

Decoding Photons: Physics in the Latent Space of a BIB-AE Generative Network

Erik Buhmann^{1,*}, Sascha Diefenbacher^{1,2}, Engin Eren², Frank Gaede², Gregor Kasieczka¹, Anatolii Korol³, and Katja Krüger²

¹Institut für Experimentalphysik, Universität Hamburg, Germany

²Deutsches Elektronen-Synchrotron, Germany

³Taras Shevchenko National University of Kyiv, Ukraine

Abstract. Given the increasing data collection capabilities and limited computing resources of future collider experiments, interest in using generative neural networks for the fast simulation of collider events is growing. In our previous study, the Bounded Information Bottleneck Autoencoder (BIB-AE) architecture for generating photon showers in a high-granularity calorimeter showed a high accuracy modeling of various global differential shower distributions. In this work, we investigate how the BIB-AE encodes this physics information in its latent space. Our understanding of this encoding allows us to propose methods to optimize the generation performance further, for example, by altering latent space sampling or by suggesting specific changes to hyperparameters. In particular, we improve the modeling of the shower shape along the particle incident axis.

1 Introduction

High-quality simulations of fundamental processes and particle interactions with complex detectors are crucial to data analysis in high energy physics. Especially in the context of increasing data volumes from upcoming runs of the Large Hadron Collider (LHC) and future experiments, the production of datasets using Monte-Carlo-based simulators is increasingly becoming a computing bottleneck [1].

A way to accelerate simulations is based on generative machine learning models that leverage recent advances in computer vision and are implemented parallelizable on graphic processing hardware. Such fast simulations based on Generative Adversarial Networks (GANs) [2] for calorimeter physics were first introduced in Ref. [3] and have seen active development in recent years [4–11]. This approach starts with a small dataset obtained using classical simulation techniques and aims to amplify its usable statistics by training a generative model. The principal feasibility of amplification was shown in Ref. [12].

Inspired by Ref. [13], we have previously implemented an improved Bounded Information Bottleneck Autoencoder (BIB-AE) architecture and shown its generation accuracy for various differential distributions of photon shower data in a high granularity calorimeter [14]. The BIB-AE architecture unifies ideas from different generative approaches, including GANs

*e-mail: erik.buhmann@uni-hamburg.de

and Variational Autoencoders (VAE) [15]. As an autoencoder, the model encodes input photon showers into a latent space from which in turn newly generated showers are sampled. The information bottleneck (IB) [16] refers in this context to the principle that the model optimizes the latent encoding while maximizing the mutual information between input and output. This contribution explores methods to understand the physics encoded in the latent space and introduces optimizations for improved generation fidelity. As opposed to Ref. [17] we do not aim to explicitly shape the latent space to match physical distributions but rather investigate how the deviations and correlations of the optimally Gaussian normal latent space features correspond to physically important observables. Compared to Ref. [18] we focus on an information-theoretic perspective and investigate correlations with physical observables instead of the topological structure of the latent space.

In the following, we first briefly introduce the data (Sec. 1.1) and our BIB-AE architecture (Sec. 1.2). We then investigate the connection between generative performance and information encoded in the latent space in Sec. 2, the correlation between learned latent space distributions and physical observables in Sec. 3, and see how this can be used to improve generative performance in Sec. 4. We close with a summary of results and draw our conclusions in Sec. 5.

1.1 Photon showers in a high granularity calorimeter

Calorimeters are an essential part of detectors used at high-energy particle colliders. They measure the energy particles deposit when interacting with material. Particles interacting with the matter in the calorimeter can produce secondary particles resulting in cascades or *showers*. Such a particle shower is created for example by an initial electromagnetically interacting photon.

Modern sampling calorimeters are built in a sandwich structure of measuring active layers interspersed with dense passive material. The active material of modern high granularity calorimeters consists of many small cells that are read out separately, yielding high resolution 3-dimensional measurements of particle showers.

We created our photon shower dataset using the GEANT4 [19] toolkit and a simulation of the SiW electromagnetic calorimeter in the International Large Detector (ILD) concept [20]. The simulated calorimeter section comprises 30 active layers with each 900 5x5 mm² calorimeter cells in a rectangular grid resulting in 3d images of 30x30x30 = 27,000 pixels. Our dataset consists of 950k photon showers with incident energies uniformly distributed between 10 and 100 GeV. This is the same dataset as used for Ref. [14] and we refer to that publication for additional details.¹

1.2 The BIB-AE model

The BIB-AE architecture consists of several building blocks: An encoder network mapping the input calorimeter images into a latent representation; a decoder network transforming the latent representation back into calorimeter images; a Post-Processor network refining the pixel values of the decoded image; a reconstruction critic network calculating the Wasserstein-distance between encoded and decoded image; and a latent critic network to regularize the latent space. The whole model is trained in two stages: First, the encoder, decoder, and critics are trained until sufficient fidelity is reached; afterwards, the whole model is trained in conjunction with the Post-Processor network to improve the accuracy of the cell

¹A fraction of the dataset as well as our implementation of the BIB-AE model are available at https://github.com/FLC-QU-hep/getting_high.

energy generation. To generate energy dependent samples, the BIB-AE is conditioned on the incident particle energy. An overview of the architecture is shown in Figure 1. A more detailed discussion of the network is provided in [14].

