VERSATILE ENERGY-BASED MODELS FOR HIGH ENERGY PHYSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy-Based Models (EBMs) have the natural advantage of flexibility in the form of the energy function. Recently, EBMs have achieved great success in modeling high-dimensional data in computer vision and natural language processing. In accordance with these signs of progress, we build a versatile energy-based model for High Energy Physics events at the Large Hadron Collider. This framework builds on a powerful generative model and describes higher-order inter-particle interactions. It suits different encoding architectures and decomposes clearly. As for applicational aspects, it can serve as a powerful parameterized event generator, a generic anomalous signal detector, and an augmented event classifier.

1 INTRODUCTION

Energy-based Models (EBMs) (Hopfield, 1982; Ackley et al., 1985; LeCun et al., 2006), being a classical generative framework, leverage the energy function for learning dependencies between input variables. With an energy function $E(\mathbf{x})$ and constructing the un-normalized probabilities through the exponential $\tilde{p}(\mathbf{x}) = \exp(-E(\mathbf{x}))$, the energy model naturally yields a probability distribution. Despite the flexibility in the modeling, the training of EBMs has been cumbersome and unstable due to the intractable partition function and the corresponding Monte Carlo sampling involved. More recently, EBMs have been succeeding in high-dimensional modeling (Nijkamp et al., 2020; 2019a;b; Du & Mordatch; Du et al., b; Song & Ermon, 2020; Deng et al., 2020; Naskar et al., 2021) for computer vision and natural language processing. At the same time, it has been revealed that neural classifiers are naturally connected with EBMs (Xie et al., 2016; Grathwohl et al.; 2021), combining the discriminative and generative learning processes in a common learning regime. More interestingly, compositionality can be easily incorporated within the framework of EBMs by simply summing up the energy functions (Du et al., a; 2021).

On the other hand, statistical physics originally inspired the invention of EBMs. This natural connection makes EBMs appealing in modeling physical systems. In physical sciences, EBMs have been used to simulate condensed-matter systems and protein molecules (Noé et al., 2019). They have also been shown great potential in structure biology (Du et al., 2020), in a use-case of protein conformation.

Given the flexibility in the architecture and the compatibility with different tasks, we explore the potential of EBMs in modeling elementary particle radiation patterns. The Large Hadron Collider (LHC) (Eva, 2008), being the most energetic particle collider in human history, is colliding highly-energetic protons to examine the underlying physics of subatomic particles. After the great success in observing the Higgs boson (Aad et al., 2012; et al., 2012), the most important task of searching for new physics signals remains challenging. High Energy Physics (HEP) events produced at the LHC have the properties of high dimensionality, high complexity, and enormous data size. Deep neural classifiers and generative models have been explored to meet the needs for more effective data selection and physics analysis. Neural net-based unsupervised learning of physics events (Paganini et al., 2018; Kansal et al., 2020; Butter et al., 2019; Touranakou et al., 2022) have been explored in the usual generative modeling methods. In comparison, Variational Autoencoders (VAEs) Kingma & Welling (2014) need a well-designed reconstruction loss, which might be difficult for sophisticated network architectures and complex input features. Generative Adversarial Networks (GANs) Goodfellow et al. (2020) employ separate networks, which need to be carefully tuned, for the generation process. They usually suffer from unstable training and high computation demands.

Topic	Practice
Generative modeling	Parameterized event generation
OOD detection	Model-independent new physics search
Hybrid modeling	Classifier combined with EBMs

Table 1: Application aspects for Energy-based Models for High Energy Physics.

Furthermore, EBMs provide a convenient mechanism to simulate high-order interactions between particles. The energy function can be flexible enough to incorporate sophisticated architectures. Aside from image generation, applications for point cloud data (Xie et al., 2021), graph neural networks (Liu et al., 2021) for molecule generation are also explored. In particle physics, we leverage the self-attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017), to mimic the complex interactions between elementary particles.

As an applicational practice, out-of-distribution (OOD) detection comes naturally in the form of energy comparison. More importantly, EBMs incur fewer spurious correlations in OOD detection. This plays a slightly different role in the context of signal searches at the LHC. There are correlations that are real and useful, but at the same time handicaps effective signal detection. As we will see in Section 4, the EBMs are free from the notorious correlation observed in many anomaly detection methods in HEP, in both the generative and the discriminative approaches.

As summarized in Table 1, we build a framework of physics-inspired EBMs. We construct an energy-based model of the fundamental interactions of elementary particles to simulate the resulting radiation patterns. We especially employ the short-run Markov Chain Monte Carlo for the EBM training, which is improved with an upper-bounded of the Kullback–Leibler divergence correction to the usual Contrastive Divergence objective. The EBMs are able to generate realistic event patterns and can be used as generic anomaly detectors free from spurious correlations.

2 Methods

2.1 Energy Based Models

Energy-based models are constructed to model the un-normalized data probabilities. They leverage the property that any exponential $\exp(-E(\mathbf{x}))$ is non-negative and thus can serve as an un-normalized probability naturally. The data distribution is modelled through the Boltzmann distribution: $p_\theta(\mathbf{x}) = \exp(-E_\theta(\mathbf{x}))/Z(\theta)$ with the energy model $E_\theta(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ mapping $\mathbf{x} \in \mathcal{X}$ to a scalar. And the partition function $Z = \int \tilde{p}(\mathbf{x})d\mathbf{x} = \int \exp(-E_\theta(\mathbf{x}))d\mathbf{x}$ integrates over all the possible states.

