A Experimental Details

Datasets For the standard EBM, we train on 300,000 simulated QCD jets. For the hybrid model EBM-CLF, we train on 300,000 simulated Standard Model jets (100,000 QCD jets, 100,000 boosted jets originating from the W boson, and 100,000 boosted jets originating from the top quark). For OOD detection test sets, we employ the hypothetical Higgs boson (in the decay mode of $H \to hh \to (b\bar{b})(b\bar{b})$) 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. Each test set consists of 10,000 samples that are p_T-refined to the range of [550, 650] GeV. 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 [4] for LHC 13 TeV, followed by Pythia8 [61] and Delphes [18] for parton shower and fast detector simulation. All jets are clustered using the anti-k_T algorithm [9] with cone size $R = 1.0$ and a selection cut in the jet transverse momentum $p_T > 450$ 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 4. Hyper-parameters for the transformer are chosen according to a jet classification task. We scan over the following hyper-parameter ranges:

`step size` $\in$ {0.01, 0.1, 1.0, 10}
`steps` $\in$ {5, 10, 24, 60}
`lr` $\in$ {1e-3, 1e-4}

We found that fewer steps {5, 10} make training unstable. Thus the model prefers a relatively larger number of steps and a smaller step size. However, 60 steps MCMC takes much longer time to train and no significant improvement was observed for even longer chains. We balance training efficiency and effective learning by choosing 24 steps.

The MCMC chains are initialized with Gaussian noises, where the constituent features are sampled from the following distributions: $\log p_{Ti} \sim \mathcal{N}(2, 1)$, $\eta_i \sim \mathcal{N}(0, 0.1^2)$, and $\phi_i \sim \mathcal{N}(0, 0.2^2)$.

Data	
input features	$\{(\log(p_T), \eta, \phi)_i\}_{i=1}^N$
input length	N=40 with zero-padding
Energy Function	
Number of layers N_L	8
Model dimension d_{model}	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	
Batch size	128
Optimizer	Adam ($\beta_1 = 0.0, \beta_2 = 0.999$)
Learning rate	1e-4 (decay rate $\gamma = 0.98$)

Table 4: Model settings.

B Additional Results

B.1 Ablation Study

We explore the most crucial aspects of the model to test their functionality:

- With an energy function approximated with a Multi-Layer-Perceptron (MLP) net, we were not able to achieve quality generation.
- We also tried out the Hamiltonian Monte Carlo (HMC) [52] for the MCMC procedure. However, we were not able to achieve good performance in these experiments.

Results measured in the Jensen-Shannon Divergence of the high-level observables (p_T and M) are recorded in Table 5.

Ablation	JSD (p_T) / 10^{-4}	JSD (M) / 10^{-4}
	Energy Function	
MLP	Fig. 5	Fig. 5
	MCMC Dynamics	
HMC	24	30

Table 5: Ablation study on different components of the model, training strategies, and training techniques. Since the MLP-based model is not able to produce high-quality samples, we instead show the observable distributions (Fig. 5) to visually show the failure patterns.

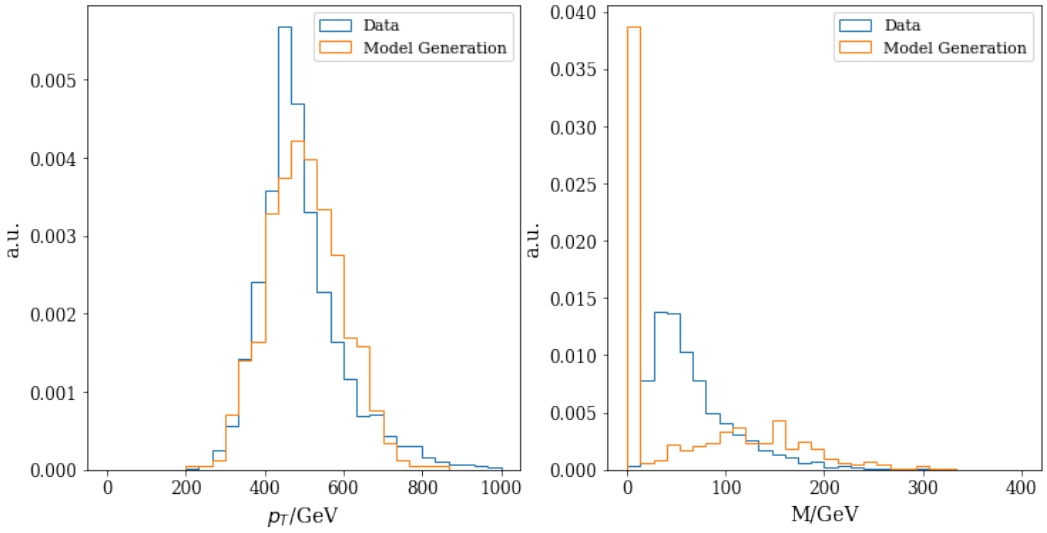

Figure 5: Typical high-level observable distributions for MLP-based models.

B.2 Classification Performance of EBM-CLF

The classification accuracies for EBM-CLF (QCD/W/Top 3-way classification) are recorded in Table 6. EBM-CLF performs on par with the fully supervised classifier ParticleNet [59], while EBM-CLF is trained on a much smaller dataset. The corresponding confusion matrices are displayed in Fig. 6. When we empirically down-weight the term $\log p(y|\mathbf{x})$ (decrease κ in Eq. 7), the classification performance drops correspondingly.

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

B.3 Additional Plots

In Fig. 7, we show the generated jet samples displayed in images on the (η, ϕ) plane. In Fig. 8, we show the high-level observable distributions of generated jet samples of EBM-CLF. In Fig. 9, we show the background mass distributions under different acceptance rates ϵ after cutting on the energy score of the standard EBM.

