# Anomaly detection with flow-based fast calorimeter simulators

Claudius Krause,[1,2,*] Benjamin Nachman,[3,4,†] Ian Pang,[5,‡] David Shih,[5,§] and Yunhao Zhu[5,6,‖]

[1]*Institut für Theoretische Physik, Universität Heidelberg,
Philosophenweg 12, 69120 Heidelberg, Germany*
[2]*Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences (OeAW),
Dominikanerbastei 16, A-1010 Vienna, Austria*
[3]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[4]*Berkeley Institute for Data Science, University of California, Berkeley, California 94720, USA*
[5]*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*
[6]*Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA*

(Received 30 January 2024; accepted 19 July 2024; published 29 August 2024)

Recently, several normalizing flow-based deep generative models have been proposed to accelerate the simulation of calorimeter showers. Using CaloFlow as an example, we show that these models can simultaneously perform unsupervised anomaly detection with no additional training cost. As a demonstration, we consider electromagnetic showers initiated by one (background) or multiple (signal) photons. The CaloFlow model is designed to generate single-photon showers, but it also provides access to the shower likelihood. We use this likelihood as an anomaly score and study the showers tagged as being unlikely. As expected, the tagger struggles when the signal photons are nearly collinear but is otherwise effective. This approach is complementary to a supervised classifier trained on only specific signal models using the same low-level calorimeter inputs. While the supervised classifier is also highly effective at unseen signal models, the unsupervised method is more sensitive in certain regions, and, thus, we expect that the ultimate performance will require a combination of these approaches.

DOI: 10.1103/PhysRevD.110.035036

## I. INTRODUCTION

In 2012, the final piece of the Standard Model, the Higgs boson, was discovered at the Large Hadron Collider (LHC) by the ATLAS [1] and CMS [2] Collaborations. Despite this milestone, compelling theoretical and experimental reasons continue to drive the search for physics beyond the Standard Model (BSM). Regrettably, the extensive search programs conducted by ATLAS [3–5], CMS [6–8], and LHCb [9] at the LHC have yet to yield conclusive evidence for new BSM physics. Given the impossibility of conducting dedicated searches for every conceivable BSM scenario, most LHC searches target specific signals derived from theoretical priors, leaving substantial portions of the LHC phase space unexplored. This limitation has spurred the development of more model-agnostic search strategies, with the hope of detecting BSM physics at the LHC. The advent of deep learning has given rise to various model-agnostic anomaly detection methods designed to explore uncharted territories of the LHC phase space—for reviews and original references, see, e.g., [10–14].

Normalizing flows [15–20] represent a class of deep learning models particularly valuable for generative modeling and density estimation tasks. A normalizing flow is characterized by a parametric diffeomorphism $f_\theta$ mapping between a latent space, with a known distribution $\pi(z)$, and a data space of interest with an analytically unknown distribution $p(x)$. In the context of a conditional normalizing flow, this transformation becomes $f_\theta(x|c)$, where $c$ denotes the conditional inputs to the flow. It is defined through a series of invertible functions, parametrized by $\theta$, that can be trained by maximizing the log-likelihood of the data following the change of variables formula:

$$\log(p(x|c)) = \log(\pi(f_\theta(x|c))) + \log\left|\det\left(\mathcal{J}(f_\theta(x|c))\right)\right|,$$

where $\mathcal{J}(f_\theta(x|c))$ represents the Jacobian of the transformation $f_\theta(x|c)$. The allowable transformations must also have a computationally tractable Jacobian, ideally efficient to compute, and the probability density of the base

[*] Contact author: Claudius.Krause@oeaw.ac.at
[†] Contact author: bpnachman@lbl.gov
[‡] Contact author: ian.pang@physics.rutgers.edu
[§] Contact author: shih@physics.rutgers.edu
[‖] Contact author: zhu.yunha@northeastern.edu

distribution $\pi(z)$ must be known. A common choice for $\pi(z)$ is the standard normal distribution.

Recently, normalizing flows have found successful applications in fast calorimeter simulation tasks [21–27]. Moreover, normalizing flows have demonstrated comparably excellent performance across various tasks within high-energy physics [28–53]. In this paper, we demonstrate the utility of these flow-based generative models as unsupervised anomaly detectors for identifying BSM physics in calorimeter shower data. Specifically, we apply CaloFlow [21], a flow-based fast calorimeter simulation model, to single-photon showers from a new sampling calorimeter version [54] of the CaloGAN dataset [55–57]. In this context, single-photon showers serve as the background events, while the signal events consist of photon showers originating from a generic BSM particle $\chi$ that undergoes the decay $\chi \to \gamma\gamma$. The $\chi$ particle is taken to be invisible and interacts only indirectly with the calorimeter through its decay products. By training CaloFlow to maximize the log-likelihood when evaluated on background events, we are able to detect the signal events based on a cut on the log-likelihood. We focus on achieving signal sensitivity, but the approach could be combined with a number of background estimation strategies [11,28].

While we believe this is the first unification of simulation and anomaly detection, both subjects have been well studied with machine learning. Many deep generative models have been studied for calorimeter simulation [21–25,27,56–81], including a number of proposals developed on the CaloChallenge datasets [82], and they are also now being integrated into experimental work flows [71]. We focus on normalizing flows, since they give direct access to the likelihood. This information can also be extracted from diffusion models [83], and it would be interesting to compare approaches in the future. For anomaly detection, unsupervised methods have been extensively studied (e.g., Refs. [84–86], and many others) and also include density-based approaches [12]. Like the density-based methods, we use the likelihood directly as the anomaly score.[1]

There may be other ways of reusing the generative model for BSM searches, including fine-tuning supervised models based on particular signal hypotheses.

This paper is organized as follows. In Sec. II, we describe the calorimeter setup and the datasets that were used during training and evaluation. In Sec. III, we explain how CaloFlow is used as an unsupervised anomaly detector by placing cuts on the log-likelihood of background and signal showers evaluated using CaloFlow. In Sec. IV, we include the results of performing unsupervised anomaly detection with CaloFlow. Finally, we conclude in Sec. V.

## II. DATASET

For this study, we decided to create a new, more realistic sampling calorimeter version [54] of the CaloGAN dataset. The original dataset included energy contributions from both active and inactive calorimeter layers, whereas our new dataset includes only energy contributions from the active layers as would be available in practice. The sampling fraction of our new calorimeter setup is ~20%. The simple detector is a three-layer, liquid argon (LAr) sampling calorimeter cube with 480 mm side length that is inspired by the ATLAS LAr electromagnetic calorimeter [90]. It is simulated as follows: Geant4 [91–93] is used to generate particles and simulate their interaction with our calorimeter using the Ftfp_Bert physics list based on the FRITIOF [94–97] and Bertini intranuclear cascade [98–100] models with the standard electromagnetic physics package [101]. While we use this new simulator to create a dataset for anomaly detection, we expect it should be more generally useful for a broad variety of tasks.

The calorimeter showers are represented as three-dimensional images that are binned in position space. In this representation, the calorimeter shower geometry is made up of voxels (volumetric pixels), and the details of the calorimeter voxel dimensions are included in Table I. Figure 1 (taken from [56,57]) shows the three-dimensional representation of a shower in the CaloGAN calorimeter. The three longitudinal layers are separated in the figure for visualization purposes. In this work, the center of the detector is taken to be at $z = 0$ m, while the front face of layer 0 is positioned at $z = 1$ m, which is consistent with an ATLAS-like configuration.

For the background dataset (single-photon calorimeter showers), we generate 100 000 showers with incident energies $E_{\mathrm{inc}}$ uniformly distributed in the range [1, 100] GeV. This dataset was used as the training dataset for CaloFlow where we used a train/validation split of 70%/30%. A second independent dataset of 100 000 showers with the same range of $E_{\mathrm{inc}}$ was generated and used for evaluation.

To obtain the signal datasets, we defined the hypothetical $\chi$ particle in Geant4, which has the same properties as the $\eta^0$ particle except having a different mass and being invisible

TABLE I. Dimensions of a calorimeter voxel. The positive $z$ axis (radial direction in full detector) is the direction of particle propagation, the $\eta$ direction is along the proton beam axis, and $\phi$ is perpendicular to $z$ and $\eta$. For the number of voxels, the first (second) number is the number of bins in the $\phi$ ($\eta$) direction (e.g., $12 \times 6$ refers to 12 $\phi$ bins and six $\eta$ bins).

| Layer index | $z$ length (mm) | $\eta$ length (mm) | $\phi$ length (mm) | Number of voxels |
|---|---|---|---|---|
| 0 | 90 | 5 | 160 | $3 \times 96$ |
| 1 | 347 | 40 | 40 | $12 \times 12$ |
| 2 | 43 | 80 | 40 | $12 \times 6$ |

[1] Note that this is not unique and is sensitive to how the data are represented or preprocessed [87–89].

