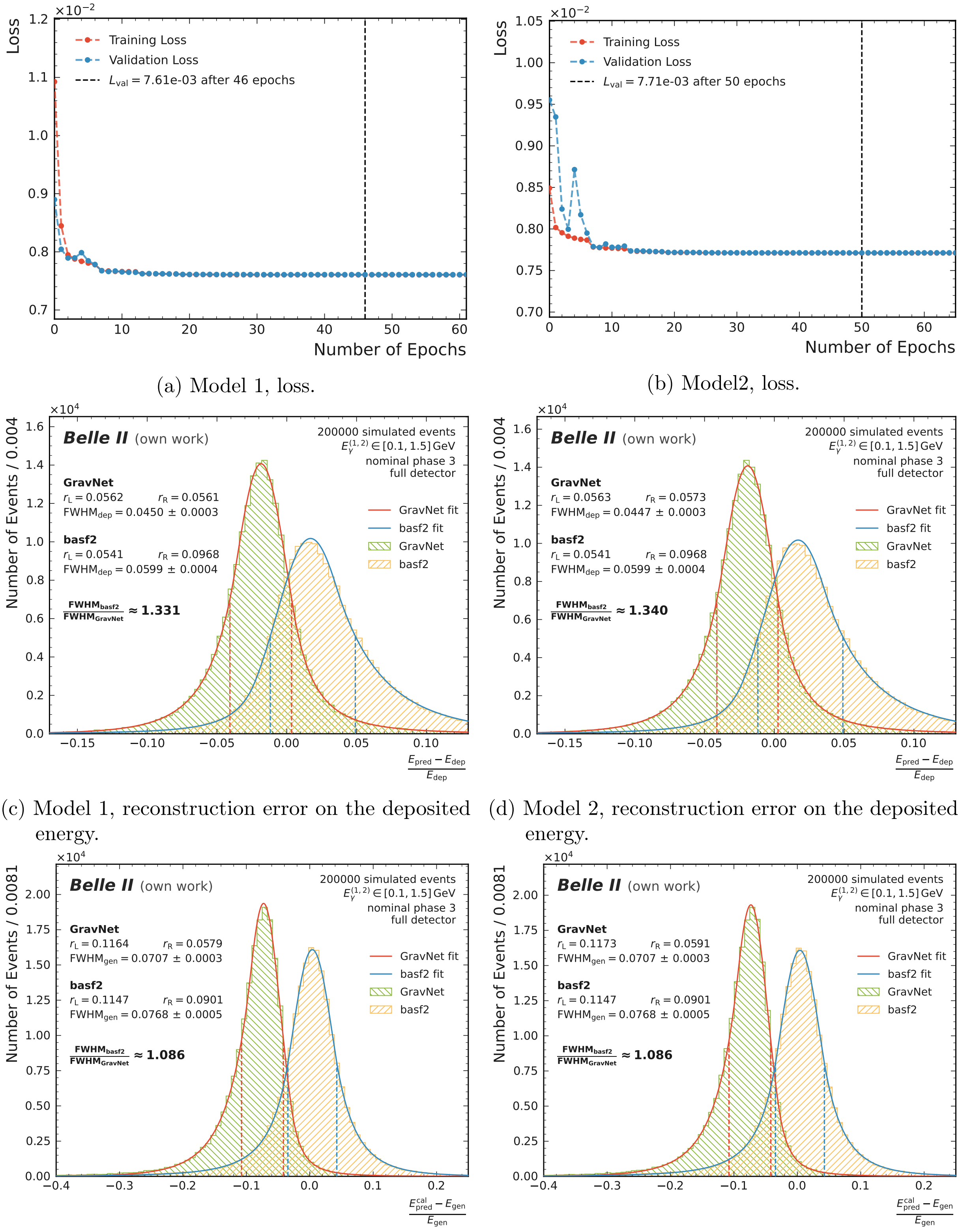


Figure .1.: Losses with conditional stopping for the early phase 3 background versions of GravNet. The left plot depicts the one-cluster GravNet, the right plot the two overlapping cluster GravNet.

A.2. Feature Analysis



(a) Model 1, loss.

(b) Model2, loss.

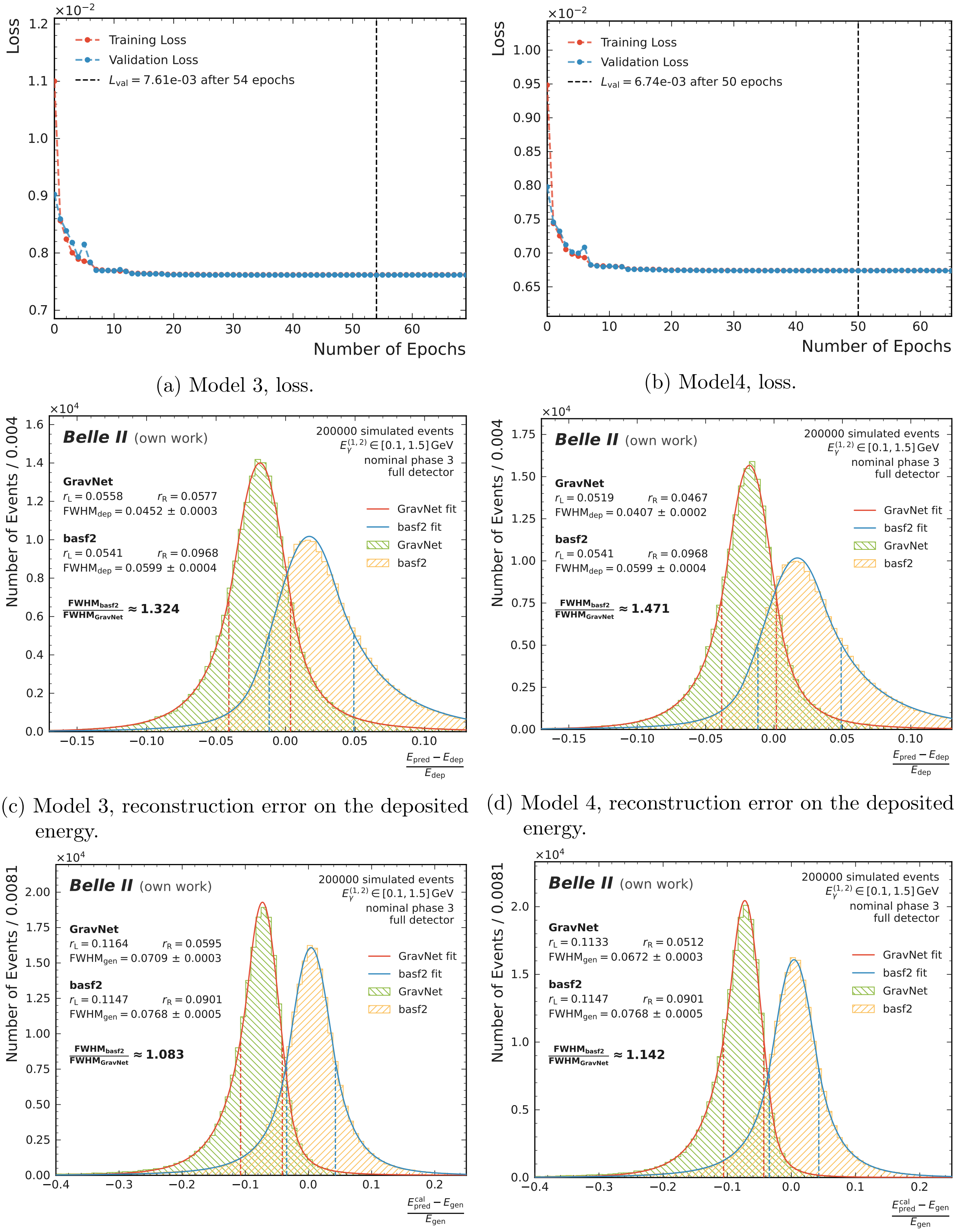
(c) Model 1, reconstruction error on the deposited energy.

(d) Model 2, reconstruction error on the deposited energy.

(e) Model 1, reconstruction error on the generated energy.

(f) Model 2, reconstruction error on the generated energy.

Figure .2.: Loss progression and fits for the resolutions on the deposited and the generated photon energy of two GravNet models with different features. The plots to the left show model 1 with global coordinates, local maxima, and recorded energy. The plots to the right show model 2 with local coordinates, local maxima, and recorded energy.



(a) Model 3, loss.

(b) Model4, loss.

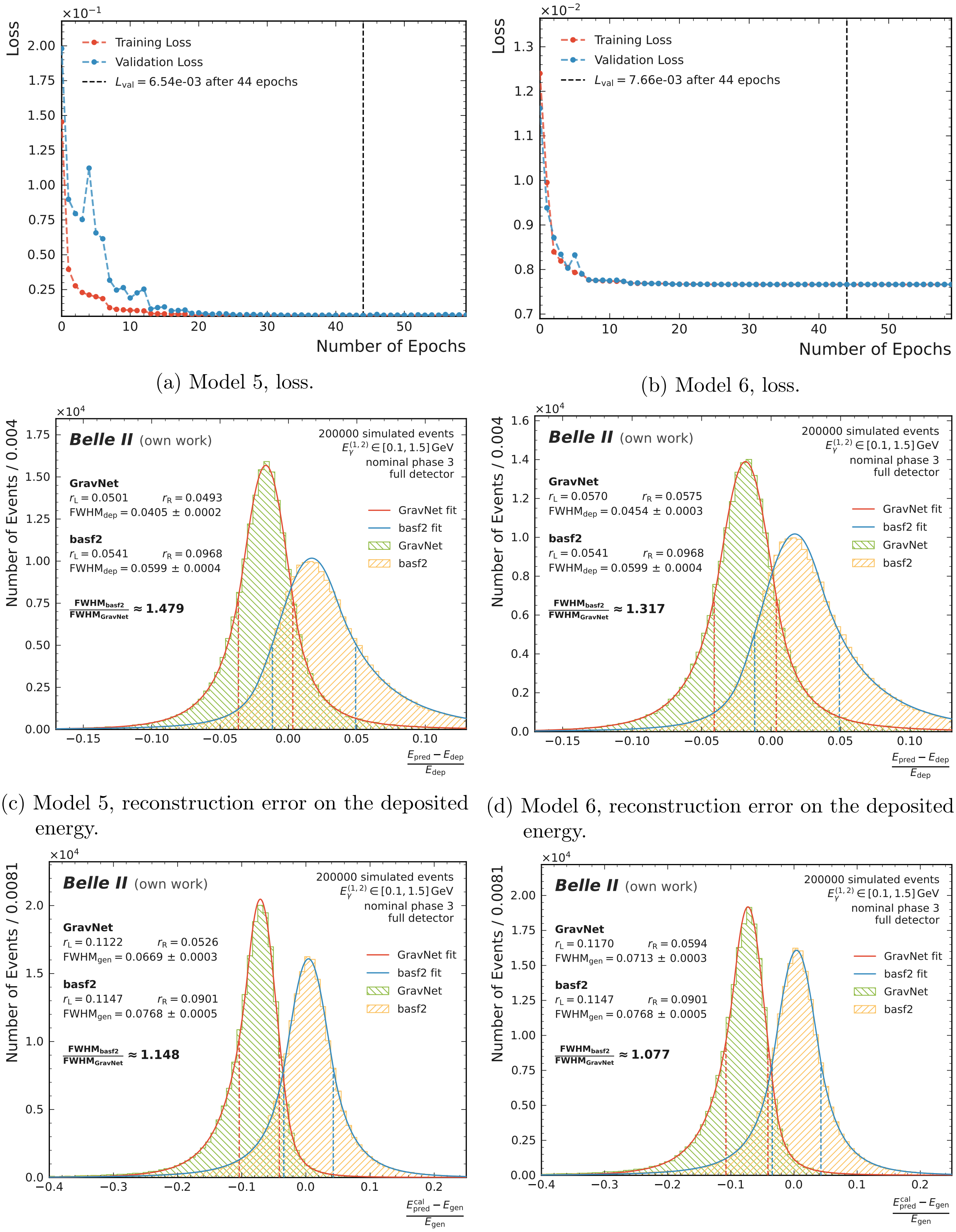
(c) Model 3, reconstruction error on the deposited energy.

(d) Model 4, reconstruction error on the deposited energy.

(e) Model 3, reconstruction error on the generated energy.

(f) Model 4, reconstruction error on the generated energy.

Figure .3.: Loss progression and fits for the resolutions on the deposited and the generated photon energy of two GravNet models with different features. The plots to the left show model 3 with global coordinates, local maxima, recorded energy, and pulse shape discrimination information. The plots to the right show model 4 with global coordinates, local maxima, recorded energy, and recorded time.



(a) Model 5, loss.

(b) Model 6, loss.

(c) Model 5, reconstruction error on the deposited energy.

(d) Model 6, reconstruction error on the deposited energy.

(e) Model 5, reconstruction error on the generated energy.

(f) Model 6, reconstruction error on the generated energy.

Figure .4.: Loss progression and fits for the resolutions on the deposited and the generated photon energy of two GravNet models with different features. The plots to the left show model 5 with global coordinates, local coordinates, local maxima, recorded energy, recorded time, and pulse shape discrimination information. The plots to the right show model 6 with non-cyclical global coordinates, local maxima, and recorded energy.

B. Metric Definitions

B.1. Linear Interpolation of Cluster Radii

ξ_m is defined as the ratio

$$\xi_m = \frac{\sum_i^m E_i^{\text{dep}}}{E_{\text{dep}}} \tag{.1}$$

for crystal m with E_{dep} from equation (5.1). Let ξ_n be the ratio of the first crystal n where $\xi_n \geq 0.96$ and $\xi_{n'}$ the ratio of the last crystal n' with $\xi_{n'} < 0.96$. The distances of the two crystals to the cluster center are defined as d_n and $d_{n'}$ respectively. Set up a linear system of equations:

$$\begin{aligned} \alpha\xi_{n'} + \beta &= d_{n'} \\ \alpha\xi_n + \beta &= d_n. \end{aligned} \tag{.2}$$

Given the desired ratio $\xi_R = 0.96$, this leads to

$$R = \alpha\xi_R + \beta, \tag{.3}$$

with

$$\alpha = \frac{d_n - d_{n'}}{\xi_n - \xi_{n'}} \quad \text{and} \quad \beta = d_{n'} - \left(\alpha \cdot \xi_{n'}\right). \tag{.4}$$

B.2. Analytical Calculation of the Full Width Half Maximum

Because the DCB, as well as the underlying distribution, are asymmetric, left and right $\text{FWHM}_{\text{L,R}}$ are calculated separately. Depending on the transition parameters $\alpha_{\text{L,R}}$ of the DCB defined in equation (5.16), each $\text{FWHM}_{\text{L/R}}$ falls either within the Gaussian part of the function or within the exponential tails.

The case $\alpha_{\text{L/R}} > \sqrt{\log 4}$ results in half of the Gaussian FWHM:

$$\text{FWHM}_{\text{L/R}} = \sqrt{2 \cdot \log 2} \cdot \sigma. \tag{.5}$$

For $\alpha_{\text{L/R}} \leq \sqrt{\log 4}$ the following applies to the FWHM in the tails:

$$\text{FWHM}_{\text{L/R}} = \left| \mu + \left(\frac{\sigma}{\alpha_{\text{L/R}}} \left(|\alpha_{\text{L/R}}|^2 + S_{\text{L/R}} - n_{\text{L/R}} - \mu|\alpha_{\text{L/R}}|\sigma^{-1} \right) \right) \right|, \tag{.6}$$

with

$$S_{\text{L/R}} = \left(\frac{1}{2} \cdot n_{\text{L/R}}^{-n_{\text{L/R}}} \cdot \exp\left(\alpha_{\text{L/R}}^2/2\right) \right)^{-1/n_{\text{L/R}}}. \tag{.7}$$

Finally, the two partial $\text{FWHM}_{\text{L,R}}$ are added to

$$\text{FWHM} = \text{FWHM}_{\text{L}} + \text{FWHM}_{\text{R}}. \tag{.8}$$

B.3. Fuzzy Clustering Agreement Index

Given $u \in \{1, 2\}$ are clusters associated with the classes cluster one and cluster two. Then $w_i^{(u)}$ is the membership strength (or weight) of cluster u in data point (crystal) i as given by a clustering (algorithm) W. The size of a cluster according to clustering W is given by the sum of weights (see section 5.2.2) and noted $o_W^{(u)} = \sum_i w_i^{(u)}$ in this context. Considering another clustering V with the membership strength $v_i^{(u)}$, then the pairwise accordance of any two classes u and u' as evaluated by W and V is given by:

$$o_{WV}^{(uu')} = \sum_i w_i^{(u)} \cdot v_i^{(u')}. \tag{.9}$$

The sum over all classes in both clusterings concludes the calculation. The pairwise accordance $\mathcal{O}_{WV}$ between W and V, the self-accordance of W $\mathcal{O}_{WW}$, the self-accordance of V $\mathcal{O}_{VV}$, and the expected accordance of two clusterings by chance $\mathcal{E}_{WV}$ are then defined as:

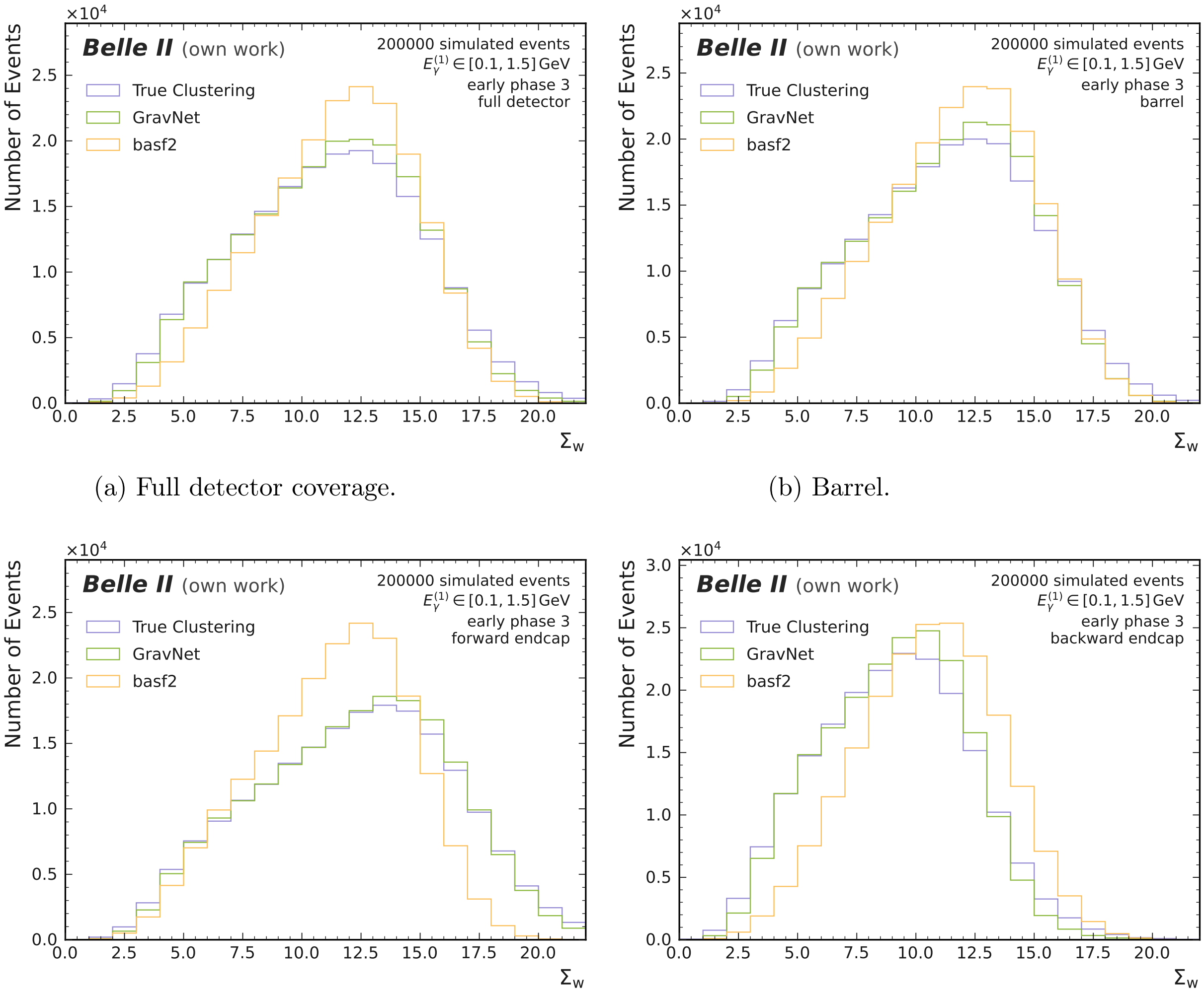
$$\begin{aligned} \mathcal{O}_{WW} &= \sum_{u \in W} \sum_{u' \in W} \phi\left(o_{WW}^{(uu')}\right) \ , \qquad & \mathcal{O}_{VV} &= \sum_{u \in V} \sum_{u' \in V} \phi\left(o_{VV}^{(uu')}\right) \ , \\ \mathcal{O}_{WV} &= \sum_{u \in W} \sum_{u' \in V} \phi\left(o_{WV}^{(uu')}\right) \ , \qquad & \mathcal{E}_{WV} &= \sum_{u \in W} \sum_{u \in V} \phi\left(\frac{o_W^{(u)} o_V^{(u)}}{n}\right) \ . \end{aligned} \tag{.10}$$

Herein n is the total number of data points (crystals) in the event and the scaling function is chosen to be $\phi(x) = x \log x$. The same formula is valid for two clusters and background with $u \in \{1, 2, 3\}$, since FCAI treats background as equal clustering class [3].

C. Single Photon Cluster Metrics

C.1. Event Properties

Sum of Weights



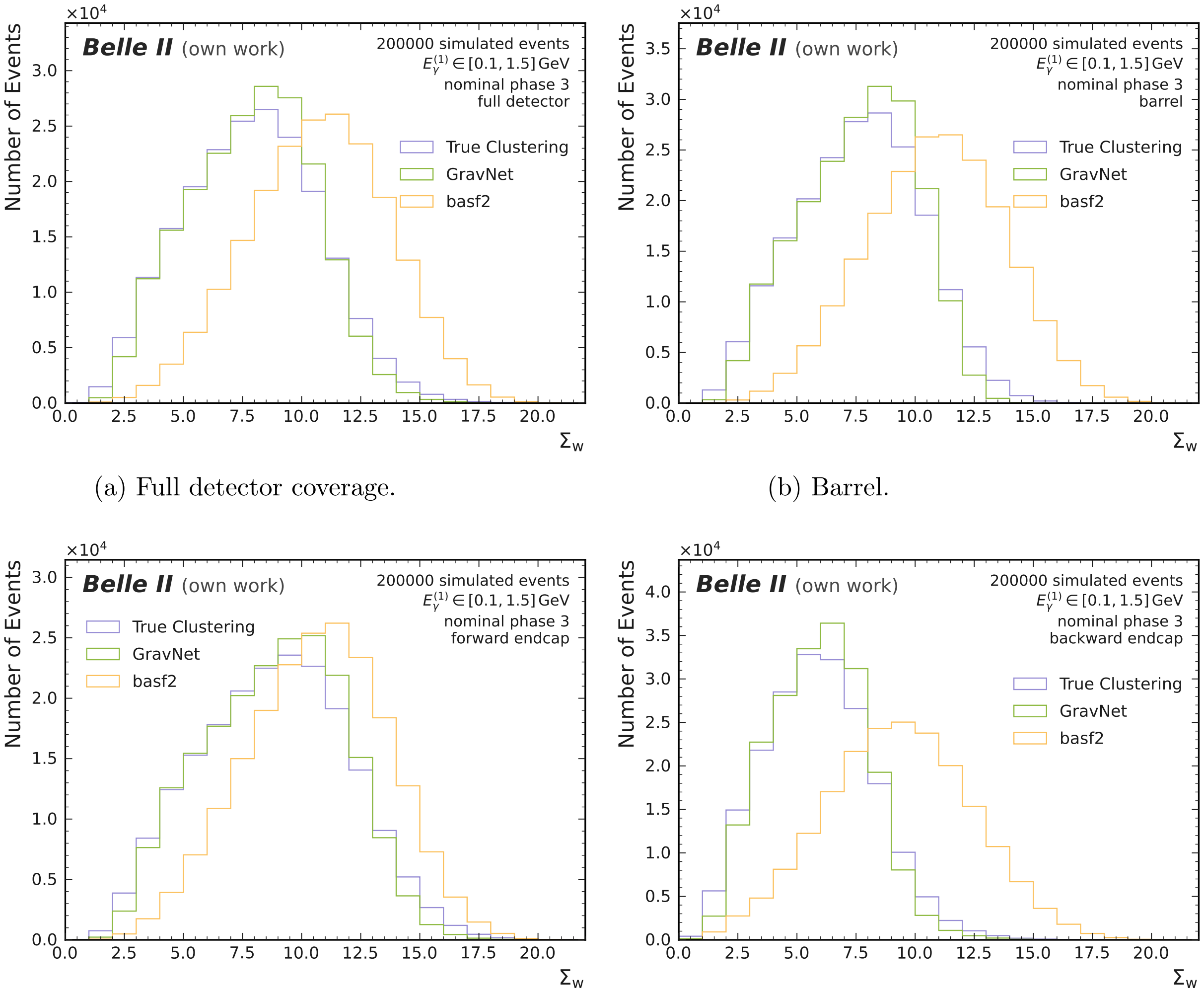
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .5.: Distribution in the sum of weights Σ_w for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

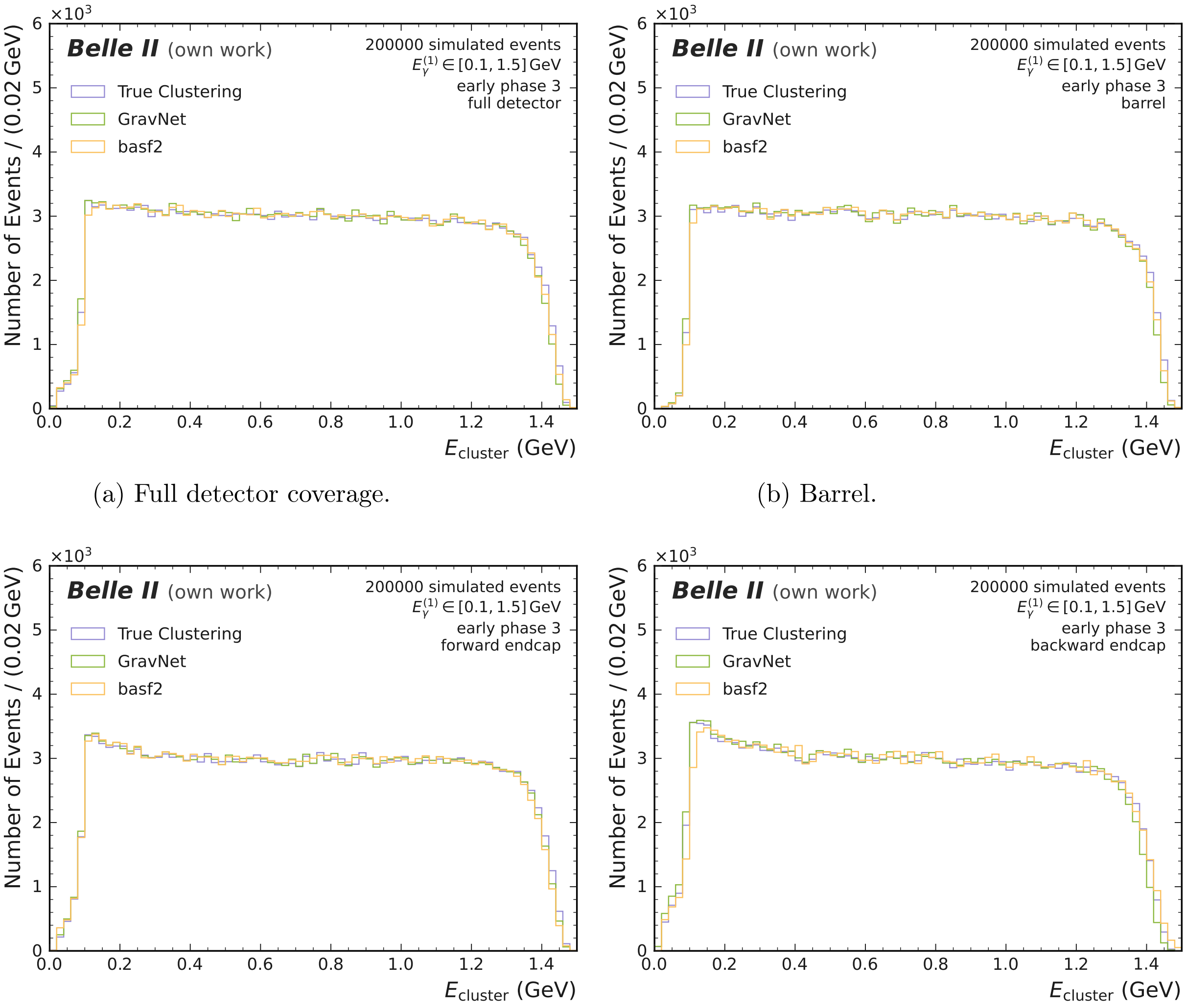
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .6.: Distribution in the sum of weights Σ_w for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Energy



