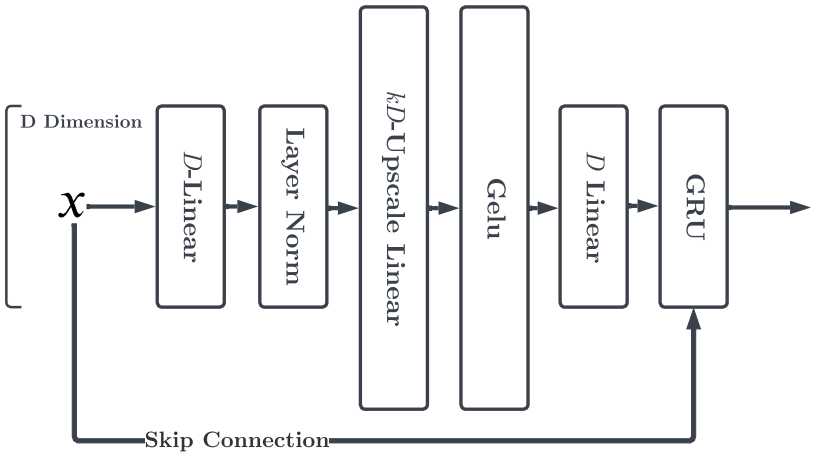


Figure 5: Block diagram from the gated inverse-bottleneck feed forward block used for the networks.

A Network Architecture

A.1 Feed-Forward Block

A custom feed-forward block, derived from the successful ConvNeXt [55] and transformer models [59], is the foundation of the networks in this study. This block features an inverted bottleneck formed by three linear layers, a GELU activation [60], layer normalization [61] for regularization, and a gated residual connection inspired by GTRXL [62]. The block uses a GRU recurrent layer instead of a traditional skip connection, where the regular output is the input to the GRU and the skip-connection value is the hidden input. A complete block diagram of the feed-forward block is presented in Figure 5.

A.2 Detector Encoder

The detector encoder network, as defined in SPANet [63, 54], processes variable-length sets of detector observations into a fixed-size D-dimensional vector. It uses the gated transformer architecture from SPANet version 2, with N_S transformer encoder blocks. Instead of utilizing the tensor attention layers, an event-level representation of the detector observations is extracted from the central transformer encoder.

A.3 Parton Encoder

The encoder starts with an embedding layer, transforming the fixed-size 55 dimensional parton representation into a D-dimensional vector via a linear layer. The encoder's body comprises N_E feed-forward blocks arranged in series. The input can be the embedded D-dimensional parton vector, or its concatenation with the D-dimensional encoded detector data. Two independent networks, each accepting identical input and sharing the same block structure, predict the mean, $\mu_\theta(x; z_y)$, and the log standard deviation, $\log \sigma_\theta(x; z_y)$, respectively. Normalizing and then scaling the mean also helped prevent the encoder from learning very small-valued components: $\frac{\sqrt{D}}{\|\mu_\theta(x;z_y)\|}\mu_\theta(x; z_y)$

A.4 Parton Decoder

The decoder retains the encoder's linear block structure. As a deterministic decoder, it comprises a single stack of N_D feed-forward blocks and a concluding linear layer mapping D dimensions back to 55. Conditional decoders also append the D-dimensional detector vector to the input before processing.

A.5 Denoising Network

The denoising network, employing the same linear block structure, consists of N_ϵ feed-forward blocks. It maps the D-dimensional latent sample z_t to the approximate noise that produced it, $\epsilon_\theta(z_t, t, z_y)$. The time, $t \in [0, 1]$, is encoded with a 32-dimensional sinusoidal position encoding as detailed in [59]. The latent vector, position encoding, and conditioning are concatenated and fed through the feed-forward network to produce the noise estimate.

B Hyperparameters

We present a full table of hyperparameters used throughout the experiments. All models use the same set of parameters, ignoring any that do not apply to particle methods. Parameters were not tuned using rigorous search. The detector transformer parameters were extract from experiments presented in SPANet [54], and the other networks where tuned to contain a similar number of parameters as the detector encoder.

Parameter	**Value**
Latent Dimensionality (D)	96
Attention Heads	4
Inverse Bottleneck Expansion (k)	2
Detector Transformer Encoder Layers (N_S)	8
Parton Encoder Blocks (N_E)	6
Parton Decoder Blocks (N_D)	6
Denoising Network Blocks (N_ϵ)	10
Primary Learning Rate	$5 \cdot 10^{-4}$
Fine-tuning Learning Rate	$1 \cdot 10^{-4}$
L_2 Gradient Clipping Limit	1.0
Consistency Loss Scale (λ_C)	0.1
Batch Size (Per GPU)	4096

Table 2: Table of complete hyperparameters used for training all generative models

C Distance Metrics

As the parton global distributions do not have a known family of distributions to describe their components, model-free measures of distribution distance must be used to evaluate the models. Three different families of distance measures are used. These non-parametric distances are only defined for 1-dimensional distributions. As there is no commonly accepted way of measuring distance for N-dimensional distributions, the 1-dimensional distances are simply summed across the components. Although not ideal, it is enough to compare different models and rank them based on performance.