EBMs can be learned through maximum likelihood $\mathbb{E}_{p_D}(\log p_\theta(\mathbf{x}))$. However, the training of EBMs can be difficult due to the intractable partition function in $\log p_\theta(\mathbf{x}) = -E_\theta(\mathbf{x}) - \log Z(\theta)$. Though the partition function is intractable, the gradients of the log-likelihood do not involve the partition function directly. Thus when taking gradients w.r.t. the model parameters θ, the partition function is canceled out. The gradient of the maximum likelihood loss function can be written as:

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{p_D(\mathbf{x})}[\nabla_\theta \log p_\theta(\mathbf{x})] \tag{1}$$

$$\simeq \mathbb{E}_{p_D(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x}^+)] - \mathbb{E}_{p_\theta(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x}^-)]\,, \tag{2}$$

where $p_D(\mathbf{x})$ is the data distribution and $p_\theta(\mathbf{x})$ is the model distribution. The training objective is thus composed of two terms, corresponding to two different learning phases (i.e., the *positive phase* to fit the data $\mathbf{x}^+$, and the *negative phase* to fit the model distribution $\mathbf{x}^-$). When parameterizing the energy function with feed-forward neural networks (Ngiam et al., 2011), the positive phase is straightforward. However, the negative phase requires sampling over the model distribution. This leads to various Monte Carlo sampling strategies for estimating the maximum likelihood.

Contrasting the energies of the data and the model samples as proposed *Contrastive Divergence* (CD) (Hinton, 2002) leads to an effective strategy to train EBMs with the following CD objective:

$$D_{\mathrm{KL}}(p_D(\mathbf{x}) \| p_\theta(\mathbf{x})) - D_{\mathrm{KL}}(T p_D(\mathbf{x}) \| p_\theta(\mathbf{x}))\,, \tag{3}$$

where T denotes the one-step Monte Carlo Markov Chain (MCMC) kernel imposed on the data distribution. In more recent approaches for high-dimensional modeling, we can directly initialize from random noises to generate the MCMC samples. More specifically, we employ gradient-based MCMC generation in the training process, which is handled by Langevin Dynamics (Welling & Teh, 2011). As written in Eq. 4, Langevin dynamics uses gradients w.r.t. the data points to generate a sequence of negative samples $\{\mathbf{x}_k^-\}_{k=1}^K$.

$$\mathbf{x}_{k+1}^- = \mathbf{x}_k^- - \frac{\lambda^2}{2}\nabla_{\mathbf{x}} E_\theta(\mathbf{x}_k^-) + \lambda \cdot \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0,1) \tag{4}$$

KL Divergence-Improved EBM Training In the original *Contrastive Divergence*, the precise gradient of the loss function is as follows (Du et al., b):

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{p_D(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x}^+)] - \mathbb{E}_{q_\theta(\mathbf{x})}[\nabla_\theta E_\theta(\mathbf{x}^-)] - \frac{\partial q_\theta(\mathbf{x})}{\partial \theta}\frac{\partial D_{\mathrm{KL}}(q_\theta(\mathbf{x})||p_\theta(\mathbf{x}))}{\partial q_\theta(\mathbf{x})}, \tag{5}$$

where $q_\theta(\mathbf{x})$ denotes the Monte Carlo estimation of the model distribution $p_\theta(\mathbf{x})$. There is then a gap between $p_D(\mathbf{x})$ and $q_\theta(\mathbf{x})$ not taken into account in the usual EBM training process. Thus the full loss (Eq. 6) consists of the usual CD loss term $\mathcal{L}_{\mathrm{CD}}$ (as in Eq. 2) and an extra KL term $\mathcal{L}_{\mathrm{KL}}$ which is ignored in most cases.

$$\mathcal{L} = \mathcal{L}_{\mathrm{CD}} + \mathcal{L}_{\mathrm{KL}}, \text{ with } \mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{q(\mathbf{x})}[E_{\hat{\theta}}(\mathbf{x})] + \mathbb{E}_{q_\theta(\mathbf{x})}[\log(q_\theta(\mathbf{x}))] \tag{6}$$

The KL term is then further decomposed into two terms: the energy of the MCMC samples $\mathbb{E}_{q(\mathbf{x})}[E_{\hat{\theta}}(\mathbf{x})]$ and the entropy of the MCMC samples $-\mathbb{E}_{q_\theta(\mathbf{x})}[\log(q_\theta(\mathbf{x}))]$. While estimating the energy term is relatively straightforward with Langevin dynamics, the entropy term can get involved with non-parametric methods. In this work, we ignore the entropy term since it's always non-negative. Thus we are actually trying to minimize the upper bound of the KL divergence term in our model training.

MC Convergence The training anatomy (Nijkamp et al., 2020; 2019a;b) for short-run non-convergent MCMC and long-run convergent MCMC shows that short-run (5-100 steps) MCMC initialized from random distributions is able to generate realistic samples.

To improve mode coverage, we use random noise to initialize MCMC chains. To accelerate training, we employ a relatively small number of MCMC steps. In practice, we can reuse the generated samples as initial samples of the following MCMC chains to accelerate mixing, similar to *Persistent Contrastive Divergence* (Tieleman, 2008). Following the same procedure in (Du & Mordatch), we use a randomly initialized buffer that is consistently updated from previous MCMC runs as the initial samples. (As empirically shown, a Metropolis-Hastings step is not necessary. So we ignore this rejection update in our experiments.)

