Hadrons, better, faster, stronger

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 025014 https://doi.org/10.1088/2632-2153/ac7848

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
15 February 2022

REVISED
2 May 2022

ACCEPTED FOR PUBLICATION
13 June 2022

PUBLISHED
1 July 2022

PAPER

Hadrons, better, faster, stronger

Erik Buhmann[1], Sascha Diefenbacher[1,*], Daniel Hundhausen[1], Gregor Kasieczka[1], William Korcari[1], Engin Eren[2,*], Frank Gaede[2], Katja Krüger[2], Peter McKeown[2] and Lennart Rustige[3]

[1] Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany
[2] Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany
[3] Center for Data and Computing in Natural Sciences and Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany
[*] Authors to whom any correspondence should be addressed.

E-mail: sascha.daniel.diefenbacher@uni-hamburg.de **and** engin.eren@desy.de

Keywords: machine learning, Wasserstein generative adversarial networks, bounded information bottleneck autoencoder, generative models, calorimeter simulation, high energy physics

Abstract

Motivated by the computational limitations of simulating interactions of particles in highly-granular detectors, there exists a concerted effort to build fast and exact machine-learning-based shower simulators. This work reports progress on two important fronts. First, the previously investigated Wasserstein generative adversarial network and bounded information bottleneck autoencoder generative models are improved and successful learning of hadronic showers initiated by charged pions in a segment of the hadronic calorimeter of the International Large Detector is demonstrated for the first time. Second, we consider how state-of-the-art reconstruction software applied to generated shower energies affects the obtainable energy response and resolution. While many challenges remain, these results constitute an important milestone in using generative models in a realistic setting.

1. Introduction

Precise simulations of interactions between fundamental particles and complex detectors are a prerequisite for carrying out modern particle physics research. The high computational cost of these simulations is well established [1] and has—sparked by [2]—resulted in a large scale effort to develop surrogate simulations based on generative machine learning models. Trained on a small initial dataset—produced either using classical, Monte-Carlo-based simulators such as Geant4 [3] or potentially taken from data—these generators amplify the effectively available statistics [4].

Motivated by the large fraction of resources already consumed by calorimeter simulation [5], and the expected increase due to higher granularities and luminosities, the precise and fast simulation of calorimeters is a primary topic of research [2, 6–23][4].

In the quest to simulate calorimeters with high accuracy, standard generative machine learning methods—generative adversarial networks (GANs) [26], variational autoencoders (VAEs) [27] and normalizing flows [28–31]—have all been considered. However, several obstacles need to be overcome before these tools can be deployed to simulate highly granular calorimeters with high resolution. In [14, 15] we showed a precise modeling of differential distributions over many orders of magnitude for electromagnetic showers. The present work extends this level of precision for the first time to the more challenging hadron-induced showers in a highly granular hadronic calorimeter.

Another important aspect is the downstream processing of generated showers. While so far the focus has been on the *raw* output of the generative model, in a realistic environment, the generated energy deposits are

[4] For reviews, including also other applications of generative models to particle physics, see [24, 25].

processed by several reconstruction steps. We investigate the quality of generated showers after processing with state-of-the-art reconstruction algorithms[5].

The remainder of this paper is organized as follows: in section 2 we introduce the concrete problem, the training data, and the particle flow reconstruction algorithm, followed by the considered network architectures in section 3. Section 4 provides a quantitative evaluation of generated showers and section 5 gives conclusions and an outlook.

2. Data and reconstruction

We first introduce the dataset (section 2.1) of hadron showers generated for this study. As generated showers are compared both using the raw distributions as well as considering the output of standard particle-flow-based reconstruction algorithms, these are discussed in section 2.2.

2.1. Dataset

The International Large Detector (ILD) [32] is a proposed next generation particle detector at the International Linear Collider. It features highly granular sampling calorimeters optimized for the use of the particle flow reconstruction scheme. In sampling calorimeters the incident particle energy is measured by recording the energy that is deposited (visible) by secondary shower particles in the sensitive layers, which are placed alternating with insensitive absorber layers. The incident particle is destroyed in this shower process.

Particle showers occurring in these calorimeters are the simulation targets. However, unlike in previous work, which focused on photon showers in the SiW ECal [14], we now consider showers initiated by positively charged pions in the highly-granular analogue hadron calorimeter (AHCal). It consists of 48 layers with stainless steel absorber plates and scintillator tiles of $3 \times 3\,\mathrm{cm}^2$ individually read out by silicon photomultipliers (SiPMs) as active material. The AHCal concept has been developed by the CALICE Collaboration [33], and extensive beam tests [34] have allowed a thorough validation of the simulation with real data. For this study, the ECal part of the calorimeter system, lying in front of the AHCal, is removed in order to avoid the additional complexity of handling different detector geometries and materials. While the shape and behavior of photon showers is almost exclusively governed by electromagnetic interactions, charged pion showers also include hadronic interactions. This results in a much greater variety of shower topologies, as is illustrated in figure 1. The larger variety presents an increased challenge for generative networks.

While we focus on the ILD calorimeter, the generative methods that are investigated are not specific to the ILD. Therefore all methods should be adaptable to highly granular calorimeter setups of other detectors such as the high-granularity calorimeter of CMS (HG-Cal) or those of potential future circular collider (FCC) detectors.

ILD uses the iLCSoft [35] software ecosystem for detector simulation, reconstruction and analysis. For the full simulation with Geant4 [3], a detailed and realistic detector model implemented in DD4hep [36] is used. The training data consists of pion showers in the AHCal, which are simulated with Geant4 version 10.4[6] and DD4hep version 1.11. The iLCSoft ecosystem is fully containerized and simulation jobs are deployed into a Kubernetes [37] cluster hosted at DESY.

A virtual *particle gun* is placed at $(x', y', z') = (3\,\mathrm{cm}, 100\,\mathrm{cm}, 100\,\mathrm{cm})$, shooting charged pions pointing along the y'-axis. Together with a strong axial magnetic field of 3.5 T, this leads to a roughly perpendicular incident angle onto the AHCal which is 106 cm away from the pion gun. The *primed* coordinate system of ILD is oriented such that the z'-coordinate points along the beam axis and the y'-axis points horizontally upwards. For the resulting training data, a coordinate system with the z-axis pointing along the depth of the calorimeter is used. While the ECal is removed for the current study, the tracking system, which has a very low thickness in terms of radiation and interaction lengths, is not removed. Combined simulation of the ECal and HCal will be considered in future work.

Our training data set consists of 500k charged pion showers with incident particle energies uniformly distributed between 10 and 100 GeV, and—for each fixed incident energy—a constant impact position and angle[7]. We project these calorimeter hits into a regular grid of $x \times y \times z = 25 \times 25 \times 48$ pixels. Here, the z-axis is parallel to the particle trajectory and the 48 layers of the tensor correspond to the 48 layers of the calorimeter. The grid is centered such that high-energy pions arrive at $(x, y, z) = (12, 12, 0)$, whereas lower

[5] Different from e.g. [10], the goal here is not to replace reconstruction algorithms with another network, but to probe whether 'seamless' integration of showers generated using a generative model into standard workflows would be possible.

[6] Using the QGSP_BERT physics list.

[7] The training and test datasets are available under https://doi.org/10.5281/zenodo.6491116.