C.1 Wasserstein Distance

The Wasserstein distance, often referred to as the earth-mover distance, quantifies the amount of work it takes to reshape one distribution into another. This concept originated from the field of optimal transport and has found wide applications in many areas, including machine learning. An equivalent definition defines this distance as the minimum cost to move and transform the mass of one distribution to match another distribution. For a pair of 1-dimensional distribution samples, denoted u and v, the Wasserstein distance can be computed in a bin-independent manner. This is achieved by computing the integral of the absolute difference between their empirical cumulative distribution functions (CDFs), $U(x)$ and $V(x)$.

$$D_{\text{Wasserstein}}(u, v) = \int_{-\infty}^{\infty} |U(x) - V(x)| dx$$

C.2 Energy Distance

Energy distance is another statistical measure used to quantify the difference between two probability distributions based on empericial CDFs. It compares the expected distance between random variables

drawn from the same distribution (intra-distribution) with the expected distance between random variables drawn from different distributions (inter-distribution). The Energy distance may be defined as the squared variant of the Wasserstein distance.

$$D_{\text{Energy}}(u, v) = \sqrt{2 \int_{-\infty}^{\infty} (U(x) - V(x))^2 \, dx}$$

C.3 Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (K-S) test is a non-parametric statistical hypothesis test used to compare the underlying probability distributions of two independent samples. It is particularly useful in machine learning applications where the goal is to assess whether two datasets come from the same distribution or if they differ significantly, without making any assumptions about the underlying distribution shape. It is also based on empirical CDFs.

C.4 KL-Divergence

An alternative approach to empirical CDF approaches is to bin the data into histograms and compute discrete distribution distances from these histograms. The common Kullback–Leibler distance is used with three different bin sizes. After finding the histograms with N bins for $1 \leq i \leq N$, $P_N(i)$ and $Q_N(i)$, the discrete KL divergence is computed as

$$D_{KL,N} = \sum_{i=1}^{N} P_N(i) \log \left(\frac{P_N(i)}{Q_N(i)} \right)$$

D Particle Distance Tables

Tables 3 to 6 present the distance metrics for each parton and model. The general trends in Table 1 remain generally consistent across partons. The neutrino reconstruction can prove difficult for Latent diffusion models, likely due to its very peaked components.

E Component Distance Tables

Tables 7 to 10 present the distance metrics for each component and model. The mass component seems to vary the most for for many of the distance functions, indicating that many models struggle reconstructing the peaked mass distributions. However, the overall results remain consistent with Table 1. We again see the clear benefit of both latent diffusion and end-to-end training.

F Global Distribution Plots

Figures 6 through 16 present a collection of global distributions for the three primary classes of generative models for every particle and component. The proposed method (VLD) closely matches the truth distributions across all components, including the mass which is slightly smoothed but peaks in the correct location. Baseline models struggle with capturing the peaks and shapes of the distributions.

G Posterior Distribution Plots

Figures 17 and 18 present a collection of posterior distributions for four testing events, along with several models and the brute-force empirical approach.

H Loss Function Derivation

We provide a derivation for the loss function presented in Equation 8, adhering to the generative model displayed in Figure 2. Here, data is generated from a latent representation, $p(x|z_x)$; the latent

data from the diffusion end-point, $p(z_x|z_0)$; a diffusion process to generate the sampled configuration $p(z_0|z_1)$; and a simple standard normal distribution prior at the base of the hierarchy $p(z_1)$. All these functions are further conditioned on the (embedded) detector observations z_y. We start with the full conditional hierarchical VAE loss function over L variational layers as per [64].

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(z_x|x,z_y)}\left[-\log p(x|z_x, z_y)\right] + \sum_{l=1}^{L} D_{KL}(q(z_l|x, z_y, z_{i\neq l}) \parallel p(z_l|z_{<l})) \tag{11}$$

Next, we substitute the layers we defined for the sum to expand the expression for our generative model. For this step, we employ a three-stage hierarchy with the following substitutions: $z_1 \leftarrow z_1$, $z_2 \leftarrow z_0$, and $z_3 \leftarrow z_x$. We derive each of these components in the following sections.

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q(z_x|x,z_y)}\left[-\log p(x|z_x, z_y)\right] & \tag{12}\\
+ D_{KL}(q(z_1|x, z_y, z_x) \parallel p(z_1)) & \tag{13}\\
+ D_{KL}(q(z_0|x, z_y, z_1, z_x) \parallel p(z_0|z_1)) & \tag{14}\\
+ D_{KL}(q(z_x|x, z_y, z_1, z_0) \parallel p(z_x|z_1, z_0)) & \tag{15}
\end{aligned}$$

H.1 Prior Loss

Equation 13 establishes the prior loss and the base layer in the hierarchy. In accordance with the VP framework, the correct prior distribution for the final latent representation is the standard normal $p(z_1) \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. We learn the noise schedule via $\log SNR(t) = -\gamma_\phi(t)$, as defined in VDM [41]. As such, we must ensure the terminal state in the forward diffusion process aligns with the prior distribution. Substituting the VP noise schedule yields the following distribution for the posterior: $q(z_1|x, z_y, z_x) = \mathcal{N}(\alpha_1 z_x, \sigma_1 \mathbb{I})$, where $\sigma_t = \sqrt{\sigma(amma_\phi(1))}$, and $\alpha_1 = \sqrt{\sigma(-\gamma_\phi(1))}$ as obtained from Section 2.4. We estimate this KL divergence through Monte-Carlo sampling:

$$D_{KL}(q(z_1|x, z_y) \parallel p(z_1)) = \mathbb{E}_{z_x \sim q(z_x|x,z_y)}\left[(\alpha_1 z_x)^2 + \sigma_1^2 - \log(\sigma_1^2) - 1\right] \tag{16}$$

