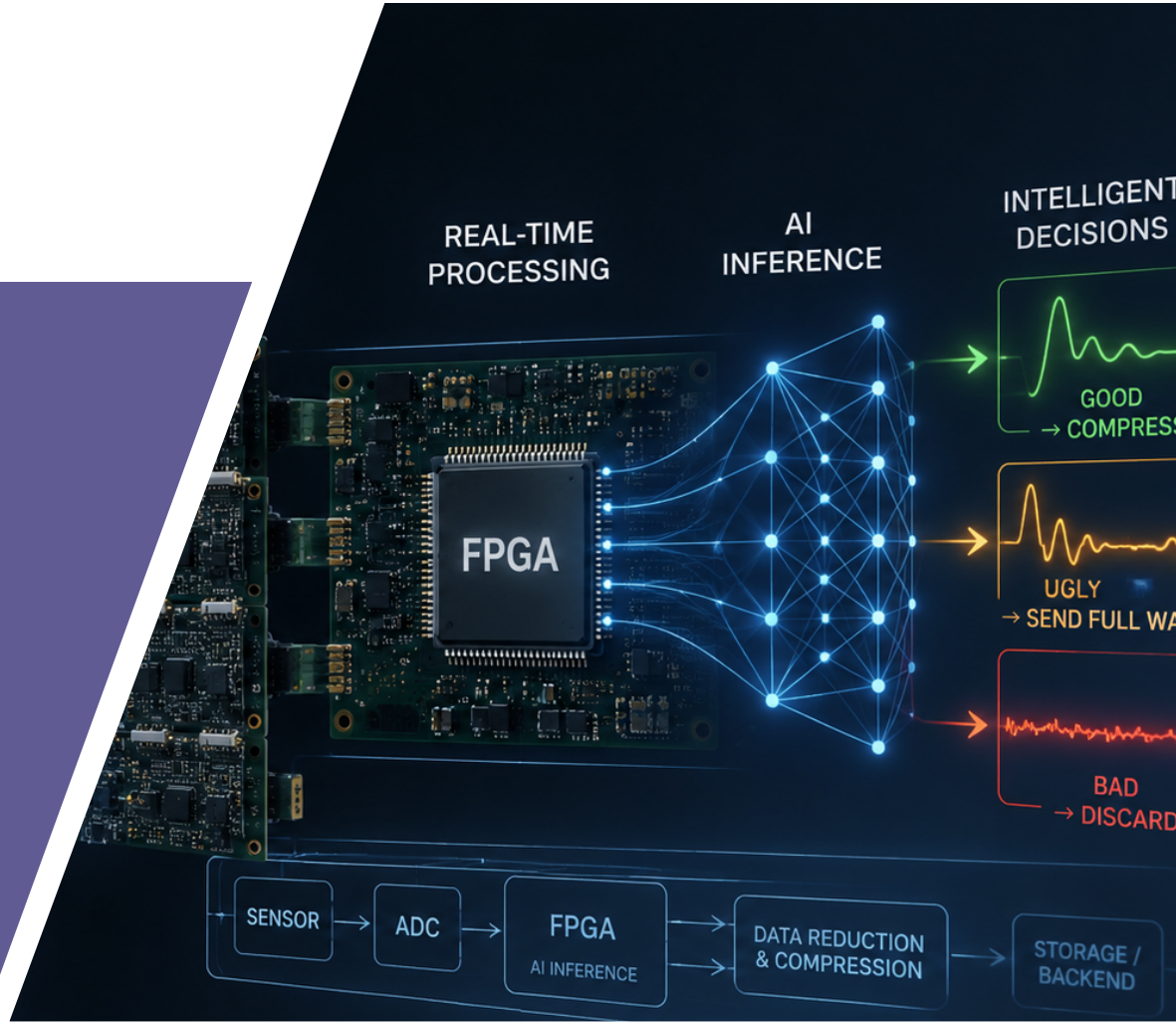
Universität Siegen

HEP Experimental High Energy Physics

CPPS Center for Particle Physics Siegen

From Detector Signals to Real-Time Inference: Intelligent Hardware for Particle Physics

Qader Dorosti



18.05.2026

JÜLICH Forschungszentrum

Integrated Computing Architectures ICA

color meets flavor

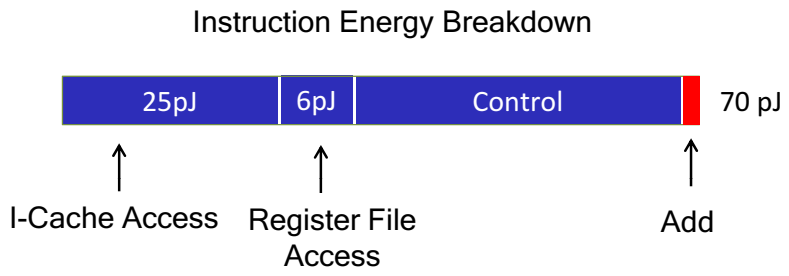
Where Hardware Pays: Movement and Switching

Energy per Operation vs. Memory Access (45nm CMOS)

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

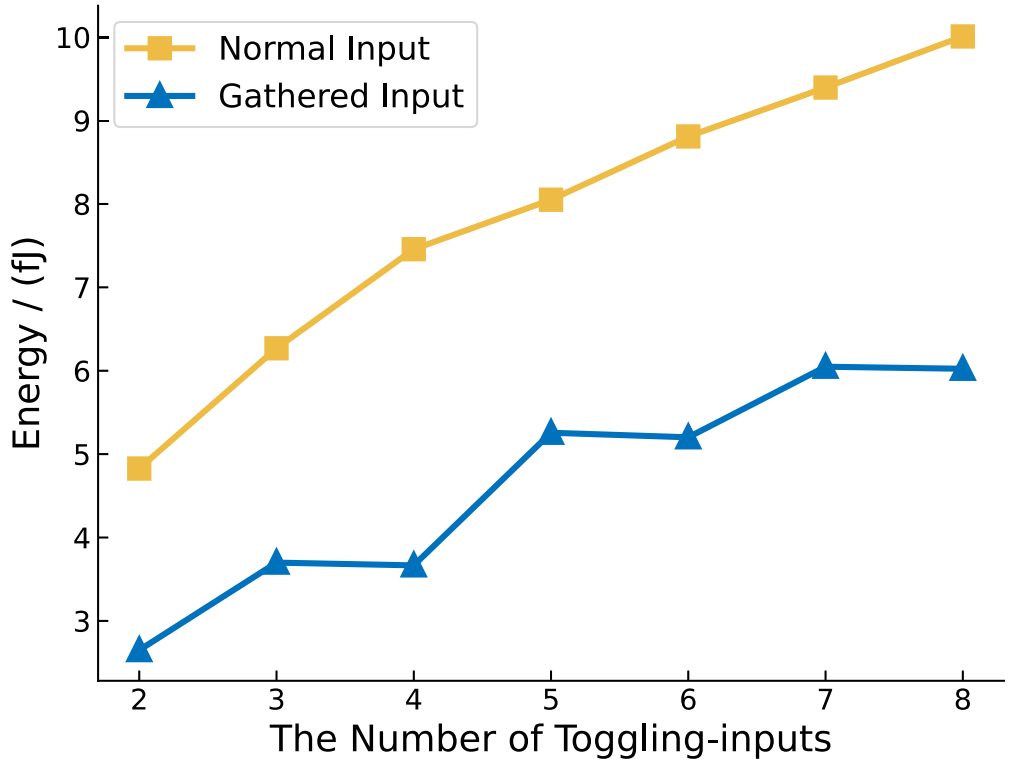
FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

