# A First Full Physics Benchmark for Highly Granular Calorimeter Surrogates

Thorsten Buss,[1,2] Henry Day-Hall,[2] Frank Gaede,[2] Gregor Kasieczka,[1] Katja Krüger,[2] Anatolii Korol,[2,*] Thomas Madlener,[2] and Peter McKeown[2,3,†]

[1]*Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany*
[2]*Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany*
[3]*CERN, 1211 Geneva 23, Switzerland*

(Dated: November 20, 2025)

The physics programs of current and future collider experiments necessitate the development of surrogate simulators for calorimeter showers. While much progress has been made in the development of generative models for this task, they have typically been evaluated in simplified scenarios and for single particles. This is particularly true for the challenging task of highly granular calorimeter simulation. For the first time, this work studies the use of highly granular generative calorimeter surrogates in a realistic simulation application. We introduce DDML, a generic library which enables the combination of generative calorimeter surrogates with realistic detectors implemented using the DD4hep toolkit. We compare two different generative models – one operating on a regular grid representation, and the other using a less common point cloud approach. In order to disentangle methodological details from model performance, we provide comparisons to idealized simulators which directly sample representations of different resolutions from the full simulation ground-truth. We then systematically evaluate model performance on post-reconstruction benchmarks for electromagnetic shower simulation. Beginning with a typical single particle study, we introduce a first multi-particle benchmark based on di-photon separations, before studying a first full-physics benchmark based on hadronic decays of the tau lepton. Our results indicate that models operating on a point cloud can achieve a favorable balance between speed and accuracy for highly granular calorimeter simulation compared to those which operate on a regular grid representation.



## I. INTRODUCTION

Accurate and efficient simulations of particle detectors are essential for modern collider experiments, where they are used for detector design and physics analysis. Traditionally, simulations rely on Monte Carlo (MC) methods, which, despite their high accuracy, incur substantial computational costs [1, 2]. Especially high computational demands arise from the simulation of particle showers in calorimeter systems, where a single incident particle can trigger a large cascade of particles and interactions.

In recent years, generative surrogate models have emerged as a promising solution. These models aim to replicate the complex patterns of particle showers while significantly accelerating simulation speed. To this end, various generative paradigms have been applied, including generative adversarial networks (GANs) [3–18], variational autoencoders (VAEs) [15, 19–26], classical normalizing flows (NFs) [27–38], auto-regressive models [39], and diffusion and continuous flow models [40–52]. For a recent taxonomy see [53].

These models have been trained and evaluated on a wide range of datasets, making it difficult to draw general conclusions about their performance. To address this, the recent CaloChallenge 2022 [54] was started. It provided a valuable, large-scale comparison of generative models for calorimeter simulation on common datasets.

However, the evaluation of generative models has typically been limited to simplified scenarios, such as single particle showers with fixed impact angles (usually normal to the calorimeter surface) and fixed positions within the detector volume. While these benchmarks are useful for assessing the basic capabilities of generative models, they do not show how these models perform in a realistic experimental setup.

A notable exception is the ATLAS experiment, where GANs are already used in production [14, 15], with NFs and diffusion models being explored for use in its next generation of fast simulation [55]. However, sufficient performance of generative calorimeter surrogates in realistic experimental setups has yet to be demonstrated for highly granular calorimeters, such as those designed for the CMS HGCAL [56] and future collider experiments [57, 58]. These detectors are able to resolve significantly finer substructure in calorimeter showers, and therefore require generative surrogates to operate on orders of magnitude higher data dimensionality and sparsity.

To address this gap, we introduce DDML [59], a flexible library designed to integrate generative models into the full simulation chain of various particle physics experiments. It is built on top of the DD4HEP [60] toolkit which is widely adopted in the HEP community. Using DDML, we present the first full physics benchmark for highly granular calorimeter surrogates, studying the International Large Detector (ILD) [57] as an example. While ILD is used as a case study of a highly granular detector, the DDML library already supports several detector concepts for future colliders, and could easily be adapted to other detector designs. We integrate two state-of-the-art generative models, CONVL2LFLOWS [38] and CALOCLOUDS3 [61]. These models are trained solely on Geant4 photon showers in an idealized detector geometry. We use these surrogate models and the traditional Geant4 simulation and compare the simulation and reconstruction results with three physics benchmark datasets: single-photon showers, di-photon separation, and $\tau$-pair events. In addition, we

* anatolii.korol@desy.de
† peter.mckeown@cern.ch

provide comparisons to three idealized models that directly sample from the ground-truth full simulation to analyze the effects that the choice of data representation and modeling assumptions have.

We systematically evaluate these models on three post-reconstruction benchmarks, specifically chosen to test highly granular calorimeter surrogates for electromagnetic showers. Beginning with a typical evaluation of single particle performance, we then introduce a first multi-particle benchmark in this context by studying di-photon separations, before performing a full physics benchmark based on hadronic decays of the tau lepton.

The remainder of the paper is structured as follows. Section II describes the datasets and benchmarks used in this work. Section III provides descriptions of the ConvL2LFlows and CaloClouds3 models which are the subjects of this study. Section IV presents the results of our benchmarks for single particles, di-photon events, and tau decay events. Finally, Section V provides a discussion.

## II. DATASETS AND BENCHMARKS

This section describes all datasets used in this study. Section II A introduces the ILD detector concept, which relies on highly granular sampling calorimeters. The different calorimeter shower representations studied are introduced in Sections II B, together with Geant4 reference generators for each representation. The training dataset for the surrogate models is introduced in Section II C. Section II D introduces the approach to benchmarking adopted in this study, followed by the datasets used for single-particle validation, the di-photon benchmark, and the tau physics benchmark.

### A. The International Large Detector

This work focuses on the International Large Detector (ILD) [57], a next generation detector for a future $e^+e^-$ Higgs factory, originally proposed for the International Linear Collider (ILC), and currently also under investigation for use at the FCC-ee. ILD is optimized for the Particle Flow approach to reconstruction, and as such features highly hermetic detector systems, highly granular calorimeters and a minimal material budget in front of the calorimeters. The ILD detector model studied in this work consists of a polyhedral barrel geometry with a total radius of 7.8 m and a total length of 13 m. The tracking system consists of a number of silicon pixel detectors, which are encapsulated in a time projection chamber (TPC). Outside of the tracking system are placed highly granular sampling calorimeters, separated into an electromagnetic calorimeter (ECAL) system and a hadronic calorimeter (HCAL) system. Both of these calorimeter systems consist of an octagonal barrel region with the ends closed by flat endcap disks. The primary focus of this study is the Si-W option for the ECAL [62]. This calorimeter consists of 30 layers of passive tungsten absorbers and active silicon sensors. The first 20 tungsten absorbers have a thickness of 2.1 mm, while the last 10 layers have a thickness of 4.2 mm. The total thickness of the calorimeter therefore corresponds to approximately 24 radiation lengths. The silicon layers feature cells of size $5 \times 5$ mm$^2$, and have a thickness of 0.525 mm. Behind the ECAL is placed the analogue hadronic calorimeter (AHCAL) [63]. It consists of 48 layers with stainless steel absorbers, each with a thickness of 17.2 mm, and 3 mm thick active layers. The active layers feature $3 \times 3$ cm$^2$ scintillator tiles, each individually read out by a silicon photomultiplier. These detector systems sit inside a superconducting solenoid coil, which produces a magnetic field of strength 3.5 T orientated along the beam axis. An iron return yoke with integrated muon system and tail catcher calorimeter is placed outside of the coil.

For this study, we employ the Key4hep [64] software ecosystem, using Geant4 [65] version 11.2.2 with the QGSP_BERT physics list, and DD4hep [60] version 1.30. A realistic and detailed model of the ILD detector geometry[1], described using the DD4hep toolkit, is used. Of particular relevance, is the geometrical structure of the sensitive layers of the ECAL, a visualization of which is shown for two active layers in Figure 1 (left). This illustrates that the readout geometry of the detector is irregular, featuring two types of insensitive volumes. The smaller insensitive volumes are present only in the active layers, and lie at the edge of the silicon wafers. A staggering effect is present in their positioning between layers. The larger insensitive volumes arise from the presence of gaps between sensors or structural supports in the calorimeter. These are therefore aligned between layers, and present in both the absorber and the active layers. Such an irregular readout geometry, which will be present in every realistic calorimeter, creates a number of difficulties when conceiving a scheme to allow a model to be used to simulate showers at varying positions on and angles to the calorimeter face. For models relying on a regular grid representation of calorimeter layers, it would require a means of removing projection artifacts [19] for all possible incident positions and angles, which would be infeasible. An additional problem that would affect models generally, is the variation in the fraction of the active layer that is sensitive depending on the local geometry near the incidence position. Since Geant4 discards energy depositions that do not land in a region of the detector that is not assigned to be sensitive, this would result in the potential loss of information when trying to simulate a shower at a different position than the one the model was originally trained at.

To combat this challenge, a modified DD4hep description of the ILD ECAL with a regularized readout geometry was created, as shown in Figure 1 (right). In this geometry, the readout segmentation of the sensitive layers was altered such that the layer contained no insensitive volumes, meaning that all energy depositions in these layers are recorded. This alteration leaves the structure of the detector, both in terms of material composition and longitudinal layer placement (i.e. into

[1] The version of the ILD detector geometry used in this study is `ILD_l5_o1_v02`

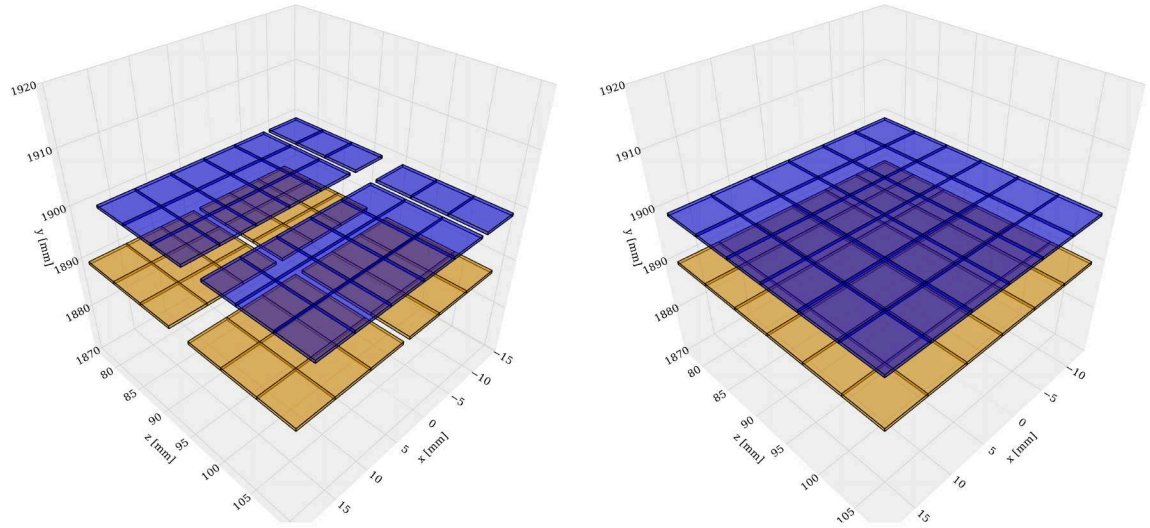


FIG. 1. Visualization of geometry maps for (left) the physical geometry and (right) the regularized geometry for a section of two sensitive layers in the calorimeter. The physical geometry includes gaps between the cells arising from insensitive volumes such as structural supports and readout electronics, as well as a staggering effect between layers. The regularized geometry consists purely of sensitive material, with the cells being perfectly aligned from one layer to the next. Figure from [66].

the depth of the calorimeter), unaffected. By using this regularized version of the calorimeter to create a training dataset, maximum information about energy depositions in the sensitive layers can be retained, allowing them to be dropped selectively during simulation (see Section A), depending on the local readout geometry and thereby avoiding artifacts from sensitive gaps in the training showers.

### B. Shower Representations

GEANT4 produces individual energy deposits, so called GEANT4 *steps*, with a much higher spatial resolution than the physical detector readout. If a generative model is to be used to simulate showers across different incident positions, intuitively having a mechanism for generating showers at a higher resolution than the detector readout should reduce artifacts and edge effects. To better understand the consequences and potential limitations of projecting showers into a realistic detector geometry, we consider three different shower representations, each accompanied by a corresponding truth-based reference, which we denote as *optimal shower generators*: OPTIMUM (X1), OPTIMUM (X9), and OPTIMUM (STEPS).

Each of these optimal shower generators is derived from simulations run with GEANT4 on the regularized ILD ECAL introduced in section II, from which all individual GEANT4 steps within the sensitive layers are extracted. This enables us to quantify the effect of projecting a regular grid with a given granularity onto the actual detector geometry, and allows us to isolate the effects of the data representation from the performance of a given generative model.

The first representation $\mathcal{R}_{\times 1}$ features a granularity identical to that of the ILD ECAL ($5 \times 5$ mm$^2$), and has a corresponding optimal shower generator OPTIMUM (X1) shown in Figure 2 (left). Here, each simulated step from GEANT4 is projected onto a virtual regular grid, exactly matching the realistic detector cell sizes. Thus, OPTIMUM (X1), provides the ideal reference for the performance achievable with any model trained on the readout geometry of the physical detector.

The second representation $\mathcal{R}_{\times 9}$ increases the lateral granularity by a factor of three in each dimension, resulting in nine times more cells ($1.67 \times 1.67$ mm$^2$) per layer compared to $\mathcal{R}_{\times 1}$. This representation has an optimal shower generator OPTIMUM (X9), shown in Figure 2 (center). An increased granularity such as this allows for a finer spatial resolution of showers, reducing projection-related effects. It serves as a reference for understanding the impact of increasing the granularity of the data representation on the fidelity of the projected showers.

Finally, representation $\mathcal{R}_{(\text{steps})}$, with corresponding generator OPTIMUM (STEPS) shown in Figure 2 (right), provides the most detailed reference scenario by avoiding any spatial projection onto a predefined grid altogether. Instead, it directly utilizes the ultimate resolution – GEANT4 simulation steps – as they occur within the sensitive material. This results in the highest achievable spatial resolution, completely free from projection artifacts, and represents the ultimate benchmark for evaluating the accuracy and potential information loss associated with any spatial discretization scheme in the regularized detector geometry.

For all optimal shower generators, a wide bounding box is used to select shower hits, with a side length of 800 mm. This box cut is necessary to exclude rare low energy hits resulting from backscatter that typically occur at the opposite end of the detector to which the showers occur. This ensures that all

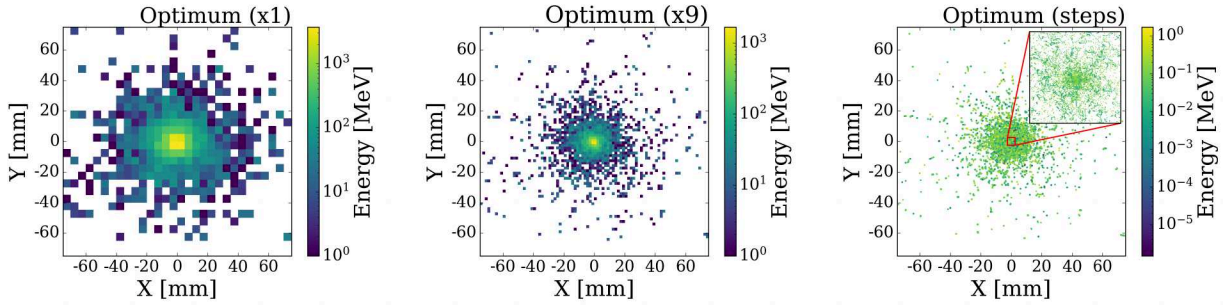


FIG. 2. Visualization of the same 90 GeV electromagnetic shower in lateral projection of the ILD ECAL, represented using the three optimal shower generators: OPTIMUM (x1) (left), OPTIMUM (x9) (center), and OPTIMUM (STEPS) (right).

relevant hits in the shower are contained.

By leveraging these optimal shower generators – OPTIMUM (x1), OPTIMUM (x9), and OPTIMUM (steps), we gain insight into the intrinsic limits imposed by projection artifacts, independent of generative model performance. Moreover, they provide invaluable baselines against which generative models can be assessed.

### C. Training Dataset

The training dataset used in the study consisted of $\sim 3$ million samples, created by photons with incident energies uniformly distributed in the range of 1-126 GeV. The photons were created at a position of $[x = 0,\ y = 1804.7$ mm, $z = -50$ mm] at the front face of the ECAL. Here the ILD global coordinate system is defined with the $z$ axis along the beam direction, the $x$ axis pointing horizontally, and the $y$ axis pointing vertically upwards. The incident angles were varied within a cone of up to 60° in $\theta'$ (relative to the normal to the calorimeter layers), with an azimuthal angle $\varphi'$ uniformly distributed in $[0°, 360°)$. These angles are defined in the local frame of reference, where the $z'$ axis is aligned with the normal to the calorimeter layers.