H.2 VAE Loss

Equation 15 delineates our contribution to this unified variational model. We derive the base distributions for the right-hand distribution in Equation 9: $p(z_x|z_0) \sim \mathcal{N}\left(\frac{1}{\alpha_0} z_0, \frac{\sigma_0}{\alpha_0}\mathbb{I}\right)$. We further describe that the posterior distribution in this KL term is merely the regular VAE posterior $q(z_x|x, z_y) \sim \mathcal{N}(\mu_\theta(x, z_y), \sigma_\theta(x, z_y))$. We define the posterior over z_0 given z_x by adhering to the definition of the VP diffusion process $q(z_0|z_x) \sim \mathcal{N}(z_x, \sigma_0 \mathbb{I})$. As all these distributions are normal, we can provide an explicit form for this loss:

$$\begin{aligned}
&D_{KL}(q(z_x|x, z_y) \parallel p(z_x|z_1)) = \\
&\quad \mathbb{E}_{z_x \sim q(z_x|x,z_y), z_0 \sim q(z_0|z_x)}\left[D_{KL}(\mathcal{N}(\mu_\theta(x, z_y), \sigma_\theta(x, z_y) \parallel \mathcal{N}\left(\frac{1}{\alpha_0} z_0, \frac{\sigma_0}{\alpha_0}\mathbb{I}\right))\right]
\end{aligned}$$

where the KL term is the regular normal distribution KL.

$$D_{KL}(\mathcal{N}(\mu_0, \sigma_0) \parallel \mathcal{N}(\mu_1, \sigma_1)) = \frac{1}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} + \frac{(\mu_0 - \mu_1)^2}{\sigma_1^2} + \log \sigma_1^2 - \log \sigma_0^2 - 1\right]$$

H.3 Diffusion Loss

Equation 14 defines the final diffusion loss term for the denoising network. We follow the derivation from Kingma et. al. [41] for a continuous-time diffusion process. The key insight is to interpret the diffusion process as infinitely deep hierarchical variation model. The VP framework defines intermediate steps as: $q(z_t|x, y) \sim \mathcal{N}(\alpha_t x, \sigma_t \mathbb{I})$. Following [41], we derive a Monte-Carlo estimate

of the integral loss for a noise prediction network:

$$
\begin{aligned}
D_{KL}(q(z_0|x, z_y, z_1, z_x) \parallel p(z_0|z_1)) &= -\frac{1}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbb{I})}\left[\int_0^1 \mathit{SNR}'(t)\,\|z_x - \hat{z_x}(z_t, z_y, t)\|_2^2\,dt\right] \\
&= -\frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbb{I})}\left[\mathit{SNR}'(t)\,\|z_x - \hat{z_x}(z_t, z_y, t)\|_2^2\right] \\
&= \frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbb{I})}\left[\gamma'_\phi(t)\,\|\epsilon - \hat{\epsilon_\theta}(z_t, z_y, t)\|_2^2\right]
\end{aligned}
$$

H.4 Reconstruction Loss

Equation 12 is the final and simplest aspect of this model. Since we are regressing the parton components, we use a Normal decoder on the VAE and use a simple Mean Squared Error loss as the primary reconstruction loss.

$$\mathcal{L}_{MSE} = \|\textsc{Decoder}(z_x, z_y) - x\|_2^2$$

As we explain in the text, this MSE loss works well for most components but fails to capture the peaked nature of the mass term. Therefore, we add the physics consistency loss described by Equation 10 in order to assist with this mass reconstruction.

$$\mathcal{L}_C = \lambda_C \left|\hat{M}^2 - \left(\hat{E}^2 - \|\hat{p}\|^2\right)\right|$$

The total reconstruction loss is simply the sum of these two components.

$$\mathbb{E}_{q(z_x|x,z_y)}\left[-\log p(x|z_x, z_y)\right] = \mathbb{E}_{q(z_x|x,z_y)}\left[\mathcal{L}_{MSE} + \mathcal{L}_C\right]$$

	VLD	**UC-VLD**	**C-VLD**	LDM	VDM	CVAE	CINN
Leptonic b	11.96	5.23	9.08	16.74	212.29	33.85	143.49
Leptonic ν_e / ν_μ	10.33	7.40	36.69	29.63	117.37	36.95	149.22
Leptonic e / μ	2.95	2.76	3.07	11.13	155.87	8.72	96.89
Hadronic b	9.52	4.72	11.38	15.66	226.58	39.02	132.69
Hadronic q_1	8.88	7.19	24.26	35.33	187.24	58.33	180.42
Hadronic q_2	8.56	5.93	53.15	43.56	99.16	46.42	123.12
Leptonic W	8.18	8.19	43.63	37.87	213.51	32.96	260.12
Hadronic W	7.55	6.94	39.97	50.28	215.10	53.85	308.33
Leptonic t	9.57	9.12	46.56	43.27	323.00	46.51	408.39
Hadronic t	15.12	6.57	36.25	48.82	343.31	59.08	444.66
$t\bar{t}$ System	16.15	9.52	85.58	70.04	384.93	68.85	761.74