Memory	
Cache	(64bit)
8KB	10pJ
32KB	20pJ
1MB	100pJ
DRAM	1.3-2.6nJ



Ref: Horowitz, ISSCC 2014, doi:10.1109/ISSCC.2014.6757323

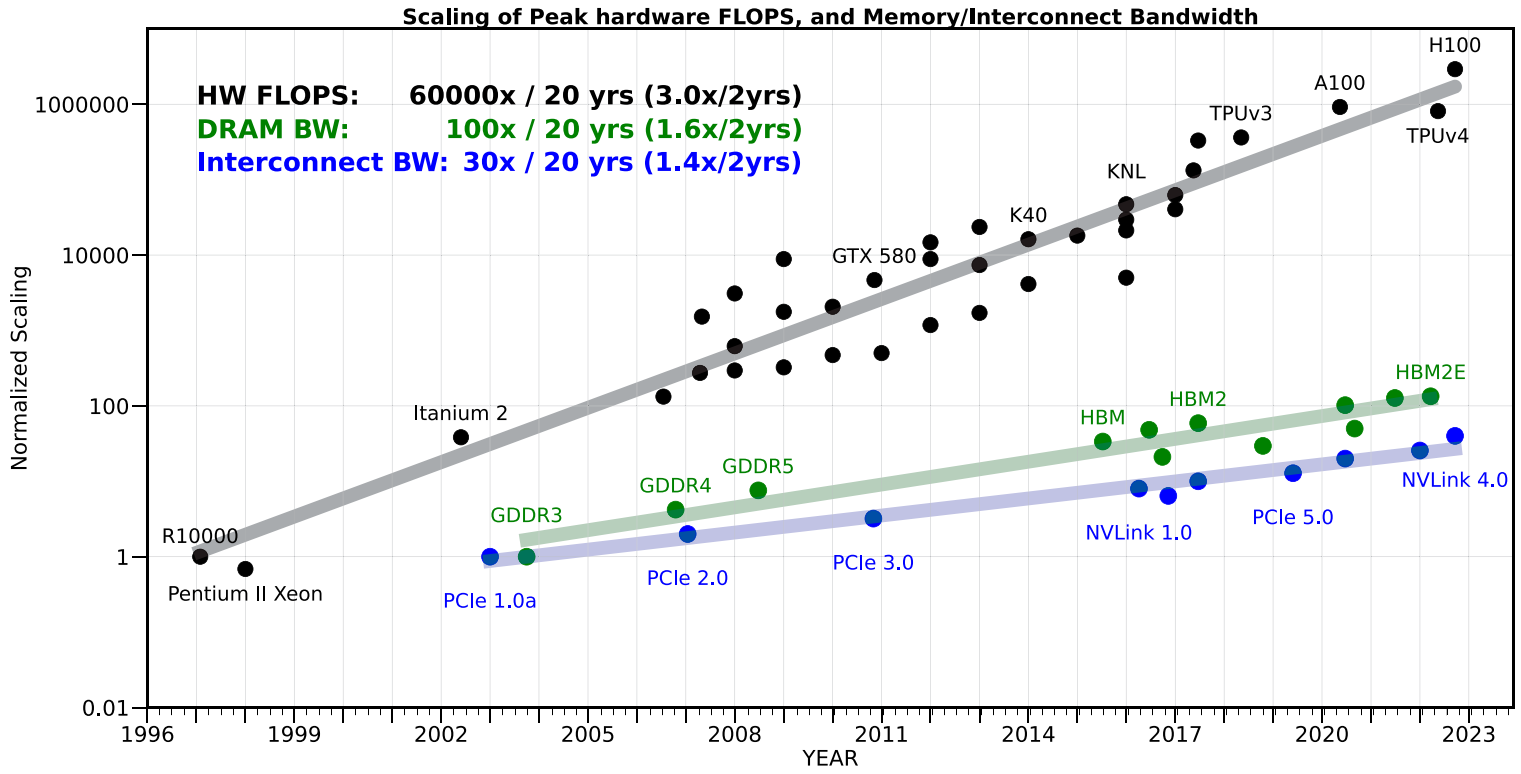
Compute Energy vs. Input Toggle Rate



Ref: Wang et al., ASPDAC 2025 — TRIFP-DCIM, DOI: 10.1145/3658617.3697577

Compute and Memory Do Not Scale Together

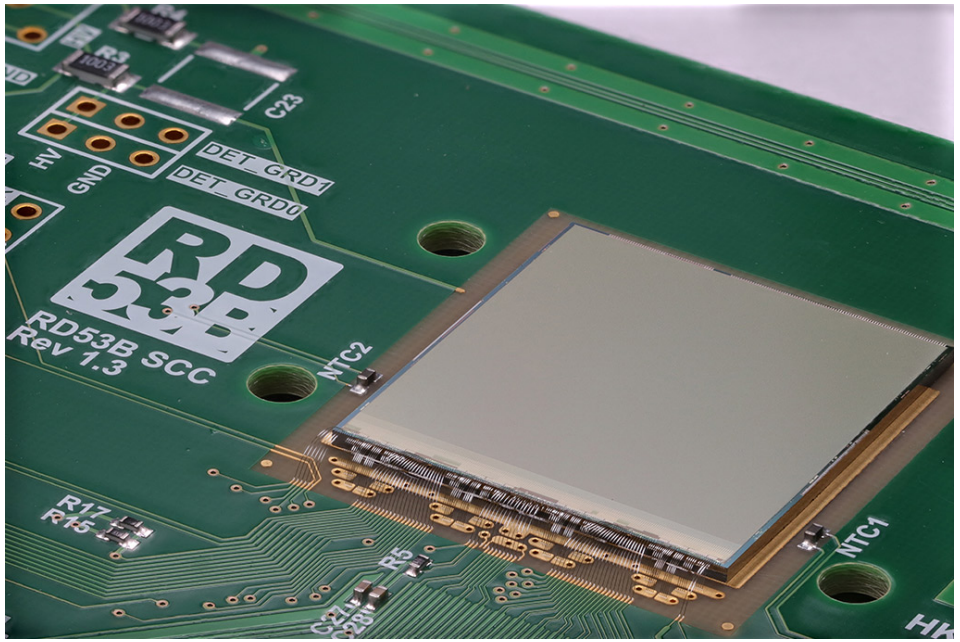
- Compute scales faster than memory transport
- Systems become transport-limited
- Early inference must respect this limit



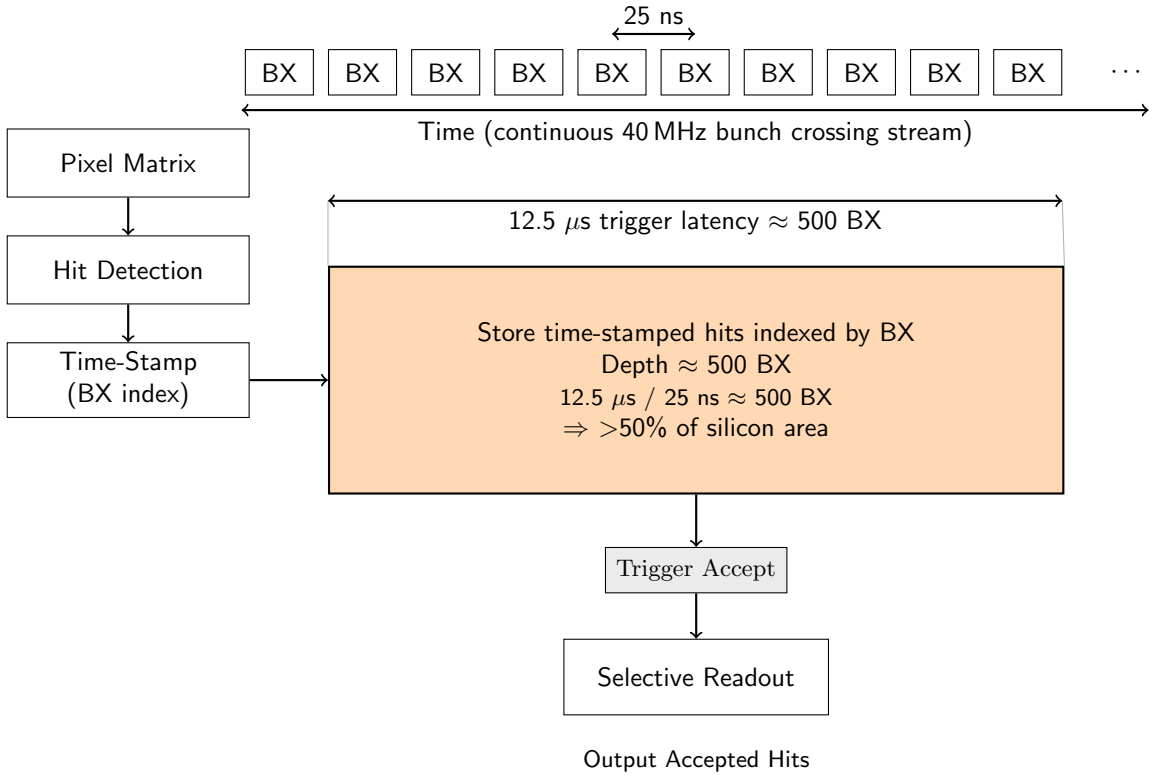
color meets flavor

Full Architectural Specialisation (ASIC)

ITkPixV2: Latency-Driven On-Chip Buffering



Ref: L. Le Pottier et. al., arxiv:2502.05097v3

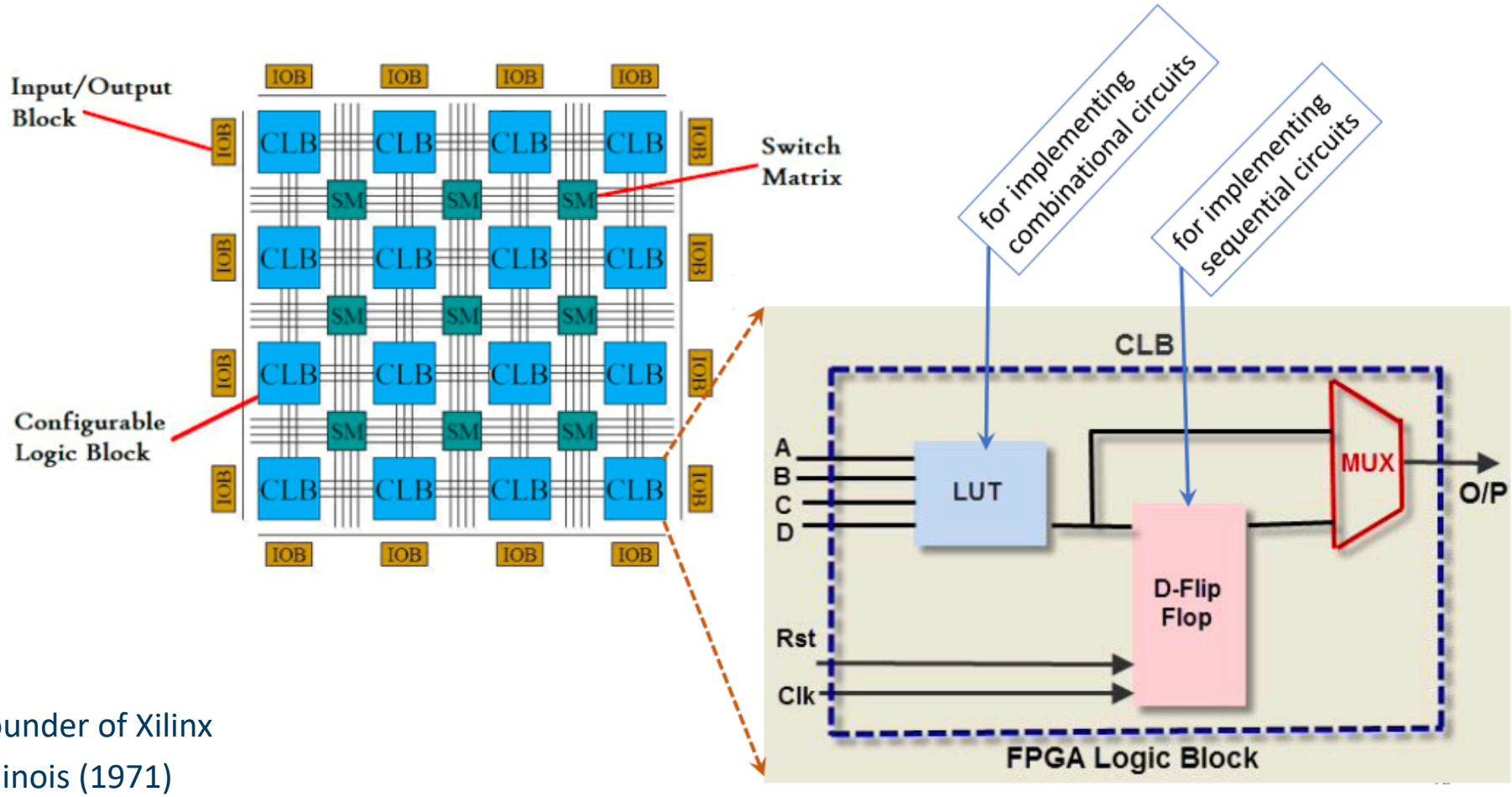


Specialisation translates physics constraints into locality.