This configuration enabled uniform sampling over the space of possible incident directions on the calorimeter face, and conforms with the coordinate system convention used in the DDML library, which is described in Appendix A 1[2].

### D. Benchmark datasets

We introduce the generic library DDML [59], which allows the inclusion of different generative models designed for fast calorimeter shower simulation in full simulation applications using the DD4HEP toolkit [60]. This library and the details of the implementation for the ILD detector used in this study are described in Appendix A. Crucially, this allows the generation of particle showers in the standard software chain of ILD. This makes it possible to run the standard reconstruction of ILD, in particular particle flow reconstruction with PANDORAPFA [67], enabling realistic physics benchmarking of generative models. All benchmarking samples used in this paper have gone through the complete software chain including event reconstruction.

#### 1. *Benchmarking Methodology*

While studying single-shower observables, as is now standard in the literature, is important to gauge the performance of a model, in real physics events showers from multiple particles may overlap. This significantly increases the complexity of evaluating the performance of a model, as the breadth of the phase space makes disentangling the interplay of overlapping showers with reconstruction algorithms a challenging task. For this reason we take a step-by-step approach, ultimately building towards benchmarking the model in a full physics setting.

We begin by studying the performance of the model in terms of single particle observables. Next, we study the simplest scenario for a multi-particle test – two photons fired into the face of the ECAL. This is a standard benchmark which has also been used for the development of reconstruction algorithms, such as PANDORAPFA [68]. This approach provides an isolated and controllable means of probing generative model performance, as well as allowing connections to be drawn to the performance on single particle observables.

Finally, we study the performance of generative models for the simulation of photon showers in a full physics process. As the performance required of a fast simulation tool will depend heavily on the physics process for which the tool is used, we desire a physics process that provides a stringent test of a fast simulation tool for electromagnetic showers. To this end, we choose hadronic decay modes of the tau lepton in the process $e^+e^- \to \tau^+\tau^-$.

The tau lepton is of interest for many precision studies planned for future $e^+e^-$ collider experiments [69–73]. As a result of its high mass, it has the strongest coupling to the Higgs of any lepton, and is the only lepton in the Standard Model which decays to hadrons. This, combined with the ability to

[2] Throughout this paper, we use primed coordinates to represent the local calorimeter coordinate system, otherwise coordinates are assumed to be in the global ILD coordinate system.

precisely reconstruct the spin state of the tau from its decay products [74, 75], makes it a prime candidate for probing the Higgs sector, including its CP structure [76, 77].

Approximately 65% of tau decays involve hadrons, with the reduced number of neutrinos involved compared to purely leptonic decays meaning that hadronic decay modes are preferred for precision measurements. Hadronic decay modes of the tau frequently involve one or more neutral pions, often occurring via the $\rho(770)$ or $a1(1200)$ intermediate resonances [78]. Reconstructing the correct decay mode of the tau, which is essential when determining its spin state, therefore often involves correctly reconstructing the number of $\pi^0$s produced, each of which almost always decay via $\pi^0 \to \gamma\gamma$ [78]. This is challenging, due to the high boost and collimation of the decay products, making the reconstruction of hadronic tau decays a standard benchmark of the performance of an electromagnetic calorimeter [74, 75]. The presence of numerous overlapping photon showers in such events therefore makes them ideal for exposing any flaws in the performance of a fast simulation tool for electromagnetic calorimeters. Correctly reconstructing these $\pi^0 s$ involves not only distinguishing the number of photons from overlapping showers (the performance of which can be linked to the aforementioned isolated di-photon benchmark), but also correctly inferring the kinematics from shower-level observables (which can be linked to the aforementioned single particle benchmark).

#### 2. Single Particle Dataset

In order to validate the generalization capability of the trained models, independent single shower test samples were generated at multiple positions, uniformly distributed over a single ECAL barrel segment. Test samples were produced at fixed photon energies between 10 and 100 GeV in 10 GeV increments. For each test energy, a sample consisted of 3,000 photon showers with positions selected randomly on the front surface of the ECAL. The incident directions were chosen such that they mimicked particles originating from the interaction point (IP), across the angular ranges $43^\circ < \theta < 137^\circ$ and $79^\circ < \varphi < 109^\circ$ in the ILD global coordinate system, thereby effectively covering one complete stave of the barrel. This configuration ensures full coverage of the region where models are expected to operate, exposing them to a wide variety of incident directions and local geometries of the sensitive layers – thereby enabling the evaluation of model performance under realistic conditions.

#### 3. Di-photon Dataset

The di-photon benchmark dataset consists of three sets of 15,000 samples containing two photons. The three sets of samples have different incident photon energies of 5 GeV, 20 GeV and 100 GeV. Within each sample, the energies of each of the two photons are identical.

The photons are produced directly at the face of the ILD ECAL barrel, with their direction of flight being orientated such that they appear to have been produced at the IP[3].

The photon positions are randomly sampled to expose the di-photon system to different local geometries of the sensitive layers, while ensuring that the separation between the two photons varies uniformly from 0 to 90 mm. The photons impact the upper barrel module, within a narrow angular range of $\theta \in [83^\circ, 87^\circ]$ and $\phi \in [88^\circ, 92^\circ]$, corresponding to a small localized patch of the detector. To ensure that all photon pairs remain fully contained within the fast simulation trigger region (see Section A 1) and avoid contamination from showers simulated with Geant4, the center of this patch is placed well inside the boundaries of the trigger region. These samples are created for Geant4, and each of the shower generation approaches under study. In each sample, both showers are produced with the respective generator.

#### 4. Tau Physics Dataset

The dataset used in this study consists of samples of the process $e^+e^- \to \tau^+\tau^-$ in an ILC running scenario at a center-of-mass energy of 250 GeV. MC Generator samples, provided by the ILD Software Working Group in the MC-2020 production [79], were created with WHIZARD [80] version 2.8.5. A realistic ILC beam energy spectrum and crossing angle, as well as the effects of bremsstrahlung and initial state radiation were included. All samples contained beams of 100% polarized left-handed electrons and right-handed positrons ($e_L^- e_R^+$). The decay of the tau leptons in the samples was simulated with the TAUOLA library [81].

In order to enable a direct comparison between the various shower representations and models described in Sections II B and III, the same set of MC generator inputs were used in all cases. This means that for each case, only the detector simulation differs, removing any differences in underlying event topologies or physics processes. In addition, no background is overlaid onto events. These two choices reduce event dilution that would only serve to obscure the performance of the calorimeter shower simulators. However, it should be noted that the detector simulation itself can significantly alter the signature of the event depending on what interactions occur prior to the calorimeter. A particularly pertinent example for this study is the case in which a photon converts into an electron-positron pair. This case provides a potentially easier reconstruction scenario, as the charged electron and positron have an associated track and larger separation at the calorimeter face.

Several selection criteria were put on the events, to further enhance the sensitivity of the analysis to the performance of the calorimeter shower simulators. It was required that all events contained at least one $\pi^0$ produced in a tau decay, with the $\pi^0$ then decaying into two photons. Both of the photons were required to have an energy above 5 GeV and to satisfy

[3] This mimics photons coming from the IP, e.g. from $\pi^0$ decays, without having to address material interactions, like pair-creation, in the tracking detector on the way to the calorimeter.

the geometry region trigger described in Appendix A 2. Detector simulation was then performed for each of the various shower representations and models described in Sections II B and III. The software configuration described in Section II A was employed, using the DDML implementation described in Appendix A 2. All photons with an energy above 5 GeV incident on the electromagnetic calorimeter which passed the geometry region trigger were then simulated with the appropriate simulator.

For each calorimeter shower simulator, samples were created with three different random seeds for the detector simulation. This provides a means of estimating the uncertainties on post-reconstruction physics observables that would otherwise be difficult to estimate, given the high level of correlation arising from the use of identical generator level input. In total this meant that samples, consisting of 3 random seeds each containing 6791 events, were generated for GEANT4, each of the optimal shower generators OPTIMUM (x1), OPTIMUM (x9), OPTIMUM (STEPS) described in Section II B and for both the CALOCLOUDS3 and CONVL2LFLOWS models, which will be described in Section III. The standard ILD reconstruction chain, as described was then applied to each sample.

## III. MODELS

We compare two state-of-the-art generative models operating on different data representations – CONVL2LFLOWS, a regular grid-based architecture, and CALOCLOUDS3, a point cloud model. Each model is trained on different shower representations described in Section II B.

### A. Convolutional L2LFlows

CONVL2LFLOWS [38] is a generative surrogate based on normalizing flows [82, 83], designed for fast and accurate simulation of electromagnetic showers in calorimeters. It operates on a fixed-grid representation, where each shower is discretized into a three-dimensional grid. This voxelized representation allows for convolution-based architectures, such as U-Nets [84], to effectively model the complex spatial dependencies in calorimeter showers.

CONVL2LFLOWS generates calorimeter responses sequentially, layer by layer, where the generation of each layer is conditioned on the previous ones. The underlying flow-based architecture allows for single-shot sampling. Further details of the model architecture can be found in [38].

As demonstrated in the CaloChallenge 2022 [54], CONVL2LFLOWS achieves one of the best trade-offs between accuracy and generation speed among the submitted models on a fixed-grid dataset, making it well-suited for fair comparison between fixed-grid and point-cloud-based models.

While the original CONVL2LFLOWS model was restricted to a fixed incident point and angle, we have extended its capabilities with respect to these conditions to enable its application in this study. Firstly, the model is trained on the $\mathcal{R}_{\times 9}$ representation of showers described in Section II B. Given that the granularity present in this representation is significantly finer than the detector readout, it was necessary to impose a bounding box of side length 150 mm. While this produces a noticeable cut in the tails of the shower, it is necessary to handle the sparsity present in the shower and constrain the model size within reasonable limits. The resulting grid has $90 \times 90 \times 30$ voxels. Secondly, conditioning on the incident angle enables the model to generate showers for a range of impact angles, rather than being limited to a 90-degree impact angle. Technically, the angle is given as a unit vector to preserve the topological structure. To prevent showers from developing outside the bounding box, all layers are shifted to center the shower core. This effectively results in a tilted bounding box. These improvements broaden the applicability of CONVL2LFLOWS, enabling it to be used for a large fraction of highly energetic photon showers in the ILD ECAL.

### B. CaloClouds

The second model is CALOCLOUDS [41], a point cloud generative model that was developed to address the challenges induced by the irregular structure of the detector readout layers described in Section II, and to more efficiently handle the very high sparsity present in highly granular calorimeter showers. In this study we employ a third iteration of the model, denoted as CALOCLOUDS3 [61], which constitutes an extension of the CALOCLOUDS II [45] architecture.

CALOCLOUDS3 utilizes a grid with a 25x higher granularity than the original cells, which we refer to as the $\mathcal{R}_{\times 25}$ representation. Additional preprocessing is applied to dequantize the hit positions, together with cuts which enforce a bounding box of side width 500 mm for shower hit selection. This box cut only removes the outer tails of the shower very far from the shower core. For more details on preprocessing, see [61].

Like the improved CONVL2LFLOWS model, this version of CALOCLOUDS also incorporates angular conditioning to generate showers for a range of incident directions. This is achieved through a similar data preprocessing and conditioning methodology as used for CONVL2LFLOWS. The angular conditioning is achieved by explicitly providing the unit vector of the particle's momentum direction as additional conditioning parameter, enabling the model to account for the direction of the incident particle during generation. Such a capability is essential for realistic applications within the simulation chain, where different directions of incoming particles are expected.

In addition to this functional advancement, CALOCLOUDS3 features a simplified architecture. Both the POINTWISENET of the diffusion model and the SHOWERFLOW components have been optimized, resulting in a reduction in the inference time by a factor of $\sim 2$ compared to the previous iteration, CALOCLOUDS II, while preserving the generative fidelity. CALOCLOUDS3 is described in more detail in a dedicated paper [61].

## IV. RESULTS

We present the results of simulator performance for each benchmark, beginning with the single particle performance in Section IV A, followed by the di-photon performance in Section IV C, and finally the results of the full physics benchmark using tau decays in Section IV D. We conclude the section by presenting the results for single shower generation times in Section IV B.

### A. Single Particle Performance

A detailed validation of individual particle showers is crucial to ensure that these models correctly capture essential physics characteristics before being used in more complex, multi-particle scenarios or physics analyses.

We present an evaluation of key calorimetric observables for electromagnetic showers generated by single photons using the generative models described previously. We study radial and longitudinal shower profiles, energy resolution, linearity, and the intrinsic shower angle reconstruction. These observables are usually chosen in the literature because they encapsulate the essential features of electromagnetic shower development, and significantly influence the performance of particle identification and reconstruction algorithms.

We further benchmark our generative models against the optimal shower generators introduced in Section II B. These optimal shower generators represent idealized performance scenarios that isolate and quantify the intrinsic limitations arising from spatial discretization effects and detector irregularities, independent of the generative model itself. By comparing the performance of the generative models against both these optimal shower generators and the standard GEANT4 simulation, we are able to clearly distinguish between artifacts arising from the data representation and the intrinsic capabilities of the approaches to generative modeling explored.

This structured approach enables a detailed assessment of generative model fidelity and highlights areas requiring further improvement.

#### 1. Shower Profiles

We begin the performance evaluation of the generative models by examining their ability to reproduce the characteristic EM shower shapes, assessed through comparisons of radial and longitudinal energy profiles, which are key observables reflecting the spatial energy deposition pattern within the calorimeter.

The following observables are computed at the reconstruction level, after the generative models have been fully integrated into the simulation pipeline. This ensures that any potential geometric or systematic effects introduced by the integration framework are accounted for in the performance evaluation.

The radial energy profile, shown in Figure 3 (left), illustrates the mean deposited energy as a function of orthogonal distance from the axis aligned with the direction of the incident particle to the center of the cell. The longitudinal energy profile, displayed in Figure 3 (right), represents the average energy deposited per calorimeter layer along the depth of the detector.

Both models reproduce the longitudinal profile with good accuracy. The CALOCLOUDS3 model achieves the closest agreement with GEANT4, with deviations typically within a few percent. The CONVL2LFLOWS model performs similarly well, although it exhibits deviations of up to 15% near the start and end of the calorimeter. Most of these deviations are due to the finite simulation volume required for fixed grid models.

At first inspection, the radial energy profile appears to be well reproduced only by the optimal shower generators – OPTIMUM (x1), OPTIMUM (x9), and OPTIMUM (STEPS), while the CALOCLOUDS3 and CONVL2LFLOWS models have notable discrepancies at larger radial distances. However, it is important to note that the radial profile of electromagnetic showers is inherently steep and narrowly peaked, with the energy density rapidly decreasing with distance from the shower axis. In fact, more than 90% of the total energy of the shower is contained within a radius of 30 mm, indicating that deviations at large radii have a limited impact on the overall shower description.

To better assess the fidelity of the models in the region that dominates the shower energy density, Figure 4 zooms into the first 30 mm from the shower axis. Note that the dip in the first bin arises from the fact that the binning here is applied at a distance less than the width of a cell, whereas the hits in the shower are necessarily at the center of a cell after reconstruction. This is the most relevant part of the shower which plays a crucial role in the separation of overlapping showers during particle reconstruction.

In this region, the CALOCLOUDS3 and CONVL2LFLOWS models show good agreement with GEANT4, with relative deviations generally remaining below 10%. By contrast, OPTIMUM (x1) underestimates the energy density by up to 20% in the innermost bins. This demonstrates the intrinsic limitation imposed by the $\mathcal{R}_{\times 1}$ representation, where the coarser lateral granularity fails to resolve the sharply peaked energy profile near the shower axis. The lack of detail in this region can significantly impact the reconstruction of particle showers with a large degree of overlap. This highlights a significant disadvantage of generative models trained on fixed grids using the true detector granularity.

#### 2. Resolution and Linearity

We now investigate the energy resolution and linearity of the generative models, two critical performance metrics for calorimeter simulation. The energy resolution quantifies the model's ability to accurately reproduce the fluctuations in the deposited energy, directly affecting the precision with which particle energies can be measured. Linearity assesses how accurately the reconstructed energy scales with the true particle energy, essential for ensuring unbiased energy measurements across a large range of incident particle energies.