Table 3: **Particle Distance**: Wasserstein Distances

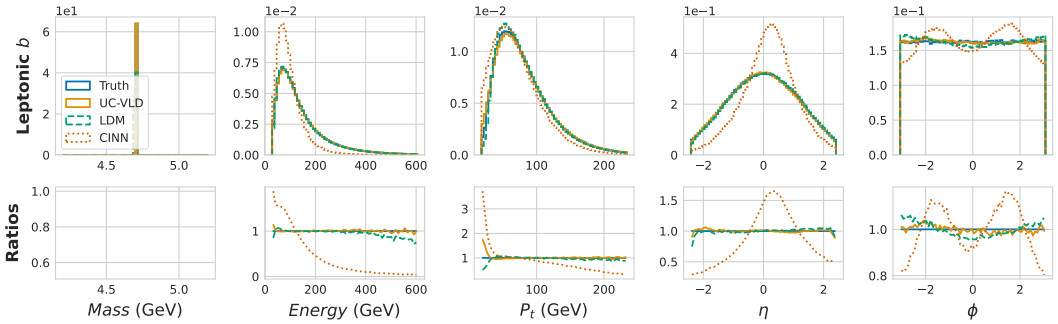


Figure 6: **Global Distribution**: Leptonic b Quark

	VLD	**UC-VLD**	**C-VLD**	LDM	VDM	CVAE	CINN
Leptonic b	0.93	0.59	0.62	1.01	17.04	2.53	10.33
Leptonic ν_e / ν_μ	0.79	0.73	2.31	1.76	11.02	2.72	12.15
Leptonic e / μ	0.24	0.29	0.26	0.70	13.99	0.59	7.56
Hadronic b	0.61	0.43	0.94	1.00	17.83	2.75	9.06
Hadronic q_1	0.90	0.77	2.71	2.75	16.30	4.82	14.73
Hadronic q_2	0.79	0.57	3.90	2.93	10.00	4.09	9.85
Leptonic W	0.54	0.66	2.94	2.37	16.48	2.09	17.58
Hadronic W	0.49	0.45	2.70	2.99	16.18	3.38	20.81
Leptonic t	0.64	0.64	2.61	2.42	21.72	2.52	23.27
Hadronic t	0.86	0.44	2.03	2.43	22.16	3.32	24.81
$t\bar{t}$ System	0.80	0.78	4.36	3.72	18.63	3.47	34.98

Table 4: **Particle Distance**: Energy Distances

	VLD	**UC-VLD**	**C-VLD**	LDM	VDM	CVAE	CINN
Leptonic b	0.85	0.07	0.06	0.59	1.91	1.01	1.30
Leptonic ν_e / ν_μ	0.97	1.06	0.73	0.88	1.76	0.77	1.71
Leptonic e / μ	0.48	0.33	0.52	0.35	1.47	0.43	1.02
Hadronic b	0.05	0.05	0.10	1.06	1.95	0.90	1.30
Hadronic q_1	0.40	0.40	0.60	0.63	1.63	0.81	1.50
Hadronic q_2	0.90	1.00	0.83	0.76	1.64	0.95	1.59
Leptonic W	0.10	0.11	0.46	0.37	1.39	0.30	1.27
Hadronic W	0.07	0.09	0.42	0.42	1.36	0.32	1.46
Leptonic t	0.09	0.10	0.35	0.32	1.61	0.30	1.43
Hadronic t	0.10	0.09	0.32	0.28	1.60	0.37	1.48
$t\bar{t}$ System	0.06	0.09	0.26	0.24	0.82	0.21	1.68

Table 5: **Particle Distance**: Kolmogorov-Smirnov Test Statistics

	VLD	**UC-VLD**	**C-VLD**	LDM	VDM	CVAE	CINN
Leptonic b	1.50	0.01	0.01	0.49	5.91	0.27	1.44
Leptonic ν_e / ν_μ	0.14	3.45	0.87	1.43	0.76	0.88	2.48
Leptonic e / μ	0.14	1.35	0.44	0.93	3.66	1.04	3.43
Hadronic b	0.00	0.01	0.01	3.57	5.99	0.29	1.09
Hadronic q_1	0.34	2.07	0.46	1.19	3.80	1.72	6.15
Hadronic q_2	1.54	0.08	0.67	0.96	2.31	0.98	2.67
Leptonic W	0.02	0.03	2.94	2.59	2.39	0.69	2.13
Hadronic W	0.01	0.02	2.43	2.54	2.11	1.25	2.53
Leptonic t	0.02	0.03	1.22	1.47	2.50	0.98	2.24
Hadronic t	0.03	0.04	0.99	1.14	2.39	1.05	2.45
$t\bar{t}$ System	0.01	0.01	0.06	0.04	0.48	0.04	3.59

Table 6: **Particle Distance**: KL Divergence with 128 bins.

	VLD	UC-VLD	C-VLD	LDM	VDM	CVAE	CINN
mass	2.06	2.61	17.76	19.12	87.24	20.91	113.29
pt	7.18	7.65	39.39	33.44	429.70	74.56	257.20
eta	0.37	0.26	0.47	0.40	0.44	0.51	4.35
phi	0.22	0.19	0.20	0.38	0.17	0.30	1.29
energy	27.33	16.26	140.80	152.24	772.94	155.16	1098.99
px	13.89	10.48	25.86	25.37	272.43	50.46	166.93
py	8.90	7.95	26.48	22.95	270.28	47.97	168.37
pz	48.80	28.16	138.66	148.42	645.15	134.69	1198.66

Table 7: **Component Distance**: Wasserstein Distances

	VLD	UC-VLD	C-VLD	LDM	VDM	CVAE	CINN
mass	0.51	0.76	3.77	3.50	15.30	3.00	9.70
pt	0.77	1.04	4.27	3.42	44.65	7.71	25.36
eta	0.26	0.17	0.28	0.25	0.25	0.33	2.76
phi	0.15	0.13	0.14	0.25	0.11	0.21	0.86
energy	1.40	0.96	7.50	7.36	53.11	8.32	70.01
px	1.25	1.05	2.23	2.20	20.68	4.13	12.49
py	0.87	0.81	2.23	1.88	20.54	3.90	13.15
pz	2.38	1.43	4.97	5.23	26.71	4.70	50.79

