


# Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation

Claudius Krause<sup>\*</sup> and David Shih<sup>†</sup>

*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*

 (Received 25 July 2022; accepted 19 April 2023; published 28 June 2023)

Recently, we introduced CaloFlow, a high-fidelity generative model for GEANT4 calorimeter shower emulation based on normalizing flows. Here, we present CaloFlow v2, an improvement on our original framework that speeds up shower generation by a further factor of 500 relative to the original. The improvement is based on a technique called probability density distillation, originally developed for speech synthesis in the machine learning literature, and which we develop further by introducing a set of powerful new loss terms. We demonstrate that CaloFlow v2 preserves the same high fidelity as the original using qualitative (average images, histograms of high-level features) and quantitative (classifier metric between GEANT4 and generated samples) measures. The result is a generative model for calorimeter showers that matches the state of the art in speed (a factor of  $10^4$  faster than GEANT4) and greatly surpasses the previous state of the art in fidelity.

DOI: [10.1103/PhysRevD.107.113004](https://doi.org/10.1103/PhysRevD.107.113004)

## I. INTRODUCTION

The enormously successful physics program at the LHC relies heavily on the availability of copious amounts of highly accurate simulated data. However, the use of GEANT4 [1–3] for full detector simulations is a major computational bottleneck and severely limits the analysis capabilities of the LHC. This is forecast to worsen significantly with future LHC upgrades and the HL-LHC [4–8].

Recently, deep generative modeling has demonstrated great potential to speed up the most computationally expensive part of detector simulations, namely, calorimeter showers [8–19]. By fitting the generative model to GEANT4 shower images, the generative model learns (often implicitly) the underlying distribution that the GEANT4 showers are drawn from and can then sample from it quickly. Most of the current approaches [8–18] are based on generative adversarial network (GAN) or variational autoencoder (VAE) architectures. Very recently, in [19], we proposed a fresh alternative, dubbed CaloFlow, based on normalizing flows (for recent reviews and original references, see, e.g., [20,21]). Flows have many advantages over GANs and VAEs—stable, convergent, principled training and model selection, based on minimizing the negative log-likelihood (negative LL, NLL) objective, no issues with mode collapse, explicitly learning a differentiable likelihood function, and fully bijective mapping to latent space. Correspondingly, we found that the quality of the generated samples was astonishing; unlike the baseline GAN that we compared

with [9,10], the samples produced with CaloFlow were almost indistinguishable from GEANT4 samples [19].<sup>1</sup>

However, the approach presented in [19] had one major downside compared to the other deep generative models: sampling speed. While CaloFlow was  $\sim 50$  times faster than GEANT4, it was considerably slower (by a factor of  $\sim 500$ ) than the GAN-based alternative. The disadvantage in generation speed was because of the masked autoregressive flow (MAF) architecture [23] used in [19], which is only fast in density estimation, but a factor 500 (given by the dimensionality of the readout channels of the detector) slower in sampling. Normalizing flows can also be constructed to be fast in the other direction (fast sampling, slow density estimation); this goes by the name of inverse autoregressive flow (IAF) [24]. Indeed, using an IAF would bring CaloFlow fully in line with the speed of the GAN-based approaches. However, what sounds like the solution does not work in practice, as IAFs with such a high dimensionality cannot be trained using the NLL objective, due to time and memory limitations, as we will elaborate on.

In this paper, we overcome the challenges of training the IAF, by building on a method alternately called “probability density distillation” or “teacher-student training” in the machine learning (ML) literature. Originally developed for the purposes of speech synthesis in [25], the basic idea is that the IAF (called the “student”) can be trained

<sup>1</sup>It is possible that the fidelity of the GAN could be improved with more sophisticated architectures compared to the original CaloGAN [9,10]; however, see [22] for an example of a state-of-the-art GAN architecture that also produces samples that can be easily distinguished from the training data with a classifier.

<sup>\*</sup>Claudius.Krause@rutgers.edu

<sup>†</sup>shih@physics.rutgers.edu

efficiently not on the original target data, but on the output of a trained MAF model (called the “teacher”). The MAF can quickly map target data points to the latent space  $x \rightarrow z$ , and the IAF inverse can quickly map the latent space back to the target data space  $z \rightarrow x'$ . By requiring this loop to close ( $x' = x$ ), i.e., by requiring the MAF and IAF to describe the same transformation or, in other words, the fast passes to be each others inverses, one can, in principle, get an arbitrarily good fit of the IAF to the MAF (and by extension, to the original target data).

The original idea of probability density distillation presented in [25] relied on minimizing the Kullback-Leibler (KL) divergence between the IAF and MAF. However, even in that work, they found that this training objective was insufficient, and they added additional, *ad hoc* loss terms based on high-level features in order to get the IAF to converge to the MAF. Subsequently, [26] explored more well-motivated and general alternatives to the KL divergence loss, based on the distance between  $x$  and  $x'$  in the MAF-IAF loop or, analogously,  $z$  and  $z'$  in the IAF-MAF loop. These provide an alternative measure of closure that has improved convergence and convexity properties compared to the KL divergence. Other works in the ML literature that explore variants on the idea of probability density distillation include [27–29].

Here we build on the version of probability density distillation presented in [26]. We found that, while using  $x-x'$  and  $z-z'$  distances offered significant improvements to the KL divergence, still they were insufficient for achieving a good fit of the IAF to the MAF. But by adding additional measures of the MAF-IAF closure, involving the intermediate layers and actual transformation parameters of the invertible normalizing flow (see Sec. II B for details), we were able to obtain an excellent fit of the IAF to the MAF. Along the way, we also devise a new method for model selection of the IAF. Since the NLL is too expensive to compute for every epoch, we instead use the much cheaper KL divergence. Although the KL divergence is not a suitable objective for training the IAF, we show that it nevertheless tracks the NLL very closely and therefore can serve as an effective proxy for selecting the best model epoch.

The result is a new version of CaloFlow (and a new version of probability density distillation) that is just as fast as the GAN baseline, while just as high fidelity as the MAF used in [19]. We demonstrate this using the same qualitative measures as in [19], as well as the classifier metric introduced in [19].

The outline of our paper is as follows. Section II A reviews the construction of the MAF and IAF normalizing flows and explains why the former is fast to density estimate and slow to sample, while the latter is the opposite. Section II B explains the idea behind probability density distillation and describes our new loss terms that greatly improve the matching of the student to the teacher.

Section III very briefly describes the dataset used for this work; for more details see [9,10,19]. Section IV describes the architecture and training procedure of CaloFlow v2. Section V contains the results of the teacher-student training—average images, histograms, classifier metric, and timing. Finally, we summarize and conclude in Sec. VI. In the Appendix, we present plots of nearest neighbors between GEANT4 and CaloFlow student samples, providing further evidence against mode collapse in the latter.

## II. DENSITY ESTIMATION AND PROBABILITY DISTILLATION WITH NORMALIZING FLOWS

### A. MAFs vs IAFs: Fast and slow directions

Our work uses autoregressive flows to learn the invertible transformation between the data space  $x \in \mathbb{R}^d$  and the latent space  $z \in \mathbb{R}^d$ . These autoregressive flows take the form

$$z_i = f(x_i; \vec{\kappa}_i), \quad x_i = f^{-1}(z_i; \vec{\kappa}_i), \quad i = 1, \dots, d, \quad (1)$$

where  $f$  is an invertible 1D transformation [we use rational quadratic splines (RQSs) [30,31]], and  $\vec{\kappa}_i$  are a set of coordinate-dependent parameters for the transformation of the  $i$ th coordinate. To preserve the autoregressive property, the parameters should only depend on the previous coordinates (i.e., those with index less than  $i$ ). But we have a choice as to whether we make the parameters explicit functions of the  $x$  coordinates or the  $z$  coordinates. That is, we must have either

$$\vec{\kappa}_i = \vec{\kappa}_i(x_1, \dots, x_{i-1}) \quad (2)$$

or

$$\vec{\kappa}_i = \vec{\kappa}_i(z_1, \dots, z_{i-1}). \quad (3)$$

In our setup, all these parameters are the output of a neural network (using the masked autoencoder for distribution estimation (MADE) block [32]), and the autoregressive property is accomplished through the application of a binary mask on the internal hidden layers.

The first choice, Eq. (2), defines the MAF architecture [23]. When the parameters are explicit functions of  $x_i$ , then  $x \rightarrow z$  (the “forward” pass for “inference,” also known as density estimation) is fast, requiring just a single evaluation of the neural networks. However, to perform the inverse transformation  $z \rightarrow x$  (for sampling), one must compute

$$x_i = f^{-1}(z_i, \vec{\kappa}_i(x_1, \dots, x_{i-1})). \quad (4)$$

Now the  $x_1, \dots, x_d$  are only known recursively, i.e.,

$$\begin{aligned}
x_1 &= f^{-1}(z_1, \vec{\kappa}_1), \\
x_2 &= f^{-1}(z_2, \vec{\kappa}_2(x_1)) = f^{-1}(z_2, \vec{\kappa}_2(f^{-1}(z_1, \vec{\kappa}_1))), \\
x_3 &= f^{-1}(z_3, \vec{\kappa}_3(x_1, x_2)) = f^{-1}(z_3, \vec{\kappa}_3(f^{-1}(z_1, \vec{\kappa}_1), \\
&\quad f^{-1}(z_2, \vec{\kappa}_2(f^{-1}(z_1, \vec{\kappa}_1))))), \\
&\dots
\end{aligned} \tag{5}$$

So to evaluate the inverse transformation  $z \rightarrow x$  requires  $d$  successive evaluations of the neural network. The MAF is fast to density estimate but a factor of  $d$  slower to sample.

The second choice, Eq. (3), defines the IAF architecture [24]. In that case, the opposite is true: sampling  $z \rightarrow x$  is fast, while density estimating  $x \rightarrow z$  is a factor of  $d$  slower.<sup>2</sup> Unfortunately, this also means that training an IAF with the LL objective (which requires  $x \rightarrow z$  density estimation) would take a factor  $d$  more time than training the MAF.<sup>3</sup> Given that the MAF takes approximately  $\sim \mathcal{O}(1 \text{ hr})$  to train, and  $d \sim 500$  in our setup, this means that the IAF would be prohibitively time consuming for us to train. Instead, training an IAF generative model for calorimeter showers requires a very different approach.

## B. Probability density distillation and teacher-student training

The key idea for how to train an IAF efficiently, which we build upon in this work, was introduced in [25] in the context of speech synthesis and given the name of probability density distillation or teacher-student training. The idea is that while fitting the IAF directly to data is practically prohibitive, fitting the IAF to the MAF is not. By starting from a sample  $z$  in the latent space and mapping it to data space via the student (IAF), we get a set of data samples  $x$  and their likelihood  $s(x)$  under the student efficiently. Mapping it back to latent space with the teacher (MAF) yields the log-likelihood of the same sample under the teacher  $t(x)$ . Every pass is then the fast one under its respective autoregressive flow.

In [25], the loss function was initially taken to be the KL divergence between these two probability densities,

$$\text{KL} = \int s(x) \log \frac{s(x)}{t(x)} dx = \sum_{x \sim S} \log \frac{s(x)}{t(x)}. \tag{6}$$

<sup>2</sup>Note that the forward direction of the IAF also refers to density estimation.

<sup>3</sup>In addition, since each coordinate transformation depends on the result of a pass through the network and the previous coordinates, which in turn depend on passes through the network [see Eq. (5)], the memory needed to store the gradients exceeds the memory requirement of a MAF by a large factor. In principle, one could be able to optimize the storage of these gradients similar to backpropagation, but this is currently not implemented in any of the available ML code frameworks.