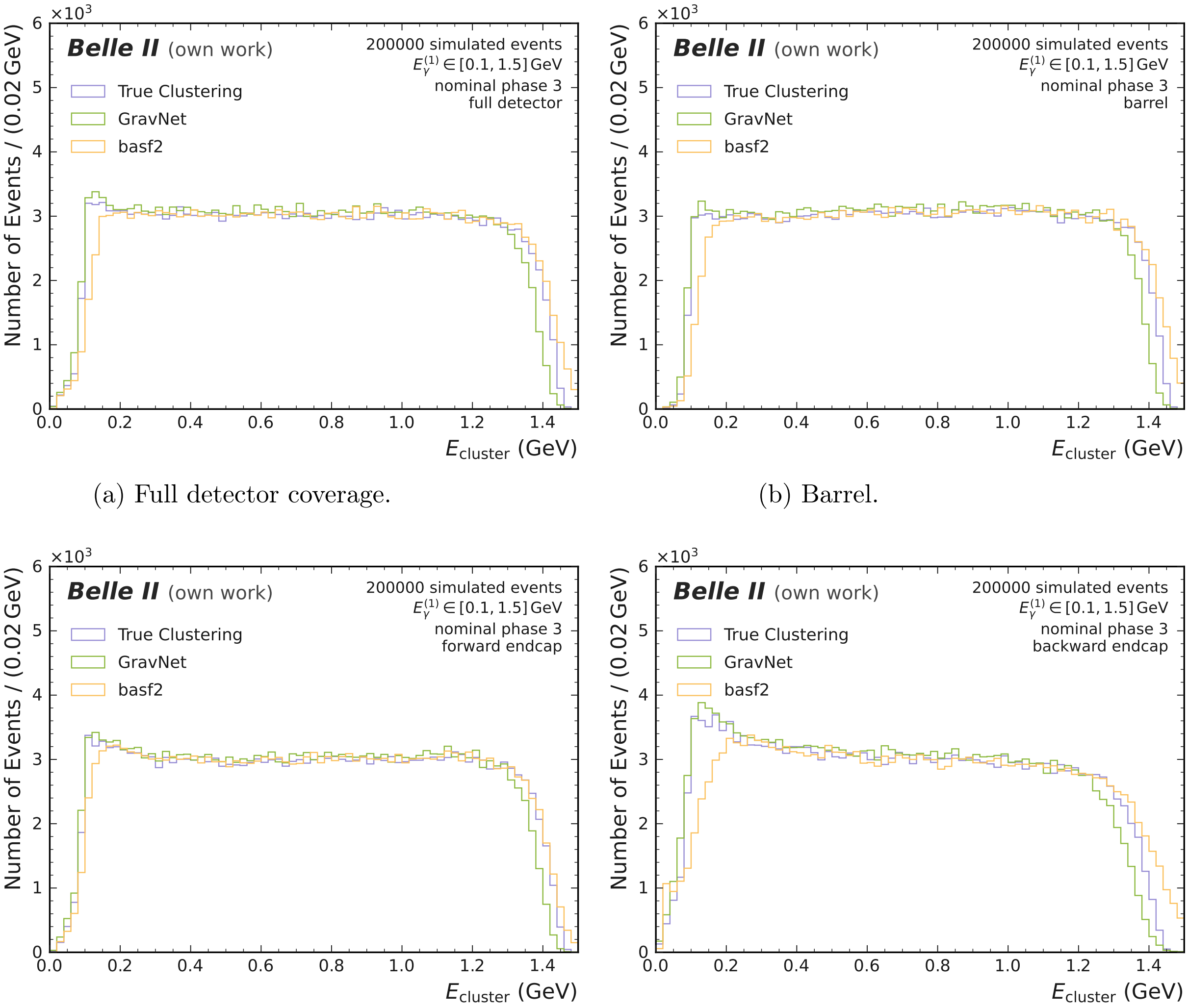
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .7.: Distribution in cluster energies E_{cluster} for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

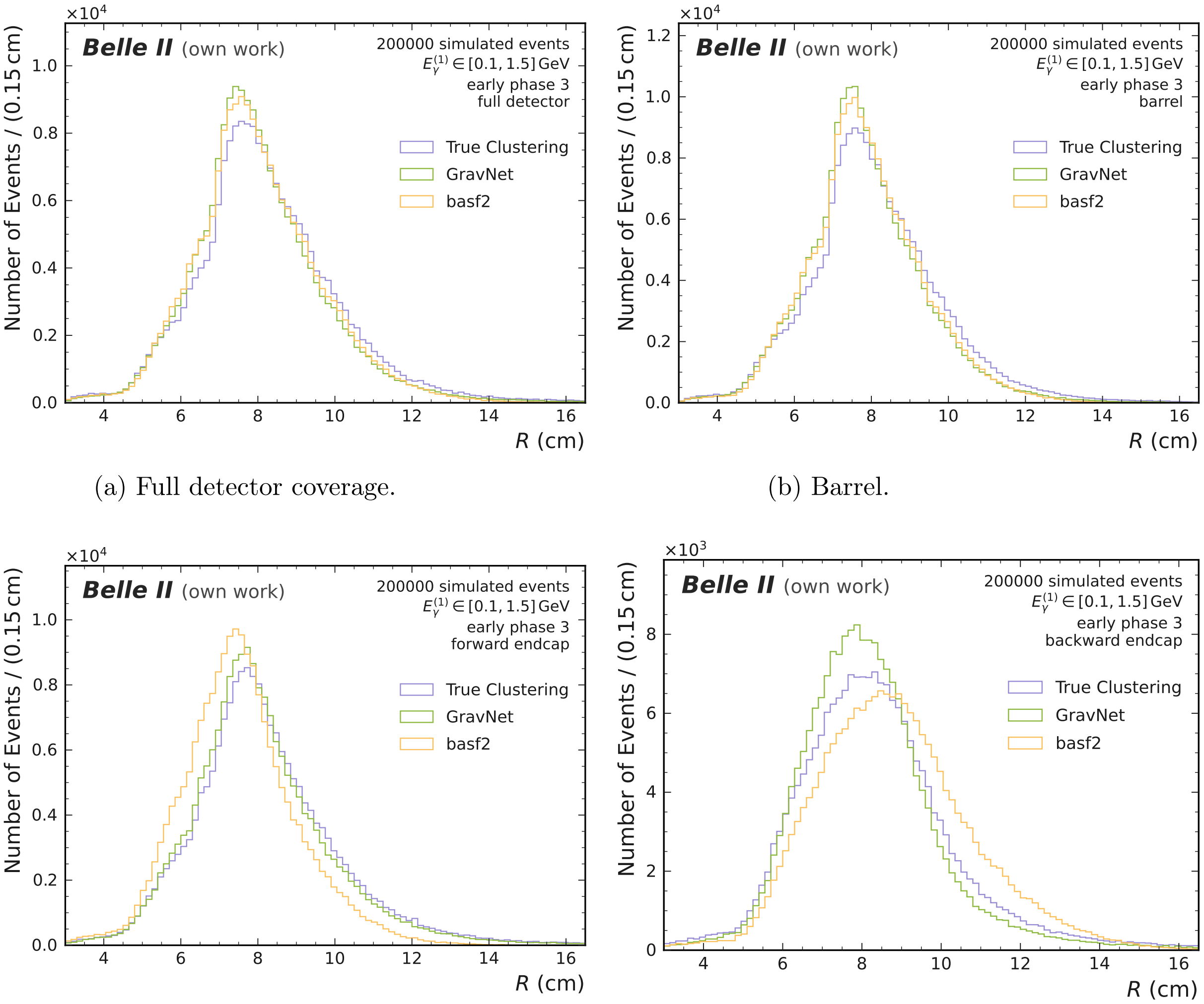
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .8.: Distribution in cluster energies E_{cluster} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Radius



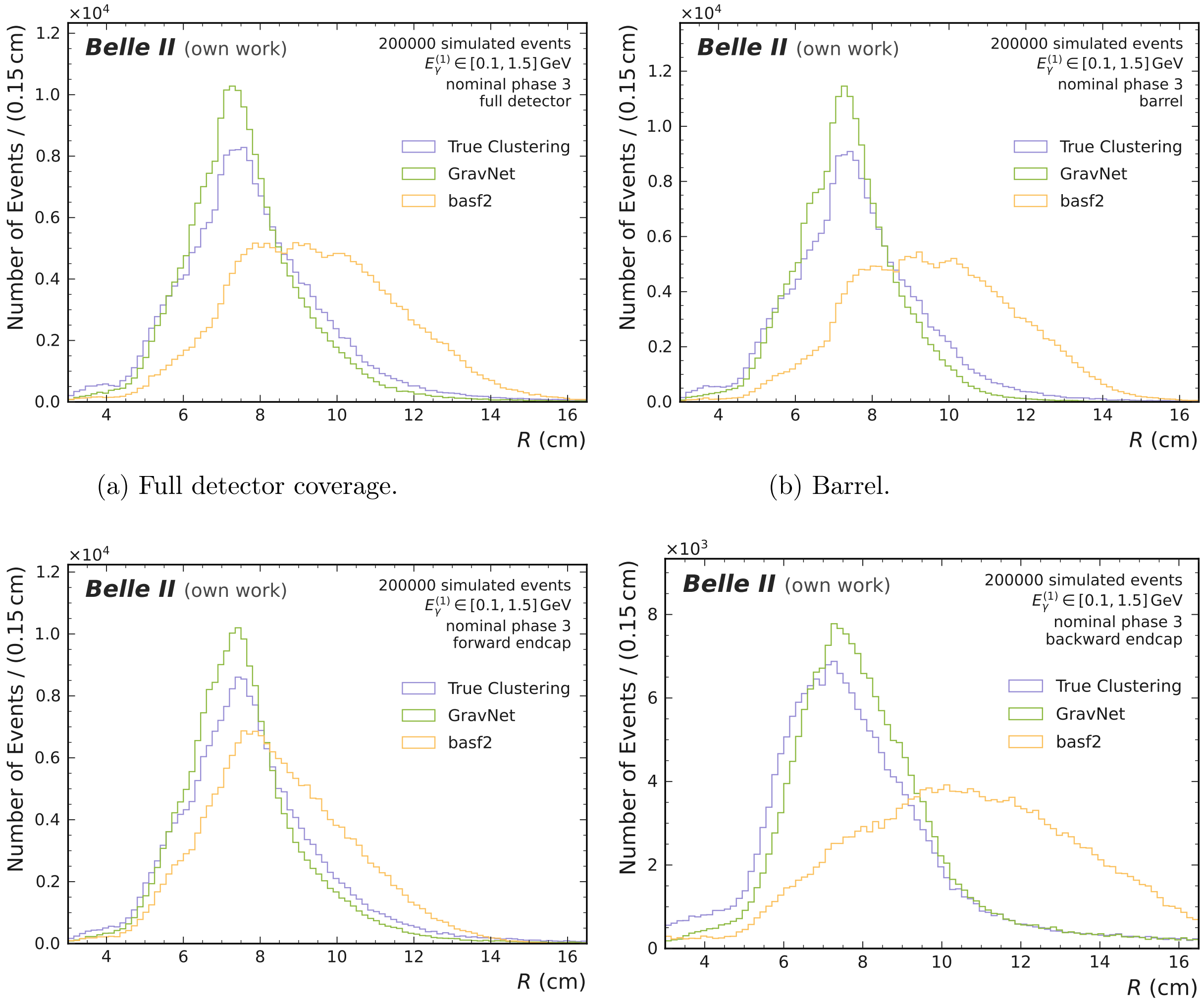
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .9.: Distribution in radii R for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

(b) Barrel.

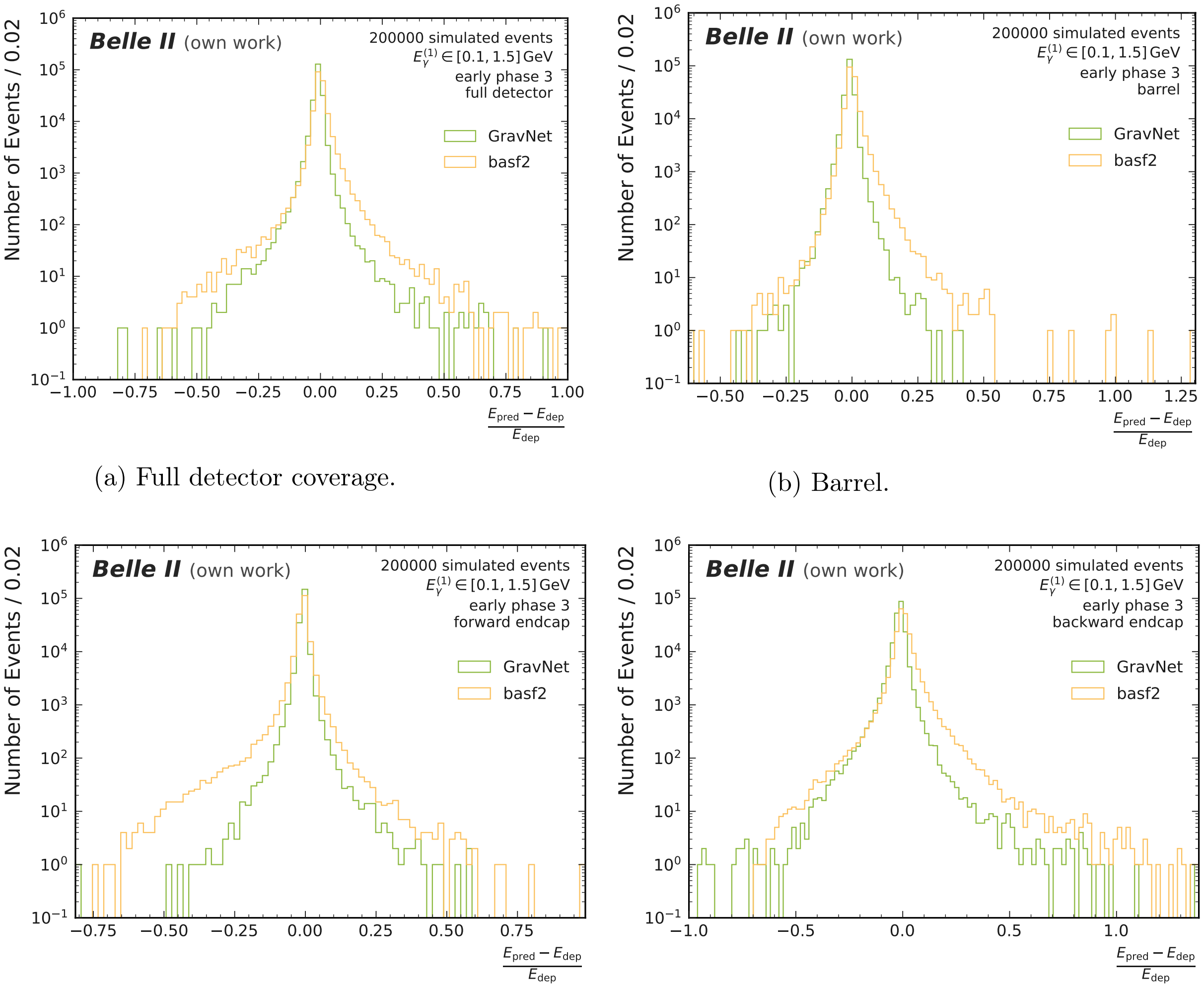
(c) Forward endcap.

(d) Backward endcap.

Figure .10.: Distribution in radii R for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

C.2. Performance Evaluation

Energy Resolution



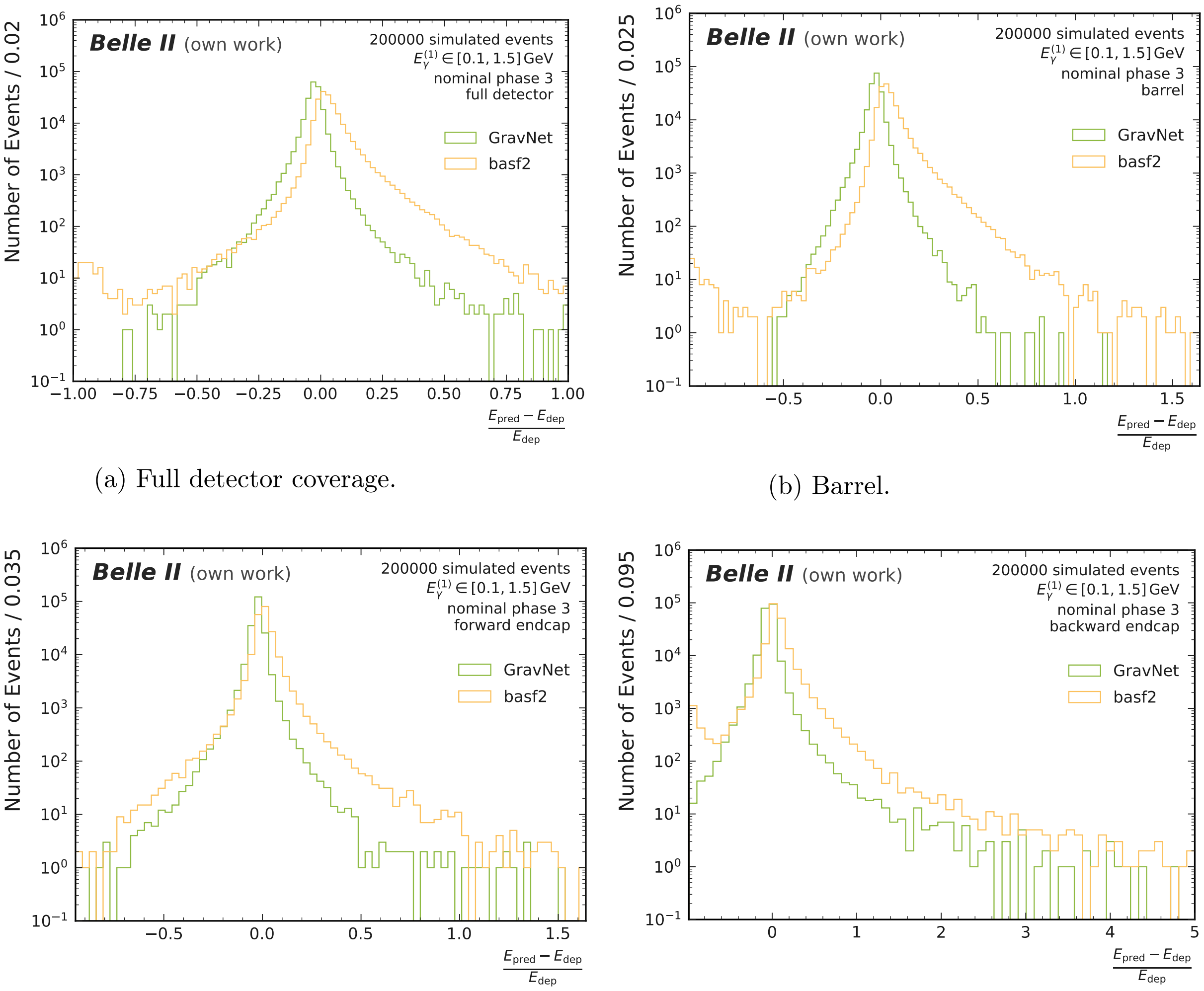
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .11.: Distribution in deposited errors η_dep for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



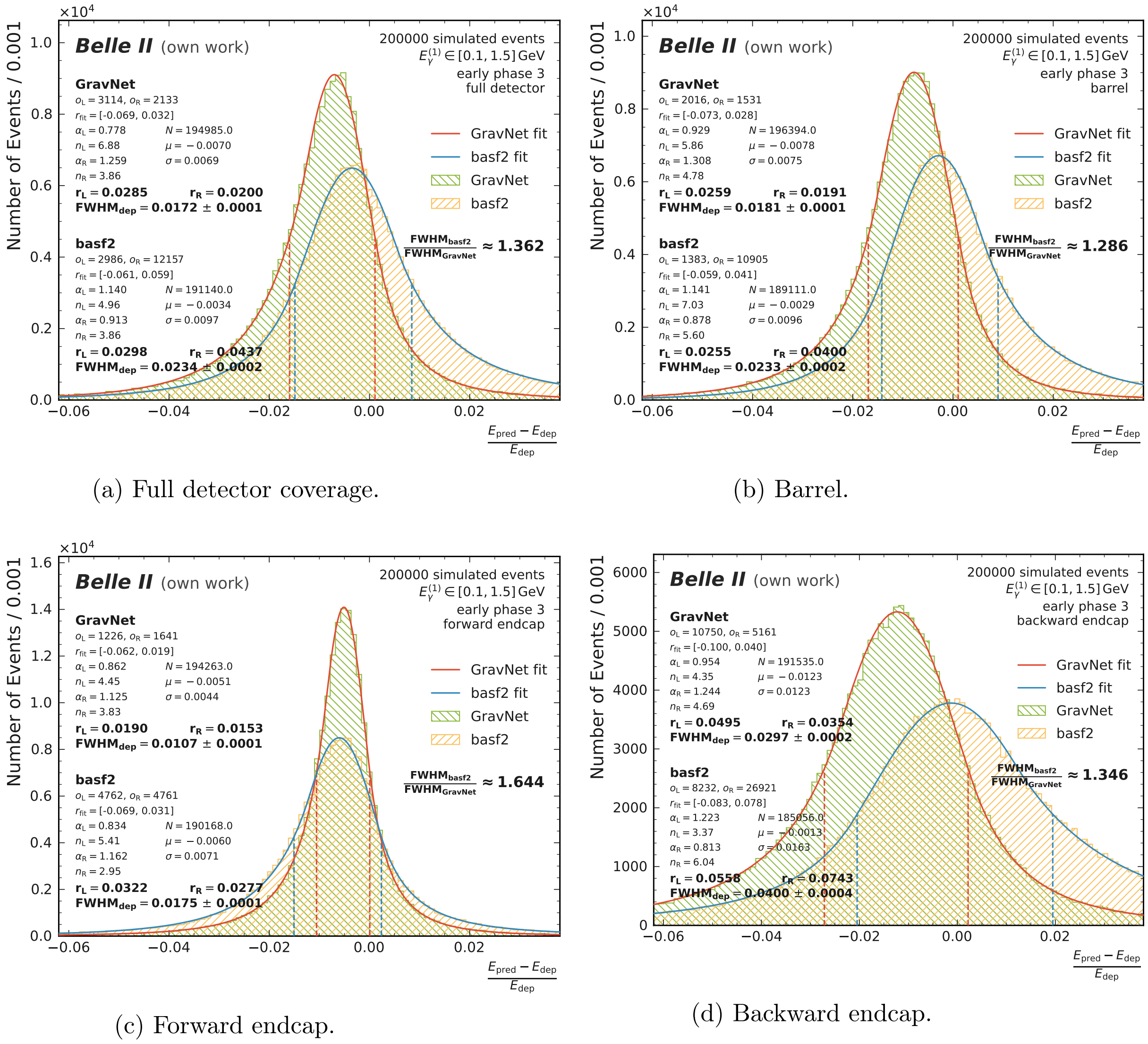
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .12.: Distribution in deposited errors η_{dep} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



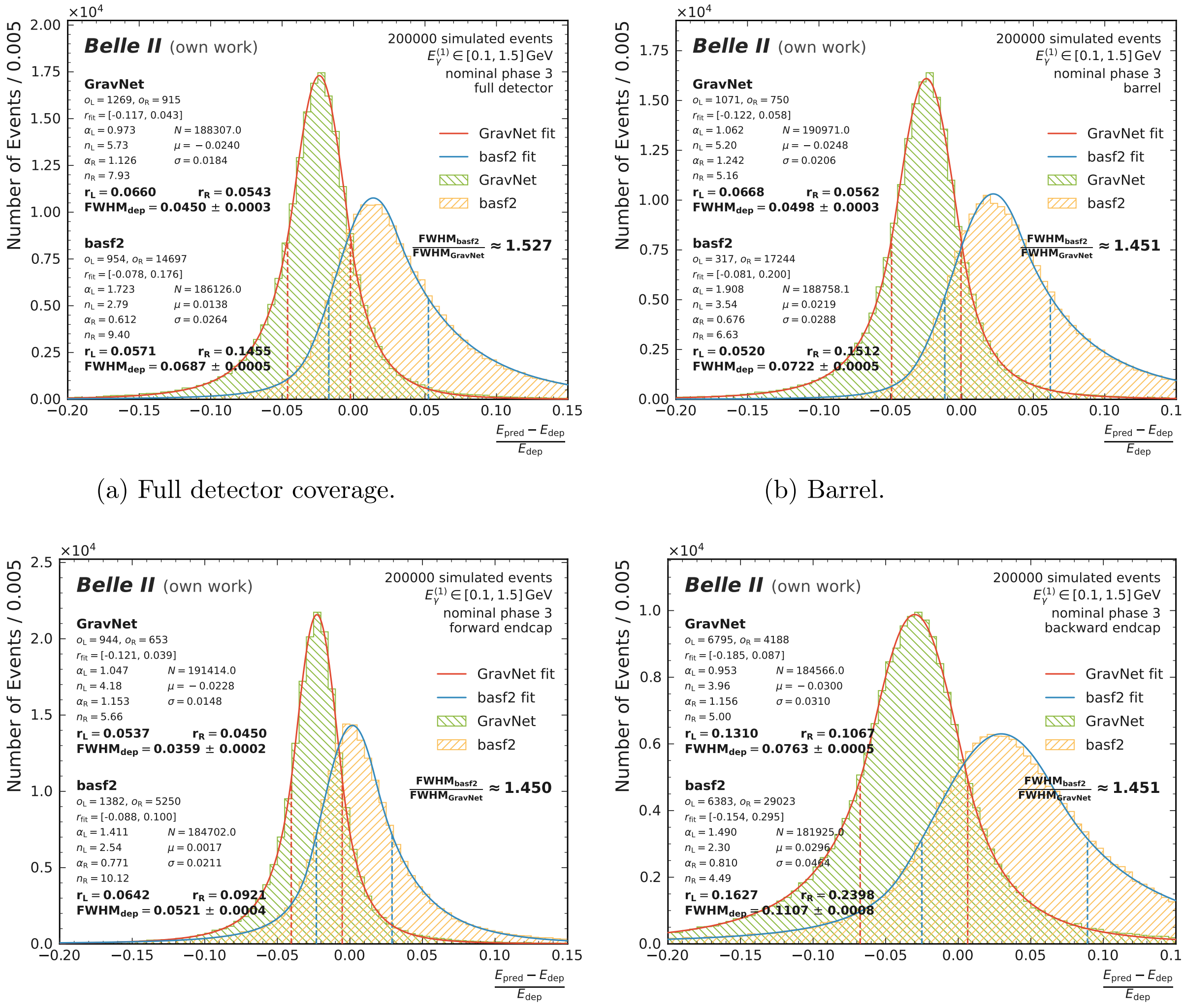
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .13.: Fit for the distribution in deposited errors η_{dep} for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



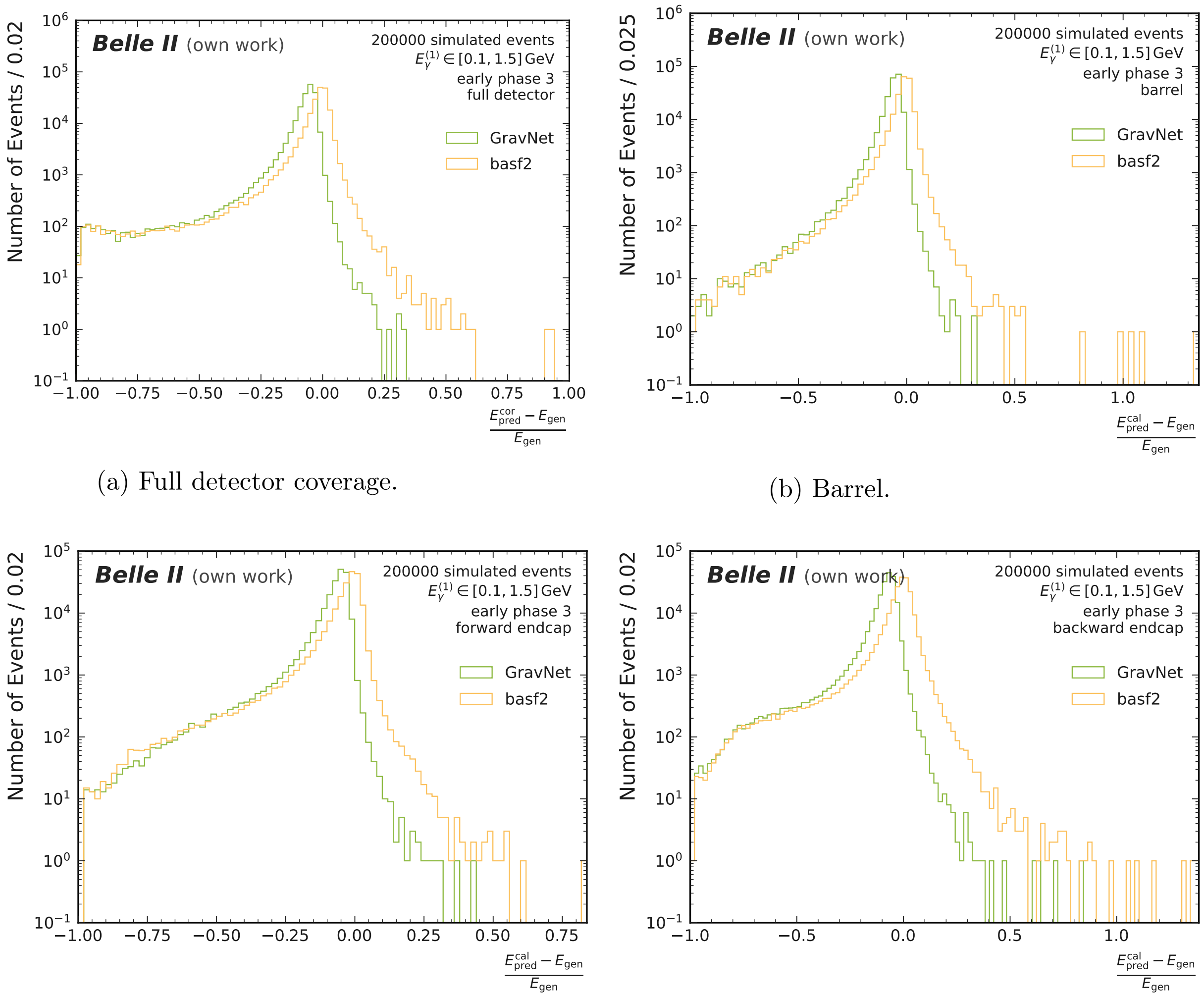
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .14.: Fit for the distribution in deposited errors η_{dep} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



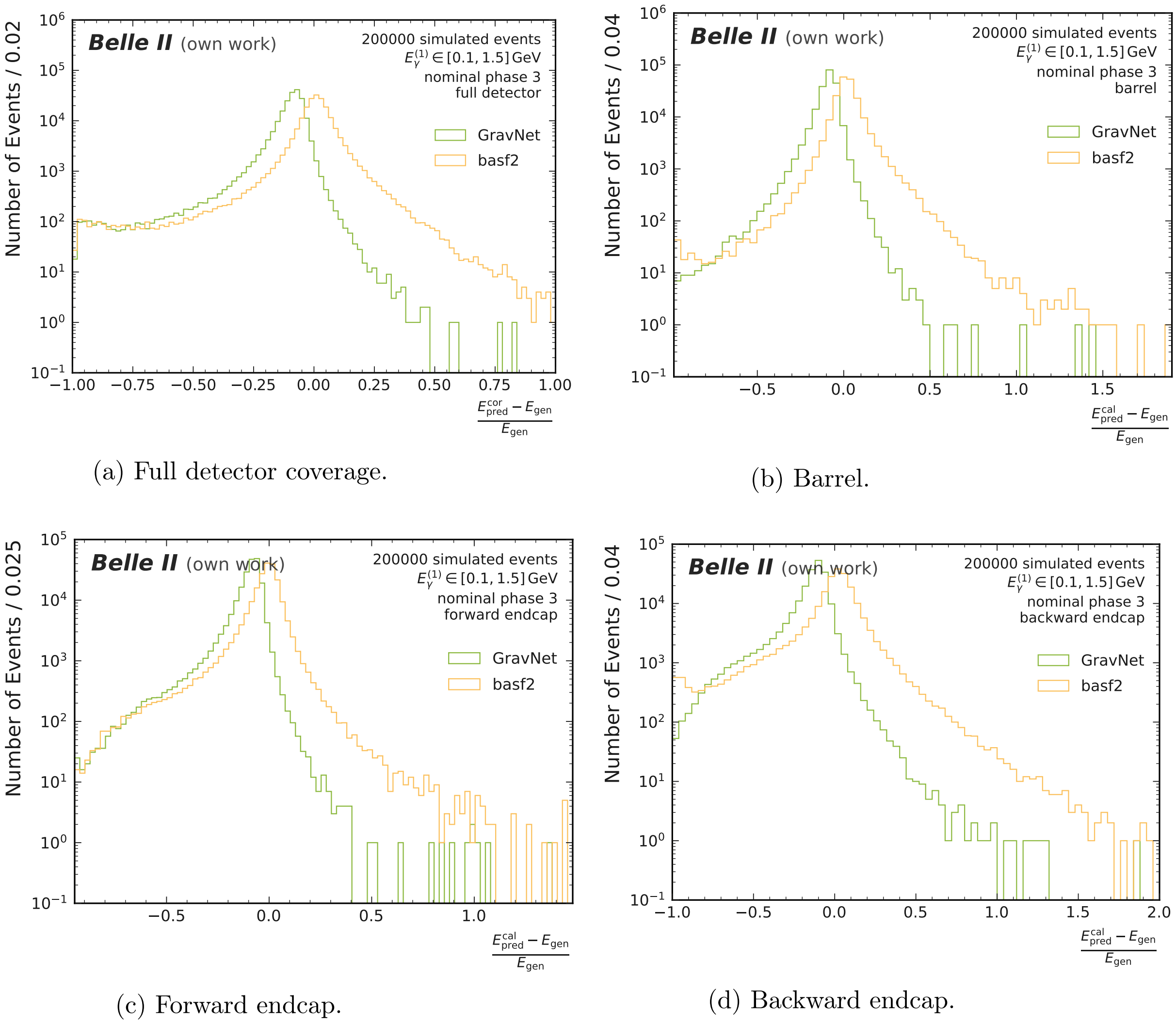
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .15.: Distribution in generated errors η_{gen} for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