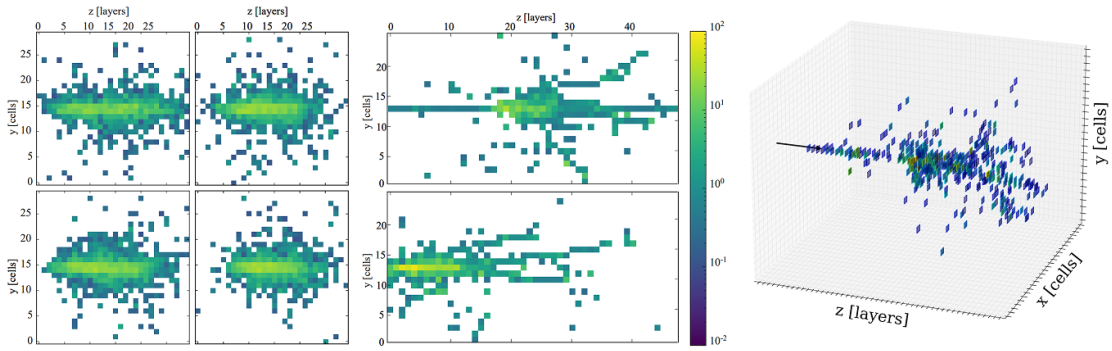

Figure 1. Left: four example images of photon showers. Center: two examples of pion showers. Right: 3D image of a pion shower. On average the pion showers have significantly more variance in their shapes.

energetic particles (i.e. 10–20 GeV) enter at slightly shifted positions due to the magnetic field. The size of the grid in the transverse direction was chosen as a trade-off between containing most of the deposited energy and keeping the image size (sparsity) from being too high (low). For a 40 GeV pion, on average 96% of the deposited energy is captured by the choice of 25×25. We further correct for any artifacts caused by the slightly irregular calorimeter structure such that each cell in this grid corresponds to exactly one sensor. In order to remove particles that pass through the detector without showering, we reject any shower with less than 70 hits above 0.25 MeV. This requirement removes approximately 1% of events. In addition to the training set, we generate an independent test set to compare with the showers produced by the generative models. This consists of 49k showers with uniformly distributed particle energies. While this is equivalent to only about 10% of the training set, we found this to be sufficient to describe the marginal distributions we are interested in reproducing. Additionally, for each particle energy in steps of 10 GeV from 20 to 90 GeV, smaller single-energy samples of 8k showers are produced.

2.2. Particle flow reconstruction

Typical calorimeters at future e^+e^- colliders are optimized for particle flow reconstruction. This means exploiting the high granularity of calorimeters combined with measurements from tracking detectors to reconstruct all individual particles created in an event. As this will determine the physics performance that can ultimately be achieved, the quality of generated showers needs to be evaluated after passing through such a reconstruction algorithm. In the following, we consider the state-of-the-art pattern recognition particle flow algorithm PandoraPFA [38], as used by ILD.

The detailed full simulation with Geant4 produces hit objects which consist of energy deposits in individual calorimeter cells as well as their exact positions in space. In a digitization step [32], all effects such as readout electronics, light yield of the scintillator tiles, and the statistical effects of the number of SiPM pixels are taken into account and applied to these simulated hits. This is followed by a two step calibration procedure:

(a) The deposited energy is normalized to the most probable energy deposited by a minimum ionizing particle (MIP).
(b) This energy, in units of MIP, is converted into a total energy in GeV, such that the sum of all hit energies corresponds to the incident particle's energy.

After digitization and calibration, PandoraPFA is run to cluster the digitized calorimeter hits by iteratively applying a number of sophisticated clustering algorithms, using information of reconstructed charged particle tracks where applicable. The output of PandoraPFA is then a list of reconstructed particles, typically referred to as particle flow objects (PFOs), which contain important information such as four-momentum, energy and particle ID. This list of PFOs is then directly used in subsequent analysis steps.

3. Generative models

We investigate the use of two distinct generative network architectures. The first setup is a rather lightweight Wasserstein GAN (WGAN) (section 3.1) and the second is a significantly more complex bounded

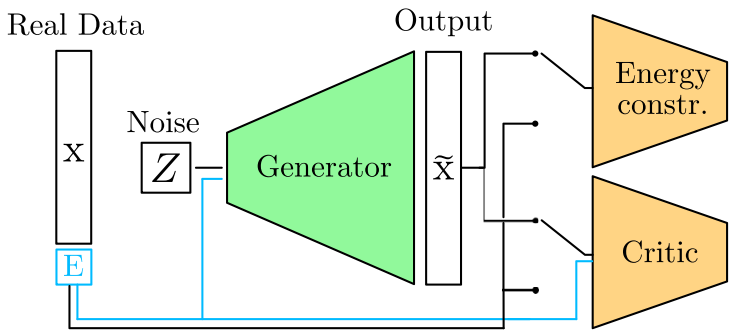

Figure 2. Schematic Illustration of our WGAN setup.

information bottleneck autoencoder (BIB-AE) (section 3.2). Both models are implemented using PyTorch 1.8.0 [39][8].

3.1. WGAN

The WGAN uses the Wasserstein-1 distance, also known as the earth mover's distance, as a loss function for better convergence and stability compared to classical GAN training [40, 41]. This distance evaluates the dissimilarity between two multi-dimensional distributions and informally gives the cost expectation for moving a mass of probability along optimal transportation paths [42]. Using the Kantorovich–Rubinstein duality, the Wasserstein loss can be expressed as

$$L = \sup_{f \in \mathrm{Lip}_1} \{\mathbb{E}[f(x)] - \mathbb{E}[f(\tilde{x})]\}. \tag{1}$$

The supremum is over all 1-Lipschitz functions f, and is approximated by a discriminator network D during the adversarial training. This discriminator is called the *critic* since it is trained to estimate the Wasserstein distance between real and generated images.

Furthermore, we need to ensure that the generated showers accurately resemble real showers of the requested energy E. This is achieved by parametrizing the generator and critic networks as functions of E and by adding a constrainer [9] network which returns the energy of a given shower. As this constrainer aims to provide a consistent metric of how well the shower energy is modeled, independent of the WGAN training itself, it is trained prior to the WGAN in a fully supervised fashion, using only Geant4 showers. The 100 million constrainer weights are then frozen during WGAN training.

The WGAN presented here is similar to the one used for photon shower generation in [14], however two architectural changes are applied in order to improve its generative performance for hadronic showers.

- The convolutional layers previously used in the critic network are replaced by 3D-residual blocks [43]. This change improves the expressiveness of the critic, i.e. the maximal complexity of the problem that the critic can describe. This increases the associated number of trainable parameters approximately four-fold to 4 million.
- A fully-connected network is employed instead of the convolutional layers previously used in the energy constrainer, as the shower patterns after preprocessing explicitly break translational symmetry.

The WGAN is trained for a total of 207k weight updates of the generator, which corresponds to 82 epochs. The Adam [44] optimizer is used with an initial learning rate of 10^{-4} (10^{-5}) for the generator (critic) networks.

3.2. BIB-AE

The BIB-AE unifies the most commonly used generative networks into an overarching theoretical framework [45]. For this reason, many elements normally associated with these models can be found in the BIB-AE. Fundamentally speaking, it is an autoencoder that maps from data-space to a lower-dimensional latent space and then back to data-space.