FPGA/eFPGA: Structured Flexibility Under Constraint

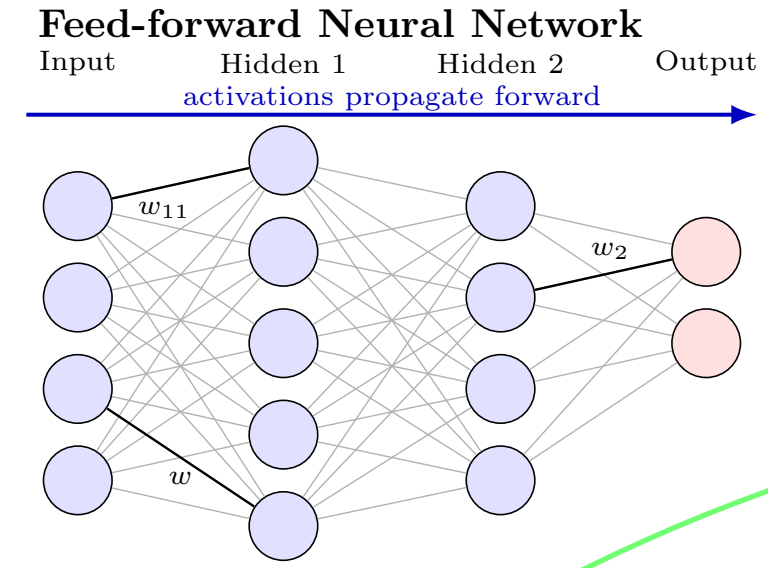
Ross Freeman (1948–1989); Founder of Xilinx

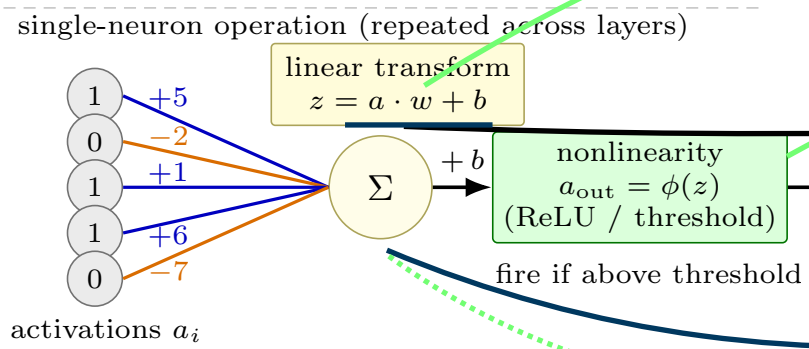
M.S. in physics, University of Illinois (1971)