Like in any VAE-based model, the trained BIB-AE can be used to generate calorimeter shower images by sampling the latent space from Standard Normal distributions. To achieve good generation results, the latent space needs to be regularized towards such a Normal distribution. For this regularization the BIB-AE model employs several loss terms during training: A Kullback-Leibler divergence (KLD) loss L_{KLD} , the output of a latent critic network $L_{\text{latent-critic}}$, and a latent Maximum Mean Discrepancy (MMD) [21] term $L_{\text{latent-MMD}}$. Each latent regularizer contribution is scaled with a weight β yielding a combined latent loss of

$$L_{\text{total-latent}} = \beta_{\text{KLD}} \cdot L_{\text{KLD}} + \beta_{\text{latent-critic}} \cdot L_{\text{latent-critic}} + \beta_{\text{latent-MMD}} \cdot L_{\text{latent-MMD}} \quad (1)$$

with the KL divergence of two discrete probability distributions P and Q defined as

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \quad (2)$$

and calculated via

$$D_{\text{KL},i} = D_{\text{KL}}(\mathcal{Z}_i \parallel \mathcal{N}(0, 1)) = -\frac{1}{2} \left(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2 \right). \quad (3)$$

In the context of this publication, *latent variables* \mathcal{Z}_i are Gaussian distributions with two trainable parameters μ_i and σ_i ($\mathcal{Z}_i \equiv \mathcal{N}(\mu_i, \sigma_i^2)$) regularized towards a Standard Normal distribution. Its sampled values are $z_i \sim \mathcal{Z}_i$.

In previous work we have implemented the BIB-AE architecture with 24 trainable latent variables and an additional 488 variables that are not encoded but sampled straight from a Standard Normal distribution during training [14]. We term the number of trainable latent space variables the *latent space size* n . Hence the total KLD loss is given by $L_{\text{KLD}} = \sum_{i=1}^n D_{\text{KL},i}$.

The loss weight β_{KLD} has the highest impact on the latent regularization as its scaling defines the magnitude of the KL divergence. Here the KL divergence measures the information content of the latent space [22, 23].

2 Different latent space sizes

Intuitively, for fixed β_{KLD} , higher latent space sizes n should yield an increased total information in the latent space until a maximum corresponding to the showers' intrinsic relevant information is reached. We test this by re-training a BIB-AE architecture with latent space sizes between 2 and 512 for fixed $\beta_{\text{KLD}} = 0.05$. The number of additional Standard Normal sampled variables is adjusted such that the total number of 512 decoded variables stays constant.

In Fig. 2 (left) we have sorted the trainable latent space variables by their individual KLD calculated via Eq. 3. On the vertical axis, the total information (i.e. the sum of KLD values up to and including latent variable i) is shown. Indeed, we observe increasing total encoded information with increasing latent space size until a saturation at approximately 45 nats (≈ 64 bits) is reached around a latent space size of 64. After that, a larger latent space does not substantially increase the learned information.

Next, we consider the KLD per latent space variable in Fig. 2 (right). All models follow a similar pattern: Only a few variables contain a high amount of information (high KLD). In

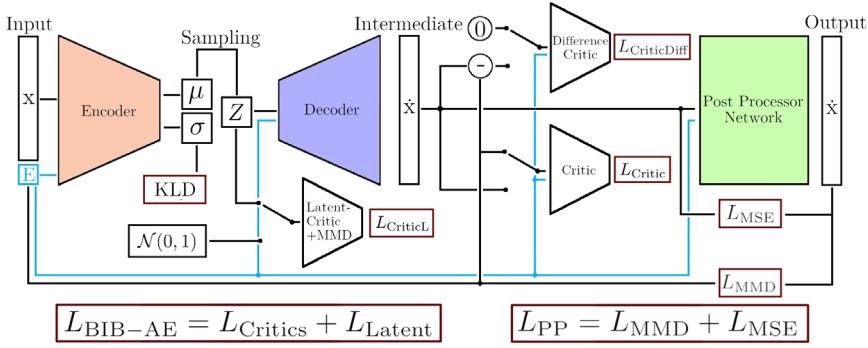


Figure 1. Overview of the BIB-AE generative model including the Post-Processor (PP) network and their respective loss terms. The model consists of multiple networks: An encoder, a decoder, a Post-Processor network as well as two reconstruction critics and a latent space critic. The latent critic, the Kullback-Leibner divergence (KLD) and a latent MMD regularize the latent space towards a Standard Normal distribution. The BIB-AE PP is conditioned on the incident particles’ energy (blue lines).

