Pacific Northwest
NATIONAL LABORATORY

LEAP-FWD: Low-latency Embedded AI for Physics

—

From Waveforms to Discovery

Samuel L. Watkins, Richard Saldanha

CPAD Workshop 2026

PNNL-SA-228562



U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Genesis Mission + TDAQ

- How can we enable enhanced discovery via AI/ML?
- National Challenge: “Unifying Physics from Quarks to the Cosmos”
  - 14A: Foundation Models of Particle Interactions and Cosmic Physics (HEP, NP)
  - “…Discovery science potential will be maximized by addressing such technical challenges as sparse-data domains and **real-time data acquisition of high-dimensional petabyte-scale datasets** with associated scalability challenges and interpreting the experimental signals using theoretical knowledge.”
- As we scale up our experiments, we will need to navigate the firehose of physics data
  - We could save everything, process offline
  - Or move algorithms to the sensor-edge for intelligent decision making

U.S. DEPARTMENT of ENERGY

Office of Science (SC), Office of Critical Minerals and Energy Innovation (CMEI), Office of Environmental Management (EM), Office of Electricity (OE), Hydrocarbons and Geothermal Energy Office (HGEO), and Office of Nuclear Energy (NE)

Genesis Mission

The Genesis Mission: Transforming Science and Energy with AI

Notice of Request for Application (RFA) Number: DE-FOA-0003612

DE-FOA-0003612-000003

LEAP-FWD: Low-Latency Embedded AI for Physics — From Waveforms to Discovery

Phase I | DOE Genesis Mission (DE-FOA-0003612) | 2026 – 2027

CHALLENGE

LEGEND-200 DAQ energy threshold currently at ~ 50 keV

Limited by high rate of cross-talk and noise events

Offline selection of physical events (“data cleaning”) is labor intensive

GOALS

Lower energy threshold of the DAQ

Reduce human workload requirements for data cleaning

Enable BSM physics searches and spatiotemporal anomaly detection

APPROACH

Develop AI algorithms deployable on FPGAs for real-time low-latency event classification

Build on previous AP-SVM work to automate data cleaning, adapting methods for low-latency deployability on FPGAs

DECISION-GATE OBJECTIVES (Month 6 Go/No-Go)

Four quantitative metrics define AI advantage

Energy Threshold ≤ 10 keV

(5× reduction, same latency, similar data storage)

Signal Efficiency ≥ 97%

(Retain physics events while rejecting all spurious noise)

Data Cleaning ≤ 8 hrs/wk

(~10× fewer person-hours for data curation)

Anomaly Recall ≥ 75%

(Detect AI-injected synthetic anomalies)

PNNL: Richard Saldanha, Sam Watkins, Antonino Tumeo
UCSD: Aobo Li
UNC: Reyco Henning, Julieta Gruszko, Himal Acharya

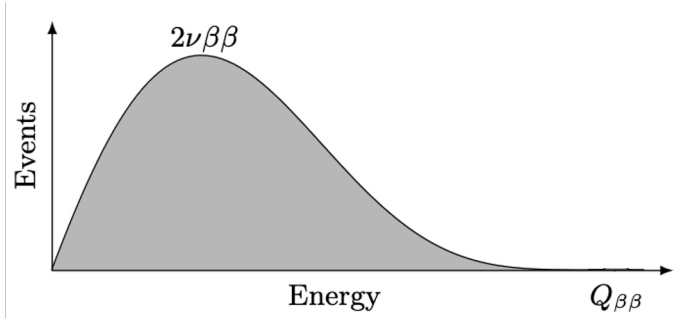
Pacific Northwest NATIONAL LABORATORY

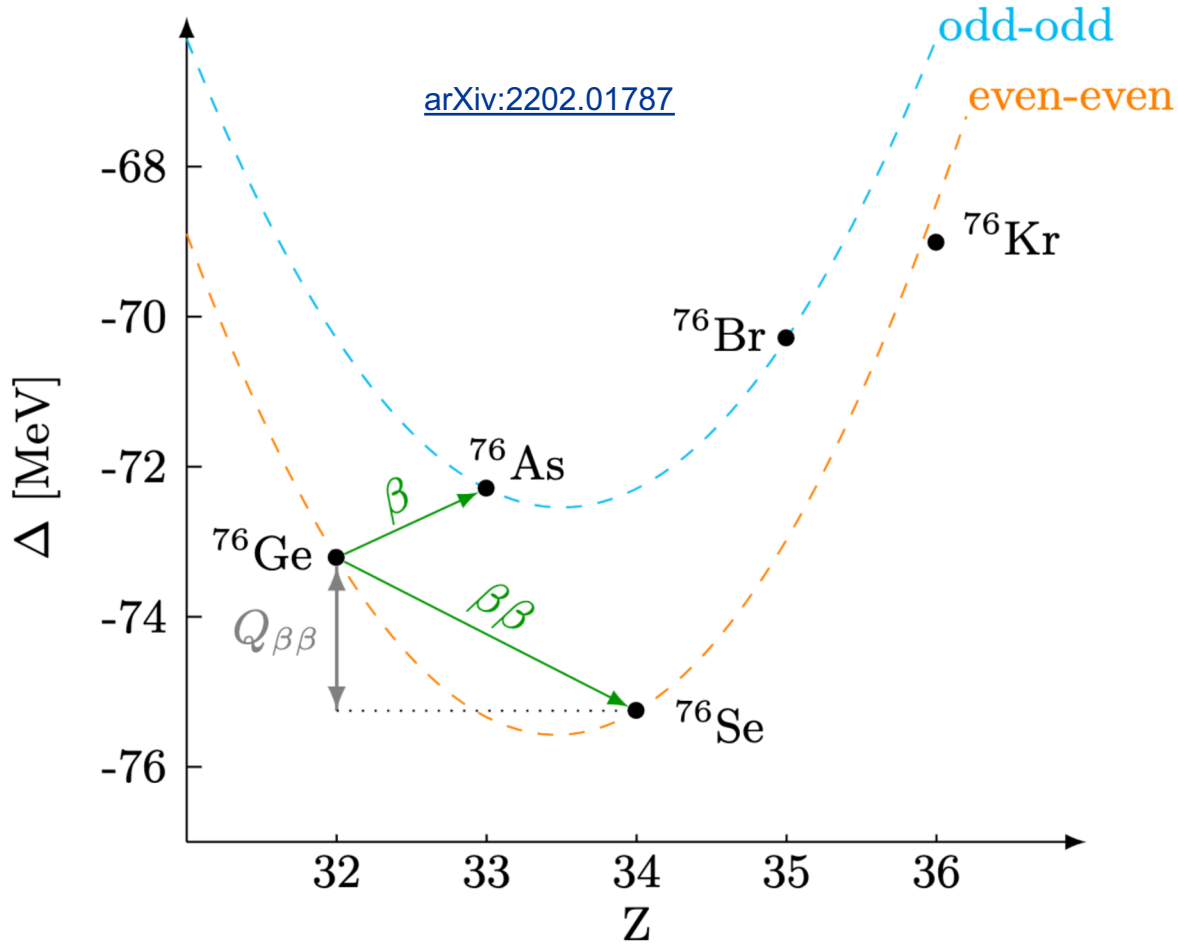
UC San Diego

The University of North Carolina at Chapel Hill

Two-Neutrino Double-Beta Decay

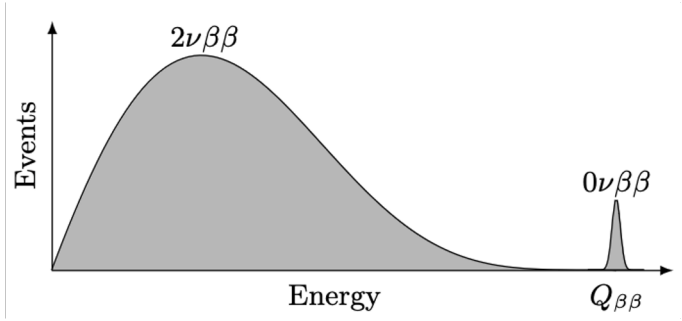
- Double beta decay
  - Energetically allowed
  - $2\nu\beta\beta$ conserves lepton number
  - Continuous beta spectrum up to $Q_{\beta\beta}$

