BitHEP – The Limits of Low-Precision ML in HEP

Daohan Wang

Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences (OeAW)

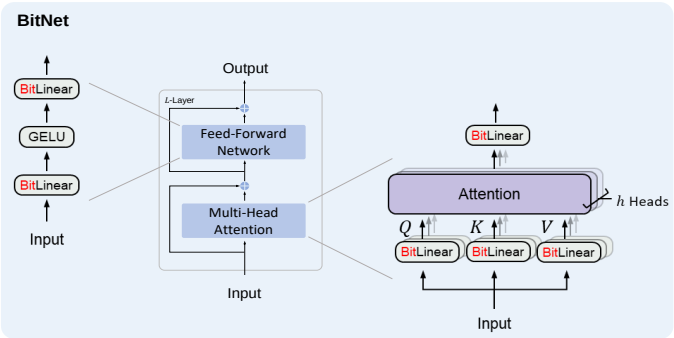
July 8, 2025

Collaborated with Claudius Krause and Ramon Winterhalder

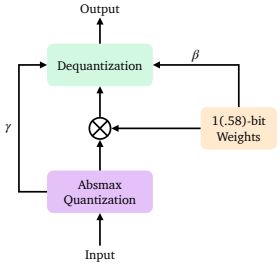
arXiv 2504.03387

Motivation

- In NLP and LLM, a recent proposal of only using 2 or 3 discrete states in the weights matrix called BITNET has gained significant attention. In this way, the matrix multiplication reduces to addition only. This significantly reduces storage and computational costs of transformer-based large language models while maintaining same level performance.
- Transformer-based LLMs: large size, high energy consumption BitNet: 1-bit Transformer



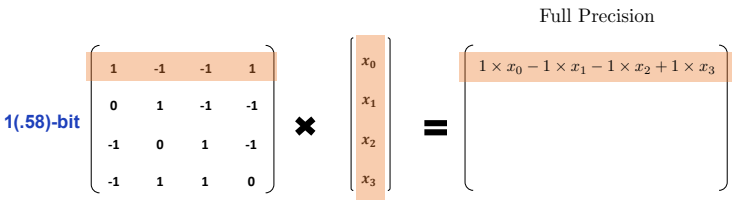
Bitlinear Layer



- Weight Quantization: $\widetilde{W}_{1b} = \text{sign}\left(W - \langle W \rangle\right) \Rightarrow \{1, -1\}, \quad \widetilde{W}_{1.58b} = \max\left(-1, \min\left(1, \text{round}\left(\frac{W}{\beta}\right)\right)\right), \beta = \langle |W| \rangle \Rightarrow \{1, 0, -1\}.$
- Pre-Activation Quantization: $\widetilde{x} = \max\left(-Q_b, \min\left(Q_b, \text{round}\left(\frac{x\, Q_b}{\gamma}\right)\right)\right), \qquad \gamma = \max(|x|)\,.$
- Quantized outputs during forward propagation: $y = \widetilde{W}\widetilde{x}$.
- Rescaled outputs during back propagation: $y = \widetilde{W}\widetilde{x} \times \frac{\beta\gamma}{Q_b}$
- We employ the 1.58-bit weights and choose an 8-bit input quantization, i.e. $b = 8$ and $Q_b = 128$.

Model Implementation

Binarizing other architectures also holds significant potential. We explore this potential by applying 1.58b-BITNET to benchmark the performance of various HEP applications.



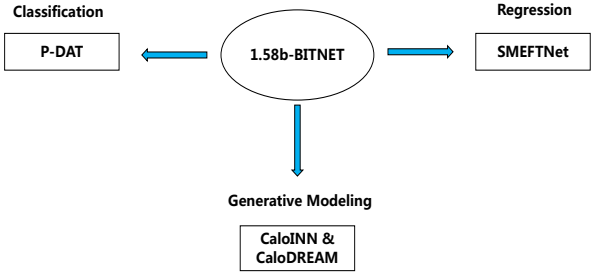
Ensure that the gradient calculation in back propagation is stable and accurate

Performance ✓ *Timing* ✗ *Energy Cost* ✗

HEP Applications

Benchmark models with 1.58b-BITNET implemented:

Linear Layer $\Rightarrow$ BitLinear Layer



Classification Application: P-DAT

Particle-Dual Attention Transformer (2307.04723)

