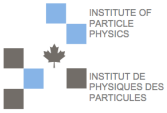


Summer Project Report

21st August 2026

ATLAS Level-0 Trigger Jet Energy Calibration with DeepSets

Philippe J. R. Joly[1,2,3,4]

[1]*McGill University, Montreal, Canada,* [2]*University of British Columbia, Vancouver, Canada,*
[3]*TRIUMF, Vancouver, Canada,* [4]*CERN (ATLAS), Geneva, Switzerland*

The High-Luminosity upgrade of the Large Hadron Collider will deliver an average of 200 inelastic proton-proton collisions per bunch crossing, placing stringent demands on the jet energy calibration. Even so, the ATLAS hardware trigger must reach a decision within a fixed latency using the limited resources of programmable firmware. A permutation-invariant DeepSets architecture is evaluated as a candidate calibration for the ATLAS Level-0 Trigger, taking topological clusters of calorimeter energy deposits as inputs. The study is based entirely on Monte Carlo simulation of dijet events at a centre-of-mass energy of 14 TeV with an average of 200 interactions per bunch crossing. A binned median estimator representative of the average-response corrections leaves the interquartile range (16-84 %) of the jet energy response unchanged, whereas the trained models reduce it 14.7 % and sharpen the efficiency turn-on of a single-jet selection by 8.8 GeV (50-99 %). The improvement is resilient under compression and post-training quantization, with a network of 281 parameters using fixed-point 16-bit weights and activations, and 12-bit inputs improving interquartile range by 13.2 % over the baseline. A firmware implementation of a network of 3k parameters occupies at most 1.4 % of the resources of the device foreseen to host the Level-0 Global Trigger, within the five 5 % target; however, the latency of the implementation is not measured.

List of contributions

Colin Gay	Supervision of the analysis.
Maximilian Swiatlowski	Supervision of the analysis.
Kelvin Leong	Supervision of the analysis.

Contents

1 Introduction

Through its first three runs, the Large Hadron Collider (LHC) has delivered an integrated luminosity of approximately 520 fb^{-1} to the ATLAS experiment. The High-Luminosity LHC (HL-LHC) is designed to increase this figure to 4 000 fb^{-1}, reaching an instantaneous proton–proton luminosity of up to $7.5 \times 10^{34}\ \text{cm}^{-2}\,\text{s}^{-1}$ [1]. This increase introduces significant challenges for the ATLAS data acquisition and trigger systems, one of the most demanding of which is resilience to pile-up. At the HL-LHC, an average of 200 inelastic collisions per bunch crossing is expected, ten times more than in LHC Run 3 conditions [2]. Maintaining trigger performance in such an environment requires substantial improvements to the reconstruction algorithms implemented in the trigger chain.

The ATLAS trigger reduces the 40 MHz bunch-crossing rate to a rate of 1 MHz at the hardware level, and to 10 kHz after the software selection [2]. The hardware decision must be produced within a fixed latency of 9.61 μs, which constrains both the complexity of the algorithms that can be deployed and the amount of information available to them. Every reconstruction step performed at this level is therefore subject to strict timing and resource requirements.

Jets are among the most common objects used to select events, and the trigger decision is typically based on thresholds applied to the jet transverse momentum. The sharpness of the resulting trigger efficiency curve is governed directly by the accuracy and the resolution of the jet energy estimate available at trigger level. A biased or poorly resolved energy estimate broadens the turn-on region, which reduces the selection efficiency for signal events close to the threshold and increases the rate of accepted background events. An accurate estimate of the jet energy is thus essential for trigger operation.

Jet energies can be estimated from the energy deposited in the electromagnetic and hadronic calorimeters. The measured energy is systematically lower than the energy of the incoming particle for several reasons: energy is deposited in dead material upstream of and between the calorimeter layers, part of the shower escapes the jet area or the calorimeter volume, and noise-suppression thresholds remove low-energy deposits. These effects fluctuate from shower to shower, so they contribute both a bias and a resolution term to the reconstructed energy. The ATLAS calorimeters are non-compensating, meaning that the response to the electromagnetic component of a shower is larger than the response to its hadronic component. Since the electromagnetic fraction of a hadronic shower varies strongly from jet to jet, this non-compensation is a dominant source of the observed energy fluctuations.

Low-energy jets are disproportionately affected. The relative contribution of noise thresholds and of energy lost in dead material grows as the jet energy decreases. Pile-up aggravates this further, as energy deposits from additional interactions overlap with the jet of interest and must be subtracted.

The current reconstruction proceeds in four stages. Calorimeter cells with a signal significantly above the expected noise are first grouped into clusters of energy deposits using the topological clustering algorithm [3]. The energy of each cluster is then estimated and calibrated to the hadronic scale. The calibrated clusters are subsequently combined into jets with the anti-k_t algorithm [4] using a radius parameter of $R = 0.4$. Finally, a jet-level calibration corrects the summed cluster energies to the particle-level jet energy scale.

The resources available for this chain at the hardware trigger level are limited. The target for new energy calibration implementations, motivated by anti-k_t algorithm performance [4] and ATLAS Level-0 (L0) Trigger goals [2], is roughly within a latency of 2 μs with a 240 MHz clock frequency and fit within 5 % of the Versal Premium VP1802 Field Programmable Gate Array (FPGA) on which the L0 Global Trigger will be hosted. This corresponds to 336 090 Flip Flops (FF), 168 045 Look Up Tables (LUT), 247 Block RAM (BRAM) blocks, and 718 digital signal processing (DSP) [5]. This analysis uses a Versal Premium VPK180 Evaluation Kit equipped with a VP1802 adaptive System on Chip (SoC) [5] to explore the hardware compatibility of DeepSets jet energy calibration.

Within these constraints, the existing approach has known limitations. The calibration applied to clusters and jets corrects the average response, but provides little improvement to the energy resolution, and its performance degrades for low-energy jets, precisely the regime in which the trigger thresholds are most sensitive to the quality of the energy estimate.

This note explores the potential of a DeepSets architecture [6] to improve the calibration of both clusters of energy deposits and jets. The permutation invariance of this architecture makes it naturally suited to the variable-size, unordered collections of cells and clusters that constitute the inputs at each stage, and its modest size makes it a plausible candidate for deployment in firmware.

2 Monte Carlo Simulations and Event Selection

2.1 Simulation and Sample Selection

The DeepSets architectures were trained and benchmarked using a dedicated Monte Carlo (MC) simulation designed to replicate High-Luminosity LHC (HL-LHC) conditions. Di-jet events were generated with Pythia 8 using the `mc21_14TeV` setup at $\sqrt{s}$ = 14 TeV, incorporating an average pileup of $\langle\mu\rangle$ = 200 and applying the latest Run 3 detector alignment and conditions. The full list of MC dataset identifiers is provided in Appendix A.

A total of 249 900 events were generated, yielding 3 370 399 reconstructed jets matched to truth jets. Each jet comprises up to 48 constituent clusters with total transverse momentum (p_{T}) reaching up to 800 GeV. Inputs to the model consist of topological clusters (topo-clusters) constructed from calorimeter cells using cell signal-to-noise ratio thresholds of $|E/\sigma_{\mathrm{noise}}| > 4, 2, 2$ for cluster seeding, growth, and boundary inclusion, respectively [3].

Throughout this paper, $p_{\mathrm{T}}^{\mathrm{EM}}$ refers to the uncalibrated jet transverse momentum measured at the electromagnetic (EM) scale by summing constituent topo-cluster energies, and $p_{\mathrm{T}}^{\mathrm{True}}$ denotes the matched particle-level truth-jet p_{T}. A selection requirement of $p_{\mathrm{T}}^{\mathrm{True}} > 10$ GeV was applied, yielding 1 613 538 selected jets (a 47.9% selection efficiency).

