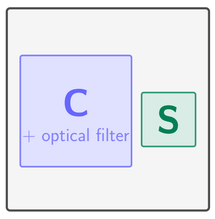
Machine Learning Enables Real-Time Waveform Decomposition for Dual-Readout Calorimetry

Liangyu Wu [1,2], Qibin Liu [1], Marco T. Lucchini [3,4], Julia Gonski [1], Marcello Campajola [5,6], Stefano Moneta [7]

[1]SLAC National Accelerator Laboratory [2]Stanford University [3]Università degli Studi di Milano-Bicocca [4]INFN Sezione di Milano-Bicocca
[5]Università degli Studi di Napoli Federico II [6]INFN Sezione di Napoli [7]INFN Sezione di Perugia
liangyu@slac.stanford.edu arXiv:2604.26090

Introduction & Motivation

The **FCC IDEA** calorimeter design [1] incorporates **dual-readout crystals** [2] with separate Cherenkov (C) and scintillation (S) SiPM readout [3] for event-by-event EM fraction correction. Full waveform digitization is required to separate the overlapping C and S components, but the fine detector granularity creates prohibitive data rates at the front-end.



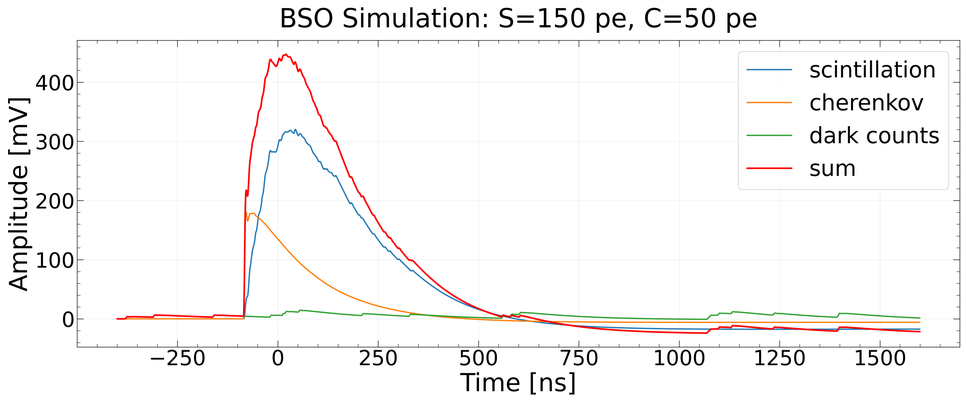
Crystal Dual-readout SiPM layout
(Rear crystal segment)

We demonstrate that **compact neural networks deployed in the front-end electronics** can perform this decomposition in real time, achieving superior performance to template-based fitting even at **significantly reduced ADC sampling rates**.

Simulation & Signal Model

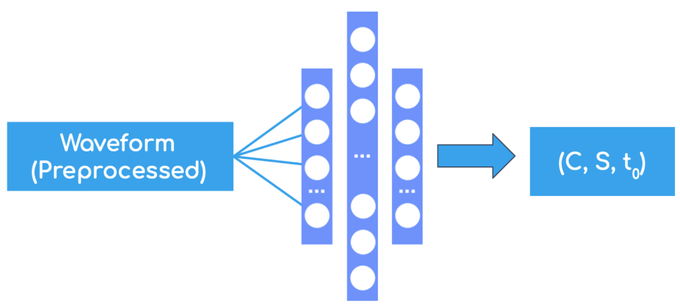
Composite waveforms are synthesized by combining Geant4-simulated shower profiles with crystal-specific scintillation decay constants and a measured single-photon response (SPR) for SiPM pulses [4]. Simulations cover e^- and K_L^0 showers at 1, 5, 10, and 30 GeV, digitized at 3.125 GHz over a 2000 ns window. Realistic SiPM dark counts (10 MHz) and timing jitter (2.5 ns) are included to match expected detector conditions.

Crystal	Scint. Yield	Decay Time	Cherenkov Yield
BGO	170 phe/GeV	τ_1=50 ns, τ_2=320 ns	50 phe/GeV
BSO	30 phe/GeV	τ_1=22 ns, τ_2=98 ns	50 phe/GeV
PWO	30 phe/GeV	τ=7.5 ns (single exp.)	50 phe/GeV



Edge-ML Design & (e)FPGA Deployment

A fully-connected DNN is chosen for its fixed-latency, fully parallelizable inference — ideal for real-time front-end deployment. The network directly regresses (c, s, t_0) from pre-processed waveform inputs.



