BitHEP – Are 1-Bit Networks all we need?

Daohan Wang

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

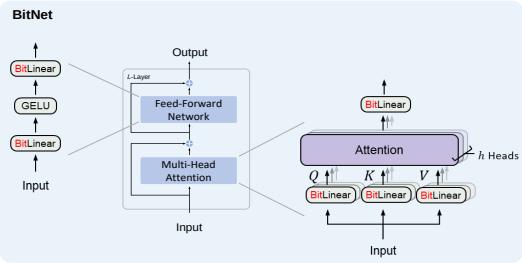
November 5, 2024

Collaborated with Claudius Krause and Ramon Winterhalder

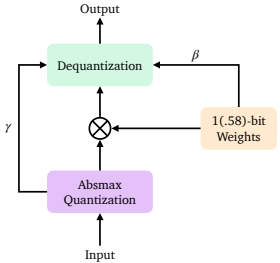
To be posted on arXiv 2412.XXXXX

Motivation

- In NLP and LLM, a recent proposal of only using 2 or 3 discrete states in the weights matrix has garnered significant attention. It is only considered for classification in fast triggers in HEP.
- Transformer-based LLMs: large size, high energy consumption BitNet: 1-bit Transformer



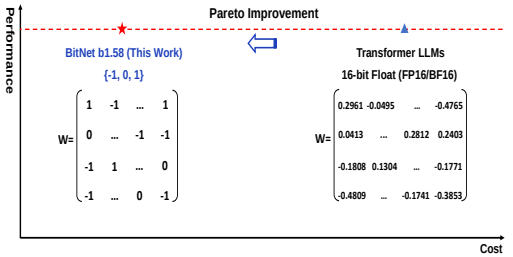
Bitlinear Layer

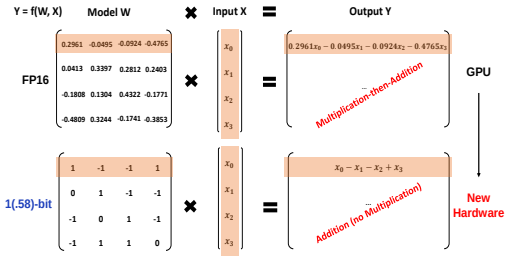


- Weight Quantization: $\widetilde{W}_{1b} = \text{sign}\left(W - \langle W \rangle\right) \Rightarrow \{1, -1\}, \quad \widetilde{W}_{1.58b} = \max\left(-1, \min\left(1, \text{round}\left(\frac{W}{\beta}\right)\right)\right), \beta = \langle |W| \rangle \Rightarrow \{1, 0, -1\}.$
- Pre-Activation Quantization: $\widetilde{x} = \max\left(-Q_b, \min\left(Q_b, \text{round}\left(\frac{x\, Q_b}{\gamma}\right)\right)\right), \qquad \gamma = \max(|x|)\,.$
- 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 $Q_b = 128$.

BITNET (2310.11453)

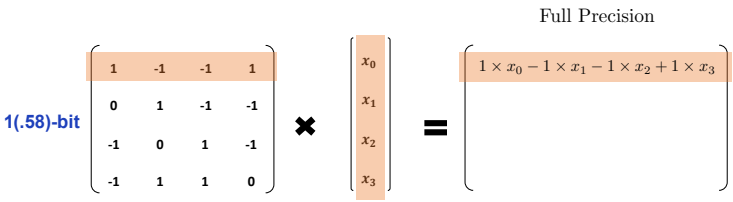
- The first 1-bit Transformer architecture for LLMs, aiming to scale efficiently in terms of both memory and computation.
- Employs low-precision binary weights and quantized activations, while maintaining high precision for the optimizer states and gradients during training.





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.