Figure 2.1 illustrates the kinematic distributions of the sample. The selected jets exhibit uniform distributions across azimuthal angle (ϕ) and pileup (μ), a exponentially falling spectrum in $p_{\mathrm{T}}^{\mathrm{True}}$, and a pseudorapidity ($\eta$) distribution symmetric around $\eta = 0$ with higher density at central rapidities ($|\eta| < 2.5$). For model training and evaluation, 15 % of the selected jets were reserved as an independent test set (accessed only

for final performance reporting), while another 15 % of the remaining sample was set aside as a validation set for hyperparameter tuning.

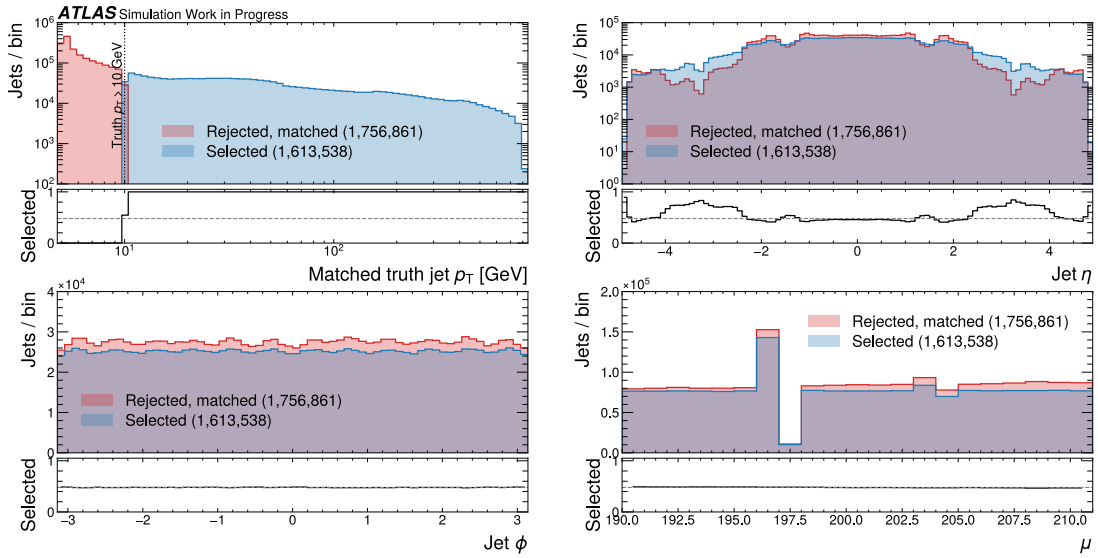


Figure 2.1: Kinematic distributions of the MC simulation dataset across $p_{\mathrm{T}}^{\mathrm{True}}$ (top left), jet η (top right), jet ϕ (bottom left), and pileup parameter μ (bottom right). Each panel shows the selected ($p_{\mathrm{T}}^{\mathrm{True}} > 10$ GeV, blue) and rejected (red) jet distributions, along with the corresponding binned selection efficiency shown in the lower subpanels.

2.2 Uncalibrated Energy Response

The baseline uncalibrated jet energy response, obtained by summing raw constituent topo-cluster energies at the EM scale, is shown in Figure 2.2. The response exhibits a significant energy-dependent offset relative to unity (left panel), driven by non-compensating calorimeter effects, inactive material losses, and out-of-cone shower leakage. This offset is accompanied by substantial response broadening and asymmetric low-energy tails below $p_{\mathrm{T}}^{\mathrm{True}} \approx 50$ GeV. The relation between $p_{\mathrm{T}}^{\mathrm{EM}}$ and $p_{\mathrm{T}}^{\mathrm{True}}$ (right panel) demonstrates that this resolution breakdown is a physical feature of energy measurement at low momentum rather than a mathematical artifact of the response ratio definition.

Energy resolution throughout this paper is quantified using the Interquartile Range (IQR = $Q_{84\%} - Q_{16\%}$) of the response distribution, which corresponds to 2σ for a Gaussian distribution while remaining robust against non-Gaussian tails. When comparing across energy scales, the relative resolution is defined as IQR$/(2 \times \text{median})$.

As shown in the middle panel of Figure 2.2, response degradation worsens significantly at extreme values of $|\eta|$ due to increased detector material and forward calorimeter geometry. Mitigating these severe low-energy and forward-region resolution losses under extreme $\langle\mu\rangle = 200$ pileup conditions serves as the primary benchmark for the DeepSets calibration architecture evaluated in subsequent sections.

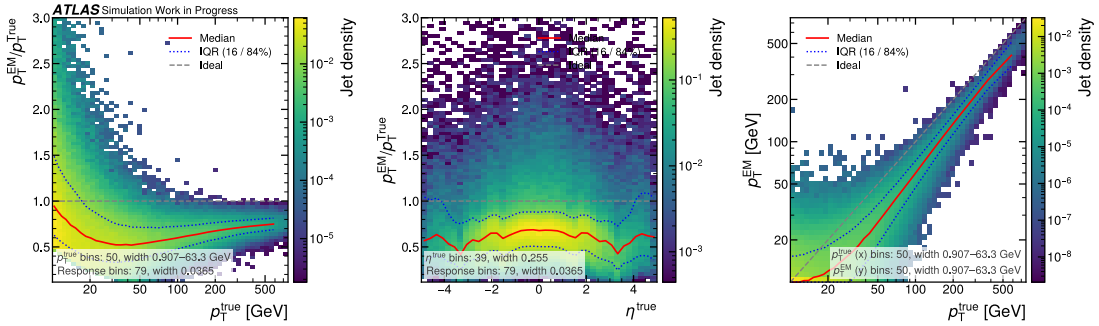


Figure 2.2: Uncalibrated jet energy response ($p_{\mathrm{T}}^{\mathrm{EM}}/p_{\mathrm{T}}^{\mathrm{True}}$) as a function of $p_{\mathrm{T}}^{\mathrm{True}}$ (left) and η (middle), alongside $p_{\mathrm{T}}^{\mathrm{EM}}$ versus $p_{\mathrm{T}}^{\mathrm{True}}$ (right). The ideal response, binned median, and interquartile range (IQR) across bins are overlaid in grey, red, and blue, respectively.

3 Methods

This work evaluates the viability of embedding a DeepSets-based calibration model in the jet energy calibration pipeline of the ATLAS Level-0 trigger for the HL-LHC. Feasibility is assessed in two domains: offline physics performance and real-time resource utilization on a FPGA.

3.1 The DeepSets architecture

The DeepSets architecture [6] is a permutation-invariant neural network designed to process variable-length sets of constituent inputs, $X = \{x_1, \ldots, x_m\}$. It comprises two multilayer perceptrons (MLPs) [7] separated by a set-wise aggregation operation. A per-element mapping subnetwork Φ first projects each fixed-dimensional constituent input x_i into a latent space representation. These latent vectors are then summed across the set, which enforces strict permutation invariance. A set-level subnetwork F subsequently processes the aggregated latent representation together with an optional set of global event features, $g = (g_1, \ldots, g_n)$, to produce the final prediction h. This composite mapping is illustrated in Figure 3.1 and defined in Eq. (3.1),

$$h\Big(X,\, g_1, \ldots, g_n\Big) = F\left(\sum_{x_i \in X} \Phi(x_i),\, g_1, \ldots, g_n\right). \tag{3.1}$$

DeepSets are well suited to high-energy physics applications because they naturally ingest inputs of variable cardinality, such as varying numbers of jets, tracks, topological clusters, or calorimeter cells [8]. Neither an artificial sequence ordering, as required by recurrent neural networks [9], nor zero-padding to a fixed input dimension, as required by standard MLPs [7], is needed.

The architecture is also attractive for real-time trigger applications. Its low computational complexity allows resource-efficient firmware synthesis on FPGAs, while the modular separation between Φ and F facilitates fine-grained hardware parallelization. Because set aggregation is deferred until after Φ has evaluated the latent representations (Figure 3.1), the per-constituent Φ operations can execute concurrently with upstream event reconstruction and global feature calculations in the hardware pipeline.