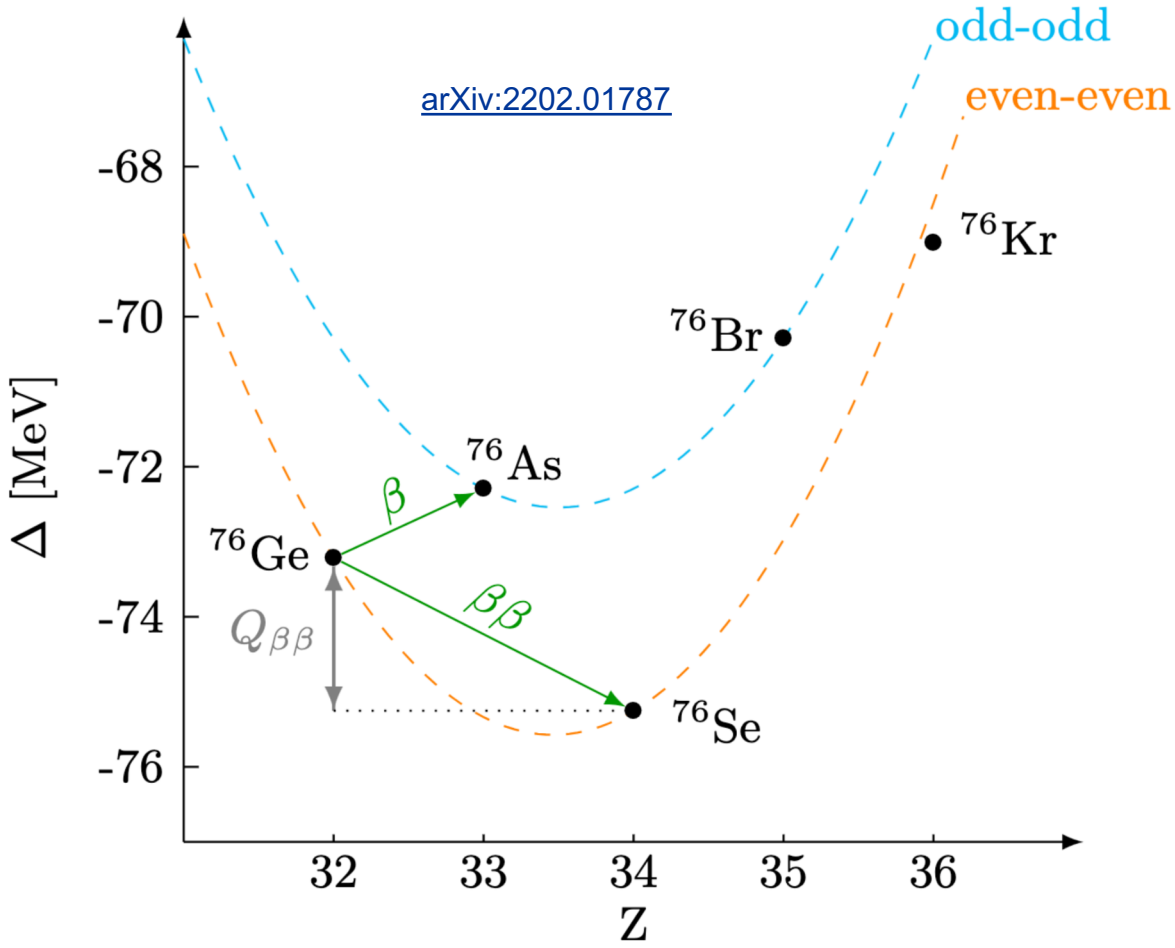




Neutrinoless Double-Beta Decay

- Double beta decay
  - Energetically allowed
  - $2\nu\beta\beta$ conserves lepton number
  - Continuous beta spectrum up to $Q_{\beta\beta}$
- But what if the neutrino was its own anti-particle?





LEGEND Concept

- **L**arge **E**nriched **G**ermanium **E**xperiment for **N**eutrinoless $\beta\beta$ **D**ecay
- Merger of two successful programs: **GERDA** and **MAJORANA**
  - Each based on high-purity Ge detectors enriched in ^{76}Ge

MAJORANA

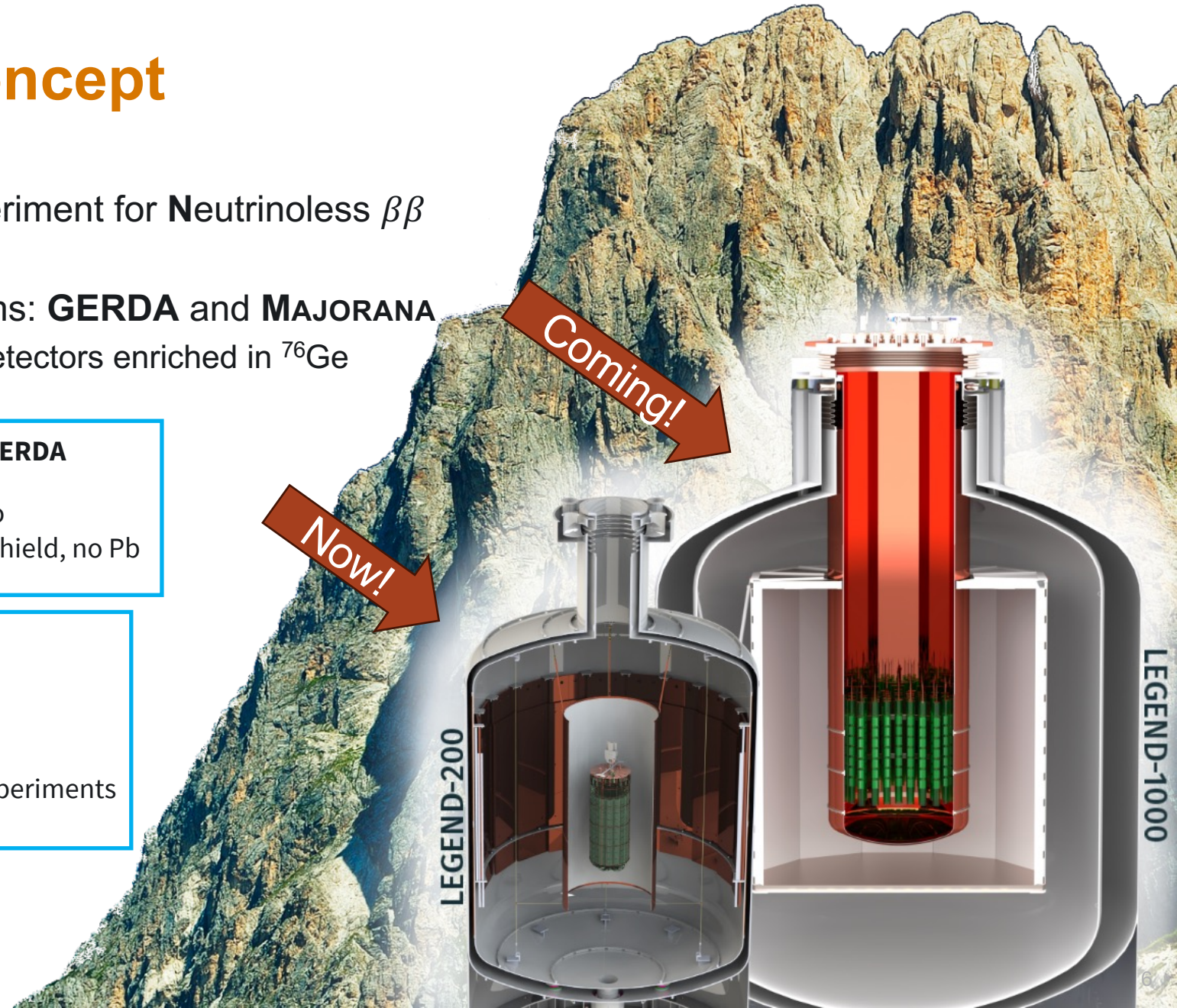
- Radiopurity of nearby parts
- Low noise electronics improves PSD
- Low energy threshold

GERDA

- LAr veto
- Low-A shield, no Pb

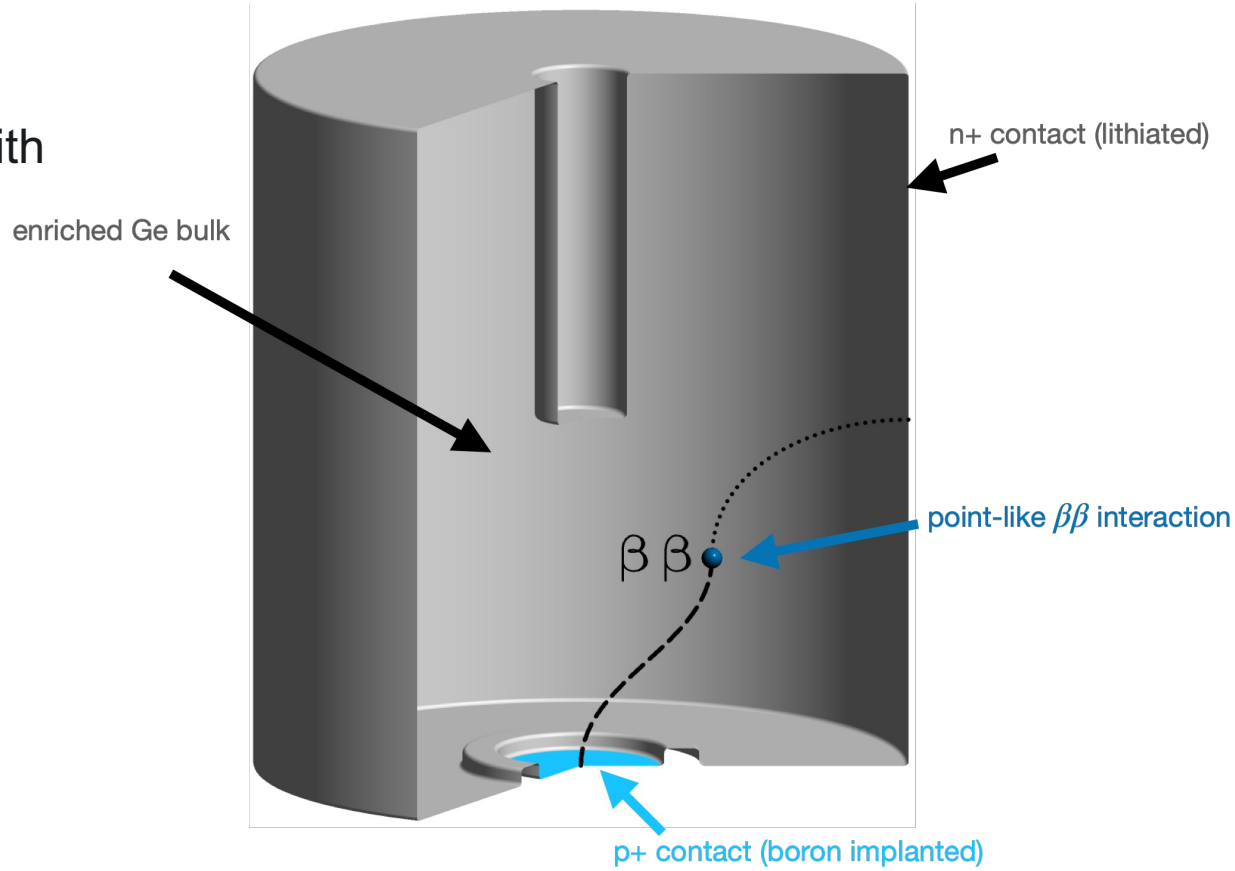
Both

- Clean fabrication techniques
- Control of surface exposure
- Development of large point-contact detectors
- Lowest background and best resolution $0\nu\beta\beta$ experiments