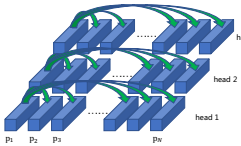
M. He & D. Wang

Quark/Gluon Discrimination

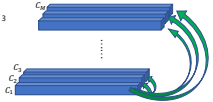
Dual Attention Mechanism

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_{N_h})$$

$$\text{where head}_i = \text{softmax}\left[\frac{\mathbf{Q}_i(\mathbf{K}_i)^{\mathrm{T}}}{\sqrt{C_h}} + \mathbf{U_1}\right]\mathbf{V}_i$$



Particle Attention Map $P \times P$

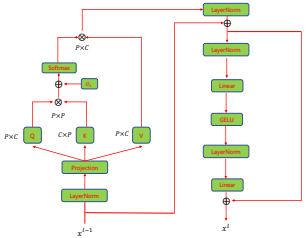


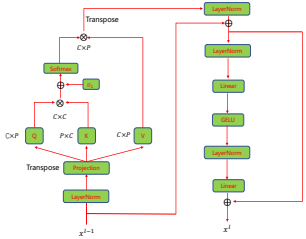
Particle Feature Attention Map $C \times C$

$$\mathcal{A}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = \text{softmax}\left[\frac{\mathbf{Q}_i^{\mathrm{T}}\mathbf{K}_i}{\sqrt{C}} + \mathbf{U_2}\right]\mathbf{V}_i^{T}$$

Particle interaction matrix $\mathbf{U_1}$:

$$\Delta R = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2},$$
$$k_{\mathrm{T}} = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})\Delta,$$
$$z = \min(p_{\mathrm{T},a}, p_{\mathrm{T},b})/(p_{\mathrm{T},a} + p_{\mathrm{T},b}),$$
$$m^2 = (E_a + E_b)^2 - \|\mathbf{p}a + \mathbf{p}_b\|^2,$$
$$\Delta p_{\mathrm{T}} = |p_{\mathrm{T},a} - p_{\mathrm{T},b}|$$





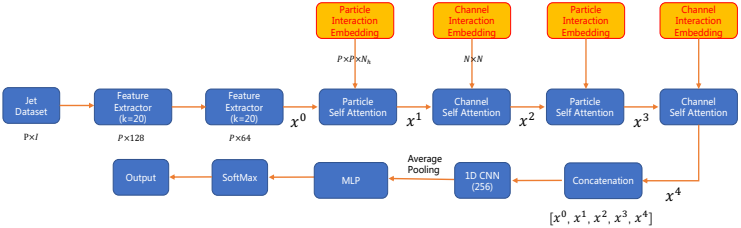
Channel interaction matrix $\mathbf{U_2}$:

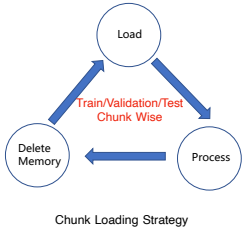
Straightforward ratios of $\{\mathrm{E_J}, p_{\mathrm{T_J}}, \sum p_{Tf}, \sum E_f, \overline{\Delta\eta}, \overline{\Delta\phi}, \overline{\Delta R}, \mathrm{PID}\}$

where $\overline{\Delta\eta}$, $\overline{\Delta\phi}$ and $\overline{\Delta R}$ correspond to the transverse momentum weighted sum of the $\Delta\eta$, $\Delta\phi$, ΔR of all the constituent particles inside the input jet, respectively. Here $\Delta\eta$, $\Delta\phi$ and ΔR refer to the distances in the $\eta - \phi$ space between each constituent particle and the input jet.

P-DAT Model Architecture

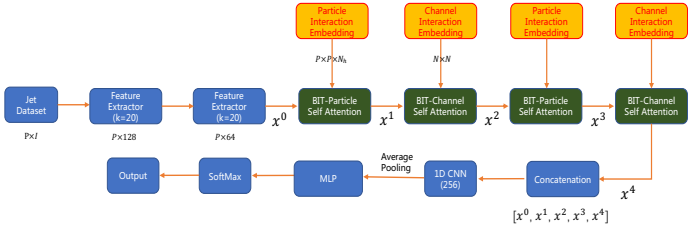
- Input features: $\log E$, $\log p_{\mathrm{T}}$, $\frac{p_{\mathrm{T}}}{p_{\mathrm{TJ}}}$, $\frac{E}{E_J}$, $\Delta\eta$ $\Delta\phi$, ΔR, PID of leading 100 particles.
- The particle attention module ($P \times P$ attention map) and the channel attention module ($C \times C$ attention map) are stacked while maintaining a consistent feature dimension of $N = 64$ and they can complement each other.
- Particle - Dual Attention Transformer: 2 Feature Extractor (1 EdgeConv + 3 Conv2D + 1 AvgPool) + 2 Particle Attention modules + 2 Channel Attention modules + 1D CNN + MLP.