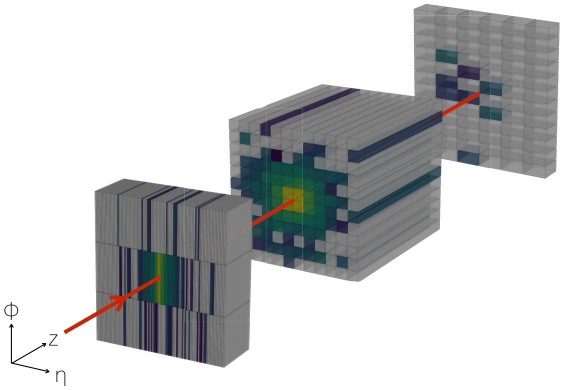

FIG. 1. Three-dimensional representation of the shower in the CaloGAN calorimeter; figure taken from [56,57]. Not-to-scale separation among the longitudinal layers is added for visualization purposes.

to the detector. Next, we generate multiple sets of 100 000 showers that originate from $\chi \to \gamma\gamma$ decay at ten chosen fixed displacements from the center of detector along the $z$ axis. The chosen fixed displacements are $z_i \in \{0.33, 0.66, 1.00, 1.04, 1.08, 1.16, 1.24, 1.32, 1.40, 1.44\}$ m. Note that the first two displacements are located in front of the calorimeter, while the last eight displacements are located at distinct positions within the calorimeter. The energy of the $\chi$ particle was fixed at 50 GeV. Such a hypothetical scenario might arise from the decay of a 100 GeV particle (close to, e.g., the Higgs boson mass), which is at rest, to a pair of $\chi$ particles.

To study how the mass $m_\chi$ affects the anomaly detection performance, we generated signal datasets with different $m_\chi \in \{5 \times 10^{-3}, 5 \times 10^{-2}, 5 \times 10^{-1}, 5\}$ GeV. For each choice of $m_\chi$, we generate 100 000 showers at each of the ten fixed displacements. The results for these fixed displacement signal datasets are shown in Sec. IV A.

To consider particles with fixed lifetimes, we construct new signal datasets using the fixed displacement samples. In particular, we use the probability that the $\chi$ particle decays at various positions along the $z$ axis to determine the proportion of showers originating from decays at each of the ten fixed displacements. Generating directly with fixed lifetimes would have been too inefficient, since we are discarding decays after the calorimeter volume, which can happen often for these lifetimes that we consider. More details of the fixed lifetime signal datasets are included in Sec. IV B.

We also generated 100 000 signal showers with kinetic energy[2] uniformly distributed in the range [1, 100] GeV for a $\chi$ particle with $m_\chi = 5 \times 10^{-3}$ GeV and lifetime $\tau = 1.00$ ns. A second set of 100 000 signal showers was generated with the same range of kinetic energies but for a $\chi$ particle with $m_\chi = 5$ GeV and lifetime $\tau = 1.00$ ns. Each of these two datasets (together with the background dataset) was used to train a supervised classifier described in Sec. IV C. Using a range of kinetic energies ensured that we obtain showers with a range of $E_{\text{inc}}$ such that the $E_{\text{inc}}$ would not be used as a discriminating factor by the supervised classifier. This allows for a fairer comparison between performance of the supervised classifier and our unsupervised method, since CaloFlow learns the likelihood conditioned on $E_{\text{inc}}$.

TABLE II. The conditional inputs for each flow and the features whose probability distributions are the output of each flow.

| | Conditionals | Dimension of conditional | Output | Dimension of output |
|---|---|---|---|---|
| Flow I | $E_{\text{inc}}$ | 1 | $E_0$, $E_1$, $E_2$ | 3 |
| Flow II | $E_0, E_1, E_2, E_{\text{inc}}$ | 4 | $\hat{\mathcal{I}}$ | 504 |

## III. CaloFlow

CaloFlow [21,22] is an approach to fast calorimeter simulation based on conditional normalizing flows. In the context of fast calorimeter simulation, CaloFlow uses a two-flow method to learn to generate the voxel-level shower energies $\vec{\mathcal{I}}$ conditioned on the corresponding incident energies of the showers $E_{\text{inc}}$ denoted by $p(\vec{\mathcal{I}}|E_{\text{inc}})$. Flow I is constructed to learn the probability density of calorimeter layer energies[3] $E_i$ conditioned on the incident energy $p_1(E_0, E_1, E_2|E_{\text{inc}})$, while flow II is designed to learn the probability density of the normalized voxel-level shower energies conditioned on the calorimeter layer energies and incident energies $p_2(\hat{\mathcal{I}}|E_0, E_1, E_2, E_{\text{inc}})$. By normalized, we mean that the voxel energies in each layer are made to sum to unity. The dimensions of the conditional inputs and outputs of the two flows are shown in Table II. Importantly, these flows were trained using only single-photon showers.

One important difference from the original CaloFlow is that in this application to anomaly detection we do not have direct empirical access to $E_{\text{inc}}$. For a given shower, we do not know *a priori* what the corresponding $E_{\text{inc}}$ is and would instead have to use a reconstructed estimate of $E_{\text{inc}}$ which we denote as $E_{\text{inc}}^{(\text{rec})}$. In this work, we use a simple regression method to reconstruct $E_{\text{inc}}$ given the total deposited energy in the calorimeter $E_{\text{dep}}$. The reconstructed incident energy is

[2] We distinguish between kinetic energy and incident energy for massive particles. The incident energy is the sum of the kinetic energy and rest mass energy of the particle. This distinction becomes noticeable only in the case of $m_\chi = 5$ GeV.

[3] The layer energy of a given calorimeter layer is the sum of all the voxel energies in that layer.

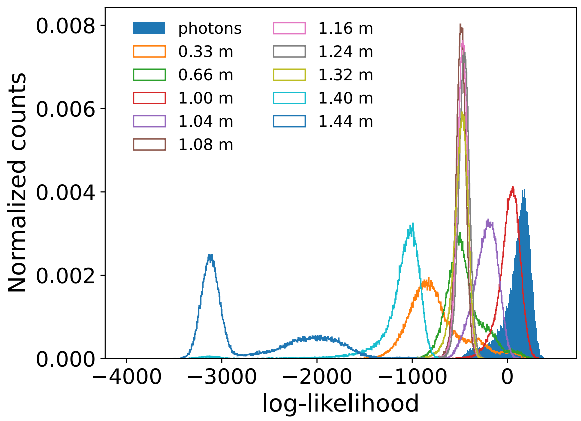

FIG. 2. Plot of log-likelihood comparison between signal showers from $\chi$ decays at the ten fixed displacements (lines) and background photon showers (filled). This example is for a $\chi$ particle with $m_\chi = 5$ GeV.

defined as $E_{\mathrm{inc}}^{(\mathrm{rec})} = \lambda E_{\mathrm{dep}}$, where $\lambda$ is the mean of $E_{\mathrm{inc}}/E_{\mathrm{dep}}$ in the single-photon training dataset. Note that the true $E_{\mathrm{inc}}$ is still used for training CaloFlow, while $E_{\mathrm{inc}}^{(\mathrm{rec})}$ is used when performing anomaly detection.

The architecture and training of CaloFlow are outlined in the Appendix. Some modifications were made to the original CaloFlow, while most of the main algorithm remains unchanged.

### A. Unsupervised anomaly detection with CaloFlow

After training CaloFlow on background single-photon showers, we evaluate the log-likelihood $\log p(\hat{\mathcal{I}}, E_i|E_{\mathrm{inc}}^{(\mathrm{rec})})$ of the background and signal showers by using the trained flows. Using both flow I and flow II, we are able to obtain the full log-likelihood:

$$\log p(\hat{\mathcal{I}}, E_i|E_{\mathrm{inc}}^{(\mathrm{rec})}) = \underbrace{\log p_1(E_i|E_{\mathrm{inc}}^{(\mathrm{rec})})}_{\text{flow I}} + \underbrace{\log p_2(\hat{\mathcal{I}}|E_i, E_{\mathrm{inc}}^{(\mathrm{rec})})}_{\text{flow II}}. \tag{1}$$

Figure 2 shows an example of a plot of the full log-likelihood of the signal showers from $\chi$ decays at the ten fixed displacements and background photon showers.