High-Purity Germanium Detectors

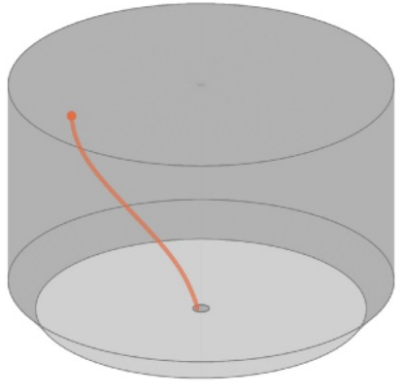
- Long used for radiation detectors
  - State-of-the-art HPGe detectors operate with various geometries
- Used in neutrinoless double beta decay searches and dark matter searches
- Point-contact design
  - Apply a high voltage, charge is collected in the p+ contact
- kg-scale detectors have O(100-1000 eV) baseline energy resolution

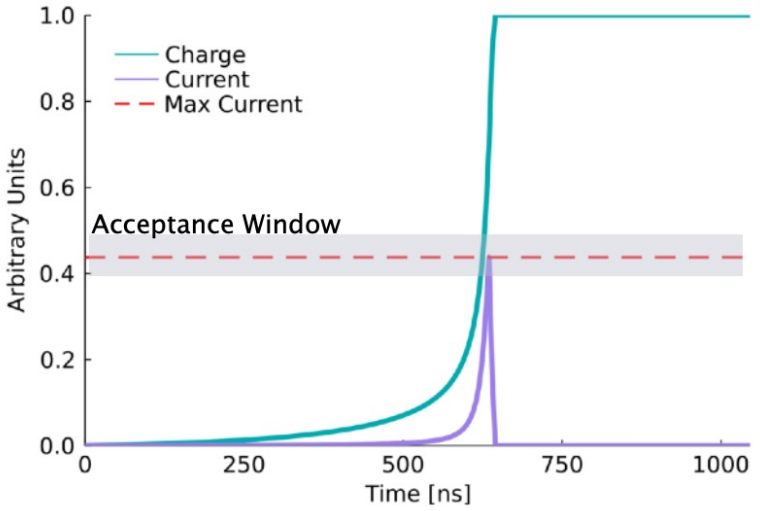


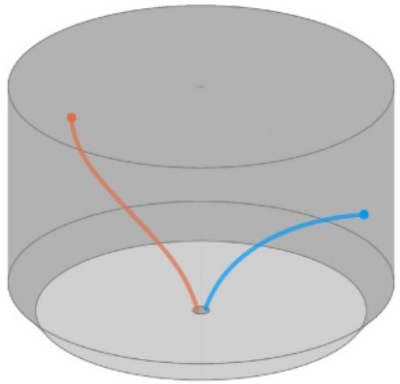
Example detector geometry from LEGEND experiment

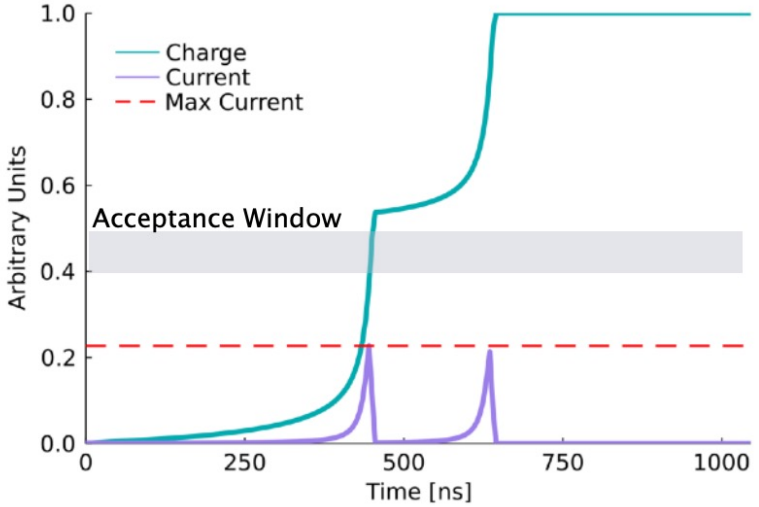
LEGEND Waveforms

- The charge signal at the p-type contact is digitized and recorded
- The time profile of the signal is used to distinguish between different types of events