P-DAT-BIT Model Architecture

- Replacing all the linear layers with BitLinear layers in the four attention modules of the P-DAT model (60% of the total parameters).
- All hyperparameters are identical to the non-binarized version.

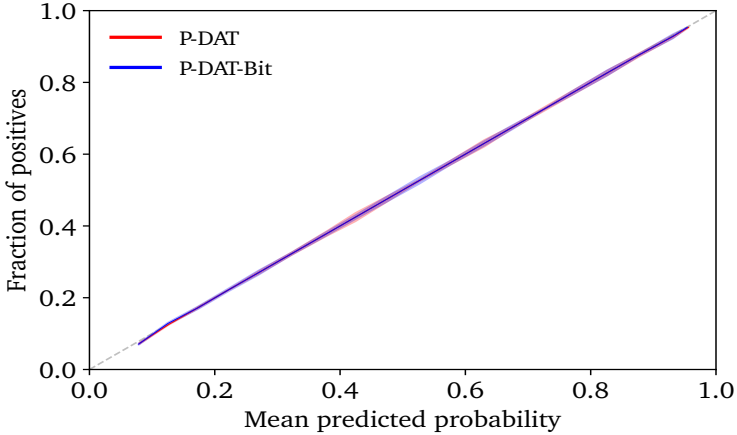


Performance

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$	Parameters	FLOPS
ParticleNet	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3	370k	540M
PCT	0.841	0.9140	43.2 ± 0.7	118.0 ± 2.2	193.3k	266M
LorentzNet	0.844	0.9156	42.4 ± 0.4	110.2 ± 1.3	224k	-
ParT	0.849	0.9203	47.9 ± 0.5	129.5 ± 0.9	2.13M	260M
P-DAT	0.839	0.9092	39.2 ± 0.6	95.1 ± 1.3	498k	144M
P-DAT-Bit	0.834	0.9040	35.0 ± 0.3	83.3 ± 1.2	498k	144M

Table: Comparison among the performance reported for P-DAT, P-DAT-Bit, and some existing classification algorithms on the quark-gluon discrimination dataset. The uncertainties in rejection rates are calculated by taking the standard deviation of 5 training runs with different random weight initialization.

Calibration Curves



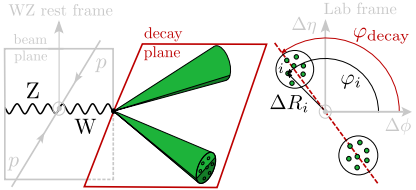
Regression Application: SMEFTNet

IRC-safe and Rotation-Equivariant Graph Neural Network (2401.10323)

S. Chatterjee, S. S. Cruz, R. Schöfbeck & D. Schwarz

Decay Plane Angle Regression

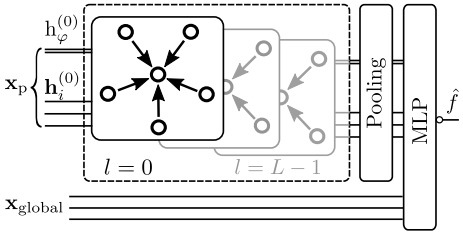
Motivation



- Equivariant SMEFTNet targets the linear SM–SMEFT interference by using dedicated angular analyses of W/Z decay planes to resolve helicity structures modified by CP-even ($\mathcal{O}_W$) and CP-odd ($\tilde{\mathcal{O}}_W$) gauge self-interaction operators.
- $pp \to W(\to q\bar{q})Z(\to l\bar{l})$ MG5+Pythia+Delphes Events with $H_T > 300$ GeV are retained. anti-k_T algorithm with R=0.8
- The decay plane angle changes as the W jet rotates. To study the hadronic final states of W boson, SMEFTNet is constructed to be equivariant to azimuthal rotations of the boosted jet's constituents around the jet axis, maintaining SO(2) symmetry.
- The particle features of each event inherently encode information about the decay plane for each event, hidden within the radiation patterns mapped to the variable-length constituent vector.