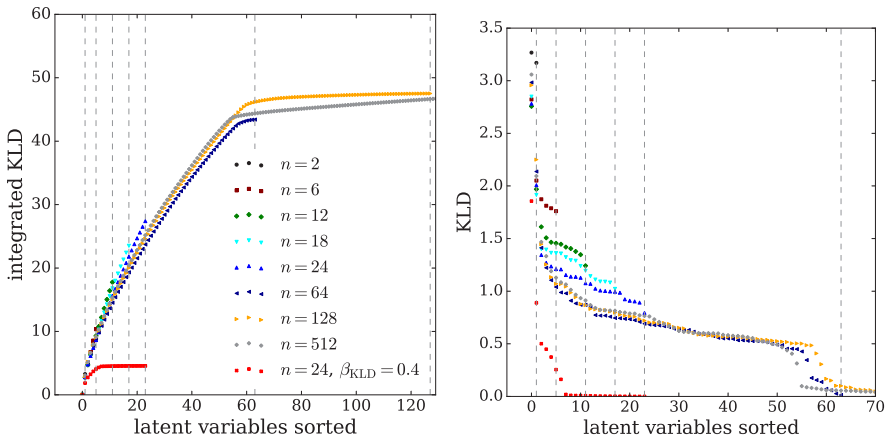


Figure 2. **Left:** Integrated Kullback-Leibler divergence (KLD) for latent variables sorted by highest KLD for models with different latent space sizes. **Right:** KL divergence of individual latent space variables sorted with decreasing KLD for different latent space sizes. All models are trained with a baseline weight $\beta_{\text{KLD}} = 0.05$, except stated otherwise.

particular, there are always two variables that encode significantly more information than the remaining ones. Furthermore, about 60 variables contain > 0.3 nats of information depending on the latent space size.

Naively, we would expect the most efficient use of the latent space at a size of $n = 64$ to yield optimal performance. Evaluating the performance of generative models is not straightforward and constitutes an active topic of research. Methods such as *Inception Score* [24] were proposed to evaluate models which produce photographic images. However, such scores are typically domain-specific and cannot directly be applied to our dataset. We therefore define a problem specific *fidelity score* S_{JS} that summarizes the performance across a number of physically relevant observables. The score is calculated by combining the Jensen–Shannon

Table 1. Fidelity score S_{JSD} for the best epoch of multiple model configurations with different latent space sizes n . For $n = 24$ the best score out of multiple training runs is given, while the mean score for those trainings is: $\bar{S}_{\text{JSD},24} = 1.02 \pm 0.12$. Only one training each was performed for sizes $n \neq 24$.

latent size	2	6	12	18	24	64	128	512
S_{JSD}	1.64	1.12	1.11	0.95	0.83	0.88	0.94	0.98

distance (JSD) between the GEANT4 truth and generation results of the six one dimensional histograms shown in Fig. 6, namely the visible cell energy, the total shower energy, the occupancy, the center of gravity in z as well as the radial and longitudinal energy distributions. These are some of the most relevant global differential distributions for photon shower analysis and were applied previously to judge model performances [14]. Additional details on how the score is calculated are given in Appendix A.

In Table 1 we show the fidelity score for different values of the latent space size n . Lower values correspond to better agreement with the underlying slow simulation. For very small n , the performance increases with n until the best observed value at $n = 24$. Seven trainings with identical network setup but different random weight initialization were performed for this point to obtain an estimate of the associated uncertainty (calculated as the standard deviation of individual results at $n = 24$). Further hyperparameters were kept the same as in Ref. [14]. In the table the best score out of those trainings is given. Limited computing resources due to several days of training needed per model did not allow for a wider estimation of the fidelity scores. For larger n the performance is approximately stable within the uncertainty observed for $n = 24$. This implies that maximum information content of ≈ 45 nats encoded in the latent space is not needed for optimal generative performance.

3 Correlations between latent space and physics

As only a few variables seem to encode most of the shower information, we investigate what kind of physics information is learned by these variables. In Fig. 3 the Pearson correlation coefficients ρ between different shower physics distributions and the distributions of the sampled z_i for the five highest KLD latent variables as well as the incident particle energy (which is used for conditioning and is included as a latent variable in the BIB-AE) are shown for four different model configurations. In this case the sampled z_i values are obtained from the encoded latent space via $\mathcal{N}(\mu_i, \sigma_i^2)$, not from a Standard Normal distribution $\mathcal{N}(0, 1)$. The physics distributions include the first and second moment in each spatial dimension² — the first moment corresponds to the center of gravity — the visible energy, the incident particle energy, the number of hits, and the fraction of deposited energy in each third of the calorimeter (in the z -direction).

Regardless of the model configuration, it is apparent that the highest KLD latent variable strongly correlates (approx. $\rho = 0.9$) with the center of gravity along the shower direction z (and in turn to the fraction of deposited energy in the first and last third of the calorimeter). Another variable is correlated (approx. $\rho = 0.5$) to the second moment in z (and to the energy fraction in the middle of the calorimeter). It appears that of all the shower variables, the center of gravity in z (CoG-Z) of each shower is encoded into these two latent variables. This is important as the CoG-Z is not as much correlated to the incident particle energy (approx. $\rho = 0.4$) on which the BIB-AE is conditioned. Hence the BIB-AE learns the CoG-Z of each shower and uses it in the decoding step for reconstruction. Interestingly, this pattern is very stable over multiple independent training runs and even different latent space sizes n .