The BIB-AE presented here is a significant extension of the setup used for photon shower generation in [14]. We therefore first describe the baseline model followed by a detailed discussion of the extensions. The

[8] The code for both the WGAN and BIB-AE including the hyperparameter settings used for training are available on https://github.com/FLC-QU-hep/neurIPS2021_hadron.

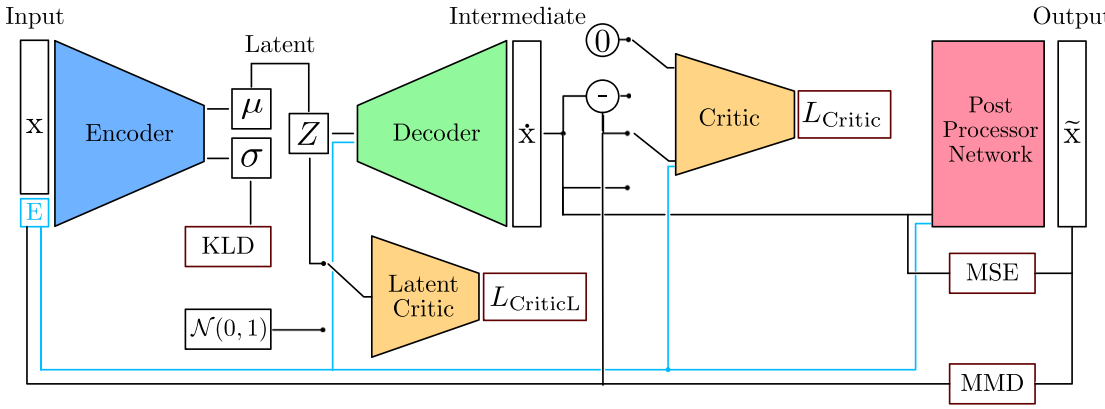

Figure 3. Schematic illustration of our BIB-AE setup including the post processor network in the final step.

baseline BIB-AE consists of two major components: the encoder/decoder chain that maps data to the latent space and back, and a series of auxiliary networks and loss terms used to train the main encoder and decoder. These auxiliary parts are:

- A dual purpose, WGAN like critic (C) that simultaneously judges whether the reconstructed output looks realistic and compares the output and input image to facilitate reconstruction.
- A Kullback–Leibler divergence (KLD) that regularizes the latent space to ensure it has a shape close to a normal distribution.
- A second WGAN like critic (C_{L}) trained to differentiate between latent space and a normal Gaussian.
- A maximum mean discrepancy (MMD) [46] term comparing the latent space to a normal distribution to further regularize it.

The contributions of these components are combined into a total loss term, each weighted by its own tuneable hyperparameter β_α,

$$L_{\text{BIB-AE}} = -\beta_{\text{C}} \cdot \mathbb{E}_{x\sim p_{\text{data}(x)}}[C(D(E(x)))] \tag{2}$$

$$-\beta_{\text{C}_\text{L}} \cdot \mathbb{E}_{x\sim p_{\text{data}(x)}}[\text{C}_\text{L}(E(x))] \tag{3}$$

$$+\beta_{\text{KLD}} \cdot \text{KLD}(E(x)) \tag{4}$$

$$+\beta_{\text{MMD}} \cdot \text{MMD}(E(x), \mathcal{N}(0,1)). \tag{5}$$

This loss function fills two rolls. The first role is filled by the critic loss, which ensures the quality of the reconstructed showers is high. The second is encapsulated by the remaining loss terms, which aim to regularize the latent space to be close to a normal Gaussian. This regularization is vital in a VAE, where a fully Gaussian latent space is assumed during generation. In contrast, the BIB-AE used here does not rely on having a perfectly Gaussian latent space. However, having well-regularized latent distributions close to Gaussians eases the burden on the kernel density estimator (KDE) latent sampling described in the following.

Both the BIB-AE network used in this work and the one previously used in [14] make use of a dedicated post processing network to fine tune certain shower features. However, the specific training process has been significantly modified.

3.2.1. BIB-AE with KDE latent space sampling

The underlying idea of generative autoencoders is to map the data space to a well-known latent space and back from this latent space to the data. When designing this latent space, one needs to strike a balance between regularization and expressiveness. A perfectly regularized latent space cannot contain information, while a very expressive latent space is difficult to sample from during generation. The standard approach in a VAE is to have a regularization loss term that is balanced with the reconstruction loss. This ensures that the latent space is both regularized and expressive. However for our BIB-AE setup the adversarial reconstruction loss makes this balance highly non-trivial. Instead we found larger success using a buffer-VAE [47] inspired approach.

After training the main BIB-AE model, the encoder is used to translate the training data set into latent space samples. This latent space data provides a good description of the underlying latent space distribution. Our goal is now to draw new samples from this latent distribution. These samples can then, in turn, be passed to the BIB-AE decoder to generate new shower samples. We found several viable options to perform this latent sampling. Both a GAN and a normalizing flow are capable of reproducing the latent space very closely, however training these setups correctly is once again non-trivial. For this reason a KDE [48] is fitted to the latent space. As the latent space distribution depends on the energy, energy labels are included in the fit. During generation, we sample both the particle energy label and the input latent noise from this KDE. In order to generate specific energy labels we make use of a rejection sampling method. A more in-depth discussion on the latent space sampling for the BIB-AE is provided in [15].

3.2.2. Minibatch discrimination

A major concern for physics applications of generative models is ensuring that the composition and overall properties of the generated data match that of the training set. Minibatch discrimination [49] is a vital tool in this effort. The underlying idea is to give the discriminator network information about a whole batch of data, in addition to the information about the individual samples in that batch. This allows the discriminator to more easily spot outliers or mode collapse.

Minibatch discrimination is implemented by first calculating the sum and standard deviation of each discriminator input sample. We then define a difference matrix between all of these sums and do the same for the standard deviations. These matrices are subsequently passed through an embedding network, in essence allowing the discriminator to learn the most important features of this batch-information. Finally, the outputs of this embedding are aggregated and passed to the fully connected main section of the discriminator. The above operations are carried out both for the un-scaled and log-scaled discriminator inputs.

3.2.3. Dual and resetting critics

One noticeable property of the pion shower data set is its sparsity. Despite having 30k possible pixel positions, the average number of active pixels above the MIP threshold (see section 4.1) is only 400, or about 1%. This means, that over the course of the adversarial training, the critic network can become blind to certain pixel positions, which in turn gives rise to artifacts at these positions. One way to remedy this is to reinitialize the critic network. As this reinitialized network has no training history, it will easily spot these artifacts and force the decoder to correct them. However this has the downside that the critic never fully converges, making it harder to learn more subtle features. Therefore we replace every critic used in the BIB-AE with a set of two networks with identical architectures. One of these network is trained continuously, while the other has its weights and optimizer reset after each epoch.

3.2.4. Improved post processing

Much like the BIB-AE setup for photon showers, the pion setup also uses a dedicated post processor network. The training procedure has, however, been heavily modified. The first difference is that, while the photon post processor was trained parallel to the main BIB-AE, we now train it on a frozen version of the BIB-AE networks, after they have already finished training. While this does mean that the main BIB-AE model can no longer improve during the post processor training, it significantly stabilizes the post processor training and allows the network to be easily trained to convergence.

Furthermore, the loss function of the post processor was also refined. In addition to the mean squared error (MSE) and sorted kernel MMD (SK-MMD) described in [14] we implemented a series of auxiliary loss terms.