The energy resolution is evaluated by measuring the rela-

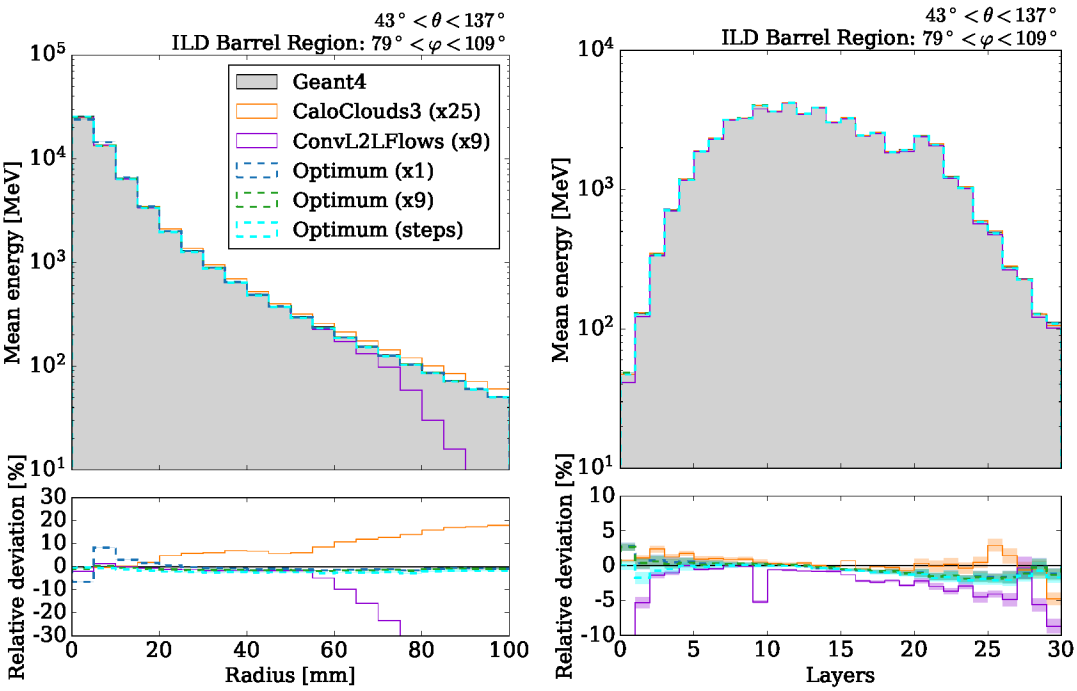


FIG. 3. Radial (left) and longitudinal (right) energy profiles of electromagnetic showers, computed at the reconstruction level after integration of the generative models. The radial profile shows the mean reconstructed energy as a function of distance from the shower axis, while the longitudinal profile shows the mean reconstructed energy per calorimeter layer. These observables provide a detailed characterization of the transverse and longitudinal shower structure and are critical benchmarks for assessing how well generative models replicate GEANT4 showers in realistic detector geometry settings. The color coding corresponds to the different generative models: CALOCLOUDS3 (orange), CONVL2LFLOWS (violet), OPTIMUM (x1) (blue), OPTIMUM (x9) (green), and OPTIMUM (STEPS) (cyan). The GEANT4 reference is shown in the light grey filled histogram. The color coding is consistent across all figures in this section. Shaded bands indicate statistical uncertainties; lower panels show relative deviations with respect to the GEANT4 baseline.

tive width of the reconstructed energy distribution for photons at various fixed energies, defined as $\frac{\sigma_{90}}{\mu_{90}}$, where $\sigma_{90}$ and $\mu_{90}$ are standard deviation and mean of the central 90% of the distribution, shown as a function of the incident photon energy in Figure 5 (left). Figure 5 (right) displays the linearity, expressed as the mean reconstructed energy divided by the true incident energy.

Similar to the radial profile results, the resolution plot shows the same trend. Among the optimal shower generators, OPTIMUM (x1) performs the worst, showing significant deviation from the GEANT4 baseline. This demonstrates that the coarse $\mathcal{R}_{\times 1}$ representation lacks sufficient granularity to accurately capture the intrinsic energy fluctuations of electromagnetic showers. Despite being derived from full GEANT4 simulation, this representation inherently limits the achievable fidelity due to the loss of fine spatial information.

With increased granularity, OPTIMUM (x9) shows improvement, more closely tracking the GEANT4 resolution. Finally, OPTIMUM (STEPS), which directly uses the individual GEANT4 steps without any spatial discretization, comes closest to reproducing the full simulation, representing the maximum achievable performance.

Importantly, the two generative models follow the trends of their respective representations. The CALOCLOUDS3 model, trained on de-quantized $\mathcal{R}_{\times 25}$, almost reaches the performance of OPTIMUM (STEPS). CONVL2LFLOWS trained on the regular $\mathcal{R}_{\times 9}$ data representation, performs comparably to OPTIMUM (x9), although the deviations from the optimal representation are larger than for CALOCLOUDS3.

These findings demonstrate that both models have successfully learned from their respective training data representations, achieving performance that closely matches the optimal shower generators derived from the same representation.

The linearity is reproduced well by all models being within a $\sim$ 3% relative deviation from GEANT4, over the entire energy range tested – from 10 to 100 GeV with 10 GeV steps. Notably, the CALOCLOUDS3 model exhibits the best agreement, showing negligible deviation from GEANT4. This is a result of a simple scaling applied during integration, where each generated shower is rescaled by a constant factor determined from the difference in average visible energy at 50 GeV between the model and GEANT4.

In principle, similar scaling could be applied to all models. However, for CONVL2LFLOWS, and any grid-based model in general, such calibration is more challenging due to its restricted generation region, where an increasing fraction the shower's energy leaks out with increasing energy of the incident particle. As a result, the discrepancy between generated and true visible shower energy becomes more pronounced at higher energies, limiting the effectiveness of global rescaling.

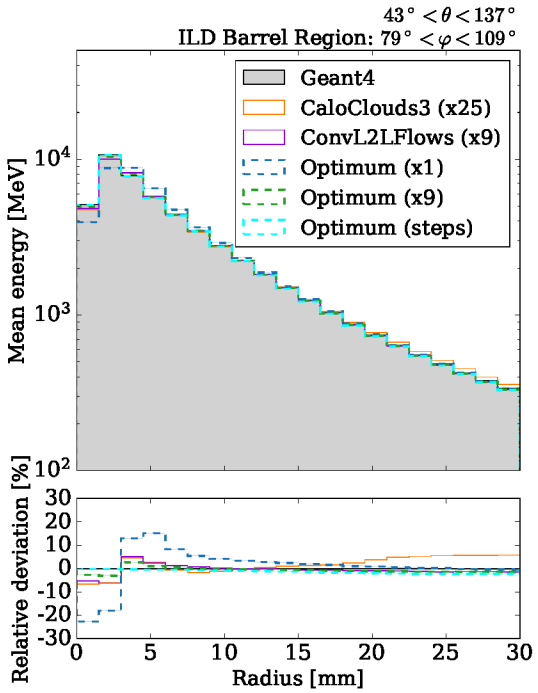


FIG. 4. Radial energy profile of the showers, zoomed in to the first 30 mm from the shower axis. The shaded error bands correspond to the statistical uncertainty in each bin. The lower subplot shows the relative deviation of the radial energy profile with respect to the GEANT4 reference. The color coding is consistent with Figure 3.

This highlights a fundamental trade-off faced by grid-based models – they must balance performance against computational efficiency. Increasing the size of the generation volume can improve accuracy by reducing energy leakage, but it also significantly increases memory usage and inference time. As a result, achieving high fidelity with grid-based models requires careful tuning of the generation volume to remain computationally feasible while minimizing physical artifacts.

#### 3. *Intrinsic Angle Reconstruction*

In a similar fashion to previous studies using the BIBAE model [24, 66], the previous state-of-the-art generative model applied to the ILD detector, we evaluate the angular response of the generative models by comparing the reconstructed intrinsic angles of showers simulated with GEANT4 to those generated by the models. A principal component analysis (PCA) is applied to all reconstructed hits of each shower to determine its principal axis. The resulting angular distributions of the azimuthal and polar angles, denoted as *iPhi* and *iTheta*, respectively, are presented as the difference between the reconstructed and true angles of the incoming particle direction. These distributions are shown in Figure 6.

As shown in Figure 6, both GEANT4 and the generative models produce angular distributions that are centered around zero with comparable widths for both *iPhi* and *iTheta*, indicating that the models qualitatively reproduce the angular response observed in detailed simulation.

The CALOCLOUDS3 and optimal shower generators, however, show a tendency to overestimate the polar angle, resulting in a double-peaked structure in the *iTheta* distribution. This indicates that this double-peak effect is likely related to the methodology of placing showers generated in the regularized detector into the real detector readout (see Appendix C). This effect is more pronounced for CALOCLOUDS3, and is likely compounded by the fact that CALOCLOUDS3 slightly overestimates the energy for larger radii (i.e. further from the shower axis).

CONVL2LFLOWS on the other hand shows a noticeably sharper peak in both the polar and azimuthal angle distributions. This observation, combined with the fact that CONVL2LFLOWS generates showers within a tightly constrained spatial region, as described in Section III A, suggests that a simple PCA applied to all hits in the shower may be a suboptimal procedure for reconstructing the intrinsic shower angle. This is likely due to the high sensitivity of PCA to hits located far from the shower core, which can disproportionately influence the estimated principal axis.

To address this, we apply an energy-based hit selection, retaining only the top 4% most energetic hits before running PCA. This suppresses noise from peripheral hits and improves the stability of the reconstructed direction, meaning that this approach results in an improved angular reconstruction algorithm[4]. The results are shown in Figure 7. All optimum shower generators, including CALOCLOUDS3 and CONVL2LFLOWS, now demonstrate angular resolutions that closely match the GEANT4 reference. It is notable that the distribution produced by the OPTIMUM (X1) shower generator shows the largest discrepancy, producing a slightly wider distribution than GEANT4. The improved reconstruction shows better alignment with GEANT4, and the double-peak structure previously observed in the polar angle distribution disappears entirely.

### B. Timing Benchmark

To quantify the speed advantage of the fast simulators in the full ILD software chain, we measure the wall-clock time per shower on a single CPU core of an AMD EPYC™ 7402. The measurements for GEANT4 and for all generative models were taken on the same machine and software setup. As a result of the integration of the models using the DDML library described in Appendix A, all timing measurements are directly comparable to those of GEANT4, thanks to the use of an identical software configuration, including overheads such as hit placement in the detector geometry. This enables us to perform a fair and realistic timing benchmark, which is not possible without model integration.

Figure 8 (left, top) reports the time per shower as a function of photon energy 10-100 GeV. The GEANT4 baseline lies

[4] For more details on this reconstruction improvement, see [61]

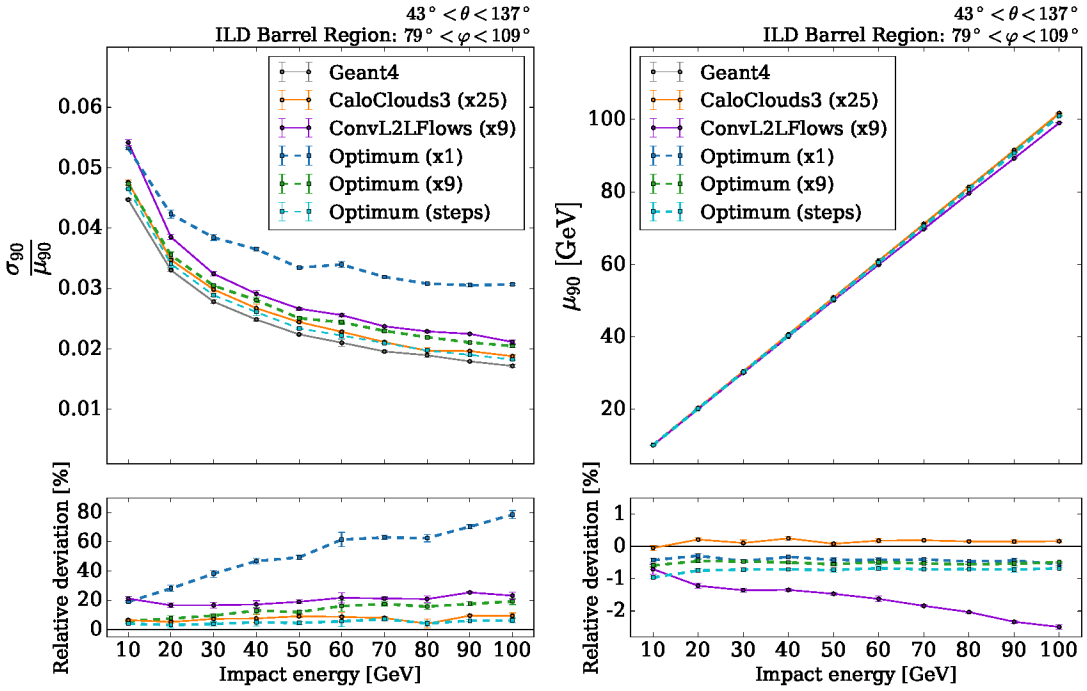


FIG. 5. Energy resolution (left) and linearity (right) of reconstructed photon showers in the ILD ECAL. The resolution is defined as the relative width $\sigma_{90}/\mu_{90}$ of the central 90% interval of the reconstructed energy distribution. The linearity is given by the mean $\mu_{90}$ of this central interval as a function of the incident photon energy.

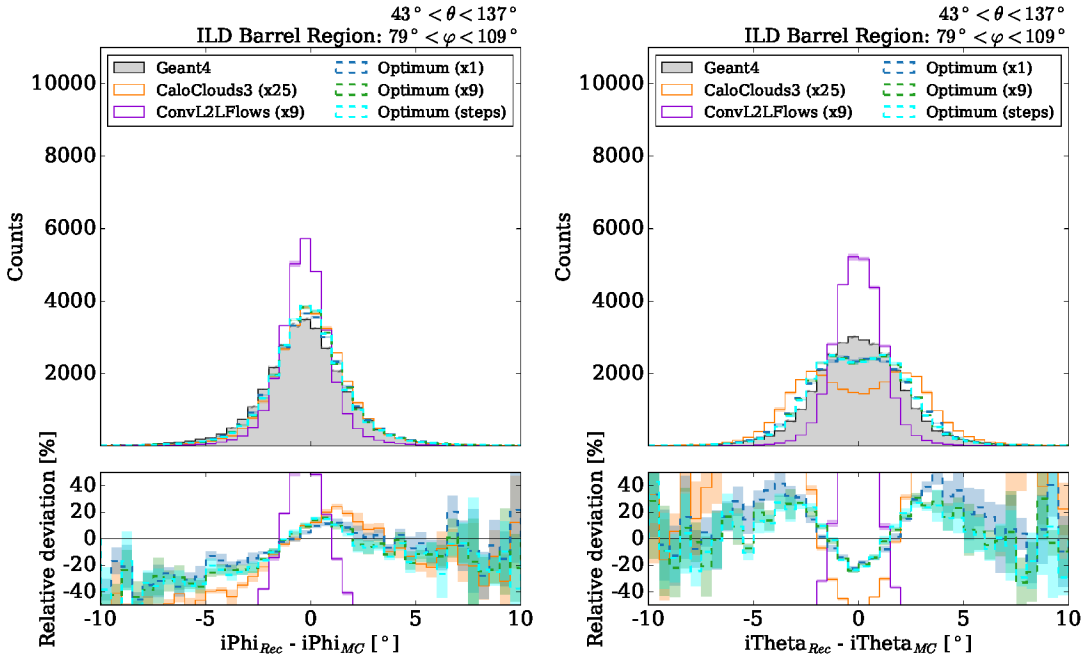


FIG. 6. Distributions of the differences between reconstructed and true intrinsic angles for showers in the ILD barrel region ($43° < \theta < 137°$, $79° < \varphi < 109°$), comparing GEANT4 to various generative models. Left: azimuthal angle (*iPhi*). Right: polar angle (*iTheta*). The top panel shows the angle residual distributions, while the bottom panel presents the relative deviation with respect to GEANT4. PCA is applied to all reconstructed hits to extract the principal axis of the shower.

at the few second level and grows linearly with energy, reflecting the increasing number of interaction steps. CALOCLOUDS3 shows a similar trend, but with a much smaller slope. By contrast, the grid-based model CONVL2LFLOWS, shows no energy

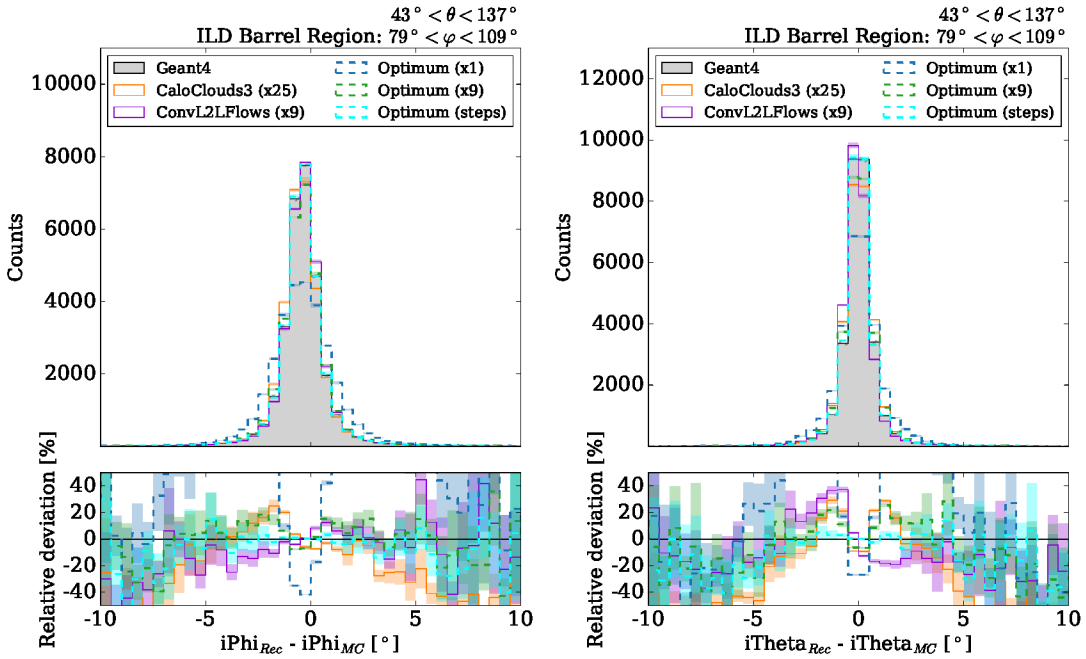


FIG. 7. Distributions of the differences between reconstructed and true intrinsic angles for showers in the ILD barrel region ($43° < \theta < 137°$, $79° < \varphi < 109°$), comparing GEANT4 to various generative models. Left: azimuthal angle (*iPhi*). Right: polar angle (*iTheta*). The top panel shows the angle residual distributions, while the bottom panel presents the relative deviation with respect to GEANT4. Here, only the top 4% most energetic hits are used for PCA-based extraction of the shower axis, resulting in significantly improved angular resolution compared to using all hits.