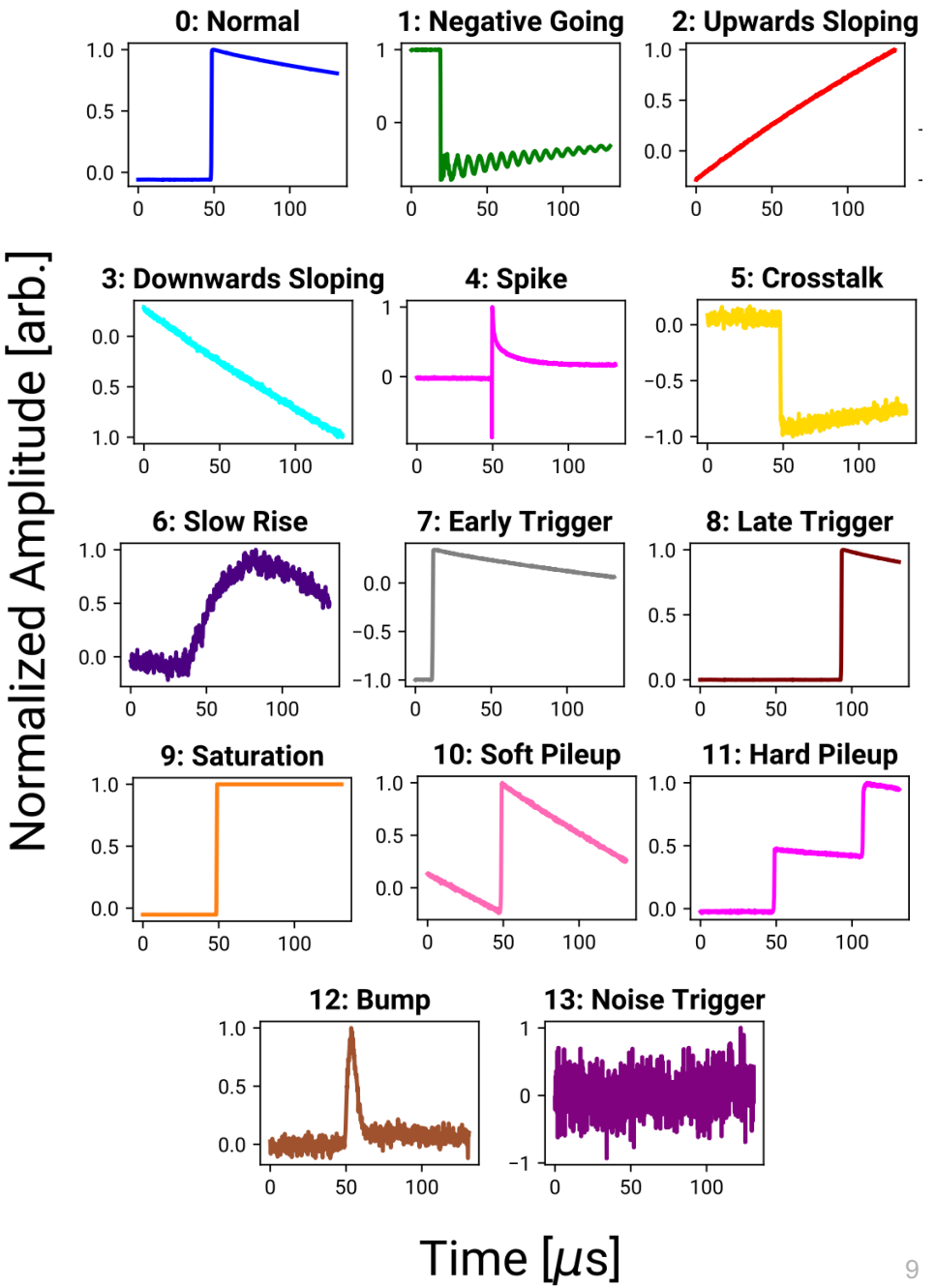




Traditional Data Cleaning

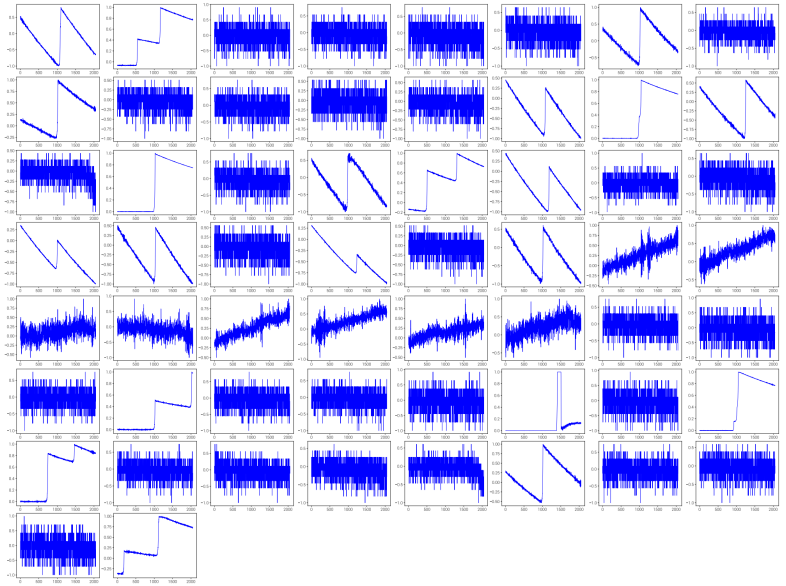
- Goal: Remove waveforms that are obviously bad and not worth keeping in physics datasets
- Common workflow:
  - Analyzers apply offline algorithms and tune various cuts to remove these waveforms
  - ~80 person-hours/week of work
- AI/ML is excellent at pattern recognition
  - Initial offline semi-supervised ML algorithm development has achieved similar/better results
  - Let's build off of this

E León *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 015064

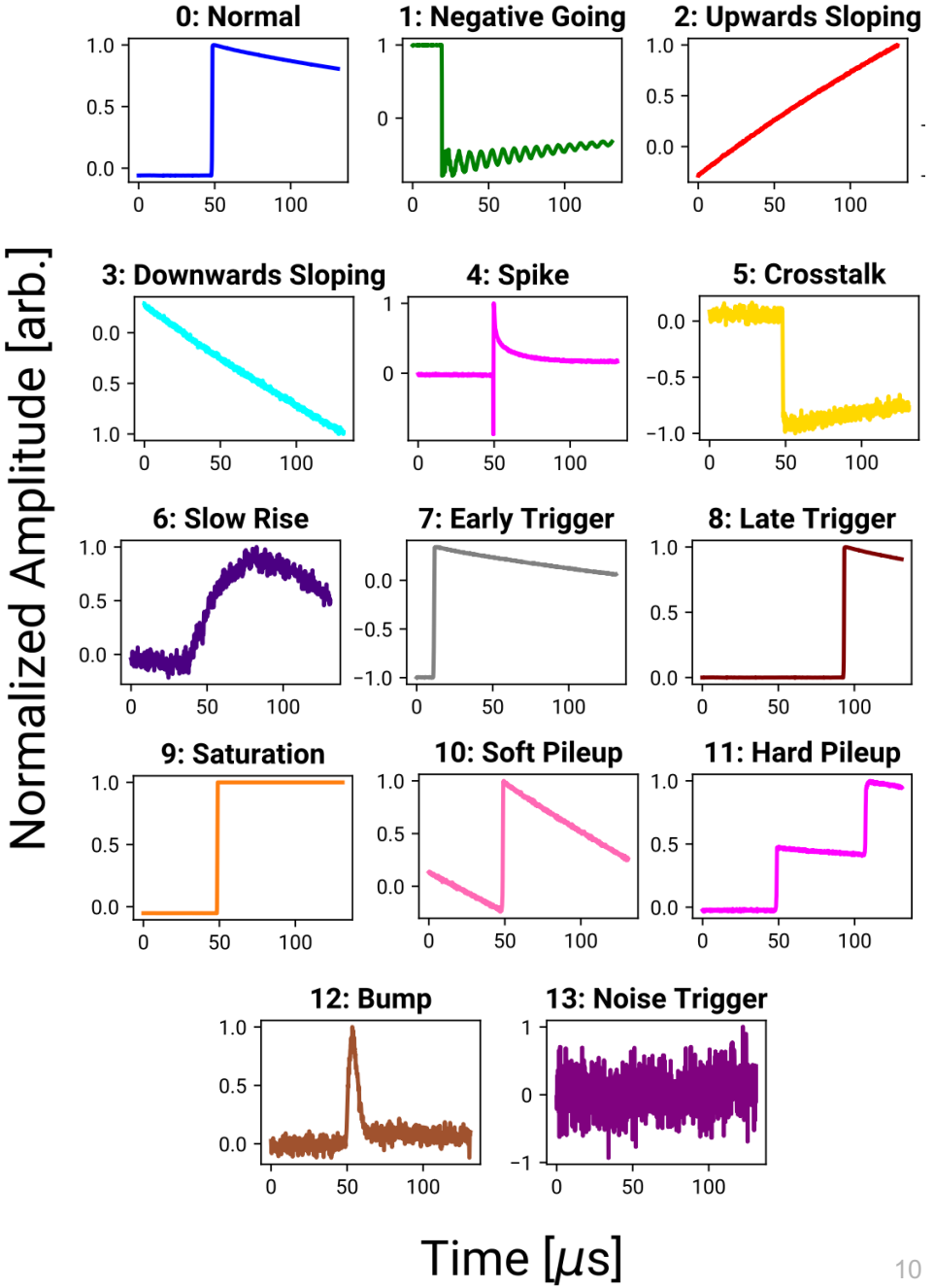


The AP-SVM Algorithm

- Affinity Propagation (AP) to define exemplar events of various classes
  - Unsupervised algorithm for classification
- Hand label each class and then train a nonlinear Support Vector Machine (SVM) on those classes
- Number of features is reduced through application of a discrete wavelet transform when training