Table 8: **Component Distance**: Energy Distances

	VLD	UC-VLD	C-VLD	LDM	VDM	CVAE	CINN
mass	3.39	2.78	3.20	4.54	7.55	4.21	4.66
pt	0.08	0.14	0.36	0.31	3.30	0.67	1.92
eta	0.14	0.09	0.13	0.12	0.11	0.18	1.34
phi	0.07	0.06	0.07	0.12	0.05	0.10	0.39
energy	0.07	0.08	0.36	0.30	2.94	0.45	3.81
px	0.10	0.10	0.18	0.18	1.16	0.31	0.78
py	0.08	0.08	0.17	0.16	1.14	0.29	0.92
pz	0.15	0.09	0.18	0.17	0.88	0.17	1.91

Table 9: **Component Distance**: Kolmogorov-Smirnov Test Statistics

	VLD	UC-VLD	C-VLD	LDM	VDM	CVAE	CINN
mass	3.69	7.04	9.78	15.80	22.78	8.53	12.34
pt	0.01	0.02	0.07	0.07	3.08	0.22	1.66
eta	0.01	0.01	0.03	0.04	0.04	0.03	1.87
phi	0.00	0.00	0.00	0.01	0.00	0.01	0.07
energy	0.01	0.01	0.08	0.24	2.99	0.15	6.53
px	0.01	0.01	0.03	0.03	1.28	0.10	0.92
py	0.00	0.01	0.04	0.03	1.27	0.09	0.92
pz	0.01	0.00	0.05	0.13	0.86	0.05	5.89

Table 10: **Component Distance**: KL Divergence with 128 bins.