Note that this KL divergence is based on the same  $x$ . Starting from data and closing the loop through the teacher first and then through the student gives different  $x$  and  $x'$  at the beginning and the end of the chain, so instead one would need to calculate KL in  $z$  space, which is not meaningful, since the base distributions of teacher and student are identical.

Although, in principle, the KL divergence of Eq. (6) is a good loss—it is non-negative and zero if and only if the IAF and MAF densities agree—it was already found in [25] to not converge well to the desired result. The authors of [25] added additional, *ad hoc* high-level-feature-based loss terms to enhance the quality of their generated audio sample.

Reasons for why the KL divergence has poor convergence properties as a loss function were given in [26]. Since the IAF fast pass should be the inverse of the MAF fast pass, two alternative loss functions based on mean squared errors (MSE) were proposed,

$$L_x \equiv \text{MSE}(x, x') \tag{7}$$

and

$$L_z \equiv \text{MSE}(z, z'). \tag{8}$$

Here, one can start from data  $x$ , map it to latent space  $z$  with the MAF, and map it back to  $x'$  in the data space with the IAF; every pass is then the fast one under its respective autoregressive flow. Starting from noise and mapping back to noise,  $z \rightarrow x \rightarrow z'$  via IAF and then MAF is also possible. Either way, requiring that the transformation closes forces the IAF to conform to the MAF. These measure more directly the closure of the MAF-IAF loop.<sup>4</sup>

In this work, we go beyond the loss functions of Eqs. (7) and (8), as we observed that, while they do improve on the KL divergence (which generally does not converge at all), they still lead to poor overall agreement between the IAF and the MAF and a bad NLL of the trained student flow.

- (i) First, we found that using both  $L_x$  and  $L_z$  together (we tried the simple average of the two) was much better than using each one separately, as was considered in [26].
- (ii) The MAF and IAF actually parametrize a series of invertible autoregressive transformations. Looking at the fast passes through the flows, we can think of them as

$$x \rightarrow y_t^{(1)} \rightarrow y_t^{(2)} \rightarrow \dots \rightarrow y_t^{(N)} = z \tag{MAF} \tag{9}$$

and

<sup>4</sup>In principle, any distance measure in the data space and latent space could be used; for simplicity, we used simple Euclidean distance in image space and in the latent space and empirically this worked well.

$$z \rightarrow y_s^{(1)} \rightarrow y_s^{(2)} \rightarrow \dots \rightarrow y_s^{(N)} = x \quad (\text{IAF inverse}). \quad (10)$$

Here,  $y_t^{(i)}$  is understood to include the permutation after the transformation and  $y_s^{(i)}$  is understood to include the permutation before the transformation.<sup>5</sup> Since each step is an invertible transformation, we could require that the MAF and IAF to agree with each other at every step and not just at the end points. In other words, we could require that

$$y_s^{(1)} = y_t^{(N-1)}, y_s^{(2)} = y_t^{(N-2)}, \dots, y_s^{(N-1)} = y_t^{(1)}. \quad (11)$$

If we start from  $x$ , the full chain of transformations in the loop is

$$\begin{aligned} x &\rightarrow y_t^{(1)} \rightarrow \dots \rightarrow y_t^{(N)} = z, \\ z &\rightarrow y_s^{(1)} \rightarrow y_s^{(2)} \rightarrow \dots \rightarrow y_s^{(N)} = x', \end{aligned} \quad (12)$$

and we could enforce stepwise agreement with the losses

$$L_{x^{(i)}} = \text{MSE}(y_t^{(i)}(x), y_s^{(N-i)}(x)), \quad i = 1, \dots, N-1, \quad (13)$$

where we have explicitly indicated here that the  $y_t^{(i)}$  and  $y_s^{(j)}$  are functions of (originated from)  $x$ . Similarly, we can perform the loop starting with  $z$  and enforce stepwise agreement with

$$L_{z^{(i)}} = \text{MSE}(y_s^{(i)}(z), y_t^{(N-i)}(z)), \quad i = 1, \dots, N-1. \quad (14)$$

Including the sum of these losses in the training (in addition to  $L_x$  and  $L_z$ ) improved the student-teacher matching further.

- (iii) Finally, we could further exploit the constraint that the MAF and IAF parametrize the same transformation at each step and require that the parameters output by each MADE block agree between the IAF and MAF,

$$L_{\kappa_x^{(i)}} = \text{MSE}(\kappa_t^{(i)}(x), \kappa_s^{(N-i+1)}(x)), \quad i = 1, \dots, N, \quad (15)$$

and

$$L_{\kappa_z^{(i)}} = \text{MSE}(\kappa_s^{(i)}(z), \kappa_t^{(N-i+1)}(z)), \quad i = 1, \dots, N, \quad (16)$$

where the former (latter) set of  $\kappa_i$  is understood to be coming from a pass that started with data  $x$  (noise  $z$ ). Since this MSE captures the full spline and not just the bin the coordinate falls into, this has the potential to drive the student even closer to the teacher than the MSEs based on Eqs. (13) and (14). Indeed, we found that including the sum of these in the loss (together with those above) led to the best result.

The various loss terms are illustrated in Fig. 1. In Table I, we demonstrate the successive improvements to the NLL for  $e^+$  showers due to including these loss terms, after training for 150 epochs as described in Sec. IV B. We observe a clear improvement of the NLL the more terms are added to the loss. Evidently, the student does best if it is guided as closely as possible.<sup>6</sup>

In summary, our final objective function for the teacher-student training is as follows:

$$\begin{aligned} L = 0.5 &\left( \underbrace{\text{MSE}(z, z') + \sum_i^{N-1} \text{MSE}(y_s^{(i)}(z), y_t^{(N-i)}(z)) + \sum_i^N \text{MSE}(\kappa_s^{(i)}(z), \kappa_t^{(N-i+1)}(z))}_{z \text{ loss}} \right) \\ &+ 0.5 \left( \underbrace{\text{MSE}(x, x') + \sum_i^{N-1} \text{MSE}(y_t^{(i)}(x), y_s^{(N-i)}(x)) + \sum_i^N \text{MSE}(\kappa_t^{(i)}(x), \kappa_s^{(N-i+1)}(x))}_{x \text{ loss}} \right). \end{aligned} \quad (17)$$

We will refer to training with the objective given by Eq. (17) as “fully guided” student training.

<sup>5</sup>The permutation between the latent space and the adjacent MADE block is absorbed in the permutation-invariant base distribution.

<sup>6</sup>In principle, the student does not have to be an IAF, it could also be a simple, fully connected neural network [27]. However, in this case, we would not have access to the LL as a measure of quality and we would not be able to train it with the additional loss terms of Eqs. (13)–(16).

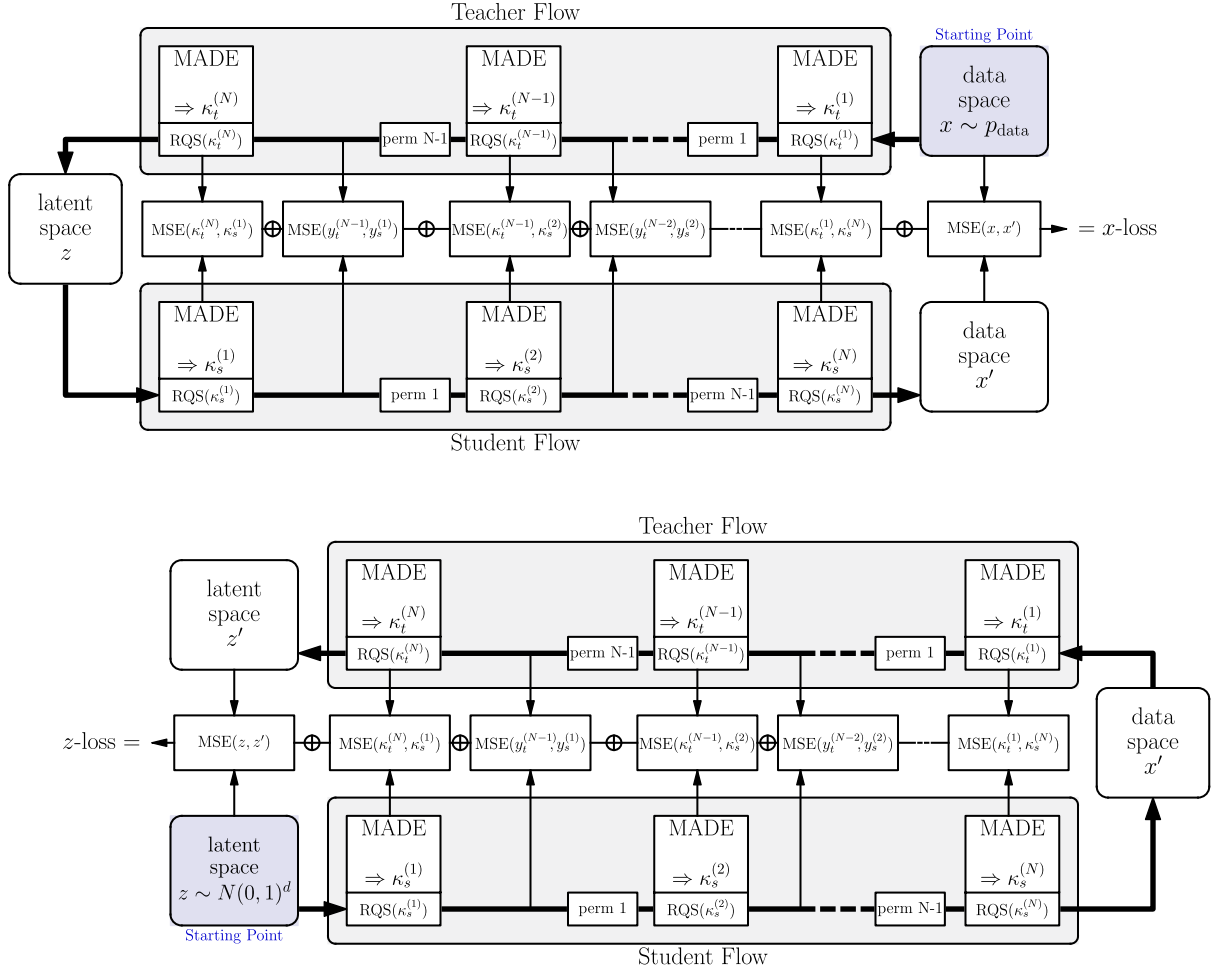


FIG. 1. Schematic view of the construction of the loss function. Top:  $x$  loss, actual data  $x$  is fed through the teacher and student starting from the top right (indicated in blue). Bottom:  $z$  loss, generated noise  $z$  is fed through the student and teacher starting from the lower left (indicated in blue). Intermediate coordinates (top,  $y_t^{(i)}$ ; bottom,  $y_s^{(i)}$ ) and RQS parameters (top,  $\kappa_t^{(i)}$ ; bottom,  $\kappa_s^{(i)}$ ) are compared in a MSE loss and then summed.

### III. CALORIMETER DATA

Since this is an improvement of CaloFlow v1 [19], we use the same calorimeter setup as there, which was based on CaloGAN [9,10]. Here we provide a very brief description; we refer the reader to [19] for details. The calorimeter is a simplified version of the ATLAS electromagnetic calorimeter. It has three layers of sizes  $3 \times 96$ ,  $12 \times 12$ , and  $12 \times 6$  voxels, respectively. The training data are showers of  $e^+$ ,  $\gamma$ , and  $\pi^+$  with energies uniformly sampled from 1–100 GeV and perpendicularly incident on the calorimeter simulated with GEANT4. These are the exact same samples [33] that were used to train and evaluate CaloFlow v1. For each particle, we have a set of 70,000 showers to train the flow, a set of 30,000 showers for model selection and validation of the flow, and additional sets of 60,000/20,000/20,000 showers to train, validate/calibrate, and test the classifier metric of Sec. VD.