- The first SK-MMD is complemented with a second SK-MMD term with a larger kernel size. This allows for a larger coverage of the cell energy spectrum.
- An MSE/mean absolute error loss between the sorted hit-energy values of the original input and processed shower. On the one hand this helps the post processor to better learn the energy spectrum, whilst on the other hand it also greatly improves the energy conditioning of the full setup.
- A loss term comparing the average shower image of an original input batch with the average of a processed shower batch. Previous iterations of the post processor were prone to adding artifacts to the shower, which were manifested in some pixels being significantly more likely to be active than others. These artifacts cannot be seen directly form individual images, but only in the averages over many showers. This loss term strongly reduces these artifacts.

The BIB-AE was trained for a total of 37 epochs. Using that model as a basis, the post processor was trained for 105 epochs. All networks used the Adam [44] optimizer with exponential learning-rate decay. During training a threshold function was used to map every value below 10% of a MIP to 0.

In the following section, we will observe the performance of the improved WGAN and BIB-AE networks for shower generation. Due to the large computational cost of training, a detailed ablation study of the specific gains afforded by each modification was not possible. However, without these improvements, no adequate description of hadronic shower distributions was possible. Furthermore, the extensive training times of both the investigated generative models meant that no systematic hyperparameter optimization was performed.

4. Results

In standard applications of generative models, the quality and fidelity of individual generated images are often the most important metrics. However, when applied to physics simulation, the statistical distributions of specific quantities over many generated images become increasingly relevant. This means we not only have to ensure that the individual showers look realistic, but also that the overall statistical properties of a set of showers agree with the training data.

In the following, we consider two different representation levels for comparisons between the output of generative models and the ground truth. The first one is termed *generator level*. As the name implies, at this stage, the direct output of simulation tools is compared, yielding insight into how well a specific model captures the training data. This is the default level used e.g. in [14] as well as other studies of generative models.

In practice, however, most physics analyses do not look directly at the raw measured or simulated data. Instead, a series of pattern recognition and data extraction algorithms are first applied, in a process referred to as reconstruction (see section 2.2). The result of this processing, the so-called *reconstruction level*, constitutes the second data representation used for comparisons here. Considering a realistic reconstruction scenario is a major improvement of this work over previous results, as it allows us to judge which *relevant* properties of the data a generative model learns.

Generator level results are presented in section 4.1, followed by reconstruction level results in section 4.2, and studies of the inference-time in section 4.3.

4.1. Generator level

We first compare the output of the generative models with that of Geant4 without additional reconstruction steps. The only additional processing performed is the removal of all cells with a pixel energy below half the energy deposition of a minimal ionizing particle, at 0.25 MeV. In a real calorimeter, this selection removes noise produced by the detector components and any simulation therefore has to employ the same cutoff.

Figure 4 shows comparisons of physically relevant shower-shape observables. Each plot uses the test-set generated using Geant4 as the baseline (gray, filled) and compares it to the BIB-AE (blue, solid) and the WGAN (orange, dashed). The 10–100 GeV pion energy range of the dataset is used throughout.

In the top left plot we see the visible cell energy distribution. The hatched region of the plot indicates the cutoff at 0.5 MIP. As described above, any hit below this threshold is discarded for further comparisons. The main feature of interest is the peak located at 1.0 MIP. We can see that the BIB-AE setup, largely thanks to its post processor, accurately replicates this feature. The WGAN setup, on the other hand, seems unable to capture the peak. This is in line with other work [14, 16] that has also demonstrated this difficulty in getting GANs to learn low energy features.

The center top figure shows the calculated shower start position [50]. Here we see a good agreement between Geant4 and the BIB-AE, while the WGAN seems to produce a significant abundance of early starting showers. In the top right, we show the center of gravity of the shower. This is equivalent to the first moment along the depth of the calorimeter, i.e. the direction the particle is traveling in. Once again, the BIB-AE distribution overlaps almost perfectly with the Geant4 one. The WGAN shows a significant overlap, but fails to correctly model the tail regions. The bottom left, center and right plots show the average energy profiles of the showers in the x, y, and z direction respectively. For all three profiles, the BIB-AE showers line up with the Geant4 ones, save for minor fluctuations in the tail regions. Specifically, the asymmetry in the x profile caused by the magnetic field in the detector is also correctly modeled. The WGAN on the other hand exhibits significant artifacts.

In addition to the marginal distributions of individual observables, a successful generative model also needs to learn higher-dimensional correlations. To this end, we explicitly verify the correlations between all pairs of considered observables. The top row of figure 5 shows two correlation matrices, with each entry corresponding to the Pearson correlation coefficient between the following quantities: first moment in x, y,

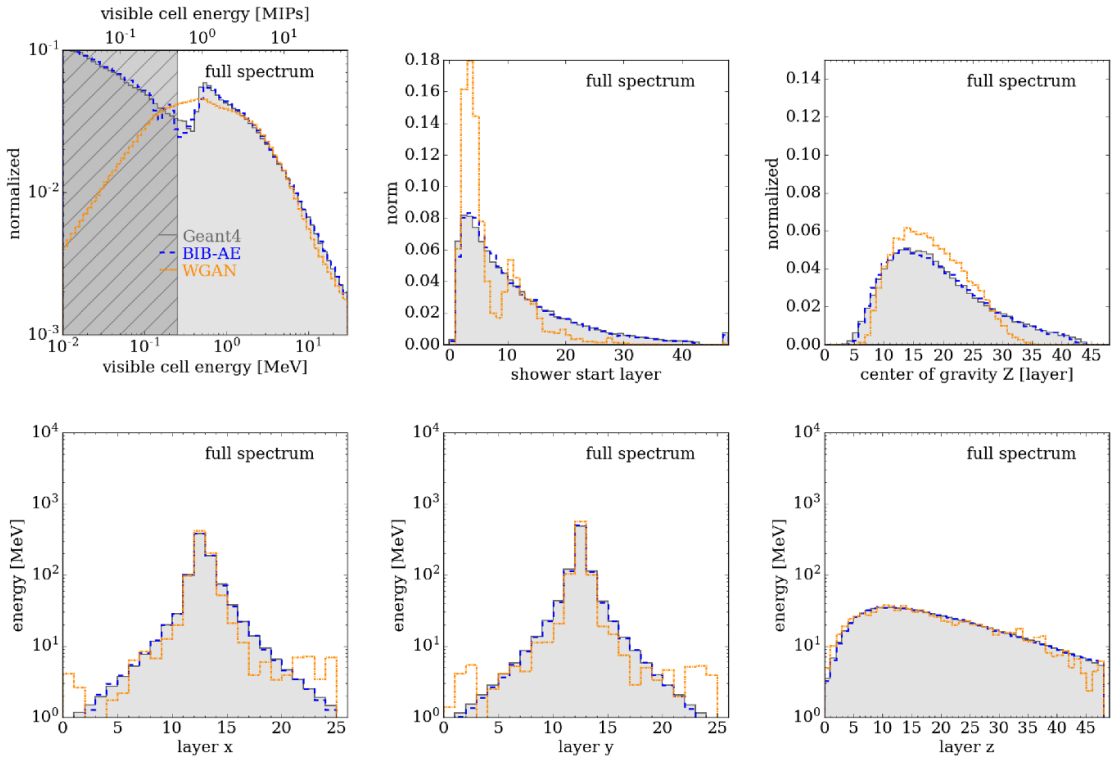

Figure 4. Differential distributions comparing physics quantities between Geant4 and the different generative models for charged pion showers. In the top left, the energy per-cell is given in MeV on the bottom axis, and in multiples of the expected energy deposit of a minimum ionizing particle (MIP) on the top axis. The grayed out area indicates the 0.5 MIP cutoff. All plots are for showers generated with uniform energies in the 10–100 GeV range.

and z direction (m_1), second moment in x, y, and z direction (m_2), the total visible energy (E_{vis}), the incident particle energy (E_{inc}), the number of hits above the 0.5 MIP threshold (n_{hit}) and the energy fractions deposited in the first, second, and last third along the depth of the calorimeter (E_i/E_{vis}).