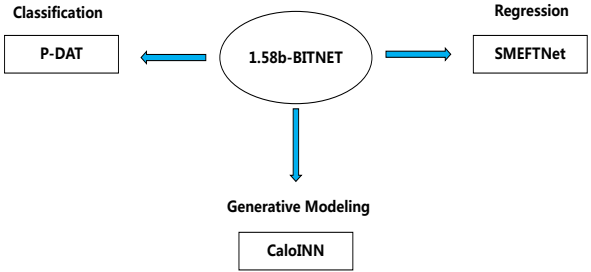
Ensure that the gradient calculation in back propagation is stable and accurate

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)

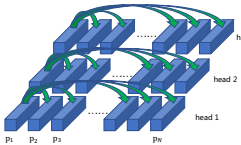
M. He & D. Wang

Quark/Gluon Discrimination

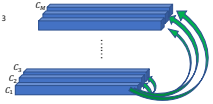
Dual Attention Mechanism

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

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



Particle Attention Map $P \times P$

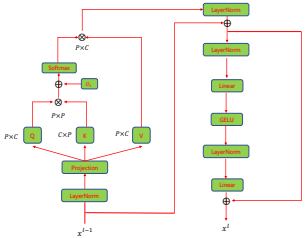


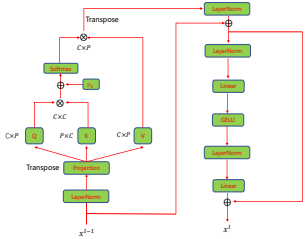
Particle Feature Attention Map $C \times C$

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

Particle interaction matrix $\mathbf{U_1}$:

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





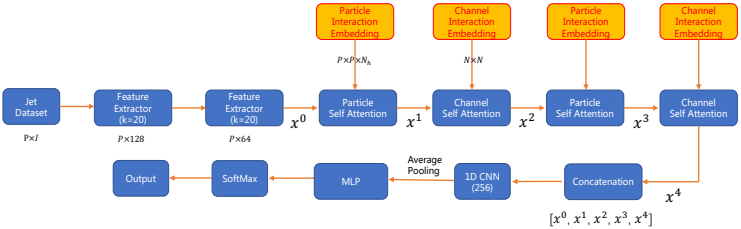
Channel interaction matrix $\mathbf{U_2}$:

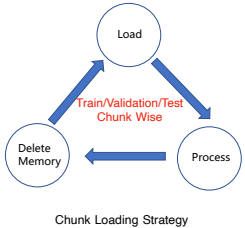
Straightforward ratios of $\{\mathrm{E_J}, p_{\mathrm{T_J}}, \sum p_{Tf}, \sum E_f, \overline{\Delta\eta}, \overline{\Delta\phi}, \overline{\Delta R}, \mathrm{PID}\}$

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

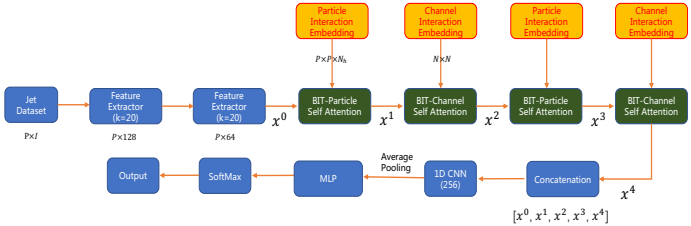
- Input features: $\log E$, $\log p_{\mathrm{T}}$, $\frac{p_{\mathrm{T}}}{p_{\mathrm{TJ}}}$, $\frac{E}{E_J}$, $\Delta\eta$ $\Delta\phi$, ΔR, PID of leading 100 particles.
- The particle attention module ($P \times P$ attention map) and the channel attention module ($C \times C$ attention map) are stacked 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

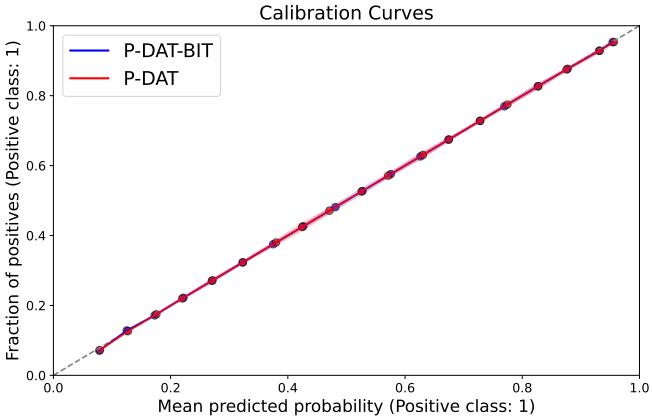
- Replacing all the linear layers with BitLinear layers in the four attention modules of the P-DAT model (60% of the total parameters).
- All hyperparameters are identical to the non-binarized version.