TABLE I. NLL (smaller is better) after training CaloFlow for 150 epochs on  $e^+$  showers with different loss functions (sums over  $i$  are understood for all applicable cases). As a comparison, the teacher model has a NLL of 142.2.

Loss	NLL
$L_x$	1596.3
$L_z$	256.6
$L_x + L_z$	198.7
$L_x + L_z + L_{x^{(i)}} + L_{z^{(i)}}$	170.6
$L_x + L_z + L_{\kappa_t^{(i)}} + L_{\kappa_s^{(i)}}$	147.4
$L_x + L_z + L_{x^{(i)}} + L_{z^{(i)}} + L_{\kappa_t^{(i)}} + L_{\kappa_s^{(i)}}$	146.4

[Equation (17)]

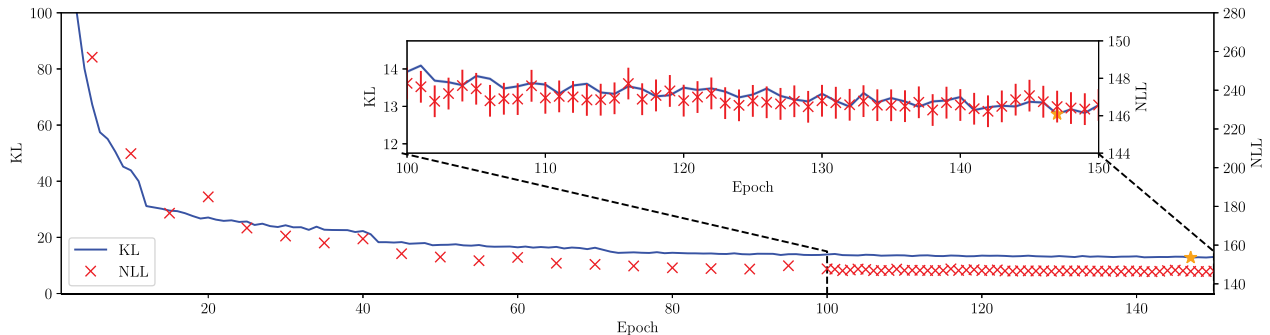


FIG. 2. KL divergence and NLL during training of the student IAF for  $e^+$  showers. The KL divergence is computed using 70,000 noise samples, while the NLL is computed using 10,000 GEANT4 samples from the validation set. Error bars on the latter show the standard error of the NLL estimate, with the error of the KL divergence being at or below its linewidth. The orange star marks the epoch with smallest KL divergence. The inset shows an enlargement of epochs 100 to 150.

#### IV. CaloFlow

##### A. Architecture

As in [19], we preserve the two-step structure of CaloFlow. In the first step, we use a small normalizing flow, called flow I, to learn the distribution of deposited energies conditioned on the input energy  $p_1(E_0, E_1, E_2|E_{\text{inc}})$ . In the second step, we use a much larger flow, called flow II, to learn the shower shapes conditioned on the energies  $p_2(\vec{\mathcal{I}}|E_0, E_1, E_2, E_{\text{inc}})$ .

Flow I is exactly as it was in [19]. In fact, we use the saved weights of [19] throughout this paper. We did not bother to train an IAF for flow I, since the time to sample from flow I is significantly smaller than the time to sample from flow II, so a factor of  $\sim 3$  speedup of flow I would have a negligible effect on the overall sampling time of CaloFlow v2.

Instead, we focus our attention in this work on training a student IAF for flow II, based on the teacher MAF for flow II from [19]. We use the same hyperparameters (8 blocks, 378 hidden neurons, 8 RQS bins) for the teacher as we used in [19]. (In fact, we use the saved weights of the MAF trainings from [19] for training the student here.) Since the student is much faster to evaluate, we could, in principle, make it bigger than the teacher. However, the fully guided blockwise loss that we introduced in Sec. II B requires the same number<sup>7</sup> of MADE blocks between the teacher and student. We are left with making the hidden layers wider, and we chose 504 nodes (to match the dimensionality of the voxel space), as we found that this improved the performance of the student. Another modification we considered was to make the neural network inside the MADE blocks deeper, but an initial study showed no improvement from this. Finally, we also considered making the teacher bigger (and hence slower) than in [19]; this showed no

<sup>7</sup>One could consider adding MADE blocks to the student and match groups of them to a single teacher MADE block, but we did not pursue this strategy here.

improvement in the LL of the teacher, probably due to overfitting. The IAF permutation  $i$  is taken to be the same as the MAF permutation  $N - i$ .

##### B. Training

We train the student of flow II as follows. We use the same training and validation datasets as for the teacher in CaloFlow v1, as we saw no sign of overfitting in this setup. For every epoch, we shuffle and divide up the 70,000 samples of the target GEANT4 training data into minibatches of 175 events. We feed these minibatches through the teacher MAF and back through the student IAF and obtain the  $x$  loss and gradients with respect to the student weights. During each minibatch, we also sample 175 events from the latent space. We feed these through the student IAF and back through the teacher MAF and obtain the  $z$  loss and gradients with respect to student weights. These are finally all combined together, and total loss is minimized with respect to the student weights via the Adam [34] optimizer for 150 epochs.

We found that increasing the minibatch size in training improves the convergence. To overcome memory constraints with too large minibatch sizes, we train the student using the gradient accumulation technique: the gradients of several minibatches are stored before a parameter update step is performed, effectively increasing the minibatch size.

We also found that optimizing with a rather sophisticated learning rate schedule helped the student converge better to

TABLE II. NLLs of the teacher and student flows, evaluated on the same validation set (lower is better).

Particle	CaloFlow v1 (teacher) NLL from [19]	CaloFlow v2 (student) NLL
$e^+$	142.159	146.393
$\gamma$	194.064	197.347
$\pi^+$	637.265	639.678

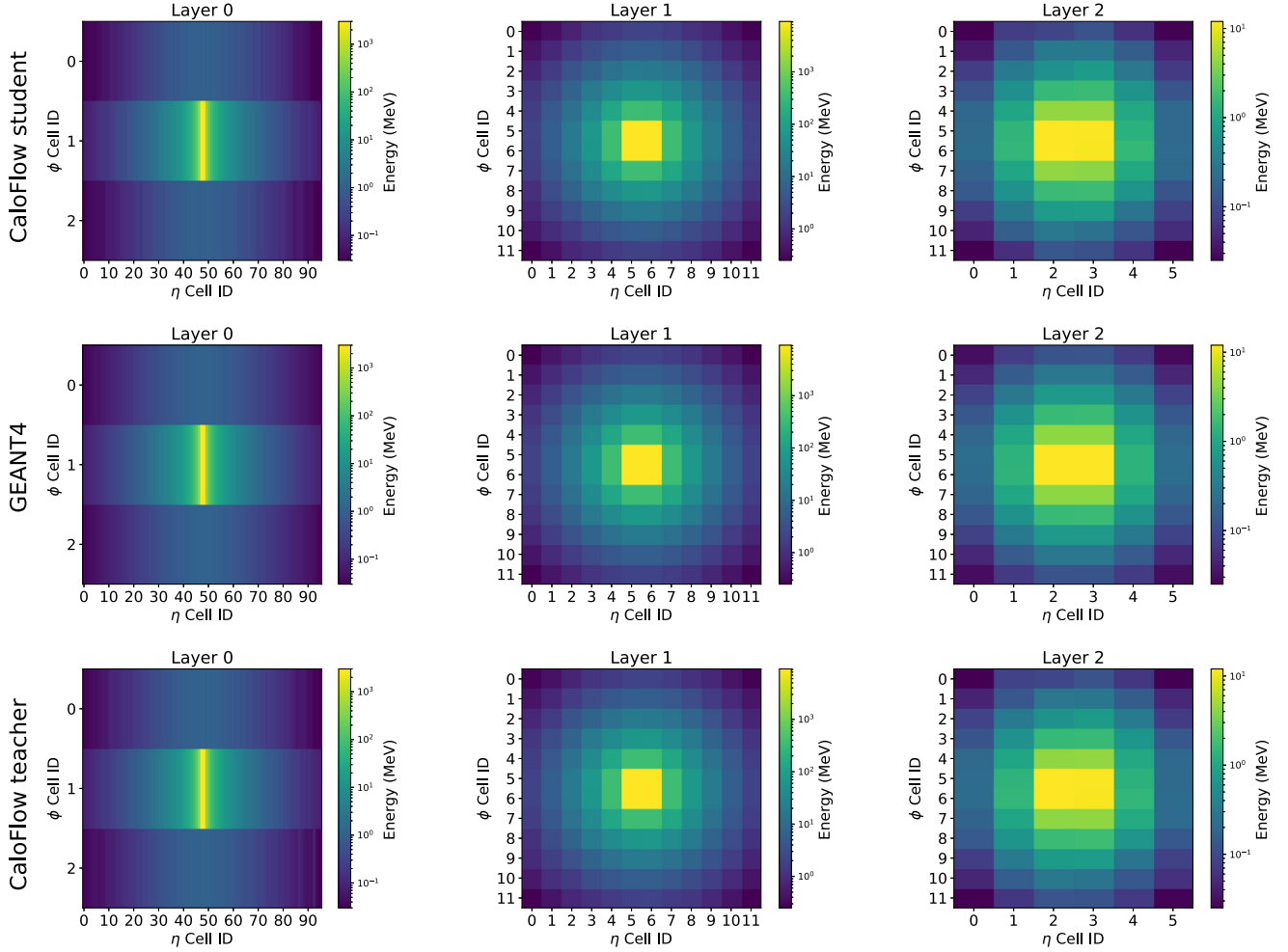


FIG. 3. Average shower shapes for  $e^+$ . Columns are calorimeter layers 0–2, top row shows CaloFlow student, center row GEANT4, and bottom row CaloFlow teacher.

the teacher. We start with a learning rate of  $2.5 \times 10^{-3}$  and a minibatch size of 175. In the first epochs, the gradients of two such minibatches are accumulated before the weight update is performed. After epochs 10, 40, and 70, we apply a factor of 0.5 to the learning rate and at the same time multiply the number of accumulated minibatches by 2. For example, after epoch 70, we accumulate 16 minibatches before the gradient update.

Finally, model selection for the flow II student is less straightforward than for the teacher. Again, since the NLL on the GEANT4 validation set is very expensive to compute for the student, we cannot use this to directly select the best model. However, as described in Sec. II, the KL divergence (6) between the teacher and student densities is very efficient to compute. Even though the KL divergence does not constitute a good loss term due to problems with its gradient [26], it is still a useful metric to judge the convergence between student and teacher. In

particular, we observe a strong correlation between the KL divergence and the NLL. This is illustrated in Fig. 2 for the training of  $e^+$  showers. We therefore compute the KL divergence of every training epoch and select the model state with the smallest KL divergence for the subsequent sampling and evaluation of the flow.<sup>8</sup>

## V. RESULTS

### A. Log-likelihood

We start with a comparison of total NLL between the student IAF and the teacher MAF. They are evaluated on

<sup>8</sup>A computationally more expensive approach would be to save the model state based on the five or ten smallest KL values and evaluate the NLL of these model states after the training. Since the weights will not be updated anymore, this evaluation could be run in parallel on different machines.