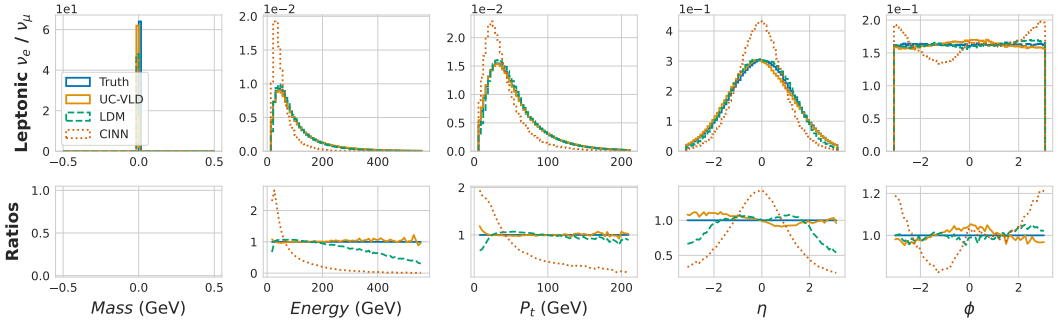


Figure 7: **Global Distribution**: Leptonic Neutrino ν_e / ν_μ

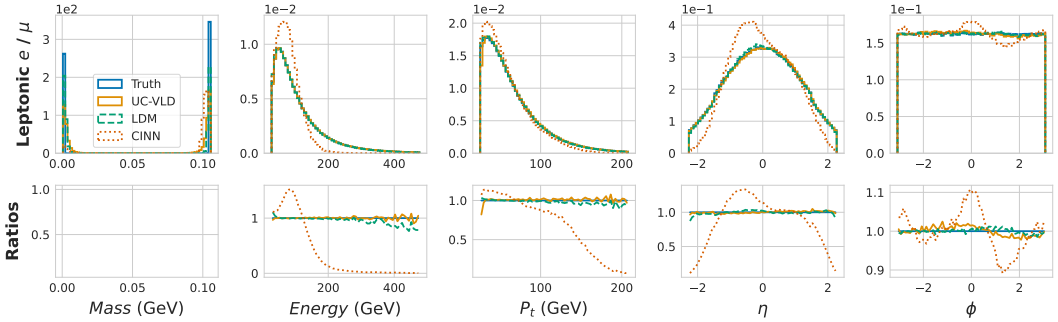


Figure 8: **Global Distribution**: Leptonic Lepton e / μ

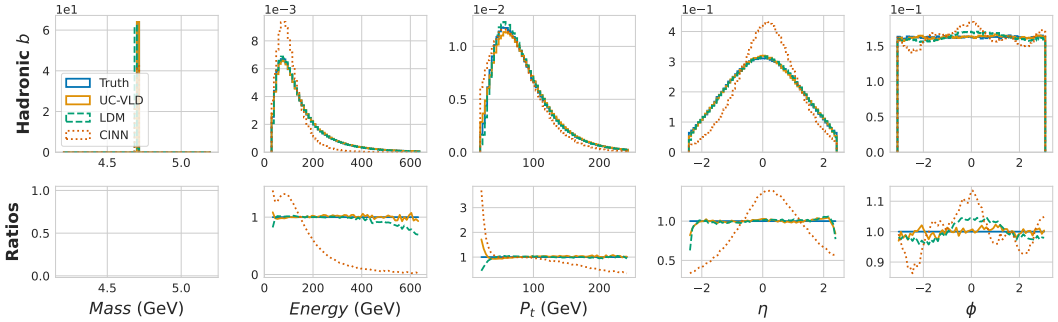


Figure 9: **Global Distribution**: Hadronic b Quark

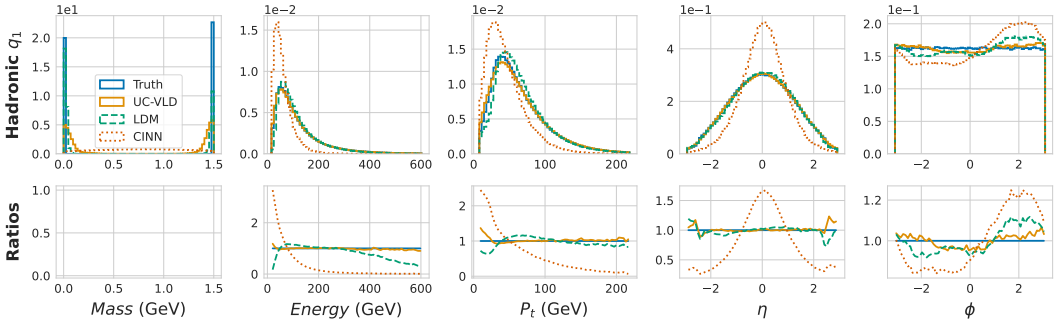


Figure 10: **Global Distribution**: Hadronic Light Quark

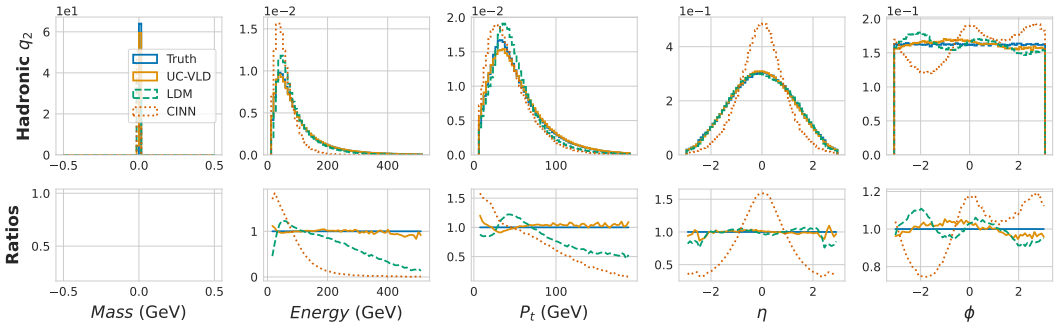


Figure 11: **Global Distribution**: Hadronic Light Quark