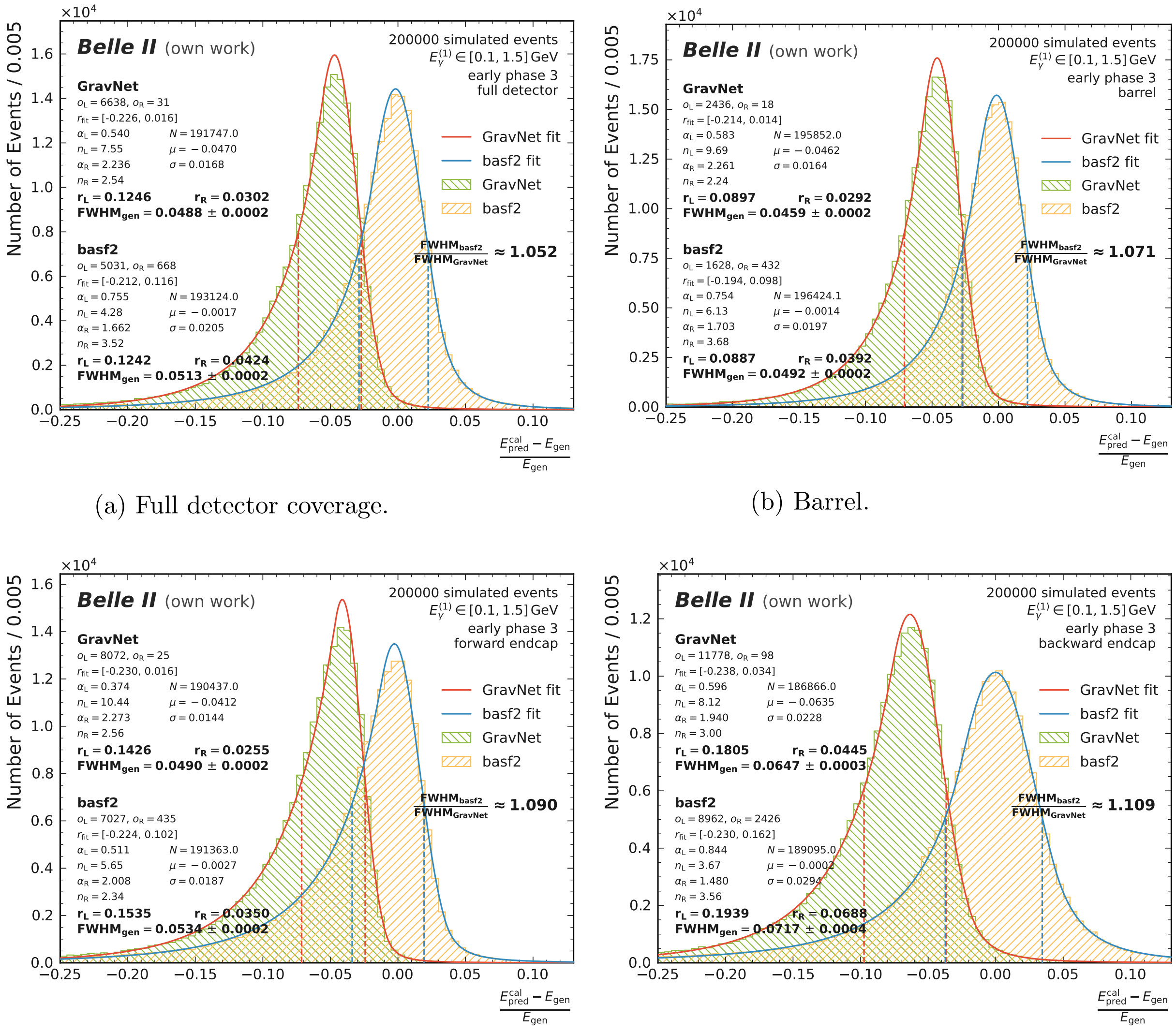
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .16.: Distribution in generated errors η_{gen} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



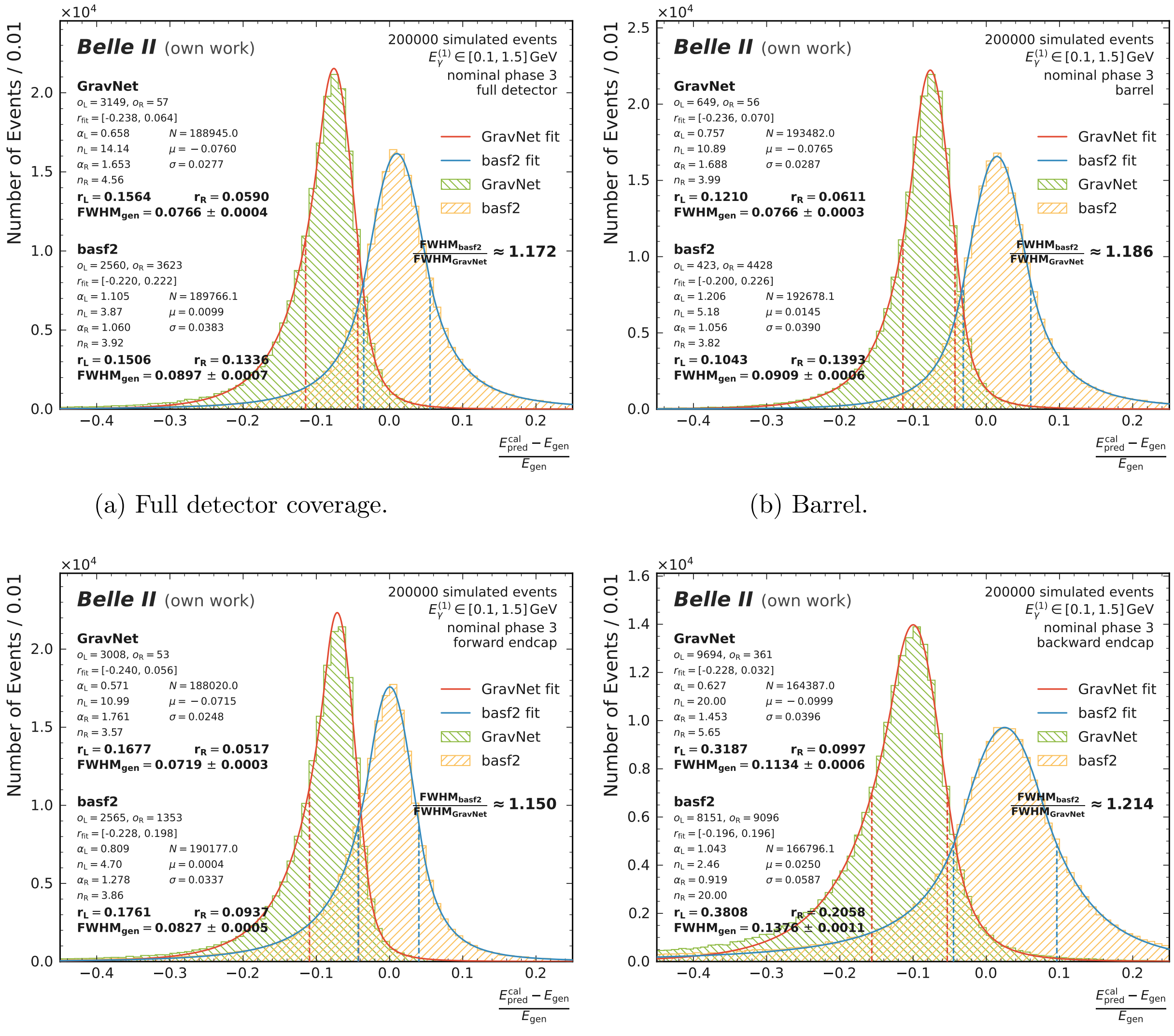
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .17.: Fit for the distribution in generated errors η_{gen} for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately.Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

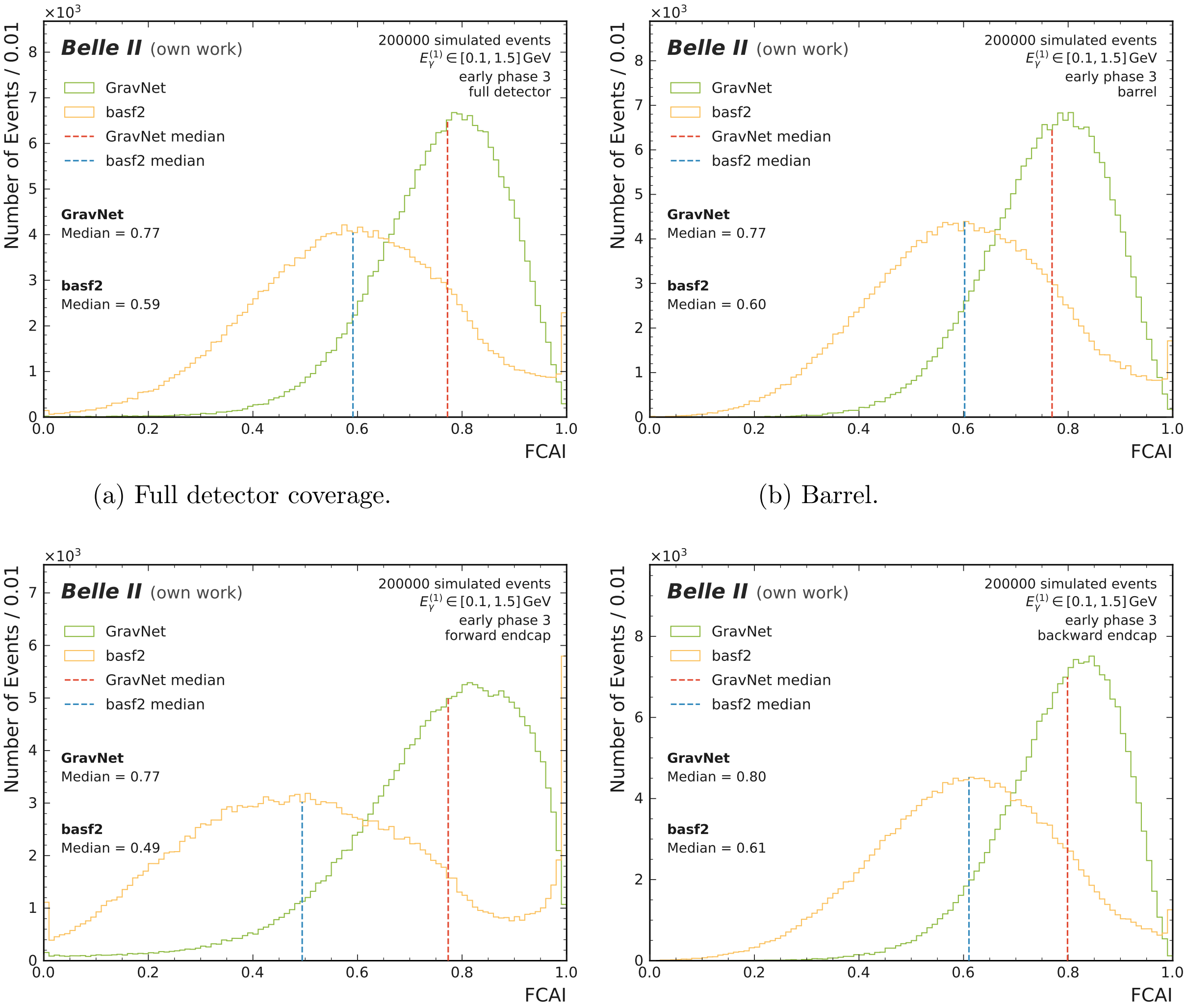
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .18.: Fit for the distribution in generated errors η_{gen} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

Fuzzy Clustering Agreement Index



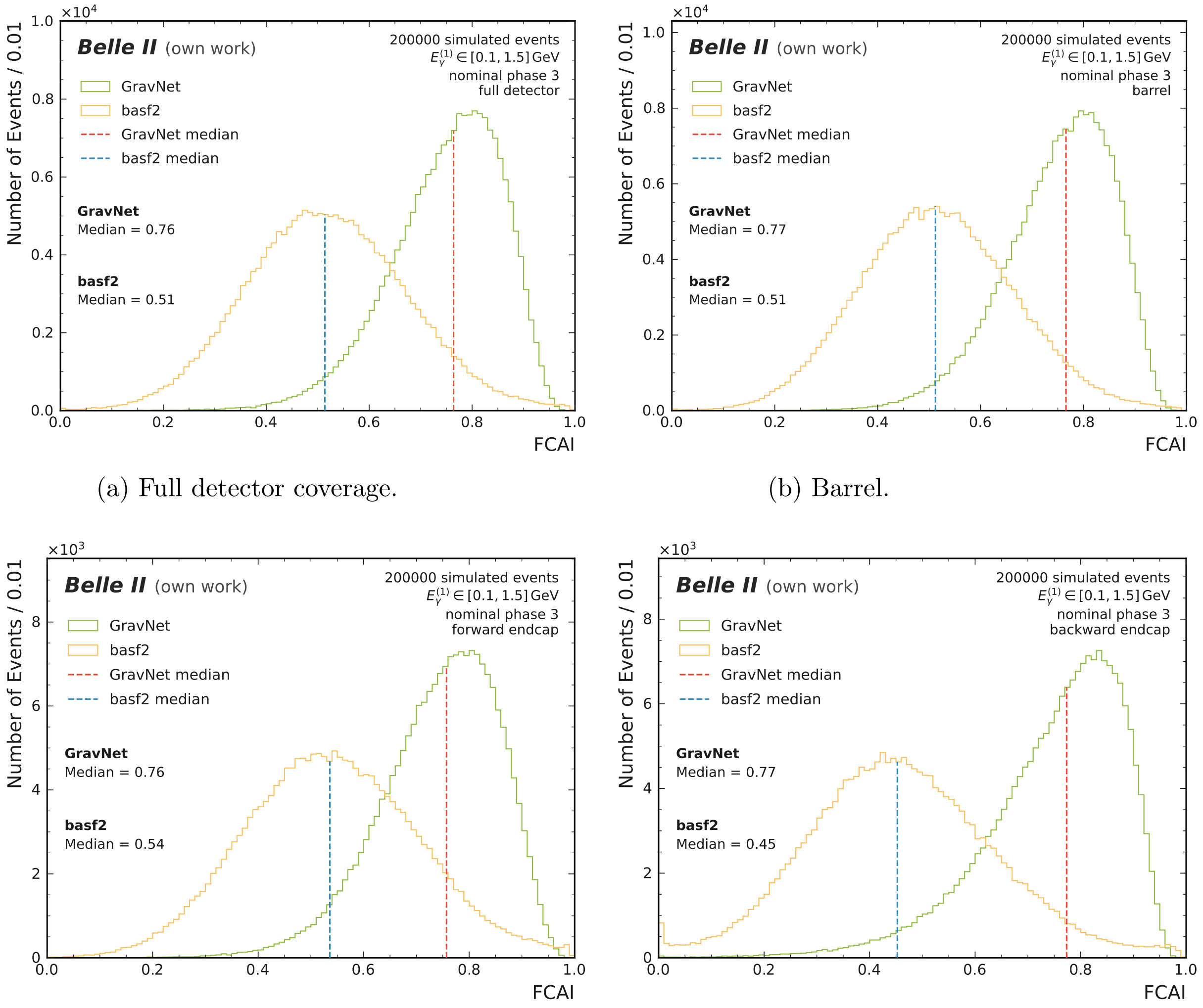
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .19.: Distribution in FCAI for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

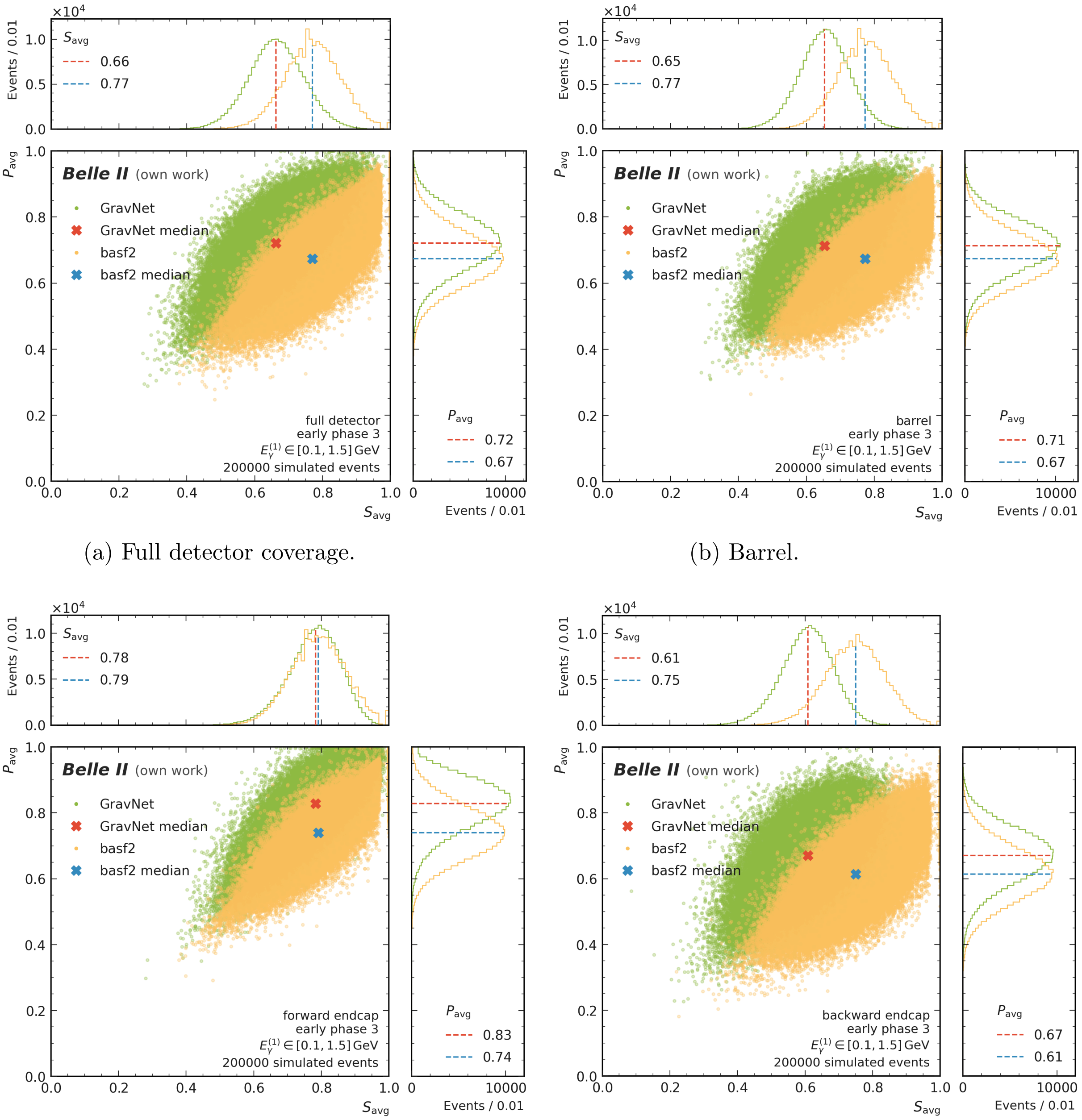
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .20.: Distribution in FCAI for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

Sensitivity and Precision



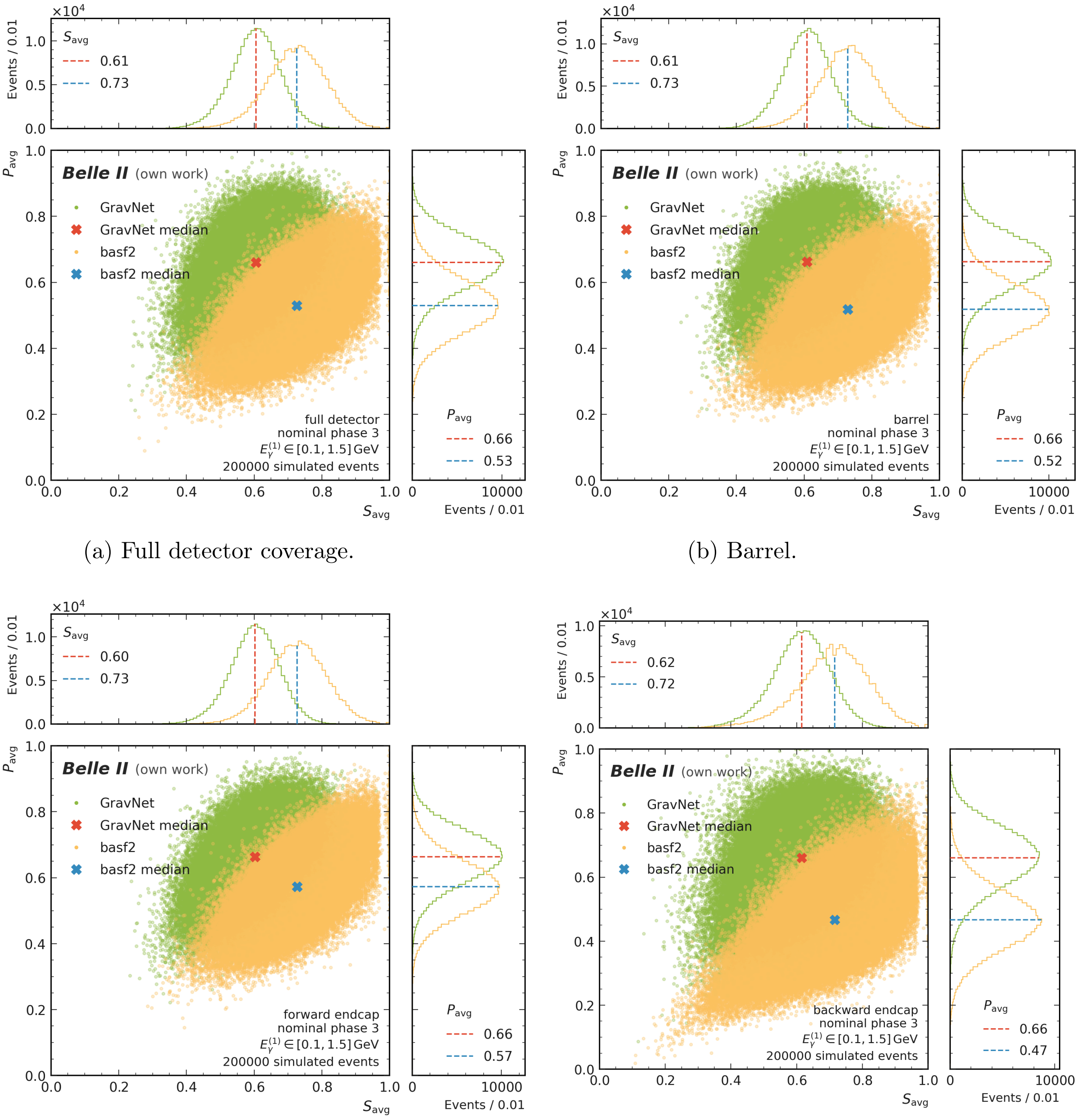
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .21.: Distribution in average sensitivity S_{avg} and precision P_{avg} for the one-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

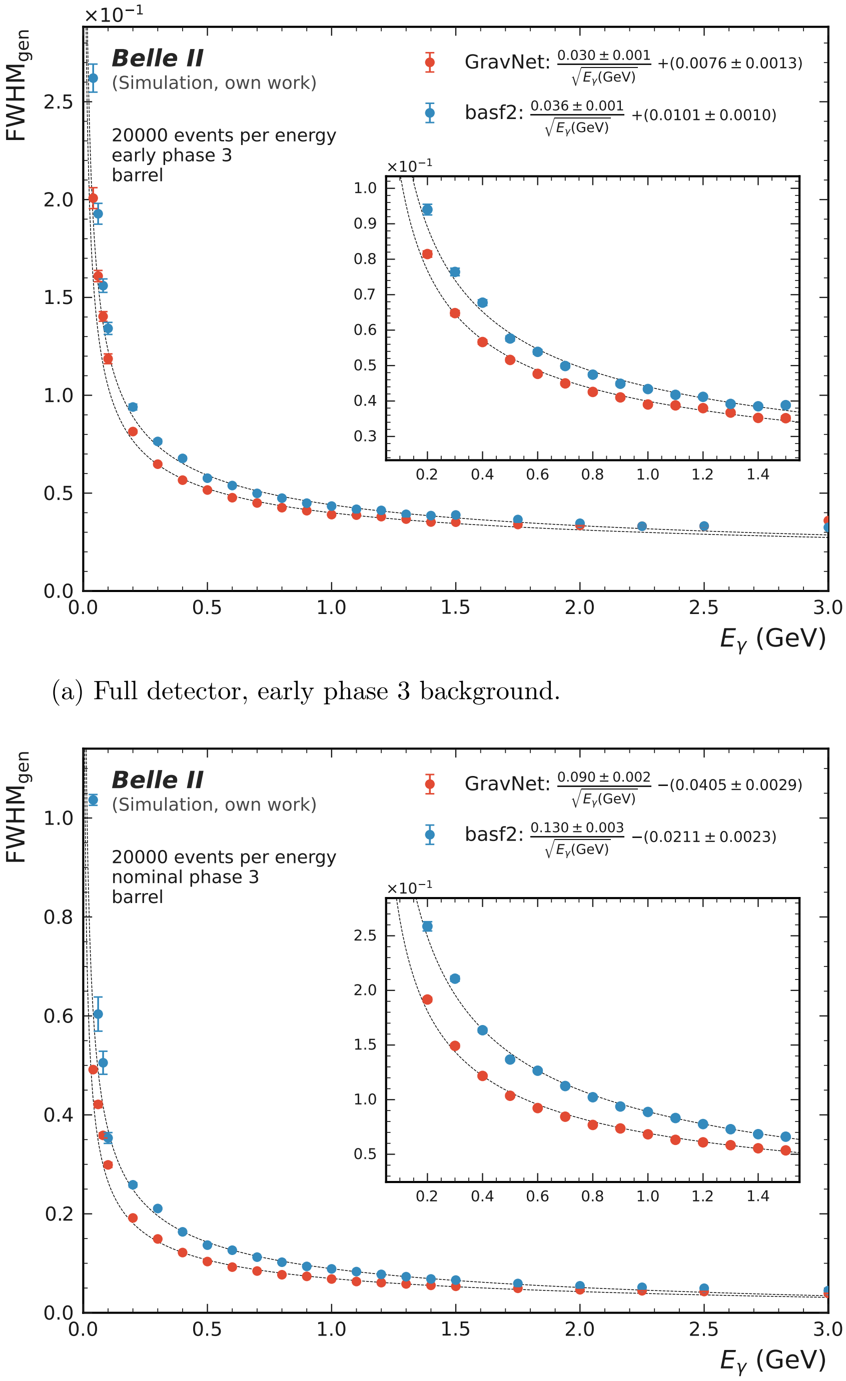
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .22.: Distribution in average sensitivity S_{avg} and precision P_{avg} for the one-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

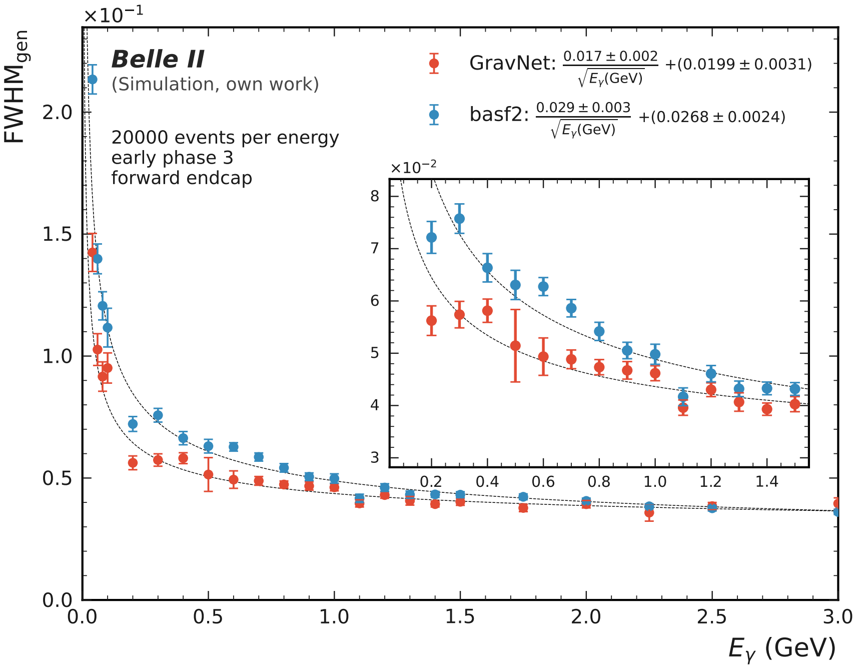
C.3. Energy Dependence



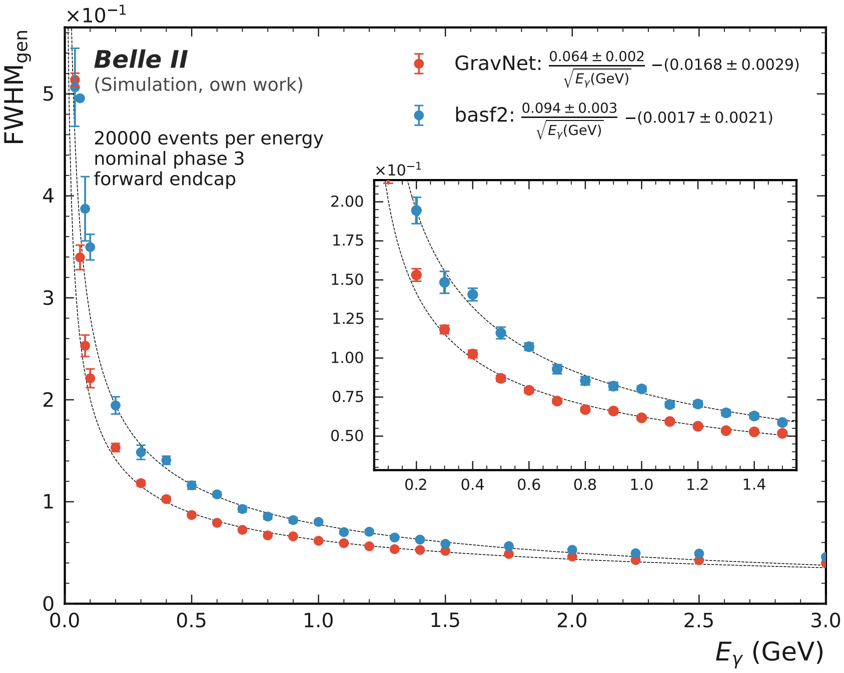
(a) Full detector, early phase 3 background.