SMEFTNet's goal is to serve as a surrogate model to provide an optimal observable for detecting SMEFT effects from LHC collision data. Our focus, however, is on testing BITNET's performance in regression tasks, so we limit our study to the regression of the decay plane angle.

Sketch of the Network Configuration

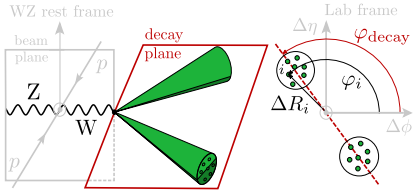


- Input features for l=0: Four-vectors p_i of the particles. $h_{\varphi,i}^{(0)} = \varphi_i, \quad \mathbf{h}_i^{(0)} = \Delta R_i$
- Message passing function: ${}^i\mathbf{m}_j^{(l)} = \omega_j^{\mathcal{N}(i)} f_{\mathrm{m}}^{(l)}(\hat{p}_i, \hat{p}_j)$ with $\omega_j^{\mathcal{N}} = \frac{p_{\mathrm{T},j}}{\sum_{k\in\mathcal{N}} p_{\mathrm{T},k}}$. Particle $j \in \mathcal{N}(i)$ of a particle i with $\Delta R_{ij} \leq \Delta R$.
- We demand SO(2) equivariance: $S_{\Delta\varphi}(h_\varphi, \mathbf{h}) = (h_\varphi + \Delta\varphi, \mathbf{h})$:
  - $\mathbf{h}_i^{(l+1)} = \sum_{j\in N(i)} \omega_j^{(N(i))} \boldsymbol{f}_{\mathrm{h}}^{(l)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, h_{\varphi,i}^{(l)} - h_{\varphi,j}^{(l)}\right)$ Invariance
  - $e^{ih_{\varphi,i}^{(l+1)}} = e^{ih_{\varphi,i}^{(l)} + i\sum_{j\in N(i)} \omega_j^{(N(i))} f_{\Phi}^{(l)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, h_{\varphi,i}^{(l)} - h_{\varphi,j}^{(l)}\right)}$ Equivariance
- After L iterations, the global pooling is applied to sum over all the constituents with the energy-weighting. The results along with the global features $\mathbf{x}_{\mathrm{global}}$ are fed into a final MLP.

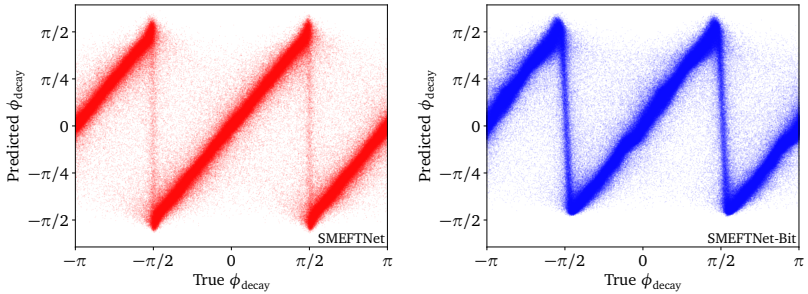
Loss Function

- Regression Target: Decay plane angle of the W boson's parton-level decay products $\phi_{j,\,\text{decay}}$
- Inputs: $\mathbf{x}_j = \{p_{T,i}, \phi_i, \Delta R_i\}_{i=1}^{N_j}$ with AK8 jet $p_T > 500$ GeV. 80%/20% of WZ data as train/test dataset.
- $L = \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{sim}}} \sin^2\left(\hat{f}(\mathbf{x}_j) - \varphi_{j,\text{decay}}\right)$.