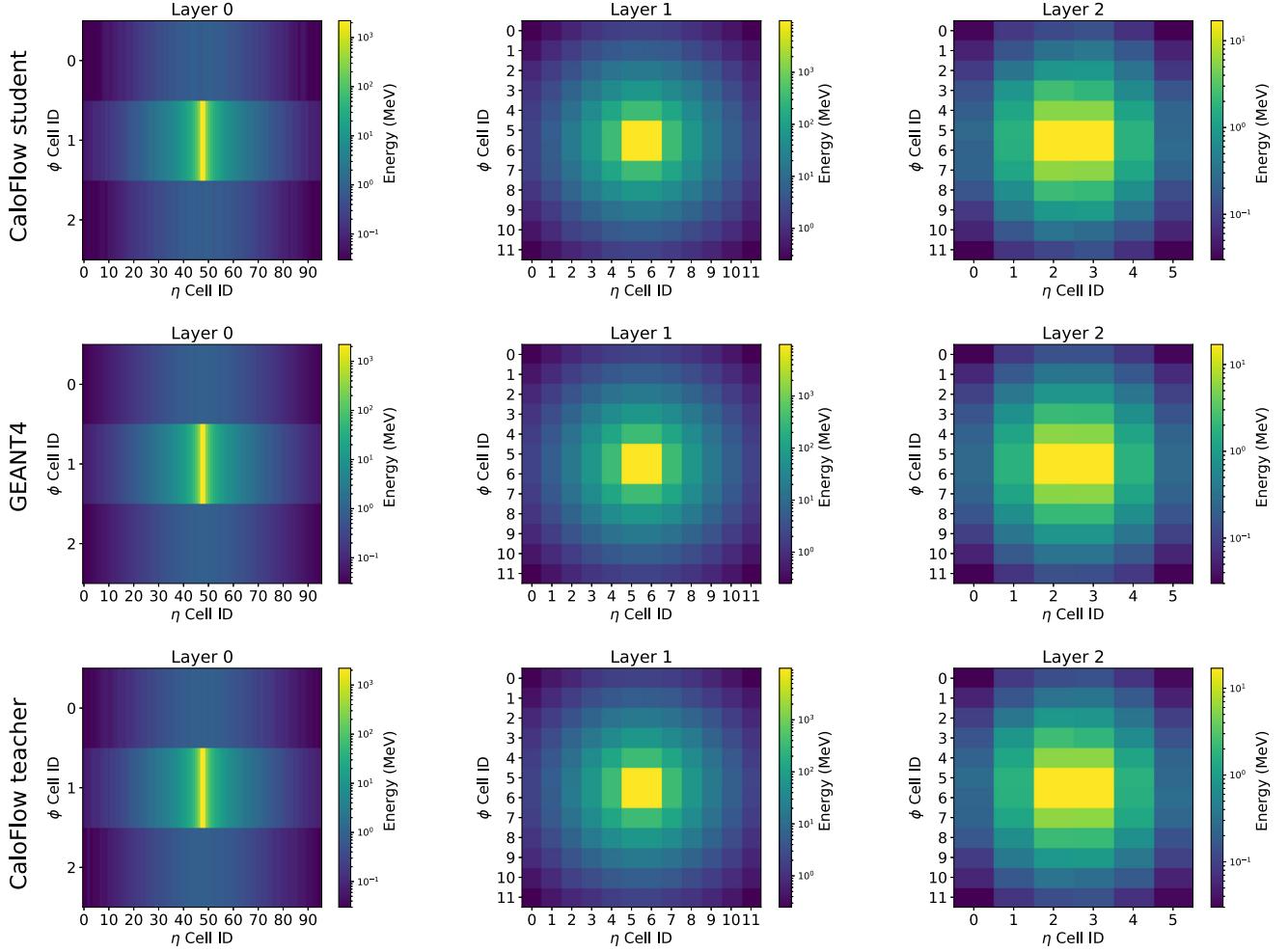


FIG. 4. Average shower shapes for  $\gamma$ . Columns are calorimeter layers 0–2, top row shows CaloFlow student, center row GEANT4, and bottom row CaloFlow teacher.

the same validation set, containing 30,000 events drawn from the GEANT4 simulation. The results are shown in Table II (also compare to Fig. 2) and demonstrate that the student NLL nearly saturates the teacher NLL for all three particle types.

### B. Average images

Next we turn to the same qualitative comparisons of average and individual images that we conducted in [19], following [9,10]. Shown in Figs. 3–5 are the average calorimeter shower images for  $e^+$ ,  $\gamma$ , and  $\pi^+$ , respectively, for GEANT4, CaloFlow teacher, and CaloFlow student. We see excellent agreement between all three; they are nearly indistinguishable by eye. There is no sign of mode collapse.

### C. Flow II histograms

Figures 6–11 show histograms of the same features as in [19] relevant for flow II, for the three different particle

types.<sup>9</sup> These are the two brightest voxels in each layer, the difference of those two divided by their sum (called  $E_{\text{ratio}}$ ), the fraction of voxels with an energy deposition (called sparsity), the centroid in  $\phi$  and  $\eta$  direction, and the standard deviation of the  $\eta$  centroid (called  $\sigma_i$ ); see [19] for more details. Each histogram compares the GEANT4 reference sample with the flow II teacher (taken from [19]) and the new flow II student. We again see in nearly all cases that the teacher and student are basically indistinguishable from one another. The largest differences between student and teacher are visible in the distributions of the brightest and second brightest pixels of layer 0 and layer 1 for  $\pi^+$ . Smaller differences between student and teacher can be seen in  $E_{2,\text{brightest,layer0}}$  and  $E_{\text{ratio},0}$  for  $e^+$  and  $\gamma$ . Curiously, in one histogram ( $E_{2,\text{brightest,layer0}}$  for  $e^+$ ) the student actually matches the GEANT4 reference *better* than the teacher. This

<sup>9</sup>We do not show histograms that are only sensitive to flow I, as flow I of CaloFlow v1 and CaloFlow v2 are identical and the only differences in histograms would be of statistical nature.



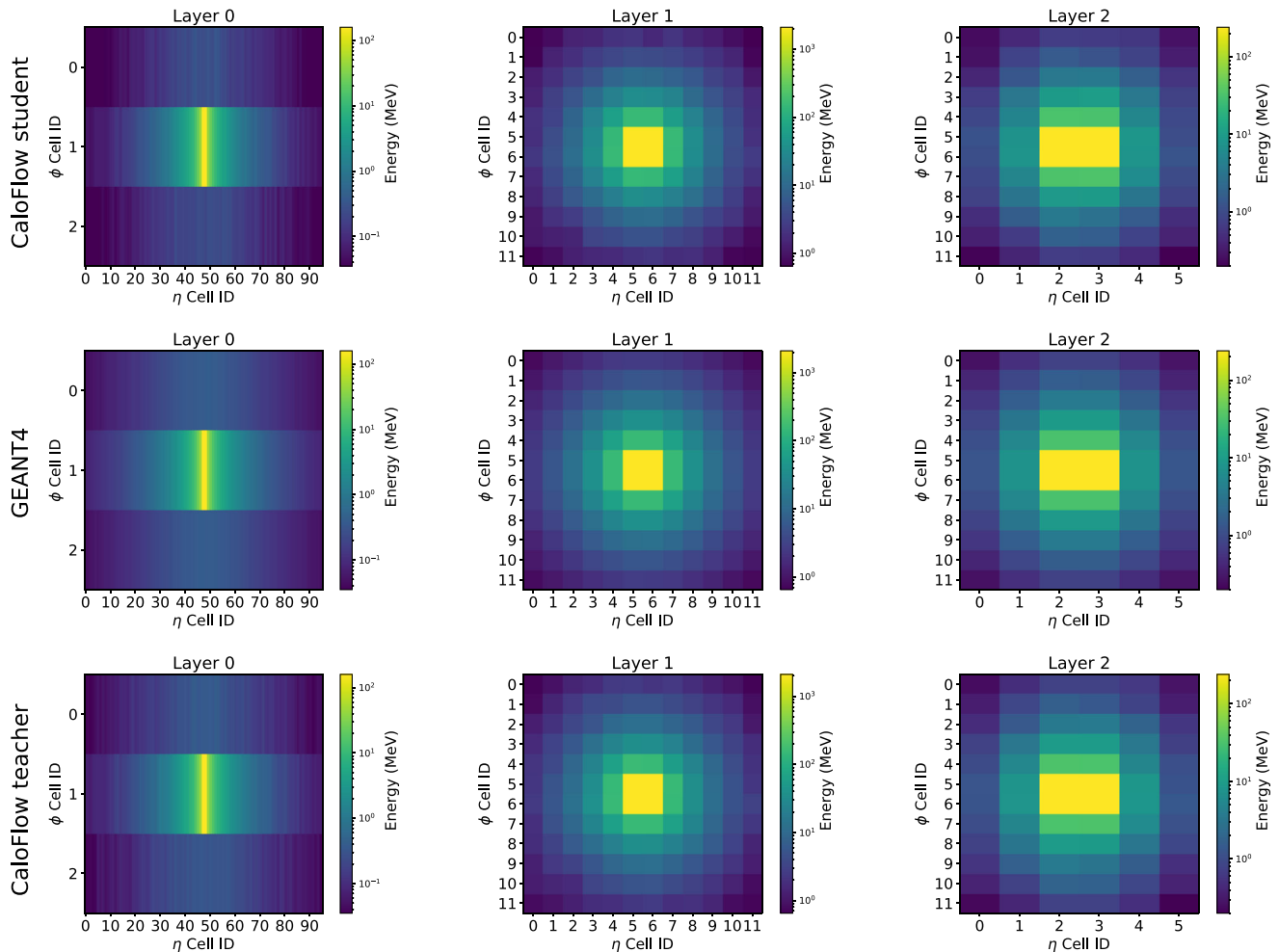


FIG. 5. Average shower shapes for  $\pi^+$ . Columns are calorimeter layers 0–2, top row shows CaloFlow student, center row GEANT4, and bottom row CaloFlow teacher.

is possible if the teacher is off from the data, and the student has not fully converged to the teacher, leading to an accidentally better agreement with the data. Finally, we observe that for  $\pi^+$  showers, there are also some small differences in the energy-weighted shower mean of the student in the top two rows of Fig. 11; these appear to be slightly narrower than their teacher’s equivalent. However, these differences are off peak and subleading.

In addition to these histograms, the Appendix collects nearest neighbor comparisons of samples from CaloFlow v2 and GEANT4. As already seen in [19], we do not observe any sign of mode collapse in these.

#### D. Classifier metrics

Next, we exhibit the result of the classifier metric introduced in [19]. It gauges the quality of the generative model through the score of a binary classifier trained to discriminate between the reference data sample and the generative model sample, approximating the Neyman-Pearson classifier.

Unlike in [19], here we focus only on a simple deep neural network (DNN) classifier trained on either all of the pixels of the calorimeter shower or on a set of high-level features.<sup>10</sup> For simplicity, we do not consider a convolutional neural network classifier (which takes considerably longer to train and was a less sensitive metric than the DNN in [19]), but we do consider the same two preprocessing approaches for the low-level features. These are “unnormalized,” i.e., using the showers as they were generated as input to the classifier, and “normalized,” i.e., using showers that are normalized such that they sum to 1 in each calorimeter layer as input to the classifier. In addition to the energy depositions of each voxel, we give the incident energy and the energy deposition per calorimeter layer to the classifier. The detailed list of high-level features and their preprocessing can be found in [19]. Before the final evaluation, we calibrate the classifiers using

<sup>10</sup>The list of high-level features, DNN architecture, and training procedure is the same as in [19], we train for 150 epochs with a learning rate of  $10^{-3}$ .