We perform this correlation calculation for Geant4, the BIB-AE and the WGAN. In order to best compare the correlations produced by the generative models and those of Geant4, the BIB-AE and WGAN matrices are subtracted from the Geant4 matrix. These results are shown in the bottom of figure 5. The closer to zero the differences are, the better the correlations are reproduced. We see that for the BIB-AE the largest deviation is around 0.13, while most do not exceed 0.05. This is indicative of the ability of the BIB-AE to reproduce the Geant4 correlations. For the WGAN we see significantly larger deviations, up to 0.31. This is in line with previous results showing the difficulties the WGAN has in learning properties such as the shower energy profiles.

The third set of comparison plots is presented in figure 6. The color coding remains the same as in the first set, however only discrete pion energies of 20, 50 and 80 GeV are used. This allows for explicit testing of whether the energy conditioning is learned correctly or not. The leftmost plot shows the total energy deposited in the active regions of the calorimeter. Both the energy sums of the BIB-AE and the WGAN largely match those present in Geant4. Furthermore, while both networks seem to slightly mismodel the very sharp 20 GeV peak, they perform very well for the 50 and 80 GeV peaks, where they correctly model the means and widths of the peaks. The plot on the right shows the total number of pixels with values above the MIP threshold. One vital aspect of modeling this quantity is correctly capturing the visible cell energy spectrum around the cutoff (figure 4, top left), as even small shifts in this region can have large impacts on how many points end up above or below the cutoff. This explains why the BIB-AE is more successful in reproducing the total number of hits compared to the WGAN, which shows some significant deviation, especially for the 50 GeV distribution.

Figure 7 shows a more in-depth look at the energy conditioning. For fixed energies between 20 and 90 GeV we produce a set of showers using both Geant4 and the generative models. We then calculate the visible energy sums of these showers and determine the mean and root-mean-square of the 90% core of these distributions, labeled μ_{90} and σ_{90} respectively, for all energies. Finally we plot the means and relative widths

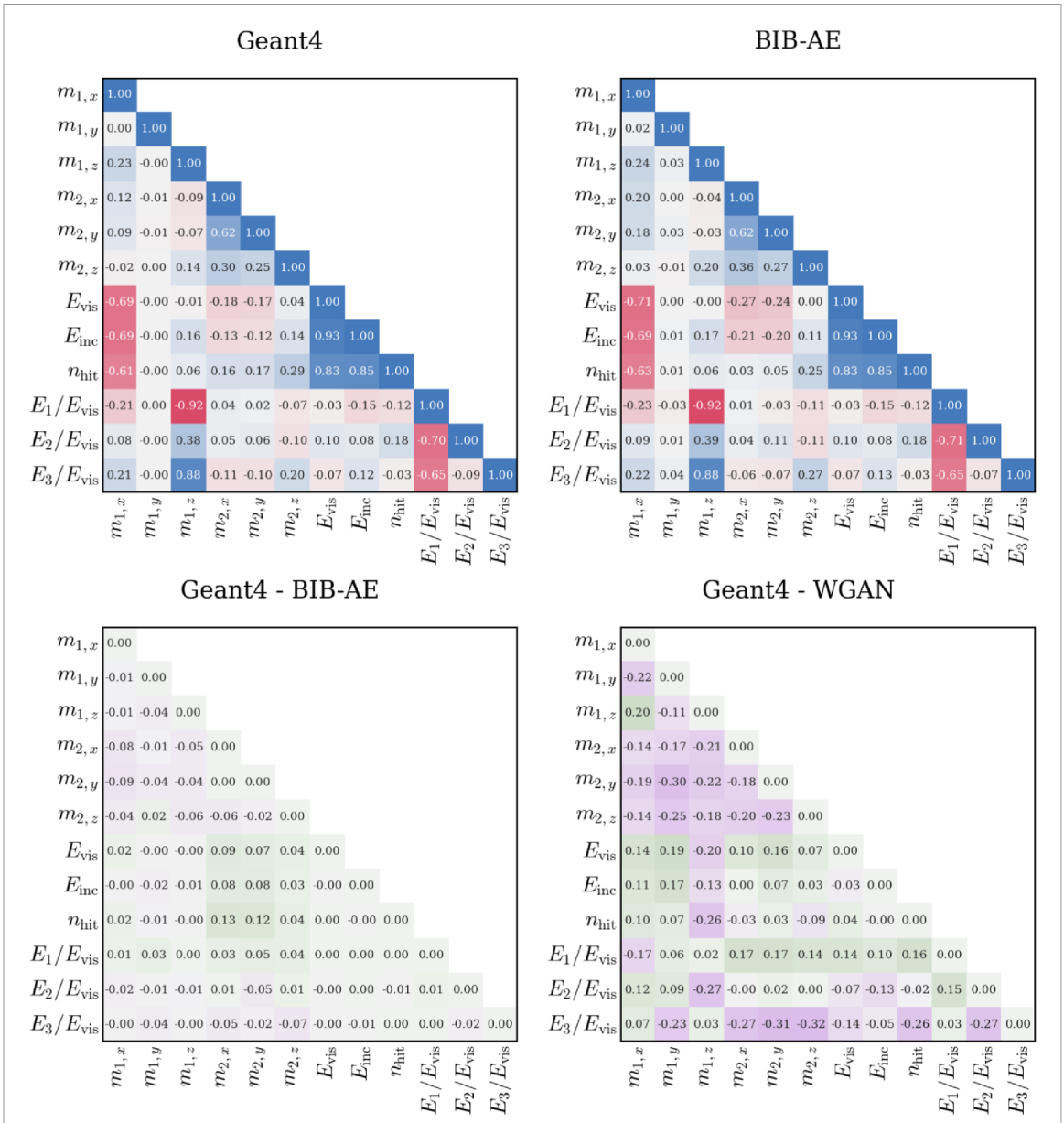

Figure 5. Top: Pearson correlation coefficients for selected physical parameters of showers simulated with Geant4 (left) and generated with the BIB-AE (right). See the text for specific definitions of the individual parameters. Bottom: differences between these correlation coefficients and Geant4 for the BIB-AE (left) and WGAN (right).

as a function of the incoming particle energy. Note that 10 and 100 GeV were omitted from this study as these points lie right at the phase space boundaries. For energies above 40 GeV the resolution does not improve with energy due to leakage effects becoming important. The leftmost curve shows that the position of the mean is especially well captured by the WGAN, with a maximum deviation of 2%. The BIB-AE exhibits some larger discrepancies, up to 3% in the high and low energy sections, but still provides a reasonable agreement. Note that a calibration factor has been applied to the WGAN-generated single-energy showers to improve the linearity. The relative width in the right plot is not modeled as well, exhibiting differences up to the 10% level for the BIB-AE and up to the 30% level for the WGAN in the edge regions.