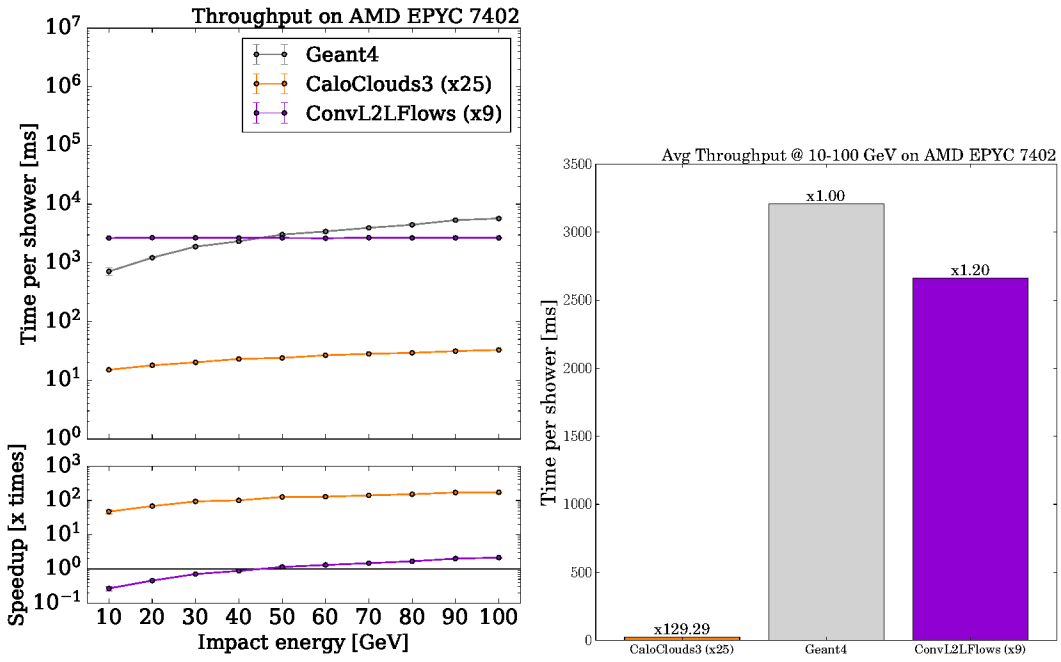


FIG. 8. Single particle timing on a single core of an AMD EPYC™ 7402 CPU within the full ILD simulation chain. Left: wall-clock time per shower (top) and speed-up vs. GEANT4 (bottom) for photons with incident energies in the range of 10 to 100 GeV with 10 GeV steps. Right: average single core throughput, with the speed-up relative to GEANT4 highlighted.

dependence and remains flat across the full range.

Complementary to the time per shower as a function of photon energy, Fig. 8 (right) shows the average single core throughput over 10-100 GeV, normalized to the GEANT4 baseline (×1). We observe single-shower simulation speed-up factors of ×129.29 for CALOCLOUDS3, and ×1.20 for CONVL2LFLOWS across this range of incident photon energies. The striking gain for CALOCLOUDS3 reflects its lightweight point-wise inference.

Overall, the scaling behavior is favorable for grid-based models as their inference time is constant with respect to shower energy because the computational cost is fixed by the voxelized volume. As a consequence, the relative speed-up over GEANT4 grows with energy. At the low-energy end of the spectrum, where GEANT4 showers contain fewer steps, the absolute latency gap narrows and the grid models can approach GEANT4 in wall-clock time. Toward higher energies, where GEANT4 scales approximately linearly with the number of interaction steps, the flat cost of CONVL2LFLOWS (and other grid models such as the BIBAE [19, 21, 24] becomes increasingly advantageous. By contrast, the point-cloud based CALOCLOUDS3 shows a mild energy dependence, with its simulation time rising with energy as the number of generated points grows with incident energy. However, the slope remains well below that of GEANT4, yielding a more uniform though less asymptotically dramatic speed-up across the full range. For completeness, we note that model initialization (loading weights, caching etc.) is excluded from the timings shown, as including it only affects small sample sizes and does not change the observed scaling trends.

In addition to inference speed, the memory footprint and model size are important practical factors for deployment within large-scale simulation workflows. The compiled CALOCLOUDS3 model has a total weight size of approximately 27 MB, corresponding to a memory footprint of 198 MB during inference. By contrast, the grid-based CONVL2LFLOWS model is substantially larger, with a compiled weight size of 2.5 GB and a peak memory footprint of about 4.4 GB. This reflects the higher parameter count and larger activation maps inherent to convolutional grid-based architectures.

While the scaling behavior is favorable for grid-based models – their inference time is essentially constant with shower energy because the compute is fixed by the voxelized volume. They also exhibit larger physics performance deviations at higher energies due to leakage from the bounded generation volume. As energy increases, a growing fraction of the shower can reach the box boundary and leak out, degrading containment and biasing observables (see Sec. IV A 2 and the single particle profiles in Sec. IV A). Mitigations such as expanding the generation volume reduce leakage but increase memory footprint and latency, reducing part of the speed advantage.

### C. Di-photon Benchmark

The di-photon benchmark allows the performance of simulators to be studied in a scenario where multiple showers overlap. This emphasizes the relevance of particular shower characteristics that are not directly probed by studying single shower observables, while maintaining a controlled environment that prevents any contamination which may be present in a realistic physics process. To this end, the number of reconstructed photons is plotted against the separation between the two photons in Figure 9, for symmetric di-photon energies of 5 GeV (left), 20 GeV (middle) and 100 GeV (right). The performance is shown for GEANT4, CALOCLOUDS3, CONVL2LFLOWS, OPTIMUM (x1), OPTIMUM (x9). Note that OPTIMUM (STEPS) is not included in order to aid visibility in the plot, as it performed on a similar level to GEANT4. The error bars represent the binomial error in each case. The red line present in each plot represents two photons being reconstructed on average, which is the optimal case for this reconstruction scenario.

For all energies, at separations of less than approximately 6 mm, only a single photon is reconstructed on average. This corresponds to slightly more than one cell's worth of separation, meaning that the two individual shower cores cannot be resolved, with all hits being clustered into a single photon. In addition, it is easier for two photons to be resolved as their incident energy is increased. For photon pairs with incident energies of 100 GeV, the average number of reconstructed photons quickly rises to two for separations of only 10–20 mm, while for incident photon pairs with energies of 5 GeV the average number of reconstructed photons rises much slower, and only reaches an average of two for separations of around 60–65 mm. This is because incident photons with a higher energy have a higher energy core, which is also more densely populated than for lower energy photon showers, making them easier to distinguish [85].

Due to the relative simplicity of separating higher energy photons, both models and the optimal shower generators agree well with GEANT4 in terms of the distribution of the average number of reconstructed photons for photon pairs with incident energies of 100 GeV. Relative deviations in this instance appear around the level of a few percent, with the largest differences arising for the CONVL2LFLOWS model. At lower photon pair energies of 5 GeV and 20 GeV, more significant deviations occur. Here, the OPTIMUM (x1) shower generator shows significant discrepancies across a large range of separations, reaching relative deviations around the 10–15% level respectively. Both the CALOCLOUDS3 and the CONVL2LFLOWS models show more contained relative deviations, typically less than 5%, with the most noticeable exception being for the CALOCLOUDS3 model at separations of around 10-20 mm where the relative deviation slightly exceeds this level.

These results indicate that training directly on the detector readout causes major discrepancies in the reconstruction performance. This is likely due to artefacts created when placing hits back into the detector geometry. The radial profile is of particular importance when separating such showers, as the reconstruction is especially sensitive to how the profiles of the two showers interfere. The effects observed around the core of the radial profile for photon showers produced with OPTIMUM (x1) in Figure 4 are therefore considered a major factor in the poor performance of this generator. The differences observed for CONVL2LFLOWS and CALOCLOUDS3 are linked to

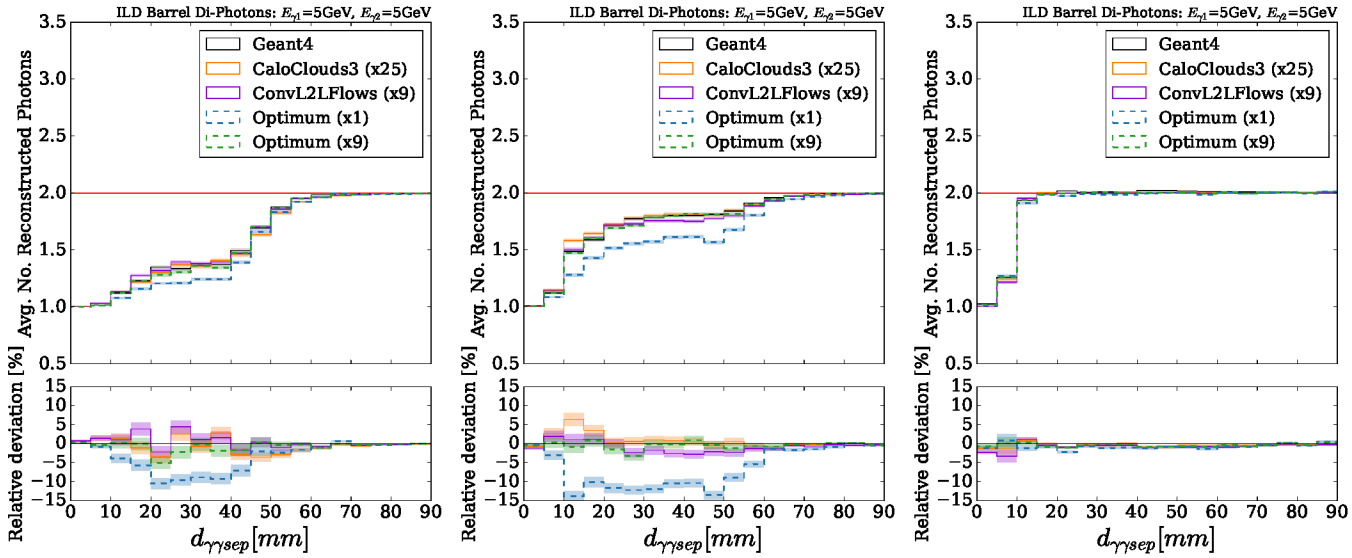


FIG. 9. Average number of reconstructed photons against the di-photon separation for identical incident photon pair energies of 5 GeV (left), 20 GeV (middle) and 100 GeV (right). Curves are shown for GEANT4 (light gray), the CALOCLOUDS3 (orange) and CONVL2LFLOWS (violet) models and the OPTIMUM (x1) (blue) and OPTIMUM (x9) (green) generators. Note that OPTIMUM (STEPS) is not included to aid visibility, as the performance was found to be comparable to that of GEANT4.

the deviations in the radial profile observed in Figure 3 (left), although these are less influential as they occur further out from the shower core. The steep fall in the radial profile of the CONVL2LFLOWS model, predicated by the constrained cut required for a regular grid, model seems to have little influence on the reconstruction performance in this test, as at large separations it is easy to distinguish the showers from their core alone.

### D. Full Physics Benchmark

We now study the post-reconstruction performance of the various simulation approaches for the full physics benchmark based on $\pi^0$s produced in hadronic decays of the tau lepton. These results are split into an investigation into the global performance of the $\pi^0$ reconstruction in the process $e^+e^- \to \tau^+\tau^-$ in Section IV D 1, followed by a study of the modeling of key $\pi^0$ physics observables in Section IV D 2.

#### 1. *Global Reconstruction Performance*

We begin by studying the overall quality of the reconstruction of $\pi^0$s produced in the tau decays for each shower simulator in comparison to GEANT4. For this evaluation, it is necessary to create a relation between Monte Carlo (MC) Truth particles and reconstructed particles. The weight of each relation is determined by the energy weighted contribution of each MC Truth particle to a reconstructed particle. If multiple MC Truth particles have a relation to a given reconstructed particle (or vice versa), the one with the highest weight is taken.

We now use these relations to define four different categories of reconstruction quality into which a given reconstructed $\pi^0$ may fall. Firstly, we define nGood as the number of correctly reconstructed $\pi^0$s. Secondly, we define nFake as the number of $\pi^0$s reconstructed without being linked to an MC Truth $\pi^0$. Thirdly, nConfused is defined as the number of $\pi^0$s reconstructed where only one of the constituent photons was reconstructed correctly. Finally nMissed represents the number of MC Truth $\pi^0$s for which there was no reconstructed $\pi^0$. We also calculate the total number of $\pi^0$s reconstructed, nRecoPi0. nFake is constrained on a per-event basis by the equation nFake = nRecoPi0 − nGood − nConfused. The results for each of the optimum generators OPTIMUM (x1), OPTIMUM (x9), OPTIMUM (STEPS), together with the CONVL2LFLOWS and CALOCLOUDS3 models, are shown in Figure 10. In each case, the results are plotted in comparison to the GEANT4 result. For each category, the results show the average over the 3 different GEANT4 random seeds, with the error calculated from the standard deviation across seeds.

For the optimum generators, the cell-level generator OPTIMUM (x1) shows the clearest mismatches. Large deviations from the GEANT4 samples are observed in all categories except nFake. This includes on average almost 10% fewer $\pi^0$s being reconstructed in total. On average more than 10% fewer $\pi^0$s are correctly reconstructed, while many more $\pi^0$s are reconstructed incorrectly (nConfused is on average more than 10% greater, while nMissed is on average more than 5% larger) with respect to GEANT4. It should be noted that while nConfused exhibits a high relative deviation, nConfused itself is low meaning that the absolute deviation is small.

