ETH zürich[1] Imperial College London[2] UC San Diego[3] MIT[4] UERJ[5] CERN

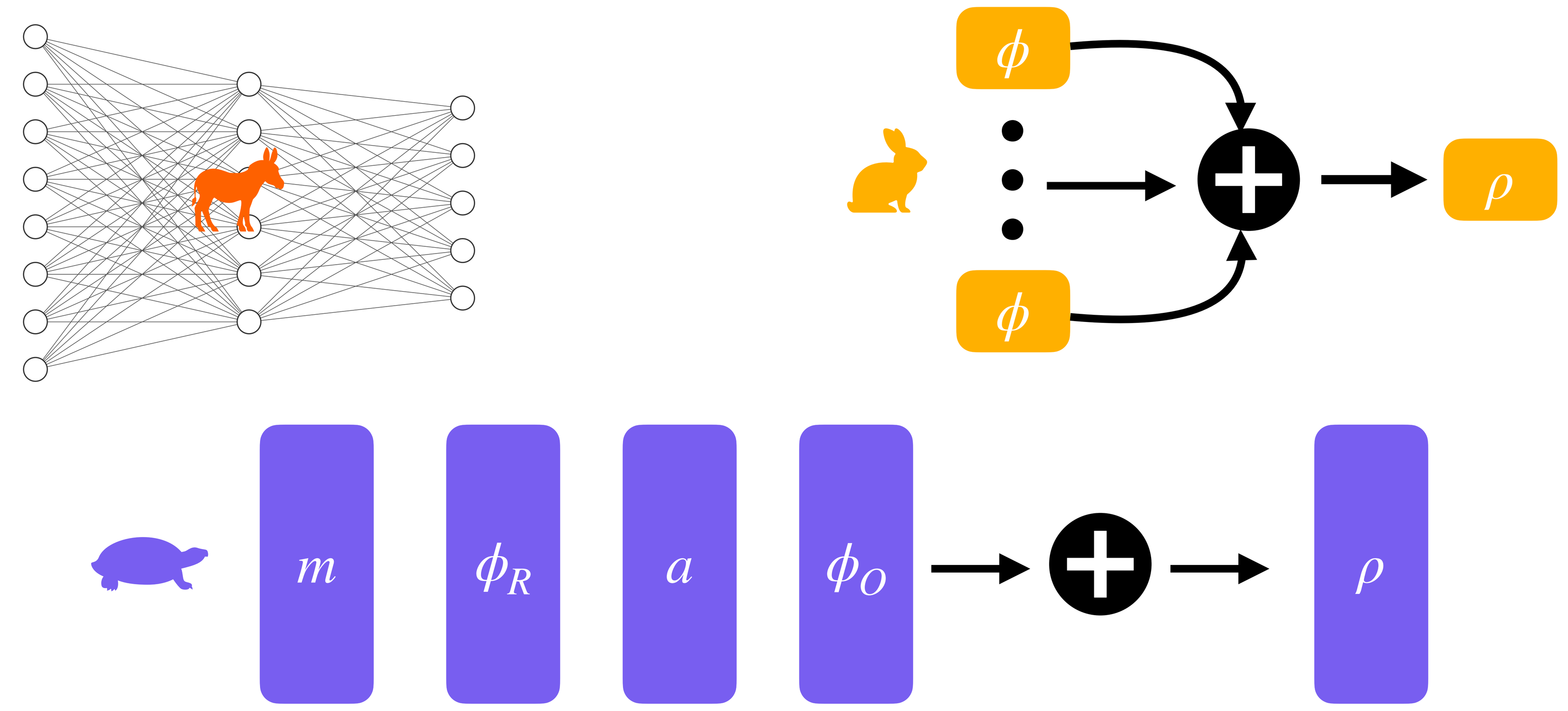
IAIFI Summer Workshop August 12–August 16 2024

Geometric Learning for Ultrafast Jet Classification at the HL-LHC

P. Odagiu[1], Z. Que[2], J. Duarte[3], V. Loncar[4], A. Sznajder[5], T. Aarrestad[1]

Introduction

Three machine learning models are used to perform jet origin classification, increasing the sensitivity at the HL-LHC. These models are optimised for deployment on a specific field-programmable gate array device. Through quantisation-aware training and efficient synthetisation, we show that O(100) of geometric ML architectures such as Deep Sets and Interaction Networks is feasible at a relatively low computational cost.