²The incident photon enters the calorimeter in the center of the x - y plane at $z=0$ and traverses along the z -axis.

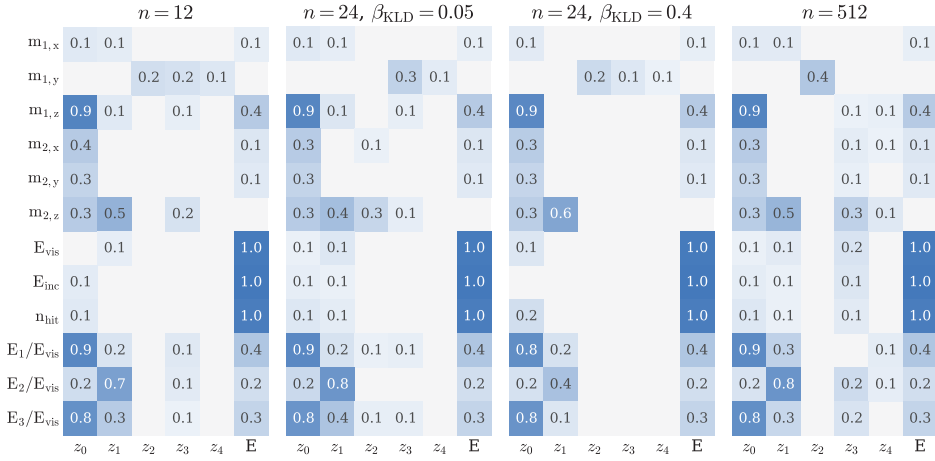


Figure 3. Pearson correlation coefficients between various physics variables and the sampled z_i of the five highest KLD latent variables as well as the incident particle energy E for multiple model sizes n . The baseline latent weight is $\beta_{KLD} = 0.05$ except for one training with $\beta_{KLD} = 0.4$. Only non-zero values of the correlations are shown.

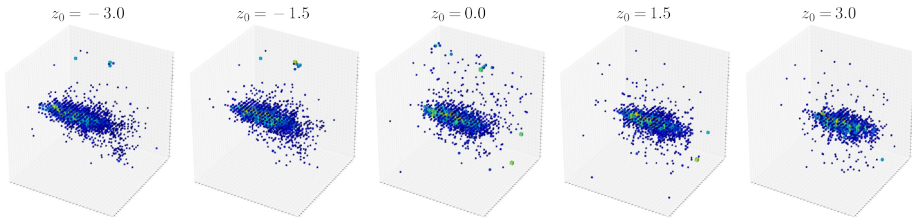


Figure 4. 3d image of generated showers decoded from a latent space with all variables $z_i = 0$, except the highest KLD latent z_0 variable which is set to values between -3 and 3. The color coding corresponds to the cells' energy deposition. The highest KLD variable z_0 correlates to the CoG-Z distribution, hence an evolution of the shower start is visible (projection found in Fig. 5).

We can use this observation to improve the CoG-Z distribution in the generated events (see Fig. 6 (bottom left)). This distribution was previously not particularly well-modeled since the generation did not take this latent space correlation into account. In addition, the targeted sampling of a subspace of these latent variables allows to generate showers with a specific shower start. This is visualized in Fig. 4 with multiple 3d images of a decoded calorimeter shower in which only the highest KLD latent z variable was altered. This variable change leads to a different shower start and hence to an altered center of gravity in the z-axis. Figure 5 visualizes the energy deposition per layer in z-direction of these five decoded shower images.

4 Improving generative performance

Understanding the encoded shower information, particularly the center of gravity, in the latent space helps us make educated optimization choices for improving model performance.

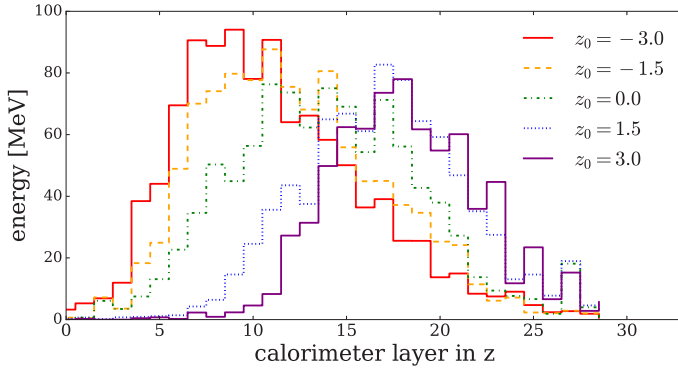


Figure 5. Deposited energy per layer in z-direction for showers which are decoded with all latent variables $z_i = 0$, except the highest KLD latent z_0 variable which is set to values between -3 and 3.