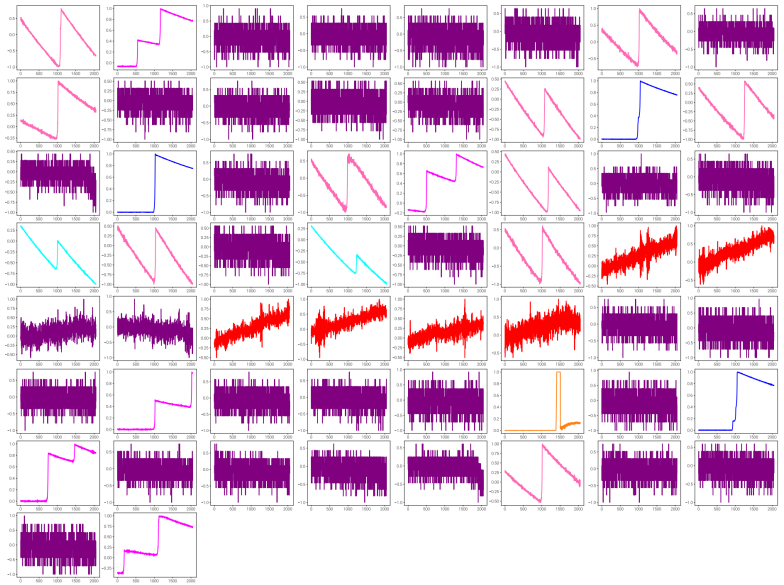


arXiv:2410.14701

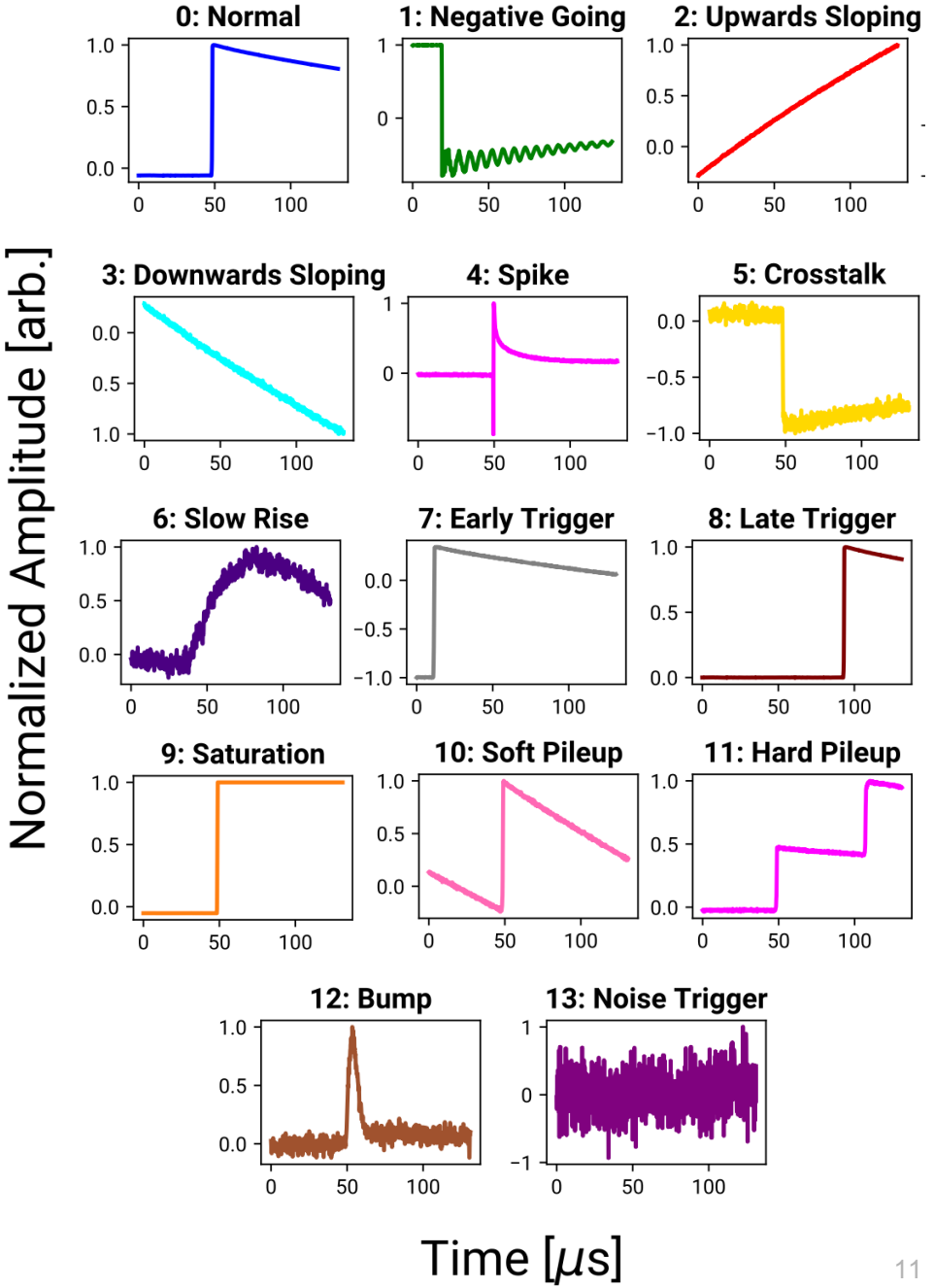


The AP-SVM Algorithm

- Affinity Propagation (AP) to define exemplar events of various classes
  - Unsupervised algorithm for classification
- Hand label each class and then train a nonlinear Support Vector Machine (SVM) on those classes
- Number of features is reduced through application of a discrete wavelet transform when training



arXiv:2410.14701



The LEGEND DAQ System

- Based on the Cherenkov Telescope Array DAQ FlashCam system
  - 16-bit mode optimized for 62.5 MHz readout of LEGEND detectors
- Built upon the Xilinx Spartan 6 FPGA chip
  - Baseline chip for the LEGEND-1000 experiment
- Simple threshold trigger on trapezoidal filtered waveform
  - Current thresholds at ~50 keV
- Two triggering schemes:
  - Global, all detectors are saved (physics data)
  - Local, only above threshold detectors are saved (calibration data)

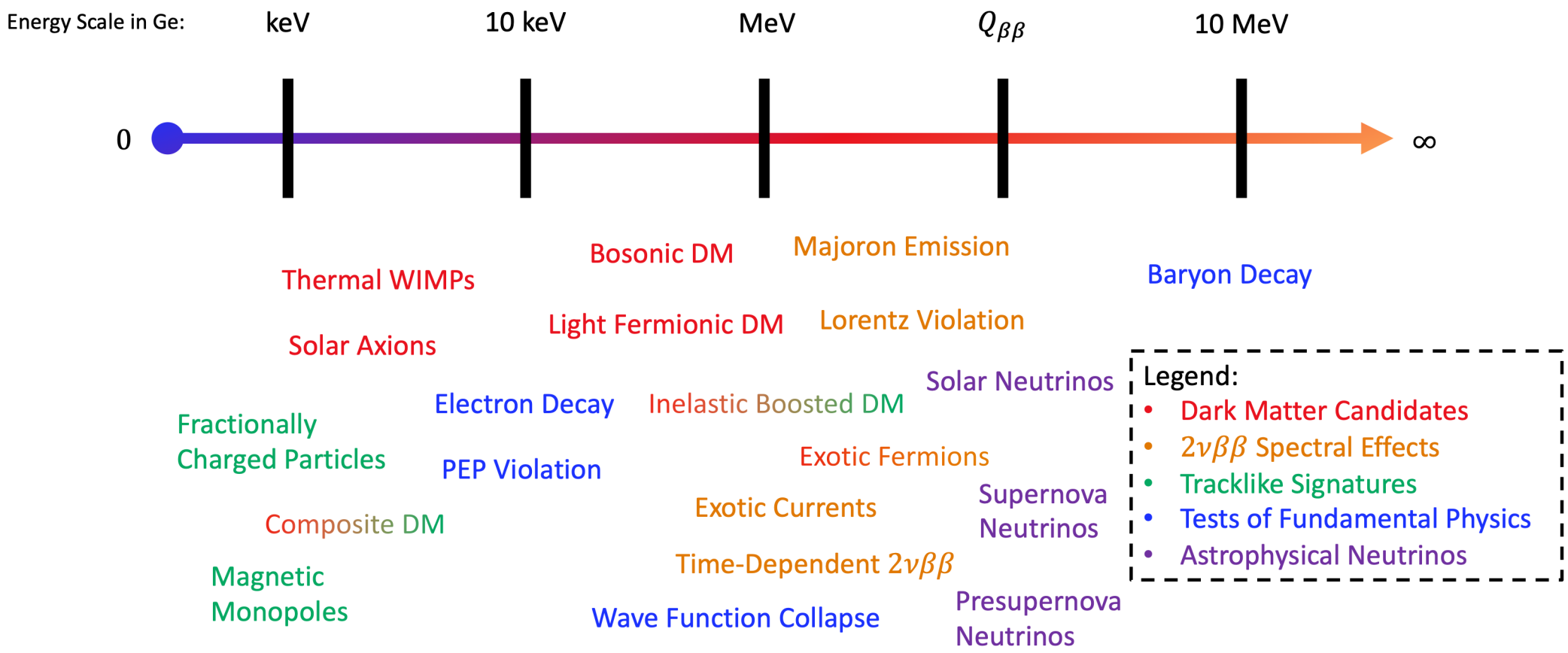
S. Sailer Ph.D. thesis



(A) FlashCam mother board. (B) FlashCam trigger board. (C) ADC piggy board with 12 readout channels. Two per mother board.

Figure 2.5: Images of a FlashCam mother board (A) equipped with a trigger card (B) and a separate ADC piggy board (C). The modular design make use of the same underlying motherboard using a Xilinx Spartan-6 FPGA.

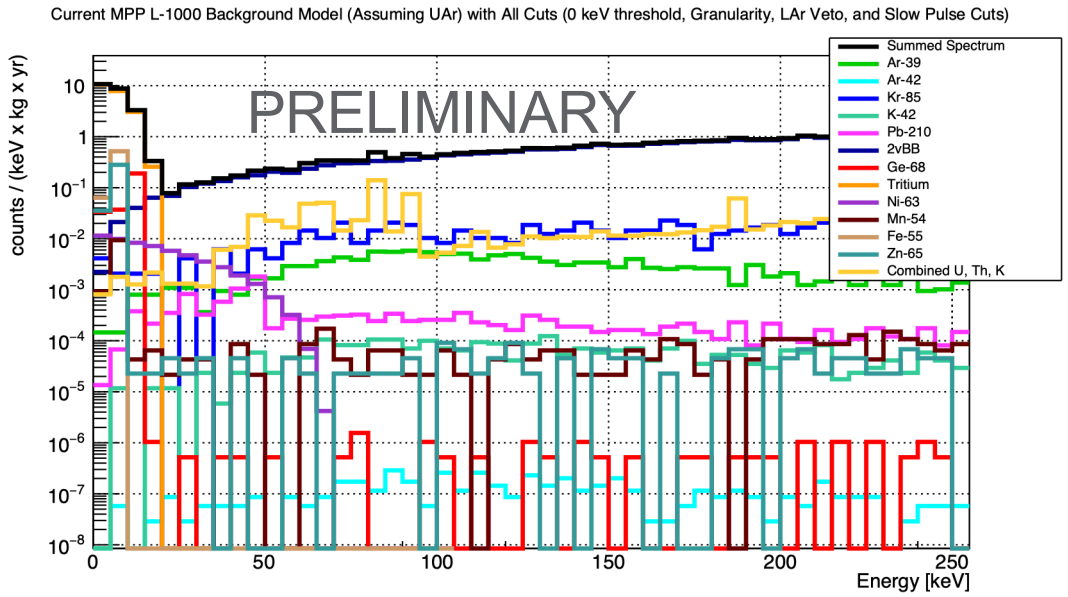
Why Lower Thresholds?
BSM Physics Opportunities beyond $0\nu\beta\beta$



What's limiting threshold?