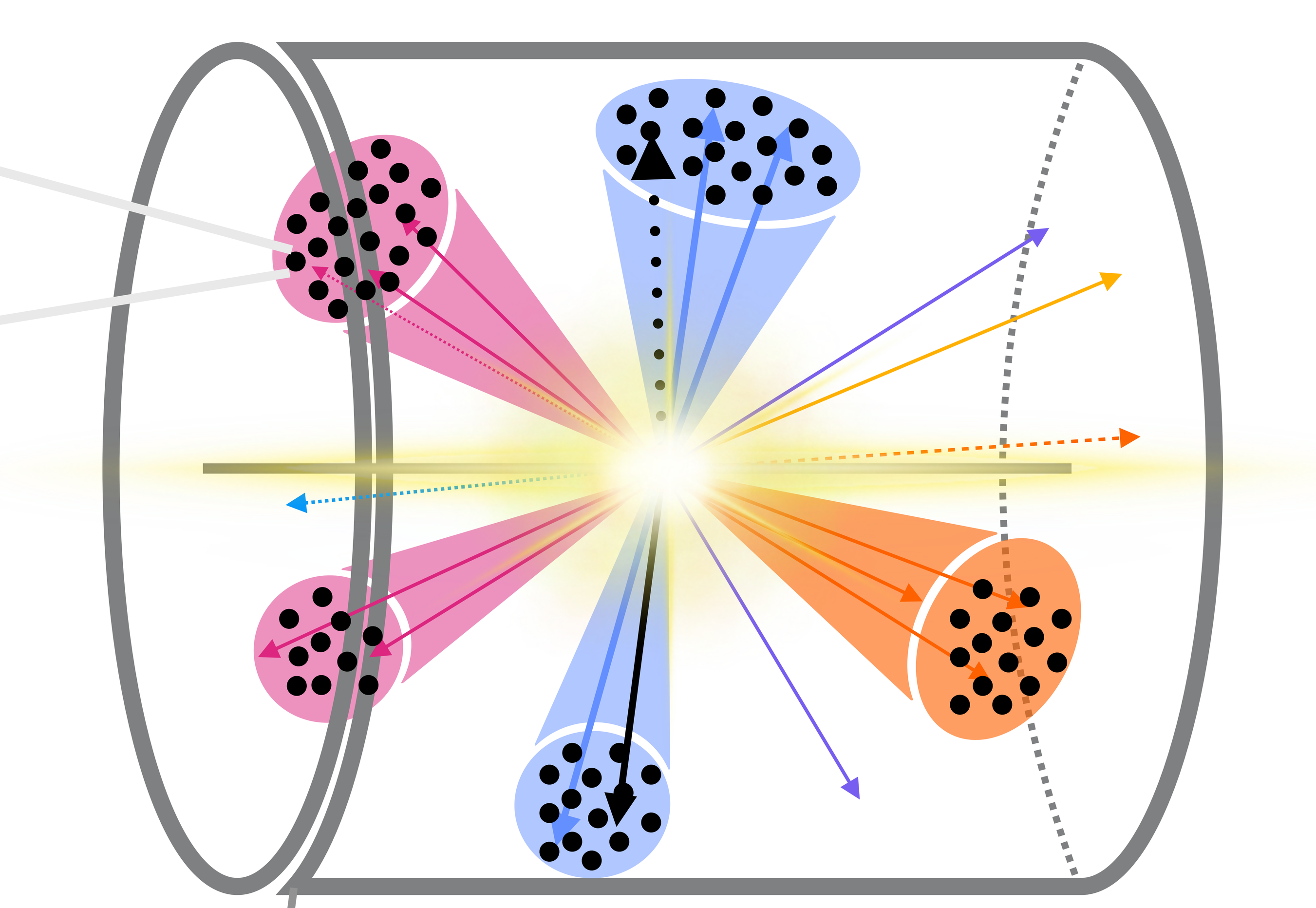


Data

The data set consists of particle jets as measured at the HL-LHC. The data has a dyadic structure: each jet has a number of particles, and each particle is specified by its kinematic properties: $p_T, \eta_{\rm rel}, \phi_{\rm rel}$. We truncate each jet to the N most energetic particles, and consider three realistic scenarios: 8, 16, 32.

