ns AI for anomaly detection with decision trees on FPGA

Efficient real-time trigger design to save new phenomena without knowing it a priori

Tae Min Hong on behalf of co-authors

University of Pittsburgh

SCAN ME

Intro

Read our open-access paper [Nat. Commun. **15**, 3527 (2024)]

nature communications

Article

https://doi.org/10.1038/s41467-024-47704-8

Nanosecond anomaly detection with decision trees and real-time application to exotic Higgs decays

Received: 23 May 2023

Accepted: 9 April 2024

Published online: 25 April 2024

Check for updates

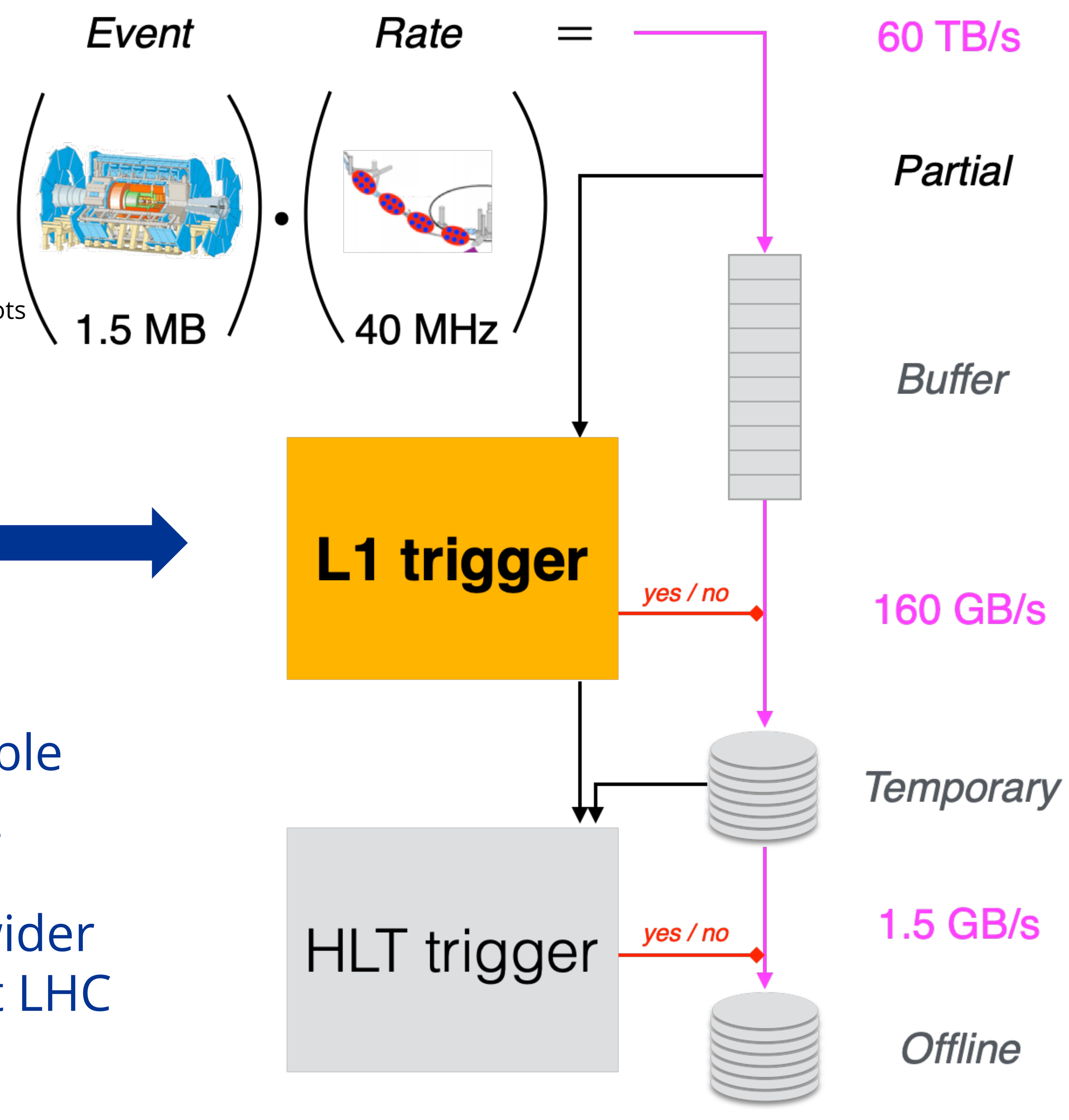
S. T. Roche[1,2], Q. Bayer[2], B. T. Carlson[2,3], W. C. Ouligian[2], P. Serhiayenka[2], J. Stelzer[2] & T. M. Hong[2]