(b) Full detector, nominal phase 3 background.

Figure .23.: Shown is the resolution on the generated energy FWHM_{gen} in dependence of the generated photon energy E_γ for the one-cluster toy studies. The top plot shows events with early phase 3 background in the barrel, the bottom plot nominal phase 3 background in the barrel. Both plots compare GravNet and the basf2 baseline. Each data point marks the FWHM_{gen} of 20000 events, each at a fixed energy $E_\gamma \in [0.01, 3.0]$ GeV. For each algorithm a fit models the relation with an inverse square root, the one sigma band is highlighted.

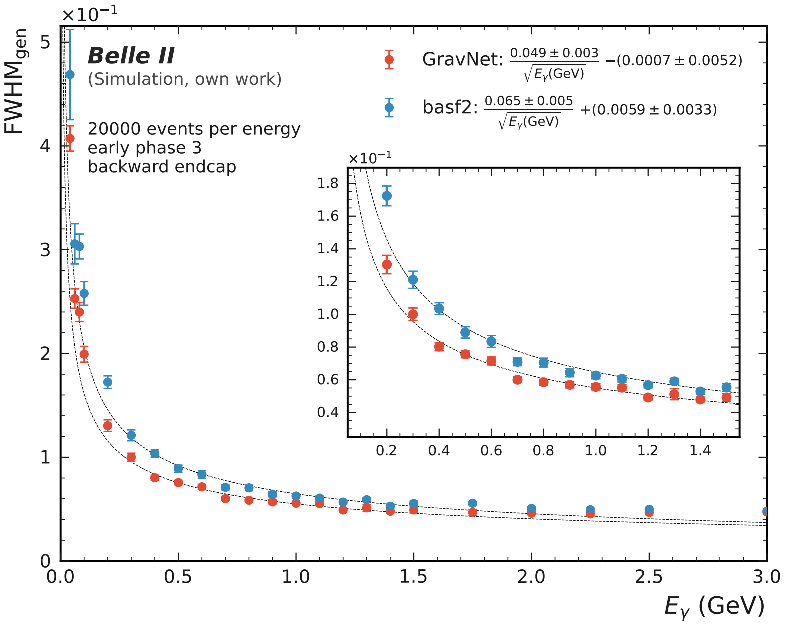


(a) Full detector, early phase 3 background.

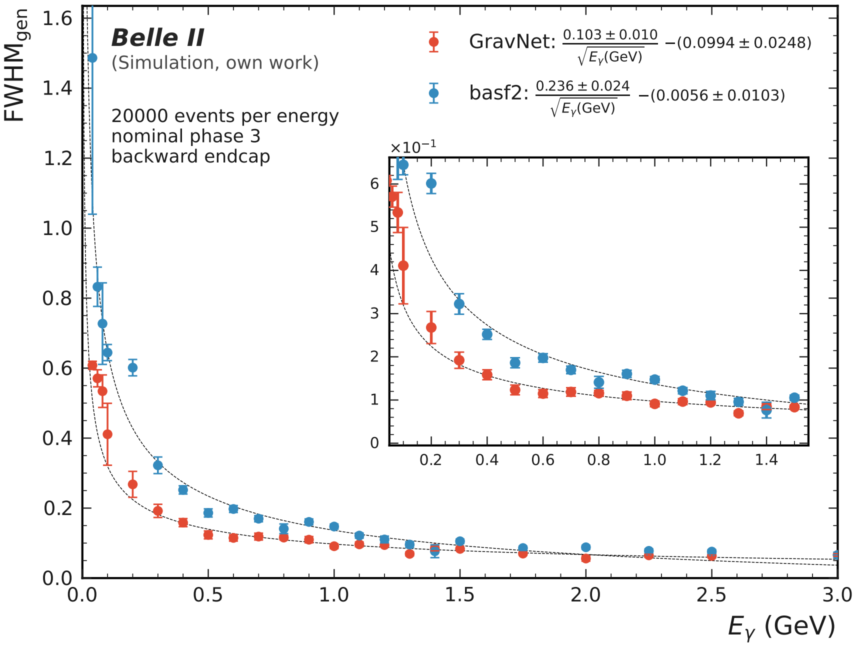


(b) Full detector, nominal phase 3 background.

Figure .24.: Shown is the resolution on the generated energy $\mathrm{FWHM}_{\mathrm{gen}}$ in dependence of the generated photon energy E_γ for the one-cluster toy studies. The top plot shows events with early phase 3 background in the forward endcap, the bottom plot nominal phase 3 background in the forward endcap. Both plots compare GravNet and the basf2 baseline. Each data point marks the $\mathrm{FWHM}_{\mathrm{gen}}$ of 20000 events, each at a fixed energy $E_\gamma \in [0.01, 3.0]$ GeV. For each algorithm a fit models the relation with an inverse square root, the one sigma band is highlighted.



(a) Full detector, early phase 3 background.



(b) Full detector, nominal phase 3 background.

Figure .25.: Shown is the resolution on the generated energy FWHM$_{\mathrm{gen}}$ in dependence of the generated photon energy E_γ for the one-cluster toy studies. The top plot shows events with early phase 3 background in the backward endcap, the bottom plot nominal phase 3 background in the backward endcap. Both plots compare GravNet and the basf2 baseline. Each data point marks the FWHM$_{\mathrm{gen}}$ of 20000 events, each at a fixed energy $E_\gamma \in [0.01, 3.0]$ GeV. For each algorithm a fit models the relation with an inverse square root, the one sigma band is highlighted.

D. Two Overlapping Photon Cluster Metrics

D.1. Event Properties

Leakage

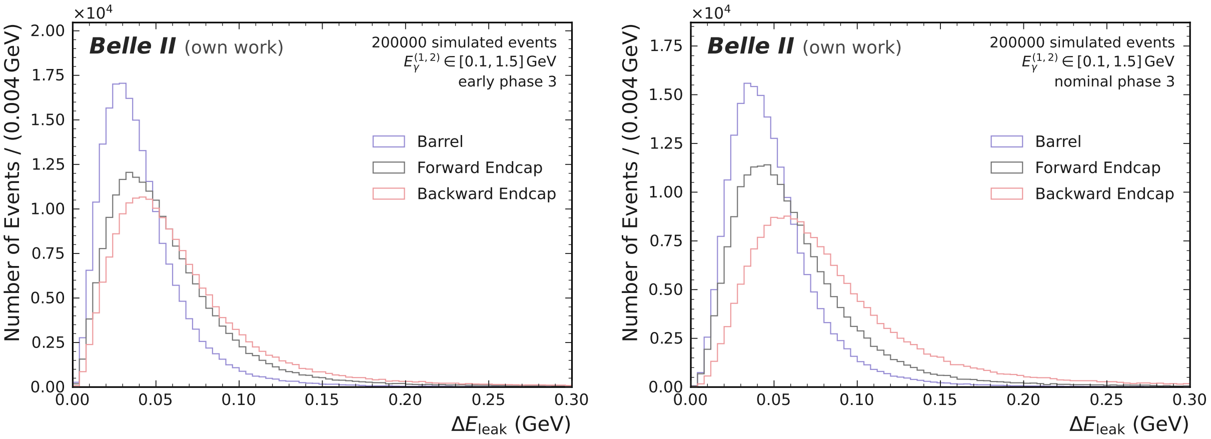
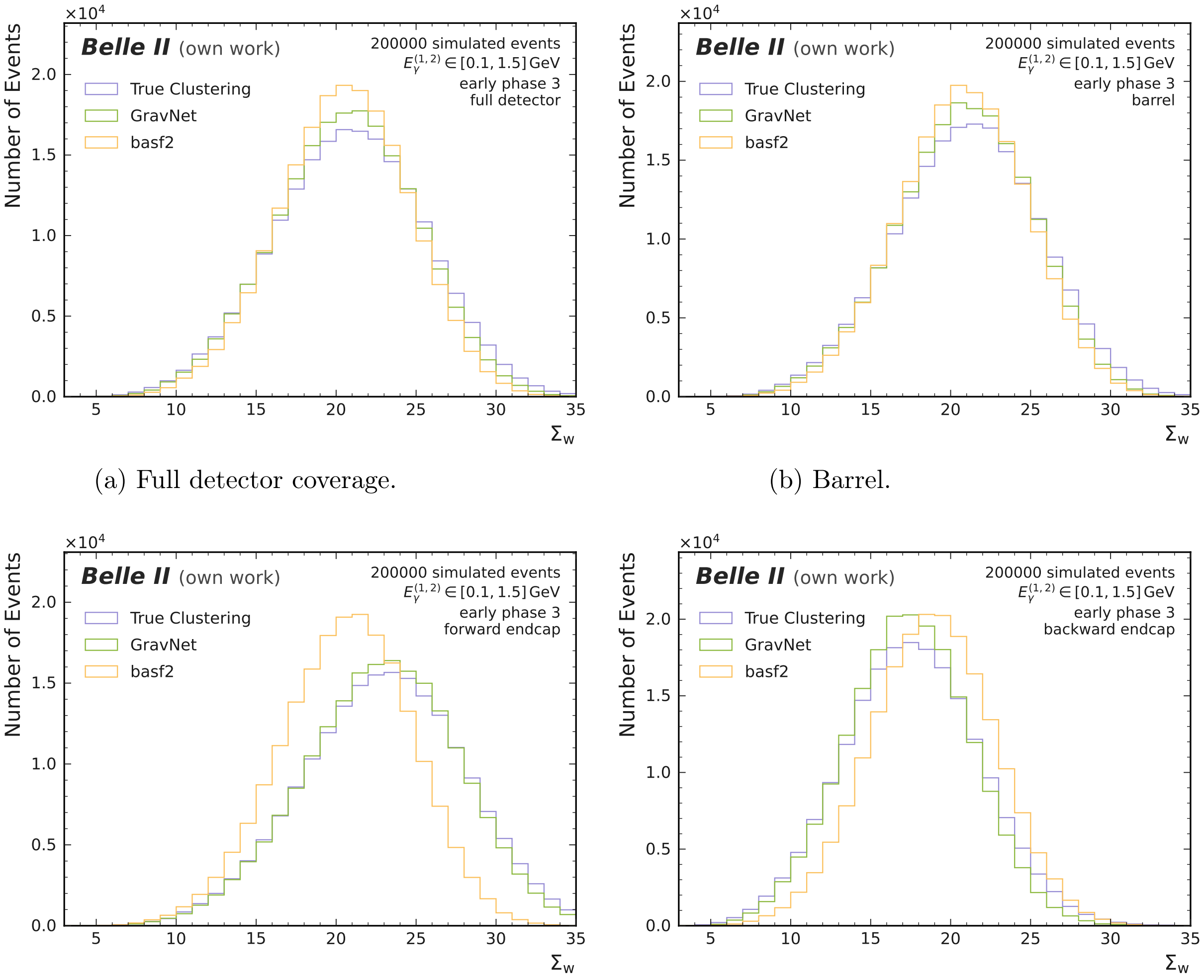


Figure .26.: Distribution in leakage ΔE_{leak} for the two-cluster toy studies with early and nominal phase 3 background. Barrel, forward, and backward endcaps are shown in comparison.

Sum of Weights



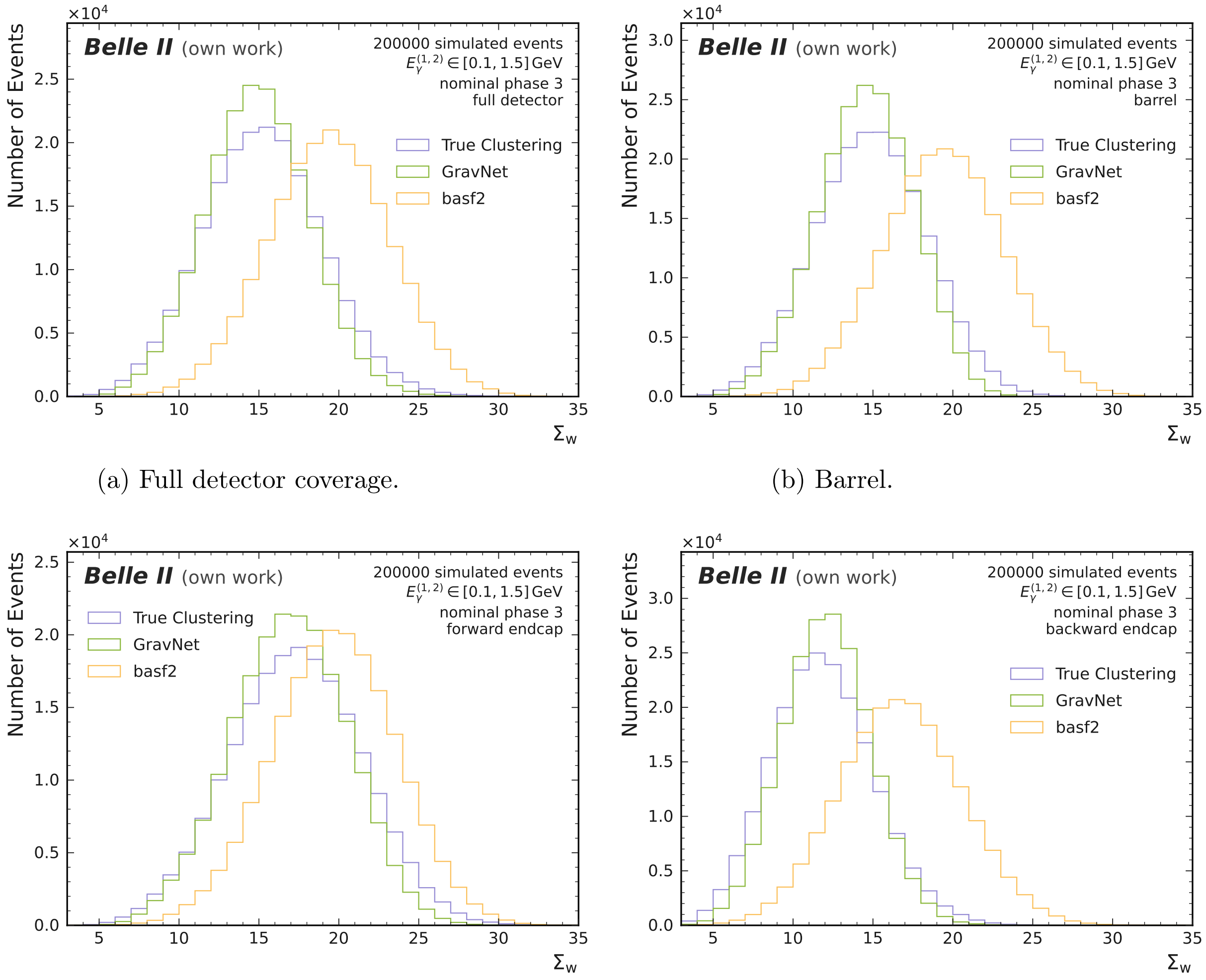
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .27.: Distribution in the sum of weights Σ_w for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

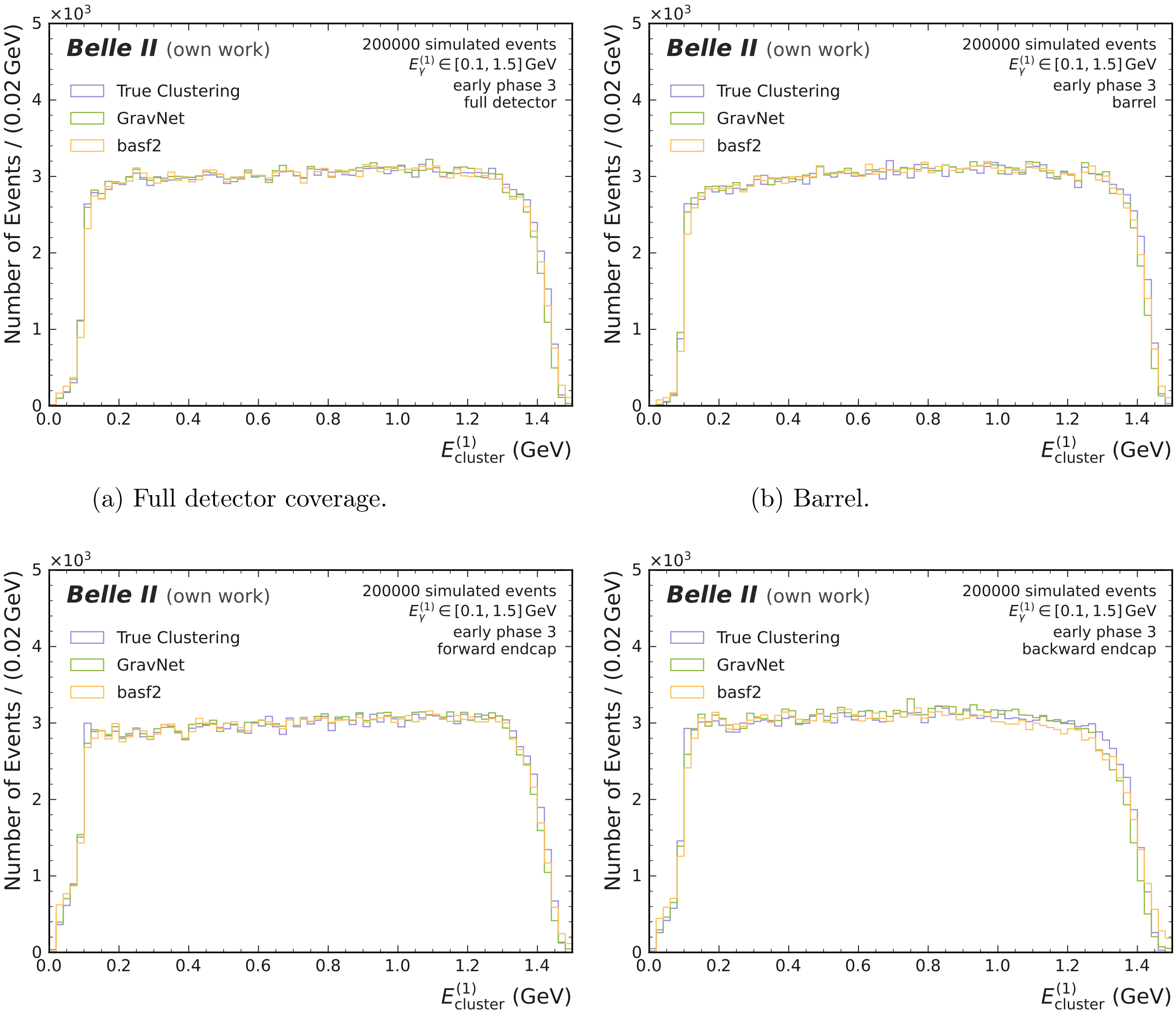
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .28.: Distribution in the sum of weights Σ_w for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Energy



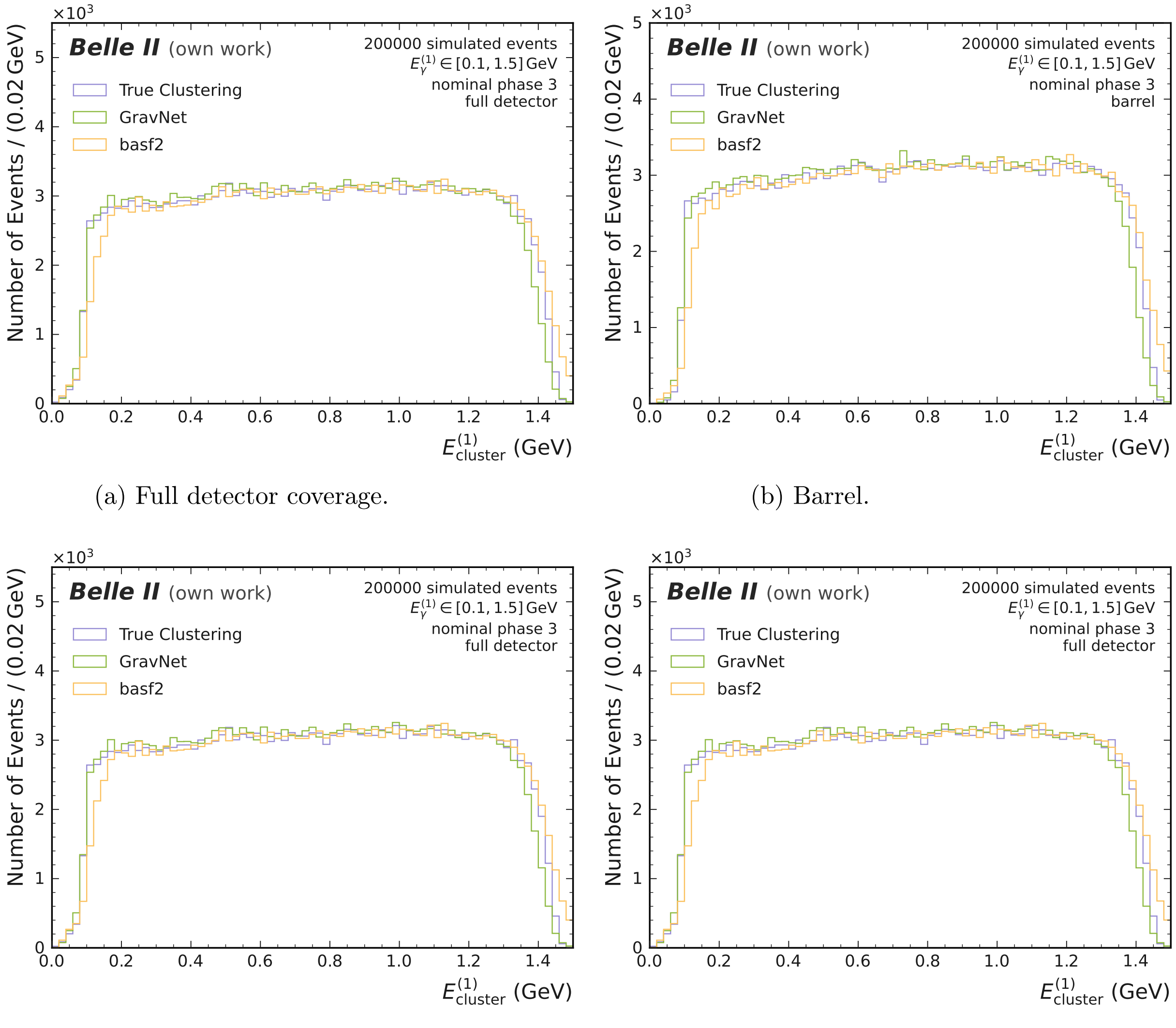
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .29.: Distribution in cluster 1 energies $E_{\text{cluster}}^{(1)}$ for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



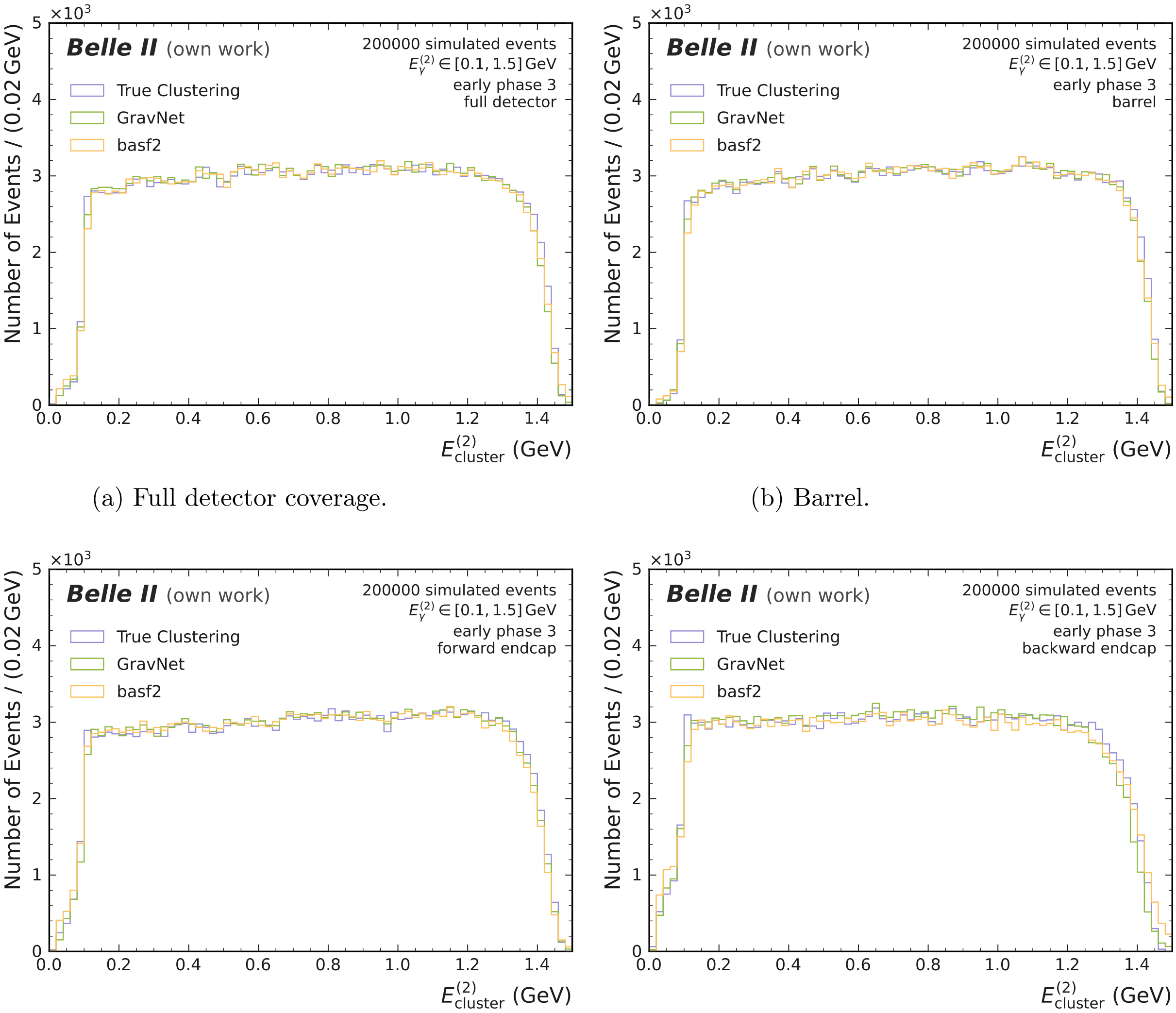
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .30.: Distribution in cluster 1 energies $E_{\text{cluster}}^{(1)}$ for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



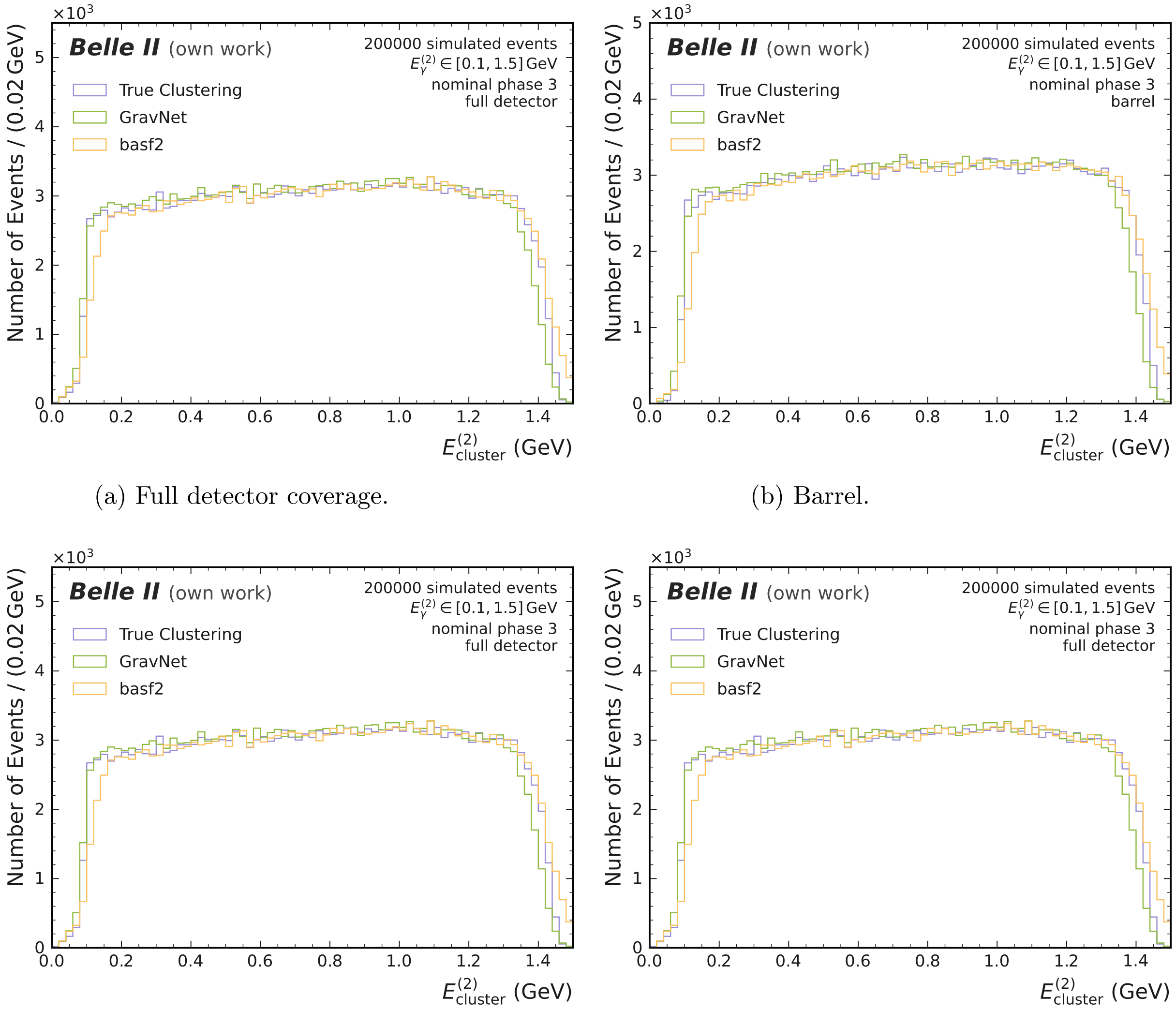
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .31.: Distribution in cluster 2 energies $E_{\mathrm{cluster}}^{(2)}$ for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