Specifically, we can increase generation fidelity by either regularizing the latent space more strongly or by leveraging and sampling from the information rich non-Gaussian distributions. Either optimization path can be approached in different ways. We have chosen one exemplary method for each: (1) By increasing β_{KLD} the overall KLD in the latent space is reduced, yielding latent distributions stronger regularized towards Standard Normal distributions and therefore more accurate generative sampling from such a $\mathcal{N}(0, 1)$ distribution; or (2) keeping the already trained model but using a second density estimator — such as Kernel Density Estimation (KDE) [25] — on the latent variables and sampling directly from the encoded latent space. The former approach is motivated by [26] while the latter mirrors a method for the *Buffer-VAE* from Ref. [27].

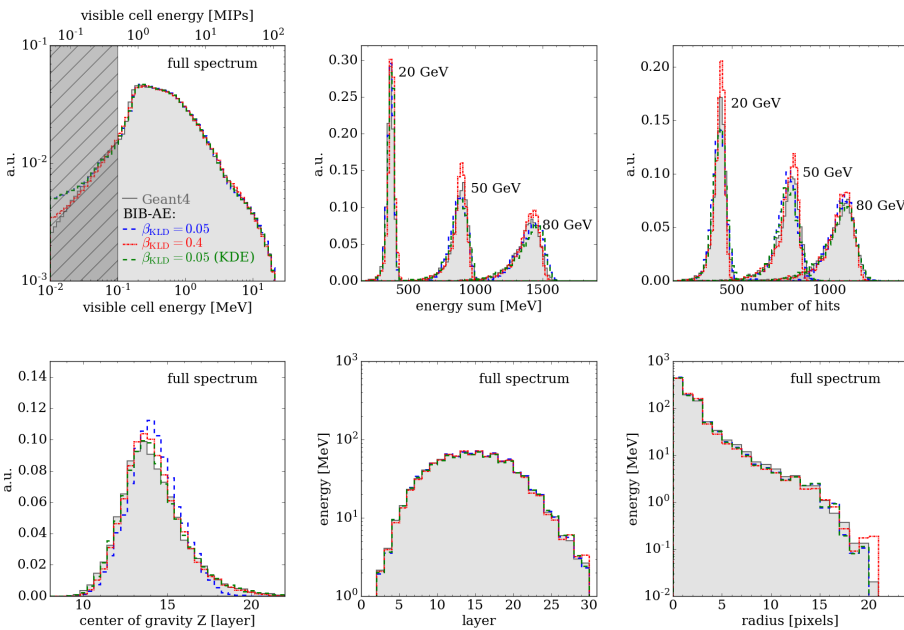


Figure 6. Differential distributions comparing physics quantities between GEANT4 and BIB-AE models with $\beta_{KLD} = 0.05$, $\beta_{KLD} = 0.4$ and $\beta_{KLD} = 0.05$ with the KDE sampling approach.

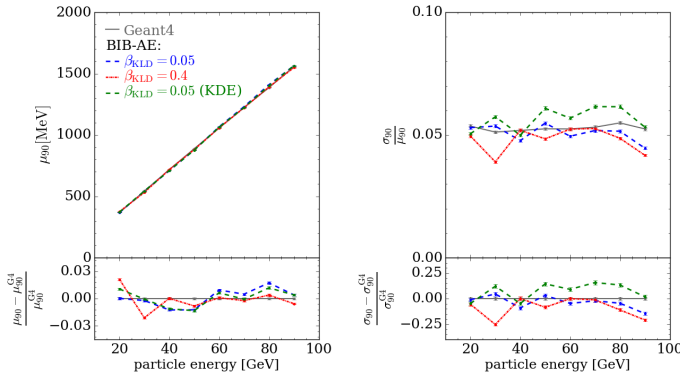


Figure 7. Mean and relative width of the energy deposited in the calorimeter for various incident particle energies for `GEANT4` and BIB-AE models with $\beta_{KLD} = 0.05$, $\beta_{KLD} = 0.4$ and $\beta_{KLD} = 0.05$ with the KDE sampling approach.

Table 2. Fidelity score S_{JSD} for the best epochs for multiple model and sampling configurations of BIB-AE models with a latent size of 24. For $\beta_{KLD} = 0.05$ the best score out of multiple training runs is given, while the mean score for those trainings is: $\bar{S}_{JSD,24} = 1.02 \pm 0.12$. For $\beta_{KLD} = 0.4$ only one training was performed.

config.	$\beta_{KLD} = 0.05$	$\beta_{KLD} = 0.4$	$\beta_{KLD} = 0.05$ +KDE sampling
S_{JSD}	0.83	0.88	0.67

4.1 Adjusting the Kullback-Leibler divergence

Our baseline model uses a latent KLD weight of $\beta_{KLD} = 0.05$. However, as a higher value for β_{KLD} leads to a lower KLD value, less information is encoded in the latent space. Therefore, the latent space more closely approaches a Standard Normal distribution and sampling from $\mathcal{N}(0, 1)$ in the generation step should yield showers resembling the `GEANT4` truth more closely. As shown in Fig. 6 (bottom left) this improves the CoG-Z distribution compared to the baseline. However, there is a trade-off for other distributions, such as the total energy or energy sum (top center) and the number of hits (top right) which become narrower than the baseline and truth distributions. This can also be seen in Fig. 7: Except for low energies the energy linearity is better, but the relative width of the energy distributions is on average narrower than the baseline model.