Increasing the granularity of the resolution drastically decreases the deviations with respect to the GEANT4 samples, with the OPTIMUM (x9) generator showing much closer agree-

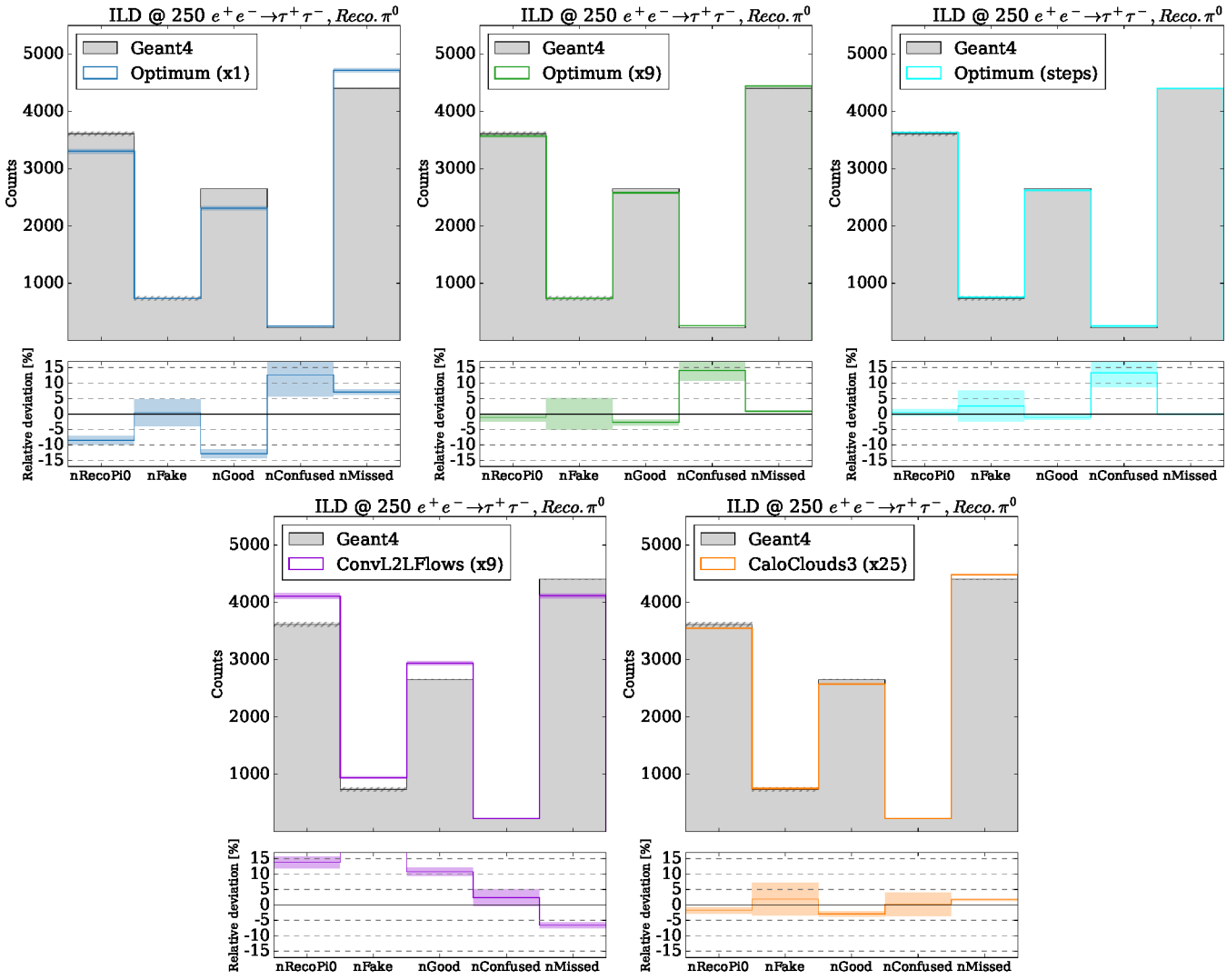


FIG. 10. Overall reconstruction quality of $\pi^0 s$ produced by tau decays in the process $e^+e^- \to \tau^+\tau^-$ for the Optimum (x1) (top left, blue), Optimum (x9) (top middle, green) and Optimum (steps) (top right, cyan) generators, and the ConvL2LFlows (bottom left, violet) and CaloClouds3 (bottom right, orange) models in comparison to Geant4 (grey). The quality of the reconstruction is characterized by five different categories; nRecoPi0, the total number of $\pi^0$s reconstructed, nFake, the number of $\pi^0$s reconstructed without a corresponding $\pi^0$ in the MC Truth, nGood, the number of correctly reconstructed $\pi^0$s, nConfused, the number of $\pi^0$s reconstructed only partially correctly and nMissed, the number of MC Truth-level $\pi^0$s that were not reconstructed. The errors on each category are derived from the standard deviation across three different random seeds used for the detector simulation. In each case, the lower panel represents the relative deviation from Geant4 in each category.

ment. However, deviations are still clearly visible, with on average fewer nGood $\pi^0$s relative to Geant4 and a similar excess in nConfused as was observed for Optimum (x1).

At the step-level resolution in Optimum (steps), the average for each category is consistent with Geant4 within error except for nConfused. Again, the deviation of nConfused relative to Geant4 is of a similar magnitude to that found for the Optimum (x1) and Optimum (x9) representations. These results demonstrate that increasing the granularity of the representation past that of the detector readout results in improved physics performance for this use case.

Turning to the performance of the generative models, ConvL2LFlows shows major differences to Geant4 in a large number of categories, including relative deviations in excess of 10% for nRecoPi0 and nGood, and in excess of 15% for nFake. The deviation with respect to Geant4 for nMissed exceeds the 5% level, while nConfused shows the closest agreement with a relative deviation of only a few percent. These deviations are consistent with the significant discrepancies observed for single particle observables produced with the ConvL2LFlows model in Section IV A.

By contrast, the CaloClouds3 model shows much closer agreement with the Geant4 baseline, with all categories showing relative deviations only at the level of a few percent. While at first appearance it may be surprising that CaloClouds3 is able to achieve better performance than Optimum (steps), for

example in the nConfused observable, it should be noted that an extra calibration was applied to the CaloClouds3 model as described in Section IV A 2.

### 2. Physics Observable Performance

We now study the performance of the individual simulators in terms of key $\pi^0$ physics observables. In order to perform a fair comparison, we require that all $\pi^0$s selected have been correctly reconstructed (i.e those that fell into the nGood category described in Section IV D 1). As before, the results show the average over 3 different Geant4 random seeds, with the error calculated from the standard deviation across seeds. It should be noted that some of these distributions are already constrained by the selection of $\pi^0$s which fall into the nGood category. This is due to quality criteria imposed by the $\pi^0$ reconstruction procedure, which includes performing a constrained kinematic fit [86].

Firstly, we calculate the invariant mass of the reconstructed $\pi^0$, $M_{\pi^0 Rec}$, which is equivalent to the invariant mass of the di-photon system $M_{\gamma\gamma}$ given by

$$M_{\gamma\gamma} = \sqrt{2E_iE_j(1-\cos(\eta))}, \tag{1}$$

where $E_i$ is the reconstructed energy of photon $i$, $E_j$ is the reconstructed energy of photon $j$, and $\eta$ is the opening angle between their reconstructed flight directions. The invariant mass distributions for Geant4, each of the optimum generators, and both the ConvL2LFlows and CaloClouds3 models are shown in Figure 11. All models show broad agreement within the stated uncertainties around the bulk of the distribution. Larger relative deviations appear for all optimum shower generators and both models in the tails of the distribution, although the increasing errors on the ratio make the exact discrepancy less clear.

The next observable studied is the difference between the reconstructed energy of the $\pi^0$ and the energy of the corresponding MC particle, $E_{\pi^0 Rec} - E_{\pi^0 MC}$, which is shown in Figure 12. It can be seen that in comparison to Geant4 the reconstructed energy tends to be slightly biased towards lower energies than that of the MC particle, although this deviation is heavily suppressed by the large magnitude of the uncertainties. This effect appears to be strongest for Optimum (steps) and ConvL2LFlows, which correlates with the shifts in linearity for single photons observed for these approaches in Section IV A.

Finally, the reconstructed angular differences for the $\pi^0$ in both the $\theta$, $\theta_{\pi^0 Rec} - \theta_{\pi^0 MC}$, and $\phi$, $\phi_{\pi^0 Rec} - \phi_{\pi^0 MC}$, directions are shown in Figure 12. For the distribution in $\theta$, the clearest deviation occurs for the Optimum (x1) representation at the level of the detector readout, which produces a noticeably broader distribution with relative deviations quickly exceeding the 30% level away from the core of the distribution. Clear mismodelings from the ConvL2LFlows model are also present in both the $\theta$ and $\phi$ distributions, with relative deviations around the 20% level around the core of the distributions.

## V. DISCUSSION

Developing tools for the fast simulation of showers in highly granular calorimeters is essential to be able to meet the demands of future collider experiments. We have introduced the DDML library for integrating generative models for fast calorimeter shower simulation into the DD4hep toolkit, enabling model benchmarking in a production-ready software suite and providing access to reconstruction-level physics benchmarks. Two different generative models for fast calorimeter simulation, ConvL2LFlows and CaloClouds3, were integrated into this library. This enabled realistic timing benchmarks of model performance, as well as post reconstruction benchmarks performed using the actual detector geometry. By comparing results using shower representations of different granularities, we were able to disentangle methodological details related to dataset construction from the actual performance of the generative models. Furthermore, we have presented new reconstruction-level benchmarks for the evaluation of generative models designed for electromagnetic shower simulation in highly granular calorimeters, including a first multi-particle benchmark involving di-photon separations, and a full physics benchmark based on hadronic decays of the tau lepton in the process $e^+e^- \to \tau^+\tau^-$.

We have demonstrated that building a dataset which uses a shower representation directly at the level of the detector readout granularity results in significant distortions in key physics observables when using realistic detector geometries. These distortions are visible in all levels of post reconstruction observables presented – from single particle observables through to higher level $\pi^0$ observables in the process $e^+e^- \to \tau^+\tau^-$ that would propagate directly through to down-stream analysis[5]. As the quality of shower representation fundamentally limits the maximum achievable performance of a generative model trained on a given dataset, it is essential that datasets use optimized representations that operate on shower information at a lower level than the detector readout. This remains true beyond highly granular calorimeters, as demonstrated by recent work which optimized a voxelized representation for photons in the barrel region of the calorimeter of the ATLAS experiment, showing significant improvement over the current fast simulation tool [55]. While the ATLAS calorimeter system has a significantly lower granularity than that studied in this work, combined they emphasize the importance of producing optimized representations across the field of fast calorimeter simulation.

In order to be able to perform these optimizations, it is essential that fast simulation models are integrated back into the software ecosystems used in particle physics experiments-both to study placement into the actual detector geometry and to gain access to reconstruction level observables. The DDML library introduced as part of this work provides this functionality for detector geometries implemented in DD4hep. Given

[5] For a summary of all of the observables studied in this paper, see Appendix B

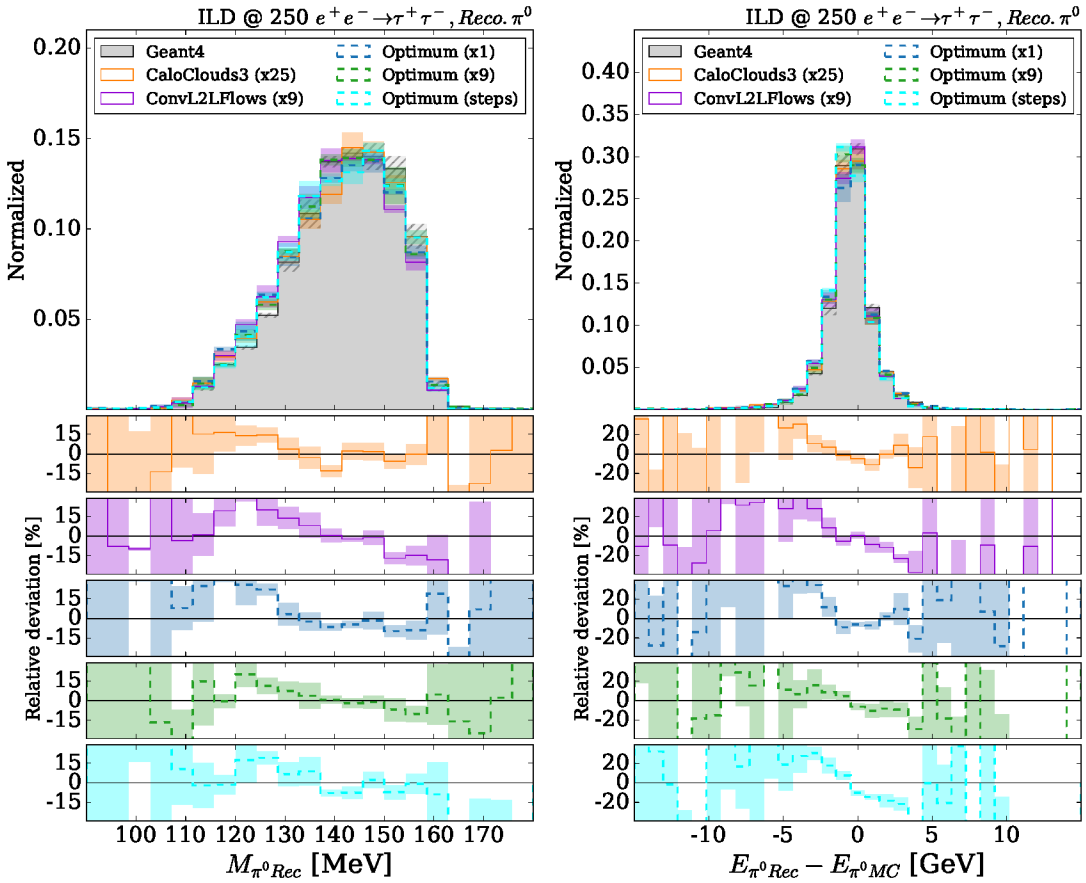


FIG. 11. Key physics observables for $\pi^0$s produced by tau decays in the process $e^+e^- \to \tau^+\tau^-$ : the reconstructed $\pi^0$ invariant mass $M_{\pi^0 Rec}$ distribution (left), and the difference between the reconstructed energy of the $\pi^0$ $E_{\pi^0 Rec}$ and the corresponding MC Truth $\pi^0$ $E_{\pi^0 MC}$ (right). Each of the optimal shower generators OPTIMUM (x1) (blue), OPTIMUM (x9) (green) and OPTIMUM (STEPS) (cyan), as well as the two generative models CALOCLOUDS3 (orange) and CONVL2LFLOWS (violet) are shown, with the GEANT4 reference shown in grey. In both cases, the errors are derived from the standard deviation across three different random seeds used for the detector simulation. The lower panels in each case shows the relative deviation form GEANT4.

that this library is designed to be generic, the addition of new ML models and detector geometries should be as straightforward as possible. While this work focused on the ILD detector, other detectors proposed for future colliders such as the FCC-ee are already supported in the library, including both the CLD detector [58] and the ALLEGRO detector [87].