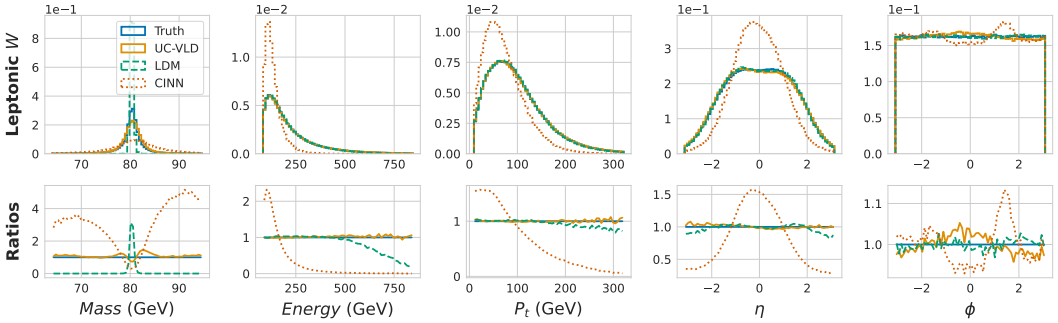


Figure 12: **Global Distribution**: Leptonic W Boson

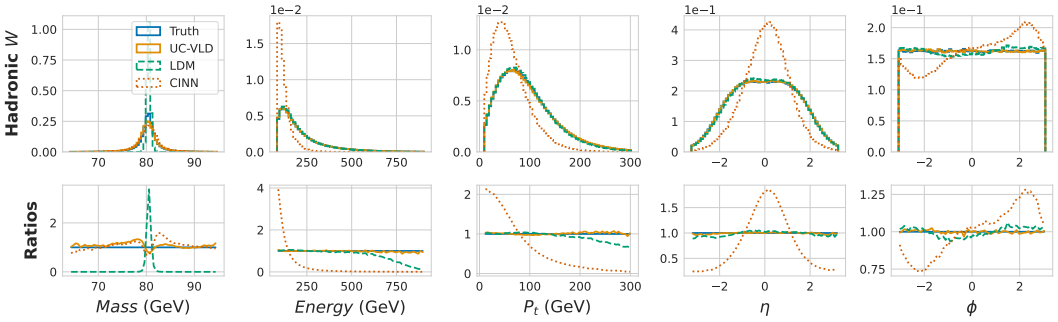


Figure 13: **Global Distribution**: Hadronic W Boson

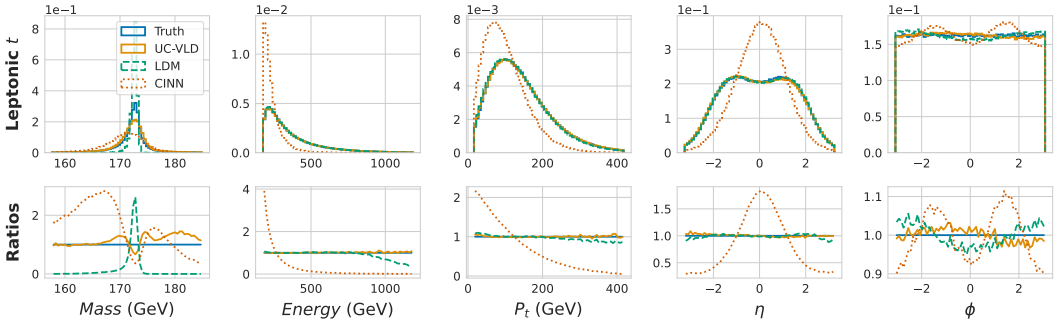


Figure 14: **Global Distribution**: Leptonic Top Quark

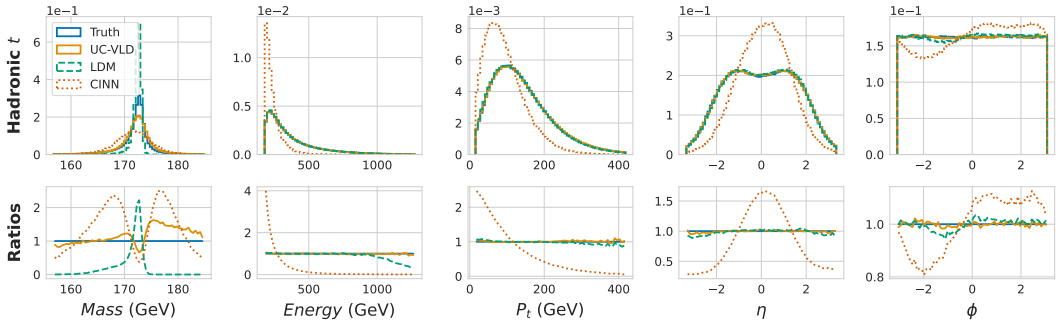


Figure 15: **Global Distribution**: Hadronic Top Quark

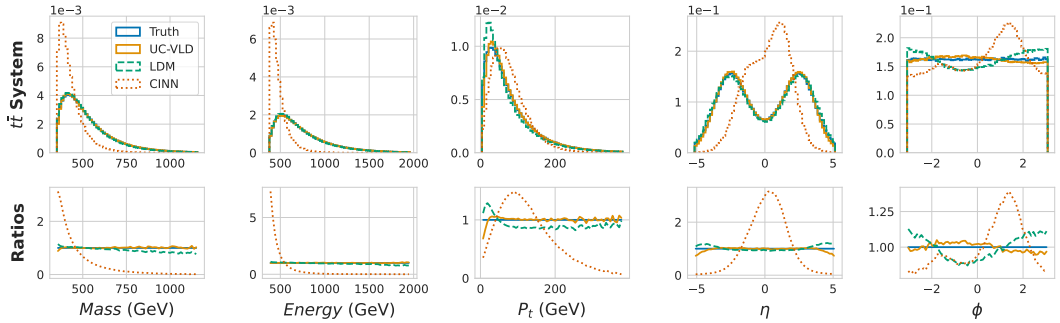
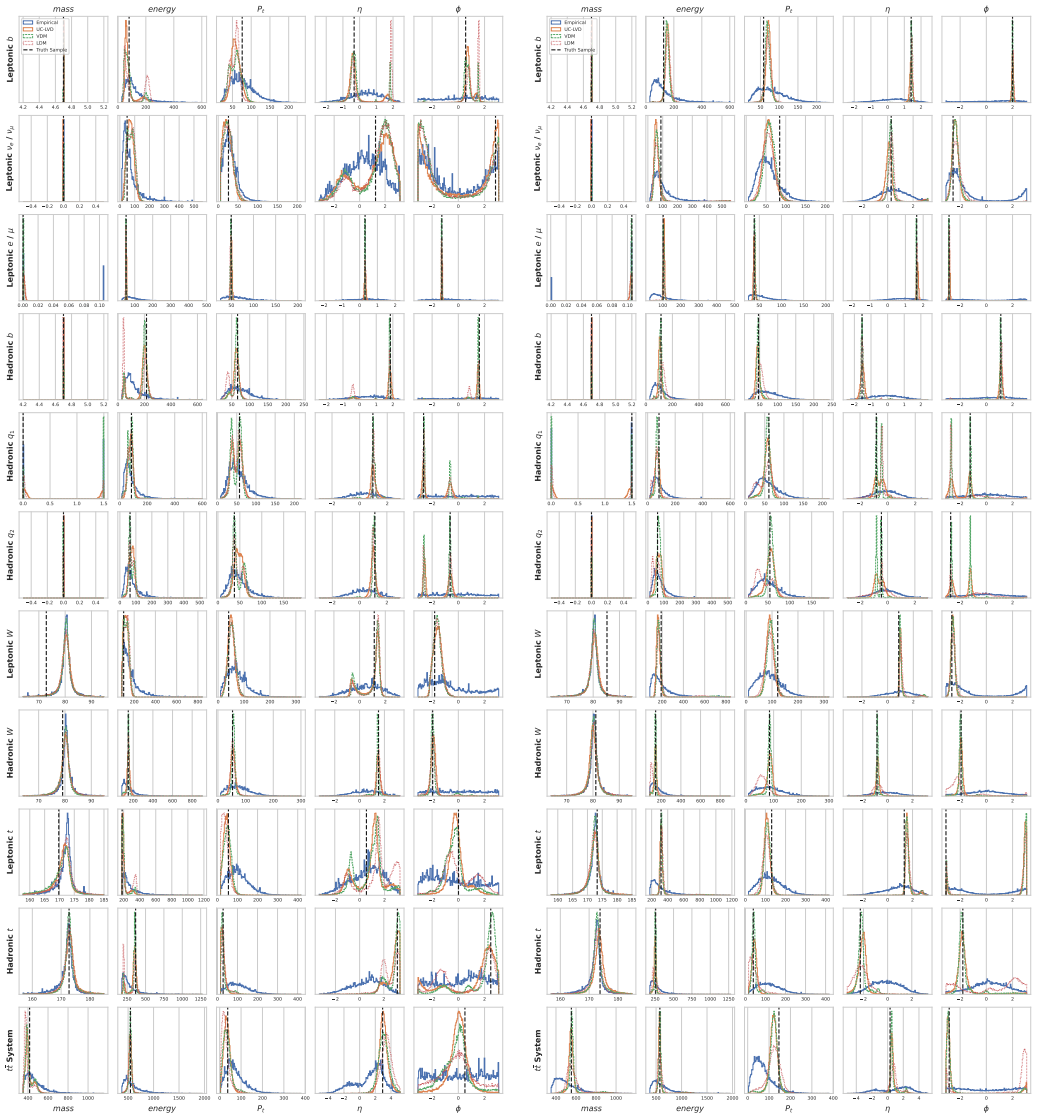