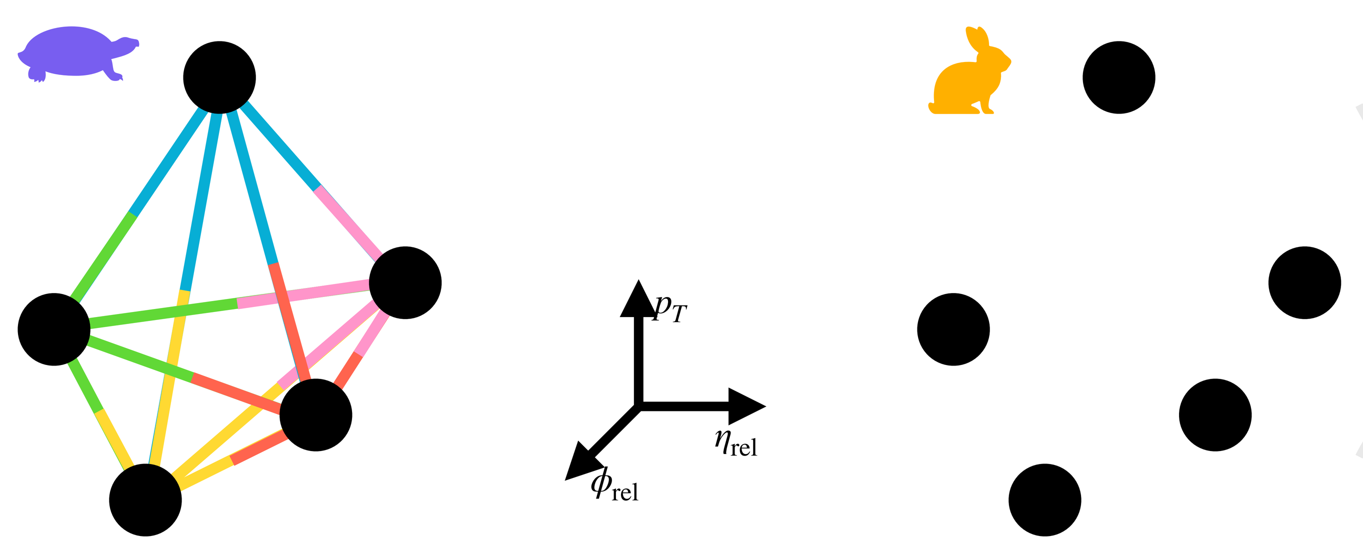
$$p_T\ \eta_{\rm rel}\ \phi_{\rm rel}$$

$$N \in \{8, 16, 32\}$$

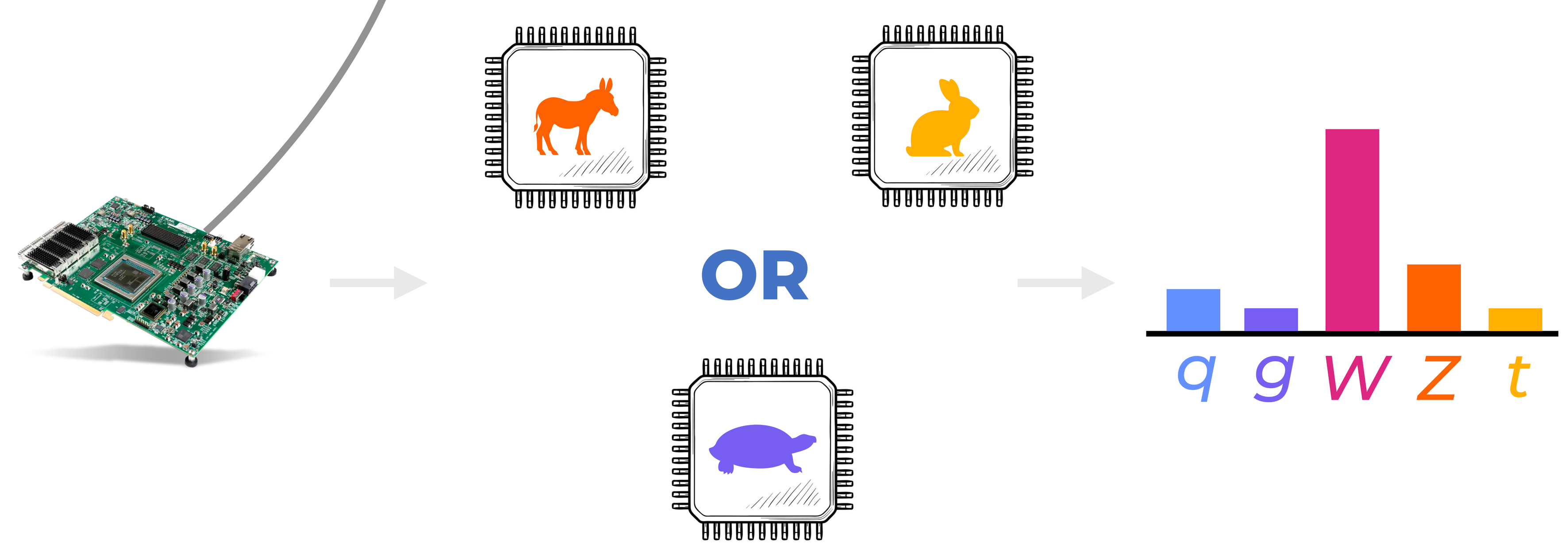


Representations

$$[p_T^1, \eta_{\rm rel}^1, \phi_{\rm rel}^1, \ldots, p_T^N, \eta_{\rm rel}^N, \phi_{\rm rel}^N]$$



The data can be represented in $p_T, \eta_{\rm rel}, \phi_{\rm rel}$ space in three ways: tabular, set, and fully-connected graph. These representations correspond to different deep learning architectures: multi-layer perceptron, deep sets, and interaction network, respectively. There exist other representations of the data; however, these three are arguably the simplest.



Odagiu P., Que Z., Duarte J., Loncar V., Sznajder A., Aarrestad T., et. al., Ultrafast Jet Classification at the HL-LHC, Machine Learning: Science and Technology.

Quantisation

Real-time jet tagging imposes a challenging constraint on the three architectures considered in this work: they need to perform inference in approximately 100 ns. For this reason, the models are synthesised on field-programmable gate arrays. Hence, the weights of the models are quantised to fit into the resource constraints of the FPGA, and quantisation-aware training is performed, along with pruning in some cases.