We see from Fig. 2 that the signal showers generally have log-likelihoods that are distinguishable from that of the background showers. Hence, we can use CaloFlow as an unsupervised anomaly detector by placing cuts on the full log-likelihood to discriminate signal from background showers. The results are detailed in Sec. IV. Though not explained here, the main features found in Fig. 2 can be understood from the discussion in Sec. IV A.

Despite the possibility of data-MC differences affecting the sensitivity of our approach, we note that the accuracy would not be affected, since one would presumably apply standard downstream background estimation. Furthermore, experiments usually calibrate their fast simulation and those calibrations could be applied to improve the sensitivity.

## IV. RESULTS

### A. Decay at fixed displacement

In this section, we study the effect of varying the displacement from the center of the detector at which the decay occurs on the evaluated likelihood. In reality, particles do not decay at fixed positions, but instead the probability of a particle decaying at a given displacement from where it was created is related to its lifetime $\tau$. Nevertheless, studying the showers produced by the particle $\chi$ at fixed decay positions is interesting from an experimental perspective, and doing so also helps us interpret the more physical results in Sec. IV B.

Our performance metric for anomaly detection will be the significance improvement which is defined as the signal efficiency (i.e., true positive rate or tpr) divided by the square root of the background efficiency (i.e., false positive rate or fpr). The maximum significance improvement corresponds to the best possible[4] multiplicative factor by which the signal significance can be improved with a cut on the anomaly score. Figure 3 shows a heat map of maximum significance improvement for each of the four different particle masses $m_\chi$ and the ten fixed displacements where the decay occurs. In general, we find that showers from decays at larger $z$ are more anomalous. However, there is a clear exception in the case of $m_\chi = 5$ GeV, where the showers originating from decays that occur before the $\chi$ particle reaches the calorimeter (e.g., $z = 0.33$, $0.66$ m) are more anomalous than those from decays occurring at the front face of the calorimeter (i.e., $z = 1.0$ m). This is due to the fact that the 5 GeV particle is less boosted compared to the other lighter particles that we consider in this study. As a result, the decay of the 5 GeV particle often results in a wider angle between the produced pair of photons which CaloFlow is better able to distinguish from the background single-photon showers. See Fig. 4 for an example of such a shower and comparison with a regular photon shower. On the other hand, if a decay occurs right at the front face of the calorimeter, there is insufficient time for the pair of photons to propagate and create two distinct blobs of energy in the calorimeter. The sudden jump in maximum significance improvement for all $m_\chi$ when going from $z = 1.32$ m to $z = 1.4$ m is due to the discretization of the calorimeter voxel geometry in the longitudinal direction described in Sec. II. The transition between the second and third longitudinal layers occurs at $z = 1.437$ m. Hence, showers from decays at $z = 1.4$ m, which have more energy

[4] This is signal model dependent but still provides a useful bound on the achievable performance.

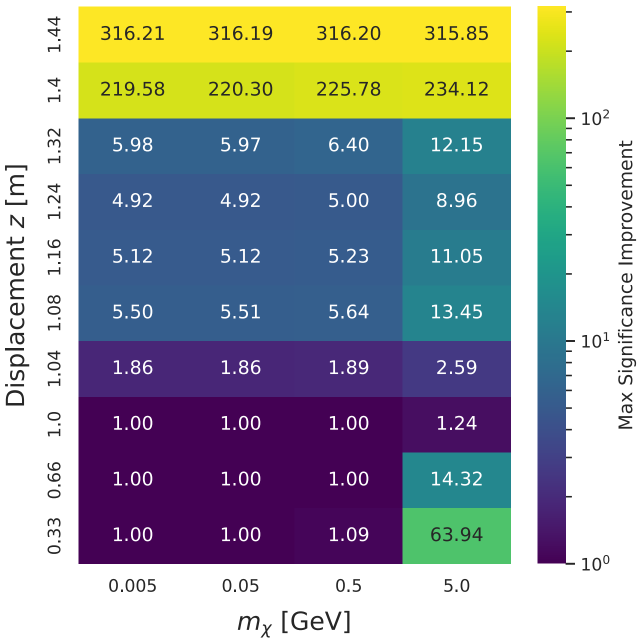

FIG. 3. Heat map of maximum significance improvement ($\mathrm{tpr}/\sqrt{\mathrm{fpr}}$) for different masses of the $\chi$ particle $m_\chi$ and the ten fixed displacements where the decay occurs.

deposited in the third layer, are flagged as significantly more anomalous.

For the three lowest $m_\chi$, we find that CaloFlow struggles to detect signal showers that originate from decays occurring before the front face of the calorimeter (i.e., $z \leq 1.0$ m). These lower-mass particles are highly boosted, which results in showers from highly collimated photon pairs that are more similar to the background photon showers.

### B. Varying lifetime

To assess the performance of our unsupervised anomaly detector on realistic scenarios, as mentioned in Sec. II, we construct new datasets consisting of showers from $\chi$ particles with fixed rest frame lifetimes. In particular, for each chosen mass $m_\chi$ and lifetime $\tau$, we have a total of 100 000 events that are made up of showers from $\chi$ decays at fixed displacements. The proportion of showers associated to decays at each fixed displacement is determined based on the lifetime $\tau$. Particles that decay after the calorimeter volume are not included in the events, since they are not detected within the calorimeter.

The probability for a particle to survive for time $t$ before decaying is given by $P_s(t) = \exp(-\frac{t}{\gamma\tau})$, where the relativistic boost factor $\gamma = E_{\mathrm{particle}}/M_{\mathrm{particle}}$. Equivalently, the probability that a particle decays before reaching displacement $z$ is given by $P_d(z) = 1 - \exp\left(-\frac{z}{c\tau\sqrt{\gamma^2-1}}\right)$. Since we consider ten fixed displacements indexed by $i \in \{1, 2, \ldots, 10\}$, the fraction of the total showers that originates from $\chi$ decay at displacement $z_i$ is set to be $\hat{w}_i$, which is defined by