- Low energy electronics noise
  - HPGe micro-discharges, crosstalk, electronics glitches, etc.
- Low energy backgrounds
  - Tritium, Ar-39, etc.
- Data throughput:
  - Event rate becomes high enough that significant dead time will be accrued
- Basic question: how can we extract the most physics from our flagship experiments?
  - LEGEND: maximize dynamic range, best energy resolution, minimize dead time

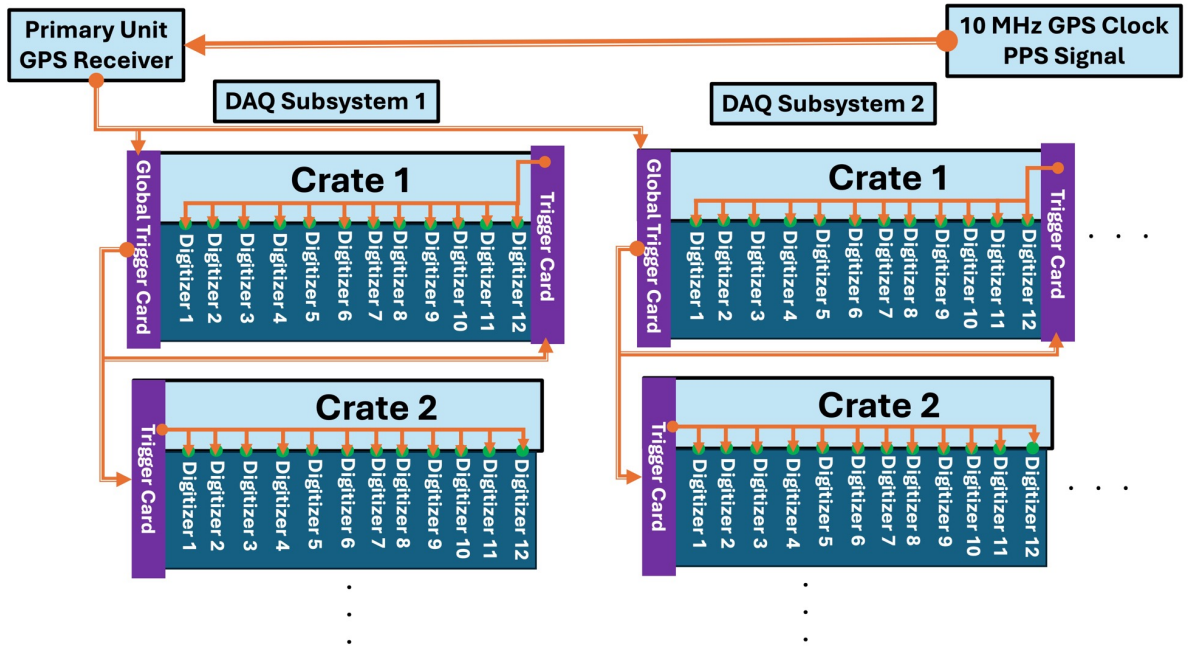
L-1000 low-energy background model



J. Waters, M.Sc. Thesis

Towards online, intelligent triggering

- Move the data cleaning onto the FPGA
- Simplify the SVM for online processing:
  - Switch to linear SVM
  - Act directly on the time domain waveform
- Basic idea:
  - At low-energies, throw out non-physics data in real time
  - At high-energies, retain data-cleaning tags for more efficient cleaning
- Allows for lower thresholds with same data throughput



LEAP-FWD: Low-Latency Embedded AI for Physics — From Waveforms to Discovery

Phase I | DOE Genesis Mission (DE-FOA-0003612) | 2026 – 2027

CHALLENGE

LEGEND-200 DAQ energy threshold currently at ~ 50 keV

Limited by high rate of cross-talk and noise events

Offline selection of physical events (“data cleaning”) is labor intensive

GOALS

Lower energy threshold of the DAQ

Reduce human workload requirements for data cleaning

Enable BSM physics searches and spatiotemporal anomaly detection

APPROACH

Develop AI algorithms deployable on FPGAs for real-time low-latency event classification

Build on previous AP-SVM work to automate data cleaning, adapting methods for low-latency deployability on FPGAs

DECISION-GATE OBJECTIVES (Month 6 Go/No-Go)

Four quantitative metrics define AI advantage

Energy Threshold ≤ 10 keV

(5× reduction, same latency, similar data storage)

Signal Efficiency ≥ 97%

(Retain physics events while rejecting all spurious noise)

Data Cleaning ≤ 8 hrs/wk

(~10× fewer person-hours for data curation)

Anomaly Recall ≥ 75%

(Detect AI-injected synthetic anomalies)

PNNL: Richard Saldanha, Sam Watkins, Antonino Tumeo
UCSD: Aobo Li
UNC: Reyco Henning, Julieta Gruszko, Himal Acharya

Pacific Northwest NATIONAL LABORATORY

UC San Diego

The University of North Carolina at Chapel Hill

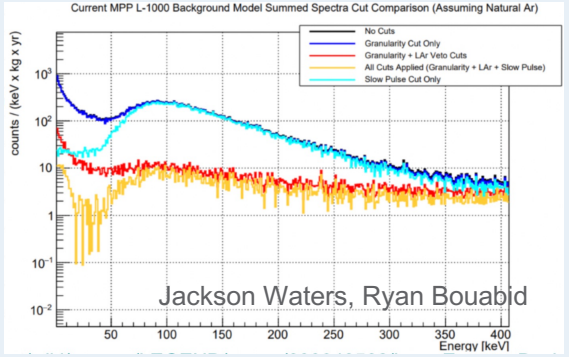
Approach & Expected Outcomes

Four integrated tasks over 9 months

T1 Data & Baseline

Lead: UNC (Henning)

Acquire low-threshold LEGEND-200 dataset.
Apply existing data pipeline to establish baseline metrics
(threshold, efficiency, person-hours)
for direct comparison.

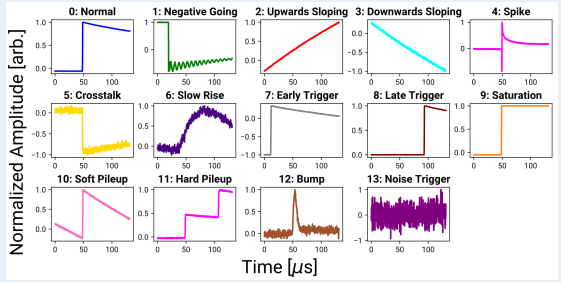


T2 AP-SVM++ Development

Lead: UCSD (Li)

Adapt León et al. 2025 AP-SVM for FPGA deployment.

Evaluate signal sacrifice and contamination on T1 validation data.



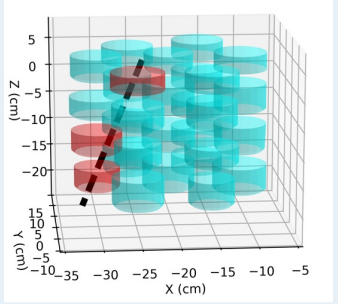
León et al. 2025, Mach. Learn.: Sci. Technol. 6, 015064

T3 Spatiotemporal Anomaly Detection

Lead: PNNL (Watkins)

Detect correlated events across detector array
(lightly ionizing particles, supernova neutrinos, BSM)

Evaluate recall on AI-injected synthetic anomalies
first at label, then waveform level.

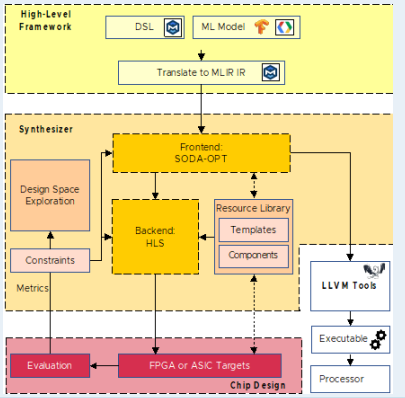


T4 FPGA Synthesis & Emulation

Lead: PNNL (Tumeo)

Translate AP-SVM++ to hardware using PNNL's SODA synthesizer.

Evaluate latency, throughput, quantization trade-offs.



MILESTONE TIMELINE (9 months)

Month 1	Month 3	Month 4	Month 5	Month 5	Month 6	Month 8	Month 9
Low-threshold dataset	Baseline pipeline done	AP-SVM++ developed	Pipeline comparison	Anomaly alg. down-select	FPGA eval + GO/NO-GO	Online AP-SVM++	Waveform→ anomaly demo

OUTCOMES & BROADER IMPACT

- New physics reach: dark matter, BSM, supernovae
- Open-source AP-SVM++ & anomaly algorithms on American Science Cloud