4.2. Reconstruction level

To apply the PandoraPFA reconstruction software, the outputs of the WGAN and BIB-AE, which yield (batches of) $25 \times 25 \times 48$ NumPy [51] arrays, need to be converted back into actual cell positions and hit energies in the ILD calorimeter. These hits are then provided as input to PandoraPFA for reconstructing corresponding PFOs[9]. The resulting PFOs are then compared to those created with the standard Geant4

[9] As no track reconstruction is considered, PandoraPFA will reconstruct neutral PFOs using calorimeter information only.

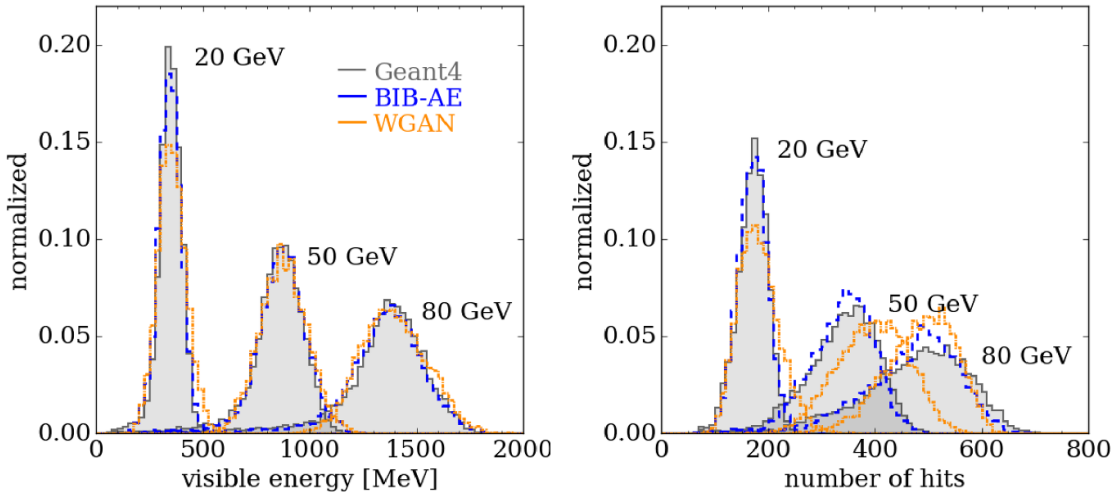

Figure 6. Comparison of the visible energy and number of active hits between Geant4 and the different generative models for three selected pion energies. The number of hits in the right plot is calculated after applying a cutoff at 0.5 MIP.

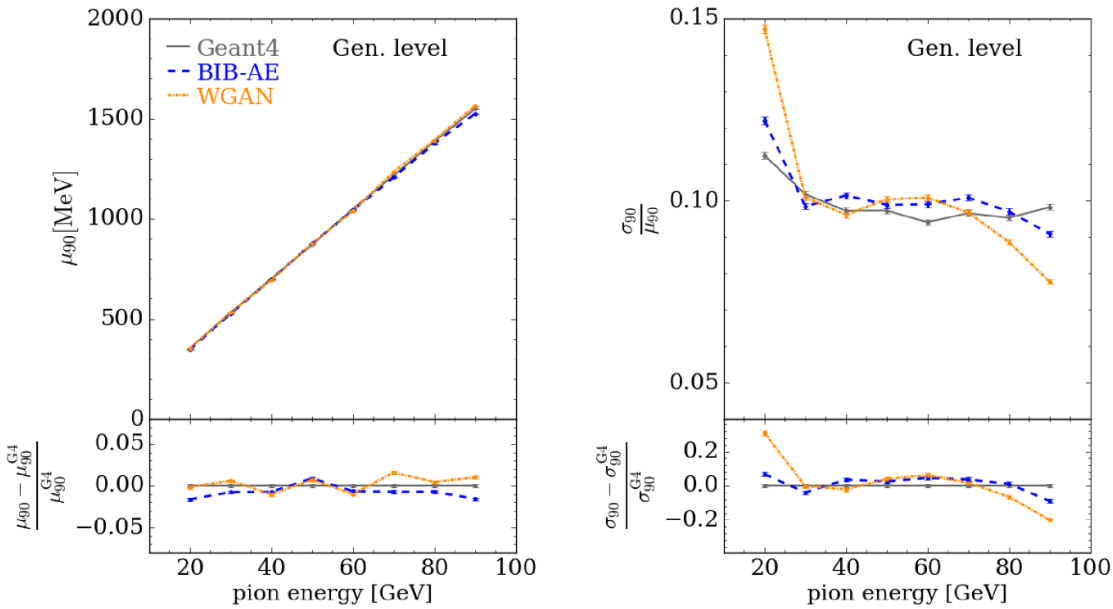

Figure 7. Mean (μ_{90}, left) and relative width (σ_{90}/μ_{90}, right) at generator level for pions with various incident energies. In order to avoid edge effects, the phase space boundary regions of 10 and 100 GeV are removed for the response and resolution studies. In the bottom panels, the relative offset of these quantities with respect to the Geant4 simulation is shown.

simulation-reconstruction chain. To ensure a consistent comparison, the Geant4 data undergoes the same projection/conversion operation as the WGAN and BIB-AE.

Figure 8 shows the same quantities presented in figure 7, but now at the reconstruction level. The leftmost plot shows that the position of the mean is well captured in the middle range of energies by both the models. Likewise, both models display some larger discrepancies, up to 3%–5% in the high and low energy sections, but still have a reasonable agreement with Geant4. The relative width on the right plot shows a fairly good agreement for the WGAN for the middle incident energies. On the edge regions, however, up to 20% differences for the BIB-AE and up to 40% for the WGAN are present. It is worth noting that our models and Geant4 have better relative width compared to the generator level as PandoraPFA uses a software compensation algorithm [52] that improves the energy reconstruction of clusters by weighting hits depending on their hit energy density.

4.3. Computing times

The prime objective for using generative models in particle physics is to reduce the time and cost per simulated sample. To do so, we benchmark the per-shower generation time both on CPU and GPU hardware