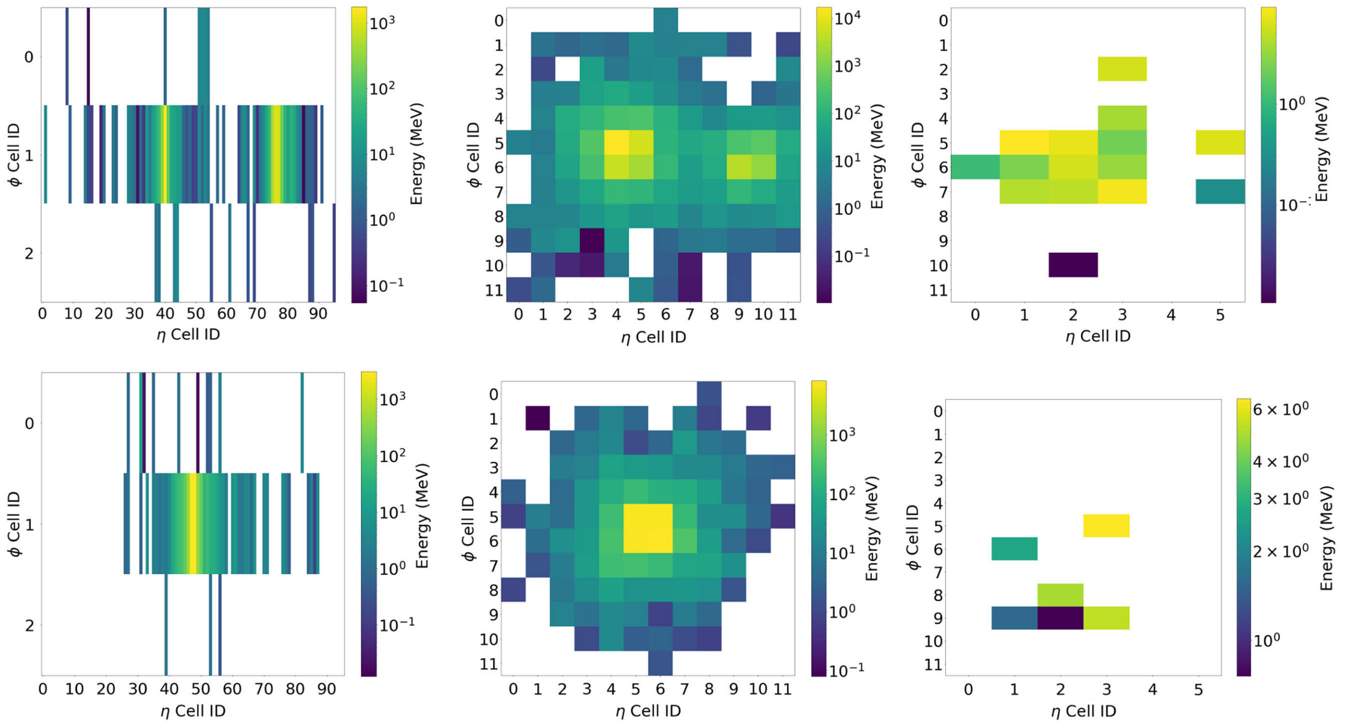

FIG. 4. Top row: example of a shower with two distinct blobs of energy originating from a 5 GeV $\chi$ particle that decayed at $z < 1.0$ m. Bottom row: example of a typical photon shower with centralized energy core. The energy deposition in each of the three layers is shown here with layer 0 on the left, layer 1 in the center, and layer 2 on the right.

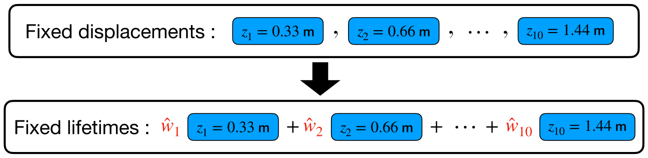

FIG. 5. Visualization of how the fixed lifetime datasets are constructed by sampling from the fixed displacement datasets. The fraction of total showers in each fixed lifetime dataset originating from $\chi$ decay at displacement $z_i$ is set to be $\hat{w}_i$.

$$w_i = \begin{cases} P_d(z_1), & i = 1, \\ P_d(z_i) - P_d(z_{i-1}), & \text{otherwise} \end{cases}$$
$$\text{and} \quad \hat{w}_i = \frac{w_i}{\sum_{j=1}^{10} w_j}.$$

In other words, the number of showers originating from the decay at $z = z_i$ is equal to $10^5 \times \hat{w}_i$.[5] In this work, we consider three possible lifetimes $\tau \in \{0.01, 0.1, 1\}$ ns for each choice of $m_\chi$. A visualization of how the fixed lifetime datasets are constructed from showers in the fixed displacement datasets is shown in Fig. 5. In Table III, we show the percentage of $\chi$ decays that occur before and within the calorimeter volume and also the average flight length of the $\chi$ particle for different $m_\chi$ and lifetimes $\tau$.

Figure 6 shows the significance improvement of our CaloFlow anomaly detector as a function of the background rejection (1/false positive rate) for various types of signal showers originating from $\chi$ particles with different $m_\chi$ and $\tau$.

Overall, we find that CaloFlow is able to achieve significantly better performance compared to a random classifier which we take to be the baseline. One exception is for $(m_\chi, \tau) = (0.5\ \text{GeV}, 0.01\ \text{ns})$, where the performance closely matches that of the random baseline. The poorer performance here is due to the lower boost factor $\gamma$ for the 0.5 GeV particle compared to the two lower particles masses which results in the majority of the decays occurring close to the center of the detector ($z = 0$ m). Meanwhile, the mass is still low enough that the photons are not widely separated by the time they reach the calorimeter. As we have seen in Fig. 3, such showers are mostly not detected as anomalous by CaloFlow. At longer lifetimes, the showers are more anomalous, as a larger proportion of them originate from decays occurring within the calorimeter.

For the three lowest $m_\chi$, the significance improvement generally increases with background rejection. Also, we find that the significance improvement curves are similar across lifetimes for the two lowest $m_\chi$. In these cases, $\frac{z}{c\tau\sqrt{\gamma^2-1}}$ is small, which implies that $P_d(z) \approx \frac{z}{c\tau\sqrt{\gamma^2-1}}$.

TABLE III. Percentage of $\chi$ decays (left number) that occur before and within the calorimeter volume and the average flight length (right number) of $\chi$ particles for different $m_\chi$ and lifetimes $\tau$.

| | $m_\chi$ | | | |
|---|---|---|---|---|
| $\tau$ (ns) | $5 \times 10^{-3}$ GeV | $5 \times 10^{-2}$ GeV | $5 \times 10^{-1}$ GeV | 5 GeV |
| 0.01 | 5% / 30 m | 38% / 3 m | 99% / 0.3 m | 100% / 0.03 m |
| 0.1 | 0.5% / 300 m | 5% / 30 m | 38% / 3 m | 99% / 0.3 m |
| 1 | 0.05% / 3000 m | 0.5% / 300 m | 5% / 30 m | 38% / 3 m |

Hence, the lifetime $\tau$ cancels out when computing $\hat{w}_i$. In other words, at fixed $m_\chi$, the large boost of these particles results in a similar proportion of particles decaying at a given fixed displacement before and within the calorimeter for different lifetimes $\tau$.

The best performance among all the (fixed mass and lifetime) signal models we considered in this study was achieved in the cases with the largest mass of $m_\chi = 5$ GeV. As explained in Sec. IV A, showers from early decay of these more massive particles are quite anomalous according to CaloFlow due to the wider angle between the produced pair of photons. In this case, going to higher lifetimes actually makes these showers slightly *less* anomalous (which is opposite to the trend seen at lower masses), since it gives the photons less time to separate before interacting with the detector.

There is a local maximum in each of the significance improvement curves for $m_\chi = 5$ GeV. To understand the local maximum, we have to look at the log-likelihood plot for $m_\chi = 5$ GeV shown in Fig. 7. The peaks at the log-likelihoods of $-1000$ and $-500$ are due to showers from decays occurring at $z = 0.33$ and 0.66 m, respectively, which can also be seen in Fig. 2. Sliding the cut on log-likelihood in the direction of decreasing log-likelihood is equivalent to increasing the background rejection (1/fpr). Notice that sliding the cut in the direction of decreasing log-likelihood across peaks in the $\chi$ shower log-likelihood curve would create a local maximum in the plot of significance improvement vs background rejection (1/fpr), since the tpr decreases faster than the increase in $1/\sqrt{\text{fpr}}$.

## C. Comparison with supervised anomaly detection

The previous section showed that the unsupervised anomaly detector has broad coverage across the various model parameters. An important question to ask is how this compares to a dedicated search for the $\chi \to \gamma\gamma$ signal. For a particular signal model, we would expect the dedicated search to outperform the unsupervised approach. However, it is not possible to have a dedicated search for every possible signal, and so the key question to ask is how well a supervised model trained on one signal would perform on other signals not seen during training.

[5] In some cases, we had to round $\hat{w}_i$ to ensure that the total number of showers is equal to 100 000.