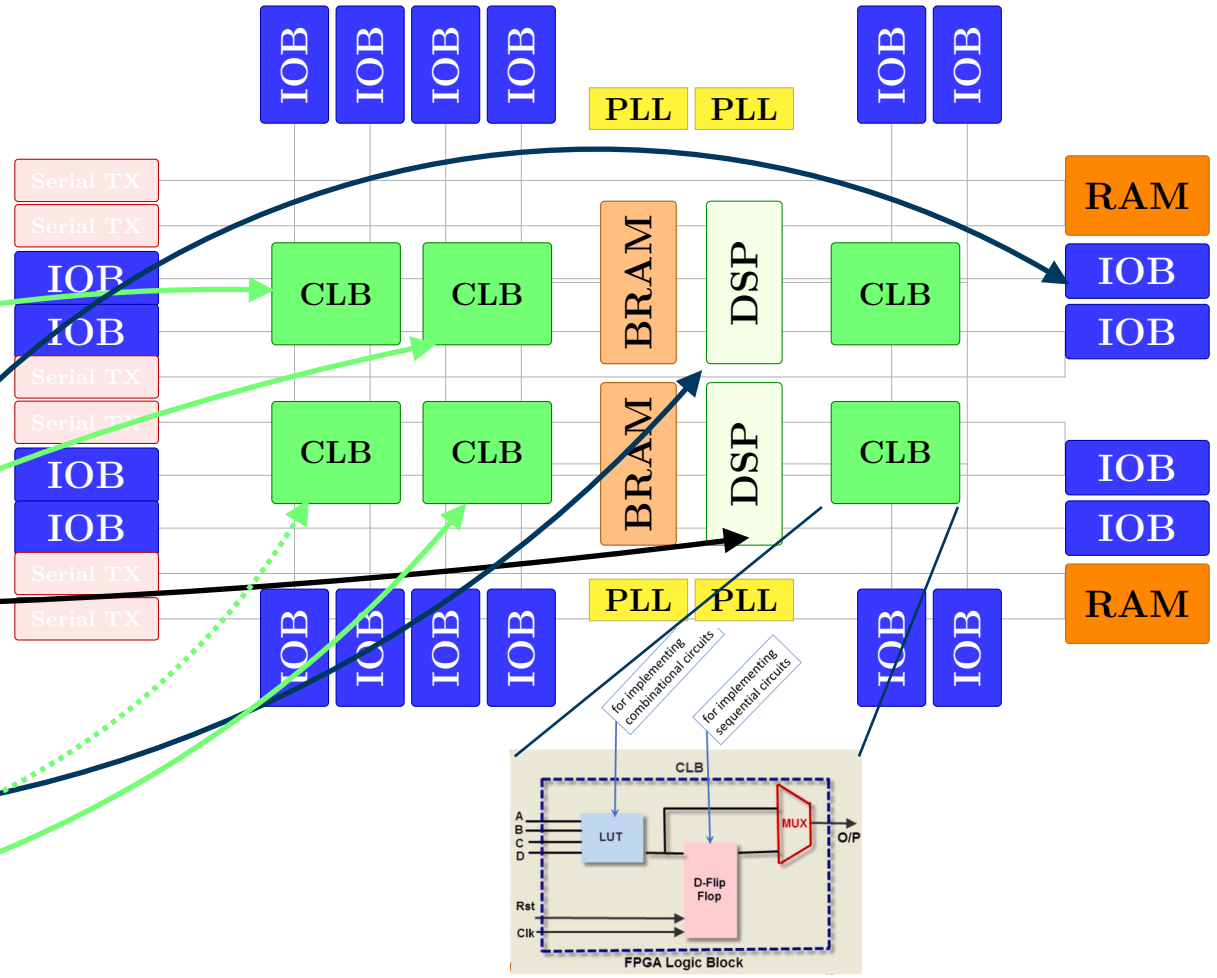


color meets flavor

From Neural Network to Silicon



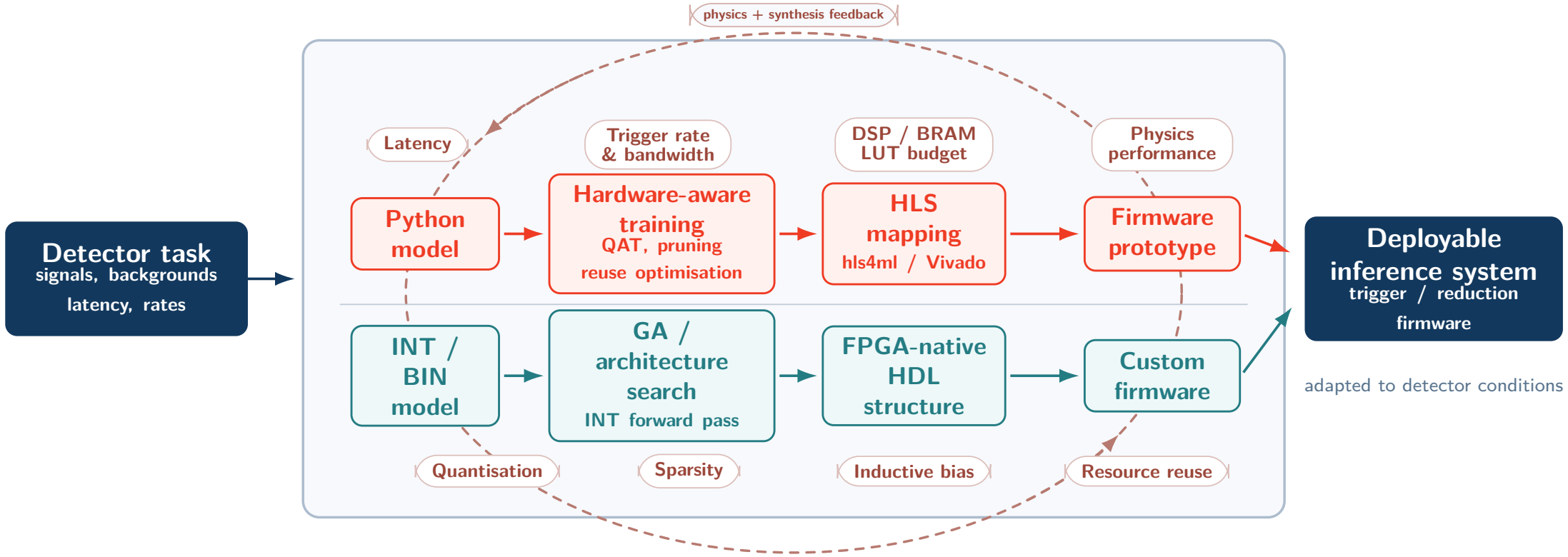




color meets flavor

Physics-Aware AI–Hardware Co-Design

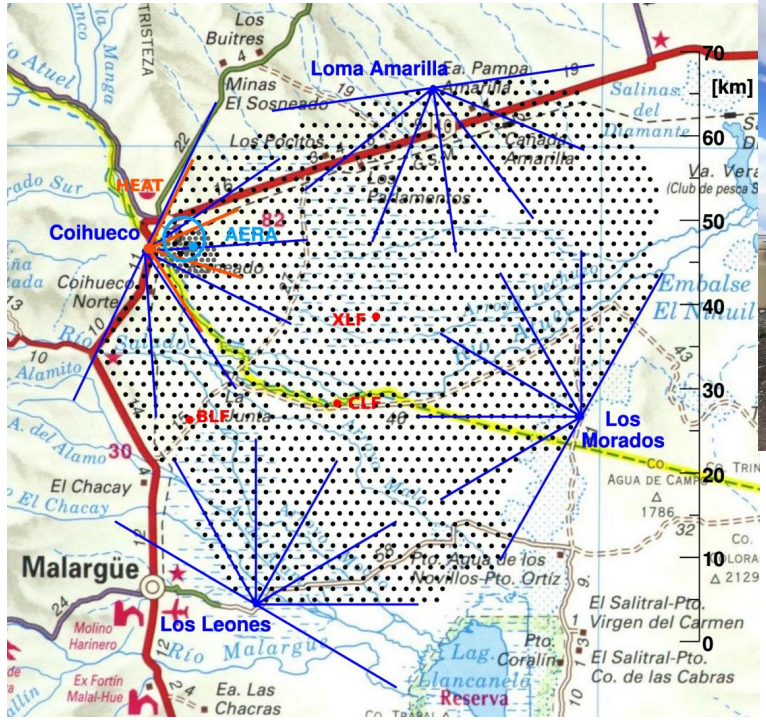
- Detector-specific physics goals and constraints
- Triggering and reduction as inference problems
- Generalisable co-design strategy
- Physics-driven inference under hardware constraints



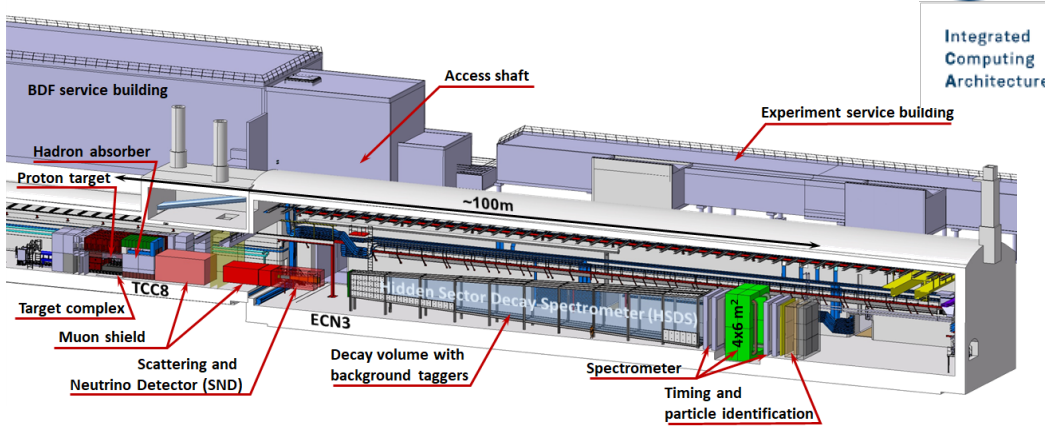
color meets flavor

Joint AI–FPGA Co-Design Activities (FZJ ↔ Siegen)

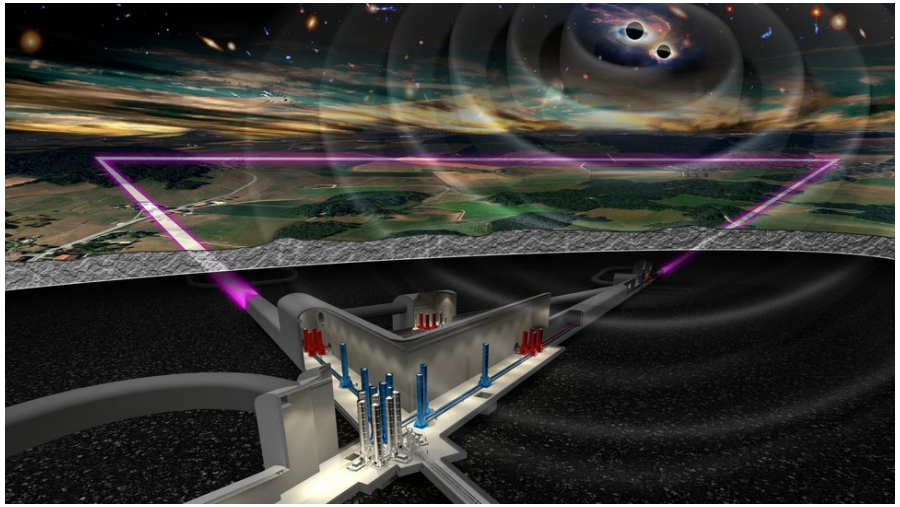
Auger Radio Self-Trigger (Real-Time AI Inference)



SHiP – Online Compression for SiPM Data



Einstein Telescope (Active Noise Mitigation)



Structured Collaboration

- Continuous technical exchange
- Shared FPGA/AI toolchain
- Joint prototyping and validation

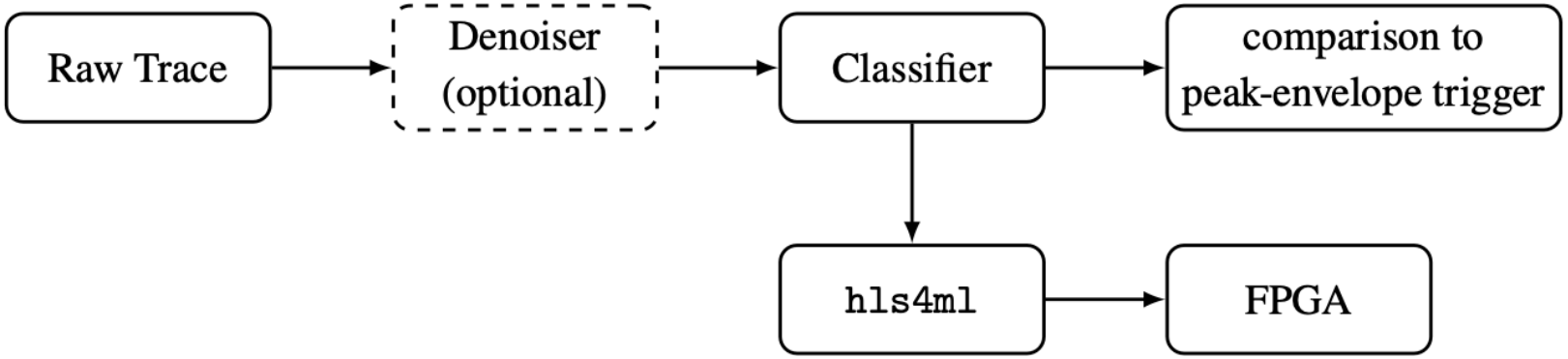
Human Capital

- Joint supervision of student projects
- Cross-site hardware training

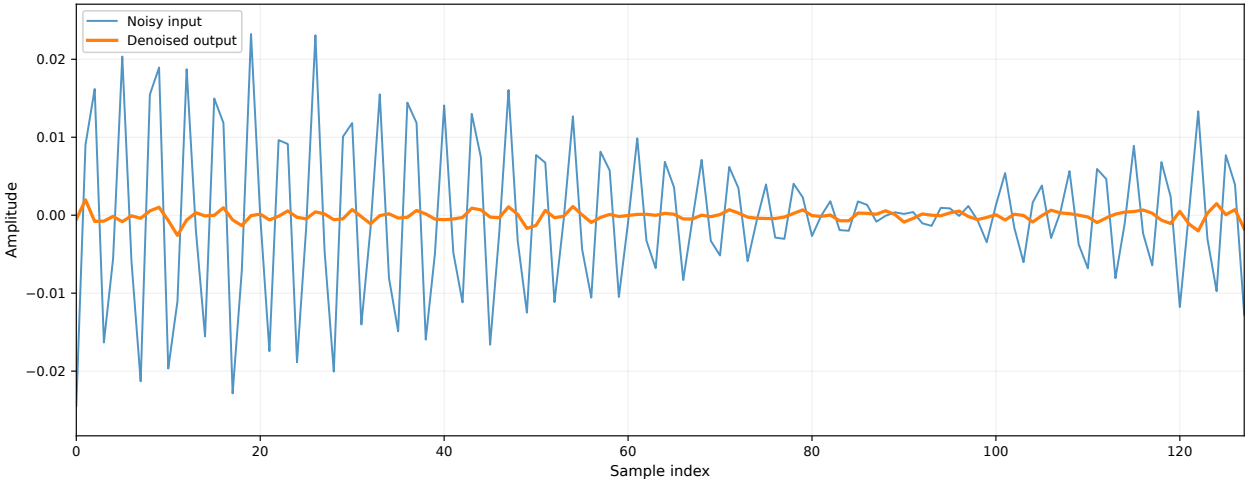
Universität Siegen

color meets flavor

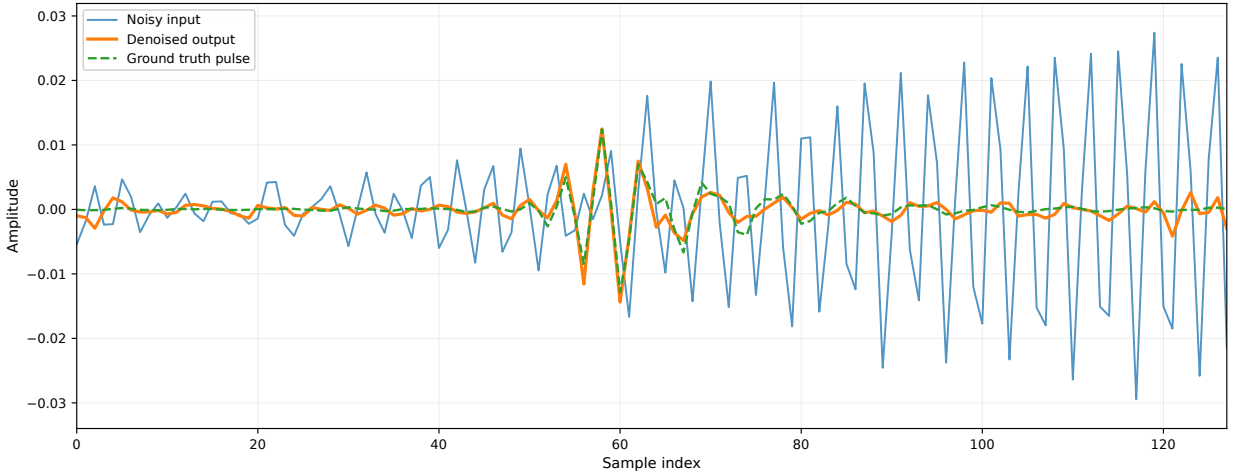
Ongoing Activities: Radio Self-Trigger Under Power and Latency Constraints