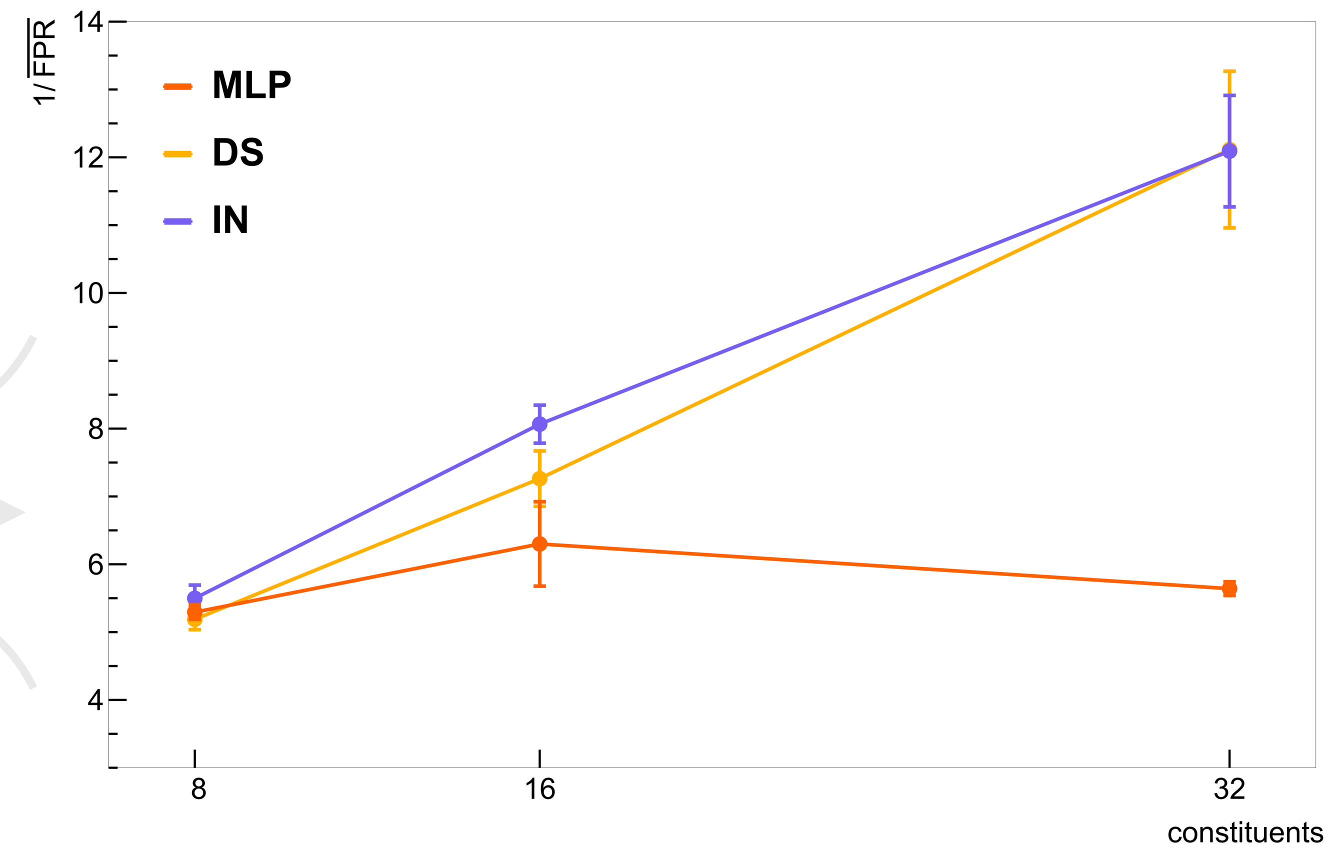
±	8 bits	23 bits
	exponent	mantissa

Results

Architecture	Constituents	Latency [ns] (cc)	LUT
	8	105 (21)	155,080 (9.0%)
	16	100 (20)	146,515 (8.5%)
	32[a]	105 (21)	155,080 (7.2%)
	8	95 (19)	386,294 (22.3%)
	16	115 (23)	747,374 (43.2%)
	32[a]	130 (26)	903,284 (52.3%)
	8	160 (32)	472,140 (27.3%)
	16	180 (36)	1,387,923 (80.3%)
	32[a]	205 (41)	1,162,104 (67.3%)

[a] Pruning

proxy for FPGA resources



Conclusion

The tabular representation loses useful information, while the fully-connected graph representation introduces additional structure that makes the associated network too cumbersome. The set representation is ideal for fast jet classification.