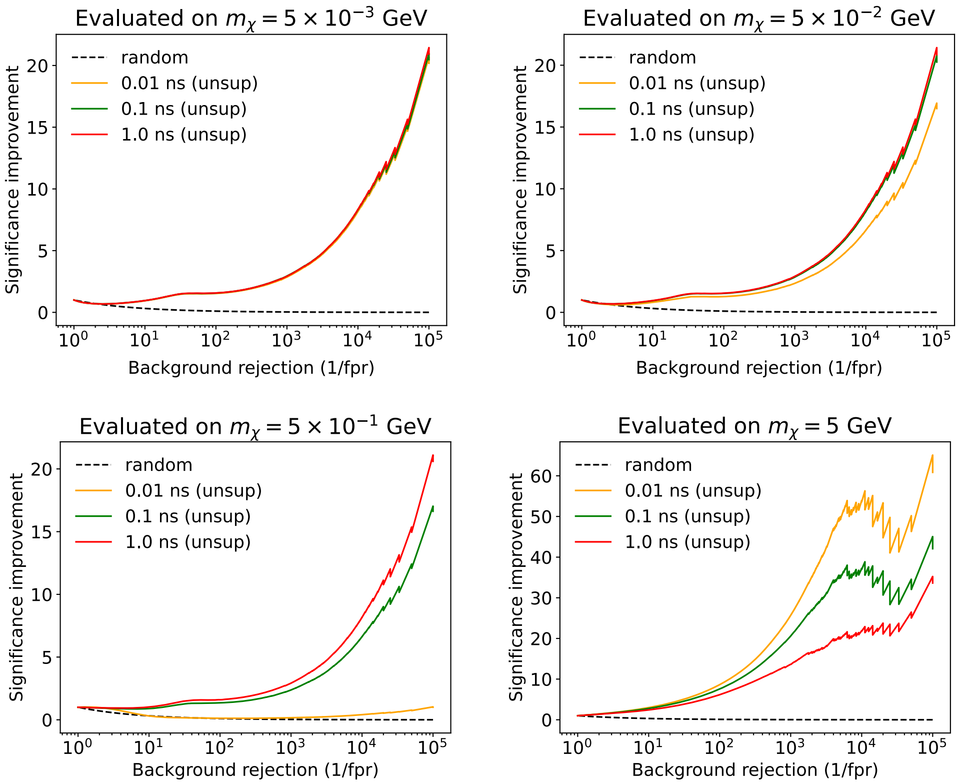

FIG. 6. Plots of significance improvement of our unsupervised (unsup) CaloFlow anomaly detector as a function of background rejection (1/false positive rate) for various types of signal showers originating from $\chi$ particles with different $m_\chi$ and $\tau$. Top left: $m_\chi = 5 \times 10^{-3}$ GeV; top right: $m_\chi = 0.05$ GeV; bottom left: $m_\chi = 0.5$ GeV; bottom right: $m_\chi = 5$ GeV. The performance of a random classifier is drawn with black dashed lines to serve as a baseline.

In this section, we compare the performance of our method against two supervised classifiers. Each supervised classifier was trained on a combined dataset with 100 000 signal showers and 100 000 background showers. The showers originate from particles with kinetic energy uniformly distributed in the range [1, 100] GeV. The signal showers originate from $\chi$ particles with lifetime $\tau = 1.00$ ns. The first (second) supervised classifier was trained on a dataset with $m_\chi = 5 \times 10^{-3} (5)$ GeV. These models were chosen because they are sufficiently different that they would likely be covered by different dedicated searches. It is, thus, interesting to ask if a search optimized for one of the models would still be sensitive to the other, since the unsupervised approach has some sensitivity to both models.

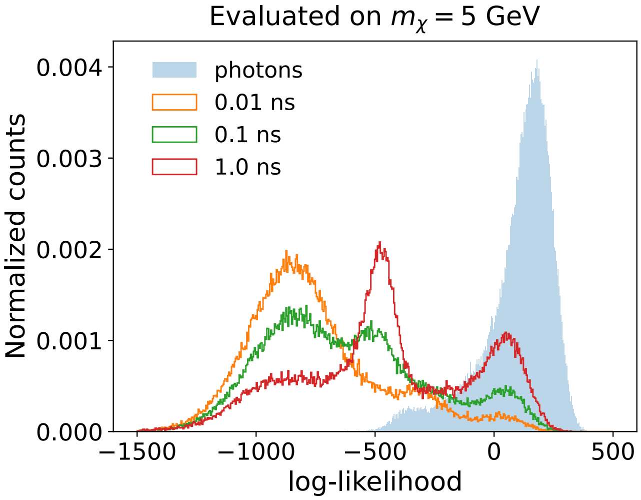

FIG. 7. Plot of log-likelihood of $\chi$ showers for $m_\chi = 5$ GeV at the different $\chi$ particle lifetimes.

The supervised classifier is a fully connected neural network with two hidden layer with 512 nodes each. We have a 505-dimensional input consisting of the voxel energies normalized by the reconstructed incident energy $\vec{\mathcal{I}}/E_{\mathrm{inc}}^{(\mathrm{rec})}$ (504-dim) and the reconstructed incident energy $E_{\mathrm{inc}}^{(\mathrm{rec})}$ (1-dim). The output layer returns a single number which is passed through a sigmoid activation function. All other activation functions are rectified linear units. The supervised classifier was trained for a total of 50 epochs with a train/test/validation split of 60%/20%/20%. These parameters were not extensively optimized, but we found little gain from small variations in the setup. CaloFlow has a total of ∼38 000 000 parameters, while the supervised

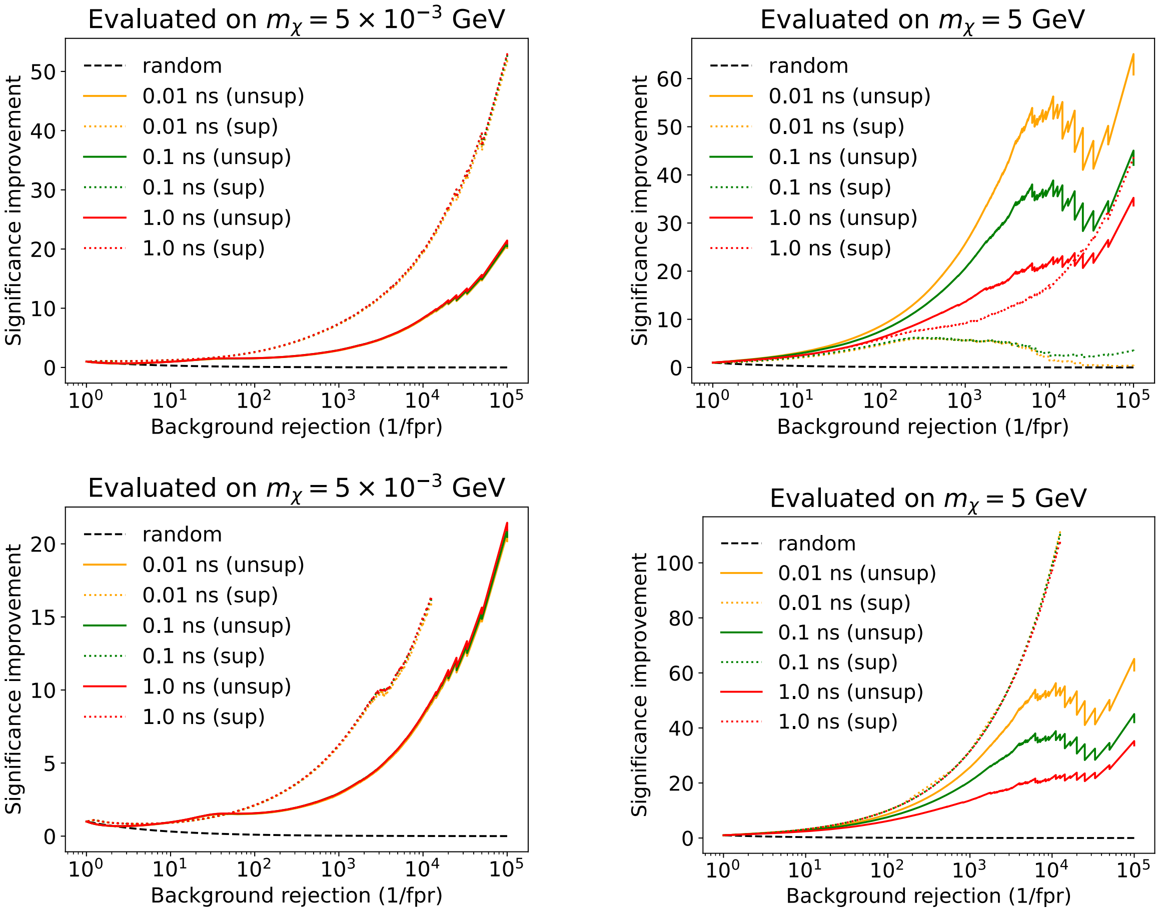

FIG. 8. Comparison plots of significance improvement of our CaloFlow anomaly detector as a function of background rejection (1/false positive rate). The plots for the performance for the $m_\chi = 5 \times 10^{-3}$ GeV and $m_\chi = 5$ GeV cases are shown on the left and right, respectively. Top row: a supervised classifier was trained on signal showers originating from $\chi$ particles with $m_\chi = 5 \times 10^{-3}$ GeV and lifetime $\tau = 1.00$ ns. Bottom row: a supervised classifier was trained on signal showers originating from $\chi$ particles with $m_\chi = 5$ GeV and lifetime $\tau = 1.00$ ns. The performance of our unsupervised CaloFlow anomaly detector (unsup) is shown as solid lines, while that of the supervised classifier (sup) is shown as dotted lines. The performance of a random classifier is drawn with black dashed lines to serve as a baseline.

classifier only has ~522 000 parameters. However, we found that increasing the number of parameters for the supervised classifier does not help to increase its performance. CaloFlow has a larger number of parameters because it has a more difficult task of learning the likelihood of the data.