Energy Function Since there is no explicit generator in EBMs, we have much freedom in designing the architectures of the energy function. This also connects with the fast-paced development of supervised neural classifiers. We can directly reuse the architectures from supervised classifiers in the generative modeling of EBMs. We use a self-attention-based transformer to parameterize the energy function $E_\theta(\cdot)$. We defer the detailed description to Sec. 3.

The full algorithm for training EBMs is described in Algorithm 1.

2.2 Hybrid Modeling

Neural Classifier as an EBM As shown in (Grathwohl et al.), a classical classifier can be re-interpreted in the framework of EBMs, with the logits $\mathbf{g}(\mathbf{x})$ corresponding to negative energies of the joint distribution $p(\mathbf{x}, y) = \frac{\exp(\mathbf{g}(\mathbf{x})_y)}{Z}$, where $\mathbf{g}(\mathbf{x})_y$ denotes the logit corresponding to label y . Thus the probability marginalized over y can be written as $p(\mathbf{x}) = \frac{\sum_y \exp(\mathbf{g}(\mathbf{x})_y)}{Z}$, with the energy of $\mathbf{x}$ as $-\log \sum_y \exp(\mathbf{g}(\mathbf{x})_y)$.

This viewpoint provides a novel method for jointly training a supervised classifier and an unsupervised generative model. The joint log-likelihood can be decomposed into two terms:

$$\log p(\mathbf{x}, y) = \log p(\mathbf{x}) + \log p(y|\mathbf{x}) \,. \tag{7}$$

Thus we can maximize $\log p(\mathbf{x})$ with the contrastive divergence of the EBM, and maximize $\log p(y|\mathbf{x})$ with the usual cross-entropy of the classification.

Algorithm 1 EBM training with KL-Divergence-Corrected Contrastive Divergence and MCMC by Langevin Dynamics

Input: training samples $\{\mathbf{x}_i^+\}_{i=1}^N$ from $p_\mathrm{D}(\mathbf{x})$, parameterized energy function $E_\theta(\cdot)$, initial buffer $\mathcal{B} \leftarrow \varnothing$, Langevin dynamics step size $\lambda_\mathbf{x}$, number of MCMC steps K, model parameter learning rate λ_θ, regularization strength α

for Gradient descent step l = 0...L-1 **do**

$\mathbf{x}_i^+ \sim p_\mathrm{D}(\mathbf{x})$

$\mathbf{x}_{i,0}^- \sim 0.95 * \mathcal{B} + 0.05 * \mathcal{U}$ ▷ Reinitialize the samples in the buffer with random noise in the probability of 0.05

for Langevin dynamics step k=0...K-1 **do**

$\mathbf{x}_{i,k+1}^- = \mathbf{x}_{i,k}^- - \lambda_\mathbf{x} \nabla_\mathbf{x} E_\theta(\mathbf{x}_{i,k}^-) + 0.005 \cdot \epsilon_k,\ \epsilon_k \sim \mathcal{N}(0,1)$ ▷ Langevin Dynamics taking gradients w.r.t. input dimensions

end for

$\mathbf{x}_i^- \leftarrow \mathbf{x}_{i,K}^-$

$\mathcal{L}_\mathrm{CD} = \frac{1}{N} \sum_i (E_\theta(\mathbf{x}_i^+) - E_\theta(\mathbf{x}_i^-))$

$\mathcal{L}_\mathrm{KL} = E_{\hat{\theta}}(\mathbf{x}_i^-)$ (+ ...) ▷ $\hat{\theta}$ denotes stopping gradient for back-propagating in the energy function parameters

$\mathcal{L}_\mathrm{reg} = \frac{1}{N} \sum_i (E_\theta(\mathbf{x}_i^+)^2 + E_\theta(\mathbf{x}_i^-))^2$ ▷ L_2 Regularization

$\theta \leftarrow \theta - \lambda_\theta \nabla_\theta (\mathcal{L}_\mathrm{CD} + \mathcal{L}_\mathrm{KL} + \alpha \mathcal{L}_\mathrm{reg})$ ▷ Update model parameters with gradient descent

$\mathcal{B} \leftarrow \mathbf{x}_{i,K}^- \cup \mathcal{B}$ ▷ Update the buffer with generated samples

end for

3 Problem Statement

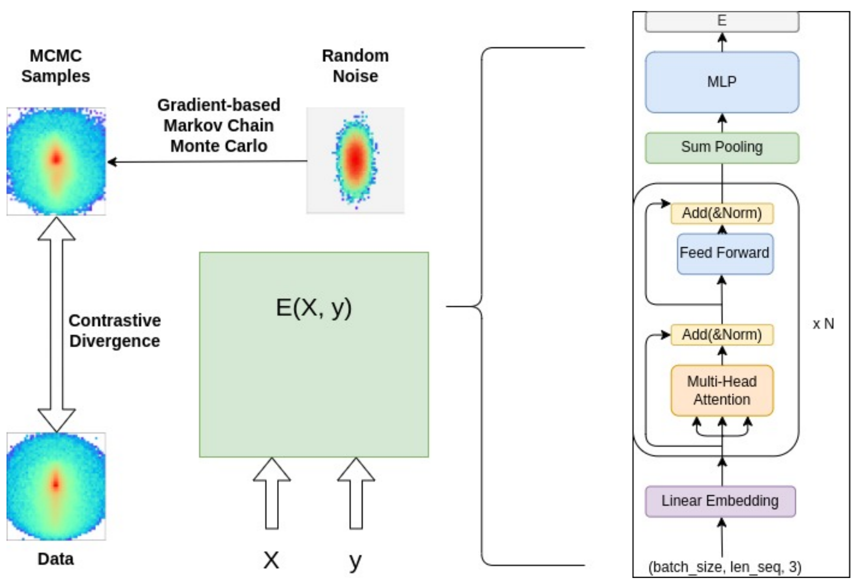

Figure 1: **Left:** EBM model training schematic. **Right:** Energy function estimated with a transformer.