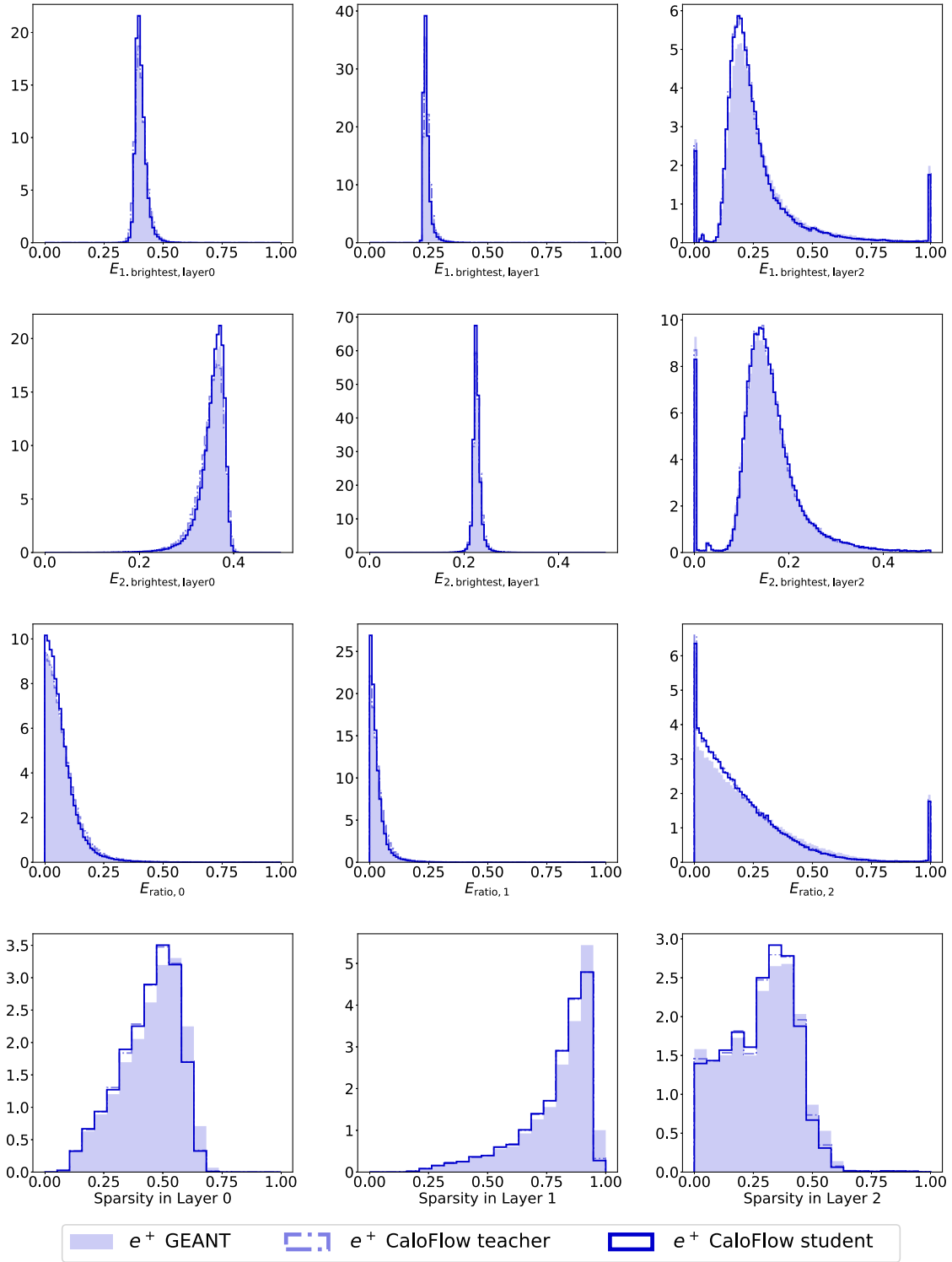


FIG. 6. Distributions that are sensitive to flow II for  $e^+$ . Top row: energy of brightest voxel compared to the layer energy. Second row: energy of second brightest voxel compared to the layer energy. Third row: difference of brightest and second brightest voxel, normalized to their sum. Last row: sparsity of the showers. See [19] for detailed definitions.

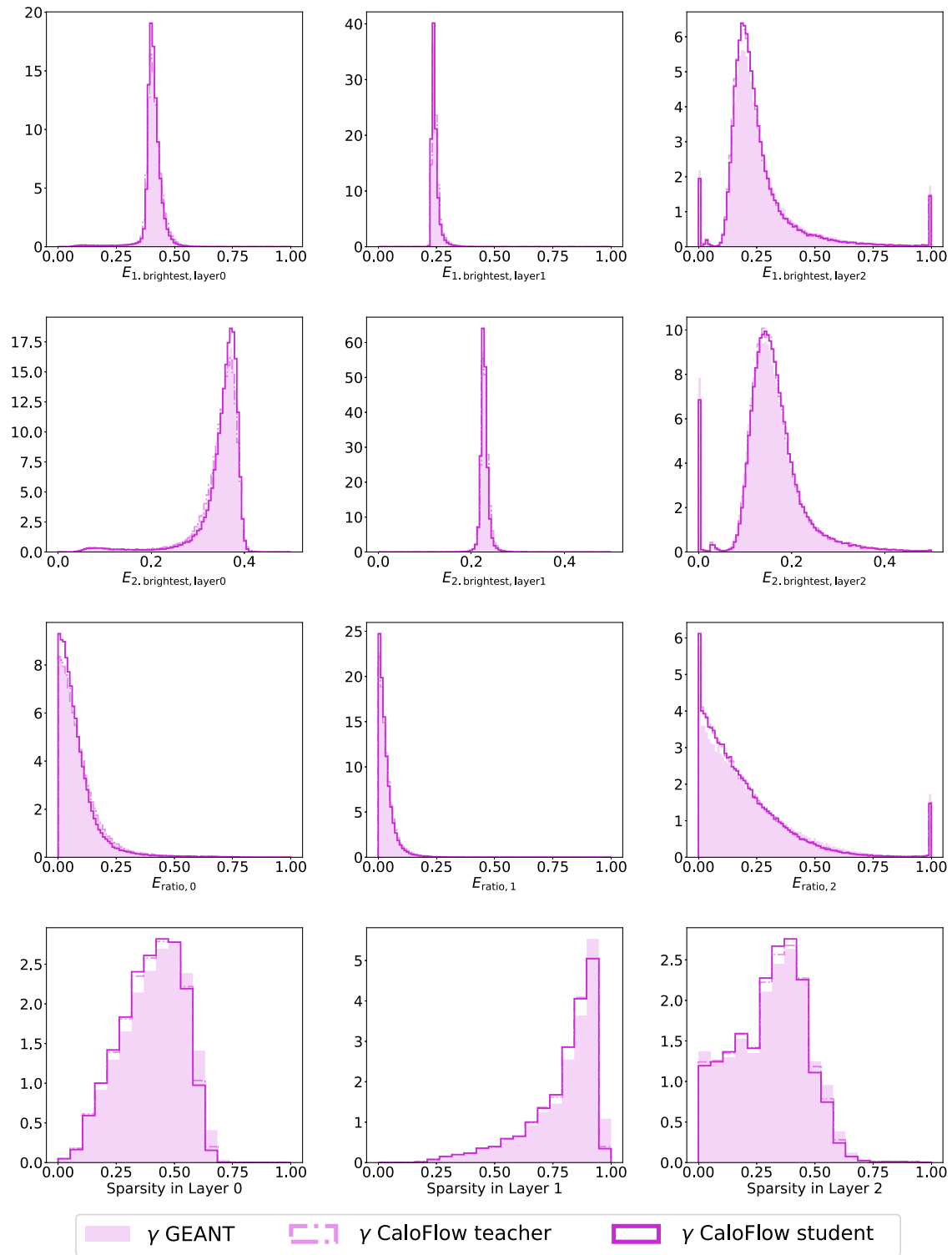


FIG. 7. Distributions that are sensitive to flow II for  $\gamma^+$ . Top row: energy of brightest voxel compared to the layer energy. Second row: energy of second brightest voxel compared to the layer energy. Third row: difference of brightest and second brightest voxel, normalized to their sum. Last row: sparsity of the showers. See [19] for detailed definitions.

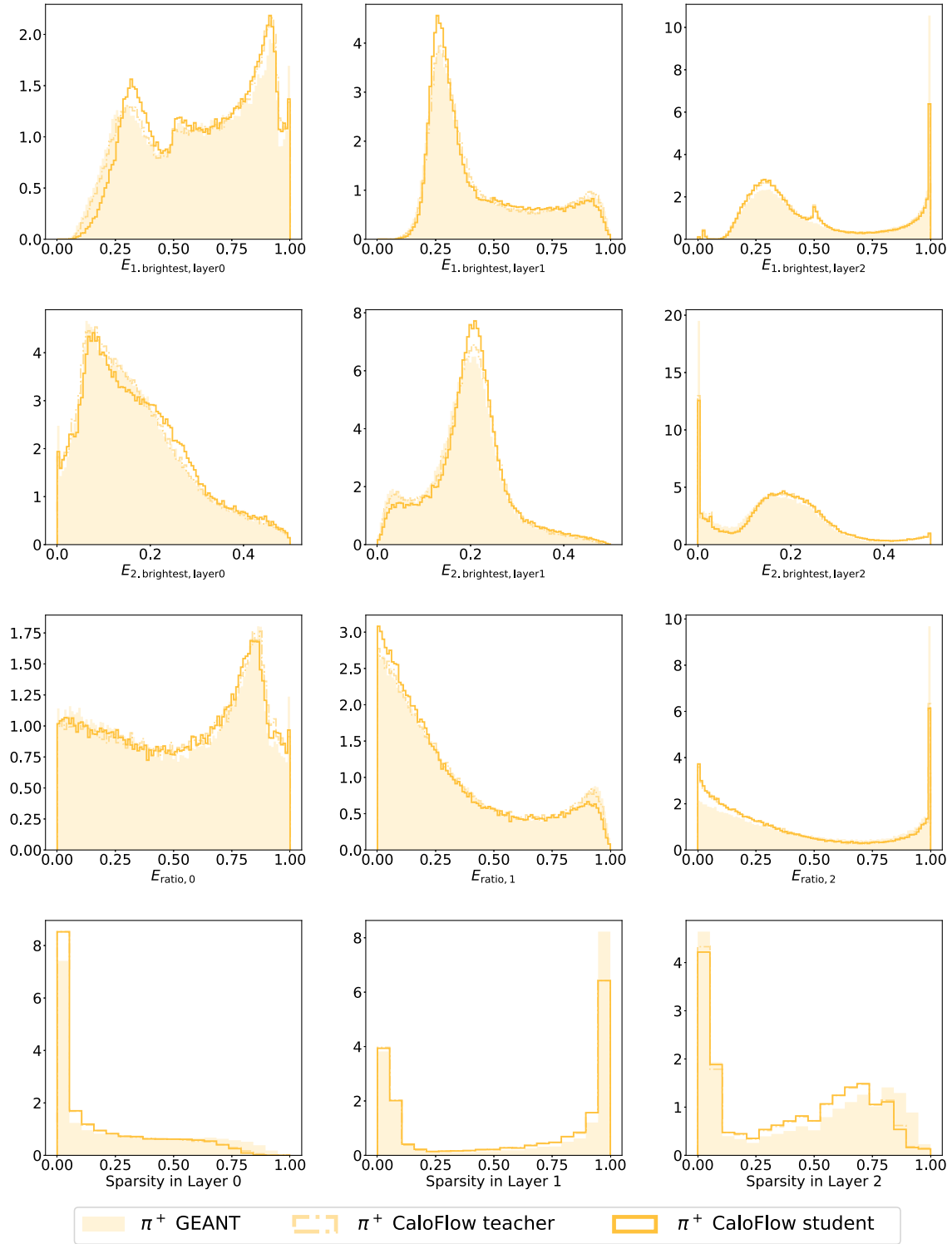


FIG. 8. Distributions that are sensitive to flow II for  $\pi^+$ . Top row: energy of brightest voxel compared to the layer energy. Second row: energy of second brightest voxel compared to the layer energy. Third row: difference of brightest and second brightest voxel, normalized to their sum. Last row: sparsity of the showers. See [19] for detailed definitions.