Performance

	Accuracy	AUC	$\text{Rej}_{50\%}$	$\text{Rej}_{30\%}$	Parameters	FLOPs
ResNeXt-50	0.821	0.9060	30.9	80.8	1.46M	-
P-CNN	0.827	0.9002	34.7	91.0	354k	15.5M
PFN	-	0.9005	34.7 ± 0.4	-	86.1k	4.62M
ParticleNet-Lite	0.835	0.9079	37.1	94.5	26k	-
ParticleNet	0.840	0.9116	39.8 ± 0.2	98.6 ± 1.3	370k	540M
ABCNet	0.840	0.9126	42.6 ± 0.4	118.4 ± 1.5	230k	-
SPCT	0.815	0.8910	31.6 ± 0.3	93.0 ± 1.2	7k	2.4M
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

Calibration Curves



Regression Application: SMEFTNet

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

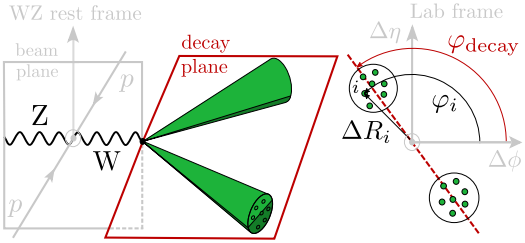
S. Chatterjee, S. S. Cruz, R. Schöfbeck & D. Schwarz

Decay Plane Angle Regression

Motivation

- The linear term in the polynomial describes the SM-SMEFT interference is the unambiguous harbinger of dimension-6 SMEFT effects.
- A dedicated angular analysis can extract SMEFT sensitivity where the orientation of the decay planes of the W or Z boson provides crucial sensitivity because it can resolve the amplitudes' helicity configuration which is altered in the SMEFT.
- Equivariant SMEFTNet focuses on the interference contribution to the differential cross-section from the operators $\mathcal{O}_W = \varepsilon^{ijk} W_\mu^{iv} W_\nu^{j\rho} W_\rho^{i\mu}$ and $\mathcal{O}_{\tilde{W}} = \varepsilon^{ijk} \widetilde{W}_\mu^{iv} W_\nu^{j\rho} W_\rho^{i\mu}$, with $\mathcal{O}_W$ and $\mathcal{O}_{\tilde{W}}$ induce CP-even and CP-odd modifications of the gauge boson self-interactions.
- A dedicated multivariate analysis of the fully leptonic decay mode extracts SMEFT sensitivity from global event kinematics.
- SMEFTNet exploits the SMEFT sensitivity from the hadronic final states $\mathcal{D} = \left\{ \boldsymbol{x}_{\text{global},j}, \{ \boldsymbol{x}_{\text{p}} \}_{i=1}^{N_p(j)} \right\}_{j=1}^{N_{\text{events}}}$.

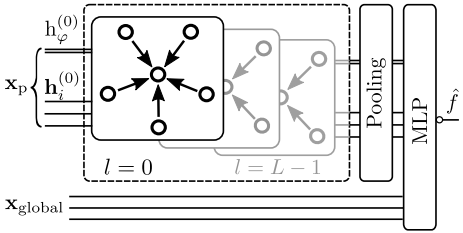
Motivation



- $pp \to W(\to q\bar{q})Z(\to l\bar{l})$ 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