Physics events produced at the LHC leave energy deposits in the detectors. Along with the spatial coordinates, we can precisely identify the collision products. The substructures of these particle traces and energy deposits manifest the underlying physics and the corresponding high-energetic elementary particles.

Describing HEP Events Most particle interactions happening at the LHC are governed by Quantum chromodynamics (QCD), due to the hadronic nature of the proton-proton collision. Thus jets are enormously produced by these interactions. A jet is formed by collimated radiations originating from highly-energetic elementary particles (e.g., quarks, gluons, and sometimes highly-boosted electroweak bosons). The tracks and energy deposits left in the particle detectors reveal the underlying physics in complex patterns. These patterns have been used to identify different types of particles, assisting in more precise data analysis and signal detection. Identifying, classifying, and sometimes reconstructing these elementary particles manifest in raw data is critical for ongoing physics analysis at the LHC. Deep neural nets are the perfect candidates for encoding this low-level information and modeling the high-dimensional distributions of the constituents within a jet.

Specifically, each particle within a jet has $(\log p_T, \eta, \phi)_i$ as the descriptive coordinates in the detector's reference frame, with p_T denoting the transverse momentum perpendicular to the beam axis, and (η, ϕ) is the spatial coordinates within the cylindrical detector. More details about the datasets can be found in Appendix A.

Energy-based Models for Elementary Particles We would like to construct an Energy-based Model for describing jets and their inner structures. In conceiving the energy function for these elementary particles, we consider the following constraints and characteristics: 1) permutation invariance – the energy function should be invariant to jet constituents permutations, and 2) high-order interactions – we would like to energy function to be powerful enough to simulate the complex inter-particle interactions.

Thus, we leverage the self-attention-based transformer (Vaswani et al., 2017) to approximate the energy function, which takes into account the *higher-order* interactions between the component particles. As indicated in Eq. 8b, the encoding vector of each constituent is connected with all other constituents through the self-attention weights in Eq. 8a.

$$A = \text{softmax}(Q \cdot K^T / \sqrt{d_{\text{model}}}) \tag{8a}$$

$$W = A \cdot V \tag{8b}$$

Moreover, we can easily incorporate particle permutation invariance (Zaheer et al., 2017) in the transformer, by simply summing up the encodings of each jet constituent. The architecture is shown in Fig. 1. The coordinates $(\log p_T, \eta, \phi)_i$ are first embedded through a linear layer, then fed into N self-attention blocks sequentially. After that, a sum-pooling layer is used to sum up the features of the jet constituents. Finally, a multi-layer-perceptron projector maps the features into the energy score. Model parameters are recorded in Table 3 of Appendix A.

Model Validation In monitoring the likelihood, the partition function can be estimated with Annealed Importance Sampling (AIS) (Neal, 2001). However, these estimates can be erroneous and consume a lot of computing resources. Fortunately for physics events, we have well-designed high-level features as a handle for monitoring the generation quality. Especially, we employ Lorentz-invariants jet transverse momentum p_T and jet mass M as the validation observables. And we calculate the Jensen–Shannon divergence, between these high-level observable distributions of the data and the model generation, as the metric. In contrast to the short-run MCMC in the training steps, we instead use longer MCMC chains for generating the validation samples.

When we focus on downstream tasks such as OOD detection, it's reasonable to employ the (i.e., Standard Model top jets as the benchmark) Area Under the ROC Curve (AUC) as the validation metric.

4 Experiments

Training Details We employ the KL-improved training of EBMs. To speed up the training, we ignore the entropy term and instead back-propagate the gradients through all the Langevin dynamics steps for the $\mathcal{L}_{\text{KL}}$ term. We have 10,000 samples in the buffer and reinitialize the random samples with a probability of 0.05 in each iteration. The training set consists of 50,000 QCD jets. To fit in the GPU memory, we use a relatively small number of steps (e.g., 24) for the MCMC chains, since we back-propagate through the full MCMC chains for estimating the KL divergence term in Eq. 6. The step size $\lambda_{\mathbf{x}}$ is set to 0.1 according to standard deviations of the input features. The noise magnitude ϵ within the Langevin dynamics is set to 0.005. The number of steps used in validation steps is set to 128 for better mixing.

We use Adam (Kingma & Ba, 2015) for optimization, with the momenta $\beta_1 = 0.0$ and $\beta_2 = 0.999$. The initial learning rate is set to 1e-4, with a decay rate of 0.98 for each epoch. We use a batch size of 128, and train the model for 50 epochs. More details can be found in Appendix A.

Generation Test-time generation is achieved in MCMC transition steps from the proposal random (Gaussian) distribution. We use a smaller step size of 0.05 to ensure stable generation. And more steps (e.g., 200) are taken to achieve realistic generation.

And as a common finding for different methods considered, the step size is the most important parameter that predominantly determines the generation quality.

OOD Detection for New Physics Searches Despite the great efforts in searching for new physics signals at the LHC, there has been no hint of beyond-Standard-Model physics. Given the large amount of data produced at the LHC, it has been increasingly challenging to cover all the search channels. We thus shift to model-independent searches which are data-oriented rather than theory-guided.