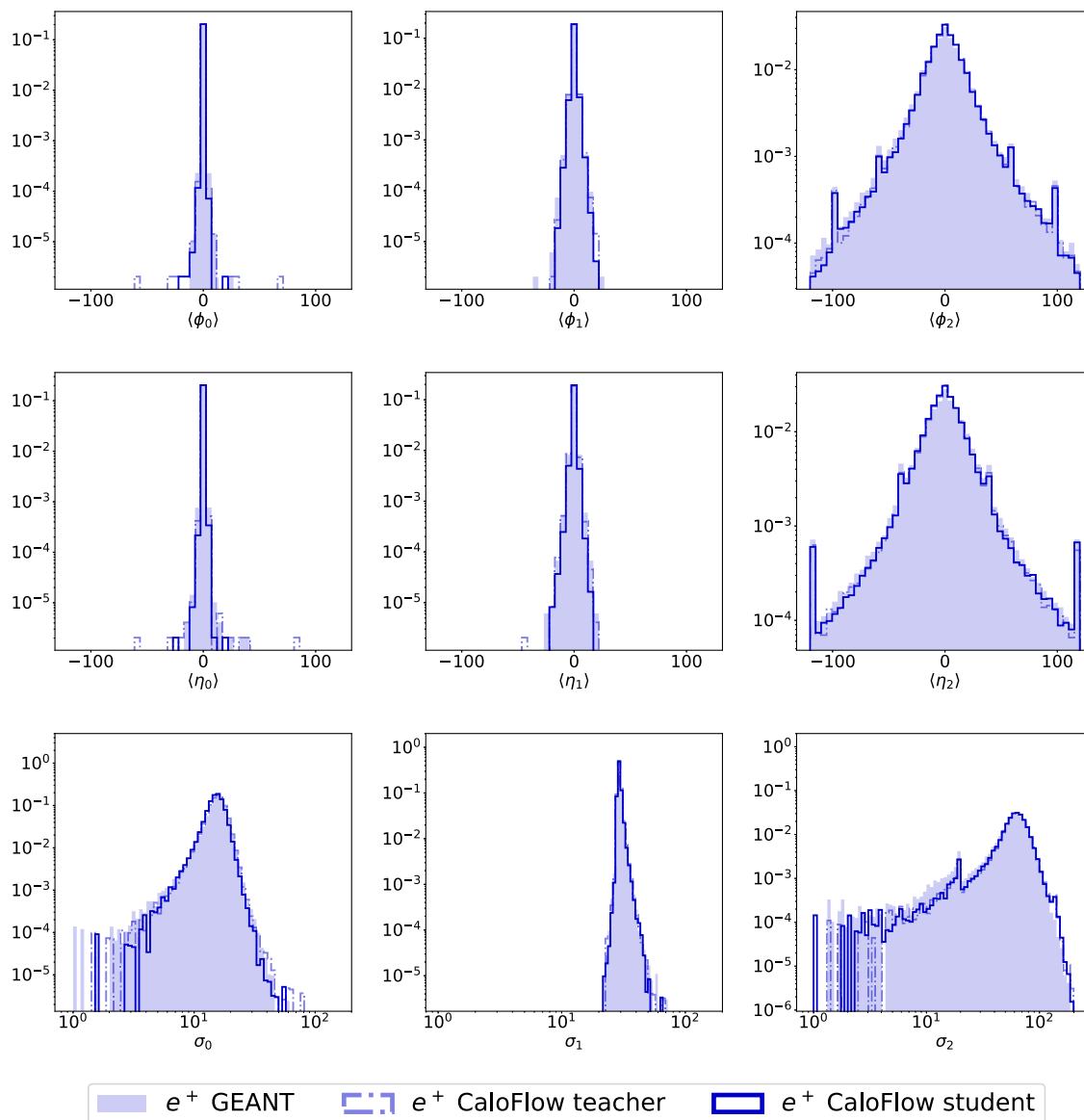


FIG. 9. Further distributions that are sensitive to flow II for  $e^+$ , as learned by flow II. Top and center rows show the location of the deposition centroid in  $\phi$  and  $\eta$  direction; the bottom row shows the standard deviation of the  $\eta$  centroid.

isotonic regression [35] of SKLEARN [36] based on the validation dataset, see [19] for more details.

We see in Table III that, in all cases, the classifier scores of the student are in line with those of the teacher, sometimes slightly worse, and sometimes even slightly better. Most importantly, they are always significantly different from unity, which indicates that they always remain much higher fidelity than the GAN. (Recall, in [19], we showed the DNN trained on GAN vs GEANT4 achieved  $AUC = 1$  for all three particle types.) The fact that the student quality surpasses the teacher's in some cases can be explained by the observation we made in Sec. VC: Some features that are not perfectly modeled by the teacher can get accidentally better in a student that does not exactly follow the teacher.

### E. Timing benchmarks

Finally, we come to the main reason for the student IAF: realizing the factor of  $d \sim 500$  gain in sampling speed compared to the MAF. We summarize training and generation times of CaloFlow v1, CaloFlow v2, CaloGAN, and GEANT4 in Table IV. Timings are evaluated on our TITAN V graphics processing unit (GPU), except for the GEANT4 run-time, which is taken from [10]. The training of the student is understood as being in addition to training the teacher. The difference in generation times for different batch sizes in CaloGAN is due to Keras-TensorFlow constructing a graph at the beginning of the execution, whereas CaloFlow is based on PyTorch [37] and does the batching only with a Python for loop with no additional speedups. We see that with the largest batch sizes, CaloFlow

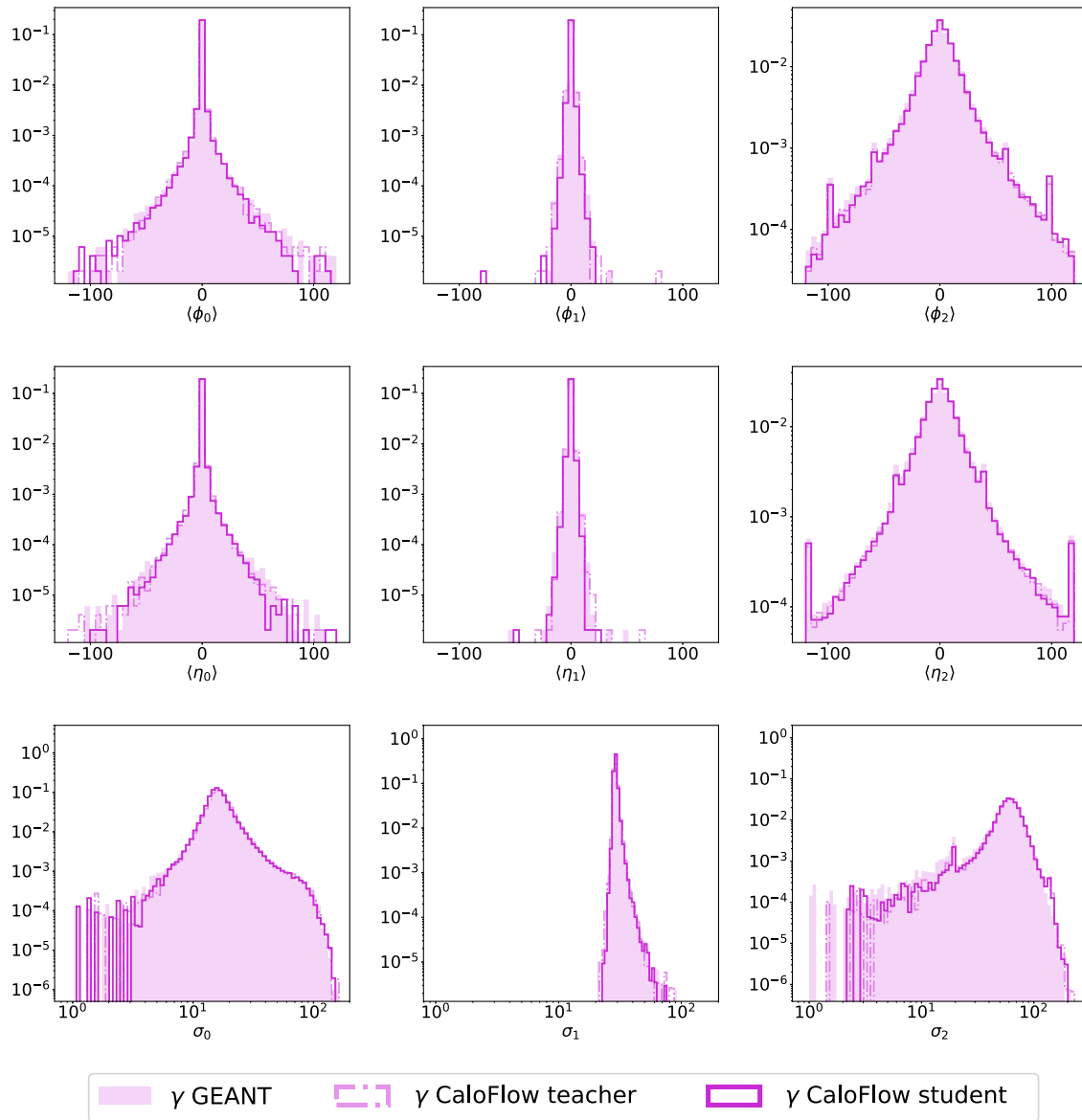


FIG. 10. Further distributions that are sensitive to flow II for  $\gamma$ , as learned by flow II. Top and center rows show the location of the deposition centroid in  $\phi$  and  $\eta$  direction; the bottom row shows the standard deviation of the  $\eta$  centroid.

TABLE III. Receiver operating characteristic (ROC) and Jensen-Shannon divergence (JSD) metrics for the classification of GEANT4 vs CaloFlow student showers (lower numbers are better). Classifiers were trained on each particle type ( $e^+$ ,  $\gamma$ ,  $\pi^+$ ) separately. All entries show mean and standard deviation of 10 classifier retrainings on the same sample and are rounded to three digits. For comparison, we also give the classifier scores of the CaloFlow teacher of [19].