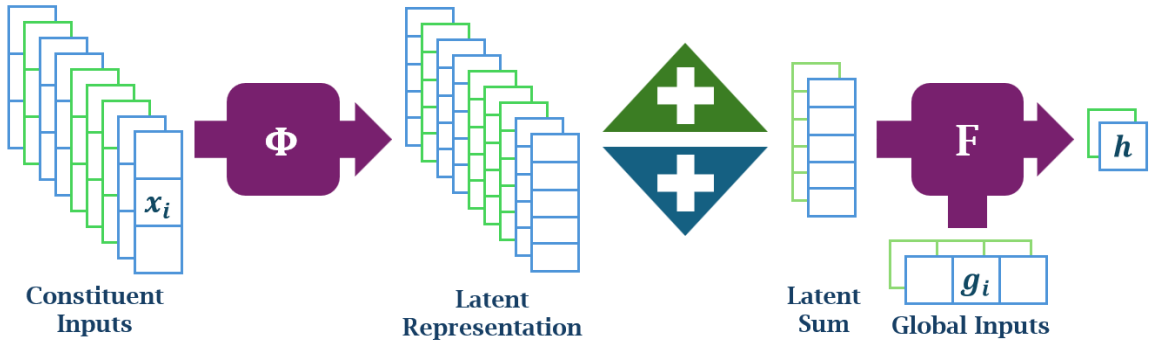


Figure 3.1: Schematic diagram of the DeepSets architecture processing two variable-length input sets (blue and green). Latent constituent representations are aggregated via summation according to set membership prior to evaluation by the set-level network F.

3.2 Training setup

Four combinations of constituent and global input features are considered throughout this study. They are defined in Table 3.1 and are referred to by the names given there.The `OfflineRef` set is the most complete configuration considered and is used for the full-precision offline model, while the `Trig` set is used for all firmware-oriented models.

Table 3.1: Definition of the global and constituent input sets. Here, z denotes the ratio of cluster p_T to jet p_T, $\Delta\phi$ and $\Delta\eta$ represent the angular distances between the jet and the cluster, and m is the reconstructed invariant jet mass.

Name	Constituent inputs	Global inputs
`Global4`	–	$\log p_\mathrm{T}$, $\lvert\eta\rvert$, η, $\log(m/p_\mathrm{T})$
`Log2`	$\log p_\mathrm{T}$, η	–
`OfflineRef`	$\log p_\mathrm{T}$, $\Delta\eta$, $\cos\Delta\phi$, $\sin\Delta\phi$, $\log z$	$\log p_\mathrm{T}$, $\lvert\eta\rvert$, η, $\log(m/p_\mathrm{T})$
`Trig`	$\log p_\mathrm{T}$, η	$\log p_\mathrm{T}$, $\lvert\eta\rvert$, η, $\log(m/p_\mathrm{T})$

All models are implemented and trained in TensorFlow [10], using rectified linear unit (ReLU) activations [11] in the intermediate layers and a linear output node. The network parameters are optimized with the Huber loss function [12] using a batch size of 1024 for up to 300 epochs, with early stopping applied after a patience of 40 epochs based on the validation loss. The regression target is the logarithmic energy correction factor, $\log\left(p_\mathrm{T}^\mathrm{True}/p_\mathrm{T}^\mathrm{EM}\right)$.

3.3 Baseline estimators

Two baselines are used for comparison. The first is the uncalibrated electromagnetic-scale response, $p_{\mathrm{T}}^{\mathrm{EM}}$. The second is a binned median estimator, which partitions the training dataset into 40 logarithmically spaced $p_{\mathrm{T}}^{\mathrm{EM}}$ bins and computes the median $p_{\mathrm{T}}^{\mathrm{True}}$ in each bin. At inference time, the estimator returns the precomputed median corresponding to the bin of the input jet. The binned median estimator serves as a proxy for the standard offline anti-k_t calibration pipeline, whose full execution was not feasible within the scope of this trigger study. In common with the anti-k_t algorithm, it corrects the average energy scale but has minimal effect on the energy resolution.

3.4 Performance metrics

Performance is quantified in terms of the jet energy response, defined as the ratio of the calibrated to the true jet transverse momentum. The energy resolution is characterized by the interquartile range (IQR) of the response distribution, and the accuracy of the energy scale by the median response and by the mean absolute error of the logarithmic correction factor (log MAE). Trigger performance is characterized by the efficiency turn-on curve for a single-jet 100 GeV threshold selection. The quantity $p_{\mathrm{T}}^{\mathrm{EFF}}$ denotes the p_{T} threshold at which the trigger efficiency reaches EFF %, and the turn-on sharpness is defined as $(p_{\mathrm{T}}^{99} - p_{\mathrm{T}}^{50})$. Differences between models are considered statistically significant if the performance metrics differ by more than $2\sigma_{\mathrm{seed}}$, where σ_{seed} is the standard deviation across independent training runs of a given architecture using different random seeds.

3.5 Post-training quantization

To meet the memory and execution constraints of the online trigger environment, post-training quantization (PTQ) is applied to the network weights, activations, and input features of all candidate trigger architectures. Fixed-point formats are denoted by $(\text{total bits}, \text{integer bits})$ where the number of integer bits include the $\pm$ bit.

The optimal fixed-point representation is derived in two steps. The dynamic range of each quantity is first analysed to determine the minimum number of integer bits required to prevent overflow. The fractional precision, and hence the total bit width, is then scanned systematically to identify the threshold at which the physics performance begins to degrade. Although raw p_{T} and $\log p_{\mathrm{T}}$ yield equivalent offline performance (Section 4.1), $\log p_{\mathrm{T}}$ is selected for the constituent inputs of the trigger models because its smaller dynamic range permits fewer integer bits in a fixed-point representation.

3.6 Firmware design

A dedicated firmware implementation was developed to establish whether a model of this class can satisfy the resource constraints of the Level-0 trigger. The implemented network uses 2 layers in Φ, with 25 and 9 nodes respectively, and 3 layers in F, with 30, 40, and 25 nodes respectively, and takes 4 constituent inputs

in fixed-point format (18,11). It contains 3 032 parameters. The fixed-point weights and biases use (18,5), while the layer activations, use (18,11).

This network differs in one respect from those discussed in the preceding sections: it provides both a regression and a classification output. The terminal three-node layer applies a softmax activation over its first two components to produce a classification score, and a linear activation on the third to produce the cluster energy regression. This configuration follows a pre-existing DeepSets implementation developed by Kelvin Leong at the University of British Columbia that was available for this study.

The hardware design and its implementation, illustrated in Appendix D, were built using high-level synthesis (HLS) with AMD Vitis [13] and Vivado [14]. The Φ network is implemented as a single HLS block that processes one cluster per pass, while the F network is split into two blocks: F-p1, comprising its first and second layers, and F-p2, comprising the third and terminal layers. These blocks are built as finite state machines whose states are driven by the PortCtrl block. Between Φ and F, the Latent-buffer stores the latent vectors of the clusters in an event until they are summed by the KSum block. SmartConnect blocks handle the routing of all control signals and the data flow between the hardware blocks and the static boundary. The Producer, Consumer, and AddrProducer blocks exist only to emulate the trigger pipeline and would not be included in a final implementation.