After training the supervised classifiers, they were evaluated on the signal showers discussed in Sec. IV B. Figure 8 shows the comparison plots between the performance of our CaloFlow anomaly detector (unsup) and the supervised classifiers (sup). We see from the figure that the supervised classifier always outperforms CaloFlow on showers originating from $\chi$ particles that have the same mass as what it was trained on (see top left and bottom right plots). However, the performance of the supervised classifier usually decreases when evaluated on signal showers originating from $\chi$ particles that have very different masses from what it was trained on (see top right and bottom left plots). Whether the supervised classifier or CaloFlow achieves better performance depends on $m_\chi$.

(i) For the upper right plot (trained on $m_\chi = 5 \times 10^{-3}$ GeV and evaluated on $m_\chi = 5$ GeV), CaloFlow easily outperforms the supervised classifier, as the supervised classifier did not see training examples of anomalous showers that are characteristic of larger $m_\chi$ (e.g., early decay resulting in two blobs) and is unable to generalize its performance.

(ii) However, in the lower left plot (trained on $m_\chi = 5$ GeV and evaluated on $m_\chi = 5 \times 10^{-3}$ GeV), CaloFlow does not outperform the supervised classifier. This is likely due to CaloFlow not being able to fully discriminate against signal showers that decay only after the first longitudinal layer (i.e., $E_0 = 0$ GeV), whereas such showers are usually perfectly distinguished by the supervised classifier.[6] As noise is added to the voxel energies when using CaloFlow, this artificially causes $E_0 > 0$ GeV[7] and prevents such showers from being perfectly distinguished by CaloFlow.

Finally, let us also comment on some other interesting features in Fig. 8. When the classifier or anomaly detector is evaluated on $m_\chi = 5 \times 10^{-3}$ GeV (left column in Fig. 8), the performance is the same for all the lifetimes considered. This is because, for $m_\chi = 5 \times 10^{-3}$ GeV, the particles are highly boosted, so there is a similar proportion of the total number particles decaying at a given fixed displacement before and within the calorimeter for different lifetimes. (Keep in mind that we consider only particles decaying before and within the calorimeter.) Interestingly, the fully supervised classifier is also the same for all lifetimes for the bottom right plot (trained and evaluated on $m_\chi = 5$ GeV). Here, the reason is that the signal showers look extremely different from the background photon showers and are perfectly distinguished by the classifier. Thus, we see that the significance improvement is equal to $1/\sqrt{\mathrm{fpr}}$.

This comparison of our unsupervised anomaly detection method against a supervised classifier highlights the potential limitation of model-specific anomaly detection, as the supervised model is unable to generalize its excellent performance to signal that is too different from what it was trained on. We note that it is possible to train a supervised classifier on all the types of signal showers we have considered here. Doing so would likely result in the supervised classifier outperforming CaloFlow when evaluated on all the signal showers. However, the point is that, even if one is to train a supervised classifier on a large number of signal types, it is impossible to exhaust the space of all possible signals. Hence, there may be an advantage in using model-agnostic, unsupervised anomaly detection methods such as the one we proposed in this work. This is especially true when using flow-based fast calorimeter simulators, as no additional training has to be performed to use them as unsupervised anomaly detectors.

## V. CONCLUSION AND OUTLOOK

Using CaloFlow as an example, we demonstrated how fast calorimeter surrogate models with access to the data likelihood can double up as unsupervised anomaly detectors.

By studying the anomaly detection performance of CaloFlow on showers from $\chi$ particle decays occurring at fixed displacements in the detector, we found that CaloFlow is generally more sensitive to signal showers from decays that occur deeper in the calorimeter. However, in the case of more massive, less highly boosted $\chi$ particles, CaloFlow still has significant discriminative power for showers from decays occurring in front of the calorimeter.

By reweighting the proportion of showers originating from decays at each fixed displacement, we constructed signal datasets corresponding to fixed particle lifetimes. We found that CaloFlow has discriminative power for most of the models we tested. In particular, CaloFlow achieves the best performance in the case with $m_\chi = 5$ GeV and $\tau = 0.01$ ns where the particle is less highly boosted and the majority of the particles decay close to the center of the detector.

Finally, we compared the performance of our unsupervised CaloFlow anomaly detector against a supervised classifier. We found that a supervised classifier trained on signal showers from highly boosted $\chi$ particles performed significantly poorer on showers from more massive, less highly boosted particles compared to our unsupervised method. When trained on signal showers from more massive $\chi$ particles and applied to signal showers from less massive $\chi$ particles, the supervised classifier still performs well. This highlights the complementarity of different approaches and reaffirms the need to have a diversity of methods in order to achieve broad sensitivity.

The datasets used in this study can be found at [54], and the software to generate these datasets are located at [102]. The machine learning software is at [103].

## ACKNOWLEDGMENTS

C. K. thanks the Baden-Württemberg-Stiftung for financing through the program *Internationale Spitzenforschung*, project *Uncertainties—Teaching AI its Limits* (BWST_IF2020-010). I. P. and D. S. are supported by the U.S. Department of Energy (DOE), Office of Science Grant No. DOE-SC0010008, and B. N. is supported by the DOE under Contract No. DE-AC02-05CH11231.

## APPENDIX: ARCHITECTURE AND TRAINING

Here, we briefly describe the architecture and training procedure used for CaloFlow (see [21,22] for more details). There are some differences compared to the implementation in the original CaloFlow papers [21,22], but most of the main algorithm remains the same.

Both flow I and flow II are masked autoregressive flows [104] with compositions of rational quadratic splines (RQS) [105] as transformations. The RQS transformations are parametrized using neural networks known as Masked Autoencoder for Distribution Estimation (MADE) blocks [106]. The details of the architecture of flow I and flow II are summarized in Table IV.

[6] Here, the point is that CaloFlow is not able to distinguish showers from late decays as well as the supervised classifier can. Nevertheless, for $m_\chi = 5 \times 10^{-3}$ GeV, CaloFlow still has a better anomaly detection performance on late decay (larger $z$) showers compared to early decay (smaller $z$) showers.

[7] We checked that a different treatment of the layers with zero energy deposition does not improve the performance.

TABLE IV. Summary of architecture of flow I and flow II. For the hidden layer sizes, the first number is the number of hidden layers in each MADE block and the second number is the number of nodes in each hidden layer (e.g., 2 × 64 refers to two hidden layers per MADE block with 64 nodes per hidden layer).

| | | | Layer sizes | | | | |
|---|---|---|---|---|---|---|---|
| Model | Dimension of base distribution | Number of MADE blocks | Input | Hidden | Output | Number of RQS bins | RQS tail bound |
| Flow I | 3 | 6 | 64 | $2\times 64$ | 69 | 8 | 14 |
| Flow II | 504 | 8 | 378 | $1\times 378$ | 11592 | 8 | 14 |

The incident energy of the incoming photon is preprocessed as

$$E_{\text{inc}} \to \log_{10}(E_{\text{inc}}/10\ \text{GeV}). \tag{A1}$$

The inputs to the flows are preprocessed as follows.

(i) Flow I: $E_i \to 2(\log_{10}(E_i + 1\ \text{keV}) - 1)$.

(ii) Flow II:

$E_i \to \log_{10}(E_i + 1\ \text{keV}) - 2$,

$\mathcal{I}_{ia} \to u_{\text{logit},ia}(\mathcal{I}_{ia}/E_i)$,

where $u_{\text{logit},ia} = \log \frac{\tilde{u}_{ia}}{1-\tilde{u}_{ia}}$,

$\tilde{u}_{ia} = \alpha + (1 - 2\alpha)u_{ia}$, and $\alpha = 10^{-6}$.

The index $i$ denotes the layer number, while the index $a$ specifies the voxel within the given layer. In the original CaloFlow, a different preprocessing was used for the layer energies $E_i$ in flow I where $E_i$ were transformed to unit space (see [21]).