Background: Noise trace (denoised: orange)



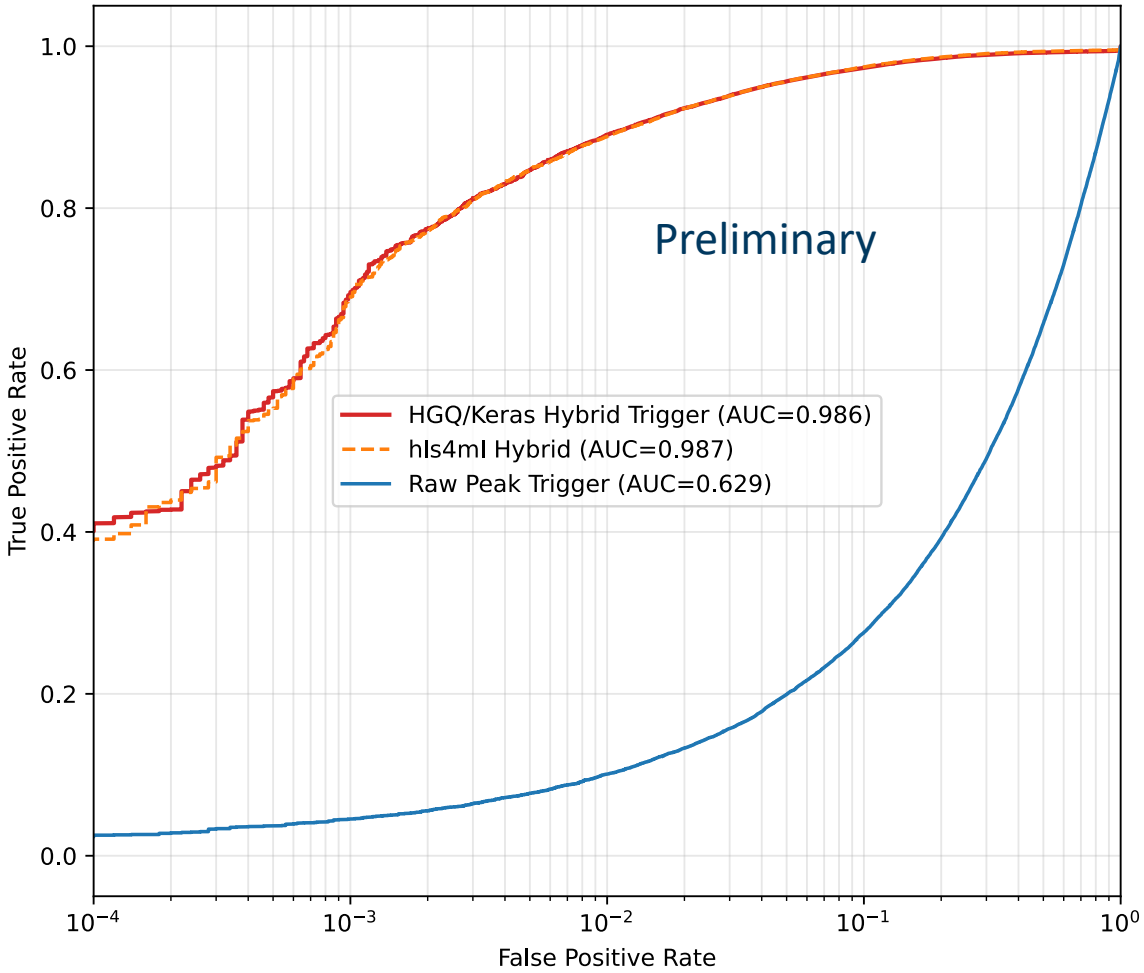
Signal: EAS pulse in noise (denoised: orange)



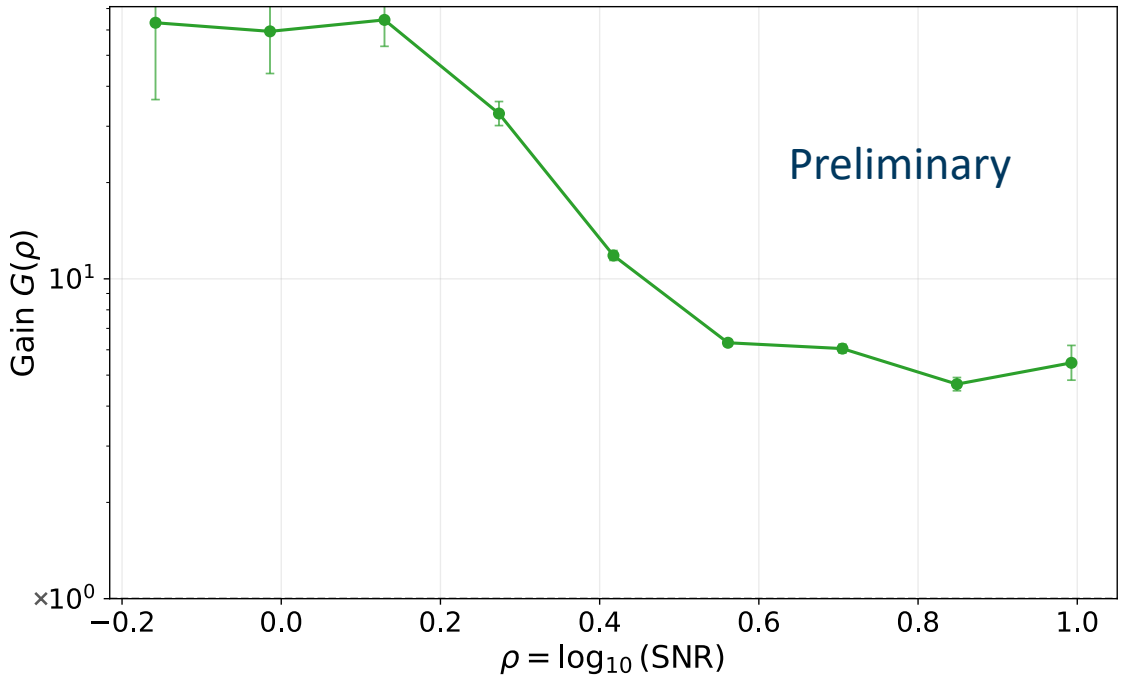
color meets flavor

Ongoing Activities: Radio Self-Trigger Under Power and Latency Constraints

Classification performance: AI Trigger vs Threshold-base trigger



AI-trigger gain for weak signals at fixed background rate



FPGA resource and power comparison: published [1] vs. current hybrid trigger

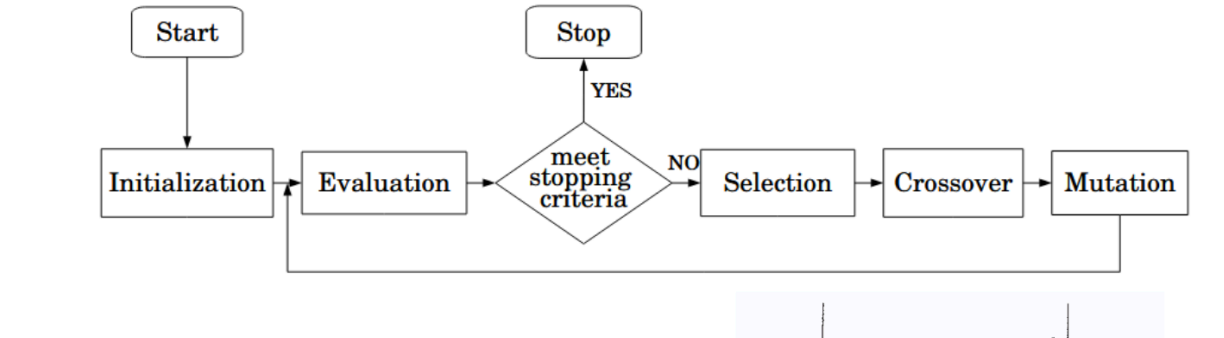
Design	Resource report	DSP	LUT	BRAM	Z-7020 dyn. power [W]	U250 dyn. power [W]
Published FPGA study [1]	post-synthesis on Z-7020	2119	191243	780	–	5.51
Current design	post-implementation on Z-7020	42	22920	50	0.562	0.270

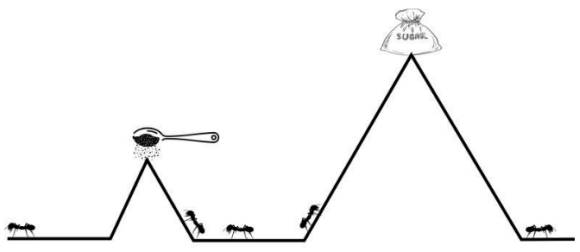
[1] Q. Dorosti, "AI-enhanced self-triggering for extensive air showers: performance and FPGA feasibility", JINST 20(2025) P10010.