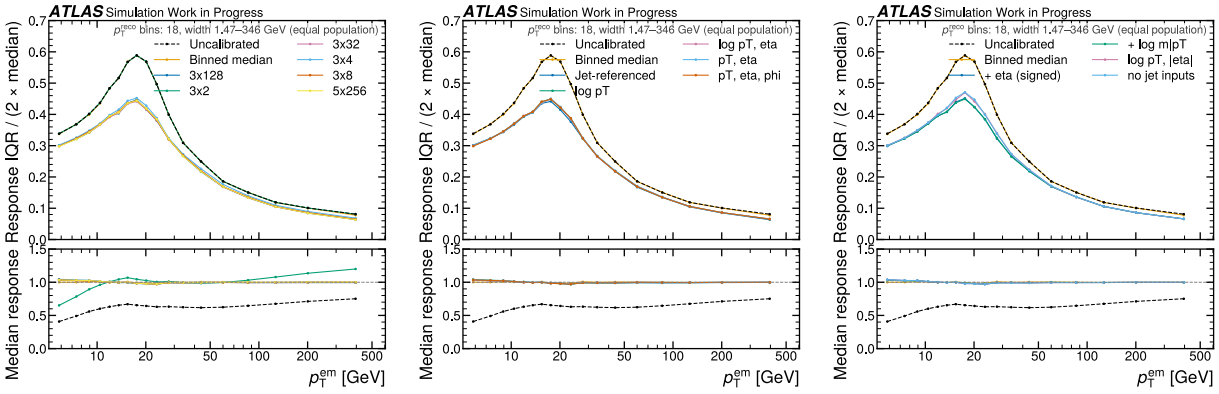
4 Results

The accuracy and precision of the models are evaluated across a range of input configurations and network architectures in order to quantify both the offline calibration performance and the prospects for trigger-level execution. Section 4.1 establishes the required network capacity and input feature content, Section 4.2 determines the fixed-point precision at which performance is preserved, Section 4.3 reports the final benchmarks on an independent test dataset, and Section 4.4 reports the resource utilization of the firmware implementation.

4.1 Architecture sizing and input feature ablation

Preliminary offline studies determine the impact of network capacity and feature selection on the model performance, benchmarked against the uncalibrated response and the binned median estimator. As shown in Figure 4.1, the binned median baseline yields a resolution comparable to that of the uncalibrated input. The trained models, by contrast, achieve marked improvements in energy resolution across the entire p_{T} spectrum, with the exception of the minimal (3×2) network layout, which shows no measurable resolution gain. While the neural networks regress the median response substantially better than the uncalibrated approach, they exhibit slightly worse median closure than the binned median baseline, particularly below 20 GeV. Reducing the network depth and width from (5×256) to (3×128) and then to (3×32) leads to a continuous degradation of the resolution. Performance plateaus below this threshold, with the (3×32), (3×8), and (3×4) architectures yielding equivalent offline performance within uncertainties.

Ablation scans over the constituent inputs indicate that replacing p_{T} with $\log p_{\mathrm{T}}$, or incorporating ϕ, provides no significant gain, whereas the inclusion of η consistently improves the resolution. The full `OfflineRef` constituent feature set yields measurable improvements over the simpler constituent inputs. For the global variables, the systematic addition of each input feature leads to incremental improvements in calibration precision.



(a) (b) (c)

Figure 4.1: Impact of (a) network size, (b) constituent inputs, and (c) global inputs on the energy resolution, given by the response IQR (top), and on the median response regression (bottom). The scans over network size, constituent inputs, and global inputs use the `OfflineRef`, `Global4`, and `Log2` input sets respectively, as defined in Table 3.1. The scanned models use symmetric Φ and F subnetworks. The input ablation analysis uses models fixed at (3×64) for each subnetwork. All metrics are evaluated on the validation dataset.

4.2 Post-training quantization

Figure 4.2 shows the bit-width optimization scan described in Section 3.5 for the `Tiny` DeepSets model, whose configuration is given in Table 4.1. Performance degradation remains negligible until the bit width drops below 16 bits for the weights and activations, and below 10 bits for the inputs. Comprehensive studies of the quantization fidelity (Figure B.1) and of the per-input bit allocations (Figure B.2) are presented in Appendix B.

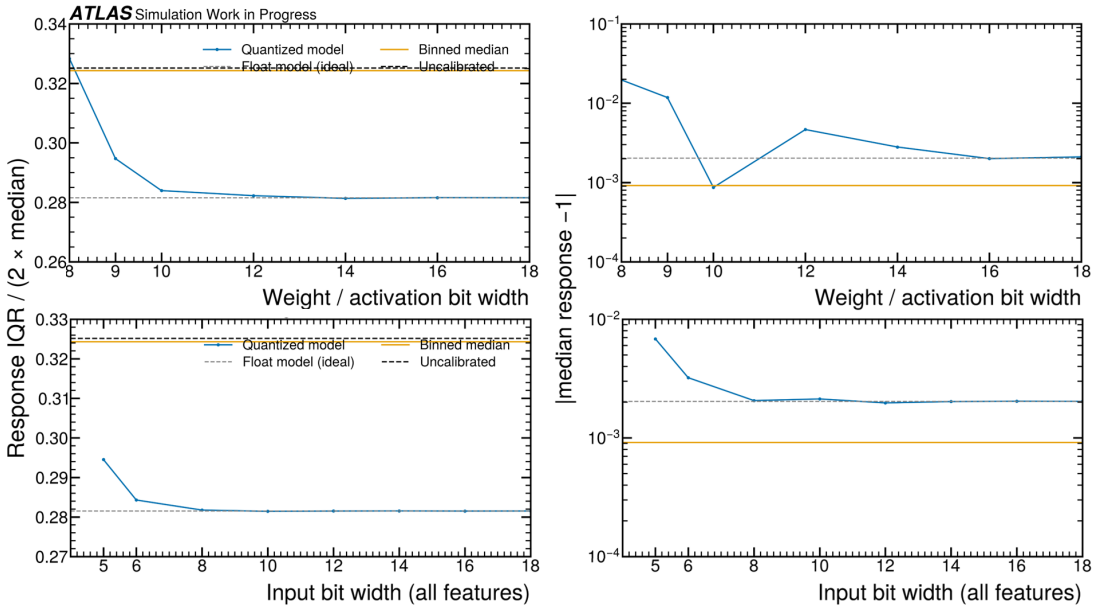


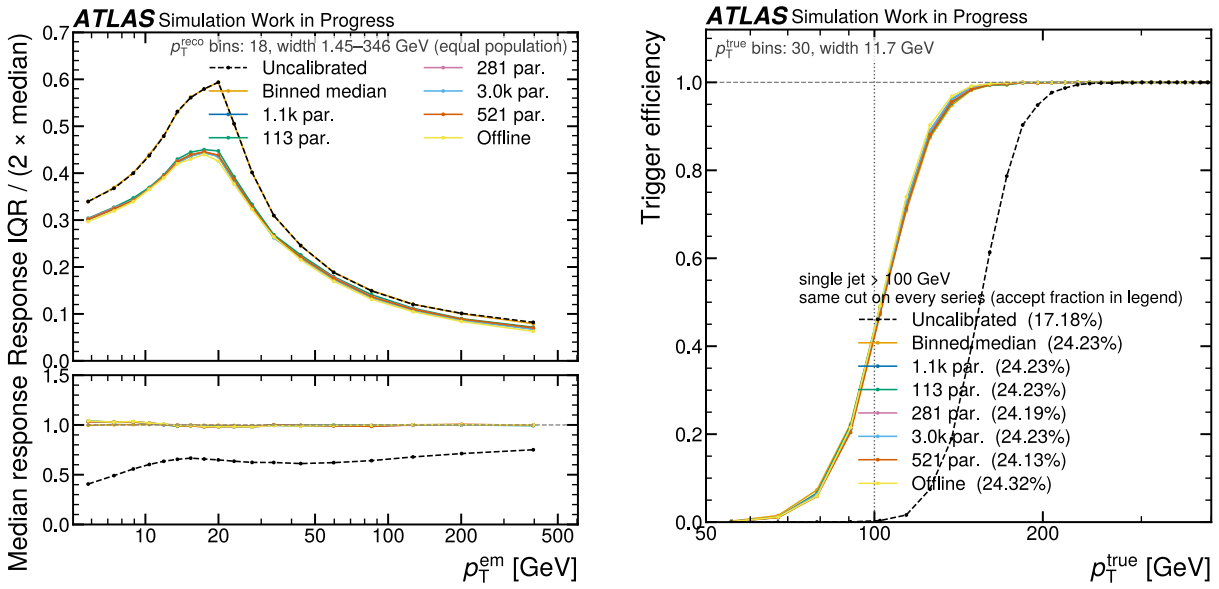
Figure 4.2: Effect of weight and activation quantization (top) and of input feature quantization (bottom) on the performance of the `Tiny` DeepSets model, whose configuration is given in Table 4.1. The energy resolution (IQR) and the median response are shown as a function of the total bit width, compared with the uncalibrated response, the binned median baseline, and the floating-point reference model.