Figure 8 illustrates that for the highest (left) and second-highest (right) KLD latent variables, the sampled z distributions for $\beta_{KLD} = 0.4$ are very similar to Normal distributions while they deviate significantly for the baseline value of $\beta_{KLD} = 0.05$. Although improving the CoG-Z distribution, the overall fidelity score given in Table 2 is slightly worse for $\beta_{KLD} = 0.4$.

4.2 Sampling from a Kernel Density Estimate

Another way to improve the generative performance, particularly the CoG-Z distribution, is to utilize latent variables highly correlated to the CoG-Z distribution. Using exactly the same model as in Ref. [14] ($\beta_{KLD} = 0.05$, $n = 24$) without retraining one can see in Fig. 8 that the encoded distribution deviates from a Standard Normal distribution. In the usual VAE-like setup one would regardless sample these variable from $\mathcal{N}(0, 1)$ to generate new samples, thereby ignoring the correlations between the latent space and the shower physics. Instead one could sample those latent variables from the distribution of the sampled z_i values, which are sample from the encoded $\mathcal{N}(\mu_i, \sigma_i^2)$ distributions. Since at least two variables as

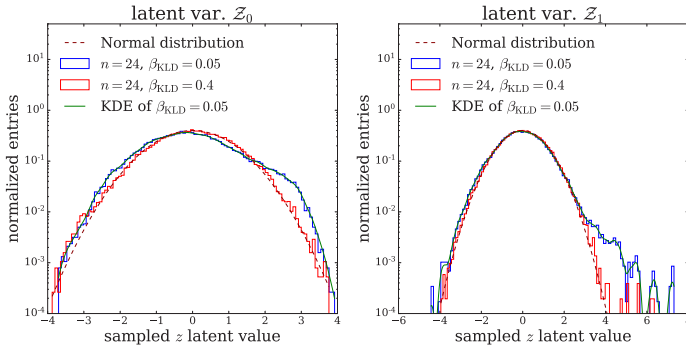


Figure 8. Sampled z values of the highest (left) and second-highest (right) KLD latent variables for 50k shower images for models with a latent size of 24 and $\beta_{KLD} = 0.05$ or $\beta_{KLD} = 0.4$. For reference added lines for a Normal distributions and the Kernel Density Estimate of the $\beta_{KLD} = 0.05$ histograms.

well as the incident energy are correlated to the CoG-Z distribution, one needs to account for correlations between latent variables when sampling. This can be done by encoding a sufficiently large number of showers (i.e. 500k) from the training set, applying a density estimation method such KDE, and then sampling new latent variables from it. In the BIB-AE case with 24 encoded latent variables plus energy conditioning, this leads to training a KDE of a 25-dimensional space. The resulting KDE kernel can be used as a probability density function for sampling the latent z variables for improved shower generation.

As shown in Fig. 6 this KDE sampling approach yields global differential distributions very similar to the GEANT4 truth; superior results for the CoG-Z distribution and the number of hits distributions in comparison to the other two models. The linearity in Fig. 7 closely resembles the baseline model, however the relative width of the energy distributions is on average slightly overestimated. The fidelity score in Table 2 is the best of all tested model configurations. The score for the model of $\beta_{KLD} = 0.05$ was chosen as the best out of seven training runs and the same model was used to simply add the KDE sampling step. This illustrates another benefit of the KDE approach: It can be applied to any already trained VAE-like model without expensive re-training.

5 Summary & conclusions

Improving the simulation of calorimeter showers with generative models is an active topic of research motivated by these tasks' large resource consumption. As such generative models still require substantial training efforts and preclude large hyperparameter scans for optimization, we investigate how a better understanding of the latent space can be used to increase performance. While a BIB-AE architecture was used for these studies, the developed strategies should readily transfer to other generative models with an encoded latent space (i.e. VAE-like but not GAN-like architectures).

We first quantify the information encoded in the latent space and note that for a fixed value of $\beta_{KLD} = 0.05$, it saturates at ≈ 45 nats. However, generative performance — as measured by a metric defined to take the relevant physical distributions into account — achieves its best value at a latent space of $n = 24$ with ≈ 28 nats. Put differently, more information encoded in the latent space will not necessarily translate into better generative performance.