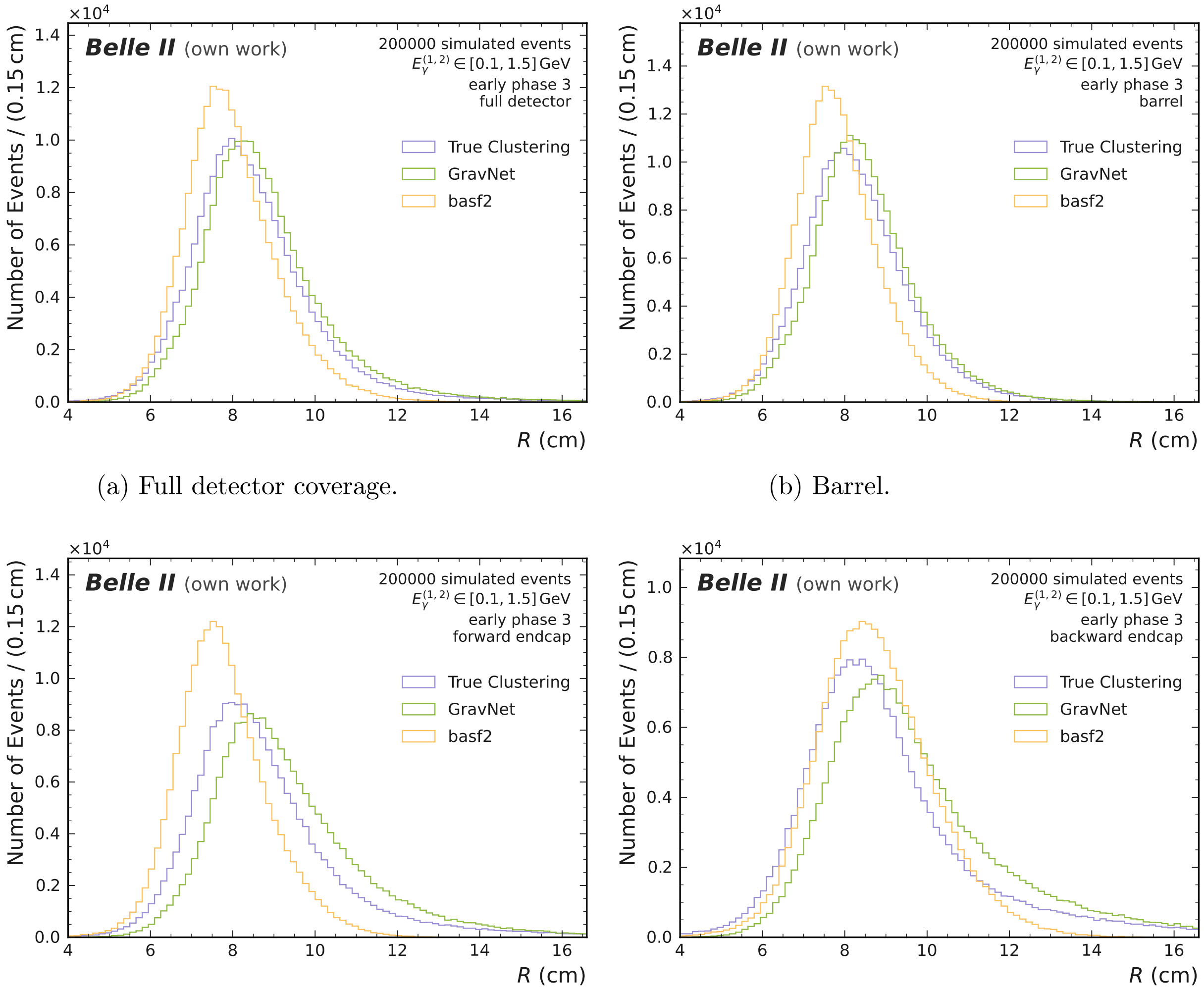
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .32.: Distribution in cluster 2 energies $E_{\text{cluster}}^{2)}$ for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Radius



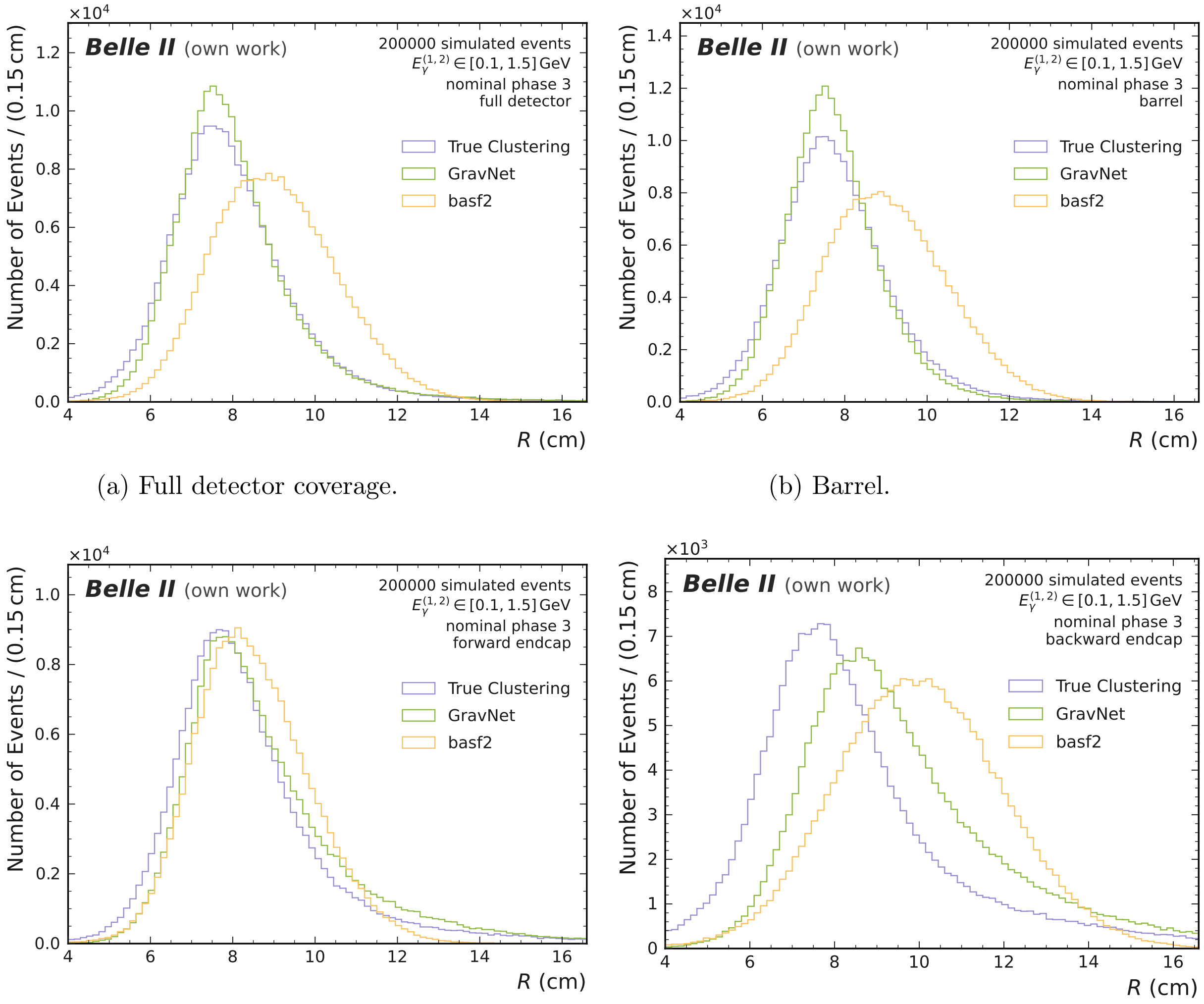
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .33.: Distribution in radii R for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

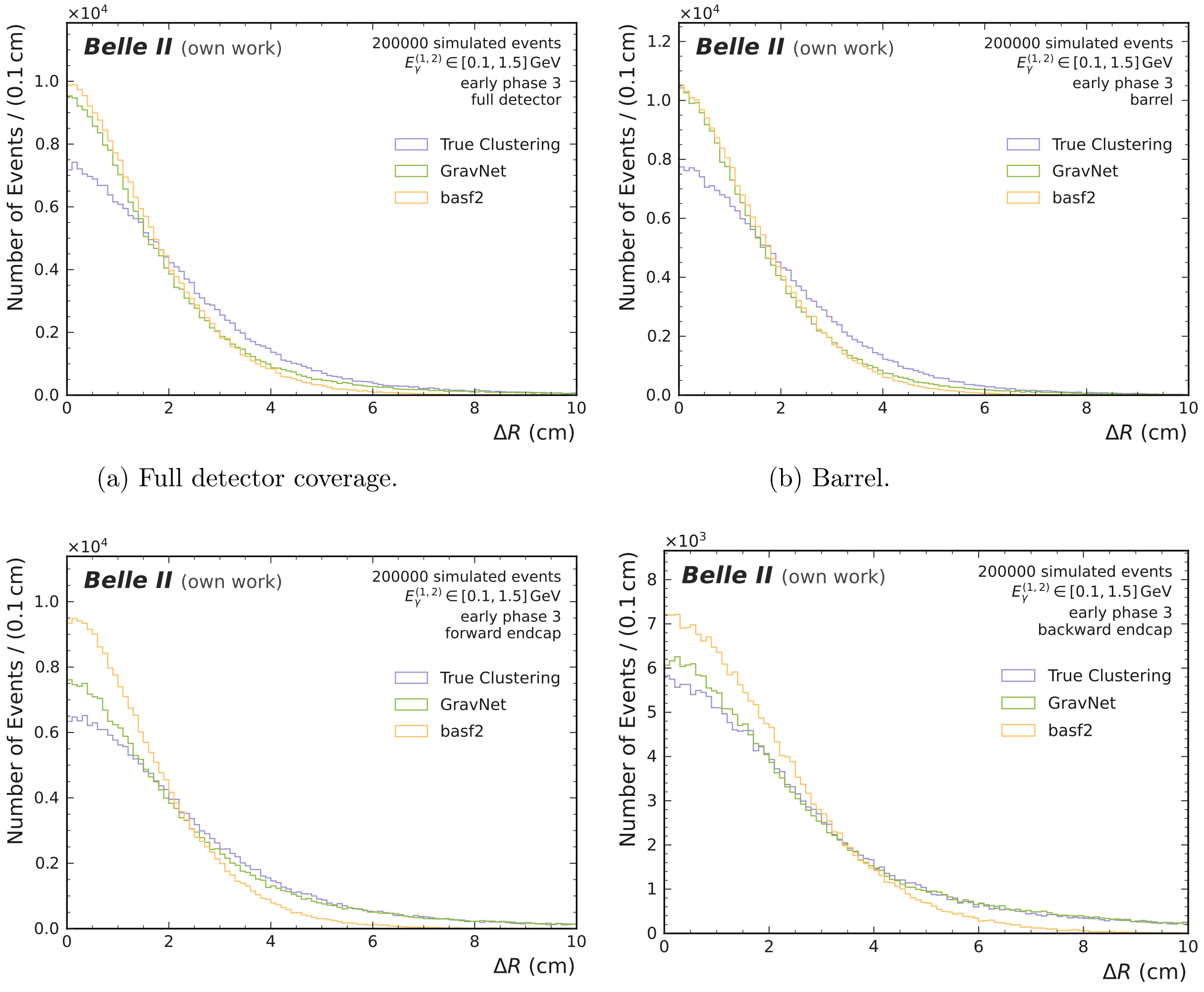
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .34.: Distribution in radii R for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Radius Difference



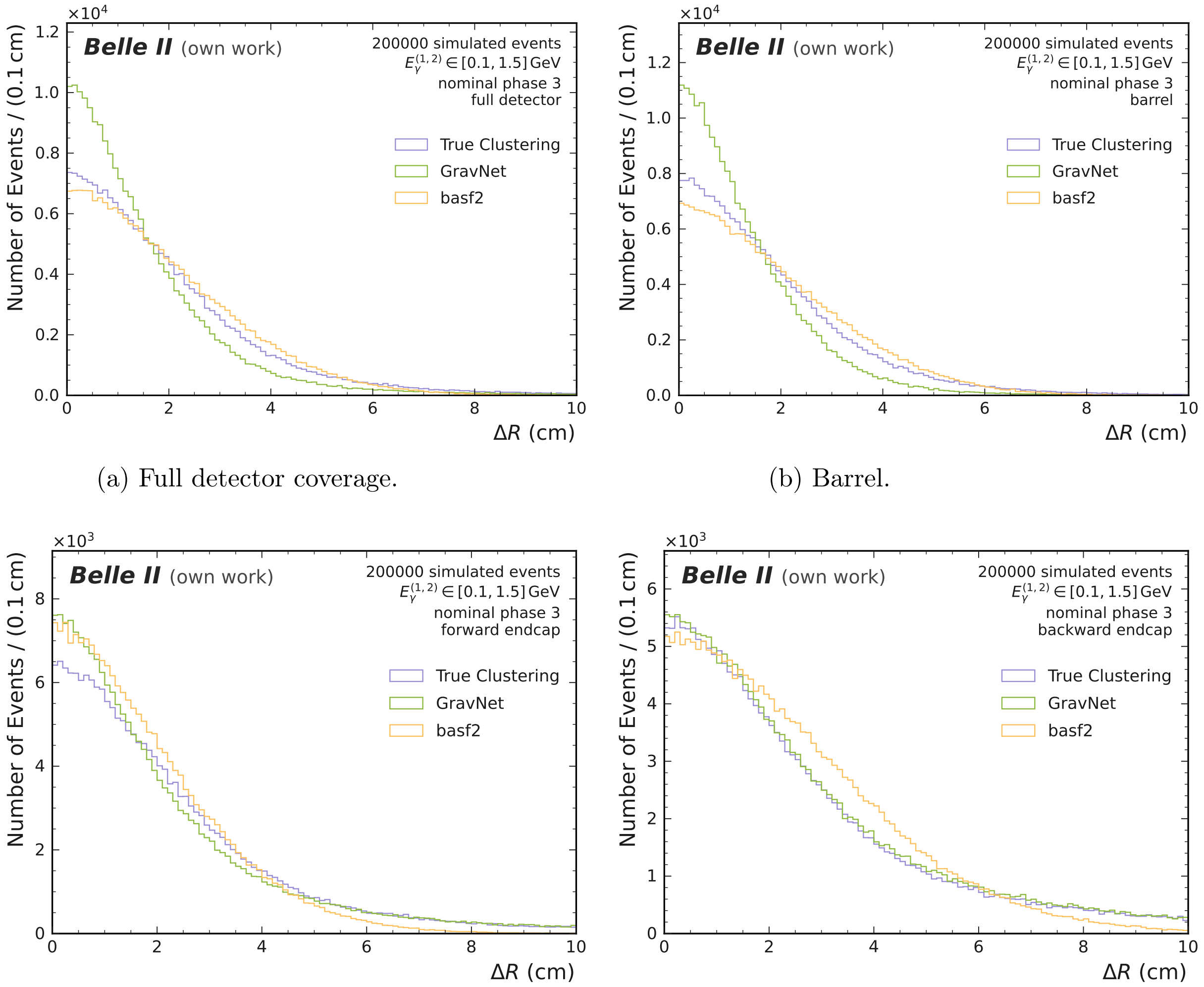
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .35.: Distribution in cluster radius differences ΔR for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

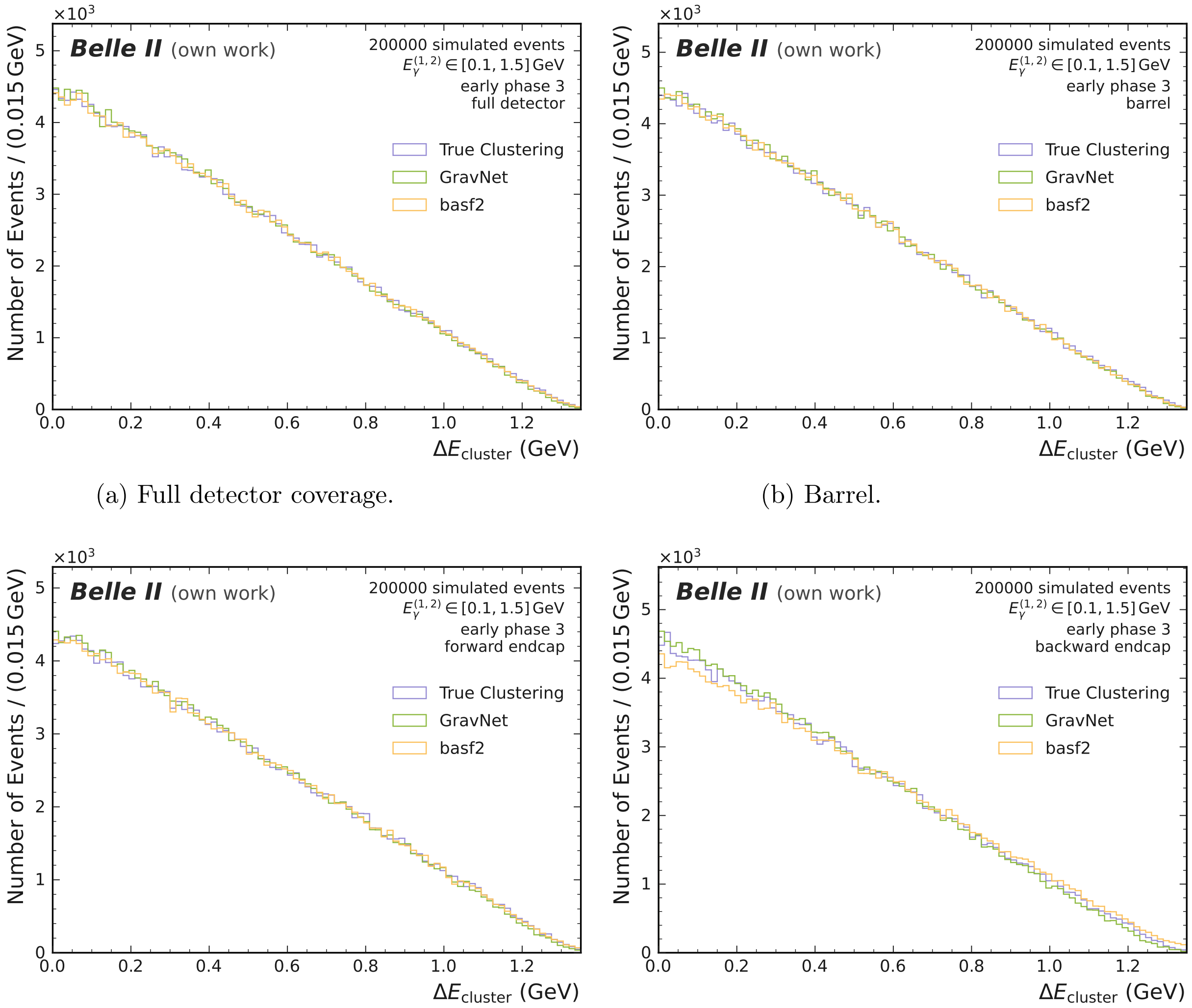
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .36.: Distribution in cluster radius differences ΔR for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Energy Difference



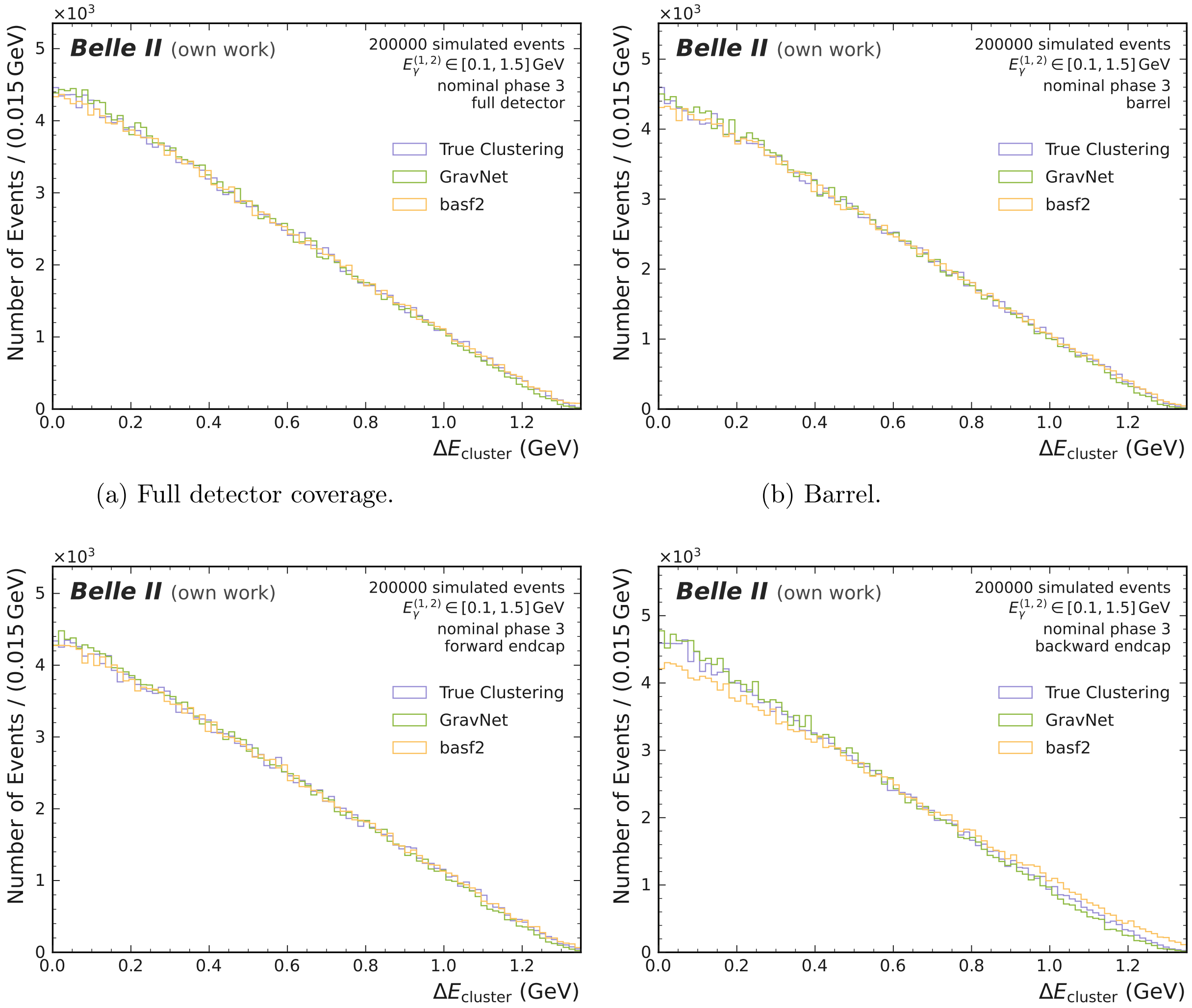
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .37.: Distribution in energy differences ΔE for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

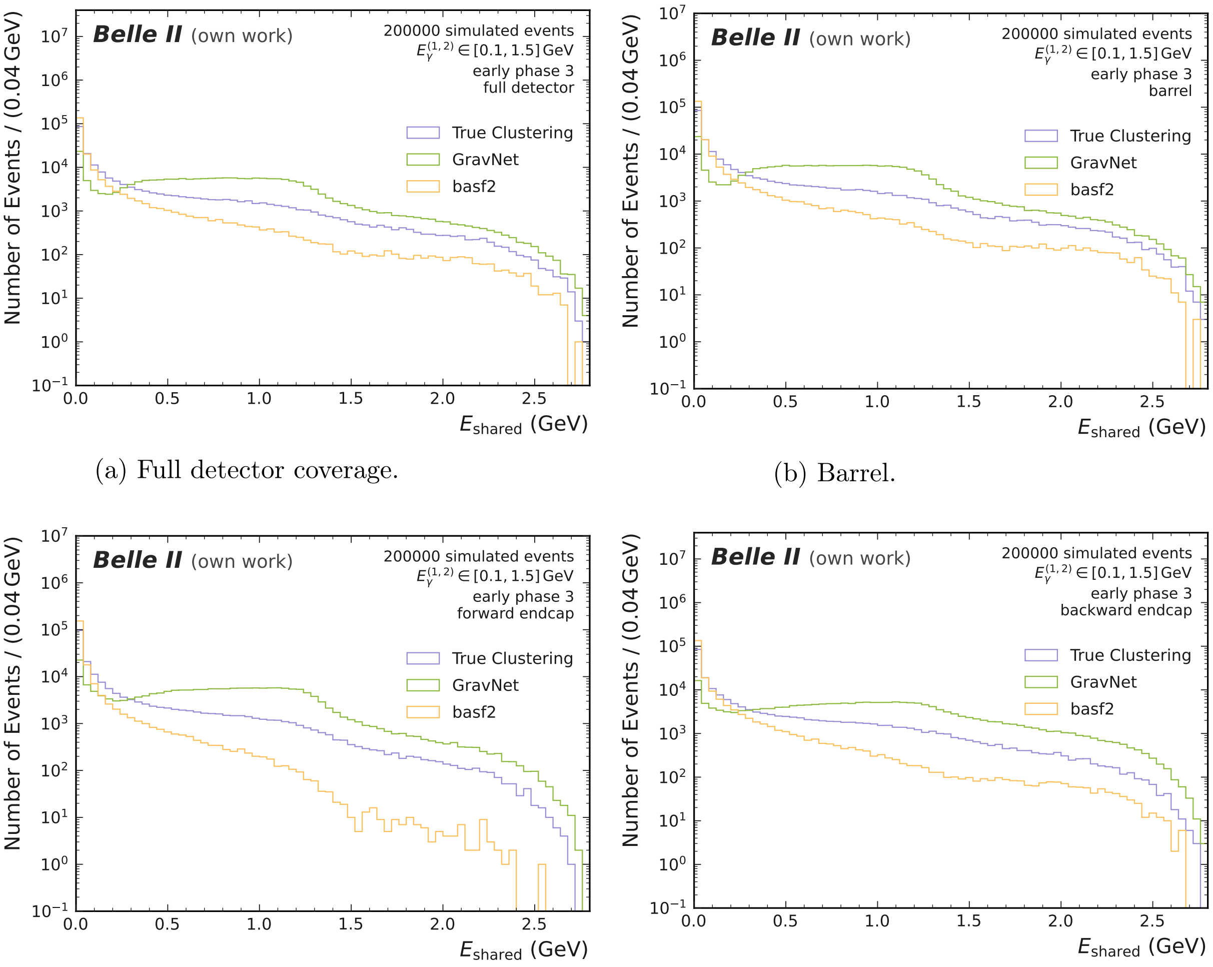
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .38.: Distribution in energy differences ΔE for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Shared Energy



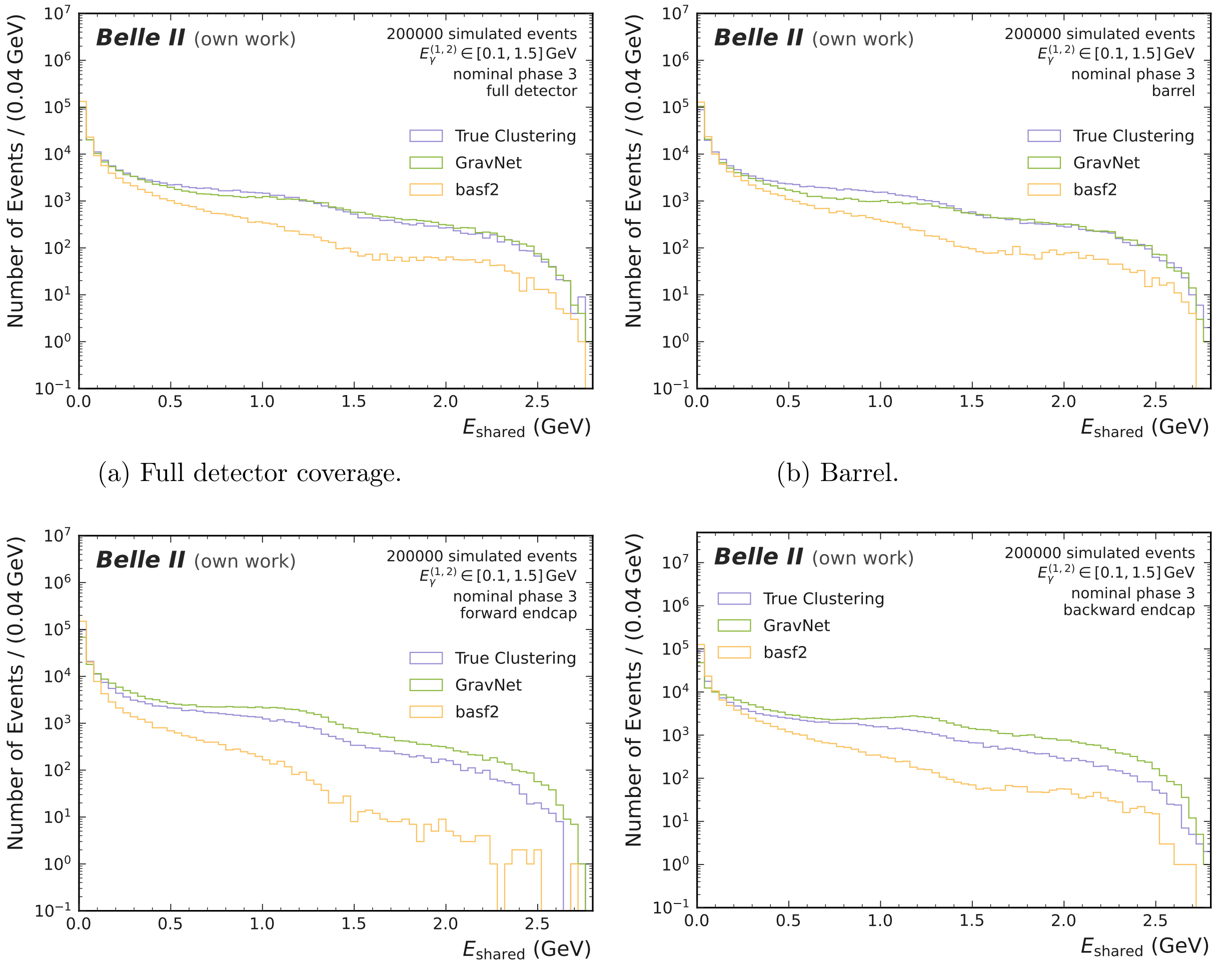
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .39.: Distribution in shared energies E_{shared} for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