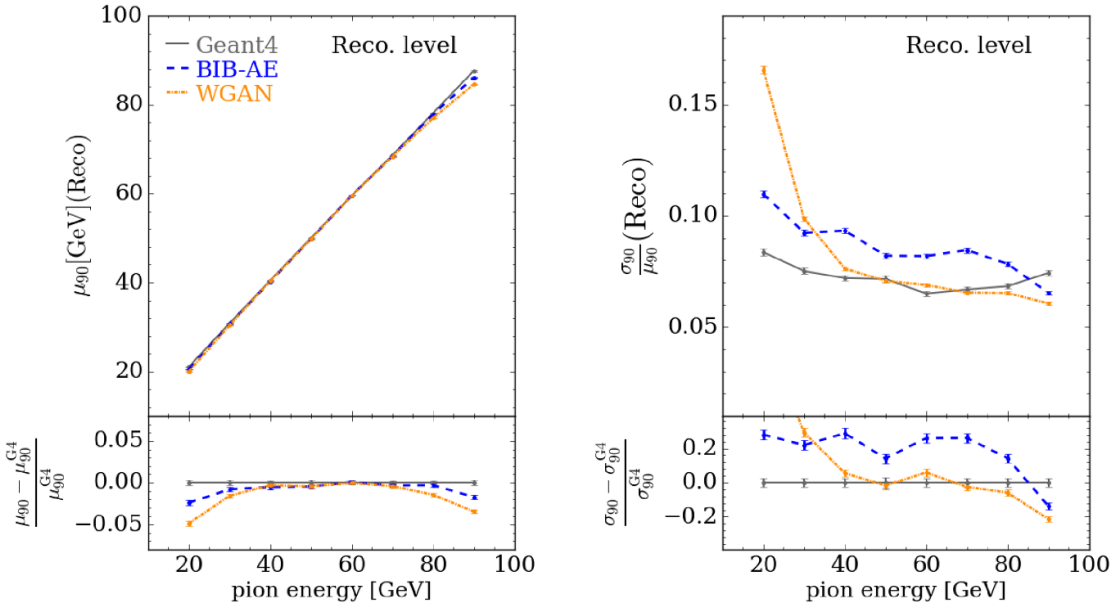

Figure 8. Mean (μ_{90}, left) and relative width (σ_{90}/μ_{90}, right) at reconstruction level for pions with various incident energies. In order to avoid edge effects, the phase space boundary regions of 10 and 100 GeV are removed for the response and resolution studies. In the bottom panels, the relative offset of these quantities with respect to the Geant4 simulation is shown.

Table 1. Computational performance of WGAN and BIB-AE generators on a single core of an Intel® Xeon® CPU E5-2640 v4 (CPU) and NVIDIA® A100 with 40 GB of memory (GPU) compared to Geant4. For the generative models, the best performing batch size is shown and given by the mean and standard deviation obtained for sets of 10 000 showers.

Hardware	Simulator	Time / shower (ms)	Speed-up
CPU	Geant4	2684 ± 125	$\times 1$
	WGAN	47.923 ± 0.089	$\times 56$
	BIB-AE	350.82 ± 0.57	$\times 8$
GPU	WGAN	0.264 ± 0.002	$\times 10\,167$
	BIB-AE	2.051 ± 0.005	$\times 1309$

architectures. Fixed factors, such as initial sample generation and network training time, are not included in this accounting, as they are expected to be small compared to the overall number of showers to be generated. Table 1 shows the average time to generate a shower with an energy in the 10–100 GeV range using Geant4, the WGAN, and the BIB-AE. Both models offer significant speedups compared to classical generation methods. Furthermore, we also see trade-offs between these models. While the BIB-AE produces overall better quality showers than the WGAN, it also is one order of magnitude slower.

5. Conclusions

The strategy of using generative models to augment classical simulations has made rapid progress since its inception. This paper advances the state-of-the-art in two regards: learning more complex hadronic showers in a high-granularity calorimeter, and considering the effects of particle-flow reconstruction algorithms. It is worth mentioning here, that while we used charged pions in this study, any other charged hadron will give rise to showers that are nearly indistinguishable from these charged pion showers, and those created from neutral particles will look almost identical except for the spurious minimum ionizing missing particle (MIP) hits deposited in the early layers before the first nuclear interaction occurs. Therefore the results can easily be extended to hadronic showers in general and eventually full events can be simulated by overlaying electromagnetic and hadronic showers accordingly. Yet this requires future work, e.g. on conditioning the showers also with incident angles and addressing irregular cell structures and gaps in real calorimeters.

We observe that the modified BIB-AE achieves excellent agreement with all physical observables at generator-level, both over the full spectrum and for specific incident particle energies. The largest disagreement of $\approx 10\%$ is seen for the hit multiplicity distribution, while most other observables agree to the percent-level or better. This is made possible by using a second-stage density estimator to sample from the

learned latent space distributions, utilizing batch-level information, adding a resetting critic network, and a more convergent post-processing network.

Still on generator level, the performance of the much simpler WGAN architecture considered is clearly worse. For example, it does not learn the correct energy distribution around the MIP-value which leads to a mismodeling of the distribution of the number of hits. Similarly, the longitudinal shower profile shows unphysical structure.

However, considering the offset and width of reconstructed energies, the difference is much smaller. Both BIB-AE and WGAN achieve excellent linearity with the largest deviation of 5% observed at the boundaries of the considered energy range. For incident particle energies between 40 and 80 GeV the WGAN-generated showers also accurately track the width of the underlying GEANT4 simulation. Only the mean energy and width, as the most relevant quantities for physics analysis, were so far considered at the reconstruction level. Future work will extend the investigation also to other properties of the shower.

Finally, we again confirm the previously observed speed-up over the initial Geant4 generation when sampling from generative models. The more complex BIB-AE architecture is slower than the simpler WGAN by approximately a factor of eight. The main cause for this is that BIB-AE uses transpose convolutions to get to the full image size and then runs another set of convolutions at that full size, while the WGAN applies no additional convolutional layers once the full image size has been reached. The maximum speed-up of four orders of magnitude is observed when comparing WGAN (GPU) to GEANT4 executed on CPU.

It is interesting to observe that while generative models overall offer a trade-off between precise simulation and resource consumption, this also holds true when comparing different generative architectures. The successful and accurate simulation of hadronic showers is another major milestone towards application-ready generative models for highly-granular calorimeters.

Data availability statement

The data that support the findings of this study are openly available. Both the training and test datasets can be found under the DOI 10.5281/zenodo.6491116 or via https://doi.org/10.5281/zenodo.6491116

Acknowledgments

This research was supported in part through the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. E Buhmann and W Korcari are funded by the German Federal Ministry of Science and Research (BMBF) via *Verbundprojekts 05H2018–R&D COMPUTING (Pilotmaßnahme ErUM-Data) Innovative Digitale Technologien für die Erforschung von Universum und Materie*. S Diefenbacher is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2121 'Quantum Universe'–390833306. E Eren is funded through the Helmholtz Innovation Pool project ACCLAIM that provided a stimulating scientific environment for parts of the research done here. L Rustige was supported by DESY and HamburgX Grant LFF-HHX-03 to the Center for Data and Computing in Natural Sciences (CDCS) from the Hamburg Ministry of Science, Research, Equalities and Districts. This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 101004761.

ORCID iD

Sascha Diefenbacher https://orcid.org/0000-0003-4308-6804

References

[1] Albrecht J *et al* 2019 A roadmap for HEP software and computing R&D for the 2020s *Comput. Softw. Big Sci.* **3** 7
[2] Paganini M, de Oliveira L and Nachman B 2018 Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters *Phys. Rev. Lett.* **120** 042003
[3] Agostinelli S *et al* 2003 Geant4-a simulation toolkit *Nucl. Instrum. Methods Phys. Res.* A **506** 250–303
[4] Butter A, Diefenbacher S, Kasieczka G, Nachman B and Plehn T 2020 GANplifying event samples (arXiv:2008.06545)
[5] Jansky R 2015 The ATLAS fast Monte Carlo production chain project *J. Phys. Conf. Ser.* **664** 072024
[6] de Oliveira L, Paganini M and Nachman B 2017 Learning particle physics by example: location-aware generative adversarial networks for physics synthesis *Comput. Softw. Big Sci.* **1** 4
[7] Paganini M, de Oliveira L and Nachman B 2018 CaloGAN: simulating 3D high energy particle showers in multi-layer electromagnetic calorimeters with generative adversarial networks *Phys. Rev.* D **97** 014021
[8] Erdmann M, Geiger L, Glombitza J and Schmidt D 2018 Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks *Comput. Softw. Big Sci.* **2** 4
[9] Erdmann M, Glombitza J and Quast T 2019 Precise simulation of electromagnetic calorimeter showers using a Wasserstein generative adversarial network *Comput. Softw. Big Sci.* **3** 4