We present an interpretable implementation of the autoencoding algorithm, used as an anomaly detector, built with a forest of deep decision trees on FPGA, field programmable gate arrays. Scenarios at the Large Hadron Collider at CERN are considered, for which the autoencoder is trained using known physical processes of the Standard Model. The design is then deployed in real-time trigger systems for anomaly detection of unknown physical processes, such as the detection of rare exotic decays of the Higgs boson. The latency is made with a latency value of 30 ns at percent-level resource usage using the Xilinx Virtex UltraScale+ VU9P FPGA. Our method offers anomaly detection at low latency values for edge AI users with resource constraints.

Context for low-latency real-time 40 MHz trigger

Example: ATLAS at LHC

cern.ch/twiki/pub/AtlasPublic/ApprovedPlotsDAQ/tdaq-run3-schematic-withoutFTK.pdf



Put AI on FPGA

IP & testbench available online at fwx.pitt.edu

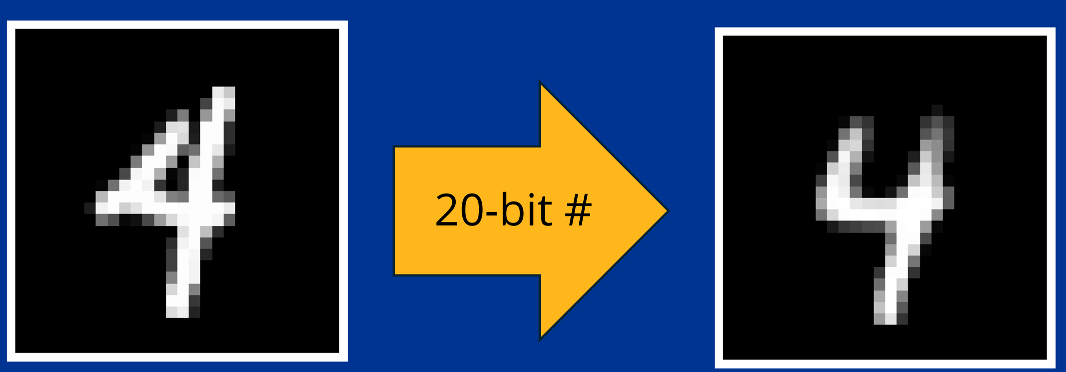
NB. Our work is for wider audiences, not just at LHC