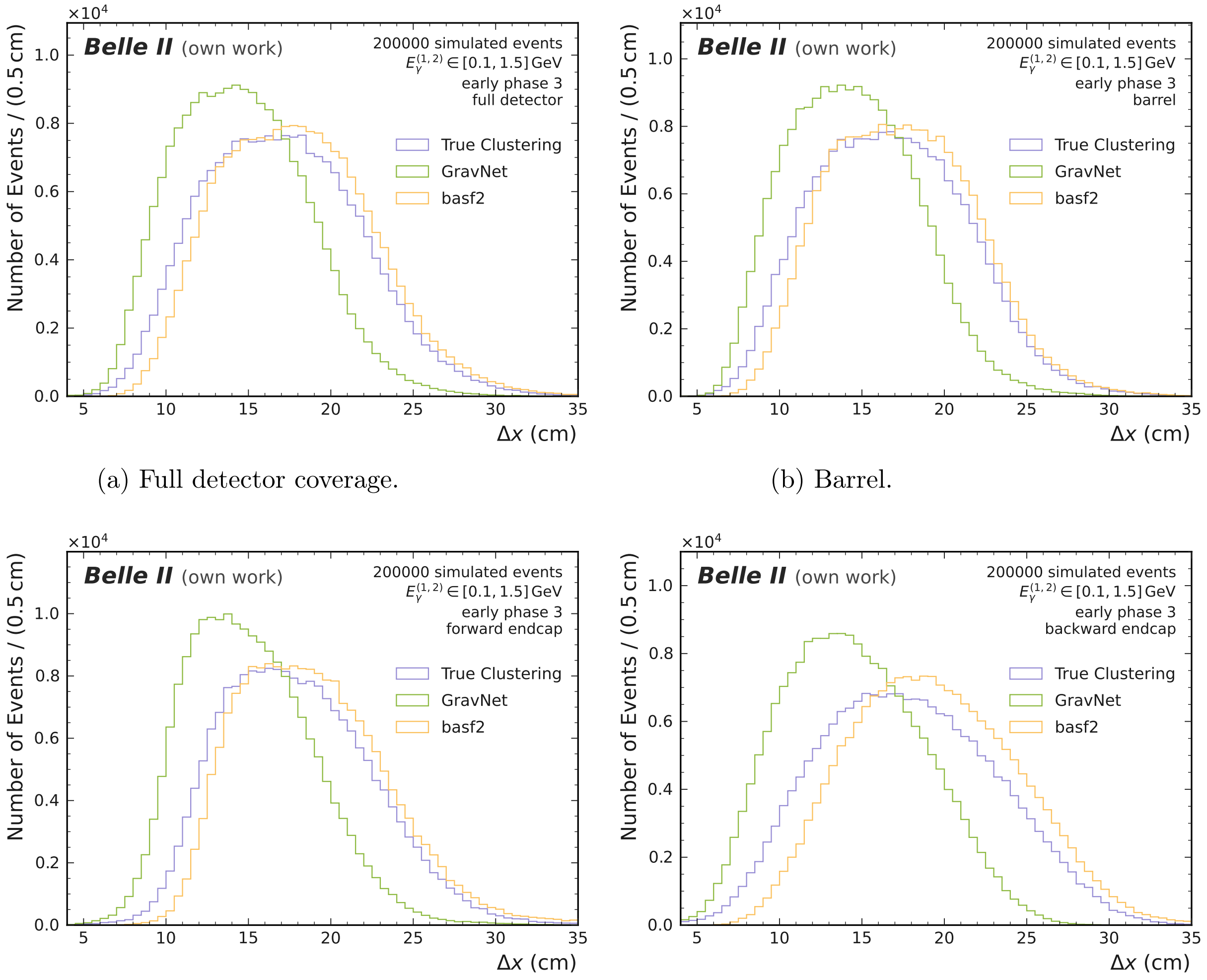
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .40.: Distribution in shared energies E_{shared} for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

Cluster Center Distance



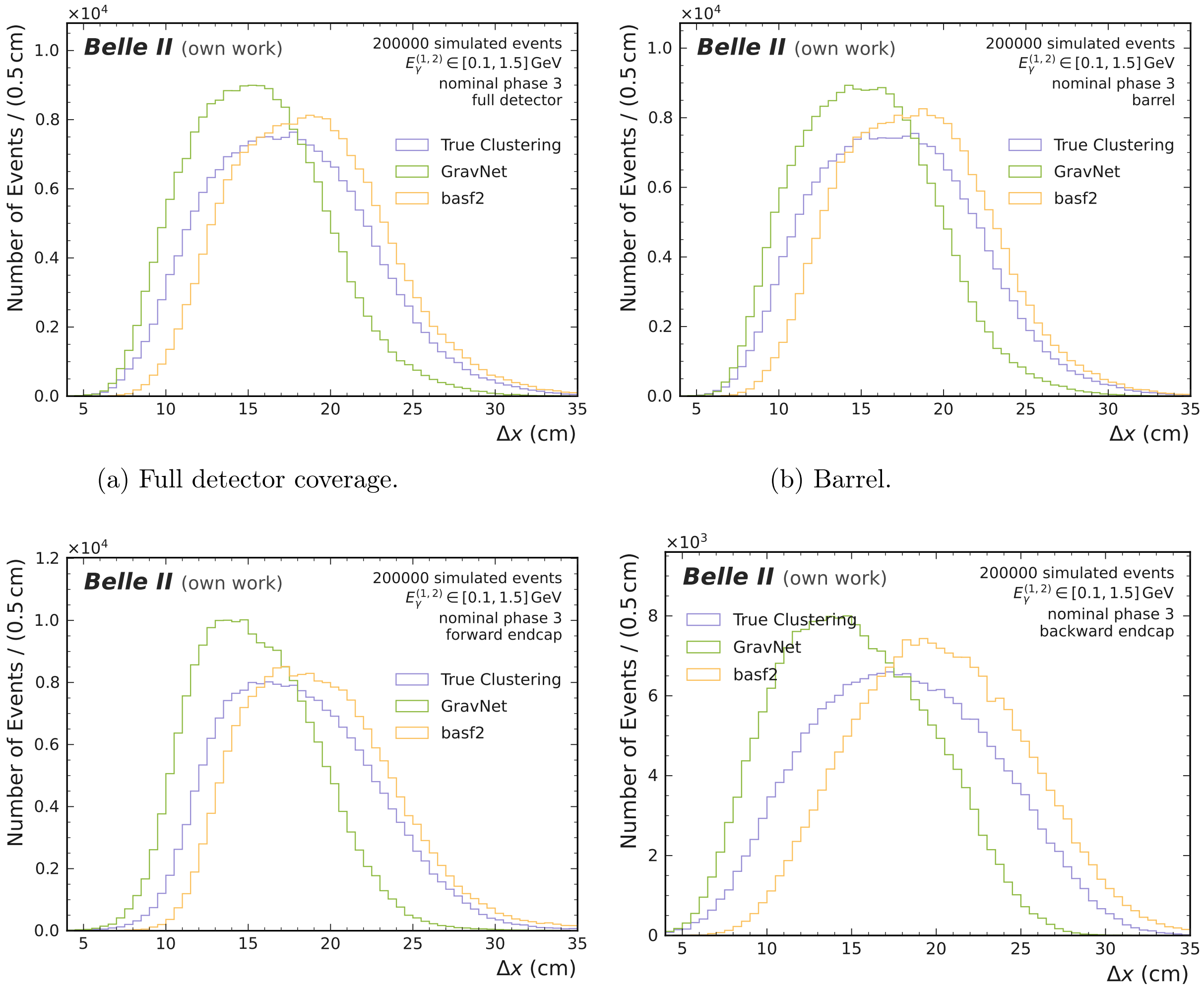
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .41.: Distribution in cluster center distances Δx for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.



(a) Full detector coverage.

(b) Barrel.

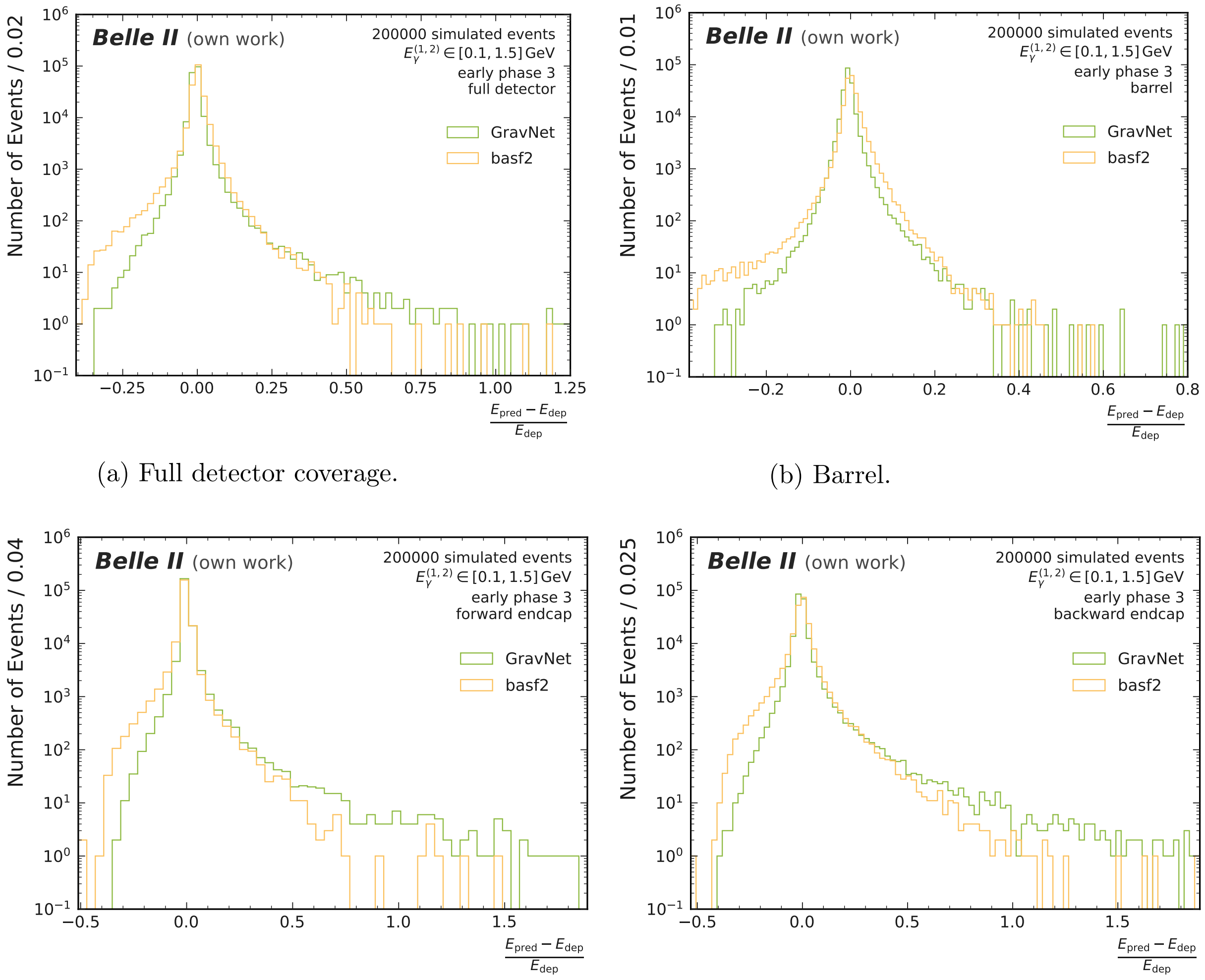
(c) Forward endcap.

(d) Backward endcap.

Figure .42.: Distribution in cluster center distances Δx for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares the true clustering, GravNet and the basf2 baseline.

D.2. Performance Evaluation

Energy Resolution



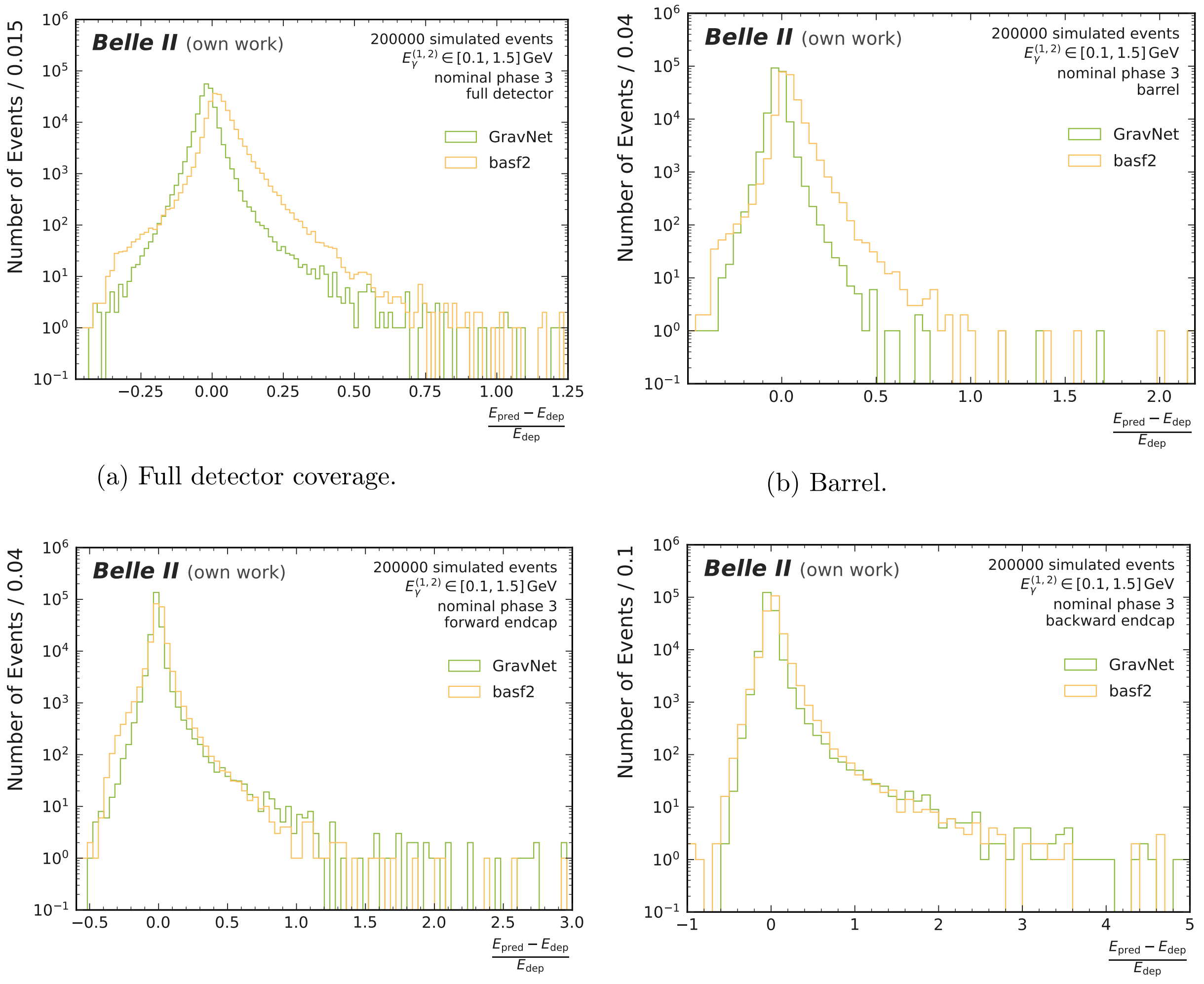
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .43.: Distribution in deposited errors η_dep for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



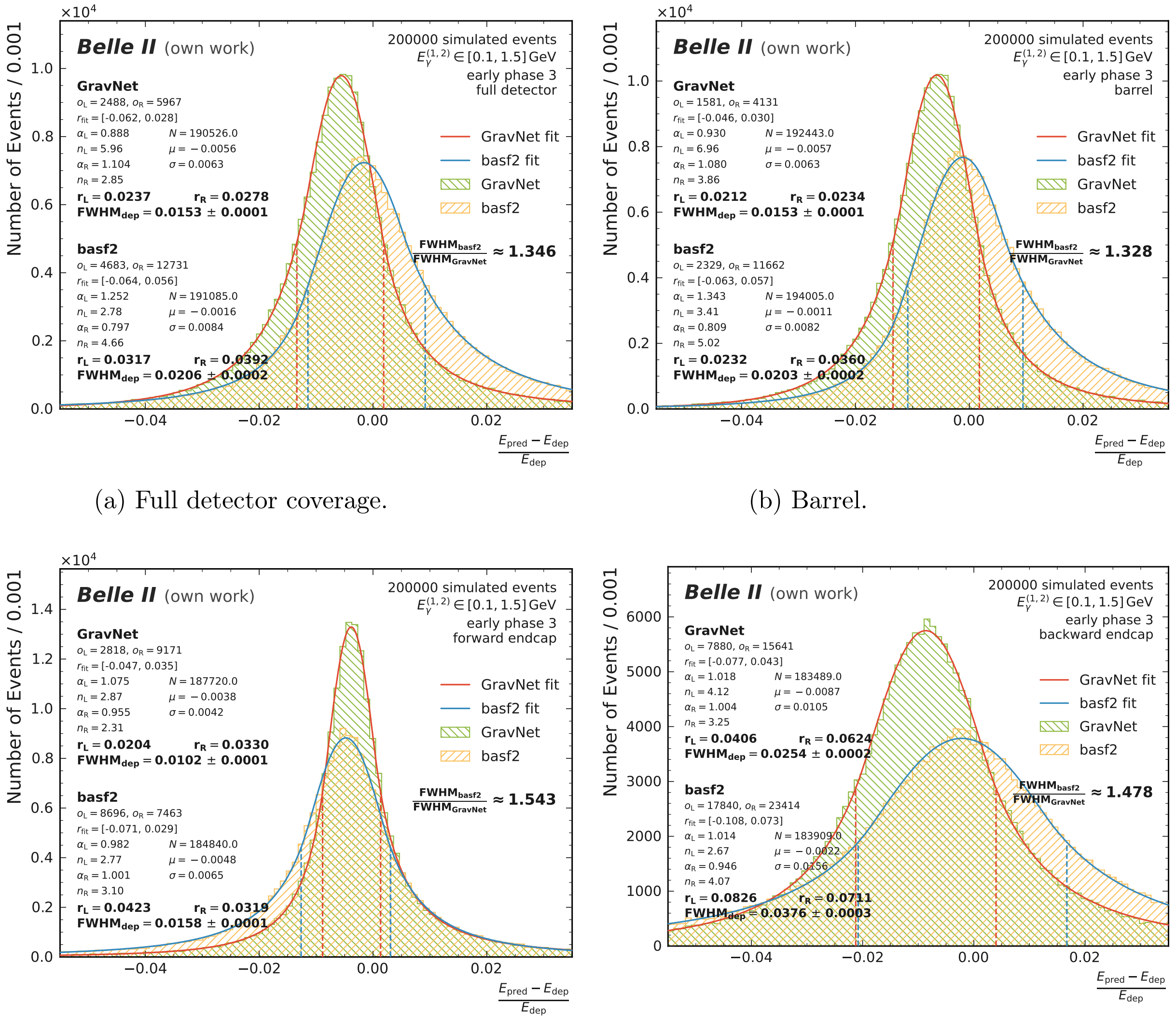
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .44.: Distribution in deposited errors η_dep for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



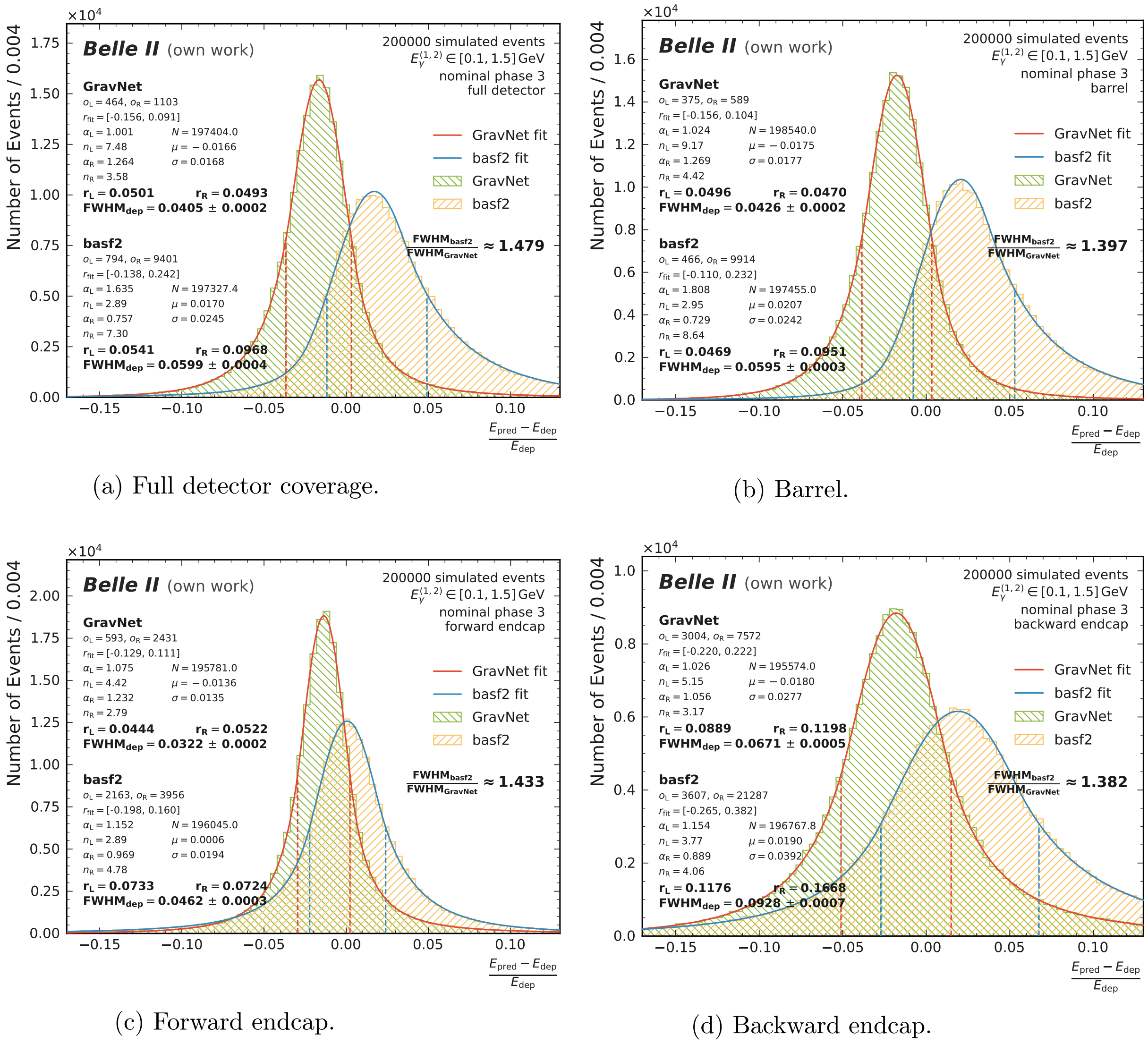
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .45.: Fit for the distribution in deposited errors η_{dep} for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



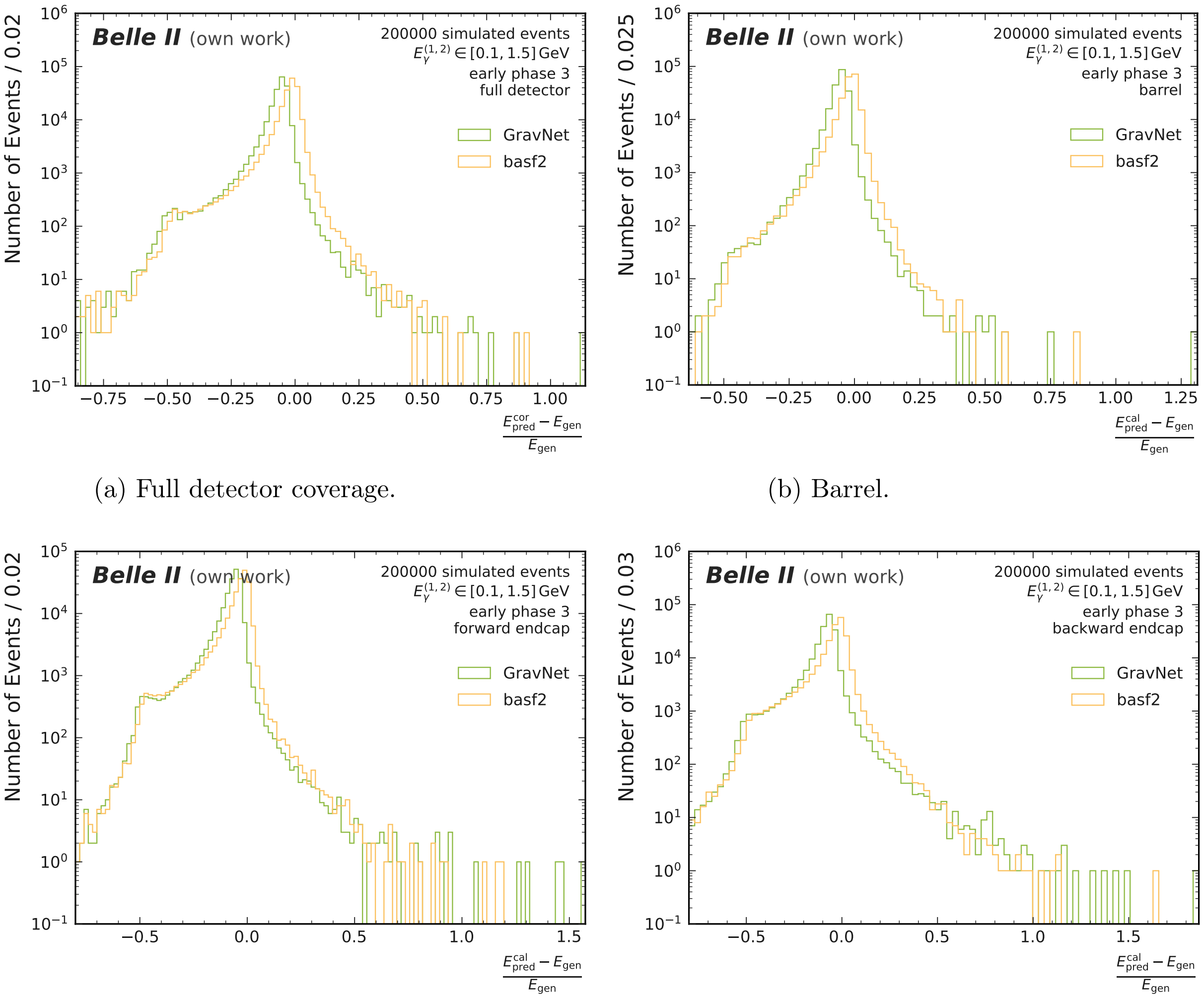
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .46.: Fit for the distribution in deposited errors η_{dep} for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



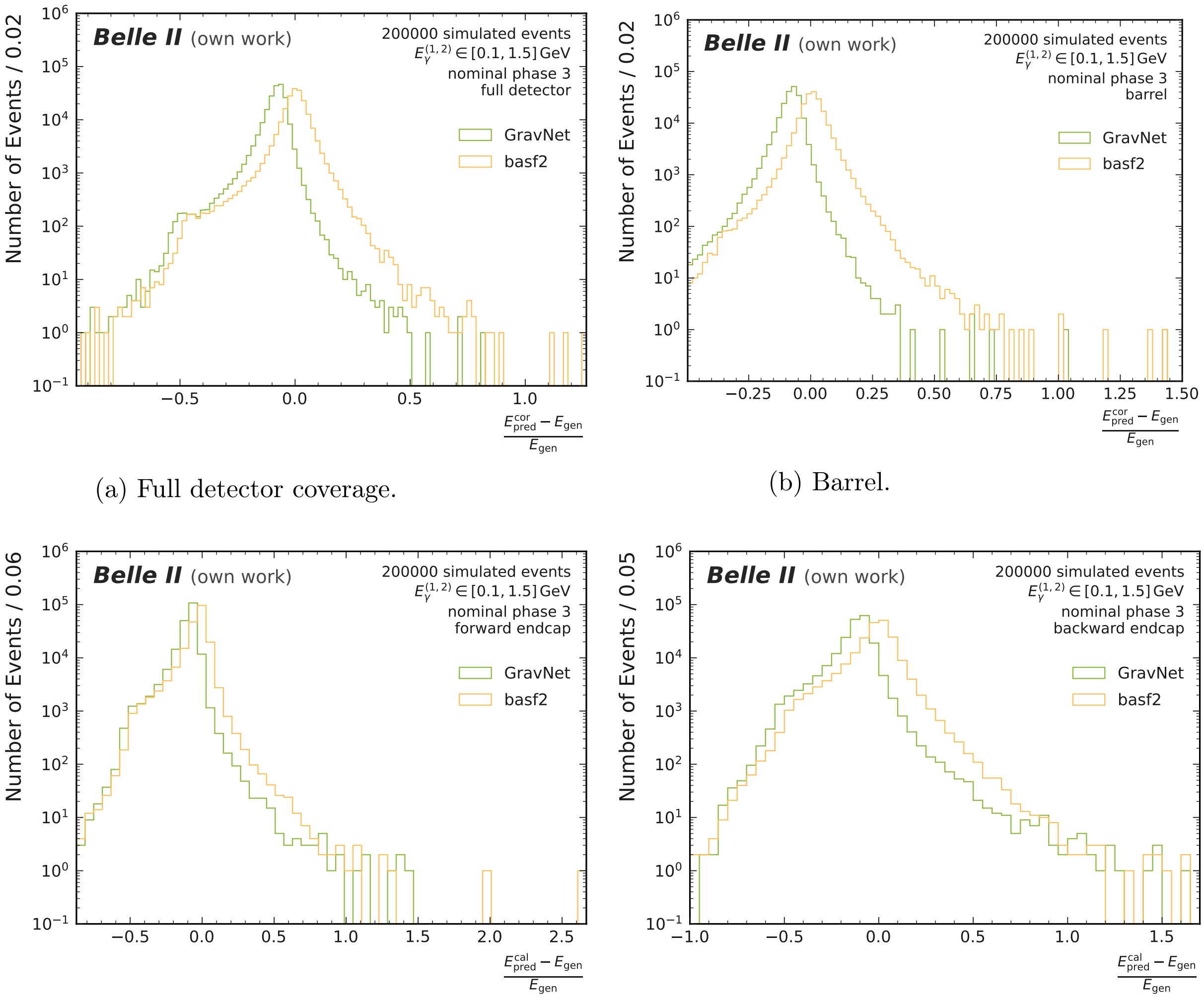
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .47.: Distribution in generated errors η_{gen} for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