The generative models included in this study are trained on two distinct data representations. CONVL2LFLOWS is a model designed to operate on a fixed grid structure, while the CALOCLOUDS3 model is point cloud based. Due to the limitations of training directly on the detector readout, as previously discussed, both models are trained on a more granular shower representation. This requirement has greater impact on the CONVL2LFLOWS model, which has to be trained on a representation with a restricted granularity due to the limitations imposed by the use of a regular grid. As a result of using a regular grid of a higher granularity than the detector readout, the CONVL2LFLOWS model shows a poor simulation throughput increase for single particles relative to GEANT4, as well as showing several deviations in observables, including for higher-level physics quantities. A significant factor in the larger deviations observed for the CONVL2LFLOWS model for single particle observables results from the constrained bounding box necessitated by the used of a regular grid. This is a major reason why this approach to generative modeling exhibited consistently larger deviations than the corresponding OPTIMUM (x9) generator, which included a far less restrictive bounding box. By contrast, the CALOCLOUDS3 model is able to operate on a more granular, and therefore more accurate shower representation. Despite this, the model is able to achieve more than two orders of magnitude faster simulation throughput than GEANT4 for single showers with energies between $10 - 100$ GeV on a single CPU core. This highlights that a point cloud representation of showers can not only offer a more efficient solution than a regular grid, but also enables a superior speed-accuracy trade-off for highly granular calorimeter simulation in realistic applications. While there are some deviations present in CALOCLOUDS3 observables, in particular for single particle showers, these are suppressed in subsequent reconstruction for the physics case studied. This emphasizes that while studying model performance at the level of single showers provides useful model insight, model

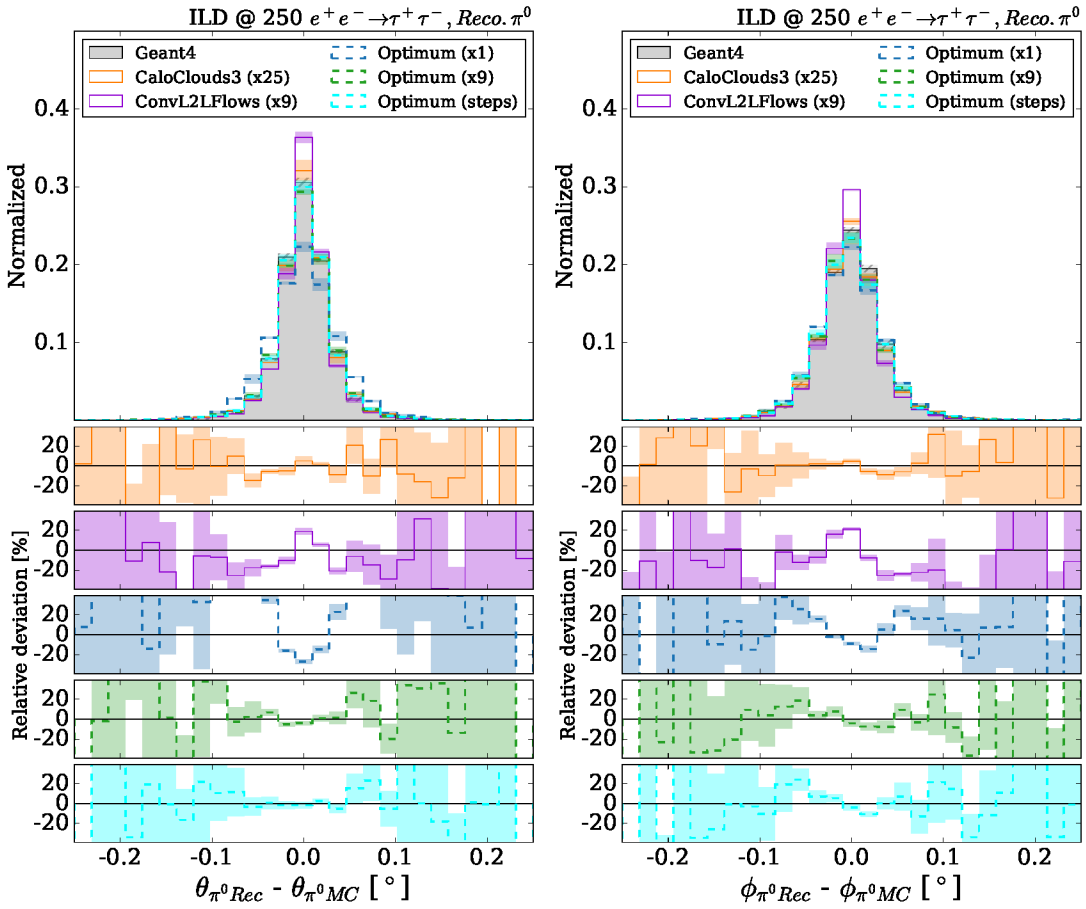


FIG. 12. Reconstructed angular deviations for $\pi^0 s$ produced by tau decays in the process $e^+e^- \to \tau^+\tau^-$: the difference between the reconstructed $\theta_{\pi^0 Rec}$ and MC Truth $\theta_{\pi^0 MC}$ direction in $\theta$ (left), and the difference between the reconstructed $\phi_{\pi^0 Rec}$ and MC Truth $\phi_{\pi^0 MC}$ direction in $\phi$ (right). Each of the optimal shower generators OPTIMUM (x1) (blue), OPTIMUM (x9) (green) and OPTIMUM (STEPS) (cyan), as well as the two generative models CALOCLOUDS3 (orange) and CONVL2LFLOWS (violet) are shown, with the GEANT4 reference shown in gray. In both cases, the errors are derived from the standard deviation across three different random seeds used for the detector simulation. The lower panels in each case shows the relative deviation form GEANT4.

performance must ultimately be judged on post-reconstruction level physics observables. This also means that the relative importance of certain shower features depends on the target down-stream analysis. For example, the deviations observed in the intrinsic cluster angles[6] in Section IV A may be more relevant for certain searches for physics beyond the standard model [88].

It should be noted that some deviations remain, even for the OPTIMUM (STEPS) generator (see for example the resolution plot in Figure 5), fixing an upper limit for CALOCLOUDS3 performance. These deviations are caused by shower placement into isolated regions in the geometry which feature greater irregularities – more details are presented in Appendix C. While the reconstructed energy here could be corrected by dedicated calibrations, improving the modeling of shower structure in these regions would require the addition of dedicated model trainings.

This work has studied the performance of generative models for highly granular electromagnetic shower simulation for the case of photons. Future work could investigate the case of electron or positron showers. As these particles are charged, they also require track-cluster associations to be performed as part of the reconstruction, and this may add increased importance to particular shower observables. In order to achieve significantly faster simulation throughput at the level of full physics events for detectors with highly granular calorimeters in general, it will be necessary to address the challenge of highly granular hadron shower simulation. Work in [51] has recently demonstrated that the use of a diffusion-transformer mechanism in a point cloud model can provide an accurate modeling of hadron showers, including for single particle observables at the post-reconstruction level. However, further work is still required to develop a model which can achieve a significantly faster simulation throughput, as well as to develop benchmarks similar to the ones shown here targeted at

[6] The extent to which these deviations appear for CALOCLOUDS3 depend on the approach used to determine the intrinsic angle. See [61] for more details.

hadronic showers.

## ACKNOWLEDGMENTS

We thank Dirk Krücker for valuable comments on the manuscript. This research was supported in part by the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 101004761. We acknowledge support by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC 2121 Quantum Universe – 390833306 and via the KISS consortium (05D23GU4, 13D22CH5) funded by the German Federal Ministry of Research, Technology and Space (BMFTR) in the ErUM-Data action plan. A.K. has received support from the Helmholtz Initiative and Networking Fund's initiative for refugees as a refugee of the war in Ukraine. P.M. has benefited from support by the CERN Strategic R&D Programme on Technologies for Future Experiments [89]. Some figures and text in this paper have previously appeared in a doctoral thesis [66].

[1] J. Albrecht *et al.* (HEP Software Foundation), A Roadmap for HEP Software and Computing R&D for the 2020s, Comput. Softw. Big Sci. **3**, 7 (2019), arXiv:1712.06982 [physics.comp-ph].

[2] A. Boehnlein, C. Biscarat, A. Bressan, D. Britton, R. Bolton, F. Gaede, C. Grandi, F. Hernandez, T. Kuhr, G. Merino, F. Simon, and G. Watts, *HL-LHC Software and Computing Review Panel, 2nd Report*, Tech. Rep. CERN-LHCC-2022-007, LHCC-G-183 (CERN, Geneva, 2022).

[3] M. Paganini, L. de Oliveira, and B. Nachman, Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters, Phys. Rev. Lett. **120**, 042003 (2018), arXiv:1705.02355 [hep-ex].

[4] M. Paganini, L. de Oliveira, and B. Nachman, CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks, Phys. Rev. D **97**, 014021 (2018), arXiv:1712.10321 [hep-ex].

[5] L. de Oliveira, M. Paganini, and B. Nachman, Controlling Physical Attributes in GAN-Accelerated Simulation of Electromagnetic Calorimeters, J. Phys. Conf. Ser. **1085**, 042017 (2018), arXiv:1711.08813 [hep-ex].

[6] M. Erdmann, L. Geiger, J. Glombitza, and D. Schmidt, Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks, Comput. Softw. Big Sci. **2**, 4 (2018), arXiv:1802.03325 [astro-ph.IM].

[7] M. Erdmann, J. Glombitza, and T. Quast, Precise simulation of electromagnetic calorimeter showers using a Wasserstein Generative Adversarial Network, Comput. Softw. Big Sci. **3**, 4 (2019), arXiv:1807.01954 [physics.ins-det].

[8] F. Carminati, A. Gheata, G. Khattak, P. Mendez Lorenzo, S. Sharan, and S. Vallecorsa, Three dimensional Generative Adversarial Networks for fast simulation, J. Phys. Conf. Ser. **1085**, 032016 (2018).

[9] P. Musella and F. Pandolfi, Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks, Comput. Softw. Big Sci. **2**, 8 (2018), arXiv:1805.00850 [hep-ex].

[10] D. Belayneh *et al.*, Calorimetry with deep learning: particle simulation and reconstruction for collider physics, Eur. Phys. J. C **80**, 688 (2020), arXiv:1912.06794 [physics.ins-det].

[11] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, and T. Plehn, GANplifying event samples, SciPost Phys. **10**, 139 (2021), arXiv:2008.06545 [hep-ph].

[12] ATLAS collaboration, *Fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks*, Tech. Rep. ATL-SOFT-PUB-2020-006 (CERN, Geneva, 2020).

[13] A. Ghosh (ATLAS), Deep generative models for fast shower simulation in ATLAS, J. Phys. Conf. Ser. **1525**, 012077 (2020).

[14] G. Aad *et al.* (ATLAS), AtlFast3: The Next Generation of Fast Simulation in ATLAS, Comput. Softw. Big Sci. **6**, 7 (2022), arXiv:2109.02551 [hep-ex].

[15] G. Aad *et al.* (ATLAS), Deep Generative Models for Fast Photon Shower Simulation in ATLAS, Comput. Softw. Big Sci. **8**, 7 (2024), arXiv:2210.06204 [hep-ex].

[16] B. Hashemi, N. Hartmann, S. Sharifzadeh, J. Kahn, and T. Kuhr, Ultra-high-granularity detector simulation with intra-event aware generative adversarial network and self-supervised relational reasoning, Nature Commun. **15**, 4916 (2024), [Erratum: Nature Commun. 115, 5825 (2024)], arXiv:2303.08046 [physics.ins-det].

[17] M. Faucci Giannelli and R. Zhang, CaloShowerGAN, a generative adversarial network model for fast calorimeter shower simulation, Eur. Phys. J. Plus **139**, 597 (2024), arXiv:2309.06515 [physics.ins-det].

[18] E. Simsek, B. Isildak, A. Dogru, R. Aydogan, A. B. Bayrak, and S. Ertekin, CALPAGAN: Calorimetry for Particles Using Generative Adversarial Networks, PTEP **2024**, 083C01 (2024), arXiv:2401.02248 [hep-ex].

[19] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, and K. Krüger, Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed, Comput. Softw. Big Sci. **5**, 13 (2021), arXiv:2005.05334 [physics.ins-det].

[20] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, and K. Krüger, Decoding Photons: Physics in the Latent Space of a BIB-AE Generative Network, EPJ Web Conf. **251**, 03003 (2021), arXiv:2102.12491 [physics.ins-det].

[21] E. Buhmann, S. Diefenbacher, D. Hundhausen, G. Kasieczka, W. Korcari, E. Eren, F. Gaede, K. Krüger, P. McKeown, and L. Rustige, Hadrons, better, faster, stronger, Mach. Learn. Sci. Tech. **3**, 025014 (2022), arXiv:2112.09709 [physics.ins-det].

[22] J. C. Cresswell, B. L. Ross, G. Loaiza-Ganem, H. Reyes-Gonzalez, M. Letizia, and A. L. Caterini, CaloMan: Fast generation of calorimeter showers with density estimation on learned manifolds, in *36th Conference on Neural Information Processing Systems: Workshop on Machine Learning and the Physical Sciences* (2022) arXiv:2211.15380 [hep-ph].

[23] S. Bieringer, A. Butter, S. Diefenbacher, E. Eren, F. Gaede, D. Hundhausen, G. Kasieczka, B. Nachman, T. Plehn, and M. Trabs, Calomplification — the power of generative calorimeter models, JINST **17** (09), P09028, arXiv:2202.07352 [hep-ph].

[24] S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Ko-

rol, K. Krüger, P. McKeown, and L. Rustige, New angles on fast calorimeter shower simulation, Mach. Learn. Sci. Tech. **4**, 035044 (2023), arXiv:2303.18150 [physics.ins-det].

[25] S. Hoque, H. Jia, A. Abhishek, M. Fadaie, J. Q. Toledo-Marín, T. Vale, R. G. Melko, M. Swiatlowski, and W. T. Fedorko, CaloQVAE: Simulating high-energy particle-calorimeter interactions using hybrid quantum-classical generative models, Eur. Phys. J. C **84**, 1244 (2024), arXiv:2312.03179 [hep-ex].

[26] Q. Liu, C. Shimmin, X. Liu, E. Shlizerman, S. Li, and S.-C. Hsu, Calo-VQ: Vector-Quantized Two-Stage Generative Model in Calorimeter Simulation (2024), arXiv:2405.06605 [physics.ins-det].

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

[28] C. Krause and D. Shih, Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation, Phys. Rev. D **107**, 113004 (2023), arXiv:2110.11377 [physics.ins-det].

[29] S. Schnake, D. Krücker, and K. Borras, Generating calorimeter showers as point clouds, in *Machine Learning and the Physical Sciences, Workshop at the 36th conference on Neural Information Processing Systems (NeurIPS)* (2022).

[30] C. Krause, I. Pang, and D. Shih, CaloFlow for CaloChallenge dataset 1, SciPost Phys. **16**, 126 (2024), arXiv:2210.14245 [physics.ins-det].

[31] S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, C. Krause, I. Shekhzadeh, and D. Shih, L2LFlows: generating high-fidelity 3D calorimeter images, JINST **18** (10), P10017, arXiv:2302.11594 [physics.ins-det].

[32] A. Xu, S. Han, X. Ju, and H. Wang, Generative machine learning for detector response modeling with a conditional normalizing flow, JINST **19** (02), P02003, arXiv:2303.10148 [hep-ex].

[33] M. R. Buckley, C. Krause, I. Pang, and D. Shih, Inductive simulation of calorimeter showers with normalizing flows, Phys. Rev. D **109**, 033006 (2024), arXiv:2305.11934 [physics.ins-det].

[34] I. Pang, D. Shih, and J. A. Raine, Calorimeter shower superresolution, Phys. Rev. D **109**, 092009 (2024), arXiv:2308.11700 [physics.ins-det].

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

[36] S. Schnake, D. Krücker, and K. Borras, CaloPointFlow II Generating Calorimeter Showers as Point Clouds (2024), arXiv:2403.15782 [physics.ins-det].

[37] H. Du, C. Krause, V. Mikuni, B. Nachman, I. Pang, and D. Shih, Unifying simulation and inference with normalizing flows, Phys. Rev. D **111**, 076004 (2025), arXiv:2404.18992 [hep-ph].

[38] T. Buss, F. Gaede, G. Kasieczka, C. Krause, and D. Shih, Convolutional L2LFlows: generating accurate showers in highly granular calorimeters using convolutional normalizing flows, JINST **19** (09), P09003, arXiv:2405.20407 [physics.ins-det].

[39] J. Birk, F. Gaede, A. Hallin, G. Kasieczka, M. Mozzanica, and H. Rose, OmniJet-$\alpha$_C: learning point cloud calorimeter simulations using generative transformers, JINST **20** (07), P07007, arXiv:2501.05534 [hep-ph].

[40] V. Mikuni and B. Nachman, Score-based generative models for calorimeter shower simulation, Phys. Rev. D **106**, 092009 (2022), arXiv:2206.11898 [hep-ph].

[41] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, W. Korcari, K. Krüger, and P. McKeown, CaloClouds: fast geometry-independent highly-granular calorimeter simulation, JINST **18** (11), P11025, arXiv:2305.04847 [physics.ins-det].

[42] F. T. Acosta, V. Mikuni, B. Nachman, M. Arratia, B. Karki, R. Milton, P. Karande, and A. Angerami, Comparison of point cloud and image-based models for calorimeter fast simulation, JINST **19** (05), P05003, arXiv:2307.04780 [cs.LG].

[43] V. Mikuni and B. Nachman, CaloScore v2: single-shot calorimeter shower simulation with diffusion models, JINST **19** (02), P02001, arXiv:2308.03847 [hep-ph].

[44] O. Amram and K. Pedro, Denoising diffusion models with geometry adaptation for high fidelity calorimeter simulation, Phys. Rev. D **108**, 072014 (2023), arXiv:2308.03876 [physics.ins-det].

[45] E. Buhmann, F. Gaede, G. Kasieczka, A. Korol, W. Korcari, K. Krüger, and P. McKeown, CaloClouds II: ultra-fast geometry-independent highly-granular calorimeter simulation, JINST **19** (04), P04020, arXiv:2309.05704 [physics.ins-det].

[46] C. Jiang, S. Qian, and H. Qu, Choose your diffusion: Efficient and flexible ways to accelerate the diffusion model in fast high energy physics simulation, SciPost Phys. **18**, 195 (2025), arXiv:2401.13162 [physics.ins-det].

[47] D. Kobylianskii, N. Soybelman, E. Dreyer, and E. Gross, Graph-based diffusion model for fast shower generation in calorimeters with irregular geometry, Phys. Rev. D **110**, 072003 (2024), arXiv:2402.11575 [hep-ex].