New AI

Tree autoencoder on MNIST images

yann.lecun.com/exdb/mnist

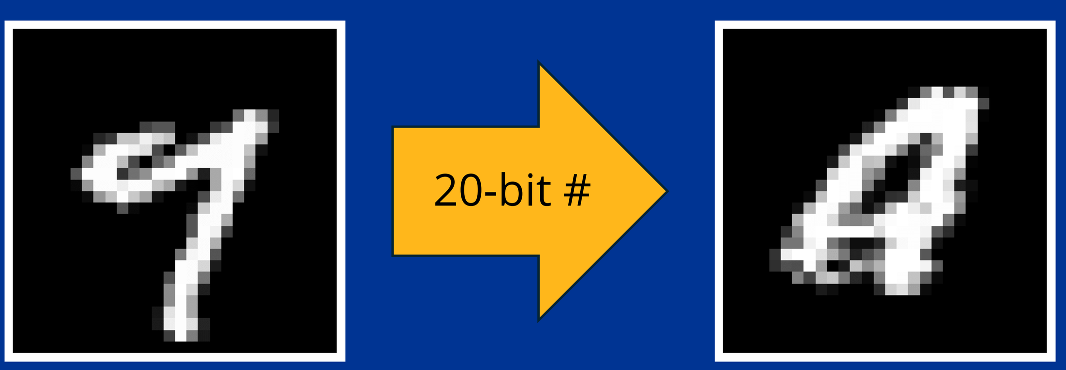
- Use handwritten 28×28 pixels of 8-bit greyscale, teach it 0–4
- Compress by 300x then decompress two images: "4" and "9"