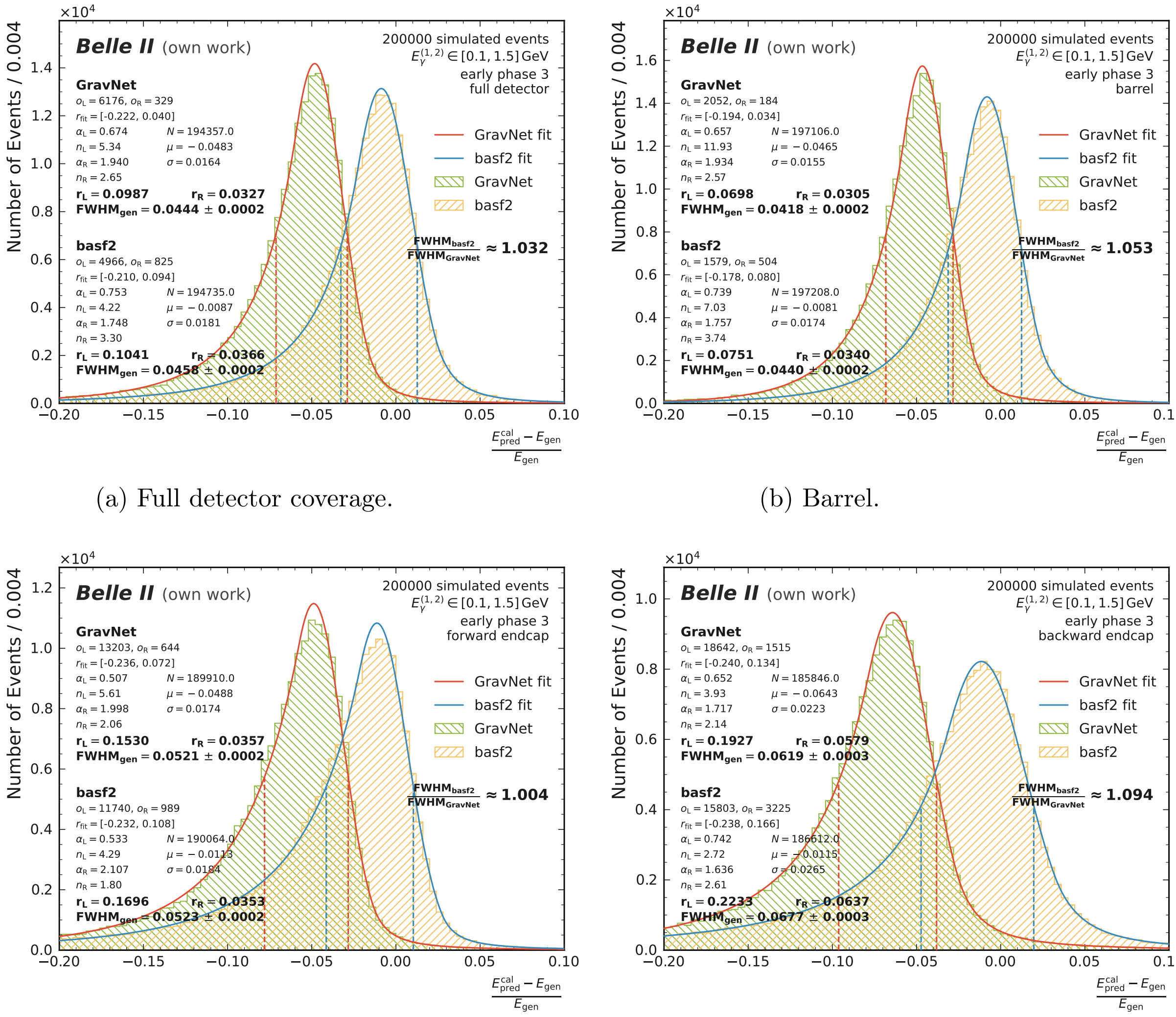
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .48.: Distribution in generated errors η_{gen} for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



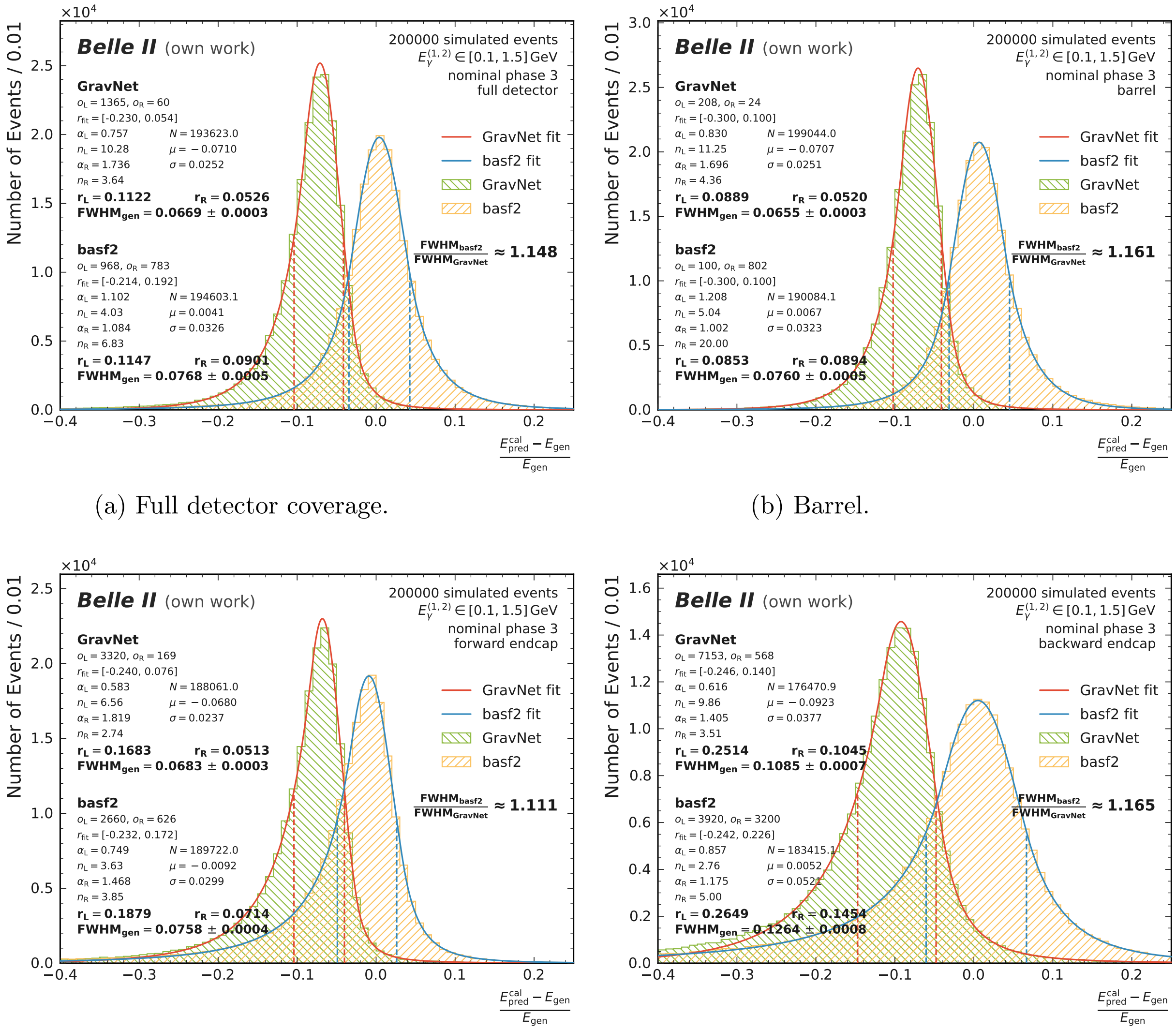
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .49.: Fit for the distribution in generated errors η_{gen} for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

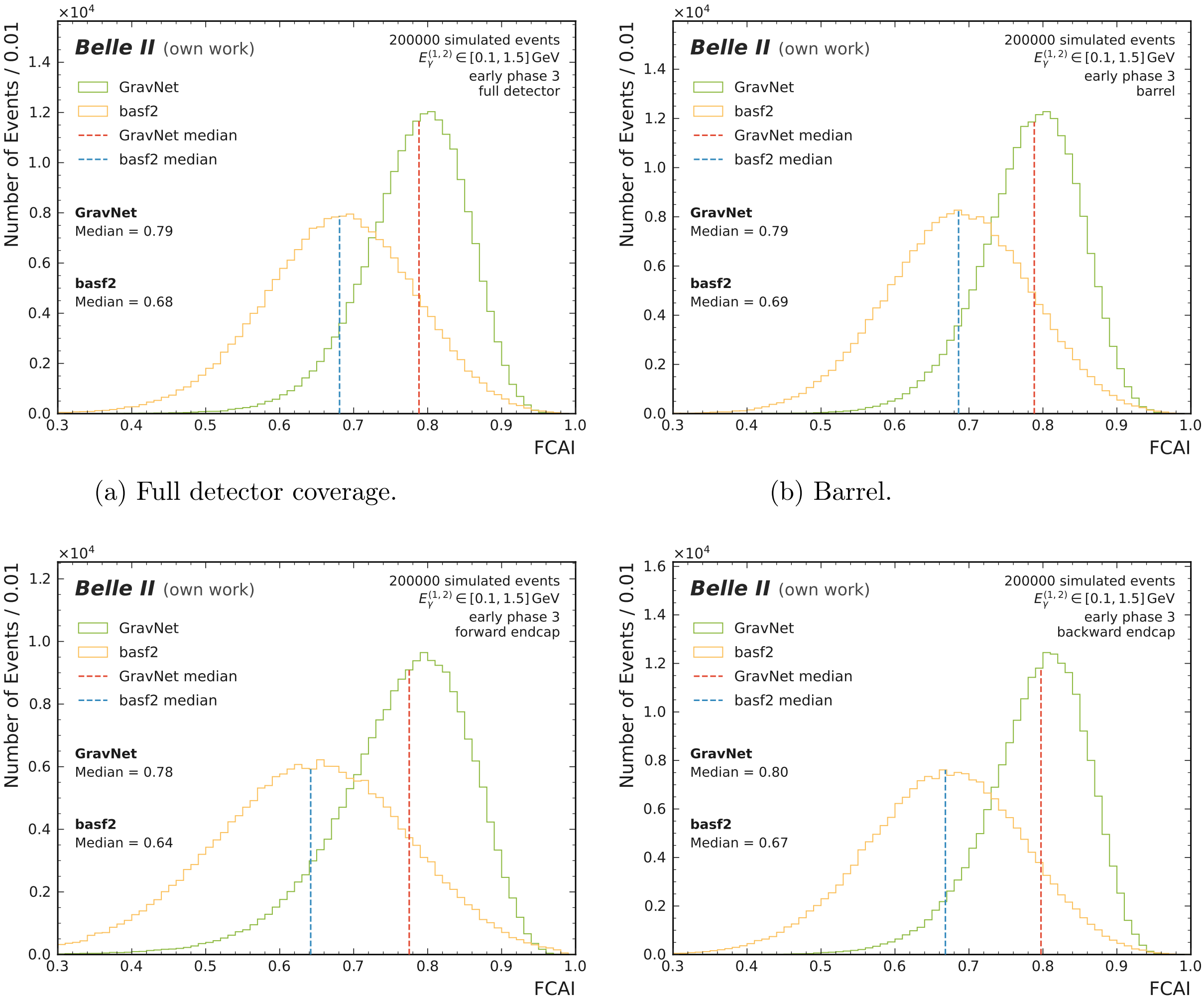
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .50.: Fit for the distribution in generated errors η_{gen} for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

Fuzzy Clustering Agreement Index



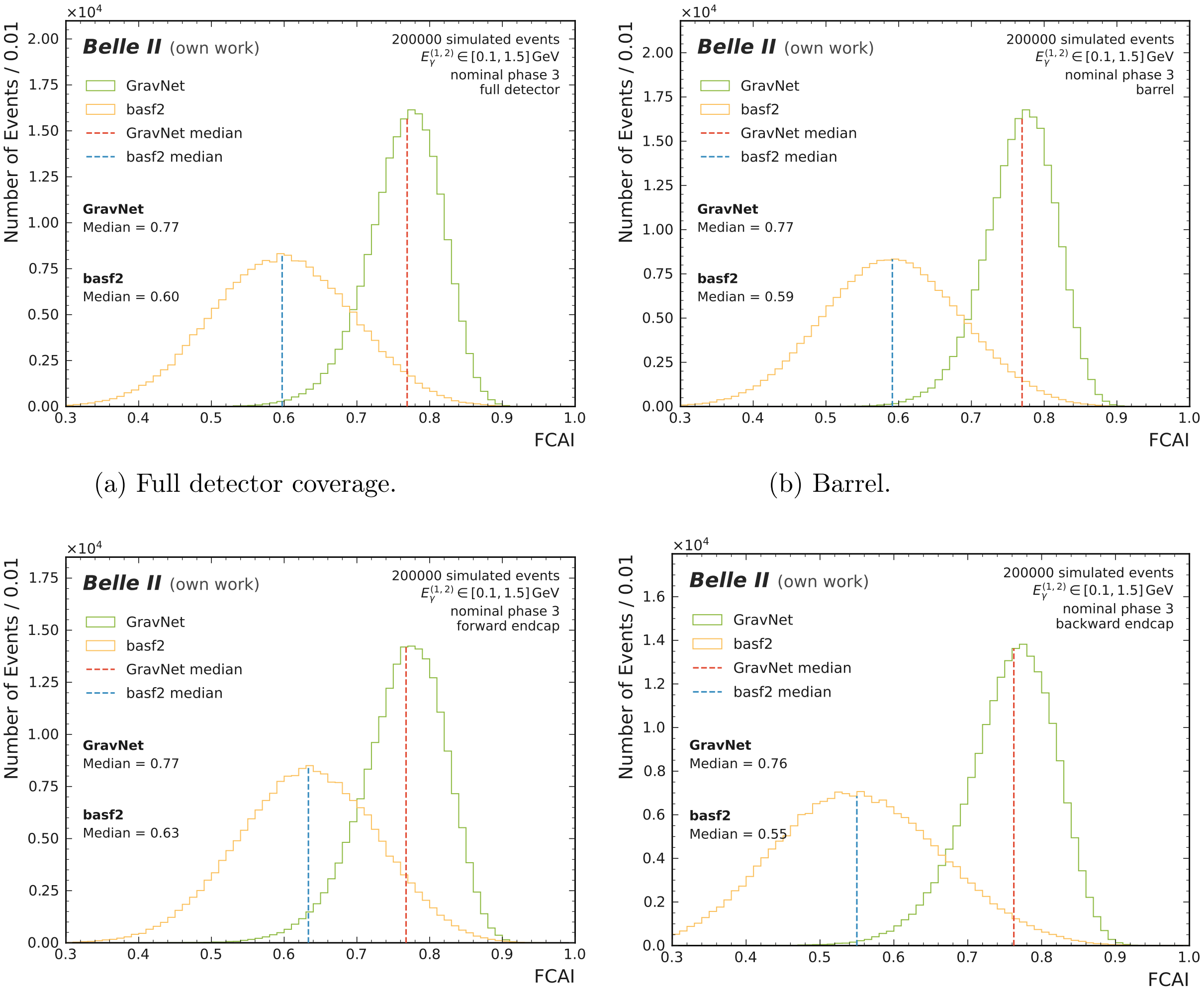
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .51.: Distribution in FCAI for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

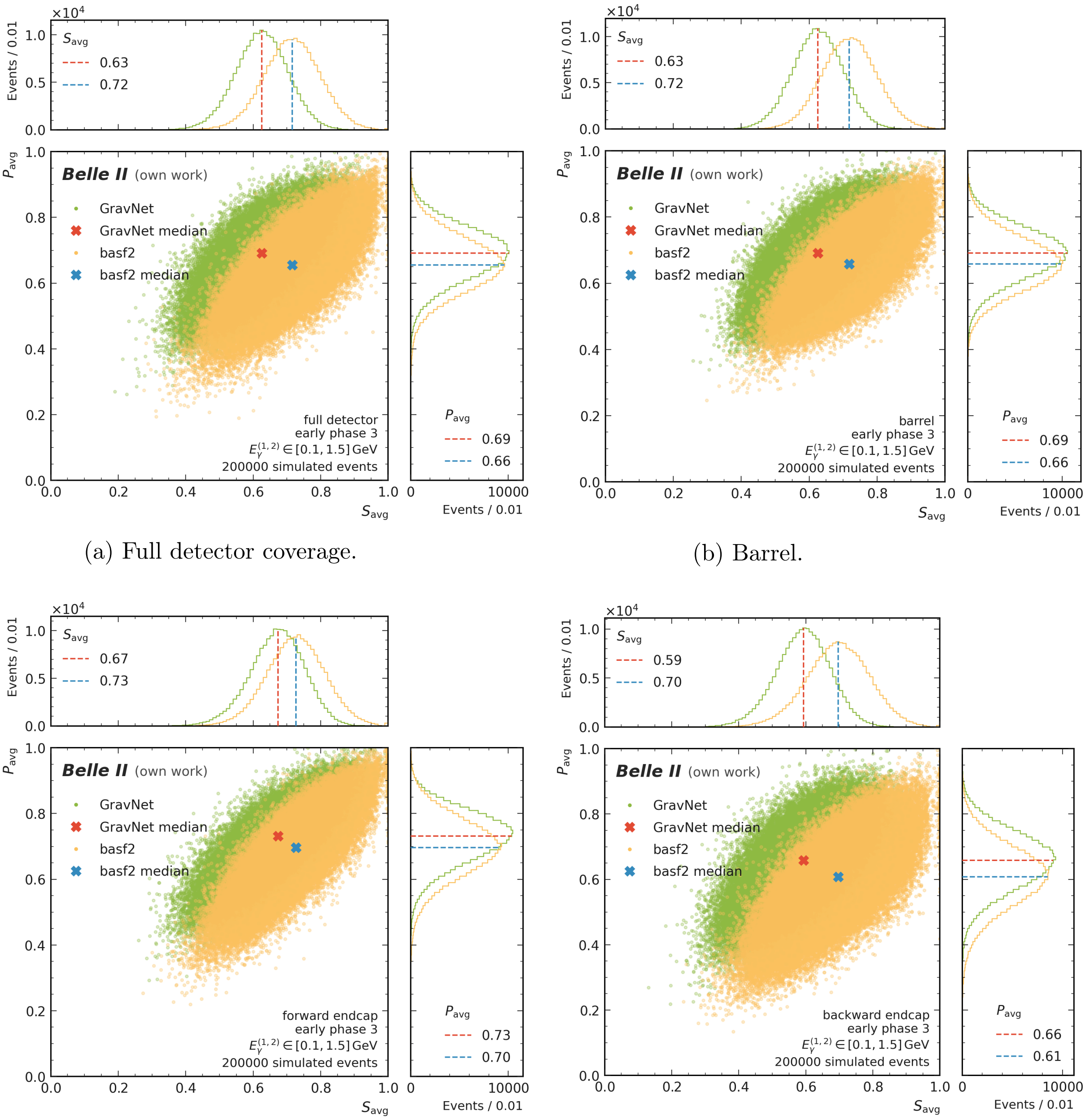
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .52.: Distribution in FCAI for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

Sensitivity and Precision



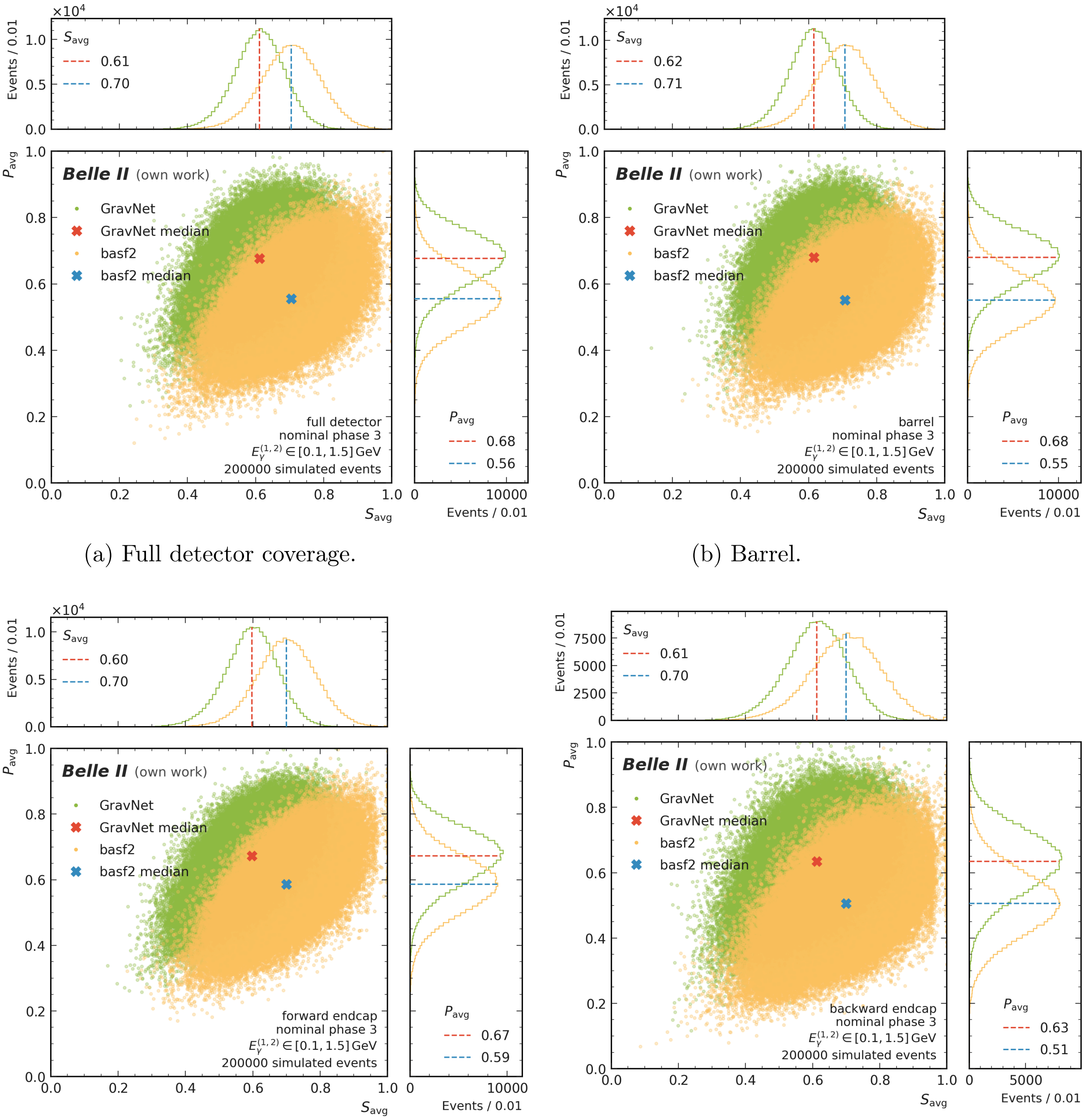
(a) Full detector coverage.

(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .53.: Distribution in average sensitivity S_{avg} and precision P_{avg} for the two-cluster toy study with early phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.



(a) Full detector coverage.

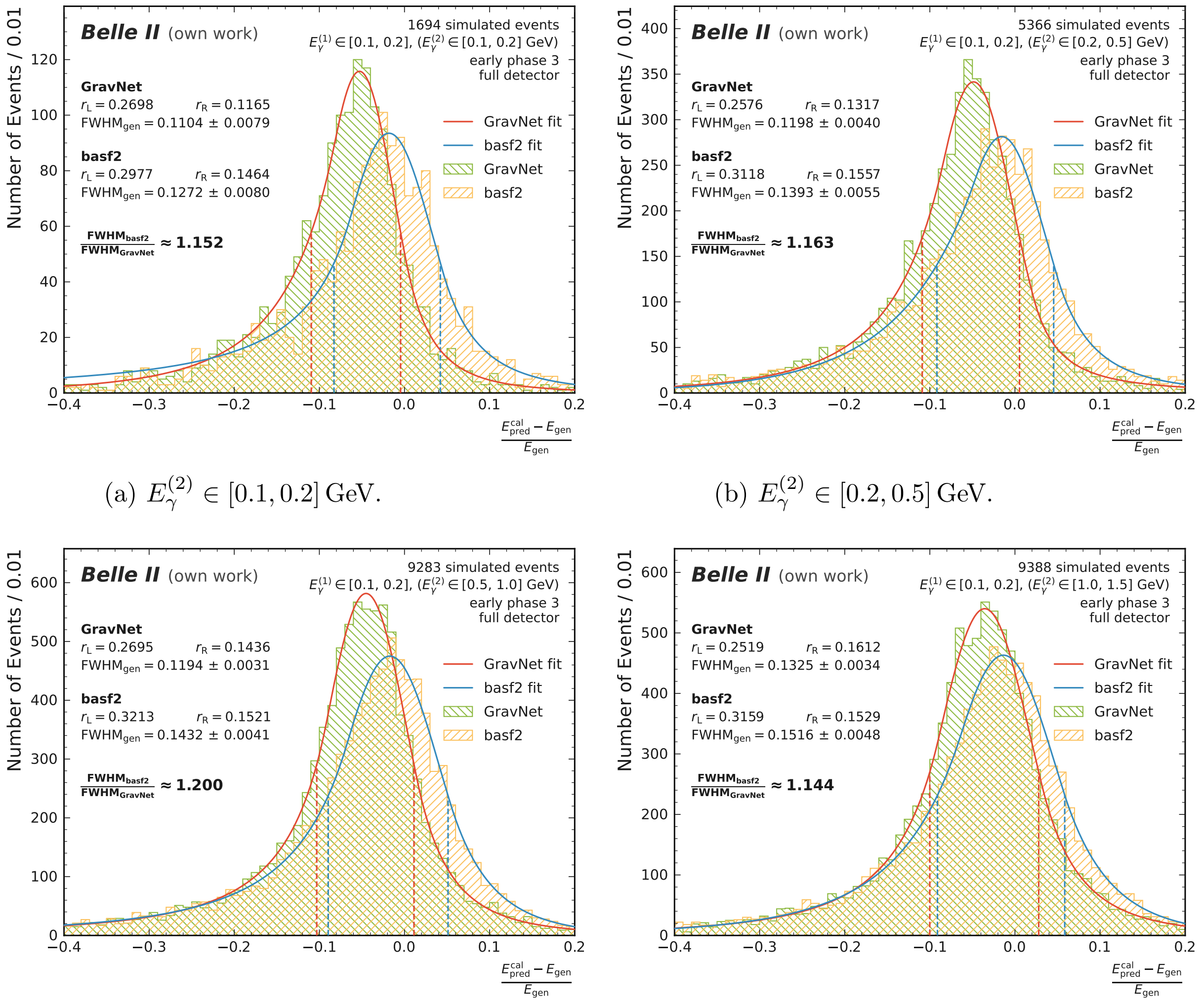
(b) Barrel.

(c) Forward endcap.

(d) Backward endcap.

Figure .54.: Distribution in average sensitivity S_{avg} and precision P_{avg} for the two-cluster toy study with nominal phase 3 background. Full detector, barrel, forward, and backward endcaps are displayed separately. Each plot compares GravNet to the basf2 baseline.

D.3. Energy Dependence



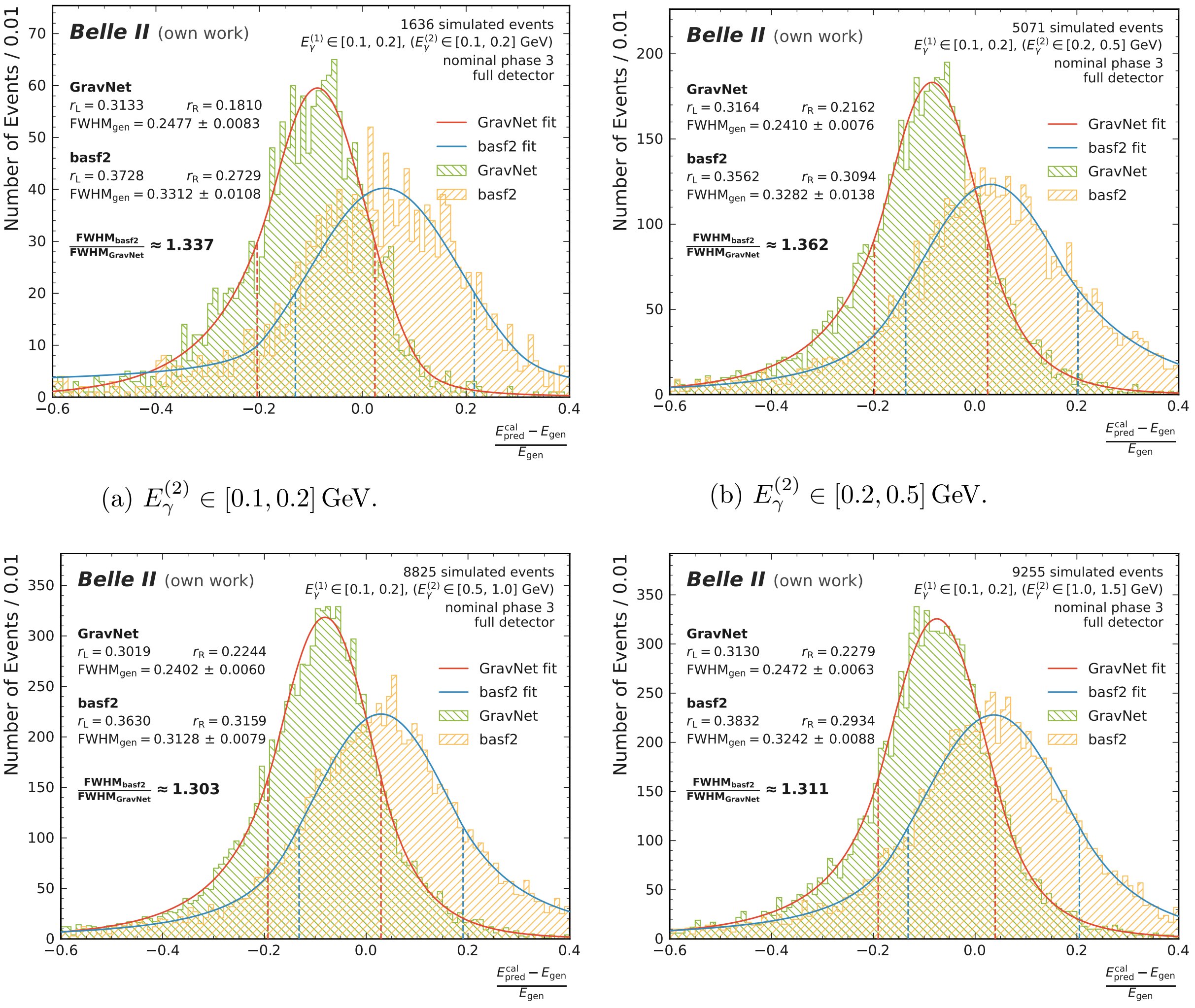
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .55.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.1, 0.2]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with early phase 3 background in full detector coverage.