As in Refs. [21,22], uniform noise in the range [0, 1] keV was applied to the voxel energies during training and evaluation. The addition of noise was found to prevent the flow from fitting unimportant features. The training of both flows in this work is optimized using independent Adam optimizers [107]. Flow I was trained by minimizing $-\log p(E_0, E_1, E_2|E_{\text{inc}})$ for 75 epochs with a batch size of 200. Flow II was trained by minimizing $-\log p(\hat{\mathcal{I}}|E_0, E_1, E_2, E_{\text{inc}})$ for 100 epochs with a batch size of 200. The initial learning of $1 \times 10^{-4}$ was chosen for the two flows, and a multistep learning schedule was used when training the flows which halves the learning rate after each selected epoch milestone during the training.

[1] ATLAS Collaboration, Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B **716**, 1 (2012).

[2] CMS Collaboration, Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Phys. Lett. B **716**, 30 (2012).

[3] ATLAS Collaboration, Exotic physics searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ExoticsPublicResults.

[4] ATLAS Collaboration, Supersymmetry searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults.

[5] ATLAS Collaboration, Higgs and Diboson searches (2023), https://twiki.cern.ch/twiki/bin/view/AtlasPublic/HDBSPublicResults.

[6] CMS Collaboration, CMS exotica public physics results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO.

[7] CMS Collaboration, CMS supersymmetry physics results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS.

[8] CMS Collaboration, CMS beyond-two-generations (B2G) public physics results (2023), https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsB2G.

[9] LHCb Collaboration, Publications of the QCD, Electroweak and Exotica Working Group (2023), http://lhcbproject.web.cern.ch/lhcbproject/Publications/LHCbProjectPublic/Summary_QE E.html.

[10] HEP ML Community, A living review of machine learning for particle physics, https://iml-wg.github.io/HEPML-LivingReview/.

[11] G. Kasieczka *et al.*, The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics, Rep. Prog. Phys. **84**, 124201 (2021).

[12] T. Aarrestad *et al.*, The dark machines anomaly score challenge: Benchmark data and model independent event classification for the Large Hadron Collider, SciPost Phys. **12**, 043 (2022).

[13] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, Machine learning in the search for new fundamental physics, arXiv:2112.03769.

[14] M. Feickert and B. Nachman, A living review of machine learning for particle physics, arXiv:2102.02770.

[15] E. G. Tabak and C. V. Turner, A family of nonparametric density estimation algorithms, Commun. Pure Appl. Math. **66**, 145 (2013).

[16] L. Dinh, D. Krueger, and Y. Bengio, Nice: Non-linear independent components estimation, arXiv:1410.8516.

[17] D. J. Rezende and S. Mohamed, Variational inference with normalizing flows, arXiv:1505.05770.

[18] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using real NVP, arXiv:1605.08803.

[19] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, https://www.jmlr.org/papers/v22/19-1028.html.

[20] C. Winkler, D. E. Worrall, E. Hoogeboom, and M. Welling, Learning likelihoods with conditional normalizing flows, arXiv:1912.00042.

[21] C. Krause and D. Shih, CaloFlow: Fast and accurate generation of calorimeter showers with normalizing flows, Phys. Rev. D **107**, 113003 (2023).

[22] C. Krause and D. Shih, CaloFlow II: Even faster and still accurate generation of calorimeter showers with normalizing flows, Phys. Rev. D **107**, 113004 (2023).

[23] C. Krause, I. Pang, and D. Shih, CaloFlow for CaloChallenge dataset 1, SciPost Phys. **16**, 126 (2024).

[24] M. R. Buckley, C. Krause, I. Pang, and D. Shih, Inductive CaloFlow, Phys. Rev. D **109**, 033006 (2024).

[25] S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, C. Krause, I. Shekhzadeh, and D. Shih, L2LFlows: Generating high-fidelity 3D calorimeter images, J. Instrum. **18**, P10017 (2023).

[26] I. Pang, J. A. Raine, and D. Shih, SuperCalo: Calorimeter shower super-resolution, Phys. Rev. D **109**, 092009 (2024).

[27] F. Ernst, L. Favaro, C. Krause, T. Plehn, and D. Shih, Normalizing flows for high-dimensional detector simulations, arXiv:2312.09290.

[28] B. Nachman and D. Shih, Anomaly detection with density estimation, Phys. Rev. D **101**, 075042 (2020).

[29] C. Gao, J. Isaacson, and C. Krause, i-flow: High-dimensional integration and sampling with normalizing flows, Mach. Learn. Sci. Technol. **1**, 045023 (2020).

[30] E. Bothmann, T. Janßen, M. Knobbe, T. Schmale, and S. Schumann, Exploring phase space with neural importance sampling, SciPost Phys. **8**, 069 (2020).

[31] C. Gao, S. Höche, J. Isaacson, C. Krause, and H. Schulz, Event generation with normalizing flows, Phys. Rev. D **101**, 076002 (2020).

[32] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe, Invertible networks or partons to detector and back again, SciPost Phys. **9**, 074 (2020).

[33] B. Stienen and R. Verheyen, Phase space sampling and inference from weighted events with autoregressive flows, SciPost Phys. **10**, 038 (2021).

[34] S. Bieringer, A. Butter, T. Heimel, S. Höche, U. Köthe, T. Plehn, and S. T. Radev, Measuring QCD splittings with invertible networks, SciPost Phys. **10**, 126 (2021).

[35] M. Bellagente, M. Haußmann, M. Luchmann, and T. Plehn, Understanding event-generation networks via uncertainties, SciPost Phys. **13**, 003 (2022).

[36] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih, and M. Sommerhalder, Classifying anomalies through outer density estimation (CATHODE), Phys. Rev. D **106**, 055006 (2022).

[37] T. Bister, M. Erdmann, U. Köthe, and J. Schulte, Inference of cosmic-ray source properties by conditional invertible neural networks, Eur. Phys. J. C **82**, 171 (2022).

[38] A. Butter, T. Heimel, S. Hummerich, T. Krebs, T. Plehn, A. Rousselot, and S. Vent, Generative networks for precision enthusiasts, SciPost Phys. **14**, 078 (2023).

[39] R. Winterhalder, V. Magerya, E. Villa, S. P. Jones, M. Kerner, A. Butter, G. Heinrich, and T. Plehn, Targeting multi-loop integrals with neural networks, SciPost Phys. **12**, 129 (2022).

[40] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, T. Plehn, D. Shih, and R. Winterhalder, Ephemeral learning—augmenting triggers with online-trained normalizing flows, SciPost Phys. **13**, 087 (2022).

[41] R. Verheyen, Event generation and density estimation with surjective normalizing flows, SciPost Phys. **13**, 047 (2022).

[42] M. Leigh, J. A. Raine, and T. Golling, $\nu$-Flows: Conditional neutrino regression, SciPost Phys. **14**, 159 (2023).

[43] A. Butter, T. Heimel, T. Martini, S. Peitzsch, and T. Plehn, Two invertible networks for the matrix element method, SciPost Phys. **15**, 094 (2023).

[44] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, Resonant anomaly detection without background sculpting, Phys. Rev. D **107**, 114012 (2023).

[45] T. Heimel, R. Winterhalder, A. Butter, J. Isaacson, C. Krause, F. Maltoni, O. Mattelaer, and T. Plehn, MadNIS—Neural multi-channel importance sampling, SciPost Phys. **15**, 141 (2023).

[46] M. Backes, A. Butter, M. Dunford, and B. Malaescu, An unfolding method based on conditional Invertible Neural Networks (cINN) using iterative training, SciPost Phys. Core **7**, 007 (2024).

[47] J. A. Raine, M. Leigh, K. Zoch, and T. Golling, $\nu^2$-Flows: Fast and improved neutrino reconstruction in multi-neutrino final states with conditional normalizing flows, Phys. Rev. D **109**, 012005 (2024).

[48] D. Sengupta, S. Klein, J. A. Raine, and T. Golling, CURTAINs flows for flows: Constructing unobserved regions with maximum likelihood estimation, arXiv:2305.04646.

[49] J. Ackerschott, R. K. Barman, D. Gonçalves, T. Heimel, and T. Plehn, Returning *CP*-observables to the frames they belong, SciPost Phys. **17**, 001 (2024).

[50] T. Heimel, N. Huetsch, R. Winterhalder, T. Plehn, and A. Butter, Precision-machine learning for the matrix element method, arXiv:2310.07752.