[48] C. Jiang, S. Qian, and H. Qu, BUFF: Boosted Decision Tree based Ultra-Fast Flow matching (2024), arXiv:2404.18219 [physics.ins-det].

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

[50] J. Brehmer, V. Bresó, P. de Haan, T. Plehn, H. Qu, J. Spinner, and J. Thaler, A Lorentz-Equivariant Transformer for All of the LHC (2024), arXiv:2411.00446 [hep-ph].

[51] T. Buss, F. Gaede, G. Kasieczka, A. Korol, K. Krüger, P. McKeown, and M. Mozzanica, CaloHadronic: a diffusion model for the generation of hadronic showers (2025), arXiv:2506.21720 [physics.ins-det].

[52] P. Raikwar, A. Zaborowska, P. McKeown, R. Cardoso, M. Piorczynski, and K. Yeo, A Generalisable Generative Model for Multi-Detector Calorimeter Simulation (2025), arXiv:2509.07700 [physics.ins-det].

[53] B. Hashemi and C. Krause, Deep generative models for detector signature simulation: A taxonomic review, Rev. Phys. **12**, 100092 (2024), arXiv:2312.09597 [physics.ins-det].

[54] O. Amram *et al.*, CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation (2024), arXiv:2410.21611 [physics.ins-det].

[55] *Photon showers in the ATLAS fast calorimeter simulation: A voxelized dataset with minimized information loss and improved ML models*, Tech. Rep. (CERN, Geneva, 2025).

[56] The CMS collaboration, *The Phase-2 Upgrade of the CMS Endcap Calorimeter*, Tech. Rep. (CERN, Geneva, 2017).

[57] H. Abramowicz *et al.* (ILD Concept Group), International Large Detector: Interim Design Report (2020), arXiv:2003.01116 [physics.ins-det].

[58] N. Bacchetta *et al.*, *CLD – A Detector Concept for the FCC-ee*, Tech. Rep. (2019) arXiv:1911.12230 [physics.ins-det].

[59] F. Gaede, T. Madlener, P. McKeown, T. Buss, A. Zaborowska, and A. Korol, key4hep/ddml: v0.2.0 (2025).

[60] M. Frank, F. Gaede, C. Grefe, and P. Mato, DD4hep: A Detector Description Toolkit for High Energy Physics Experiments, J. Phys. Conf. Ser. **513**, 022010 (2014).

[61] T. Buss, H. Day-Hall, F. Gaede, G. Kasieczka, K. Krüger,

A. Korol, T. Madlener, P. McKeown, M. Mozzanica, and L. Valente, Caloclouds3: Ultra-fast geometry-independent highly-granular calorimeter simulation (2025), arXiv:2511.01460 [physics.ins-det].

[62] J. Repond *et al.* (CALICE), Design and Electronics Commissioning of the Physics Prototype of a Si-W Electromagnetic Calorimeter for the International Linear Collider, JINST **3**, P08001, arXiv:0805.4833 [physics.ins-det].

[63] C. Adloff *et al.* (CALICE), Construction and Commissioning of the CALICE Analog Hadron Calorimeter Prototype, JINST **5**, P05004, arXiv:1003.2662 [physics.ins-det].

[64] V. Völkl *et al.*, The Key4hep turnkey software stack, PoS **ICHEP2022**, 234 (2022).

[65] S. Agostinelli *et al.* (GEANT4), GEANT4 - A Simulation Toolkit, Nucl. Instrum. Meth. A **506**, 250 (2003).

[66] P. McKeown, *Development and Performance of a Fast Simulation Tool for Showers in High Granularity Calorimeters based on Deep Generative Models*, Ph.D. thesis, Hamburg U., Hamburg (2024).

[67] J. S. Marshall and M. A. Thomson, The Pandora Software Development Kit for Pattern Recognition, Eur. Phys. J. C **75**, 439 (2015), arXiv:1506.05348 [physics.data-an].

[68] B. Xu, Improvement of photon reconstruction in PandoraPFA, in *International Workshop on Future Linear Colliders* (2016) arXiv:1603.00013 [physics.ins-det].

[69] M. Dam, The $\tau$ challenges at FCC-ee, Eur. Phys. J. Plus **136**, 963 (2021).

[70] L. Calibbi, X. Marcano, and J. Roy, Z lepton flavour violation as a probe for new physics at future $e^+e^-$ colliders, Eur. Phys. J. C **81**, 1054 (2021), arXiv:2107.10273 [hep-ph].

[71] S. Jahedi and A. Sarkar, Exploring optimal sensitivity of lepton flavor violating effective couplings at the e+e- colliders, Phys. Rev. D **110**, 095021 (2024), arXiv:2408.00190 [hep-ph].

[72] A. Gutiérrez-Rodríguez, C. Pérez-Mayorga, and A. González-Sánchez, Sensitivity Estimates on the Electromagnetic Dipole Moments of the $\tau$-Lepton at Future $e^+e^-$ Linear Colliders, Int. J. Theor. Phys. **61**, 132 (2022).

[73] J. de Blas *et al.*, Higgs Boson Studies at Future Particle Colliders, JHEP **01**, 139, arXiv:1905.03764 [hep-ph].

[74] T. H. Tran, V. Balagura, V. Boudry, J.-C. Brient, and H. Videau, Reconstruction and classification of tau lepton decays with ILD, Eur. Phys. J. C **76**, 468 (2016), arXiv:1510.05224 [physics.ins-det].

[75] D. Jeans, Tau lepton reconstruction at collider experiments using impact parameters, Nucl. Instrum. Meth. A **810**, 51 (2016), arXiv:1507.01700 [hep-ex].

[76] S. Berge, W. Bernreuther, and H. Spiesberger, Higgs CP properties using the $\tau$ decay modes at the ILC, Phys. Lett. B **727**, 488 (2013), arXiv:1308.2674 [hep-ph].

[77] D. Jeans and G. W. Wilson, Measuring the CP state of tau lepton pairs from Higgs decay at the ILC, Phys. Rev. D **98**, 013007 (2018), arXiv:1804.01241 [hep-ex].

[78] S. Navas *et al.* (Particle Data Group), Review of particle physics, Phys. Rev. D **110**, 030001 (2024).

[79] H. Ono and A. Miyamoto, Status of ILD new 250 GeV common MC sample production, in *International Workshop on Future Linear Colliders* (2021) arXiv:2105.06040 [physics.acc-ph].

[80] W. Kilian, T. Ohl, and J. Reuter, WHIZARD: Simulating Multi-Particle Processes at LHC and ILC, Eur. Phys. J. C **71**, 1742 (2011), arXiv:0708.4233 [hep-ph].

[81] S. Jadach, J. H. Kuhn, and Z. Was, TAUOLA: A Library of Monte Carlo programs to simulate decays of polarized tau leptons, Comput. Phys. Commun. **64**, 275 (1990).

[82] L. Dinh, D. Krueger, and Y. Bengio, NICE: Non-linear Independent Components Estimation (2014) arXiv:1410.8516 [cs.LG].

[83] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural Spline Flows (2019) arXiv:1906.04032 [stat.ML].

[84] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, LNCS, Vol. 9351 (Springer, 2015) pp. 234–241, arXiv:1505.04597 [cs.CV].

[85] B. Xu, *Detectors and Physics at a Future Linear Collider*, Ph.D. thesis, Cambridge U. (2017).

[86] B. List, J. List, and DESY, MarlinKinfit: An Object–Oriented Kinematic Fitting Package, LC Notes 10.3204/PHPPUBDB-10294 (2009).

[87] M. Mlynarikova, Design and performance of the calorimeter system for allegro fcc-ee detector concept, EPJ Web Conf. **320**, 00022 (2025).

[88] L. Lee, C. Ohm, A. Soffer, and T.-T. Yu, Collider Searches for Long-Lived Particles Beyond the Standard Model, Prog. Part. Nucl. Phys. **106**, 210 (2019), [Erratum: Prog.Part.Nucl.Phys. 122, 103912 (2022)], arXiv:1810.12602 [hep-ph].

[89] C. Joram *et al.*, *Extension of the R&D Programme on Technologies for Future Experiments*, Tech. Rep. (CERN, https://cds.cern.ch/record/2850809, 2023).

[90] J. Allison *et al.* (GEANT4), Par04 example, `https://github.com/Geant4/geant4/tree/master/examples/extended/parameterisations/Par04` (2025), accessed: 06.08.2025.

[91] A. Paszke *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems 32 , 8024 (2019), Advances in Neural Information Processing Systems 32 pp. 8024–8035.

[92] O. R. developers, ONNX Runtime, `https://onnxruntime.ai/` (2021), accessed: 30.12.2023.

[93] A. Boehnlein, C. Biscarat, A. Bressan, D. Britton, R. Bolton, F. Gaede, C. Grandi, F. Hernandez, T. Kuhr, G. Merino, F. Simon, and G. Watts, *HL-LHC Software and Computing Review Panel, 2nd Report*, Tech. Rep. (CERN, Geneva, 2022).

[94] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, Pyro: Deep Universal Probabilistic Programming, J. Mach. Learn. Res. **20**, 28:1 (2019).

## Appendix A: Integration into Standard Software Chains

In order to study the performance of the generative models described in Section III in a realistic physics simulation including event reconstruction, it is necessary for the models to be combined with the experiment's standard software ecosystems, typically implemented in C++. This appendix provides additional details on this generic library in Section A 1, and on the specific implementation for ILD in Section A 2.

### 1. The DDML Library

To integrate generative models for fast calorimeter simulation into full simulation applications, we introduce the DDML[7] library [59] as part of the Key4hep software stack. It follows the Geant4 Par04 example for running machine learning inference in fast simulation models [90]. Our library is built on top of the Geant4 and DD4hep toolkits, providing access to a broad suite of functionality. In particular, this includes an interface to a trigger mechanism present in Geant4 for terminating physics-based full simulation in favor of an alternative simulation approach. A trigger is associated with a particular geometrical region of the detector, and is activated if an impinging particle satisfies certain criteria (particle type, energy etc.). This provides a seamless way of incorporating a generative model based fast calorimeter simulation tool into a full simulation application.

The DDML library is designed to support a generic approach to fast simulation with generative models. This can be split into three key requirements:

1. Allow the use of different kinds of generative model, in terms of their structure, the inputs they require and the outputs they generate.

2. Allow different engines for model inference to be employed.

3. Allow the use of different detector geometries implemented in DD4hep.

This requires two conventions to be chosen for the library. The first is the adoption of a local (right-handed) coordinate system, defined such that the origin is placed at the point of entry into the calorimeter, with the $z'$ axis orthogonal to the calorimeter face, and pointing into the calorimeter system. The $x'$ axis is aligned with the direction of the magnetic field. This allows the model output to be handled in the same manner independent of where in the detector a particular particle is incident. The second is the interpretation of the model outputs as local space points in this coordinate system. This provides a generic means of interpreting the output of a model, independent of its architectural details.

[7] https://github.com/key4hep/DDML

The library is split into a number of different interfaces via a class template. This decouples the three aspects detailed above as far as possible, making it easier to extend and maintain the library. The interfaces are outlined below, with a class diagram of DDML shown in Figure 13, and the order of operations being shown in Algorithm 1.

**Algorithm1** Pseudocode illustrating the order of operations for the core components of the DDML library.

```
if Trigger.checkTrigger(track) == True then
    Kill full simulation of particle
    localDir = Geometry.getLocalDir(track)
    inputs = Model.prepareInputs(track, localDir)
    outputs = Inference.runInference(inputs)
    localSPs = Model.convertOutput(track, localDir, outputs)
    globalSPs = Geometry.localToGlobal(track, localSP)
    for (sp in globalSPs) do
        HitMaker.makeHit(sp, track)
    end for
else
    Full simulation of particle with Geant4
end if
```

**Trigger**
The Trigger Interface sits on top of the trigger on particle type and energy that exists in DD4hep and Geant4. This interface allows a particular fast simulation model to be excluded from running in certain regions of the detector (*checkTrigger*). This provides a simple means of handling regions of a calorimeter with an irregular structure, where either full simulation or a separate generative model can be run instead.

**Model**
The Model Interface is used to provide a model specific implementation. The role of this interface is two-fold. The first role is the preparation of the input (*prepareInputs*) in the form expected by the model. As part of this, a *localDir* object is available to provide information about the local direction at the calorimeter face. This can then be used as conditioning input for the generative model. The second role is the interpretation of the output of the model such that it can be converted into local space points (*convertOutput*).

**Inference**
The Inference Interface provides a simple means of calling the inference library for a model (*runInference*). Currently both the LibTorch [91] and OnnxRuntime [92] inference libraries are supported. Alternatively, functionality is provided for loading a pre-simulated shower library from a HDF5 file, which is intended for model prototyping and representation investigations.

**Geometry**
The Geometry Interface performs two separate roles. The first is to compute the local direction (*getLocalDir*), which is provided to the Model Interface as a consistent means of model conditioning. The second is to place the local space points produced as output of the Model Interface into the

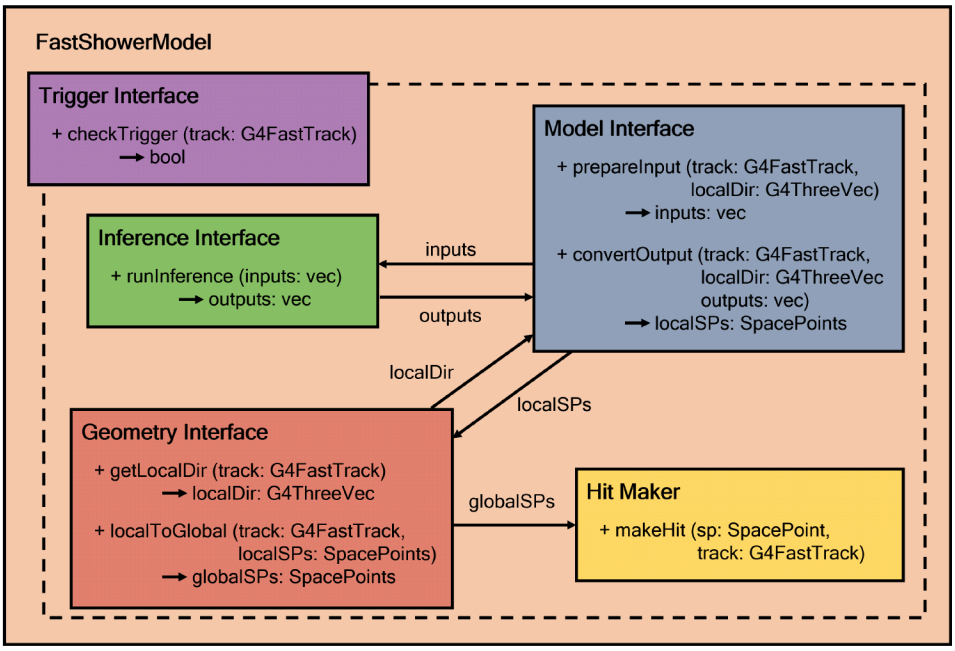


FIG. 13. Class diagram illustrating the core components of the DDML library. See the text for a detailed description of the interfaces. Figure from [66].

geometry of the detector (*localToGlobal*). This includes both the conversion from local calorimeter coordinates into global (envelope) coordinates, and the placement of hits onto sensitive detector elements. It must be implemented on a per-geometry basis, with disk endcap, polyhedral and cylindrical barrel calorimeter geometries already supported.

**HitMaker**
A helper class provided by GEANT4 to allow the placement of energy deposits produced by the fast simulation model, given that their position lies within a sensitive element of the detector (*makeHit*).

Currently, the DDML library only supports single shower generation with a generative model (inference with batch size of one) on a CPU, which still represents the dominant hardware available in high energy physics computing infrastructure [93]. Future support for batched shower generation and the addition of GPU support is foreseen. Such a development would then make the most significant simulation speed-ups relative to GEANT4 accessible via parallelization.

### 2. DDML Implementation for ILD

The DDML implementation for the ILD detector used in this study includes both the CALOCLOUDS3 and CONVL2LFLOWS models described in Section III. The models were converted to a format suitable for use in C++, with the components that composed the architectures being serialized by a combination of tracing and scripting each of the individual operations. This was achieved predominantly by using the utilities provided by TORCHSCRIPT in PYTORCH [91], with the SHOWERFLOW component of the CALOCLOUDS3 architecture making use of the POUTINE effect handlers present in the PYRO [94] deep probabilistic programming library in which the model was implemented.