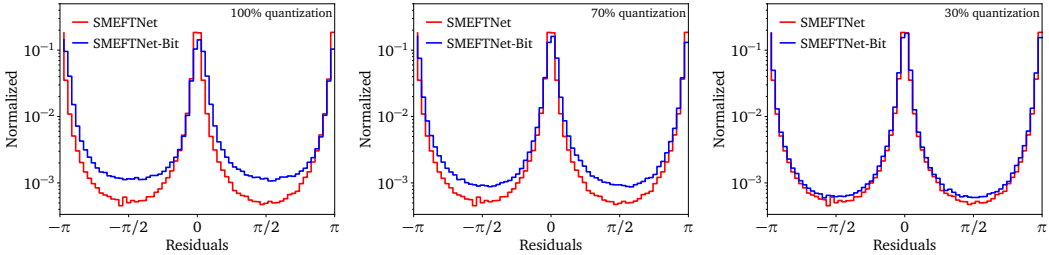
Since the simulated data lacks the information to distinguish constituents originating from up-type and down-type quarks, although switching the positions of the two partons alters the decay plane angle by π, the underlying simulated data remains unchanged. Consequently, the sine function is specifically employed to speed up learning the periodicity.



Scatter Plots



Probability Density Histograms of Residuals



Model Comparison	Wasserstein Distance	Separation Power
SMEFTNet	0.0021	0.0001
SMEFTNet-Bit	0.4546	1.4860
SMEFTNet-Bit70	0.2528	0.6138
SMEFTNet-Bit30	0.1040	0.1503

Generation Application: CaloINN and CaloDREAM

Normalizing Flows for High-Dimensional Detector Simulations (2312.09290)

F. Ernst, L. Favaro, C. Krause, T. Plehn & D. Shih

Detector Response Emulation via Attentive flow Matching (2405.09629)

L. Favaro, A. Ore, S. Schweitzer & T. Plehn

Fast Calorimeter Shower Simulations

Motivation

Deep generative networks based on normalizing flow provide fast and accurate surrogates for simulations in high-dimensional phase spaces by learning the underlying probability distribution of calorimeter showers from a reference dataset and then generating new samples based on this learned distribution.

The CaloChallenge was a data challenge in the HEP community, with four different datasets, increasing in their dimensionality. The goal of the challenge was to train generative networks on the datasets and to generate artificial samples as fast and precise as possible.

We consider two well-performing submissions of the CaloChallenge

- CaloINN based on normalizing flow.
- CaloDREAM based on conditional flow matching.

Evaluation Metrics

- Low-level classifier: Phase space of the voxels in each layer.
- High-level classifier: $\mathcal{I}_{ia}$, $E_i = \sum_a \mathcal{I}_{ia}$, $\frac{E_{\text{dep}}}{E_{\text{inc}}} = \frac{\sum_{a,i} \mathcal{I}_{ia}}{E_{\text{inc}}}$, $\langle l \rangle = \frac{\sum_a l_a \mathcal{I}_{ia}}{\sum_a \mathcal{I}_{ia}}$, $\sqrt{\frac{\sum_a l_a^2 \mathcal{I}_{ia}}{\sum_a \mathcal{I}_{ia}} - \left(\frac{\sum_a l_a \mathcal{I}_{ia}}{\sum_a \mathcal{I}_{ia}}\right)^2}$, λ_i.

CaloINN

- INN (coupling layer based normalizing flow) is applied for dataset 1 & 2 to sample $p_{\text{model}}(\mathbf{x})$ from $p_{\text{latent}}(\mathbf{r})$.
- INN uses rational quadratic splines for dataset 1 and cubic splines for dataset 2 & 3.
- Loss Function:

$$\mathcal{L}_{\text{INN}} = -\langle \log p_{model}(x) \rangle_{\mathcal{P}_d} = -\left\langle \log p_{latent}\left(\overline{G}_\theta(x)\right) + \log \left| \frac{\partial \overline{G}_\theta(x)}{\partial x} \right| \right\rangle_{\mathcal{P}_d} . \tag{1}$$

- The INN is trained on the full data, conditioned on the logarithm of the incident energies.
- CaloINN consists of multiple coupling layers, each containing an independent small neural network that models an invertible transformation for high-dimensional distributions.

Quantization Strategies for CaloINN

Strategy	Quantized Layers	Permutation Scheme	Notes
Default	None	Random	Baseline from original paper. Confirms consistency with prior work.
Exchange Permutation	None	**Fixed (Exchange)**	Shares permutation with BlockCentral. Ensures each dimension is transformed exactly once.
NNCentral	Central layers in each NN	Random	Only quantizes central layers in each NN. Minimal impact on model performance.
BlockCentral	All layers in central bijectors	**Fixed (Exchange)**	Float $\rightarrow$ quantized $\rightarrow$ float structure. Balances expressiveness and compression.
All	All layers in all bijectors	Random	Most aggressive compression strategy. Highest risk of performance degradation.