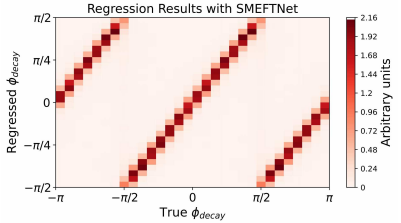
- Input features for l=0: Four-vectors p_i of the particles. $h_{\varphi,i}^{(0)} = \varphi_i, \quad \mathbf{h}_i^{(0)} = \Delta R_i$
- Message passing function: ${}^i\mathbf{m}_j^{(l)} = \omega_j^{(\mathcal{N}(i))} f_{\mathrm{m}}^{(l)}(\hat{p}_i, \hat{p}_j)$ with $\omega_j^{\mathcal{N}} = \frac{p_{\mathrm{T},j}}{\sum_{k\in\mathcal{N}} p_{\mathrm{T},k}}$. Particle $j \in \mathcal{N}(i)$ of a particle i with $\Delta R_{ij} \leq \Delta R$.
- We demand SO(2) equivariance: $S_{\Delta\varphi}(\mathrm{h}_\varphi, \mathbf{h}) = (\mathrm{h}_\varphi + \Delta\varphi, \mathbf{h})$:
  - $\mathbf{h}_i^{(l+1)} = \sum_{j\in N(i)} \omega_j^{(N(i))} \boldsymbol{f}_{\mathrm{h}}^{(l)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, h_{\varphi,i}^{(l)} - h_{\varphi,j}^{(l)}\right)$ 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.

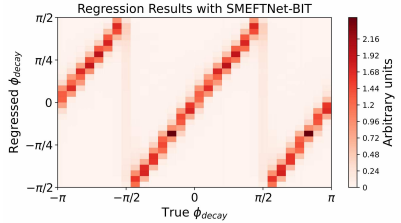
Loss Function

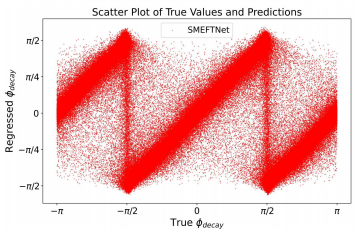
- Regression Target: Decay plane angle of the W boson's parton-level decay products $\phi_{\text{j, decay}}$
- Inputs: $\mathbf{x}_j = \{p_{T,i}, \phi_i, \Delta R_i\}_{i=1}^{N_j}$ with AK8 jet $p_T > 500$ GeV. 80%/20% of WZ data as train/test dataset.
- $L = \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{sim}}}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

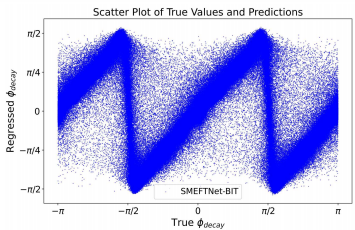
Since the simulated data lacks the information to distinguish constituents originating from up-type and down-type quarks, although switching the positions of the two partons alters the decay plane angle by π, the underlying simulated data remains unchanged. Consequently, the sine function is specifically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

2D Density and Scatter Plots

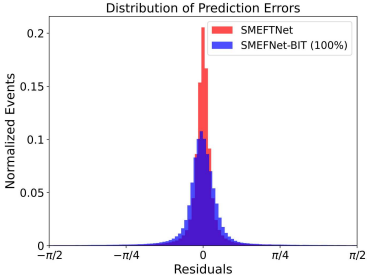


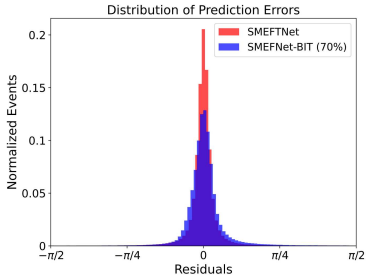


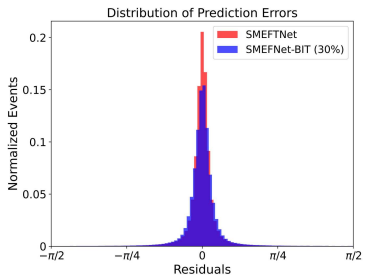




Probability Density Histograms of Residuals







Generation Application: CaloINN

Normalizing Flows for High-Dimensional Detector Simulations (2312.09290)