4.3 Final model benchmarks

The physics performance of the baseline algorithms, of the full-precision offline model, and of five quantized trigger-optimized networks, ranging from 3 017 down to 113 parameters, is summarized in Figure 4.3 and Table 4.1. The evaluation is performed on an independent, secluded test dataset.

All neural network models outperform both the uncalibrated response and the binned median baseline for every metric evaluated, and the performance scales smoothly with parameter count. The best-performing offline model achieves an IQR of 0.2769, compared with 0.3248 for the binned median estimator. The quantized trigger variants retain most of this gain; even the ultra-compact `Tiny` network, with 281 parameters, yields an IQR of 0.2820.

The trigger efficiency turn-on curves for a single-jet 100 GeV threshold selection are shown in Figure 4.3(b) and parameterized in Table 4.1. The full offline model reduces the 99 % efficiency point, p_{T}^{99}, from 157.3 GeV for the binned median estimator to 148.2 GeV, and improves the turn-on sharpness from 54.3 GeV to 45.5 GeV. The quantized trigger models show consistent turn-on profiles, with sharpness ranging between 46.7 GeV for the `Large` network up to 53.2 GeV for the `Mini` network with the `Tiny` network achieving 50.9 GeV. Additional performance distributions across η (Figure C.2) and across bins of the p_{T} spectrum (Figure C.1) are provided in Appendix C. These demonstrate good closure across the detector acceptance, with minor residual deviations limited to the high-$|\eta|$ region and to low p_{T}.



(a) (b)

Figure 4.3: (a) Energy resolution and median response, and (b) single-jet 100 GeV trigger turn-on curves, both evaluated on the independent test dataset. The network configurations, input sets, bit widths, and trigger metrics are reported in Table 4.1.

4.4 Hardware Implementation

The resource utilization of the firmware design described in Section 3.6 is reported in Table 4.2 for a Versal Premium VP1802 FPGA driven at a clock speed of 240 MHz [5]. The complete design occupies 47 604 FFs, 29 959 LUTs, 68 BRAMs, and 106 DSPs, corresponding to at most 1.38% of any single resource category available on the device. The network itself accounts for a smaller fraction still: the Φ and F blocks together use 22 734 FF and 12 191 LUT, while the remainder is dominated by the SmartConnect routing infrastructure and by the emulation blocks, which would not be present in a final implementation. The asymmetry between Φ and F reflects the larger layer widths of the set-level network.

Since Φ is evaluated once per cluster, its throughput sets the rate at which an event can be processed, and multiple copies of the block can be instantiated to evaluate clusters concurrently. Table 4.3 reports the resource usage as a function of the Φ parallelization factor. The DSP usage scales approximately linearly with the parallelization factor, reaching 5.82% of the available slices at a factor of 64, whereas the BRAM usage remains constant at 7 blocks because the network parameters are shared between the instantiated copies. 55 is the highest factor so that the total DSP usage remains below the device limit of 5 %.

Table 4.1: Performance summary of the offline and trigger models evaluated on the test dataset, compared with the baseline estimators. Fixed-point quantization formats are denoted by (total bits, integer bits). The quantity $p_{\mathrm{T}}^{\mathrm{EFF}}$ corresponds to the p_{T} threshold at which the trigger efficiency reaches EFF%, and the turn-on sharpness is defined as $(p_{\mathrm{T}}^{99} - p_{\mathrm{T}}^{50})$.

Metric / configuration	Uncalibrated	Binned median	Offline	Large	Medium	Small	Tiny	Mini
Parameter count	–	–	594 945	3 017	1 081	521	281	113
Φ architecture	–	–	(5×256)	(2×26)	(2×15)	(2×10)	(2×7)	(2×4)
F architecture	–	–	(5×256)	(3×26)	(3×15)	(3×10)	(3×7)	(3×4)
Input feature set	$p_{\mathrm{T}}^{\mathrm{EM}}$	$p_{\mathrm{T}}^{\mathrm{EM}}$	`OfflineRef`	`Trig`	`Trig`	`Trig`	`Trig`	`Trig`
Weight format	–	–	`float32`	(16,3)	(16,4)	(14,3)	(16,3)	(16,2)
Activation format	–	–	`float32`	(16,7)	(16,6)	(14,8)	(16,7)	(16,7)
Input format	`float32`	`float32`	`float32`	(10,4)	(10,4)	(10,4)	(12,4)	(12,4)
IQR	0.3251	0.3248	0.2769	0.2796	0.2804	0.2814	0.2820	0.2847
log MAE	0.5225	0.2647	0.2290	0.2307	0.2313	0.2325	0.2325	0.2353
p_{T}^{50} [GeV]	154.6	103.0	102.7	103.3	103.4	103.7	103.2	102.8
p_{T}^{95} [GeV]	196.4	138.1	134.1	135.3	135.5	136.5	136.5	137.0
p_{T}^{99} [GeV]	223.7	157.3	148.2	150.0	149.5	155.2	154.1	156.0
Sharpness [GeV]	69.1	54.3	45.5	46.7	46.1	51.5	50.9	53.2

Table 4.2: Total resource usage and its breakdown for the Versal Premium VP1802 FPGA implementation, driven at a clock speed of 240 MHz [5]. The values in parentheses are the percentages of the resource of the given type available on the device.

Resource type	FF	LUT	BRAM	DSP
Versal VP1802 limit	6 721 729	3 360 896	4 941	14 352
Producer	1 563 (0.02%)	1 001 (0.03%)	3 (0.06%)	0 (0.0%)
AddrProd.	1 227 (0.02%)	762 (0.02%)	1.5 (0.03%)	0 (0.0%)
Φ	3 996 (0.06%)	2 206 (0.07%)	7 (0.14%)	13 (0.09%)
Latent-buffer	1 165 (0.02%)	506 (0.02%)	9 (0.18%)	0 (0.0%)
KSum	1 216 (0.02%)	547 (0.02%)	0 (0.0%)	0 (0.0%)
F-p1	9 146 (0.14%)	4 450 (0.13%)	25.5 (0.52%)	50 (0.35%)
F-p2	9 592 (0.14%)	6 535 (0.19%)	22 (0.45%)	43 (0.30%)
Consumer	240 (<0.01%)	409 (<0.01%)	0 (0.0%)	0 (0.0%)
Control SmartConnect	13 883 (0.21%)	10 550 (0.31%)	0 (0.0%)	0 (0.0%)
Memory SmartConnect	4 858 (0.07%)	2 829 (0.08%)	0 (0.0%)	0 (0.0%)
Total usage of Φ	3 996 (0.06%)	2 206 (0.07%)	7 (0.14%)	13 (0.09%)
Total usage of F	18 738 (0.28%)	10 985 (0.32%)	47.5 (0.97%)	93 (0.65%)
Total usage	47 604 (0.71%)	29 959 (0.89%)	68 (1.38%)	106 (0.74%)

Table 4.3: Resource usage for different Φ parallelization factors on the Versal Premium VP1802 FPGA implementation, driven at a clock speed of 240 MHz [5]. The values in parentheses are the percentages of the resource of the given type available on the device. The first two rows are taken from the Vivado implementation report [14], whereas the remaining rows are taken from the Vitis C synthesis report [13]. The FF and LUT usages from Vitis are omitted as they were found to be unreliable.