AUC/JSD		GEANT4 vs CaloFlow v2 (student)	GEANT4 vs CaloFlow v1 (teacher) [19]
$e^+$	Unnormalized	0.786(7)/0.201(11)	0.859(10)/0.365(14)
	Normalized	0.824(4)/0.257(8)	0.870(2)/0.378(5)
	High-level	0.762(3)/0.164(5)	0.795(1)/0.229(3)
$\gamma$	Unnormalized	0.758(14)/0.162(18)	0.756(48)/0.174(68)
	Normalized	0.760(3)/0.158(4)	0.796(2)/0.216(4)
	High-level	0.739(2)/0.139(3)	0.727(2)/0.131(3)
$\pi^+$	Unnormalized	0.729(2)/0.144(3)	0.649(3)/0.060(2)
	Normalized	0.807(1)/0.230(3)	0.755(3)/0.153(3)
	High-level	0.893(2)/0.410(5)	0.888(1)/0.401(4)

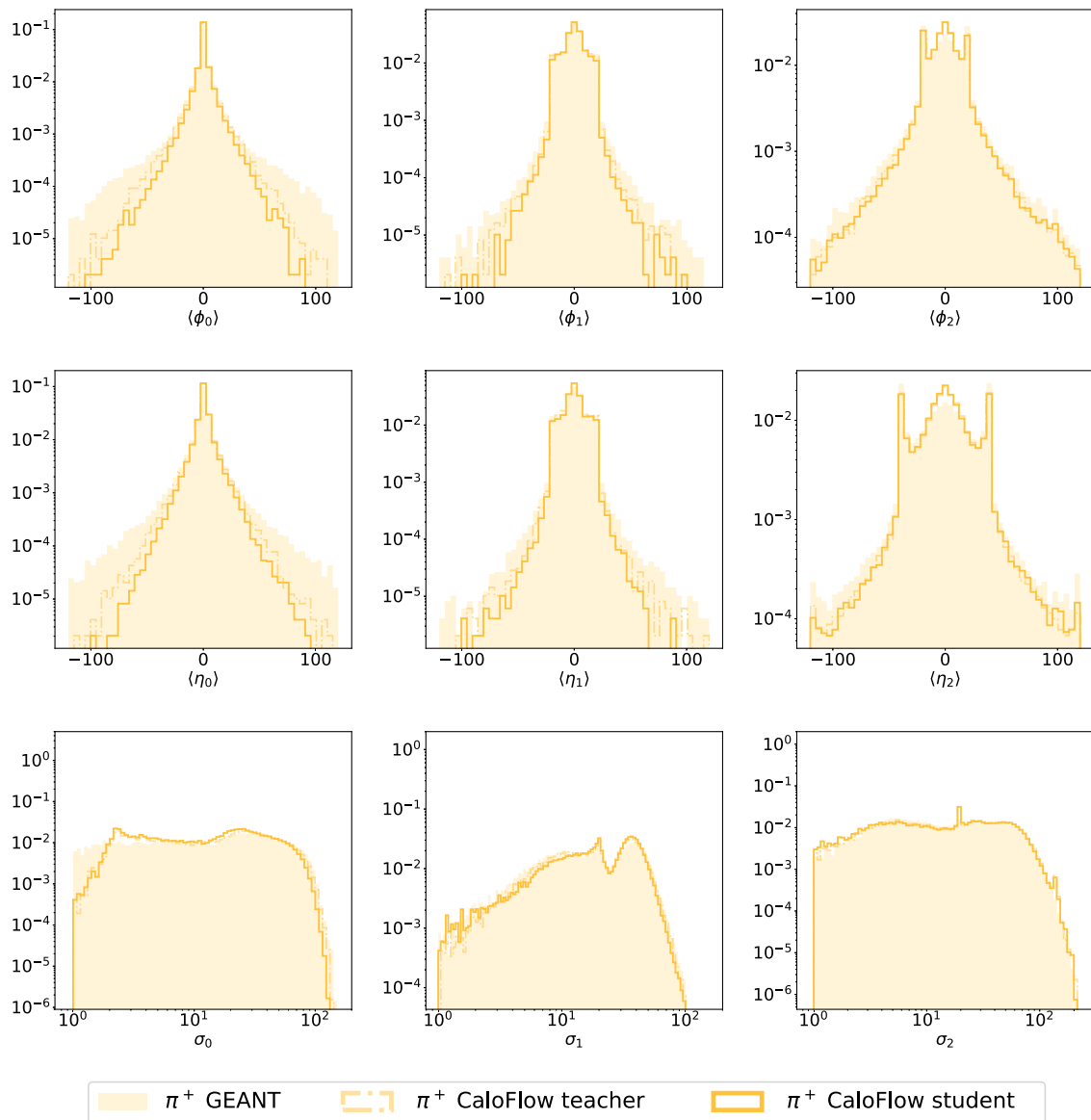


FIG. 11. Further distributions that are sensitive to flow II for  $\pi^+$ , as learned by flow II. Top and center rows show the location of the deposition centroid in  $\phi$  and  $\eta$  direction; the bottom row shows the standard deviation of the  $\eta$  centroid.

TABLE IV. Training and evaluation times of CaloFlow and CaloGAN. These are evaluated on a TITAN V GPU, the GEANT4 run-time is taken from [10].

	CaloFlow		CaloGAN	GEANT4	
	v1 (teacher) [19]	v2 (student)			
Training	22 + 82 min	+480 min	210 min	0 min	
Time per shower					
Generation batch size	(ms)	(ms)	Batch size requested (ms)	100,000 requested (ms)	(ms)
10	835	5.81	455	2.2	1772
100	96.1	0.60	45.5	0.3	1772
1000	41.4	0.12	4.6	0.08	1772
10000	36.2	<b>0.08</b>	0.5	<b>0.07</b>	1772

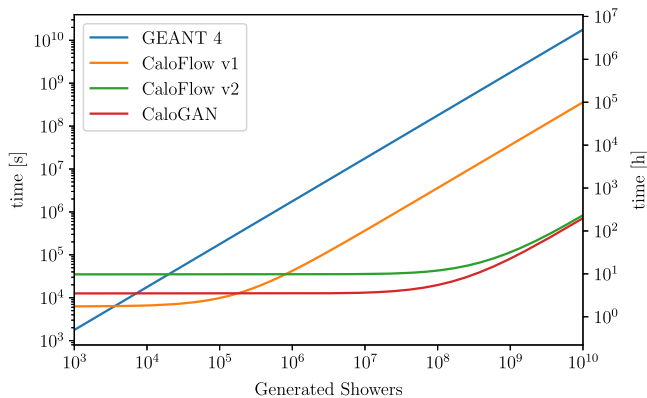


FIG. 12. Comparison of shower generation times, using the fastest CaloGAN numbers for comparison.

v2 fully matches the impressive speed of CaloGAN (0.08 vs 0.07 ms per shower).

In Fig. 12, we show the time needed to generate the samples vs the size of the requested dataset, including the times needed to train the generative models (visible by the plateau at low number of generated showers).<sup>11</sup> Given that many millions (or even billions) of simulated events are required by the LHC Collaborations for their analyses, with each event typically involving hundreds or thousands of showers, this figure demonstrates that the initial computational cost of training the generative models will barely matter when generating samples for actual LHC data analysis. It is clear that fast and accurate GEANT4 emulation is an extremely worthwhile endeavor at the LHC [8].

## VI. CONCLUSIONS

In this work, we have presented CaloFlow v2, a faster-sampling normalizing flow for GEANT4 calorimeter shower emulation that matches the speed of CaloGAN yet retains the superior fidelity of CaloFlow v1 [19]. To achieve this impressive performance, CaloFlow v2 is based on the fast-sampling IAF architecture, whereas CaloFlow v1 was based on the alternative MAF architecture. We overcame fundamental

<sup>11</sup>Note that Fig. 12 and Table IV do not include the time needed to generate the GEANT4 training data for the deep generative models.

obstacles in training IAFs for high-dimensional datasets using the novel technique of probability density distillation to fit the student IAF to the teacher MAF instead of directly to the GEANT4 data. We also improved and innovated beyond the existing ML literature for probability density distillation, inventing several new loss terms that greatly improve the matching of the IAF to the MAF. We expect there could be many applications of this fully guided teacher-student training to other domains in fundamental physics and beyond.

Through [19] and the present work, we have demonstrated that normalizing flows are an extremely promising method for fast and accurate generative modeling of high-dimensional datasets. With regards to calorimeter emulation, many interesting future directions remain, including generalizing this work to even higher-dimensional calorimeters (e.g., International Linear Detector [16,17] and CMS high-granularity calorimeter [13,38]), generalizing beyond perpendicular and central incident particles [8,11–15,18], and including simulations of both electromagnetic calorimeter and hadronic calorimeter showers.

In this work, we used the NumPy 1.16.4 [39], Matplotlib 3.1.0 [40], PANDAS 0.24.2 [41], SKLEARN 0.21.2 [36], H5PY 2.9.0 [42], PyTorch 1.7.1 [37], and nFlows 0.14 [43] software packages.

Our code is available at [44].

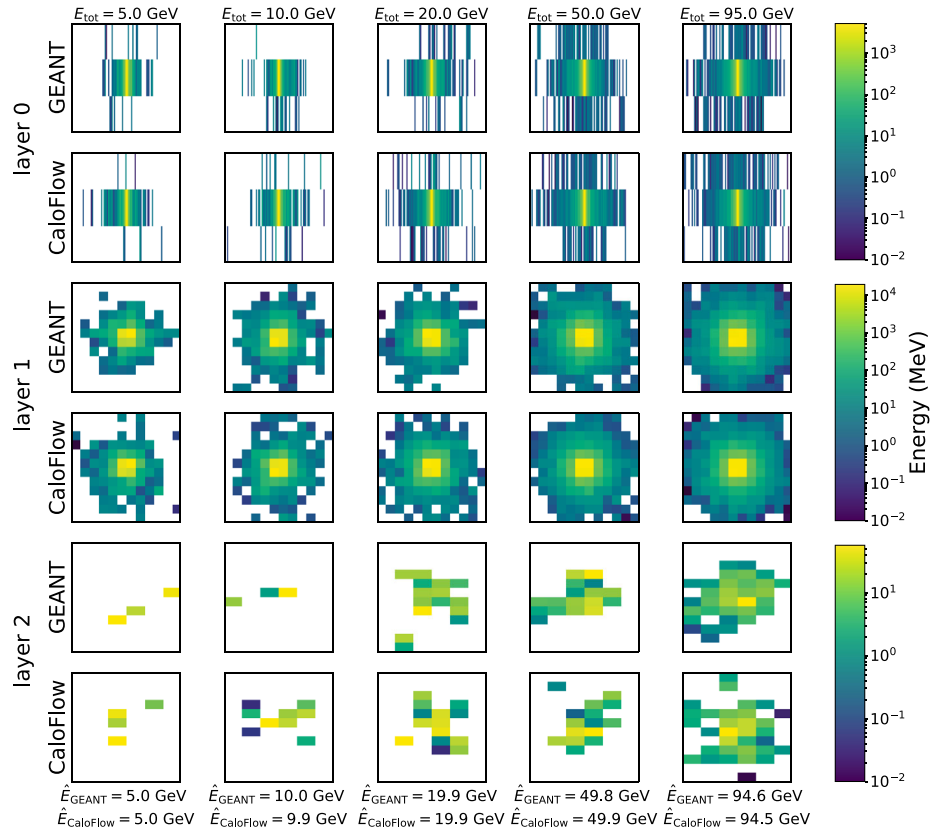
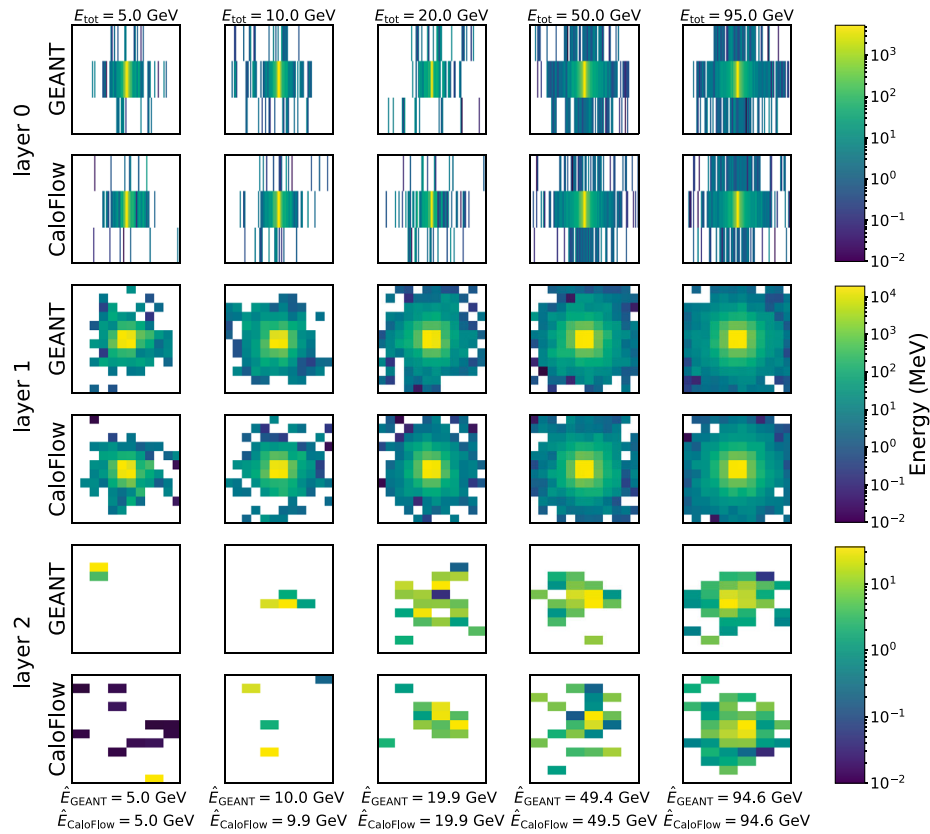
## ACKNOWLEDGMENTS

We are grateful to Ben Nachman for helpful discussions and comments on the draft. This work was supported by DOE Award No. DOE-SC0010008.