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

Fast Calorimeter Shower Simulations

Motivation

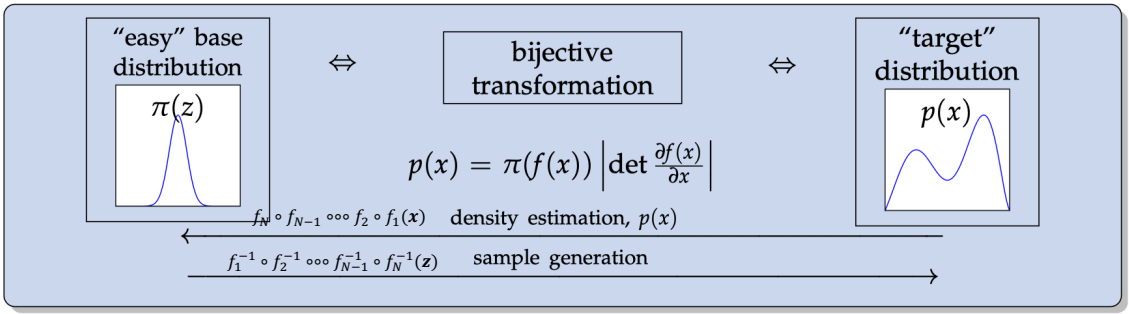
Deep generative networks based on normalizing flow provide fast and accurate surrogates for simulations in high-dimensional phase spaces by learning the underlying probability distribution of calorimeter showers from a reference dataset and then generating new samples based on this learned distribution.

CaloINN builds fast and accurate surrogate models for calorimeter shower simulation, using the technology of normalizing flows and VAEs.

Showers simulated with GEANT4 for different incident particles.

- The detector volume is segmented into layers in the direction of the incoming particle.
- Each layer is segmented along polar coordinates in radial (r) and angular (α) bins.
- A shower is given as the incident energy of the incoming particle and the energy depositions in each voxel.
- Dataset 1: Calorimeter showers for central photons and charged pions with incident energy ranging from 256 MeV to 4.2 TeV.
- Dataset 2 & 3: 100,000 positron showers with log-uniform incident energy ranging from 1 GeV to 1 TeV.

Normalizing Flows learn a change-of-coordinates efficiently.



INN (Invertible Neural Network)

- 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.
- Following the methodology presented in 2312.09290, we train a binary classifier D(x) on the voxels and a binary classifier trained on a set of high-level features, to distinguish GEANT4 showers from generated showers.
  - Low-level classifier: Phase space of the voxels in each layer
  - High-level classifier: $\langle \eta \rangle_i$, $\sigma_{\langle \eta \rangle_i}$, λ_i, E_i, E_{tot}/E_{inc}
- We use AUC (area under the calibration curve) to analyze the quality of the generative networks.

Classification Results

dataset	setup	AUC low-level	AUC high-level
ds1 photon	regular	0.619(2)	0.650(4)
	partial bitnet (10%)	0.646(2)	0.669(3)
	full bitnet	0.884(2)	0.855(3)
ds1 pion	regular	0.787(3)	0.735(2)
	partial bitnet (10%)	0.848(3)	0.787(2)
	full bitnet	0.906(2)	0.900(2)

Table: Performance of regular and bitnet using the classifier metric of CALOCHALLENGE WRITEUP. Uncertainties show the standard deviation over 10 random initializations and trainings of the classifier on the same sample.

Summary and Outlook

- BITNET with ternary precision (weights ±1, 0) performs excellently across various tasks in high-energy physics while significantly reducing computational resources.
- It has been successfully applied to practical tasks, including fast calorimeter simulations, quark/gluon discrimination and decay plane angle regression.

A typical usecase at the LHC: the Trigger system

- Collisions occur at 40 MHz, then we trigger with hardware down to 100 kHz and with software to a few kHz.
- If we use NNs, they need to be really fast!
- Current state-of-the-art: put NNs directly on a chip: FPGA, but size is limited!
- ML models can be compressed and converted into HLS projects using tools like hls4ml or conifer, leading to the creation of custom firmware designs for FPGAs that enable high parallelism for low latency and high throughput.
- Does a BitNet fits efficiently on an FPGA? If yes, we can use it to fit better NNs on the chip, allowing better triggers.