EBM naturally has the handle for discriminating between in-distribution and out-of-distribution examples. While in-distribution data points are trained to have lower energy, energies of OOD examples are pushed up in the learning process. This property has been used for OOD detection in computer vision (Du & Mordatch). This indicates great potential for EBM-based new physics detection at the LHC.

Compared with the common practice in computer vision, there is specificity in EBM-based OoD detection for HEP. The simple approach comparing two different datasets such as CIFAR10 and SVHN ignores the complex real-world applicational environments. Adapting OOD detection to scientific discovery at the LHC, we reformulate and tailor the decision process as follows: if we train on Standard Model datasets, we focus on class-conditional model evaluation for discriminating between the unseen signals and the most copious background events (i.e., QCD jets rather than all the Standard Model jets).

4.1 Generative Modeling – Energy-based Event Generator

We present the generated jets transformed from initial random noises with the Langevin dynamics MCMC. Due to the non-human-readable nature of physics events (e.g., low-level raw records at the particle detectors), we are not able to examine the generation quality through formats such as images directly. However, it has a long history that expert-designed high-level observables can serve as strong discriminating features. In the first row of Fig. 2, we first show the distributions of input features for the data and the model generation. Meanwhile, in the second row, we plot the distributions of high-level expert observables including the jet transverse momentum p_{T} and the jet mass M. Through modeling low-level features in the detector space, we achieve precise recovery of the high-level physics observables in the theoretical framework. For better visualization, we map the jets onto the (η, ϕ) plane, with pixel intensities associated with the corresponding energy deposits. We show the average generated jet images in Fig. 3, comparing to the real jet images (*Right*) in the (η, ϕ) plane.

At the Large Hadron Collider, event simulation serves as an important handle for background estimation and data analysis. For many years, physics event simulators (Campbell et al., 2022) are build on Monte Carlo methods based on physics rules. These generators are slow and need to be tuned to the data frequently. Deep neural networks provide us with an efficient parameterized generative approach to event simulation for the coming decades.

4.2 Anomaly Detection – Anomalous Jet Tagging

Since EBMs naturally provide an energy score for each jet, for which the in-distribution samples should have lower scores while OOD samples are expected to incur higher energies. Furthermore, a classifier, when interpreted as an energy-based model, the transformed energy score can also serve as an OOD identifier (Grathwohl et al.; Liu et al., 2020).

EBM In HEP, the *in-situ* energy score can be used to identify potential new physics signals. Experiments at the LHC over the past decades have been focused on model-oriented searches, such as searching for the Higgs boson (Englert & Brout, 1964; Higgs, 1964). The null results up to now from

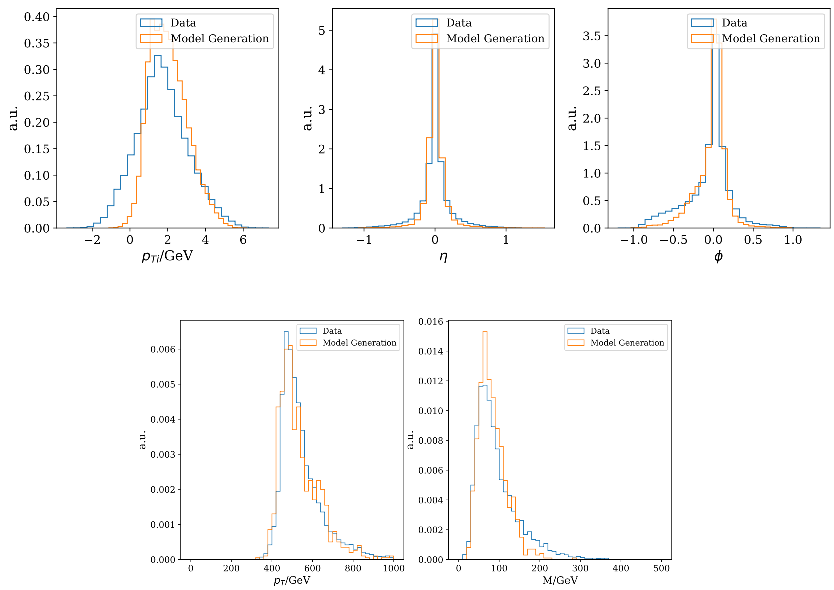

Figure 2: **Top:** Input feature distributions of jet constituents for the data and the model generation. **Bottom:** High-level feature distributions for the data and the model generation.

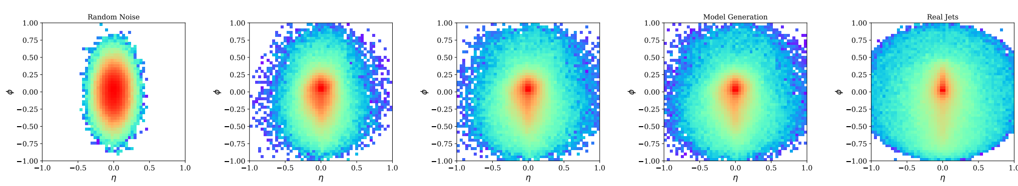

Figure 3: Jet images averaged over 10000 jet samples. **Left:** Random noises. **Middle:** EBM-generated jet samples by the MCMC chains in intervals. **Right:** Real jets.

model-driven searches call for novel solutions. Model-independent and data-driven search strategies are thus under investigation.