Model	Top-1 Accuracy	Top-2 Accuracy
ParticleNet[59]	0.871	0.976
EBM-CLF ($\kappa = 1.0$)	0.850	0.969
EBM-CLF ($\kappa = 0.5$)	0.708	0.906
EBM-CLF ($\kappa = 0.1$)	0.679	0.852

Table 6: Classification performance of EBM-CLF on QCD/W/Top 3-way classification. κ denotes the weight of the discriminative log-likelihood.

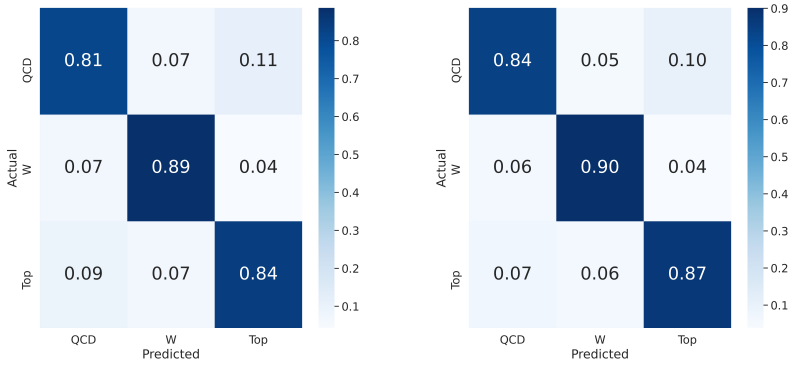

Figure 6: Confusion matrices for EBM-CLF(*left*) and ParticleNet(*right*).

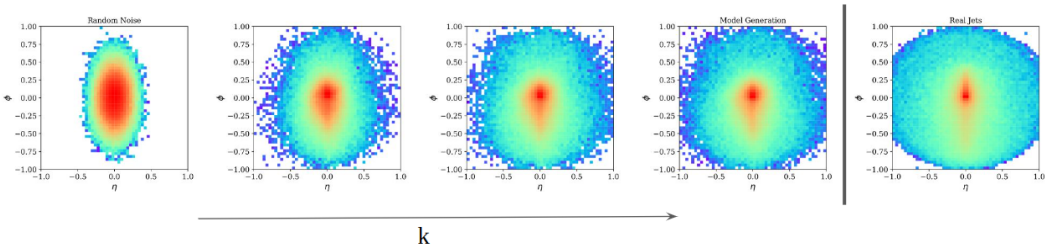

Figure 7: Jet images averaged over 10000 jet samples. From left to right, we show the initial random noises (*left-most*), EBM-generated jet samples by the MCMC chains in intervals (*middle*), and Real jets (*right-most*).

C Reproducibility Statement

To ensure reproducibility and encourage open-source practice in the HEP community, we release the code implementation in `https://github.com/taolicheng/EBM-HEP`. The training sets and test sets are accessible at [47, 14]. Due to difficulties in aligning model comparison protocols for different research groups, we thus focus on methods with code publicly available, that serve as credible baselines, for model comparison.

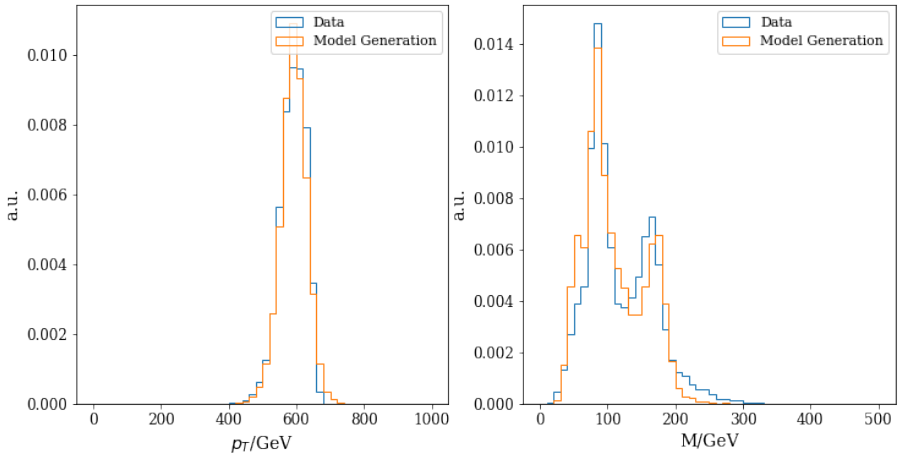

Figure 8: High-level observable distributions for the generated samples of EBM-CLF (*orange*) and the data (*blue*).

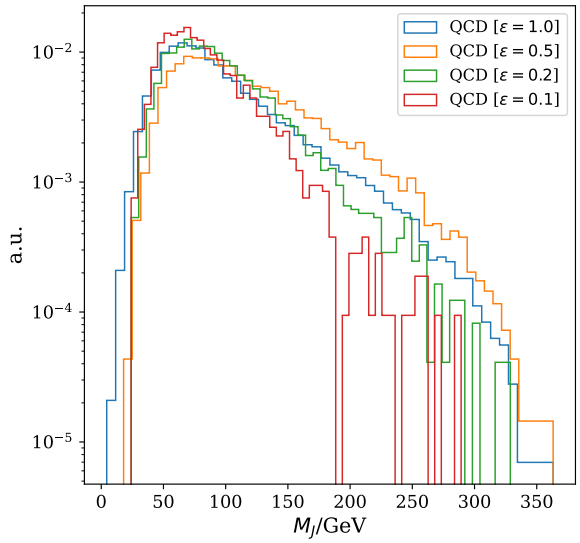

Figure 9: Background mass distributions under different acceptance rates ϵ after cutting on the energy score from the EBM.