This observation offers an interesting parallel to the information bottleneck principle [13, 16]. It proposes that for a supervised classification task, the latent space \mathcal{Z} should maximise its mutual information I with the true class labels \mathcal{C} but minimise information irrelevant for classification between data examples \mathcal{X} and latent space:

$$\mathcal{L}_S(\phi) = I_\phi(\mathcal{X}; \mathcal{Z}) - \beta I(\mathcal{Z}; \mathcal{C}). \tag{4}$$

Here \mathcal{L}_S is the supervised optimisation target, we minimise over parametric mappings ϕ from data to latent space, and the Lagrange multiplier β denotes the trade-off between the two goals.

For unsupervised tasks, no class labels are available, and the problem becomes:

$$\mathcal{L}_U(\phi) = I_\phi(\mathcal{X}; \mathcal{Z}) - \beta I(\mathcal{Z}; \mathcal{X}) \quad (5)$$

which is also the core of the BIB-AE loss formulation [13]. It is a much more challenging compression problem as the entropy of a small number of class labels will, in general, be much smaller — and therefore easier to encode — than the entropy of the data distribution. We observed that without additional constraints, such as restricting the latent space size n , more information than needed for good generative performance is encoded in the latent space, suggesting the need for additional regularising constraints. An interesting open question for future research is therefore how the *useful* encoded information might be quantified.

Regardless of the model configuration, only a few latent variables of the BIB-AE contain most of the shower information. Correlating the latent variables with various shower physics metrics reveals that the center of gravity in z-direction is always encoded into the two highest KLD latent variables. This encoding can be leveraged for targeted shower generation of photon showers with a specific shower start by sampling from a subspace of the highest KLD variable.

Furthermore, this observations can help improve the generative fidelity of the BIB-AE model. This can be achieved either by lowering the encoded KLD or by sampling directly from the encoded latent space density distribution, e.g. learned via Kernel Density Estimation. Forcing the latent distributions closer to unit Normal naturally improves physical observables most strongly correlated with the corresponding latent space variables with the highest-KLD values, and decreases the performance of the others. The latter approach yields the best results with the additional benefit of applying to the already previously trained BIB-AE model (or any other VAE-like model).

The increasing use of generative machine learning models motivates a closer look into their learned encoding. Especially in particle physics, the needed precision for many differential distributions over many orders of magnitude offers a rich laboratory to study the connection between generation fidelity and latent space. On the one hand, this offers several methods to probe and improve generative performance, for example by identifying poorly modeled distributions for which a discrepancy between encoded-into and sampled-from latent space exists. Resolving this discrepancy yields better-generated showers. On the other hand, the observed difference between maximum-information and best-performance latent space capacity raises an interesting problem for future studies.

Acknowledgements

We would like to thank the Maxwell and National Analysis Facility (NAF) computing centers at DESY for the smooth operation and technical support. E. Buhmann is funded by a scholarship of the Friedrich Naumann Foundation for Freedom and 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 was funded through the Helmholtz Innovation Pool project AMALEA that provided a stimulating scientific environment for parts of the research done here.

References

- [1] R. Jansky, *The ATLAS Fast Monte Carlo Production Chain Project* (2015), J. Phys. Conf. Ser., Vol. 664, No. 7
- [2] I.J. Goodfellow et al., *Generative Adversarial Nets*, in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA, 2014), NIPS'14, p. 2672–2680, 1406.2661, <https://dl.acm.org/doi/10.5555/2969033.2969125>
- [3] M. Paganini, L. de Oliveira, B. Nachman, *Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters* (2018), 1705.02355
- [4] L. de Oliveira, M. Paganini, B. Nachman, *Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis* (2017), 1701.05927
- [5] M. Paganini, L. de Oliveira, B. Nachman, *CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks* (2018), 1712.10321
- [6] M. Erdmann, L. Geiger, J. Glombitza, D. Schmidt, *Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks* (2018), 1802.03325
- [7] M. Erdmann, J. Glombitza, T. Quast, *Precise simulation of electromagnetic calorimeter showers using a Wasserstein Generative Adversarial Network* (2019), 1807.01954
- [8] ATLAS Collaboration, Tech. Rep. ATL-SOFT-PUB-2018-001, CERN, Geneva (2018), <http://cds.cern.ch/record/2630433>
- [9] ATLAS Collaboration, Tech. Rep. ATL-SOFT-SIM-2019-007, CERN (2019), <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-007/>
- [10] A. Ghosh (ATLAS Collaboration), Tech. Rep. ATL-SOFT-PROC-2019-007, CERN, Geneva (2019), <https://cds.cern.ch/record/2680531>
- [11] D. Belayneh et al., *Calorimetry with Deep Learning: Particle Simulation and Reconstruction for Collider Physics* (2019), 1912.06794
- [12] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, *GANplifying Event Samples* (2020), 2008.06545
- [13] S. Voloshynovskiy, M. Kondah, S. Rezaeifar, O. Taran, T. Holotyak, D.J. Rezende, *Information bottleneck through variational glasses* (2019), 1912.00830
- [14] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, K. Krüger, *Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed* (2021), 2005.05334
- [15] D.P. Kingma, M. Welling, *Auto-Encoding Variational Bayes* (2014), 1312.6114
- [16] N. Tishby, F.C. Pereira, W. Bialek, *The information bottleneck method* (2000), arXiv preprint physics/0004057 [physics.data-an]
- [17] J.N. Howard, S. Mandt, D. Whiteson, Y. Yang, *Foundations of a Fast, Data-Driven, Machine-Learned Simulator* (2021), 2101.08944
- [18] J. Batson, C.G. Haaf, Y. Kahn, D.A. Roberts, *Topological Obstructions to Autoencoding* (2021), 2102.08380
- [19] S. Agostinelli et al., *Geant4—a simulation toolkit* (2003), Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**, 250, <http://www.sciencedirect.com/science/article/pii/S0168900203013688>

