Global HENP event reconstruction with Foundation Models using real experimental data

Principal Investigator(s): Joe Osborn (PI, PO) and David Park (Co-PI,CDS)

List of the proposal participants and their organizations if other than NPP:

Yeonju Go (PO), Yi Huang (CDS)

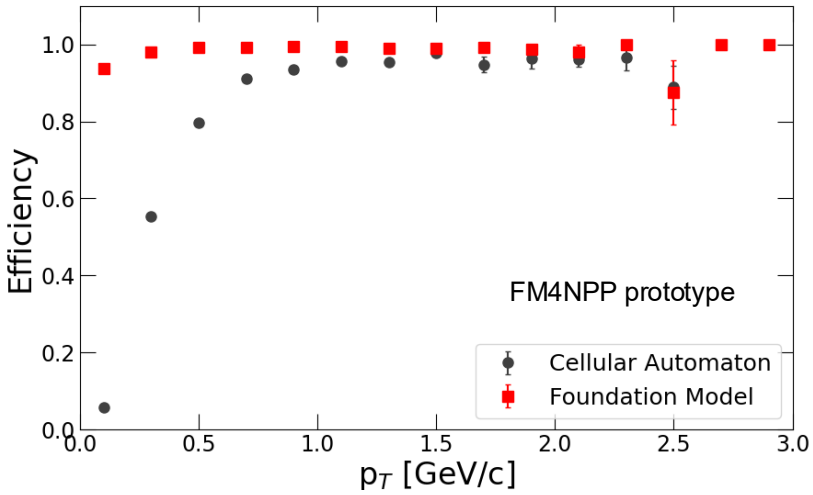
ECA eligibility: Yes

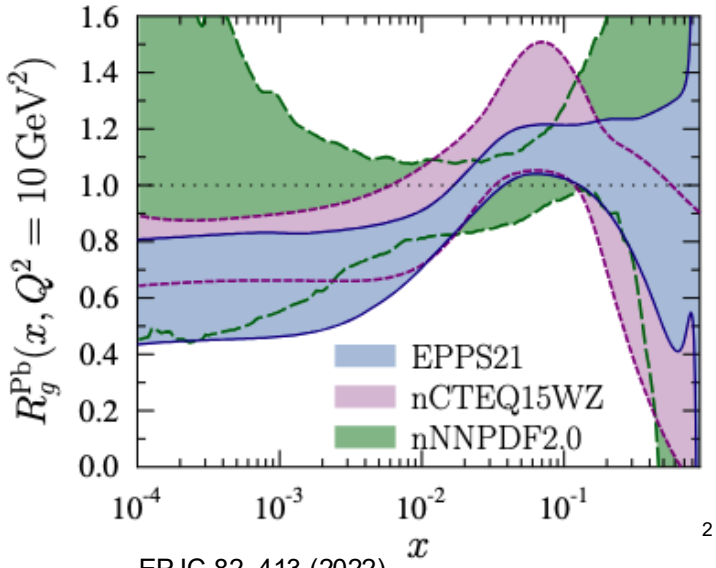
Proposal term from: 10/01/2026 to: 09/30/2028

Annual funding: FY27 $250k FY28 $250k

Foundation Models (FM) for HENP

- Prototype FMs have been developed for single detectors (e.g. sPHENIX TPC, ePIC DIRC)
- LDRD goal
  - Train a FM on entire sPHENIX detector and demonstrate global event reconstruction in simulation
  - Apply to real data
- Benefits if successful
  - Advanced analysis techniques may uncover hidden patterns traditional algorithms are blind to
  - Improved reconstruction performance
  - Develop complex tokenization methods to advance understanding of AI applications for HENP data
  - New physics opportunities at RHIC and EIC e.g. low p_T heavy flavor production





EPJC 82, 413 (2022)

Brookhaven National Laboratory

Milestones and Personnel

- Milestones
  1. FY27Q1 - Prep realistic pp simulation for model development
  2. FY27Q4 - Develop downstream adapters to include tokenized silicon and calorimeter data
  3. FY28Q1 - Fine tune adapter models
  4. FY28Q2 - Interface adapters to a production environment
  5. FY28Q3 - Test on real data
- Personnel
  - Joe and Yeonju bring necessary sPHENIX physics, reconstruction, and data/simulation expertise
  - David and Yi bring necessary AI model development and data preparation expertise
- Same team responsible for 2508.14087, recently accepted by ICLR, a top AI conference