| Known orig. – Est. | = small distance

Model knows "4"
Identifies known physics

Original 784 input var. Compress 300x *Decompressed estimate*

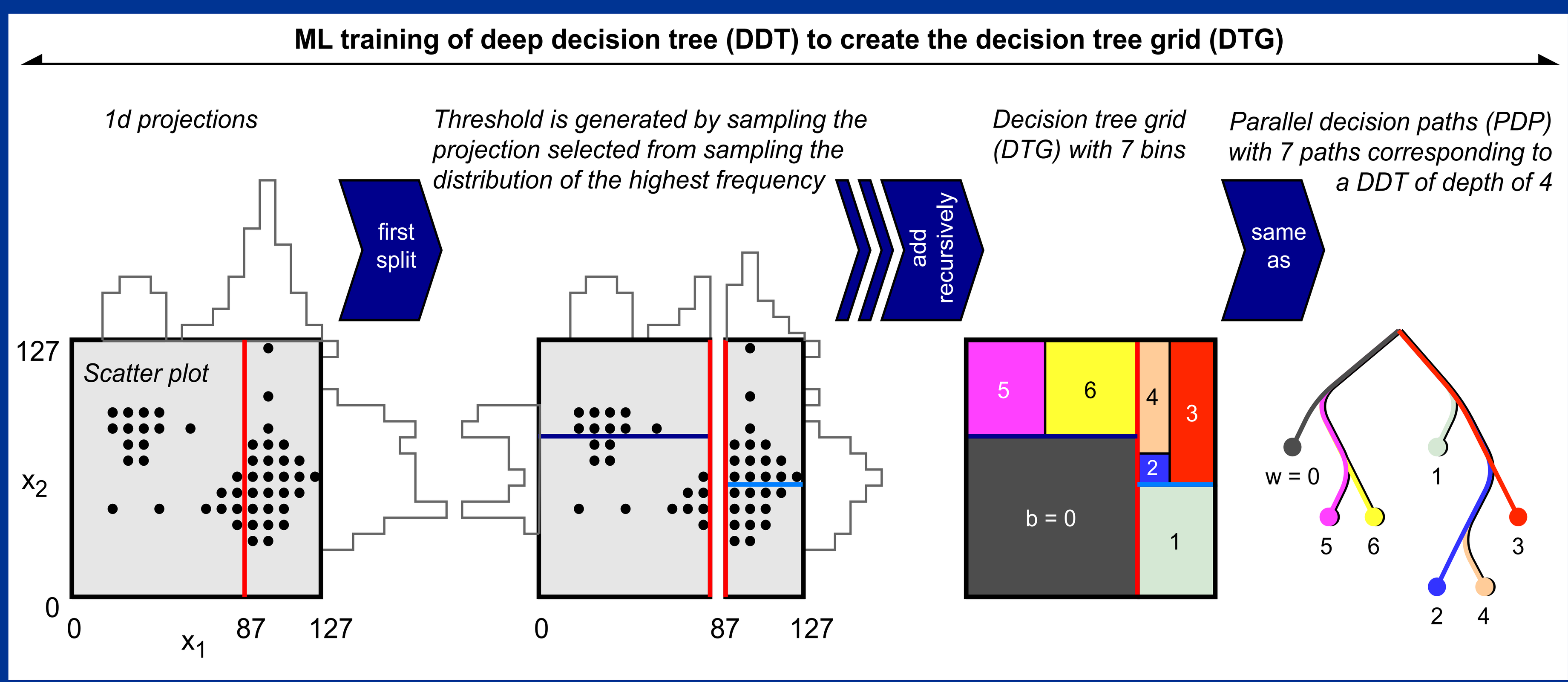


| Unknown orig. – Est. | = large distance

Model doesn't know "9"
Identifies new phenomena

Tree training by sampling 1d PDFs

- Unsupervised training using one sample, e.g., using known physics
- Iteratively split the training data by sampling its 1d projections