Significant gain in weak-signal detection with AI-based triggering

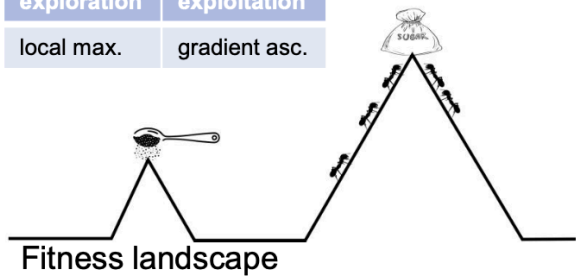
color meets flavor

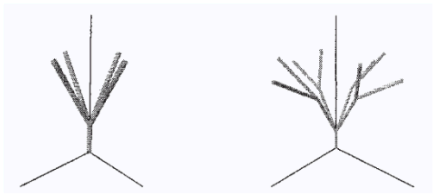
Ongoing Activities: Hardware-Constrained Learning with Genetic Algorithms

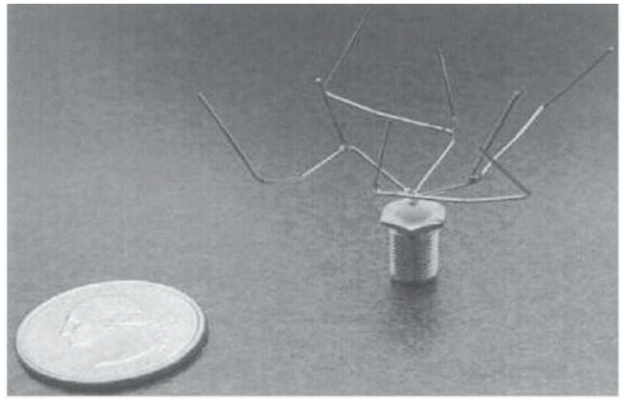




exploration	exploitation
local max.	gradient asc.







The 2006 NASA ST5 spacecraft antenna.

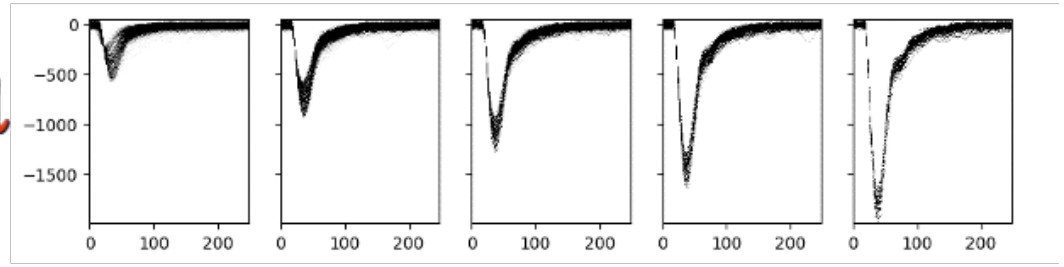
- FPGA-native discrete inference structures
- Non-differentiable optimisation problem
- Joint optimisation of:
  - accuracy
  - sparsity
  - latency
  - resource usage
- Hardware constraints included during learning

The FPGA fabric becomes part of the optimisation process itself.

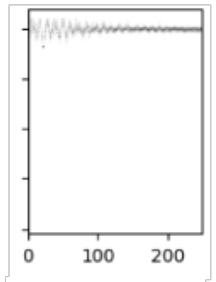
color meets flavor

Ongoing Activities: Online Compression for SiPM Data

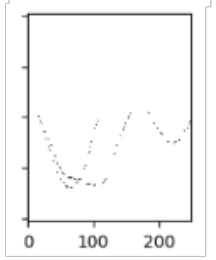
the good



the bad



the ugly



- ADC readout flexible but waveforms have significant data volume
- Idea: Classify ADC data
  - Good: parameterizable → compress
  - Bad: electronics noise → count
  - Ugly: double hits, others → send

→ reduce data volume

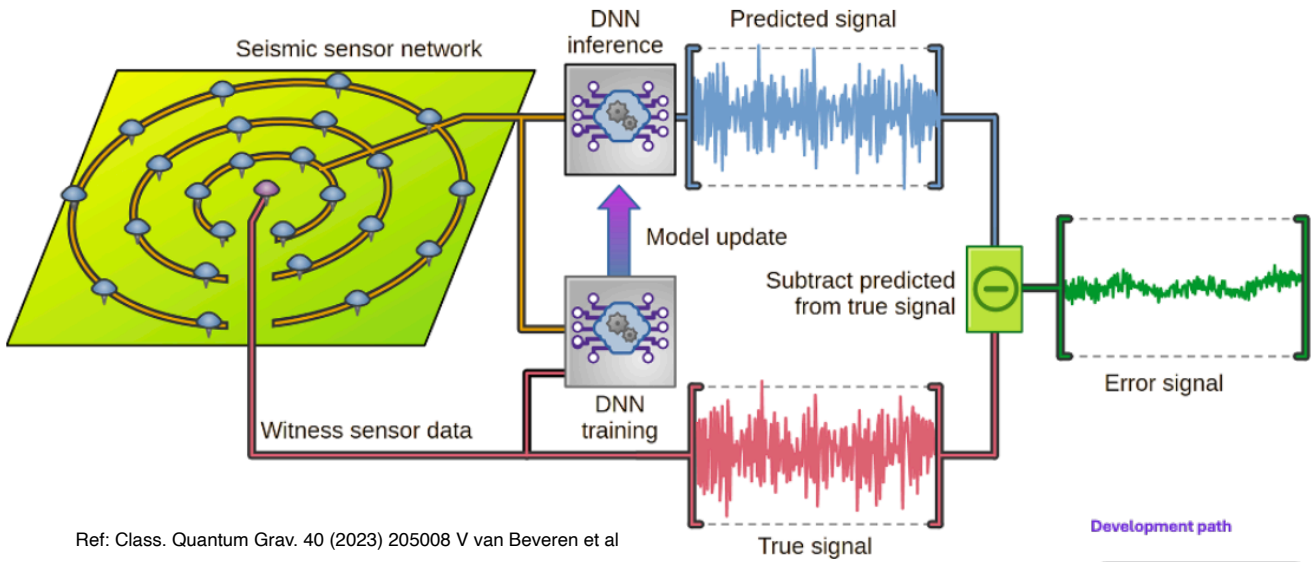
Compare NN approaches [1]:

- Binary NN + Genetic algorithm
  128(int7) - 32 - 32 - 2
- HLS4ML 2DCNN
  128 - 4 - 6 - 8 - 2

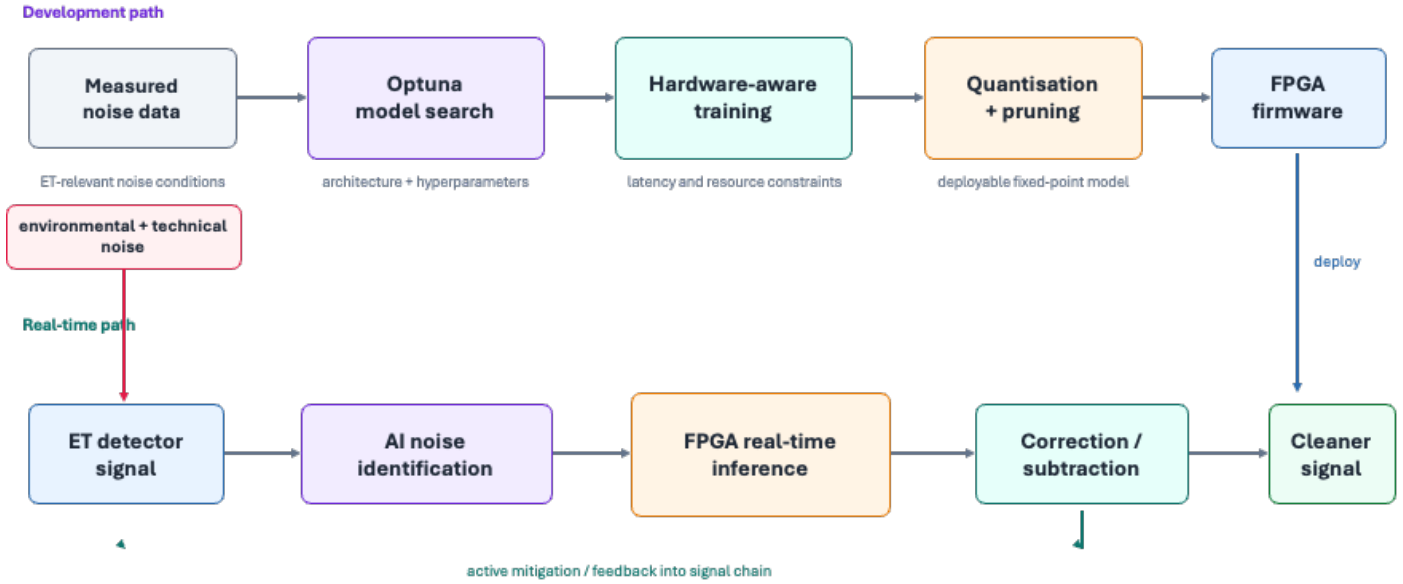
BNN+GA	hls4ml 2DCNN
74 % accuracy	95 % accuracy
16640 weights * 2bit	2696 weights * 8bit
58k LUT + 1.5k FF no DSP, BRAM	186k LUT + 112k FF 556 DSP, 120 BRAM
10 ns latency	3 µs latency
105 min * 90 cpu train.	2 min * 16 cpu training

[1] I. Bekman, et.al, „FPGA-Based Real-Time Waveform Classification“, https://arxiv.org/abs/2511.05479 (also presented at TWEPP 25)

Active AI-Based Noise Mitigation for Einstein Telescope (ET)



Ref: Class. Quantum Grav. 40 (2023) 205008 V van Beveren et al



color meets flavor

Summary & Outlook

- Intelligent real-time detector systems under strict hardware constraints
- AI–hardware co-design from model development to deployment
- Low-latency inference and adaptive signal processing on FPGA platforms
- Resource-aware architectures for triggering, filtering, and data reduction
- Transferable methodologies across heterogeneous detector environments

Outlook:

- Radiation-aware AI inference
- Intelligent front-end ASICs
- Shared FPGA/AI infrastructure and workflows
- Toward scalable intelligent detector electronics for future experiments

Possible IHL Project Directions

1. Neuromorphic / memristor computing

- memristor-based neuromorphic computing
- analog / in-memory AI architectures
- radiation-hard memristor technologies

2. AI inference on FPGA / Intelligent DAQ and edge processing

- low-latency neural networks
- AI triggers / intelligent data reduction
- real-time signal classification
- FPGA-based edge processing