Intellectual Merit Summary

- This project intersects the forefront of AI and HENP data processing research
  - Can new tools in AI be utilized for scientific advancement in HENP?
  - Can these tools find correlations traditional algorithms cannot?
  - Can models trained on simulation be applied to real data?
- Potential access to new physics measurements through improved reconstruction
- Determine complex tokenization methods that will enable exploration of AI applications to HENP data

Return on Investment or Potential Future Funding

- Would maintain BNL at forefront of AI for HENP research
  - e.g. LLNL has reached out to collaborate, other labs collaborating on AmSC led by BNL, etc.
- Would position our team to be ready for NOFOs from DOE related to model development and/or AI for science
  - e.g. through Genesis project
- If this project were successful, we would seek funding to expand to multi-detector and multi-facility studies
  - e.g. how does a model trained on sPHENIX data perform for ePIC at EIC?

The Broader Impact on the Laboratory

- Impactful for collider based data reconstruction in both NP and HEP
- Cross division collaboration at BNL, supporting applied AI research
- Supports laboratory mission to understand building blocks of the universe **and** developing next generation information science
- Project has ample directions to grow and be impactful for both NP and HEP

Names of Suggested BNL Reviewers

Yuewei Lin

Michael Begel

Viviana Cavaliere

Shinjae Yoo

If ECA Eligible – How this LDRD Benefits Your Application

- This LDRD would enable a product to be developed and tested on real experimental data
  - Current demonstrations e.g. with TPC or DIRC are prototypes
    - Single subsystem, with a few tasks – proof of principle
- Provides a springboard for further development that an ECA could be based upon
  - Develop an AI product that could demonstrate holistic reconstruction using multi-modal data
  - Determine how to practically deploy in an active experiment
  - Expand physics benefit and gain
    - e.g. low p_T heavy flavor measurements

LDRD Funding Table from BOM

Title: "Global HENP event reconstruction with foundation models"
PI: Joseph Osborn
David Park

Resource Category	DESCRIPTION		FY27		FY28
050	Salary - Scientific		53,254		49,656
051	Salary - Research Assoc		15,162		15,768
050	Salary - Professional		65,078		67,681
050	Salary -Technical		0		0
050	Salary - Management & Admin.		0		0
	Total FTEs		**0.70**		**0.67**
TOTAL SALARY/WAGE & FRINGE			**133,494**		**133,105**
various	Contracts - Low Value		0		0
280	Foreign Travel		0		0
290	Domestic Travel		8,000		8,000
various	Purchases		0		0
TOTAL MSTC			**8,000**		**8,000**
TOTAL DIRECT COSTS			**141,494**		**141,105**
251	Electric Distributed (Electric Power Burden)		1,335		1,331
700/701/481	Organizational Burden		26,819		27,210
TOTAL ORGANIZATIONAL BURDEN			**28,154**		**28,541**
745	Procurement (Material Handling)		560		560
735	G&A Burden		0		0
730	Common Institutional Support		79,792		79,794
722	Safeguards & Security Assess		0		0
TOTAL LABORATORY BURDEN			**80,352**		**80,354**
705	LDRD Burden		0		0
TOTAL PROGRAM COSTS			**250,000**		**250,000**
740	Full Cost Recovery		0		0
TOTAL PROGRAM COSTS			**250,000**		**250,000**

Labor Band	Name	FY27 FTE	FY27 Amount	FY28 FTE	FY28 Amount
SCI1	J. Osborn	0.25	53,254	0.22	49,656
PROF2	D. Park	0.25	46,484	0.25	48,344
PROF2	Y. Huang	0.10	18,594	0.10	19,337
RA4	Y. Go	0.10	15,162	0.10	15,768
	Total	0.70	133,494	0.67	133,105

Backup

Foundation Model Challenges for Experimental PP Collision

Challenge 1: Tokenization. Mixed trajectories from different particles: the prior physics-informed binning and tokenization strategy would lose its merit.

- However, we have preliminary results on pure space-filling curve based (e.g., Hilbert, Z-order) tokenization.
- This allows to extend the existing FM4NPP on the new, more complex data.

Challenge 2: Extensive number of tokens. A naïve extension would quickly face GPU memory issues

- Group-level tokenization would resolve this: essentially taking every n points as a group.