[10] Belayneh D *et al* 2019 Calorimetry with deep learning: particle simulation and reconstruction for collider physics (arXiv:1912.06794)
[11] ATLAS Collaboration 2018 Deep generative models for fast shower simulation in ATLAS *Technical Report ATL-SOFT-PUB-2018-001 CERN* (Geneva) (available at: http://cds.cern.ch/record/2630433)
[12] ATLAS Collaboration 2019 VAE for photon shower simulation in ATLAS *Technical Report ATL-SOFT-SIM-2019-007 CERN* (available at: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-007/)
[13] Ghosh A (ATLAS Collaboration) 2019 Deep generative models for fast shower simulation in ATLAS *Technical Report* ATL-SOFT-PROC-2019-007 CERN (Geneva) (available at: https://cds.cern.ch/record/2680531)
[14] Buhmann E, Diefenbacher S, Eren E, Gaede F, Kasieczka G, Korol A and Krüger K 2021 Getting high: high fidelity simulation of high granularity calorimeters with high speed *Comput. Softw. Big Sci.* **5** 13
[15] Buhmann E, Diefenbacher S, Eren E, Gaede F, Kasieczka G, Korol A and Krüger K 2021 Decoding photons: physics in the latent space of a BIB-AE generative network *EPJ Web Conf.* **251** 03003
[16] Khattak G R, Vallecorsa S, Carminati F and Khan G M 2021 Fast simulation of a high granularity calorimeter by generative adversarial networks (arXiv:2109.07388)
[17] Carminati F *et al* 2020 Generative adversarial networks for fast simulation *J. Phys. Conf. Ser.* **1525** 012064
[18] Hariri A, Dyachkova D and Gleyzer S 2021 Graph generative models for fast detector simulations in high energy physics (arXiv:2104.01725)
[19] Rehm F, Vallecorsa S, Saletore V, Pabst H, Chaibi A, Codreanu V, Borras K and Krücker D 2021 Reduced precision strategies for deep learning: a high energy physics generative adversarial network use case (arXiv:2103.10142)
[20] Rehm F, Vallecorsa S, Borras K and Krücker D 2021 Validation of deep convolutional generative adversarial networks for high energy physics calorimeter simulations (arXiv:2103.13698)
[21] Rehm F, Vallecorsa S, Borras K and Krücker D 2021 Physics validation of novel convolutional 2D architectures for speeding up high energy physics simulations *EPJ Web Conf.* **251** 03042
[22] Krause C and Shih D 2021 CaloFlow: fast and accurate generation of calorimeter showers with normalizing flows (arXiv:2106.05285)
[23] Krause C and Shih D 2021 CaloFlow II: even faster and still accurate generation of calorimeter showers with normalizing flows (arXiv:2110.11377)
[24] Alanazi Y, Sato N, Ambrozewicz P, Blin A N H, Melnitchouk W, Battaglieri M, Liu T and Li Y 2021 A survey of machine learning-based physics event generation (arXiv:2106.00643)
[25] Butter A and Plehn T 2020 Generative networks for LHC events (arXiv:2008.08558)
[26] Goodfellow I J *et al* 2014 Generative adversarial nets *Proc. 27th Int. Conf. on Neural Information Processing Systems (NIPS'14) (Cambridge, MA)* vol 2 pp 2672–80
[27] Kingma D P and Welling M 2014 Auto-encoding variational bayes (arXiv:1312.6114)
[28] Dinh L, Krueger D and Bengio Y 2014 NICE: non-linear independent components estimation (arXiv:1410.8516)
[29] Dinh L, Sohl-Dickstein J and Bengio S 2016 Density estimation using real NVP (arXiv:1605.08803)
[30] Rezende D J and Mohamed S 2015 Variational inference with normalizing flows *Proc. 32nd Int. Conf. on Int. Conf. on Machine Learning (ICML'15)* vol 37 pp 1530–8
[31] Papamakarios G, Nalisnick E, Rezende D J, Mohamed S and Lakshminarayanan B 2019 Normalizing flows for probabilistic modeling and inference (arXiv:1912.02762)
[32] Abramowicz H *et al* 2020 International large detector: interim design report (arXiv:2003.01116)
[33] Adloff C *et al* 2010 Construction and commissioning of the CALICE analog hadron calorimeter prototype *J. Instrum.* **5** 05004
[34] Adloff C *et al* 2013 Validation of GEANT4 Monte Carlo models with a highly granular scintillator-steel hadron calorimeter *J. Instrum.* **8** 07005
[35] iLCSoft Authors 2016 iLCSoft project page (available at: https://github.com/iLCSoft)
[36] Frank M, Gaede F, Grefe C and Mato P 2014 DD4hep: a detector description toolkit for high energy physics experiments *J. Phys. Conf. Ser.* **513** 022010
[37] Joe B, Kelsey H and Brendan B 2017 *Kubernetes: Up and Running* (Sebastopol, CA: O'Reilly Media, Inc.)
[38] Marshall J S and Thomson M A 2015 The Pandora software development kit for pattern recognition (available at: http://dx.doi.org/10.1140/epjc/s10052-015-3659-3)
[39] Paszke A *et al* 2019 PyTorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32 pp 8024–35
[40] Arjovsky M, Chintala S and Bottou L 2017 Wasserstein GAN (arXiv:1701.07875)
[41] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A 2017 Improved training of Wasserstein GANs *Advances in Neural Information Processing Systems* vol 30 pp 5767–77
[42] Cédric V 2009 *Optimal Transport: Old and New* (Berlin: Springer)
[43] Hara K, Kataoka H and Satoh Y 2018 Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA: IEEE Computer Society) pp 6546–55
[44] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization (arXiv:1412.6980)
[45] Voloshynovskiy S, Kondah M, Rezaeifar S, Taran O, Holotyak T and Rezende D J 2019 Information bottleneck through variational glasses (arXiv:1912.00830)
[46] Gretton A, Borgwardt K M, Rasch M J, Schölkopf B and Smola A J 2008 A kernel method for the two-sample problem *CoRR* (arXiv:0805.2368)
[47] Otten S, Caron S, de Swart W, van Beekveld M, Hendriks L, van Leeuwen C, Podareanu D, de Austri R R and Verheyen R 2019 Event generation and statistical sampling for physics with deep generative models and a density information buffer (arXiv:1901.00875)
[48] Parzen E 1962 On estimation of a probability density function and mode *Ann. Math. Stat.* **33** 1065
[49] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A and Chen X 2016 Improved techniques for training GANs (arXiv:1909.10578)
[50] Adloff C *et al* 2011 Tests of a particle flow algorithm with CALICE test beam data *J. Instrum.* **6** 07005
[51] Harris C R *et al* 2020 Array programming with NumPy *Nature* **585** 357–62
[52] Tran H L, Krüger K, Sefkow F, Green S, Marshall J, Thomson M and Simon F 2017 Software compensation in particle flow reconstruction (available at: https://doi.org/10.1140/epjc/s10052-017-5298-3)