Φ parallelization factor	FF	LUT	BRAM	DSP
1	3 996 (0.06%)	2 206 (0.07%)	7 (0.14%)	13 (0.09%)
2	5 534 (0.08%)	3 297 (0.1%)	7 (0.14%)	27 (0.19%)
4	–	–	7 (0.14%)	54 (0.38%)
8	–	–	7 (0.14%)	106 (0.74%)
16	–	–	7 (0.14%)	211 (1.47%)
32	–	–	7 (0.14%)	419 (2.92%)
64	–	–	7 (0.14%)	835 (5.82%)
Rest of the model	43 608 (0.65%)	27 753 (0.82%)	61 (1.24%)	93 (0.65%)
Versal VP1802 limit	6 721 729	3 360 896	4 941	14 352

5 Discussion

The central result is that a DeepSets calibration improves the jet energy resolution, whereas the binned median estimator does not. The binned median estimator returns an IQR of 0.3248, statistically indistinguishable from the uncalibrated value of 0.3251, while the full-precision offline model reaches 0.2769. This is expected: the binned median estimator, like the algorithm used in the standard anti-k_t calibration, applies a single multiplicative correction per $p_{\mathrm{T}}^{\mathrm{EM}}$ bin and can only shift the centre of the response distribution. The DeepSets model has access to the internal structure of the jet through its constituent topo-clusters, and the ablation scans confirm that this is where the gain originates, since constituent-level inputs improve on jet-level information alone. The improvement translates into the trigger behaviour anticipated, with the turn-on sharpness for a single-jet 100 GeV selection improving from 54.3 GeV to 45.5 GeV and the 99 % efficiency point moving from 157.3 GeV to 1148.2 GeV. One feature runs counter to this picture: the networks show slightly worse median closure than the binned median estimator below 20 GeV, which is unsurprising given that the baseline is median-unbiased by construction. The trade of closure for spread is favourable well above threshold, but would need revisiting for selections operating at low jet p_{T}.

What matters for a trigger application is how cheaply this gain can be retained, and the answer is that it is very cheap. Performance plateaus below the (3×32) layout, and the `Tiny` network reproduces the IQR of the offline model to within 0.0051 using 281 parameters rather than 569k. Post-training quantization costs little further, with performance minimally affected above 16 bits for the weights and activations and 12 bits for the inputs.

The firmware implementation bears this out, occupying 47 604 FF, 29 959 LUT, 68 BRAM, and 106 DSP on the Versal Premium VP1802, comfortably within the target budget of 336 086 FF, 168 044 LUT, 247 BRAM, and 717 DSP. The true margin is larger still, since the Producer, AddrProducer, and Consumer blocks emulate the trigger pipeline and would be absent from a deployed design. Throughput is set by Φ, which processes one cluster per pass and can be evaluated close to 40 times per jet. Replicating it scales the DSP usage approximately linearly while leaving the BRAM usage fixed, since the network parameters are shared between copies. This would allow for a parallelization of Φ by a factor of 55 before exceeding

resource requirement. This threshold aligns with the expected cluster-to-jet ratio, beyond which the F layer would become the new throughput bottleneck.

Resource utilization alone, however, cannot establish feasibility. Neither the end-to-end latency nor the initiation interval of the design is measured against the 2 μs target, and since the latency budget is the constraint that most sharply distinguishes trigger-level from offline calibration, the feasibility argument is incomplete without it. A related gap separates the two halves of the study: the network implemented in firmware is not one of those benchmarked for physics performance, differing in parameter count, bit width, layer configuration, and output structure. The study shows that a network of this scale (3 032 parameters) fits the budget, and separately that a network of this scale (from 3 017 to 113 parameters) delivers performance improvements.

Several further qualifications bound the conclusions. The binned median estimator is a proxy for the standard offline anti-k_t pipeline rather than the pipeline itself, so the quoted improvements cannot be read as improvements relative to the calibration deployed in ATLAS. Notably, the simulation uses Run 3 detector conditions rather than the HL-LHC geometry, and a single dijet sample from a single generator for which no systematic uncertainties are evaluated. Moreover, trigger performance is only characterized for single-jet jet threshold. Methodologically, only post-training quantization is applied, so the reported bit widths are an upper bound. Quantization aware training (QAT) and pruning are not explored.

The priorities for future work follow directly. A single quantized trigger model should be carried through to firmware and characterized for latency as well as resources, and benchmarked against the full offline pipeline. Training on HL-LHC detector conditions and a wider range of processes would test the generality of the calibration. Extending the method to cluster-level calibration would allow both stages of the existing chain to be treated in one framework, and evaluating trigger selections operating at low jet p_{T}, such as those based on jet multiplicity. Quantization-aware training and pruning are expected to reduce the resource footprint further before integration with the surrounding L0 Global Trigger firmware.

6 Conclusion

The HL-LHC will deliver an average of 200 inelastic collisions per bunch crossing to ATLAS, and the Level-0 Trigger must select events under those conditions within a fixed latency and with the limited resources available in firmware. Jet selections depend on the accuracy and resolution of the jet energy estimate available at trigger level, and the calibration applied at present corrects the average response while leaving the resolution largely unchanged.

This note has evaluated a DeepSets architecture as a candidate calibration, using simulated dijet events at a centre-of-mass energy of 14 TeV with an average of 200 interactions per bunch crossing. The permutation-invariant structure matches the variable-size, unordered collections of topological clusters that constitute a jet, allowing constituent-level information to inform the correction. The trained models improve the energy resolution across the transverse momentum spectrum, which the binned median baseline does not, and sharpen the efficiency turn-on for a single-jet selection. The performance is remarkably resilient to compression and quantization, and a firmware implementation of a network of comparable scale occupies well under the target of five per cent of the device that will host the Level-0 Global Trigger.

The physics performance and the resource footprint are therefore each compatible with the requirements of the Level-0 Trigger. It remains to be shown that a single network satisfies both simultaneously, and that the latency target is met, since the timing of the implemented design was not measured. Extending the evaluation to HL-LHC detector conditions, to a wider range of processes, and to an explicit assessment of pile-up robustness are the necessary steps before the method could be considered for deployment.

7 Acknowledgments

I would like to thank Prof. Colin Gay, Prof. Max Swiatlowski, and Master's candidate Kelvin Leong of the University of British Columbia and TRIUMF for their constant support and for the opportunity to participate in their research group. I am also grateful to the Canadian Institute of Particle Physics and CERN for organizing and partially funding this opportunity. I also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Bibliography

[1] ATLAS Collaboration, *The ATLAS Upgrade for the HL-LHC*, ATL-UPGRADE-PUB-2025-001, 2025, URL: https://cds.cern.ch/record/2928798 (cit. on p. 3).

[2] ATLAS Collaboration,
Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System,
CERN-LHCC-2017-020, ATLAS-TDR-029, 2017,
URL: https://cds.cern.ch/record/2285584 (cit. on pp. 3, 4).

[3] ATLAS Collaboration,
Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1,
The European Physical Journal C **77** (2017), ISSN: 1434-6052,
URL: http://dx.doi.org/10.1140/epjc/s10052-017-5004-5 (cit. on pp. 3, 4).

[4] ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s}$ = 13TeV with the ATLAS detector*, The European Physical Journal C **81** (2021), ISSN: 1434-6052,
URL: http://dx.doi.org/10.1140/epjc/s10052-021-09402-3 (cit. on pp. 3, 4).

[5] Advanced Micro Devices Inc., *VPK180 Evaluation Board User Guide, v. 1.3*,
Accessed: August 19, 2026, 2026,
URL: https://docs.amd.com/r/en-US/ug1582-vpk180-eval-bd (cit. on pp. 4, 12–14).

[6] M. Zaheer et al., *Deep Sets*, 2018, arXiv: 1703.06114 [cs.LG],
URL: https://arxiv.org/abs/1703.06114 (cit. on pp. 4, 6).

[7] F. Rosenblatt,
The perceptron: A probabilistic model for information storage and organization in the brain,
Psychological Review **65(6)** (1958), ISSN: 386-408,
URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0042519 (cit. on p. 6).