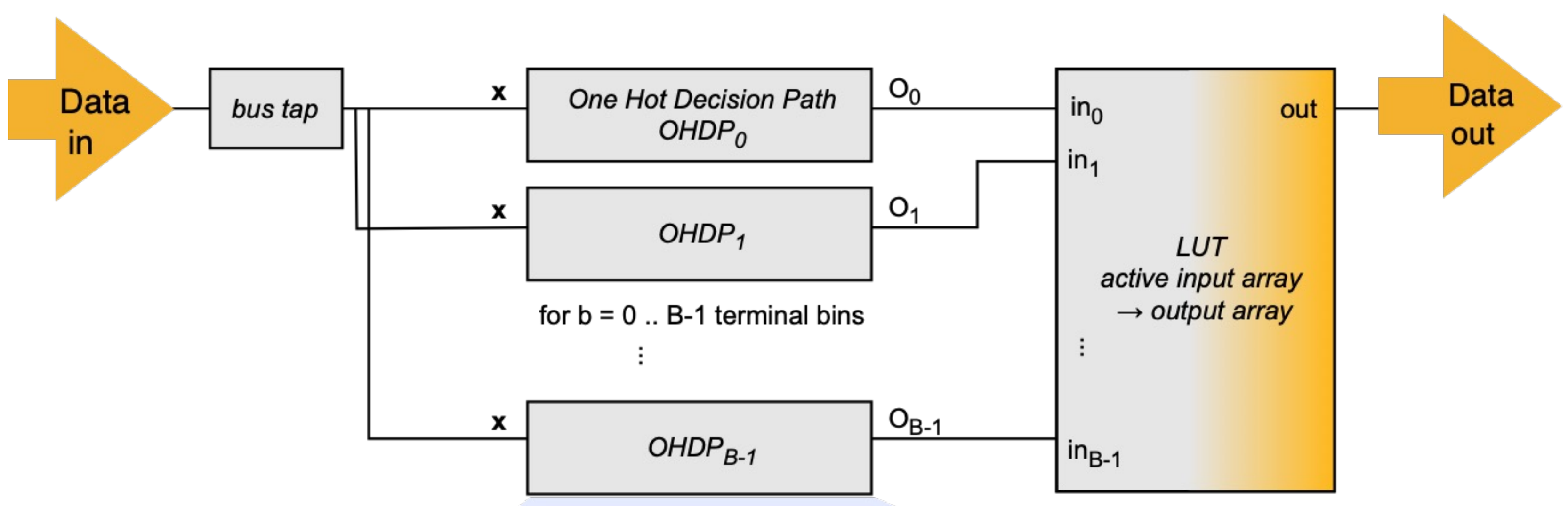


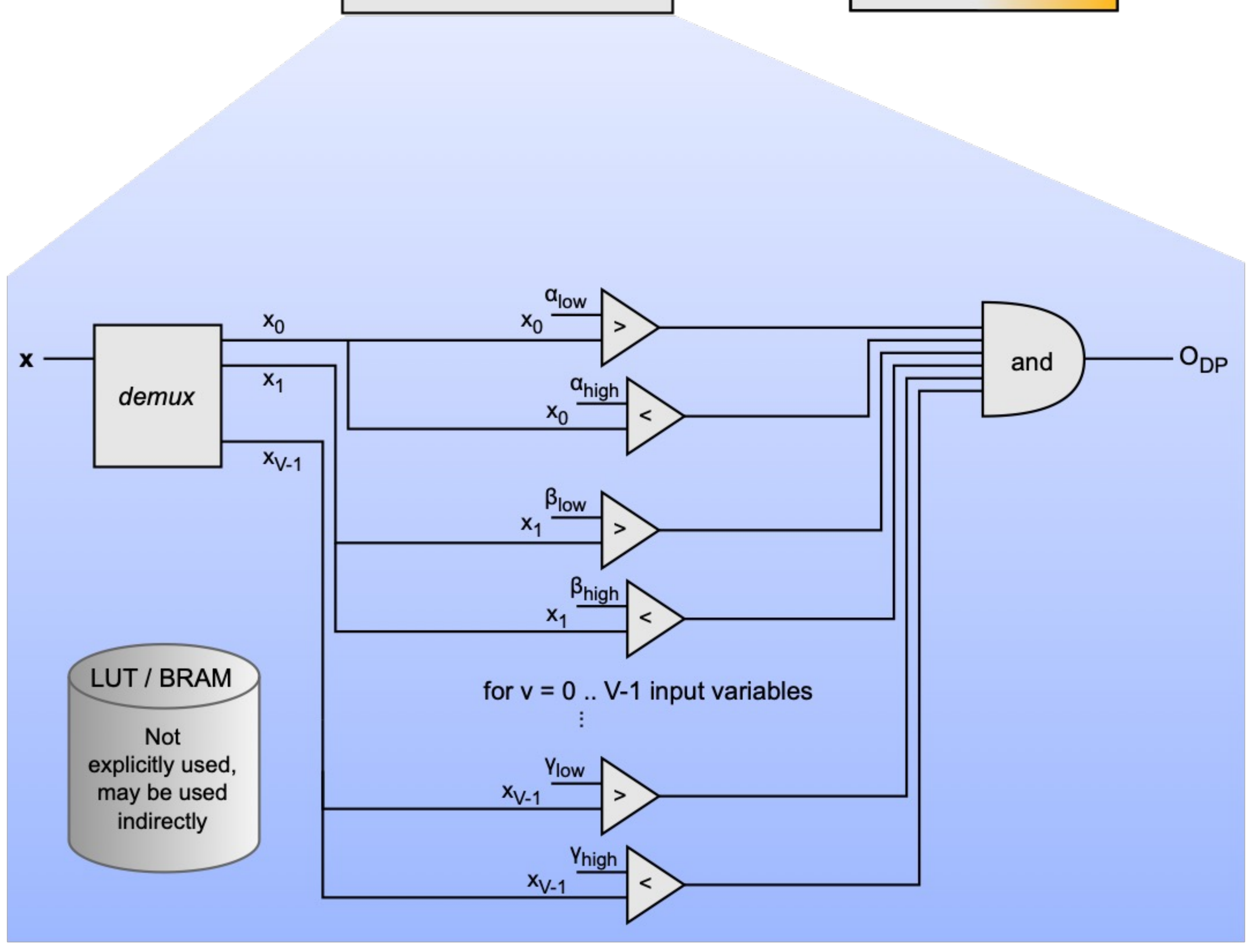
- For each bin, median value of the training data is the estimate
- Anomaly score = $\sum |\mathbf{x}_{\text{original}} - \mathbf{x}_{\text{estimate}}|$, $\mathbf{x}$ is the set of input variables

Application

Design

- Parallel decision path (PDP) [J. Instrum. **17**, P09039 (2022)]
  Composed to combinatoric logic (mostly threshold comparisons)