3. Intelligent front-end ASICs

- embedded ML in ASICs
- feature extraction and clustering in front-end electronics
- eFPGA architectures integrated in ASICs for flexible processing

4. FPGA–ASIC co-design

- prototyping algorithms on FPGA and migrating them to ASIC
- hardware-aware ML design flows
- design frameworks for intelligent detector electronics

5. Radiation-hard intelligent electronics

- rad-hard AI hardware
- fault-tolerant inference architectures
- resilient detector electronics

6. Cologne Chip ecosystem & Sensor–FPGA integration concepts

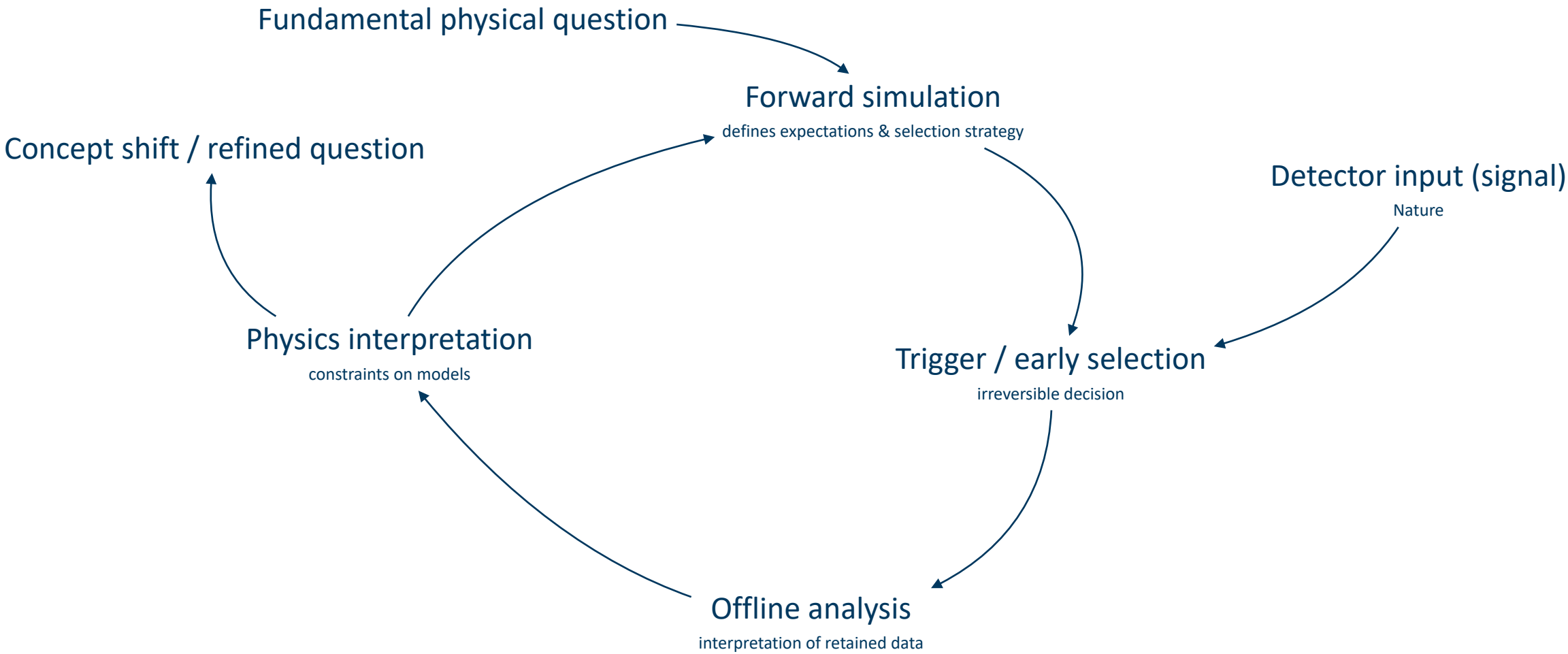
- collaboration with the Cologne Chip initiative
- chip design and prototyping infrastructure
- development of application-specific detector electronics
- direct integration of FPGA with sensors
- bump-bonding FPGA or processing ASICs to sensors
- evaluation boards and demonstrator platforms for intelligent detectors

7. CIPix / in-pixel intelligence

- on-pixel processing
- intelligent sensor architectures
- event-driven pixel detectors

Physics Discovery Loop:

Detectors, Forward Modelling, and Real-Time Inference



color meets flavor

Universität Siegen

HEP Experimental High Energy Physics

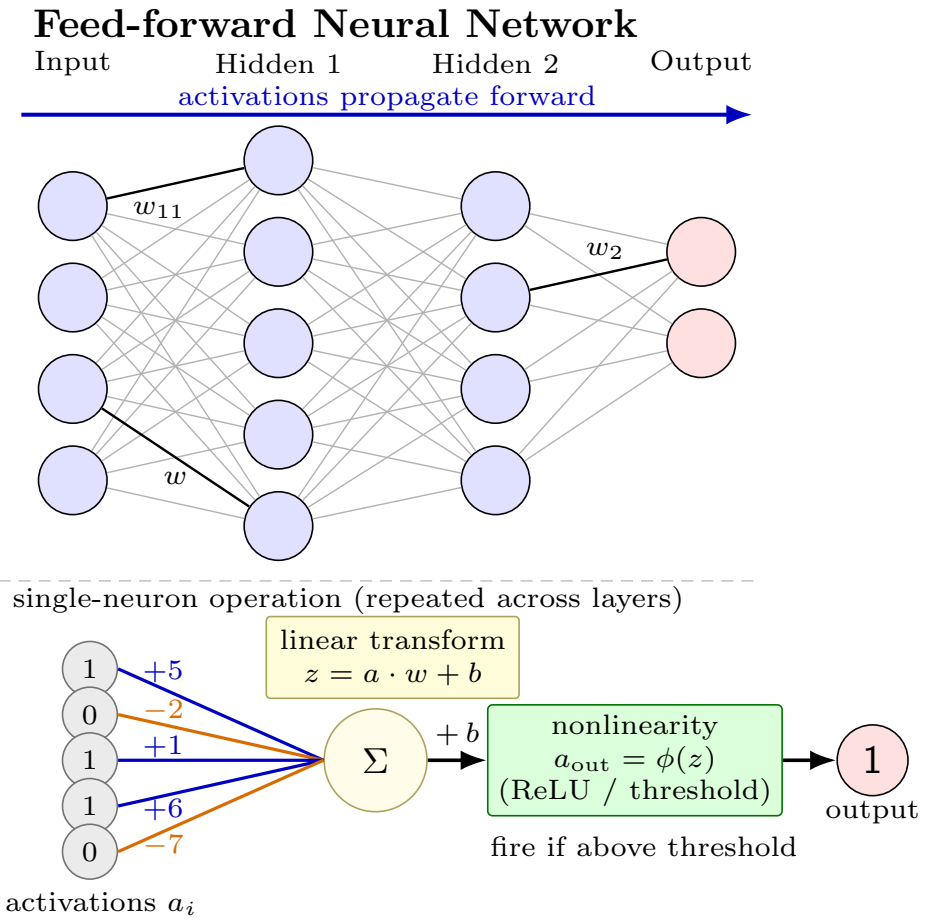
CPPS Center for Particle Physics Siegen

From physical cause to observed signal

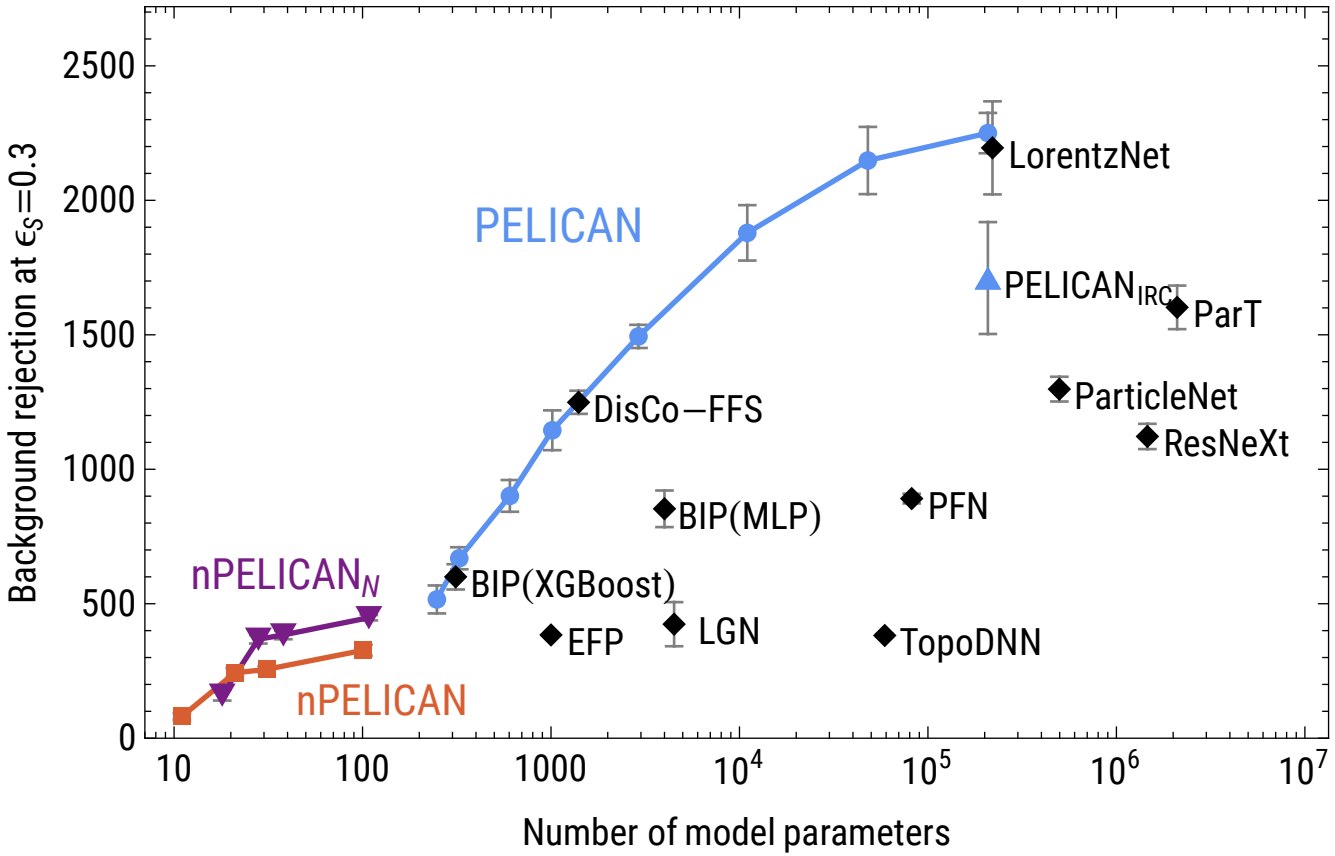
Discovery is always an inference problem under incomplete information.