Training uses 140k waveforms per energy with a weighted MSE loss and Adam optimizer with linear warmup. A single model trained on the combined 1–30 GeV dataset generalizes across energies without retraining. Performance is evaluated across ADC rates from 312.5 MHz down to 10.4 MHz.

Models are compressed with magnitude pruning and quantization-aware training, then synthesized to FPGA firmware using `hls4ml` [5, 6]. All configurations achieve $\leq$25 ns inference latency with **zero DSP** usage.

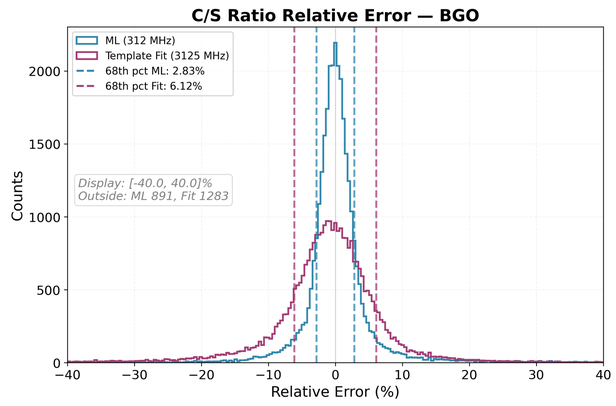
Crystal	Rate	Compression	Freq [MHz]	LUT	FF	DSP	ML err_{68}	Fit err_{68}
BGO	10 MHz	40%, ⟨10,1⟩	57.5	36,879	213	0	5.84%	6.12%
	312 MHz	30%, ⟨10,2⟩	46.7	312,914	3,933	0	**3.08%**	6.12%
BSO	10 MHz	50%, ⟨10,2⟩	57.7	30,471	213	0	8.14%	6.85%
	312 MHz	30%, ⟨10,4⟩	46.6	328,953	3,933	0	**4.18%**	6.85%
PWO	10 MHz	50%, ⟨10,2⟩	57.7	30,321	213	0	11.47%	11.41%
	312 MHz	30%, ⟨10,4⟩	46.6	311,282	3,933	0	**7.40%**	11.41%

The $\leq$25 ns latency at 46–58 MHz clock is compatible with potential future collider bunch crossing rates. The modest LUT requirement of $\mathcal{O}$(30k) can be accommodated in compact FPGAs such as embedded FPGA (eFPGA) ASICs.

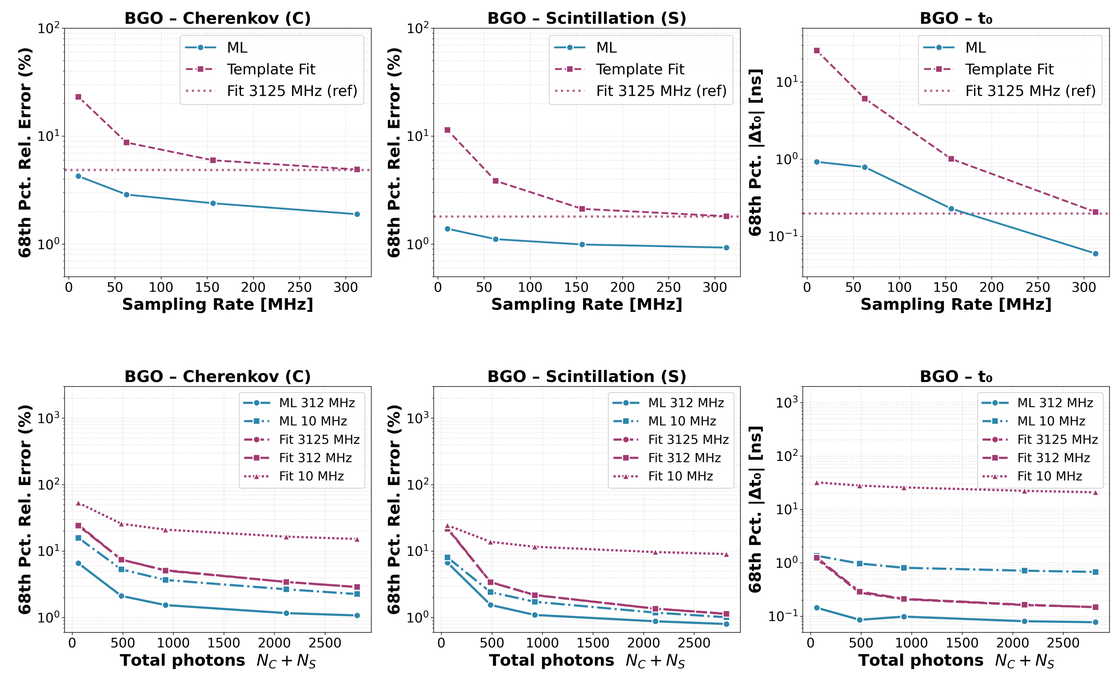
Performance: ML vs. Template Fitting

ML achieves **superior C/S separation** at substantially reduced sampling rates. At 312.5 MHz (10× reduction), the ML model outperforms template fitting for all 3 crystals at the full 3.125 GHz.