- [20] H. Abramowicz et al. (ILD Concept Group), *International Large Detector: Interim Design Report* (2020), **2003.01116**
- [21] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A.J. Smola, *A Kernel Method for the Two-Sample Problem* (2008), **0805.2368**
- [22] C.E. Shannon, *A mathematical theory of communication*. (1948), Bell Syst. Tech. J. **27**, 379, <http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48>
- [23] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959)
- [24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, *Improved Techniques for Training GANs* (2016), **1606.03498**
- [25] E. Parzen, *On Estimation of a Probability Density Function and Mode* (1962), The Annals of Mathematical Statistics **33**, pp. 1065, <http://www.jstor.org/stable/2237880>
- [26] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, in *ICLR* (2017)
- [27] S. Otten, S. Caron, W. de Swart, M. van Beekveld, L. Hendriks, C. van Leeuwen, D. Podareanu, R.R. de Austri, R. Verheyen, *Event Generation and Statistical Sampling for Physics with Deep Generative Models and a Density Information Buffer* (2019), **1901.00875**

A Fidelity score

Comparing histograms of shower variables such as total energy, number of hits, shower profile and center of gravity as shown in Fig. 6 is a way to determine the generation performance of the generative model in comparison to the GEANT4 simulation. It is however difficult to quantify the model improvement by manually observing these plots. A quantification of the 'generation performance' or 'fidelity' can be calculated via the difference between the histograms of generated and GEANT4 observables. This can be done for example by calculating the Jensen-Shannon distance (JSD) by considering each histogram as a discrete probability density distribution. As an alternative we have calculated a fidelity score based on the area difference between the histograms. This score was comparable to our fidelity score S_{JSD} . A similar fidelity metric was calculated in Ref. [27].

The JSD can be calculated for each of the six histograms in Fig. 6. To have one score combining all six histograms one needs to weight each individual histograms' JSD in comparison to all other JSDs of the same model. This weighting is done in the following way:

1. Calculate JSD for each of the six plots for each model configuration and epoch: $\text{JSD}_{i,m,e}$ with i for 1 in 6 plots, m for 1 in x models, and e for 1 in y epochs that are compared in the score
2. Calculate the $\bar{6}$ weighting factor for the JSD of each i plot:
 $\langle \text{JSD}_i \rangle = \overline{\text{JSD}_{i,m,e}}$ for each plot i
3. Calculate the fidelity score S_{JSD} for each model m and epoch e :

$$S_{\text{JSD},m,e} = \langle \text{JSD}_{m,e} \rangle = \frac{1}{6} \sum_i \text{JSD}_{i,m,e} \cdot \frac{1}{\langle \text{JSD}_i \rangle}$$

An example of this weighted S_{JSD} score is shown in Fig. 9 for an epoch-wise scan during the training of two models with different β_{KLD} weights; each with and without the Post-Processor network. Note that the KLD is increasing with each epoch and saturates over time. However, a higher KLD does not necessarily correlate with a lower fidelity score.

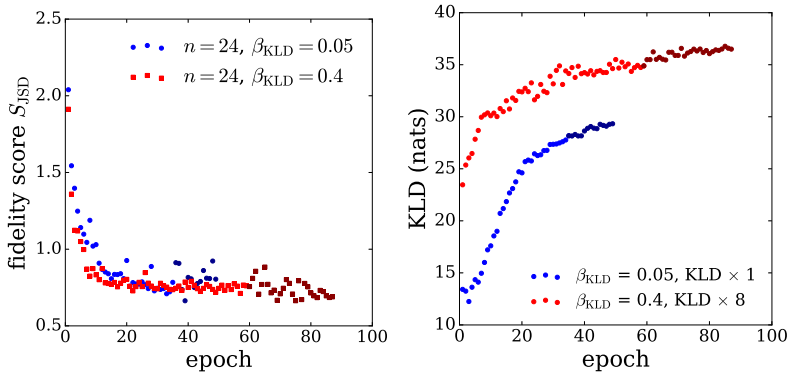


Figure 9. Evolution of the fidelity score S_{JSD} and the KL divergence over the course the of training for the two models with $\beta_{KLD} = 0.05$ and $\beta_{KLD} = 0.4$. Based on the fidelity score the best epochs were chosen (epoch 39 and epoch 87 respectively). Color brightness implies training with or without the Post-Processor network (see Sec. 1.2).