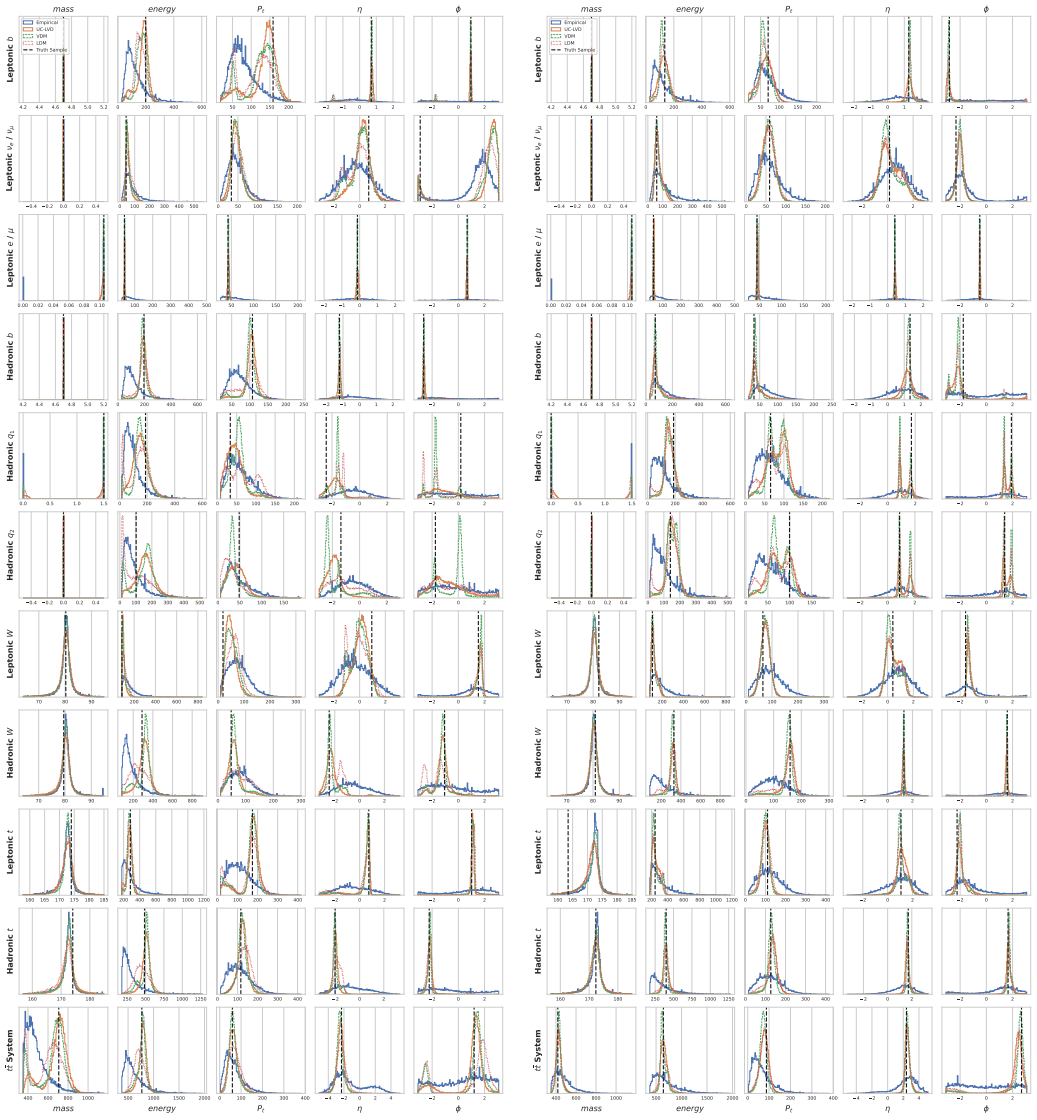
Figure 16: **Global Distribution**: Complete $t\bar{t}$ System



(a) Example Event 1

(b) Example Event 2

Figure 17: Posterior distributions for example events. Included is an empirical posterior distribution calculated from the training dataset.



(a) Example Event 3

(b) Example Event 4

Figure 18: Posterior distributions for example events. Included is an empirical posterior distribution calculated from the training dataset.