[51] T. Heimel, N. Huetsch, F. Maltoni, O. Mattelaer, T. Plehn, and R. Winterhalder, The MadNIS reloaded, arXiv:2311.01548.

[52] C. Bierlich, P. Ilten, T. Menzo, S. Mrenna, M. Szewc, M. K. Wilkinson *et al.*, Towards a data-driven model of hadronization using normalizing flows, arXiv:2311.09296.

[53] R. Das, G. Kasieczka, and D. Shih, Residual ANODE, arXiv:2312.11629.

[54] C. Krause, B. Nachman, I. Pang, D. Shih, and Y. Zhu, Electromagnetic sampling calorimeter shower images, 10.5281/zenodo.10393540 (2023).

[55] B. Nachman, L. de Oliveira, and M. Paganini, Electromagnetic calorimeter shower images, Mendeley Data (2017).

[56] 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).

[57] 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).

[58] 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).

[59] 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).

[60] 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).

[61] ATLAS Collaboration, Deep generative models for fast shower simulation in ATLAS, Report No. ATL-SOFT-PUB-2018-001, 2018.

[62] D. Belayneh *et al.*, Calorimetry with deep learning: Particle simulation and reconstruction for collider physics, Eur. Phys. J. C **80**, 688 (2020).

[63] S. Vallecorsa, F. Carminati, and G. Khattak, 3D convolutional GAN for fast simulation, EPJ Web Conf. **214**, 02010 (2019).

[64] SHiP Collaboration, Fast simulation of muons produced at the SHiP experiment using generative adversarial networks, J. Instrum. **14**, P11028 (2019).

[65] V. Chekalina, E. Orlova, F. Ratnikov, D. Ulyanov, A. Ustyuzhanin, and E. Zakharov, Generative models for fast calorimeter simulation. LHCb case, EPJ Web Conf. **214**, 02034 (2019).

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

[67] S. Vallecorsa, Generative models for fast simulation, J. Phys. Conf. Ser. **1085**, 022005 (2018).

[68] P. Musella and F. Pandolfi, Fast and accurate simulation of particle detectors using generative adversarial networks, Comput. Software Big Sci. **2**, 8 (2018).

[69] K. Deja, T. Trzcinski, and L. Graczykowski, Generative models for fast cluster simulations in the TPC for the ALICE experiment, EPJ Web Conf. **214**, 06003 (2019).

[70] ATLAS Collaboration, Deep generative models for fast photon shower simulation in ATLAS, Comput. Software Big Sci. **8**, 7 (2024).

[71] ATLAS Collaboration, AtlFast3: The next generation of fast simulation in ATLAS, Comput. Software Big Sci. **6**, 7 (2022).

[72] 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).

[73] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, D. Hundhausen, G. Kasieczka *et al.*, Hadrons, better, faster, stronger, Mach. Learn. Sci. Technol. **3**, 025014 (2022).

[74] V. Mikuni and B. Nachman, Score-based generative models for calorimeter shower simulation, Phys. Rev. D **106**, 092009 (2022).

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

[76] H. Hashemi, N. Hartmann, S. Sharifzadeh, J. Kahn, and T. Kuhr, Ultra-high-resolution detector simulation with intra-event aware GAN and self-supervised relational reasoning, Nat. Commun. **15**, 5825 (2024).

[77] J. Liu, A. Ghosh, D. Smith, P. Baldi, and D. Whiteson, Generalizing to new geometries with geometry-aware autoregressive models (GAAMs) for fast calorimeter simulation, J. Instrum. **18**, P11003 (2023).

[78] S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, K. Krüger *et al.*, New angles on fast calorimeter shower simulation, Mach. Learn. Sci. Technol. **4**, 035044 (2023).

[79] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol, W. Korcari, K. Krüger, and P. McKeown, CaloClouds: Fast geometry-independent highly-granular calorimeter simulation, J. Instrum. **18**, P11025 (2023).

[80] V. Mikuni and B. Nachman, CaloScore v2: Single-shot calorimeter shower simulation with diffusion models, J. Instrum. **19**, P02001 (2024).

[81] F. T. Acosta, V. Mikuni, B. Nachman, M. Arratia, K. Barish, B. Karki, R. Milton, P. Karande, and A. Angerami, Comparison of point cloud and image-based models for calorimeter fast simulation, J. Instrum. **19**, P05003 (2024).

[82] M. Faucci Giannelli, G. Kasieczka, C. Krause, B. Nachman, D. Salamani, D. Shih, and A. Zaborowska, Fast calorimeter simulation challenge (2022), https://calochallenge.github.io/homepage.

[83] V. Mikuni and B. Nachman, High-dimensional and permutation invariant anomaly detection, SciPost Phys. **16**, 062 (2024).

[84] J. Hajer, Y.-Y. Li, T. Liu, and H. Wang, Novelty detection meets collider physics, Phys. Rev. D **101**, 076015 (2020).

[85] M. Farina, Y. Nakai, and D. Shih, Searching for new physics with deep autoencoders, Phys. Rev. D **101**, 075021 (2020).

[86] T. Heimel, G. Kasieczka, T. Plehn, and J. M. Thompson, QCD or what?, SciPost Phys. **6**, 030 (2019).

[87] P. Kirichenko, P. Izmailov, and A. G. Wilson, Why normalizing flows fail to detect out-of-distribution data, arXiv:2006.08545.

[88] C. Le Lan and L. Dinh, Perfect density models cannot guarantee anomaly detection, Entropy **23**, 1690 (2021).

[89] G. Kasieczka, R. Mastandrea, V. Mikuni, B. Nachman, M. Pettee, and D. Shih, Anomaly detection under coordinate transformations, Phys. Rev. D **107**, 015009 (2023).

[90] ATLAS Collaboration, ATLAS liquid-argon calorimeter: Technical Design Report, Technical Design Report, ATLAS, CERN, Geneva, 1996.

[91] GEANT4 Collaboration, Geant4–A simulation toolkit, Nucl. Instrum. Methods Phys. Res., Sect. A **506**, 250 (2003).

[92] 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).

[93] 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).

[94] B. Ganhuyag and V. Uzhinsky, Modified FRITIOF code: Negative charged particle production in high energy nucleus nucleus interactions, Czech. J. Phys. **47**, 913 (1997).

[95] B. Nilsson-Almqvist and E. Stenlund, Interactions between hadrons and nuclei: The Lund Monte Carlo, FRITIOF Version 1.6, Comput. Phys. Commun. **43**, 387 (1987).

[96] B. Andersson, G. Gustafson, and B. Nilsson-Almqvist, A model for low p(t) hadronic reactions, with generalizations to hadron—nucleus and nucleus-nucleus collisions, Nucl. Phys. **B281**, 289 (1987).

[97] B. Andersson, A. Tai, and B.-H. Sa, Final state interactions in the (nuclear) FRITIOF string interaction scenario, Z. Phys. C **70**, 499 (1996).

[98] M. P. Guthrie, R. G. Alsmiller, and H. W. Bertini, Calculation of the capture of negative pions in light elements and comparison with experiments pertaining to cancer radiotherapy, Nucl. Instrum. Methods **66**, 29 (1968).

[99] H. W. Bertini and M. P. Guthrie, News item results from medium-energy intranuclear-cascade calculation, Nucl. Phys. **A169**, 670 (1971).

[100] V. A. Karmanov, Light front wave function of relativistic composite system in explicitly solvable model, Nucl. Phys. **B166**, 378 (1980).

[101] H. Burkhardt, V. Grichine, P. Gumplinger, V. Ivanchenko, R. Kokoulin, M. Maire *et al.*, Geant4 standard electromagnetic package for HEP applications, in *IEEE Symposium Conference Record Nuclear Science 2004* (2004), Vol. 3, pp. 1907–1910, 10.1109/NSSMIC.2004.1462617.

[102] https://github.com/hep-lbdl/CaloGAN/tree/samplingEM.

[103] https://github.com/Ian-Pang/AD_with_CF.

[104] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, arXiv:1705.07057.

[105] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, Adv. Neural Inf. Process. Syst. **32**, 7511 (2019).

[106] M. Germain, K. Gregor, I. Murray, and H. Larochelle, Made: Masked autoencoder for distribution estimation, in *Proceedings of the International Conference on Machine Learning, PMLR* (2015), pp. 881–889, https://proceedings.mlr.press/v37/germain15.html.

[107] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.