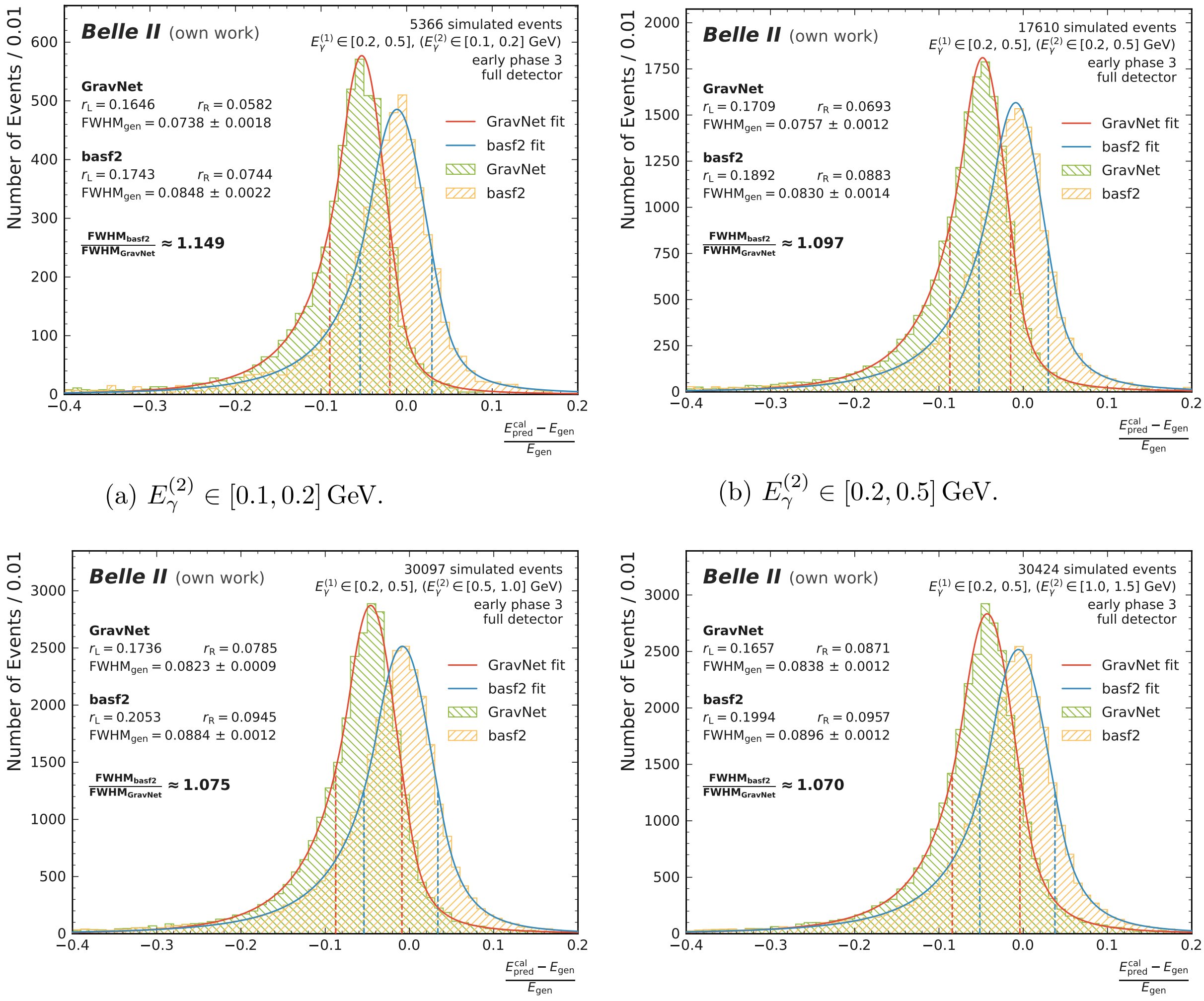
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .56.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.1, 0.2]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with nominal phase 3 background in full detector coverage.



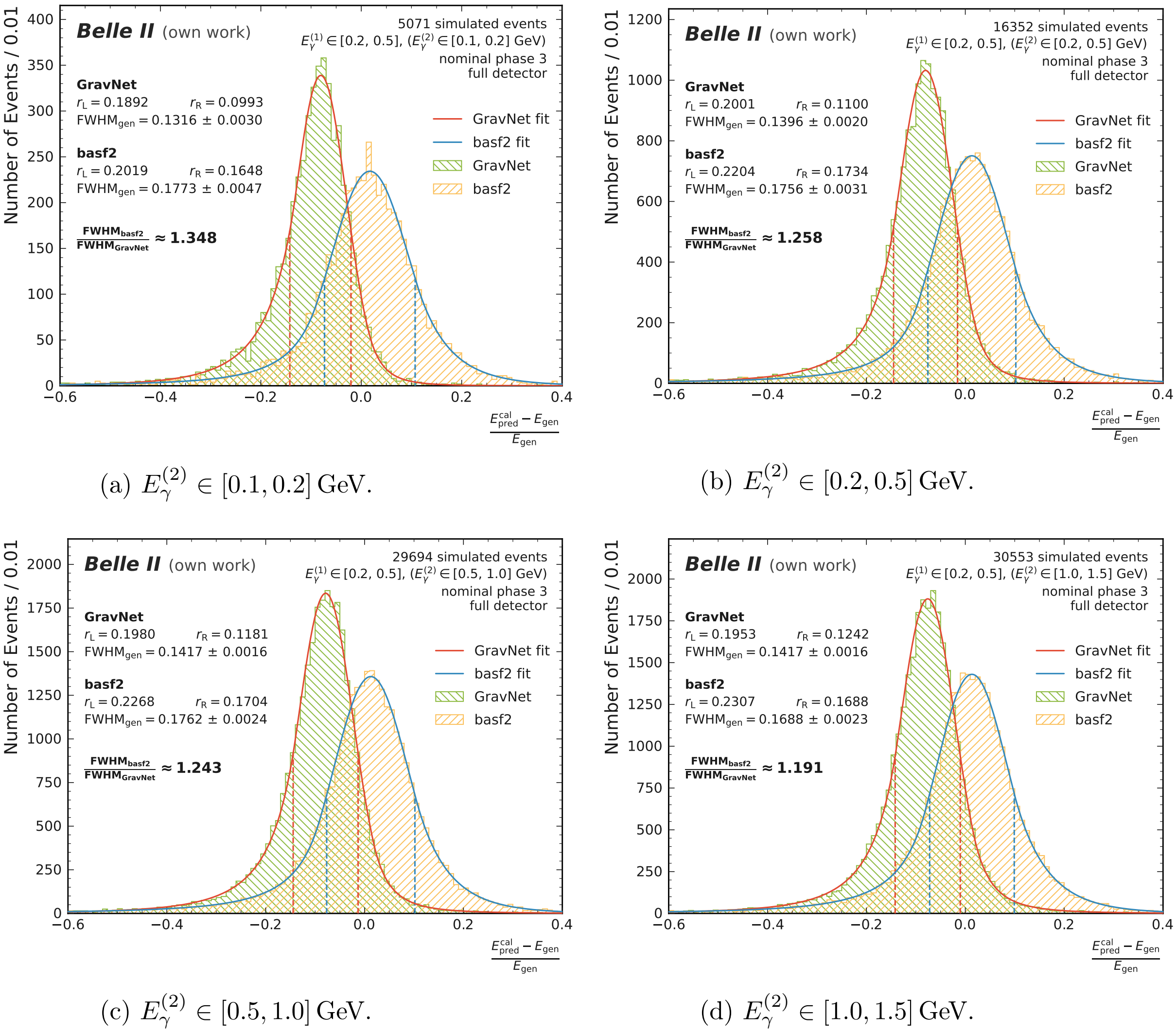
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .57.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.2, 0.5]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with early phase 3 background in full detector coverage.



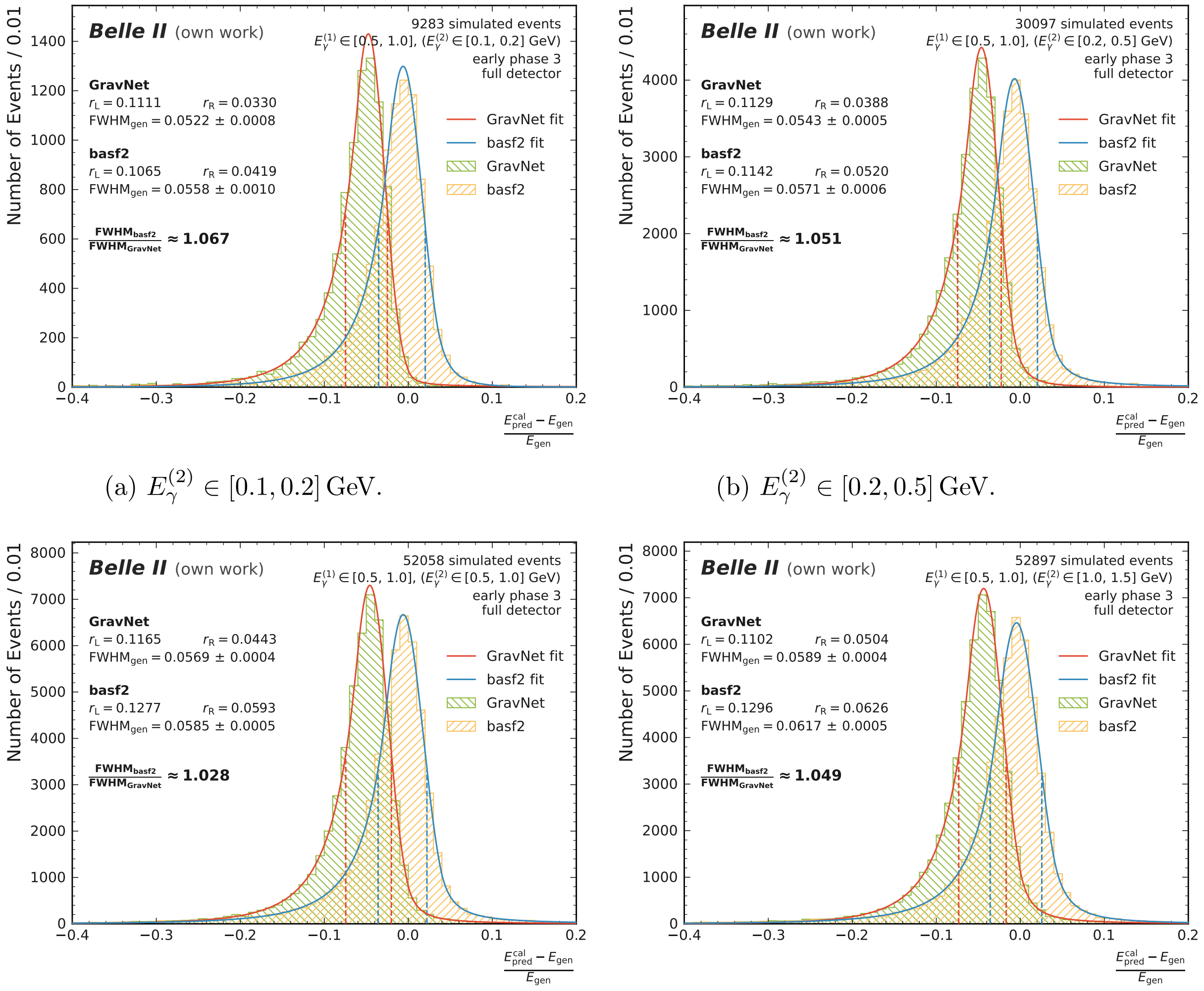
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .58.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.2, 0.5]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with nominal phase 3 background in full detector coverage.



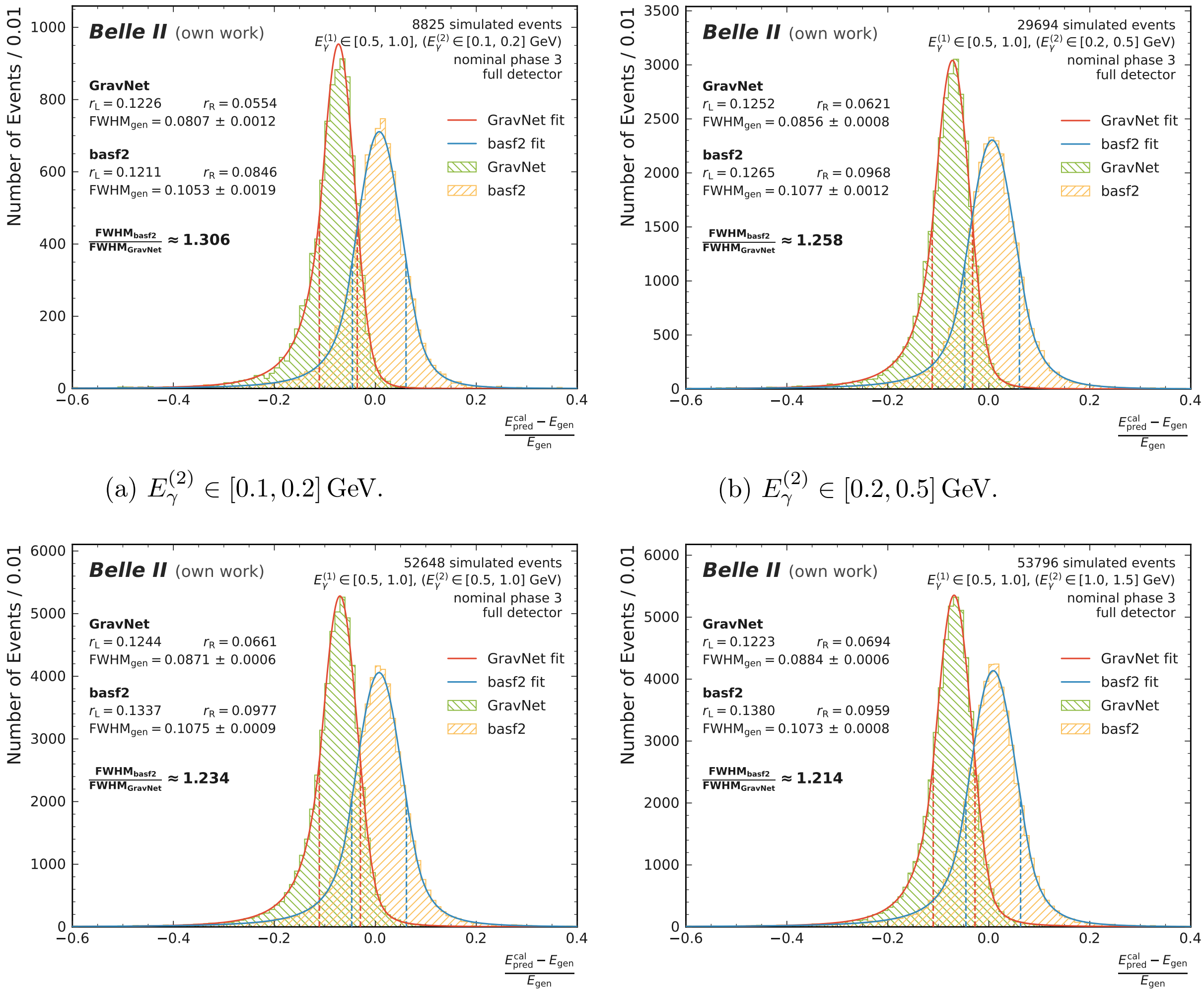
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .59.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.5, 1.0]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with early phase 3 background in full detector coverage.



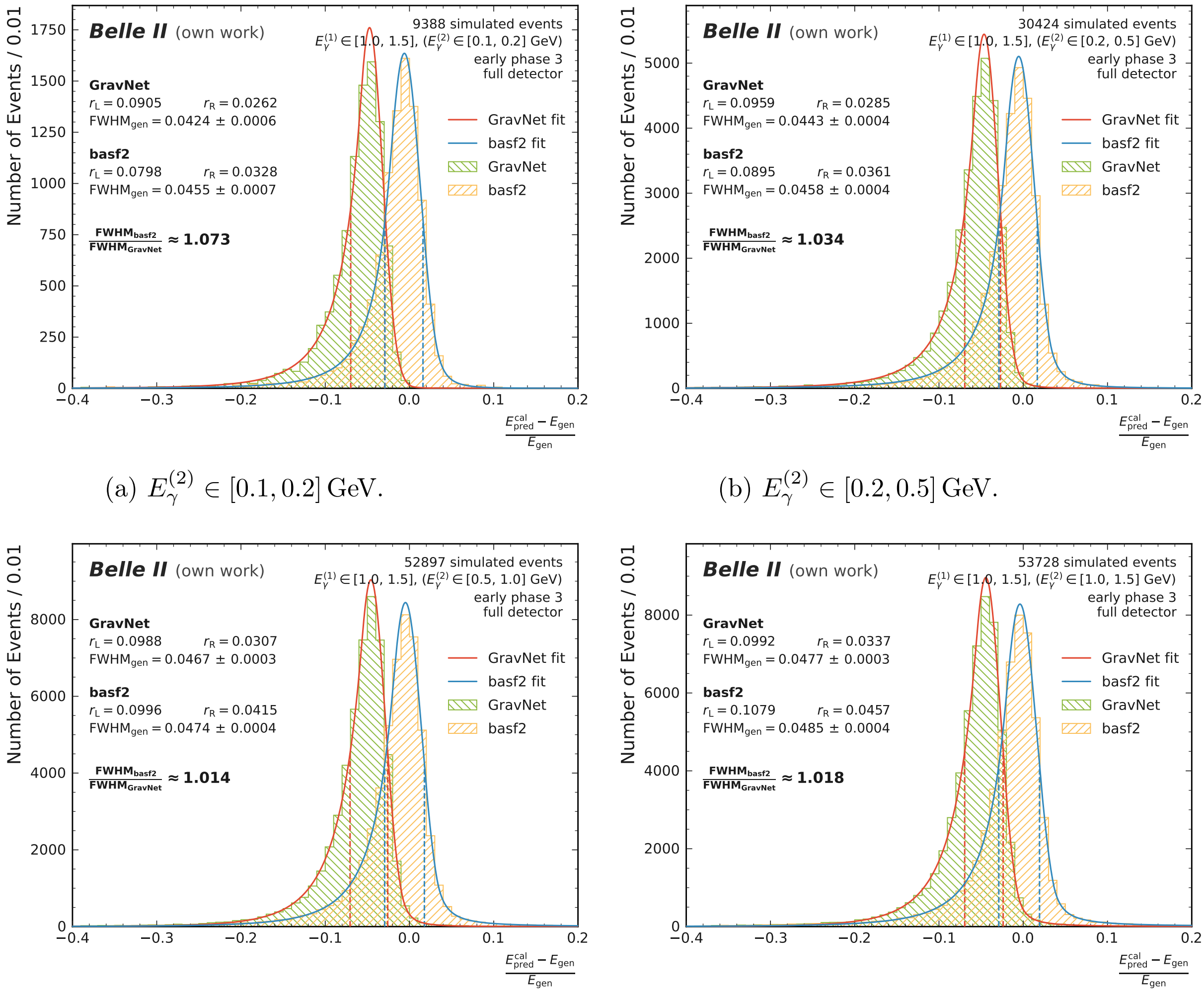
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .60.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [0.5, 1.0]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with nominal phase 3 background in full detector coverage.



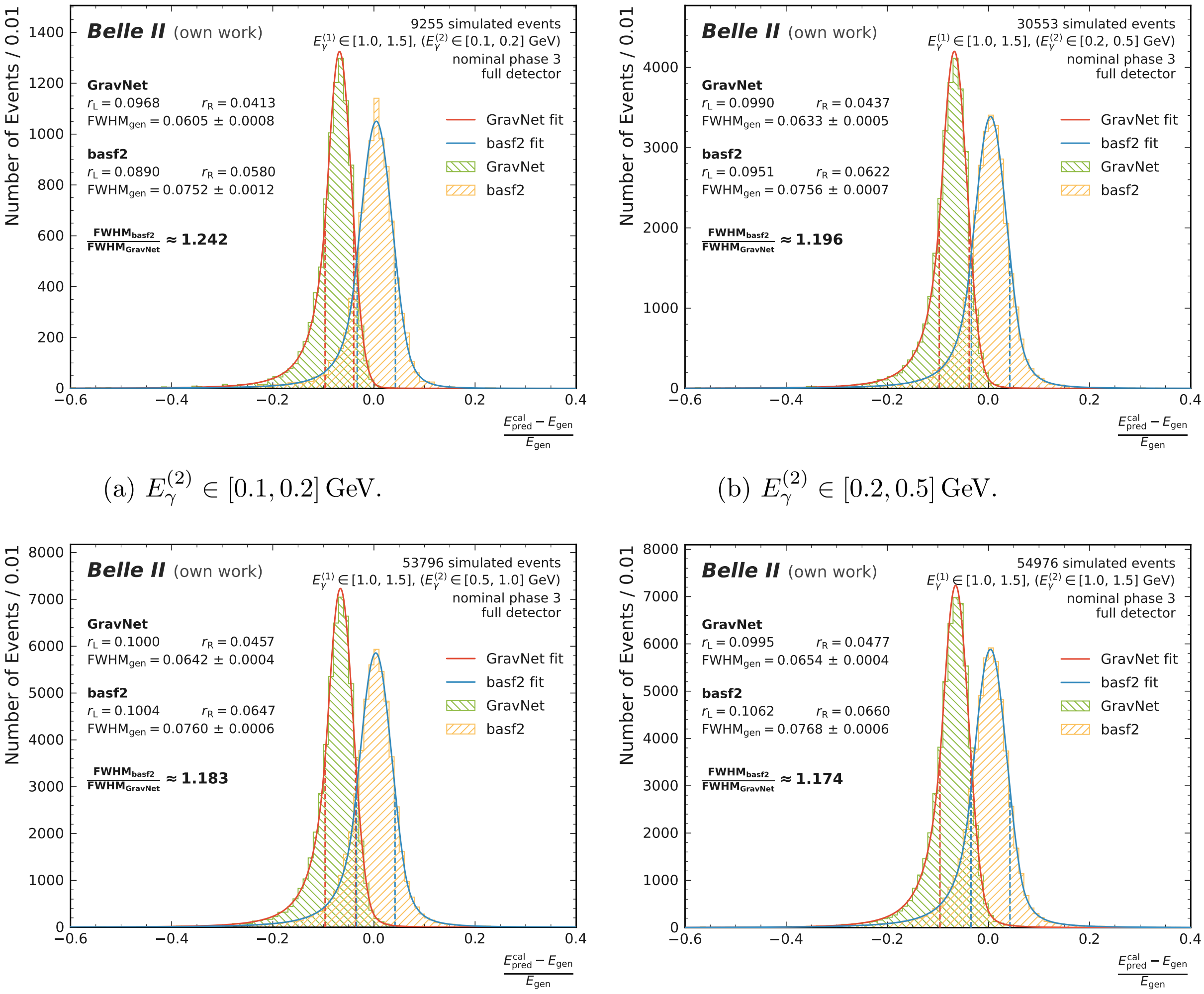
(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .61.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [1.0, 1.5]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with early phase 3 background in full detector coverage.



(a) $E_\gamma^{(2)} \in [0.1, 0.2]$ GeV.

(b) $E_\gamma^{(2)} \in [0.2, 0.5]$ GeV.

(c) $E_\gamma^{(2)} \in [0.5, 1.0]$ GeV.

(d) $E_\gamma^{(2)} \in [1.0, 1.5]$ GeV.

Figure .62.: Fit for the photon resolution $\mathrm{FWHM_{gen}}$ of one photon with photon energy $E_\gamma^{(1)} \in [1.0, 1.5]$ GeV in dependence of the second photon energy $E_\gamma^{(2)}$. The results are for the two-cluster toy study events with nominal phase 3 background in full detector coverage.

Table .1.: $\mathrm{FWHM}_{\mathrm{gen}} \times 10^{-4}$ of one photon with photon energy $E_{\gamma}^{(1)}$ in dependence of the second photon energy $E_{\gamma}^{(2)}$. The improvement to the basf2 baseline is stated in percent for each energy interval. The results are for the two-cluster toy study events with nominal phase 3 background in full detector coverage.

$E_{\gamma}^{(1)}$ (GeV) ↓	$E_{\gamma}^{(2)}$ (GeV) →	[0.1, 0.2]	[0.2, 0.5]	[0.5, 1.0]	[1.0, 1.5]
[0.1, 0.2]	GravNet	1104	1198	1194	1325
	basf2	1272	1393	1432	1516
	Improvement	**15.2 %**	**16.3 %**	**20.0 %**	**14.4 %**
[0.2, 0.5]	GravNet	738	757	823	838
	basf2	848	830	884	896
	Improvement	**14.9 %**	**9.7 %**	**7.5 %**	**7.0 %**
[0.5, 1.0]	GravNet	522	543	569	589
	basf2	558	571	585	617
	Improvement	**6.7 %**	**5.1 %**	**2.8 %**	**4.9 %**
[1.0, 1.5]	GravNet	424	443	467	477
	basf2	455	458	474	485
	Improvement	**7.3 %**	**3.4 %**	**1.4 %**	**1.8 %**

Table .2.: $\mathrm{FWHM}_{\mathrm{gen}} \times 10^{-4}$ of one photon with photon energy $E_{\gamma}^{(1)}$ in dependence of the second photon energy $E_{\gamma}^{(2)}$. The improvement to the basf2 baseline is stated in percent for each energy interval. The results are for the two-cluster toy study events with early phase 3 background in full detector coverage.

$E_{\gamma}^{(1)}$ (GeV) ↓	$E_{\gamma}^{(2)}$ (GeV) →	[0.1, 0.2]	[0.2, 0.5]	[0.5, 1.0]	[1.0, 1.5]
[0.1, 0.2]	GravNet	2477	2410	2402	2472
	basf2	3312	3282	3128	3242
	Improvement	**33.7 %**	**36.2 %**	**30.3 %**	**31.1 %**
[0.2, 0.5]	GravNet	1316	1396	1417	1417
	basf2	1773	1756	1762	1688
	Improvement	**34.8 %**	**25.8 %**	**24.3 %**	**19.1 %**
[0.5, 1.0]	GravNet	807	856	871	884
	basf2	1053	1077	1075	1073
	Improvement	**30.6 %**	**25.8 %**	**23.4 %**	**21.4 %**
[1.0, 1.5]	GravNet	605	633	642	654
	basf2	752	756	760	768
	Improvement	**24.2 %**	**19.6 %**	**18.3 %**	**17.4 %**

Bibliography

[1] **Belle II Framework Software Group**, T. Kuhr, C. Pulvermacher, M. Ritter, *et al.*, "The Belle II Core Software," *Computing and Software for Big Science.* **3** no. 1, (2019) . https://arxiv.org/abs/1809.04299v2.

[2] S. R. Qasim, J. Kieseler, Y. Iiyama, and M. Pierini, "Learning Representations of Irregular Particle-Detector Geometry With Distance-Weighted Graph Networks," *The European Physical Journal C* **79** no. 7, (2019) . https://doi.org/10.1140/epjc/s10052-019-7113-9.

[3] R. Rabbany and O. Zaïane, "A General Clustering Agreement Index: For Comparing Disjoint and Overlapping Clusters," *Proceedings of the AAAI Conference on Artificial Intelligence* **31** no. 1, (2017) . https://ojs.aaai.org/index.php/AAAI/article/view/10905.

[4] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering* **9** no. 3, (2007) 90–95. 10.1109/MCSE.2007.55.

[5] R. Oerter, "The Theory of Almost Everything: The Standard Model, the Unsung Triumph of Modern Physics," *Penguin Publishing Group,* (2006) . https://doi.org/10.1063/1.2337829.

[6] E. Kou, P. Urquijo, W. Altmannshofer, *et al.*, "The Belle II Physics Book," *Progress of Theoretical and Experimental Physics* **2020** no. 2, (2020) . https://doi.org/10.1093/ptep/ptaa008.

[7] T. Abe, I. Adachi, K. Adamczyk, *et al.*, "Belle II Technical Design Report," (2010). https://arxiv.org/abs/1011.0352.

[8] T. Ferber, "Electromagnetic Calorimeter Reconstruction in Belle II," (2019). Conference talk given at ACAT 2019. https://indico.cern.ch/event/708041/contributions/3269704.

[9] H. Ikeda, "Development of the CsI(Tl) Calorimeter for the Measurement of CP Violation at KEK B-Factory," (1999). https://opac2.lib.nara-wu.ac.jp/webopac/TD00003756.

[10] V. Aulchenko, A. Bobrov, T. Ferber, *et al.*, "Time and Energy Reconstruction at the Electromagnetic Calorimeter of the Belle II Detector," *Journal of Instrumentation* **12** no. 08, (2017) C08001. https://dx.doi.org/10.1088/1748-0221/12/08/C08001.

[11] S. Longo, J. Roney, C. Cecchi, *et al.*, "CsI(Tl) Pulse Shape Discrimination With the Belle II Electromagnetic Calorimeter as a Novel Method to Improve Particle Identification at Electron-Positron Colliders," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **982** (2020) 164562. https://doi.org/10.1016/j.nima.2020.164562.

[12] T. Ferber, C. Hearty, and M. Roney, "Design of the ECL Software for Belle II," (2016). BELLE2-NOTE-TE-2016-001 (Belle II Internal Note). https://docs.belle2.org/record/316.

[13] S. Sugihara and H. Aihara, "Design Study of Belle II Interaction Region," (2011). https://docs.belle2.org/record/217.

[14] **The Belle II Collaboration**, "Belle II Analysis Software Framework (basf2)." https://github.com/belle2/basf2.

[15] K. Miyabayashi, "Belle II Electromagnetic Calorimeter and its Performance During Early SuperKEKB Operation," (2019). Conference talk given at CHEF2019. https://indico.cern.ch/event/818783/contributions/3598493.

[16] S. Agostinelli, J. Allison, K. Amako, *et al.*, "GEANT4 - A Simulation Toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** no. 3, (2003) 250–303. https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[17] Z. Liu and J. Zhou, "Introduction to Graph Neural Networks," *Springer International Publishing* (2020) . https://doi.org/10.1007/978-3-031-01587-8_15.

[18] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," (2015). https://arxiv.org/abs/1511.07289.

[19] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Computing Research Repository* **abs/1912.01703** (2019) . http://arxiv.org/abs/1912.01703.

[20] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," *Computing Research Repository* **abs/1903.02428** (2019) . http://arxiv.org/abs/1903.02428.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12** (2011) 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195.

[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," (2014). https://arxiv.org/abs/1412.6980.

[23] T. Akiba, S. Sano, T. Yanase, *et al.*, "Optuna: A Next-generation Hyperparameter Optimization Framework," (2019). https://arxiv.org/abs/1907.10902.

[24] J. Eschle, A. Puig Navarro, R. Silva Coutinho, and N. Serra, "zfit: Scalable Pythonic Fitting," *SoftwareX* **11** (2020) 100508. https://www.sciencedirect.com/science/article/pii/S2352711019303851.

[25] E. O. Lebigot, "Uncertainties: A Python Package for Calculations With Uncertainties,". http://pythonhosted.org/uncertainties.

[26] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness and Correlation," *Computing Research Repository* **abs/2010.16061** (2020) . https://arxiv.org/abs/2010.16061.

[27] M. Wakai, "Cluster Position Reconstruction With a Neural Network Approach," (2022). Conference talk given at 43rd B2GM. https://indico.belle2.org/event/7476/contributions/46280.