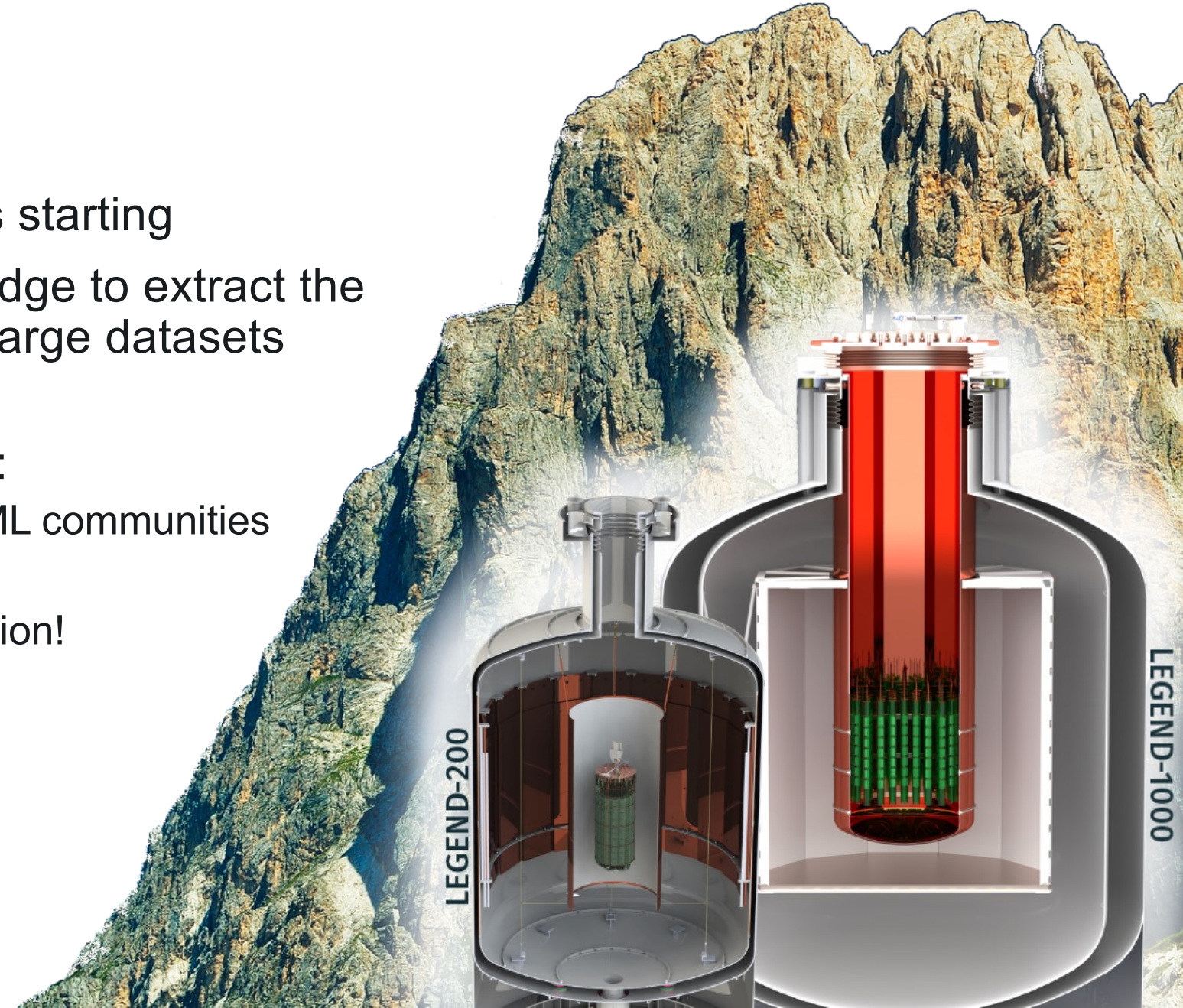
PHASE II

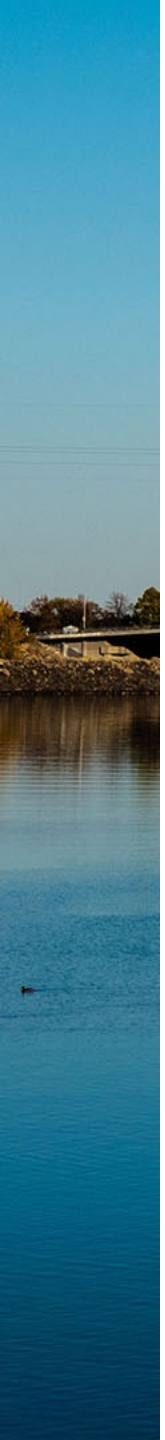
- LEGEND-200 DAQ integration with AMD FPGAs/SoCs
- Transferable pipeline for next-gen experiments (XLZD, EIC, other HENP)

Pacific Northwest
NATIONAL LABORATORY

Summary

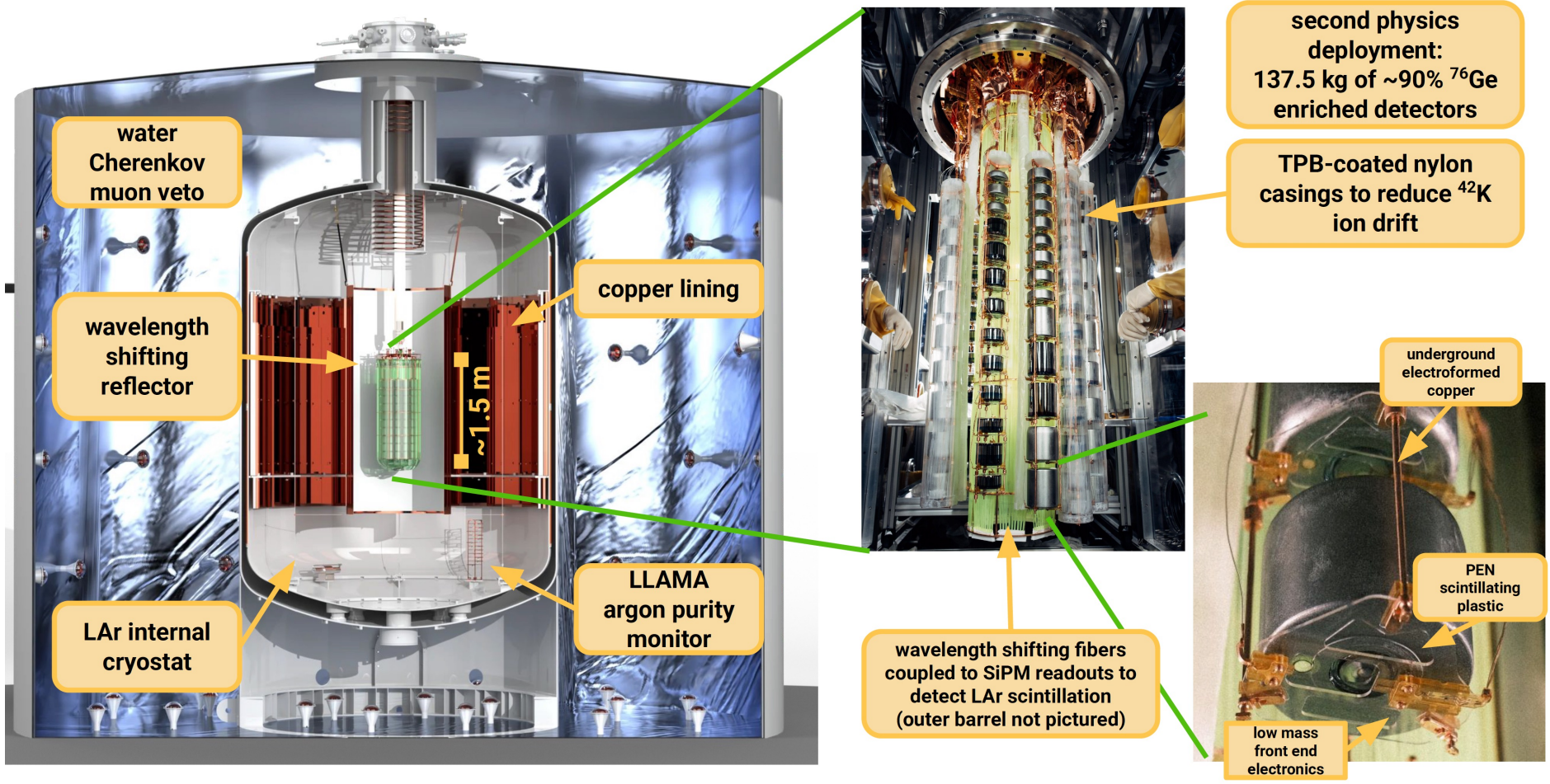
- The LEAP-FWD project is starting
- Moving AI to the sensor edge to extract the most physics from these large datasets
  - Applications are general
- Looking towards Phase II:
  - Lots of overlap with FastML communities
  - EIC, LHC, etc.
  - We are open to collaboration!





Pacific Northwest
NATIONAL LABORATORY

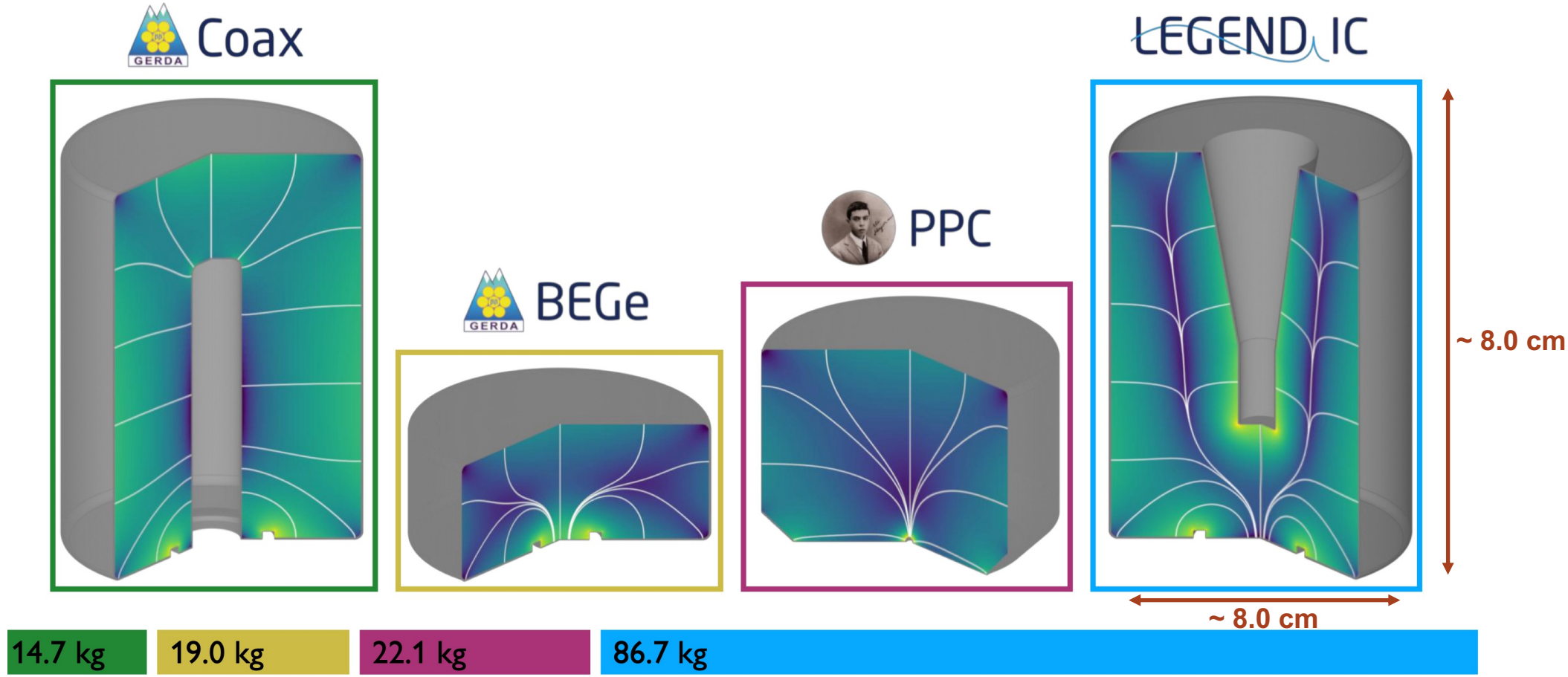
LEGEND-200 Experiment (ongoing)