This study focuses on the barrel region of the ILD detector. In order to leverage the symmetry present in the ILD detector, we generate showers at different positions in the detector using models trained on showers in a single location, as described in Section II. This is possible because the regularized calorimeter used for creating the training dataset has no dead material within an active layer, and no gaps. This means that the model is a valid simulator for the vast majority of the detector, while in regions which are particularly irregular, full simulation is run instead using the trigger interface described in Section A 1.

The first type of region excluded is the transition between the edge of the barrel and the endcap. In this region, there is a gap between the barrel and the endcap, and a change in orientation of the calorimeter layers. For this reason, the edges of the barrel at $\theta < 40$ degrees and $\theta > 140$ degrees in the global ILD coordinate system are excluded from fast simulation. The second type of transition region is the intersection between staves of the octagonal barrel, where an asymmetrical change in orientation of the calorimeter layers occurs. These 8 transition regions are excluded by 8 cuts in the global ILD coordinate system, each of which consists of an 8.01 degree range in $\varphi$.

### Appendix B: Observable Comparison Summary

Table I provides a concise overview of the agreement between the evaluated surrogate models and the GEANT4 reference across all benchmark observables. The comparison employs the Jensen-Shannon (JS) divergence and the mean absolute error ($L_1$) as quantitative metrics, both measuring the deviation of a model's reconstructed distributions from those obtained with GEANT4. Lower values correspond to better agreement.

Uncertainties are estimated from repeated evaluations using identical event samples simulated with different random seeds, for both GEANT4 and the surrogate models.

Among the reference optimal shower generators, OPTIMUM (x1)- trained directly at detector readout granularity- shows systematically larger deviations, indicating that coarse spatial representations limit achievable fidelity. Progressively finer geometrical resolution in OPTIMUM (x9) and OPTIMUM (STEPS) significantly improves the match, providing an effective upper bound on achievable performance for surrogates.

### Appendix C: Systematics resulting from simulation methodology

Figure 14 shows the average reconstructed energy pattern for photon showers, visualized as YZ projections of the mean energy per voxel across 90,000 events. The vertical direction Y corresponds to the calorimeter layers, while Z indicates the position within the upper ECAL barrel in mm. All optimal shower generators and generative models reproduce the overall longitudinal and transverse structure observed in the GEANT4 reference, including layer-dependent modulations and geometrical features of the detector segmentation. No significant deviations are visible at this scale, demonstrating that all models are properly integrated into the detector geometry and capture the overall event topology with high fidelity.

In contrast, the relative per-voxel energy difference maps shown in Fig. 15 reveal small but systematic biases. To the right of each panel, zoomed-in views highlight regions near module boundaries and absorber gaps. Because all surrogate models were trained on idealized geometries without inactive regions, they tend to overestimate the early energy deposition. The showers in the surrogate models start slightly earlier since the absorber material is present in every layer homogeneously in the idealized geometry. In the realistic ILD geometry, however, the presence of structural gaps in the absorber delays the shower onset, allowing the energy to penetrate deeper. This results in a characteristic underestimation of energy in the later layers for all surrogates when compared to the full GEANT4 simulation.

These effects originate from the interplay between the idealized local training setup and the complex global detector structure. They represent geometry-dependent systematic biases that persist even for the most accurate surrogate, OPTIMUM (STEPS), and define a practical upper limit on achievable agreement when a single conditional model is reused across the barrel.

Because the affected regions are spatially confined, extending the approach with localized trainings or lightweight region-specific calibrations offers a straightforward path to further reduce the remaining discrepancies.

| Metric | Optimum (x1) | Optimum (x9) | Optimum (steps) | ConvL2LFlows (x9) | CaloClouds3 (x25) |
|---|---|---|---|---|---|
| Single shower observables | | | | | |
| $\mathrm{JS}^{E_{\mathrm{radial}(100)}}$ ($\times 10^{-4}$) | 4.71 ± 0.04 | **0.04 ± 0.01** | 0.06 ± 0.01 | *8.37 ± 0.03* | 1.11 ± 0.02 |
| $\mathrm{JS}^{E_{\mathrm{radial}(30)}}$ ($\times 10^{-4}$) | *23.33 ± 0.16* | 0.46 ± 0.02 | **0.04 ± 0.01** | 1.92 ± 0.06 | 2.24 ± 0.06 |
| $\mathrm{JS}^{E_{\mathrm{long}}}$ ($\times 10^{-4}$) | **0.049 ± 0.013** | 0.050 ± 0.013 | **0.049 ± 0.013** | *0.313 ± 0.023* | 0.050 ± 0.011 |
| $\mathrm{JS}^{iPhi_{\mathrm{res}(100\%)}}$ ($\times 10^{-4}$) | **13.84 ± 1.58** | 22.72 ± 2.07 | 24.68 ± 2.02 | *342.18 ± 7.96* | 48.64 ± 3.62 |
| $\mathrm{JS}^{iPhi_{\mathrm{res}(4\%)}}$ ($\times 10^{-4}$) | *299.91 ± 7.08* | 9.33 ± 1.20 | **1.37 ± 0.53** | 6.58 ± 1.20 | 16.85 ± 1.93 |
| $\mathrm{JS}^{iTheta_{\mathrm{res}(100\%)}}$ ($\times 10^{-4}$) | 44.40 ± 3.15 | 37.72 ± 2.71 | **37.52 ± 2.82** | *553.91 ± 10.26* | 374.08 ± 8.83 |
| $\mathrm{JS}^{iTheta_{\mathrm{res}(4\%)}}$ ($\times 10^{-4}$) | *195.27 ± 5.54* | 13.23 ± 1.32 | **2.02 ± 0.66** | 39.12 ± 2.49 | 32.10 ± 2.16 |
| $\mathrm{L}_1{}^{\frac{\sigma_{90}}{\mu_{90}}}$ | *114.33 ± 5.43* | 29.84 ± 5.13 | **11.63 ± 5.08** | 49.31 ± 5.35 | 17.98 ± 5.26 |
| $\mathrm{L}_1{}^{\mu_{90}}$ | 0.25 ± 0.04 | 0.29 ± 0.03 | 0.40 ± 0.03 | *1.06 ± 0.04* | **0.09 ± 0.04** |
| Multi-particle observables | | | | | |
| $\mathrm{JS}^{\gamma\gamma_{rec.}}$ ($\times 10^{-6}$)@5GeV | *182.69 ± 27.87* | **18.09 ± 10.28** | – | 28.52 ± 13.56 | 36.30 ± 14.21 |
| $\mathrm{JS}^{\gamma\gamma_{rec.}}$ ($\times 10^{-6}$)@20GeV | *380.74 ± 32.14* | **10.54 ± 6.04** | – | 19.56 ± 8.86 | 34.21 ± 12.79 |
| $\mathrm{JS}^{\gamma\gamma_{rec.}}$ ($\times 10^{-6}$)@100GeV | 5.57 ± 2.94 | **2.17 ± 1.24** | – | *7.28 ± 4.48* | 2.95 ± 1.77 |
| $\mathrm{JS}^{M_{\pi^0}}$ ($\times 10^{-4}$) | 23.12 ± 7.38 | **13.91 ± 6.89** | 25.83 ± 8.58 | *32.06 ± 8.74* | 22.08 ± 7.40 |
| $\mathrm{JS}^{E_{\pi^0}}$ ($\times 10^{-4}$) | 31.40 ± 9.93 | **19.62 ± 6.74** | *35.88 ± 8.02* | 27.24 ± 8.22 | 26.87 ± 8.82 |
| $\mathrm{JS}^{\theta_{\pi^0 res}}$ ($\times 10^{-4}$) | *143.62 ± 16.41* | 10.97 ± 4.94 | **7.04 ± 4.27** | 34.97 ± 7.63 | 10.69 ± 4.90 |
| $\mathrm{JS}^{\phi_{\pi^0 res}}$ ($\times 10^{-4}$) | 25.07 ± 6.98 | 13.90 ± 6.05 | 16.11 ± 6.38 | *49.74 ± 10.14* | **8.68 ± 4.91** |
| $\mathrm{JS}^{\pi^0_{rec}}$ ($\times 10^{-4}$) | 9.90 ± 1.66 | 0.73 ± 0.28 | **0.49 ± 0.33** | *12.75 ± 2.3* | 0.51 ± 0.28 |

TABLE I. Quantitative comparison of models' performance relative to Geant4 for single photon, di-photon and tau samples. Metrics are JS divergence or $L_1$ distance as indicated, lower values correspond to better agreement. Bold entries denote the best agreement with Geant4 per observable, italics denote the worst.

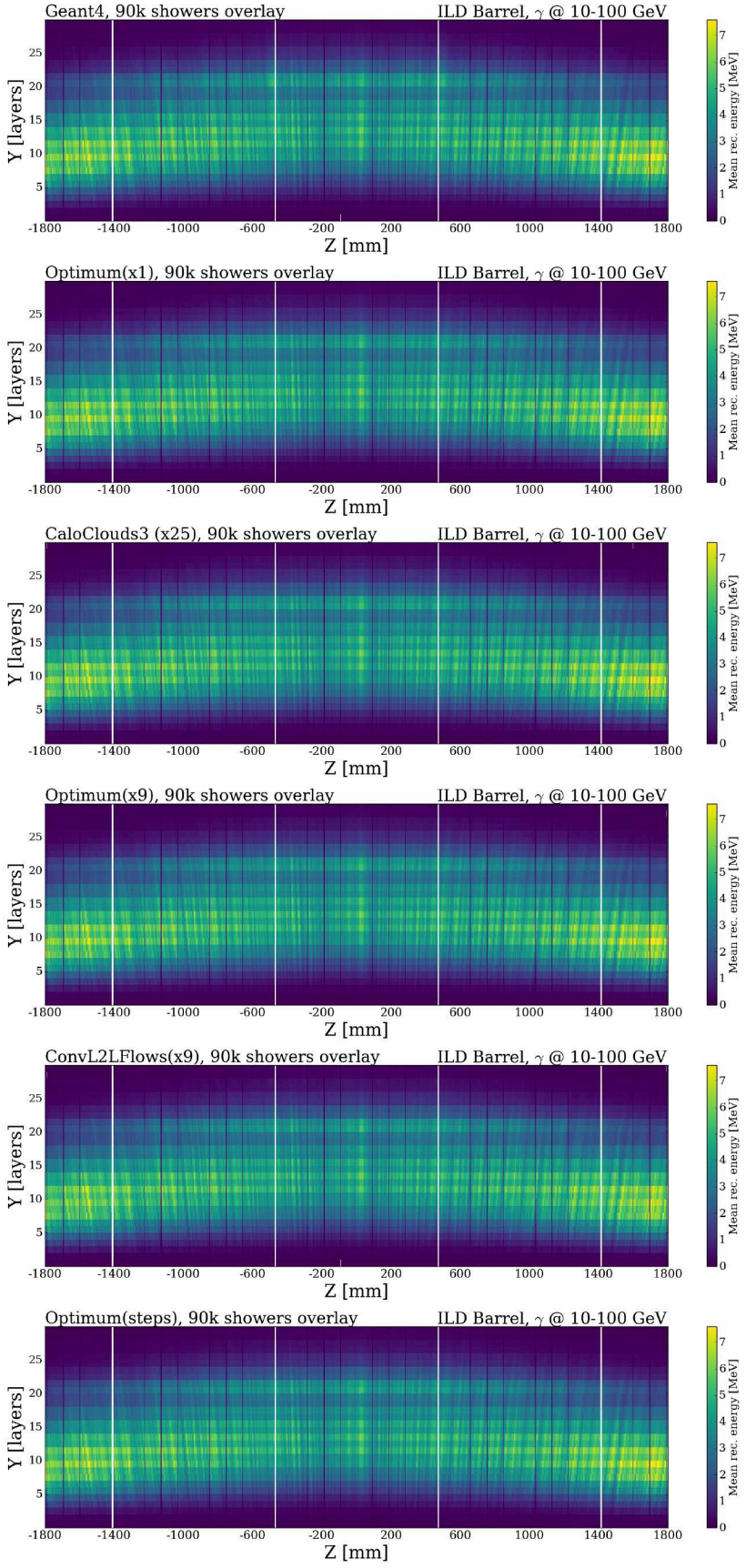


FIG. 14. Average reconstructed energy of 90,000 photon showers incident on the upper barrel segment of the ILD ECAL, projected onto the YZ plane. Each panel shows one generator or model (name of the model in the title of each panel), with Y denoting the calorimeter depth (layer index) and Z denoting the global detector coordinate in mm. The color scale corresponds to the mean deposited energy per voxel.

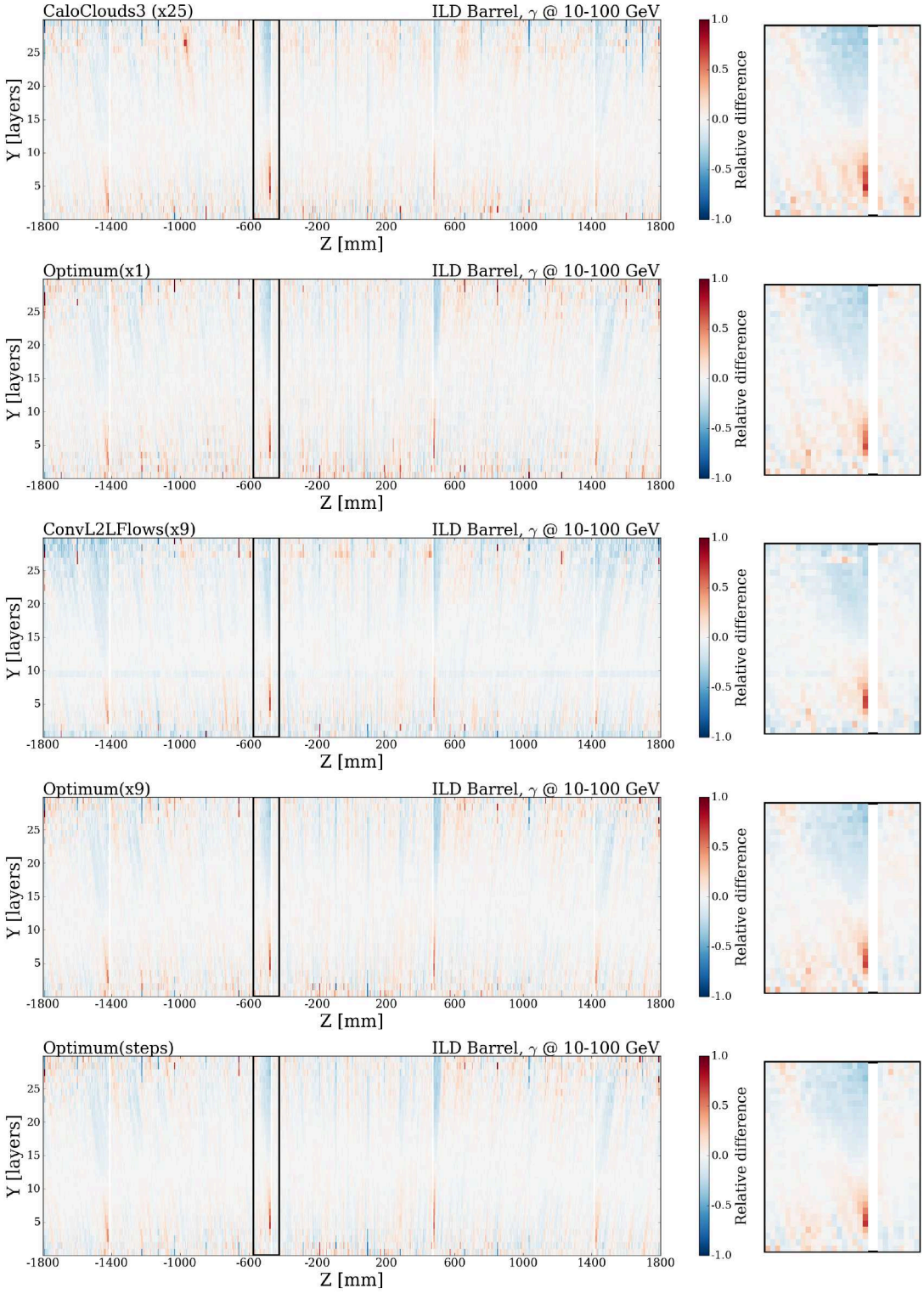


FIG. 15. Relative per-voxel deviation in mean reconstructed energy in the ILD ECAL with respect to GEANT4 for each surrogate model (name of the model in the title of each panel). The accompanying zoomed-in panels (on the right of each panel) show the regions near calorimeter module gaps. The consistent overestimation at shower onset and underestimation at larger depths arise from the absence of absorber gaps in the idealized training geometry. These localized geometry-dependent effects set a practical limit on surrogate accuracy when transferring models trained at a single reference position to the full ILD geometry.