## APPENDIX: NEAREST NEIGHBORS

Shown in Figs. 13–15 are five randomly selected events from the GEANT4 datasets for  $e^+$ ,  $\gamma$ , and  $\pi^+$  at incident energies  $E_{\text{inc}} = 5, 10, 20, 50,$  and  $95$  GeV and the Euclidean nearest neighbors in the CaloFlow student samples. We use the exact same setup as previously in [19], with 2000 CaloFlow samples at each of the incident energies and the exact same GEANT4 references. Again, we observe nearest neighbors that are close to the GEANT4 samples at all energies, suggesting that no mode collapse occurred.



FIG. 13. Five randomly selected  $e^+$  events of GEANT4 and their nearest neighbors in the CaloFlow student samples.FIG. 14. Five randomly selected  $\gamma$  events of GEANT4 and their nearest neighbors in the CaloFlow student samples.

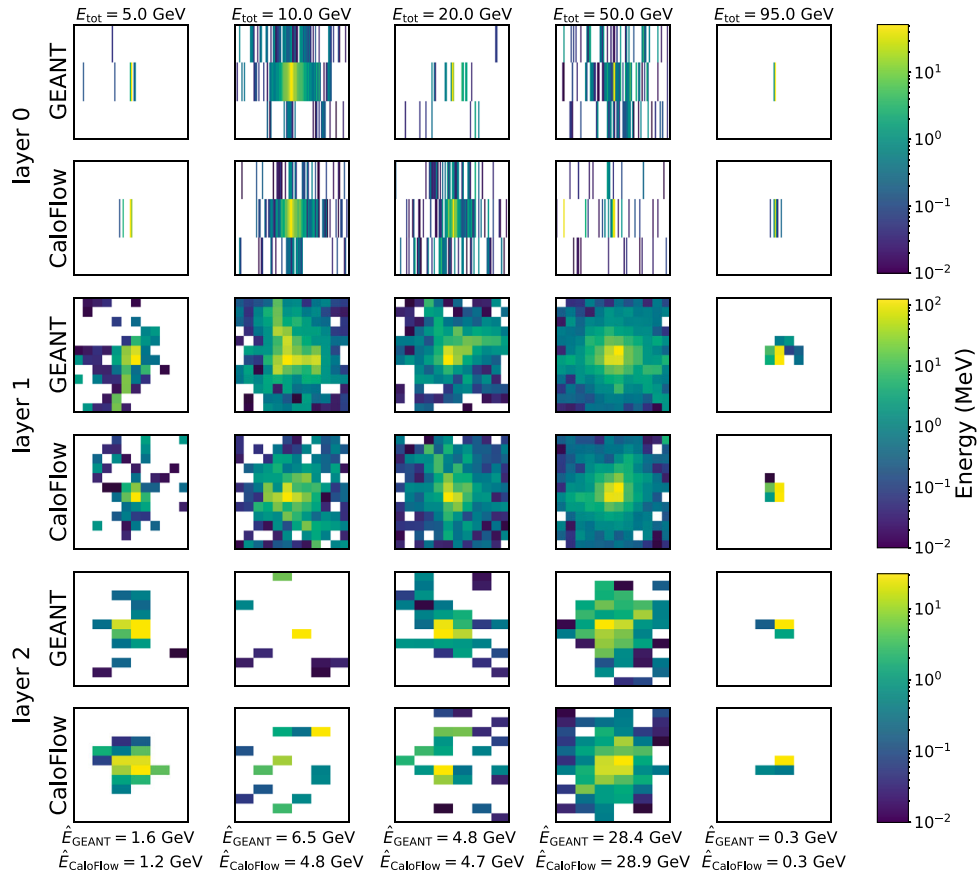


FIG. 15. Five randomly selected  $\pi^+$  events of GEANT4 and their nearest neighbors in the CaloFlow student samples.

- [1] GEANT4 Collaboration, GEANT4—A simulation toolkit, *Nucl. Instrum. Methods Phys. Res., Sect. A* **506**, 250 (2003).
- [2] J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce Dubois, M. Asai *et al.*, GEANT4 developments and applications, *IEEE Trans. Nucl. Sci.* **53**, 270 (2006).
- [3] J. Allison, K. Amako, J. Apostolakis, P. Arce, M. Asai, T. Aso *et al.*, Recent developments in GEANT4, *Nucl. Instrum. Methods Phys. Res., Sect. A* **835**, 186 (2016).
- [4] HEP Software Foundation, HEP Software Foundation Community White Paper Working Group—Detector simulation, [arXiv:1803.04165](https://arxiv.org/abs/1803.04165).
- [5] HEP Software Foundation, HL-LHC computing review: Common tools and community software, in *2022 Snowmass Summer Study*, edited by P. Canal *et al.* (2020), [10.5281/zenodo.4009114](https://zenodo.org/record/4009114).
- [6] P. Calafiura, J. Catmore, D. Costanzo, and A. Di Girolamo, ATLAS HL-LHC computing conceptual design report, Technical Report No. CERN-LHCC-2020-015, CERN, Geneva, 2020.
- [7] CMS Collaboration, CMS offline and computing public results, Technical Report Approved HL-LHC resource projections, CERN, Geneva, 2020, <https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults>.
- [8] ATLAS Collaboration, AtlFast3: The next generation of fast simulation in ATLAS, *Comput. Softw. Big Sci.* **6**, 7 (2022).
- [9] 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).
- [10] 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).
- [11] 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).
- [12] M. Erdmann, L. Geiger, J. Glombitza, and D. Schmidt, Generating and refining particle detector simulations using

- the Wasserstein distance in adversarial networks, *Comput. Software Big Sci.* **2**, 4 (2018).
- [13] M. Erdmann, J. Glombitza, and T. Quast, Precise simulation of electromagnetic calorimeter showers using a Wasserstein generative adversarial network, *Comput. Software Big Sci.* **3**, 4 (2019).
- [14] ATLAS Collaboration, Deep generative models for fast shower simulation in ATLAS, Technical Report No. ATL-SOFT-PUB-2018-001, CERN, Geneva, 2018.
- [15] D. Belayneh *et al.*, Calorimetry with deep learning: Particle simulation and reconstruction for collider physics, *Eur. Phys. J. C* **80**, 688 (2020).
- [16] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol *et al.*, Getting high: High fidelity simulation of high granularity calorimeters with high speed, *Comput. Software Big Sci.* **5**, 13 (2021).
- [17] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol *et al.*, Decoding photons: Physics in the latent space of a BIB-AE generative network, *EPJ Web Conf.* **251**, 03003 (2021).
- [18] ATLAS Collaboration, Fast simulation of the ATLAS calorimeter system with generative adversarial networks, Technical Report No. ATL-SOFT-PUB-2020-006, CERN, Geneva, 2020.
- [19] C. Krause and D. Shih, following paper, Fast and accurate simulations of calorimeter showers with normalizing flows, *Phys. Rev. D* **107**, 113003 (2023).
- [20] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, Normalizing flows: An introduction and review of current methods, *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3964 (2021).
- [21] G. Papamakarios, E. Nalisnick, D. Jimenez Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, *J. Mach. Learn. Res.* **22**, 1 (2021), <https://jmlr.org/papers/v22/19-1028.html>.
- [22] S. Diefenbacher, E. Eren, G. Kasieczka, A. Korol, B. Nachman, and D. Shih, DCTRGAN: Improving the precision of generative models with reweighting, *J. Instrum.* **15**, P11004 (2020).
- [23] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, [arXiv:1705.07057](https://arxiv.org/abs/1705.07057).
- [24] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Improving variational inference with inverse autoregressive flow, [arXiv:1606.04934](https://arxiv.org/abs/1606.04934).
- [25] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu *et al.*, Parallel WaveNet: Fast high-fidelity speech synthesis, [arXiv:1711.10433](https://arxiv.org/abs/1711.10433).
- [26] C.-W. Huang, F. Ahmed, K. Kumar, A. Lacoste, and A. Courville, Probability distillation: A caveat and alternatives, in *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, edited by R. P. Adams and V. Gogate, Proceedings of Machine Learning Research Vol. 115 (2020), pp. 1212–1221, <http://proceedings.mlr.press/v115/huang20c.html>.
- [27] D. Baranchuk, V. Aliev, and A. Babenko, Distilling the knowledge from conditional normalizing flows, [arXiv:2106.12699](https://arxiv.org/abs/2106.12699).
- [28] R. Yamamoto, E. Song, and J.-M. Kim, Probability density distillation with generative adversarial networks for high-quality parallel waveform generation, in *Proceedings of the Interspeech 2019* (2019), pp. 699–703, [10.21437/Interspeech.2019-1965](https://doi.org/10.21437/Interspeech.2019-1965).
- [29] W. Ping, K. Peng, and J. Chen, ClariNet: Parallel wave generation in end-to-end text-to-speech, [arXiv:1807.07281](https://arxiv.org/abs/1807.07281).
- [30] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32, <https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf>, [arXiv:1906.04032](https://arxiv.org/abs/1906.04032).
- [31] J. A. Gregory and R. Delbourgo, Piecewise rational quadratic interpolation to monotonic data, *IMA J. Numer. Anal.* **2**, 123 (1982).
- [32] M. Germain, K. Gregor, I. Murray, and H. Larochelle, MADE: Masked autoencoder for distribution estimation, *Proc. Mach. Learn. Res.* **37**, 881 (2015), <https://proceedings.mlr.press/v37/germain15.html>.
- [33] C. Krause and D. Shih, Electromagnetic calorimeter shower images of CaloFlow, Zenodo (2021), [10.5281/zenodo.5904188](https://doi.org/10.5281/zenodo.5904188).
- [34] D. P. Kingma and J. Ba, ADAM: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [35] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, On calibration of modern neural networks, [arXiv:1706.04599](https://arxiv.org/abs/1706.04599).
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al.*, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011), <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>.
- [37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), pp. 8024–8035, <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [38] CMS Collaboration, The CMS HGCal detector for HL-LHC upgrade, in *5th Large Hadron Collider Physics Conference* (2017), p. 8, [arXiv:1708.08234](https://arxiv.org/abs/1708.08234).
- [39] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau *et al.*, Array programming with NumPy, *Nature (London)* **585**, 357 (2020).
- [40] J. D. Hunter, Matplotlib: A 2d graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [41] The Pandas Development Team, pandas-dev/pandas: PANDAS (2020), [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134).
- [42] A. Collette, *Python and HDF5* (O'Reilly, North Sebastopol, CA, 2013).
- [43] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, nFlows: Normalizing flows in PyTorch (2020), [10.5281/zenodo.4296287](https://doi.org/10.5281/zenodo.4296287).
- [44] C. Krause and D. Shih, CaloFlow, <https://gitlab.com/claudius-krause/caloflow>.