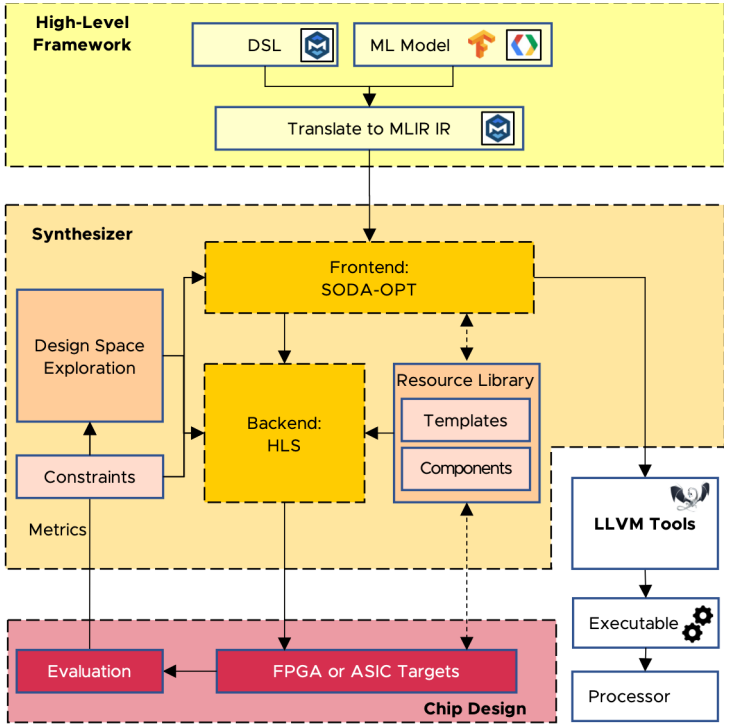
Latest physics result: [Phys. Rev. Lett. 136, 022701 (2026)]

LEGEND Detector Geometries

High-purity Germanium (HPGe) detectors, enriched to 86-92% in ^{76}Ge
p-type detectors with lithium-diffused n+ electrodes around the surface



SODA Synthesizer: Overview



- A modular, multi-level, interoperable, extensible, **open-source hardware compiler** from **high-level programming frameworks to silicon**
- Compiler-based frontend, leveraging the MultiLevel Intermediate Representation (MLIR)
- **Compiler-based backend**, leveraging state-of-the-art High-Level Synthesis (HLS) techniques
- Generates **synthesizable Verilog** for a variety of targets, from Field Programmable Gate Arrays (FPGAs) to Application Specific Integrated Circuits (ASICs)
- Optimizations at all levels are performed as **compiler optimization** passes

[M. Minutoli, V. G. Castellana, C. Tan, J. Manzano, V. Amatya, A. Tumeo, D. Brooks, G-Y. Wei: SODA: a New Synthesis Infrastructure for Agile Hardware Design of Machine Learning Accelerators. ICCAD 2020: 98:1-98:7]

[J. Zhang, N. Bohm Agostini, S. Song, C. Tan, A. Limaye, V. Amatya, J. Manzano, M. Minutoli, V. G. Castellana, A. Tumeo, G-Y. Wei, D. Brooks: Towards Automatic and Agile AI/ML Accelerator Design with End-to-End Synthesis. ASAP 2021: 218-225]

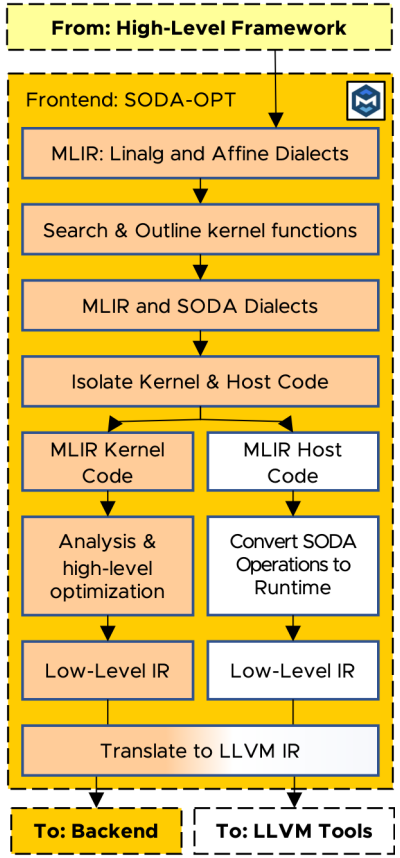
[N. Bohm Agostini, S. Curzel, J. Zhang, A. Limaye, C. Tan, V. Amatya, M. Minutoli, V.G. Castellana, J. Manzano, A. Tumeo: Bridging Python to Silicon: The SODA Toolchain. IEEE Micro Magazine 2022]
[N. Bohm Agostini, S. Curzel, V. Amatya, C. Tan, M. Minutoli, V. G. Castellana, J. Manzano, D. Kaeli, A. Tumeo : An MLIR-based Compiler Flow for System-Level Design and Hardware Acceleration. ICCAD 22]
[Fabrizio Ferrandi, Vito Giovanni Castellana, Serena Curzel, Pietro Fezzardi, Michele Fiorito, Marco Lattuada, Marco Minutoli, Christian Pilato, Antonino Tumeo: Invited: Bambu: an Open-Source Research Framework for the High-Level Synthesis of Complex Applications. DAC 2021: 1327-1330]

SODA-OPT: Frontend and High-Level IR

- **SODA-OPT: S**earch, **O**utline, **D**ispatch, **A**ccelerate frontend **opt**imizer “generates” the SODA High-Level IR
- Employs and embraces the MLIR framework
  - MLIR: Multi-Level Intermediate Representation
  - Used in TensorFlow, TFRT, ONNX-MLIR, NPComp, others
  - Several architecture independent dialects (Linalg, Affine, SCF) and optimizations
- Interfaces with high-level ML frameworks through MLIR “bridges” (e.g., libraries, rewriters)
- Defines the SODA MLIR dialect and related compiler passes to:
  - Identify dataflow segments for hardware generation
  - Perform high-level optimizations (dataflow transformations, data-level and instruction-level parallelism extraction)
  - Generate interfacing code and runtime calls for microcontroller

[N. Bohm Agostini, S. Curzel, V. Amatya, C. Tan, M. Minutoli, V. G. Castellana, J. Manzano, D. Kaeli, A. Tumeo : An MLIR-based Compiler Flow for System-Level Design and Hardware Acceleration. ICCAD 22]

[N. Bohm Agostini, S. Curzel, D. Kaeli, A. Tumeo: SODA-OPT an MLIR based flow for co-design and high-level synthesis. CF 2022: 201-202 - **Best Poster Award**.]

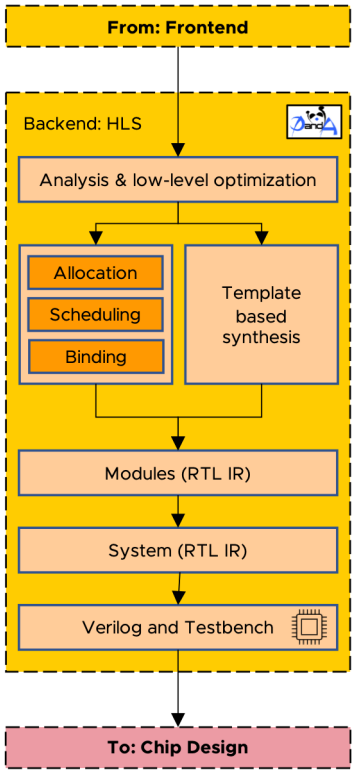


SODA-OPT: System Overview

https://github.com/pnnl/soda-opt

SODA Synthesizer: HLS Backend

- The synthesizer backend take as input the properly optimized low-level IR and generate the hardware descriptions of the accelerators
- The HLS backend is PandA-Bambu, an open-source state-state-of-the-art High-Level Synthesis (HLS) framework
  - Key features: **parallel accelerator designs**, **modular HLS**, and **ASIC support**
- The HLS backend provides automated testing and verification of the generated designs



Bambu HLS

https://panda.dei.polimi.it

[Fabrizio Ferrandi, Vito Giovanni Castellana, Serena Curzel, Pietro Fezzardi, Michele Fiorito, Marco Lattuada, Marco Minutoli, Christian Pilato, Antonino Tumeo: Invited: Bambu: an Open-Source Research Framework for the High-Level Synthesis of Complex Applications. DAC 2021: 1327-1330]