By reducing the main QCD background events, generic anti-QCD jet taggers facilitate effective model-independent searches for new physics signals. Thus with an energy-based model, which is trained on the QCD background events or directly on the slightly signal-contaminated data, we expect unseen signals have higher energies, correspondingly lower likelihoods.

In Fig. 4, we compare the energy distributions of in-distribution QCD samples, out-of-distribution signal examples (hypothesized Heavy Higgs boson which decays into four QCD sub-jets), and random samples drawn from the Gaussian distribution. We observe that random samples unusually have the highest energies. Signal jets have relatively higher energies compared with the QCD background jets, making model-independent new physics searches possible.

A more intriguing property of EBMs is that spurious correlations can be better handled. Spurious correlations in jet tagging might result in distorted background distributions and obscure effective signal detection. For instance, VAEs in OOD detection can be highly correlated with the input particle numbers (Cheng et al., 2020), similar to the spurious correlation with image pixel numbers in image recognition (Nalisnick et al., 2019; Ren et al., 2019). In the right panel of Fig. 4, we plot the correlation between energy scores and jet masses. Unlike other generative strategies for model-independent anomaly detection, EBMs are largely free from the spurious correlation between the energy $E(\mathbf{x})$ and the jet mass M. This makes EBMs a promising candidate for model-independent new physics search.

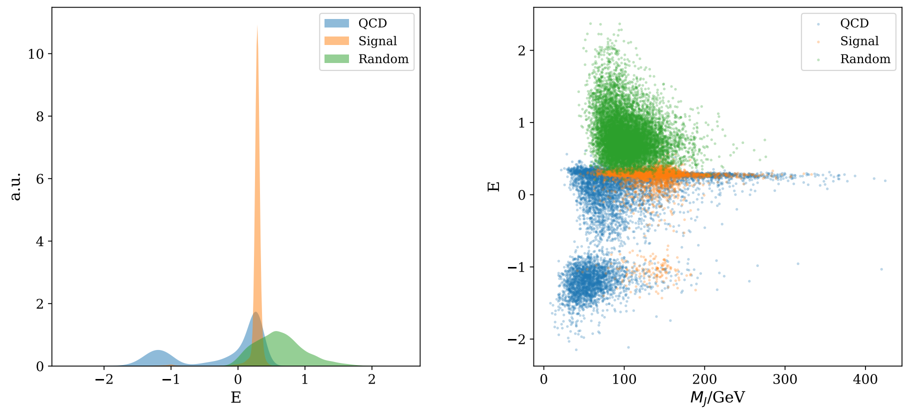

Figure 4: **Left:** Energy distributions for random samples, background QCD jets, and novel signals. **Right:** Correlation between the jet mass M_J and the energy E.

EBM-CLF To better suit the goal of OOD detection, we employ the hybrid learning scheme (Ngiam et al., 2011; Grathwohl et al.) combining the discriminate and the generative approaches. The jointly trained jet classifier and EBM (EBM-CLF) according to Eq. 7 maintain the classification accuracy. The associated EBM is thus augmented by the discriminative task, and thus assisted with better inductive biases.

We train a multi-class classifier for discriminating different Standard Model jets (QCD jets, boost W jets, and boost top jets), along with the associated EBM. The resulting generative sampling results are shown in Fig. 5. We measure the OOD detection performance in AUCs in the binary classification of QCD background samples and the signal jets. Table 2 records the AUCs for tagging Standard Model Top jets and hypothesized Higgs boson by different models. THe jointly trained model has even better anomaly tagging performance compared with the naive EBM. Corresponding ROC curves are shown in the left panel of Fig. 6, in terms of the signal efficiency ϵ_S and the background rejection rate $1/\epsilon_B$. In the right panel, we plot the background mass distributions under different cuts on the energy scores. We observe excellent jet mass decorrelation/invariance for energy score-based anomalous jet tagging.

We also record the AUCs for the class-conditional softmax probability-based jet tagging in Table 2. We employ the $p(y|\mathbf{x})$ corresponding to the QCD class as the anomaly score. However, without further decorrelation strategies, this anomaly score is usually strongly correlated with the masses of the in-distribution classes and distorts the background distributions. Thus we list the results here only for reference.

Model	AUC (Top)	AUC(OOD H)
EBM ($E(\mathbf{x})$)	0.681	0.782
EBM-CLF ($E(\mathbf{x})$)	0.711	0.817
EBM-CLF ($p(y\|\mathbf{x})$)	0.929	0.870

Table 2: Anomaly detection performance measured in AUCs.

5 Conclusion

We present a versatile generative framework for modeling the behavior of elementary particles. By mimicking the inter-particle interactions with a self-attention-based transformer, we map the correlations in the detector space to a probabilistic space with an energy function. The energy model is used for the implicit generation of physics events. Despite the difficulty in training EBMs, we employ adapted training strategies to balance learning efficiency and training stability. This

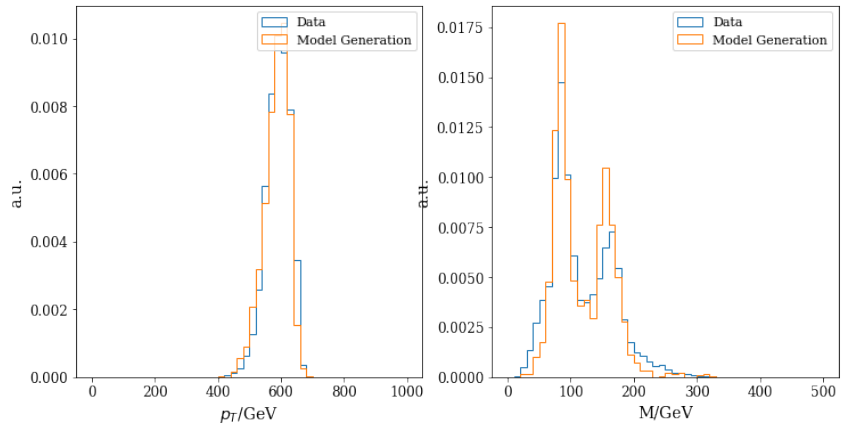

Figure 5: High-level observables for the generated samples from EBM-CLF.