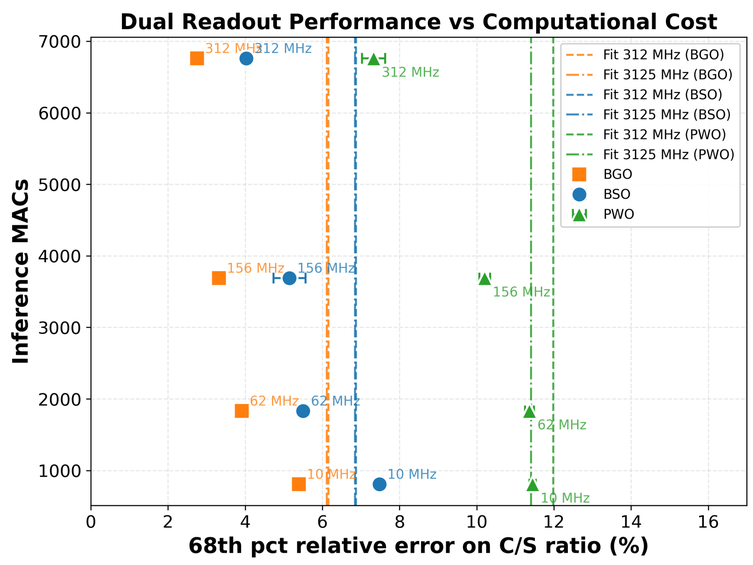
The advantage is consistent across all three crystals and the full ADC rate scan. Even at 10.4 MHz (300× reduction), ML remains competitive with template fitting at full rate.



A single model trained across 1–30 GeV generalizes robustly across all three crystals without per-energy retraining. This generalization capability is particularly valuable for applications where beam energy is not precisely known a priori.



The performance-resource trade-off demonstrates that ML models can maintain competitive C/S separation while dramatically **reducing hardware footprint requirements**.



Summary & Outlook

Compact neural networks effectively extract Cherenkov and scintillation components from dual-readout calorimeter waveforms, achieving better C/S resolution than template fitting at lower sampling rates. A single model trained across 1–30 GeV generalizes robustly without per-energy retraining, and FPGA synthesis delivers $\leq$25 ns inference latency with zero DSP usage — demonstrating **real-time feasibility** for FCC-ee front-end readout.

Future work will focus on test beam data validation and hardware characterization of eFPGA-based readout ASICs [7] to demonstrate end-to-end deployment in the IDEA detector [1] readout chain.

References

[1] M. Abbrescia *et al.*, "The IDEA detector concept for FCC-ee," 2025.
[2] M. T. Lucchini, W. Chung, S. C. Eno, Y. Lai, L. Lucchini, M.-T. Nguyen, and C. G. Tully, "New perspectives on segmented crystal calorimeters for future colliders," *JINST*, vol. 15, no. 11, p. P11005, 2020.
[3] A. Benaglia, F. Cetorelli, M. T. Lucchini, E. Auffray, L. Roux, and J. Delenne, "Characterization of thin optical filters for high purity Cherenkov light readout from scintillating crystals," *JINST*, vol. 21, no. 03, p. P03025, 2026.
[4] M. Alviggi *et al.*, "Cherenkov and scintillation light separation in BGO and BSO crystals coupled to SiPMs for dual-readout electromagnetic calorimetry at future colliders," 2026.
[5] J. Duarte *et al.*, "Fast inference of deep neural networks in FPGAs for particle physics," *JINST*, vol. 13, no. 07, p. P07027, 2018.
[6] FastML Team, "fastmachinelearning/hls4ml," 2024.
[7] J. Gonski, A. Gupta, H. Jia, H. Kim, L. Rota, L. Ruckman, A. Dragone, and R. Herbst, "Embedded FPGA developments in 130 nm and 28 nm CMOS for machine learning in particle detector readout," *Journal of Instrumentation*, vol. 19, no. 08, p. P08023, 2024.