Performance of default and BitNet CaloINN

Dataset		Setup	Quantization	Low-Level AUC	High-Level AUC
ds1–γ	reg.	Default	-	0.633(3)	0.656(3)
		Exchange Perm.	-	0.640(4)	0.651(3)
	quant.	NNCentral	8.4%	0.640(3)	0.650(2)
		BlockCentral	66.6%	0.680(3)	0.669(3)
		All	99.9%	0.759(2)	0.828(2)
ds1–π^+	reg.	Default	-	0.793(3)	0.742(3)
		Exchange Perm.	-	0.784(2)	0.736(3)
	quant.	NNCentral	5.9%	0.801(2)	0.751(3)
		BlockCentral	66.6%	0.852(1)	0.807(2)
		All	99.9%	0.882(2)	0.907(2)
ds2	reg.	Default	-	0.738(4)	0.859(2)
		Exchange Perm.	-	0.728(6)	0.857(3)
	quant.	NNCentral	0.3%	0.780(3)	0.876(4)
		BlockCentral	71.4%	0.950(2)	0.979(1)
		All	99.9%	0.993(1)	0.998(0)

CaloDREAM

CaloDream combines Conditional Flow Matching (CFM) with transformers, consisting of:

- **Energy Network**: an autoregressive transformer predicting 45 layer energies.
- **Shape Network**: a vision transformer learning normalized shower shapes.

- Uses **patching** to group nearby voxels, enabling scalability to large detectors like DS3.
- Self-attention captures spatial correlations and sparsity in calorimeter showers.
- CFM enables efficient training of continuous normalizing flows.

- Unlike CaloINN, which uses many small NNs, CaloDream uses two large networks.
- This design makes it **less sensitive to quantization**, as fewer components are involved.
- The two networks can be quantized **independently and flexibly**.

Quantization Strategies for CaloDREAM

- Embedding: 4-head transformer, 64-dim embeddings.
- CFM network: 8-layer MLP with 256 hidden units.
- **Quantization Options**:
  - **Regular**: No quantization (baseline).
  - **Quantized**: Quantize 6 out of 8 hidden layers in the MLP (CFM), resulting in 66.09% of the energy network being quantized, corresponding to 5.54% of total CaloDREAM parameters.

- Operates on 135 patches of 48 voxels (total: 6480 voxels).
- Uses time, position, and conditional embeddings $\rightarrow$ 6 ViT blocks $\rightarrow$ final MLP.
- **Quantization Options**:
  - **Regular**: No quantization (baseline).
  - **No Embedding**: Quantize QKV, projections, and MLPs in ViT blocks (63.8% quantized).
  - **Full**: Additionally quantize embedding layers (66.22% quantized).

Performance of regular and BitNet CaloDREAM

Energy net	Shape net	Quantization	low-level AUC	high-level AUC
regular	regular	-	0.531(3)	0.523(3)
quantized		5.5%	0.532(3)	0.525(3)
regular	no embedding	58.5%	0.611(2)	0.543(3)
quantized		63.9%	0.610(5)	0.545(2)
regular	full	60.7%	0.735(4)	0.942(2)
quantized		66.2%	0.738(4)	0.944(3)

Table: Performance of regular and BitNet CaloDREAM using the classifier metric of the CaloChallenge. Uncertainties show the standard deviation over 10 random initializations and trainings of the classifier on the same CaloDREAM sample.

Summary and Outlook

- We demonstrated that **quantization-aware training (QAT)** enables competitive performance across classification, regression, and generative tasks in HEP, especially for classification tasks such as quark–gluon tagging.
- Larger networks can be quantized more easily, maintaining performance even with **60%** of weights quantized.
- **Selective quantization** (targeting specific layers) significantly improves performance compared to full model quantization.
- **Self-attention layers** in transformer-based architectures are particularly robust to low-bit quantization.
- QAT aligns with future **hardware and energy constraints**, offering a promising direction for scalable and efficient ML at the HL-LHC and beyond.
- Future directions include **optimizing quantization configurations**, **integrating QAT into large-scale models**, and adapting BITNET for fully quantized, low-precision inference—with **measurable gains in energy, memory, and latency**. Extending QAT to **real-time tasks**, such as the LHC trigger system, may enable fast, accurate inference on resource-constrained hardware like FPGAs.