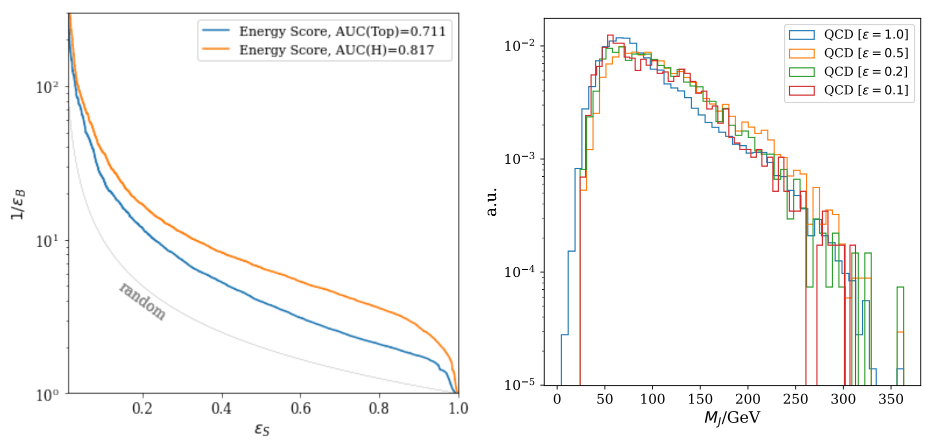

Figure 6: **Left:** ROC Curves for the EBM-CLF with the energy $E(\mathbf{x})$ as the anomaly score. The grey line denotes the case of random guessing. **Right:** Background mass distributions under different acceptance rates ϵ after cutting on the energy score from the EBM-CLF.

framework thus provides us with flexible tools for parameterized physics event simulation and spurious-correlation-free model-independent signal detection.

References

LHC Machine. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001.

Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(85)80012-4. URL `https://www.sciencedirect.com/science/article/pii/S0364021385800124`.

Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, and Tim Stelzer. Madgraph 5: going beyond. *Journal of High Energy Physics*, 2011(6), Jun 2011. ISSN 1029-8479. doi: 10.1007/jhep06(2011)128. URL `http://dx.doi.org/10.1007/JHEP06(2011)128`.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Anja Butter, Tilman Plehn, and Ramon Winterhalder. How to GAN LHC Events. *SciPost Phys.*, 7(6):075, 2019. doi: 10.21468/SciPostPhys.7.6.075.

Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008. ISSN 1029-8479. doi: 10.1088/1126-6708/2008/04/063. URL `http://dx.doi.org/10.1088/1126-6708/2008/04/063`.

J. M. Campbell et al. Event Generators for High-Energy Physics Experiments. In *2022 Snowmass Summer Study*, 3 2022.

Taoli Cheng, Jean-François Arguin, Julien Leissner-Martin, Jacinthe Pilette, and Tobias Golling. Variational autoencoders for anomalous jet tagging, 2020. URL `https://arxiv.org/abs/2007.01850`.

J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. Delphes 3: a modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics*, 2014(2), Feb 2014. ISSN 1029-8479. doi: 10.1007/jhep02(2014)057. URL `http://dx.doi.org/10.1007/JHEP02(2014)057`.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=B1l4SgHKDH`.

Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. URL `https://arxiv.org/abs/1903.08689v6`.

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation and inference with energy based models. a. doi: 10.48550/arXiv.2004.06030. URL `https://arxiv.org/abs/2004.06030v3`.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. b. URL `http://arxiv.org/abs/2012.01316`.

Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1e_9xrFvS`.

Yilun Du, Shuang Li, Yash Sharma, Joshua B. Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. In *NeurIPS*, 2021.

F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964. doi: 10.1103/PhysRevLett.13.321. URL `https://link.aps.org/doi/10.1103/PhysRevLett.13.321`.

S. Chatrchyan et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2012.08.021. URL `https://www.sciencedirect.com/science/article/pii/S0370269312008581`.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. doi: 10.48550/arXiv.1912.03263. URL `https://arxiv.org/abs/1912.03263v3`.

Will Grathwohl, Jacob Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Kristjanson Duvenaud. No mcmc for me: Amortized sampling for fast and stable training of energy-based models. *ArXiv*, abs/2010.04230, 2021.

Peter W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys. Lett.*, 12:132–133, 1964. doi: 10.1016/0031-9163(64)91136-9.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8): 1771–1800, aug 2002. ISSN 0899-7667. doi: 10.1162/089976602760128018. URL `https://doi.org/10.1162/089976602760128018`.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL `https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554`.

Raghav Kansal, Javier Duarte, Breno Orzari, Thiago Tomei, Maurizio Pierini, Mary Touranakou, Jean-Roch Vlimant, and Dimitrios Gunopulos. Graph Generative Adversarial Networks for Sparse Data Generation in High Energy Physics. In *34th Conference on Neural Information Processing Systems*, 11 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Yann LeCun, Sumit Chopra, Raia Hadsell, Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. 2006.