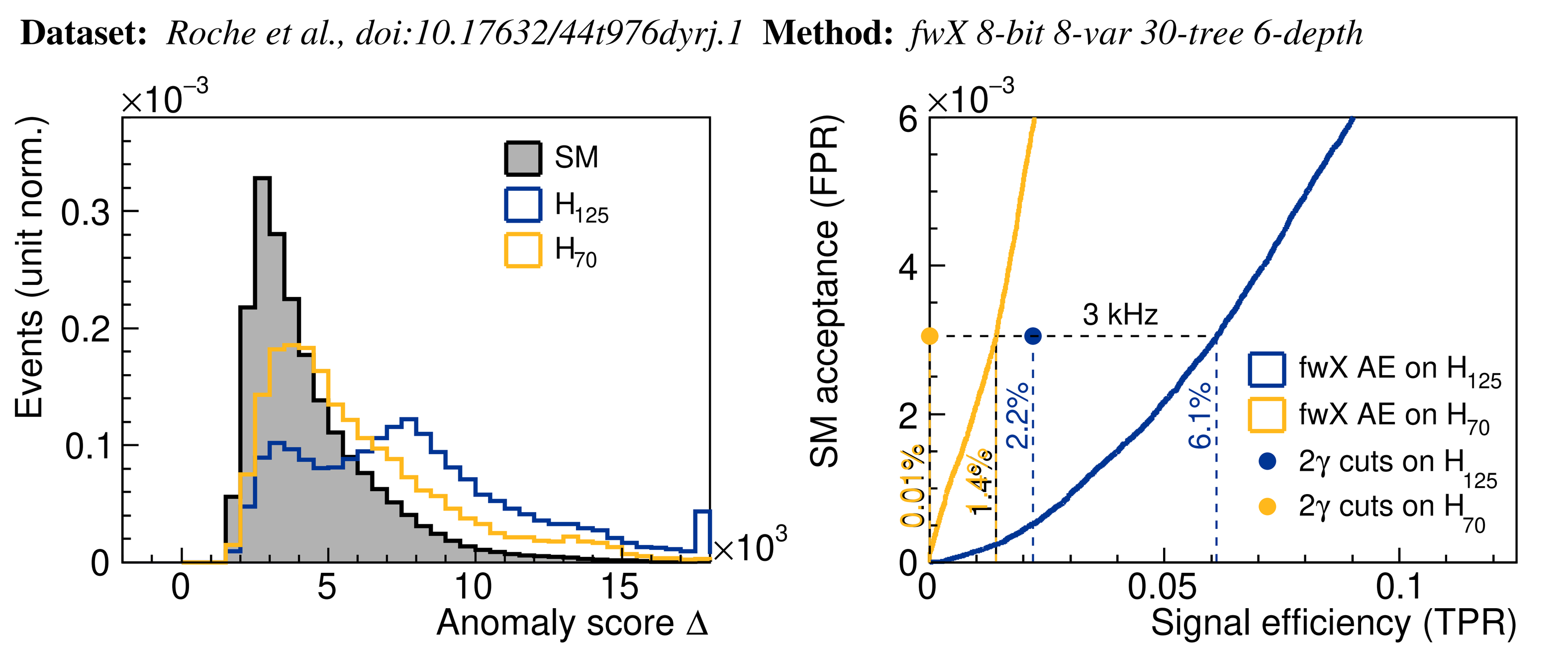


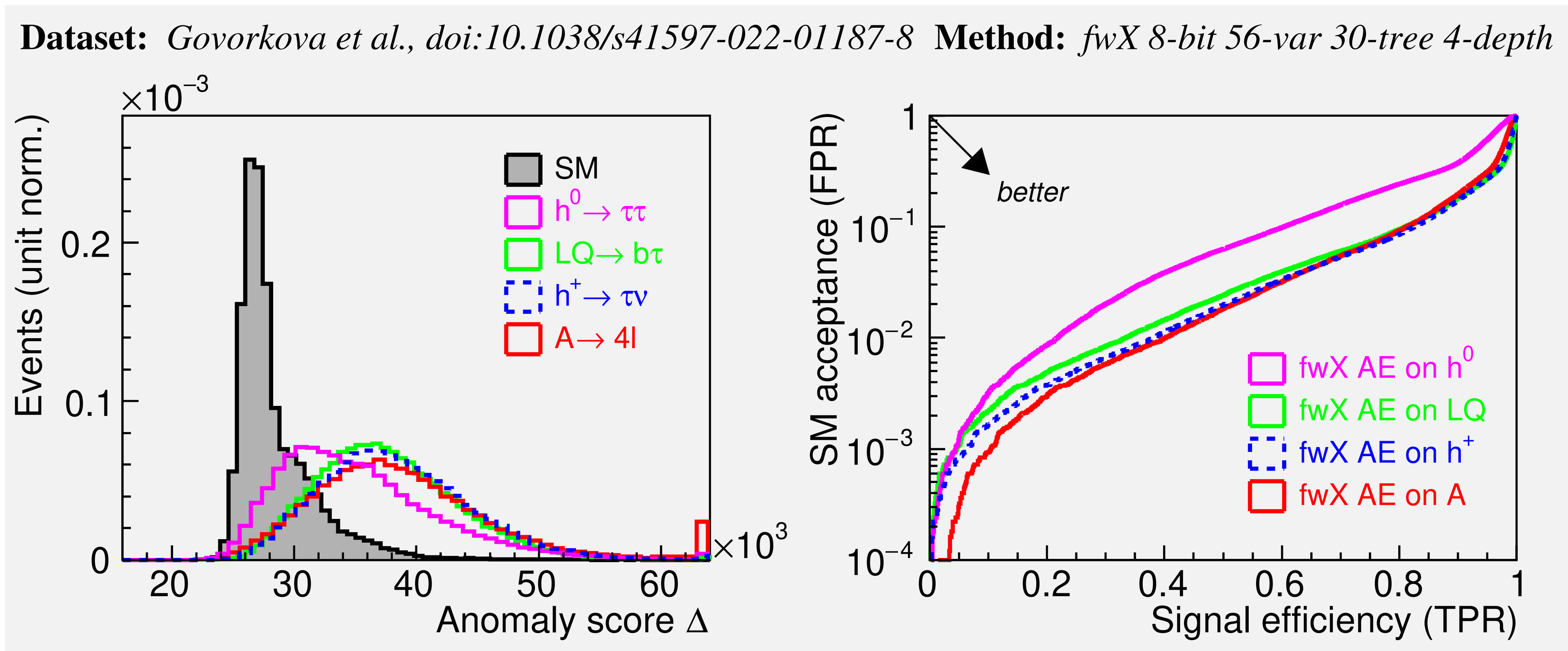
- Encoding *is* decoding with ★-coder technology
  Put the estimate of **x** inside the terminal bin → No need for a latent space!
  But the latent data (it's the bin #) is retrievable if desired

Performance

- Featured physics
  $H_{125} \to a_{70}\, a_{10} \to bb\,\gamma\gamma$ (also for H_{70})
  8 input variables
  30-trees, 6-deep, 8-bit

Physics : eff. 3x higher vs. ATLAS-inspired 3kHz
Timing : 30 ns latency, 5 ns interval
Resource: O(1)% on Xilinx Virtex UltraScale+ VU9P





- Comparison
  LHC anomaly dataset [Sci. Data **9**, 118 (2022)]
  54 input variables, 30-trees 4-deep 8-bits

  hls4ml-based [Nat. Mach. Intell. **4**, 154 (2022)]
  Physics perf. : Comparable AUC to ours
  Firmware perf. : See table on right for hls4ml DNN VAE

	our work	hls4ml
Latency	30 ns	80 ns
FF	0.6%	0.5%
LUT	9%	3%
DSP	0.8%	1%
BRAM	0	0.3%