https://commons.wikimedia.org/wiki/File:An_Illustration_of_The_Allegory_of_the_Cave,_from_Plato's_Republic.jpg

Neural Networks as Dataflow



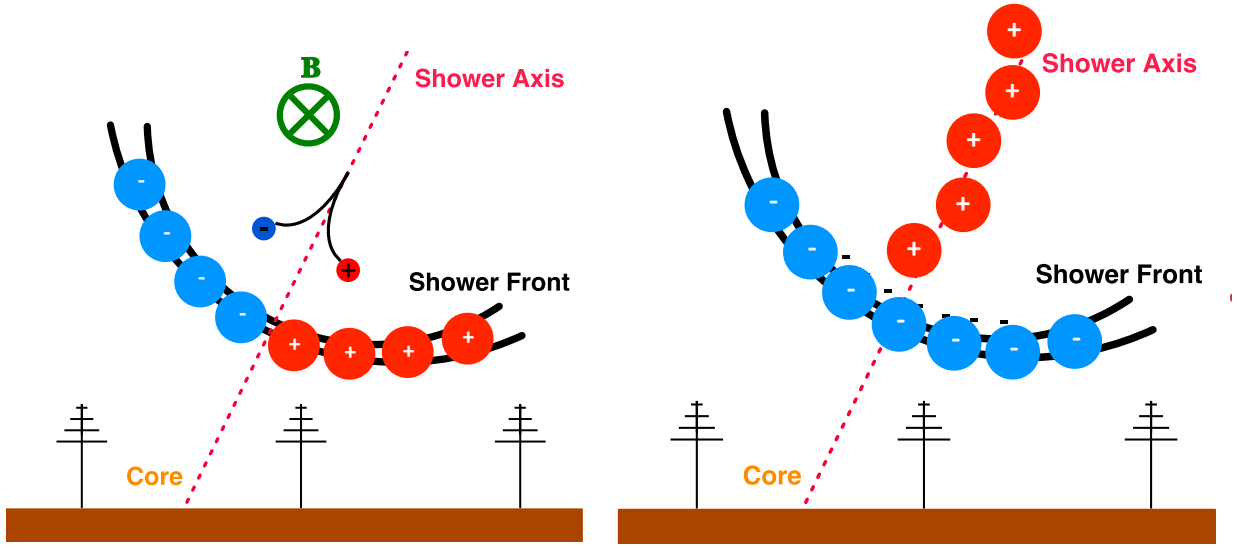
Top-Tagging Performance: Background Rejection vs Model Size



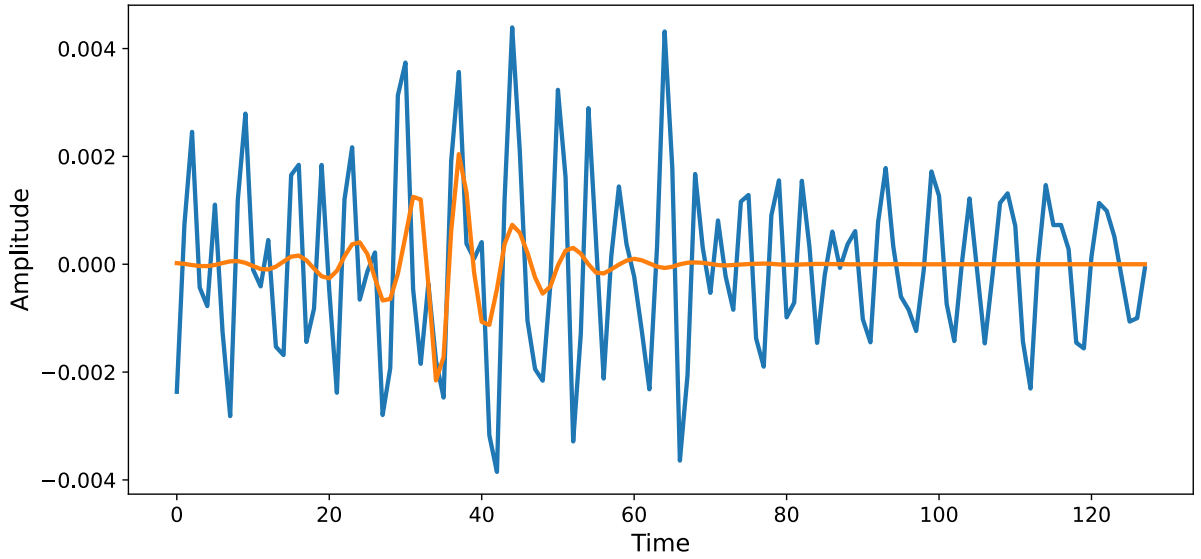
Ref: A. Bogatskiy et. al., NeurIPS 2023, arXiv:2310.16121

color meets flavor

Ongoing Activities: Radio Self-Trigger Under Power and Latency Constraints



Extensive air shower radio pulse (orange) buried in noise (blue)



- **Emission:** Geomagnetic + Askaryan → nanosecond-scale radio pulse.
- **Problem:** Noise/RFI dominates → threshold-base self-triggering fails.
- **Solution:** Deep learning exploits full pulse morphology.

color meets flavor

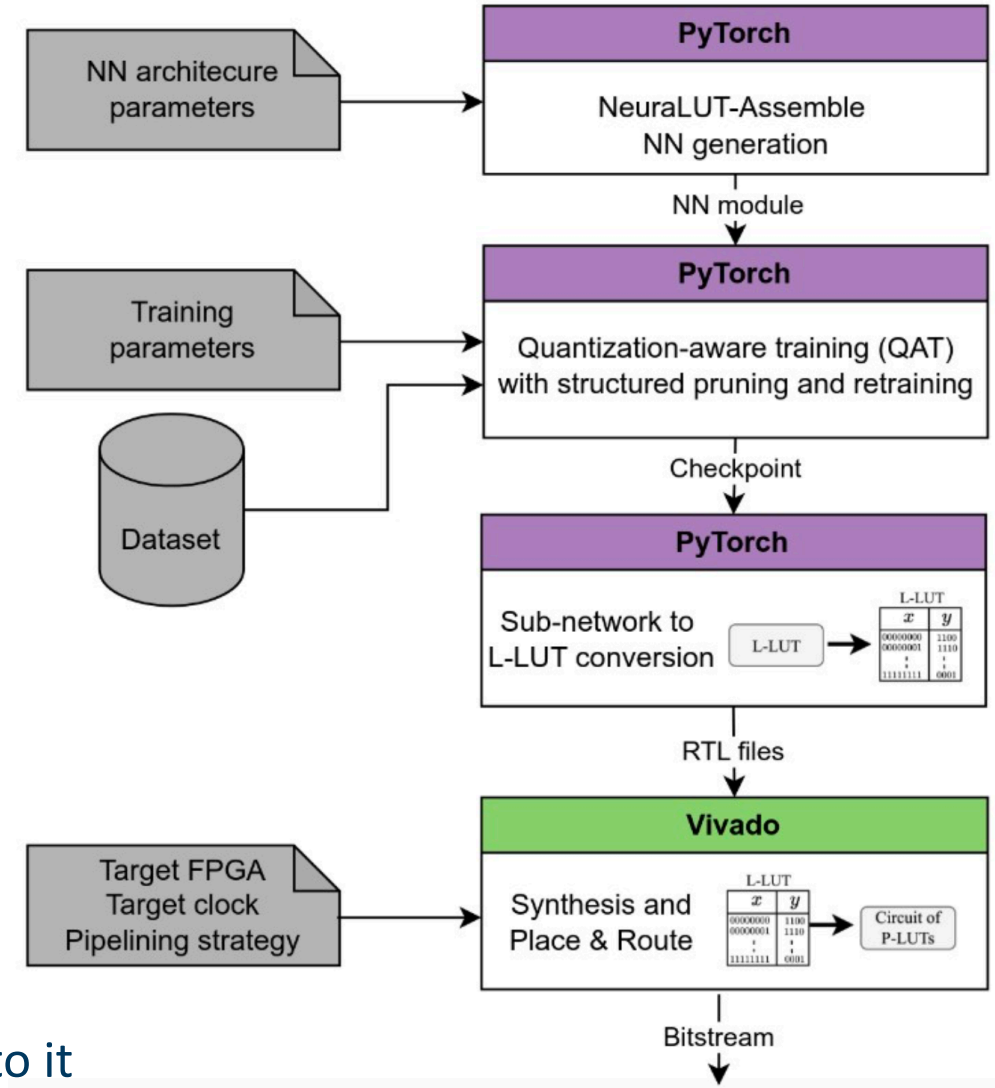
Ongoing Activities: Native LUT-Based Neural Inference on FPGA

Concept

- Assemble large neurons from small LUT units
- Hierarchical L-LUT structure
- Skip connections for expressivity

Why It Matters

- Native FPGA structure (no DSP blocks)
- Ultra-low latency (~2 ns)
- Strong area × delay efficiency
- Ideal for extreme power/latency constraints



Neural networks designed for FPGA fabric — not mapped onto it

color meets flavor

The Intelligent Hardware Lab

Technical Focus:

- Hardware/software co-design for advanced AI/ML algorithms
- Automated optimisation and hardware deployment
- Radiation-tolerant hardware (eFPGAs)
- Advanced packaging and integration