[8] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*,
Journal of High Energy Physics **2019** (2019), ISSN: 1029-8479,
URL: http://dx.doi.org/10.1007/JHEP01(2019)121 (cit. on p. 6).

[9] M. I. Jordan,
Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986,
tech. rep., California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, 1986,
URL: https://www.osti.gov/biblio/6910294 (cit. on p. 6).

[10] Martín Abadi et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*,
Software available from tensorflow.org, 2015, URL: https://www.tensorflow.org/
(cit. on p. 7).

[11] V. Nair and G. E. Hinton, 'Rectified linear units improve restricted boltzmann machines',
ICML 2010, 2010 807, URL: https:
//www.bibsonomy.org/bibtex/2a0b7deb2839b69a52fcfcc15b7277a2a/georgheyer
(cit. on p. 7).

[12] P. J. Huber, *Robust Estimation of a Location Parameter*,
Annals of Mathematical Statistics **35** (1964) 492,
URL: https://api.semanticscholar.org/CorpusID:121252793 (cit. on p. 7).

[13] Advanced Micro Devices Inc., *Vitis High-Level Synthesis (HLS), v. 2025.2.1*,
Software. Accessed: August 19, 2026,
URL: https://www.amd.com/en/products/software/adaptive-socs-and-fpgas/vitis/vitis-hls.html (cit. on pp. 9, 14).

[14] Advanced Micro Devices Inc., *AMD Vivado Design Suite, v. 2025.2.1*,
Software. Accessed: August 19, 2026, URL: https://www.amd.com/en/products/software/adaptive-socs-and-fpgas/vivado.html
(cit. on pp. 9, 14, 23).

[15] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*,
SciPost Phys. Codebases (2022) 8,
URL: https://scipost.org/10.21468/SciPostPhysCodeb.8 (cit. on p. 20).

[16] C. Bierlich et al., *Codebase release 8.3 for PYTHIA*, SciPost Phys. Codebases (2022) 8,
URL: https://scipost.org/10.21468/SciPostPhysCodeb.8-r8.3 (cit. on p. 20).

Appendices

A Monte Carlo dijet samples

The simulated dijet samples used in this study are listed in Table A.1. All samples were generated with Pythia 8 [15, 16] and are divided into slices of the leading-jet transverse momentum, denoted JZ0 to JZ4, in order to populate the full spectrum with adequate statistical precision.

Table A.1: Simulated dijet samples used in this study.

Slice	Dataset identifier
JZ0	`mc21_14TeV.801165.Py8EG_A14NNPDF23LO_jj_JZ0`
JZ1	`mc21_14TeV.801166.Py8EG_A14NNPDF23LO_jj_JZ1`
JZ2	`mc21_14TeV.801167.Py8EG_A14NNPDF23LO_jj_JZ2`
JZ3	`mc21_14TeV.801168.Py8EG_A14NNPDF23LO_jj_JZ3`
JZ4	`mc21_14TeV.801169.Py8EG_A14NNPDF23LO_jj_JZ4`

B Quantization studies

This appendix supplements the bit-width optimization presented in Section 4.2. Figure B.1 quantifies the agreement between the quantized and floating-point implementations of the `Tiny` model at the operating point selected in Table 4.1, and Figure B.2 isolates the sensitivity of the model to the precision of each input feature individually.

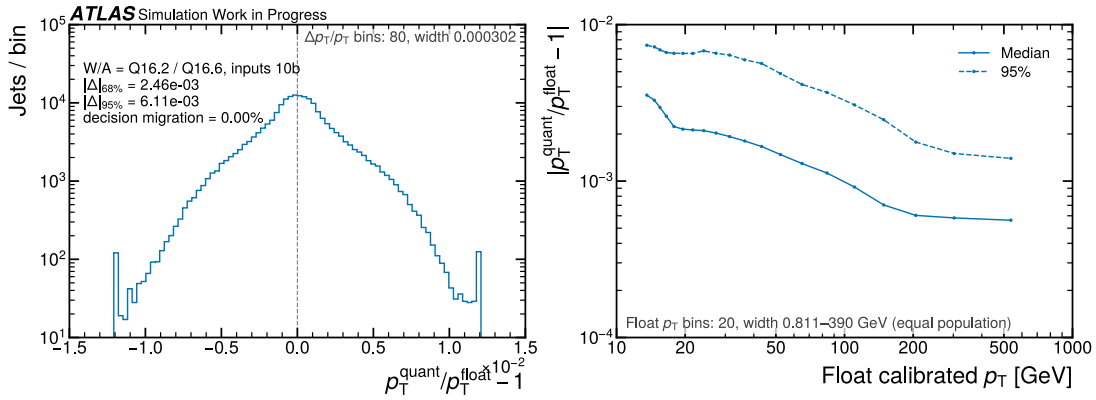


Figure B.1: Agreement between the fixed-point and floating-point implementations of the `Tiny` DeepSets model, with 281 parameters, under post-training quantization at $(16, 3)$ weights, $(16, 7)$ activations, and $(12, 4)$ inputs. The left panel shows the per-jet deviation $|p_{\mathrm{T}}^{\mathrm{quant}}/p_{\mathrm{T}}^{\mathrm{float}} - 1|$, together with its 68 % and 95 % quantiles, denoted $|\Delta|_{68\%}$ and $|\Delta|_{95\%}$, and the fraction of jets whose trigger decision migrates at an accept fraction of 15 %. The right panel shows the median and the 95 % quantile of $|\Delta|$ in equal-population bins of the floating-point calibrated p_{T}.

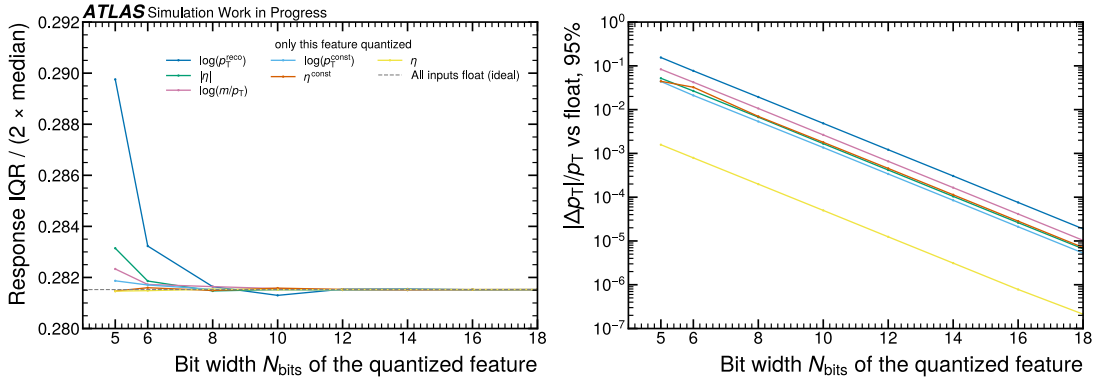


Figure B.2: Sensitivity of the `Tiny` DeepSets model, with 281 parameters, to the input precision of each feature. At each point only the named feature is quantized to the quoted width, with all other inputs and the model itself left in floating point, so that each curve isolates the tolerance of a single input to precision loss. The integer widths are fixed from the measured dynamic range of each feature. The left panel shows the energy resolution (IQR), with the all-floating-point value shown for reference. The right panel shows the 95 % quantile of the per-jet deviation from the floating-point prediction.

C Model characterization

This appendix supplements the final benchmarks of Section 4.3 by resolving the performance across the transverse momentum spectrum and the detector acceptance.