M. Liu, Keqiang Yan, Bora Oztekin, and Shuiwang Ji. Graphebm: Molecular graph generation with energy-based models. *ArXiv*, abs/2102.00546, 2021.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *ArXiv*, abs/1810.09136, 2019.

Subhajit Naskar, Pedram Rooshenas, Simeng Sun, Mohit Iyyer, and Andrew McCallum. Energy-based reranking: Improving neural machine translation using energy-based models. In *ACL*, 2021.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

Jiquan Ngiam, Zhenghao Chen, Pang Wei Koh, and Andrew Y. Ng. Learning deep energy models. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1105–1112, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. In *NeurIPS*, 2019a.

Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. On learning non-convergent short-run mcmc toward energy-based model. *ArXiv*, abs/1904.09770, 2019b.

Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *AAAI*, 2020.

Frank Noé, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365, 2019.

Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. *Phys. Rev. Lett.*, 120(4): 042003, 2018. doi: 10.1103/PhysRevLett.120.042003.

J. Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *ArXiv*, abs/1906.02845, 2019.

Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to pythia 8.1. *Computer Physics Communications*, 178(11):852–867, Jun 2008. ISSN 0010-4655. doi: 10.1016/j.cpc.2008.01.036. URL http://dx.doi.org/10.1016/j.cpc.2008.01.036.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1064–1071, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390290. URL https://doi.org/10.1145/1390156.1390290.

Mary Touranakou, Nadezda Chernyavskaya, Javier Duarte, Dimitrios Gunopulos, Raghav Kansal, Breno Orzari, Maurizio Pierini, Thiago Tomei, and Jean-Roch Vlimant. Particle-based fast jet simulation at the lhc with variational autoencoders, 2022. URL https://arxiv.org/abs/2203.00520.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pp. 2635–2644. PMLR, 2016.

Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative pointnet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14971–14980, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alex Smola. Deep sets. In *NIPS*, 2017.

A EXPERIMENTAL DETAILS

Datasets For the simple EBM, we train on 50,000 simulated QCD jets. For the hybrid model EBM-CLF, we train on 300,000 simulated Standard Model jets (QCD jets, boosted jets originating from the W boson, and boosted jets originating from the top quark). For OOD detection test sets, we employ the hypothetical Higgs boson with a mass of 174 GeV, which decays into two lighter Higgs bosons of 80 GeV. The intermediate light Higgs boson decays into two b quarks. All the jet samples are generated with a pipeline of physics simulators.

Event Generation QCD jets are extracted from QCD di-jet events that are generated with MadGraph (Alwall et al., 2011) for LHC 13 TeV, followed by Pythia8 (Sjöstrand et al., 2008) and Delphes (de Favereau et al., 2014) for Parton shower and fast detector simulation. All jets are clustered using the anti-k_T algorithm (Cacciari et al., 2008) with cone size $R = 0.8$ and the selection cut in the jet transverse momentum $p_T \in [550, 650]$ GeV. We use the particle flow objects for jet clustering.

Input Preprocessing Jets are preprocessed before being fed into the neural models. Jets are longitudinally boosted and centered at $(0, 0)$ in the (η, ϕ) plane. The centered jets are then rotated so that the jet principal axis $(\sum_i \frac{\eta_i E_i}{R_i}, \sum_i \frac{\phi_i E_i}{R_i})$ (with $R_i = \sqrt{\eta_i^2 + \phi_i^2}$) and E_i is the constituent energy) is vertically aligned on the (η, ϕ) plane.

Hyper-parameters Hyper-parameters are recorded in Table 3.

Data	
input features	$\{(\log(p_T), \eta, \phi)_i\}_{i=1}^N$
input length	N=40
Energy Function	
Number of layers	8
Model dimension	128
Number of heads	16
Feed-forward dimension	1024
Dropout rate	0.1
Normalization	`None`
MCMC	
Number of steps	24
Step size	0.1
Buffer size	10000
Resample rate	0.05
Noise	$\epsilon = 0.005$
Regularization	
L2 Regularization	0.1
Training	
Optimizer	Adam ($\beta_1 = 0.0, \beta_2 = 0.999$)
Learning rate	1e-4 (decay rate $\gamma = 0.98$)

Table 3: Model settings.

B ADDITIONAL RESULTS

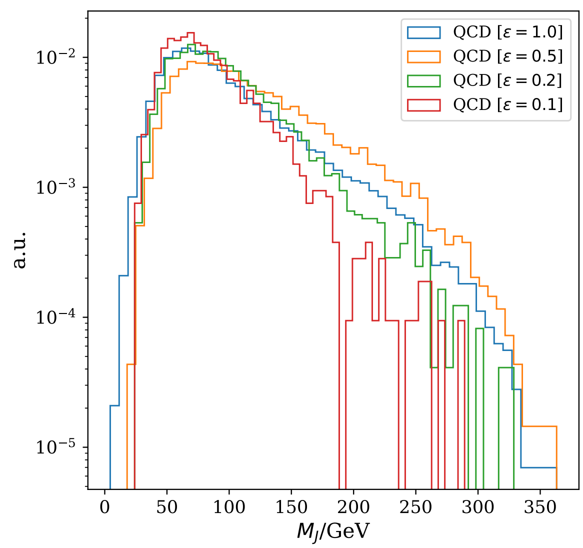

Figure 7: Background mass distributions under different acceptance rates ϵ after cutting on the energy score from the EBM.