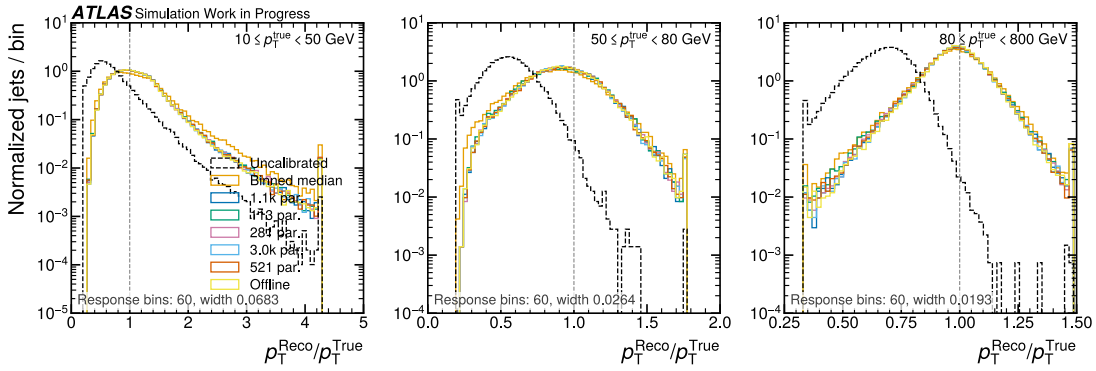


Figure C.1: Normalized jet response, $p_{\mathrm{T}}^{\mathrm{DeepSets}}/p_{\mathrm{T}}^{\mathrm{True}}$, in windows of truth jet transverse momentum of 10–50, 50–80, and 80–800 GeV, for the uncalibrated response, the binned median baseline, and the models defined in Table 4.1.

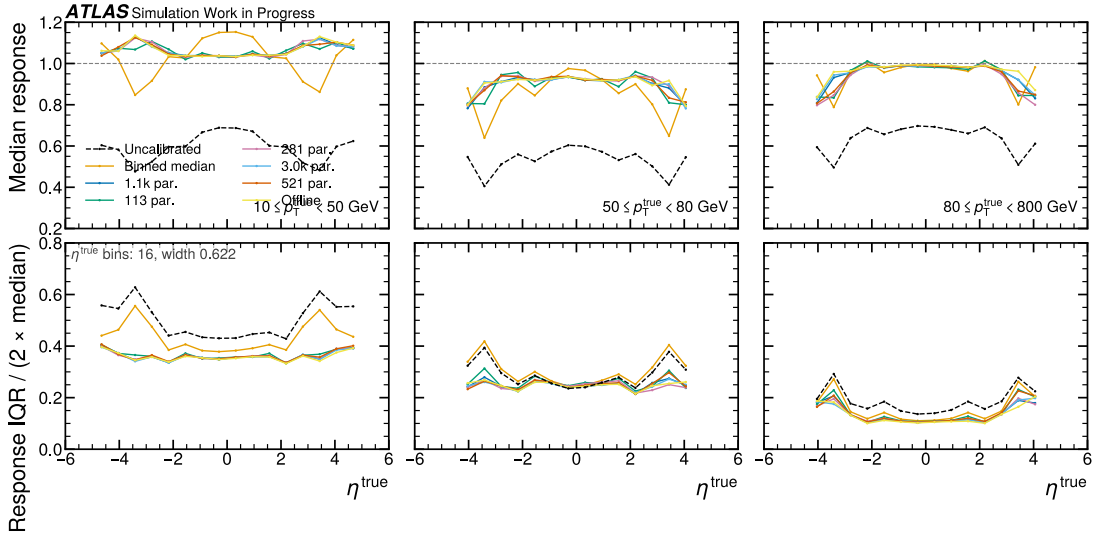


Figure C.2: Median response (top) and energy resolution (bottom) as a function of the truth jet η, in windows of truth jet transverse momentum of 10–50, 50–80, and 80–800 GeV, for the uncalibrated response, the binned median baseline, and the models defined in Table 4.1.

D Block Design

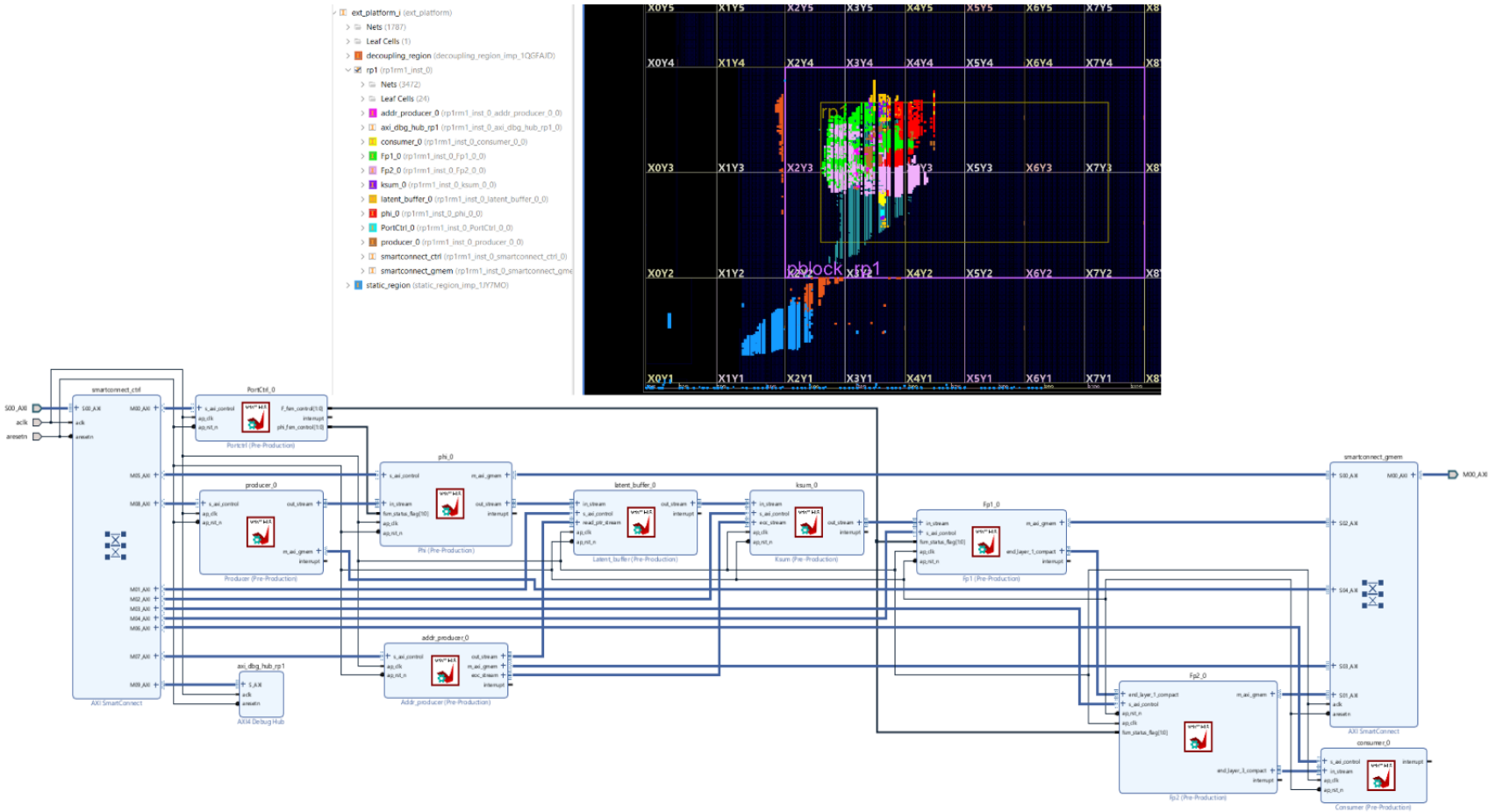


Figure D.1: Vivado [14] implementation floor plan (top) and block design (bottom) of the programmable logic for the DeepSets model implementation. Each hardware block is connected by black lines which propagate the clock and the active-low reset and by blue lines which account for control signals as well as the flow of data through the pipeline. For development, this partition is attached to a larger network of static Versal Premium VPK180 Development Kit HLS blocks (not shown here). The static and programmable logic regions are decoupled by DFx AXI Shutdown Managers (not shown here).